PhyloGFN: Phylogenetic inference with generative flow networks

Ming Yang Zhou[1], Zichao Yan[2], Elliot Layne[1,2], Nikolay Malkin[2,3], Dinghuai Zhang[2,3], Moksh Jain[2,3], Mathieu Blanchette[1], Yoshua Bengio[2,3,4]

[1]McGill University [2]Mila – Quebec AI Institute, [3]Universite de Montreal, ´ [4]CIFAR

Phylogenetic inference

Infer evolution history and relationship among a set of species

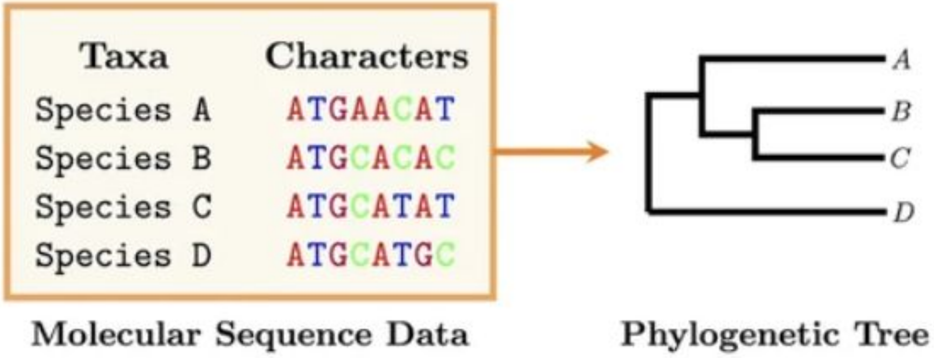
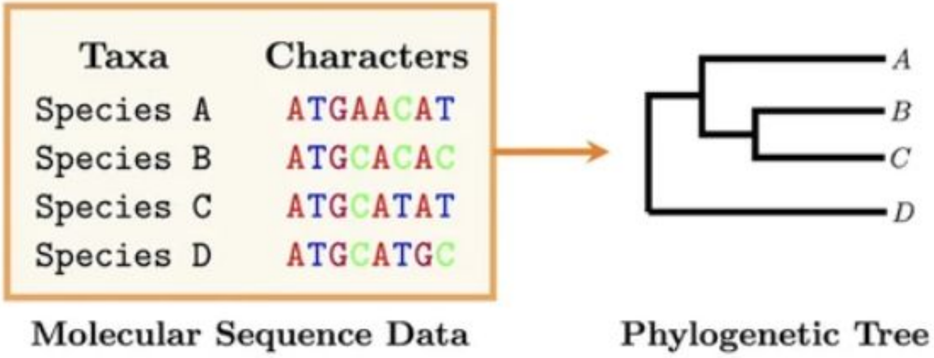


Image credit: Zhang Cheng Neurips 2020 presentation “Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows”
https://neurips.cc/virtual/2020/public/poster_d96409bf894217686ba124d7356686c9.html

Phylogenetic inference

Infer evolution history and relationship among a set of species



Input: multi sequence alignment of studied species

$$Y = \{y_1, y_2 \dots y_n\} \in \Sigma^{n \times m}$$

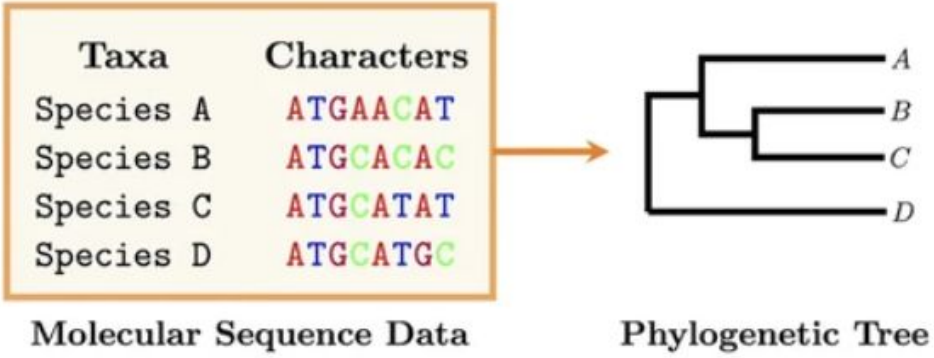
Σ: characters set

- {A,C,G,T} for DNA data
- {A,C,G,U} for RNA data

Image credit: Zhang Cheng Neurips 2020 presentation “Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows” https://neurips.cc/virtual/2020/public/poster_d96409bf894217686ba124d7356686c9.html

Phylogenetic inference

Infer evolution history and relationship among a set of species



Output: phylogenetic trees

- Leaves labeled by studied species
- Two components:
  - Tree topology z
  - Branch lengths b

Image credit: Zhang Cheng Neurips 2020 presentation "Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows"
https://neurips.cc/virtual/2020/public/poster_d96409bf894217686ba124d7356686c9.html

Bayesian phylogenetic inference

Given observed sequences Y, infer the posterior distribution of weighted phylogenetic trees (z,b)

Likelihood Prior

$$P(z,b|Y) = \frac{P(Y|z,b)P(z,b)}{P(Y)}$$

Posterior

Marginal

A pre-defined evolution model is employed to calculate likelihood and prior:

- Likelihood is calculated using Felsenstein's algorithm

Challenges:

