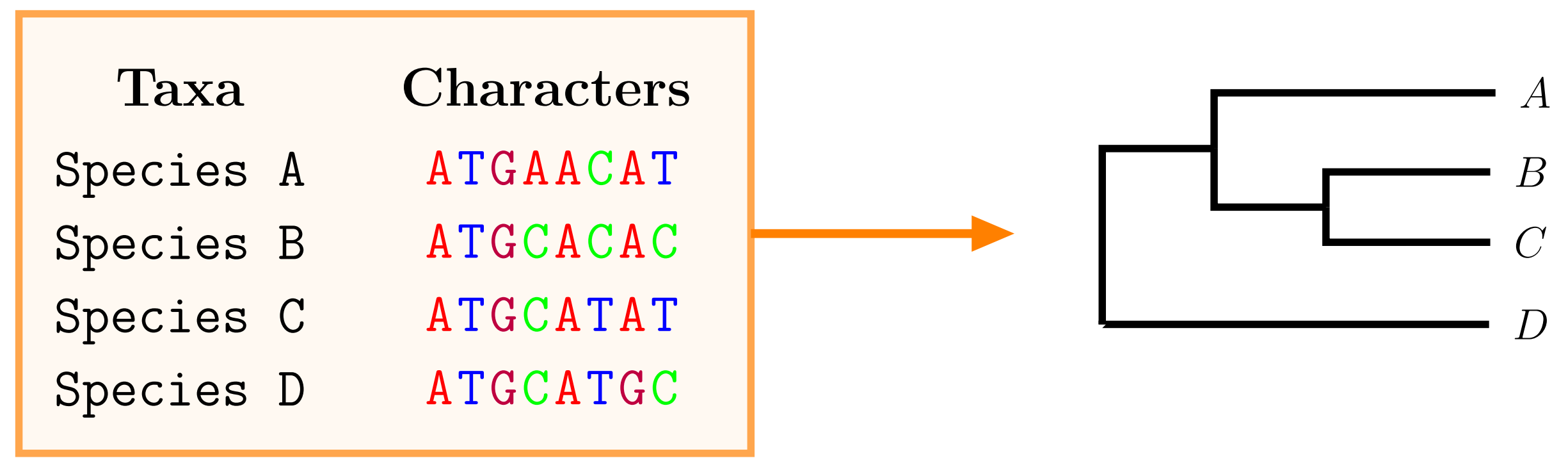
Variational Bayesian Phylogenetic Inference

Cheng Zhang and Frederick A. Matsen IV

czhang23@fredhutch.org   Program in Computational Biology, Fred Hutchinson Cancer Research Center

Introduction

Phylogenetic inference: reconstruct the evolution history (e.g., phylogenetic trees) from molecular sequence data (e.g., DNA, RNA or protein sequences).



Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data   Phylogenetic Tree

Likelihood and posterior: A phylogenetic tree $(\tau, \boldsymbol{q})$ contains both the discrete tree topology τ and continuous branch lengths $\boldsymbol{q}$. Let $\boldsymbol{Y} = \{Y_i\}_{i=1}^M$ be the observed characters.

$$p(\boldsymbol{Y}|\tau, \boldsymbol{q}) = \prod_{i=1}^{M} \sum_{a^i} \eta(a_\rho^i) \prod_{(u,v)\in E(\tau)} P_{a_u^i a_v^i}(q_{uv})$$

