Massively scalable inference of level-1 phylogenetic networks

Nathan Kolbow[1,2], Sungsik Kong[1], Claudia Solis-Lemus[1,2,3]

[1]Wisconsin Institute for Discovery, University of Wisconsin-Madison

[2]Department of Statistics, University of Wisconsin-Madison

[3]Department of Plant Pathology, University of Wisconsin-Madison

Abstract

Recent advancements in sequencing technologies have enabled large-scale phylogenomic analyses. While these analyses often rely on phylogenetic trees, increasing evidence suggests that non-treelike evolutionary events, such as hybridization and horizontal gene transfer, are prevalent in the evolutionary histories of many species, and in such cases, tree-based models are insufficient. Phylogenetic networks can capture such complex evolutionary histories, but current methods for accurately inferring them lack scalability. For instance, state-of-the-art model-based approaches are limited to around 30 taxa. Implicit network inference methods like NeighborNet and Consensus Networks are fast but lack biological interpretability. Here, we introduce a novel method called InPhyNet that merges a set of non-overlapping, independently inferred networks into a unified topology, achieving linear scalability while maintaining high accuracy under the multispecies network coalescent model. Our simulations show that InPhyNet matches the accuracy of SNaQ on datasets with 30 taxa while drastically decreasing the overall network inference time. InPhyNet is also more accurate than implicit network methods on large datasets while maintaining computational feasibility. Re-analyzing a phylogeny of 1,158 land plants with InPhyNet, we recover known reticulate events and provide evidence for the controversial placement of Order Gnetales within gymnosperms. These results demonstrate that InPhyNet enables biologically meaningful network inference at previously unprecedented scales.

1 Introduction

dvancements in sequencing technologies have increased the availability of biological genomic data for phylogenetic analyses. These analyses are foundational to our understandings of many bioloigcal processes such as cancerous tumour progression [35], infectious disease spread [9], drug discovery [36], conservation biology [43], and crop domestication in agriculture [34]. To date, the backbone of these analyses has been phylogenetic trees: binary tree models that represent evolutionary relationships based on two descendant species sharing a single common ancestor. However, growing bodies of evidence suggest that some evolutionary events, such as hybridization and horizontal gene transfer, are prevelant in the histories of many species [4]. These phenomena diverge from the strictly bifurcating structure enforced by phylogenetic trees, rendering such trees inadequate for modeling them.

Phylogenetic networks can accurately model non-treelike, reticulate evolution, but are much more computationally intensive to infer due to their increased model complexity, higher-dimensional parameter space, and a larger set of possible topologies. State-of-the-art maximum pseudo-likelihood (also sometimes referred to as composite likelihood) inference methods such as Species Network applying Quartets (SNaQ) [40], Phylogenetic Network Estimation using SiTe patterns (PhyNEST) [22], or PhyloNet's maximum pseudolikelihood (MPL) option [46] offer reasonable speeds and high accuracies but can only handle upwards of 30 taxa. More recent research to increase this upper bound has been fruitful, leading to methods like FastNet [17] or PhyloNet's divide-and-conquer option [45], the latter of which can handle up to 81 taxa. However, these methods still have the fundamental issue of polynomial runtime complexity.

This runtime issue can be circumvented by inferring implicit networks, a class of networks that do not directly model biological phenomena. Methods for inferring implicit networks such as NeighborNet [7], Consensus Networks [18], or PhyloFusion [48] are often employed when researchers wish to analyze hundreds or thousands of taxa with a network model [16]. However, such analyses are often inappropriate given their lack of biological grounding [23, 42].

Here, we present InPhyNet, a novel method for constructing arbitrarily large semi-directed phylogenetic networks. We then use InPhyNet to develop a phylogenetic network inference pipeline that scales *linearly* with respect to the number of taxa in the network. Additionally, we show that such a pipeline can be statistically consistent for the underlying species network under the multi-species network coalescent model. The practicality of this theory is verified with a rigorous simulation study that evaluates InPhyNet's performance under realistically complex sampling conditions. We compare our method against SNaQ [40] in small simulation settings, and against NeighborNet [7] and Consensus Networks [18] in the SplitsTree software [19] in larger simulation settings where more accurate methods are unavailable.

Lastly, we re-evaluate the phylogeny of 1,158 land plants [29] by inferring reticulations in portions of the phylogeny that were poorly resolved in their estimated phylogenetic tree. We found support for previously well known instances of reticulate evolution as well as evidence for a controversial placement of Gnetales, a division of plant, within gymnosperms. Our method is available as open-source software at https://github.com/NathanKolbow/InPhyNet.jl and all scripts and results used in the simulations presented here are available at https://github.com/NathanKolbow/InPhyNetSimulations.

2 Definitions

Definition 1. A **semi-directed, binary, labelled, metric phylogenetic network** $(N, \{l_e\}_{e \in E}, \{\gamma_e\}_{e \in E_H})$ on taxa set X is a connected, acyclic graph with nodes $V = V_L \cup V_H \cup V_T$ and edges $E = E_T \cup E_H$ with corresponding edge lengths $\{l_e\}_{e \in E}$, inheritance parameters $\{\gamma_e\}_{e \in E_H}$, and the following characteristics:

1. The set of *leaf nodes* V_L map bijectively to X.

2. Each leaf node $v \in V_L$ has in-degree 1 and out-degree 0.

3. Each *tree node* $v \in V_T$ has in-degree 1 and out-degree 2.

4. Each *hybrid node* $v \in V_H$ has in-degree 2 and out-degree 1.

5. A *tree edge* $e = (u, v) \in E_T$ is an undirected edge connecting u and v.

6. A *hybrid edge* or *reticulation* $e = (u, v) \in E_H$ is a directed edge pointing from u to v where v is a hybrid node.

7. Each tree edge $e \in E_T$ has length $0 < l_e < \infty$.

8. Each hybrid edge $e \in E_H$ has length $0 \leq l_e < \infty$ and hybrid parameter $0 < \gamma_e < 1$.

9. Each hybrid node has two incoming hybrid edges e, e' where $\gamma_e + \gamma_{e'} = 1$.

Inheritance parameters γ_e describe the proportion of genes that are inherited through edge e. The inheritance parameters $\gamma_e + \gamma'_e = 1$ for a pair of hybrid edges e, e' that point to the same hybrid node because each gene tree generated according to network N must be inherited through exactly one of e or e'.

Definition 2. An **unrooted, binary, labelled, metric phylogenetic tree** $(T, \{l_e\}_{e \in E})$ is a semi-directed, binary, labelled, metric phylogenetic network network where $E_H = \emptyset$.

For brevity, we refer to these as "unrooted trees" and "semi-directed networks" or simply as "trees" and "networks". Additionally, because a tree is a subclass of a network, statements about networks may be generally applied to trees as well. Specifically, unless explicitly specified, sets of networks may in fact consist in part, or entirely, of trees.

Definition 3. The **displayed trees** of network $N = (V, E)$ are the set of $2^{|V_H|}$ possible trees obtained by, for each hybrid node $v \in V_H$ with incoming hybrid edges e and e' and removing one e or e'. Such a tree may have trivial nodes of in-degree 1 and out-degree 1, so these trees are subsequently simplified to comply with Definition 2 by removing such nodes and "fusing" each of their adjacent edges.

See Figure 1 for an example of a semi-directed network and some of its displayed trees. Note that for a tree T, $2^{|V_H|} = 1$, and we say that T is its own displayed tree.

Definition 4. A **path** on $G = (V, E)$ is an ordered set of nodes $\{v_i\}_{i=1}^{k}$ where each $v_i \in V$ is unique, $v_1 = u$, $v_k = w$, and $(v_j, v_{j+1}) \in E$ for all $j = 1, 2, \ldots, k-1$. The set of all such paths in G is given by $\mathbf{p} = u \overset{G}{\rightsquigarrow} w$, or simply $u \rightsquigarrow w$ when there is no ambiguity in G. Two nodes u, w are **neighbors** in the unrooted tree T if there exists a path $p = u \rightsquigarrow w$ with $|p| \leq 3$. Additionally, u, w are neighbors in the semi-directed network N if the previous statement is true for *any displayed tree of* N.

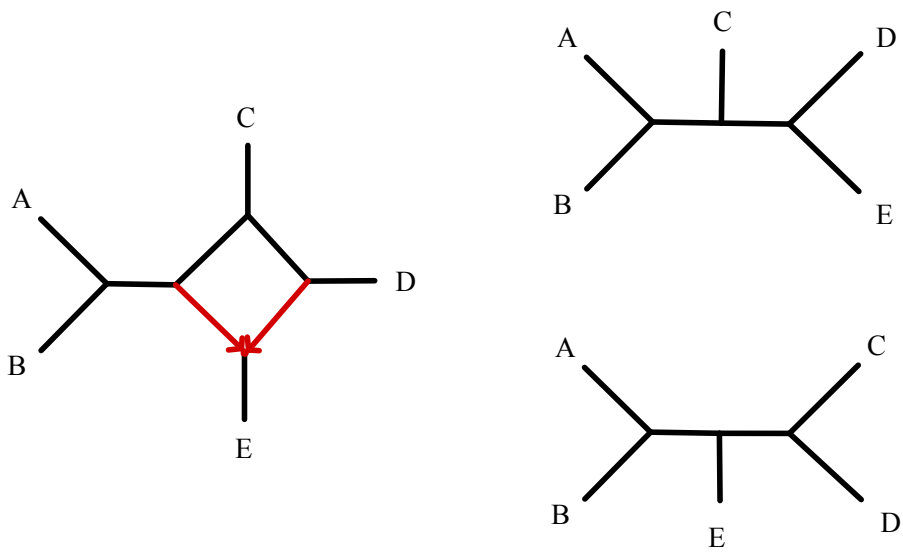

Figure 1: A semi-directed network and its two displayed trees. Red nodes and edges are hybrid nodes and edges.

The condition $|p| \leq 3$ is necessary as opposed to $|p| = 3$ to cover the case where $T = (V, E)$ is a trivial 2-taxa tree with $V = \{u, v\}$ and $E = \{(u, v)\}$. In this case, u and v are clearly neighbors, but their only path in T has length 2.

Definition 5. The path $p^*_{N_k}(L_i, L_j)$ for nodes L_i, L_j in the semi-directed network N_k (just p^* when unambiguous) is the shortest path belonging to $\mathbf{p} = L_i \overset{N_k}{\rightsquigarrow} L_j$, unless $\mathbf{p}$ contains a path of length 4 and a path of length 5. Then, $p^*_{N_k}(L_i, L_j)$ is the path of length 5. This exception is in place specifically for case (5) in Figure 2.

Definition 6. A **level-k network** N is a semi-directed network in which each biconnected component contains at most k hybrid nodes.

Definition 7. Given n gene trees $\mathcal{G}_n$ from species network $\mathcal{N}$ following the multi-species network coalescent (MSNC) model, M_D is statistically consistent for T, an arbitrary displayed tree of $\mathcal{N}$, if

$$Q_i^{(T)}\left(M_D(\mathcal{G}_n)\right) \xrightarrow{p} c_i < 0, \text{ for } i = 1, 2 \text{ as } n \to \infty$$

Given n gene trees $\mathcal{G}_n$ from species network $\mathcal{N}$ following the MSNC, M_C is statistically consistent for $\mathcal{N}$ if

$$M_C(\mathcal{G}_n) \xrightarrow{p} \mathcal{N} \text{ as } n \to \infty$$

Definition 8. For network $\mathcal{N}$ on taxa Y, a subnetwork **spanned** by $X \subseteq Y$ is the minimal non-reticulated subgraph of $\mathcal{N}$ necessary to connect all taxa in X plus any reticulations in $\mathcal{N}$ that are fully contained within

this subgraph. A network N on taxa X is said to **agree with** network $\mathcal{N}$ on taxa Y if $X \subseteq Y$ and N is topologically equivalent to the subnetwork of $\mathcal{N}$ spanned by X.

3 Model

3.1 The InPhyNet Algorithm

InPhyNet takes a pairwise dissimilarity matrix D on a set of taxa X and a set of binary level-1 networks (hereafter "constraint networks") $\mathbf{N} = \{N_i\}_{i=1}^k$ on disjoint subsets of taxa $\mathbf{S} = \{S_i\}_{i=1}^k \subseteq X$ as input, and outputs a network $\widehat{\mathcal{N}}$ on X. The approach of InPhyNet stems from TreeMerge [28] which itself is derived from Neighbor-Joining [33]. InPhyNet begins with a network $\widehat{\mathcal{N}}$ containing n terminal vertices (i.e., taxa) corresponding to each entry in X, without any edges or internal vertices. Then, the algorithm iteratively conducts the following steps:

1. Compute Q, the rate-adjusted pairwise dissimilarity matrix, from D, as described in [33].
2. Select the two most similar nodes L_i and L_j in $\widehat{\mathcal{N}}$ according to Q that (a) do not have parent nodes, (b) are siblings in all $N_k \in \mathbf{N}$ where $L_i, L_j \in N_k$, and (c) satisfy the validity heuristic from [27].
3. Connect L_i and L_j in $\widehat{\mathcal{N}}$ to a new, shared parent node L_p.
4. For each constraint network $N_k \in \mathbf{N}$, if portions of any reticulations occur along the path $L_i \overset{N_k}{\rightsquigarrow} L_j$, reflect those relationships in $\widehat{\mathcal{N}}$ (see subsection Retaining Reticulate Relationships below).
5. For each constraint network $N_k \in \mathbf{N}$, if only $L_i \in N_k$, rename L_i and L_p. If only $L_j \in N_k$, rename L_j as L_p. If both $L_i, L_j \in N_k$, find the path $p^*_{N_k}(L_i, L_j)$ in N_k (see Definition 11 below) and remove all nodes and edges in p^* except L_p.
6. Add approximate entries for the new hypothetical taxa L_p to D (see subsection Approximate Taxa Imputation below) and remove entries for L_i and L_j from D.
7. Repeat (1)-(6) until the only parent-less node in $\widehat{\mathcal{N}}$ is the root.

Finally, $\widehat{\mathcal{N}}$ is a semi-directed network (i.e. a network where all reticulations are directed towards their respective hybrid node and all other edges are undirected) so the root is suppressed. Note that the set of level-1 networks used as inputs *constrain* the set of possible node pairs in condition (b) of step 1, so we refer to them as "constraint networks". Further details on each step of this algorithm can be found in the Appendices.

