Admixed populations in the neighbor-joining algorithm: a geometric analysis with five taxa

Joy Z. Zhang[1], Wai Tung 'Jack' Lo[2], Michael Stillman[3†], Jaehee Kim[2*†]

[1*]Department of Applied Mathematics, Cornell University, 136 Hoy Rd, Ithaca, 14850, NY, USA.
[2]Department of Computational Biology, Cornell University, 237 Tower Rd, Ithaca, 14850, NY, USA.
[3]Department of Mathematics, Cornell University, 212 Garden Ave, Ithaca, 14850, NY, USA.

*Corresponding author(s). E-mail(s): jaehee.kim@cornell.edu;
Contributing authors: zz547@cornell.edu; wl428@cornell.edu; mes15@cornell.edu;
†These authors contributed equally to this work.

Abstract

The Neighbor-Joining (NJ) algorithm is a widely used method for constructing phylogenetic trees from genetic distances. While NJ is known to perform well with tree-like data, its behavior under admixture remains understudied. In this work, we present a geometric framework for analyzing the NJ algorithm under a linear admixture model. We focus on three key properties related to clustering order, distance, and topological path length in the resulting NJ trees involving five taxa. Our approach leverages polyhedral geometry to define NJ cones, which correspond to distinct cherry-picking orders and partition the space of dissimilarity vectors. We project dissimilarity vectors with admixture into a lower-dimensional space without admixture, defining polyhedral regions induced by NJ cones that satisfy specified properties. We compute the exact probabilities that these properties hold by directly calculating the volumes of the induced NJ cones and compare them with Monte Carlo integration and standard NJ simulation methods. Our results show that the property on clustering order is always satisfied, while the other properties are highly probable but depend on the admixture fraction. We also prove that certain induced NJ cones have zero volume, indicating that the corresponding NJ tree topologies are infeasible under admixture. We have implemented our methods as a publicly available module `NeighborJoining`

within `Macaulay2`, providing an efficient tool for analyzing NJ cones and their properties. This work provides new insights into the geometric structure inherent to the NJ algorithm in the presence of admixture, identifying the conditions under which admixture influences the resulting phylogenetic trees.

Keywords: Admixture, Neighbor-joining, Phylogenetics, Polyhedral geometry

1 Introduction

The Neighbor-Joining (NJ) algorithm [1] is a widely used distance-based method for inferring phylogenetic trees. Given a pairwise distance matrix, or equivalently a dissimilarity vector, NJ constructs an unrooted binary tree by iteratively merging pairs of nodes according to a specific criterion. When the input distances are additive [2] or nearly additive [3], NJ accurately reconstructs the underlying true tree [4, 5]. However, distances derived from empirical data often deviate from additivity, especially in cases of non-tree-like evolution, such as admixture events resulting from recent gene flow between distinct source populations. Admixed populations, which result from recent mixtures of distinct source populations, often exhibit unique behaviors in NJ trees. Empirical studies across diverse species and genetic datasets [6–12] have demonstrated that an admixed population appears as a short branch along the path connecting its two source populations in an inferred NJ tree. However, the rigorous theoretical understanding of these observed behaviors, including the exact probabilities and the specific conditions under which such patterns arise in NJ trees, remains poorly understood.

Under two-way linear admixture, Kopelman et al. [13] formally defined three key properties that characterize the clustering order, branch lengths, and topological structure of the NJ tree in the presence of admixture: (1) *antecedence of clustering*, where the admixed population clusters with one of its source populations before the two source populations cluster; (2) *intermediacy of distances*, where the distance between the admixed population and each source population is less than the distance between the two source populations; and (3) *intermediacy of topological path lengths*, where the number of edges between the admixed population and each source population is less than or equal to the number of edges separating the two source populations. Kim et al. [14] further investigated these properties through systematic simulations, estimating the approximate probabilities that a random admixed dissimilarity vector satisfies each property. They concluded that, although these properties can be violated, the presence of admixture leads to the properties being satisfied more frequently than expected by chance.

In this work, we introduce a formal geometric framework for studying the behavior of the NJ algorithm under a two-way linear admixture model, with a special focus on the case of five taxa. Unlike previous approaches that rely on empirical or simulation-based methods, our framework enables exact computation of the probabilities that NJ trees satisfy specified properties involving admixture. Our method leverages NJ cones [15, 16], each corresponding to a distinct cherry-picking order, partitioning the space of dissimilarity vectors. By projecting admixed dissimilarity

vectors onto a lower-dimensional space without admixture, we define the polyhedral structure induced by the NJ cones. The volumes of these induced NJ cones provide a direct measure of the probability that a random admixed dissimilarity vector satisfies the three specified properties. We have implemented our computational framework as a module, `NeighborJoining`, within `Macaulay2` [17] for public use. Our geometric approach extends the theoretical understanding of the NJ algorithm, offering deeper insights into the structural constraints on the NJ trees imposed by admixture.

2 Background and definitions

Table 1: **Summary of notations and definitions.** This table lists the key symbols and terms used in this work, with references to their formal definitions.

Symbol	Description	Reference
$\mathbf{D}_n$	Dissimilarity matrix	Definition 1
$\mathbf{d}^{(n)}$	Dissimilarity vector	Definition 1
$\mathbf{A}^{(n)}$	A-matrix	Definition 2
$\mathbf{R}^{(n)}$	R-matrix	Definition 3
ℓ_{ij}	Minimum path length between nodes	Definition 4
$\boldsymbol{\delta}$	Tree metric	Definition 5
$C_{\mathcal{O}}$	NJ cone	Definition 6
$\mathbf{M}_{C_{\mathcal{O}}}$	NJ cone matrix	Definition 6
ψ_σ	Assignment map	Definition 8
$\mathbf{\Pi}_\sigma$	Linear map from dissimilarity vectors to tree metrics	Definition 9
$\mathbf{U}$	Tree metric entry comparison matrix	Definition 10
$\tilde{C}_{\mathcal{O}}$	Property 2 cone	Definition 11
$\mathbf{M}_{\tilde{C}_{\mathcal{O}}}$	Property 2 cone matrix	Definition 11
ι_α	Embedding map for dissimilarity vector with admixture	Definition 12
$\pi_\alpha(C)$	Induced cone	Definition 13

2.1 Neighbor-joining algorithm

The NJ algorithm [1] is an iterative procedure for constructing a phylogenetic tree from an input dissimilarity map of samples. At each iteration of the NJ algorithm, a pair of nodes is selected and merged into a new internal node, which then replaces the original pair. This iterative process continues until an unrooted binary tree, with the samples as tips, is fully constructed. In this work, we restrict our attention to unrooted labeled binary trees, hereafter referred to simply as trees.

The pair selection process ("cherry picking") in the NJ algorithm corresponds to the formation of half-spaces, which define NJ cones [15]. Each distinct cherry-picking sequence corresponds to an NJ cone. As a result, the output of the NJ algorithm—a specific cherry-picking order for each input dissimilarity map—partitions the space of dissimilarity maps into their respective NJ cones. This section provides an overview of the NJ algorithm within the framework of polyhedral geometry.

2.1.1 Dissimilarity map

The NJ algorithm, as a distance-based method, requires a dissimilarity matrix as input. Each entry in this matrix represents the pairwise dissimilarity between samples, defined by a dissimilarity map on the sample space. This map satisfies the properties of non-negativity, identity, and symmetry, while potentially relaxing the triangle inequality, thus operating as a semi-metric.

Definition 1 (Dissimilarity matrix and dissimilarity vector) We denote the total number of initial samples by N and let $n \in [N]$ represent the number of taxa remaining at a given iteration. Here, $[N]$ denotes the set $\{1, \ldots, N\}$ for a positive integer N. The index set corresponding to these n taxa is denoted by $I^{(n)}$, where $n = |I^{(n)}|$. The *dissimilarity matrix* $\mathbf{D}_n$ for n taxa is an $n \times n$ symmetric matrix with a zero diagonal, where each off-diagonal entry represents the dissimilarity between a pair of taxa.

We define the *dissimilarity vector* $\mathbf{d}^{(n)}$ for n taxa as a column vector of length $\binom{n}{2}$, composed of the lower-triangular entries of the dissimilarity matrix $\mathbf{D}_n$, arranged in lexicographical order based on the taxon indices. The entry in the i-th row and j-th column $(i > j)$ of $\mathbf{D}_n$ is bijectively mapped to the $\left(\frac{(i-1)(i-2)}{2} + j\right)$-th position in $\mathbf{d}^{(n)}$ [16].

For example, when $n = 5$ with $I^{(n)} = [5]$,

$$\mathbf{D}_5 = \begin{bmatrix} 0 & d_{21}^{(5)} & d_{31}^{(5)} & d_{41}^{(5)} & d_{51}^{(5)} \\ d_{21}^{(5)} & 0 & d_{41}^{(5)} & d_{42}^{(5)} & d_{52}^{(5)} \\ d_{31}^{(5)} & d_{32}^{(5)} & 0 & d_{43}^{(5)} & d_{53}^{(5)} \\ d_{41}^{(5)} & d_{42}^{(5)} & d_{43}^{(5)} & 0 & d_{54}^{(5)} \\ d_{51}^{(5)} & d_{52}^{(5)} & d_{53}^{(5)} & d_{54}^{(5)} & 0 \end{bmatrix},$$

and

$$\mathbf{d}^{(5)} = \left[d_{21}^{(5)},\, d_{31}^{(5)},\, d_{32}^{(5)},\, d_{41}^{(5)},\, d_{42}^{(5)},\, d_{43}^{(5)},\, d_{51}^{(5)},\, d_{52}^{(5)},\, d_{53}^{(5)},\, d_{54}^{(5)}\right]^{\mathsf{T}}.$$

2.1.2 The Q-criterion

At each step of the NJ algorithm, a pair of taxa is selected based on the Q-criterion, a linear transformation mapping each entry $d_{ab}^{(n)}$ of the dissimilarity vector to its corresponding Q-value q_{ab}, as defined below.

$$q_{ab} = (n-2)d_{ab}^{(n)} - \sum_{k \in I^{(n)}} d_{ak}^{(n)} - \sum_{k \in I^{(n)}} d_{kb}^{(n)}.$$

This linear transformation can be represented by a $\binom{n}{2} \times \binom{n}{2}$ matrix, which we refer to as the A-matrix.

Definition 2 (A-matrix) Let the indices i and j correspond to the pairs of taxa (a, b) and (c, d), respectively. The *A-matrix* $\mathbf{A}^{(n)}$ is defined as a $\binom{n}{2} \times \binom{n}{2}$ matrix, where each entry

$A_{ij}^{(n)}$ is given by:

$$A_{ij}^{(n)} = \begin{cases} n-4 & i = j, \\ -1 & i \neq j \text{ and } \{a,b\} \cap \{c,d\} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

For example, when $n = 5$,

$$\mathbf{A}^{(5)} = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 21 & 31 & 32 & 41 & 42 & 43 & 51 & 52 & 53 & 54 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 0 & -1 & -1 & 0 & 0 \\ -1 & 1 & -1 & -1 & 0 & -1 & -1 & 0 & -1 & 0 \\ -1 & -1 & 1 & 0 & -1 & -1 & 0 & -1 & -1 & 0 \\ -1 & -1 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & -1 \\ -1 & 0 & -1 & -1 & 1 & -1 & 0 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & -1 & 0 & 0 & 1 & -1 & -1 & -1 \\ -1 & 0 & -1 & 0 & -1 & 0 & -1 & 1 & -1 & -1 \\ 0 & -1 & -1 & 0 & 0 & -1 & -1 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix} \end{array}. \quad (1)$$

We define the Q-vector $\mathbf{q}^{(n)}$ as $\mathbf{q}^{(n)} = \mathbf{A}^{(n)}\mathbf{d}^{(n)}$. The pair of taxa to be merged in the current iteration corresponds to the index of the Q-vector with the minimum value: $\underset{i \in [\binom{n}{2}]}{\operatorname{argmin}} q_i$.

2.1.3 Updating the dissimilarity vector and tree construction

After selecting a pair of taxa for merging, a new node representing the pair is created. The newly created nodes are indexed as $N+1, N+2, \ldots, 2N-2$, with node $N+i$ introduced in the i-th iteration. The original pair is removed from the set of taxa in the subsequent iteration, decreasing the number of taxa to be processed by one. The dissimilarity vector $\mathbf{d}^{(n)}$ is then updated to $\mathbf{d}^{(n-1)}$ by removing the entries associated with the original pair and introducing new entries that represent the dissimilarities between the newly created node and all remaining taxa.

Formally, let $a, b \in I^{(n)}$ be the pair of taxa selected at a given step, with u representing the newly created node for the pair. Then, $I^{(n-1)} = \{u\} \cup I^{(n)} \setminus \{a, b\}$, and the entries of the updated dissimilarity vector $\mathbf{d}^{(n-1)}$ are defined as follows:

$$d_{ux}^{(n-1)} = \frac{1}{2}(d_{ax}^{(n)} + d_{bx}^{(n)} - d_{ab}^{(n)}), \quad (2)$$

$$d_{xy}^{(n-1)} = d_{xy}^{(n)}, \quad (3)$$

where $x, y \in I^{(n)} \setminus \{a, b\}$. This process of updating the dissimilarity vector at each step of the algorithm can be formalized as a linear transformation [16].

Definition 3 (R-matrix) We define the *R-matrix* as a $\binom{n-1}{2} \times \binom{n}{2}$ matrix $\mathbf{R}^{(n)}$, such that $\mathbf{d}^{(n-1)} = \mathbf{R}^{(n)}\mathbf{d}^{(n)}$. Let k_{cd} and k_{ef} denote the positions of the pairs (c, d) and (e, f), respectively, in the lexicographically ordered set $\{(x, y) \mid x, y \in [5], x > y\}$. The entry $R^{(n)}_{k_{cd}k_{ef}}$ in the k_{cd} row and the k_{ef} column is defined as:

$$R^{(n)}_{k_{cd}k_{ef}} = \begin{cases} 1 & \{c,d\} = \{e,f\} \text{ and } \{e,f\} \cap \{a,b\} = \emptyset, \\ \frac{1}{2} & \text{if } |\{c,d\} \cap [N]| = 1 \text{ with } \gamma = \{c,d\} \cap [N] \text{ and if } \{e,f\} = \gamma \cup \{\lambda\} \text{ with } \lambda \in \{a,b\}, \\ -\frac{1}{2} & \{c,d\} = \{a,b\} \text{ and } |\{e,f\} \cap \{a,b\}| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

For example, if we start with $n = 5$, and taxa 2 and 3 are selected in the first iteration, the resulting dissimilarity vector $\mathbf{d}^{(4)}$ is:

$$\mathbf{d}^{(4)} = \left[d^{(4)}_{41},\, d^{(4)}_{51},\, d^{(4)}_{54},\, d^{(4)}_{61},\, d^{(4)}_{64},\, d^{(4)}_{65}\right]^{\mathsf{T}},$$

and the matrix $\mathbf{R}^{(5)}$ that updates the dissimilarity vector from $\mathbf{d}^{(5)}$ to $\mathbf{d}^{(4)}$ is given by:

$$\mathbf{R}^{(5)} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

After merging taxa a and b into a newly created node u, the node pairs (a, u) and (b, u) define edges e_{au} and e_{bu}, respectively, on the inferred NJ tree, with their lengths β_{au} and β_{bu} given by:

$$\beta_{au} = \frac{1}{2}d^{(n)}_{ab} + \frac{1}{2(n-2)} \sum_{k \in I^{(n)}} (d^{(n)}_{ak} - d^{(n)}_{bk}), \tag{4}$$

$$\beta_{bu} = d^{(n)}_{ab} - \beta_{au}. \tag{5}$$

The final NJ tree with N leaves is obtained after $N - 2$ iterations. To quantify distances within this tree, we introduce the tree metric (or additive metric) [2, 5] that defines the pairwise distances between the leaves.

Definition 4 (Minimum path length between nodes) Let $V = [2N - 2]$ denote the set of nodes in the final NJ tree. We define the function $\ell : V \times V \to \mathbb{R}_{\geq 0}$ as the map that assigns to each pair of nodes in the tree the *minimum path length* between them. Specifically, for any two nodes $i, j \in V$, $\ell(i, j)$—denoted by ℓ_{ij}—represents the sum of edge lengths along the shortest path connecting i and j within the tree. If i and j are adjacent, ℓ_{ij} is precisely β_{ij}, the length of the edge directly connecting them.

Definition 5 (Tree metric) By restricting ℓ to the set of leaves $[N]$, we define the *tree metric* $\boldsymbol{\delta}$ as a column vector of length $\binom{N}{2}$, where each entry δ_{ij} corresponds to the minimum path length ℓ_{ij} between leaves $i, j \in [N]$ of the tree.

2.1.4 NJ cones

In Sections 2.1.2 and 2.1.3, we formalized the NJ algorithm as a sequence of linear transformations applied to the input dissimilarity vector, with the cherry-picking order determined by the Q-criterion, represented as a set of linear inequalities at each iteration. This process defines a geometric structure known as NJ cones [15], where each cone corresponds to a unique NJ tree, thereby partitioning the space of input dissimilarity vectors. This section provides a mathematical description of this structure.

