GeoPhy: Differentiable Phylogenetic Inference via Geometric Gradients of Tree Topologies

Takahiro Mimori [1 2] **Michiaki Hamada** [1 3]

Abstract

Phylogenetic inference, grounded in molecular evolution models, is essential for understanding evolutionary relationships in biological data. While variational Bayesian methods offer scalable models for biological analysis, reliable inference for latent tree topology and branch lengths remains challenging due to the vast possibilities for topological candidates. In response, we introduce GeoPhy, a novel approach that employs a fully differentiable formulation of phylogenetic inference, representing topological distributions in continuous geometric spaces without limiting topological candidates. In experiments using real benchmark datasets, GeoPhy significantly outperformed other approximate Bayesian methods that considered whole topologies. [1]

1. Introduction

Phylogenetic inference, the reconstruction of tree-structured evolutionary relationships between biological units, such as genes, cells, individuals, and species ranging from viruses to macro-organisms, is a fundamental problem in biology. As the phylogenetic relationships are often indirectly inferred from molecular observations, including DNA, RNA and protein sequences, Bayesian inference has been an essential tool to quantify the uncertainty of phylogeny. However, due to the complex nature of the phylogenetic tree object, which involve both a discrete topology and dependent continuous variables for branch lengths, the default approach for the phylogenetic inference has typically been an Markov-chain Monte Carlo (MCMC) method (Ronquist et al., 2012), enhanced with domain-specific techniques such as a mixed strategy for efficient exploration of tree topologies.

As an alternative approach to the conventional MCMCs, Zhang & Matsen IV (2019) proposed a variational Bayesian approach termed VBPI, which has subsequently been improved in the expressive powers of topology-dependent branch length distributions (Zhang, 2020; 2023). Although these methods have presented accurate joint posterior distributions of topology and branch lengths for real datasets, they required reasonable preselection of candidate tree topologies to avoid a combinatorial explosion in the number of weight parameters beforehand. There have also been proposed variational approaches (Moretti et al., 2021; Koptagel et al., 2022) on top of the combinatorial sequential Monte-Carlo method (CSMC; Wang et al. (2015)), where topologies and their weights were iteratively updated without the need for the preselection steps. However, the fidelity of the joint posterior distributions was still largely behind MCMC and VBPI as reported in Koptagel et al. (2022).

In this work, we propose a simple yet effective scheme for parameterizing a binary tree topological distribution with a transformation of continuous distributions. We further formulate a novel differentiable variational Bayesian approach named GeoPhy to optimize a variational distribution of the tree topology and branch lengths to approach the posterior distribution, without preselection of candidate topologies. In our experiments using real biological datasets, we demonstrate that GeoPhy significantly outperforms other approaches, all without topological restrictions, in terms of the fidelity of the marginal log-likelihood (MLL) estimates to gold-standard provided with long-run MCMCs.

2. Background

2.1. Phylogenetic models

Let τ represent an unrooted binary tree topology with N leaf nodes (tips), and let B_τ denote a set of evolutionary distances defined on each of the branches of τ. A phylogenetic tree (τ, B_τ) represents an evolutionary relationship between N species, which is inferred from molecular data, such as DNA, RNA or protein sequences obtained for the species. Let $Y = \{Y_{ij} \in \Omega\}_{1\le i\le N, 1\le j\le M}$ be a set of aligned sequences with length M from the species, where Y_{ij} denote a character (base) of the i-th sequence at j-th

[1] Waseda University, Tokyo, Japan [2] RIKEN AIP, Tokyo, Japan [3] AIST, Tokyo, Japan. Correspondence to: Takahiro Mimori <takahiro.mimori@aoni.waseda.jp>, Michiaki Hamada <mhamada@waseda.jp>.

Published at the Differentiable Almost Everything Workshop of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. July 2023. Copyright 2023 by the author(s).

[1] *This study's complete version is currently under review.*

site, and is contained in a set of possible bases Ω. For DNA sequences, Ω represents a set of 4-bit vectors, where each bit represents 'A', 'T', 'G', or 'C'. A likelihood model of the sequences $P(Y|\tau, B_\tau)$ is determined based on evolutionary assumptions. In this study, we follow a common practice for method evaluations (Zhang & Matsen IV, 2019; Zhang, 2023) as follows: Y is assumed to be generated from a Markov process along the branches of τ in a site-independent manner; The base mutations are assumed to follow the Jukes-Cantor model (Jukes et al., 1969). The log-likelihood $\ln P(Y|\tau, B_\tau)$ can be calculated using Felsenstein's pruning algorithm (Felsenstein, 1973), which is also known as the sum-product algorithm, and differentiable with respect to B_τ.

2.2. Variational Bayesian phylogenetic inference

The variational inference problem for phylogenetic trees, which seeks to approximate the posterior probability $P(\tau, B_\tau|Y)$, is formulated as follows:

$$\min_Q D_{\mathrm{KL}}\left(Q(\tau)Q(B_\tau|\tau)\|P(\tau, B_\tau|Y)\right), \tag{1}$$

where D_{KL}, $Q(\tau)$ and $Q(B_\tau|\tau)$ denote the Kullback-Leibler divergence, a variational tree topology distribution, and a variational branch length distribution, respectively. The first variational Bayesian phylogenetic inference method (VBPI) was proposed by Zhang & Matsen IV (2019), which has been successively improved for the expressiveness of $Q(B_\tau|\tau)$ (Zhang, 2020; 2023). For the expression of variational topology mass function $Q(\tau)$, they all rely on a subsplit Bayesian network (SBN) (Zhang & Matsen IV, 2018), which represents tree topology mass function $Q(\tau)$ as a product of conditional probabilities of splits (i.e., bipartition of the tip nodes) given their parent splits. However, SBN necessitates a preselection of the likely set of tree topologies, which hinders an end-to-end optimization strategy of the distribution over all the topologies and branch lengths.

3. Proposed methods

3.1. Geometric representations of tree topology ensembles

Considering the typically infeasible task of parameterizing the probability mass function of unrooted tree topologies $\mathcal{T}$, which requires $(2N-5)!!-1$ degrees of freedom, we propose an alternative approach. We suggest constructing the mass function $Q(\tau)$ through a transformation of a certain probability density $Q(z)$ over a continuous domain $\mathcal{Z}$, as follows:

$$Q(\tau) := \mathbb{E}_{Q(z)}[\mathbb{I}[\tau = \tau(z)]], \tag{2}$$

where $\tau : \mathcal{Z} \to \mathcal{T}$ denotes a deterministic link function that maps N coordinates to the corresponding tree topology (Fig. 1). Note that we have overloaded τ to represent both a variable and function for notational simplicity. An example of the representation space $\mathcal{Z}$ is a product of Euclidean spaces $\mathbb{R}^{N\times d}$ or hyperbolic spaces $\mathbb{H}^{N\times d}$ (Appendix A), where d denotes the dimension of each tip's representation coordinate. For the link function, we can use $\tau(z) = T_{\mathrm{NJ}} \circ D(z)$, where $D : \mathcal{Z} \to \mathbb{R}^{N\times N}$ denotes a function that takes N coordinates and provides a distance matrix between those based on a geometric measure such as the Euclidean or hyperbolic distance. $T_{\mathrm{NJ}} : \mathbb{R}^{N\times N} \to \mathcal{T}$ denotes a map that takes this distance matrix and generates an unrooted binary tree topology of their phylogeny, determined using the Neighbor-Joining (NJ) algorithm (Saitou & Nei, 1987). While the NJ algorithm offers a rooted binary tree topology accompanied by estimated branch lengths, we only use the topology information and remove the root node from it to obtain the unrooted tree topology $\tau \in \mathcal{T}$.

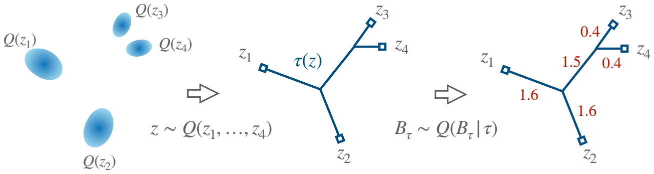

Figure 1. The proposed scheme for constructing a variational distribution $Q(\tau, B_\tau)$ of a tree topology and branch lengths by using a distribution $Q(z)$ defined on the continuous space. **Left**: The marginal distributions of four tip nodes $Q(z_1), \ldots, Q(z_4)$. **Middle**: Coordinates of the tip nodes $z = \{z_1, \ldots, z_4\}$ sampled from the distribution $Q(z)$, and a tree topology $\tau(z)$ determined from z. **Right**: A set of branch lengths B_τ (red figures) of the tree τ is sampled from a topology dependent distribution $Q(B_\tau|\tau)$.

3.2. Derivation of variational lower bound

Given a distribution of tip coordinates $Q(z)$ and an induced tree topology distribution $Q(\tau)$ according to equation (2), the variational lower bound (1) is evaluated as follows:

$$\mathcal{L}[Q] = \mathbb{E}_{Q(z)Q(B_\tau|\tau(z))}\left[\ln \frac{P(Y, B_\tau|\tau(z))P(\tau(z))}{Q(B_\tau|\tau(z))Q(\tau(z))}\right]. \tag{3}$$

Thanks to the deterministic mapping $\tau(z)$, we can obtain an unbiased estimator of $\mathcal{L}[Q]$ by sampling from $Q(z)$ without summing over the combinatorial many topologies $\mathcal{T}$. However, even when the density $Q(z)$ is computable, the evaluation of $\ln Q(\tau)$ remains still infeasible with small samples according to the definition (2). We resolve this issue by introducing the second lower bound with respect to a conditional variational distribution $R(z|\tau)$ as follows:

$$\mathcal{L}[Q,R] = \mathbb{E}_{Q(z)Q(B_\tau|\tau(z))}\left[\ln \frac{P(Y,B_\tau|\tau(z))P(\tau(z))R(z|\tau(z))}{Q(B_\tau|\tau(z))Q(z)}\right]. \tag{4}$$

Proposition 3.1. *The relation* $\ln P(Y) \geq \mathcal{L}[Q] \geq \mathcal{L}[Q,R]$ *holds, where the first and second equality holds when* $Q(\tau, B_\tau) = P(\tau, B_\tau|Y)$ *and* $R(z|\tau) = Q(z|\tau)$*, respectively.*

The proof is provided in Appendix C. Similar to Burda et al. (2016), we can also derive a tractable importance-weighted lower-bound of the model evidence (IW-ELBO), which is used for estimating the marginal log-likelihood (MLL), $\ln P(Y)$, or an alternative lower-bound objective for maximization. For the MLL estimates, we use 1000 Monte Carlo samples, similar to (Zhang & Matsen IV, 2019; Zhang, 2023). The details and derivations are described in Appendix C.

3.3. Differentiable phylogenetic inference with GeoPhy

Here, we introduce the parameterized variational distributions, $Q_\theta(z), Q_\phi(B_\tau|\tau), R_\psi(z|\tau)$, to facilitate the gradient-based maximization of the variational objective $\mathcal{L}[Q_{\theta,\phi}, R_\psi]$. We term this framework GeoPhy, since it allows us to optimize the distribution over entire phylogenetic trees within continuous geometric spaces.

