Rooted NNI moves on tree-based phylogenetic networks

Péter L. Erdős[1,*]
peter.erdos@renyi.hu
ORCiD: 0000-0002-1139-2316

Andrew Francis[2]
A.Francis@westernsydney.edu.au
ORCiD: 0000-0002-9938-3499

Tamás Róbert Mezei[1,*,†]
tamas.robert.mezei@renyi.hu
ORCiD: 0000-0002-7608-3215

[1]Alfréd Rényi Institute of Mathematics (a Hungarian Academy of Sciences Centre of Excellence), Reáltanoda u. 13–15, 1053 Budapest, Hungary
[2] Centre for Research in Mathematics and Data Science, Western Sydney University, Sydney, Australia

September 18, 2020

Abstract

We show that the space of rooted tree-based phylogenetic networks is connected under rooted nearest-neighbor interchange (*rNNI*) moves.

Keywords: nearest-neighbor interchange, rooted NNI, rooted phylogenetic network, directed acyclic graph

1 Introduction

Phylogenetic networks are a generalisation of phylogenetic trees that have become widely used as ways to represent evolutionary histories, because they are able to either capture uncertainty in the inference, or represent non-tree-like evolutionary processes [1, 14] (see also the texts [9, 17]). Such processes include hybridization, in which two species combine to produce a third, and horizontal gene transfer, in which genetic material from one species is incorporated into that of another (common in bacteria).

Despite these non-tree-like evolutionary events, evolution can still appear "tree-like", in the sense that it may be representable as having a broad, underlying tree, with additional arcs (directed edges) between the arcs of the tree. This sense motivated the definition of a "tree-based network" [6].

Tree-based networks have become an active area of research because they capture biological intuition and have many mathematical characterisations [5, 12, 18] and connections to other well-studied properties (for example they are precisely the "tree-child" networks for which every embedded tree is a base tree [16]).

For many applications, it is important to be able to randomly move around a set of phylogenetic networks, for instance when searching for a network that maximises a likelihood, or has the highest parsimony score. Mechanisms that allow such movement are important, as without them such sampling is very difficult.

The *nearest neighbour interchange (NNI)* is a local operation on a graph that is widely used for moving around the space of trees or networks. It was introduced for phylogenetic trees in 1971 [15], generalised to unrooted phylogenetic networks in 2016 [8], and for rooted networks shortly after [7] (where the move is called rNNI). The spaces of such trees and networks are connected under the relevant rNNI moves, and this allows random walks within those spaces to search for optimal trees or networks.

In this paper we prove that the space of *rooted tree-based networks* is connected under rNNI moves. This is the rooted analogue of the result of Fischer and Francis [4] showing the connectedness of (unrooted) tree-based networks under NNI moves. A similar result has been proved for rooted phylogenetic networks that allow parallel edges (which we do not), in Klawitter [13]. We also show that the space is connected under the recently introduced *tail*-moves [11].

*PLE and TRM were supported in part by the National Research, Development and Innovation Office — NKFIH grant K 116769, KH 126853, and K 132930

†Corresponding author

The paper begins by introducing necessary concepts in Section 2. We then explore the effect of rNNI moves on a tree-based network with some technical results in Section 3, and finally prove connectedness in Section 4.

There are many opportunities to extend this work. For instance, extending the understanding of tree-based networks as a space, it would be interesting to extend the notion of *tree-based rank*, introduced for unrooted networks in [3], to the rooted case. This would be another generalization of the proximity measures for rooted tree-based networks discussed in [5]. Finally, there are many other useful classes of network that might be connected under such rearrangements, including well-studied classes such as the tree-child network, and the recently introduced *orchard* networks [2].

2 Preliminaries

A *rooted phylogenetic network* N on X is a directed, simple acyclic graph with the following types of vertices:

- a single vertex of out-degree 1 or 2 and in-degree 0 called the *root*;
- vertices of in-degree 1 and out-degree 0 called *leaves*, which are labelled bijectively by the elements of X;
- vertices of in-degree 1 and out-degree 2, called *tree vertices*;
- vertices of in-degree 2 and out-degree 1, called *reticulation vertices*.

