A method for massively scalable inference of phylogenetic networks

NATHAN KOLBOW[1,2], SUNGSIK KONG[1,3], AND CLAUDIA SOLÍS-LEMUS[1,4,*]

[1] *Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, 53706, USA*
[2] *Department of Statistics, University of Wisconsin-Madison, Madison, 53706, USA*
[3] *RIKEN iTHEMS, Wako, Saitama, 351-0198, Japan*
[4] *Department of Plant Pathology, University of Wisconsin-Madison, Madison, 53706, USA*
[*] *Email: solislemus@wisc.edu*

Abstract — Recent advances 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, in which case tree-based models are insufficient. Phylogenetic networks can capture such complex evolutionary histories, but current methods for accurately inferring them lack scalability. Implicit network inference methods are fast but lack biological interpretability. Here, we introduce a novel method called InPhyNet that merges a set of non-overlapping, independently inferred level-1 networks into a unified topology, achieving linear scalability while maintaining high accuracy under the multispecies network coalescent model. We prove that a statistically consistent and scalable pipeline for inferring phylogenetic networks can be constructed with InPhyNet. Using simulation, we infer networks with up to 200 taxa and show that divide-and-conquer pipelines utilizing InPhyNet allow for accurate network inference at scales and speeds previously unseen. Re-analyzing a phylogeny of 1,158 land plants with InPhyNet, we recover well-accepted reticulate events and illustrate how InPhyNet enables large-scale analyses of biologically meaningful reticulate phylogenies at previously unprecedented scales.

Key words: Phylogenetic network inference, divide-and-conquer, scalable network inference, statistical consistency

INTRODUCTION

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 (Schwartz and Schäffer, 2017), infectious disease spread (Ciccozzi et al., 2019), drug discovery (Searls, 2003), conservation biology (Véron et al., 2019), and crop domestication in agriculture (Sang and Ge, 2007). To date, the backbone of these analyses has been phylogenetic trees: binary and bifurcating 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 (Boto, 2009). 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 composite likelihood (sometimes referred to as pseudo-likelihood) inference methods such as Species Network applying Quartets (SNaQ) (Solís-Lemus and Ané, 2016), Phylogenetic Network Estimation using SiTe patterns (PhyNEST) (Kong et al., 2025), or PhyloNet's maximum pseudolikelihood (MPL) option (Yu and Nakhleh, 2015) 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 (Hejase et al., 2018) or PhyloNet's divide-and-conquer option (Yu et al., 2020), the latter of which can handle up to 81 taxa. However, these methods still have the fundamental issue of high-degree polynomial runtime complexity.

This runtime issue can be circumvented by inferring implicit networks, a type of networks that do not directly model biological phenomena but summarize conflicting signals in data. Methods for inferring implicit networks such as NeighborNet (Bryant and Huson, 2023) or Consensus Networks (Holland et al., 2004) are often employed when researchers wish to analyze hundreds or thousands of taxa with a network model (Forster et al., 2020). However, such analyses are often inappropriate given their lack of biological grounding (Kong et al., 2016; Sánchez-Pacheco et al., 2020).

Here, we present InPhyNet, a novel method for constructing arbitrarily large semi-directed phylogenetic networks that scales *linearly* with respect to the number of taxa in the network. InPhyNet works by merging a set of level-1, semi-directed, binary phylogenetic networks with the guidance of a pairwise dissimilarity matrix. This results in a semi-directed, binary, tree-based phylogenetic network that is not necessarily level-1. This approach facilitates the accurate, efficient inference of phylogenetic networks with over

1,000 species while leveraging highly accurate maximum likelihood (ML) and maximum composite likelihood (MCL) methods.

We show that such a pipeline can be statistically consistent for the underlying species network under the multispecies network coalescent model (Yu et al., 2014). The practicality of this theory is verified with a rigorous simulation study that evaluates InPhyNet's performance under realistically complex conditions.

Lastly, we re-evaluate the phylogeny of 1,158 land plant species (One Thousand Plant Transcriptomes Initiative, 2019) by inferring reticulations in portions of the phylogeny that were poorly resolved in their estimated phylogenetic tree. We find support for previously well-accepted 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 an open-source Julia package at https://github.com/NathanKolbow/InPhyNet.jl and all scripts and results used in the simulations presented here are available at github.com/NathanKolbow/InPhyNetSimulations.

Materials and Methods

Scalable Inference Framework

This framework aims to make phylogenetic network inference feasible for larger numbers of input taxa. To achieve this, we use a divide-and-conquer strategy: the original, intractable problem is divided into smaller, tractable subproblems, and a novel method is then applied to merge the solutions of these subproblems into a final result. For a set of input data G of arbitrary type on the set of taxa X, the framework works as follows:

1. Decompose X into disjoint subsets $\mathbf{S} = \{S_i\}_{i=1}^k$ where $\cup_{i=1}^k S_i = X$.
2. Infer a level-1 network C_i on each subset of taxa S_i, giving $\mathbf{C} = \{C_i\}_{i=1}^k$.
3. Infer a pairwise dissimilarity matrix D on X.
4. Utilize InPhyNet to combine the information in D and $\mathbf{C}$ to infer a network $\widehat{\mathcal{N}}$ on all taxa in X.

Each of these steps is agnostic to the type of input data used or specific method used, so they allow researchers to utilize whatever methods best fit their own use cases. If the networks in $\mathbf{C}$ and the dissimilarity matrix D are inferred in a statistically consistent manner, and the subset decomposition $\mathbf{S}$ meets certain criteria, we prove in Section 6 that such a framework is statistically consistent for inferring the true underlying species network $\mathcal{N}$. The main contribution of our work is in developing a novel method for step 4 which we call InPhyNet. One drawback of our approach is that reticulations cannot be inferred between the disjoint subsets $\mathbf{S}$ constructed in step 1, but this does not mean that interclade or ancestral hybridizations cannot be inferred with InPhyNet. Instead, it means that great care needs to be taken when constructing the subset decomposition. Below, we propose a practical method for performing such a decomposition while maintaining the identifiability of such reticulate relationships.

Preliminaries

We use the definitions and notation for phylogenetic networks and their components presented in Ané et al. (2024). Additionally, we assume going forward that inferred inheritance parameters γ are never exactly equal to 0.5.

Definition 1 A *semi-directed phylogenetic network* on taxon set X is an unrooted, connected, acyclic, binary, semi-directed graph with nodes $V = V_L \cup V_H \cup V_T$ and edges $E = E_T \cup E_H$ with corresponding edge lengths $\{l_e\}_{e \in E}$ and inheritance parameters $\{\gamma_e\}_{e \in E_H}$. V_L are the network's set of labelled leaves in bijection with X, V_T are internal *tree vertices* without any incoming directed edges, and V_H are internal *hybrid vertices* with two incoming directed edges and an additional undirected edge. In this paper we assume that all vertices in V_T and V_H have degree 3 whereas those in V_L have degree 1. Additionally, we assume that each $\gamma_e \in (0, 1)$ and $\gamma_e + \gamma_{e'} = 1$ where e and e' are directed edges incident to the same hybrid vertex.

Networks in this paper are assumed to be semi-directed and binary unless otherwise stated and do not necessarily contain any hybrid vertices.

Definition 2 A *pairwise dissimilarity matrix* is a symmetric matrix with zeros along its diagonal where entry D_{AB} is a measure of dissimilarity between taxa $A \in X$ and $B \in X$. Values are in the interval $(0, \infty)$ and higher values indicate that A and B are more dissimilar.

Definition 3 A *level-k* network is a semi-directed network in which each biconnected component contains at most k hybrid nodes.

Definition 4 The *displayed trees* of network $N = (V, E)$ are the set of $2^{|V_H|}$ trees obtained by, for each hybrid node $v \in V_H$ with incoming hybrid edges e and e', removing either e or e' and fusing edges incident to degree-2 nodes. We denote the set of displayed trees of N as $\text{disp}(N)$. The *major tree* of N is the displayed tree obtained by removing e if $\gamma_e < \gamma_{e'}$, else by removing e'.

Definition 5 For semi-directed network N on taxon set X, a subnetwork *spanned by* $Y \subseteq X$ is the minimal subgraph of N, constructed exclusively from edges with $\gamma > 0.5$, necessary to connect all taxa in Y, along

with any edges with $\gamma < 0.5$ whose vertices are contained in this subgraph. This subgraph can be constructed by taking the major tree of N *without* fusing edges incident to degree-2 nodes, then adding any minor edges whose incident nodes are both contained in the subgraph and fusing edges incident to degree-2 nodes. A network N' on taxa Y is said to *agree with* N if $Y \subseteq X$ and N' is topologically equivalent to the subnetwork of N spanned by Y.

DEFINITION 6 A *path* from leaf u to leaf w in network $N = (V, E)$ is an ordered set of vertices $\{v_i\}_{i=1}^k$ where each $v_i \in V$ is unique, $v_1 = u$, $v_k = w$, and an edge exists in E that connects either v_j to v_{j+1} or v_{j+1} to v_j for $j = 1, 2, \ldots, k-1$. Two leaves u, w in network N are *neighbors* if there exists a path p from u to w in any displayed tree of N where $|p| \leqslant 3$.

InPhyNet

The InPhyNet algorithm takes as input a pairwise dissimilarity matrix D on each taxon in X along with a set of inferred level-1 networks $\mathbf{C} = \{C_i\}_{i=1}^k$ on disjoint subsets of taxa $\mathbf{S} = \{S_i\}_{i=1}^k$ and outputs a network $\widehat{\mathcal{N}}$ on X. We denote this algorithm by $\phi(D, \mathbf{C}) = \widehat{\mathcal{N}}$. The approach of InPhyNet stems from TreeMerge (Molloy and Warnow, 2019) which itself is derived from Neighbor-Joining (Saitou and Nei, 1987). The algorithm begins with an incomplete network $\widehat{\mathcal{N}}$ consisting of N leaves, one for each taxon in X, and zero edges. Then, iteratively, the algorithm builds $\widehat{\mathcal{N}}$ into a full rooted tree via a series of "merge" operations before using the reticulate information contained in the input networks to add reticulations to $\widehat{\mathcal{N}}$. Finally, the root of the network is suppressed and $\widehat{\mathcal{N}}$ is a semi-directed network. At each iteration, the algorithm can be summarized as follows:

1. Find the set of all pairs of nodes (u, v) that either (i) are neighbors in every network C_i where both $u, v \in C_i$ or (ii) $\nexists C_i | u, v \in C_i$. Call this set P.

   (Algorithm 2)

2. Compute a rate-adjusted matrix Q from D as in Saitou and Nei (1987) to find the most "similar" pair of nodes $(u_{\min, v_{\min})}$ in P. In the case of a tie, $u_{\min}$ and $v_{\min}$ are chosen randomly.

   (Part of Algorithm 1)

3. Merge $u_{\min}$ and $v_{\min}$ in $\widehat{\mathcal{N}}$ via a new shared parent w, expanding the inferred network.

   (Part of Algorithm 1)

4. In each input network C_i, shrink the subnetwork spanned by $u_{\min}$ and $v_{\min}$ to a single vertex w. Record reticulate relationships in this subnetwork so that they may be reflected in $\widehat{\mathcal{N}}$ later.

   (Algorithms 3 and 4)

5. Remove $u_{\min}$ and $v_{\min}$ from D and add an approximation for their new parent node w.

   (Part of Algorithm 1)

These steps occur $|X|-1$ times, after which the reticulate relationships recorded in step 4 are added to $\widehat{\mathcal{N}}$ and $\widehat{\mathcal{N}}$ is unrooted, yielding a semi-directed network. The root is removed at this step because it is in some sense artificial, without biological meaning. The set of input networks constrains the set of pairs of nodes that are allowed to be merged in step 1, so, hereafter, we refer to these networks as *constraint networks*.

In the case where $|\mathbf{C}| = 2$, this process will always successfully conclude so long as the constraints are constructed on disjoint subsets of taxa. However, when $|\mathbf{C}| > 2$, certain sequences of merges may lead to reduced constraint networks that conflict with one another, leading to $|P| = 0$ in step 1. In this case, the algorithm is adapted as in Molloy and Warnow (2019) as follows:

1. Select $C_1, C_2 \in \mathbf{C}$.
2. Denote $D^{(12)}$ as the portions of D that only pertain to leaves in C_1 and C_2.
3. Compute $\widehat{\mathcal{N}}_{12} = \phi\left(D^{(12)}, \{C_1, C_2\}\right)$.
4. Replace C_1 with $\widehat{\mathcal{N}}_{12}$ and remove C_2.
5. Repeat until only 1 network remains in $\mathbf{C}$. This network is $\widehat{\mathcal{N}}$.

This procedure is formalized in Algorithm 5. This ensures that a valid network will be constructed at the end of the algorithm. While all of the constraint networks must be level-1 and we present statistical consistency results later for level-1 species networks, this algorithm does not necessarily construct a level-1 network.

Runtime Complexity

Let N be the total number of taxa, r the total number of reticulations, and m the maximum constraint network size. The aforementioned inference framework has three main costs: (i) constraint inference, (ii) pairwise dissimilarity computations, and (iii) the InPhyNet algorithm itself. If the inference method used in (i) has runtime complexity $O(\mathcal{F}(N))$, then the runtime complexity to infer $\frac{N}{m}$ constraint networks, each of size m, is $O\left(\left\lceil \frac{N}{m} \right\rceil \mathcal{F}(m)\right)$. Additionally, (ii) and (iii) have $O(N^2)$ and $O(N^3 + Nr)$ runtime, respectively (see Appendix 3). Thus, the theoretical runtime complexity of such a pipeline is $O\left(\left\lceil \frac{N}{m} \right\rceil \mathcal{F}(m) + N^3\right)$. In practice, though, when an accurate inference method is used in (i), the runtime of (ii) and (iii) are negligible. This means that we can expect the runtime of such pipelines to scale *linearly* with respect to the number of taxa in the full network, a claim that we investigate later.

Recording Reticulate Relationships

For each constraint network C_i, if $u_{\min}$ and $v_{\min}$ are both contained in the network, the subnetwork that they span can only take the topologies shown in Figure 1. When this subnetwork is shrunk, all vertices shown (except for w) and all non-dashed edges are removed from the constraint network, except for subnetwork (1) which is simply reduced to a single vertex w.

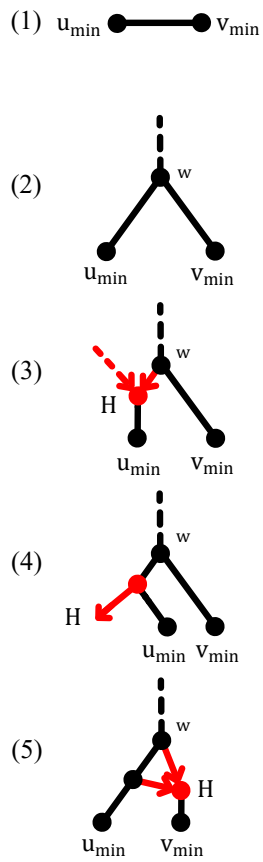

Fig. 1. All possible subnetworks in a given constraint network that can be spanned by nodes $u_{\min}$ and $v_{\min}$ selected in step (3) of the InPhyNet algorithm. H is an arbitrary hybrid vertex in the network and w is the node that all non-dashed elements of the subnetwork will be shrunk down to in step (4) of the InPhyNet algorithm. The dashed line in subnetwork (3) may be present in the constraint network, or may be missing as a result of a subnetwork of type (4) being shrunk previously.