Let a and b be the pair of taxa selected at a given step with n taxa remaining, and let $i \in \left[\binom{n}{2}\right]$ denote the index corresponding to the position of the pair (a, b) within $\mathbf{d}^{(n)}$. Denote by $\mathbf{e}_k^{(n)}$ the k-th standard basis column vector in $\mathbb{R}^{\binom{n}{2}}$, with a 1 in the k-th position and 0 in all other positions. Then,

$$(a, b) = i = \underset{k \in \left[\binom{n}{2}\right]}{\operatorname{argmin}} \{q_k\} = \underset{k \in \left[\binom{n}{2}\right]}{\operatorname{argmin}} (\mathbf{e}_k^{(n)\mathsf{T}} \mathbf{A}^{(n)} \mathbf{d}^{(n)}) = \underset{k \in \left[\binom{n}{2}\right]}{\operatorname{argmin}} (\mathbf{A}^{(n)} \mathbf{e}_k^{(n)} \cdot \mathbf{d}^{(n)}),$$

where the final step follows directly from the symmetry of the A-matrix.

Since i corresponds to the index of the minimum Q-criterion value, the following inequality holds for all $j \in \left[\binom{n}{2}\right]$:

$$\mathbf{A}^{(n)} \mathbf{e}_i^{(n)} \cdot \mathbf{d}^{(n)} \leq \mathbf{A}^{(n)} \mathbf{e}_j^{(n)} \cdot \mathbf{d}^{(n)}. \tag{6}$$

For given i, and for each $j \in I^{(n)}$, Eq. 6 defines a half-space in the space of input dissimilarity vectors $\mathbb{R}_{\geq 0}^{\binom{n}{2}}$:

$$H_{ij} = \left\{ \mathbf{d}^{(n)} \in \mathbb{R}_{\geq 0}^{\binom{n}{2}} \;\middle|\; \mathbf{A}^{(n)} \left(\mathbf{e}_i^{(n)} - \mathbf{e}_j^{(n)} \right) \cdot \mathbf{d}^{(n)} \leq 0 \right\}. \tag{7}$$

If an input dissimilarity vector $\mathbf{d}^{(n)}$ satisfies the inequality in Eq. 7, it lies within the corresponding half-space H_{ij}. When $i = j$, the inequality holds trivially, resulting in $\binom{n}{2} - 1$ non-trivial half-spaces for each cherry-picking among n nodes. Thus, the i-th pair is selected if and only if $\mathbf{d}^{(n)}$ lies within the intersection of all non-trivial half-spaces generated in that step, formally expressed as

$$\mathbf{d}^{(n)} \in \bigcap_{j \neq i} H_{ij}.$$

The total number of half-spaces required to reconstruct a tree with N samples is obtained by summing the number of half-spaces generated at each step of the algorithm: $\sum_{k=4}^{N} \left(\binom{k}{2} - 1 \right)$. Note that at the iteration where $n = 3$, all entries of the Q-vector are identical because $\mathbf{A}^{(3)}$ has rank 1 by construction:

$$\mathbf{A}^{(3)} = \begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{bmatrix}. \tag{8}$$

Therefore, only the iterations preceding $n = 3$ contribute to the total set of half-spaces. We now formally define the NJ cone using these half-planes.

Definition 6 (NJ cone) Let $\mathcal{O} = (o_1, o_2, \ldots, o_{N-3})$ represent a cherry-picking order on a set of N taxa, specifying the sequential selection of $N-3$ cherries, with o_k denoting the k-th cherry. At each iteration k, define $n_k = N - k + 1$ as the number of taxa remaining at the beginning of the iteration, and let $i_k \in \left[\binom{n_k}{2}\right]$ be the index of o_k in the corresponding Q-vector. For each $j_k \in \left[\binom{n_k}{2}\right]$ with $j_k \neq i_k$, define the row vector $\mathbf{h}_{i_k j_k}$ of length $\binom{N}{2}$, constructed from the comparison of the Q-values q_{i_k} and q_{j_k} as follows:

$$\mathbf{h}_{i_k j_k} = \left(\mathbf{e}_{i_k}^{(n_k)} - \mathbf{e}_{j_k}^{(n_k)} \right)^{\mathsf{T}} \mathbf{A}^{(n_k)} \left(\mathbf{R}^{(n_k+1)} \cdots \mathbf{R}^{(N)} \right).$$

At iteration k, there are $\binom{n_k}{2} - 1$ hyperplanes, defined by the vectors $\mathbf{h}_{i_k j_k}$, for each $j_k \in \left[\binom{n_k}{2}\right] \setminus \{i_k\}$. We define the reindexing map $\phi_k : \{i_k\} \times \left[\binom{n_k}{2}\right] \setminus \{i_k\} \to \{k\} \times \left[\binom{n_k}{2} - 1\right]$, which reassigns the indices i_k and j_k as follows:

$$\phi_k(i_k, j_k) = \begin{cases} (k, j_k) & \text{if } j_k < i_k, \\ (k, j_k - 1) & \text{if } j_k > i_k. \end{cases} \tag{9}$$

This reindexes each hyperplane $\mathbf{h}_{i_k j_k}$ to $\mathbf{h}_{km}$, where $m \in \left[\binom{n_k}{2} - 1\right]$.

We define the *NJ cone matrix* $\mathbf{M}_{C_\mathcal{O}}$ as a $\left(\sum_{x=4}^{N} \left(\binom{x}{2} - 1 \right) \right) \times \binom{N}{2}$ matrix, where rows correspond to hyperplanes associated with the cherry-picking order $\mathcal{O}$:

$$\mathbf{M}_{C_\mathcal{O}} = \left[\mathbf{h}_{11}, \ldots, \mathbf{h}_{1(\binom{N}{2}-1)}, \mathbf{h}_{21}, \ldots, \mathbf{h}_{2(\binom{N-1}{2}-1)}, \ldots, \mathbf{h}_{(N-3)1}, \ldots, \mathbf{h}_{(N-3)5} \right]^{\mathsf{T}}.$$

The *NJ cone* $C_\mathcal{O}$ associated with the cherry-picking order $\mathcal{O}$ is then defined as:

$$C_\mathcal{O} = \left\{ \mathbf{d}^{(N)} \in \mathbb{R}_{\geq 0}^{\binom{N}{2}} \;\middle|\; \mathbf{M}_{C_\mathcal{O}} \mathbf{d}^{(N)} \leq 0 \right\}.$$

In other words, the NJ cone $C_\mathcal{O}$ represents the region in the space of dissimilarity vectors where the NJ algorithm, following the order $\mathcal{O}$, produces a unique unrooted labeled binary tree topology, distinguishing reflection-symmetric topologies by assigning them to distinct cones (Table 2).

Each input dissimilarity vector $\mathbf{d}^{(N)}$ either lies strictly within the interior of a single NJ cone or on a boundary shared by multiple NJ cones. The latter occurs if and only if there exists an NJ cone $C_\mathcal{O}$ and a cherry-picking iteration k such that, at the k-th iteration, a row vector $\mathbf{h}_{i_k j_k}$ in $\mathbf{M}_{C_\mathcal{O}}$ satisfies:

$$\mathbf{h}_{i_k j_k}^{\mathsf{T}} \cdot \mathbf{d}^{(N)} = 0. \tag{10}$$

Eq. 10 implies that the entries q_{i_k} and q_{j_k} of the Q-vector at the k-th iteration are both the minimum among all entries of the Q-vector at that iteration. So the NJ algorithm cannot distinguish picking the cherry i_k or the cherry j_k at the k-th iteration. Let m_k denote the number of cherries whose Q-vector entries have the same minimum value at the k-th iteration, i.e., there are m_k pairs of nodes satisfying Eq. 10. The cherry-picking order for $\mathbf{d}^{(N)}$ can then select any of these m_k cherries, as their Q-vector entries are all minimum.

Since each iteration determines a single cherry, and there are $N-3$ iterations that define the half-spaces bounding an NJ cone, the total number of distinct NJ cones containing $\mathbf{d}^{(N)}$ on the boundary is given by multiplying the number of equivalent choices at each iteration, adjusting for the three pairs of identical rows in the A-matrix at the $(N-3)$-th iteration (Eq. 8):

$$(m_{N-3}-2)\prod_{k=1}^{N-4} m_k.$$

For example, for $N=5$, if $\mathbf{d}^{(5)}$ has $m_1=5$ and $m_2=3$, it lies on the boundary shared by $(m_2-2)m_1=(3-2)\times(5-0)=5$ distinct NJ cones.

2.1.5 The equivalence relation among cherry-picking orders

While each cherry-picking order belongs to an NJ cone, a single NJ cone can correspond to multiple cherry-picking orders due to the structural properties of the A-matrix (Definition 2). If two rows of $\mathbf{A}^{(n)}$ are identical, the corresponding Q-vector entries are equal for any input dissimilarity vector (Section 2.1.2). In this case, the NJ algorithm arbitrarily selects a cherry from node pairs with the minimum Q-value. This multiplicity induces an equivalence class of cherry-picking orders. The proof of this equivalence relation is straightforward and is therefore omitted.

Definition 7 (Equivalence class of cherry-picking orders) Two cherry-picking orders are *equivalent* if and only if, at every iteration of the NJ algorithm, the corresponding rows of the A-matrix are identical.

For $N=5$, the set of cherry-picking orders forms thirty distinct equivalence classes, with each class containing six orders [15]. This results from the combinatorial properties of tree topologies under the NJ algorithm. There is one unrooted unlabeled binary tree topology, which can be labeled to form $(2N-5)!!=15$ unrooted labeled binary tree topologies. However, the cherry-picking order introduces asymmetry relative to the central edge, thereby doubling the total number of unique NJ cones to thirty. Thus, each NJ cone corresponds to a unique unrooted labeled binary tree topology that accounts for asymmetry with respect to the central node, resulting in a total of thirty distinct NJ cones (Table 2).

The equivalence of six cherry-picking orders per class is due to the ties in the Q-criterion, as defined by the matrices $\mathbf{A}^{(4)}$ and $\mathbf{A}^{(3)}$. At the first iteration, the Q-vector is given by $\mathbf{q}^{(5)}=\mathbf{A}^{(5)}\mathbf{d}^{(5)}$. Since the rows of $\mathbf{A}^{(5)}$ are distinct (Eq. 1), no two

cherries are guaranteed to have the same Q-value at this step. Assuming, without loss of generality, that nodes 4 and 5 are selected as a cherry in the first step, with node 6 added, the Q-vector at the second step is computed as $\mathbf{q}^{(4)} = \mathbf{A}^{(4)}\mathbf{d}^{(4)}$, where $\mathbf{A}^{(4)}$ is given by:

$$\mathbf{A}^{(4)} = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 61 \\ 62 \\ 63 \end{array} \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 61 & 62 & 63 \end{array} \\ \begin{bmatrix} 0 & -1 & -1 & -1 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & -1 & 0 \end{bmatrix} \end{array}. \tag{11}$$

Regardless of the cherry selected in the first step, $\mathbf{A}^{(4)}$ contains three pairs of identical rows, leading to ties in the Q-values. In the third and final cherry-picking step, $\mathbf{A}^{(3)}$ has identical rows (Eq. 8), resulting in ties across all node pairs. These ties in both the second and third steps establish the equivalence of six cherry-picking orders.

This equivalence allows for a systematic representation of each NJ cone for $N = 5$ based on the first two cherries. Since each NJ cone corresponds to six equivalent cherry-picking orders, any of these orders can be selected to represent the cone. Among these six, three have their first two cherries composed solely of the original taxa indices $\{1, \ldots, 5\}$. We represent the cone by the cherry-picking order where these cherries consist exclusively of the original indices. Let (a, b) and (c, d) be the first two cherries of an NJ cone, where $a, b, c, d \in [5]$. The NJ cone is then denoted by $C_{(a,b)(c,d)}$. This notation will be used to represent NJ cones with $N = 5$ throughout the remainder of this work.

2.2 Populations with admixture

Consider a set of N populations, indexed by $I^{(N)} = [N]$ and labeled $\{t_1, t_2, \ldots, t_N\}$. Let t_N represent an admixed population from a two-way admixture between source populations t_1 and t_2. Define the admixture fraction $\alpha \in (0, 1)$ as the proportion of contribution from source population t_1 to the admixed population t_N, with the remaining $1 - \alpha$ representing the contribution from the other source population t_2. We employ the linear admixture model as described in Kopelman et al. [13] and Kim et al. [14], where the pairwise genetic distances between the admixed population t_N and other populations are expressed as a linear combination of the distances involving the source populations. For each $i \in [N]$, these distances are given by:

$$d_{Ni}^{(N)} = \alpha d_{1i}^{(N)} + (1 - \alpha) d_{2i}^{(N)}. \tag{12}$$

The corresponding dissimilarity matrix $\mathbf{D}^{(N)}$, incorporating the admixed population t_N is given by:

$\mathbf{D}^{(N)} =$

$$\begin{bmatrix} 0 & d_{21}^{(N-1)} & \cdots & d_{(N-1)1}^{(N-1)} & (1-\alpha)d_{21}^{(N-1)} \\ d_{21}^{(N-1)} & 0 & \cdots & d_{(N-1)2}^{(N-1)} & \alpha d_{21}^{(N-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{(N-1)1}^{(N-1)} & d_{(N-1)2}^{(N-1)} & \cdots & 0 & \alpha d_{(N-1)1}^{(N-1)} + (1-\alpha)d_{(N-1)2}^{(N-1)} \\ (1-\alpha)d_{21}^{(N-1)} & \alpha d_{21}^{(N-1)} & \cdots & \alpha d_{(N-1)1}^{(N-1)} + (1-\alpha)d_{(N-1)2}^{(N-1)} & 0 \end{bmatrix}. \tag{13}$$

We assume all off-diagonal elements in $\mathbf{D}^{(N)}$ are strictly positive.

2.2.1 Property 1: antecedence of clustering

In the NJ algorithm, Property 1 is satisfied if the admixed taxon clusters with one of the source taxa before the two source taxa are clustered together. Formally, let Γ_i denote the clade containing taxon i, and (Γ_i, Γ_j) represent the agglomeration of two clades. The clustering order satisfies Property 1 in the last two of the following three cases:

$$((\Gamma_1, \Gamma_2), \Gamma_N), \quad ((\Gamma_1, \Gamma_N), \Gamma_2), \quad ((\Gamma_2, \Gamma_N), \Gamma_1).$$

For example, when $N = 5$, Property 1 is satisfied if source taxon 1 and admixed taxon 5 are clustered in the first step, forming a new node 6. In the second step, node 6 clusters with taxon 2, such that the admixed taxon 5 is clustered with source taxon 1 before the source taxa 1 and 2 are joined. Conversely, Property 1 is violated if source taxa 1 and 2 are clustered in the first step.

2.2.2 Property 2: intermediacy of distances

Property 2 states that the distances on the inferred NJ tree (the tree metric; Definition 5) between the admixed taxon and each source taxon are less than or equal to the distance between the two source taxa. Since we are working within the framework of closed polyhedra, the strict inequality "$<$" from Kopelman et al. [13] is replaced by "$\leq$". This adjustment does not affect the probability of satisfying or violating the property, since the boundary of a polyhedron has measure zero. Thus, the property requires:

$$\delta_{1N} \leq \delta_{12}, \qquad \delta_{2N} \leq \delta_{12}.$$

2.2.3 Property 3: intermediacy of topological path lengths

Unlike Property 2, which focuses on branch lengths, Property 3 concerns the topological structure of the NJ tree, which is determined by the cherry-picking order. This property requires that the number of edges along the shortest path between the admixed taxon and each source taxon be less than or equal to the number of edges along the shortest path between the two source taxa. Formally, let τ_{ij} denote the number of edges on the shortest path between taxa i and j. Property 3 is then expressed as:

$$\tau_{1N} \leq \tau_{12}, \qquad \tau_{2N} \leq \tau_{12}.$$

3 Methods

Our objective is to compute the probability that a random dissimilarity vector with admixture produces an NJ tree satisfying the three properties defined in Section 2.2. Given that NJ cones partition the space of dissimilarity vectors (Section 2.1.4), associating each NJ cone with these properties enables the identification of dissimilarity vectors that satisfy them. In this section, we present a theoretical framework for computing these probabilities based on the volumes of NJ cones, employing two approaches: (1) Monte Carlo integration and (2) direct calculation. While the methods are demonstrated for $N = 5$, they can be generalizable to cases where $N > 5$.

3.1 Classification of NJ cones based on the three properties

Properties 1 and 3 are topological, determined solely by the cherry-picking order associated with the input dissimilarity vector. In contrast, Property 2 is not necessarily satisfied across an entire NJ cone, as it further depends on the branch lengths of the final NJ tree. In this section, we identify the NJ cones whose equivalence classes of cherry-picking orders satisfy Property 1 and, separately, those that satisfy Property 3. We also construct the cones associated with Property 2 using a linear transformation that maps input dissimilarity vectors to the corresponding tree metrics in the final NJ tree.

3.1.1 NJ cones satisfying Property 1 (antecedence of clustering)

We first identify NJ cones satisfying Property 1 for $N = 5$. Property 1 holds if the admixed taxon clusters with either source taxon before the source taxa cluster with each other. Within an equivalence class of an NJ cone, some cherry-picking orders may satisfy Property 1, while others may not. We classify NJ cones into three categories based on the number of cherry-picking orders in their equivalence class that satisfy Property 1. Since ties in the Q-criterion define these equivalence classes, we deem an NJ cone to satisfy Property 1 if at least one cherry-picking order within its equivalence class satisfies it.