Write $V = V(N)$ for the set of vertices of N, and $E = E(N)$ for the set of arcs (directed edges) of N. For an arc $e = (u, v) \in E$, write $s(e) := u$ and $t(e) := v$ for the *source* and *target* of e, respectively. If $(u, v) \in E(N)$, we say N has *an arc on* $\{u, v\}$.

Rooted phylogenetic networks with the above properties are commonly called *binary*. Denote the set of rooted phylogenetic networks on X by $RP(X)$. Throughout this paper phylogenetic networks will be taken to be both rooted and binary unless otherwise stated.

A rooted phylogenetic network without reticulation vertices is actually a rooted tree, hence it is called a rooted phylogenetic X-tree.

An arc $e = (u, v)$ of $N \in RP(X)$ may be *subdivided* by removing e, and adding a new vertex w and new arcs (u, w) and (w, v). A network with a subdivided arc is no longer a phylogenetic network because it contains a vertex of degree 2. In the other direction, a vertex w of degree 2 may be *suppressed* by deleting it and its two incident arcs (u, w) and (w, v), and adding the arc (u, v) to the network.

A rooted phylogenetic network that has a spanning tree T whose leaves are precisely the leaves of N, is a *tree-based network* [6]. Such a spanning tree for a tree-based network N is called a *support tree* for N. Note that a support tree for N is generally not a phylogenetic X-tree, because it will have vertices of degree 2 (unless N is itself a tree, in which case $T = N$). By "suppressing" the vertices of degree 2 in T, one obtains a phylogenetic X-tree $\widehat{T}$ that is called a *base tree* for N, in the sense that N may be obtained from $\widehat{T}$ by "subdividing" arcs of T and adding additional arcs, as in the original definition in [6].

The set of tree-based networks is denoted $TBN(X) \subseteq RP(X)$.

Nearest neighbour interchange (NNI) operations defined on phylogenetic trees have been used to explore the space of trees for half a century [15]. They have recently been generalised to unrooted phylogenetic networks [8], and to rooted networks [7], as in Definition 2.1.

Definition 2.1. Suppose $N \in RP(X)$ has arcs on $\{a, b\}, \{b, c\}, \{c, d\}$, for distinct vertices $a, b, c, d \in V(N)$. A *rooted nearest neighbour interchange* (*rNNI*) move on $\{a, b\}, \{b, c\}, \{c, d\}$, replaces those arcs with arcs on $\{a, c\}, \{b, c\}, \{b, d\}$, with the following conditions:

1. the in-degrees and out-degrees of a and d are unchanged;
2. the in-degrees and out-degrees of b and c remain 1 or 2;
3. the network remains an acyclic phylogenetic network.

Note that (3) precludes the network N from containing arcs on the arcs $\{a, c\}$ and $\{b, d\}$.

An *rNNI* move is a local operation on a subgraph of N of four vertices and three arcs. If P and Q are subgraphs of N such that $|V(P)| = |V(Q)| = 4$ and $|E(P)| = |E(Q)| = 3$, we say that an *rNNI* move switches P to Q if it changes N to N' where $E(N') = (E(N) \setminus E(P)) \cup E(Q)$.

For the proof of the main result we will need the notion of a "burl-rooted tree", defined as follows. This is a rooted version of the networks with "k-burls" used in [10].

Definition 2.2. A *burl-rooted tree* is a rooted phylogenetic network N with reticulation vertices $b_1, \dots, b_k$ and root ρ with the following properties:

- there is a path from ρ to a leaf ℓ_1 that consists only of the vertices $(\rho, b_1, \dots, b_k, \ell_1)$;
- all paths from ρ to other leaves begin $(\rho, a_1, \dots, a_k, u, \dots)$ for tree vertices $a_1, \dots, a_k$; and
- N contains arcs (a_i, b_i) for $i = 1, \dots, k$.

The structure of a burl-rooted tree is illustrated in Figure 1.