At the same time, a new vertex w is added to the inferred network $\widehat{\mathcal{N}}$ along with edges $E_{wu_{\min}}$ and $E_{wv_{\min}}$ connecting w to the other respective vertices. In the cases of subnetworks (1) and (2), no additional work is done. For the remaining subnetworks, information corresponding to the reticulate structures in the subnetwork is logged with respect to the new edges in $\widehat{\mathcal{N}}$ so that this reticulate structure can be reconstructed in $\widehat{\mathcal{N}}$. For subnetwork (3) this means that edge $E_{wu_{\min}}$ is marked as "incoming" for hybrid vertex H, for subnetwork (4) $E_{wu_{\min}}$ is marked as "outgoing" for H, and for subnetwork (5) $E_{wu_{\min}}$ is marked outgoing for H and $E_{wv_{\min}}$ is marked incoming for H. In each case, the associated γ values in the constraint network are recorded as well.

For each hybrid vertex H found in a given constraint network, there will be one edge in $\widehat{\mathcal{N}}$ marked as "outgoing" and another marked as "incoming". In the post-processing phase of InPhyNet, each of these edges are split in two with a new vertex being added between them, say Z and H, respectively. The network $\widehat{\mathcal{N}}$ is still rooted at this point, so the edge parent to the hybrid vertex H is assigned the respective γ value as found in its associated constraint network. Then, a new directed edge is created from Z to H with an inheritance parameter of $1-\gamma$. In cases where a single edge in $\widehat{\mathcal{N}}$ is marked as incoming or outgoing for multiple hybrid nodes, the order in which these markings occurred with respect to the InPhyNet algorithm is stored. Then, the edge is split once for each such respective hybrid with earlier hybrids being assigned lower splits, preserving the ordering observed in the InPhyNet algorithm.

Statistical Consistency

THEOREM 1. *Given true species tree T, Neighbor Joining (NJ, Saitou and Nei (1987)) will construct T from a distance matrix D if $Q_1^{(T)}(D) < 0$ and $Q_2^{(T)}(D) < 0$ for each pair of true neighbors in T $(1,2)$ and arbitrary other taxa (i,j).*

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

$$Q_2^{(T)}(D) = D_{12} + D_{ij} - D_{1j} - D_{2i} \quad (2)$$

Proof. Presented in Appendix C of Saitou and Nei (1987). □

Note that Theorem 1 makes no assumption on the underlying data used to generate D, but instead asserts a property that D must uphold in order to construct T. We examine whether such a matrix D constructed under the average gene tree internode distance (AGID) metric upholds this property under the MSNC model later.

DEFINITION 7 Given input data $\mathcal{G}_n$ of size n (i.e. n gene trees, MSA with n base pairs, etc.) generated under true species network $\mathcal{N}$, a metric for computing the pairwise dissimilarity matrix M_D is said to be statistically consistent if the following is true for some displayed tree T of $\mathcal{N}$ where c_i is an arbitrary constant less than 0:

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

Again given $\mathcal{G}_n$ generated under true species network $\mathcal{N}$, a method for inferring species networks M_C is statistically consistent if:

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

where $\equiv$ indicates topological equivalence.

LEMMA 1 Given the true species network $\mathcal{N}$, distance matrix D, and constraint networks $\mathbf{C}$, if D satisfies Equations 1 and 2 for some $T \in \text{disp}(\mathcal{N})$ (the set of displayed trees of $\mathcal{N}$) and all constraint networks agree with $\mathcal{N}$, then $T \in \text{disp}\left(\phi(D, \mathbf{C})\right)$ where $\phi(D, \mathbf{C})$ is the network inferred by InPhyNet.

Proof. Notice that the final inferred network of InPhyNet is constructed by adding reticulate edges to a tree phylogeny, call that phylogeny $\tau(\widehat{\mathcal{N}})$. For each constraint network $C_i \in \mathbf{C}$, because each C_i agrees with $\mathcal{N}$, any reticulations that are suppressed in $\mathcal{N}$ to obtain T can similarly be suppressed in C_i, if they exist in C_i, to obtain a displayed tree of C_i that agrees with T. It follows that at each iteration of InPhyNet, the pair of nodes u^*, v^* for which the rate-adjusted matrix Q is minimized will belong to P because, by Definition 6, they will be considered neighbors in every C_i where they both appear. Thus, at each iteration, the global pair of nodes minimizing Q will be merged, which is exactly the NJ algorithm, so $\tau\left(\widehat{\mathcal{N}}\right) = T$. Whether $\tau\left(\widehat{\mathcal{N}}\right)$ is the major tree of $\widehat{\mathcal{N}}$ or simply some minor tree depends on the specific γ values inferred in the constraint networks $\mathbf{C}$. □

THEOREM 2. *Given the true species network $\mathcal{N}$, constraint networks $\mathbf{C}$ on non-overlapping sets of taxa $\mathbf{S}$, and distance matrix D, the network inferred by InPhyNet $\widehat{\mathcal{N}} = \phi(D, \mathbf{C})$ is exactly $\mathcal{N}$ if the following are true:*

1. *D satisfies Equations 1 and 2 for some $T \in$ disp$(\mathcal{N})$,*
2. *all C_i are level-1 and agree with $\mathcal{N}$ (see Definition 5), and*
3. *for each hybrid vertex in $\mathcal{N}$, at least one subset in $\mathbf{S}$ contains at least one individual from each of the clades A, B, C, D, and H as shown in Figure 3.*

Proof. From assumption (1) and (2), if InPhyNet's initial merging proposals are always accepted, then according to Lemma 1 $T \in \text{disp}\left(\widehat{\mathcal{N}}\right)$. Then, 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 hybrid vertex $v \in V_H$ with incoming hybrid edges e and e', we refer to the hybrid vertex's *reticulation* as whichever edge e or e' is *not* displayed in T. Then, 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 3. Notice that by assumptions (2) and (3) the cyclic structure shown in Figure 3 will also appear in at least one constraint network. For the remainder of the proof, we refer to the corresponding vertices in this cyclic structure as they appear in the constraint networks as $\tilde{A}, \tilde{B}, E_{\tilde{A,H}}$, and so on.

Notice that in order for a hybrid node to be created in $\widehat{\mathcal{N}}$, two nodes in some constraint network C_k must be neighbors connected by a path that either (a) contains a hybrid node, or (b) 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 traverse in this path 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 one of them will be suppressed in the displayed tree along with V_H. If clades B and C are merged after each of their members are merged, though, such a node can be created in the inferred network by splitting edge E_B (as in Figure 2). When $\tilde{B}$ and $\tilde{C}$ are merged, $E_{\tilde{B}}$ is traversed, whereafter it is created in $\widehat{\mathcal{N}}$ and marked as outgoing to H. Similarly, when $\tilde{A}$ and $\tilde{H}$ are merged, $E_{\tilde{H}}$ is traversed, whereafter it is created in $\widehat{\mathcal{N}}$ and marked as incoming to H. Then, in the post-processing step of InPhyNet, the exact aforementioned process needed to place the hybrid properly of splitting these edges is conducted.

Thus, it follows that V_B and thereby $E_{B,H}$ 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, edges $E_{\tilde{B}}$ and $E_{\tilde{B},\tilde{C}}$ occur in the subnetwork spanned by $\tilde{B}$ and $\tilde{C}$. Condition (i) follows from assumption (1). Given that we took $\tilde{A}$ and $\tilde{H}$ to have been merged first without loss of generality, any merging of $\tilde{B}$ and $\tilde{C}$ must go through edges $E_{\tilde{B}}$ and $E_{\tilde{B},\tilde{C}}$, 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 vertices in the constraint network is given by assumption (3), and their status as neighbors is given by assumption (2). It follows, then, that each hybrid vertex and its respective hybrid edges will be placed correctly in $\widehat{\mathcal{N}}$. □

COROLLARY 1 Given input data $\mathcal{G}_n$ of size n on taxa X generated under true species network $\mathcal{N}$ and disjoint subset decomposition $\mathbf{S} = \{S_i\}_{i=1}^k$, if M_C and M_D are both statistically consistent in the sense of Definition 7 and $\mathbf{S}$ satisfies condition (3) in Theorem 2, then InPhyNet is statistically consistent with respect to n for the true species network $\mathcal{N}$.

Proof. From consistency of M_D, we know that $\forall \epsilon_i > 0, \exists n_i$ such that the probability that $M_D(\mathcal{G}_{n_i})$ satisfies the conditions in Theorem 2 for $T \in \text{disp}(\mathcal{N})$ for each quartet is at least $1 - \sum_{i=1}^{\binom{|X|}{2}} \epsilon_i$. Similarly, by assumption of consistency of M_C, $\forall \epsilon_l, \exists n_l$ such that the probability

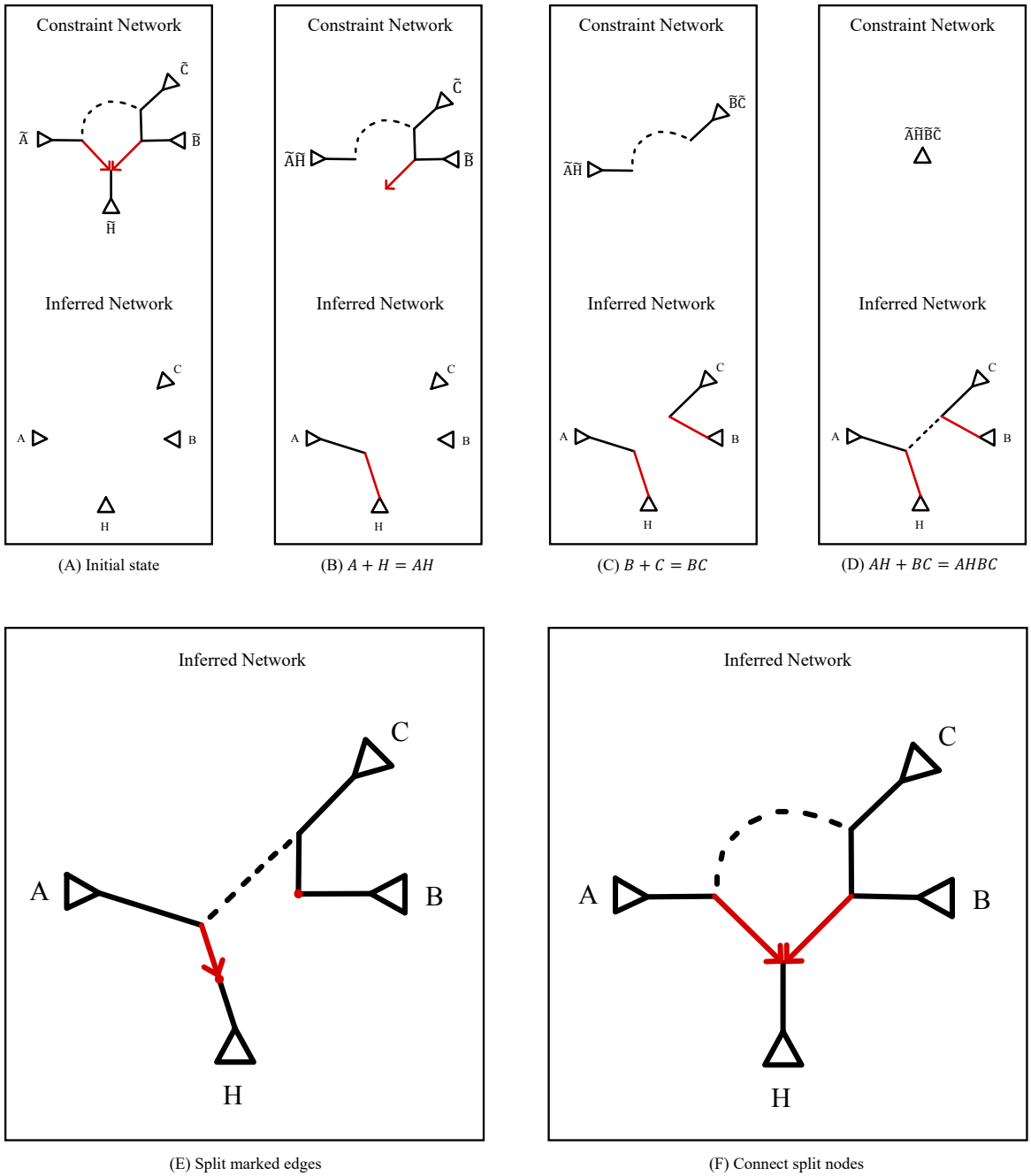

Fig. 2. **A:** the initial relevant state of the algorithm at which point all taxa within each clade have been merged. **B:** nodes A and H are merged. The path p^* connecting them in the constraint network traverses one reticulation, so its corresponding edge is marked in the inferred network. **C:** 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.

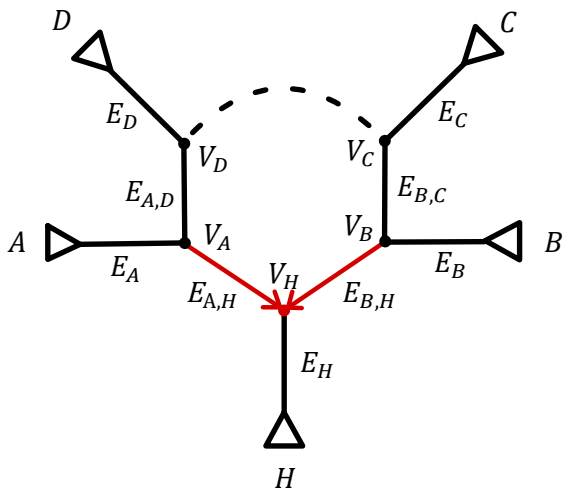

Fig. 3. Cycle corresponding to hybrid node V_H in the true species network with labellings corresponding to those used in Theorem 2. This structure is also displayed in any constraint that meets condition 3 of Theorem 2 but with labels corresponding to the relevant subsets of each clade $\tilde{A} \subseteq A, \tilde{B} \subseteq B$, and so on.

that the constraint network inferred by M_C on subset S_l 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:

$$\left(1 - \sum_i \epsilon_i\right) \prod_l (1 - \epsilon_l) \tag{3}$$

$$\geqslant \left(1 - 2\binom{|X|}{4} \max_i \{\epsilon_i\}\right) \left(1 - \max_l \{\epsilon_l\}\right)^k \tag{4}$$

$$\geqslant \left(1 - 2\binom{|X|}{4} \max_i \{\epsilon_i\}\right) \left(1 - k \max_l \{\epsilon_l\}\right) \tag{5}$$

$$\geqslant 1 - 2\binom{|X|}{4} \max_i \{\epsilon_i\} - k \max_l \{\epsilon_l\} \tag{6}$$

where Equation 3 assumes independence between M_D and M_C and Equation 5 is given by Bernoulli's Inequality. Then, $\forall \epsilon > 0$, choose $\epsilon_i, \epsilon_l > 0$ such that $2\binom{N}{4} \max_i\{\epsilon_i\} + k \max_l\{\epsilon_l\} < \epsilon$ and choose $n = \max(\max_i\{n_i\}, \max_l\{n_l\})$. Then, Equation 6 is at least $1 - \epsilon$ and therefore, by Theorem 2, the probability that InPhyNet constructs $\mathcal{N}$ is at least $1 - \epsilon$, so InPhyNet is statistically consistent for $\mathcal{N}$ with respect to n. □

Performance Study

Overview

We quantify the estimation error of the inference framework presented above with various amounts of input data (number of gene trees or sequence length), numbers of taxa, levels of incomplete lineage sorting (ILS), and sizes of constraint networks (m). See Table 1 for all parameter values. We evaluated the topological accuracy of the framework by calculating the unrooted hardwired cluster distance (HWCD, (Huson et al., 2010)) between the inferred species network and the true species network, namely HWCD($\mathcal{N}, \widehat{\mathcal{N}}$). These calculations were done using the `hardwiredclusterdistance` function in the PhyloNetworks Julia package (Solís-Lemus et al., 2017). HWCD is a metric on *rooted* networks, but can be calculated on unrooted networks by finding the rootings of each network that minimize the rooted HWCD. Additionally, we evaluated the runtime of the framework across each simulation setting. The original data used in these simulations can be found on Dryad, and the scripts used to analyze this data can be found at github.com/NathanKolbow/InPhyNetSimulations

Parameter	Values
Number of taxa	25, 50, 100, 200
Number of base pairs	100, 1000
Number of loci	100, 1000
Maximum constraint size (m)	10, 20
Incomplete lineage sorting (ILS)	low, high

Table 1. Simulation parameters and values used in the simulation study. Low and high ILS correspond to average network branch lengths of 1.5 and 0.15, respectively. Ten replicates were conducted for each parameter combination.