Gradient estimators and variance reduction An unbiased estimator of the gradient $\nabla_\theta \mathcal{L}$ is derived as follows:

$$\widehat{g}_\theta = \nabla_\theta \ln Q_\theta(z) \cdot f(z, B_\tau) - \nabla_\theta \ln Q_\theta(h_\theta(\epsilon_z)) \tag{5}$$

where we denote $\epsilon_z^{(k)} \sim p_z\, z^{(k)} = h_\theta(\epsilon_z^{(k)})$, $\epsilon_B^{(k)} \sim p_B$, $B_\tau^{(k)} = h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))$, and

$$f(z, B_\tau) := \ln \frac{P(Y, B_\tau|\tau(z))}{Q_\phi(B_\tau|\tau(z))} + \ln P(\tau(z))R_\psi(z|\tau(z)).$$

Note that we explicitly distinguish z and $h_\theta(\epsilon_z)$ to indicate the target of differentiation with respect to θ. Unlike the gradient terms associated with ϕ and ψ, the optimization of $\mathcal{L}$ suffers from the high variance of the gradient $\widehat{g}_\theta$, which arises from the term being proportional to the score function $\nabla_\theta Q_\theta(z)$. To address this high gradient variance issue, we have explored several control variates. These are based on a leave-one-out (LOO) baseline estimation that uses multiple Monte-Carlo samples (Kool et al., 2019; Richter et al., 2020) and surrogate functions (Grathwohl et al., 2018), as detailed in Appendix C.

Variational distributions To investigate the basic effectiveness of GeoPhy algorithm, we employ simple constructions for the variational distributions $Q_\theta(z)$, $Q_\phi(B_\tau|\tau)$, and $R_\psi(z|\tau)$. We use an independent distribution for each tip node coordinate, i.e. $Q_\theta(z) = \prod_{i=1}^N Q_{\theta_i}(z_i)$, where we use a d-dimensional normal or wrapped normal distribution (Appendix A) for the coordinates of each tip node z_i. For the conditional distribution of branch lengths given tree topology, $Q_\phi(B_\tau|\tau)$, we use the diagonal lognormal distribution whose location and scale parameters are parameterized as functions of the unique features defined for each topology τ (Appendix B), as proposed in Zhang (2023). For the model of $R_\psi(z|\tau)$, we also employ an independent distribution: $R_\psi(z|\tau) = \prod_{i=1}^N R_{\psi_i}(z_i|\tau)$, where, we use the same type of distribution as $Q_{\theta_i}(z_i)$, independent of τ.

4. Related work

Differentiability for discrete optimization Discrete optimization problems often suffer from the lack of informative gradients of the objective functions. To address this issue, continuous relaxation for discrete optimization has been actively studied, such as a widely-used reparameterization trick with the Gumbel-softmax distribution (Jang et al., 2016; Maddison et al., 2016). Beyond categorical variables, recent approaches have further advanced the continuous relaxation techniques to more complex discrete objects, including spanning trees (Struminsky et al., 2021). However, it is still nontrivial to extend such techniques to the case with binary tree topologies. As outlined in equation (2), we have introduced a distribution over binary tree topologies $\mathcal{T}$ derived from continuous distributions $Q(z)$. This method facilitates a gradient-based optimization further aided by variance reduction techniques.

Gradient-based algorithms for tree optimization For the hierarchical clustering (HC), which reconstructs a tree relationship based on the distance measures between samples, gradient-based algorithms (Monath et al., 2019; Chami et al., 2020; Chien et al., 2022) have been proposed based on Dasgupta's cost function (Dasgupta, 2016). In particular, Chami et al. (2020) proposed to decode tree topology from hyperbolic coordinates while the optimization is performed for a relaxed cost function, which is differentiable with respect to the coordinates. However, these approaches are not readily applicable to more general problems, including phylogenetic inference, as their formulations depend on the specific form of the cost functions.

Phylogenetic analysis in hyperbolic space The approach of embedding phylogenetic trees into hyperbolic spaces has been explored for visualization and an interpretation of novel samples with existing phylogeny (Matsumoto et al., 2021; Jiang et al., 2022). For the inference task, a maximum-

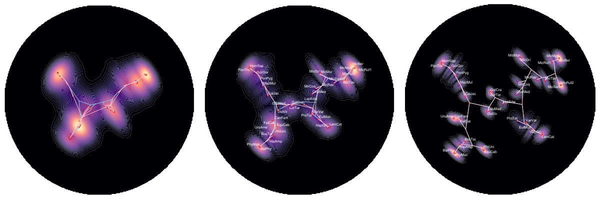

Figure 2. Superimposed probability densities of topological distributions $\sum_{i=1}^{N} Q(z_i)$ up to 100,000, 200,000, and 500,000 steps (MC samples) from the left. For $Q(z)$, we employed a wrapped normal distribution with a two-dimensional full covariance matrix. The experiments used the DS3 dataset ($N = 36$). The majority-rule consensus phylogenetic tree obtained with MrBayes and each step of GeoPhy are shown in blue and red lines, respectively. The center area is magnified by transforming the radius r of the Poincaré coordinates into $\tanh 2.1r$.

likelihood approach was proposed in (Wilson, 2021), which however assumed a simplified likelihood function of pairwise distances. A recent study (Macaulay et al., 2023) proposed an MCMC-based algorithm for sampling (τ, B_τ), which were linked from coordinates z using the NJ algorithm (Saitou & Nei, 1987). However, there remained the issue of an unevaluated Jacobian determinant, which posed a challenge in evaluating inference objectives. Given that we only use topology τ as described in equation (2), the variational lower bound for the inference can be unbiasedly evaluated through sampling, as shown in Proposition 3.1.

5. Experiments

We applied GeoPhy for an approximate posterior inference of phylogenetic models given biological sequence data of species. We trained for GeoPhy until one million Monte-Carlo tree samples were consumed for the gradient estimation of the lower bound objectives. This number equals the number of likelihood evaluations (NLEs) and is used for a standardized comparison of experimental runs (Wang et al., 2015; Zhang & Matsen IV, 2019). More details of the experimental setup are found in Appendix D.

To demonstrate the inference performance of GeoPhy, we compared the marginal log-likelihood (MLL) estimates for the eight real datasets (DS1-8) compiled in (Lakner et al., 2008) against gold-standard values obtained in the study (Zhang & Matsen IV, 2019) using MrBayes stepping-stone (SS) algorithm (Ronquist et al., 2012; Xie et al., 2011). Alongside the training steps, a majority consensus tree procured with GeoPhy progressively aligns with those obtained via MCMC, as demonstrated in Fig. 2.

In Table 1, we compiled the MLL estimates for GeoPhy and other approximate Bayesian inference approaches across eight datasets. Specifically, we present GeoPhy results for two $Q(z)$ configurations: a wrapped normal distribution $\mathcal{WN}$ with a full 4-dimensional covariance matrix and a 2-dimensional diagonal matrix, along with three control variate choices. The consistently superior performance of the larger model, $\mathcal{WN}$(full,4), compared to $\mathcal{WN}$(diag,2) is evident across all datasets. Details of other settings, including those of the normal distributions in Euclidean spaces, are provided in Table 2. There, we noted a slight advantage of wrapped normal distributions over their Euclidean counterparts. While VBPI-GNN (Zhang, 2023) employs a preselected set of tree topologies as its support set before execution, it is known to provide reasonable MLL estimates near the reference values. Other methods including CSMC (Wang et al., 2015), VCSMC (Moretti et al., 2021), ϕ-CSMC (Koptagel et al., 2022), and GeoPhy (our approach), address the more challenging general problem of model optimization by considering all candidate topologies without preselection. Among these methods, GeoPhy consistently outperforms other CSMC-based methods, even when a less performant $\mathcal{WN}$(full,2) is used for $Q(z)$. This demonstrates the stability and efficiency of GeoPhy.

6. Conclusion

We developed a differential phylogenetic inference algorithm named GeoPhy that performed an approximate Bayesian inference without the preselection of candidate topologies. In experiments conducted with real sequence datasets, GeoPhy consistently outperformed other methods that took whole topologies into consideration.

Acknowledgements

We appreciate the valuable comments on our study provided by Dr. Tsukasa Fukunaga of Waseda University and the anonymous reviewers.

References

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance weighted autoencoders. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076, 2020.

Chien, E., Tabaghi, P., and Milenkovic, O. Hyperaid: Denoising in hyperbolic spaces for tree-fitting and hierarchical clustering. *arXiv preprint arXiv:2205.09721*, 2022.

Dasgupta, S. A cost function for similarity-based hierarchi-

Table 1. Comparison of the MLL estimates with different approaches in eight benchmark datasets. The MLL values for MrBayes SS and VBPI-GNN are sourced from (Zhang, 2023), while those for CSMC, VCSMC, and ϕ-CSMSC are referenced from (Koptagel et al., 2022). Following the literature (Zhang & Matsen IV, 2019), we refer to the dataset DS7 in (Koptagel et al., 2022) as DS8. For each configuration of GeoPhy ($\mathcal{WN}$(full,4) and $\mathcal{WN}$(diag,2)), we used three different sets of CVs: LAX with $K = 1$, LOO with $K = 3$ which we labeled LOO(3), and a combination of LOO and LAX, denoted as LOO(3)+. For GeoPhy, the mean MLL values of five independent runs are shown with the standard deviation indicated in parentheses. The bold figures are the best (highest) values obtained with GeoPhy and the tree CSMC-based methods, all of which perform an approximate Bayesian inference without the preselection of topologies. We underlined GeoPhy's MLL estimates that outperformed the other CSMC-based methods, demonstrating superior performance across various configurations.