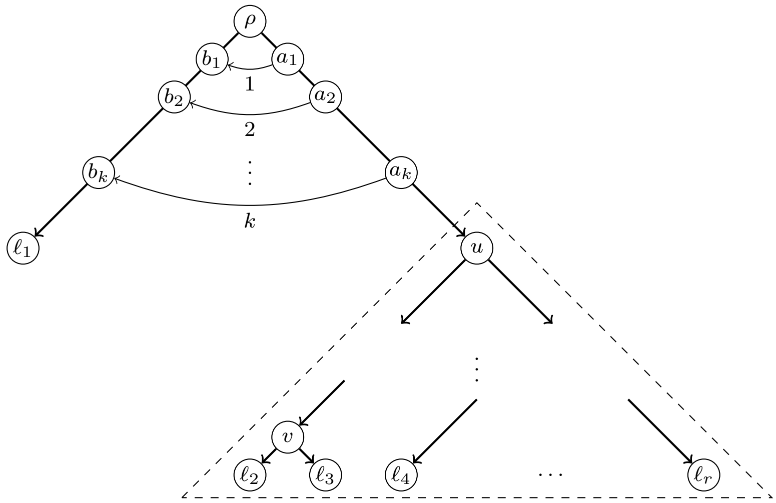

Figure 1: The structure of a burl-rooted tier-k tree. The k reticulation arcs join the vertices between $\rho - \ell_1$ and $\rho - u$. The vertices contained inside the dashed triangle induce a rooted binary tree with $r - 1$ leaves.

Finally we recall the definition of head and tail moves, introduced in [11].

Definition 2.3. Let $e = (u, v)$ and f be arcs in a rooted phylogenetic network N. A *tail move* of e to f involves: deleting e; subdividing f with a new node u'; suppressing u; and adding the arc (u', v). A *head move* of e to f involves: deleting e; subdividing f with a new node v'; suppressing v; and adding the arc (u, v').

3 The impact of *rNNI* moves on tree-based-ness

Lemma 3.1. *Let N be a rooted tree-based phylogenetic network with support tree T. Suppose $P : u \to v \to w \to z$ is a path of length 3 in T. Let $e, f \in E(N) \setminus E(P)$ be arcs incident to v and w, respectively. If either*

(a) *f is oriented away from w and $e \neq vz$, or*

(b) *e is oriented towards v and $f \neq uw$, or*

(c) *$f \neq uw$ and $e \neq vz$, and N does not contain a directed $t(e) \to s(f)$ path,*

then the rNNI move switching the path P to $Q : u \to w \to v \to z$ is valid and the resulting network N' is still tree-based.

Remark 3.2. The *rNNI* move $P \to Q$ simply relocates w onto the uv arc. Depending on the orientation of f, this *rNNI* move is equivalent (up to isomorphism) to a distance-1 head-move or tail-move.

Proof. Both $zv, wu \notin E(N)$, because either arc would make N cyclic. If f is oriented away from w, then $uw \notin E(N)$, because w has total degree 3. If e is oriented towards v, then $vz \notin E(N)$, because v has total degree 3.

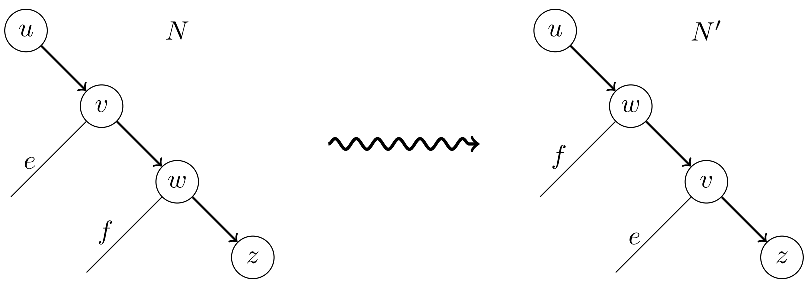

Figure 2: An *rNNI* move which is valid if one of the three conditions of Lemma 3.1 hold.