Simulated Data

Specific settings and parameters used for each of the software mentioned below can be found in Appendix B. The ground truth species networks were simulated according to a birth-death-hybridization model with the SiPhyNetwork package in R (Justison et al., 2023). This package simulates networks of arbitrary level, so in each case networks were simulated with fixed parameters until a level-1 network was obtained. Then, gene trees were simulated according to the multispecies network coalescent model using the PhyloCoalSimulations package in Julia (Fogg et al., 2023). Finally, multiple sequence alignments (MSAs) were simulated with seq-gen (Rambaut and Grass, 1997) under the F81 model of evolution (Felsenstein, 1981) with state frequencies $\pi_A = \pi_T = 0.3$ and $\pi_C = \pi_G = 0.2$.

Framework Methods

We used SNaQ (Solís-Lemus and Ané, 2016), PhyloNet's maximum likelihood option (PhyloNet-ML) (Yu et al., 2014), PhyloNet's maximum pseudolikelihood option (PhyloNet-MPL) (Yu and Nakhleh, 2015),

and Squirrel (Holtgrefe et al., 2025) to infer the constraint networks. It should be noted, though, that Squirrel is not a coalescent-aware method, and so these simulations are explicit model violations. Analyses with PhyloNet-ML were restricted to only using $m = 10$ due to prohibitively long runtimes. In addition, eight analyses utilizing PhyloNet-ML with $m = 10$ were excluded due to a failure to complete within 1 month.

The average gene tree internode distance (AGID, see Definition 9 for a formal definition) metric was used to compute the pairwise dissimilarity matrix D in each simulation, and the methodology presented in Materials and Methods was used for subset decomposition. We established in the section Statistical Consistency that this framework can be statistically consistent when M_C and M_D are statistically consistent and subset decomposition is done in an appropriate manner, but we do not claim that this specific implementation of the framework meets these criterion. In particular, AGID is *not* statistically consistent in general in the sense of Definition 7 (see Appendix 1).

Subset Decomposition The primary goal of a subset decomposition method in this framework is to (a) identify areas where hybrids may be present, and (b) choose subsets such that the conditions in Theorem 2 are met. In doing so, we infer the *tree of blobs* of the species network

1. Infer the tree of blobs $\mathcal{T}$ of the true species network with TINNiK (Allman et al., 2024).
2. For each blob B_i in $\mathcal{T}$, the blob is connected to edges $E_1, E_2, \ldots, E_k$. For each edge E_j, imagine splitting $\mathcal{T}$ in half by E_j, and randomly selecting a single leaf node that is *not* in the same split half as the blob center. Add every such leaf to a set $\mathcal{S}_i$.
3. Infer a species tree $\widehat{T}$ on the full set of taxa with ASTRAL (Zhang et al., 2018).
4. For each set $\mathcal{S}_i$, remove all taxa in $\mathcal{S}_i$ from $\widehat{T}$ except for the first entry of $\mathcal{S}_i$, denote this taxa as $\mathcal{S}_i^{(1)}$.
5. Select the edge in $\widehat{T}$ that splits $\widehat{T}$ into equally sized subtrees (or as close to equally sized as possible). Repeat this recursively on the resulting subtrees until each subtree contains no more than m leaves.
6. Find each taxa $\mathcal{S}_i^{(1)}$ and its corresponding subtree and expand $\mathcal{S}_i^{(1)}$ into all of $\mathcal{S}_i$.

The taxa contained in each subtree now correspond to subsets in the decomposition.

We found that when a tree of blobs is inferred with more than 50 taxa, step (1) is computationally infeasible. In this case, we identify blobs in batches by first inferring a species tree with ASTRAL and then decomposing the set of taxa into subsets of size 50 in both manners depicted in Figure 4. Then, steps (1) and (2) are conducted on every such subset of size 50 constructed in this manner, generating a single set $\mathcal{S}$ which is then used for each of steps (4)-(6). This approach allows us to identify hybridization events in the true species network whether they occur between distant or local lineages.

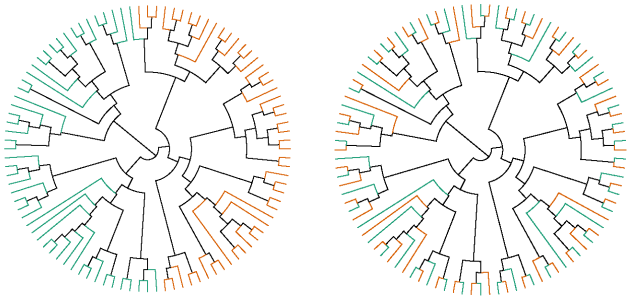

Fig. 4. An illustration of a species tree $\widehat{T}$ inferred on 100 taxa being split into subsets of 50 individuals (one subset in green, the other in orange) in two different ways, as described in Subset Decomposition. **Left:** two subsets of 50 taxa split such that the majority of any given clade tends to appear in the same subset. **Right:** two subsets of 50 taxa split such that distant taxa tend to appear in the same subset.

Inference with SNaQ Analyses with the SNaQ method were conducted with the SNaQ package in Julia (Kolbow et al., 2025). SNaQ takes estimated quartet concordance factors as input, which require estimated gene trees to be computed. So, we inferred gene trees using IQ-TREE and used ModelFinder for model selection. Further, each instance of SNaQ requires a maximum number of hybrids (`hmax`) to be specified before a network can be inferred. In practice, for a fixed set of taxa, a set of networks with ascending `hmax` values are inferred before model selection is conducted. This significantly increases the computational burden of each simulation, so instead, for a given subset of taxa S_i from true species network $\mathcal{N}$, we selected `hmax` to be the number of hybrids present in the subnetwork of $\mathcal{N}$ spanned by S_i.

Inference with Squirrel Analyses with the Squirrel method were conducted with the `physquirrel` package in Python. Dense sets of quarnets were inferred from the respective MSA with the `delta_heuristic` method, and we inferred level-1 networks from these dense sets of quarnets with the `squirrel` method.

Results

Accuracy Total output error tended to decrease as the number of gene trees and number of

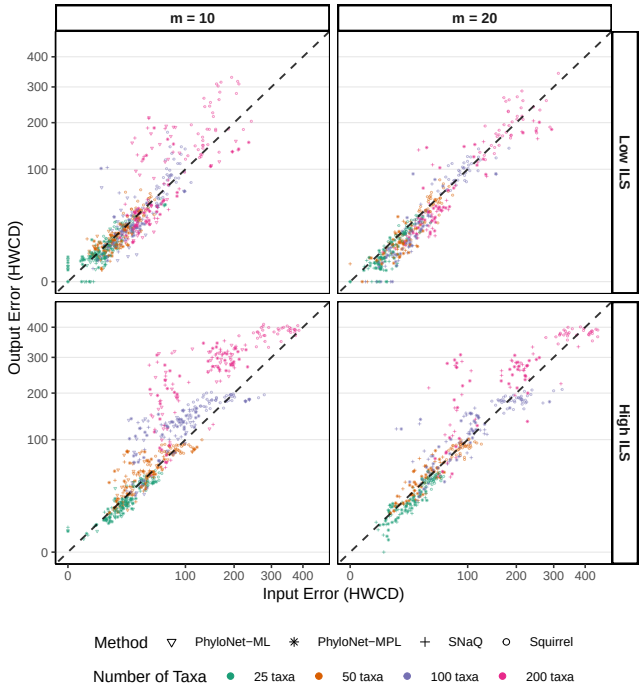

Fig. 5. Output error of the species network inferred by InPhyNet in unrooted hardwired cluster distance (HWCD) vs. the total input error to InPhyNet. Input error is calculated by summing the inference errors of each constraint network used as input to InPhyNet. The inference error of each constraint is calculated by taking the HWCD between the constraint network and the subnetwork of the true species network that is spanned by the taxa contained in the constraint network.

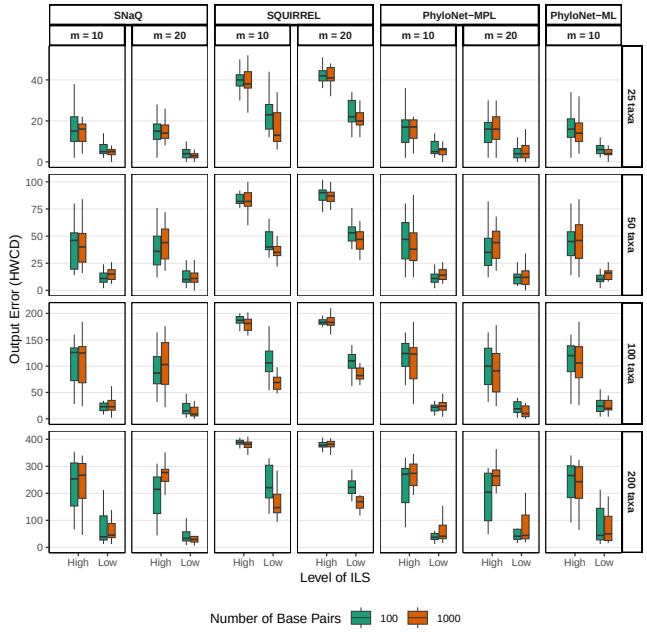

Fig. 6. Output error of the pipeline across various simulation settings in hardwired cluster distance (HWCD). m denotes the maximum number of tips included in each constraint network. The labels running along the rightmost subplots indicate the number of tips in the true species network.

base pairs increased in the simulations (Figure 6). We observed comparable accuracy when utilizing SNaQ, PhyloNet-ML, and PhyloNet-MPL in tandem with InPhyNet, but observed significantly worse performance when utilizing Squirrel. In cases with low ILS, median error across all simulation settings was near zero with all methods except for Squirrel. Simulations with high ILS did not perform as well, however. We found that the maximum constraint network size (m) did not have a noticeable impact on accuracy in any simulations with any underlying inference method.

We observed a very strong dependence between the total error in the input networks given to InPhyNet and the output error of the network output by InPhyNet (Figure 5). This indicates that InPhyNet is highly robust to error in the distance matrix, but highly sensitive to error in the input networks.

Runtime For a fixed maximum constraint network size m, total runtime increased approximately linearly with respect to the number of taxa in the full network N (Figure 7). Runtimes were significantly longer for simulations with $m = 20$ than their counterparts with $m = 10$. Specifically, simulations utilizing Squirrel, SNaQ, and PhyloNet-MPL with $m = 20$ took 2.6, 10.9, and 8.9 times longer than their counterparts with $m = 10$. Inference was only conducted with PhyloNet-ML with $m = 10$. All simulations utilizing Squirrel took at most 10 minutes to complete, whereas simulations with SNaQ and PhyloNet took on the order of tens or hundreds of hours to complete, depending on m.

Phylogeny of Green Plants

We re-analyze the phylogeny of 1,158 green plants presented in One Thousand Plant Transcriptomes Initiative (2019). Phylogenetic tree analyses led to observed quartet concordance frequencies that conflicted with what would be expected under the inferred model. Specifically, there was significant conflict between concordance factors involving four major gymnosperm clades and between concordance factors involving four major fern clades. These areas of the phylogeny have previously been shown to exhibit evidence of reticulate evolution, so we utilized InPhyNet to incorporate such reticulations into the phylogeny of green plants. Portions of the tree phylogeny outside of the gymnosperms and ferns were well resolved, so we utilized the subnetwork of the phylogeny inferred in One Thousand Plant Transcriptomes Initiative (2019)

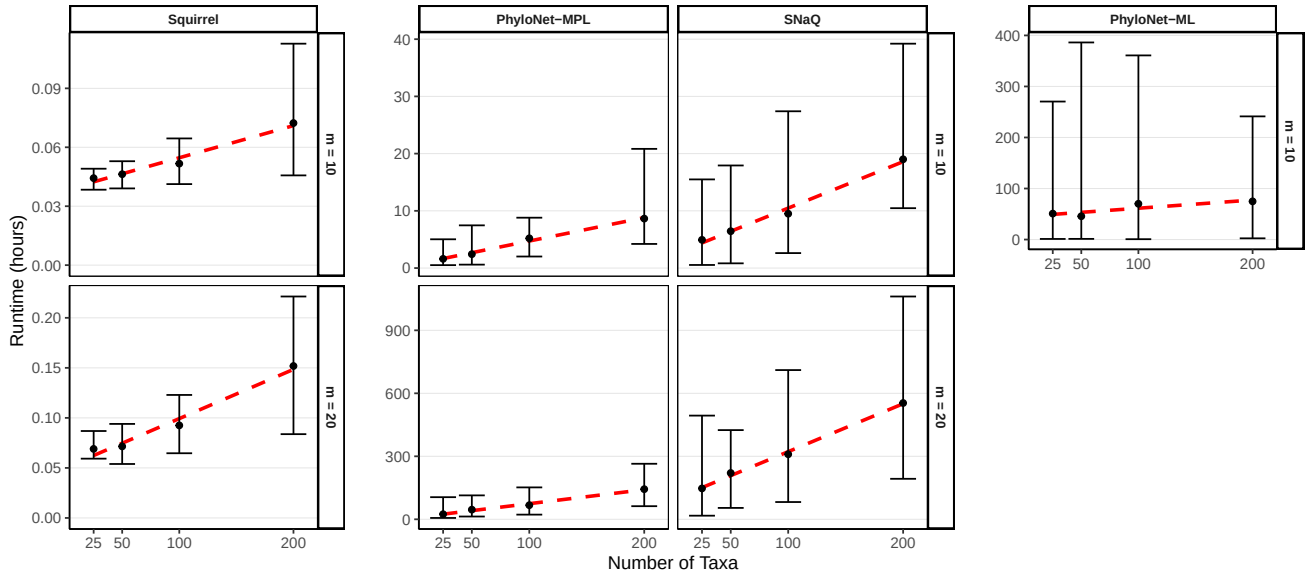

Fig. 7. 95% confidence intervals for total runtime (in hours) to infer constraint networks, estimate the pairwise dissimilarity matrix D, and infer the final network with InPhyNet across various inference methods in settings with low levels of incomplete lineage sorting (ILS). Each column of subplots corresponds to a different inference methods whereas each row corresponds to different maximum constraint network sizes (m).

spanned by all non-gymnosperm and non-fern species as a single constraint network in InPhyNet. Constraint networks for the remaining species were inferred using the aforementioned inference pipeline with a maximum constraint size of $m = 20$. The subset decomposition step was done manually to ensure that all apparent instances of reticulate evolution would be identifiable by SNaQ. For further details, see Appendix 5.

