# PhyloVAE: Unsupervised Learning of Phylogenetic Trees via Variational Autoencoders

Tianyu Xie  Harry Richman  Jiansi Gao  Frederick A. Matsen IV  Cheng Zhang

PEKING UNIVERSITY

Fred Hutch Cancer Center

## Representations of Phylogenetic Trees


Figure 1. An example of phylogenetic tree.

A **phylogenetic tree** is described by a bifurcating tree topology $\tau$ and the associated non-negative branch lengths $\boldsymbol{q}$.

Our interest is the following problem arising from the practice of phylogenetics:

> Given a collection of tree topologies $\mathcal{T} = \{(\tau^i, w^i)\}_{i=1}^M$, how can we find a **low-dimensional representation** for each tree topology $\tau^i$?
>
> - Distilling crucial information from phylogenetic trees
> - Clustering/visualization the diverse evolutionary solutions
> - Diagnosing multiple phylogenetic inference analyses
> - etc ...

## PhyloVAE: An Overview

Our training set is a collection of trees $\mathcal{T} = \{(\tau^i, w^i)\}_{i=1}^M$, where

- $\tau^i$ is a tree topology, coming from scientific findings, phylogenetic analysis results, etc.
- $w^i$ is the weight for the tree topology $\tau^i$ and $\sum_{i=1}^M w^i = 1$.

Our PhyloVAE comprises of:

- A **generative model** $p_{\boldsymbol{\theta}}(\tau|\boldsymbol{z})$, modeling the probability of generating $\tau$ given a latent variable $\boldsymbol{z}$. We assume Gaussian prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}_d, \boldsymbol{I}_d)$.
- A **inference model** $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\tau)$ that approximates the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\tau)$.

Our training objective is the multi-sample lower bound

$$\mathbb{E}_{p_{\text{data}}(\tau)} L_K(\tau; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{p_{\text{data}}(\tau)} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}^1|\tau)\cdots q_{\boldsymbol{\phi}}(\boldsymbol{z}^K|\tau)} \log\left(\frac{1}{K}\sum_{i=1}^K \frac{p_{\boldsymbol{\theta}}(\tau, \boldsymbol{z}^i)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}^i|\tau)}\right) \le \mathbb{E}_{p_{\text{data}}(\tau)} \log p_{\boldsymbol{\theta}}(\tau), \tag{1}$$

which can be efficiently optimized using the reparametrization trick.

## A Tree Encoding Mechanism

To model the generative model $p_{\boldsymbol{\theta}}(\tau|\boldsymbol{z})$, we aim to build an encoding map for each tree:

$$\boldsymbol{s}(\tau) : \tau \mapsto \boldsymbol{s} = (s_3, s_4, \ldots, s_{N-1})$$

where $\boldsymbol{s}$ records the position indexes of the newly added edges when growing the tree topologies [XZ23].

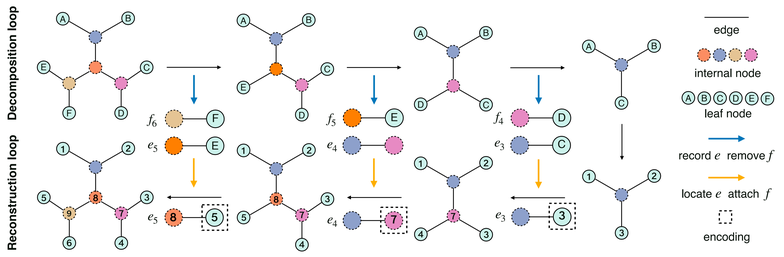


Figure 2. The decomposition loop and reconstruction loop for computing the encoding.

**Decomposition Loop** From the upper left, we

- remove the pendant edges $f_6, f_5, f_4$ (associated with the leaf nodes F, E, and D) sequentially;
- record the corresponding edge decisions $e_5, e_4, e_3$.

**Reconstruction Loop** From the lower right, we

- add back $f_4, f_5, f_6$ and index these nodes (except for the root) sequentially;
- record the indexes (3,7,5) associated with $e_5, e_4, e_3$.

**Theorem** Given a tree topology $\tau$ with $N$ leaf nodes, the time complexity of computing its encoding vector $\boldsymbol{s}(\tau)$ is $O(N)$.

## Generative Model and Inference Model in PhyloVAE

**Generative Model** Similar to the diagonal Gaussian distribution used in standard VAEs, we assume the elements in $\boldsymbol{s}(\tau)$ are conditionally independent given $\boldsymbol{z}$, i.e.,

$$p_{\boldsymbol{\theta}}(\tau|\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{s}(\tau)|\boldsymbol{z}) = \prod_{n=3}^{N-1} p_{\boldsymbol{\theta}}(s_n|\boldsymbol{z}). \tag{2}$$

The probability components are parametrized using MLPs.

**Inference Model** We use a diagonal normal distribution for the conditional distribution of the latent variable $\boldsymbol{z}$ whose mean and standard deviation are defined based on $\boldsymbol{f}_\tau$ as follows

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\tau) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_\tau, \text{diag}(\boldsymbol{\sigma}_\tau^2)\right), \ \boldsymbol{\mu}_\tau = \text{MLP}_\mu(\boldsymbol{f}_\tau), \ \log \boldsymbol{\sigma}_\tau = \text{MLP}_\sigma(\boldsymbol{f}_\tau), \tag{3}$$

where $\text{MLP}_\mu$ and $\text{MLP}_\sigma$ are two multi-layer perceptrons, and $\boldsymbol{f}_\tau$ is the outputs of a GNN that takes $\tau$ as input.