3.2 Retaining Reticulate Relationships

Definition 9. A **displayed tree** T of a semi-directed network N is any unrooted tree that can be obtained by, for each pair of reticulation edges [11] $r_{\text{major}}, r_{\text{minor}}$ incident to hybrid node v_H, conducting the following:

1. Remove either r_{major} or r_{minor} from N.

2. Undirect the reticulation edge that was not removed.
3. Remove any nodes remaining in N with two attached edges and fuse these two edges into a single edge.

Definition 10. A **path** from node u to node w in the semi-directed network N with vertices and edges (V, E) is an ordered set of nodes $\{v_i\}_{i=1}^{k}$ where all $v_i \in V$ are unique, $v_1 = u$, $v_k = w$, and E contains an edge between (v_j, v_{j+1}) for each $j = 1, 2, \ldots, k-1$. The set of all such paths in N is given by $\mathbf{p} = u \overset{N}{\rightsquigarrow} w$. Two nodes are **siblings** in N if there exists some path $p \in u \overset{T}{\rightsquigarrow} w$ with $|p| \leq 3$ in any displayed tree T of N.

Definition 11. The path $p^*_{N_k}(L_i, L_j)$ for nodes L_i, L_j in the semi-directed network N_k (just p^* when unambiguous) is the shortest path belonging to $\mathbf{p} = L_i \overset{N_k}{\rightsquigarrow} L_j$, unless $\mathbf{p}$ contains a path of length 4 and a path of length 5. Then, $p^*_{N_k}(L_i, L_j)$ is the path of length 5. This exception is in place specifically for case (5) in Figure 2.

We say that each p^* contains a node L_p as depicted in Figure 2. This node is the point of reference that allows reticulations to be properly placed in $\widehat{\mathcal{N}}$.

When two nodes L_i, L_j are connected in the inferred network $\widehat{\mathcal{N}}$, we reflect this in each constraint network N_k by collapsing L_i and L_j into L_p along the path $p^*_{N_k}(L_i, L_j)$. In doing so, we simplify the constraint networks and reflect the new structure of the inferred network. The neighbor relationship between L_i and L_j is well represented in $\widehat{\mathcal{N}}$ at this point, but reticulation edges incident to any portion of p^* may not be.

Notice that when p^* traverses through a reticulation, the parent edges of L_i and L_j in $\widehat{\mathcal{N}}$ can correspond to arcs that contain multiple edges in the constraint N_k. In order to accurately reflect this structure in $\widehat{\mathcal{N}}$, we adapt $\widehat{\mathcal{N}}$ as shown in Figure 2 based on the structure of p^*. After these operations, the reticulate structure of N_k is accurately reflected in $\widehat{\mathcal{N}}$.

3.3 Approximate Taxa Imputation

After two nodes L_i, L_j are connected in the inferred network $\widehat{\mathcal{N}}$ by a new, shared parent node L_p, subsequent iterations of InPhyNet will consider L_p in step (1) above but will not consider L_i or L_j anymore. InPhyNet quantifies the similarity of two nodes based on their entries in D, so we need to approximate dissimilarities for L_p relative to all other remaining nodes in order to proceed.

We approximate these distances with the metric presented in [33] and given in Equation 1. Here, p is the new index in D corresponding to the new parent node, i and j are the existing entries corresponding to nodes L_i and L_j, and o is an arbitrary index corresponding to all other indices in D.

$$D_{p,o} = D_{o,p} = \frac{D_{o,i} + D_{o,j} - D_{i,j}}{2} \tag{1}$$

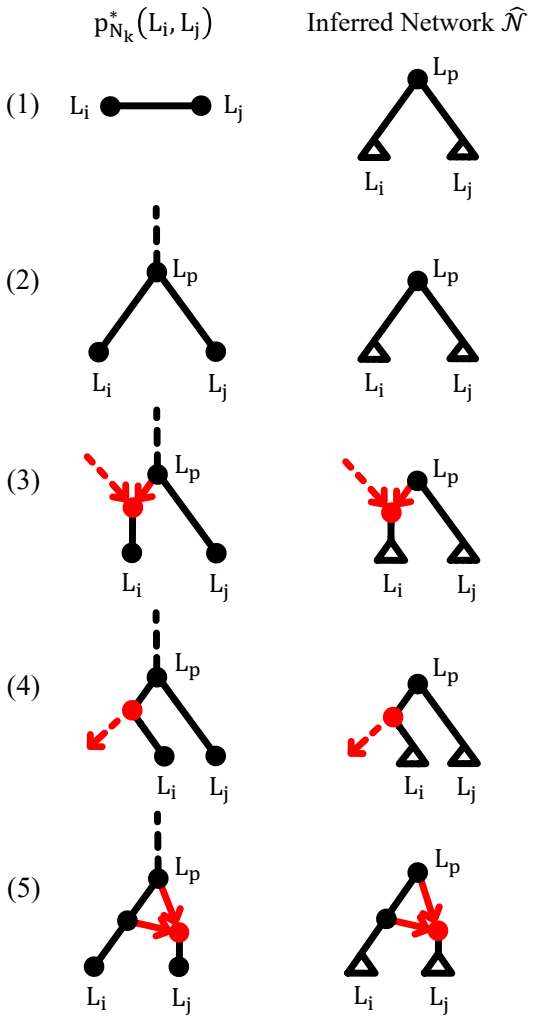

Figure 2: All possible neighbor relationships between two nodes L_i, L_j in a constraint network N_k along with their corresponding $p^*_{N_k}(L_i, L_j)$ paths and L_p nodes.

3.4 Species Network Inference Pipeline

Given input data I on a set of taxa X (e.g. multiple sequence alignments or estimated gene trees), the following pipeline can be utilized to efficiently infer species networks:

1. Estimate a pairwise dissimilarity matrix D on X from I.
2. Decompose X into disjoint subsets $\mathbf{S} = \{S_i\}_{i=1}^k$ where $\cup_{i=1}^k S_i \subseteq X$.
3. For each subset S_i, infer a constraint network N_i.
4. Construct species network $\widehat{\mathcal{N}}$ on taxa X by using D and $\mathbf{N} = \{N_i\}_{i=1}^k$ as input to InPhyNet.

Such a pipeline is statistically consistent under the multi-species network coalescent (MSNC; see [10]) given that certain assumptions are met in steps 1-3 (see Theorem 5). Most importantly, the methods used in step 1 and 3 must be statistically consistent under the MSNC and all reticulations in the underlying true species network must be estimable in at least one $S_i \in \mathbf{S}$.

Given that the network inference method used in step 3 has runtime complexity $O(f(n))$ with respect to the number of taxa in the inferred network (n), and that InPhyNet has runtime complexity $O(\phi(n))$, this pipeline has runtime complexity $O(kf(m) + \phi(n))$ where $k = |\mathbf{S}|$ and $m = \max\{|S_i| : S_i \in \mathbf{S}\}$. This is clear given that f is monotonically increasing because $\sum_{i=1}^k f(|S_i|) \leq \sum_{i=1}^k f(m) = kf(m)$. In practice, when maximum likelihood or maximum pseudo-likelihood methods are used to infer constraint networks, $kf(m) >> \phi(n)$, so this complexity is approximately $O(kf(m))$.

4 Results

In this section, we quantify the estimation error of InPhyNet and other methods by computing the normalized hardwired cluster distance (HWCD) to compare the topologies between the true and inferred networks. Normalized HWCD is a novel extension of normalized Robinson-Foulds distance [32] to network space. See the Performance Metrics subsection of Materials and Methods for further details.

4.1 Simulation Study

We compared the performance of the pipeline utilizing InPhyNet in conjunction with SNaQ to infer the constraint networks to that of SNaQ on its own when $n = 30$ and to the implicit network inference methods NeighborNet [7] and Consensus Networks [18] when $n = 200$ (see Figure 4) in terms of accuracy and efficiency. We also attempted to infer networks with PhyloFusion [48], however, the software did not complete any analyses in a feasible timeframe. InPhyNet matched SNaQ's accuracy when $m = 20$ (number of taxa in the constraint networks) and outperformed the implicit methods across all settings in our simulations (see Figure 4).

Across all simulations, we observed decreased HWCD as ILS decreased, as the number of gene trees

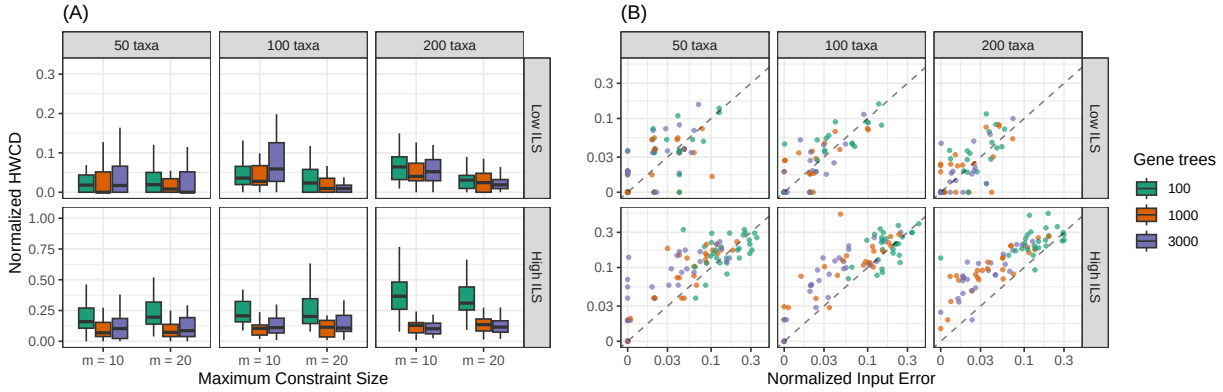

Figure 3: Normalized hardwired cluster distance (HWCD, see Equation 9) of InPhyNet across 900 simulations (25 per parameter combination). Simulations with differing levels of input data (gene trees) are represented with different colors. **A:** Normalized HWCD observed across various network sizes (columns; 50, 100, and 200 taxa) and simulation settings (rows; low and high incomplete lineage sorting (ILS)). Note the difference in y-axis scale between rows. **B:** InPhyNet inference error (y-axis) vs. total error of the distance matrix and constraint networks used as input to InPhyNet (x-axis; see Equation 8). The dotted line is $y = x$, where input and output error are equivalent. We present results that used a maximum constraint network size $m = 20$ here, however similar trends with overall higher levels of both input and output error were observed in simulations where $m = 10$ as well.

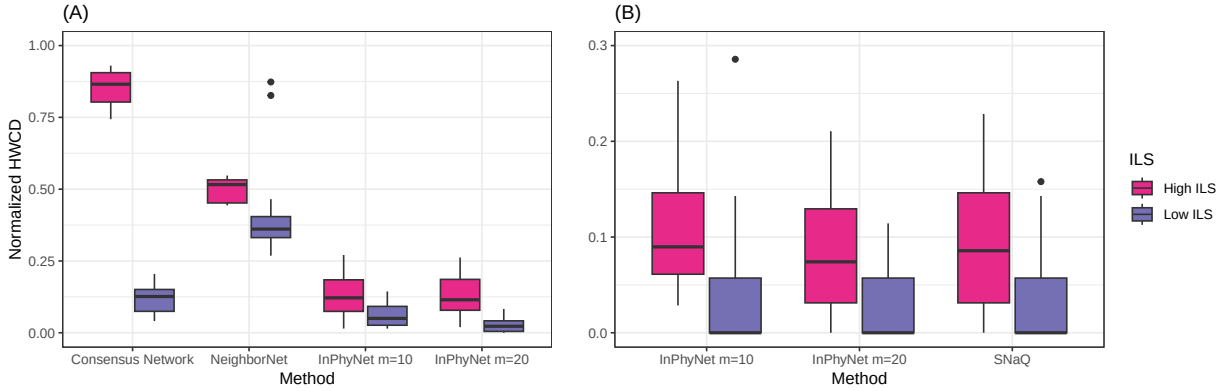

Figure 4: Normalized hardwired cluster distance (HWCD, see Equation 9) of the networks obtained from Consensus Network, NeighborNet, SNaQ, and InPhyNet analyses (x-axis). m refers to the maximum size of constraint networks used as input to InPhyNet. **A:** Normalized HWCD for the networks inferred with 200 taxa with high and low incomplete lineage sorting (ILS). **B:** Networks inferred with $n = 30$, the maximum that SNaQ could reasonably handle, with high and low ILS. Note the y-axis scale differs between the two plots presented.

increased, and as m increased. InPhyNet's accuracy was highly sensitive to the input data's accuracy in all simulation settings (see Figure 3).

Simulations with $m = 20$ had considerably higher runtimes than those with $m = 10$ (Figure 5). This is expected, as InPhyNet's runtime scales linearly with n but *polynomially* with m when using SNaQ. The median single-core runtimes to infer a 30 taxa network using InPhyNet were 1.2 hours with $m = 10$, 15 hours with $m = 20$, but 287 hours with SNaQ. The runtimes for NeighborNet and Consensus Networks were on the order of minutes with little to no change when parameters were varied.

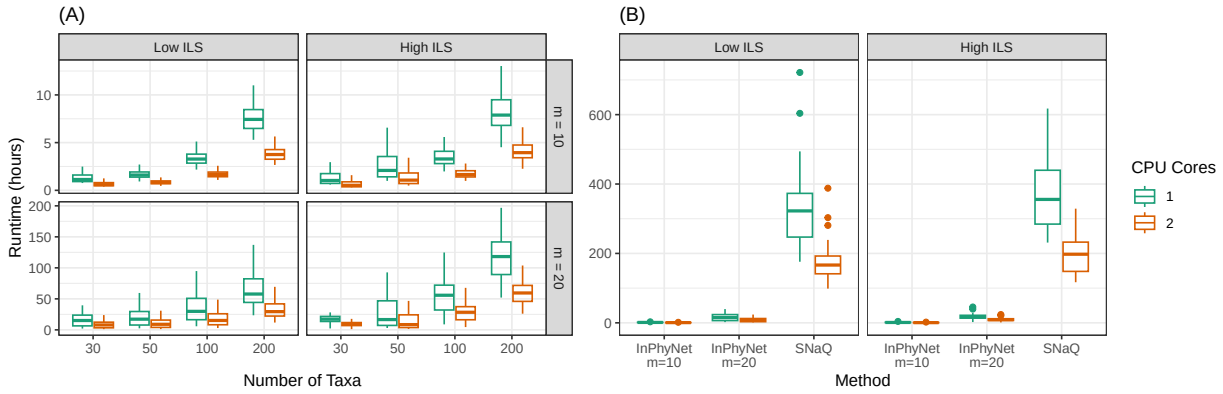

Figure 5: Total wall-clock runtime in hours to infer all of the constraint networks for a given simulation replicate. Runtimes with various cores were computed after the fact but are not simple averages of overall runtime (i.e. $\frac{\text{total runtime}}{\text{number of cores}}$). Instead, they are actual runtimes that would be observed with a given number of cores while taking into account work that cannot be parallelized across cores. For example, SNaQ can parallize *multiple inference replicates* across cores, but cannot parallelize *an individual replicate* across cores. So, if SNaQ is run with 10 inference replicates, no runtime improvement would be seen with greater than 10 cores. m refers to the maximum size of constraint networks used as input to InPhyNet. **A:** InPhyNet runtimes across various simulation settings. Each boxplot corresponds to results obtained from 25 unique network topologies. **B:** Runtime of InPhyNet with $m = 10$, $m = 20$, and SNaQ on its own. All simulations here were conducted on networks with low levels of incomplete lineage sorting (ILS).