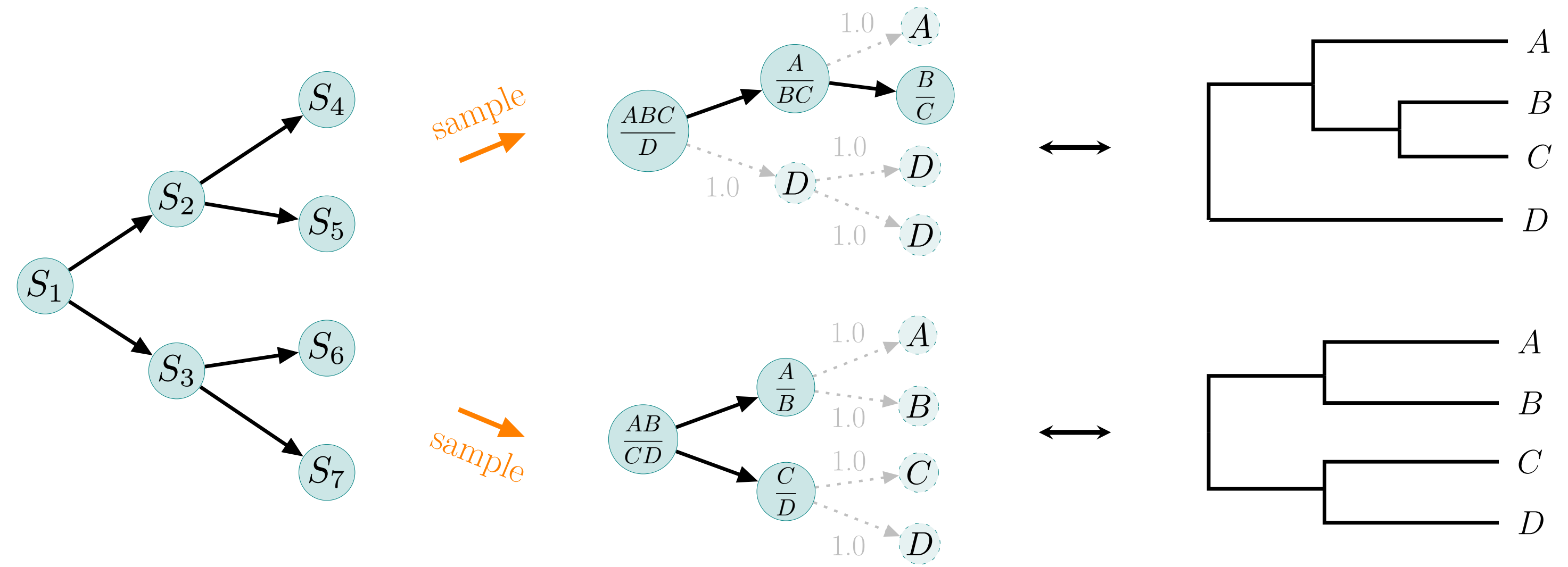
Given a proper prior distribution $p(\tau, \boldsymbol{q})$, the posterior is $p(\tau, \boldsymbol{q}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\tau, \boldsymbol{q})p(\tau, \boldsymbol{q})$.

Motivation: Current random walk MCMC based Bayesian phylogenetic methods do not handle the intertwined parameter space of phylogenetic models with ease, leading to low exploration efficiency and expensive computation.

Based on **subsplit Bayesian networks** (SBNs), a recent framework that provides flexible distributions of trees, we propose the first general variational Bayes formulation of phylogenetic inference for both tree topologies and branch lengths that

- provides comparable posterior estimates to MCMC methods with much less computation.
- results can be used for further statistical analysis such as marginal likelihood estimation.

