A method for massively scalable phylogenetic network inference

Nathan Kolbow[a,b], **Sungsik Kong**[a], **and Claudia Solis-Lemus**[a,c,1]

This manuscript was compiled on May 5, 2025

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.

Phylogenetics | Semi-directed network inference | Land plants | Divide-and-conquer

Significance Statement

The development of sequencing technologies has led to an unprecedented availability of genomic data, but scalable methods for analyzing such data have not kept up with their demand. Phylogenetic trees can be inferred with thousands of taxa, but model-based phylogenetic network inference is only computationally feasible for a few dozen taxa. Here, we present a novel method for accurate semi-directed phylogenetic network inference that scales linearly with the number of taxa in the input data by merging independently estimated, non-overlapping networks.

Advancements in sequencing technologies have increased the availability of biological genomic data for phylogenetic analyses. These analyses are foundational to our understandings of many biological processes such as cancerous tumour progression (1), infectious disease spread (2), drug discovery (3), conservation biology (4), and crop domestication in agriculture (5). 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 prevalent in the histories of many species (6). 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) (7), Phylogenetic Network Estimation using SiTe patterns (PhyNEST) (8), or PhyloNet's maximum pseudolikelihood (MPL) option (9) 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 (10) or PhyloNet's divide-and-conquer option (11), 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 (12), Consensus Networks (13), or PhyloFusion (14) are often employed when researchers wish to analyze hundreds or thousands of taxa with a network model (15). However, such analyses are often inappropriate given their lack of biological grounding (16, 17).

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

Author affiliations: [a]Wisconsin Institute for Discovery; [b]University of Wisconsin-Madison, Department of Statistics; [c]University of Wisconsin-Madison, Department of Plant Pathology

NK developed software, performed simulation study, developed theory, and wrote the manuscript. SK helped design the simulation study, edited the manuscript, and provided oversight throughout. CSL edited the manuscript and provided oversight throughout.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: solislemuswisc.edu

DRAFT

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 (7) in small simulation settings, and against NeighborNet (12) and Consensus Networks (13) in the SplitsTree software (18) in larger simulation settings where more accurate methods are unavailable.

Lastly, we re-evaluate the phylogeny of 1,158 land plants (19) 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.

Model

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 (20) which itself is derived from Neighbor-Joining (21). 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 (21).
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 (22).
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 3 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 Supporting Information.

Retaining Reticulate Relationships.

Definition 1 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 (23) $r_{\text{major}}, r_{\text{minor}}$ 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 2 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 3 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 1.

We say that each p^* contains a node L_p as depicted in Figure 1. 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 1 based on the structure of p^*. After these operations, the reticulate structure of N_k is accurately reflected in $\widehat{\mathcal{N}}$.

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 (21) 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.