Results

The network of green plants that we inferred is presented in Figure 8. While this species network generally agrees with previously inferred phylogenies (One Thousand Plant Transcriptomes Initiative, 2019; Bowe et al., 2000; Liu et al., 2022; Wickett et al., 2014), it also provides new insights into regions of persistent discordance. For example, the placement of asterids, rosids, Caryophyllales, and Saxifragales remains unresolved, which is a particularly consequential issue because together these clades represent approximately 75% of all angiosperm species (Wang et al., 2009). Other recent large-scale phylogenomic studies (Zuntini et al., 2024; Kates et al., 2024) have also highlighted ongoing uncertainty in these relationships, underscoring that this lack of resolution is not unique to our re-analysis. In addition, the network recovered reticulations that help reconcile competing hypotheses regarding the placement of Order Gnetales within gymnosperms, and highlights evidence of reticulate evolution within Pinaceae and Polypodiidae. In the following subsections, we examine each of these cases in detail, drawing on prior research to substantiate the inferred reticulations.

Gnetales The placement of Gnetales in the phylogeny of gymnosperms is controversial. Some research suggests that they are sister to Pinaceae (the "Gnepine" hypothesis), others suggest that they are sister to conifers (the "Gnetifer" hypothesis) and others still suggest that they are sister to cupressophytes (the "Gne-cup" hypothesis) or to all conifers together (Bowe et al., 2000; Soltis et al., 2002; Ran et al., 2018; Chaw et al., 2000; Zhong et al., 2011; Braukmann et al., 2009; Magallón and Sanderson, 2002; Coiro et al., 2022). Treelike analyses of these species have led to significant conflict between these hypotheses, but our analyses reconcile them by providing partial support for both the Gnetifer and Gnepine hypotheses. However, because the inferred phylogeny of Gnetales is known to be highly sensitive to the type of data analyzed and the handling of systematic biases (Zhong et al., 2011), this result should be regarded with caution and warrants further investigation.

Pinaceae High rates of polyploidy have been identified in genus *Juniperus* Farhat et al. (2019), and similar evidence has been found in *Cupressales* and *Platycladus* Farhat et al. (2022). We are unaware of such direct evidence for *Calocedrus decurrens*, *Microbiota decussata*, or *Tetraclinis spp.*, but research suggests that

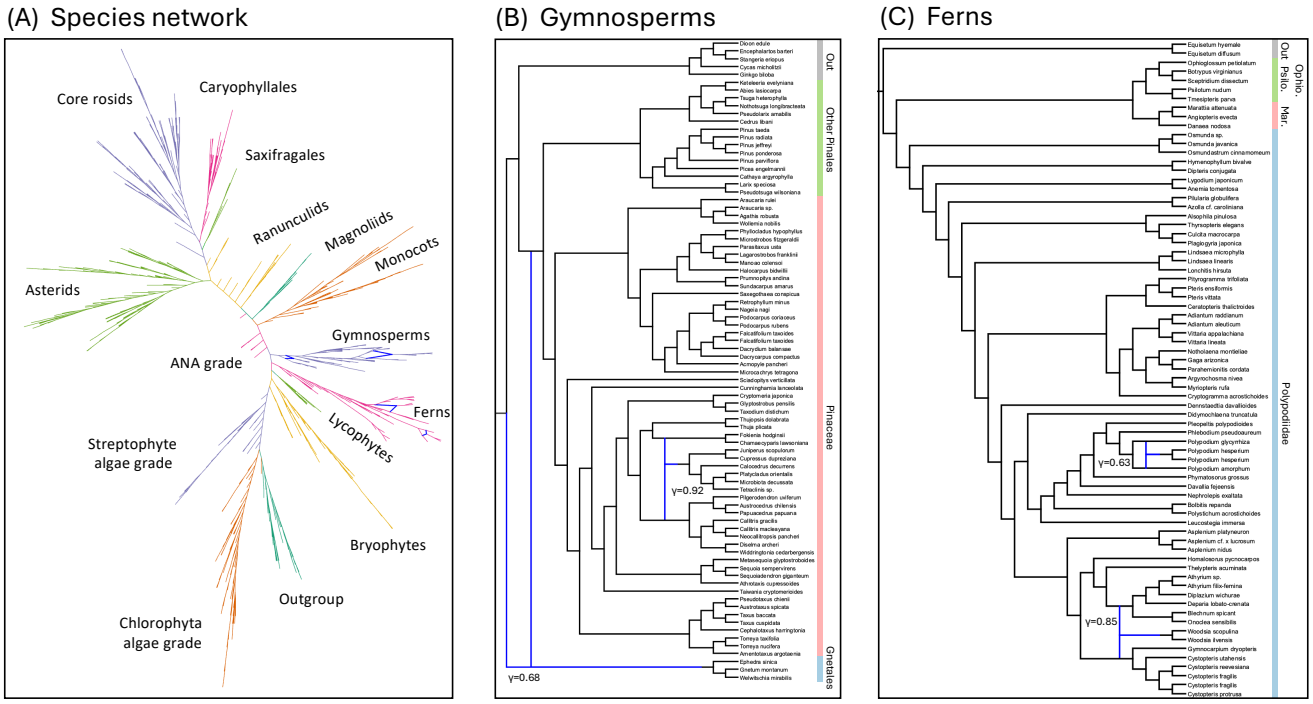

Fig. 8. **A:** Semi-directed phylogenetic network of green plants obtained using InPhyNet. Various colored branches signify different clades within the phylogeny. Branch lengths are not scaled. **B:** Magnified portion of the phylogeny corresponding to gymnosperms. The reticulation ancestral to Gnetales 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 Farhat et al. (2019, 2022); Zonneveld (2012). 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 Sigel et al. (2014); Smith et al. (2006); Shao et al. (2015).

these three species are closely related to *Platycladus orientalis* Zonneveld (2012), corroborating their placement in this phylogeny.

While SNaQ does not directly model polyploidization, PhyloNet (Wen et al., 2018), a similar likelihood-based approach, has been shown to accurately infer polyploidy events in genus *Fragaria* despite not explicitly modelling polyploidization (Yan et al., 2022). Accordingly, these reticulations in the inferred network may reflect hybridization events, including allopolyploidy, but they could also arise from other sources of conflict, and their interpretation remains uncertain.

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* Sigel et al. (2014). 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 Smith et al. (2006), or that have otherwise found generally weak support for treelike backbones between these species Shao et al. (2015).

Discussion

Performance Study

Our performance study demonstrates that the accuracy of InPhyNet is strongly associated with the accuracy of the input networks (see Figure 5). The magnitude of this relationship, however, depends markedly on the number of taxa in the network and on the value of m. In particular, simulations with fewer taxa more frequently produced output errors that were less than or equal to the corresponding input errors. For datasets of 25, 50, 100, and 200 taxa, this outcome was observed in 82.3%, 61.4%, 52.4%, and 39.2% of simulations, respectively. This trend also seems to depend on the chosen value of m and the level of ILS present in the true network. This outcome was observed in simulations with $m = 10$ and $m = 20$ 51.6% and 68.3% of the time, respectively, and 69.0% and 48.4% in simulations

with low and high ILS, respectively. This trend demonstrates that InPhyNet is sensitive to error in the input networks, with the dependence on both the number of taxa and the value of m further strengthening this conclusion: larger values of m and smaller numbers of taxa enable InPhyNet to rely more heavily on the input networks and less on the pairwise distance matrices.

This makes the decision of m difficult in practice, though, as higher values of m drastically increase runtime, but also tend to improve inference accuracy. The most important factor in determining output error, though, is input error, which does not depend on m. Thus, when utilizing hill-climbing algorithms that leverage the results of multiple runs such as SNaQ, PhyloNet-ML or PhyloNet-MPL, we recommend choosing smaller values of m and instead investing more computational resources into increasing the number of independent runs performed by the optimization software. We expect that this will provide the best tradeoff between method accuracy and total runtime.

Surprisingly, little to no difference in output accuracy was observed between PhyloNet-ML, PhyloNet-MPL, and SNaQ. Differences in runtimes between the three methods were drastic, though, indicating that the fastest method PhyloNet-MPL should be preferred. It is worth noting that these simulations were performed with the version 1.0 of the Julia package `SNaQ.jl`, whose runtime has since been improved dramatically in version 1.1, so this result may no longer hold (Kolbow et al., 2025).

The results in Figure 6 show a consistent positive relationship between the amount of input data used (number of gene trees and number of base pairs in this case) and output accuracy. This trend reflects the statistical consistency results outlined above and highlights the robustness InPhyNet as well as each of the underlying inference methods.

Our simulations utilized the AGID metric to compute a pairwise dissimilarity matrix. We show in Appendix 1 that, for generic level-1 networks, as the number of input gene trees goes to ∞ this metric does not converge on a dissimilarity matrix D that constructs a displayed tree of a given network as in Theorem 1. It is unclear whether other metrics may be statistically consistent in this sense or whether they would generally yield more favorable results. Future work could investigate the role that this metric plays in the overall accuracy of InPhyNet.

Additionally, one limitation of InPhyNet is that reticulations cannot be inferred between the disjoint subsets of taxa that are selected at the beginning of the inference framework. This can be accounted for by utilizing tree of blob analyses with methods like TINNiK Allman et al. (2024) or TREE-QMC Han and Molloy (2025) during the subset decomposition phase. Future work could alleviate this issue by relaxing the requirement for strictly disjoint subsets of taxa, or by testing for the presence of hybridizations as the network is being built.

Phylogeny of Green Plants

Green plants are a large, complicated group comprised of many important clades, and their analysis warrants extreme care. Our goal in this empirical analysis is not to provide novel empirical evidence on the phylogeny of green plants, but instead to illustrate how InPhyNet can be integrated into an empirical workflow to synthesize information in areas of the phylogeny that exhibit discordance.

We re-analyzed the 1,158-taxon green plant phylogeny from the One Thousand Plant Transcriptomes Initiative (1KP), focusing on regions with strong quartet concordance factor discordance that are poorly modeled by treelike phylogenies. We used InPhyNet to incorporate reticulate hypotheses where quartet-based signals conflicted, particularly among gymnosperms and ferns, into the broader phylogeny of green plants. Outside these groups, there was relatively little observed discordance, and so we treated the subtree of the 1KP inferred species tree containing these taxa as a single constraint network. We performed subset decomposition on the gymnosperms and ferns manually, utilizing past research identifying areas of discordance and the trees of blobs inferred by TINNiK (Allman et al., 2024) as guides. The resulting semi-directed network generally agrees with prior large-scale phylogenies, and provides a single, coherent hypothesis for the full reticulate phylogeny of green plants under the multispecies network coalescent model.

Within gymnosperms, the recovered network yielded a reticulate event ancestral to Gnetales that partially supports both the Gnetifer (Gnetales sister to conifers) and Gnepine (Gnetales sister to Pinaceae) hypotheses. This placement requires further analyses to validate, especially due to the systematic error in similar analyses that has been by Zhong et al. (2011). The reticulate event depicted in Pinaceae comprises taxa characterized by high levels of allopolyploidy. It is unclear whether SNaQ would infer a hybridization event in the presence of allopolyploidy, but PhyloNet's maximum parsimony method is capable of inferring hybridizations in such cases (Yan et al., 2022). Whether this reticulation resolves the discordance induced by allopolyploidy or discordance induced by another source entirely requires further examination.

Within ferns, reticulate events were only identified with Polypodiidae, each of which are well-founded by prior research. The discordance among quartet concordance factors observed among ferns by One Thousand Plant Transcriptomes Initiative (2019) was primarily between the three clades depicted in Figure 8 rather than within each of the clades, so more research is warranted to further resolve this portion of the phylogeny.

While the results presented here require further research to corroborate, this analysis, particularly with respect to gymnosperms, exhibits the distinct advantage of being able to interpret reticulate events that span large portions of the phylogeny with the full context of the entire phylogeny. By preserving a well-supported

backbone and targeting regions of known discordance, InPhyNet provides a coherent picture of the evolutionary history of all green plants that explicitly accounts for instances of reticulate evolution.

Conclusions

We presented a novel method, InPhyNet, for merging constraint networks inferred on disjoint subsets of taxa. We established that InPhyNet yields correct results when the input data are correct and the disjoint subsets are chosen appropriately (see Theorem 2). Furthermore, we demonstrated that the InPhyNet algorithm is statistically consistent with respect to the amount of input data provided given that its input data is inferred in a statistically consistent manner.

Through an extensive simulation study, we verified the practicality of these theoretical findings. Key results from our simulations indicate that the accuracy of InPhyNet is highly correlated with the accuracy of its input data, and that accuracy consistently improves as the amount of input data increases. The maximum constraint network size parameter m had no noticeable effect on accuracy but does have a direcct negative impact on runtime. The most accurate results were observed with PhyloNet-ML, though SNaQ and PhyloNet-MPL yielded similarly accurate results across the simulation study.

For a fixed value of m, total runtime scaled linearly with respect to the number of tips in the true species network. These results demonstrate that a pipeline utilizing accurate network inference methods, AGID metric, and InPhyNet can be used to accurately infer networks with hundreds of tips in a very feasible time frame.

Finally, we applied InPhyNet to expand the phylogeny of green plants, successfully identifying instances of reticulate evolution in regions previously poorly supported by strictly treelike phylogenies. Collectively, these results underscore the utility of InPhyNet for robust and efficient phylogenetic network inference in complex evolutionary scenarios.

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, Erin Molloy and Yunheng Han for their help in improving the simulation study, and two anonymous reviewers for their highly valuable feedback.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.h44j0zq0b.

References

Allman, E. S., H. Baños, J. D. Mitchell, and J. A. Rhodes. 2024. TINNiK: Inference of the tree of blobs of a species network under the coalescent model 19:23.

Ané, C., J. Fogg, E. S. Allman, H. Baños, and J. A. Rhodes. 2024. Anomalous networks under the multispecies coalescent: Theory and prevalence 88:29.