Subsplit Bayesian Networks



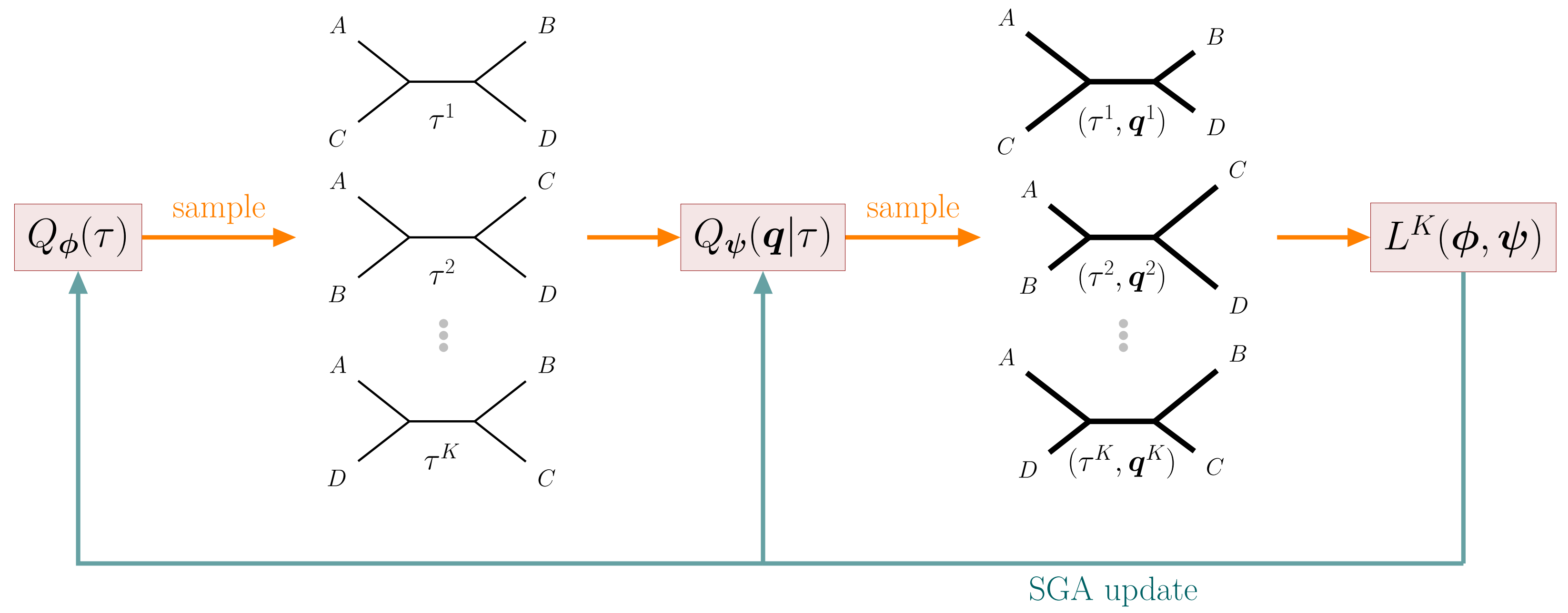
- **Rooted Trees**:

$$p_{\text{sbn}}(T = \tau) = p(S_1 = s_1) \prod_{i>1} p(S_i = s_i | S_{\pi_i} = s_{\pi_i})$$

- **Unrooted Trees**:

$$p_{\text{sbn}}(T^{\text{u}} = \tau) = \sum_{s_1 \sim \tau} p(S_1 = s_1) \prod_{i>1} p(S_i = s_i | S_{\pi_i} = s_{\pi_i})$$

Variational Bayesian Phylogenetic Inference



$Q_\phi(\tau)$
sample
τ^1
τ^2
τ^K
$Q_\psi(\boldsymbol{q}|\tau)$
sample
$(\tau^1, \boldsymbol{q}^1)$
$(\tau^2, \boldsymbol{q}^2)$
$(\tau^K, \boldsymbol{q}^K)$
$L^K(\boldsymbol{\phi}, \boldsymbol{\psi})$
A
B
C
D
SGA update

Variational Approximation. We use $Q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau, \boldsymbol{q}) = Q_{\boldsymbol{\phi}}(\tau) \cdot Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau)$ as our approximating distribution, where $Q_{\boldsymbol{\phi}}(\tau)$ is a probability distribution of tree topologies (e.g., **subsplit Bayesian networks**), and $Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau)$ is the conditional probability of branch lengths given the tree topology (e.g., Lognormal)