Thus the network N' created by the *rNNI* move transforming P into Q is well-defined, but it might contain an oriented cycle. Suppose there is an oriented cycle in N', let the shortest one be $\vec{C}$. Note that this forces $z \neq \rho$, because the root ρ has in-degree 0.

Suppose first, that $f, wv, e \in \vec{C}$: then f is oriented towards w, e is oriented away from v, and there is a directed $t(e) \to s(f)$ path in N', which is also present in N. In any case, we have a contradiction. If both $e, f \in \vec{C}$, but $wv \notin \vec{C}$, then $\vec{C}$ is not the shortest oriented cycle, because we could shortcut through wv. Because e and f cannot be both traversed by the cycle, $\vec{C}$ can be trivially shortened or extended by one arc to form an oriented cycle in N.

Lastly, observe that $T' = T - E(P) + E(Q)$ is a support tree of N'. □

Lemma 3.3. *Let N be a rooted phylogenetic network with support tree T. Suppose $P : u \leftarrow z \to v \to w$ is a subgraph of 3 arcs in T. If there is no $u \to v$ path in N and $vu \notin E(N)$, then the rNNI move switching P to $Q : u \leftarrow v \leftarrow z \to w$ is valid and the resulting network is still tree-based. The statement holds even if $z = \rho$.*

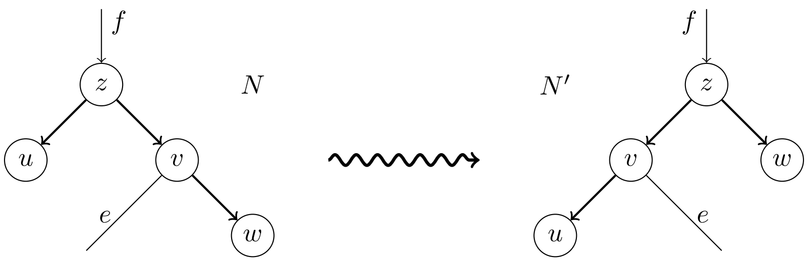

Figure 3: A child v of z can move across z into its other branch if there is no arc joining $\{u, v\}$. If $z = \rho$ then f should be omitted from the picture.

Proof. By the assumptions, there is no arc of any orientation between u and v. Because N is acyclic, $wz \notin E(N)$. Therefore the *rNNI* move switching P to Q produces a valid network.

Suppose there is an oriented cycle in N'; let the shortest one be $\vec{C}$.

Suppose first, that $e \in \vec{C}$: either e is oriented towards v and $vu \in \vec{C}$, or e is oriented away from v and $zv \in \vec{C}$. In any case, this means that there is a $u \to v$ path in N. If $f \in \vec{C}$ and $e \notin \vec{C}$, then $\vec{C}$ can be trivially shortened or extended by one arc to form an oriented cycle in N. If $e, f \notin \vec{C}$, then $\vec{C}$ is already an oriented cycle in N.

Lastly, observe that $T' = T - E(P) + E(Q)$ is a support tree of N'. □

4 The connectedness of the space of tree-based networks

Lemma 4.1. *The set of burl-rooted trees in tier k is connected under rNNI moves.*

Proof. By definition, the burls of burl-rooted trees in tier k are identical, and they only differ by the trees below the burl (vertex u in Figure 1). Since the space of trees is connected under *rNNI* moves [15], one can be transformed into the other, treating the vertex in position u as the root. □

We can now prove our main theorem.

Theorem 4.2. *TBN(X) is connected under rNNI moves.*

Proof. We show that any tree-based network can be transformed into a burl-rooted tree (Definition 2.2), and use the fact that it is possible to move between any two networks in that form (Lemma 4.1).

We fix an arbitrary tree-based network N, and a support tree T for N. In each step we need to cover four cases regarding N and T and their root ρ:

(A) ρ has out-degree 1 in T;

(B) ρ has out-degree 2 in T and on both sides of the root there are branching points in T;

(C) ρ has out-degree 2 in T, but on one side of the root there are no branching points in T;

(D) ρ has out-degree 2 in T and T is path (in this case $|X| = 2$).

