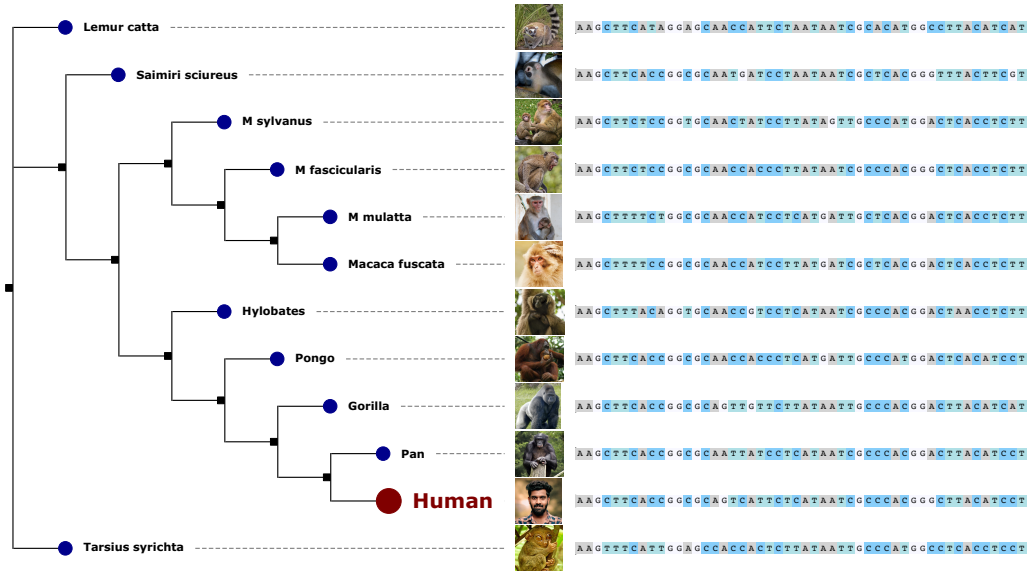
ARTree: A Deep Autoregressive Model for Phylogenetic Inference

Tianyu Xie[1], **Cheng Zhang**[2,*]

[1]School of Mathematical Sciences, Peking University
[2]School of Mathematical Sciences and Center for Statistical Science, Peking University
[*]Corresponding Author

Phylogenetic Trees



Phylogenetic Trees

Leaf nodes	$\Longleftrightarrow$	Observed species
Internal nodes	$\Longleftrightarrow$	Unobserved ancestor species
Branch length	$\Longleftrightarrow$	Evolutionary time between two species

Phylogenetic Trees

Leaf nodes	$\Longleftrightarrow$	Observed species
Internal nodes	$\Longleftrightarrow$	Unobserved ancestor species
Branch length	$\Longleftrightarrow$	Evolutionary time between two species

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

Phylogenetic Trees

Leaf nodes	$\Longleftrightarrow$	Observed species
Internal nodes	$\Longleftrightarrow$	Unobserved ancestor species
Branch length	$\Longleftrightarrow$	Evolutionary time between two species

- A phylogenetic tree is described by a bifurcating tree topology τ and the associated non-negative branch lengths $\boldsymbol{q}$.
- $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\} \in \Omega^{N \times M}$ are the observed sequences (with characters in Ω) of length M over N species. (e.g. $\Omega = \{A, C, G, T\}$ contain the nucleotides.)

Phylogenetic Trees

Leaf nodes	$\Longleftrightarrow$	Observed species
Internal nodes	$\Longleftrightarrow$	Unobserved ancestor species
Branch length	$\Longleftrightarrow$	Evolutionary time between two species

- A phylogenetic tree is described by a bifurcating tree topology τ and the associated non-negative branch lengths $\boldsymbol{q}$.
- $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\} \in \Omega^{N \times M}$ are the observed sequences (with characters in Ω) of length M over N species. (e.g. $\Omega = \{A, C, G, T\}$ contain the nucleotides.)
- $P(\boldsymbol{Y}|\tau, \boldsymbol{q})$ follows a continuous-time Markov chain.

Bayesian Phylogenetic Inference

- The core question in phylogenetic inference is:

> Given the biological sequences $\boldsymbol{Y}$ of observed species, what are the underlying phylogenetic trees?