Multi-sample Lower Bound. We use multi-sample lower bound that facilitates exploration in tree space

$$L^K(\boldsymbol{\phi}, \boldsymbol{\psi}) = \mathbb{E}_{Q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau^{1:K}, \boldsymbol{q}^{1:K})} \log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{p(\boldsymbol{Y}|\tau^i, \boldsymbol{q}^i) p(\tau^i, \boldsymbol{q}^i)}{Q_{\boldsymbol{\phi}}(\tau^i) Q_{\boldsymbol{\psi}}(\boldsymbol{q}^i|\tau^i)} \right) \leq \log p(\boldsymbol{Y})$$

Parameterization. We define the conditional probability in SBNs via softmax

$$p(S_1 = s_1) = \frac{\exp(\phi_{s_1})}{\sum_{s_r \in \mathbb{S}_r} \exp(\phi_{s_r})}, \quad p(S_i = s | S_{\pi_i} = t) = \frac{\exp(\phi_{s|t})}{\sum_{s \in \mathbb{S}_{\cdot|t}} \exp(\phi_{s|t})}$$

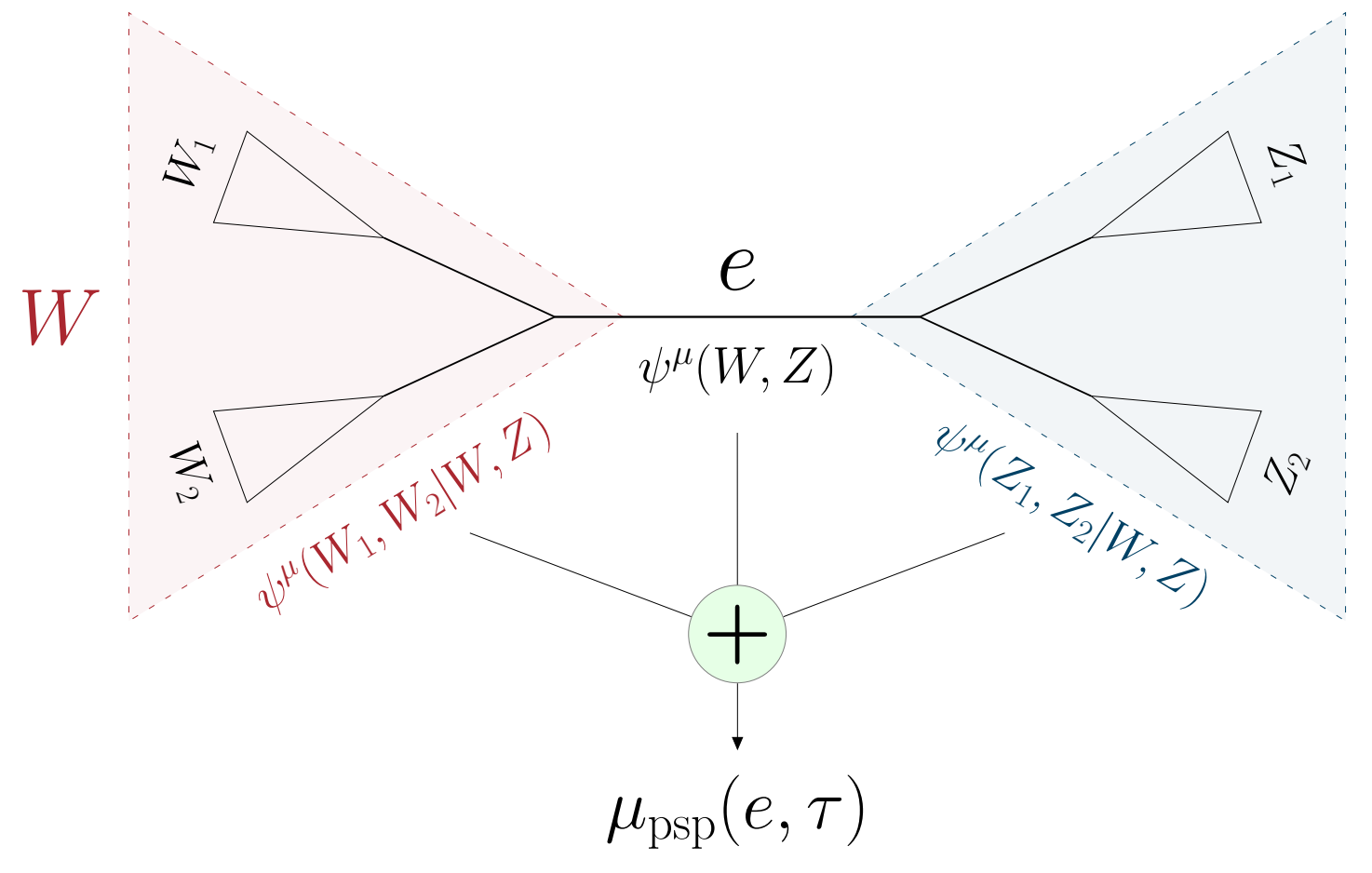
where $\mathbb{S}_r$ and $\mathbb{S}_{\text{ch|pa}}$ are supports estimated from ultrafast bootstrap trees.