Dataset	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
#Taxa (N)	27	29	36	41	50	50	59	64
#Sites (M)	1949	2520	1812	1137	378	1133	1824	1008
MrBayes SS	−7108.42	−26367.57	−33735.44	−13330.06	−8214.51	−6724.07	−37332.76	−8649.88
	(0.18)	(0.48)	(0.5)	(0.54)	(0.28)	(0.86)	(2.42)	(1.75)
VBPI-GNN	−7108.41	−26367.73	−33735.12	−13329.94	−8214.64	−6724.37	−37332.04	−8650.65
	(0.14)	(0.07)	(0.09)	(0.19)	(0.38)	(0.4)	(0.26)	(0.45)
CSMC	−8306.76	−27884.37	−35381.01	−15019.21	−8940.62	−8029.51	–	−11013.57
	(166.27)	(226.6)	(218.18)	(100.61)	(46.44)	(83.67)	–	(113.49)
VCSMC	−9180.34	−28700.7	−37211.2	−17106.1	−9449.65	−9296.66	–	–
	(170.27)	(4892.67)	(397.97)	(362.74)	(2578.58)	(2046.7)	–	–
ϕ-CSMC	−7290.36	−30568.49	−33798.06	−13582.24	−8367.51	−7013.83	–	−9209.18
	(7.23)	(31.34)	(6.62)	(35.08)	(8.87)	(16.99)	–	(18.03)
$\mathcal{WN}$(full,4)	$\underline{\mathbf{-7111.55}}$	$\underline{-26379.48}$	$\underline{-33757.79}$	$\underline{-13342.71}$	$\underline{-8240.87}$	$\underline{-6735.14}$	$\underline{-37377.86}$	$\underline{-8663.51}$
	(0.07)	(11.60)	(8.07)	(1.61)	(9.80)	(2.64)	(29.48)	(6.85)
LOO(3)	$\underline{-7119.77}$	$\underline{\mathbf{-26368.44}}$	$\underline{-33736.01}$	$\underline{-13339.26}$	$\underline{-8234.06}$	$\underline{\mathbf{-6733.91}}$	$\underline{\mathbf{-37350.77}}$	$\underline{-8671.32}$
	(11.80)	(0.13)	(0.03)	(3.19)	(7.53)	(0.57)	(11.74)	(5.99)
LOO(3)+	$\underline{-7116.09}$	$\underline{-26368.54}$	$\underline{\mathbf{-33735.85}}$	$\underline{\mathbf{-13337.42}}$	$\underline{\mathbf{-8233.89}}$	$\underline{-6735.90}$	$\underline{-37358.96}$	$\underline{\mathbf{-8660.48}}$
	(10.67)	(0.12)	(0.12)	(1.32)	(6.63)	(1.13)	(13.06)	(0.78)
$\mathcal{WN}$(diag,2)	$\underline{-7126.89}$	$\underline{-26444.84}$	−33823.74	$\underline{-13358.16}$	$\underline{-8251.45}$	$\underline{-6745.60}$	$\underline{-37516.88}$	$\underline{-8719.44}$
	(10.06)	(27.91)	(15.62)	(9.79)	(9.72)	(8.36)	(69.88)	(60.54)
LOO(3)	$\underline{-7130.67}$	$\underline{-26380.41}$	$\underline{-33737.75}$	$\underline{-13346.94}$	$\underline{-8239.36}$	$\underline{-6741.63}$	$\underline{-37382.28}$	$\underline{-8690.41}$
	(10.67)	(14.40)	(2.48)	(4.25)	(4.62)	(3.23)	(31.96)	(15.92)
LOO(3)+	$\underline{-7128.40}$	$\underline{-26375.28}$	$\underline{-33736.91}$	$\underline{-13347.32}$	$\underline{-8235.41}$	$\underline{-6742.40}$	$\underline{-37411.28}$	$\underline{-8683.22}$
	(9.78)	(11.78)	(1.91)	(4.42)	(5.70)	(1.94)	(56.74)	(13.13)

cal clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 118–127, 2016.

Felsenstein, J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

Garey, J. R., Near, T. J., Nonnemacher, M. R., and Nadler, S. A. Molecular evidence for acanthocephala as a subtaxon of rotifera. *Journal of Molecular Evolution*, 43: 287–292, 1996.

Grathwohl, W., Choi, D., Wu, Y., Roeder, G., and Duvenaud, D. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

Hedges, S. B., Moberg, K. D., and Maxson, L. R. Tetrapod phylogeny inferred from 18s and 28s ribosomal rna sequences and a review of the evidence for amniote relationships. *Molecular Biology and Evolution*, 7(6):607–633, 1990.

Henk, D. A., Weir, A., and Blackwell, M. Laboulbeniopsis termitarius, an ectoparasite of termites newly recognized as a member of the laboulbeniomycetes. *Mycologia*, 95 (4):561–564, 2003.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Jiang, Y., Tabaghi, P., and Mirarab, S. Phylogenetic placement problem: A hyperbolic embedding approach. In *Comparative Genomics: 19th International Conference, RECOMB-CG 2022, La Jolla, CA, USA, May 20–21, 2022, Proceedings*, pp. 68–85. Springer, 2022.

Jukes, T. H., Cantor, C. R., et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kool, W., van Hoof, H., and Welling, M. Buy 4 reinforce samples, get a baseline for free! 2019.

Koptagel, H., Kviman, O., Melin, H., Safinianaini, N., and Lagergren, J. Vaiphy: a variational inference based algorithm for phylogeny. In *Advances in Neural Information Processing Systems*, 2022.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Systematic biology*, 57(1):86–103, 2008.

Macaulay, M., Darling, A. E., and Fourment, M. Fidelity of hyperbolic space for bayesian phylogenetic inference. *PLoS computational biology*, 2023.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Matsumoto, H., Mimori, T., and Fukunaga, T. Novel metric for hyperbolic phylogenetic tree embeddings. *Biology Methods and Protocols*, 6(1):bpab006, 2021.

Monath, N., Zaheer, M., Silva, D., McCallum, A., and Ahmed, A. Gradient-based hierarchical clustering using continuous representations of trees in hyperbolic space. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 714–722, 2019.

Moretti, A. K., Zhang, L., Naesseth, C. A., Venner, H., Blei, D., and Pe'er, I. Variational combinatorial sequential monte carlo methods for bayesian phylogenetic inference. In *Uncertainty in Artificial Intelligence*, pp. 971–981. PMLR, 2021.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pp. 4693–4702. PMLR, 2019.

Richter, L., Boustati, A., Nüsken, N., Ruiz, F., and Akyildiz, O. D. Vargrad: a low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 33:13481–13492, 2020.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

Rossman, A. Y., McKemy, J. M., Pardo-Schultheiss, R. A., and Schroers, H.-J. Molecular studies of the bionectriaceae using large subunit rdna sequences. *Mycologia*, 93 (1):100–110, 2001.

Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

Sala, F., De Sa, C., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pp. 4460–4469. PMLR, 2018.

Struminsky, K., Gadetsky, A., Rakitin, D., Karpushkin, D., and Vetrov, D. P. Leveraging recursive gumbel-max trick for approximate inference in combinatorial spaces. *Advances in Neural Information Processing Systems*, 34: 10999–11011, 2021.

Wang, L., Bouchard-Côté, A., and Doucet, A. Bayesian phylogenetic inference using a combinatorial sequential monte carlo method. *Journal of the American Statistical Association*, 110(512):1362–1374, 2015.

Wilson, B. Learning phylogenetic trees as hyperbolic point configurations. *arXiv preprint arXiv:2104.11430*, 2021.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2): 150–160, 2011.

Yang, Z. and Yoder, A. D. Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic biology*, 52(5):705–716, 2003.

Yoder, A. D. and Yang, Z. Divergence dates for malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. *Molecular Ecology*, 13(4):757–773, 2004.

Zhang, C. Improved variational bayesian phylogenetic inference with normalizing flows. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18760–18771. Curran Associates, Inc., 2020.

Zhang, C. Learnable topological features for phylogenetic inference via graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhang, C. and Matsen IV, F. A. Generalizing tree probability estimation via bayesian networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1449–1458. Curran Associates, Inc., 2018.

Zhang, C. and Matsen IV, F. A. Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*, 2019.

Zhang, N. and Blackwell, M. Molecular phylogeny of dogwood anthracnose fungus (discula destructiva) and the diaporthales. *Mycologia*, 93(2):355–365, 2001.

A. Hyperbolic spaces and wrapped normal distributions

Hyperbolic spaces are known to be able to embed hierarchical data with less distortion in considerably fewer dimensions than Euclidean spaces (Sala et al., 2018). The wrapped normal distribution, as proposed in (Nagano et al., 2019), is a probability distribution defined on hyperbolic spaces, which is easy to sample from and evaluate its probability density at arbitrary coordinates in the hyperbolic spaces. In the following, we provide a detailed summary of the calculation involved in applying the wrapped normal distributions within the context of the Lorentz model of hyperbolic spaces.

The Lorentz model, denoted as $\mathbb{H}^d$, represents the d-dimensional hyperbolic space as a submanifold of a $d+1$ dimensional Euclidean space. Given $u, v \in \mathbb{R}^{d+1}$, we can define the pseudo-inner product and pseudo-norm as follows:

$$\langle u, v \rangle_L := -u_0 v_0 + \sum_{j=1}^{d} u_j v_j, \quad \|u\|_L := \sqrt{\langle u, u \rangle_L}. \tag{6}$$

The distance between hyperbolic coordinates $\nu, \mu \in \mathbb{H}^d$ is defined as follows:

$$\mathrm{d}(\nu, \mu) := \cosh^{-1}(-\langle \nu, \mu \rangle_L), \tag{7}$$

where $\cosh^{-1}$ denotes the inverse function of the hyperbolic cosine function. Consider hyperbolic coordinates $\nu, \mu \in \mathbb{H}^d$ and tangent vectors $u \in T_\mu \mathbb{H}^d$ and $v \in T_\nu \mathbb{H}^d$. An exponential map $\exp_\mu(u) \in \mathbb{H}^d$, a logarithm map $\log_\mu(\nu) \in T_\mu \mathbb{H}^d$, and a parallel transport map $\mathrm{PT}_{\nu \to \mu}(v) \in T_\mu \mathbb{H}^d$, can be calculated as follows:

$$\exp_\mu(u) = \cosh(\|u\|_L)\mu + \sinh(\|u\|_L)\frac{u}{\|u\|_L}, \tag{8}$$

$$\log_\mu(\nu) = \frac{\cosh^{-1}(\alpha)}{\sqrt{\alpha^2 - 1}} (\nu - \alpha\mu), \tag{9}$$

$$\mathrm{PT}_{\nu \to \mu}(v) = v + \frac{\langle \mu - \alpha\nu, v \rangle_L}{\alpha + 1}(\nu + \mu), \tag{10}$$

where we denote $\alpha = -\langle \nu, \mu \rangle_L$, and $\cosh^{-1}$ represents the inverse function of $\cosh$.

Given location and scale parameters denoted as $\mu \in \mathbb{H}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, respectively, the procedure for sampling from a wrapped normal distribution $z \sim \mathcal{WN}(\mu, \Sigma)$ defined over $\mathbb{H}^d$ is given as follows:

$$z = \exp_\mu \circ \mathrm{PT}_{\mu^o \to \mu}(u), \quad u_{1:d} \sim \mathcal{N}(0, \Sigma), \tag{11}$$

where $\mu^o = (1, 0, \ldots, 0)^\top$ denotes the origin of the $\mathbb{H}^d$. Note that we set $u_0 = 0$, and $u := u_{0:d} \in T_{\mu^o}\mathbb{H}^d$ represents a tangent vector at the origin $T_{\mu^o}\mathbb{H}^d$.

From the sampling definition in equation (11), the probability density function $\mathcal{WN}(z; \mu, \Sigma)$ can be derived as follows:

$$\log \mathcal{WN}(z;\, \mu, \Sigma) = \log \mathcal{N}(u;\, 0, \Sigma) - (d-1) \ln \left(\frac{\sinh \|u\|_L}{\|u\|_L} \right), \tag{12}$$

where u is defined as $u = \mathrm{PT}_{\mu \to \mu^o} \circ \log_\mu(z)$. For detailed derivation, we refer to Appendix A of (Nagano et al., 2019).

B. GNN-based parameterization for variational branch length distributions

In this work, we employ a variational branch length distribution $Q_\phi(B_\tau|\tau)$ parameterized with a graph neural network (GNN) as described in (Zhang, 2023). In concrete, each of the branch lengths follows an independent lognormal distribution, where its location and scale parameters are predicted with a GNN that takes the tree topology τ and the learnable topological features (LTFs) of the topology τ, which are computed with a method described in (Zhang, 2023). Below, we summarize an architecture that we use in this study.