- n species -> topology space size (2n-5)!!
- Discrete topology + continuous branch lengths

Bayesian phylogenetic inference: prior works

MCMC based algorithms:

- Popular softwares
  - MrBayes (Ronquist et al. 2012)
  - RevBayes (Höhna et al. 2016)
- Limited scalability to high dimensional distribution with multiple distanced modes.
  - n species -> (2n-5)!! tree topologies

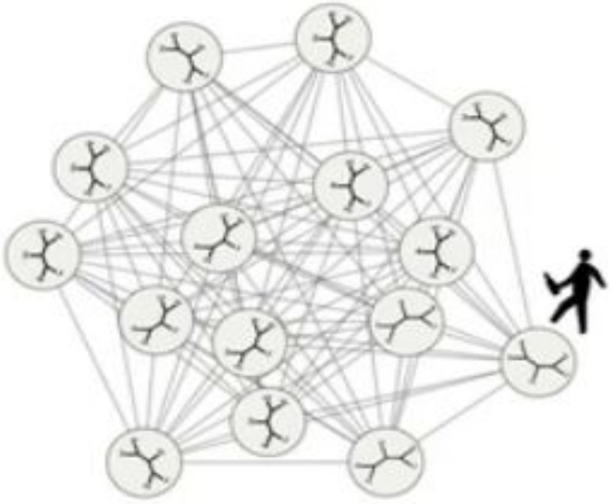
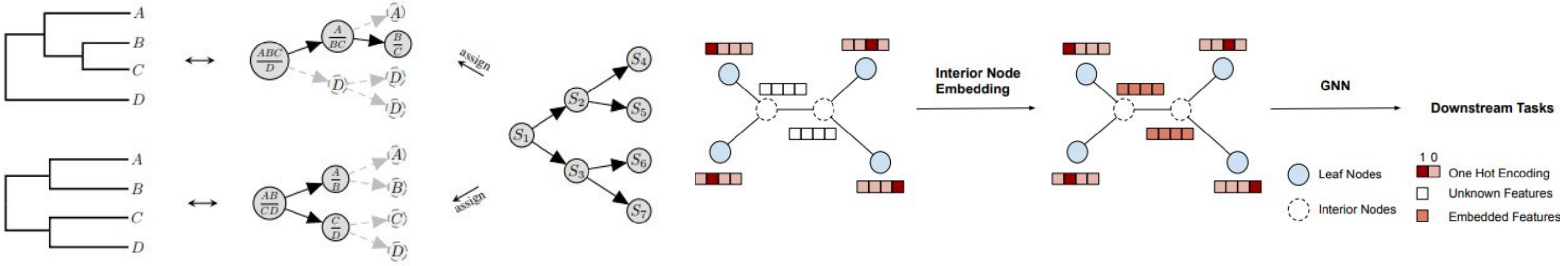


Image credit: Zhang Cheng Neurips 2020 presentation “Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows” https://neurips.cc/virtual/2020/public/poster_d96409bf894217686ba124d7356686c9.html

Bayesian phylogenetic inference: prior works

Variational Inference algorithms:

- VBPI (Zhang et al. 2018), VBPI-NF (Zhang 2020), VBPI-GNN (Zhang 2023)
  - Limited tree topology sampling Space

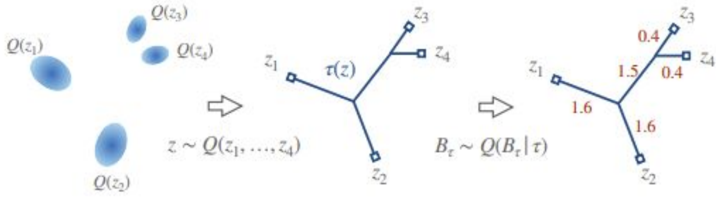


Zhang 2023

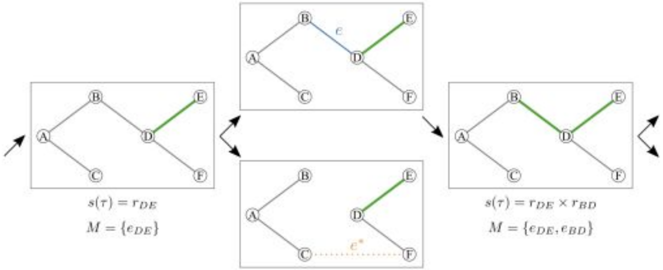
Bayesian phylogenetic inference: prior works

Variational Inference algorithms:

- VaiPhy (**koptagel et al., 2022**), GeoPhy (**Mimori & Hamada, 2023**)
  - Underperformance in marginal log likelihood (MLL) estimation



Mimori & Hamada, 2023



koptagel et al., 2022

Generative flow network (GFlowNet)

GFlowNet constructs object $x \in \mathcal{X}$ through a sequence of incremental actions based on a stochastic policy.

Construction procedures modeled by MPD

- Initial state s_0
- Terminal states $\mathcal{X}$
- Trajectory from s_0 to x represent a construction sequence of x.