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

DRAFT

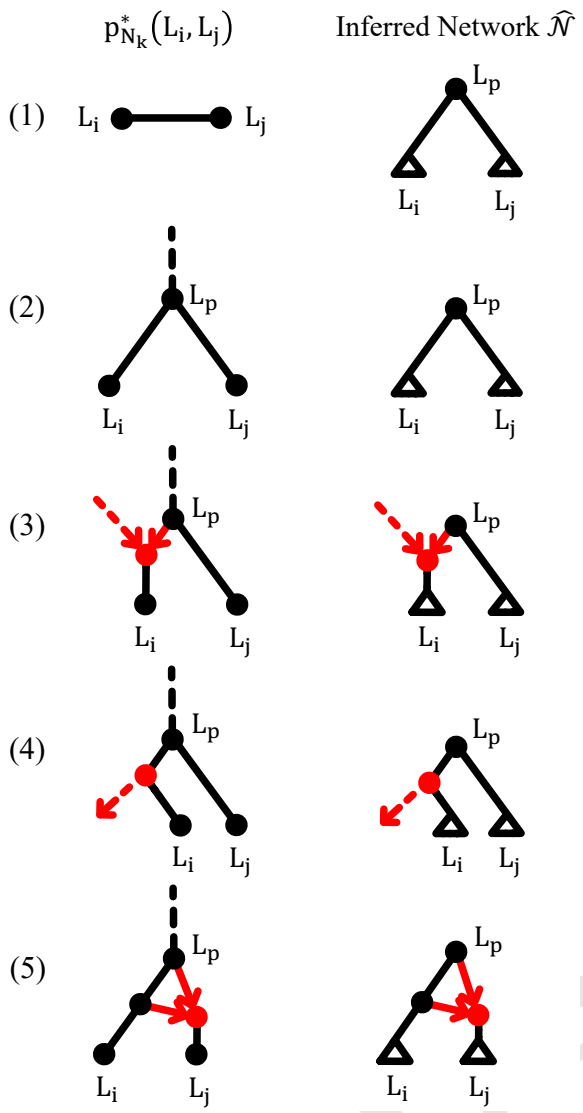

Fig. 1. All possible neighbor relationships between two nodes L_i, L_j in a constraint network N_k along with their corresponding $p^*_{N_k}(L_i, L_j)$ paths and L_p nodes.

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 (24)) given that certain assumptions are met in steps 1-3 (see Theorem 3). 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 $O(kf(m))$.

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 (25) to network space. See the Performance Metrics subsection of Materials and Methods for further details.

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 (12) and Consensus Networks (13) when $n = 200$ (see Figure 3) in terms of accuracy and efficiency. We also attempted to infer networks with PhyloFusion (14), 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 3).

Across all simulations, we observed decreased HWCD as ILS decreased, as the number of gene trees increased, and as m increased. InPhyNet's accuracy was highly sensitive to the input data's accuracy in all simulation settings (see Figure 2).

Simulations with $m = 20$ had considerably higher runtimes than those with $m = 10$ (Figure 4). 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.

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 (19). 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 (26). The phylogenetic network inferred

DRAFT

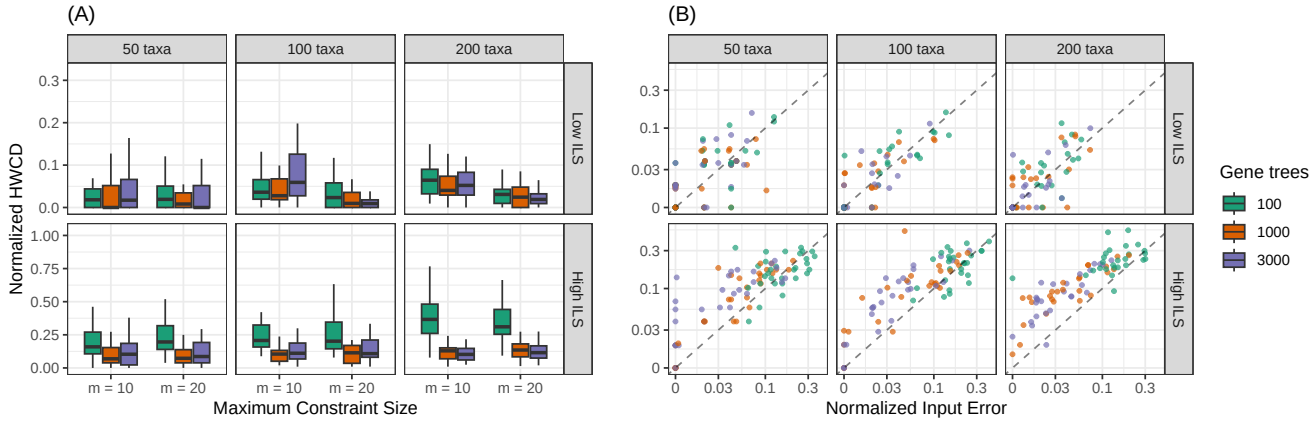

Fig. 2. Normalized hardwired cluster distance (HWCD, see Equation 5) of InPhyNet across 900 simulations (25 per parameter combination). Simulations with differing levels of input data (gene trees) are represented with different colors. **A:** Normalized HWCD observed across various network sizes (columns; 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 6). 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.

by InPhyNet had 4 reticulations, each of which are well supported by previous research (see Figure 5 and Discussion).