Bayesian Phylogenetic Inference

- The core question in phylogenetic inference is:

> Given the biological sequences $\boldsymbol{Y}$ of observed species, what are the underlying phylogenetic trees?

- Now, in a Bayesian framework: (i) The likelihood function is $P(\boldsymbol{Y}|\tau, \boldsymbol{q})$. (ii) The assumed prior distribution is $P(\tau, \boldsymbol{q})$.
  Then the above question turns into:

> How can we infer the posterior distribution:
>
> $$P(\tau, \boldsymbol{q}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\tau, \boldsymbol{q})P(\tau, \boldsymbol{q})}{p(\boldsymbol{Y})} \propto P(\boldsymbol{Y}|\tau, \boldsymbol{q})P(\tau, \boldsymbol{q})$$

Variational Bayesian Phylogenetic Inference

[Zhang and Matsen IV, 2019]

- Challenges for Bayesian phylogenetic inference:
  - Combinatorially explosive size ($(2n-5)!!$) of the tree topology space.
  - The composite structure of discrete (tree topology) and continuous (branch length) components

Variational Bayesian Phylogenetic Inference

[Zhang and Matsen IV, 2019]

- Challenges for Bayesian phylogenetic inference:
  - Combinatorially explosive size ($(2n-5)!!$) of the tree topology space.
  - The composite structure of discrete (tree topology) and continuous (branch length) components
- Variational family:

$$Q_{\phi,\psi}(\tau, \boldsymbol{q}) = \overbrace{Q_{\psi}(\boldsymbol{q}|\tau)}^{\text{branch length}} \cdot \overbrace{Q_{\phi}(\tau)}^{\text{tree topology}}$$

Variational Bayesian Phylogenetic Inference

[Zhang and Matsen IV, 2019]

- Challenges for Bayesian phylogenetic inference:
  - Combinatorially explosive size ($(2n-5)!!$) of the tree topology space.
  - The composite structure of discrete (tree topology) and continuous (branch length) components
- Variational family:

$$Q_{\phi,\psi}(\tau, \boldsymbol{q}) = \overbrace{Q_{\psi}(\boldsymbol{q}|\tau)}^{\text{branch length}} \cdot \overbrace{Q_{\phi}(\tau)}^{\text{tree topology}}$$

- Multi-sample lower bound:

$$L^K(\phi, \psi) = \mathbb{E}_{\{(\tau^i, \boldsymbol{q}^i)\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} Q_{\phi,\psi}} \log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{P(\boldsymbol{Y}|\tau^i, \boldsymbol{q}^i) P(\tau^i, \boldsymbol{q}^i)}{Q_{\phi}(\tau^i) Q_{\psi}(\boldsymbol{q}^i|\tau^i)} \right). \tag{1}$$

ARTree: Overview

- We propose ARTree, a deep autoregressive model for phylogenetic inference enjoys:
  - **Unconstrained support**;
  - **Discarding heuristic features**.

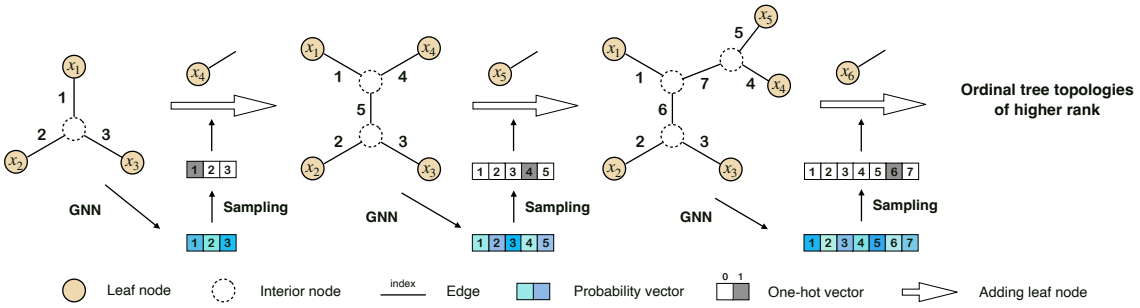
ARTree: Overview

- We propose ARTree, a deep autoregressive model for phylogenetic inference enjoys:
  - **Unconstrained support**;
  - **Discarding heuristic features**.
- Notations:
  - $\tau_n = (V_n, E_n)$: tree topology with n leaf nodes
  - V_n, E_n: nodes and edges of τ_n.
  - $\mathcal{X} = \{x_1, \ldots, x_N\}$: a pre-selected order for the leaf nodes.