Note, that multiple support trees may exist for a fixed tree-based network N. Furthermore, the degree of ρ might be 1 in one support tree, and 2 in another, which means that in the first case ρ must be the source of a reticulation arc.

4.1 Case (A), when $|X| = 1$.

Although this is a degenerate case, we need to deal with it for the sake of completeness. There is no burl-rooted tree when there is only one leaf. Let $e = \rho v$ be the reticulation arc incident on the root. Lemma 3.1(a) applies to the source of reticulation that is below v and closest to it. Therefore, we can move every source of reticulation between ρ and v one-by-one. Next, via Lemma 3.1(b), we can move every target of reticulation below v similarly, in an appropriate order. Lastly, we can freely permute the sources between ρ and v via Lemma 3.1(a), and similarly, we can permute the targets below v freely via Lemma 3.1(b). Thus it is clear that any two networks of tier-k that have exactly one leaf each are connected via *rNNI* moves.

4.2 Case (A), when $|X| \geq 2$.

Let b_1 be the branching point in T which is the closest to ρ (in both T and N). Let ℓ be an arbitrarily chosen leaf, and let b_2 be the closest branching point to it. (We may have $b_1 = b_2$). Let $d_T(x, y)$ be the undirected distance between x and y in the support tree T for N. Define the quantity

$$\Theta_{N,T} := \sum_{f \in E(N) \setminus E(T)} d_T(\rho, s(f)) + \sum_{e \in E(N) \setminus E(T)} d_T(t(e), \ell).$$

Let the number of reticulation arcs be τ. We claim that via *rNNI* moves we can reduce Θ to $\binom{\tau}{2} + \binom{\tau+1}{2} = \tau^2$ (note, that there is a reticulation arc whose source is ρ), ie., in the desired network, a vertex v is

- the source of a reticulation if and only if $v = \rho$ or v is between ρ and b_1 on the support tree,
- the target of a reticulation if and only if v is between b_2 and ℓ on the support tree.

Suppose f is a reticulation arc for which $d_T(\rho, s(f)) > d_T(\rho, b)$, and the left hand side is minimal wrt. f. Lemma 3.1(a) applies to our case, because $e = vz$ contradicts the minimality assumptions on f. For the same reason, the *rNNI* move specified by Lemma 3.1 decreases the first sum in Θ by 1.

If such an f does not exist, but $\Theta > \tau^2$, then there exists an e such that $t(e)$ is not between b_2 and ℓ. Choose the e for which $d_T(t(e), \ell) > d_T(b_2, \ell)$, and the left hand side is minimal wrt. e. We have three cases.

(i) If the undirected $t(e) \to \ell$ path in T starts on an out-arc of $t(e)$, then Lemma 3.1(b) applies to e (with the same name), because $f = uw$ contradicts the minimality assumption on e. The *rNNI* move specified by Lemma 3.1 decreases Θ by 1.

(ii) If the undirected $t(e) \to \ell$ path in T starts on an in-arc of $t(e)$, and the next arc is in opposite orientation (see the bold arcs in N in Figure 3), then we apply Lemma 3.3 to e. The conditions of the lemma are satisfied, because the sources of reticulation arcs are closer to the root than the parent of $t(e)$ in T. By the minimality assumption on e, the *rNNI* move reduces Θ by at least 1.

(iii) If the undirected $t(e) \to \ell$ path in T starts on an in-arc of $t(e)$, and the next arc is in the same orientation (see graph N' in Figure 2), then we apply the *rNNI* move of Lemma 3.1(c) to e, but with the labels of arcs e and f exchanged. Because $b_1 \neq v, w, z$ in the setup of Lemma 3.1, the conditions of the lemma are satisfied. The *rNNI* move decreases Θ by 1 (by the minimality assumption on e).

