Variational Combinatorial Sequential Monte Carlo Methods for Bayesian Phylogenetic Inference

Antonio Khalil Moretti[1]   **Liyi Zhang**[1]   **Christian A. Naesseth**[1]   **Hadiah Venner**[1]   **David Blei**[1]   **Itsik Pe'er**[1]

[1]Columbia University

Abstract

Bayesian phylogenetic inference is often conducted via local or sequential search over topologies and branch lengths using algorithms such as random-walk Markov chain Monte Carlo (MCMC) or Combinatorial Sequential Monte Carlo (CSMC). However, when MCMC is used for evolutionary parameter learning, convergence requires long runs with inefficient exploration of the state space. We introduce Variational Combinatorial Sequential Monte Carlo (VCSMC), a powerful framework that establishes variational sequential search to learn distributions over intricate combinatorial structures. We then develop nested CSMC, an efficient proposal distribution for CSMC and prove that nested CSMC is an exact approximation to the (intractable) locally optimal proposal. We use nested CSMC to define a second objective, VNCSMC which yields tighter lower bounds than VCSMC. We show that VCSMC and VNCSMC are computationally efficient and explore higher probability spaces than existing methods on a range of tasks.

1 INTRODUCTION

What is the origin of SARS-COV-II and how can we analyze the progression of its genetic variants? How do antibodies evolve and develop in response to infection and vaccination? Bayesian phylogenetic inference is a powerful statistical tool to address these and other questions of central importance in molecular evolutionary biology and epidemiology [Dhar et al., 2020, Boni et al., 2020]. Given an evolutionary model and an alignment of observed molecular sequences (DNA, RNA, PROTEIN), Bayesian methods sample latent bifurcating trees to uncover genetic history, quantify uncertainty and incorporate prior information [Huelsenbeck and Ronquist, 2001]. Phylogenetic modeling involves three distinct challenges: (i) sampling from a discrete distribution to approximate an intractable summation over tree topologies, (ii) for each tree, integrating over the continuous branch lengths that govern the stochastic process for genetic mutations, and (iii) performing parameter optimization or model learning. The *marginalization* of tree topologies and branch lengths is typically accomplished via local search algorithms such as random-walk Markov chain Monte Carlo (MCMC) [Huelsenbeck and Ronquist, 2001] or sequential search algorithms such as Combinatorial Sequential Monte Carlo (CSMC) [Bouchard-Côté et al., 2012]. Sophisticated proposal methods based on Hamiltonian Monte Carlo or particle MCMC have been suggested to simultaneously sample from composite spaces and optimize evolutionary parameters [Dinh et al., 2017a, Wang et al., 2015, Wang and Wang, 2020]. However, these methods are often difficult to implement, slow to converge requiring days or weeks of CPU time, and heavily dependent upon heuristics.

Variational Inference (VI) is a computationally efficient alternative to MCMC. VI posits an approximate posterior and then recovers parameters of both the model and approximate posterior by maximizing a lower bound to the log-marginal likelihood. One approach to learning variational distributions on phylogenetic trees is to parameterize the tree as a sequence of *subsplits*, or ordered partitions on clades, and to recast the problem as a Bayesian network [Zhang and Matsen IV, 2018]. The drawback of this setup is that the support of the conditional probability tables scales exponentially with the number of taxa [Zhang and Matsen IV, 2019]. A body of recent work has established connections between VI and sequential search by defining a variational family of distributions on hidden Markov models, where Sequential Monte Carlo (SMC) is used as the marginal likelihood estimator [Maddison et al., 2017, Le et al., 2018, Naesseth et al., 2018, Lawson et al., 2018, Moretti et al., 2019b,a, Naesseth et al., 2020, Moretti et al., 2020a,b, Moretti, 2021]. We extend these approaches by developing variational sequential search methods that learn distributions over complex combinatorial structures. Our contributions are as follows:

- We develop Variational Combinatorial Sequential Monte Carlo (VCSMC), a novel variational objective and structured approximate posterior defined on the space of phylogenetic trees. VCSMC blends CSMC and VI, providing the user with a flexible and powerful approximate inference algorithm.

- We further extend CSMC with nested SMC [Naesseth et al., 2015, 2019a], introducing a new efficient proposal distribution for CSMC. We prove that this proposal is an *exact approximation* to the (intractable) locally optimal proposal for CSMC. We use NCSMC to define a second objective, VNCSMC which yields tighter lower bounds than VCSMC.

- In empirical studies, we demonstrate the advantage of VCSMC and VNCSMC. First, we analyze a standard dataset of primate mitochondrial DNA, then the complete genomes of 17 Betacoronavirus species over 36,889 sites, and finally 7 benchmark datasets (DS1-DS7) ranging from 27 to 64 taxa. VCSMC and VNC-SMC are compared to existing benchmarks and shown to perform favorably across a range of tasks.

Related Work. Bayesian phylogenetics is often approximated using local search algorithms such as random-walk MCMC [Huelsenbeck and Ronquist, 2001] or sequential search algorithms such as CSMC [Bouchard-Côté et al., 2012]. MCMC methods can also be used for model learning, jointly estimating the phylogenetic trees and evolutionary parameters. Probabilistic path Hamiltonian Monte Carlo ($ppHMC$) [Dinh et al., 2017a] is one such method that extends Hamiltonian Monte Carlo by defining a Markov chain on the orthant complex of phylogenetic tree space. It is often the case that the likelihood term in the MCMC acceptance ratio is difficult to evaluate. The idea of Particle MCMC algorithms (PMCMC) is to use SMC as an unbiased estimate of the marginal likelihood to define a proposal for MCMC [Andrieu et al., 2010]. A PMCMC algorithm for evolutionary parameter learning was introduced in [Wang et al., 2015], and improved upon using a particle Gibbs sampler in [Wang and Wang, 2020]. In contrast to these methods, the proposed approach leverages VI for inference and introduces a new efficient proposal distribution for CSMC.

One approach to VI for phylogenetic trees is to parameterize a tree as a sequence of *subsplits*, or ordered partitions on clades and to recast the problem as a Bayesian network [Zhang and Matsen IV, 2018]. A drawback of this setup is that the support of the conditional probability tables scales exponentially with the number of taxa [Zhang and Matsen IV, 2019]. In subsequent work, the authors introduce two Variational Bayesian Phylogenetic Inference frameworks (VBPI and VBPI-NF) by using pre-computed topologies to define the support of the conditional probability tables for the approximation [Zhang and Matsen IV, 2019, Zhang, 2020]. In contrast, VCSMC does not restrict the support of the tree topologies and instead leverages CSMC to compute a lower bound.