ARTree: Overview

- We propose ARTree, a deep autoregressive model for phylogenetic inference enjoys:
  - **Unconstrained support**;
  - **Discarding heuristic features**.
- Notations:
  - $\tau_n = (V_n, E_n)$: tree topology with n leaf nodes
  - V_n, E_n: nodes and edges of τ_n.
  - $\mathcal{X} = \{x_1, \ldots, x_N\}$: a pre-selected order for the leaf nodes.
- We consider unrooted tree topologies in ARTree.

ARTree: Overview



ARTree: Sequential Generating Process

- During the generating process, the selected edges at each time step form a decision sequence $D = (e_3, \ldots, e_{N-1})$.

ARTree: Sequential Generating Process

- During the generating process, the selected edges at each time step form a decision sequence $D = (e_3, \ldots, e_{N-1})$.
- It can be proved that the generative process $g : \mathcal{D} \to \mathcal{T}$ is bijective. We call g^{-1} decomposition process.

ARTree: Sequential Generating Process

- During the generating process, the selected edges at each time step form a decision sequence $D = (e_3, \ldots, e_{N-1})$.
- It can be proved that the generative process $g : \mathcal{D} \to \mathcal{T}$ is bijective. We call g^{-1} decomposition process.
- Decompose $Q(D)$ as the product of conditional distributions:

$$Q(\tau) = Q(D) = \prod_{n=3}^{N-1} Q(e_n | e_3, \ldots, e_{n-1}). \tag{2}$$

ARTree: Sequential Generating Process

- During the generating process, the selected edges at each time step form a decision sequence $D = (e_3, \ldots, e_{N-1})$.
- It can be proved that the generative process $g : \mathcal{D} \to \mathcal{T}$ is bijective. We call g^{-1} decomposition process.
- Decompose $Q(D)$ as the product of conditional distributions:

$$Q(\tau) = Q(D) = \prod_{n=3}^{N-1} Q(e_n | e_3, \ldots, e_{n-1}). \tag{2}$$

- We use graph neural networks to parametrize $Q(e_n | e_{<n})$.

Topological node embeddings

- First find the node embeddings of $\tau_n = (V_n, E_n)$, which is a set $\{f_n(u) \in \mathbb{R}^N : u \in V_n\}$.
- For **leaf nodes**, one-hot encoding :

$$[f_n(x_i)]_j = \delta_{ij}, 1 \leq i \leq n, 1 \leq j \leq N,$$

  where δ is Kronecker delta function.
- For **interior nodes**, minimizing the Dirichlet energy

$$\ell(f_n, \tau_n) := \sum_{(u,v)\in E_n} ||f_n(u) - f_n(v)||^2$$

  using the efficient two-pass algorithm described in [Zhang, 2023].

Message passing networks

- Initialized as topological node embeddings, the node features are updated with the information from their neighborhoods in a convolutional manner [Gilmer et al., 2017].
- l-th round message passing (L round in total):

$$m_n^l(u,v) = F_{\text{message}}^l(f_n^l(u), f_n^l(v)),$$
$$f_n^{l+1}(v) = F_{\text{updating}}^l\left(\{m_n^l(u,v); u \in \mathcal{N}(v)\}\right),$$

  where $\mathcal{N}(v)$ is the neighborhood of the node v.
- In our implementations, the choices of F_{message}^l and F_{updating}^l follow the edge convolution operator [Wang et al., 2018].

Node hidden states

- The conditional distribution $Q(\cdot|e_{<n})$ has to capture the information from all the previous tree topologies.
- After obtaining the final node features of $\{f_n^L(v)\}$, a gated recurrent unit (GRU) [Cho et al., 2014] follows, i.e.

$$h_n(v) = \mathrm{GRU}(h_{n-1}(v), f_n^L(v)),$$

where $h_n(v)$ is the hidden state of v at the n-th generation step and is initialized to zero.