Type 1: NJ cones whose equivalence class consists entirely of cherry-picking orders that satisfy Property 1.

If the first cherry in a cherry-picking order is either $(5, 1)$ or $(5, 2)$, where one taxon is the source and the other the admixed taxon, the corresponding NJ cone satisfies Property 1 and is classified as Type 1. In this case, the source and the admixed taxa form a new node, which is subsequently clustered with the remaining source taxon. Thus, all cherry-picking orders within this equivalence class have this property. For example, the NJ cone corresponding to the NJ tree in Figure 1 is classified as a Type-1 cone. All six cherry-picking orders associated with this NJ cone satisfy Property 1: $(5, 1)(4, 2)(7, 6)$, $(5, 1)(4, 2)(7, 3)$, $(5, 1)(4, 2)(6, 3)$, $(5, 1)(6, 3)(7, 4)$, $(5, 1)(6, 3)(7, 2)$, and $(5, 1)(6, 3)(4, 2)$. The complete set of NJ trees corresponding to Type-1 cones is provided in Figure B1.

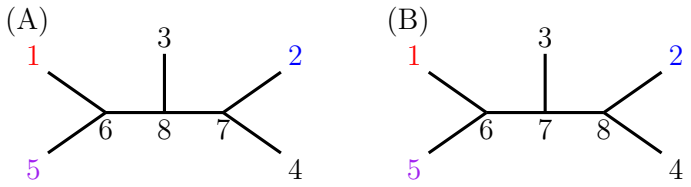


Fig. 1: **Example NJ tree corresponding to a Type-1 NJ cone satisfying Property 1.** The two trees shown are the same NJ tree but differ only in their internal node labels, resulting from different cherry-picking orders within the same equivalence class. **(A)** Labeled tree topology, with internal node labeled, from the cherry-picking orders: $(5,1)(4,2)(7,6)$, $(5,1)(4,2)(7,3)$, and $(5,1)(4,2)(6,3)$. **(B)** Labeled tree topology, with internal node labeled, from the cherry-picking orders $(5,1)(6,3)(7,4)$, $(5,1)(6,3)(7,2)$, and $(5,1)(6,3)(4,2)$.

Type 2: NJ cones whose equivalence class contains at least one, but not all, cherry-picking orders that violate Property 1.

If an equivalence class of an NJ cone includes at least one cherry-picking order that satisfies Property 1, but not all, we classify the NJ cone as Type 2 and satisfying Property 1. This convention is adopted because the equivalence class arises from ties in the Q-criterion, where any tied pairs can be selected randomly. Thus, if any cherry-picking order within the equivalence class satisfies Property 1, we designate the entire NJ cone as satisfying Property 1. For example, the NJ cone corresponding to the NJ tree in Figure 2 is a Type-2 cone. The cherry-picking order, $(3,1)(4,2)(7,6)$ violates Property 1, whereas the other cherry-picking orders in the same equivalence class—$(3,1)(4,2)(6,5)$, $(3,1)(4,2)(7,5)$, $(3,1)(6,5)(7,4)$, $(3,1)(6,5)(7,2)$, and $(3,1)(6,5)(4,2)$—satisfy it. The complete set of NJ trees corresponding to Type-2 cones is listed in Figure B2.

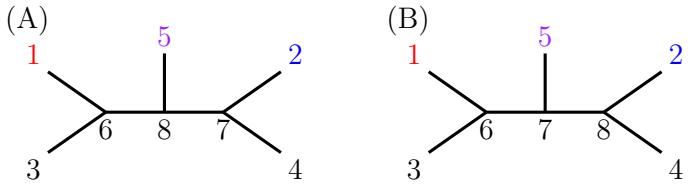


Fig. 2: **Example NJ tree corresponding to a Type-2 NJ cone satisfying Property 1.** The two trees shown are the same NJ tree but differ only in their internal node labels, resulting from different cherry-picking orders within the same equivalence class. **(A)** Labeled tree topology, with internal node labeled, from the cherry-picking orders: $(3,1)(4,2)(7,6)$, $(3,1)(4,2)(6,5)$, and $(3,1)(4,2)(7,5)$. **(B)** Labeled tree topology, with internal node labeled, from the cherry-picking orders: $(3,1)(6,5)(7,4)$, $(3,1)(6,5)(7,2)$, and $(3,1)(6,5)(4,2)$.

Type 3: NJ cones whose equivalence class consists entirely of cherry-picking orders that violate Property 1.

If the first cherry in a cherry-picking order consists of the source taxa, $(2, 1)$, the corresponding NJ cone is classified as Type 3 and necessarily violates Property 1. This violation occurs because all cherry-picking orders in this equivalence class start with $(1, 2)$ as the first cherry, and the admixed taxon 5 can only join $(2, 1)$ after these two source taxa have already been clustered. For example, the NJ cone corresponding to the NJ tree in Figure 3 is classified as a Type-3 cone. All six cherry-picking orders associated with this NJ cone violate Property 1: $(2, 1)(6, 5)(7, 3)$, $(2, 1)(6, 5)(7, 4)$, $(2, 1)(6, 5)(4, 3)$, $(2, 1)(4, 3)(6, 5)$, $(2, 1)(4, 3)(7, 6)$, and $(2, 1)(4, 3)(7, 5)$. The complete set of NJ trees corresponding to Type-3 cones is provided in Figure B3.

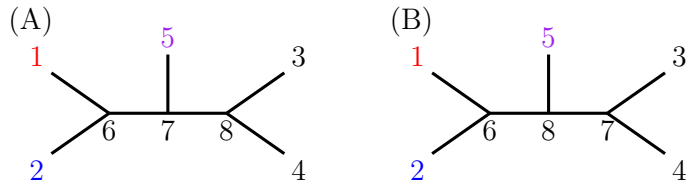


Fig. 3: **Example NJ tree corresponding to a Type-3 NJ cone violating Property 1.** The two trees shown are the same NJ tree but differ only in their internal node labels, resulting from different cherry-picking orders within the same equivalence class. **(A)** Labeled tree topology, with internal node labeled, from the cherry-picking orders: $(2, 1)(6, 5)(7, 3)$, $(2, 1)(6, 5)(7, 4)$, and $(2, 1)(6, 5)(4, 3)$. **(B)** Labeled tree topology, with internal node labeled, from the cherry-picking orders: $(2, 1)(4, 3)(6, 5)$, $(2, 1)(4, 3)(7, 6)$, and $(2, 1)(4, 3)(7, 5)$.

3.1.2 NJ cones satisfying Property 3 (intermediacy of path length)

Property 3 applies solely to the labeled tree topology of the NJ tree. Since cherry-picking orders determine the labeled tree topology of the final NJ tree, an input dissimilarity vector satisfies this property if and only if it is contained within an NJ cone whose corresponding labeled tree topology satisfies Property 3. Given that all dissimilarity vectors within an NJ cone result in the same labeled tree topology, distinguishing reflection symmetry, either all vectors in the cone satisfy Property 3 or none do. There are sixteen NJ cones whose labeled tree topologies satisfy Property 3. The complete set of trees corresponding to the NJ cones that do not satisfy Property 3 is presented in Figure B4.

3.1.3 Cones satisfying Property 2 (intermediacy of distances)

An NJ cone can contain both dissimilarity vectors that satisfy Property 2 and those that do not. Therefore, an NJ cone cannot be strictly categorized as either fully satisfying or not satisfying Property 2. Instead, we identify the subset of each NJ cone that satisfies Property 2 and compute its associated volume. To identify the subset of an NJ cone composed exclusively of dissimilarity vectors satisfying Property 2, we must map

each input dissimilarity vector $\mathbf{d}^{(5)}$ to its corresponding tree metric $\boldsymbol{\delta}$, as Property 2 applies to the tree metric of the resulting NJ tree. In this section, we show that there exists a linear transformation from an input dissimilarity vector to its corresponding tree metric and identify cones associated with Property 2.

We first show that, for $N = 5$, the dissimilarity vector $\mathbf{d}^{(3)}$, obtained after the second cherry-picking step, equals the vector of the minimum path lengths between the two remaining internal nodes and the remaining leaf node. Recall that every cherry-picking order is equivalent to one where the first two cherries consist of the original taxa $\{1, \ldots, 5\}$ (Section 2.1.5). Thus, we can assume the first cherry corresponds to the leaves on the left side of the final NJ tree, and the second cherry to the right. Under this assumption, the remaining nodes to be clustered are the two internal nodes $\{6, 7\}$—formed after selecting the first two cherries—and the remaining taxon from the original set.

Lemma 1 Given $N = 5$, let $\mathbf{d}^{(3)}$ be the dissimilarity vector obtained after the second cherry-picking step. Denote the unpaired original taxon as e, and the two newly created nodes as u and v, respectively. The NJ tree restricted to these three nodes is shown in Figure 4. Then, the following holds:

$$d^{(3)}_{eu} = \ell_{eu}, \qquad d^{(3)}_{ev} = \ell_{ev}, \qquad d^{(3)}_{uv} = \ell_{uv}.$$

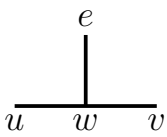


Fig. 4: **NJ tree restricted to the final three nodes for $N = 5$.** Without loss of generality, for $N = 5$, the first two cherries are assumed to be on opposite sides of the NJ tree. The figure shows the remaining three nodes to be clustered after the first two cherry-picking steps.

Proof. With three taxa remaining, all the rows of $A^{(3)}$ are identical (Section 2.1.5), and thus, any pair from the remaining cherries can be selected in the third cherry-picking step. Without loss of generality, suppose the cherry chosen at the third cherry-picking

step is (e, u). Then by Eqs. 2, 4, and 5, we have

$$\begin{aligned}
\beta_{ew} &= \frac{1}{2}\left(d^{(3)}_{eu} + d^{(3)}_{ev} - d^{(3)}_{uv}\right), \\
\beta_{uw} &= \frac{1}{2}\left(d^{(3)}_{eu} - d^{(3)}_{ev} + d^{(3)}_{uv}\right), \\
\beta_{vw} &= \frac{1}{2}\left(d^{(3)}_{ev} + d^{(3)}_{uv} - d^{(3)}_{eu}\right), \\
\ell_{eu} &= \beta_{ew} + \beta_{uw} = d^{(3)}_{eu}, \\
\ell_{ev} &= \beta_{ew} + \beta_{vw} = d^{(3)}_{ev}, \\
\ell_{uv} &= \beta_{uw} + \beta_{vw} = d^{(3)}_{uv}.
\end{aligned} \tag{14}$$

The result also holds when the cherry chosen at the third iteration is either (e, v) or (u, v); the same proof applies by permuting the node labels $\{e, u, v\}$ accordingly. □

We next construct a matrix that maps an input dissimilarity vector to its corresponding tree metric.

Definition 8 (Assignment map) Let $\sigma \in S_5$ be a permutation, where S_5 is the symmetric group on five elements. Consider the unrooted binary tree structure in Figure 5A, where the set of leaves $\{a, b, c, d, e\}$ represents the taxa, and $\{1, 2, 3, 4, 5\}$ is the index set. Define the *initial assignment* $\psi_0 : \{a, b, c, d, e\} \to [5]$, where each taxon is mapped to its corresponding index as follows:

$$\psi_0(a) = 1, \quad \psi_0(b) = 2, \quad \psi_0(c) = 3, \quad \psi_0(d) = 4, \quad \psi_0(e) = 5.$$

An *σ-assignment* is defined as a permutation-induced reassignment of taxa on the tree leaves via $\sigma : [5] \to [5]$. The function $\psi_\sigma = \sigma \circ \psi_0$ describes this reassignment according to σ, such that for each $i \in \{a, b, c, d, e\}$, $\psi_\sigma(i) = \sigma(\psi_0(i))$.

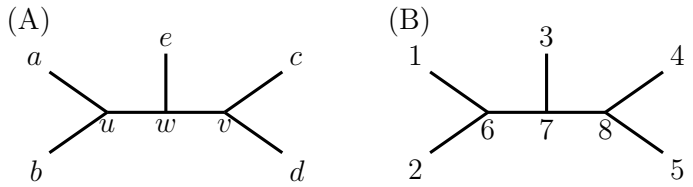


Fig. 5: **Node labels for an unrooted binary tree for five leaves and their initial assignment. (A)** The NJ tree for $N = 5$ taxa is shown with leaf labels $\{a, b, c, d, e\}$ and internal nodes $\{u, w, v\}$. The leaves $\{a, b, c, d, e\}$ correspond to the taxa indexed by $\{1, 2, 3, 4, 5\}$. The tree topology remains fixed, while the indices corresponding to each leaf label are permuted according to the σ-assignments. **(B)** The leaf labels $\{a, b, c, d, e\}$ are mapped to the taxa indices $\{1, 2, 3, 4, 5\}$ via the initial assignment ψ_0 (Definition 8). The internal nodes u, v, and w are labeled as 6, 7, and 8, respectively, according to the cherry-picking order $(2, 1)(4, 3)(6, 5)$.

Lemma 2 For a given leaf pair (ϵ, ζ) $(\epsilon > \zeta)$ in the labeled tree topology under the initial assignment (Figure 5B), we demonstrate that there exists a column vector $\boldsymbol{\nu}^{(\epsilon\zeta)} \in \mathbb{R}^{10\times 1}$ such that the tree metric $\delta_{\epsilon\zeta}$ is expressed as:

$$\delta_{\epsilon\zeta} = \boldsymbol{\nu}^{(\epsilon\zeta)} \cdot \mathbf{d}^{(5)},$$

i.e., the shortest path length between the leaves (ϵ, ζ) is a linear combination of the entries of the initial dissimilarity vector.

Proof. By Definition 8, ψ_0 defines the mapping from the leaves of the NJ tree to the taxa indices. The internal nodes u, v, and w are assigned to the indices 6, 7, and 8, respectively, under the assumption that the first cherry is $(2, 1)$, the second cherry is $(4, 3)$, and the third cherry is $(6, 5)$. This assumption holds because each equivalence class of cherry-picking orders contains at least one order where the first and second cherries consist solely of the original taxa (Section 2.1.5). Additionally, since all rows of $\mathbf{A}^{(3)}$ are identical, any remaining pair of nodes can be selected arbitrarily as the third cherry. With this cherry-picking order, the labeled tree topology corresponding to the initial assignment is shown in Figure 5B.

We classify the leaf pairs of the labeled tree topology under the initial assignment into four types based on the structure of the edges along the minimum path between each pair. Type 1 consists of pairs $(3, 1), (4, 1), (4, 2)$, and $(3, 2)$, where each minimum path traverses one edge from the first cherry-picking step, one from the second, and the edges e_{86} and e_{87}. Type 2 includes pairs $(5, 3)$ and $(5, 4)$, where the minimum paths involves edges e_{85}, e_{87}, and one edge from the second cherry-picking step. Type 3 includes pairs $(5, 1)$ and $(5, 2)$, whose minimum paths traverse edges e_{85}, e_{86}, and one edge from the first cherry-picking step. Finally, Type 4 consists of pairs $(2, 1)$ and $(4, 3)$, where the minimum path length between each pair equals their dissimilarity at the start of the NJ algorithm.

The classification of leaf pairs serves the purpose of expressing the minimum path length, equivalent to the tree metric between leaves, as a linear combination of the entries in the input dissimilarity vector for the five taxa. Two leaf pairs are classified under the same type if their respective linear combinations are related by a permutation of the indices. We claim that if the coefficient vector $\boldsymbol{\nu}^{(\epsilon\zeta)}$ is known for one leaf pair of a given type, the coefficient vector for any other leaf pair of the same type can be computed via a permutation matrix. This follows from the reasoning below.

Let (η, θ) $(\eta > \theta)$ be another leaf pair of the labeled tree topology with the initial assignment, such that (η, θ) belongs to the same type as (ϵ, ζ). Define a permutation σ' such that $\sigma'(\epsilon) = \eta$, $\sigma'(\zeta) = \theta$, $\sigma'(\eta) = \epsilon$, $\sigma'(\theta) = \zeta$, and $\sigma'(\chi) = \chi$ for all $\chi \in [5]$ with $\chi \neq \eta, \zeta, \theta, \epsilon$. Let k_{ij} denote the position of the pair $(i, j) \in \{(x, y) \mid x, y \in [5], x > y\}$ in the lexicographical ordering. Let $\boldsymbol{\nu}^{(\eta\theta)} \in \mathbb{R}^{10\times 1}$ be a column vector where each entry $\nu^{(\eta\theta)}_{k_{ij}}$ equals $\nu^{(\epsilon\zeta)}_{k_{\sigma'(i)\sigma'(j)}}$. Let $\boldsymbol{\rho}_{\sigma'} \in \mathbb{R}^{10\times 10}$ denote the permutation matrix induced by σ', permuting the indices of $\boldsymbol{\nu}^{(\epsilon\zeta)}$. Then,

$$\boldsymbol{\nu}^{(\eta\theta)} = \boldsymbol{\rho}_{\sigma'} \boldsymbol{\nu}^{(\epsilon\zeta)}.$$

From the classification, leaf pairs of the same type share minimum paths with identical edge compositions, including an equal number of edges from both the first and second

cherry-picking steps, as well as edges not involved in either step. Since edges from the same cherry-picking step correspond to the same number of unclustered nodes, their lengths are determined by the same formula, differing only by a permutation of the dissimilarity vector indices. This structural equivalence allows us to compute the tree metric between (η, θ) by permuting the known coefficient vector associated with another leaf pair of the same type. Therefore, $\delta_{\eta\theta}$ is expressed as:

$$\delta_{\eta\theta} = \boldsymbol{\nu}^{(\eta\theta)} \cdot \mathbf{d}^{(5)} = (\boldsymbol{\rho}_{\sigma'} \boldsymbol{\nu}^{(\epsilon\zeta)}) \cdot \mathbf{d}^{(5)}.$$

We proceed by constructing $\boldsymbol{\nu}^{(\epsilon\zeta)}$ for a representative pair from each type.