2 BACKGROUND

Phylogenetic Trees. We wish to infer a latent bifurcating tree that describes the evolutionary relationships among a set of observed molecular sequences. A phylogeny is defined by a tree topology τ and a set of branch lengths $\mathcal{B}$. A *tree topology* is defined as a connected acyclic graph (V, E) where V is a set of vertices and E is a set of edges. *Leaf nodes* denote vertices of degree 1 and correspond to observed taxa. *Internal nodes* designate vertices of degree 3 (one parent and two children) and represent unobserved taxa (e.g. DNA bases of ancestral species). The *root node* is of degree 2 (two children) and represents the common evolutionary ancestor of all taxa.

For each edge $e \in E$, we associate a *branch length*, denoted $b(e) \in \mathbb{R}_{>0}$, and $\mathcal{B} = \{b(e)\}_{e \in E}$. The branch length captures the intensity of the evolutionary changes between two vertices. An *ultrametric tree* is one with constant evolutionary rate along all paths from v to its descendants. *Nonclock trees* are general trees that do not require ultrametric assumptions. In this work we focus on phylogenetic inference methods for nonclock trees as these are most pertinent to biologists.

Bayesian Phylogenetic Inference. Let the matrix $\mathbf{Y} = \{Y_1, \cdots, Y_S\} \in \Omega^{N \times S}$ denote the observed molecular sequences with characters in Ω of length S over N species. Bayesian inference requires specifying the prior density and likelihood function over tree topology τ, branch length set $\mathcal{B}$ and generative model parameters θ to write the joint posterior,

$$P_\theta(\mathcal{B}, \tau | \mathbf{Y}) = \frac{P_\theta(\mathbf{Y}|\tau, \mathcal{B}) P_\theta(\tau, \mathcal{B})}{P_\theta(\mathbf{Y})}. \quad (1)$$

The prior is uniform over topologies and a product of independent exponential distributions over branch lengths with rate λ_{bl}. The evolution of each site is modeled independently using a continuous time Markov chain with rate matrix $\mathbf{Q}$. Let $\zeta_{v,s}$ denote the state of genome for species v at site s and define the evolutionary model along branch $b(v \rightarrow v')$:

$$P_\theta(\zeta_{v',s} = j | \zeta_{v,s} = i) = \exp(b(e)\mathbf{Q}_{i,j}). \quad (2)$$

The likelihood of a given phylogeny $P_\theta(\mathbf{Y}|\tau, \mathcal{B}) = \prod_{i=1}^{S} P_\theta(Y_i|\tau, \mathcal{B})$ can be evaluated in linear time using the sum-product or Felsenstein's pruning algorithm [Felsenstein, 1981] via the formula:

$$P_\theta(\mathbf{Y}|\tau, \mathcal{B}) := \prod_{i=1}^{S} \sum_{a^i} \eta(a_\rho^i) \prod_{(u,v) \in E(\tau)} \exp\left(-b_{u,v} \mathbf{Q}_{a_u^i, a_v^i}\right),$$
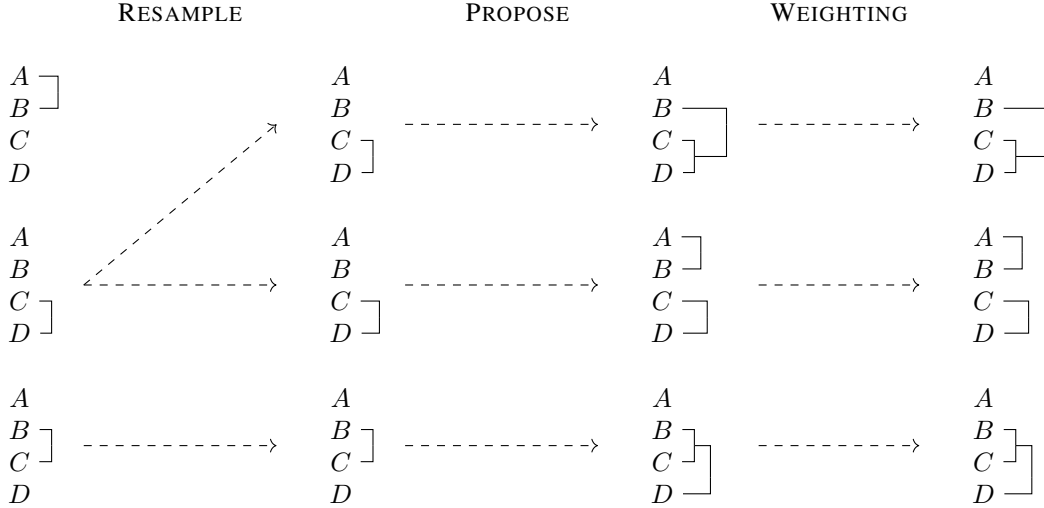
Figure 1: Overview of the CSMC framework. K partial states are maintained as forests over the set of taxa. Each iteration of Algorithm 2 involves three steps: (1) resample partial states according to their importance weights, (2) propose an extension of each partial state to a new partial state by connecting two trees in the forest, and (3) compute weights for each new partial state by using Felsenstein's pruning algorithm. In the above, three samples are shown over four taxa A, B, C, D.

where ρ is the root node, a_u^i is the assigned character of node u, $E(\tau)$ represents the set of edges in τ and η is the prior or stationary distribution of the Markov chain. The normalization constant $P_\theta(\mathbf{Y})$ requires marginalizing the $(2N-3)!!$ distinct topologies which is intractable [Semple and Steel, 2003].

Computational Challenges. We distinguish the two computational tasks required for phylogenetic inference. First, inference involves computing the normalization constant $P_\theta(\mathbf{Y})$ by marginalizing the $(2N-3)!!$ distinct topologies:

$$P_\theta(\mathbf{Y}) = \sum_{\tau \in \mathcal{T}} \int p_\theta(\mathbf{Y}|\tau, \mathcal{B}) p_\theta(\tau, \mathcal{B}) d\mathcal{B}. \quad (3)$$

A common approach used for approximating Eq. 3 is to sample tree topologies τ and branch lengths $\mathcal{B}$ via Monte Carlo methods, such as CSMC, given that θ is known.

Second, learning (*optimization*) refers to finding the set of parameters $\theta = (Q, \{\lambda_i\}_{i=1}^{|E|} \in \Theta)$ that maximize the data log-likelihood obtained by marginalizing Eq. 3:

$$\theta^\star = \underset{Q, \{\lambda_i\}_{i=1}^{|E|}}{\arg\max} \ \log P_\theta(\mathbf{Y}). \quad (4)$$

Sampling algorithms can also be used by assigning a prior to θ, then performing a local search for the parameters via MCMC methods, given that the data likelihood is available.