We use diagonal Lognormal for branch lengths distribution

$$Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}(q_e \mid \mu(e, \tau), \sigma(e, \tau))$$

- *Simple Split*: $\mu_{\text{s}}(e, \tau) = \psi^\mu_{e/\tau}, \ \sigma_{\text{s}}(e, \tau) = \psi^\sigma_{e/\tau}$.
- *Primary Subsplit Pair*:

$$\mu_{\text{psp}}(e, \tau) = \psi^\mu_{e/\tau} + \sum_{s \in e /\!\!/ \tau} \psi^\mu_s, \quad \sigma_{\text{psp}}(e, \tau) = \psi^\sigma_{e/\tau} + \sum_{s \in e /\!\!/ \tau} \psi^\sigma_s.$$



W
W_1
W_2
e
$\psi^\mu(W, Z)$
Z
Z_1
Z_2
$\psi^\mu(W_1, W_2|W, Z)$
$\psi^\mu(Z_1, Z_2|W, Z)$
$\mu_{\text{psp}}(e, \tau)$

Stochastic Gradient Estimator. We use different stochastic gradient estimators for different variational parameters. For tree topology parameters $\boldsymbol{\phi}$

- *VIMCO*.

$$\nabla_{\boldsymbol{\phi}} L^K(\boldsymbol{\phi}, \boldsymbol{\psi}) \simeq \sum_{j=1}^{K} \left(\hat{L}^K_{j|-j}(\boldsymbol{\phi}, \boldsymbol{\psi}) - \tilde{w}^j \right) \nabla_{\boldsymbol{\phi}} \log Q_{\boldsymbol{\phi}}(\tau^j) \text{ with } \tau^j, \boldsymbol{q}^j \overset{\text{iid}}{\sim} Q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau, \ \boldsymbol{q}).$$

- *RWS*.

$$\nabla_{\boldsymbol{\phi}} L^K(\boldsymbol{\phi}, \boldsymbol{\psi}) \simeq \sum_{j=1}^{K} \tilde{w}^j \nabla_{\boldsymbol{\phi}} \log Q_{\boldsymbol{\phi}}(\tau^j) \ \text{ with } \ \tau^j, \boldsymbol{q}^j \overset{\text{iid}}{\sim} Q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau, \ \boldsymbol{q})$$