The first type of pairs consist of $(3,1),(4,1),(4,2)$ ***and*** $(3,2)$**.**
For the first type, without loss of generality, we express δ_{31} in terms of the input dissimilarity vector $\mathbf{d}^{(5)}$:

$$\delta_{31} = \beta_{61} + \ell_{76} + \beta_{73} = \beta_{61} + d^{(3)}_{76} + \beta_{73}. \tag{15}$$

The second step follows from Lemma 1. Using Eqs. 2–5, we compute each term separately as follows:

$$\beta_{61} = \frac{1}{2} d^{(5)}_{21} + \frac{1}{6}\left(d^{(5)}_{31} + d^{(5)}_{41} + d^{(5)}_{51} - d^{(5)}_{32} - d^{(5)}_{42} - d^{(5)}_{52}\right), \tag{16}$$

$$\begin{aligned} d^{(3)}_{76} &= \frac{1}{2}\left(d^{(4)}_{63} + d^{(4)}_{64} - d^{(5)}_{43}\right) \\ &= \frac{1}{2}\left[\frac{1}{2}\left(d^{(5)}_{41} + d^{(5)}_{42} - d^{(5)}_{21}\right) + \frac{1}{2}\left(d^{(5)}_{31} + d^{(5)}_{32} - d^{(5)}_{21}\right) - d^{(5)}_{43}\right] \\ &= \frac{1}{4}\left(-2d^{(5)}_{21} + d^{(5)}_{31} + d^{(5)}_{41} + d^{(5)}_{32} + d^{(5)}_{42} - 2d^{(5)}_{43}\right), \end{aligned} \tag{17}$$

$$\begin{aligned} \beta_{73} &= \frac{1}{2} d^{(5)}_{43} + \frac{1}{4}\left(d^{(4)}_{63} + d^{(5)}_{53} - d^{(4)}_{64} - d^{(5)}_{54}\right) \\ &= \frac{1}{2} d^{(5)}_{43} + \frac{1}{4}\left[\frac{1}{2}\left(d^{(5)}_{31} + d^{(5)}_{32} - d^{(5)}_{21}\right) + d^{(5)}_{53} - \frac{1}{2}\left(d^{(5)}_{41} + d^{(5)}_{42} + -d^{(5)}_{21}\right) - d^{(5)}_{54}\right] \\ &= \frac{1}{8}\left(d^{(5)}_{31} + d^{(5)}_{32} - d^{(5)}_{41} - d^{(5)}_{42} + 4d^{(5)}_{43} + 2d^{(5)}_{53} - 2d^{(5)}_{54}\right). \end{aligned} \tag{18}$$

Substituting Eqs. 16–18 into Eq. 15 gives:

$$\begin{aligned} \delta_{31} &= \frac{13}{24} d^{(5)}_{31} + \frac{5}{24} d^{(5)}_{32} + \frac{7}{24} d^{(5)}_{41} - \frac{1}{24} d^{(5)}_{42} + \frac{1}{6} d^{(5)}_{51} - \frac{1}{6} d^{(5)}_{52} + \frac{1}{4} d^{(5)}_{53} - \frac{1}{4} d^{(5)}_{54} \\ &= \underbrace{\left[0, \frac{13}{24}, \frac{5}{24}, \frac{7}{24}, -\frac{1}{24}, 0, \frac{1}{6}, -\frac{1}{6}, \frac{1}{4}, -\frac{1}{4}\right]^{\mathsf{T}}}_{\boldsymbol{\nu}^{(31)}} \cdot \mathbf{d}^{(5)}. \end{aligned} \tag{19}$$

The second type of pairs consist of $(5,3)$ ***and*** $(5,4)$**.**

For the second type, without loss of generality, we express δ_{53} in terms of the input dissimilarity vector $\mathbf{d}^{(5)}$:

$$\delta_{53} = \beta_{73} + \ell_{75} = \beta_{73} + d^{(3)}_{75}. \tag{20}$$

The second step follows from Lemma 1. Using Eqs. 2–5, we compute each term separately as follows:

$$\begin{aligned} d^{(3)}_{75} &= \frac{1}{2}\left(d^{(4)}_{54} + d^{(4)}_{53} - d^{(4)}_{43}\right) \\ &= \frac{1}{2}\left(d^{(5)}_{54} + d^{(5)}_{53} - d^{(5)}_{43}\right). \end{aligned} \tag{21}$$

By Eq. 18, we have:

$$\beta_{73} = \frac{1}{8}\left(d^{(5)}_{31} + d^{(5)}_{32} - d^{(5)}_{41} - d^{(5)}_{42} + 4d^{(5)}_{43} + 2d^{(5)}_{53} - 2d^{(5)}_{54}\right). \tag{22}$$

Substituting Eqs. 21 and 22 into Eq. 20 gives:

$$\begin{aligned} \delta_{53} &= \frac{1}{8}d^{(5)}_{31} + \frac{1}{8}d^{(5)}_{32} - \frac{1}{8}d^{(5)}_{41} - \frac{1}{8}d^{(5)}_{42} + \frac{3}{4}d^{(5)}_{53} + \frac{1}{4}d^{(5)}_{54} \\ &= \underbrace{\left[0, \frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}, 0, 0, 0, \frac{3}{4}, \frac{1}{4}\right]^{\mathsf{T}}}_{\boldsymbol{\nu}^{(53)}} \cdot \mathbf{d}^{(5)}. \end{aligned} \tag{23}$$

The third type of pairs consist of $(5,1)$ ***and*** $(5,2)$**.**

For the third type, without loss of generality, we express δ_{51} in terms of the input dissimilarity vector $\mathbf{d}^{(5)}$:

$$\delta_{51} = \beta_{61} + \ell_{65} = \beta_{61} + d^{(3)}_{65}. \tag{24}$$

The second step follows from Lemma 1. Using Eqs. 14 and 16, we compute each term separately as follows:

$$\beta_{61} = \frac{1}{2}d^{(5)}_{21} + \frac{1}{6}\left(d^{(5)}_{51} + d^{(5)}_{31} + d^{5}_{41} - d^{(5)}_{52} - d^{(5)}_{32} - d^{(5)}_{42}\right), \tag{25}$$

$$d^{(3)}_{65} = d^{(4)}_{65} = \frac{1}{2}\left(d^{(5)}_{51} + d^{(5)}_{52} - d^{(5)}_{21}\right). \tag{26}$$

Substituting Eqs. 25 and 26 into Eq. 24 gives:

$$\delta_{51} = \frac{1}{6}d^{(5)}_{31} - \frac{1}{6}d^{(5)}_{32} + \frac{1}{6}d^{(5)}_{41} - \frac{1}{6}d^{(5)}_{42} + \frac{2}{3}d^{(5)}_{51} + \frac{1}{3}d^{(5)}_{52}$$

$$= \underbrace{\left[0, \frac{1}{6}, -\frac{1}{6}, \frac{1}{6}, -\frac{1}{6}, 0, \frac{2}{3}, \frac{1}{3}, 0, 0\right]^{\mathsf{T}}}_{\boldsymbol{\nu}^{(51)}} \cdot \mathbf{d}^{(5)}. \tag{27}$$

The fourth type of pairs consist of $(2,1)$ ***and*** $(4,3)$**.**

From Eq. 5, we have:

$$\delta_{21} = \beta_{61} + \beta_{62} = d_{21}^{(5)} = \underbrace{[1,0,0,0,0,0,0,0,0,0]^{\mathsf{T}}}_{\boldsymbol{\nu}^{(21)}} \cdot \mathbf{d}^{(5)},$$
$$\delta_{43} = \beta_{73} + \beta_{74} = d_{43}^{(5)} = \underbrace{[0,0,0,0,0,1,0,0,0,0]^{\mathsf{T}}}_{\boldsymbol{\nu}^{(43)}} \cdot \mathbf{d}^{(5)}.$$

Up to this point, the coefficient vector has been computed for at least one leaf pair in each of the four types. Thus, the tree metric for any leaf pair in the labeled topology, under the initial assignment, can be expressed as a linear combination of the entries in the input dissimilarity vector.

From Eq. 19, the tree metrics for all leaf pairs belonging to the first type are given as follows:

$$\delta_{31} = \left[0, \frac{13}{24}, \frac{5}{24}, \frac{7}{24}, -\frac{1}{24}, 0, \frac{1}{6}, -\frac{1}{6}, \frac{1}{4}, -\frac{1}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)},$$
$$\delta_{32} = \left[0, \frac{5}{24}, \frac{13}{24}, -\frac{1}{24}, \frac{7}{24}, 0, -\frac{1}{6}, \frac{1}{6}, \frac{1}{4}, -\frac{1}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)},$$
$$\delta_{41} = \left[0, \frac{7}{24}, -\frac{1}{24}, \frac{13}{24}, \frac{5}{24}, 0, \frac{1}{6}, -\frac{1}{6}, -\frac{1}{4}, \frac{1}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)},$$
$$\delta_{42} = \left[0, -\frac{1}{24}, \frac{7}{24}, \frac{5}{24}, \frac{13}{24}, 0, -\frac{1}{6}, \frac{1}{6}, -\frac{1}{4}, \frac{1}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)}.$$

From Eq. 23, the tree metrics for all leaf pairs belonging to the second type are given as follows:

$$\delta_{53} = \left[0, \frac{1}{8}, \frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}, 0, 0, 0, \frac{3}{4}, \frac{1}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)}$$
$$\delta_{54} = \left[0, -\frac{1}{8}, -\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, 0, 0, 0, \frac{1}{4}, \frac{3}{4}\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)}$$

From Eq. 27, the tree metrics for all leaf pairs belonging to the third type are given as follows:

$$\begin{aligned} \delta_{51} &= \left[0, \frac{1}{6}, -\frac{1}{6}, \frac{1}{6}, -\frac{1}{6}, 0, \frac{2}{3}, \frac{1}{3}, 0, 0\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)} \\ \delta_{52} &= \left[0, -\frac{1}{6}, \frac{1}{6}, -\frac{1}{6}, \frac{1}{6}, 0, \frac{1}{3}, \frac{2}{3}, 0, 0\right]^{\mathsf{T}} \cdot \mathbf{d}^{(5)} \end{aligned} \tag{28}$$

□

Lemma 2 establishes the existence of a linear transformation from the input dissimilarity vector to the corresponding tree metric, formally stated in the following proposition.

Proposition 3 For $N = 5$, there exists a linear transformation $\mathbf{\Pi}_\sigma : \mathbb{R}^{10}_{\geq 0} \to \mathbb{R}^{10}_{\geq 0}$, represented by a matrix $\mathbf{\Pi}_\sigma$, that maps the dissimilarity vector $\mathbf{d}^{(5)}$ to the tree metric $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} = \mathbf{\Pi}_\sigma \mathbf{d}^{(5)}.$$

Proof. We first construct the matrix $\mathbf{\Pi}_{id}$ corresponding to the initial assignment ψ_0, where *id* represents the identity permutation. By Eqs. 19, 23, 27–28, the matrix $\mathbf{\Pi}_{id}$ is defined such that $\boldsymbol{\delta} = \mathbf{\Pi}_{id}\mathbf{d}^{(5)}$, as follows:

$$\mathbf{\Pi}_{id} = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 21 & 31 & 32 & 41 & 42 & 43 & 51 & 52 & 53 & 54 \end{array} \\ \left[\begin{array}{cccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{13}{24} & \frac{5}{24} & \frac{7}{24} & -\frac{1}{24} & 0 & \frac{1}{6} & -\frac{1}{6} & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{5}{24} & \frac{13}{24} & -\frac{1}{24} & \frac{7}{24} & 0 & -\frac{1}{6} & \frac{1}{6} & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{7}{24} & -\frac{1}{24} & \frac{13}{24} & \frac{5}{24} & 0 & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{4} & \frac{1}{4} \\ 0 & -\frac{1}{24} & \frac{7}{24} & \frac{5}{24} & \frac{13}{24} & 0 & -\frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & -\frac{1}{6} & \frac{1}{6} & -\frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{6} & \frac{1}{6} & -\frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & \frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{array}\right] \end{array}. \quad (29)$$

The matrix $\mathbf{\Pi}_\sigma$ for any σ-assignment is constructed by permuting the row and column indices of $\mathbf{\Pi}_{id}$ according to σ. Let k_{ij} denote the position of the pair $(i, j) \in \{(x, y) \mid x, y \in [5], x > y\}$ in the lexicographical ordering. Then the row or column indexed by the pair k_{ij} is mapped to the row or column indexed by $k_{\sigma(i)\sigma(j)}$.

Formally, let $\boldsymbol{\rho}_\sigma$ be the 10×10 permutation matrix induced by σ acting on the row indices of $\mathbf{\Pi}_{id}$. The rows of $\mathbf{\Pi}_{id}$ are permuted by left-multiplication with $\boldsymbol{\rho}_\sigma$, while the columns are permuted by right-multiplication with $\boldsymbol{\rho}_\sigma^\mathsf{T}$. Then, $\mathbf{\Pi}_\sigma$ is defined as:

$$\mathbf{\Pi}_\sigma = \boldsymbol{\rho}_\sigma \mathbf{\Pi}_{id} \boldsymbol{\rho}_\sigma^\mathsf{T}.$$

□

Definition 9 (Linear map from dissimilarity vectors to tree metrics for five taxa) Let $\mathbf{\Pi}_{id}$ denote the linear map based on the initial assignment, as defined in Eq. 29. The *linear map from dissimilarity vectors to tree metrics for five taxa* $\mathbf{\Pi}_\sigma$, associated with the σ-assignment, is defined as:

$$\mathbf{\Pi}_\sigma = \boldsymbol{\rho}_\sigma \mathbf{\Pi}_{id} \boldsymbol{\rho}_\sigma^\mathsf{T},$$

where $\boldsymbol{\rho}_\sigma$ is the permutation matrix induced by σ. The tree metric $\boldsymbol{\delta}$, corresponding to the NJ tree under the σ-assignment, is given by:

$$\boldsymbol{\delta} = \boldsymbol{\Pi}_\sigma \mathbf{d}^{(5)}.$$

Recall that Property 2 holds if and only if the distance between the admixed taxon and either source taxon is less than the distance between the two source taxa in the final NJ tree. With the tree metric $\boldsymbol{\delta}$ defined as a linear transformation from the input dissimilarity space to the tree metric space (Definition 9), Property 2 is represented by the following matrix:

Definition 10 (Tree metric entry comparison matrix) Define $\mathbf{U}$ as a linear map from $\mathbb{R}^{10}$ to $\mathbb{R}^2$ given by:

$$\mathbf{U} = \begin{array}{c} \begin{matrix} 21 & 31 & 32 & 41 & 42 & 43 & 51 & 52 & 53 & 54 \end{matrix} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}.$$

For $N = 5$, the tree metric $\boldsymbol{\delta}$ satisfies Property 2 if and only if

$$\mathbf{U}\boldsymbol{\delta} \leq 0. \tag{30}$$

Thus, $\mathbf{U}$ is referred to as the *tree metric entry comparison matrix*.