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 demonstrated 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; 27), a method that explicitly requires manual adjustment of its parameters in order to attain

DRAFT

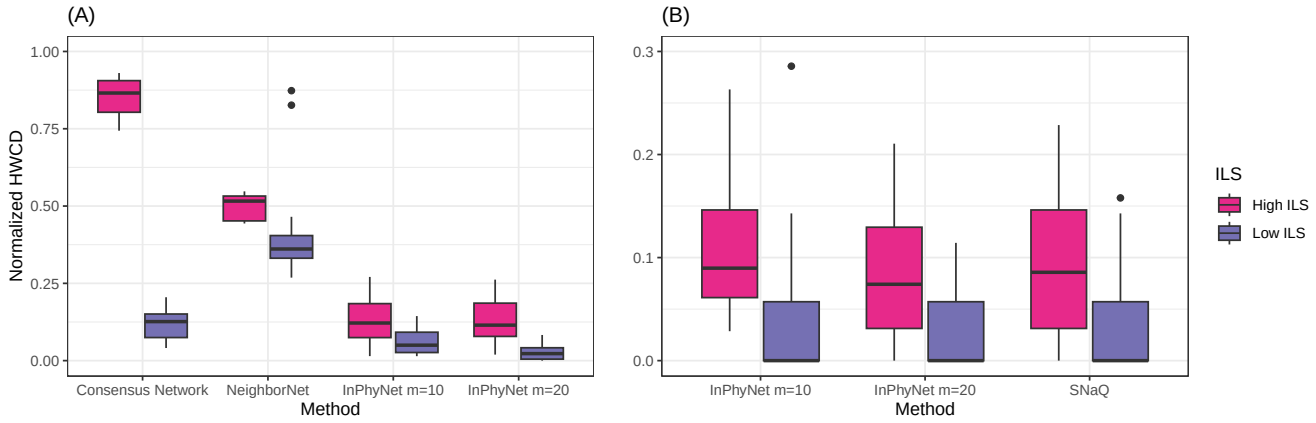

Fig. 3. Normalized hardwired cluster distance (HWCD, see Equation 5) of the networks obtained from Consensus Network, NeighborNet, SNaQ, and InPhyNet analyses (x-axis). m refers to the maximum size of constraint networks used as input to InPhyNet. **A:** Normalized HWCD for the networks inferred with 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.

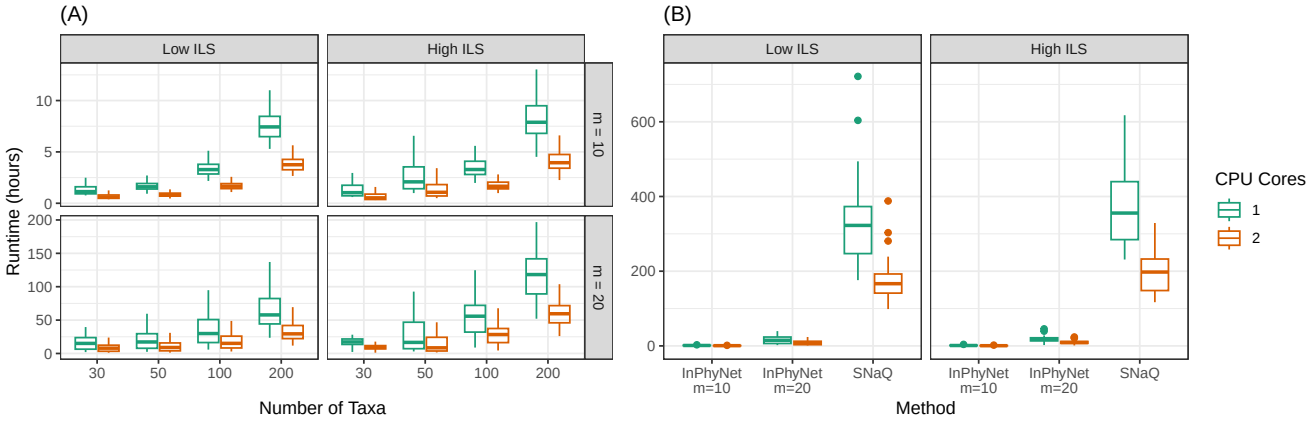

Fig. 4. Total wall-clock runtime in hours to infer all of the constraint networks for a given simulation replicate. Runtimes with various cores were computed after the fact but are not simple averages of overall runtime (i.e. $\frac{\text{total runtime}}{\text{number of cores}}$). Instead, they are actual 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).

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 4 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.

Phylogeny of Seed Plants. Our species network generally agreed with previously inferred phylogenies (19) 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 (19, 28, 29). We then focus our discussion on the reticulate relationships uncovered by InPhyNet.