Given a reward function $R(x) : \mathcal{X} \rightarrow \mathbb{R}^+$ GFlowNet learns a policy such that sampling probability $P_F^\top(z, b) \propto R(z, b)$

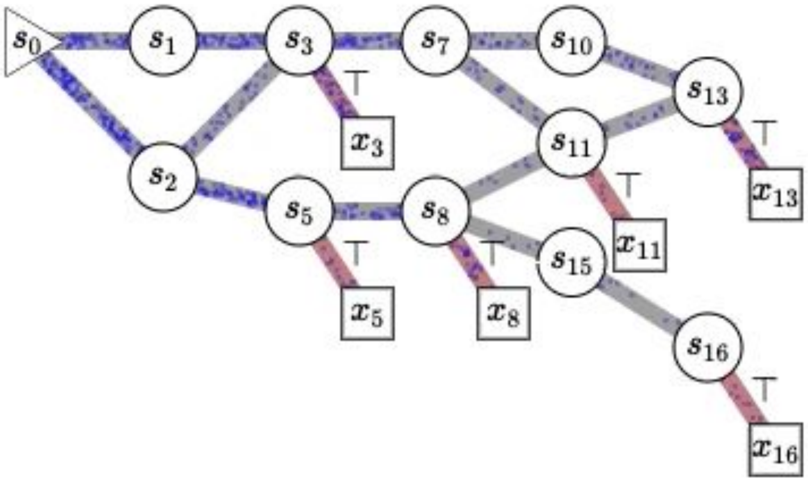


Image credit: Emmanuel bengio blog post "Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation" https://folinoid.com/w/gflownet/

Generative flow network (GFlowNet)

GFlowNet constructs object $x \in \mathcal{X}$ through a sequence of incremental actions based on a stochastic policy.

Construction procedures modeled by MPD

- Initial state s_0
- Terminal states $\mathcal{X}$
- Trajectory from s_0 to x represent a construction sequence of x.

Given a reward function $R(x) : \mathcal{X} \to \mathbb{R}^+$ GFlowNet learns a policy such that sampling probability $P_F^\top(z, b) \propto R(z, b)$

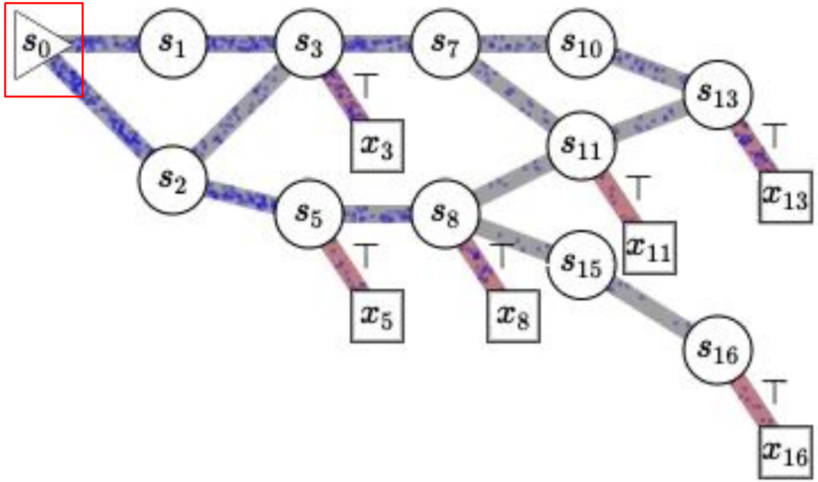


Image credit: Emmanuel bengio blog post "Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation" https://folinoid.com/w/gflownet/

Generative flow network (GFlowNet)

GFlowNet constructs object $x \in \mathcal{X}$ through a sequence of incremental actions based on a stochastic policy.

Construction procedures modeled by MPD

- Initial state s_0
- Terminal states $\mathcal{X}$
- Trajectory from s_0 to x represent a construction sequence of x.

Given a reward function $R(x) : \mathcal{X} \to \mathbb{R}^+$ GFlowNet learns a policy such that sampling probability $P_F^\top(z, b) \propto R(z, b)$

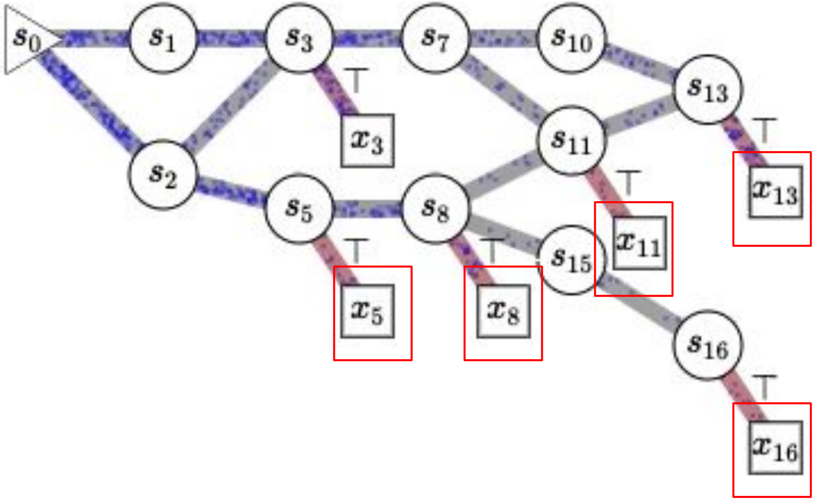


Image credit: Emmanuel bengio blog post "Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation" https://folinoid.com/w/gflownet/

Generative flow network (GFlowNet)

GFlowNet constructs object $x \in \mathcal{X}$ through a sequence of incremental actions based on a stochastic policy.