Variational Inference. VI is a technique for approximating the posterior $P_\theta(\mathcal{B}, \tau|\mathbf{Y})$ when marginalization of latent variables is not analytically feasible. By introducing a tractable distribution $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ it is possible to form a lower bound to the log-likelihood:

$$\log P_\theta(\mathbf{Y}) \geq \mathcal{L}_{\mathrm{ELBO}}(\theta, \phi, \mathbf{Y}) \coloneqq \underset{Q}{\mathbb{E}} \left[\log \frac{P_\theta(\mathbf{Y}, \mathcal{B}, \tau)}{Q_\phi(\mathcal{B}, \tau|\mathbf{Y})} \right]. \quad (5)$$

Auto Encoding Variational Bayes [Kingma and Welling, 2013] (AEVB) simultaneously trains $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ and $P_\theta(\mathbf{Y}, \mathcal{B}, \tau)$. The expectation in Eq. 5 is approximated by averaging Monte Carlo samples from $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ which are reparameterized by evaluating a deterministic function of a ϕ-independent random variable. When the ratio $P_\theta(\mathbf{Y}, \mathcal{B}, \tau)/Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ is concentrated around its mean, Jensen's inequality produces a tighter bound.

Deriving a tractable approximation $Q_\phi(\mathcal{B}, \tau|\mathbf{Y})$ for the phylogenetic tree model can be challenging so we turn to CSMC.

Combinatorial Sequential Monte Carlo. CSMC is designed for inference in phylogenetic tree models. CSMC approximates a sequence of target distributions $\bar{\pi}_r$ on increasing probability spaces such that the final target coincides with Eq. 1 [Wang et al., 2015]. The (unnormalized) target distribution π and its normalization constant $\|\pi\|$ corresponding to the numerator and denominator in Eq. 1 are approximated by sequential importance resampling in R steps. Unlike standard SMC methods, the target π is defined on a combinatorial set (the space of tree topologies) and the continuous branch lengths. This requires defining an intermediate object referred to as a *partial state*.

Definition 1 (Partial State). A partial state of rank r denoted $s_r = \{(t_i, X_i)\}$ is a collection of rooted trees that satisfies the following three conditions: (i) the set of partial states of different ranks are disjoint, $\forall r \neq s, \mathcal{S}_r \cap \mathcal{S}_s = \emptyset$; (ii)

the set of partial states of smallest rank has a single element $S_0 = \{\bot\}$; and (ii) the set of partial states at the final rank R corresponds to the target space $\mathcal{X}$.

CSMC operates by sampling K *partial states* (or *particles*) $\{s_r^k\}_{k=1}^K \in \mathcal{S}_r$ at each rank r which are used to form a distribution,

$$\widehat{\pi}_r = \|\widehat{\pi}_{r-1}\| \frac{1}{K} \sum_{k=1}^K w_r^k \delta_{s_r^k}(s) \qquad \forall s \in \mathcal{S}, \qquad (6)$$

where δ_s is the Dirac measure and w_r^k are the importance weights. Resampling ensures that particles remain in areas of high probability mass. Each resampled state $s_{r-1}^{a_{r-1}^k}$, where $a_{r-1}^k \in \{1, \ldots, K\}$ is the resampled index, of rank $r-1$ is then extended to a state of rank r, s_r^k, by simulating from a proposal distribution $q(\cdot|s_{r-1}^{a_{r-1}^k}) : \mathcal{S} \to [0, 1]$. The importance weights are computed as follows:

$$w_r^k = w(s_{r-1}^{a_{r-1}^k}, s_r^k) = \frac{\pi(s_r^k)}{\pi(s_{r-1}^{a_{r-1}^k})} \cdot \frac{\nu^-(s_{r-1}^{a_{r-1}^k})}{q(s_r^k|s_{r-1}^{a_{r-1}^k})}, \qquad (7)$$

where ν^- is a probability density over $\mathcal{S}$ correcting an over-counting problem [Wang et al., 2015]. An overview of the procedure is given in Fig. 1. An unbiased estimate for the marginal likelihood can be constructed from the weights which converges in L^2 norm,

$$\widehat{\mathcal{Z}}_{CSMC} := \|\widehat{\pi}_R\| = \prod_{r=1}^R \left(\frac{1}{K} \sum_{k=1}^K w_r^k \right) \to \|\pi\|. \qquad (8)$$

VCSMC melds VI and CSMC to approximate the posterior as well as the model parameters.

3 VARIATIONAL COMBINATORIAL SEQUENTIAL MONTE CARLO

Variational Objective. The idea of VCSMC is to simultaneously learn the model parameters and proposal parameters by maximizing a lower bound to the data marginal log-likelihood, using CSMC as an unbiased estimator of the marginal likelihood.

We begin by defining a structured approximate posterior which factorizes over rank events. Each state (or rank event) s_r is specified by a topology, a forest of trees, and their corresponding set of branch lengths. The proposal $q_{\phi,\psi}(s_r^k|s_{r-1}^{a_{r-1}^k})$ is the probability of state s_r^k given the resampled state at the previous rank $s_{r-1}^{a_{r-1}^k}$. Subscripts ϕ and ψ denote discrete and continuous proposal parameters respectively. The

approximate posterior is (written explicitly in Eq. 16):

$$Q_{\phi,\psi}\left(s_{1:R}^{1:K}, a_{1:R-1}^{1:K}\right) := \qquad (9)$$

$$\prod_{k=1}^K q_{\phi,\psi}(s_1^k) \times \prod_{r=2}^R \prod_{k=1}^K \left[\frac{w_{r-1}^{a_{r-1}^k}}{\sum_{l=1}^K w_{r-1}^l} \cdot q_{\phi,\psi}\left(s_r^k|s_{r-1}^{a_{r-1}^k}\right) \right].$$

At the final rank event $R = N - 1$, an unbiased approximation to the likelihood is formed by averaging over importance weights, which, in turn represent the sample phylogenies that are constructed iteratively. A multi-sample variational objective is formed via the lower bound:

$$\mathcal{L}_{CSMC} := \mathbb{E}_Q \left[\log \widehat{\mathcal{Z}}_{CSMC} \right]. \qquad (10)$$

The presence of the discrete distribution over partial states presents a challenge for variational reparameterization. Unlike standard variational SMC methods [Naesseth et al., 2018], states are formed by sampling from a large combinatorial set. We take two approaches, the first is to drop discrete terms from the gradient estimates. The second is to reparameterize these terms as Gumbel-Softmax random variables forming a differentiable approximation through a convex relaxation over the simplex. Continuous proposal terms are drawn by evaluating a deterministic function of a ψ-independent random variable.

Implementation Details. Constructing the objective $\mathcal{L}_{CSMC}$ is done iteratively in three steps. The proposal procedure, $q(s_r|s_{r-1})$, requires selecting two trees to coalesce by sampling without replacement. This is accomplished by defining Gumbel-Softmax random variables. The uniform log-probability for each index is perturbed by adding independent Gumbel distributed noise, after which the largest two elements are returned. For example let $U \sim \text{UNIFORM}(0, 1)$, we then form $G = \gamma - \log(-\log U)$ so that G can be reparameterized as $G' = G + \gamma$. The RESAMPLE procedure can also be reparameterized similarly by defining Gumbel-Softmax random variables.