Gnetales. The placement of Gnetales in the phylogeny of gymnosperms is controversial. Some research suggests that they are sister to Pinaceae (36–41), while others suggest that they are sister to conifers (36, 37, 42). Treelike analyses of these species lead to significant conflict between each of these hypotheses (19), 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 (40), so this result requires further investigation in the future.

Pinaceae. High rates of polyploidy have been identified in genus *Juniperus* (30), and similar evidence has been found in *Cupressales* and *Platycladus* (31). 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* (32), corroborating their placement in this phylogeny.

While SNaQ does not directly model polyploidization, PhyloNet (43), 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.

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* (33). 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 (34), or that have otherwise found generally weak support for treelike backbones between these species (35).

Statistical Consistency.

Theorem 1 Given true species tree T, Neighbor-Joining (21) will construct T from distance matrix D only if $Q_1(D) < 0$ and $Q_2(D) < 0$ (Equations 2 and 3) for each pair of true neighbors in T (labelled $1, 2$) and arbitrary other taxa (labelled i, j).

DRAFT

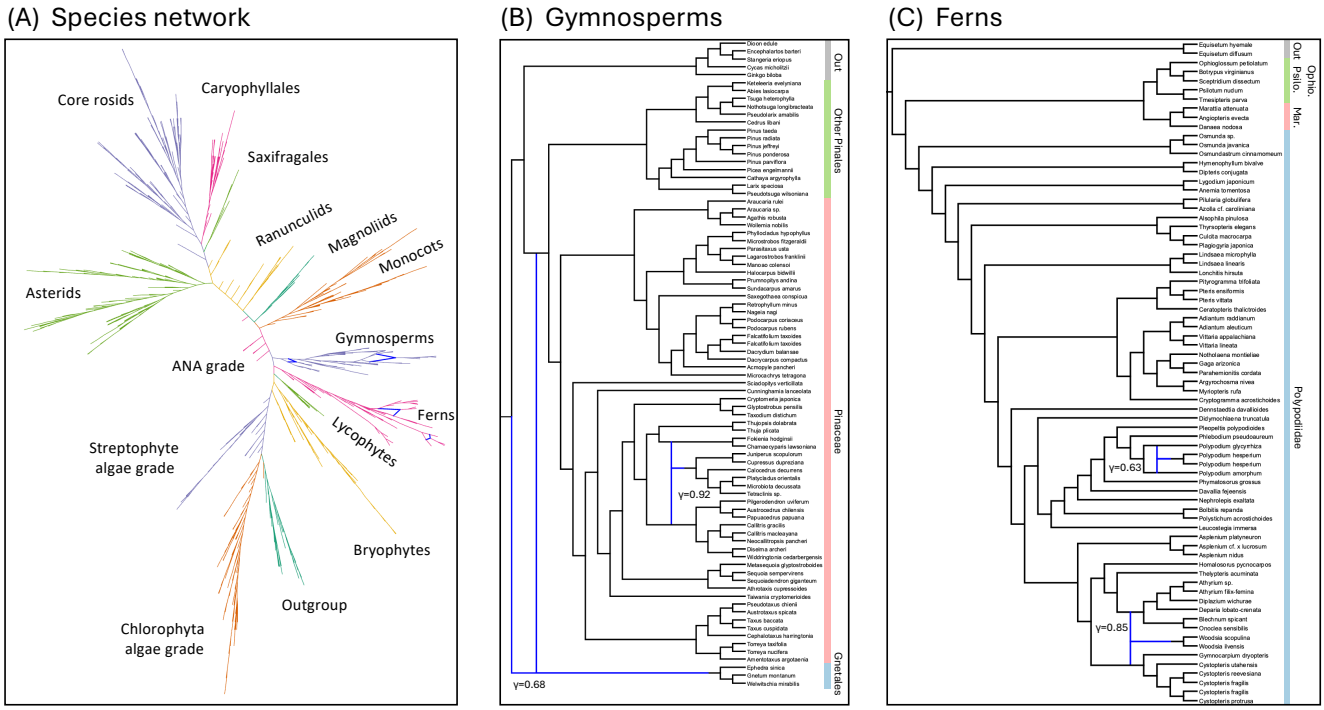

Fig. 5. **A:** Semi-directed phylogenetic network of seed plants obtained using InPhyNet. Various colored branches signify different clades within the phylogeny. Branch lengths are not scaled. **B:** Magnified portion of the phylogeny corresponding to gymnosperms. The reticulation ancestral to Gnetales 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 (30–32). 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 (33–35).