Branch length parameterizations Let V_τ and E_τ respectively represent the sets of nodes and branch edges for a given unrooted binary tree topology τ. The input to the GNN consists of node features represented by LTFs denoted as $\{h_v^{(0)}\}_{v \in V_\tau}$. These features undergo transformation L times as follows:

$$\{h_v^{(L)}\}_{v \in V_\tau} = \mathrm{GNN}(\{h_v^{(0)}\}_{v \in V_\tau}) = g^{(L)} \circ \cdots \circ g^{(1)}(\{h^{(0)}\}_{v \in V_\tau}), \tag{13}$$

where we set $L = 2$. The function $g^{(\ell)}$ represents a GNN layer. For this function, we utilize edge convolutional layers, which will be described in more detail in the following paragraph.

Next, the last node features $h^{(L)}$ are transformed to output parameters of edge length as follows:

$$\widetilde{h}_v = \mathrm{MLP}_V(h_v^{(L)}), \qquad (\forall v \in V_\tau) \tag{14}$$

$$\widetilde{m}_{(v,u)} = \mathrm{MAX}(\widetilde{h}_v, \widetilde{h}_u), \qquad (\forall (v,u) \in E_\tau) \tag{15}$$

$$\mu_{(v,u)}, \log \sigma_{(v,u)} = \mathrm{MLP}_E(\widetilde{m}_{(v,u)}) \qquad (\forall (v,u) \in E_\tau) \tag{16}$$

where MLP_N, MAX, and MLP_E denotes a multi-layer perceptron for node features with two hidden layers, the element-wise max operation, and a multi-layer perceptron with a hidden layer that outputs the location and scale parameter (μ_e, σ_e) of the lognormal distributions for each edge $e \in \mathbb{E}_\tau$. For each of the hidden layers employed in MLP_N and MLP_E, we set its width to 100 and apply the ELU activation function after the linear transformation of input values.

Edge convolutional layers In a previous study (Zhang, 2023), a GNN with edge convolutional layers, referred to as EDGE, demonstrated strong performance when predicting the posterior tree distributions. In EDGE, the function $g^{(\ell)}$ transforms node features $\{h_v^{(\ell)}\}_{v \in V_\tau}$ according to the following scheme:

$$\{h_v^{(\ell+1)}\}_{v \in V_\tau} = g^{(\ell)}(\{h_v^{(\ell)}\}_{v \in V_\tau}), \tag{17}$$

where $g^{(\ell)}$ is comprised of the edge convolutional operation with the exponential linear unit (ELU) activation function. Specifically, the transformation with the layer $g^{(\ell)}$ is computed as follows:

$$e_{u \to v}^{(\ell)} = \mathrm{MLP}^{(\ell)}\left(h_v^{(\ell)} \| h_u^{(\ell)} - h_v^{(\ell)}\right), \quad \forall u \in N_\tau(v) \tag{18}$$

$$h_v'^{(\ell+1)} = \underset{u \in N_\tau(v)}{\mathrm{AGG}^{(\ell)}} e_{u \to v}^{(\ell)}, \tag{19}$$

$$h_v^{(\ell+1)} = \mathrm{ELU}\left(h_v'^{(\ell+1)}\right), \tag{20}$$

where $N_\tau(v)$ represents a set of neighboring nodes connected to node v in the tree topology τ, $\|$ refers to the concatenation operation of elements, $\mathrm{MLP}^{(\ell)}$ denotes a full connection layer and the exponential linear unit (ELU) activation unit, and $\mathrm{AGG}^{(\ell)}$ represents an aggregation operation that takes the maximum value of neighboring edge features $e_{u \to v}^{(\ell)}, \forall u \in N(v)$ for each element.

C. Variational Lower Bounds and Gradient Estimators

C.1. Variational lower bound

In Proposition C.1 we will demonstrate that the following functional is a lower bound of the marginal log-likelihood $\ln P(Y)$:

$$\mathcal{L}[Q, R] := \mathbb{E}_{Q(z,B_\tau)}\left[\ln F'(z, B_\tau)\right] = \mathbb{E}_{Q(z)}\left[\mathbb{E}_{Q(B_\tau|z)}[\ln F(z, B_\tau)] - \ln Q(z)\right] \tag{21}$$

$$\leq \ln P(Y), \tag{22}$$

where F and F' are respectively defined as follows:

$$F(z, B_\tau) := \frac{P(Y, B_\tau|\tau(z))}{Q(B_\tau|\tau(z))} P(\tau(z)) R(z|\tau(z)), \quad F'(z, B_\tau) := \frac{F(z, B_\tau)}{Q(z)}. \tag{23}$$

Proposition C.1 (Restatement of Proposition 3.1)**.** *The relation* $\ln P(Y) \geq \mathcal{L}[Q] \geq \mathcal{L}[Q, R]$ *holds, where the first and second equality holds when* $Q(\tau, B_\tau) = P(\tau, B_\tau|Y)$ *and* $R(z|\tau) = Q(z|\tau)$*, respectively.*

Proof. The first variational lower bound of the marginal log-likelihood is given as follows:

$$\ln P(Y) \geq \ln P(Y) - D_{\mathrm{KL}}\left[Q(\tau, B_\tau) \| P(\tau, B_\tau|Y)\right] = \mathbb{E}_{Q(\tau,B_\tau)}\left[\ln \frac{P(Y, \tau, B_\tau)}{Q(\tau, B_\tau)}\right] := \mathcal{L}[Q], \tag{24}$$

where the equality condition of the first inequality holds when $Q(\tau, B_\tau) = P(\tau, B_\tau|Y)$. Since we have defined $Q(\tau)$ in equation (2), we can further transform the lower bound as $\mathcal{L}[Q]$

$$\mathcal{L}[Q] = \mathbb{E}_{Q(z)} \left[\sum_{\tau \in \mathcal{T}} \mathbb{I}[\tau = \tau(z)] \, \mathbb{E}_{Q(B_\tau|\tau)} \left[\ln \frac{P(Y, B_\tau|\tau)}{Q(B_\tau|\tau)} + \ln \frac{P(\tau)}{Q(\tau)} \right] \right], \tag{25}$$

from which equation (3) immediately follows. Hence the inequality $\ln P(Y) \geq \mathcal{L}[Q]$ and its equality condition $Q(\tau, B_\tau) = P(\tau, B_\tau|Y)$ have been proven.

Next, the entropy term $-\mathbb{E}_{Q(z)}[\ln Q(\tau(z))] = -\mathbb{E}_{Q(\tau)}[\ln Q(\tau)]$ in equation (3) can be transformed to derive further lower bound as follows:

$$\begin{aligned} -\mathbb{E}_{Q(\tau)}\left[\ln Q(\tau)\right] &\geq -\mathbb{E}_{Q(\tau)}\left[\ln Q(\tau) + D_{\mathrm{KL}}\left(Q(z|\tau) \| R(z|\tau)\right)\right] \\ &= -\mathbb{E}_{Q(\tau)Q(z|\tau)}\left[\ln \frac{Q(\tau)Q(z|\tau)}{R(\tau|z)}\right] = -\mathbb{E}_{Q(z)}\left[\ln \frac{Q(z)}{R(z|\tau(z))}\right], \end{aligned}$$

where the last equality is derived by using the relation $\mathbb{E}_{Q(\tau)Q(z|\tau)}[\cdot] = \mathbb{E}_{Q(z)}[\sum_\tau \mathbb{I}[\tau = \tau(z)] \cdot]$ and $Q(\tau)Q(z|\tau) = Q(z)\mathbb{I}[\tau = \tau(z)]$. The equality condition of the first inequality holds when $R(z|\tau) = Q(z|\tau)$. Hence, the inequality $\mathcal{L}[Q] \geq \mathcal{L}[Q, R]$ and the equality condition is proven. □

C.2. Gradient estimators for variational lower bound

The gradient of $\mathcal{L}[Q_{\theta,\phi}, R_\psi]$ with respect to θ is given by

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_{Q_\theta(z)}\left[\mathbb{E}_{Q_\phi(B_\tau|z)}[\ln F(z, B_\tau)] - \ln Q_\theta(z)\right] & (26) \\ &= \mathbb{E}_{Q_\theta(z)}\left[(\nabla_\theta \ln Q_\theta(z)) \left(\mathbb{E}_{Q_\phi(B_\tau|\tau(z))}\left[\ln \frac{P(Y, B_\tau|\tau(z))}{Q_\phi(B_\tau|\tau(z))}\right] + \ln P(\tau(z)) R_\psi(z|\tau(z))\right)\right] \\ &\quad + \nabla_\theta \mathbb{H}[Q_\theta(z)], & (27) \end{aligned}$$

where $\mathbb{H}$ denotes the differential entropy. We assume that $Q_\phi(B_\tau|\tau)$ is reparameterizable as in (Zhang, 2023): namely, B_τ can be sampled through $B_\tau = h_\phi(\epsilon_B, \tau)$, where $\epsilon_B \sim p_B(\epsilon_B)$, where $p_B(\epsilon)$ and h_ϕ denote a parameter-free base distribution and a differentiable function with ϕ, respectively. Consequently, the gradient of $\mathcal{L}$ with respect to ϕ is evaluated as follows:

$$\begin{aligned} \nabla_\phi \mathcal{L} &= \nabla_\phi \mathbb{E}_{Q_\theta(z)} \mathbb{E}_{Q_\phi(B_\tau|z)}[\ln F(z, B_\tau)] & (28) \\ &= \mathbb{E}_{Q_\theta(z)} \mathbb{E}_{p_B(\epsilon_B)}\left[\nabla_\phi \ln \frac{P(Y, B_\tau = h_\phi(\epsilon_B, \tau)|\tau(z))}{Q_\phi(B_\tau = h_\phi(\epsilon_B, \tau)|\tau(z))}\right]. & (29) \end{aligned}$$

Lastly, the gradient of $\mathcal{L}$ with respect to ψ can be evaluated with a tractable density model $R_\psi(z|\tau)$ as follows:

$$\nabla_\psi \mathcal{L} = \nabla_\psi \mathbb{E}_{Q_\theta(z)} \mathbb{E}_{Q_\phi(B_\tau|z)}[\ln F(z, B_\tau)] = \mathbb{E}_{Q_\theta(z)}\left[\nabla_\psi \ln R_\psi(z|\tau(z))\right]. \tag{30}$$

Given samples $\epsilon_z \sim p_z$ and $\epsilon_B \sim p_B$, we can compute $z = h_\theta(\epsilon_z)$, $\tau(z)$, and $B_\tau = h_\phi(\epsilon_B, \tau(z))$. Then, the below equations are estimators of gradients $\nabla_\theta \mathcal{L}$, $\nabla_\phi \mathcal{L}$, and $\nabla_\psi \mathcal{L}$, respectively:

$$\begin{aligned} \widehat{g}_\theta &= \nabla_\theta \ln Q_\theta(z) \cdot \ln F(z, B_\tau) - \nabla_\theta \ln Q_\theta(h_\theta(\epsilon_z)), & (31) \\ \widehat{g}_\phi &= \nabla_\phi \ln F(z, h_\phi(\epsilon_B, \tau(z))) = \nabla_\phi \ln \frac{P(Y, h_\phi(\epsilon_B, \tau(z))|\tau(z))}{Q_\phi(h_\phi(\epsilon_B, \tau(z))|\tau(z))}, & (32) \\ \widehat{g}_\psi &= \nabla_\psi \ln F(z, h_\phi(\epsilon_B, \tau(z))) = \nabla_\psi \ln R_\psi(z|\tau(z)). & (33) \end{aligned}$$

The gradients can be computed through the auto-gradient of the following target:

$$\widehat{\mathcal{L}}' = \ln Q_\theta(z) \cdot \mathrm{detach}[f(z, B_\tau)] + f(z, h_\phi(\epsilon_B, \tau(z))) - \ln Q_\theta(h_\theta(\epsilon_z)), \tag{34}$$

where we denote $f(z, B_\tau) = \ln F(z, B_\tau)$, and $\mathrm{detach}[\cdot]$ refers to an operation that blocks backpropagation through its argument. For clarity in terms of differentiability with respect to the parameters, we distinguish between expressions (z, B_τ) and $(h_\theta(\epsilon_z), h_\phi(\epsilon_B, \tau(z)))$.