By Proposition 3, a tree metric can be expressed as the product of $\boldsymbol{\Pi}_\sigma$ and the corresponding input dissimilarity vector. Thus, by Eq. 30, an input dissimilarity vector satisfies Property 2 if and only if the following holds:

$$\mathbf{U}\boldsymbol{\Pi}_\sigma \mathbf{d}^{(5)} \leq 0. \tag{31}$$

The system of inequalities in Eq. 31 defines a cone bounded by two half-spaces, where each row of the matrix $\mathbf{U}\boldsymbol{\Pi}_\sigma$ corresponds to a half-space. To identify the set of dissimilarity vectors in an NJ cone that satisfy Property 2, we intersect the NJ cone with the cone defined by the half-spaces of $\mathbf{U}\boldsymbol{\Pi}_\sigma$. For each NJ cone, we augment the half-spaces of $\mathbf{U}\boldsymbol{\Pi}_\sigma$ with those defined by the NJ cone matrix in Definition 6. The resulting cone, bounded by this augmented set of half-spaces, is defined as follows:

Definition 11 (Property 2 cone) The *Property 2 cone matrix* $\mathbf{M}_{\widetilde{C}_\mathcal{O}}$ is a $\left(\sum_{k=4}^{N}\left(\binom{k}{2}-1\right)+2\right) \times \binom{N}{2}$ matrix defined as:

$$\mathbf{M}_{\widetilde{C}_\mathcal{O}} = \begin{bmatrix} \mathbf{M}_{C_\mathcal{O}} \\ \mathbf{U}\boldsymbol{\Pi}_\sigma \end{bmatrix},$$

where $\mathbf{M}_{C_\mathcal{O}}$ is the NJ cone matrix associated with $C_\mathcal{O}$, $\boldsymbol{\Pi}_\sigma$ is the permutation matrix from Proposition 3, and $\mathbf{U}$ is the comparison matrix enforcing Property 2. For each NJ cone $C_\mathcal{O}$, the *Property 2 cone*, denoted $\widetilde{C}_\mathcal{O}$, is the subset of $C_\mathcal{O}$ consisting of all input dissimilarity vectors whose corresponding tree metrics satisfy Property 2. It is defined as:

$$\widetilde{C}_\mathcal{O} = \left\{ \mathbf{d}^{(N)} \in \mathbb{R}^{\binom{N}{2}}_{\geq 0} \;\middle|\; \mathbf{M}_{\widetilde{C}_\mathcal{O}} \mathbf{d}^{(5)} \leq 0 \right\}.$$

The points within the cones defined by the augmented matrices represent dissimilarity vectors that produce tree metrics satisfying Property 2. Each NJ cone corresponds to a distinct labeled tree topology (not equivalent under reflection about the central node), uniquely determining the assignment of five taxa to the leaves of the final NJ tree. The assignment defines the σ-assignment (Definition 8), which allows for the construction of the matrix $\mathbf{\Pi}_\sigma$. Using the matrices $\mathbf{M}_{C_\mathcal{O}}$ and $\mathbf{U}$, we then construct the matrix $\mathbf{M}_{\widetilde{C}_\mathcal{O}}$ (Definition 11). A dissimilarity vector $\mathbf{d}^{(5)}$ gives rise to a tree metric satisfying Property 2 if and only if there exists a cherry-picking order whose associated Property 2 cone contains the dissimilarity vector.

3.2 Embedding of dissimilarity vectors with admixture

The dissimilarity vector with admixture for $N = 5$, derived from the dissimilarity matrix (Eq. 13), is:

$$\begin{bmatrix} d_{21}^{(4)} \\ d_{31}^{(4)} \\ d_{32}^{(4)} \\ d_{41}^{(4)} \\ d_{42}^{(4)} \\ d_{43}^{(4)} \\ (1-\alpha)d_{21}^{(4)} \\ \alpha d_{21}^{(4)} \\ \alpha d_{31}^{(4)} + (1-\alpha)d_{32}^{(4)} \\ \alpha d_{41}^{(4)} + (1-\alpha)d_{42}^{(4)} \end{bmatrix}.$$

Since the admixed dissimilarity vector contains six independent and four dependent entries, with the latter expressible as linear combinations of the former, there exists an injective linear transformation mapping a dissimilarity vector in $\mathbb{R}^6$ to an admixed dissimilarity vector in $\mathbb{R}^{10}$. Formally, we define:

Definition 12 (Embedding map for dissimilarity vector with admixture) The *embedding map with admixture fraction* α, denoted by $\iota_\alpha : \mathbb{R}^6 \hookrightarrow \mathbb{R}^{10}$, is defined as

$$\iota_\alpha\left(\begin{bmatrix} d_{21}^{(4)} \\ d_{31}^{(4)} \\ d_{32}^{(4)} \\ d_{41}^{(4)} \\ d_{42}^{(4)} \\ d_{43}^{(4)} \end{bmatrix}\right) = \mathbf{E}_\alpha \begin{bmatrix} d_{21}^{(4)} \\ d_{31}^{(4)} \\ d_{32}^{(4)} \\ d_{41}^{(4)} \\ d_{42}^{(4)} \\ d_{43}^{(4)} \end{bmatrix} = \begin{bmatrix} d_{21}^{(4)} \\ d_{31}^{(4)} \\ d_{32}^{(4)} \\ d_{41}^{(4)} \\ d_{42}^{(4)} \\ d_{43}^{(4)} \\ (1-\alpha)d_{21}^{(4)} \\ \alpha d_{21}^{(4)} \\ \alpha d_{31}^{(4)} + (1-\alpha)d_{32}^{(4)} \\ \alpha d_{41}^{(4)} + (1-\alpha)d_{42}^{(4)} \end{bmatrix},$$

where

$$\mathbf{E}_\alpha = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{array}{c} \begin{array}{cccccc} 21 & 31 & 32 & 41 & 42 & 43 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1-\alpha & 0 & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha & 1-\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & 1-\alpha & 0 \end{bmatrix} \end{array}. \tag{32}$$

Since the admixed dissimilarity vector involves six independent variables, despite residing in $\mathbb{R}^{10}$, we can project NJ cones into $\mathbb{R}^6$ without loss of information, induced by the inverse of the embedding map ι_α. This is formally defined as:

Definition 13 (Induced cone) Let $C \subseteq \mathbb{R}^{10}$ be a cone, and let $\mathbf{M}_C \in \mathbb{R}^{x \times 10}$ be the matrix whose rows correspond to the inward-pointing normal vectors of the half-spaces defining C, where x denotes the number of these half-spaces. For a given admixture fraction α, define the projection map $\pi_\alpha : \mathbb{R}^{10} \to \mathbb{R}^6$ by the matrix $\mathbf{E}_\alpha \in \mathbb{R}^{10\times 6}$ (Eq. 32). The image of C under π_α, denoted $\pi_\alpha(C) \subseteq \mathbb{R}^6$, is the cone whose bounding half-spaces are determined by the rows of the matrix $\mathbf{M}_C\mathbf{E}_\alpha$. We define $\pi_\alpha(C)$ as the *induced cone* by $\mathbf{E}_\alpha$.

3.3 Computation of cone volumes

In Section 3.1, we identified the cones associated with the three properties for $N = 5$, where every dissimilarity vector in each cone satisfies the corresponding property. Since the dissimilarity vector with admixture has six independent variables, the induced cones reside in $\mathbb{R}^6$ (Section 3.2). To compute the probability that a specific property holds, we restrict the sample space of dissimilarity vectors to the hypercube $[0, 1]^6$, normalizing the maximum dissimilarity between any pair of taxa to 1. This normalization can be achieved by dividing all entries of the dissimilarity vector by its maximum value. Such rescaling is linear and does not affect the outcome of the NJ algorithm. The intersection of each induced cone with this sample space forms a bounded polytope. The probability of the property being satisfied is given by the sum of the volumes of the polytopes formed by the cones satisfying the property, relative to the total volume of the sample space $[0, 1]^6$, which is 1.

We constructed the matrices $\mathbf{M}_{C_\mathcal{O}}$ (Definition 6) for all NJ cones $C_\mathcal{O} \subseteq \mathbb{R}^{10}$ by generating the matrices $\mathbf{A}^{(5)}$, $\mathbf{A}^{(4)}$, and $\mathbf{R}^{(5)}$ for each cherry-picking order using the functions `getPairs`, `makeAMatrix`, and `makeRMatrix` from our `NeighborJoining` module implemented in `Macaulay2` (see "Data and Code Availability"). Similarly, we constructed $\mathbf{M}_{\widetilde{C}_\mathcal{O}}$ for all Property 2 cones $\widetilde{C}_\mathcal{O}$ (Definition 11) by generating the matrices $\mathbf{U}$ and $\mathbf{\Pi}_\sigma$, then appending them to $\mathbf{M}_{C_\mathcal{O}}$. For a given admixture fraction α, we

subsequently computed the induced cones (Definition 13), $\pi_\alpha(\mathbf{M}_{C_\mathcal{O}})$ and $\pi_\alpha(\mathbf{M}_{\widetilde{C}_\mathcal{O}})$, in $\mathbb{R}^6$ using the projection map π_α.

We evaluated the volumes of the intersections between the induced cones and the sample space using two approaches. First, we computed the volumes directly using `Macaulay2`. Second, we estimated the volumes via Monte Carlo integration by randomly sampling dissimilarity vectors within a bounded region and calculating the proportion that lies inside the polytope.

3.3.1 Direct volume computation

`Macaulay2` computes the volume of a polytope by first triangulating the polytope into simplices. The triangulation is computed in `Macaulay2` using the software `TOPCOM` [18, 19]. Because there is a closed formula to compute the volume of a simplex that only depends on the vertex matrix of the simplex, it is then straightforward to compute the volume of the polytope as the sum of the volumes of all of these simplices.

Denote the volume of the intersection $[0,1]^6 \cap \pi_\alpha(C)$ as $\mathrm{Vol}_{\pi_\alpha(C)}$, where C represents either an NJ cone $C_\mathcal{O}$ or a Property 2 cone $\widetilde{C}_\mathcal{O}$. We employed the `Polyhedra` module from `Macaulay2` to compute the exact volumes of these intersections. The function `coneFromHData` was used to transform each $\mathbf{M}_C$ into its corresponding cone, which was then mapped to the induced cone $\pi_\alpha(C)$. To define the bounding regions, we used `hypercube` to construct the hypercube $[-1,1]^6$ and `posOrthant` to define the positive orthant $[0,\infty)^6$. The intersection of these polyhedral components was then computed using `intersect`, and the volume $\mathrm{Vol}_{\pi_\alpha(C)}$ was obtained using the `volume` function. This process provided the exact volumes of $[0,1]^6 \cap \pi_\alpha(C)$ for both NJ and Property 2 cones.

Let Ω_{all} represent the set of all equivalence classes of cherry-picking orders, and let Ω_1 and Ω_3 denote the sets of equivalence classes of cherry-picking orders satisfying Property 1 and Property 3, respectively (Table 2). The probabilities that a random dissimilarity vector results in an NJ tree satisfying Property 1, Property 2, or Property 3, respectively, are given by:

$$\begin{aligned} P_1 &= \sum_{\mathcal{O}\in\Omega_1} \mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}, \\ P_2 &= \sum_{\mathcal{O}\in\Omega_{\text{all}}} \mathrm{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}, \\ P_3 &= \sum_{\mathcal{O}\in\Omega_3} \mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}. \end{aligned} \tag{33}$$

3.3.2 Monte Carlo integration

For each $\alpha \in \{0.01, 0.02, ..., 0.99\}$, we generated $N_{\text{sample}} = 100,000$ random vectors $\mathbf{d}^{(4)} \in [0,1]^6$, totalling $99,00,000$ samples, using the `random` function from `Macaulay2`. Each random vector $\mathbf{d}^{(4)}$ corresponds to a dissimilarity vector with admixture, $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)} \in \mathbb{R}^{10}$. For each cone $C \subseteq \mathbb{R}^{10}$, representing either an NJ cone $C_\mathcal{O}$ or a Property 2 cone $\widetilde{C}_\mathcal{O}$, we evaluated whether $\mathbf{d}^{(4)}$ lies within the induced cone $\pi_\alpha(C) \subseteq \mathbb{R}^6$ by

checking the following inequality:

$$(\mathbf{M}_C \mathbf{E}_\alpha)\, \mathbf{d}^{(4)} \leq 0.$$

For a given admixture fraction α, the probabilities that an NJ tree inferred from $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$ satisfies Property 1, Property 2, or Property 3, respectively, are computed as:

$$P_1 = \frac{1}{N_{\text{sample}}} \sum_{\mathcal{O} \in \Omega_1} \mathbb{1}\left[\mathbf{d}^{(4)} \in \pi_\alpha(C_\mathcal{O})\right],$$

$$P_2 = \frac{1}{N_{\text{sample}}} \sum_{\mathcal{O} \in \Omega_{\text{all}}} \mathbb{1}\left[\mathbf{d}^{(4)} \in \pi_\alpha(\widetilde{C}_\mathcal{O})\right],$$

$$P_3 = \frac{1}{N_{\text{sample}}} \sum_{\mathcal{O} \in \Omega_3} \mathbb{1}\left[\mathbf{d}^{(4)} \in \pi_\alpha(C_\mathcal{O})\right].$$

3.3.3 Standard NJ simulation

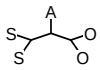
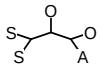
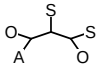
For comparison with the standard approach [14], we directly applied the NJ algorithm to each randomly generated admixed dissimilarity vector, $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$, to infer the corresponding NJ tree using our `NeighborJoining` module in `Macaulay2`. The `runNeighborJoiningClassic` function tracks the cherry-picking order throughout iterations, enabling identification of the corresponding NJ cone. To evaluate Property 1, we assessed whether a cherry containing the two source taxa $(1, 2)$ was selected in the first iteration, as those are the only cones violating Property 1 (Section 3.1.1; Type 3). Properties 2 and 3 were evaluated directly on the inferred NJ trees by first traversing the NJ tree with the `dfs` function to compute the number of edges and shortest path lengths between pairs of source and admixed taxa. These metrics were then compared to test adherence to the respective properties.

4 Results

We present the probability that a random dissimilarity vector for $N = 5$ with admixture satisfies the three properties defined in Section 2.2. For each admixture fraction $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, we computed $\text{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$, and averaged these values across all α. The probabilities of violating the properties were evaluated using three methods: direct volume computation, Monte Carlo integration, and NJ algorithm-based simulations. All methods produced consistent results.

Further analyses revealed the dependency of $\text{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ on α and its impact on the probabilities of satisfying the defined properties. We provide theoretical insights into why $\text{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0$ for certain cherry-picking orders and discuss its effect on the probabilities of satisfying Properties 1 and 3. Additionally, we proved the volume equivalence between specific $\pi_\alpha(C_\mathcal{O})$ and their corresponding $\pi_\alpha(\widetilde{C}_\mathcal{O})$, identifying the induced NJ cones that only contain dissimilarity vectors satisfying Property 2.

Table 2: **Mean $\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and mean $\mathbf{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ from direct computation.** The first column lists the indices of 30 distinct NJ cones $C_\mathcal{O}$ for $N = 5$, while the second column shows the equivalence class of cherry-picking orders for each NJ cone, following the convention in Section 2.1.5. The third column illustrates the labeled tree topology for each order, distinguishing reflection symmetry about the central taxon. The fourth and fifth columns indicate whether the corresponding NJ cone satisfies Property 1 and Property 3, respectively. "T" denotes that the property is satisfied, while "F" indicates it is not. The sixth and seventh columns report the mean volumes of $\pi_\alpha(C_\mathcal{O})$ and $\pi_\alpha(\widetilde{C}_\mathcal{O})$, respectively, averaged over 99 values of α in $\{0.01, 0.02, ..., 0.99\}$ for each $\mathcal{O} \in \Omega_{\text{all}}$. These volumes were obtained using the direct computation method (Section 3.3.1). Non-averaged volumes at $\alpha = 0.01, 0.5$, and 0.99 are provided in Table A1. The final column categorizes the 30 NJ cones by tree topology, distinguishing three node types: source taxa (S), an admixed taxon (A), and other taxa (O). This classification yields six distinct labeled tree topologies based on the arrangement of two S taxa, one A taxon, and two O taxa. These topology categories are considered equivalent if they differ only by reflection about the central taxon. Rows are ordered by increasing mean volume within each labeled tree topology category, with NJ cones further sorted by the ascending label of the central taxon.

Index	$\mathcal{O}$	**NJ Tree**	**Prop. 1**	**Prop. 3**	**Mean $\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})}$**	**Mean $\mathbf{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$**	**Tree Category**
1	(21)(43)		F	F	0	0	
2	(43)(21)		T				
3	(21)(54)		F	F	0	0	
4	(21)(53)						
5	(54)(21)		T				
6	(53)(21)						
7	(53)(42)		T	F	0	0	
8	(54)(32)						
9	(53)(41)						
10	(54)(31)						
11	(42)(53)				0.00031	0.00024	
12	(32)(54)						
13	(41)(53)						
14	(31)(54)						

Index	$\mathcal{O}$	NJ Tree	Prop. 1	Prop. 3	Mean $\mathrm{Vol}_{\pi_\alpha(C_{\mathcal{O}})}$	Mean $\mathrm{Vol}_{\pi_\alpha(\tilde{C}_{\mathcal{O}})}$	Tree Category
15	(51)(42)		T	T	0.04283	0.04259	
16	(52)(41)						
17	(51)(32)						
18	(52)(31)						
19	(42)(51)				0.05169	0.04370	
20	(41)(52)						
21	(32)(51)						
22	(31)(52)						
23	(31)(42)		T	T	0.06077	0.06073	
24	(32)(41)						
25	(42)(31)						
26	(41)(32)						
27	(52)(43)		T	T	0.05649	0.05613	
28	(51)(43)						
29	(43)(52)				0.13229	0.13229	
30	(43)(51)						

4.1 Mean $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\mathrm{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ over α

Table 2 presents the mean volumes $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\mathrm{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ for each $\mathcal{O} \in \Omega_{\text{all}}$, averaged across all values of $\alpha \in \{0.01, 0.02, ..., 0.99\}$, computed using the direct computation method (Section 3.3.1) For each $\mathcal{O} \in \Omega_{\text{all}}$, the induced NJ cones corresponding to the cherry-picking orders indexed 1–10 have volume zero. The proof of this result for all $\alpha \in (0, 1)$ is provided in Section 4.3. Since $\pi_\alpha(\widetilde{C}_\mathcal{O})$ is a subset of $\pi_\alpha(C_\mathcal{O})$, any $\pi_\alpha(\widetilde{C}_\mathcal{O})$ corresponding to a $\pi_\alpha(C_\mathcal{O})$ with volume zero also has a volume zero.