Construction procedures modeled by MPD

- Initial state s_0
- Terminal states $\mathcal{X}$
- Trajectory from s_0 to x represent a construction sequence of x.

Given a reward function $R(x) : \mathcal{X} \to \mathbb{R}^+$ GFlowNet learns a policy such that sampling probability $P_F^\top(z, b) \propto R(z, b)$

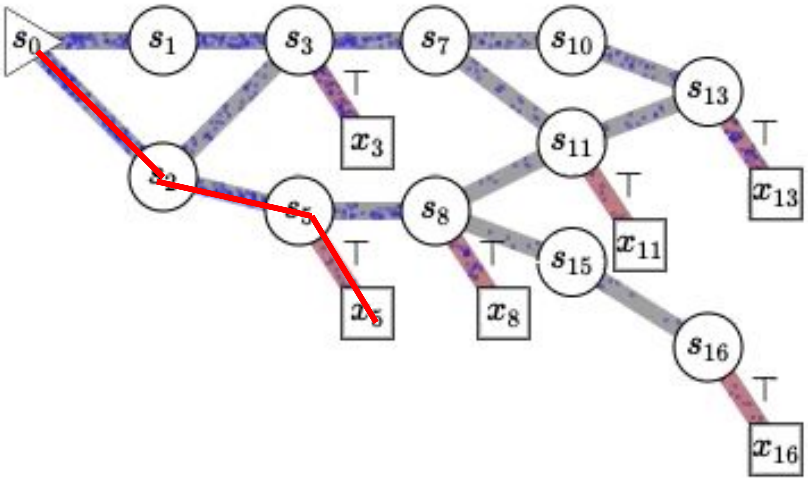


Image credit: Emmanuel bengio blog post "Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation" https://folinoid.com/w/gflownet/

PhyloGFN Bayesian: objective and reward function

Objective: given observed sequences Y, learn a GFlowNet over $\mathcal{X} = \{(z, b)\}$ such that: $P_F^{\top}(z, b) = P(Y|z, b)$

Reward function: $R(z, b) = P(Y|z, b)P(b)$

PhyloGFN Bayesian: objective and reward function

Objective: given observed sequences Y, learn a GFlowNet over $\mathcal{X} = \{(z, b)\}$ such that: $P_F^{\top}(z, b) = P(Y|z, b)$

Reward function: $R(z, b) = P(Y|z, b)P(b)$

- $P(z, b|Y) = R(z, b)\dfrac{P(z)}{P(Y)}$

PhyloGFN Bayesian: objective and reward function

Objective: given observed sequences Y, learn a GFlowNet over $\mathcal{X} = \{(z, b)\}$ such that: $P_F^\top(z, b) = P(Y|z, b)$

Reward function: $R(z, b) = P(Y|z, b)P(b)$

- $P(z, b|Y) = R(z, b)\boxed{\frac{P(z)}{P(Y)}}$ Constant

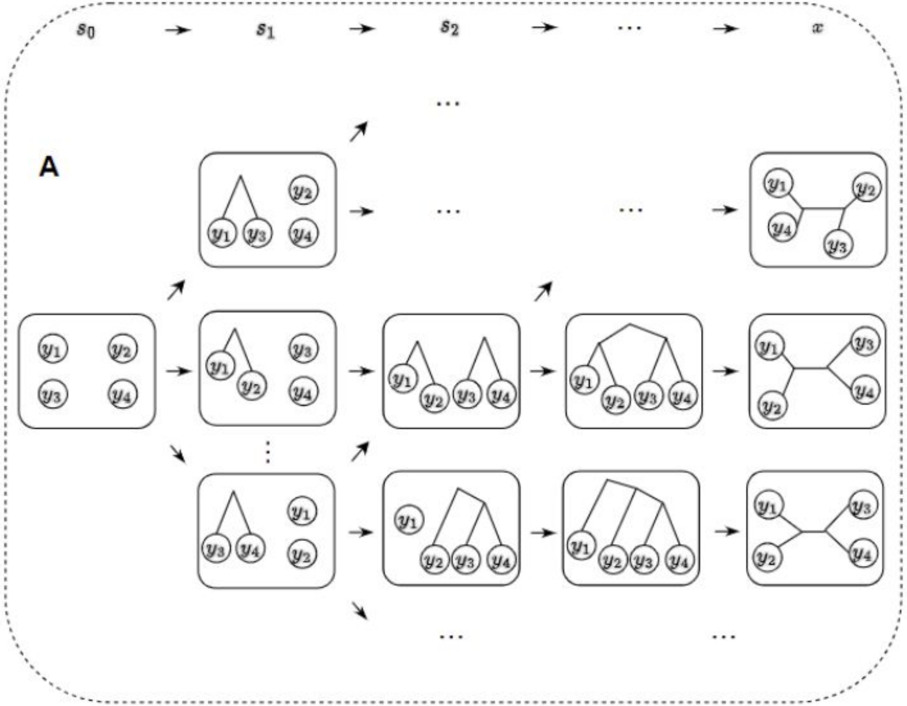
PhyloGFN Bayesian: objective and reward function