C.3. Multi-sample gradient estimators

Given a K set of Monte-Carlo (MC) samples from $Q_{\theta,\phi}(z, B_\tau)$, i.e. $\{\epsilon_Z^{(k)}, z^{(k)} = h_\theta(\epsilon_Z^{(k)})\}_{k=1}^K$ and $\{\epsilon_B^{(k)}, B_\tau^{(k)} = h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))\}_{k=1}^K$, we can simply estimate $\nabla L_\theta[Q_{\theta,\phi}, R_\psi]$ as follows:

$$\widehat{g}_\theta^{(K)} = \frac{1}{K}\sum_{k=1}^K \left(\nabla_\theta \ln Q_\theta(z^{(k)}) \cdot f(z^{(k)}, B_\tau^{(k)}) - \nabla_\theta \ln Q_\theta(h_\theta(\epsilon_z^{(k)}))\right). \tag{35}$$

As a simple extension of equation (34), the gradients are obtained through an auto-gradient computation of the following target:

$$\widehat{\mathcal{L}'}^{(K)} = \frac{1}{K}\sum_{k=1}^K \left(\ln Q_\theta(z^{(k)}) \cdot \mathrm{detach}[f(z^{(k)}, B_\tau^{(k)})] + f(z^{(k)}, h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))) - \ln Q_\theta(h_\theta(\epsilon_z^{(k)}))\right), \tag{36}$$

C.4. Leave-one-out (LOO) control variates for variance reduction

For the term of K-sample gradient estimator $\widehat{g}_\theta^{(K)}$ proportional to the score function $\nabla_\theta \ln Q_\theta$, a leave-one-out (LOO) variance reduction is known to be effective (Kool et al., 2019; Richter et al., 2020), which is denoted as follows:

$$\widehat{g}_{\mathrm{LOO},\theta}^{(K)} = \frac{1}{K}\sum_{k=1}^K \left[\nabla_\theta \ln Q_\theta(z^{(k)}) \cdot \left(f(z^{(k)}, B_\tau^{(k)}) - \overline{f_k}(z^{(\backslash k)}, B_\tau^{(\backslash k)})\right) - \nabla_\theta \ln Q_\theta(h_\theta(\epsilon_z^{(k)}))\right], \tag{37}$$

where $\overline{f_k}$ denotes:

$$\overline{f_k}(z^{(\backslash k)}, B_\tau^{(\backslash k)}) := \frac{1}{K-1}\sum_{k'=1, k'\neq k}^K f(z^{(k')}, B_\tau^{(k')}). \tag{38}$$

To employ the LOO gradient estimator for θ, the target of auto-gradient computation in equation (36) needs to be adjusted as follows:

$$\begin{aligned}\widehat{\mathcal{L}'}_{\mathrm{LOO}}^{(K)} = \frac{1}{K}\sum_{k=1}^K \Big(&\ln Q_\theta(z^{(k)}) \cdot \mathrm{detach}[f(z^{(k)}, B_\tau^{(k)}) - \overline{f_k}(z^{(\backslash k)}, B_\tau^{(\backslash k)})] \\ &+ f(z^{(k)}, h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))) - \ln Q_\theta(h_\theta(\epsilon_z^{(k)}))\Big),\end{aligned} \tag{39}$$

C.5. LAX estimators for adaptive variance reduction

The LAX estimator (Grathwohl et al., 2018) is a stochastic gradient estimator based on a surrogate function, which can be adaptively learned to reduce the variance regarding the term $\nabla_\theta \ln Q_\theta(z)$. In our case, the LAX estimator is given as follows:

$$\widehat{g}_{\mathrm{LAX},\theta} := \nabla_\theta \ln Q_\theta(z) \cdot (f(z, B_\tau) - s_\chi(z)) + \nabla_\theta s_\chi(h_\theta(\epsilon_z)). \tag{40}$$

As we assume $Q_\theta(z)$ is differentiable with respect to z, we can also use a modified estimator as follows:

$$\widehat{g}_{\mathrm{LAX},\theta} := \nabla_\theta \ln Q_\theta(z) \cdot (f(z, B_\tau) - s_\chi(z)) + \nabla_\theta s_\chi(h_\theta(\epsilon_z)) - \nabla_\theta \ln Q_\theta(h_\theta(\epsilon_z)). \tag{41}$$

Since it is favorable to reduce the variance of $\widehat{g}_{\mathrm{LAX},\theta}$, we optimize χ to minimize the following objective as proposed in (Grathwohl et al., 2018):

$$\left\langle \mathbb{V}_{Q_\theta(z)}[\widehat{g}_\theta] \right\rangle := \frac{1}{n_\theta}\sum_{i=1}^{n_\theta} \mathbb{V}_{Q_\theta(z)}[\widehat{g}_{\theta_i}] = \frac{1}{n_\theta}\sum_{i=1}^{n_\theta} \left(\mathbb{E}_{Q_\theta(z)}[\widehat{g}_{\theta_i}^2] - \mathbb{E}_{Q_\theta(z)}[\widehat{g}_{\theta_i}]^2\right), \tag{42}$$

where n_θ denotes the dimension of θ. As the gradient in equation (40) is given as an unbiased estimator of $\nabla_\theta \mathcal{L}$, which is not dependent on χ, we can use the relation $\nabla_\chi \mathbb{E}_{Q_\theta(z)}[\widehat{g}_{\mathrm{LAX},\theta_i}] = 0$. Therefore, the unbiased estimator of the gradient $\nabla_\chi \left\langle \mathbb{V}_{Q_\theta(z)}[\widehat{g}_{\mathrm{LAX},\theta}] \right\rangle$ is given as follows:

$$\widehat{g}_\chi = \frac{1}{n_\theta} \sum_{i=1}^{n_\theta} \nabla_\chi \widehat{g}^2_{\mathrm{LAX},\theta_i}. \tag{43}$$

As we require the gradient of $\nabla_\theta \mathcal{L}$ with respect to χ for the optimization, we use different objectives for auto-gradient computation with respect to θ and the other parameters ϕ and ψ as follows:

$$\widehat{\mathcal{L}'}_{\mathrm{LAX},\theta} = \ln Q_\theta(z) \cdot (\mathrm{detach}[f(z, B_\tau)] - s_\chi(z)) + s_\chi(h_\theta(\epsilon_z)) - \ln Q_\theta(h_\theta(\epsilon_z)), \tag{44}$$

$$\widehat{\mathcal{L}'}_{\phi,\psi} = f(z, h_\phi(\epsilon_B, \tau(z))). \tag{45}$$

C.6. LAX estimators with multiple MC-samples

For the cases with K MC-samples, we use LAX estimators by differentiating the following objectives:

$$\widehat{\mathcal{L}'}^{(K)}_{\mathrm{LAX},\theta} = \frac{1}{K} \sum_{k=1}^{K} \left(\ln Q_\theta(z^{(k)}) \cdot \left(\mathrm{detach}[f(z^{(k)}, B_\tau^{(k)})] - s_\chi(z^{(k)}) \right) + s_\chi(h_\theta(\epsilon_z^{(k)})) - \ln Q_\theta(h_\theta(\epsilon_z^{(k)})) \right), \tag{46}$$

$$\widehat{\mathcal{L}'}^{(K)}_{\phi,\psi} = \frac{1}{K} \sum_{k=1}^{K} f(z^{(k)}, h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))). \tag{47}$$

When we combine LAX estimators with LOO control variates. the target for auto-gradient computation changes to the following:

$$\begin{aligned} \widehat{\mathcal{L}'}^{(K)}_{\mathrm{LOO+LAX},\theta} = \frac{1}{K} \sum_{k=1}^{K} \Big(& \ln Q_\theta(z^{(k)}) \cdot \left(\mathrm{detach}[f(z^{(k)}, B_\tau^{(k)}) - \overline{f_k}(z^{(\backslash k)}, B_\tau^{(\backslash k)})] - s_\chi(z^{(k)}) \right) \\ & + s_\chi(h_\theta(\epsilon_z^{(k)})) - \ln Q_\theta(h_\theta(\epsilon_z^{(k)})) \Big). \end{aligned} \tag{48}$$

We note that $\widehat{\mathcal{L}'}^{(K)}_{\phi,\psi}$ is not affected by the introduction of LOO control variates.

C.7. Derivation of importance-weighted evidence lower bound (IW-ELBO)

An importance-weighted evidence lower bound (IW-ELBO) (Burda et al., 2016), is a tighter lower bound of the log-likelihood $\ln P(Y)$ than ELBO. For our model, a conventional K-sample IW-ELBO is given as follows:

$$\mathcal{L}^{(K)}_{\mathrm{IW}}[Q] := \mathbb{E}_{Q(z^{(1)}, B_\tau^{(1)}) \cdots Q(z^{(K)}, B_\tau^{(K)})} \left[\ln \frac{1}{K} \sum_{k=1}^{K} \frac{P(Y, B_\tau^{(k)}, \tau(z^{(k)}))}{Q(B_\tau^{(k)}, \tau(z^{(k)}))} \right]. \tag{49}$$

The fact that $\mathcal{L}^{(K)}_{\mathrm{IW}}[Q]$ is the lower bound of $\ln P(Y)$ is directly followed from Theorem 1 in (Burda et al., 2016). However, as our model cannot directly evaluate the mass function $Q(\tau)$, we must resort to considering the second lower bound, similar to the case of $K = 1$ as depicted in Proposition 3.1. We define the K-sample tractable IW-ELBO as follows:

$$\mathcal{L}^{(K)}_{\mathrm{IW}}[Q, R] := \mathbb{E}_{Q(z^{(1)}, B_\tau^{(1)}) \cdots Q(z^{(K)}, B_\tau^{(K)})} \left[\ln \frac{1}{K} \sum_{k=1}^{K} F'(z^{(k)}, B_\tau^{(k)}) \right], \tag{50}$$

where F' is defined in equation (23). We will prove in Theorem C.3 that $\mathcal{L}^{(K)}_{\mathrm{IW}}[Q, R]$ serves as a lower bound of the $\ln P(Y)$. Although this inequality holds when $K = 1$, as shown by $\ln P(Y) \geq \mathcal{L}[Q] \geq \mathcal{L}[Q, R]$ in Proposition 3.1, the relationship is less obvious when $K > 1$. Before delving into that, we prepare the following proposition.