The largest mean $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ occurs for $\mathcal{O} = (43)(52)$ and $(43)(51)$ with $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0.13229$. Notably, the mean volumes of the induced Property 2 cones corresponding to these two NJ cones are identical to those of the induced NJ cones themselves, indicating that every dissimilarity vector within these induced NJ cones results in tree metrics that satisfy Property 2. The proof for this result for all $\alpha \in (0, 1)$ is provided in Section 4.4.

Table 2 groups the thirty labeled tree topologies into six categories, each representing a distinct assignment of two S taxa, one A taxon, and two O taxa to the leaves of the NJ tree. Categories differing only by a reflection about the central node are treated as equivalent. This categorization captures the distinct effects of admixture on the NJ tree structure. The category with the largest mean volume—the average $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ for all $\mathcal{O}$ within that category—corresponds to cherry-picking orders of the form $(SA)(OO)$.

4.2 $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ as functions of α

Figure 6 presents the volumes of induced NJ cones $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$, constrained to the sample space $[0, 1]^6$, as functions of the admixture fraction α, where $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$. The volumes reported are from the direct computation method. α ranges from near-boundary values ($\alpha = 0.01$ and $\alpha = 0.99$, where one source population dominates) to the midpoint $\alpha = 0.5$, representing equal contribution from both source populations. Each subplot groups distinct cones based on their mean volume over all α-values, as indicated by the NJ cone indices in Table 2. The curves show distinct behaviors depending on the structure of $\mathcal{O}$, despite the cones in each subplot having identical mean volumes averaged across all admixture fractions. Within each subplot, the NJ cones exhibit two distinct volume trajectories with respect to α, even when four cones share the same mean volume.

$\mathrm{Vol}_{\pi_\alpha(C_{(42)(53)})}$ and $\mathrm{Vol}_{\pi_\alpha(C_{(32)(54)})}$ decrease monotonically with α, while $\mathrm{Vol}_{\pi_\alpha(C_{(41)(53)})}$ and $\mathrm{Vol}_{\pi_\alpha(C_{(31)(54)})}$ increases over the same range of α (Figure 6A). These contrasting monotonic behaviors reflect distinct structural properties of the NJ cones, highlighting the role of admixture fractions in differentiating NJ cone geometries. Cones in the $(O, A)(S, O)$ category have volumes (maximum: 0.001760227) that are two orders of magnitude smaller than those in Figures 6B–F. The largest volume, 0.1701146, occurs for cones $\pi_\alpha(C_{(43)(52)})$ and $\pi_\alpha(C_{(43)(51)})$, classified under the $(S, A)(O, O)$ category (Figures 6F).

Similarly, $\mathrm{Vol}_{\pi_\alpha(C_{(52)(41)})}$ and $\mathrm{Vol}_{\pi_\alpha(C_{(52)(31)})}$ decrease monotonically as α increases from 0.01 to 0.99, while $\mathrm{Vol}_{\pi_\alpha(C_{(51)(42)})}$ and $\mathrm{Vol}_{\pi_\alpha(C_{(51)(32)})}$ increase monotonically over the same range (Figure 6B). All cherry-picking orders in this subplot are expressed as

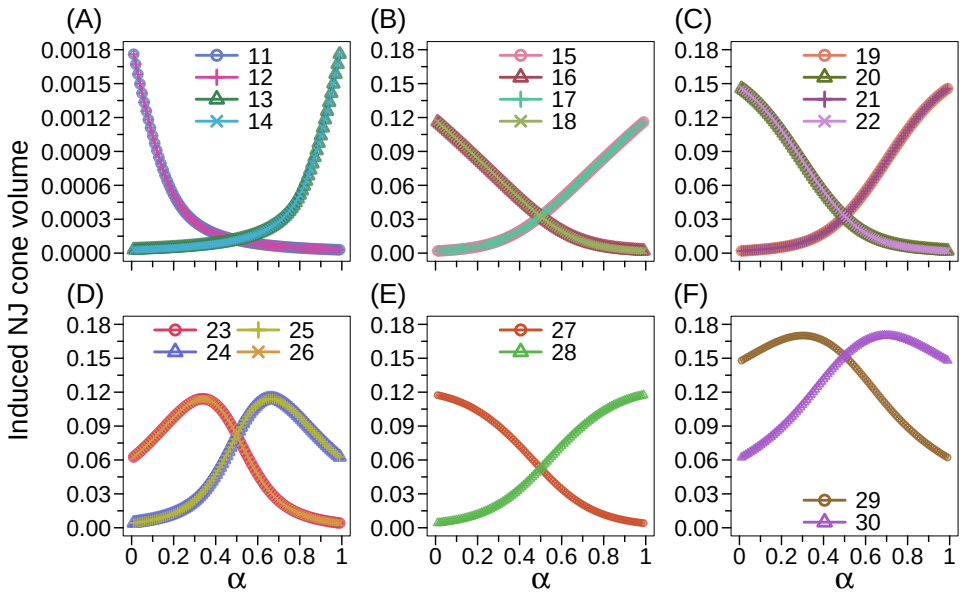


Fig. 6: **Induced NJ cone volume $\mathbf{Vol}_{\boldsymbol{\pi_\alpha(C_\mathcal{O})}}$ as a function of $\boldsymbol{\alpha}$.** The x-axis represents the admixture fraction α, where $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, and the y-axis shows the volume of the induced NJ cone constrained to the sample space $[0,1]^6$, i.e., $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})} = [0,1]^6 \cap \pi_\alpha(C)$, computed using the direct computation method (Section 3.3.1). Each curve corresponds to a distinct NJ cone $C_\mathcal{O}$, indexed as in Table 2 and grouped by mean volume over all α values. NJ cones with indices 1–10, for which $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0$ for all α, are omitted from the figure. NJ cones correspond to: **(A)** (42)(53), (32)(54), (41)(53), and (31)(54); indices 11–14. **(B)** (51)(42), (52)(41), (51)(32), and (52)(31); indices 15–18. **(C)** (42)(51), (41)(52), (32)(51), and (31)(52); indices 19–22. **(D)** (31)(42), (32)(41), (42)(31), and (41)(32); indices 23–26. **(E)** (52)(43) and (51)(43); indices 27, 28. **(F)** (43)(52) and (43)(51); indices 29, 30.

$(S,A)(S,O)$, where the first cherry consists of a source taxon and an admixed taxon. For small α, $d^{(5)}_{5i} = \alpha d^{(5)}_{1i} + (1-\alpha) d^{(5)}_{2i}$ (Eq. 12), which makes $d^{(5)}_{5i}$ closer to $d^{(5)}_{2i}$, increasing the probability of selecting $(5,2)$ as the first cherry. Conversely, for large α, $d^{(5)}_{5i}$ becomes closer to $d^{(5)}_{1i}$ than to $d^{(5)}_{2i}$, favoring the selection of $(5,1)$ as the first cherry. This analogous pattern is observed for cones in panels C and E.

Figures 6D and F exhibit a qualitative difference from the other panels, where two volumes increase while two decrease monotonically with α. In contrast, Figures 6D and F show more complex, non-monotonic relationships between α and the induced NJ cone volumes. In Figure 6D, cones $\pi_\alpha(C_{(31)(42)})$ and $\pi_\alpha(C_{(41)(32)})$ attain their maximum and minimum volumes at $\alpha = 0.34$ and $\alpha = 0.99$, respectively, while cones $\pi_\alpha(C_{(32)(41)})$ and $\pi_\alpha(C_{(42)(31)})$ reach their maximum at $\alpha = 0.66$ and minimum at $\alpha = 0.01$. All these cones share the same maximum volume of 0.1142092 and minimum volume of 0.001760227. In Figure 6F, cone $\pi_\alpha(C_{(43)(52)})$ reaches its maximum volume at $\alpha = 0.30$ and minimum volume at $\alpha = 0.99$, while cone $\pi_\alpha(C_{(43)(51)})$ attains its maximum at $\alpha = 0.70$ and minimum at $\alpha = 0.01$. Both cones share the largest

maximum volume of 0.1701146 and the largest minimum volume of 0.06219378 across all cones for any α.

Figure 7 shows the induced Property 2 cone volumes $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ as a function of α. The results qualitatively align with Figure 6, except for Figure 7C, though the volumes are consistently smaller for each $\mathcal{O}$ and α due to the induced Property 2 cone being a subset of the corresponding induced NJ cone. Unlike Figure 6C, Figure 7C does not exhibit strict monotonicity with respect to α. The volumes of $\pi_\alpha(C_{(42)(51)})$ and $\pi_\alpha(C_{(32)(51)})$ increase monotonically until $\alpha = 0.91$, reaching a maximum of 0.09632147, after which they decrease monotonically as α approaches 0.99. Conversely, the volumes of $\pi_\alpha(C_{(41)(52)})$ and $\pi_\alpha(C_{(31)(52)})$ follow an opposite trend, increasing monotonically until $\alpha = 0.09$, reaching the same maximum of 0.09632147, and decreasing monotonically for $\alpha \in (0.09, 0.99)$.

Table A1 reports the computed volumes $\text{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ for each $\mathcal{O}$ at specific values of $\alpha = 0.01, 0.5$, and 0.99. These α-values were chosen to capture the

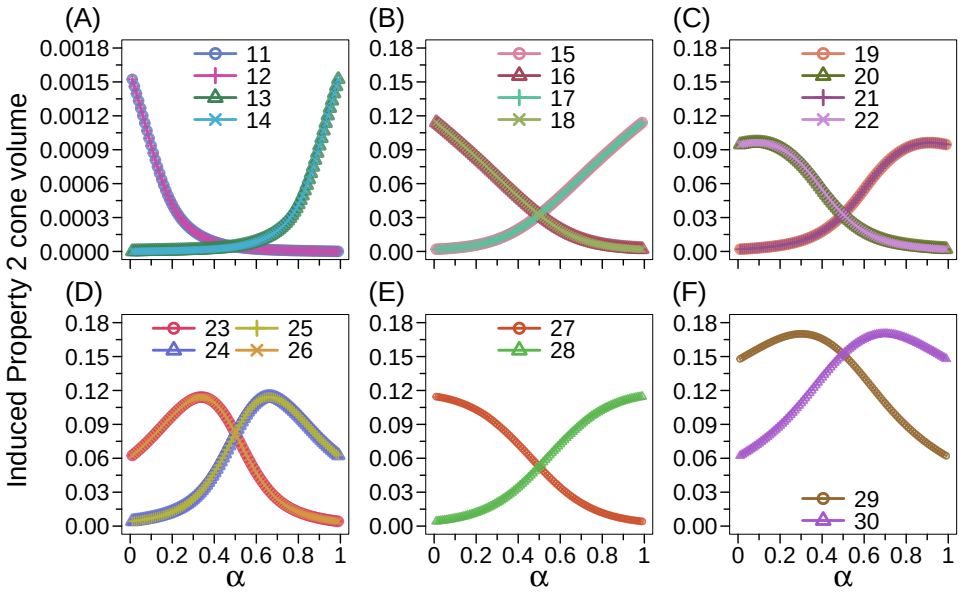


Fig. 7: Induced Property 2 cone volume $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ as a function of α. The x-axis represents the admixture fraction α, where $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, and the y-axis shows the volume of the induced Property 2 cone constrained to the sample space $[0, 1]^6$, i.e., $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})} = [0, 1]^6 \cap \pi_\alpha(\widetilde{C})$, computed using the direct computation method (Section 3.3.1). The cone indices follow those listed in Table 2. NJ cones with indices 1–10, for which $\text{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})} = 0$ for all α, are omitted from the figure. The figure layout mirrors that of Figure 6. Property 2 cones correspond to: **(A)** (42)(53), (32)(54), (41)(53), and (31)(54); indices 11–14. **(B)** (51)(42), (52)(41), (51)(32), and (52)(31); indices 15–18. **(C)** (42)(51), (41)(52), (32)(51), and (31)(52); indices 19–22. **(D)** (31)(42), (32)(41), (42)(31), and (41)(32); indices 23–26. **(E)** (52)(43) and (51)(43); indices 27, 28. **(F)** (43)(52) and (43)(51); indices 29, 30.

behavior at the extremities of the parameter space, where $\alpha = 0.01$ and $\alpha = 0.99$ correspond to near-complete contribution from a single source population, and $\alpha = 0.5$ represents an equal admixture scenario.

4.3 Cherry-picking orders with $\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0$

Both the direct volume computation method and the Monte Carlo method support the following proposition, which we prove using two independent methods.

Proposition 4 For any admixture fraction $\alpha \in (0,1)$, $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0$ if $\mathcal{O}$ corresponds to one of the following cherry-picking orders:

$$\begin{aligned}&(2,1)(4,3),\ \ (2,1)(5,3),\ \ (2,1)(5,4),\ \ (4,3)(2,1),\ \ (5,3)(2,1),\\ &(5,3)(4,1),\ \ (5,3)(4,2),\ \ (5,4)(2,1),\ \ (5,4)(3,1),\ \ (5,4)(3,2).\end{aligned} \tag{34}$$

Proof. ***Method 1: proof by contradiction***

Let $\mathcal{O}$ be a cherry-picking order from Eq. 34. By applying the Fourier–Motzkin elimination [20–22] to the system of inequalities $(\mathbf{M}_{C_\mathcal{O}}\mathbf{E}_\alpha)\,\mathbf{d}^{(4)} \leq 0$, which defines the induced NJ cone $\pi_\alpha(C_\mathcal{O})$, we obtain that there exists $i \in [6]$ such that $d_i^{(4)} \leq 0$ [14]. This contradicts the assumption that all entries of $\mathbf{d}^{(4)}$ are strictly positive. Thus, no dissimilarity vector $\mathbf{d}^{(4)}$ belongs to the induced cone $\pi_\alpha(C_\mathcal{O})$, implying its volume is $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})} = 0$.

Method 2: proof by dimensionality

The dimensions of the induced NJ cones $\pi_\alpha(C_\mathcal{O})$ were computed in `Macaulay2`. All cones corresponding to the cherry-picking orders in Eq. 34 have dimensions strictly less than 6, indicating that they reside in proper subspaces of $\mathbb{R}^6$. Since the dimension of each cone is strictly lower than that of the ambient space, their volumes in $\mathbb{R}^6$ are zero. □

Corollary 4.1 For any admixture fraction $\alpha \in (0,1)$, the probability that an admixed dissimilarity vector $\mathbf{d}^{(5)}$ violates Property 1 is zero.

Proof. Only Type-3 cherry-picking orders violate Property 1 (Section 3.1.1), and Proposition 4 lists all such Type-3 cherry-picking orders, showing that the induced cones corresponding to these orders have zero volume. Thus, for a dissimilarity vector $\mathbf{d}^{(5)}$ with admixture, no Type-3 cherry-picking orders are possible. Therefore, the probability of $\mathbf{d}^{(5)}$ violating Property 1 is zero.

□

Corollary 4.2 For any admixture fraction $\alpha \in (0,1)$, the probability that an admixed dissimilarity vector $\mathbf{d}^{(5)}$ violates Property 3 is given by the sum of volumes of the induced NJ cones corresponding to the following cherry-picking orders:

$$(3,2)(5,4),\ \ (4,2)(5,3),\ \ (3,1)(5,4),\ \ (4,1)(5,3).$$

Proof. There are 14 cherry-picking orders that violate Property 3 (Section 3.1.2). Of these, 10 correspond to volume-zero induced NJ cones as listed in Proposition 4, while the remaining four have non-zero volumes. Thus, the probability of violating Property 3 is given by the sum of the volumes of these four $\pi_\alpha(C_{\mathcal{O}})$'s not included in Proposition 4. □

4.4 Cherry-picking orders with $\mathbf{Vol}_{\pi_\alpha(C_{\mathcal{O}})} = \mathbf{Vol}_{\pi_\alpha(\widetilde{C}_{\mathcal{O}})}$

Proposition 5 For any admixture fraction $\alpha \in (0,1)$, no dissimilarity vector $\mathbf{d}^{(4)}$ within the induced NJ cones $\pi_\alpha(C_{(4,3)(5,1)})$ and $\pi_\alpha(C_{(4,3)(5,2)})$ (Figure B5) violates Property 2.

Proof. We proceed by contradiction, assuming that Property 2 is violated. This implies that either $\delta_{21} < \delta_{51}$ or $\delta_{21} < \delta_{52}$.