4.2 Phylogeny of Seed Plants

Phylogenetic tree analyses led to observed quartet concordance frequencies that conflicted with what would be expected under the inferred tree model [29]. Specifically, there was significant conflict between concordance factors involving four major gymnosperm clades as well as four major fern clades. We constructed subsets to ensure that reticulations between these clades would be estimable (see subsection Phylogeny of Seed Plants in Materials and Methods). Nevertheless, only one reticulation was estimated after model selection that spanned multiple clades (i.e., ancestral hybridization). Networks were inferred with $h = 0, 1, 2, 3$ reticulations for each subset of taxa and slope heuristics were used to select the best-fitting model for each subset [3]. The phylogenetic network inferred by InPhyNet had 4 reticulations, each of which are well supported by previous research (see Figure 6 and Discussion).

5 Discussion

To expand the limits of accurately inferring large phylogenetic networks, we developed InPhyNet, a fast method for obtaining a massive network via merging small networks. This approach integrates the high accuracy of small networks inferred using robust model-based methods with the efficiency gained by decomposing large problems into smaller subproblems. Our simulation study demontrated the positive performance of InPhyNet in comparison to existing methods. We also established empirical benchmarks and characterized its sensitivity to key input parameters, including the degree of ILS, the number of gene trees (i.e., input dataset size), and the maximum constraint network size. InPhyNet exhibits versatility: it can efficiently infer large phylogenetic networks with substantially greater accuracy than existing consensus or distance-based methods. Conversely, it enables the inference of smaller networks at orders-of-magnitude faster computational times than state-of-the-art methods without compromising accuracy.

Careful subset decomposition is an important step to optimize method accuracy as InPhyNet cannot infer reticulations between constraint networks. Given the volume of simulations conducted here, we could not perform thorough discordance analyses and manual subset decomposition for each replicate, and had to automate the process instead. Such a process requires manual adjustment and evaluation to achieve favorable results, so this likely acted as an additional source of input error in our simulations. In particular, this decomposition process utilized Tree of blobs INference for a NetworK [TINNiK; 2], a method that explicitly requires manual adjustment of its parameters in order to attain optimal results. However, manual selection of taxon subsets via biological knowledge of the group under study remains a suitable alternative.

Given that constraint networks are inferred with a method whose runtime scales at a rate of $f(n)$, the runtime complexity to infer all constraint networks for InPhyNet is $O\left(\frac{n}{m}f(m)\right)$. Despite this, Figure 5 shows runtimes that scale at rates worse than $O(n)$ with fixed m, especially with high levels of ILS. This superlinear scaling is a result of the significant leverage that outliers have on inference runtime. As n increases with a fixed m, there are more subsets being inferred, leading to a higher prevalence of outliers. Simulations with high ILS are particularly susceptible to this apparent superlinear scaling because the increased levels of gene tree discordance make convergence more difficult for SNaQ. Additionally, more reticulations in a network increased the difficulty of inference and led to higher levels of error but did not have a direct impact on runtimes.

5.1 Phylogeny of Seed Plants

Our species network generally agreed with previously inferred phylogenies [29] except for the placement of Asterids, Rosids, Caryophyllales, and Saxifragales. Their placement is due to prior inference of a four-way multifurcation between these groups, whereas InPhyNet does not allow multifurcations. The remaining relationships displayed in our species network agree with prior studies of land plants [24, 29, 44]. We then

focus our discussion on the reticulate relationships uncovered by InPhyNet.

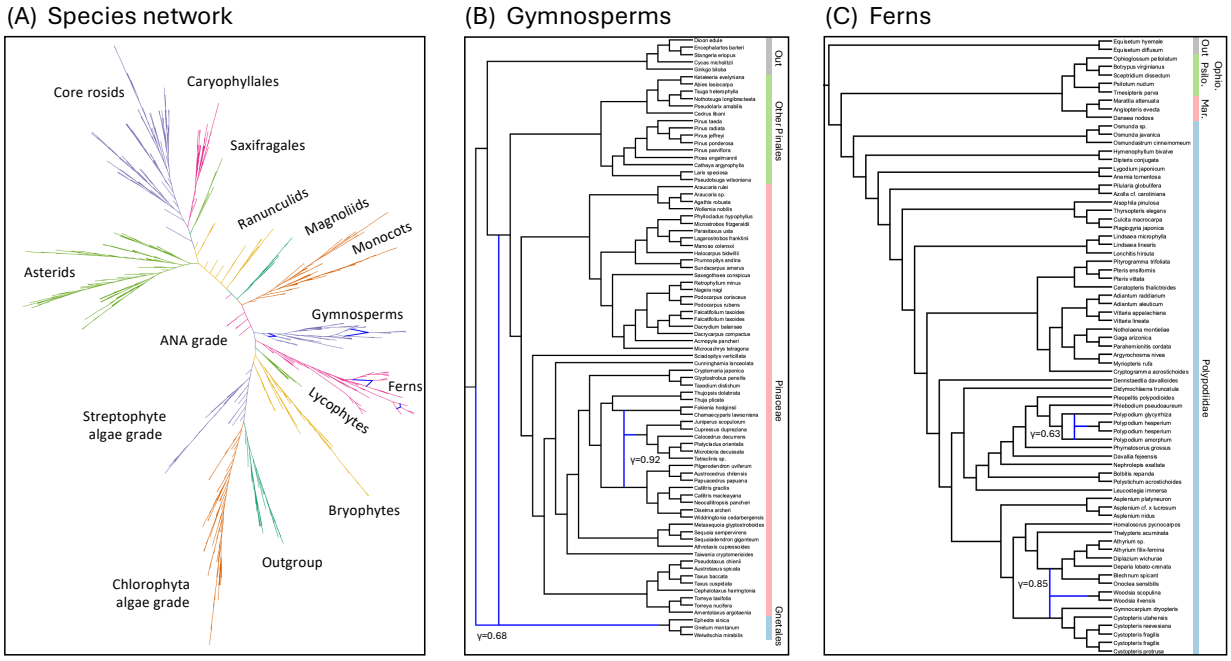

Figure 6: **A:** Semi-directed phylogenetic network of seed plants obtained using InPhyNet. Various colored branches signify different clades within the phylogeny. Branch lengths are not scaled. **B:** Magnified portion of the phylogeny corresponding to gymnosperms. The reticulation ancestral to Gnetales provides support to two controversial placements of Gnetales within the phylogeny of gymnosperms. The reticulation inferred within Pinaceae is corroborated by previous research into the histories of each its descendant species [12, 13, 50]. The inheritance probability (γ) for each reticulation is noted in the network. **C:** Portion of the phylogeny corresponding to ferns. Previously reported phylogenetic trees for this group revealed significant discordance between the four major clades shown here. However, SNaQ and InPhyNet inferred no reticulations between these clades. Our analyses resulted in two reticulations within Polypodiidae that are congruent with previous studies [38, 39, 37].

5.1.1 Gnetales

The placement of Gnetales in the phylogeny of gymnosperms is controversial. Some research suggests that they are sister to Pinaceae [5, 41, 31, 8, 49, 6], while others suggest that they are sister to conifers [5, 41, 25]. Treelike analyses of these species lead to significant conflict between each of these hypotheses [29], but our analyses reconcile these hypotheses by providing partial support for each. The major edge of this reticulation corresponds to the Gnetifer hypothesis while the minor edge corresponds to the Gnepine hypothesis. The inferred phylogeny of Gnetales is known to rely heavily on the specific type of data used and how it is handled [49], so this result requires further investigation in the future.

5.1.2 Pinaceae

High rates of polyploidy have been identified in genus *Juniperus* [12], and similar evidence has been found in *Cupressales* and *Platycladus* [13]. We are unaware of such direct evidence for *Calocedrus decurrens*, *Microbiota decussata*, or *Tetraclinis spp.*, but research suggests that these three species are closely related to *Platycladus orientalis* [50], corroborating their placement in this phylogeny.

While SNaQ does not directly model polyploidization, PhyloNet [1], a similar likelihood-based approach, has been shown to accurately infer polyploidy events in genus *Fragaria* despite not expliciting modelling polyploidization. This suggests that the reticulation in Pinaceae may be indicative of the high rates of polyploidy observed in *Juniperus*, *Cupressales*, and *Platycladus*, but more evidence is needed to corroborate this claim.

5.1.3 Genera *Polypodium* and *Woodsia*

The reticulation inferred in *Polypodium* is well-founded as *Polypodium hesperium* is an allotetraploid and its progenitors are precisely *Polypodium amorphum* and *Polypodium glycyrrhiza* [38]. The reticulation in *Woodsia* can also explain previous works that have inferred either a multifurcation between genera *Woodsia*, *Athyrium*, *Diplazium*, *Blechnum*, *Onoclea*, and other ferns [39], or that have otherwise found generally weak support for treelike backbones between these species [37].

6 Statistical Consistency

Theorem 1. Given true species tree T, Neighbor Joining (NJ) [33] will construct T from distance matrix D only if $Q_1(D) < 0$ and $Q_2(D) < 0$ for each pair of true neighbors in T $(1, 2)$ and arbitrary other taxa (i, j).

$$Q_1(D) = D_{12} + D_{ij} - D_{1i} - D_{2j} \tag{2}$$

$$Q_2(D) = D_{12} + D_{ij} - D_{1j} - D_{2i} \tag{3}$$

Proof. Presented in [33].

Theorem 2. The expected AGID for two taxa A and B in species tree T is $t_{AB} + 2$ where t_{AB} is the sum of the lengths of the edges between A and B in coalescent units.

Proof. Under the multispecies coalescent (MSC) model, two taxa A and B cannot coalesce until their lineages meet in the true species tree T. The time this takes to happen is exactly the distance by which A and B are separated in T which we call t_{AB}. Once A and B are able to coalesce, under the MSC the time until their coalescence X is modelled by $X \sim \text{Exponential}(1)$. In the time before A and B coalesce, *both* of their

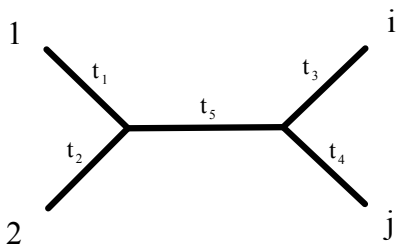

Figure 7: Treelike unrooted quartet and its associated edge lengths. Taxa 1 and 2 are true neighbors whereas taxa i and j are arbitrary.

$$\begin{aligned}\mathbb{E}\left(Q_1^{(T)}(\hat{D})\right) &= (t_1+t_2+2)+(t_3+t_4+2)- \\ &\quad (t_1+t_5+t_3+2)-(t_2+t_5+t_4+2) \\ &= -2t_5 < 0 \\ \mathbb{E}\left(Q_2^{(T)}(\hat{D})\right) &= (t_1+t_2+2)+(t_3+t_4+2)- \\ &\quad (t_1+t_5+t_4+2)-(t_2+t_5+t_3+2) \\ &= -2t_5 < 0\end{aligned}$$

branches are increasing in length, thus the total internode distance between A and B is $t_{AB}+2X$. Averaging across n gene trees, this gives us an expected AGID of $t_{AB}+2$ as shown in Equation (4).

$$\mathbb{E}\left(\text{AGID}(A,B)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left[t_{AB}+2X_i\right]\right) = \frac{1}{n}n\left[t_{AB}+2\mathbb{E}\left(X_1\right)\right] = t_{AB}+2 \tag{4}$$

Theorem 3. The average gene tree internode distance (AGID) metric is statistically consistent for the true species tree T if all internal branches in T have length greater than 0.

Proof. It can be shown that the expected AGID for two taxa A, B in species tree T is $t_{AB}+2$ where t_{AB} is the sum of the lengths of the edges connecting A and B (see Theorem 2).

Without loss of generality, consider taxa $1, 2$ which are true neighbors in T and an arbitrary pair of taxa i, j in T. The true unrooted quartet containing $1, 2, i, j$ that is displayed in T must be $12|ij$ with branch lengths shown in Figure 7 because 1 and 2 are assumed to be true neighbors.

Thus, where $\hat{D} = M_D(\mathcal{G}_n)$, we have the equations given below. It follows that, $Q_1^{(T)}(M_D(\mathcal{G}_n)) \xrightarrow{p} -2t_5 < 0$ and $Q_2^{(T)}(M_D(\mathcal{G}_n)) \xrightarrow{p} -2t_5 < 0$, so AGID is statistically consistent for T by Definition 7.

Theorem 4. Given n gene trees from level-1 species network $\mathcal{N}$ following the MSNC, if all edges in $\mathcal{N}$ have positive, non-zero length, the AGID distance metric is statistically consistent for a displayed tree of $\mathcal{N}$ outside of a parameter space of measure 0.

Proof. From Theorem 1, we need only show that Equation (2) and Equation (3) hold in the limit for some displayed tree of $\mathcal{N}$ for every possible quarnet. As outlined in [40] there are 5 possible level-1 quarnets, so we verify this fact for each. Below, we express $\mathbb{E}\left(Q_i^{(X)}(M_D(\mathcal{G}_n))\right)$ as $\mathbb{E}\left(Q_i^{(X)}\right)$ for brevity.