DRAFT

$$Q_1(D) = D_{1,2} + D_{i,j} - D_{1,i} - D_{2,j} \quad [2]$$
$$Q_2(D) = D_{1,2} + D_{i,j} - D_{1,j} - D_{2,i} \quad [3]$$

Proof. Presented in (21).

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

Proof. The inferred network $\widehat{N}$ is first constructed as a tree in Algorithm 1 (see Supporting Information) before reticulations are added in post-processing. By Theorem 1, if all initial proposals are accepted in Algorithm 2 (see Supporting Information), 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 3 Given true level-1 species network $\mathcal{N}$, constraint networks $\mathbf{C} = \{C_1, \ldots, C_k\}$ on disjoint sets of taxa $\mathbf{S} = \{S_1, \ldots, S_k\}$, and distance matrix D on taxa X where $\bigcup_{i=1}^{k} S_i \subseteq X$, $\phi(D, \mathbf{C}) = \mathcal{N}$ if the following are true:

1. D satisfies the conditions in Theorem 1 for an arbitrary displayed tree T of $\mathcal{N}$.
2. All $C_i \in \mathbf{C}$ agree with $\mathcal{N}$.
3. For each reticulation $r_j \in R(\mathcal{N})$, at least one subset $S_i \in \mathbf{S}$ contains at least one individual from each of the clades A, B, C, D, H as depicted in Figure 6.

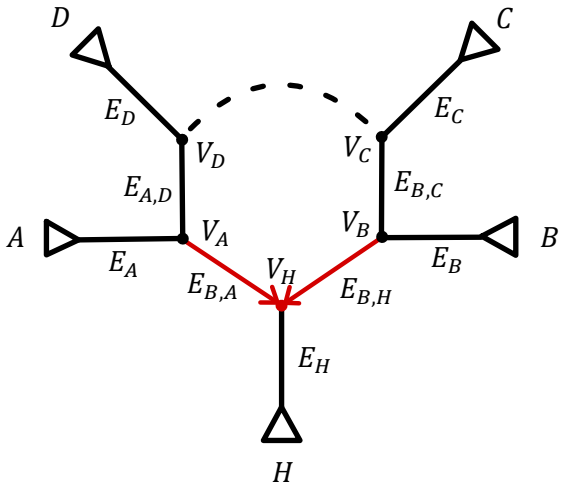

Fig. 6. Cycle corresponding to hybrid node V_H in the true species network with labellings corresponding to those used in Theorem 3. This structure is also displayed in any constraint that meets condition 3 of Theorem 3 but with labels corresponding to the relevant subsets of each clade $\tilde{A} \subseteq A$, $\tilde{B} \subseteq B$, etc.

Proof. From (1), if InPhyNet's initial merging proposals are always accepted, then according to Corollary 2 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 6. Notice that there are only two points throughout Algorithm 3 (see SI) 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, 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 part E of Figure 7). 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 (see Supporting Information).

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 $\widehat{\mathcal{N}}$. □

Materials and Methods

Simulation data.

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 (44) 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 4 in order to achieve the desired level of ILS. In Equation 4, 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.

$$L_i^{\text{new}} = \frac{\lambda L_i^{\text{old}}}{\bar{L}} \qquad [4]$$

Using the generated species network, true gene trees were simulated with the PhyloCoalSimulations (45) 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 (46). Multiple sequence alignments with 1,000 b.p. were simulated for each gene tree under the F81 model of evolution (47) with state frequencies $\pi_A = \pi_T = 0.3$ and $\pi_C = \pi_G = 0.2$ and a scaling parameter of 0.036.

Estimated data. IQ-TREE 2's ModelFinder (48) was used to select the best nucleotide substitution model to infer gene trees from the simulated sequence alignments. Note the Jukes-Cantor model (49) 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 (7) 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.

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 (50).
2. Infer a tree of blobs T_{blob} on X with TINNiK (27).
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.