The σ-assignment corresponding to the induced NJ cone $\pi_\alpha(C_{(4,3)(5,1)})$, and thereby the labeled tree topology in Figure B5A, reorders the taxa from the initial assignment (Figure 5B) as follows:

$$\sigma(1) = 4,\ \ \sigma(2) = 3,\ \ \sigma(3) = 5,\ \ \sigma(4) = 1,\ \ \sigma(5) = 2.$$

The linear map from dissimilarity vectors to tree metrics for this σ-assignment, $\mathbf{\Pi}_\sigma = \boldsymbol{\rho}_\sigma \mathbf{\Pi}_{\mathrm{id}} \boldsymbol{\rho}_\sigma^{\mathsf{T}}$ (Definition 9), is then given by:

$$\mathbf{\Pi}_\sigma = \begin{array}{c} \\ 21 \\ 31 \\ 32 \\ 41 \\ 42 \\ 43 \\ 51 \\ 52 \\ 53 \\ 54 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 21 & 31 & 32 & 41 & 42 & 43 & 51 & 52 & 53 & 54 \end{array} \\ \begin{bmatrix} \frac{3}{4} & \frac{1}{8} & 0 & \frac{1}{8} & 0 & 0 & 0 & \frac{1}{4} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{4} & \frac{13}{24} & \frac{1}{6} & \frac{5}{24} & -\frac{1}{6} & 0 & 0 & -\frac{1}{4} & \frac{7}{24} & -\frac{1}{24} \\ 0 & \frac{1}{6} & \frac{2}{3} & -\frac{1}{6} & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{6} & -\frac{1}{6} \\ \frac{1}{4} & \frac{5}{24} & -\frac{1}{6} & \frac{13}{24} & \frac{1}{6} & 0 & 0 & -\frac{1}{4} & -\frac{1}{24} & \frac{7}{24} \\ 0 & -\frac{1}{6} & \frac{1}{3} & \frac{1}{6} & \frac{2}{3} & 0 & 0 & 0 & -\frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & -\frac{1}{8} & 0 & -\frac{1}{8} & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ -\frac{1}{4} & \frac{7}{24} & \frac{1}{6} & -\frac{1}{24} & -\frac{1}{6} & 0 & 0 & \frac{1}{4} & \frac{13}{24} & \frac{5}{24} \\ -\frac{1}{4} & -\frac{1}{24} & -\frac{1}{6} & \frac{7}{24} & \frac{1}{6} & 0 & 0 & \frac{1}{4} & \frac{5}{24} & \frac{13}{24} \end{bmatrix} \end{array}. \tag{35}$$

We first consider the case where $\mathcal{O} = (4,3)(5,1)$ and $\delta_{21} < \delta_{52}$. Then by Eq. 35,

$$\delta_{52} = \frac{1}{4}d_{21}^{(5)} - \frac{1}{8}d_{31}^{(5)} - \frac{1}{8}d_{41}^{(5)} + \frac{3}{4}d_{52}^{(5)} + \frac{1}{8}d_{53}^{(5)} + \frac{1}{8}d_{54}^{(5)}, \tag{36}$$

$$\delta_{21} = \frac{3}{4}d_{21}^{(5)} + \frac{1}{8}d_{31}^{(5)} + \frac{1}{8}d_{41}^{(5)} + \frac{1}{4}d_{52}^{(5)} - \frac{1}{8}d_{53}^{(5)} - \frac{1}{8}d_{54}^{(5)}. \tag{37}$$

Subtracting Eq. 36 from Eq. 37 results in the following:

$$\begin{aligned}\delta_{21} - \delta_{52} &= \frac{1}{2}d_{21}^{(5)} + \frac{1}{4}d_{31}^{(5)} + \frac{1}{4}d_{41}^{(5)} - \frac{1}{2}d_{52}^{(5)} - \frac{1}{4}d_{53}^{(5)} - \frac{1}{4}d_{54}^{(5)}, \\ &= \frac{1}{2}(1-\alpha)d_{21}^{(4)} + \frac{1}{4}(1-\alpha)d_{31}^{(4)} + (1-\alpha)d_{32}^{(4)} + \frac{1}{4}(1-\alpha)d_{41}^{(4)} - \frac{1}{4}(1-\alpha)d_{42}^{(4)},\end{aligned}$$

where the last step follows from $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$ (Definition 12). Then $\delta_{21} < \delta_{52}$ can be written as the following inequality:

$$\left[\frac{1}{2}(1-\alpha), \frac{1}{4}(1-\alpha), 1-\alpha, \frac{1}{4}(1-\alpha), -\frac{1}{4}(1-\alpha)\right]^\mathsf{T} \cdot \mathbf{d}^{(4)} < 0.$$

Dividing both sides of the inequality by $1-\alpha$ gives:

$$\left[\frac{1}{2}, \frac{1}{4}, 1, \frac{1}{4}, -\frac{1}{4}\right]^\mathsf{T} \cdot \mathbf{d}^{(4)} < 0. \tag{38}$$

By Definition 13, every dissimilarity vector $\mathbf{d}^{(4)}$ contained in $\pi_\alpha(C_{(4,3)(5,1)})$ satisfies:

$$(\mathbf{M}_{C_{(4,3)(5,1)}}\mathbf{E}_\alpha)\mathbf{d}^{(4)} = \begin{bmatrix} 0 & -\alpha & \alpha-1 & -\alpha & \alpha-1 & 1 \\ -\alpha+2 & -2 & 0 & -\alpha & \alpha-2 & 2 \\ \alpha+1 & 0 & -2 & -\alpha-1 & \alpha-1 & 2 \\ -\alpha+2 & -\alpha & \alpha-2 & -2 & 0 & 2 \\ \alpha+1 & -\alpha-1 & \alpha-1 & 0 & -2 & 2 \\ 2\alpha & 0 & -1 & 0 & -1 & 1 \\ -2\alpha+2 & -1 & 0 & -1 & 0 & 1 \\ 1 & -2\alpha & 2\alpha-2 & -1 & -1 & 2 \\ 1 & -1 & -1 & -2\alpha & 2\alpha-2 & 2 \\ -\alpha & -1/2\alpha & 1/2\alpha & -1/2\alpha & 1/2\alpha & 0 \\ -2\alpha+1 & -1/2 & 1/2 & -1/2 & 1/2 & 0 \end{bmatrix} \mathbf{d}^{(4)} \leq 0. \tag{39}$$

Multiplying Eq. 38 by 2 and adding it to the last row of Eq. 39 yields:

$$\begin{aligned}&\left[2(1-\alpha), 0, \frac{5}{2}, 0, 0\right]^\mathsf{T} \mathbf{d}^{(4)} \leq 0, \\ \implies &\ 2(1-\alpha)d_{21}^{(4)} + \frac{5}{2}d_{32}^{(4)} \leq 0.\end{aligned} \tag{40}$$

Since $2(1-\alpha) > 0$ and $\frac{5}{2} > 0$, Eq. 40 implies that either $d_{21}^{(4)} \leq 0$ or $d_{21}^{(4)} \leq 0$, which contradicts the assumption that all entries of $\mathbf{d}^{(4)}$ are strictly positive.

In the second case, $\mathcal{O} = (4,3)(5,1)$ and $\delta_{21} < \delta_{51}$. As in the first case, if the dissimilarity vector $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$ corresponds to a tree metric where $\delta_{21} < \delta_{51}$, the

following inequality holds:

$$\left[(\frac{5}{4}\alpha - \frac{1}{4}), \frac{1}{8}(1-\alpha), -\frac{1}{8}(1-\alpha), \frac{1}{8}(1-\alpha), -\frac{1}{8}(1-\alpha), 0\right]^{\mathsf{T}} \cdot \mathbf{d}^{(4)} \leq 0. \tag{41}$$

Multiplying Eq. 41 by 8 and adding it to the last row of Eq. 39, scaled by $2(1-\alpha)$, results in the following:

$$\begin{aligned} & [2(1-\alpha)(-2\alpha+1) + 10\alpha - 2, 0, 0, 0, 0, 0]^{\mathsf{T}} \cdot \mathbf{d}^{(4)} \leq 0, \\ \implies & \left[4\alpha^2 + 4\alpha, 0, 0, 0, 0, 0\right]^{\mathsf{T}} \cdot \mathbf{d}^{(4)} \leq 0, \\ \implies & 4\alpha(\alpha+1)d_{21}^{(4)} \leq 0. \end{aligned} \tag{42}$$

Since $4\alpha > 0$ and $\alpha + 1 > 0$, Eq. 42 implies $d_{21}^{(4)} \leq 0$, contradicting the assumption that all entries of $\mathbf{d}^{(4)}$ are strictly positive.

Thus, no dissimilarity vector in $\pi_\alpha(C_{(4,3)(5,1)})$ violates Property 2. By analogous reasoning, the same conclusion holds for dissimilarity vectors in $\pi_\alpha(C_{(4,3)(5,2)})$, which corresponds to the labeled tree topology in Figure B5B. □

4.5 Probability of satisfying each of the three properties

For an admixed dissimilarity vector with $N = 5$, the probability of satisfying Property 1 is always 1, independent of α. Only Type-3 equivalence classes corresponding to the cherry-picking orders (21)(54), (21)(53), and (21)(43) violate Property 1 (Section 3.1.1). Two independent approaches confirmed that the NJ cones corresponding to these equivalence classes of cherry-picking orders have zero volume. The first approach involves direct computation from Table 2, and the second follows from Proposition 4 in Section 4.3.

The probability of satisfying Property 2 equals the total volume of all cones satisfying Property 2 (Eq. 33). Figure 8A shows how P_2 depends on the admixture parameter α, computed via direct volume computation. For $\alpha = 0.5$, representing equal contribution from both source populations, the probability reaches its maximum value of 0.9996976. As α deviates from 0.5, reflecting increasing asymmetry in the admixture proportions, the probability decreases symmetrically, reaching a minimum of 0.8903937 at the extreme values $\alpha = 0.01$ and 0.99, where one source population contributes almost entirely to the admixed population. The observed symmetry around $\alpha = 0.5$ aligns with the interchangeable roles of the two source populations in the assumed admixture model.

Similar qualitative behavior is observed for Property 3 (Figure 8B), whose probability is computed as the total volume of the induced NJ cones satisfying Property 3 (Eq. 33, Figure B4). As with Property 2, the maximum probability occurs at balanced admixture ($\alpha = 0.5$) with a value of 0.999537, which is slightly higher than that of Property 2. However, unlike Property 2, even at extreme admixture proportions ($\alpha = 0.01$ and $\alpha = 0.99$), the probability remains high, at 0.999537. This indicates that Property 3 is robust to skewed admixture, while Property 2 is more sensitive to asymmetry in the contributions from the source populations.

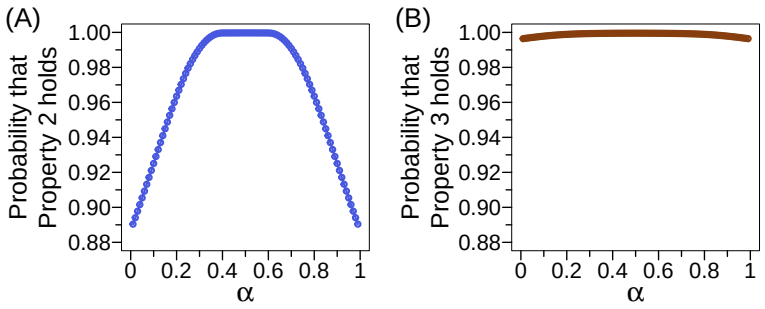


Fig. 8: P_2 **and** P_3 **as functions of** α**.** The x-axis denotes the admixture fraction α, where $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$, and the y-axis represents the probability that a given property holds, computed using the direct computation method (Section 3.3.1). For each value of α, the probability is obtained by summing the volumes of the induced cones intersecting with the sample space $[0,1]^6$. **(A)** The probability that Property 2 holds, $P_2 = \sum_{\mathcal{O}\in\Omega_{\text{all}}} \text{Vol}_{\pi_\alpha(\tilde{C}_\mathcal{O})}$, where the summation is taken over all cherry-picking orders. **(B)** The probability that Property 3 holds, $P_3 = \sum_{\mathcal{O}\in\Omega_3} \text{Vol}_{\pi_\alpha(C_\mathcal{O})}$, where the summation is restricted to the cones corresponding to cherry-picking orders that satisfy Property 3. Property 1 holds for all admixed dissimilarity vectors $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$ across all values of α, i.e., $P_1 = 1$ for all $\alpha \in (0,1)$ (Corollary 4.1). Thus, the corresponding plot for Property 1 is omitted from the figure.

To assess the accuracy of the direct volume computation and Monte Carlo methods, we compared their results with those obtained from the standard NJ simulation [14] in `Macaulay2` (Section 3.3). Table 3 summarizes the probabilities of violating each of the three properties across the methods. The results show close agreement, with differences only at the fifth decimal place.

Table 3: **Methods comparison.** This table shows the probabilities of violating each of the three properties averaged across all $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$. In the direct computation method (Section 3.3.1), the probability of violation was computed as 1 minus the total volume of the induced cones satisfying each property within the sample space $[0,1]^6$. In the Monte Carlo method (Section 3.3.2), 100,000 dissimilarity vectors $\mathbf{d}^{(4)}$ were uniformly sampled from $[0,1]^6$ for each of the 99 α values. The violation probability was computed as the proportion of vectors falling outside the induced cones, based on the total of 9,900,000 samples. In the standard NJ simulation (Section 3.3.3), the NJ algorithm was directly applied to each dissimilarity vector $\mathbf{d}^{(5)} = \mathbf{E}_\alpha \mathbf{d}^{(4)}$ to infer the corresponding NJ tree. The probability of violation was then calculated as the fraction of NJ trees (out of 9,900,000) that failed to satisfy the properties.

	Probability of violation		
Property violated	**1**	**2**	**3**
Direct computation	0	0.0341325	0.0012592
Monte Carlo	0	0.0341488	0.0012688
Standard NJ simulation	0	0.0341288	0.0012461

5 Discussion

In this study, we have investigated the geometric and probabilistic behavior of the NJ algorithm applied to distance matrices under admixture, focusing on a five-taxon case. By formulating the problem using polyhedral cones and projection maps, we have introduced a geometric framework that partitions the space of dissimilarity vectors based on their clustering, distance, and topological path length properties. This approach has enabled direct computation of the associated probabilities by evaluating the volumes of the induced cones within the bounded dissimilarity space. We have validated our analytical results via Monte Carlo integration and classical NJ simulations, confirming their accuracy. We have shown that while Property 1 is always satisfied, the probabilities of satisfying Properties 2 and 3 depend significantly on the admixture fraction. We have also proven that certain induced NJ cones have zero volume when admixture is present, indicating that these topologies are structurally incompatible with admixture under the NJ framework. Our study contributes to advancing the theoretical understanding of how admixture affects the NJ tree inference. Further, our `Macaulay2` implementation enables efficient analysis of the NJ algorithm under admixture, providing a valuable tool for studying complex evolutionary relationships involving admixed populations.

Although our analysis has focused on the five-taxon case, the geometric and combinatorial structure of NJ cones and their projections to lower-dimensional spaces can be generalized to cases with $N > 5$ taxa. Our framework can also be extended to multi-way linear admixture models beyond two-way admixture, introducing additional constraints on the dissimilarity vectors. These complexities can be addressed by extending the projection map to accommodate higher-order mixtures and defining new classes of induced NJ cones that capture the more complex admixture relationships among taxa. Beyond linear distance measures, further theoretical exploration should focus on formalizing the behavior of the NJ algorithm under non-linear genetic distances, such as F_{ST} [23] and other F-statistics [24, 25], which would require modifications to the current linear projection map.

The NJ algorithm has been shown to be a greedy heuristic [26] for the balanced minimum evolution (BME) problem [27, 28]. A natural extension of our current geometric approach would be to investigate whether the same properties involving admixture and associated probabilities observed under NJ also hold in BME. By leveraging the geometric methods from this work, we can analyze the BME cones [15, 16, 29] analogous to the NJ cones. Such an analysis would reveal structural similarities and differences between NJ and BME, providing a deeper understanding of how algorithmic choices impact phylogenetic tree construction under complex evolutionary models, including cases of admixture.

Finally, our framework provides a principled approach for studying an admixed taxon as a "rogue taxon" [30, 31]. Kim et al. [14] demonstrated via simulation that the three properties hold more frequently when the distances among $N - 1$ non-admixed taxa are additive, a phenomenon linked to the rogue taxon behavior. Since a metric is additive if and only if it satisfies the four-point condition [2], defined by a set of linear inequalities, our geometric framework is ideally suited to rigorously analyze how

the addition of an admixed taxon to an underlying set of $N-1$ populations with a tree-like evolutionary history affects the topological stability of inferred source trees.

Data and Code Availability. All code used in this manuscript, including our implementation of the NJ algorithm as the module `NeighborJoining`, is available on Zenodo at 10.5281/zenodo.13363307.

Acknowledgements. This work was supported by National Science Foundation grant DMS-2001367.