Extending the Target Measure. The WEIGHTING step requires some care. In order to compute importance weights, the likelihood of a partial state must be evaluated using Felsenstein's pruning algorithm, however the likelihood of Eq. 3 and the probability measure π are defined on the target space of trees $\mathcal{S}_R$, and not the larger sample space of partial states $\mathcal{S}_{r<R}$, which are defined on forests (trees disjoint from each other). The pruning algorithm yields a maximum likelihood estimate for an evolutionary tree, but partial states are defined as collections of disjoint trees or leaf nodes. One extension of the target measure π into a measure on $\mathcal{S}_{r<R}$ is to treat all elements of the jump chain as trees [Wang et al., 2015]. The contribution of each of the trees to the likelihood is multiplied by taking the inner product of each distribution over characters with η.
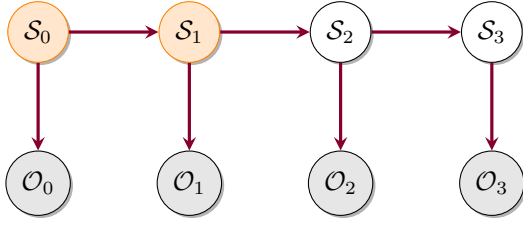
(a) State space model representation of CSMC. Latent variables consist of internal nodes and branch lengths with Markovian dependencies whereas observations are the recorded molecular sequences.

$$\pi(s) = \prod_{(t_i, X_i) \in s} \pi_{Y_i(x_i)}(t_i)$$
$$= P_{AB} \cdot \eta \times C \cdot \eta \times D \cdot \eta$$
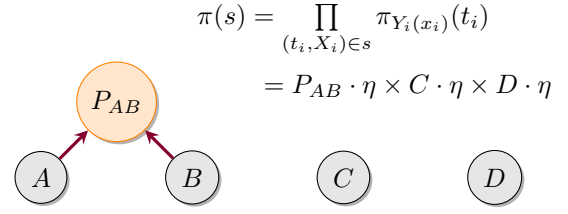
(b) Overview of the natural forest extension of the target measure. Partial state $s_r^k = \{P_{AB}, C, D\}$ is defined as a forest over leaves $\{A, B, C, D\}$.

Figure 2: Illustration of the sequence of probability spaces along with the natural forest extension used by CSMC. The probability measure π is defined on the target space of trees $\mathcal{S}_R$, and not the larger sample space of partial states $\mathcal{S}_{r<R}$, which are defined on forests. The partial state $s_r^k = \{P_{AB}, C, D\}$ corresponding to $\mathcal{S}_1$ of Fig. 2 (a) is illustrated in Fig. 2 (b) as a set of disjoint components over the four taxa $\{A, B, C, D\}$. Felsenstein's pruning algorithm is used to obtain a marginal likelihood estimate for each tree by passing messages from left and right child nodes and taking the inner product with η, the stationary state of $\mathbf{Q}$. Each distinct likelihood is then multiplied to assign probability $\pi(s_r^k)$ to partial state s_r^k.

Definition 2 (Natural Forest Extension). The natural forest extends target measure π into forests by taking a product over the trees in the forest:

$$\pi(s) := \prod_{(t_i, X_i)} \pi_{Y_i(x_i)}(t_i). \qquad (11)$$

The natural forest extension (NFE) has the advantage of passing information from the non-coalescing elements to the local weight update. Fig. 2 provides an illustration of the NFE applied to the state consisting of $PA(A, B)$ and non-coalescing singletons $\{C\}$ and $\{D\}$.

4 NESTED COMBINATORIAL SEQUENTIAL MONTE CARLO

A potential drawback of the CSMC method is that partial states are sampled to coalesce uniformly, when many of the resulting topologies correspond to areas of low probability mass. It seems natural to incorporate information from future iterations within the proposal distribution to subsequently guide the exploration of partial states. Adapting the proposal requires marginalizing the intermediate target over future topologies and branch lengths.

Locally Optimal Combinatorial SMC. Choosing a good proposal distribution is key for the effectiveness of SMC methods. The *locally optimal* SMC [Doucet et al., 2000, Naesseth et al., 2019b] chooses the proposal in such a way that all particles have equal weights. This can significantly improve the performance over the standard proposal used in CSMC. The locally optimal proposal based on the natural forest extension is

$$q(s_r | s_{r-1}) \propto \frac{\pi(s_r) \nu^-(s_{r-1})}{\pi(s_{r-1})}. \qquad (12)$$

This locally optimal proposal for the CSMC algorithm is computationally intractable, it requires us to exactly marginalize the branch lengths. We use the nested SMC [Naesseth et al., 2015, 2019a] method to overcome this problem.

Nested Combinatorial SMC. We provide an overview of Nested Combinatorial Sequential Monte Carlo before presenting a detailed description in Algorithm 1 (we have annotated the overview with steps from the algorithm). NCSMC iterates over rank events (*line 2*) to perform a standard RE-SAMPLE step also used in CSMC methods (*line 4*). For each sample, NCSMC enumerates all $\binom{N-r}{2}$ possible one-step ahead topologies and samples corresponding M sub-branch lengths (*line 7*). We evaluate importance *sub-weights* or *potential functions* for each of these $s_r^{k,m}[i]$ sampled look-ahead states (*line 8*). Then, we extend our ancestral partial state $s_{r-1}^{a_{r-1}^k}$ to the new partial state s_r^k (*line 11*) by selecting one of the topologies and a corresponding branch length according to its weight. Finally, for each sample (*line 12*), we compute its weight by averaging over all the potential functions. An illustration of the procedure is given in the Appendix.

Variational Nested CSMC Objective. The nested CSMC method described in Algorithm 1 can also be used to construct a variational objective:

$$\mathcal{L}_{NCSMC} := \mathbb{E}_Q \left[\log \hat{\mathcal{Z}}_{NCSMC} \right], \qquad (13)$$

$$\hat{\mathcal{Z}}_{NCSMC} := \prod_{r=1}^{R} \left(\frac{1}{K} \sum_{k=1}^{K} w_r^k \right). \qquad (14)$$

We refer to the resulting VI framework as VNCSMC.

Theoretical Justification. Nested CSMC is an SMC algorithm on the extended space of all random variables generated by Algorithm 1. This means it keeps keeps the favorable properties of CSMC, such as unbiasedness of the normalization constant estimate and asymptotic consistency. The key property that ensures this for NCSMC is *proper weighting* [Naesseth et al., 2015, 2019a].