Boto, L. 2009. Horizontal gene transfer in evolution: Facts and challenges. Proceedings of the Royal Society B: Biological Sciences 277:819–827.

Bowe, L. M., G. Coat, and C. W. dePamphilis. 2000. 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 97:4092–4097.

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

Bryant, D. and D. H. Huson. 2023. NeighborNet: Improved algorithms and implementation 3.

Chaw, S.-M., C. L. Parkinson, Y. Cheng, T. M. Vincent, and J. D. Palmer. 2000. 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 97:4086–4091.

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

Coiro, M., E. A. Roberts, C.-C. Hofmann, and L. J. Seyfullah. 2022. Cutting the long branches: Consilience as a path to unearth the evolutionary history of Gnetales 10.

Day, W. H. E. 1985. Optimal algorithms for comparing trees with labeled leaves 2:7–28.

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

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

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

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

Forster, P., L. Forster, C. Renfrew, and M. Forster. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences of the United States of America 117:9241–9243.

Han, Y. and E. K. Molloy. 2025. Improved Robustness to Gene Tree Incompleteness, Estimation Errors, and Systematic Homology Errors with Weighted TREE-QMC 74:758–777.

Hasegawa, M., H. Kishino, and T.-a. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Hejase, H. A., N. VandePol, G. M. Bonito, and K. J. Liu. 2018. FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data. Pages 242–259 *in* Comparative Genomics (M. Blanchette and A. Ouangraoua, eds.) Springer International Publishing, Cham.

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

Holtgrefe, N., K. T. Huber, p. u. family=Iersel, given=Leo, M. Jones, S. Martin, and V. Moulton. 2025. Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments Page msaf067.

Huson, D. H., R. Rupp, and C. Scornavacca. 2010. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press.

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

Kates, H. R., B. C. O'Meara, R. LaFrance, G. W. Stull, E. K. James, S.-Y. Liu, Q. Tian, T.-S. Yi, D. Conde, M. Kirst, J.-M. Ané, D. E. Soltis, R. P. Guralnick, P. S. Soltis, and R. A. Folk. 2024. Shifts in evolutionary lability underlie independent gains and losses of root-nodule symbiosis in a single clade of plants 15:4262.

Kolbow, N., S. Kong, T. Chafin, J. Justison, C. Ané, and C. Solís-Lemus. 2025. Snaq.jl: Improved scalability for phylogenetic network inference unpublished. Available upon request.

Kong, S., J. C. Pons, L. Kubatko, and K. Wicke. 2022. Classes of explicit phylogenetic networks and their biological and mathematical significance. Journal of Mathematical Biology 84:47.

Kong, S., D. L. Swofford, and L. S. Kubatko. 2025. Inference of Phylogenetic Networks From Sequence Data Using Composite Likelihood 74:53–69.

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

Liu, G.-Q., L. Lian, and W. Wang. 2022. The Molecular Phylogeny of Land Plants: Progress and Future Prospects. Diversity 14:782.

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

Molloy, E. K. and T. Warnow. 2019. TreeMerge: A new method for improving the scalability of species tree estimation methods. Bioinformatics 35:i417–i426.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685.

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

Ran, J.-H., T.-T. Shen, M.-M. Wang, and X.-Q. Wang. 2018. 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 285:20181012.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425.

Sang, T. and S. Ge. 2007. Genetics and phylogenetics of rice domestication. Current Opinion in Genetics & Development 17:533–538.

Schwartz, R. and A. A. Schäffer. 2017. The evolution of tumour phylogenetics: Principles and practice. Nature reviews. Genetics 18:213–229.

Searls, D. B. 2003. Pharmacophylogenomics: Genes, evolution and drug targets. Nature Reviews Drug Discovery 2:613–623.

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

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

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

Solís-Lemus, C. and C. Ané. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLOS Genetics 12:e1005896.

Solís-Lemus, C., P. Bastide, and C. Ané. 2017. PhyloNetworks: A Package for Phylogenetic Networks. Molecular Biology and Evolution 34:3292–3298.

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

Sánchez-Pacheco, S. J., S. Kong, P. Pulido-Santacruz, R. W. Murphy, and L. Kubatko. 2020. 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 117:12518–12519.

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

Wang, H., M. J. Moore, P. S. Soltis, C. D. Bell, S. F. Brockington, R. Alexandre, C. C. Davis, M. Latvis, S. R. Manchester, and D. E. Soltis. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests 106:3853–3858.

Wen, D., Y. Yu, J. Zhu, and L. Nakhleh. 2018. Inferring Phylogenetic Networks Using PhyloNet 67:735–740.

Wickett, N. J., 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. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences 111:E4859–E4868.

Yan, Z., Z. Cao, Y. Liu, H. A. Ogilvie, and L. Nakhleh. 2022. Maximum Parsimony Inference of Phylogenetic Networks in the Presence of Polyploid Complexes 71:706–720.

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

Yu, Y., J. Dong, K. J. Liu, and L. Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories 111:16448–16453.

Yu, Y. and L. Nakhleh. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16:S10.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.

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

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

Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epitawalage, E. Françoso, B. Gallego-Paramo, C. McGinnie, R. Negrão, S. R. Roy, L. Simpson, E. Toledo Romero, V. M. A. Barber, L. Botigué, J. J. Clarkson, R. S. Cowan, S. Dodsworth, M. G. Johnson, J. T. Kim, L. Pokorny, N. J. Wickett, G. M. Antar, L. DeBolt, K. Gutierrez, K. P. Hendriks, A. Hoewener, A.-Q. Hu, E. M. Joyce, I. A. B. S. Kikuchi, I. Larridon, D. A. Larson, p. u. family=Lírio, given=Elton John, J.-X. Liu, P. Malakasi, N. A. S. Przelomska, T. Shah, J. Viruel, T. R. Allnutt, G. K. Ameka, R. L. Andrew, M. S. Appelhans, M. Arista, M. J. Ariza, J. Arroyo, W. Arthan, J. B. Bachelier, C. D. Bailey, H. F. Barnes, M. D. Barrett, R. L. Barrett, R. J. Bayer, M. J. Bayly, E. Biffin, N. Biggs, J. L. Birch, D. Bogarín, R. Borosova, A. M. C. Bowles, P. C. Boyce, G. L. C. Bramley, M. Briggs, L. Broadhurst, G. K. Brown, J. J. Bruhl, A. Bruneau, S. Buerki, E. Burns, M. Byrne, S. Cable, A. Calladine, M. W. Callmander, Cano, D. J. Cantrill, W. M. Cardinal-McTeague, M. M. Carlsen, A. J. A. Carruthers, p. u. family=Castro Mateo, given=Alejandra, M. W. Chase, L. W. Chatrou, M. Cheek, S. Chen, M. J. M. Christenhusz, P.-A. Christin, M. A. Clements, S. C. Coffey, J. G. Conran, X. Cornejo, T. L. P. Couvreur, I. D. Cowie, L. Csiba,

I. Darbyshire, G. Davidse, N. M. J. Davies, A. P. Davis, p. u. family=Dijk, given=Kor-jent, S. R. Downie, M. F. Duretto, M. R. Duvall, S. L. Edwards, U. Eggli, R. H. J. Erkens, M. Escudero, p. l. u. family=Estrella, given=Manuel, F. Fabriani, M. F. Fay, P. d. L. Ferreira, S. Z. Ficinski, R. M. Fowler, S. Frisby, L. Fu, T. Fulcher, M. Galbany-Casals, E. M. Gardner, D. A. German, A. Giaretta, M. Gibernau, L. J. Gillespie, C. C. González, D. J. Goyder, S. W. Graham, A. Grall, L. Green, B. F. Gunn, D. G. Gutiérrez, J. Hackel, T. Haevermans, A. Haigh, J. C. Hall, T. Hall, M. J. Harrison, S. A. Hatt, O. Hidalgo, T. R. Hodkinson, G. D. Holmes, H. C. F. Hopkins, C. J. Jackson, S. A. James, R. W. Jobson, G. Kadereit, I. M. Kahandawala, K. Kainulainen, M. Kato, E. A. Kellogg, G. J. King, B. Klejevskaja, B. B. Klitgaard, R. R. Klopper, S. Knapp, M. A. Koch, J. H. Leebens-Mack, F. Lens, C. J. Leon, Léveillé-Bourret, G. P. Lewis, D.-Z. Li, L. Li, S. Liede-Schumann, T. Livshultz, D. Lorence, M. Lu, P. Lu-Irving, J. Luber, E. J. Lucas, M. Luján, M. Lum, T. D. Macfarlane, C. Magdalena, V. F. Mansano, L. E. Masters, S. J. Mayo, K. McColl, A. J. McDonnell, A. E. McDougall, T. G. B. McLay, H. McPherson, R. I. Meneses, V. S. F. T. Merckx, F. A. Michelangeli, J. D. Mitchell, A. K. Monro, M. J. Moore, T. L. Mueller, K. Mummenhoff, J. Munzinger, P. Muriel, D. J. Murphy, K. Nargar, L. Nauheimer, F. J. Nge, R. Nyffeler, A. Orejuela, E. M. Ortiz, L. Palazzesi, A. L. Peixoto, S. K. Pell, J. Pellicer, D. S. Penneys, O. A. Perez-Escobar, C. Persson, M. Pignal, Y. Pillon, J. R. Pirani, G. M. Plunkett, R. F. Powell, G. T. Prance, C. Puglisi, M. Qin, R. K. Rabeler, P. E. J. Rees, M. Renner, E. H. Roalson, M. Rodda, Z. S. Rogers, S. Rokni, R. Rutishauser, p. u. family=Salas, given=Miguel F., H. Schaefer, R. J. Schley, A. Schmidt-Lebuhn, A. Shapcott, I. Al-Shehbaz, K. A. Shepherd, M. P. Simmons, A. O. Simões, A. R. G. Simões, M. Siros, E. C. Smidt, J. F. Smith, N. Snow, D. E. Soltis, P. S. Soltis, R. J. Soreng, C. A. Sothers, J. R. Starr, P. F. Stevens, S. C. K. Straub, L. Struwe, J. M. Taylor, I. R. H. Telford, A. H. Thornhill, I. Tooth, A. Trias-Blasi, F. Udovicic, T. M. A. Utteridge, J. C. Del Valle, G. A. Verboom, H. P. Vonow, M. S. Vorontsova, p. u. family=Vos, given=Jurriaan M., N. Al-Wattar, M. Waycott, C. A. D. Welker, A. J. White, J. J. Wieringa, L. T. Williamson, T. C. Wilson, S. Y. Wong, L. A. Woods, R. Woods, S. Worboys, M. Xanthos, Y. Yang, Y.-X. Zhang, M.-Y. Zhou, S. Zmarzty, F. O. Zuloaga, A. Antonelli, S. Bellot, D. M. Crayn, O. M. Grace, P. J. Kersey, I. J. Leitch, H. Sauquet, S. A. Smith, W. L. Eiserhardt, F. Forest, and W. J. Baker. 2024. Phylogenomics and the rise of the angiosperms 629:843–850.

APPENDIX 1

STATISTICAL CONSISTENCY OF THE AGID METRIC

DEFINITION 8 The **internode distance** $d_T(u,v)$ between two leaves u and v in a phylogenetic tree T is the sum of the lengths of the edges along the shortest path connecting u and v in T. The **average gene tree internode distance** (AGID) between two leaves u and v in a set of gene trees $\mathcal{G}$ is the average of the internode distances between u and v taken across all trees in $\mathcal{G}$, namely $\frac{1}{|\mathcal{G}|}\sum_{i=1}^{|\mathcal{G}|} d_{\mathcal{G}_i}(u,v)$.

DEFINITION 9 The **average internode distance** (AGID) of two leaves u and v is their average internode distance across all gene trees $T \in \mathcal{G}$. Namely, $\sum_{T\in\mathcal{G}} d_T(u,v)$.

THEOREM 3. *The expected internode distance between two taxa A and B in the true gene tree T emerging from species tree $\mathcal{T}$ under the MSC is $t_{AB}+2$ where t_{AB} is the sum of the lengths of the edges between A and B in coalescent units in T.*

Proof. Under the MSC, two taxa A and B cannot coalesce until their lineages meet in the species tree $\mathcal{T}$. The time this takes to happen is exactly the distance by which A and B are separated in $\mathcal{T}$ 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 branches are increasing in length, thus the total internode distance between A and B is $t_{AB}+2X$. $\mathbb{E}(X)=1$, so $\mathbb{E}(d_T(A,B)) = t_{AB}+2$. □

COROLLARY 2 The expected AGID between two taxa A and B across a set of n true gene trees $\mathcal{G}$ emerging from species tree $\mathcal{T}$ under the MSC is $t_{AB}+2$.

Proof. This result follows from Theorem 3 and basic properties of expectations.

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

□

THEOREM 4. *For a set of n gene trees $\mathcal{G}_n$ inferred in an unbiased manner, the AGID metric is statistically consistent for the true species tree $\mathcal{T}$ in the sense of Definition 7 if all internal branches in $\mathcal{T}$ have non-zero length.*

Proof. By assuming an unbiased method of inference we have that $\mathbb{E}(d_{\mathcal{G}}(A,B)) = t_{AB}+2$ where t_{AB} is the sum of lengths of the edges in the unique path connecting A and B in $\mathcal{T}$.

Without loss of generality, consider taxa $1,2$ which are true neighbors in $\mathcal{T}$ and an arbitrary pair of taxa

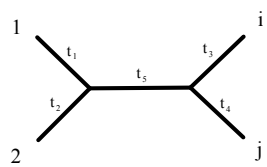

Fig. 1.1. Treelike unrooted quartet and its associated edge lengths. Taxa 1 and 2 are true neighbors whereas taxa i and j are arbitrary.

i, j in $\mathcal{T}$ where each of $1, 2, i, j$ are unique. The true unrooted quartet containing $1, 2, i, j$ that is displayed in $\mathcal{T}$ must be $12|ij$ with branch lengths shown in Figure 1.1 because 1 and 2 are assumed to be true neighbors. It follows then that $t_{12} = t_1 + t_2$, $t_{ij} = t_3 + t_4$, $t_{1i} = t_1 + t_5 + t_3$, and so on. Thus, $\mathbb{E}\left(Q_j^{(T)}(M_D(\mathcal{G}_n))\right) = -2t_5$ for $j = 1, 2$ where $M_D(\mathcal{G}_n) = d_{\mathcal{G}_n}$, so the AGID metric is statistically consistent for $\mathcal{T}$ by Definition 7.

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

□

Quarnet Derivations

We now derive $\mathbb{E}\left(Q_1^{(\mathcal{N})}\right)$ and $\mathbb{E}\left(Q_2^{(\mathcal{N})}\right)$ for each possible level-1, binary, semi-directed quarnet $\mathcal{N}$. To do so, we derive each $\mathbb{E}\left(Q_i^{(\mathcal{N})}\right)$ conditioned on each possible coalescent histories of $\mathcal{N}$ before applying the Law of Total Probability. The derivations of the probabilities of each coalescent history for quarnets $1, 3, 4, 5$, and 6 are outlined in Section 1 of the Supporting Information of Solís-Lemus and Ané (2016), and we additionally include quarnet 2 here for completeness.