Networks 1, 2, 4, 5: Each of these networks have two displayed trees, T_{major} and T_{minor}, both of which have the topology $12|ij$. From Theorem 3, as long as all non-reticulate edges in $\mathcal{N}$ have positive length, these topologies both yield $\mathbb{E}\left(Q_i^{T_j}\right) < 0$ for $i = 1, 2$ and $T_j = T_{\text{major}}, T_{\text{minor}}$, so we have:

$$\mathbb{E}\left(Q_i^{(\mathcal{N})}\right) = (1-\gamma)\mathbb{E}\left(Q_i^{(T_{\text{major}})}\right) + \gamma\mathbb{E}\left(Q_i^{(T_{\text{minor}})}\right) < 0$$

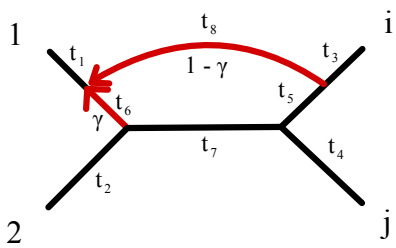

Figure 8: Network 3 in Theorem 4 and each of its edge lengths. Taxa 1 and 2 are true neighbors whereas taxa i and j are arbitrary.

$$\mathbb{E}\left(Q_1^{T_{\text{major}}}\right) = -2(1-\gamma)t_7 + 2\gamma t_5$$
$$\mathbb{E}\left(Q_1^{T_{\text{minor}}}\right) = 2(1-\gamma)t_7 - 2\gamma t_5$$
$$\mathbb{E}\left(Q_2^{T_{\text{major}}}\right) = -2(1-\gamma)t_7$$
$$\mathbb{E}\left(Q_2^{T_{\text{minor}}}\right) = -2(1-\gamma)t_5$$

Network 3: Regardless of the direction of the reticulation in this quarnet, we have $\mathbb{E}(D_{12} + D_{ij}) < \mathbb{E}(D_{1j} + D_{2i})$ and $\mathbb{E}(D_{1i} + D_{2j}) < \mathbb{E}(D_{1j} + D_{2i})$. Then, for the major and minor trees T_{major} and T_{minor} of $\mathcal{N}$ where taxa $1, 2$ and taxa $1, i$ are true neighbors, respectively, we have:

$\mathbb{E}\left(Q_1^{T_{\text{major}}}\right) = -\mathbb{E}\left(Q_1^{T_{\text{minor}}}\right)$ but $\mathbb{E}\left(Q_2^{T_j}\right) < 0$ for $T_j = T_{\text{major}}, T_{\text{minor}}$, so one of T_{major} or T_{minor} will have $\mathbb{E}\left(Q_i^{T_j}\right) < 0$ for $i = 1, 2$, so one of the displayed trees of this quarnet will be recovered so long as $\mathbb{E}\left(Q_1^{T_{\text{major}}}\right) \neq 0$, which is always the case outside of a parameter space of measure 0.

Note that if two different quarnets, say $q^{(1)}$ and $q^{(2)}$, display the same reticulation r_i, they may have the relationship shown in Equation (5). In this case, $q^{(1)}$ implies taxa being merged over the major edge of r_i while $q^{(2)}$ implies taxa being merged over the minor edge of r_i. However, outside of a space of measure 0, one of these pairs will have lower expected distance, and the edge connecting this pair will be the one to appear in the reconstructed tree.

$$\mathbb{E}\left(Q_1^{T^{(1)}_{\text{major}}}\right), \mathbb{E}\left(Q_1^{T^{(2)}_{\text{minor}}}\right) < 0 < \mathbb{E}\left(Q_1^{T^{(1)}_{\text{minor}}}\right), \mathbb{E}\left(Q_1^{T^{(2)}_{\text{major}}}\right) \tag{5}$$

Tree-like Quartet: Theorem 3 covers the case of a tree-like quartet.

We have shown that for each quarnet q in $\mathcal{N}$ with displayed quartets q_{major} and q_{minor}, $Q_i^{(q_j)} \xrightarrow{p} c_i$ where $c_i < 0$ for $i = 1, 2$ for one of $q_j = q_{\text{major}}$ or $q_j = q_{\text{minor}}$. Call T the displayed tree of $\mathcal{N}$ that contains each of these displayed quartets. For each displayed quartet q_j, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|Q_i^{(q_j)} - c_i\right| > \epsilon_{i,j}\right) \to 0, \;\; \forall \epsilon_{i,j} > 0 \text{ where } i = 1, 2$$

then, because $\mathbb{P}\left(\bigcup_j A_j\right) \leq \sum_j \mathbb{P}\left(A_j\right)$, it follows that:

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_j \left|Q_i^{(q_j)} - c_i\right| > \epsilon_{i,j}\right) \leq \lim_{n\to\infty} \sum_j \left[\mathbb{P}\left(\left|Q_i^{(q_j)} - c_i\right| > \epsilon_{i,j}\right)\right]$$
$$= \sum_j \left[\lim_{n\to\infty} \mathbb{P}\left(\left|Q_i^{(q_j)} - c_i\right| > \epsilon_{i,j}\right)\right] = 0, \ \ \forall \epsilon_{i,j} \text{ where } i = 1, 2$$

It follows from Theorem 1 that AGID is statistically consistent for T. □

Corollary 1. Given true species network $\mathcal{N}$, distance matrix D, and constraint networks $\mathbf{C}$ where D satisfies Equation (2) and Equation (3) for $\mathcal{T} \in \text{disp}(\mathcal{N})$ a displayed tree of $\mathcal{N}$ and all constraint networks agree with $\mathcal{N}$, then $\hat{T} = \text{major}(\phi(D, \mathbf{C})) = \mathcal{T}$.

Proof. The inferred network $\widehat{\mathcal{N}}$ is first constructed as a tree in Algorithm 1 before reticulations are added in post-processing. By Theorem 1, if all initial proposals are accepted in Algorithm 2 then $\widehat{T} = \mathcal{T}$. All constraint networks agree with $\mathcal{N}$ by assumption, so they must also all agree with $\mathcal{T}$. Thus, all initial merge proposals will be accepted and $\widehat{T} = \mathcal{T}$. □

Theorem 5. Given true level-1 species network $\mathcal{N}$, constraint networks $\mathbf{C} = \{C_1, \ldots, C_k\}$ on non-overlapping sets of taxa $\mathbf{S} = \{S_1, \ldots, S_k\}$, and distance matrix D, $\phi(D, \mathbf{C}) = \mathcal{N}$ if the following are true:

1. D satisfies the conditions in Theorem 1 for an arbitrary displayed tree T of $\mathcal{N}$
2. All C_i agree with $\mathcal{N}$
3. For each reticulation $r_j \in R(\mathcal{N})$, at least one subset contains at least one individual each from clades A, B, C, D, and H as shown in Figure 9

Proof. From (1), if InPhyNet's initial merging proposals are always accepted, then according to Corollary 1 the backbone tree of $\widehat{\mathcal{N}}$ will be T. From (2), all C_i agree with $\mathcal{N}$ and therefore agree with T, and from (3) each reticulation must appear in at least one constraint, so all of InPhyNet's initial merging proposals will be accepted and the backbone of $\widehat{\mathcal{N}}$ will be T. All that remains to be shown is that the reticulations in $\widehat{\mathcal{N}}$ are placed in the same positions as they appear in $\mathcal{N}$.

For each reticulation r_j the various components of the cycle formed by r_j in $\mathcal{N}$ and their corresponding components in the constraint network that contains r_j are identified in Figure 9. Notice that there are only two points throughout Algorithm 3 at which an edge is directed towards a node: when nodes V_i and V_j are merged and the path p^* connecting them in some constraint either (1) contains a hybrid node, or (2) contains a node with a directed edge pointing towards a hybrid node. We have shown that $\widehat{\mathcal{N}}$ will display T, so it follows that, for each $V_{\tilde{H}}$, one of its incoming directed edges $E_{\tilde{A},\tilde{H}}$ or $E_{\tilde{B},\tilde{H}}$ will be traversed in p^* and therein be placed correctly in $\widehat{\mathcal{N}}$. Without loss of generality, assume that this edge is $E_{\tilde{A},\tilde{H}}$. Then,

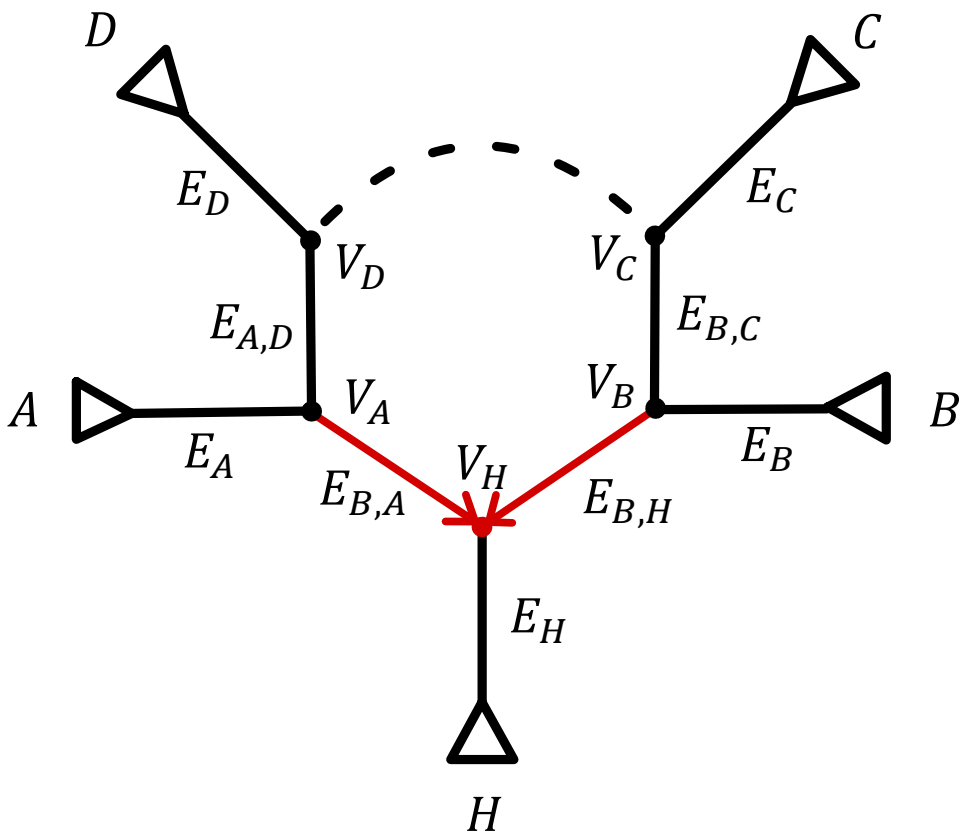

Figure 9: Cycle corresponding to hybrid node V_H in the true species network with labellings corresponding to those used in Theorem 5. This structure is also displayed in any constraint that meets condition 3 of Theorem 5 but with labels corresponding to the relevant subsets of each clade $\tilde{A} \subseteq A, \tilde{B} \subseteq B$, etc.

notice that nodes V_B and V_C would not both appear in a displayed tree of $\mathcal{N}$ because V_B leads to a hybrid node. If clades B and C are merged after each of their members are merged, though, such a node can be induced in the inferred network by splitting edge E_B (see Figure 3e). When $\tilde{B}$ and $\tilde{C}$ are merged, $V_{\tilde{B}}$ is traversed in p^*, whereafter it is marked as a which leads to this exact splitting operation being conducted in the post-processing step of Algorithm 1.

Thus, it follows that V_B and therefore $E_{B,H}$ will be placed correctly in $\widehat{\mathcal{N}}$ if (i) all members of $\tilde{B}$ and $\tilde{C}$ merge with one another before they do with any other taxa in the constraint network, and (ii) when $V_{\tilde{B}}$ and $V_{\tilde{C}}$ are merged along path p^*, edges $E_{\tilde{B}}$ and $E_{\tilde{B},\tilde{C}}$ occur along p^*. Condition (i) follows from assumption (1), and all paths in the constraint connecting $V_{\tilde{B}}$ to any other node must traverse $E_{\tilde{B}}$ and $E_{\tilde{B},\tilde{C}}$ following the merging of $V_{\tilde{A}}$ and $V_{\tilde{H}}$, so condition (ii) is satisfied as long as $V_{\tilde{B}}$ and $V_{\tilde{C}}$ are neighbors in the constraint network. The presence of these nodes in the constraint network is given by assumption (3) and their status as neighbors is given by assumption (2), so it follows that each reticulation r_j will be placed correctly in $\widehat{\mathcal{N}}$. □

