SIAM J. APPL. MATH.
Vol. 84, No. 4, pp. 1845–1867

A NOVEL ALGEBRAIC APPROACH TO TIME-REVERSIBLE EVOLUTIONARY MODELS*

MARTA CASANELLAS†, ROSER HOMS‡, AND ANGÉLICA TORRES‡

Abstract. In recent years, algebraic tools have been proven useful in phylogenetic reconstruction and model selection through the study of phylogenetic invariants. However, up to now, the models studied from an algebraic viewpoint are either too general or too restrictive (as group-based models with a uniform stationary distribution) to be used in practice. In this paper we provide a new framework to study time-reversible models, which are the most widely used by biologists. In our approach we consider algebraic time-reversible models on phylogenetic trees (as defined by Allman and Rhodes) and introduce a new inner product to make all transition matrices of the process diagonalizable through the same orthogonal eigenbasis. This framework generalizes the Fourier transform widely used to work with group-based models and recovers some of the well-known results. As illustration, we combine our technique with algebraic geometry tools to provide relevant phylogenetic invariants for trees evolving under the Tamura–Nei model of nucleotide substitution.

Key words. nucleotide substitution model, time-reversible substitution model, phylogenetic variety, phylogenetic invariant

MSC codes. 92D15, 14M99, 62R01

DOI. 10.1137/23M1605302

1. Introduction.

Phylogenetics aims at recovering the evolutionary history of a given set of biological species from certain molecular information. This evolutionary process is represented on a phylogenetic tree or network whose leaves correspond to living species and whose interior nodes represent their common ancestors. One of the most common ways of approaching phylogenetic reconstruction is by modeling the substitution of molecular units (usually nucleotides or amino acids) via a Markov process on a phylogenetic tree.

During the last twenty years, algebraic methods have been developed with the aim of helping biologists address phylogenetic reconstruction. The key is that Markov processes on phylogenetic trees parametrize algebraic varieties, and tools from algebraic geometry turn out to be relevant, as suggested by Felsenstein, Cavender, and Lake in the late eighties; see [9], [25]. They introduced the use of *phylogenetic invariants*, which are polynomial constraints satisfied by any distribution that arises as a hidden Markov process on a phylogenetic tree. These tools avoid parameter inference, which might be a tedious task, and incorporate the geometry of the algebraic varieties to detect the tree that best fits the given data, in a certain measure. Methods based

*Received by the editors September 27, 2023; accepted for publication (in revised form) April 4, 2024; published electronically August 22, 2024.
https://doi.org/10.1137/23M1605302
Funding: This work was supported by the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M) and PID2019-103849GB-I00 funded by MICIU/AEI/10.13039/501100011033. The first author has also been supported by AGAUR project 2021 SGR 00603, Geometry of Manifolds and Applications, GEOMVAP. The second author has received funding from the postdoctoral fellowships program Beatriu de Pinós (ref. 2021BP00119), funded by the Secretary of Universities and Research (Government of Catalonia).
†Departament de Matemàtiques, Universitat Politècnica de Catalunya, Barcelona, 08028, Spain, and Centre de Recerca Matemàtica, Bellaterra, Barcelona, 08193, Spain (marta.casanellas@upc.edu).
‡Centre de Recerca Matemàtica, Bellaterra, Barcelona, 08193, Spain (rhoms@crm.cat, atorres@crm.cat).

on algebraic tools such as SVDquartets [11] or Erik + 2 [16] have been implemented successfully in the phylogenetic software PAUP* [35]. These methods consider the most *general Markov* model of nucleotide substitution (GM, for short). Other models that have been studied by algebraists are G-equivariant models (see [12], [6], [4]), which are friendly models from a mathematical approach but only used by biologists in very special cases.

Markov processes on phylogenetic trees mostly used by biologists have the property of being *stationary*. The GM model on a phylogenetic tree is not a stationary process and, among G-equivariant models, those that are stationary are too simple as their stationary distribution is uniform. Another property that is commonly assumed by biologists is *time-reversibility*. Roughly speaking, a stationary Markov process is time-reversible if, at equilibrium, the rate at which transitions from state i to state j occur is the same as the rate at which transitions from j to i occur. Thus, there is a need to provide algebraic methods for time-reversible processes on phylogenetic trees for any stationary distribution. This can be especially relevant in the case of amino acid substitution models, where the GM model is too large to be of biological utility.