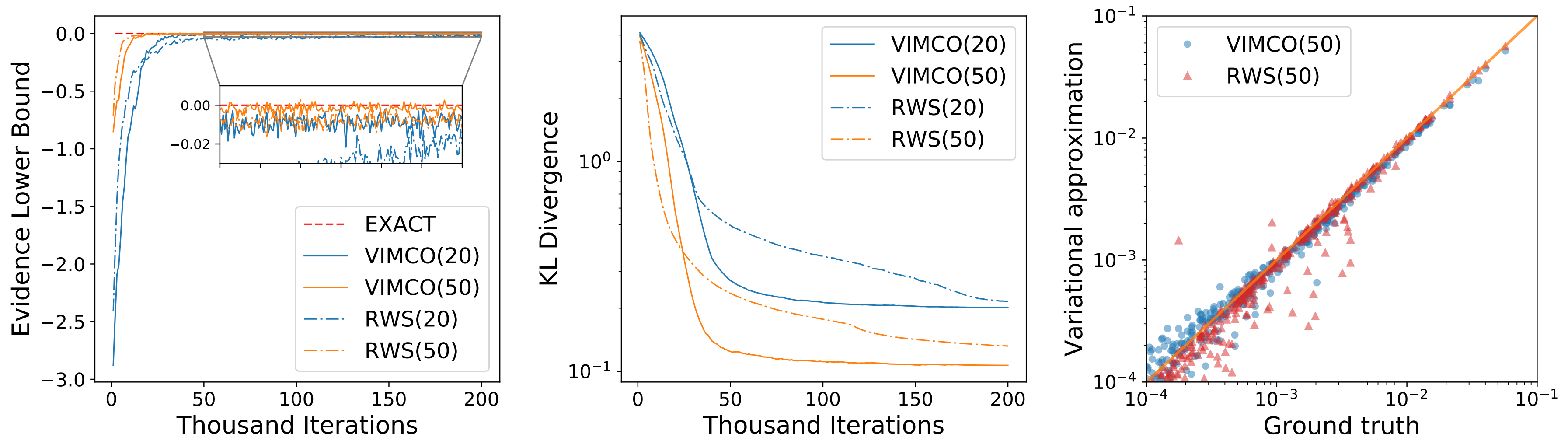
For branch length parameters $\boldsymbol{\psi}$, a simple reparameterization for Lognormal $g_{\boldsymbol{\psi}}(\boldsymbol{\epsilon}|\tau) = \exp(\boldsymbol{\mu}_{\boldsymbol{\psi},\tau} + \boldsymbol{\sigma}_{\boldsymbol{\psi},\tau} \odot \boldsymbol{\epsilon})$.

- *Reparameterization Trick*. Let $f_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau, \boldsymbol{q}) = \frac{p(\boldsymbol{Y}|\tau,\boldsymbol{q})p(\tau,\boldsymbol{q})}{Q_{\boldsymbol{\phi}}(\tau)Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau)}$.

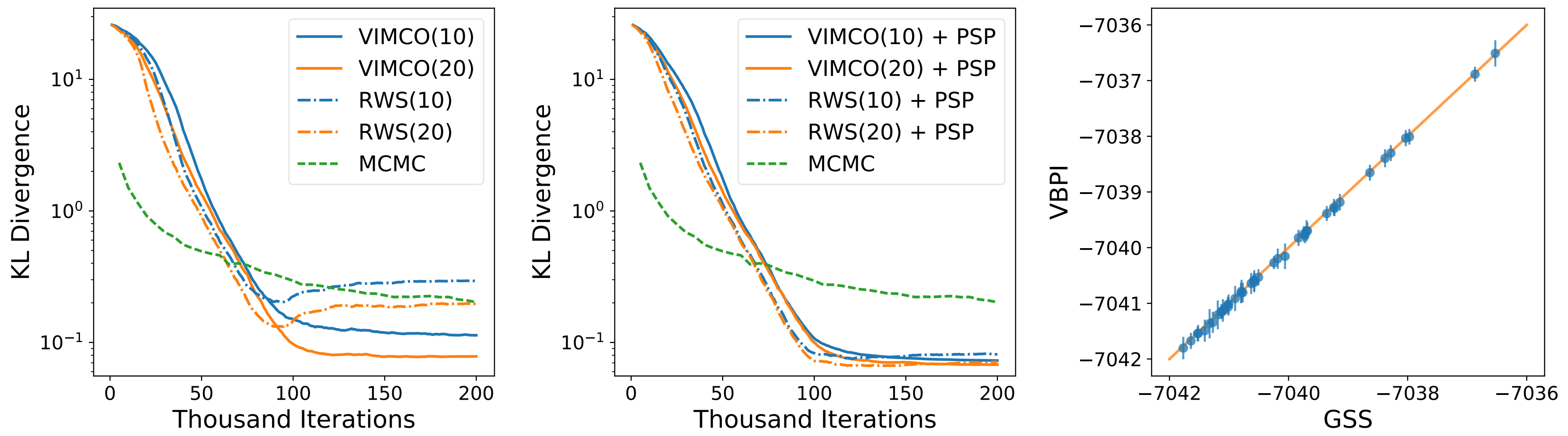
$$\nabla_{\boldsymbol{\psi}} L^K(\boldsymbol{\phi}, \boldsymbol{\psi}) \simeq \sum_{j=1}^{K} \tilde{w}^j \nabla_{\boldsymbol{\psi}} \log f_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau^j, g_{\boldsymbol{\psi}}(\boldsymbol{\epsilon}^j|\tau^j)) \text{ with } \tau^j \overset{\text{iid}}{\sim} Q_{\boldsymbol{\phi}}(\tau), \ \boldsymbol{\epsilon}^j \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}).$$