Definition 3 (Proper Weighting). We say that the random pair (s_r, w_r) are *properly weighted* for the unnormalized distribution $\frac{\pi(s_r)\nu^-(s_{r-1})}{\pi(s_{r-1})}$ if $w_r \geq 0$ almost surely, and for all measurable functions h,

$$\mathbb{E}[w_r h(s_r)] = \int h(s_r) \frac{\pi(s_r)\nu^-(s_{r-1})}{\pi(s_{r-1})} \, \mathrm{d}s_r. \quad (15)$$

We formalize the result for NCSMC, Algorithm 1, in Theorem 1. We say that nested CSMC is an *exact approximation* [Naesseth et al., 2019b] of CSMC with the locally optimal proposal.

Theorem 1. *The particles s_r^k and weights w_r^k generated by Algorithm 1 are properly weighted for $\frac{\pi(s_r)\nu^-(s_{r-1})}{\pi(s_{r-1})}$.*

Proof.

$$\mathbb{E}[w_r^k h(s_r^k)] = \mathbb{E}\left[w_r^k \cdot h(s_r^{k,J}[I]) \right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{L} \sum_{j=1}^{M} w_r^k \frac{w_r^{k,j}[i]}{\sum_l \sum_m w_r^{k,m}[l]} h(s_r^{k,j}[i]) \right]$$

$$= \frac{1}{ML} \sum_{i=1}^{L} \sum_{j=1}^{M} \mathbb{E}\left[w_r^{k,j}[i] \cdot h(s_r^{k,j}[i]) \right]$$

$$= \mathbb{E}\left[w_r^{k,j}[i] \cdot h(s_r^{k,j}[i]) \right] = \int h(s_r) \frac{\pi(s_r)\nu^-(s_{r-1})}{\pi(s_{r-1})} \, \mathrm{d}s_r$$

$\square$

5 EXPERIMENTS

We evaluate VCSMC and VNCSMC on three tasks: (i) a standard dataset of primate mitochondrial DNA, (ii) on the complete 36 kilobase genomes of 17 species of Betacoronavirus, and (iii) on 7 large taxa benchmarks datasets ranging from 27 to 64 taxa. For experiments using the same initialization of likelihood and prior, the proposed methods converge to higher log-marginal likelihood values than existing methods. Additionally, they can be more easily adopted to a variety of models, with arbitrary settings of parameters θ. VCSMC and VNCSMC also scale well with the number of sites in input sequences.

Algorithm 1 Nested Combinatorial Sequential Monte Carlo

Input: $\mathbf{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$, $\theta = (\mathbf{Q}, \{\lambda_i\}_{i=1}^{|E|})$
1: Initialization. $\forall k, s_0^k \leftarrow \perp, w_0^k \leftarrow 1/K$.
2: **for** $r = 1$ **to** $R = N - 1$ **do**
3:    **for** $k = 1$ **to** K **do**
4:      RESAMPLE $\mathbb{P}(a_{r-1}^k = i) = \frac{w_{r-1}^i}{\sum_{l=1}^{K} w_{r-1}^l}$
5:      **for** $i = 1$ **to** $L = \binom{N-r}{2}$ **do**
6:        **for** $m = 1$ **to** M **do**
7:          FORM LOOK-AHEAD PARTIAL STATE

$$s_r^{k,m}[i] \sim q(\cdot | s_{r-1}^{a_{r-1}^k})$$

8:          COMPUTE POTENTIALS

$$w_r^{k,m}[i] = \frac{\pi(s_r^{k,m}[i])}{\pi(s_{r-1}^{a_{r-1}^k})} \cdot \frac{\nu^-(s_{r-1}^{a_{r-1}^k})}{q(s_r^{k,m}[i] | s_{r-1}^{a_{r-1}^k})}$$

9:      **end for**
10:      **end for**
11:    EXTEND PARTIAL STATE

$$s_r^k = s_r^{k,J}[I],$$

$$\mathbb{P}(I = i, J = j) = \frac{w_r^{k,j}[i]}{\sum_{l=1}^{L} \sum_{m=1}^{M} w_r^{k,m}[i]}$$

12:    COMPUTE WEIGHTS

$$w_r^k = \frac{1}{ML} \sum_{i=1}^{L} \sum_{m=1}^{M} w_r^{k,m}[i]$$

13:    **end for**
14: **end for**
   Output: $s_R^{1:K}$, $w_{1:R}^{1:K}$

Primate Mitochondrial DNA. We evaluate VCSMC on a benchmark dataset of nucleotide sequences of homologous fragments of primate mitochondrial DNA [Hayasaka et al., 1988]. The dataset consists of 12 taxa $\{S_0, \cdots, S_{11}\}$ over 898 sites admitting 13,749,310,575 distinct tree topologies. The set of taxa includes five species of homonoids, four species of old world monkeys, one species of new world monkey and two species of prosimians. VCSMC is run with $K = \{4, 8, 16, 32, 64, 128\}$ particles, whereas VNCSMC is run with $K = \{4, 8, 16, 32, 64, 128\}$ and $M = 1$ particles, each averaged over 5 random seeds. Fig. 3 shows higher values of K produce larger log-marginal likelihood values (tighter ELBO values) with lower stochastic gradient noise. VCSMC (blue) with $K \geq 16$ outperforms probabilistic path Hamiltonain Monte Carlo ($ppHMC$) shown (green trace) for comparison. VNCSMC (red) requires fewer epochs than VCSMC to converge and produces tighter ELBO / larger log-
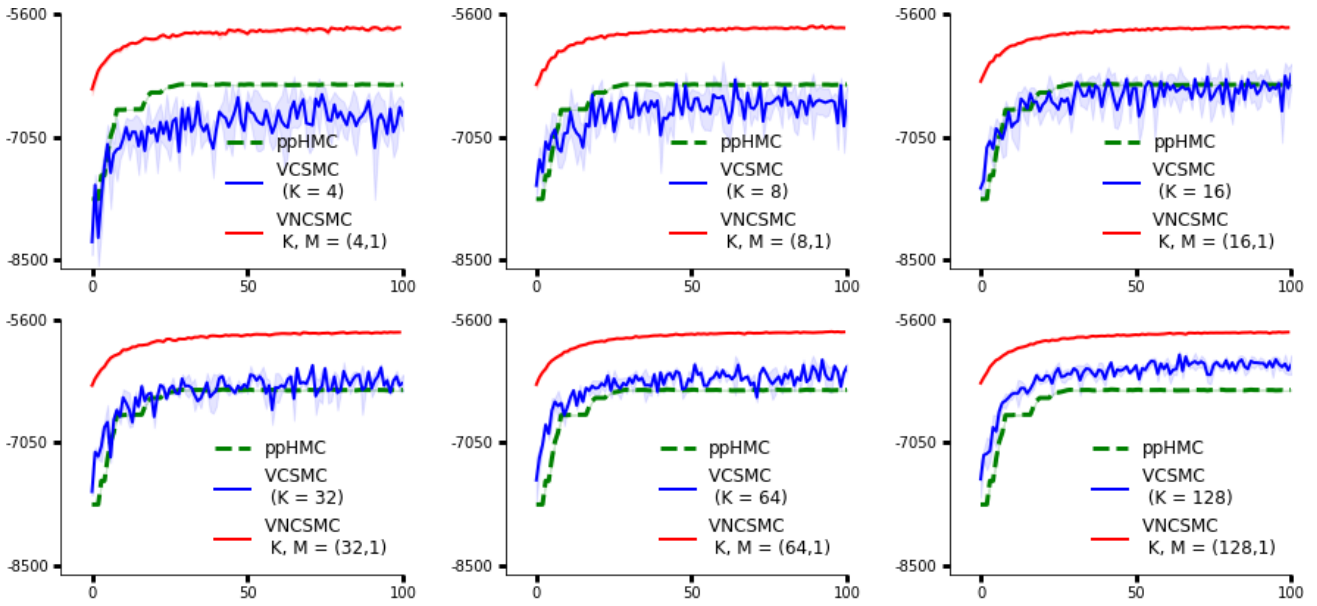
Figure 3: Log likelihood values for VCSMC (blue) with $K = \{4, 8, 16, 32, 64, 128\}$ samples and VNCSMC (red) with $K = \{4, 8, 16, 32, 64, 128\}$ and $M = 1$ samples on the primates data averaged across 5 random seeds. Higher values of K produce tighter ELBO / larger log likelihood values with lower stochastic gadient noise. VCSMC with $K \geq 16$ outperforms probabilistic path Hamiltonain Monte Carlo ($ppHMC$) which is shown (green trace) for comparison. VNCSMC requires fewer epochs than VCSMC to converge and produces tighter ELBO / larger log likelihood values with lower stochastic gadient noise. VNCSMC with $(K, M) = (4, 1)$ (top left) outperforms both $ppHMC$ and VCSMC with $K = 128$ (bottom right).