Objective: given observed sequences Y, learn a GFlowNet over $\mathcal{X} = \{(z, b)\}$ such that: $P_F^{\top}(z, b) = P(Y|z, b)$

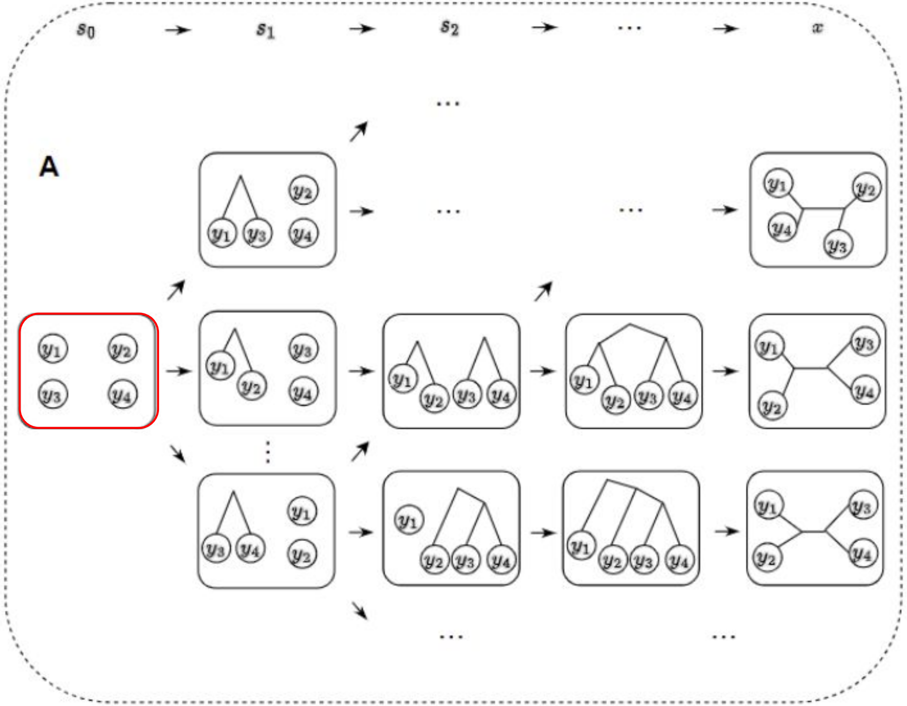
Reward function: $R(z, b) = P(Y|z, b)P(b)$

- $P(z, b|Y) = R(z, b)\boxed{\frac{P(z)}{P(Y)}}$ Constant
  - $P_F^{\top}(z, b) \propto R(z, b) \implies P_F^{\top}(z, b) = P(Y|z, b)$

PhyloGFN: phylogenetic trees construction



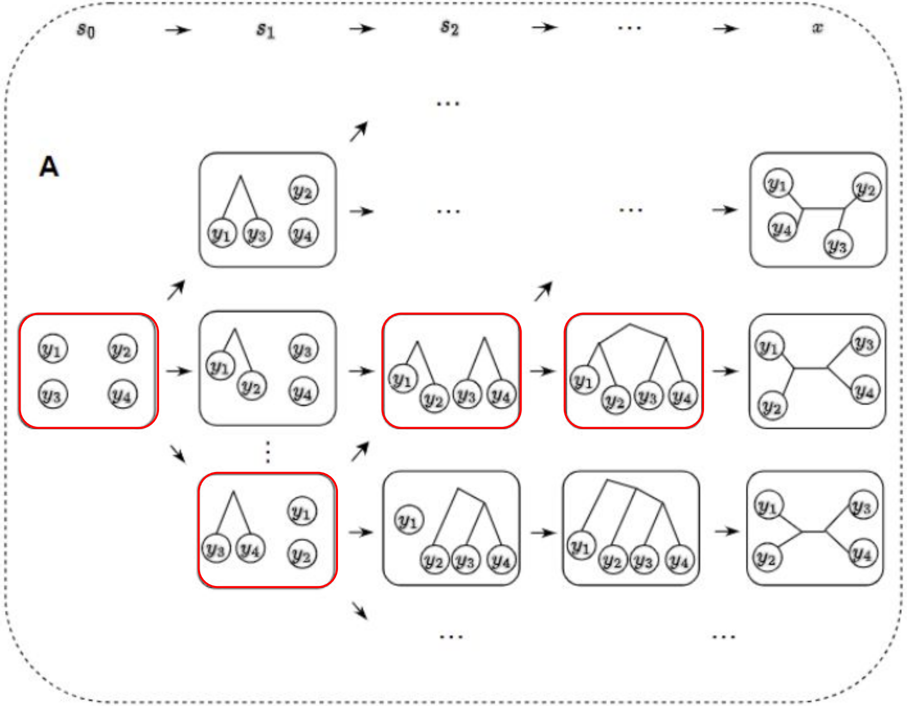
PhyloGFN: phylogenetic trees construction



Sequential construction:

- Initialize with the set of sequences as a forest of rooted trees

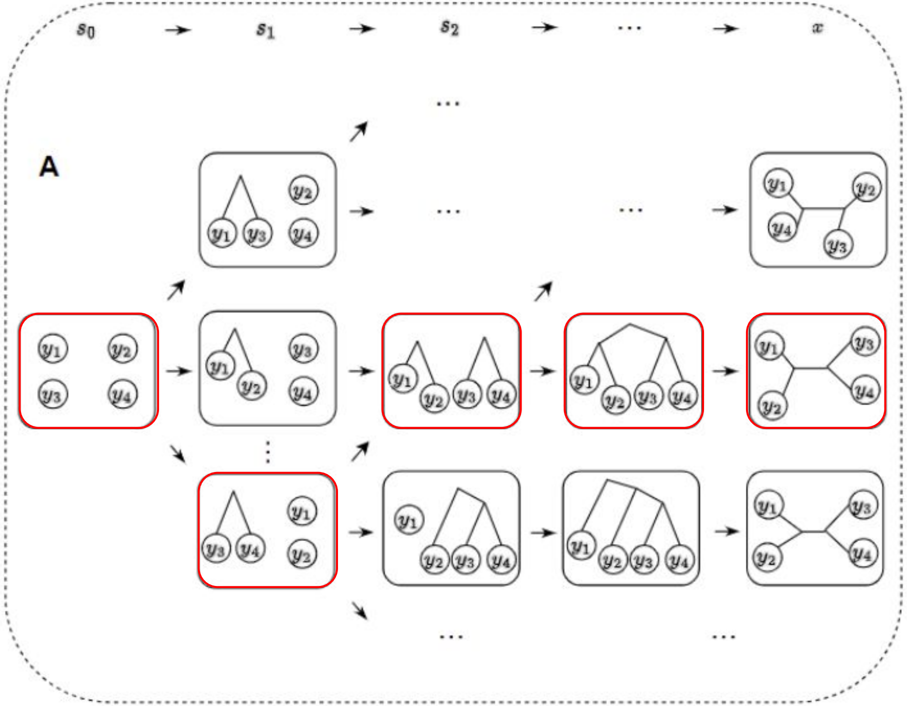
PhyloGFN: phylogenetic trees construction



Sequential construction:

- Initialize with the set of sequences as a forest of rooted trees
- Iteratively joining pair of trees until a full tree is constructed.

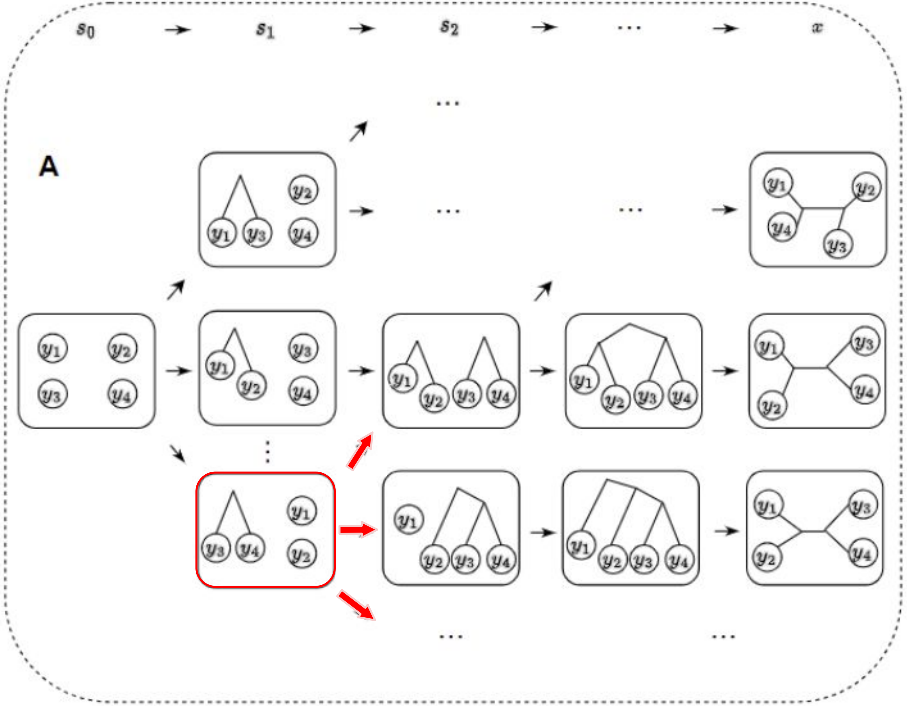
PhyloGFN: phylogenetic trees construction



Sequential construction:

- Initialize with the set of sequences as a forest of rooted trees
- Iteratively joining pair of trees until a full tree is constructed.
- Remove root node at the end if we infer unrooted trees