Proposition C.2. *Given $Q(z, \tau)$ as defined in equation 2 and an arbitrary conditional distribution $R(z|\tau)$, it follows that*

$$\mathbb{E}_{R(z|\tau)}[\mathbb{I}[\tau = \tau(z)]] \leq 1, \tag{51}$$

where setting $R(z|\tau) = Q(z|\tau)$ is a sufficient condition for the equality to hold.

Proof. The inequality immediately follows from the definition as follows:

$$\mathbb{E}_{R(z|\tau)}[\mathbb{I}[\tau = \tau(z)]] \leq \mathbb{E}_{R(z|\tau)}[1] = 1. \tag{52}$$

Next, when we set $R(z|\tau) = Q(z|\tau)$, the condition for equality is satisfied as follows:

$$\mathbb{E}_{Q(z|\tau)}[\mathbb{I}[\tau = \tau(z)]] = \frac{\mathbb{E}_{Q(z)}[\mathbb{I}[\tau = \tau(z)]^2]}{Q(\tau)} = \frac{Q(\tau)}{Q(\tau)} = 1, \tag{53}$$

where we have used the definition of $Q(\tau) := \mathbb{E}_{Q(z)}[\mathbb{I}[\tau = \tau(z)]]$ from equation (2) and the resulting relation $Q(z|\tau)Q(\tau) = \mathbb{I}[\tau = \tau(z)]Q(z)$. □

Theorem C.3. *Given $Q(z, \tau)$ as defined in equation 2 and an arbitrary conditional distribution $R(z|\tau)$, for any natural number $K > 1$, the following relation holds:*

$$\ln P(Y) \geq \mathcal{L}_{\text{IW}}^{(K)}[Q, R] \geq \mathcal{L}_{\text{IW}}^{(K-1)}[Q, R]. \tag{54}$$

Additionally, if $F'(z, B_\tau)$ is bounded and $\forall\tau, \mathbb{E}_{R(z|\tau)}[\mathbb{I}[\tau = \tau(z)]] = 1$, then $\mathcal{L}_{\text{IW}}^{(K)}[Q, R]$ approaches $\ln P(Y)$ as $K \to \infty$.

Proof. We first show that for any natural number $K > M$,

$$\mathcal{L}_{\text{IW}}^{(K)}[Q, R] \geq \mathcal{L}_{\text{IW}}^{(M)}[Q, R]. \tag{55}$$

For simplicity, we denote $Q(z^{(k)}, B_\tau^{(k)})$ and $F'(z^{(k)}, B_\tau^{(k)})$ as Q_k and F'_k, respectively, in the following discussion. Let U_M^K represent a uniform distribution over a subset with M distinct indices chosen from the K indices $\{1, \ldots, K\}$. Similar to the approach used in (Burda et al., 2016), we will utilize the following relationship:

$$\frac{1}{K}\sum_{k=1}^{K} F'_k = \mathbb{E}_{\{i_1,\ldots,i_M\}\sim U_M^K}\left[\frac{1}{M}\sum_{m=1}^{M} F'_{i_m}\right]. \tag{56}$$

Now, the inequality (55) is derived as follows:

$$\begin{aligned}
\mathbb{E}_{Q_1\cdots Q_K}\left[\ln\left(\frac{1}{K}\sum_{k=1}^{K} F'_k\right)\right] &= \mathbb{E}_{Q_1\cdots Q_K}\left[\ln \mathbb{E}_{\{i_1,\ldots,i_m\}\sim U_M^K}\left[\left(\frac{1}{M}\sum_{m=1}^{M} F'_{i_m}\right)\right]\right] && (57)\\
&\geq \mathbb{E}_{Q_1\cdots Q_K}\left[\mathbb{E}_{\{i_1,\ldots,i_m\}\sim U_M^K}\left[\ln\left(\frac{1}{M}\sum_{m=1}^{M} F'_{i_m}\right)\right]\right] && (58)\\
&= \mathbb{E}_{Q_1\cdots Q_M}\left[\ln\left(\frac{1}{M}\sum_{m=1}^{M} F'_m\right)\right], && (59)
\end{aligned}$$

where we have also used Jensen's inequality.

Next, we show that $\ln P(Y) \geq \mathcal{L}_{\text{IW}}^{(K)}[Q, R]$. We again use Jensen's inequality as follows:

$$\begin{aligned}
\mathcal{L}_{\text{IW}}^{(K)}[Q, R] &= \mathbb{E}_{Q_1\cdots Q_K}\left[\ln\frac{1}{K}\sum_{k=1}^{K} F'_k\right] && (60)\\
&\leq \ln \mathbb{E}_{Q_1\cdots Q_K}\left[\frac{1}{K}\sum_{k=1}^{K} F'_k\right] = \ln \mathbb{E}_{Q(z,B_\tau)}\left[F'(z, B_\tau)\right]. && (61)
\end{aligned}$$

The last term is further transformed as follows:

$$\begin{aligned}
\ln \mathbb{E}_{Q(z,B_\tau)}\left[F'(z,B_\tau)\right] &= \ln \mathbb{E}_{Q(z,B_\tau)}\left[\frac{P(Y,B_\tau|\tau(z))R(z|\tau(z))}{Q(B_\tau|\tau(z))Q(z)}\right] && (62)\\
&= \ln \mathbb{E}_{Q(z,B_\tau)}\left[\frac{P(Y,B_\tau|\tau(z))R(z|\tau(z))}{Q(z,B_\tau)}\right] && (63)\\
&= \ln \mathbb{E}_{Q(z)}\left[\frac{P(Y,\tau(z))R(z|\tau(z))}{Q(z)}\right] && (64)\\
&= \ln \sum_{\tau'\in\mathcal{T}} \mathbb{E}_{Q(z)}\left[\frac{P(Y,\tau')R(z|\tau')}{Q(z)}\mathbb{I}[\tau'=\tau(z)]\right] && (65)\\
&= \ln \sum_{\tau'\in\mathcal{T}} P(Y,\tau')\,\mathbb{E}_{R(z|\tau')}\left[\mathbb{I}[\tau'=\tau(z)]\right] && (66)\\
&\le \ln \sum_{\tau'\in\mathcal{T}} P(Y,\tau') = \ln P(Y), && (67)
\end{aligned}$$

where, in the transition from the first to the second row, we employed the following relation:

$$Q(B_\tau|\tau(z)) = \sum_{\tau\in\mathcal{T}} Q(B_\tau|\tau)\mathbb{I}[\tau=\tau(z)] = \sum_{\tau\in\mathcal{T}} Q(B_\tau|\tau)Q(\tau|z) = Q(B_\tau|z), \tag{68}$$

and we have used Proposition C.2 for the last inequality.

Finally, we will show that the following convergence property assuming that $F(z,B_\tau)$ is bounded:

$$\mathcal{L}^{(K)}_{\text{IW}}[Q,R] \to \ln\left(\sum_{\tau'\in\mathcal{T}} P(Y,\tau')\,\mathbb{E}_{R(z|\tau')}\,\mathbb{I}[\tau'=\tau(z)]\right) \qquad (K\to\infty). \tag{69}$$

From the strong law of large numbers, it follows that $\frac{1}{K}\sum_{k=1}^{K} F'_k$ converges to the following term almost surely as $K\to\infty$:

$$\mathbb{E}_{Q(z_k,B_k)}\left[F'(z_k,B_k)\right] = \sum_{\tau'\in\mathcal{T}} P(Y,\tau')\,\mathbb{E}_{R(z|\tau')}\,\mathbb{I}[\tau'=\tau(z)], \tag{70}$$

where we have employed the same transformations as used from equation (62) to (66). Observe that the *r.h.s* term of the equation (69) equals to $\ln P(Y)$ when $\forall\tau'\in\mathcal{T}, \mathbb{E}_{R(z|\tau')}[\mathbb{I}[\tau'=\tau(z)]]=1$, which completes the proof. □

Estimation of marginal log-likelihood For the estimation of $\ln P(Y)$, we employ $\mathcal{L}^{(K)}[Q,R]$ with $K=1,000$ similar to (Zhang, 2023). From Theorem C.3, IW-ELBO $\mathcal{L}^{(K)}[Q,R]$ is at least a better lower bound of $\ln P(Y)$ than ELBO $\mathcal{L}[Q,R]$, and converges to $\ln P(Y)$ when $\forall\tau, \mathbb{E}_{R(z|\tau)}[\mathbb{I}[\tau=\tau(z)]]=1$. According to Proposition C.2, this equality condition is satisfied when we set $R(z|\tau)=Q(z|\tau)$, which is approached by maximizing $\mathcal{L}[Q,R]$ with respect to R as indicated in Proposition 3.1.

C.8. Gradient estimators for IW-ELBO

The gradient of IW-ELBO $\mathcal{L}^{(K)}_{\text{IW}}[Q_{\theta,\phi},R_\psi]$ with respect to θ is given by

$$\begin{aligned}
\nabla_\theta \mathcal{L}^{(K)}_{\text{IW}} = {} & \mathbb{E}_{Q_{\theta,\phi}(z^{(1)},B^{(1)}_\tau)\cdots Q_{\theta,\phi}(z^{(K)},B^{(K)}_\tau)}\left[\sum_{k=1}^{K} w_k(z^{(1:K)},B^{(1:K)})\nabla_\theta \ln F'(z^{(k)},B^{(k)}_\tau)\right]\\
& + \mathbb{E}_{Q_{\theta,\phi}(z^{(1)},B^{(1)}_\tau)\cdots Q_{\theta,\phi}(z^{(K)},B^{(K)}_\tau)}\left[\sum_{k=1}^{K} \nabla_\theta \ln Q_\theta(z^{(k)})\cdot \ell(z^{(1:K)},B^{(1:K)})\right], \qquad (71)
\end{aligned}$$

where we have defined

$$w_k(z^{(1:K)}, B_\tau^{(1:K)}) := \frac{F'(z^{(k)}, B_\tau^{(k)})}{\sum_{k'=1}^{K} F'(z^{(k')}, B_\tau^{(k')})}, \tag{72}$$

$$\ell(z^{(1:K)}, B_\tau^{(1:K)}) := \ln\left(\frac{1}{K}\sum_{k'=1}^{K} F'(z^{(k')}, B_\tau^{(k')})\right). \tag{73}$$

Similarly, as F' is differentiable with respect to B_τ, and $B_\tau = h_\phi(\epsilon_B, \tau)$ is differentiable with respect to ϕ, the gradient of $\mathcal{L}_{\mathrm{IW}}^{(K)}$ with respect to ϕ can be evaluated as follows:

$$\nabla_\phi \mathcal{L}_{\mathrm{IW}}^{(K)} = \mathbb{E}_{Q_\theta(z^{(1)})\cdots Q_\theta(z^{(K)})}\,\mathbb{E}_{p_B(\epsilon_B^{(1)})\cdots p_B(\epsilon_B^{(K)})}\left[\sum_{k=1}^{K} w_k(z^{(1:K)}, B_\tau^{(1:K)})\nabla_\phi \ln F'(z^{(k)}, h_\phi(\epsilon^{(k)}, \tau))\right]. \tag{74}$$