Note that for each quarnet there are at most 8 possible labellings that preserve the neighborhood of taxa 1 and 2. These alternative labellings can be obtained by applying combinations of the following moves: (a) swap the labels for taxa 1 and 2, (b) swap the labels for taxa A and B, or (c) swap the labels for taxa A and i *and* swap the labels for taxa B and j. Notice that moves (a) and (b) simply adjust the labels such that Q_1 becomes exactly Q_2 and Q_2 becomes exactly Q_1, and move (c) simply reorders the terms in Q_1 and Q_2. Thus, Q_1 and Q_2 are invariant to the labelling of a given quarnet, so we derive their expectations for a single labelling for each quarnet.

Let $\mathbb{E}_{\mathcal{H}_j}\left(Q_i^{(\mathcal{N})}\right)$ denote the expected value of Q_i for quarnet $\mathcal{N}$ under coalescent history $\mathcal{H}_j$. Then, the results for each quarnet are as follows.

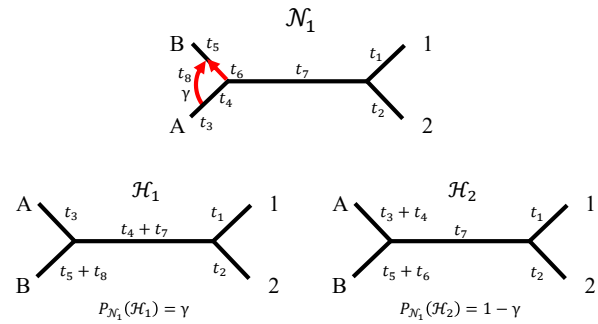

Fig. 1.2. Quarnet 1, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

Quarnet 1

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_1)}\right) &= \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_1)}\right) = -2t_7 - 2t_4 \\
\mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_1)}\right) &= \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_1)}\right) = -2t_7
\end{aligned}$$

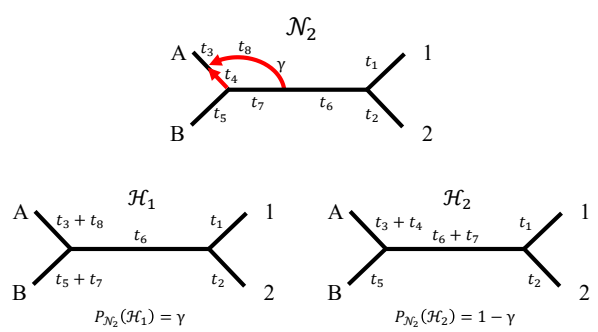

Fig. 1.3. Quarnet 2, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

Quarnet 2

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_2)}\right) &= \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_2)}\right) = -2t_6 \\
\mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_2)}\right) &= \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_2)}\right) = -2t_6 - 2t_7
\end{aligned}$$

Quarnet 3

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_3)}\right) &= \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_3)}\right) = \mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_3)}\right) \\
&= \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_3)}\right) \\
&= -2t_8
\end{aligned}$$

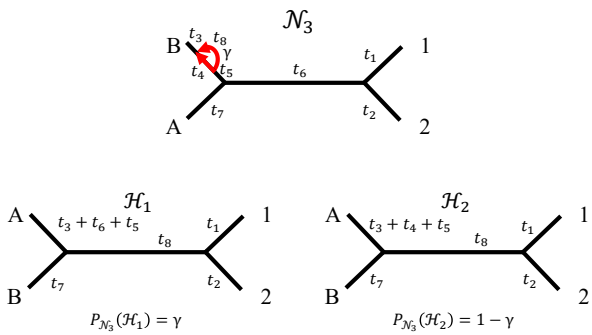

Fig. 1.4. Quarnet 3, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

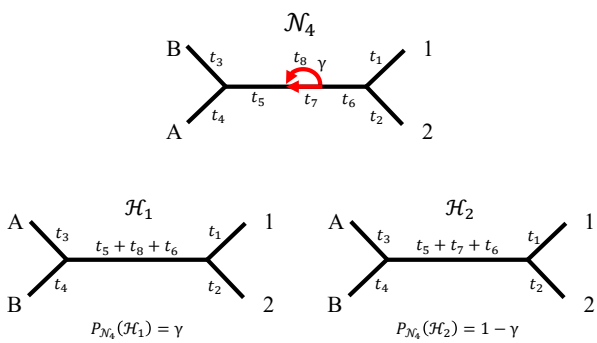

Fig. 1.5. Quarnet 4, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

Quarnet 4

$$\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_4)}\right) = \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_4)}\right) = -2t_5 - 2t_6 - 2t_8$$
$$\mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_4)}\right) = \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_4)}\right) = -2t_5 - 2t_6 - 2t_7$$
$$\mathbb{E}_{\mathcal{H}_3}\left(Q_1^{(\mathcal{N}_4)}\right) = \mathbb{E}_{\mathcal{H}_3}\left(Q_2^{(\mathcal{N}_4)}\right) = -2t_5 - 2t_6$$
$$\mathbb{E}_{\mathcal{H}_4}\left(Q_1^{(\mathcal{N}_4)}\right) = \mathbb{E}_{\mathcal{H}_4}\left(Q_2^{(\mathcal{N}_4)}\right) = -2t_5 - 2t_6$$

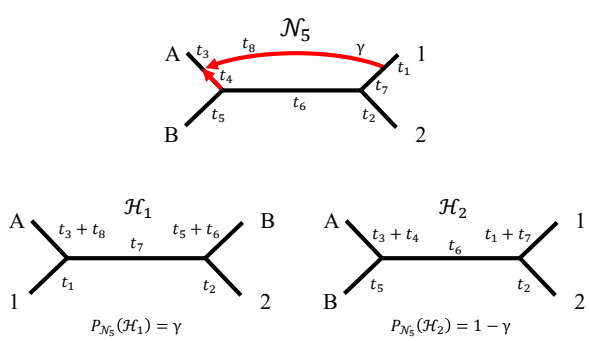

Fig. 1.6. Quarnet 5, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

Quarnet 5

$$\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_5)}\right) = \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_5)}\right) = 2\gamma t_7 - 2(1-\gamma)t_6$$
$$\mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_5)}\right) = \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_5)}\right) = -2(1-\gamma)t_6$$

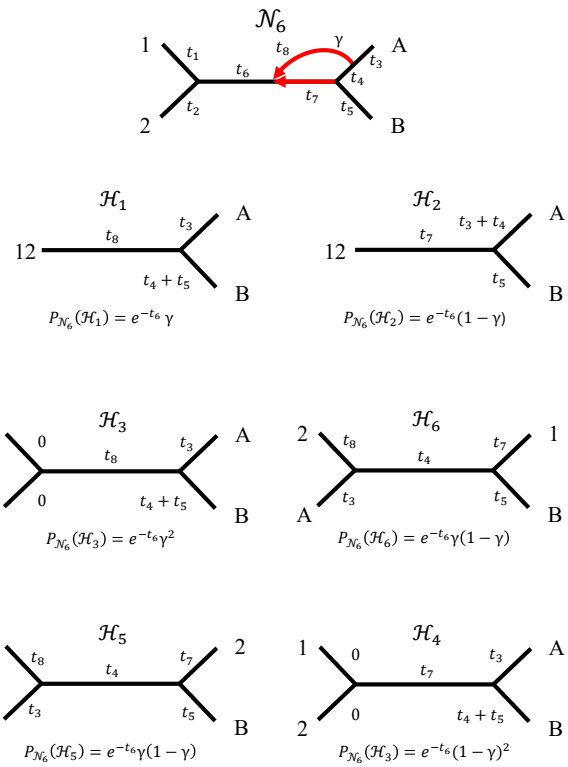

Fig. 1.7. Quarnet 6, its coalescent histories $\mathcal{H}_j$, and their associated probabilities.

Quarnet 6 Under coalescent histories $\mathcal{H}_1$ and $\mathcal{H}_2$, taxon 1 and 2 are assumed to have merged prior to reaching the hybrid node in $\mathcal{N}_6$, so $\mathbb{E}(D_{12})$ takes on the following new form:

$$\begin{aligned}\mathbb{E}_{\mathcal{H}_1}(D_{12}) &= \mathbb{E}_{\mathcal{H}_2}(D_{12}) \\ &= t_1 + t_2 + 2\int_0^{t_6} xe^{-x}dx \\ &= t_1 + t_2 + 2 - 2e^{-t_6}t_6 - 2e^{-t_6}\end{aligned}$$

With this result, we compute each $\mathbb{E}_{\mathcal{H}_j}\left(Q_i^{(\mathcal{N}_6)}\right)$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_1}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_1}\left(Q_2^{(\mathcal{N}_6)}\right) \\
&= 2 - 2t_6e^{-t_6} - 2e^{-t_6} - 2t_6 - 2\gamma t_8 \\
\mathbb{E}_{\mathcal{H}_2}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_2}\left(Q_2^{(\mathcal{N}_6)}\right) \\
&= 2 - 2t_6e^{-t_6} - 2e^{-t_6} - 2t_6 - 2(1-\gamma)t_7 \\
\mathbb{E}_{\mathcal{H}_3}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_3}\left(Q_2^{(\mathcal{N}_6)}\right) = -2t_8 \\
\mathbb{E}_{\mathcal{H}_4}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_4}\left(Q_2^{(\mathcal{N}_6)}\right) = -2t_7 \\
\mathbb{E}_{\mathcal{H}_5}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_6}\left(Q_2^{(\mathcal{N}_6)}\right) = 2t_4 \qquad (1.1) \\
\mathbb{E}_{\mathcal{H}_6}\left(Q_1^{(\mathcal{N}_6)}\right) &= \mathbb{E}_{\mathcal{H}_5}\left(Q_2^{(\mathcal{N}_6)}\right) = 0 \qquad (1.2)
\end{aligned}$$

Note the differing pattern in the indices of Equations 1.1 and 1.2.

$$\begin{aligned}
\mathbb{E}\left(Q_1^{(\mathcal{N}_6)}\right) = &\left(1 - e^{-t_6}\right)\left[2\left(2 - 2t_6e^{-t_6} - 2e^{-t_6} - 2t_6\right)\right. \\
&\left. - 2\gamma t_8 - 2(1-\gamma)t_7\right] \\
&- 2e^{-t_6}\gamma^2 t_8 - 2e^{-t_6}(1-\gamma)^2 t_7 \\
&+ 2e^{-t_6}\gamma(1-\gamma)t_4 \qquad (1.3) \\
=&\mathbb{E}\left(Q_2^{(\mathcal{N}_6)}\right)
\end{aligned}$$

Notice now that the only quarnets that are capable of having non-negative expected Q_i values are quarnets $\mathcal{N}_5$ and $\mathcal{N}_6$. For $\mathbb{E}\left(Q_1^{(\mathcal{N}_5)}\right)$ to be non-negative, which would lead to AGID not being statistically consistent for some displayed tree of the true species network, we need $\gamma t_7 \geqslant (1-\gamma)t_6$ in $\mathcal{N}_5$. This condition seems particularly lenient, but can be largely avoided by including samples from two individuals of each taxa or species. In this case, the labelling of $\mathcal{N}_5$ with 1 or 2 under the hybrid node would no longer be possible, because any leaf under the hybrid node would have a true neighbor that is also under the hybrid node. Then, only the labelling as shown in Figure 1.6 is possible. To satisfy the given inequality in this case, true neighbors 1 and 2 would need to be particularly far from one another, leading to a large t_7, while their most recent common ancestor would need to be very close to the most recent common ancestor of i and j, leading to a small t_6. Whether such a case is particularly common or plausible is unclear, and would require more research.

In the case of $\mathbb{E}\left(Q_j^{(\mathcal{N}_6)}\right)$, $j = 1, 2$, a more complicated relationship arises (see Equation 1.3). In this case, what is required for AGID to *not* be statistically consistent for some displayed tree of the true species network, in general, γ must not be close to 0 or 1, t_6, t_7, and t_8 must be very small relative to t_4, *or* t_4 must be very, very large. This reflects a situation where there is moderate to high levels of gene exchange at the hybrid node while the parents of the hybrid node in $\mathcal{N}_6$ are very far from one another relative to the distances represented by t_6, t_7, and t_8. This is highly reminiscent of the conditions necessary for a network to be anomalous (see Ané et al. (2024)), but is not one and the same due to the linear relationship between each $\mathbb{E}\left(Q_j^{(\mathcal{N}_6)}\right)$ and t_4.

APPENDIX 2

PERFORMANCE STUDY FURTHER DETAILS

Statistics on Generated Data

For each simulation, we computed the gene tree estimation error (GTEE) as the average hardwired cluster distance (HWCD) between each true gene tree and its associated estimated gene tree, divided by the maximum possible HWCD value ($2N - 6$ where N is the number of taxa in the tree). This value varied heavily across simulation settings, the distribution of which is shown in Figure 2.1.

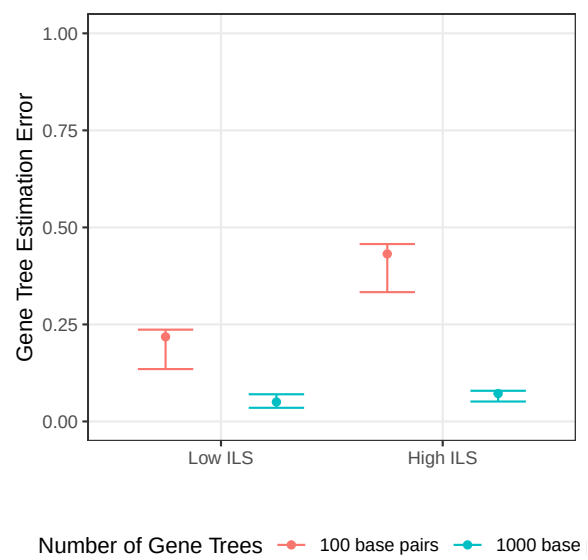

Fig. 2.1. Distribution of gene tree estimation errors (GTEEs) across various simulation settings. Values did not vary significantly across simulations with differing number of taxa or gene trees. Values are irrelevant to simulations that utilized Squirrel because Squirrel utilizes MSAs as input, not estimated gene trees.

Simulation Software Parameters

Default parameters were used for IQ-TREE, ASTRAL v5.7.1, and Squirrel. For IQ-TREE, this means that IQ-TREE's ModelFinder functionality was used and the true underlying model of evolution was never specified when inferring gene trees. Specific divergences from default parameters in other software are enumerated

NBP	ILS	Number of Taxa	Avg. GTEE
100	Low	25	0.221
100	Low	50	0.233
100	Low	100	0.233
100	Low	200	0.267
100	High	25	0.376
100	High	50	0.419
100	High	100	0.449
100	High	200	0.464
1000	Low	25	0.0824
1000	Low	50	0.102
1000	Low	100	0.147
1000	Low	200	0.152
1000	High	25	0.102
1000	High	50	0.0936
1000	High	100	0.140
1000	High	200	0.105

Table 2.1. Average gene tree estimation error (GTEE) for each combination of number of taxa, incomplete lineage sorting (ILS), and number of base pairs (NBP) in the MSA.

below. Any parameters that are unmentioned were left at their default values.

seq-gen For each true gene tree, one MSA was generated with a single sample for each tip in the gene tree. Base pairs were simulated under the HKY model of evolution Hasegawa et al. (1985) with substitution rates $\pi_A = \pi_T = 0.3$ and $\pi_C = \pi_G = 0.2$. The branch length scaling parameter (`-s`) was set to 0.05.