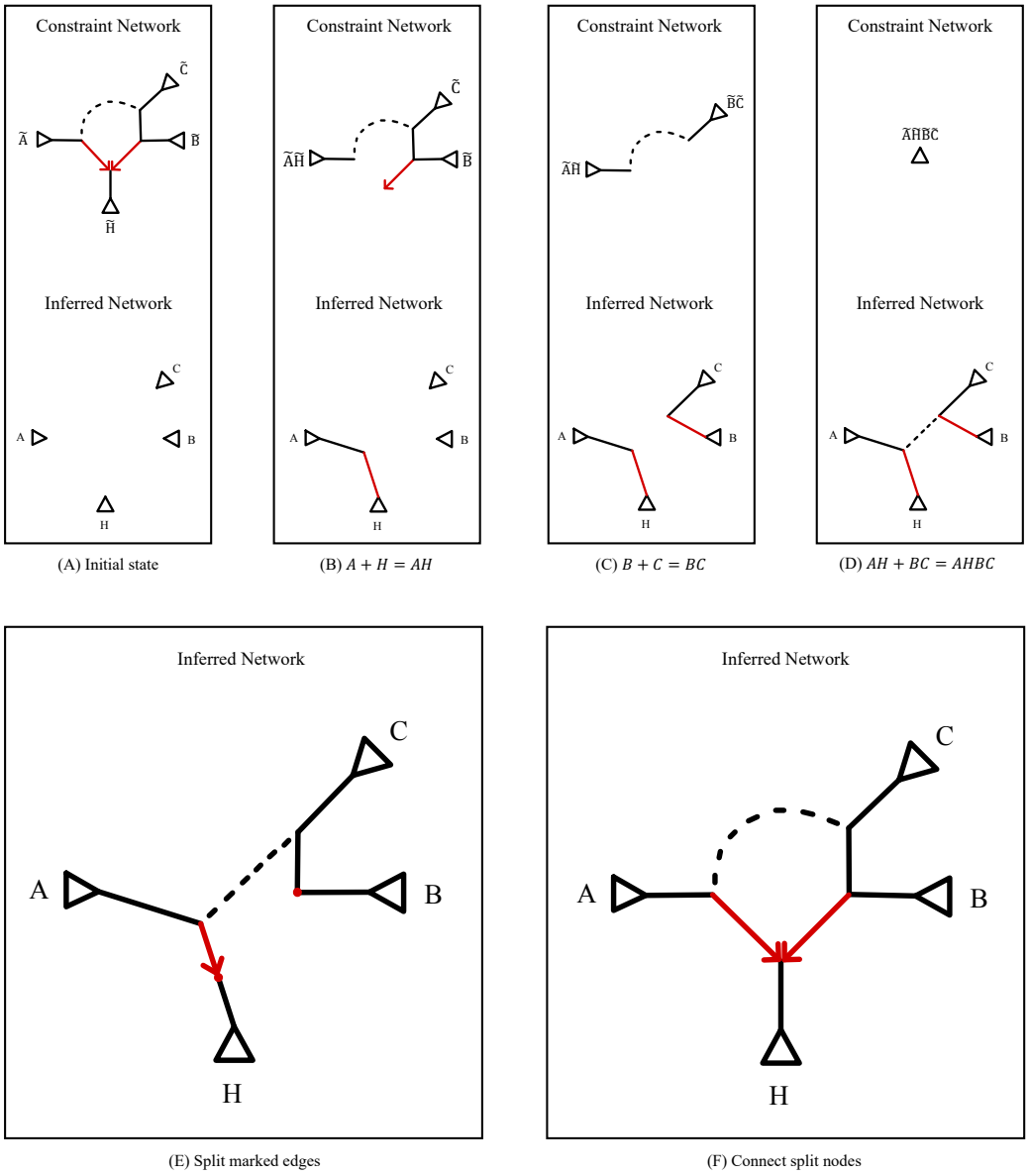

Figure 10: **A:** the initial relevant state of the algorithm at which point all taxa within each clade have been merged. **B:** nodes A and H are merged. The path p^* connecting them in the constraint network traverses one reticulation, so its corresponding edge is marked in the inferred network. **C:** nodes B and C are merged where p^* traverses a node with an outgoing reticulation, so this reticulation is also recorded in the inferred network. **D:** there may be zero, one, or multiple mergings along the dotted line before this step occurs, but those do not affect how A, B, and C are ordered around H. **E:** the edges that correspond to reticulations are split. The edges are split differently based on whether they correspond to an incoming or outgoing reticulation. **F:** a reticulation is drawn from the red node above B to the hybrid node above H to complete the inferred network.

Corollary 2. Given gene trees $\mathcal{G}_n$ from $\mathcal{N}$ following the MSNC and subset decomposition $\mathbf{S} = \{S_1, \ldots, S_k\}$, if M_C are M_D are both statistically consistent and $\mathbf{S}$ satisfies condition (3) in Theorem 5, then InPhyNet is statistically consistent under the MSNC.

Proof. From Theorem 4, $\forall \epsilon_{i,j} > 0, \exists n_{i,j}$ such that the probability of $M_D(\mathcal{G}_n)$ satisfying the conditions in Theorem 1 for an arbitrary displayed tree T of $\mathcal{N}$ is at least $1-\sum_{i,j} \epsilon_{i,j}$. Further, by assumption, $\forall \epsilon_l > 0, \exists n_l$ the probability that the constraint network inferred by M_C on subset S_i agrees with $\mathcal{N}$ is at least $1 - \epsilon_l$. Then, the probability that each of these events happen for all constraint networks is at least:

$$\begin{aligned}
\left(1 - \sum_{i,j} \epsilon_{i,j}\right) \prod_l (1 - \epsilon_l) &\geq \left(1 - 2\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\}\right) \left(1 - \max_l \{\epsilon_l\}\right)^k \\
&\geq \left(1 - 2\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\}\right) \left(1 - k \max_l \{\epsilon_l\}\right) \quad \text{by Bernoulli's Inequality} \\
&= 1 - 2\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\} - k \max_l \{\epsilon_l\} + 2k\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\} \max_l \{\epsilon_l\} \\
&\geq 1 - \left(2\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\} + k \max_l \{\epsilon_l\}\right) \qquad (6)
\end{aligned}$$

Then, $\forall \epsilon > 0$, choose $\epsilon_{i,j}, \epsilon_l > 0$ such that $2\binom{n}{4} \max_{i,j} \{\epsilon_{i,j}\} + k \max_l \{\epsilon_l\} < \epsilon$ and choose $n = \max\left(\max_{i,j}\{n_{i,j}\}, \max_l\{n_l\}\right)$. Then, Equation (6) is at least $1 - \epsilon$ and, by Theorem 5, the probability that InPhyNet constructs $\mathcal{N}$ is at least $1 - \epsilon$, so InPhyNet is statistically consistent under the MSNC.

6.1 Simulation data

6.1.1 Ground truth data

For each number of ingroup taxa in a network $n = 30, 50, 100, 200$, 25 unique level-1, metric (i.e., all edge lengths and inheritance probabilites specified), semi-directed networks were simulated under a birth-death-hybridization (BDH) model using the SiPhyNetwork [21] package in R. An outgroup taxa was added on to each simulated network, so each final network has $n + 1$ taxa. These simulations were conducted with a speciation rate of 1, extinction rate of 0.2, hybridization rate of 0.025. Also, the rates of lineage generative, lineage degenerative, and lineage neutral reticulations were set as $0.5, 0.25$, and 0.25, respectively. Each inheritance probability γ was drawn from a Beta$(10, 10)$ distribution. Edge lengths were initially determined according to this BDH model before being adjusted according to Equation 7 in order to achieve the desired level of ILS. In Equation 7, L_i^{old} and L_i^{new} refer to the length of an edge in the network prior to and after adjustment, respectively. Moreover, λ is a scaling parameter and the network's average edge length will equal λ after this adjustment. Values of $\lambda = 2.0$ and $\lambda = 0.5$ were used to low and high levels of ILS, respectively.

$$L_i^{\text{new}} = \frac{\lambda L_i^{\text{old}}}{\bar{L}} \qquad (7)$$

Using the generated species network, true gene trees were simulated with the PhyloCoalSimulations [15] package in Julia under the a model of independent inheritance. For each replicate, a unique set of gene trees

was generated for each combination of ILS and number of gene trees.

True sequence alignments were simulated using the seq-gen software [30]. Multiple sequence alignments with 1,000 b.p. were simulated for each gene tree under the F81 model of evolution [14] with state frequencies $\pi_A = \pi_T = 0.3$ and $\pi_C = \pi_G = 0.2$ and a scaling parameter of 0.036.

6.1.2 Estimated data

IQ-TREE 2's ModelFinder [26] was used to select the best nucleotide substitution model to infer gene trees from the simulated sequence alignments. Note the Jukes-Cantor model [20] was almost always selected for our data. Average gene tree estimation error (GTEE) ranged from 2–6% for low ILS simulations and 28–74% for high ILS scenarios. Gene trees with more taxa tended to have higher average GTEEs.

After subset decomposition (described below), constraint networks were inferred using SNaQ [40] with default parameters. Each constraint network was selected from the best of 10 independent searches according to their pseudo-likelihood scores. For each constraint network consisting of taxa Y, SNaQ was only run with the true number of reticulations h for Y. Inferring networks with $h = 0, 1, 2, \ldots$ followed by an automated model selection process would have been more realistic but was computationally intractible given the number of simulations conducted.

The pairwise dissimilarity matrix D was computed by taking the average gene tree internode distance of the gene trees estimated with IQ-TREE 2. The matrix D and the constraint networks inferred with SNaQ were used as input to infer a species network with InPhyNet on all taxa.

6.1.3 Subset decomposition

The main priority of subset decomposition here is to ensure that all reticulations in the true underlying species network are *estimable*. This is not the case if, for example, the true phylogeny contains a reticulation between clades A and B, but no subset contains members from both clades. In order to avoid such a situation, we implement a subset decomposition algorithm that conducts the following steps:

1. Infer a species tree T_{species} on X with ASTRAL-III [47].
2. Infer a tree of blobs T_{blob} on X with TINNiK [2].
3. Consolidate taxa in T_{species} that inscribe blobs in T_{blob}.
4. Perform centroid edge decomposition on the consolidated T_{species} to attain disjoint subsets $\mathbf{S}$.
5. Expand the taxa in $\mathbf{S}$ that were consolidated in Step (3).

TINNiK's accuracy depends on manually adjusting α and β parameters to fit the data at hand, so we expect that our use of fixed values $\alpha = 0.01$ and $\beta = 0.99$ acted as an additional source of error in these simulations that could not be accurately quantified. Additionally, TINNiK's runtime scales polynomially with respect to the number of taxa in the tree of blobs, and we found that $n > 50$ was computationally infeasible for these simulations. This meant that for networks with $n > 50$ we were required to subdivide

the taxa into sets of 50 and infer multiple trees of blobs. We did this twice for each network with $n > 50$, once grouping taxa into sets of 50 as they appear in preorder traversal order, and once by assigning taxa in preorder traversal order to alternative groups. For example, with $n = 200$, the first such subset would be $\{t_1, t_2, \ldots, t_{50}\}$, and the second would be $\{t_1, t_5, t_9, \ldots, t_{197}\}$. This allows reticulations between nearby and distant taxa to be identified in a reasonable amount of time.

The maximum subset size parameter m was imposed at the centroid edge decomposition step in order to enforce desired maximum constraint network sizes. Centroid edge decomposition is conducted by iteratively splitting a tree in half at a single edge such that both subtrees obtained afterwards are approximately equal in size. This is repeated until each subtree contains no more than m taxa.

$$\text{error}(D, \mathbf{C}) = \frac{\min\limits_{T \in \text{disp}(\mathcal{N})} \text{HWCD}\left(T, \text{NJ}(D)\right) + \sum_{i=1}^{k} \text{HWCD}\left(C_i, \hat{C}_i\right)}{|E^{(T)}| - |V^{(T)}| + E^{(\text{NJ}(D))} - V^{(\text{NJ}(D))} + \sum_{i=1}^{k} \left(|E^{(C_i)}| - |V^{(C_i)}| + |E^{(\hat{C}_i)}| - |V^{(\hat{C}_i)}|\right)} \tag{8}$$

6.2 Performance Metrics

To assess the accuracy of InPhyNet across topologies with widely varying numbers of taxa, we extended the normalized Robinson-Foulds distance metric [32] from tree space to network space. Given two networks $N_1 = (V^{(1)}, E^{(1)})$ and $N_2 = (V^{(2)}, E^{(2)})$, this metric normalizes the hardwired cluster distance (HWCD) across the space of pairs of networks (denoted $\text{HWCD}(N_1, N_2)$) with $|V^{(1)}| + |V^{(2)}|$ nodes and $|E^{(1)}| + |E^{(2)}|$ edges. In the case where N_1 and N_2 are both trees, this metric is exactly the normalized Robinson-Foulds distance.

$$\text{HWCD}_{\text{norm}}\left(N_1, N_2\right) = \frac{\text{HWCD}\left(N_1, N_2\right)}{\left|E^{(1)}\right| - \left|V_L^{(1)}\right| + \left|E^{(2)}\right| - \left|V_L^{(2)}\right|} \tag{9}$$

This metric is conservative, as it normalizes over a space of networks that is strictly smaller than the space of all level-1 networks. Normalizing across the space of all level-1 networks may be more appropriate, but we are unaware of any such methods, or of any work that identifies the maximum HWCD between two level-1 networks.

To assess the total error between the distance matrix D and constraint networks $\mathbf{C}$ used as input to InPhyNet, we extend this notion of normalized HWCD to normalize across several sources of error. The result of this is Equation 8, where $\text{NJ}(D)$ is the tree obtained from the Neighbor-Joining algorithm with D as input and where $\text{disp}(\mathcal{N})$ is the set of displayed trees associated with $\mathcal{N}$. The first term in the numerator minimizes the HWCD of $\text{NJ}(D)$ across all of the displayed trees of $\mathcal{N}$ because our statistical consistency derivations assert that we need only be able to accurately infer any of the displayed trees of $\mathcal{N}$ with the

Neighbor-Joining algorithm, not any particular displayed tree.

6.3 Phylogeny of Seed Plants

Our analyses were based on estimated gene trees provided in Leebens-Mack et al. [29]. Notably in their work, significant gene tree-species tree discordances were identified between four major clades of gymnosperms and four major clades of ferns, highlighting instances of phylogenetic conflict that could be indicative of reticulate evolution. To account for these incongruences, we ensured that at least one subset of gymnosperms and one subset of ferns included multiple taxa from each of their four major clades along with an outgroup. Outgroups were selected randomly from Chlorophyta and Rhodophyta because their relationships with gymnosperms and ferns are non-contentious. Thus, reticulations between these major clades were estimable and would be reconstructible in the full inferred species network topology with InPhyNet according to Theorem 5 (see subsection Statistical Consistency below).

We inferred trees of blobs with all gymnosperms plus an outgroup and another with all ferns plus an outgroup using TINNiK [2]. These trees were constructed with α and β parameters ranging from 0.01 to 0.001 and 0.99 to 0.999, respectively. We identified sets of taxa in these trees that inscribed blobs and performed manual subset decomposition such that reticulations between these taxa would be estimable by SNaQ [40].

For each subset, 20 networks were inferred using SNaQ [40] each with $h = 0, 1, 2, 3$ reticulations. The best network for each value of h was selected by taking the network with the lowest negative log-pseudolikelihood. The best-fitting h value was then selected according to the slope heuristic method described in Section 4.2 of Baudry et al [3]. The pseudolikelihood scores of all subsets exhibited strongly dimishing return by $h = 2$, suggesting that inferring any models with $h > 3$ would be unnecessary. In total, this procedure was conducted on 12 total subsets of taxa. Four of these subsets selected models with $h = 1$ while the remainder selected models with $h = 0$.