Since $\nabla_\theta \ln F'(z^{(k)}, B_\tau^{(k)}) = -\nabla_\theta \ln Q_\theta(z^{(k)})$ from equation (23), an unbiased estimator of the gradient $\nabla_\theta \mathcal{L}^{(K)}$ is given as follows:

$$\widehat{g}_{\mathrm{IW},\theta}^{(K)} := \sum_{k=1}^{K} \nabla_\theta \ln Q_\theta(z^{(k)}) \cdot \left[-w_k(z^{(1:K)}, B_\tau^{(1:K)}) + \ell(z^{(1:K)}, B_\tau^{(1:K)})\right]. \tag{75}$$

The remaining gradient estimators are given as follows:

$$\widehat{g}_{\mathrm{IW},\phi}^{(K)} = \sum_{k=1}^{K} w_k(z^{(1:K)}, B_\tau^{(1:K)})\nabla_\phi \ln F(z^{(k)}, h_\phi(\epsilon^{(k)}, \tau)), \tag{76}$$

$$\widehat{g}_{\mathrm{IW},\phi}^{(K)} = \sum_{k=1}^{K} w_k(z^{(1:K)}, B_\tau^{(1:K)})\nabla_\psi \ln F(z^{(k)}, h_\phi(\epsilon^{(k)}, \tau)). \tag{77}$$

In total, the target for computing auto-gradient for these gradients is given as follows:

$$\begin{aligned}\widehat{\mathcal{L}}_{\mathrm{IW}}^{\prime(K)} &= \sum_{k=1}^{K} \ln Q_\theta(z^{(k)}) \cdot \mathrm{detach}\left[-w_k(z^{(1:K)}, B_\tau^{(1:K)}) + \ell(z^{(1:K)}, B_\tau^{(1:K)})\right] \\ &+ \sum_{k=1}^{K} \mathrm{detach}[w_k(z^{(1:K)}, B_\tau^{(1:K)})] \ln F(z^{(k)}, h_\phi(\epsilon_B^{(k)}, \tau(z^{(k)}))).\end{aligned} \tag{78}$$

D. Experimental Details

D.1. Models and training

As a prior distribution of $P(\tau)$ and $P(B_\tau|\tau)$, we assumed a uniform distribution over all topologies, and an exponential distribution $\mathrm{Exp}(10)$ independent for all branches, respectively, as commonly used in the literature (Zhang & Matsen IV, 2019; Koptagel et al., 2022). For the neural network used in the parameterization of $Q_\phi(B_\tau|\tau)$, we employed edge convolutional operation (EDGE), which was well-performed architecture in (Zhang, 2023). For the stochastic gradient optimizations, we used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001.

D.2. Initialization of coordinates

We initialized the mean parameters of the tip coordinate distribution $Q_\theta(z)$ with the multi-dimensional scaling (MDS) algorithm when Q_θ was given as normal distributions. For Q_θ comprised of wrapped normal distributions, we used the hyperbolic MDS algorithm (hMDS) proposed in (Sala et al., 2018) for the initialization. For a distance matrix used for MDS and hMDS, we used the Hamming distance between each pair of the input sequences Y as similar to (Macaulay et al., 2023). For the scale parameters, we used 0.1 for all experiments. For $R_\psi(z)$, we used the same mean parameters as $Q_\theta(z)$ and 1.0 for the scale parameters.

D.3. Training process

For the training of GeoPhy, we continued the stochastic gradient descent process until a total of 1,000,000 Monte-Carlo (MC) tree samples were consumed. Specifically, if K MC-samples were used per step, we performed up to 1,000,000 / K steps. It is noteworthy that the number of MC samples equaled the number of likelihood evaluations (NLEs), which provided us with a basis for comparing convergence speed between different runs.In all of our experiments, we used Adam optimizer with an initial learning rate of 0.0001. The learning rate was then multiplied by 0.75 after every 200,000 steps. Similar to approaches taken by Zhang & Matsen IV (2019), we incorporated an annealing procedure during the initial consumption of 100,000 MC samples. Specifically, we replaced the likelihood function in the lower bound with $P(Y|B_\tau, \tau)^\beta$ and linearly increased the inverse temperature β from 0.001 to 1 throughout the iterations. Note that all the estimations of marginal log-likelihood (MLL) were performed with β set to 1.

D.4. Variational branch length distribuions

For the variational branch length distribution $Q_\phi(B_\tau|\tau)$, we followed an architecture of (Zhang, 2023); namely, each branch length independently followed a lognormal distribution which was parameterized with a graph neural network (GNN). Details are described in Appendix B.

D.5. LAX estimators

As input features of a surrogate function $s_\chi(z)$ used in the LAX estimators, we employed a flattened vector of coordinates $z \in \mathbb{R}^{N\times d}$ when z resides in Euclidean space. In cases where the coordinates were $z \in \mathbb{H}^d$, we first transformed z with a logarithm map $\log_{\mu^o} z \in T_{\mu^o}\mathbb{H}^d$, then omitted their constant value 0-th elements and subsequently flattened the result. We implemented a simple multi-layer perceptron (MLP) network with a single hidden layer of width $10Nd$ and a subsequent sigmoid linear unit (SiLU) activation function as the neural network to output $s_\chi(z)$.

D.6. Replication of MLL estimates with MrBayes SS

Given the observed discrepancies in marginal log-likelihood (MLL) estimates obtained with the MrBayes stepping-stone (SS) method between references (Zhang & Matsen IV, 2019) and (Koptagel et al., 2022), we replicated the MrBayes SS runs using MrBayes version 3.2.7a. The script we used is provided below.

```
BEGIN MRBAYES;
set autoclose=yes nowarn=yes Seed=123 Swapseed=123;
lset nst=1;
prset statefreqpr=fixed(equal);
prset brlenspr=Unconstrained:exp(10.0);
ss ngen=10000000 nruns=10 nchains=4 printfreq=1000 samplefreq=100 \
savebrlens=yes filename=mrbayes_ss_out;
END;
```

We incorporated the results in the row named Replication in Table 2, where the values aligned more closely with those found in (Zhang & Matsen IV, 2019). We deduced that the prior distribution used in (Koptagel et al., 2022) might have been set differently as the current default values of $\mathrm{brlenspr}$ are $\mathrm{Unconstrained} : \mathrm{GammaDir}(1.0, 0.100, 1.0, 1.0)$ [2], which deviates from the model assumption used for the benchmarks. We observed that the line $\mathrm{brlenspr}$ was not included in the code provided in Appendix F of (Koptagel et al., 2022). Having been able to replicate the results found in (Zhang & Matsen IV, 2019), we opted to use their values as a reference in Table 1.

D.7. Visualization of tree topologies

We visualized the sum of the probability densities for tip node distribution $\sum_{i=1}^{N} Q(z_i)$ in Fig. 2 by projecting a hyperbolic coordinate $z_i \in \mathbb{H}^d$ onto the Poincaré coordinates $\overline{z}_{ik} = z_{ik}/\left(1 + z_{i0}\right)$ $(k = 1, \ldots, d)$, then applying a transformation $\overline{z}_i \mapsto \overline{z}'_i = \tanh\left(a \left\|\overline{z}_i\right\|_2\right) \cdot \overline{z}_i / \left\|\overline{z}_i\right\|_2$ with $a = 2.1$ to close up the central region. To display the density Q in the new coordinates, the Jacobian term was also considered to evaluate the density $Q(\overline{z}'_i)$.

[2] https://github.com/NBISweden/MrBayes/blob/develop/doc/manual/Manual_MrBayes_v3.2.pdf

For the comparison of consensus tree topologies, we plotted the edges of the tree by connecting each of their end node coordinate pairs with a geodesic line. The coordinate in $\mathbb{H}^d$ of the i-th tip node was determined as the location parameter $\mu_i \in \mathbb{H}^d$ of the wrapped normal distribution $Q(z_i) = \mathcal{WN}(z_i; \mu_i, \Sigma_i)$. Let $\xi_u \in \mathbb{H}^d$ denotes the coordinate of an interior node u, we defined ξ_u by using the Lorentzian centroid operation $\mathcal{C}$ (**?**) as follows:

$$\xi_u := \mathcal{C}(\{c_s\}_{s \in \mathcal{S}_\tau(u)}, \{\nu_s\}_{s \in \mathcal{S}_\tau(u)}) = \frac{\widetilde{\xi}_u}{\sqrt{-\left\langle \widetilde{\xi}_u, \widetilde{\xi}_u \right\rangle_L}}, \tag{79}$$

where $\widetilde{\xi}_u := \sum_{s \in \mathcal{S}_\tau(u)} \nu_s c_s$ denote an unnormalized sum of weighted coordinates, $s \in \mathcal{S}_\tau(u)$ denote a subset of tip node indices partitioned by the interior node u in the tree topology τ, $c_s := \mathcal{C}(\{\mu_i\}_{i \in s}, \{1\}_{i \in s})$ denote the Lorentzian centroid of the tip nodes contained in the subset s, and $\nu_s = N - |s|$ denote the number of the tip nodes in the complement set of s where $|s|$ represents the number of tip nodes in the subset s. As an unrooted tree topology τ can be identified by the set of tip node partitions introduced by the interior nodes of τ, the same unrooted tree topologies give the same set of interior coordinates $\{\xi_u\}_{u \in V}$ according to equation (79).

E. Additional Results

E.1. Marginal log-likelihood estimates for eight datasets

We present more comprehensive results in Table 2, extending upon the data from Table 1. This table presents the marginal log-likelihood (MLL) estimates obtained with various GeoPhy configurations and other conventional methods for the datasets DS1-DS8 (Hedges et al., 1990; Garey et al., 1996; Yang & Yoder, 2003; Henk et al., 2003; Lakner et al., 2008; Zhang & Blackwell, 2001; Yoder & Yang, 2004; Rossman et al., 2001). Once again, GeoPhy demonstrates its superior performance, consistently outperforming CSMC-based approaches that do not require preselection of tree topologies across the majority of configurations and datasets. This reaffirms the stability and excellence of our approach. Additionally, we found that a $Q(z)$ configuration using a 4-dimensional wrapped normal distribution with a full covariance matrix was the most effective among the tested configurations.

Table 2. Extended results of Table 1 comparing the marginal log-likelihood (MLL) estimates with different approaches in eight benchmark datasets. The MLL values for MrBayes SS and VBPI-GNN were obtained from (Zhang, 2023), while CSMC, VCSMC, and ϕ-CSMSC are referenced from (Koptagel et al., 2022). We also included replicated results for MrBayes SS. The MLL values for our approach (GeoPhy) are presented for a variety of $Q(z)$ configurations, encompassing distribution types (normal $\mathcal{N}$ or wrapped normal $\mathcal{WN}$), embedding dimensions (2 or 4), and the covariance matrix (full or diagonal). Each result features various CVs: LAX with $K = 1$, LOO with $K = 3$ denoted as LOO(3), and a combination of LOO and LAX, denoted as LOO(3)+. The figures highlighted in bold represent the highest values obtained with GeoPhy and the three CSMC-based methods, all of which perform an approximate Bayesian inference without the preselection of topologies. We have underlined MLL estimates where GeoPhy outperformed the other CSMC-based methods.