The time-reversibility property has been studied from an algebraic point of view in [2]. Allman and Rhodes tailor time-reversibility from an algebraic approach and define the class of *algebraic time-reversible* (ATR) models. This class contains all time-reversible models that biologists use in their everyday work such as GTR [38], TN93 [37], or HKY85 [18]. In an ATR model on a tree, all transition matrices must commute. This is a natural requirement since it is satisfied by all time-homogeneous continuous-time processes, which are the most widely used in phylogenetic reconstruction.

We build upon this definition of ATR models and develop a new framework that simplifies the study of these models. This can be thought of as a generalization of the well-studied Hadamard or Fourier transform for group-based models exploited in a large list of publications: [14], [36], [19], and [32], among others. First of all, if data has reached equilibrium, the stationary distribution π can be inferred from the data and we can consider it as input data (this approach was already considered by the first author and M. Steel in the study of the Equal-Input model [7]). Then, for a fixed stationary distribution π of an ATR model on a phylogenetic tree, we introduce a new inner product $\langle , \rangle_\pi$ and prove that all transition matrices diagonalize under an orthogonal eigenbasis with respect to $\langle , \rangle_\pi$. By fixing this orthogonal eigenbasis, we are able to do a change of coordinates that simplifies the parametrization of our model. For example, we are able to recover the celebrated result of Evans and Speed [14]. With these new coordinates we can provide phylogenetic invariants for these Markov processes on trees and describe the corresponding algebraic varieties. We illustrate these tools with a deep study of the TN93 model and give phylogenetic invariants that can be used for topology reconstruction or model selection. We focus this study on quartets (i.e., trees with four leaves) because they can be used as a building block in phylogenetic reconstruction by means of quartet-based methods; see [29], for instance.

The structure of the paper is as follows. In section 2 we introduce the preliminaries on Markov processes on phylogenetic trees. In section 3 we develop the framework that allows us to disentangle ATR models on trees: we introduce the inner product $\langle , \rangle_\pi$, prove that ATR models on trees deal with transition matrices that simultaneously diagonalize in a basis that is orthogonal for this inner product, and define algebraic varieties associated to these models. In section 4 we explore the change of coordinates to this eigenbasis and prove the main technique that permits the study of these models on phylogenetic trees from the study on smaller trees (Theorem 4.6). In section 5 we delve into the study of the TN93 model: we give phylogenetic invariants for trees

evolving under this model with any number of leaves and, for quartet trees, we specify a collection of phylogenetic invariants that (locally) cut out the algebraic variety associated to this model. In other words, we give a collection of constraints that suffice to describe distributions evolving under a quartet tree under the TN93 model; see Theorem 5.14. All computations are available in the institutional CORA repository https://dataverse.csuc.cat/dataset.xhtml?persistentId=doi:10.34810/data1128

2. Preliminaries. In this section we give a brief introduction to Markov processes on phylogenetic trees and set up some notation needed throughout the paper. These concepts can be found in [30, Chapters 1 and 8].

Let $L=\{l_1,\dots,l_n\}$ be a finite set of cardinality n (in our setting these elements represent biological entities, such as homologous genes of different species). A *phylogenetic tree* T *on* L is a tree (connected acyclic graph) with leaf set L (that is, the leaf nodes of the graph are in bijection with L). We use $E(T), N(T), Int(T)$ to denote the set of edges, nodes, and interior nodes of T, respectively. We say that T is a *rooted* phylogenetic tree if we specify an interior node r of T and direct all edges away from it. If $e=u\to v$ is an edge on a rooted tree, we say that u is the *parent node* of e (denoted by $p(e)$), and v is the child node of e (denoted by $c(e)$). The set of all phylogenetic trees on L will be denoted by $\mathcal{T}_n$.