To construct the dissimilarity matrix D used with InPhyNet, we computed the internode distances between taxa in the species tree inferred in [29]. We selected this approach based on the theoretical result that one requirement for InPhyNet to infer the true species network is for the Neighbor-Joining tree inferred on D to match one of the displayed trees of the true species network (see Theorem 5). We expect that the internode distances taken from this species tree will be more accurate in this sense than the distance matrix that would be produced by averaging the internode distances of species across estimated gene trees.

7 Data and Code Availability

We have maintained records of all scripts used to generate and analyze the data in a version-controlled repository that is publicly available on GitHub [https://github.com/NathanKolbow/InPhyNetSimulations].

The code for InPhyNet is written in the Julia programming language and available as open-source software in a publicly available, version-controlled repository on GitHub alongside detailed documentation and tutorials [https://github.com/NathanKolbow/InPhyNet.jl].

8 Acknowledgements

This work was supported by the National Science Foundation (DEB-2144367 to CSL). The authors thank Jiayang Wang for his review of the proofs presented herein as well as Erin Molloy and Yunheng Han for their help in improving the simulation study.

References

[1] Inferring Phylogenetic Networks Using PhyloNet | Systematic Biology | Oxford Academic. https://academic.oup.com/sysbio/article/67/4/735/4921127.

[2] E. S. Allman, H. Baños, J. D. Mitchell, and J. A. Rhodes. TINNiK: Inference of the Tree of Blobs of a Species Network Under the Coalescent. *bioRxiv*, page 2024.04.20.590418, Apr. 2024.

[3] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: Overview and implementation. *Statistics and Computing*, 22(2):455–470, Mar. 2012.

[4] L. Boto. Horizontal gene transfer in evolution: Facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683):819–827, 2009.

[5] L. M. Bowe, G. Coat, and C. W. dePamphilis. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proceedings of the National Academy of Sciences*, 97(8):4092–4097, Apr. 2000.

[6] T. W. A. Braukmann, M. Kuzmina, and S. Stefanović. Loss of all plastid ndh genes in Gnetales and conifers: Extent and evolutionary significance for the seed plant phylogeny. *Current Genetics*, 55(3):323–337, June 2009.

[7] D. Bryant and V. Moulton. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265, 2004.

[8] S.-M. Chaw, C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proceedings of the National Academy of Sciences*, 97(8):4086–4091, Apr. 2000.

[9] M. Ciccozzi, A. Lai, G. Zehender, A. Borsetti, E. Cella, M. Ciotti, E. Sagnelli, C. Sagnelli, and S. Angeletti. The phylogenetic approach for viral infectious disease evolution and epidemiology: An updating review. *Journal of Medical Virology*, 91(10):1707–1724, 2019.

[10] J. H. Degnan. Modeling hybridization under the network multispecies coalescent. *Systematic Biology*, 67(5):786–799, 2018.

[11] J. Döcker, S. Linz, and C. Semple. Displaying trees across two phylogenetic networks. *Theoretical Computer Science*, 796:129–146, Dec. 2019.

[12] P. Farhat, O. Hidalgo, T. Robert, S. Siljak-Yakovlev, I. J. Leitch, R. P. Adams, and M. Bou Dagher-Kharrat. Polyploidy in the Conifer Genus Juniperus: An Unexpectedly High Rate. *Frontiers in Plant Science*, 10, 2019.

[13] P. Farhat, S. Siljak-Yakovlev, O. Hidalgo, K. Rushforth, J. A. Bartel, N. Valentin, I. J. Leitch, and R. P. Adams. Polyploidy in Cupressaceae: Discovery of a new naturally occurring tetraploid, Xanthocyparis vietnamensis. *Journal of Systematics and Evolution*, 60(4):824–834, 2022.

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

[15] J. Fogg, E. S. Allman, and C. Ané. PhyloCoalSimulations: A Simulator for Network Multispecies Coalescent Models, Including a New Extension for the Inheritance of Gene Flow. *Systematic Biology*, 72(5):1171–1179, Sept. 2023.

[16] P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9241–9243.

[17] H. A. Hejase, N. VandePol, G. M. Bonito, and K. J. Liu. FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data. In M. Blanchette and A. Ouangraoua, editors, *Comparative Genomics*, pages 242–259, Cham, 2018. Springer International Publishing.

[18] B. R. Holland, K. T. Huber, V. Moulton, and P. J. Lockhart. Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny. *Molecular Biology and Evolution*, 21(7):1459–1461, 2004.

[19] D. H. Huson and D. Bryant. The SplitsTree App: Interactive analysis and visualization using phylogenetic trees and networks. *Nature Methods*, 21(10):1773–1774, 2024.

[20] T. H. Jukes and C. R. Cantor. Evolution of Protein Molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, 1969.

[21] J. A. Justison, C. Solis-Lemus, and T. A. Heath. SiPhyNetwork: An R package for simulating phylogenetic networks. *Methods in Ecology and Evolution*, 14(7):1687–1698, 2023.

[22] S. Kong, D. L. Swofford, and L. S. Kubatko. Inference of Phylogenetic Networks from Sequence Data using Composite Likelihood, Nov. 2022.

[23] S. Kong, S. J. Sánchez-Pacheco, and R. W. Murphy. On the use of median-joining networks in evolutionary biology. *Cladistics*, 32(6):691–699, 2016.

[24] G.-Q. Liu, L. Lian, and W. Wang. The Molecular Phylogeny of Land Plants: Progress and Future Prospects. *Diversity*, 14(10):782.

[25] S. Magallón and M. J. Sanderson. Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. *American Journal of Botany*, 89(12):1991–2006, 2002.

[26] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020.

[27] E. K. Molloy and T. Warnow. NJMerge: A Generic Technique for Scaling Phylogeny Estimation Methods and Its Application to Species Trees. In M. Blanchette and A. Ouangraoua, editors, *Comparative Genomics*, pages 260–276. Springer International Publishing.

[28] E. K. Molloy and T. Warnow. TreeMerge: A new method for improving the scalability of species tree estimation methods. *Bioinformatics*, 35(14):i417–i426, July 2019.

[29] One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685, Oct. 2019.

[30] A. Rambaut and N. C. Grass. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, June 1997.

[31] J.-H. Ran, T.-T. Shen, M.-M. Wang, and X.-Q. Wang. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings of the Royal Society B: Biological Sciences*, 285(1881):20181012, June 2018.

[32] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, Feb. 1981.

[33] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, July 1987.

[34] T. Sang and S. Ge. Genetics and phylogenetics of rice domestication. *Current Opinion in Genetics & Development*, 17(6):533–538, Dec. 2007.

[35] R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: Principles and practice. *Nature reviews. Genetics*, 18(4):213–229, Apr. 2017.

[36] D. B. Searls. Pharmacophylogenomics: Genes, evolution and drug targets. *Nature Reviews Drug Discovery*, 2(8):613–623, Aug. 2003.

[37] Y. Shao, R. Wei, X. Zhang, and Q. Xiang. Molecular Phylogeny of the Cliff Ferns (Woodsiaceae: Polypodiales) with a Proposed Infrageneric Classification. *PLOS ONE*, 10(9):e0136318, 2015.

[38] E. M. Sigel, M. D. Windham, and K. M. Pryer. Evidence for reciprocal origins in Polypodium hesperium (Polypodiaceae): A fern model system for investigating how multiple origins shape allopolyploid genomes. *American Journal of Botany*, 101(9):1476–1485, 2014.

[39] A. R. Smith, K. M. Pryer, E. Schuettpelz, P. Korall, H. Schneider, and P. G. Wolf. A classification for extant ferns. *TAXON*, 55(3):705–731, 2006.

[40] C. Solís-Lemus and C. Ané. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLOS Genetics*, 12(3):e1005896, Mar. 2016.

[41] D. E. Soltis, P. S. Soltis, and M. J. Zanis. Phylogeny of seed plants based on evidence from eight genes. *American Journal of Botany*, 89(10):1670–1681, 2002.

[42] S. J. Sánchez-Pacheco, S. Kong, P. Pulido-Santacruz, R. W. Murphy, and L. Kubatko. Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23):12518–12519, 2020.

[43] S. Véron, V. Saito, N. Padilla-García, F. Forest, and Y. Bertheau. The Use of Phylogenetic Diversity in Conservation Biology and Community Ecology: A Common Base but Different Approaches. *The Quarterly Review of Biology*, 94(2):123–148, June 2019.

[44] N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, B. R. Ruhfel, E. Wafula, J. P. Der, S. W. Graham, S. Mathews, M. Melkonian, D. E. Soltis, P. S. Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny, A. J. Shaw, L. DeGironimo, D. W. Stevenson, B. Surek, J. C. Villarreal, B. Roure, H. Philippe, C. W. dePamphilis, T. Chen, M. K. Deyholos, R. S. Baucom, T. M. Kutchan, M. M. Augustin, J. Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, and J. Leebens-Mack. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[45] X. Yu, T. Le, S. A. Christensen, E. K. Molloy, and T. Warnow. Advancing Divide-and-Conquer Phylogeny Estimation using Robinson-Foulds Supertrees. Preprint, Bioinformatics, May 2020.

[46] Y. Yu and L. Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10):S10, Oct. 2015.

[47] C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, May 2018.

[48] L. Zhang, B. Cetinkaya, and D. H. Huson. PhyloFusion- Fast and easy fusion of rooted phylogenetic trees into rooted phylogenetic networks.

[49] B. Zhong, O. Deusch, V. V. Goremykin, D. Penny, P. J. Biggs, R. A. Atherton, S. V. Nikiforova, and P. J. Lockhart. Systematic error in seed plant phylogenomics. *Genome Biology and Evolution*, 3:1340–1348, 2011.

[50] B. J. M. Zonneveld. Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nordic Journal of Botany*, 30(4):490–502, 2012.

Appendices

A Algorithms

Algorithm 1: InPhyNet$(X, D, \mathbf{N})$

Input : X a set of n labels
D a pairwise dissimilarity matrix on X
$\mathbf{N}$ a set of k level-1 networks on disjoint subsets $\mathbf{S} = \{S_i\}_{i=1}^k$ where $\bigcup_{i=1}^k S_i \subseteq X$

Output: A network $\widehat{\mathcal{N}} = (V, E)$ on taxa X

```
Create n disconnected leaf nodes V_L where each v_i ∈ V_L corresponds to taxa X_i;
Initialize a list of nodes to consider in merges L = V_L;
Initialize a map h : ⋃_{i=1}^k V_H^(i) → E × E;
N̂ = (V, E) = (V_L, ∅);
while |L| > 1 do
    Choose Nodes To Merge u, v
        Compute each Q_ij = (n − 2) × D_ij − Σ_{k=1}^n D_ki − Σ_{k=1}^n D_kj;
        𝕍 = ValidMergingPairs(N, X) (see Algorithm 2);
        if |V| = 0 then
            return PairwiseInPhyNet(N′, D′, X′) (see Algorithm 5);
        end
        u, v = L_i, L_j where i, j = arg min_{(i,j)∈𝕍} Q[i, j];
    Merge u and v
        Add a new node p to V;
        Add two new bidirectional edges e_u = (p, u) and e_v = (p, v) to E;
        Remove u, v from L, add p to L;
        for N ∈ N do
            AddHybridInformation!(N̂, h, N, u, v) (see Algorithm 3);
        end
    Update N and D
        for N ∈ N do
            MergeNodesInConstraint!(N, u, v, p) (see Algorithm 4);
        end
        Add a row and column to D so that now D ∈ ℝ^{m×m};
        for k = 1, 2, ..., m − 1 do
            D_{k,m} = D_{m,k} = (D_{k,i} + D_{j,k} − D_{i,j})/2;
        end
        Remove rows and columns i, j from D;
end
Post-Processing
    for each (e_C, e_1, e_2) where h[e_C] = (e_1, e_2) do
        Disconnect e_1 and e_2 from each of their parents e_{1p} and e_{2p};
        Create new parent nodes for each of e_1 and e_2, namely e′_{1p} and e′_{2p};
        Create the bidirectional edge (e_1, e′_{1p}) and directed edges e_{2p} → e′_{2p} and e′_{1p} → e′_{2p};
    end
    Unroot N̂;
return N̂;
```

Algorithm 2: ValidMergingPairs($\mathbf{N}, X$)

Input : $\mathbf{N}$ a set of k level-1 networks on disjoint subsets of taxa $\mathbf{S} = \{S_i\}_{i=1}^{k}$ where $\bigcup_{i=1}^{k} S_i \subseteq X$
Output: $\mathbb{V}$ the set of all pairs of nodes $u, v \in X$ that are valid to merge

---

Initialization

```
Initialize a map f : X × X → {0,1};
For all (u,v) ∈ X, u ≠ v, set f[u,v] = 1 and f[u,u] = 0;
```

Check Neighbors

```
for (u,v) ∈ X, u ≠ v do
    for N_i ∈ N do
        if (u,v) ∈ N_i AND (u,v) not neighbors in N_i then
            f[u,v] = 0;
        end
    end
end
```

Check Pairwise Topological Isomorphism

```
for (u,v) ∈ {(u,v) : f[u,v] = 1} do
    N' = Copy(N);
    for N' ∈ N' do
        p = parent of u,v in N';
        MergeNodesInConstraint!(N', u, v, p);
    end
    for i = 1,2,...,k−1 do
        overlap = {w : w ∈ N'_i, w ∈ N'_{i+1}};
        if RF(N'_i[overlap], N'_{i+1}[overlap]) ≠ 0 then
            f[u,v] = 0;
            break;
        end
    end
end
return {(u,v) : f[u,v] = 1};
```

Algorithm 3: AddHybridInformation!($\widehat{\mathcal{N}}, h, N_i, u, v$)

Input : $\widehat{\mathcal{N}}$ the inferred network
h a map from the set all reticulations in $\mathbf{N}$ to a pair of edges in $\widehat{\mathcal{N}}$
$N_i = (V^{(i)}, E^{(i)})$ a constraint network
u, v two nodes in the process of being merged