TINNiK Parameters $\alpha = 0.01$ and $\beta = 0.99$ were used.

SNaQ The first estimated gene tree was used as the starting topology for each independent run of SNaQ. Additionally, 20 independent runs were performed for each constraint network. For a given subset of taxa, the maximum number of hybrids parameter (`hmax`) was chosen to be the number of hybrids in the subnetwork of the true species network spanned by that subset of taxa.

PhyloNet As with SNaQ, the maximum number of hybrids in each constraint network was chosen to be the number of hybrids in the subnetwork of the true species network spanned by that constraint network's subset of taxa. 10 independent runs were performed for each constraint network, and each round of inference was performed with 5 processors.

APPENDIX 3

ALGORITHMS

Below, we provide formal algorithms for each step of InPhyNet, along with informal proofs of runtime complexity upper bounds. The following definition is used in Algorithms 3 and 4 below.

DEFINITION 10 Denote the set of all paths from u to w in network N_k as $u \overset{N_k}{\rightsquigarrow} w$. The path $p^*_{N_k}(u, w)$ (or p^* when unambiguous) is the shortest path belonging to $\mathbf{p} = u \overset{N_k}{\rightsquigarrow} w$, unless $\mathbf{p}$ contains a path of length 4 and a path of length 5. Then, $p^*_{N_k}(u, w)$ is the path of length 5. This exception is in place specifically for case 5 in Figure 1 in the main text.

For the convenience of the reader, Table 3 offers descriptions of variables that are used frequently throughout the following algorithms.

Variable	Description
$\widehat{\mathcal{N}}$	The inferred network, constructed by `InPhyNet`
X	The set of taxa on which $\widehat{\mathcal{N}}$ is inferred
D	A pairwise dissimilarity matrix defined on all pairs in X
$\mathbf{S}$ or S_i	The set of disjoint subsets of X or one of these subsets
$\mathbf{C}$ or C_i	The set of constraint networks inferred over $\mathbf{S}$ or the constraint network specifically inferred over S_i
k	The number of disjoint subsets, or $\lvert S\rvert$ and $\lvert C\rvert$
u, v, w	u and v are a pair of nodes being merged, w is their artificial parent node
$\mathbb{V}$	The set of all valid pairs of nodes that can be merged at a given iteration of the algorithm
m	$\max\{\lvert S_i\rvert\}$

Table 3.1. Descriptions of variables used frequently throughout the algorithms below.

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

Input: X a set of taxa labels
D a pairwise dissimilarity matrix on X
$\mathbf{C}$ a set of k level-1, semi-directed networks on disjoint subsets $\mathbf{S} = \{S_i\}_{i=1}^k$ where $\bigcup_{i=1}^k S_i \subseteq X$
Output: A semi-directed network $\widehat{\mathcal{N}} = (V, E)$ on X

Create $|X|$ disconnected leaf nodes V_L where each $v_i \in V_L$ corresponds to taxa X_i
Initialize a list of nodes to consider in merges $L = V_L$
Initialize a map $h : \bigcup_{i=1}^k V_H^{(i)} \to E \times E$
$\widehat{\mathcal{N}} = (V, E) = (V_L, \emptyset)$
while $|L| > 1$ **do**
    # Choose Nodes To Merge u, v
    Compute each $Q_{ij} = (|L| - 2)D_{ij} - \sum_{l=1}^{|L|} D_{li} - \sum_{j=1}^{|L|} D_{lj}$
    $\mathbb{V} =$ ValidMergingPairs($\mathbf{C}, X$) (see Algorithm 2)
    if $|V| = 0$ **then**
        return PairwiseInPhyNet($\mathbf{C}, X$) (see Algorithm 5)
    end if
    $(u, v) = L_i, L_j$ where $i, j = \arg\min_{(i,j)\in\mathbb{V}} Q_{ij}$

    # 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 $C \in \mathbf{C}$ **do**
        AddHybridInformation!($\widehat{\mathcal{N}}, h, C, u, v$) (see Algorithm 3)
    end for

    # Update $\mathbf{C}$ ***and*** D
    for $C \in \mathbf{C}$ **do**
        MergeNodesInConstraint!(C, u, v, p) (see Algorithm 4)
    end for
    Add a row and column to D so that now $D \in \mathbb{R}^{m\times m}$
    for $k = 1, 2, \ldots, m - 1$ **do**
        $D_{k,m} = D_{m,k} = (D_{k,i} + D_{j,k} - D_{i,j})/2$
    end for
    Remove rows and columns i, j from D
end while

 # 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 $e_{2p} \to e'_{2p}$ and $e'_{1p} \to e'_{2p}$
end for
Suppress the root of $\widehat{\mathcal{N}}$
return $\widehat{\mathcal{N}}$

In the following, assume that all subsets are of equal size, or take m to be the size of the largest subset. In the ***Choose Nodes To Merge*** u, v section, there are at most on the order of mk valid merging pairs in $\mathbb{V}$ where k is the total number of disjoint subsets (see Algorithm 2), so we need only calculate each Q_{ij} for these pairs. This leaves us with a runtime complexity of $O(k^2m^2)$. Then, the first 3 lines of ***Merge*** u ***and*** v happen in constant time while the for loop is executed k times, and its contents have runtime complexity $O(m + r)$ (see Algorithm 3), leaving it with runtime complexity $O(km + kr)$. Finally, examining each for loop in the section ***Update*** $\mathbf{N}$ ***and*** D gives the runtime complexity $O(km + kr))$ (see Algorithm 4). Each of these steps happens $|X|$ times, so the overall runtime

complexity, excluding post-processing, given that `PairwiseInPhyNet` is never called, is $O(|X|k^2m^2 + |X|kr)$. Notice that, by definition, $\frac{|X|}{k} = m$, so this complexity reduces to $O(|X|^3 + |X|kr)$.

In the worst case, this algorithm is executed until the final iteration where $|\mathbb{V}| = 0$ on line 9 leads to `PairwiseInPhyNet` being called. The runtime complexity of `PairwiseInPhyNet` is k times the runtime complexity of this algorithm (see Algorithm 5; `PairwiseInPhyNet` is never called when `InPhyNet` is called from `PairwiseInPhyNet`). When `InPhyNet` is called from inside of `PairwiseInPhyNet`, k is fixed as 2. Further, `InPhyNet` is called a total of $k-1$ times from within `PairwiseInPhyNet`. Thus, the worst-case complexity of the `InPhyNet` algorithm is $O(|X|k^2m^2 + |X|kr + (k-1)(2^2|X|m^2 + 2|X|r)) = O(|X|^3 + |X|kr)$. Thus, the runtime complexity of the `InPhyNet` algorithm is $O(|X|^3 + |X|kr)$ or $O(|X|^3 + |X|r)$ if k is treated as fixed.

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

Input: $\mathbf{C}$ 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

1: Initialize a map $f : X \times X \to \{0, 1\}$
2: For all $(u, v) \in X, u \neq v$, set $f[u, v] = 1$ and $f[u, u] = 0$
3:
4: ***# Check Neighbors***
5: **for** $(u, v) \in X, u \neq v$ **do**
6: **for** $C_i \in \mathbf{C}$ **do**
7: **if** $(u, v) \in C_i$ AND (u, v) not neighbors in C_i **then**
8: $f[u, v] = 0$
9: **end if**
10: **end for**
11: **end for**
12:
13: ***# Check Pairwise Topological Isomorphism***
14: **for** $(u, v) \in \{(u, v) : f[u, v] = 1\}$ **do**
15: $\mathbf{C}' = \text{Copy}(\mathbf{C})$
16: **for** $C' \in \mathbf{C}'$ **do**
17: $p =$ parent of u, v in C'
18: MergeNodesInConstraint!(C', u, v, p) (see Algorithm 4)
19: **end for**
20:
21: **for** $i = 1, 2, \ldots, k-1$ **do**
22: overlap $= \{w : w \in C'_i, w \in C'_{i+1}\}$
23: **if** RF$(C'_i[\text{overlap}], C'_{i+1}[\text{overlap}]) \neq 0$ **then**
24: $f[u, v] = 0$
25: **break**
26: **end if**
27: **end for**
28: **end for**
29:
30: **return** $\{(u, v) : f[u, v] = 1\}$

The overall runtime of this algorithm can be broken down into two sections: the portion marked ***Check Neighbors*** and the portion marked ***Check Pairwise Topological Isomorphism***. Checking neighborhood status can be done more efficiently by initially marking all pairs of taxa as non-neighbors, then iterating over each leaf in each network $N_i \in \mathbf{N}$ and checking whether that leaf has a sibling. Checking for a sibling can be done in constant time, and each network has at most m taxa, so this process has runtime complexity $O(mk)$.

A given unrooted tree with m can have at most $\frac{m}{2}$ sets of neighbors (in the case where the tree is perfectly balanced), and, by our definition of neighbor, networks can have one additional neighbor per reticulation. Thus, the most neighbors that a single network can have is $\frac{m}{2} + r$. Then, Algorithm 4 has runtime complexity $O(m + r)$, and the Robinson-Foulds (RF) distance metric can be computed in linear time (Day, 1985). Thus, the runtime complexity of the second portion of this algorithm is $O(k(m + r)^2 + (m + r)km$, bringing the overall algorithm runtime complexity to $O(k(m + r)^2)$.

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

Input: $\widehat{\mathcal{N}}$ the partially inferred network
h a map from the set of all reticulations in $\mathbf{C}$ to a pair of edges in $\widehat{N}$
$C_i = (V^{(i)}, E^{(i)})$ a constraint network
u, v two nodes in the process of being merged
Output: Nothing, h is updated in-place

1: **if** $u \notin V^{(i)}$ OR $v \notin V^{(i)}$ **then**
2: **return**
3: **end if**
4:
5: Find $p^* \in u \overset{C_i}{\rightsquigarrow} v$
6: **if** $|p^*| \leqslant 3$ **then**
7: Nothing; # Case 1 in Figure 1
8: **else if** $|p^*| = 5$ AND $p_4^* \in V_H^{(i)}$ **then**
9: $h[p_4^*] = (e_u, e_v)$; # Case 4 in Figure 1
10: **else if** $|p^*| = 5$ AND $p_2^* \in V_H^{(i)}$ **then**
11: $h[p_2^*] = (e_v, e_u)$; # Case 4 in Figure 1 with u, v swapped
12: **else if** $p_2^* \in V_H^{(i)}$ **then**
13: $h[p_2^*][1] = e_u$; # Case 2 in Figure 1
14: **else if** $p_3^* \in V_H^{(i)}$ **then**
15: $h[p_2^*][1] = e_v$; # Case 2 in Figure 1 with u, v swapped
16: **else**
17: **if** p_2^* is connected to a hybrid edge **then**
18: $h[p_2^*][2] = e_u$; # Case 3 in Figure 1
19: **else**
20: $h[p_2^*][2] = e_v$; # Case 3 in Figure 1 with u, v swapped
21: **end if**
22: **end if**

Each portion of this algorithm runs in constant time except for line 6 when $p^* \in u \overset{C_i}{\rightsquigarrow} v$ is obtained. In the worst case with a pathfinding algorithm like A*, the runtime complexity of this step is upper bounded by the number of edges in the graph.

To derive an upper bound for this value, we examine the maximum number of edges that can appear in tree-based networks, a parent class to level-1 networks (see Kong et al. (2022) for further details). All tree-based networks can be constructed by taking a tree and attaching a series of reticulations to the tree. To construct a semi-directed level-1 network, we will begin this process with an unrooted phylogenetic tree, which has $2m - 3$ total edges when m is the number of tips in the network. Then, each time a reticulation is added to this topology, the most new edges that can be created is three (one from splitting the edge at the origin of the reticulation, one from splitting the edge at the destination of the reticulation, and the third being the reticulation itself). Thus, each reticulation adds at most 3 new edges to the network, and we have an upper bound of $2m + 3r - 3$ where m is the number of tips in the network and r is the number of reticulations in the network. This gives an upper-bounded runtime to Algorithm 4 of $O(m + r)$.

Algorithm 4 MergeNodesInConstraint!(C_i, u, v, P)

Input: $C_i = (V, E_T \cup E_H)$ a constraint network
u, v nodes being merged
w the new parent of u and v
Output: Nothing, C is updated in-place

if $u \in C_i \cap v \notin C_i$ **then**
  Replace u with p
else if $u \notin C_i \cap v \in C_i$ **then**
  Replace v with p
else if $u \in C_i \cap v \in C_i$ **then**
  Find $p^* \in u \overset{C_i}{\rightsquigarrow} v$
  if $|p^*| = 2$ **then**
    Set $C_i = (w, \emptyset)$ in-place
  else if $|p^*| = 3$ **then**
    Remove nodes p_1^*, p_3^* from C_i
    Remove all edges along path p^* from C_i
    Replace p_2^* with w
  else if $|p^*| = 4$ **then**
    if $p_2^* \in E_H$ OR p_2^* is connected to a hybrid edge **then**
      Remove nodes p_1^*, p_2^*, p_4^* from C_i
      Remove all edges along path p^* from C_i
      Replace p_3^* with w
    else
      Remove nodes p_1^*, p_3^*, p_4^* from C_i
      Remove all edges along path p^* from C_i
      Replace p_2^* with w
    end if
  else
    Remove nodes $p_1^*, p_2^*, p_4^*, p_5^*$ from C_i
    Remove all edges along path p^* from C_i
    Replace p_3^* with w
  end if
end if

Following the same reasoning as for Algorithm 3, this algorithm's runtime is $O(m + r)$.

Algorithm 5 PairwiseInPhyNet(**C**, D, X)

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

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

1: Create FIFO queue Q
2: **for** $i = 1, 2, \ldots, k$ **do**
3: Push (C_i, S_i) to Q
4: **end for**
5:
6: **while** $|Q| > 1$ **do**
7: Pop (C_1, S_1) from Q
8: Pop (C_2, S_2) from Q
9: **if** $|Q| = 2$ **then**
10: $I = \{1, 2, \ldots, |X|\}$
11: **else**
12: $I = \{i : X_i \in S_1 \cup S_2\}$
13: **end if**
14: $\widehat{\mathcal{N}}_{12}$ =InPhyNet$(\{C_1, C_2\}, D_{I,I}, X_I)$
15: Push $(\widehat{\mathcal{N}}_{12}, S_1 \cup S_2)$ to Q
16: **end while**
17: Pop $\widehat{\mathcal{N}}, X$ from Q
18: **return** $\widehat{\mathcal{N}}$

The **while** loop on line 6 will execute $k - 1$ times, so the runtime complexity of this function will be $k - 1$ times the runtime complexity of the `InPhyNet` function. See the discussion under Algorithm 1 for further details.