**Representation** The mean of the inference model, $\boldsymbol{\mu}_\tau \in \mathbb{R}^d$, is a deterministic low-dimensional representation of $\tau$.

## Representation Learning Task

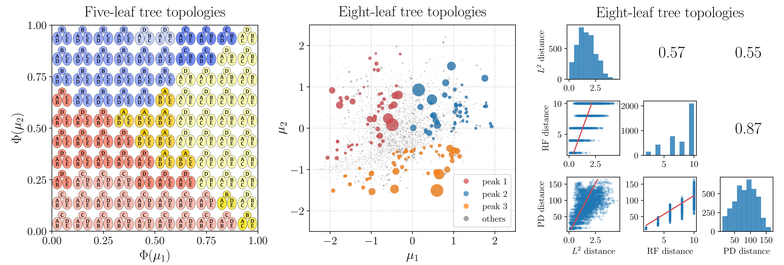


Figure 3. Performance of PhyloVAE for structural representation on small simulated data sets.

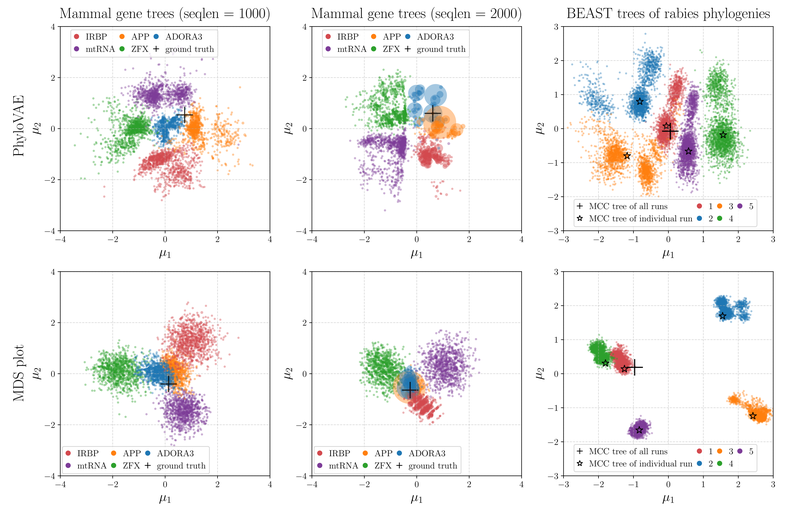


Figure 4. Performances of PhyloVAE and MDS plot [HY13] for representing **clusters of different gene trees** and **different phylogenetic inference analyses**.

## Generative Modeling Task

Table 1. KL divergences to the ground truth of different methods across eight benchmark data sets. "PhyloVAE ($d$)" means PhyloVAE with latent dimension $d$. The "# Training set" and "# Ground truth" columns show the number of unique tree topologies in the training sets and ground truth respectively.

| Sequence set | # Leaves | # Training set | # Ground truth | KL divergence to ground truth | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | SBN-EM | SBN-EM-$\alpha$ | ARTree | PhyloVAE (2) | PhyloVAE (10) |
| DS1 | 27 | 1228 | 2784 | 0.0136 | 0.0130 | **0.0045** | 0.0273 | 0.0189 |
| DS2 | 29 | 7 | 42 | 0.0199 | 0.0128 | **0.0097** | 0.0100 | 0.0098 |
| DS3 | 36 | 43 | 351 | 0.1243 | 0.0882 | 0.0548 | 0.0529 | **0.0477** |
| DS4 | 41 | 828 | 11505 | 0.0763 | 0.0637 | **0.0299** | 0.0619 | 0.0469 |
| DS5 | 50 | 33752 | 1516877 | 0.8599 | 0.8218 | 0.6266 | 0.7985 | **0.5744** |
| DS6 | 50 | 35407 | 809765 | 0.3016 | 0.2786 | 0.2360 | 0.3241 | **0.2207** |
| DS7 | 59 | 1125 | 11525 | 0.0483 | 0.0399 | **0.0191** | 0.0591 | 0.0370 |
| DS8 | 64 | 3067 | 82162 | 0.1415 | 0.1236 | **0.0741** | 0.1372 | 0.1061 |

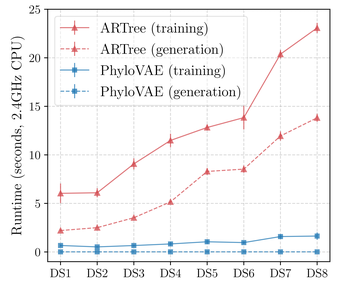


Figure 5. Runtime comparison between ARTree and PhyloVAE ($d = 10$) with $K = 32$ particles. Training means 10 training iterations. Generation means generating 100 tree topologies.

## References

[HY13] Robert M Hamer and Forrest W Young. *Multidimensional scaling: History, theory, and applications*. Psychology Press, 2013.

[XZ23] Tianyu Xie and Cheng Zhang. ARTree: A deep autoregressive model for phylogenetic inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Zha23] Cheng Zhang. Learnable topological features for phylogenetic inference via graph neural networks. In *International Conference on Learning Representations*, 2023.