Output: Nothing, h is updated in-place

```
if u ∉ V^(i) OR v ∉ V^(i) then
    return;
end

Find p* ∈ u ⇝^{N_i} v ;
if |p*| ≤ 3 then
    Nothing; # Case 1 in Figure 2
else if |p*| = 5  AND  p*_4 ∈ V_H^(i) then
    h[p*_4] = (e_u, e_v); # Case 4 in Figure 2
else if |p*| = 5  AND  p*_2 ∈ V_H^(i) then
    h[p*_2] = (e_v, e_u); # Case 4 in Figure 2 with u, v swapped
else if p*_2 ∈ V_H^(i) then
    h[p*_2][1] = e_u; # Case 2 in Figure 2
else if p*_3 ∈ V_H^(i) then
    h[p*_2][1] = e_v; # Case 2 in Figure 2 with u, v swapped
else
    if p*_2 is connected to a hybrid edge then
        h[p*_2][2] = e_u; # Case 3 in Figure 2
    else
        h[p*_2][2] = e_v; # Case 3 in Figure 2 with u, v swapped
    end
end
```

Algorithm 4: MergeNodesInConstraint!(N, u, v, P)

Input : $N = (V, E_T \cup E_H)$ a constraint network
u, v nodes being merged
P the new parent of u and v

Output: Nothing, N is updated in-place

```
if u ∈ N ∩ v ∉ N then
    Replace u with p;
else if u ∉ N ∩ v ∈ N then
    Replace v with p;
else if u ∈ N ∩ v ∈ N then
    Find p* ∈ u ⇝^{N_i} v;
    if |p| = 2 then
        N = (P, ∅);
    else if |p*| = 3 then
        Remove nodes p*_1, p*_3 from N;
        Remove all edges along path p* from N;
        Replace p*_2 with P;
    else if |p*| = 4 then
        if p*_2 ∈ E_H OR p*_2 is connected to a hybrid edge then
            Remove nodes p*_1, p*_2, p*_4 from N;
            Remove all edges along path p* from N;
            Replace p*_3 with P;
        else
            Remove nodes p*_1, p*_3, p*_4 from N;
            Remove all edges along path p* from N;
            Replace p*_2 with P;
        end
    else
        Remove nodes p*_1, p*_2, p*_4, p*_5 from N;
        Remove all edges along path p* from N;
        Replace p*_3 with P;
    end
end
```

Algorithm 5: PairwiseInPhyNet($\mathbf{N}$, D, X)

Input : X a set of n labels
D a pairwise dissimilarity matrix on X
$\mathbf{N}$ a set of k level-1 networks on disjoint subsets of taxa $\mathbf{S} = \{S_i\}_{i=1}^k$ where $\bigcup_{i=1}^k S_i \subseteq X$

Output: A network $\widehat{\mathcal{N}} = (V, E)$ on taxa X

Initialization
  Create FIFO queue Q;
  for $i = 1, 2, \ldots, k$ **do**
    Push (N_i, S_i) to Q;
  end

while $|Q| > 1$ **do**
  Pop (N_1, S_1) from Q;
  Pop (N_2, S_2) from Q;
  if $|Q| = 2$ **then**
    $I = \{1, 2, \ldots, n\}$;
  else
    $I = \{i : X_i \in S_1 \cup S_2\}$;
  end
  $\widehat{N}_{12}$ =InPhyNet($\{N_1, N_2\}, D_{I,I}, X_I$);
  Push $(\widehat{N}_{12}, S_1 \cup S_2)$ to Q;
end
Pop $\widehat{\mathcal{N}}, X$ from Q;
return $\widehat{\mathcal{N}}$;

B Phylogeny of Seed Plants

Here, we enumerate the subsets (including outliers) with which we inferred constraint networks using SNaQ [40]. Networks were inferred with $h = 0, 1, 2, 3$ reticulations for each subset. We provide the tree of blobs inferred with TINNiK [2] that were used to perform manual subset decomposition in Figures 11 and 13. Additionally, we provide figures displaying the model selection decisions for these subsets in Figures 12 and 14.

B.1 Gymnosperms

- Subset 1 (21 taxa): *Sundacarpus amarus*, *Prumnopitys andina*, *Halocarpus bidwillii*, *Microstrobos fitzgeraldii*, *Phyllocladus hypophyllus*, *Parasitaxus usta*, *Manoao colensoi*, *Lagarostrobos franklinii*, *Saxegothaea conspicua*, *Microcachrys tetragona*, *Acmopyle pancheri*, *Podocarpus coriaceus*, *Podocarpus rubens*, *Nageia nagi*, *Retrophyllum minus*, *Dacrycarpus compactus*, *Dacrydium balansae*, *Falcatifolium taxoides*, *Falcatifolium taxoides*, *Botryococcus braunii*, *Chloromonas perforata*

- Subset 2 (15 taxa): *Sciadopitys verticillata*, *Cephalotaxus harringtonia*, *Amentotaxus argotaenia*, *Torreya taxifolia*, *Torreya nucifera*, *Austrotaxus spicata*, *Pseudotaxus chienii*, *Taxus baccata*, *Taxus cus-*

pidata, *Araucaria rulei*, *Wollemia nobilis*, *Agathis robusta*, *Araucaria sp.*, *Coccomyxa pringsheimii*, *Pleurastrum insigne*

- Subset 3 (15 taxa): *Cunninghamia lanceolata*, *Metasequoia glyptostroboides*, *Sequoiadendron giganteum*, *Taxodium distichum*, *Juniperus scopulorum*, *Calocedrus decurrens*, *Platycladus orientalis*, *Thuja plicata*, *Austrocedrus chilensis*, *Pilgerodendron uviferum*, *Callitris macleayana*, *Callitris gracilis*, *Taiwania cryptomerioides*, *Trebouxia arboricola*, *Stephanosphaera pluvialis*

- Subset 4 (16 taxa): *Athrotaxis cupressoides*, *Sequoia sempervirens*, *Cryptomeria japonica*, *Glyptostrobus pensilis*, *Tetraclinis sp.*, *Microbiota decussata*, *Cupressus dupreziana*, *Fokienia hodginsii*, *Thujopsis dolabrata*, *Papuacedrus papuana*, *Neocallitropsis pancheri*, *Diselma archeri*, *Widdringtonia cedarbergensis*, *Chamaecyparis lawsoniana*, *Leptosira obovata*, *Brachiomonas submarina*

- Subset 5 (19 taxa): *Ginkgo biloba*, *Encephalartos barteri*, *Stangeria eriopus*, *Picea engelmannii*, *Abies lasiocarpa*, *Keteleeria evelyniana*, *Cystopteris reevesiana*, *Huperzia squarrosa*, *Selaginella wallacei*, *Pinus radiata*, *Pinus taeda*, *Amborella trichopoda*, *Nuphar advena*, *Nymphaea sp.*, *Pseudotsuga wilsoniana*, *Ephedra sinica*, *Welwitschia mirabilis*, *Microthamnion kuetzingianum*, *Haematococcus pluvialis*

- Subset 6 (20 taxa): *Cycas micholitzii*, *Dioon edule*, *Cedrus libani*, *Pseudolarix amabilis*, *Nothotsuga longibracteata*, *Tsuga heterophylla*, *Nephrolepis exaltata*, *Dennstaedtia davallioides*, *Selaginella moellendorffii*, *Pinus parviflora*, *Pinus ponderosa*, *Pinus jeffreyi*, *Austrobaileya scandens*, *Kadsura heteroclita*, *Illicium floridanum*, *Cathaya argyrophylla*, *Larix speciosa*, *Gnetum montanum*, *Botryococcus terribilis*, *Chlamydomonas noctigama*,

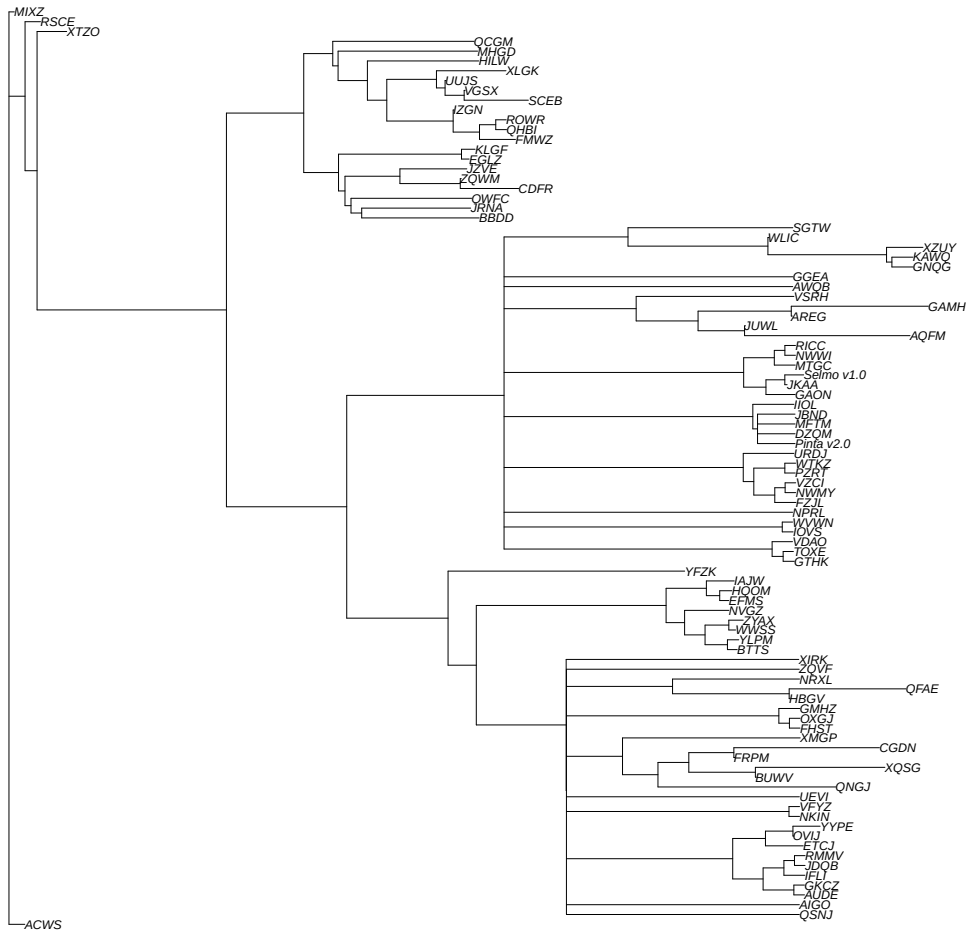

Figure 11: Gymnosperm tree of blobs inferred by TINNiK ($\alpha = 0.01$, $\beta = 0.99$). See Table 1 for code conversions.

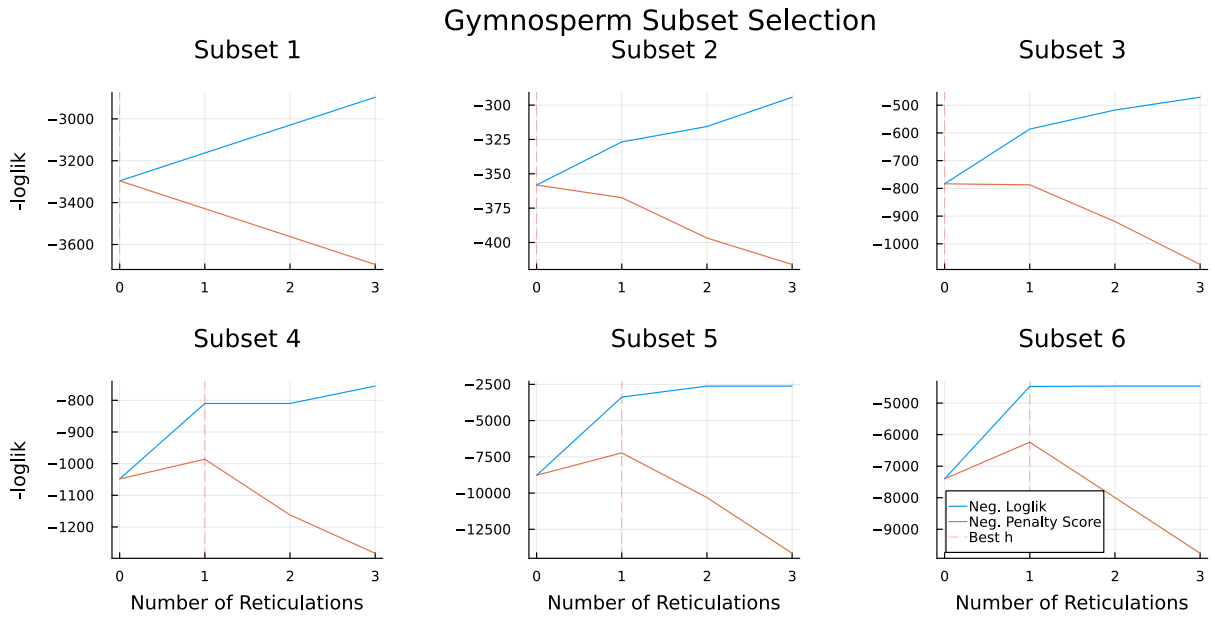

Figure 12: Model selection figures corresponding to each of the 6 subsets listed above. The blue line is the negative log pseudo-likelihood of each model (higher is better), the orange line is the penalized score for each model (higher is better), and the vertical dashed red line is placed at the value of h with the largest penalized score. This is the model that was selected.