APPENDIX 4

MERGE CONFLICT EXAMPLE

Here, we provide an example of when the `InPhyNet` algorithm would fail without Algorithm 5. Below we provide a dissimilarity matrix D (Table 4.1) and a set of constraints $\mathbf{C}$ that leads to this result. Additionally, we show how the constraints change throughout the algorithm, eventually reaching a point where there do not exist any pairs of nodes that can be merged.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
t_1	0.0	0.83	0.99	0.54	0.19	0.49	0.22	0.61	0.28	0.21	0.94	0.73
t_2	-	0.0	0.46	0.86	0.45	0.47	0.08	0.17	0.61	0.94	0.43	0.74
t_3	-	-	0.0	0.5	0.92	0.47	0.47	0.35	0.46	0.03	0.18	0.1
t_4	-	-	-	0.0	0.5	0.69	0.07	0.5	0.31	0.67	0.67	0.86
t_5	-	-	-	-	0.0	0.67	0.83	0.27	0.5	0.21	0.19	0.96
t_6	-	-	-	-	-	0.0	0.93	0.98	0.93	0.28	0.65	0.17
t_7	-	-	-	-	-	-	0.0	0.46	0.92	0.03	0.8	0.21
t_8	-	-	-	-	-	-	-	0.0	0.46	0.24	0.1	0.65
t_9	-	-	-	-	-	-	-	-	0.0	0.12	0.2	0.2
t_{10}	-	-	-	-	-	-	-	-	-	0.0	0.65	0.7
t_{11}	-	-	-	-	-	-	-	-	-	-	0.0	0.81
t_{12}	-	-	-	-	-	-	-	-	-	-	-	0.0

Table 4.1. Example distance matrix that leads to conflict.

We use the notation $(t_i, t_j) \rightarrow t_k$ to indicate that nodes t_i and t_j have been chosen to be merged, with the new artificial parent that will be replacing them in the constraint networks and dissimilarity matrix denoted t_j. Then, this distance matrix along with the constraint networks shown in the top-left of Figure 4.1 lead to a merge conflict.

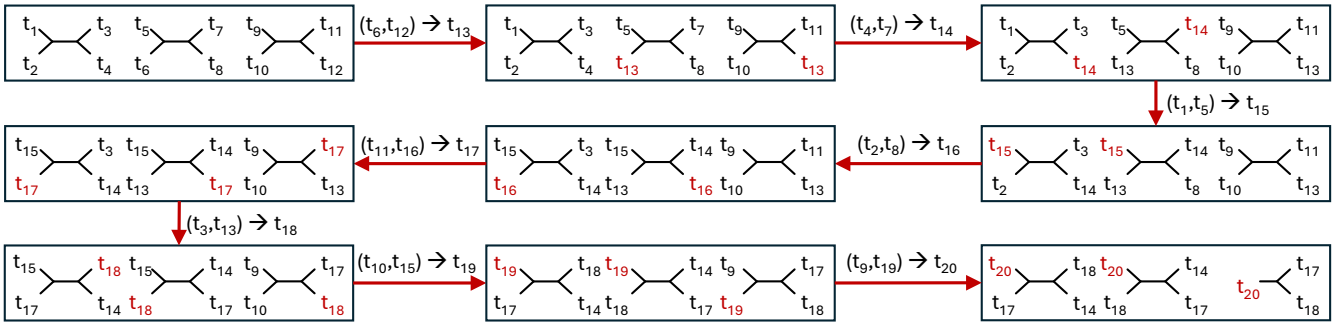

Fig. 4.1. Example of a conflict that can arise in `InPhyNet` when more than 2 constraint networks are used as input. The set of constraint networks at each step of the algorithm is outlined in a black rectangle. Red arrows between these rectangles indicate the algorithm progressing forward one step. The label on each of these arrows $(t_i, t_j) \rightarrow t_k$ indicates that nodes t_i and t_j are being merged and replaced by their new artificial parent t_k. After each step of the algorithm, nodes that were modified are given red font. The algorithm begins in the top left and ends in the bottom right. At the end of the algorithm, no valid pairs of nodes exist that can be merged.

Notice at the end of the algorithm example shown in Figure 4.1 that there are 4 unique artificial taxa remaining among the constraints: t_{14}, t_{17}, t_{18}, and t_{20}. The third constraint is a triplet, so it does not restrict the set of valid nodes. The first and second constraints, on the other hand, contain the same 4 tips but with differing topologies. This means that every possible pair of these remaining 4 taxa will be valid in one of these networks, but invalid in the other.

APPENDIX 5

GREEN PLANTS SUBSET DECOMPOSITION

Here, we enumerate the subsets (including outliers) with which we inferred constraint networks using SNaQ Solís-Lemus and Ané (2016). Networks were inferred with $h = 0, 1, 2, 3$ reticulations for each subset. We provide the tree of blobs inferred with TINNiK Allman et al. (2024) that were used to perform manual subset decomposition in Figures 5.1 and 5.3. Additionally, we provide figures displaying the model selection decisions for these subsets in Figures 5.2 and 5.4.

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 cuspidata*, *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*,

Fig. 5.1. Gymnosperm tree of blobs inferred by TINNiK ($\alpha = 0.01$, $\beta = 0.99$).

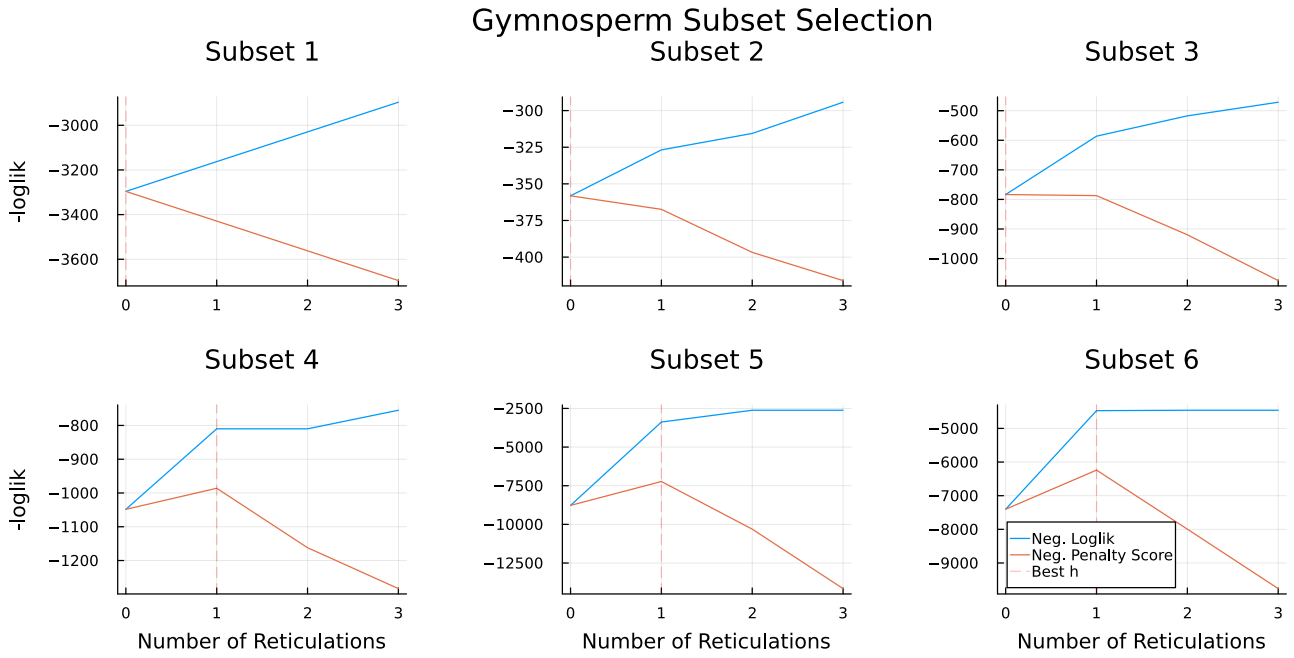

Fig. 5.2. Model selection figures corresponding to each of the 6 subsets listed above. The blue line is the negative log pseudolikelihood of each model (higher is better), the orange line is the penalized score for each model (higher is better), and the vertical dashed red line is placed at the value of h with the largest penalized score. This is the model that was selected.

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*,

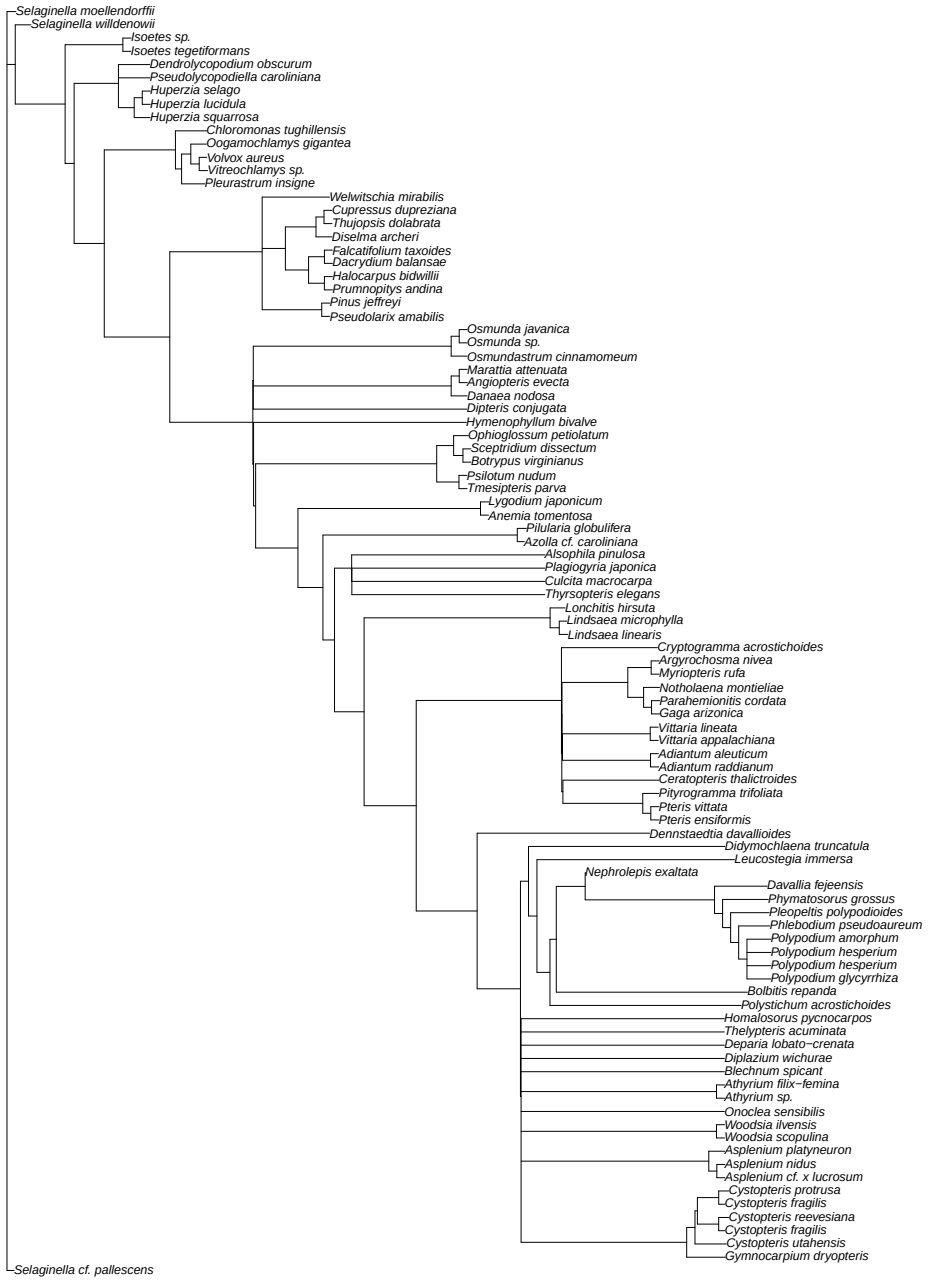

Fig. 5.3. Fern tree of blobs inferred by TINNiK ($\alpha = 0.001$, $\beta = 0.999$).

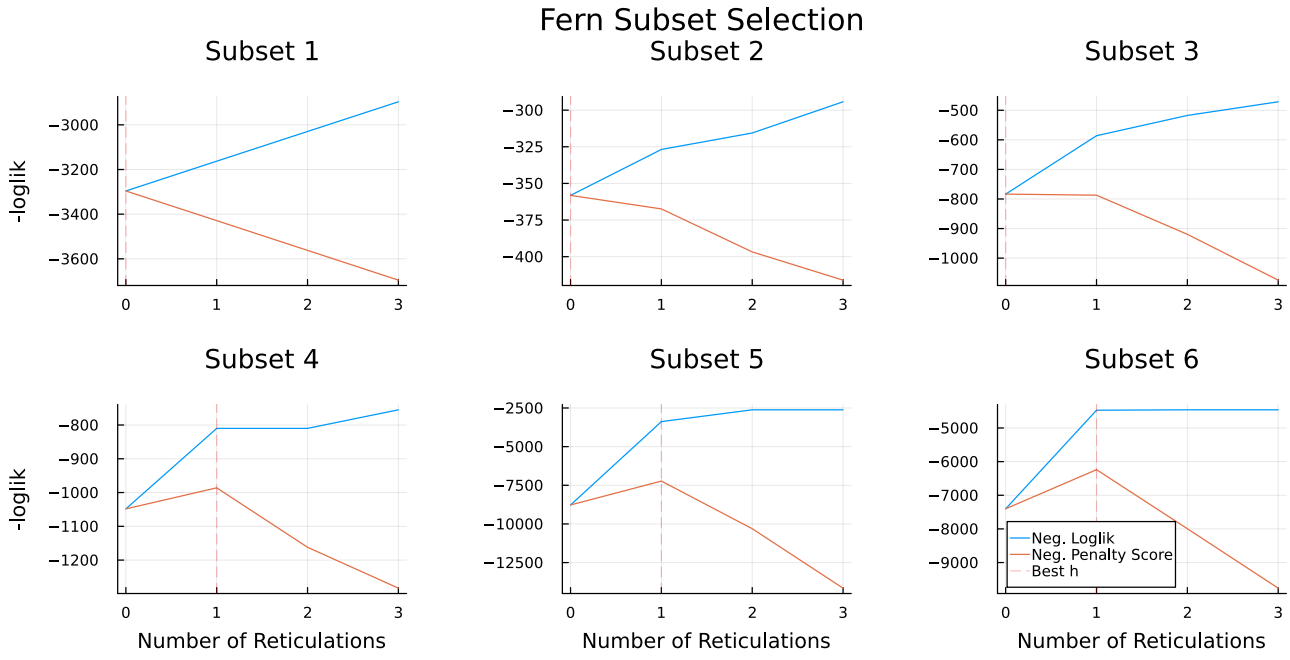

Fig. 5.4. Model selection figures corresponding to each of the 6 subsets listed above. The blue line is the negative log pseudolikelihood of each model (higher is better), the orange line is the penalized score for each model (higher is better), and the vertical dashed red line is placed at the value of h with the largest penalized score. This is the model that was selected.