We may assume now that $\Theta_{N,T} = \tau^2$. Let e be the reticulation arc whose source is the root ρ. Via a couple of *rNNI* moves provided by Lemma 3.1(c), we may assume that $t(e)$ is the child of b_2 in T (while keeping $\Theta = \tau^2$). Change the tree base while keeping the network N intact: let $T' = T - b_2t(e) + e$. This a support tree for N, because b_2 is a branching vertex in T.

Although e is no longer a reticulation arc, $b_2t(e)$ becomes one. If $b_1 = b_2$, we have a burl-rooted tree. Otherwise, b_2 can be moved to the path between ρ and b_1 via Lemma 3.1(a). Lastly, note that the choice of leaf ℓ on the burl has been arbitrary.

4.3 Cases (B) and (C).

Suppose b_1 is a branching vertex closest to ρ in T. Let ℓ be an arbitrary leaf below b_1 in T, and let b_2 be the branching point it is joined to in T. On the branch of the root containing b_1 we may perform the procedure outlined in the previous Case (A) until Θ is reduced to its minimum (counting only sources or targets of reticulations on the branch of b_1). We have to check that reticulation arcs that join the two main branches (originating at the root) do not interfere with the previously described *rNNI* sequence. This is easily seen to be the case.

We will reduce this case to Case (A) now. Let u, v be the children of ρ such that v is on the same branch as b_1 in T. Let $t(e)$ be the target of reticulation which is the child of b_2 on T.

If $vu \in E(N)$, then we change the support tree of N to $T' = T - \rho u + vu$, and the reduction to Case (A) is done.

If $vu \notin E(N)$, then the root can be moved down along the $\rho \to t(e)$ path in T until ρ is between b_2 and $t(e)$; all we need to do is check that the conditions of Lemma 3.3 are satisfied at each step. Because a directed path cannot traverse the root and all of the targets of reticulation arcs are below b_2 in the branch of v, the conditions are satisfied. Once we have $t(e) \leftarrow \rho \to b_2$, we change the support tree to $T' = T - \rho t(e) + e$. We have completely reduced this case to Case (A).

4.4 Case (D)

Let u, v be the two children of ρ. Without loss of generality, we may assume that v or one of the vertices below it in the support tree is a target of reticulation. If $vu \in E(N)$, we can rewire the support tree through vu and reduce this case to Case (A). If there is a $u \to v$ path in N, then v must be the target of a reticulation arc e, such that $s(e)$ is below u in T (otherwise N would contain an oriented cycle). Via Lemma 3.1(a), we may assume that $e = uv$, and we can rewire the support tree through uv to reduce this case to Case (A).

If $vu \notin E(N)$ and there is no $u \to v$ path in N, we can perform the *rNNI* move described by Lemma 3.3. By repeating the argument, we may assume that v is the target of a reticulation arc, in which case we are done (as above). □

4.5 Connectedness under distance-1 tail-moves

Theorem 4.3. *TBN*(X) *is connected under distance-1 tail-moves.*

Proof. In the proof of Theorem 4.2, each *rNNI*-move performed falls into the scope of either Lemma 3.1 or Lemma 3.3. We claim that all of these *rNNI*-moves performed during the proof are either already distance-1 tail-moves, or they can be simulated with tail-moves (see Definition 2.3).

The *rNNI*-move described by Lemma 3.3 is a distance-1 tail-move if v is the tail of e. In Section 4.3, this is always the case for applications of Lemma 3.3. In Section 4.2, however, it is possible that in terms of the labeling used in Lemma 3.3, v is the head of e if z is not the root. In both of these cases the *rNNI*-move can be simulated with two distance-1 tail-moves: first, move the tail of zu onto vw, and then move the tail of zw onto the incoming arc of v which is different from e. The intermediate graph is a phylogenetic network, because the first tail-move is an *rNNI*-move to which Lemma 3.1(c) applies. Moreover, it trivially has a tree-base, because f, zu, zv, and vw are all supporting the tree-base of N.