Performance Metrics. To assess the accuracy of InPhyNet across topologies with widely varying numbers of taxa, we extended the

DRAFT

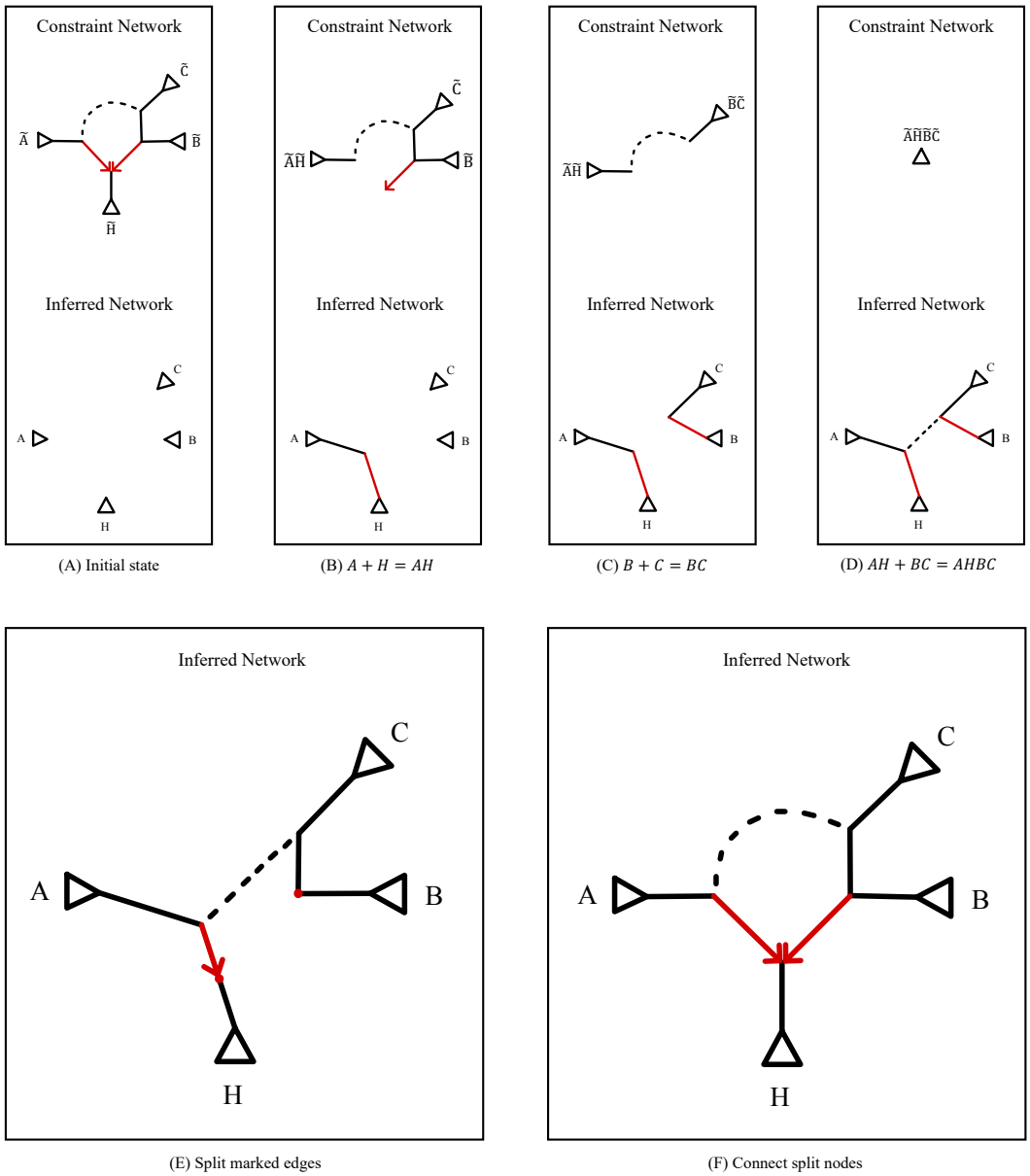

Fig. 7. A: the initial relevant state of the algorithm at which point all taxa within each clade have been merged. **B:** nodes A and H are merged. The path p^* connecting them in the constraint network traverses one reticulation, so its corresponding edge is marked in the inferred network. **C:** nodes 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.

normalized Robinson-Foulds distance metric (25) 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.

$$\text{HWCD}_{\text{norm}}(N_1, N_2) = \frac{\text{HWCD}(N_1, N_2)}{\left|E^{(1)}\right| - \left|V_L^{(1)}\right| + \left|E^{(2)}\right| - \left|V_L^{(2)}\right|} \quad [5]$$