Time Guided Readout

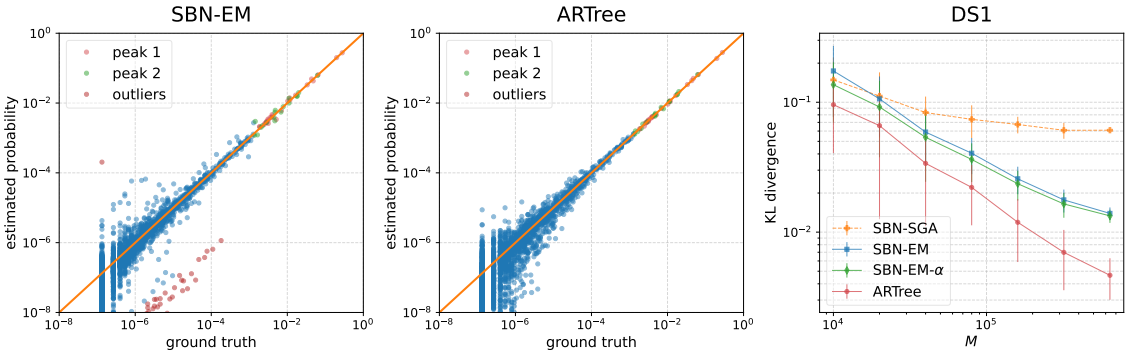
- A main difference from other graph autoregressive models: **the topological node embedding $f_n^0(v)$ depends the time step n.**
- Time guided readout step:

$$p_n(e) = F_{\text{pooling}}\left(h_n(u) + b_n, h_n(v) + b_n\right),$$
$$r_n(e) = F_{\text{readout}}\left(p_n(e) + b_n\right),$$

  where b_n is the sinusoidal positional embedding [Vaswani et al., 2017] of time step n.
- Edge decision probability:

$$Q(\cdot|e_{<n}) \sim \text{Discrete}\left(q_n\right), \quad q_n = \text{softmax}\left(\{r_n(e)\}_{e \in E_n}\right),$$

Experiments: Tree Topology Density Estimation



Given a training data set $\mathcal{M} = \{\tau_m\}_{m=1}^{M}$, we train ARTree via **maximum likelihood estimation**. In each iteration, the stochastic gradient is obtained by $\nabla_{\phi} L(\phi; \mathcal{M}) = \frac{1}{B} \sum_{b=1}^{B} \nabla_{\phi} \log Q_{\phi}(\tau_{m_b})$, where a minibatch $\{\tau_{m_b}\}_{b=1}^{B}$ is randomly sampled from $\mathcal{M}$.

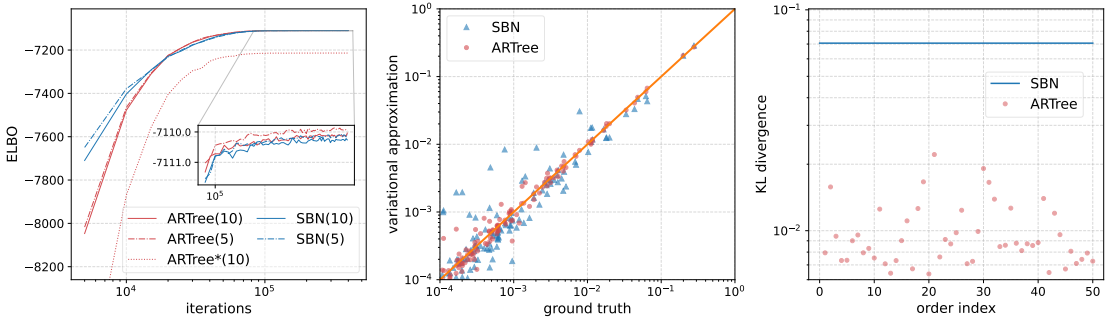
Experiments: Tree Topology Density Estimation

Table 1: KL divergences to the ground truth of different methods across 8 benchmark data sets. Sampled trees column shows the numbers of unique tree topologies in the training sets formed by MrBayes runs. The results are averaged over 10 replicates.