Experiments

SBNs can learn accurate approximation of phylogenetic tree distributions via our **variational Bayesian phylogenetic inference** (VBPI) framework.



With guided tree space exploration enabled by SBNs, VBPI provides competitive performance to MCMC methods and arrives at good approximation with much less computation. Moreover, the variational approximations can be readily used for marginal likelihood computation via importance sampling.



For a wide range of real data sets, VBPI provides more reliable marginal likelihood estimates than one of the state-of-the-art methods, stepping-stone (SS), and requires much less samples (VBPI: 1000 vs SS: 100,000).

Data set (#Taxa, #Sites)		Marginal Likelihood (NATs)				
		VIMCO(10)	VIMCO(20)	VIMCO(10)+PSP	VIMCO(20)+PSP	SS
DS1	(27, 1949)	-7108.43(0.26)	-7108.35(0.21)	-7108.41(0.16)	**-7108.42(0.10)**	-7108.42(0.18)
DS2	(29, 2520)	-26367.70(0.12)	-26367.71(0.09)	**-26367.72(0.08)**	-26367.70(0.10)	-26367.57(0.48)
DS3	(36, 1812)	-33735.08(0.11)	-33735.11(0.11)	**-33735.10(0.09)**	-33735.07(0.11)	-33735.44(0.50)
DS4	(41, 1137)	-13329.90(0.31)	-13329.98(0.20)	**-13329.94(0.18)**	-13329.93(0.22)	-13330.06(0.54)
DS5	(50, 378)	-8214.36(0.67)	-8214.74(0.38)	-8214.61(0.38)	-8214.55(0.43)	**-8214.51(0.28)**
DS6	(50, 1133)	-6723.75(0.68)	-6723.71(0.65)	-6724.09(0.55)	**-6724.34(0.45)**	-6724.07(0.86)
DS7	(59, 1824)	-37332.03(0.43)	-37331.90(0.49)	-37331.90(0.32)	**-37332.03(0.23)**	-37332.76(2.42)
DS8	(64, 1008)	-8653.34(0.55)	-8651.54(0.80)	**-8650.63(0.42)**	-8650.55(0.46)	-8649.88(1.75)

Conclusion

We introduced VBPI, a general variational framework for Bayesian phylogenetic inference, based on subsplit Bayesian networks and efficient structured parameterizations for branch length distributions. Thanks to SBNs, VBPI exhibits guided exploration in tree space and provides competitive performance to MCMC methods with much less computation. Moreover, variational approximations provided by VBPI can be readily used for further statistical analysis such as marginal likelihood estimation. We hope that these ideas will draw more attention to challenging structural learning problems in computational biology and inspire new representation learning methods for phylogenetic models.