Dataset	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
#Taxa (N)	27	29	36	41	50	50	59	64
#Sites (M)	1949	2520	1812	1137	378	1133	1824	1008
MrBayes SS	−7108.42	−26367.57	−33735.44	−13330.06	−8214.51	−6724.07	−37332.76	−8649.88
	(0.18)	(0.48)	(0.5)	(0.54)	(0.28)	(0.86)	(2.42)	(1.75)
(Replication)	−7107.81	−26366.45	−33732.79	−13328.40	−8209.17	−6721.54	−37331.85	−8646.18
	(0.25)	(0.40)	(0.63)	(0.48)	(0.46)	(0.77)	(3.08)	(1.19)
VBPI-GNN	−7108.41	−26367.73	−33735.12	−13329.94	−8214.64	−6724.37	−37332.04	−8650.65
	(0.14)	(0.07)	(0.09)	(0.19)	(0.38)	(0.4)	(0.26)	(0.45)
CSMC	−8306.76	−27884.37	−35381.01	−15019.21	−8940.62	−8029.51	–	−11013.57
	(166.27)	(226.6)	(218.18)	(100.61)	(46.44)	(83.67)	–	(113.49)
VCSMC	−9180.34	−28700.7	−37211.2	−17106.1	−9449.65	−9296.66	–	–
	(170.27)	(4892.67)	(397.97)	(362.74)	(2578.58)	(2046.7)	–	–
ϕ-CSMC	−7290.36	−30568.49	−33798.06	−13582.24	−8367.51	−7013.83	–	−9209.18
	(7.23)	(31.34)	(6.62)	(35.08)	(8.87)	(16.99)	–	(18.03)
$\mathcal{N}$(diag,2)	$\underline{-7126.79}$	$\underline{-26440.54}$	−33814.98	$\underline{-13356.21}$	$\underline{-8251.99}$	$\underline{-6747.15}$	$\underline{-37526.41}$	$\underline{-8727.93}$
	(10.06)	(28.78)	(20.31)	(9.48)	(9.43)	(15.21)	(66.28)	(43.33)
LOO(3)	$\underline{-7130.60}$	$\underline{-26375.10}$	$\underline{-33737.71}$	$\underline{-13345.55}$	$\underline{-8236.99}$	$\underline{-6747.46}$	$\underline{-37375.93}$	$\underline{-8716.61}$
	(10.66)	(11.75)	(2.32)	(3.26)	(6.29)	(6.90)	(28.86)	(26.32)
LOO(3)+	$\underline{-7128.36}$	$\underline{-26369.93}$	$\underline{-33735.91}$	$\underline{-13346.03}$	$\underline{-8236.13}$	$\underline{-6751.97}$	$\underline{-37430.82}$	$\underline{-8691.38}$
	(9.77)	(0.25)	(0.13)	(4.54)	(5.71)	(11.24)	(67.19)	(10.70)
$\mathcal{N}$(diag,4)	$\underline{-7123.95}$	$\underline{-26382.91}$	$\underline{-33762.45}$	$\underline{-13341.62}$	$\underline{-8241.07}$	$\underline{-6735.78}$	$\underline{-37396.04}$	$\underline{-8679.48}$
	(12.03)	(16.89)	(10.68)	(3.64)	(6.56)	(5.25)	(26.39)	(27.55)
LOO(3)	$\underline{-7120.88}$	$\underline{-26368.53}$	$\underline{-33736.04}$	$\underline{-13338.99}$	$\underline{-8238.16}$	$\underline{-6735.59}$	$\underline{-37357.86}$	$\underline{-8665.54}$
	(13.15)	(0.05)	(0.09)	(6.08)	(0.52)	(4.51)	(10.76)	(5.66)
LOO(3)+	$\underline{-7119.81}$	$\underline{-26368.49}$	$\underline{-33735.92}$	$\underline{-13339.79}$	$\underline{-8236.69}$	$\underline{-6736.74}$	$\underline{-37353.08}$	$\underline{-8665.99}$
	(11.71)	(0.10)	(0.15)	(4.55)	(4.70)	(3.55)	(16.97)	(7.53)
$\mathcal{N}$(full,2)	$\underline{-7129.70}$	$\underline{-26487.71}$	−33807.05	$\underline{-13353.30}$	$\underline{-8251.01}$	$\underline{-6750.00}$	$\underline{-37487.49}$	$\underline{-8736.81}$
	(6.14)	(54.79)	(22.97)	(5.92)	(10.34)	(11.91)	(50.43)	(52.38)
LOO(3)	$\underline{-7132.35}$	$\underline{-26391.00}$	$\underline{-33736.98}$	$\underline{-13347.17}$	$\underline{-8237.75}$	$\underline{-6752.46}$	$\underline{-37462.07}$	$\underline{-8684.38}$
	(6.89)	(11.88)	(1.89)	(7.77)	(6.08)	(8.64)	(54.40)	(7.98)
LOO(3)+	$\underline{-7122.76}$	$\underline{-26380.59}$	$\underline{-33736.93}$	$\underline{-13343.21}$	$\underline{-8239.96}$	$\underline{-6753.84}$	$\underline{-37419.02}$	$\underline{-8691.96}$
	(10.81)	(14.39)	(2.10)	(2.14)	(4.84)	(14.30)	(35.94)	(13.51)
$\mathcal{N}$(full,4)	$\underline{-7120.03}$	$\underline{-26378.55}$	$\underline{-33753.20}$	$\underline{-13342.27}$	$\underline{-8237.33}$	$\underline{-6734.51}$	$\underline{-37373.32}$	$\underline{-8662.53}$
	(11.92)	(11.05)	(3.03)	(2.71)	(5.41)	(1.95)	(10.08)	(4.58)
LOO(3)	$\underline{-7124.62}$	$\underline{-26368.49}$	$\underline{-33736.03}$	$\underline{-13337.74}$	$\underline{-8234.18}$	$\underline{-6734.49}$	$\underline{-37347.46}$	$\underline{-8666.63}$
	(12.33)	(0.13)	(0.16)	(1.71)	(6.11)	(3.14)	(11.93)	(7.86)
LOO(3)+	$\underline{-7123.37}$	$\underline{-26368.51}$	$\underline{-33735.99}$	$\underline{\mathbf{-13337.06}}$	$\underline{-8241.25}$	$\underline{-6734.63}$	$\underline{-37352.30}$	$\underline{-8666.39}$
	(11.28)	(0.09)	(0.05)	(1.45)	(8.15)	(2.18)	(12.32)	(7.54)
$\mathcal{WN}$(diag,2)	$\underline{-7126.89}$	$\underline{-26444.84}$	−33823.74	$\underline{-13358.16}$	$\underline{-8251.45}$	$\underline{-6745.60}$	$\underline{-37516.88}$	$\underline{-8719.44}$
	(10.06)	(27.91)	(15.62)	(9.79)	(9.72)	(8.36)	(69.88)	(60.54)
LOO(3)	$\underline{-7130.67}$	$\underline{-26380.41}$	$\underline{-33737.75}$	$\underline{-13346.94}$	$\underline{-8239.36}$	$\underline{-6741.63}$	$\underline{-37382.28}$	$\underline{-8690.41}$
	(10.67)	(14.40)	(2.48)	(4.25)	(4.62)	(3.23)	(31.96)	(15.92)
LOO(3)+	$\underline{-7128.40}$	$\underline{-26375.28}$	$\underline{-33736.91}$	$\underline{-13347.32}$	$\underline{-8235.41}$	$\underline{-6742.40}$	$\underline{-37411.28}$	$\underline{-8683.22}$
	(9.78)	(11.78)	(1.91)	(4.42)	(5.70)	(1.94)	(56.74)	(13.13)
$\mathcal{WN}$(diag,4)	$\underline{-7122.10}$	$\underline{-26381.84}$	$\underline{-33759.19}$	$\underline{-13342.81}$	$\underline{-8243.92}$	$\underline{\mathbf{-6733.38}}$	$\underline{-37369.36}$	$\underline{-8666.85}$
	(12.29)	(17.18)	(9.95)	(3.45)	(6.74)	(0.79)	(13.45)	(10.63)
LOO(3)	$\underline{-7120.94}$	$\underline{-26368.52}$	$\underline{-33735.98}$	$\underline{-13339.77}$	$\underline{-8236.42}$	$\underline{-6735.12}$	$\underline{\mathbf{-37341.92}}$	$\underline{-8673.15}$
	(13.11)	(0.03)	(0.08)	(3.84)	(3.63)	(2.52)	(9.15)	(0.97)
LOO(3)+	$\underline{-7125.78}$	$\underline{-26368.51}$	$\underline{-33736.00}$	$\underline{-13342.38}$	$\underline{-8235.03}$	$\underline{-6736.20}$	$\underline{-37345.80}$	$\underline{-8666.68}$
	(13.10)	(0.10)	(0.18)	(6.35)	(5.36)	(1.91)	(11.13)	(5.78)
$\mathcal{WN}$(full,2)	$\underline{-7124.63}$	$\underline{-26458.50}$	−33804.63	$\underline{-13358.16}$	$\underline{-8251.09}$	$\underline{-6748.78}$	$\underline{-37484.98}$	$\underline{-8717.27}$
	(8.09)	(32.42)	(20.76)	(1.20)	(7.46)	(7.38)	(34.91)	(28.49)
LOO(3)	$\underline{-7125.94}$	$\underline{-26391.02}$	$\underline{-33736.98}$	$\underline{-13344.13}$	$\underline{-8236.90}$	$\underline{-6753.86}$	$\underline{-37416.00}$	$\underline{-8684.90}$
	(13.07)	(11.89)	(1.96)	(0.22)	(5.13)	(10.68)	(3.12)	(12.81)
LOO(3)+	$\underline{-7115.19}$	$\underline{-26385.21}$	$\underline{-33736.97}$	$\underline{-13343.95}$	$\underline{-8239.55}$	$\underline{-6747.61}$	$\underline{-37431.76}$	$\underline{-8683.54}$
	(8.16)	(13.77)	(1.93)	(1.76)	(4.72)	(6.87)	(43.65)	(3.57)
$\mathcal{WN}$(full,4)	$\underline{\mathbf{-7111.55}}$	$\underline{-26379.48}$	$\underline{-33757.79}$	$\underline{-13342.71}$	$\underline{-8240.87}$	$\underline{-6735.14}$	$\underline{-37377.86}$	$\underline{-8663.51}$
	(0.07)	(11.60)	(8.07)	(1.61)	(9.80)	(2.64)	(29.48)	(6.85)
LOO(3)	$\underline{-7119.77}$	$\underline{\mathbf{-26368.44}}$	$\underline{-33736.01}$	$\underline{-13339.26}$	$\underline{-8234.06}$	$\underline{-6733.91}$	$\underline{-37350.77}$	$\underline{-8671.32}$
	(11.80)	(0.13)	(0.03)	(3.19)	(7.53)	(0.57)	(11.74)	(5.99)
LOO(3)+	$\underline{-7116.09}$	$\underline{-26368.54}$	$\underline{\mathbf{-33735.85}}$	$\underline{-13337.42}$	$\underline{\mathbf{-8233.89}}$	$\underline{-6735.90}$	$\underline{-37358.96}$	$\underline{\mathbf{-8660.48}}$
	(10.67)	(0.12)	(0.12)	(1.32)	(6.63)	(1.13)	(13.06)	(0.78)