Data set	#Taxa	#Sites	Sampled trees	KL divergence to ground truth			
				SBN-EM	SBN-EM-α	SBN-SGA	ARTree
DS1	27	1949	1228	0.0136	0.0130	0.0504	**0.0045**
DS2	29	2520	7	0.0199	0.0128	0.0118	**0.0097**
DS3	36	1812	43	0.1243	0.0882	0.0922	**0.0548**
DS4	41	1137	828	0.0763	0.0637	0.0739	**0.0299**
DS5	50	378	33752	0.8599	0.8218	0.8044	**0.6266**
DS6	50	1133	35407	0.3016	0.2786	0.2674	**0.2360**
DS7	59	1824	1125	0.0483	0.0399	0.0301	**0.0191**
DS8	64	1008	3067	0.1415	0.1236	0.1177	**0.0741**

Experiments: Variational Bayesian Phylogenetic Inference

For both ARTree and SBN, the collaborative branch lengths are parametrized using learnable topological features with GNNs [Zhang, 2023]. VBPI is done by maximizing the multi-sample lower bound with $K = 10$.



Results: VBPI on DS1. **Left**: The number of particles K are in the brackets. The ARTree^* method refers to ARTree without time guidance ($b_n = 0$). **Right**: KL divergences across 50 random taxa orders.

Experiments: Variational Bayesian Phylogenetic Inference

Table 2: Results: VBPI on 8 benchmarks (KL, ELBO, 10-sample lower bound (LB-10), and marginal likelihood (ML)). GT trees row shows the number of unique tree topologies in the ground truth. The ML estimates are obtained via importance sampling using 1000 samples. The results of ϕ-CSMC are from [Koptagel et al., 2022].

	Data set	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
	# Taxa	27	29	36	41	50	50	59	64
	# Sites	1949	2520	1812	1137	378	1133	1824	1008
	GT trees	2784	42	351	11505	1516877	809765	11525	82162
KL	SBN	0.0707	0.0144	0.0554	0.0739	1.2472	0.3795	0.1531	0.3173
	ARTree	**0.0097**	**0.0004**	**0.0064**	**0.0219**	**0.8979**	**0.2216**	**0.0123**	**0.1231**
ELBO	SBN	-7110.24(0.03)	-26368.88(0.03)	-33736.22(0.02)	-13331.83(0.03)	-8217.80(0.04)	**-6728.65(0.06)**	-37334.85(0.04)	-8655.05(0.05)
	ARTree	**-7110.09(0.04)**	**-26368.78(0.07)**	**-33736.17(0.08)**	**-13331.82(0.05)**	**-8217.68(0.04)**	**-6728.65(0.06)**	**-37334.84(0.13)**	**-8655.03(0.05)**
LB-10	SBN	-7108.69(0.02)	-26367.87(0.02)	-33735.26(0.02)	-13330.29(0.02)	-8215.42(0.04)	**-6725.33(0.04)**	-37332.58(0.03)	-8651.78(0.04)
	ARTree	**-7108.68(0.02)**	**-26367.86(0.02)**	**-33735.25(0.02)**	**-13330.27(0.03)**	**-8215.34(0.03)**	**-6725.33(0.04)**	**-37332.54(0.03)**	**-8651.73(0.04)**
ML	ϕ-CSMC	-7290.36(7.23)	-30568.49(31.34)	-33798.06(6.62)	-13582.24(35.08)	-8367.51(8.87)	-7013.83(16.99)	N/A	-9209.18(18.03)
	SBN	**-7108.41(0.15)**	-26367.71(0.08)	**-33735.09(0.09)**	-13329.94(0.20)	-8214.62(0.40)	**-6724.37(0.43)**	-37331.97(0.28)	-8650.64(0.50)
	ARTree	-7108.41(0.19)	**-26367.71(0.07)**	**-33735.09(0.09)**	**-13329.94(0.17)**	**-8214.59(0.34)**	-6724.37(0.46)	**-37331.95(0.27)**	**-8650.61(0.48)**

References i

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014).
Learning phrase representations using RNN encoder-decoder for statistical machine translation.
arXiv preprint arXiv:1406.1078.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017).
Neural message passing for quantum chemistry.
ArXiv, abs/1704.01212.

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

References ii

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
  Attention is all you need.
  In *Advances in Neural Information Processing Systems*, volume 30.
- Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2018).
  Dynamic graph CNN for learning on point clouds.
  ACM Transactions on Graphics (TOG), 38:1 – 12.
- Zhang, C. (2023).
  Learnable topological features for phylogenetic inference via graph neural networks.
  In *International Conference on Learning Representations*.

References iii


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

Thank you!