PhyloGFN: phylogenetic trees construction



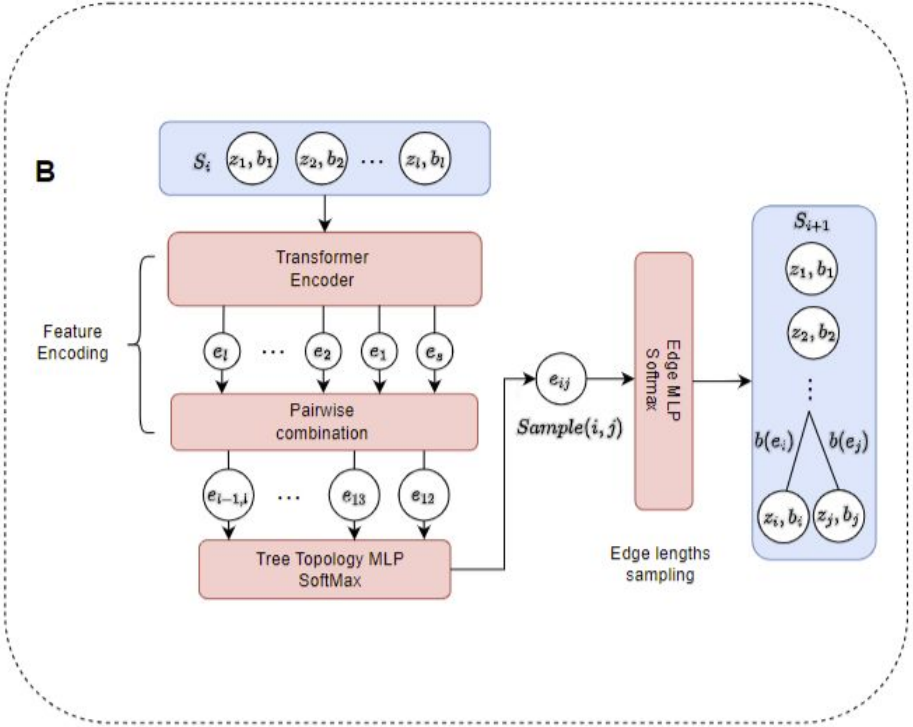
Sequential construction:

- Initialize with the set of sequences as a forest of rooted trees
- Iteratively joining pair of trees until a full tree is constructed.
- Remove root node at the end if we infer unrooted trees

Two steps action:

- Choose a pair of trees to join
- Generate branch lengths on new edges

PhyloGFN: Forward policy model



Transformer based architecture

- order-equivariant model

For an n trees state, generate probability logits for $\binom{n}{2}$ pairs of trees

Branch lengths modeling:

- Discrete: multinomial distribution of fixed bin size
- Continuous: mixture of log-normal distributions

PhyloGFN - Training

Trajectory balance loss with uniform backward policy (Malkin et al. 2022)

$$\mathcal{L}_{\mathrm{TB}}(\tau) = \left(\log \frac{Z_\theta \prod_{i=0}^{n-1} P_F(s_{i+1} \mid s_i; \theta)}{R(x) P_B(\tau \mid x)}\right)^2, \quad P_B(\tau \mid x) := \prod_{i=1}^{n} \frac{1}{|\mathrm{Pa}(s_i)|}$$

Exploration strategies:

- Eps-Greedy
- Temperature annealing
- Replay buffer storing best trees seen

Evaluation - Dataset

Table S1: Statistics of the benchmark datasets from DS1 to DS8.

Dataset	# Species	# Sites	Reference
DS1	27	1949	Hedges et al. (1990)
DS2	29	2520	Garey et al. (1996)
DS3	36	1812	Yang & Yoder (2003)
DS4	41	1137	Henk et al. (2003)
DS5	50	378	Lakner et al. (2008)
DS6	50	1133	Zhang & Blackwell (2001)
DS7	59	1824	Yoder & Yang (2004)
DS8	64	1008	Rossman et al. (2001)

Evaluation - Bayesian inference

For bayesian inference, performance is evaluated with estimated marginal log likelihood (MLL) lower bound

$$\log P(Y) \geq \mathbb{E}_{\tau_1,\ldots,\tau_k \sim P_F} \log \left(P(z) \frac{1}{K} \sum_{\substack{\tau_i \\ \tau_i : s_0 \rightarrow \cdots \rightarrow (z_i, b_i)}}^{k} \frac{P_B(\tau_i | z_i, b_i) R(z_i, b_i)}{P_F(\tau_i)} \right)$$

Methods in comparison:

- MCMC based algorithm: MrBayes SS (Xie et al., 2011, Ronquist et al., 2012)
- VI algorithm
    - VBPI-GNN (Zhang, 2023)
    - VaiPhy (koptagel et al., 2022)
    - GeoPhy (Mimori & Hamada, 2023)

Evaluation - Bayesian inference

MLL estimation

Table S4: Marginal log-likelihood estimation with different methods on real datasets DS1-DS8. PhyloGFN-C(ontinuous) now outperforms ϕ-CSMC, GeoPhy and PhyloGFN-B(ayesian) across all datasets and it is effectively performing on par with the state of the arts MrBayes and VBPI-GNN.

	MCMC		ML-based / amortized, full tree space			
Dataset	MrBayes SS	VBPI-GNN*	ϕ-CSMC	GeoPhy	PhyloGFN-B	PhyloGFN-C
DS1	−7108.42 ±0.18	−7108.41 ±0.14	−7290.36 ±7.23	−7111.55 ±0.07	−7108.95 ±0.06	−7108.40 ±0.04
DS2	−26367.57 ±0.48	−26367.73 ±0.07	−30568.49 ±31.34	−26368.44 ±0.13	−26368.90 ±0.28	−26367.70 ±0.04
DS3	−33735.44 ±0.5	−33735.12 ±0.09	−33798.06 ±6.62	−33735.85 ±0.12	−33735.6 ±0.35	−33735.11 ±0.02
DS4	−13330.06 ±0.54	−13329.94 ±0.19	−13582.24 ±35.08	−13337.42 ±1.32	−13331.83 ±0.19	−13329.91 ±0.02
DS5	−8214.51 ±0.28	−8214.64 ±0.38	−8367.51 ±8.87	−8233.89 ±6.63	−8215.15 ±0.2	−8214.38 ±0.16
DS6	−6724.07 ±0.86	−6724.37 ±0.4	−7013.83 ±16.99	−6733.91 ±0.57	−6730.68 ±0.54	−6724.17 ±0.10
DS7	−37332.76 ±2.42	−37332.04 ±0.12		−37350.77 ±11.74	−37359.96 ±1.14	−37331.89 ±0.14
DS8	−8649.88 ±1.75	−8650.65 ±0.45	−9209.18 ±18.03	−8660.48 ±0.78	−8654.76 ±0.19	−8650.46 ±0.05