(a) MrBayes phylogeny

(b) VNCSMC phylogeny

Figure 4: MrBayes vs VNCSMC phylogeny on the primate mitochondrial DNA dataset. The data consists of 12 taxa $\{S_0, \cdots, S_{11}\}$ over 898 sites on the genome. The maximum likelihood topology returned by VNCSMC corresponds to that of Mr Bayes. The bottom clade partitions monkeys, while the central and top clades partition hominids and prosimians. MrBayes uses 20,000 iterations of MCMC in contrast to VNCSMC which uses 256 samples.

marginal likelihood values with lower stochastic gadient noise. VNCSMC with $(K, M) = (4, 1)$ (top left) outperforms both $ppHMC$ and VCSMC with $K = 128$ (bottom right).

Fig. 4 provides a single maximum likelihood phylogeny selected from a run of VNCSMC using $K, M = (256, 1)$ particles, along with a phylogeny from Mr Bayes on the same dataset. The topology returned by VNCSMC corre-

Table 1: Log-marginal likelihood estimates of different variational inference techniques across 7 benchmark datasets for Bayesian phylogenetic inference. Results reported by VBPI [Zhang and Matsen IV, 2019] and VBPI-NF [Zhang, 2020] were obtained by (i) using 10 replicates of 10,000 maximum likelihood bootstrap trees [Minh et al., 2013] to obtain topologies defining the support of the conditional probability tables and (ii) performing 400,000 parameter updates. VCSMC does not require bootstrapped or MCMC tree topologies in order to learn parameters. We give VCSMC 2048 particles and evaluate the likelihood after 100 parameter updates. Results for VCSMC and VCSMC (JC) are averaged over three random seeds. VCSMC consistently explores higher probability phylogenies than VBPI and VBPI-NF without the use of preloaded topologies.

Dataset	Reference	# Taxa (N)	# Sites (S)	VBPI	VBPI-NF	VCSMC	VCSMC (JC)
				Log Marginal Likelihood			
DS1	Hedges et al. [1990]	27	1949	-7108.4	-7108.4	-5929.8	-6906.7
DS2	Garey et al. [1996]	29	2520	-26367.7	-26367.7	-14160.4	-23252.6
DS3	Yang and Yoder [2003]	36	1812	-33735.1	-33735.1	-17460.5	-33177.8
DS4	Henk et al. [2003]	41	1137	-13329.9	-13329.9	-11251.9	-12232.6
DS5	Lakner et al. [2008]	50	378	-8214.5	-8214.5	-5797.1	-7921.2
DS6	Zhang and Blackwell [2001]	50	1133	-6724.3	-6724.3	-5216.5	-6575.5
DS7	Rossman et al. [2001]	64	1008	-8650.6	-8650.4	-5847.5	-6781.5

sponds to that of Mr Bayes. The bottom clade partitions monkeys, while central and top clades partition hominids and prosimians respectively.

Betacoronavirus Data. The evolutionary origin of SARS-COV-II and the development of its genetic variants is an open question of paramount importance in both virology and in public health. At a high level, the species SARS-COV-II belongs to the genera of betacoronaviruses, which include OC43 and HKU1 (which cause the common cold) of lineage A, SARS-COV and SARS-COV-II (which causes the disease COVID-19) of lineage B, and MERS-COV-II (which causes the disease MERS) of lineage C [Boni et al., 2020]. The exact origin of SARS-COV-II however is unknown; different approaches to phylogenetic inference produce statistically incompatible results [Pipes et al., 2020]. Coronaviruses have relatively large genomes ranging from 26-32 kilobases, and performing analyses on the full genomes is often a challenge. Recently, it has been argued that viral recombination in betacoronaviruses often encompasses the receptor binding domain (RBD) of the spike gene [Patiño-Galindo et al., 2020]. This process is thought to have produced a recombination event at least 11 years ago in an ancestor of SARS-COV-II [Patiño-Galindo et al., 2020]. We use VNCSMC to analyze the complete genomes for 17 species of Betacoronavirus downloaded from the NCBI Viral Genomes Resource [Brister et al., 2014]. Multiple Sequence Aligmnent using Clustal was performed and each nucleotide was one-hot encoded as a vector, producing input sequences with 36,889 sites. Fig. 6 of the Appendix provides the maximum likelihood phylogeny from a VNCSMC run using $K, M = (256, 1)$ particles. The result shows that the phylogeny partitions four lineages into clades: Embecovirus (*lineage A*), Sarbecovirus (*lineage B* including SARS-COV and SARS-COV-II), Merbecovirus (*lineage C*), and Nobecovirus (*lineage D*) [Cotten

et al., 2013, Woo et al., 2010, Geldenhuys et al., 2018].