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 6, 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.

$$\text{error}(D, \mathbf{C}) = \frac{\min\limits_{T \in \text{disp}(\mathcal{N})} \text{HWCD}\left(T, \text{NJ}(D)\right) + \sum_{i=1}^{k} \text{HWCD}\left(C_i, \hat{C}_i\right)}{|E^{(T)}| - |V^{(T)}| + E^{(\text{NJ}(D))} - V^{(\text{NJ}(D))} + \sum_{i=1}^{k} \left(|E^{(C_i)}| - |V^{(C_i)}| + |E^{(\hat{C}_i)}| - |V^{(\hat{C}_i)}|\right)} \quad [6]$$

Phylogeny of Seed Plants. Our analyses were based on estimated gene trees provided in Leebens-Mack et al. (19). 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 3 (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 (27). 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 (7).

For each subset, 20 networks were inferred using SNaQ (7) 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 (26). 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 (19). 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 3). 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.

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].

ACKNOWLEDGMENTS. 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.

DRAFT

1. R Schwartz, AA Schäffer, The evolution of tumour phylogenetics: Principles and practice. *Nat. reviews. Genet.* **18**, 213–229 (2017).
2. M Ciccozzi, et al., The phylogenetic approach for viral infectious disease evolution and epidemiology: An updating review. *J. Med. Virol.* **91**, 1707–1724 (2019).
3. DB Searls, Pharmacophylogenomics: Genes, evolution and drug targets. *Nat. Rev. Drug Discov.* **2**, 613–623 (2003).
4. S Véron, V Saito, N Padilla-García, F Forest, Y Bertheau, The Use of Phylogenetic Diversity in Conservation Biology and Community Ecology: A Common Base but Different Approaches. *The Q. Rev. Biol.* **94**, 123–148 (2019).
5. T Sang, S Ge, Genetics and phylogenetics of rice domestication. *Curr. Opin. Genet. & Dev.* **17**, 533–538 (2007).
6. L Boto, Horizontal gene transfer in evolution: Facts and challenges. *Proc. Royal Soc. B: Biol. Sci.* **277**, 819–827 (2009).
7. C Solís-Lemus, C Ané, Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLOS Genet.* **12**, e1005896 (2016).
8. S Kong, DL Swofford, LS Kubatko, Inference of Phylogenetic Networks from Sequence Data using Composite Likelihood (2022).
9. Y Yu, L Nakhleh, A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* **16**, S10 (2015).
10. HA Hejase, N VandePol, GM Bonito, KJ Liu, FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data in *Comparative Genomics*, eds. M Blanchette, A Ouangraoua. (Springer International Publishing, Cham), pp. 242–259 (2018).
11. X Yu, T Le, SA Christensen, EK Molloy, T Warnow, Advancing Divide-and-Conquer Phylogeny Estimation using Robinson-Foulds Supertrees, (Bioinformatics), Preprint (2020).
12. D Bryant, V Moulton, Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).
13. BR Holland, KT Huber, V Moulton, PJ Lockhart, Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny. *Mol. Biol. Evol.* **21**, 1459–1461 (2004).
14. L Zhang, B Cetinkaya, DH Huson, PhyloFusion- Fast and easy fusion of rooted phylogenetic trees into rooted phylogenetic networks (2024).
15. P Forster, L Forster, C Renfrew, M Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. United States Am.* **117**, 9241–9243 (2020).
16. S Kong, SJ Sánchez-Pacheco, RW Murphy, On the use of median-joining networks in evolutionary biology. *Cladistics* **32**, 691–699 (2016).
17. SJ Sánchez-Pacheco, S Kong, P Pulido-Santacruz, RW Murphy, L Kubatko, Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. *Proc. Natl. Acad. Sci. United States Am.* **117**, 12518–12519 (2020).
18. DH Huson, D Bryant, The SplitsTree App: Interactive analysis and visualization using phylogenetic trees and networks. *Nat. Methods* **21**, 1773–1774 (2024).
19. JH Leebens-Mack, et al., One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
20. EK Molloy, T Warnow, TreeMerge: A new method for improving the scalability of species tree estimation methods. *Bioinformatics* **35**, i417–i426 (2019).
21. N Saitou, M Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
22. EK Molloy, T Warnow, NJMerge: A Generic Technique for Scaling Phylogeny Estimation Methods and Its Application to Species Trees in *Comparative Genomics*, eds. M Blanchette, A Ouangraoua. (Springer International Publishing), pp. 260–276 (2018).
23. J Döcker, S Linz, C Semple, Displaying trees across two phylogenetic networks. *Theor. Comput. Sci.* **796**, 129–146 (2019).
24. JH Degnan, Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* **67**, 786–799 (2018).
25. DF Robinson, LR Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
26. JP Baudry, C Maugis, B Michel, Slope heuristics: Overview and implementation. *Stat. Comput.* **22**, 455–470 (2012).
27. ES Allman, H Baños, JD Mitchell, JA Rhodes, TINNiK: Inference of the Tree of Blobs of a Species Network Under the Coalescent. *bioRxiv* p. 2024.04.20.590418 (2024).
28. GQ Liu, L Lian, W Wang, The Molecular Phylogeny of Land Plants: Progress and Future Prospects. *Diversity* **14**, 782 (2022).
29. NJ Wickett, et al., Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* **111**, E4859–E4868 (2014).
30. P Farhat, et al., Polyploidy in the Conifer Genus Juniperus: An Unexpectedly High Rate. *Front. Plant Sci.* **10** (2019).
31. P Farhat, et al., Polyploidy in Cupressaceae: Discovery of a new naturally occurring tetraploid, Xanthocyparis vietnamensis. *J. Syst. Evol.* **60**, 824–834 (2022).
32. BJM Zonneveld, Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nord. J. Bot.* **30**, 490–502 (2012).
33. EM Sigel, MD Windham, KM Pryer, Evidence for reciprocal origins in Polypodium hesperium (Polypodiaceae): A fern model system for investigating how multiple origins shape allopolyploid genomes. *Am. J. Bot.* **101**, 1476–1485 (2014).
34. AR Smith, et al., A classification for extant ferns. *TAXON* **55**, 705–731 (2006).
35. Y Shao, R Wei, X Zhang, Q Xiang, Molecular Phylogeny of the Cliff Ferns (Woodsiaceae: Polypodiales) with a Proposed Infrageneric Classification. *PLOS ONE* **10**, e0136318 (2015).