Molecular sequences can be thought of as ordered sequences of a finite set of characters or states. We call Σ this finite set of κ states and assume that different positions on the sequence are independent and identically distributed, so that we only model the evolution on one site. For example, we use $\Sigma=\{A,G,C,T\}$ if we consider nucleotide sequences or $\kappa=20$ if we consider amino acid sequences. We denote the elements of Σ by $\{1,\dots,\kappa\}$ for convenience.

We recall how to describe a Markov process on a phylogenetic tree T to model the evolution of molecular sequences along T. At each node v of a rooted phylogenetic tree T we assign a random variable X_v taking values in Σ. We introduce a Markov process on T by defining a parametric statistical model which assumes that each random variable is conditionally independent of its nondescendants given its parent variable [40, section 3]. The parameters of these models are the distribution π^r at the root node r and a Markov or transition matrix M^e for each edge $e=u\to v\in E(T)$. The entry i,j of the Markov matrix M^e stands for the conditional probability of observing state $j\in\Sigma$ at X_v given the observation of state $i\in\Sigma$ at X_u.

By definition, the entries of a Markov matrix are conditional probabilities. However, in this work we extend this term to allow for negative entries. That is, by a $\kappa\times\kappa$ Markov matrix we mean a real square matrix whose rows sum to one. We denote by $\mathcal{M}^1_{\mathbb{R}}$ the set of $\kappa\times\kappa$ Markov matrices (κ will be understood from the context and 1 denotes that the sum of rows is equal to one) and $\mathcal{M}^1_{\mathbb{C}}$ is defined analogously.

A *character* $\boldsymbol{i}$ *on* $N(T)$ is an assignment of states at the nodes of T, that is, $\boldsymbol{i}=(i_v)_{v\in N(T)}$, $i_v\in\Sigma$. If all Markov matrices M^e are nonnegative, the probability of observing a character $\boldsymbol{i}$ at the nodes of T is

$$p^T_{\boldsymbol{i}}=\pi^r_{i_r}\prod_{e\in E(T)}M^e_{i_{p(e)},i_{c(e)}}, \tag{2.1}$$

and this expression can be extended to matrices in $\mathcal{M}^1_{\mathbb{R}}$ or in $\mathcal{M}^1_{\mathbb{C}}$.

If A is a subset of the set of nodes and $\boldsymbol{j}_A=(j_v)_{v\in A}$, $j_v\in\Sigma$, is a collection of states at the nodes of A, we say that a character $\boldsymbol{i}=(i_v)$ on $N(T)$ *extends* $\boldsymbol{j}_A$ if $i_v=j_v$ for all $v\in A$. The set of all characters on $N(T)$ that extend $\boldsymbol{j}_A$ is denoted by $ext(\boldsymbol{j}_A)$. One defines

$$p^T_{\boldsymbol{j}_A} = \sum_{\boldsymbol{i} \in ext(\boldsymbol{j}_A)} p^T_{\boldsymbol{i}} \tag{2.2}$$

as the marginalization over $N(T) \setminus A$. When $A = L(T)$, we denote $p^T_{\boldsymbol{i}_A}$ by $p^T_{i_1 \dots i_n}$ and in this case expression (2.2) can be rewritten as

$$p^T_{i_1 \dots i_n} = \sum_{\substack{v \in Int(T) \\ i_v \in \Sigma}} \pi^r_{i_r} \prod_{e \in E(T)} M^e_{i_{p(e)}, i_{c(e)}}.$$

If F is a field (either $\mathbb{R}$ or $\mathbb{C}$), we call W the F-vector space of dimension κ and call $e^1, \dots, e^\kappa$ its standard basis. This induces a standard basis in the tensor product $\otimes^l W$, $l \geq 1$: $e^{i_1} \otimes \cdots \otimes e^{i_l}$, $i_j \in \Sigma$. We often identify a tensor in $\otimes^l W$ with the column vector formed by its coordinates in the standard basis. In this way, a joint distribution $(p_{1\dots1}, \dots, p_{\kappa\dots\kappa})$ can be identified with a *n-way* tensor in $\otimes^n W$.