Large Taxa Benchmarks. We evaluate VCSMC on 7 large benchmark datasets for Bayesian phylogenetic inference [Hedges et al., 1990, Garey et al., 1996, Yang and Yoder, 2003, Henk et al., 2003, Lakner et al., 2008, Zhang and Blackwell, 2001, Rossman et al., 2001]. Each dataset ranges from 27 to 64 eukaryote species with 378 to 2520 sites. Table 1 provides the marginal likelihood values for various methods. VBPI [Zhang and Matsen IV, 2019] and VBPI-NF [Zhang, 2020] both learn a simplified model of molecular evolution referred to as Jukes-Cantor (JC), which fixes the transition matrix [Jukes and Cantor, 1969]. For a fair comparison, we report VCSMC (JC) results in addition to the harder task of also learning the transition matrix. Results reported by VBPI and VBPI-NF were obtained by (i) using 10 replicates of 10,000 maximum likelihood bootstrap trees [Minh et al., 2013] to obtain topologies defining the support of the conditional probability tables and (ii) performing 400,000 parameter updates. Without bootstrap trees, the conditional probability tables for VBPI and VBPI-NF scale exponentially with the number of taxa [Zhang and Matsen IV, 2019]. VCSMC does not restrict the support of the tree topologies and instead leverages CSMC to compute a lower bound. We give VCSMC 2048 particles and evaluate the likelihood after 100 parameter updates, averaged over three random seeds. Both VCSMC and VCSMC (JC) explore higher probability spaces than VBPI and VBPI-NF.

Empirical Running Times. We report the empirical running times of VCSMC and VNCSMC on the primates dataset and highlight the results in Table 2 of the Appendix. Experiments were performed on a 2.4GHz 8-core Intel i9 processor Macbook pro with 64 GB memory and no GPU utilization. We note that alternative methods are designed for solving

simpler problems in both inference and learning making any runtime comparisons indirect. For instance, VBPI and VBPI-NF use precomputed topologies, while $ppHMC$ support Jukes-Cantor models. VCSMC runs on the primates dataset at an average speed of 19.34 iterations per second (*it/s*) with $K = 4$ and an average of 2.25 seconds per iteration (*s/it*) with $K = 256$. VNCSMC runs in 3.89 seconds per iteration with $K = 4$ and 21.77 seconds per iteration with $K = 256$. These numbers can be improved by leveraging GPU utilization. In contrast, MrBayes in Figure 4 takes 12 seconds with 20,000 iterations, and the minimum it would take to converge on the primates dataset is $\sim$2,000 iterations, implying a runtime of $\sim$1.2 seconds. We observe that the first epoch of VCSMC and VNCSMC is equivalent to the inference task and runs faster than MrBayes.

6 DISCUSSION

Computational Complexity. The locally optimal proposal in NCSMC requires additional computational complexity to marginalize the intermediate target densities in exchange for a more informed exploration of partial states. NCSMC costs $\mathcal{O}(KN^3M)$ in contrast to $\mathcal{O}(KNM)$ for CSMC. Empirically, NCSMC with small K, M produces a more accurate posterior approximation than CSMC with larger K (see Fig. 3). NCSMC can accommodate a large number of particles with low memory overhead, however maintaining the computational graph and applying the sum-product algorithm symbolically for each of the K samples and M sub-samples, along with evaluating gradients for each of these terms places a practical restriction on the values of K, M and N used with VNCSMC without GPU utilization. Alternative implementations of Bayesian phylogenetic inference are computationally intensive. While the process of enumerating the $\binom{N-r}{2}$ topologies across rank events cannot be avoided, we find that choosing K as large as possible and $M = 1$ is a useful heuristic for producing good results. For example, $K, M = (256, 1)$ can be run on the betacoronavirus data with $N = 17$ and 36,889 sites (see Fig. 6 without GPU utilization. One advantage of VCSMC and VNCSMC is the ability to use minibatch iteration to speed up training. The experiments were trained using ADAM with a batch size $B = S/4$. Opportunities exist to parallelize VCSMC and leverage GPU optimization which we expect would produce significant performance gains on DS1-DS7 as K increases.

Effective Sample Size. One pertinent theoretical question concerns the relationship between the effective sample size (ESS), the number of samples K and the number of taxa N. The ESS measures the diversity among samples and is defined as $\text{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$ where w_i are the unnormalized weights. We report ESS values on the primates data in Table 2 of the Appendix. While an ESS close to K is not sufficient to ensure a good approximation, it is

a necessary condition. We find near optimal ESS values across all choices of K for both VCSMC and VNCSMC. The theoretical foundations for developing lower bounds on ESS for a given value of K have only been developed in the context of *online inference*, where the posterior distribution is updated as new sequence data becomes available [Dinh et al., 2017b]. We leave theoretical questions of ESS and online extensions of VCSMC for future work.

Contacts Outside of Phylogenetic Inference. VCSMC and VNCSMC may be adapted to a wide class of problems outside of phylogenetic inference. In principle, any generative model of data simulated by a Markov tree can be fit using VCSMC and VNCSMC. Coalescent models for heirarchical Bayesian clustering and diffusion trees [Teh et al., 2009, Boyles and Welling, 2012, Knowles and Ghahramani, 2011] are examples of alternative probabilistic approaches involving distributions over latent trees that may be suited for VCSMC. The nested CSMC algorithm may also be used to simulate approximate solutions to other combinatorial optimization tasks. Combinatorial Monte Carlo methods are used to approximate the number of self-avoiding random walks on the lattice [Sokal, 1996, Shirai and Kikuchi, 2013]. Another point of contact is the reconstruction of jet structures in particle physics for the analysis of data from experiments at the Large Hadron Collider at CERN [Höche, 2014]. Jet reconstruction algorithms are typically based on greedy approximation methods [Cacciari et al., 2008, Dokshitzer et al., 1997], however VCSMC and VNCSMC may be particularly suited for this domain. The aforementioned extensions are open directions for further development.

Conclusion. We have introduced VCSMC, a powerful framework for both inference and learning in Bayesian phylogenetics. VCSMC is the first method to establish the use of variational sequential search to learn distributions over intricate combinatorial structures, uncovering connections between VI and SMC. We have introduced NCSMC, and proved that it provides an exact approximation to the locally optimal proposal for CSMC. We have used NCSMC to define a second objective, VNCSMC which yields tighter lower bounds than VCSMC. VCSMC and VNCSMC outperform existing methods on a range of tasks. A TensorFlow implementation of both VCSMC and VNCSMC is available online at https://github.com/amoretti86/phylo.

Author Contributions

Antonio Khalil Moretti and Liyi Zhang contributed equally to this paper.

Acknowledgements

References

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3): 269–342, 2010.

M. Boni, P. Lemey, X. Jiang, T. Lam, B. Perry, T. Castoe, A. Rambaut, and D. Robertson. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology*, 5(11): 1408–1417, Nov. 2020.

A. Bouchard-Côté, S. Sankararaman, and M. Jordan. Phylogenetic inference via sequential Monte Carlo. *Systematic biology*, 61:579–93, 01 2012.