Evaluation - Bayesian inference

MLL estimation

Table S4: Marginal log-likelihood estimation with different methods on real datasets DS1-DS8. PhyloGFN-C(ontinuous) now outperforms ϕ-CSMC, GeoPhy and PhyloGFN-B(ayesian) across all datasets and it is effectively performing on par with the state of the arts MrBayes and VBPI-GNN.

	MCMC	ML-based / amortized, full tree space				
Dataset	MrBayes SS	VBPI-GNN*	ϕ-CSMC	GeoPhy	PhyloGFN-B	PhyloGFN-C
DS1	−7108.42 ±0.18	−7108.41 ±0.14	−7290.36 ±7.23	−7111.55 ±0.07	−7108.95 ±0.06	−7108.40 ±0.04
DS2	−26367.57 ±0.48	−26367.73 ±0.07	−30568.49 ±31.34	−26368.44 ±0.13	−26368.90 ±0.28	−26367.70 ±0.04
DS3	−33735.44 ±0.5	−33735.12 ±0.09	−33798.06 ±6.62	−33735.85 ±0.12	−33735.6 ±0.35	−33735.11 ±0.02
DS4	−13330.06 ±0.54	−13329.94 ±0.19	−13582.24 ±35.08	−13337.42 ±1.32	−13331.83 ±0.19	−13329.91 ±0.02
DS5	−8214.51 ±0.28	−8214.64 ±0.38	−8367.51 ±8.87	−8233.89 ±6.63	−8215.15 ±0.2	−8214.38 ±0.16
DS6	−6724.07 ±0.86	−6724.37 ±0.4	−7013.83 ±16.99	−6733.91 ±0.57	−6730.68 ±0.54	−6724.17 ±0.10
DS7	−37332.76 ±2.42	−37332.04 ±0.12		−37350.77 ±11.74	−37359.96 ±1.14	−37331.89 ±0.14
DS8	−8649.88 ±1.75	−8650.65 ±0.45	−9209.18 ±18.03	−8660.48 ±0.78	−8654.76 ±0.19	−8650.46 ±0.05

Evaluation - Bayesian inference running time

Running time:

- Reported results take 3-7 days
- Achieves similar performance with24% training data (<2 days for all datasets)

Compare with VI methods on DS1

	VBPI-GNN	GeoPhy	ϕ-CSMC	PhyloGFN Full	PhyloGFN - 40%	PhyloGFN - 24%
Running Time	16h10min	8h10min	~ 2h	62h40min	20h40min	15h40min
MLL	-7108.41 (0.14)	-7111.55 (0.07)	-7290.36 (7.23)	-7108.40 (0.04)	-7108.39 (0.09)	-7108.42 (0.05)

Evaluation - Bayesian inference running time

Running time:

- Reported results take 3-7 days
- Achieves similar performance with24% training data (<2 days for all datasets)

Compare with VI methods on DS1

	VBPI-GNN	GeoPhy	ϕ-CSMC	PhyloGFN Full	PhyloGFN - 40%	PhyloGFN - 24%
Running Time	16h10min	8h10min	~ 2h	62h40min	20h40min	15h40min
MLL	-7108.41 (0.14)	-7111.55 (0.07)	-7290.36 (7.23)	-7108.40 (0.04)	-7108.39 (0.09)	-7108.42 (0.05)

Evaluation - Bayesian inference running time

Running time:

- Reported results take 3-7 days
- Achieves similar performance with24% training data (<2 days for all datasets)

Compare with VI methods on DS1

	VBPI-GNN	GeoPhy	ϕ-CSMC	PhyloGFN Full	PhyloGFN - 40%	PhyloGFN - 24%
Running Time	16h10min	8h10min	~ 2h	62h40min	20h40min	15h40min
MLL	-7108.41 (0.14)	-7111.55 (0.07)	-7290.36 (7.23)	-7108.40 (0.04)	-7108.39 (0.09)	-7108.42 (0.05)

Evaluation - Bayesian inference running time

Running time:

- Reported results take 3-7 days
- Achieves similar performance with24% training data (<2 days for all datasets)

Compare with VI methods on DS1

	VBPI-GNN	GeoPhy	ϕ-CSMC	PhyloGFN Full	PhyloGFN - 40%	PhyloGFN - 24%
Running Time	16h10min	8h10min	~ 2h	62h40min	20h40min	15h40min
MLL	-7108.41 (0.14)	-7111.55 (0.07)	-7290.36 (7.23)	-7108.40 (0.04)	-7108.39 (0.09)	-7108.42 (0.05)