References

[1] Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4), 406–425 (1987) https://doi.org/10.1093/oxfordjournals.molbev.a040454

[2] Buneman, P.: A note on the metric properties of trees. Journal of Combinatorial Theory, Series B **17**(1), 48–50 (1974) https://doi.org/10.1016/0095-8956(74)90047-1

[3] Atteson, K.: The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica **25**(2), 251–278 (1999) https://doi.org/10.1007/PL00008277

[4] Mihaescu, R., Levy, D., Pachter, L.: Why neighbor-joining works. Algorithmica **54**(1), 1–24 (2009) https://doi.org/10.1007/s00453-007-9116-4

[5] Steel, M.: Phylogeny: Discrete and Random Processes in Evolution. Society for Industrial and Applied Mathematics, Philadelphia (2016)

[6] Bowcock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K., Cavalli-Sforza, L.L.: Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. Proceedings of the National Academy of Sciences of the United States of America **88**(3), 839–843 (1991) https://doi.org/10.1073/pnas.88.3.839

[7] Mountain, J.L., Cavalli-Sforza, L.L.: Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proceedings of the National Academy of Sciences of the United States of America **91**(14), 6515–6519 (1994) https://doi.org/10.1073/pnas.91.14.6515

[8] Bian, Y., Zhang, S., Zhou, W., Zhao, Q., Siqintuya, Zhu, R., Wang, Z., Gao, Y., Hong, J., Lu, D., Li, C.: Analysis of genetic admixture in Uyghur using the 26 Y-STR loci system. Scientific Reports **6**(1), 19998 (2016) https://doi.org/10.1038/srep19998

[9] Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L., Badii, R.,

Al-Nabet Al-Marri, A., Abi Khalil, C., Zirie, M., Jayyousi, A., Salit, J., Keinan, A., Clark, A.G., Crystal, R.G., Mezey, J.G.: Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome Research **26**(2), 151–162 (2016) https://doi.org/10.1101/gr.191478.115

[10] Kennedy, J.P., Pil, M.W., Proffitt, C.E., Boeger, W.A., Stanford, A.M., Devlin, D.J.: Postglacial expansion pathways of red mangrove, *Rhizophora mangle*, in the Caribbean Basin and Florida. American Journal of Botany **103**(2), 260–276 (2016) https://doi.org/10.3732/ajb.1500183

[11] Bergland, A.O., Tobler, R., González, J., Schmidt, P., Petrov, D.: Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. Molecular Ecology **25**(5), 1157–1174 (2016) https://doi.org/10.1111/mec.13455

[12] Feng, X., Merilä, J., Löytynoja, A.: Complex population history affects admixture analyses in nine-spined sticklebacks. Molecular Ecology **31**(20), 5386–5401 (2022) https://doi.org/10.1111/mec.16651

[13] Kopelman, N.M., Stone, L., Gascuel, O., Rosenberg, N.A.: The behavior of admixed populations in neighbor-joining inference of population trees. Pacific Symposium on Biocomputing **18**, 273–284 (2013) https://doi.org/10.1142/9789814447973_0027

[14] Kim, J., Disanto, F., Kopelman, N.M., Rosenberg, N.A.: Mathematical and simulation-based analysis of the behavior of admixed taxa in the neighbor-joining algorithm. Bulletin of Mathematical Biology **81**(2), 452–493 (2018) https://doi.org/10.1007/s11538-018-0444-0

[15] Eickmeyer, K., Huggins, P., Pachter, L., Yoshida, R.: On the optimality of the neighbor-joining algorithm. Algorithms for Molecular Biology **3**, 5 (2008) https://doi.org/10.1186/1748-7188-3-5

[16] Eickmeyer, K., Yoshida, R.: The geometry of the neighbor-joining algorithm for small trees. In: Horimoto, K., Regensburger, G., Rosenkranz, M., Yoshida, H. (eds.) Algebraic Biology: AB 2008. Lecture Notes in Computer Science Vol. 5147, pp. 81–95. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-85101-1_7

[17] Grayson, D.R., Stillman, M.E.: Macaulay2, a software system for research in algebraic geometry. Available at http://www2.macaulay2.com

[18] Rambau, J.: TOPCOM: Triangulations of point configurations and oriented matroids. In: Mathematical Software, pp. 330–340. World Scientific, River Edge, NJ (2002). https://doi.org/10.1142/9789812777171_0035

[19] De Loera, J.A., Rambau, J., Santos, F.: Triangulations: Structures for Algorithms and Applications. Algorithms and Computation in Mathematics. Springer,

Heidelberg, Germany (2010). https://doi.org/10.1007/978-3-642-12971-1

[20] Dantzig, G.B., Eaves, B.C.: Fourier-Motzkin elimination and its dual. Journal of Combinatorial Theory, Series A **14**(3), 288–297 (1973) https://doi.org/10.1016/0097-3165(73)90004-6

[21] Schrijver, A.: Theory of Linear and Integer Programming. Wiley, Chichester (1986)

[22] Ziegler, G.M.: Lectures on Polytopes. Springer, New York, NY (1995)

[23] Wright, S.: Isolation by distance. Genetics **28**(2), 114–138 (1943) https://doi.org/10.1093/genetics/28.2.114

[24] Holsinger, K.E., Weir, B.S.: Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews Genetics **10**(9), 639–650 (2009) https://doi.org/10.1038/nrg2611

[25] Meirmans, P.G., Hedrick, P.W.: Assessing population structure: F_{ST} and related measures. Molecular Ecology Resources **11**(1), 5–18 (2011) https://doi.org/10.1111/j.1755-0998.2010.02927.x

[26] Gascuel, O., Steel, M.: Neighbor-joining revealed. Molecular Biology and Evolution **23**, 1997–2000 (2006) https://doi.org/10.1093/molbev/msl072

[27] Desper, R., Gascuel, O.: Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. In: Guigó, R., Gusfield, D. (eds.) Algorithms in Bioinformatics, pp. 357–374. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/3-540-45784-4_27

[28] Desper, R., Gascuel, O.: Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Molecular Biology and Evolution **21**(3), 587–598 (2004) https://doi.org/10.1093/molbev/msh049

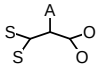
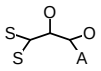
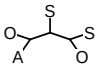
[29] Haws, D.C., Hodge, T.L., Yoshida, R.: Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. Bulletin of Mathematical Biology **73**(11), 2627–2648 (2011) https://doi.org/10.1007/s11538-011-9640-x

[30] Sanderson, M.J., , Shaffer, H.B.: Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics **33**(1), 49–72 (2002) https://doi.org/10.1146/annurev.ecolsys.33.010802.150509

[31] Cueto, M.A., Matsen, F.A.: Polyhedral geometry of phylogenetic rogue taxa. Bulletin of Mathematical Biology **73**(6), 1202–1226 (2011) https://doi.org/10.1007/s11538-010-9556-x

Appendix A Supplementary Tables

Table A1: **$\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\mathbf{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ for $\boldsymbol{\alpha = 0.01, 0.5}$ and 0.99.** The table mirrors the structure of Table 2, with mean volumes replaced by those computed for fixed values of $\alpha = 0.01, 0.5$, and 0.99. For each fixed α and cherry-picking order $\mathcal{O} \in \Omega$, the volumes $\mathrm{Vol}_{\pi_\alpha(C_\mathcal{O})}$ and $\mathrm{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$ were obtained using the direct computation method (Section 3.3.1). The final column, labeled "TC", stands for "Tree Category".

#	$\mathcal{O}$	NJ Tree	$\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})}$			$\mathbf{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$			TC
			$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.99$	$\alpha = 0.01$	$\alpha = 0.5$	$\alpha = 0.99$	
1	(21)(43)		0	0	0	0	0	0	
2	(43)(21)								
3	(21)(54)		0	0	0	0	0	0	
4	(21)(53)								
5	(54)(21)								
6	(53)(21)								
7	(53)(42)		0	0	0	0	0	0	
8	(54)(32)								
9	(53)(41)								
10	(54)(31)								
11	(42)(53)		0.00176	0.00012	0.00003	0.00152	0.00004	0	
12	(32)(54)								
13	(41)(53)		0.00003	0.00012	0.00176	0	0.00004	0.00152	
14	(31)(54)								

#	$\mathcal{O}$	NJ Tree	$\mathbf{Vol}_{\pi_\alpha(C_\mathcal{O})}$			$\mathbf{Vol}_{\pi_\alpha(\widetilde{C}_\mathcal{O})}$			TC
			$\alpha = \mathbf{0.01}$	$\alpha = \mathbf{0.5}$	$\alpha = \mathbf{0.99}$	$\alpha = \mathbf{0.01}$	$\alpha = \mathbf{0.5}$	$\alpha = \mathbf{0.99}$	
15	(51)(42)		0.00195	0.03255	0.1161	0.00195	0.03255	0.11409	
16	(52)(41)		0.1161		0.00195	0.11409		0.00195	
17	(51)(32)		0.00195		0.1161	0.00195		0.11409	
18	(52)(31)		0.1161		0.00195	0.11409		0.00195	
19	(42)(51)		0.00196	0.0328	0.14561	0.00196	0.0328	0.09457	
20	(41)(52)		0.14561		0.00196	0.09457		0.00196	
21	(32)(51)		0.00196		0.14561	0.00196		0.09457	
22	(31)(52)		0.14561		0.00196	0.09457		0.00196	
23	(31)(42)		0.0626	0.08212	0.00417	0.06251	0.08212	0.00417	
24	(32)(41)		0.00417		0.0626	0.00417		0.06251	
25	(42)(31)		0.00417		0.0626	0.00417		0.06251	
26	(41)(32)		0.0626		0.00417	0.06251		0.00417	
27	(52)(43)		0.11734	0.05264	0.00414	0.11455	0.05264	0.00414	
28	(51)(43)		0.00414		0.11734	0.00414		0.11455	
29	(43)(52)		0.14795	0.15218	0.06219	0.14795	0.15218	0.06219	
30	(43)(51)		0.06219		0.14795	0.06219		0.14795	

Appendix B Supplementary Figure

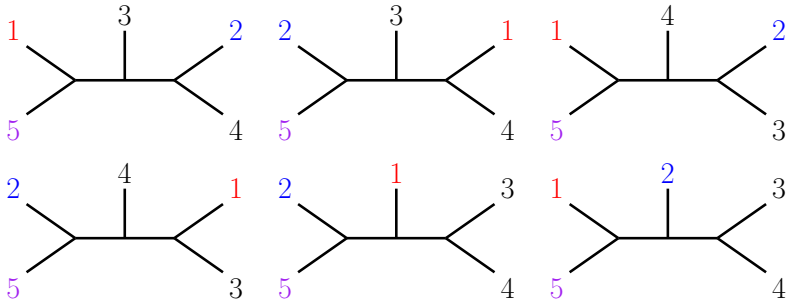


Fig. B1: **Type 1 labeled tree topologies for Property 1.** These six labeled tree topologies correspond to Type-1 NJ cones (Section 3.1.1), where each equivalence class consists entirely of cherry-picking orders that satisfy Property 1. The labeled tree topologies are distinguished based on reflection symmetry with respect to the central taxon.

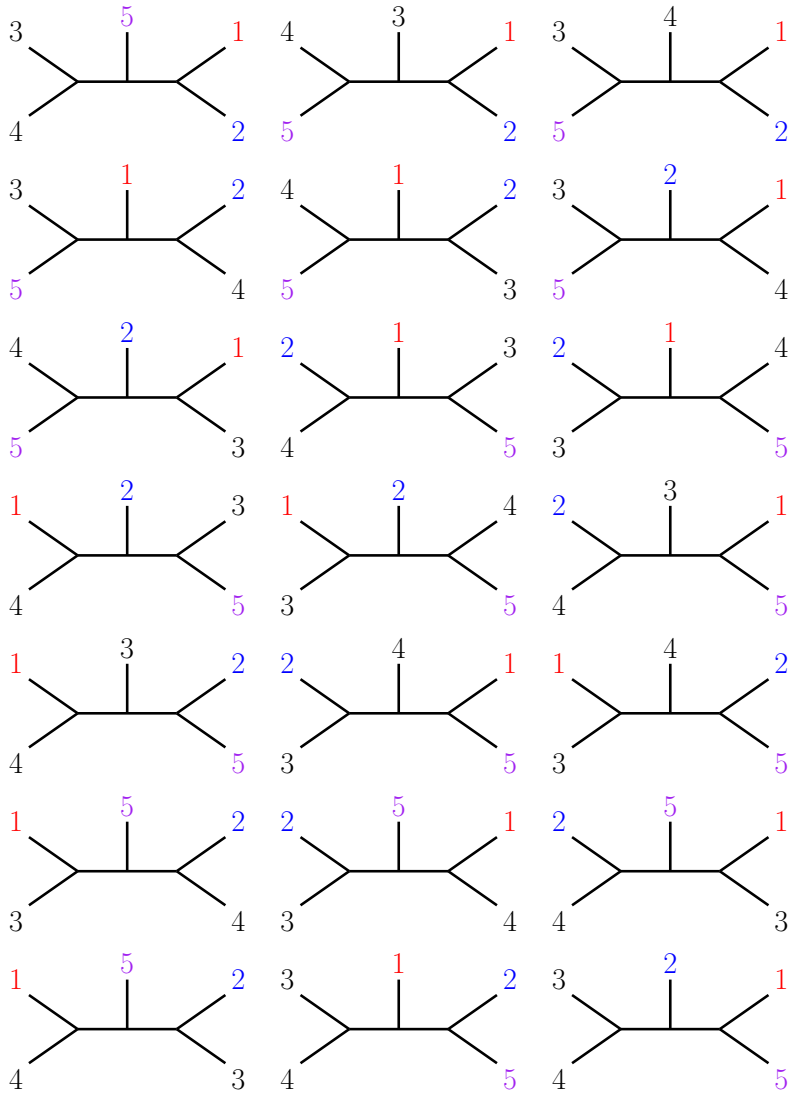


Fig. B2: **Type 2 labeled tree topologies for Property 1.** These 21 labeled tree topologies correspond to Type-2 NJ cones (Section 3.1.1), where each equivalence class contains at least one, but not all, cherry-picking orders that violate Property 1. The labeled tree topologies are distinguished based on reflection symmetry with respect to the central taxon.

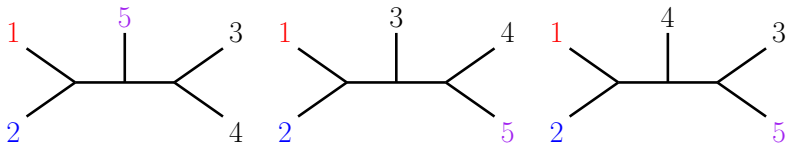


Fig. B3: **Type 3 labeled tree topologies for Property 1.** These three labeled tree topologies correspond to Type-3 NJ cones (Section 3.1.1), where each equivalence class consists entirely of cherry-picking orders that violate Property 1. The labeled tree topologies are distinguished based on reflection symmetry with respect to the central taxon.

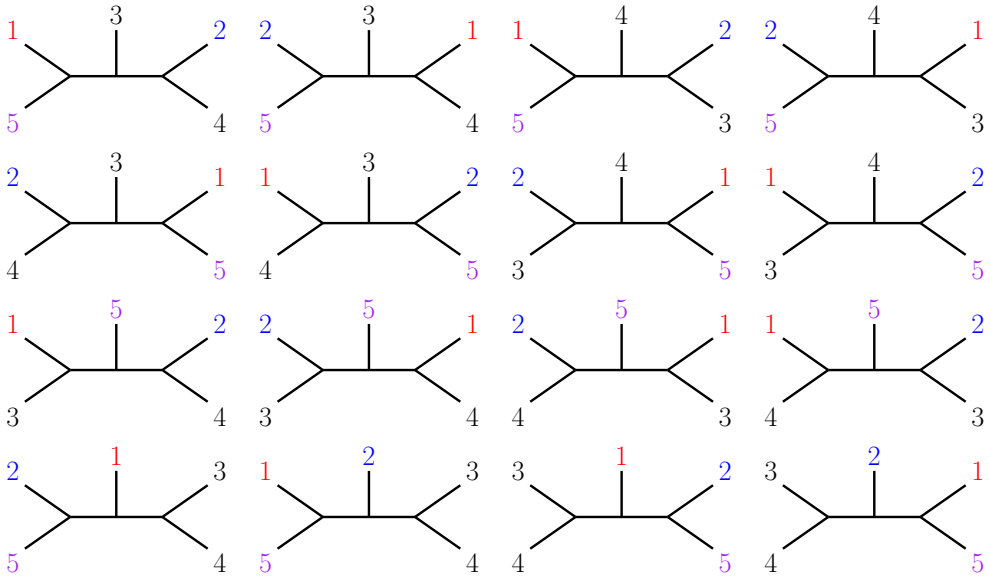


Fig. B4: **Labeled tree topologies that satisfy Property 3.** These 16 labeled tree topologies represent the complete set of NJ cones that satisfy Property 3 (Section 3.1.2). The labeled tree topologies are distinguished by their reflection symmetry relative to the central taxon.

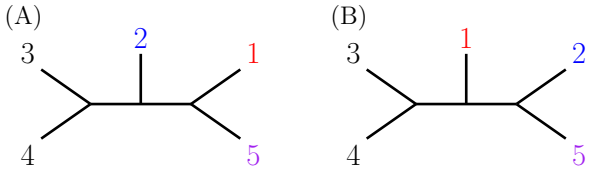


Fig. B5: **Two labeled tree topologies that always satisfy Property 2.** For any admixture fraction $\alpha \in (0, 1)$, $\pi_\alpha(C_{(4,3)(5,1)})$ and $\pi_\alpha(C_{(4,3)(5,2)})$ are the only two induced NJ cones in which every dissimilarity vector $\boldsymbol{d}^{(4)}$ satisfies Property 2. **(A)** Labeled tree topology corresponding to the cone $C_{(4,3)(5,1)}$. **(B)** Labeled tree topology corresponding to the cone $C_{(4,3)(5,2)}$. The labeled tree topologies are distinguished based on reflection symmetry with respect to the central taxon.