L. Boyles and M. Welling. The time-marginalized coalescent prior for hierarchical clustering. In *Advances in Neural Information Processing Systems*, 2012.

J. R. Brister, D. Ako-adjei, Y. Bao, and O. Blinkova. NCBI Viral Genomes Resource. *Nucleic Acids Research*, 43 (D1):D571–D577, 11 2014.

M. Cacciari, G. P. Salam, and G. Soyez. The anti-ktjet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

M. Cotten, T. T. Lam, S. J. Watson, A. L. Palser, V. Petrova, P. Grant, O. G. Pybus, A. Rambaut, Y. Guan, D. Pillay, P. Kellam, and E. Nastouli. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. *Emerging infectious diseases*, 19(5):736–42, 2013.

A. Dhar, D. K. Ralph, V. N. Minin, and F. A. Matsen. A bayesian phylogenetic hidden Markov model for B cell receptor sequence analysis. *PLOS Computational Biology*, 16(8):e1008030, Aug 2020.

V. Dinh, A. Bilge, C. Zhang, and F. A. Matsen, IV. Probabilistic path Hamiltonian Monte Carlo. volume 70 of *Proceedings of Machine Learning Research*, pages 1009–1018, International Convention Centre, Sydney, Australia, 06–11 Aug 2017a. PMLR.

V. Dinh, A. E. Darling, and F. A. Matsen IV. Online Bayesian Phylogenetic Inference: Theoretical Foundations via Sequential Monte Carlo. *Systematic Biology*, 67 (3):503–517, 12 2017b.

Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber. Better jet clustering algorithms. *Journal of High Energy Physics*, 1997(08):001–001, Aug 1997.

A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

J. Garey, T. Near, M. Nonnemacher, and S. Nadler. Molecular evidence for Acanthocephala as a subtaxon of Rotifera. *Journal of molecular evolution*, 43(3):287–292, 1996.

M. Geldenhuys, M. Mortlock, J. Weyer, O. Bezuidt, E. C. J. Seamark, T. Kearney, C. Gleasner, T. H. Erkkila, H. Cui, and W. Markotter. A metagenomic viral discovery approach identifies potential zoonotic and novel mammalian viruses in Neoromicia bats within South Africa. *PLOS ONE*, 13(3):1–27, 03 2018.

K. Hayasaka, T. Gojobori, and S. Horai. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, 5(6):626–644, 11 1988.

S. Hedges, K. Moberg, and L. Maxson. 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, 11 1990.

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

J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees . *Bioinformatics*, 17(8): 754–755, 08 2001.

S. Höche. Introduction to parton-shower event generators, 2014.

T. H. Jukes and C. R. Cantor. Chapter 24 - evolution of protein molecules. In H. MUNRO, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes, 2013.

D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534 – 1552, 2011.

C. Lakner, P. van der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. *Systematic Biology*, 57(1):86–103, 02 2008.

D. Lawson, G. Tucker, C. A. Naesseth, C. Maddison, R. P. Adams, and Y. W. Teh. Twisted variational sequential Monte Carlo. *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood. Auto-encoding sequential Monte Carlo. In *International Conference on Learning Representations*, 2018.

C. J. Maddison, D. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. W. Teh. Filtering variational objectives. 2017.

B. Q. Minh, M. A. T. Nguyen, and A. von Haeseler. Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution*, 30(5):1188–1195, 02 2013.

A. Moretti, L. Zhang, and I. Pe'er. Variational combinatorial sequential Monte Carlo for Bayesian phylogenetic inference. *Machine Learning in Computational Biology*, Nov 2020a.

A. K. Moretti. *Variational Bayesian Methods for Inferring Spatial Statistics and Nonlinear Dynamics*. PhD thesis, 2021.

A. K. Moretti, Z. Wang, L. Wu, I. Drori, and I. Pe'er. Particle smoothing variational objectives. *CoRR*, abs/1909.09734, 2019a.

A. K. Moretti, Z. Wang, L. Wu, and I. Pe'er. Smoothing nonlinear variational objectives with sequential Monte Carlo. *ICLR Workshops*, 2019b.

A. K. Moretti, Z. Wang, L. Wu, I. Drori, and I. Pe'er. Variational objectives for Markovian dynamics with backward simulation. *European Conference on Artificial Intelligence*, 2020b.

C. A. Naesseth, F. Lindsten, and T. B. Schön. Nested sequential Monte Carlo methods. In *International Conference on Machine Learning (ICML)*, pages 1292–1301, 2015.

C. A. Naesseth, S. Linderman, R. Ranganath, and D. Blei. Variational sequential Monte Carlo. volume 84 of *Proceedings of Machine Learning Research*, 2018.

C. A. Naesseth, F. Lindsten, and T. B. Schön. High-dimensional filtering using nested sequential Monte Carlo. *IEEE Transactions on Signal Processing*, 67(16):4177–4188, 2019a.

C. A. Naesseth, F. Lindsten, and T. B. Schön. Elements of sequential Monte Carlo. *Foundations and Trends® in Machine Learning*, 12(3):307–392, 2019b.

C. A. Naesseth, F. Lindsten, and D. Blei. Markovian score climbing: Variational inference with KL(p‖q). In *Advances in Neural Information Processing Systems*, 2020.

J. Á. Patiño-Galindo, I. Filip, M. AlQuraishi, and R. Rabadan. Recombination and lineage-specific mutations led to the emergence of SARS-CoV-2. 2020.

L. Pipes, H. Wang, J. P. Huelsenbeck, and R. Nielsen. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. *Molecular Biology and Evolution*, 12 2020.

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

C. Semple and M. Steel. Phylogenetics. 2003.

N. C. Shirai and M. Kikuchi. How to estimate the number of self-avoiding walks over 10100? use random walks. *Interdisciplinary Information Sciences*, 19(1):79–83, 2013.

A. D. Sokal. Monte carlo methods for the self-avoiding walk. *Nuclear Physics B - Proceedings Supplements*, 47 (1):172–179, 1996.

Y. W. Teh, H. D. III, and D. Roy. Bayesian agglomerative clustering with coalescents, 2009.

L. Wang, A. Bouchard-Côté, and A. Doucet. Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *Journal of the American Statistical Association*, 01 2015.

S. Wang and L. Wang. Particle Gibbs sampling for Bayesian phylogenetic inference, 2020.

P. C. Woo, Y. Huang, S. K. Lau, and K. Y. Yuen. Coronavirus genomics and bioinformatics analysis. *Viruses*, 2(8):1804–1820, 2010.

Z. Yang and A. Yoder. 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, 10 2003.

C. Zhang. Improved variational Bayesian phylogenetic inference with normalizing flows, 2020.

C. Zhang and F. A. Matsen IV. Generalizing tree probability estimation via Bayesian networks. In *Advances in Neural Information Processing Systems*. 2018.

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

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