The *rNNI*-move described by Lemma 3.1 is a distance-1 tail-move if v is the tail of e or w is the tail of f. Otherwise, if $t(e) = v$ and $t(f) = w$, the *rNNI*-move can be decomposed into three tail-moves. Let $s(e) = x$ and $s(f) = y$, so that $e = xv$ and $f = yw$. By the assumptions of Lemma 3.1(c), $y \neq u$. The arcs e and f are not supporting the tree-base of N, hence x and y are not reticulation vertices. First, move the tail of f to uv and let the new incoming arc of w be $y'w$. Secondly, move the tail x of e to the original position of y. Lastly, move the tail of $y'w$ to the original position of x. The two intermediate networks are trivially acyclic, because both the starting network N and the target network N' are acyclic and $y'w$ is the only additional arc in the two intermediate networks. Both of the intermediate networks possess a tree-base, because the arcs whose tails were moved are not contained in the support tree of the chosen tree-base.

Although the three tail-moves are generally not distance-1, they can be broken up into distance-1 tail-moves, such that the tail traverses the shortest undirected path in the support tree. Let N'' be any intermediary network along these refined steps. For any vertex x, the set of vertices that are reachable through a direct path starting from x is broader in N (or alternatively, in N') than in N''. Thus N'' is acyclic, too. □

References

[1] W Ford Doolittle. "Phylogenetic classification and the universal tree." In: *Science* 284.5423 (1999), pp. 2124–2128.

[2] Péter L Erdős, Charles Semple, and Mike Steel. "A class of phylogenetic networks reconstructable from ancestral profiles." In: *Mathematical Biosciences* 313 (2019), pp. 33–40.

[3] Mareike Fischer and Andrew Francis. "How tree-based is my network? Proximity measures for unrooted phylogenetic networks." In: *Discrete Applied Mathematics* (2020, in press).

[4] Mareike Fischer and Andrew Francis. "The space of tree-based phylogenetic networks." In: *Bulletin of Mathematical Biology* (2020, in press).

[5] Andrew Francis, Charles Semple, and Mike Steel. "New characterisations of tree-based networks and proximity measures." In: *Advances in Applied Mathematics* 93 (2018), pp. 93–107.

[6] Andrew R Francis and Mike Steel. "Which phylogenetic networks are merely trees with additional arcs?" In: *Systematic Biology* 64.5 (2015), pp. 768–777.

[7] Philippe Gambette et al. "Rearrangement moves on rooted phylogenetic networks." In: *PLOS Computational Biology* 13.8 (2017), pp. 1–21. DOI: `10.1371/journal.pcbi.1005611`.

[8] Katharina T Huber, Vincent Moulton, and Taoyang Wu. "Transforming phylogenetic networks: Moving beyond tree space." In: *Journal of Theoretical Biology* 404 (2016), pp. 30–39.

[9] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[10] Remie Janssen and Jonathan Klawitter. "Rearrangement operations on unrooted phylogenetic networks." In: *arXiv preprint arXiv:1906.04468* (2019).

[11] Remie Janssen et al. "Exploring the tiers of rooted phylogenetic network space using tail moves." In: *Bulletin of Mathematical Biology* 80.8 (2018), pp. 2177–2208.

[12] Laura Jetten and Leo van Iersel. "Nonbinary tree-based phylogenetic networks." In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2016).

[13] Jonathan Klawitter. "Spaces of phylogenetic networks." PhD thesis. University of Auckland, 2020.

[14] Eugene V Koonin. "The turbulent network dynamics of microbial evolution and the statistical tree of life." In: *Journal of molecular evolution* 80.5-6 (2015), pp. 244–250.

[15] David F Robinson. "Comparison of labeled trees with valency three." In: *Journal of Combinatorial Theory, Series B* 11.2 (1971), pp. 105–119.

[16] Charles Semple. "Phylogenetic networks with every embedded phylogenetic tree a base tree." In: *Bulletin of Mathematical Biology* 78.1 (2016), pp. 132–137.

[17] Mike Steel. *Phylogeny: Discrete and random processes in evolution*. SIAM, 2016.

[18] Louxin Zhang. "On tree-based phylogenetic networks." In: *Journal of Computational Biology* 23.7 (2016), pp. 553–565.