If we call $\mathcal{P}$ the set of free parameters of the Markov process, the *(hidden) Markov process on* T is the map

$$\begin{array}{ccc} \mathcal{P} & \xrightarrow{\phi_T} & \bigotimes^n W \\ \pi^r, (M^e)_{e \in E(T)} & \mapsto & p^T = \sum_{i_1, \dots, i_n} p^T_{i_1 \dots i_n} e^{i_1} \otimes \cdots \otimes e^{i_n}, \end{array} \tag{2.3}$$

which assigns the joint distribution at the leaves of the tree to each set of Markov matrices and each π^r. If no further restrictions on the Markov matrices or on the distribution at the root are assumed, then this is called a *general Markov* process on T. We omit the superscript T in p^T when it is understood from the context, and even if we could distinguish whether $F = \mathbb{R}$ or $F = \mathbb{C}$, we talk about the same map ϕ_T.

One of the main constructions for studying the general Markov model from an algebraic viewpoint is the flattening of a tensor. We recall the definition below.

DEFINITION 2.1. *Let $A|B$ be a bipartition of the set of leaves L. Assume that leaves are ordered so that $A = \{1, \dots, m\}$, $B = \{m+1, \dots, n\}$. If $p \in \otimes^n W$, the flattening of p according to the bipartition $A|B$ is the $\kappa^{|A|} \times \kappa^{|B|}$ matrix* $\mathrm{Flat}_{A|B}(p)$ *whose $(i_1 \dots i_m, i_{m+1} \dots i_n)$ entry is $p_{i_1 \dots i_n}$. For any other order of the set of leaves,* $\mathrm{Flat}_{A|B}(p)$ *is defined analogously.*

The following theorem is one of the main tools in algebraic phylogenetics.

THEOREM 2.2 (see [3]). *Let $T \in \mathcal{T}_n$ and let $p \in \mathrm{Im}(\phi_T)$. Then, if $A|B$ is a bipartition induced by removing an edge of T,* $\mathrm{rank}\left(\mathrm{Flat}_{A|B}(p)\right)$ *is bounded above by κ. On the contrary, if $A|B$ cannot be induced by removing any edge of T, the rank of* $\mathrm{Flat}_{A|B}(p)$ *is larger than κ if the parameters that gave rise to p were sufficiently general.*

3. Time-reversible evolutionary models. Let $\Delta^{\kappa-1}$ be the standard simplex in $\mathbb{R}^\kappa$, fix a distribution $\pi \in \Delta^{\kappa-1}$ with nonzero entries, and call D_π the diagonal matrix $diag(\pi)$. Any positive Markov matrix has a unique stationary distribution (a left-eigenvector of eigenvalue 1), and we say that a $\kappa \times \kappa$ Markov matrix M is *π-stationary* if $\pi^t M = \pi^t$.

A Markov matrix M is *π-time-reversible* if $D_\pi M = M^t D_\pi$, that is, $\pi_i M_{i,j} = \pi_j M_{j,i}$ for any i, j. Note that if M is π-time-reversible, then M is π-stationary. In terms of probabilities, the time-reversibility condition means that the probability of observing state i at the parent node and j at the child node of the process governed by M is the same as observing i at the child node and j at the parent node. We introduce an inner product that gives another way of expressing time-reversibility.

DEFINITION 3.1. *The* π-inner product *of* $u, v \in W$ *is*

$$\langle u, v\rangle_\pi := \sum_i \frac{1}{\pi_i} u_i v_i = u^t D_\pi^{-1} v,$$

where u_i *and* v_i*, for* $i = 1, \ldots, \kappa$*, are the coordinates of* u, v *in the standard basis.*

Then M is π-time-reversible if and only if M^t is a self-adjoint matrix with respect to this inner product (that is, $\langle M^t u, v\rangle_\pi = \langle u, M^t v\rangle_\pi$ for any u, v). In particular, thanks to the Spectral Theorem, all eigenvalues of M are real and there exists a basis of eigenvectors of M^t which is orthogonal with respect to $\langle\ ,\ \rangle_\pi$. An orthogonal basis with respect to $\langle\ ,\ \rangle_\pi$ will be called a π*-orthogonal basis.*

Remark 3.2. Be aware that in [26, section 12] a similar inner product $\langle\ ,\ \rangle'$ is defined but using D_π instead of D_π^{-1}. We have that M^t is self-adjoint with respect to $\langle\ ,\ \rangle_\pi$ if and only if M is self-adjoint with respect to $\langle\ ,\ \rangle'$. We defined the inner product this way because we are interested in eigenvectors of M^t instead of M (as Markov matrices act to the right of row vectors).