B.2 Ferns

- Subset 1 (22 taxa): *Dendrolycopodium obscurum*, *Pseudolycopodiella caroliniana*, *Huperzia selago*, *Huperzia lucidula*, *Huperzia squarrosa*, *Chloromonas tughillensis*, *Oogamochlamys gigantea*, *Volvox aureus*, *Vitreochlamys sp.*, *Pleurastrum insigne*, *Welwitschia mirabilis*, *Cupressus dupreziana*, *Thujopsis dolabrata*, *Diselma archeri*, *Falcatifolium taxoides*, *Dacrydium balansae*, *Halocarpus bidwillii*, *Prumnopitys andina*, *Pinus jeffreyi*, *Pseudolarix amabilis*, *Ettlia oleoabundans*, *Dunaliella tertiolecta*
- Subset 2 (7 taxa): *Selaginella moellendorffii*, *Selaginella willdenowii*, *Selaginella cf. pallescens*, *Isoetes sp.*, *Isoetes tegetiformans*, *Parachlorella kessleri*, *Dunaliella primolecta*
- Subset 3 (23 taxa): *Osmunda javanica*, *Osmunda sp.*, *Osmundastrum cinnamomeum*, *Marattia attenuata*, *Angiopteris evecta*, *Danaea nodosa*, *Dipteris conjugata*, *Hymenophyllum bivalve*, *Ophioglossum petiolatum*, *Sceptridium dissectum*, *Botrypus virginianus*, *Psilotum nudum*, *Tmesipteris parva*, *Lygodium japonicum*, *Anemia tomentosa*, *Pilularia globulifera*, *Azolla cf. caroliniana*, *Dennstaedtia davallioides*, *Didymochlaena truncatula*, *Leucostegia immersa*, *Nephrolepis exaltata*, *Prototheca wickerhamii*, *Dunaliella salina*
- Subset 4 (23 taxa): *Alsophila pinulosa*, *Plagiogyria japonica*, *Culcita macrocarpa*, *Thyrsopteris elegans*, *Lonchitis hirsuta*, *Lindsaea microphylla*, *Lindsaea linearis*, *Cryptogramma acrostichoides*, *Argyrochosma nivea*, *Myriopteris rufa*, *Notholaena montieliae*, *Parahemionitis cordata*, *Gaga arizonica*,

Vittaria lineata, *Vittaria appalachiana*, *Adiantum aleuticum*, *Adiantum raddianum*, *Ceratopteris thalictroides*, *Pityrogramma trifoliata*, *Pteris vittata*, *Pteris ensiformis*, *Picochlorum atomus*, *Dunaliella salina*

- Subset 5 (12 taxa): *Davallia fejeensis*, *Phymatosorus grossus*, *Pleopeltis polypodioides*, *Phlebodium pseudoaureum*, *Polypodium amorphum*, *Polypodium hesperium*, *Polypodium hesperium*, *Polypodium glycyrrhiza*, *Bolbitis repanda*, *Polystichum acrostichoides*, *Eremosphaera viridis*, *Dunaliella salina*

- Subset 6 (21 taxa): *Homalosorus pycnocarpos*, *Thelypteris acuminata*, *Deparia lobato-crenata*, *Diplazium wichurae*, *Blechnum spicant*, *Athyrium filix-femina*, *Athyrium sp.*, *Onoclea sensibilis*, *Woodsia ilvensis*, *Woodsia scopulina*, *Asplenium platyneuron*, *Asplenium nidus*, *Asplenium cf. x lucrosum*, *Cystopteris protrusa*, *Cystopteris fragilis*, *Cystopteris reevesiana*, *Cystopteris fragilis*, *Cystopteris utahensis*, *Gymnocarpium dryopteris*, *Nannochloris atomus*, *Asteromonas gracilis*,

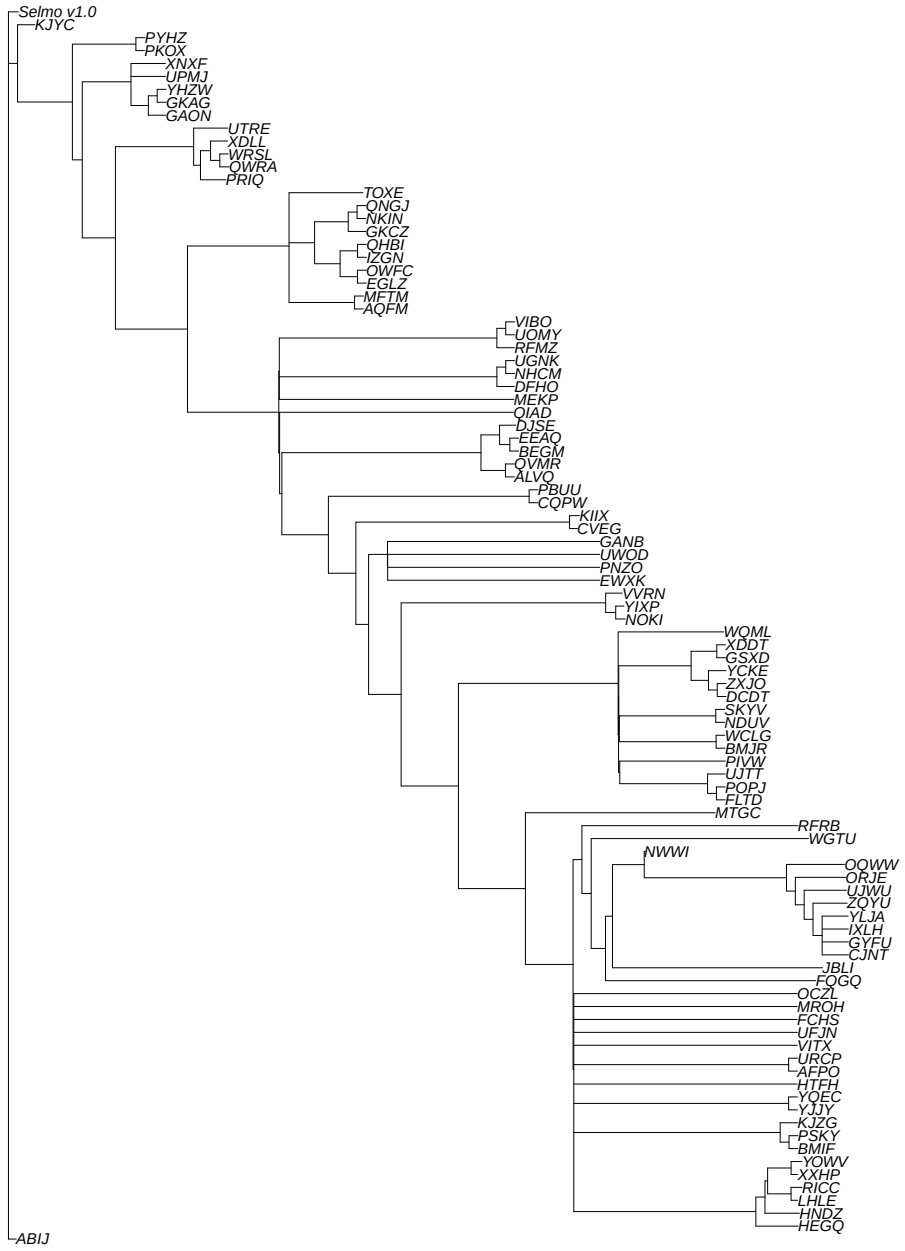

Figure 13: Fern tree of blobs inferred by TINNiK ($\alpha = 0.001$, $\beta = 0.999$). See Table 1 for code conversions.

Figure 14: Model selection figures corresponding to each of the 6 subsets listed above. The blue line is the negative log pseudo-likelihood of each model (higher is better), the orange line is the penalized score for each model (higher is better), and the vertical dashed red line is placed at the value of h with the largest penalized score. This is the model that was selected.

Taxa Codes

Code	Taxa	Code	Taxa	Code	Taxa
ABIJ	Selaginella cf. pallescens	JBLI	Bolbitis repanda	SCEB	Podocarpus coriaceus
ACWS	Araucaria sp.	JBND	Pinus ponderosa	SGTW	Ginkgo biloba
AFPO	Athyrium sp.	JDQB	Neocallitropsis pancheri	SKYV	Vittaria lineata
AIGO	Chamaecyparis lawsoniana	JKAA	Selaginella wallacei	SYJM	Dunaliella salina
AKCR	Parachlorella kessleri	JQFK	Picochlorum atomus	Selmo_v1.0	Selaginella moellendorffii
ALVQ	Tmesipteris parva	JRNA	Phyllocladus hypophyllus	TOXE	Welwitschia mirabilis
AQFM	Pseudolarix amabilis	JUWL	Keteleeria evelyniana	UEVI	Fokienia hodginsii
AREG	Nothotsuga longibracteata	JZVE	Parasitaxus usta	UFJN	Diplazium wichurae
AUDE	Widdringtonia cedarbergensis	KAWQ	Stangeria eriopus	UGNK	Marattia attenuata
AWQB	Picea engelmannii	KFEB	Haematococcus pluvialis	UJTT	Pityrogramma trifoliata
BBDD	Microstrobos fitzgeraldii	KIIX	Pilularia globulifera	UJWU	Pleopeltis polypodioides
BEGM	Botrypus virginianus	KJYC	Selaginella willdenowii	UKUC	Dunaliella salina
BILC	Prototheca wickerhamii	KJZG	Asplenium platyneuron	UOMY	Osmunda sp.
BMIF	Asplenium cf. x lucrosum	KLGF	Sundacarpus amarus	UPMJ	Pseudolycopodiella caroliniana
BMJR	Adiantum raddianum	LHLE	Cystopteris fragilis	URCP	Athyrium filix-femina
BTTS	Austrotaxus spicata	MEKP	Dipteris conjugata	URDJ	Amborella trichopoda
BUWV	Platycladus orientalis	MFTM	Pinus jeffreyi	UTRE	Chloromonas tughillensis
CDFR	Manoao colensoi	MFYC	Nannochloris atomus	UUJS	Nageia nagi
CGDN	Tetraclinis sp.	MHGD	Microcachrys tetragona	UWOD	Plagiogyria japonica
CJNT	Polypodium glycyrrhiza	MIXZ	Agathis robusta	VALZ	Chlamydomonas noctigama
CQPW	Anemia tomentosa	MNCB	Eremosphaera viridis	VDAO	Ephedra sinica
CVEG	Azolla cf. caroliniana	MROH	Thelypteris acuminata	VFYZ	Thuja plicata
DCDT	Gaga arizonica	MTGC	Dennstaedtia davallioides	VGSX	Retrophyllum minus
DFHO	Danaea nodosa	NDUV	Vittaria appalachiana	VIBO	Osmunda javanica
DJSE	Ophioglossum petiolatum	NHCM	Angiopteris evecta	VITX	Blechnum spicant
DZQM	Pinus radiata	NKIN	Thujopsis dolabrata	VSRH	Abies lasiocarpa
EATP	Microthamnion kuetzingianum	NKXU	Trebouxia arboricola	VVRN	Lonchitis hirsuta
EEAQ	Sceptridium dissectum	NOKI	Lindsaea linearis	VZCI	Illicium floridanum
EEJO	Ettlia oleoabundans	NPRL	Cathaya argyrophylla	WCLG	Adiantum aleuticum
EFMS	Torreya taxifolia	NRXL	Metasequoia glyptostroboides	WDWX	Dunaliella primolecta
EGLZ	Prumnopitys andina	NVGZ	Cephalotaxus harringtonia	WGTU	Leucostegia immersa
ETCJ	Pilgerodendron uviferum	NWMY	Kadsura heteroclita	WLIC	Dioon edule
ETGN	Botryococcus braunii	NWWI	Nephrolepis exaltata	WQML	Cryptogramma acrostichoides
EWXK	Thyrsopteris elegans	OCZL	Homalosorus pycnocarpos	WRSL	Volvox aureus
FCHS	Deparia lobato-crenata	OQWW	Davallia fejeensis	WTKZ	Nuphar advena
FHST	Taxodium distichum	ORJE	Phymatosorus grossus	WVWN	Larix speciosa
FLTD	Pteris ensiformis	OVIJ	Papuacedrus papuana	WWSS	Taxus baccata
FMWZ	Dacrycarpus compactus	OWFC	Halocarpus bidwillii	XDDT	Argyrochosma nivea
FQGQ	Polystichum acrostichoides	OXGJ	Glyptostrobus pensilis	XDLL	Oogamochlamys gigantea
FRPM	Calocedrus decurrens	PBUU	Lygodium japonicum	XIRK	Athrotaxis cupressoides
FZJL	Austrobaileya scandens	PIVW	Ceratopteris thalictroides	XLGK	Podocarpus rubens
GAMH	Tsuga heterophylla	PKOX	Isoetes tegetiformans	XMGP	Juniperus scopulorum
GANB	Alsophila pinulosa	PNZO	Culcita macrocarpa	XNXF	Dendrolycopodium obscurum
GAON	Huperzia squarrosa	POPJ	Pteris vittata	XQSG	Microbiota decussata
GGEA	Cedrus libani	PRIQ	Pleurastrum insigne	XTZO	Araucaria rulei
GKAG	Huperzia lucidula	PSKY	Asplenium nidus	XXHP	Cystopteris fragilis
GKCZ	Diselma archeri	PYHZ	Isoetes sp.	XZUY	Cycas micholitzii
GMHZ	Cryptomeria japonica	PZRT	Nymphaea sp.	YCKE	Notholaena montieliae
GNQG	Encephalartos barteri	Pinta_v2.0	Pinus taeda	YFZK	Sciadopitys verticillata
GSXD	Myriopteris rufa	QCGM	Saxegothaea conspicua	YHZW	Huperzia selago
GTHK	Gnetum montanum	QFAE	Sequoiadendron giganteum	YIXP	Lindsaea microphylla
GUBD	Brachiomonas submarina	QHBI	Falcatifolium taxoides	YJJY	Woodsia scopulina
GXBM	Coccomyxa pringsheimii	QIAD	Hymenophyllum bivalve	YLJA	Polypodium amorphum
GYFU	Polypodium hesperium	QNGJ	Cupressus dupreziana	YLPM	Pseudotaxus chienii
HBGV	Sequoia sempervirens	QRTH	Chloromonas perforata	YOWV	Cystopteris protrusa
HEGQ	Gymnocarpium dryopteris	QSNJ	Taiwania cryptomerioides	YQEC	Woodsia ilvensis
HILW	Acmopyle pancheri	QVMR	Psilotum nudum	YYPE	Austrocedrus chilensis
HNDZ	Cystopteris utahensis	QWRA	Vitreochlamys sp.	ZDIZ	Dunaliella tertiolecta
HQOM	Torreya nucifera	QYXY	Botryococcus terribilis	ZFXU	Asteromonas gracilis
HTFH	Onoclea sensibilis	RFMZ	Osmundastrum cinnamomeum	ZLQE	Stephanosphaera pluvialis
IAJW	Amentotaxus argotaenia	RFRB	Didymochlaena truncatula	ZNUM	Leptosira obovata
IFLI	Callitris gracilis	RHVC	Dunaliella salina	ZQVF	Cunninghamia lanceolata
IIOL	Pinus parviflora	RICC	Cystopteris reevesiana	ZQWM	Lagarostrobos franklinii
IOVS	Pseudotsuga wilsoniana	RMMV	Callitris macleayana	ZQYU	Phlebodium pseudoaureum
IXLH	Polypodium hesperium	ROWR	Falcatifolium taxoides	ZXJO	Parahemionitis cordata
IZGN	Dacrydium balansae	RSCE	Wollemia nobilis	ZYAX	Taxus cuspidata

Table 1: Codes used in tree of blobs Figures 11 and 13 and their corresponding taxa (arranged alphabetically according to code).