36. LM Bowe, G Coat, CW dePamphilis, Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl. Acad. Sci.* **97**, 4092–4097 (2000).
37. DE Soltis, PS Soltis, MJ Zanis, Phylogeny of seed plants based on evidence from eight genes. *Am. J. Bot.* **89**, 1670–1681 (2002).
38. JH Ran, TT Shen, MM Wang, XQ Wang, Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proc. Royal Soc. B: Biol. Sci.* **285**, 20181012 (2018).
39. SM Chaw, CL Parkinson, Y Cheng, TM Vincent, JD Palmer, Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl. Acad. Sci.* **97**, 4086–4091 (2000).
40. B Zhong, et al., Systematic error in seed plant phylogenomics. *Genome Biol. Evol.* **3**, 1340–1348 (2011).
41. TWA Braukmann, M Kuzmina, S Stefanović, Loss of all plastid ndh genes in Gnetales and conifers: Extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.* **55**, 323–337 (2009).
42. S Magallón, MJ Sanderson, Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. *Am. J. Bot.* **89**, 1991–2006 (2002).
43. D Wen, Y Yu, J Zhu, L Nakhleh, Inferring Phylogenetic Networks Using PhyloNet. *Syst. Biol.* **67**, 735–740 (2018).
44. JA Justison, C Solis-Lemus, TA Heath, SiPhyNetwork: An R package for simulating phylogenetic networks. *Methods Ecol. Evol.* **14**, 1687–1698 (2023).
45. J Fogg, ES Allman, C Ané, PhyloCoalSimulations: A Simulator for Network Multispecies Coalescent Models, Including a New Extension for the Inheritance of Gene Flow. *Syst. Biol.* **72**, 1171–1179 (2023).
46. A Rambaut, NC Grass, Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**, 235–238 (1997).
47. J Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
48. BQ Minh, et al., IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
49. TH Jukes, CR Cantor, Evolution of Protein Molecules in *Mammalian Protein Metabolism*, ed. HN Munro. (Academic Press), pp. 21–132 (1969).
50. C Zhang, M Rabiee, E Sayyari, S Mirarab, ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma.* **19**, 153 (2018).

DRAFT