A Markov process on a phylogenetic tree is π-time-reversible if all its transition matrices are π-time-reversible and π is the distribution at the root. In [2], all transition matrices of the process are assumed to commute with each other to say that the process is *algebraic time-reversible* (ATR). This extra assumption is equivalent to saying that all matrices simultaneously diagonalize (if they are diagonalizable), and it is an implicit assumption when one considers continuous time-reversible models that are homogeneous over time (that is, for any edge e, $M^e = exp(t_e Q)$ for a certain rate matrix Q). As these are the most widely used processes in phylogenetic software, in this paper we consider ATR processes on trees.

If we have an ATR process with stationary distribution π, then there exists a π-orthogonal basis

$$B = \{u^1, \ldots, u^\kappa\},$$

which diagonalizes all transpose matrices $(M^e)^t$, $e \in E(T)$. As π is a left-eigenvector with eigenvalue 1 for each π-time-reversible Markov matrix M^e, we can assume $u^1 = \pi$. In particular, $\langle u^1, u^1\rangle_\pi = 1$ and $\langle u^1, e^i\rangle_\pi = 1$ for any $i = 1, \ldots, \kappa$.

DEFINITION 3.3. *Let* $\pi \in \Delta^{\kappa-1}$ *be a fixed distribution with nonzero entries, and let* $B = \{u^1 = \pi, u^2, \ldots, u^\kappa\}$ *be a* π*-orthogonal basis in* $\mathbb{R}^\kappa$*. A phylogenetic tree* T evolves under a B-time-reversible model *if all Markov matrices* M^e*,* $e \in E(T)$*, on the Markov process on* T *have* B *as a left eigenbasis and* $\pi^r = \pi$*.*

The following lemma guarantees that a B-time-reversible model on a phylogenetic tree is an ATR process. Before proving it, we introduce some notation. Throughout this paper we denote by $\mathbf{1}$ the vector $\sum_{j=1}^{\kappa} e^j$ and let A be the change of basis matrix from B to the standard basis $e^1, \ldots, e^\kappa$, that is, $A = \begin{pmatrix} u^1 & \ldots & u^\kappa \end{pmatrix}$. As B is π-orthogonal we have

$$A^t D_\pi^{-1} A = S, \tag{3.1}$$

where S is the diagonal matrix $\mathrm{diag}(\langle u^1, u^1\rangle_\pi, \ldots, \langle u^\kappa, u^\kappa\rangle_\pi)$.

LEMMA 3.4. *Let* $\pi \in \Delta^{\kappa-1}$ *be a distribution with positive entries,* $B = \{u^1 = \pi, \ldots, u^\kappa\}$ *a* π*-orthogonal basis in* $\mathbb{R}^\kappa$*, and* M *a* $\kappa \times \kappa$ *matrix for which* B *is a left eigenbasis.*

Then $D_\pi M = M^t D_\pi$ and M has constant row sum; moreover, the first eigenvalue λ_1 is equal to one if and only if M is a π-time-reversible Markov matrix (and in this case π is a stationary distribution for M).

Proof. If B is a left eigenbasis for M, we have $M = A^{-t}\Lambda A^t$ for some diagonal matrix Λ. Thus, $D_\pi M = D_\pi A^{-t}\Lambda A^t$, which equals $AS^{-1}\Lambda A^t$ by (3.1). As S^{-1} and Λ commute, applying (3.1) again we have $D_\pi M = A\Lambda A^{-1} D_\pi$, which is $M^t D_\pi$, as we wanted to prove.

Note that as $\langle u^1, u^i\rangle_\pi = 0$ for any $i \neq 1$, u^i has sum of coordinates equal to 0. Thus, the first column of A adds to 1 and the other columns add to 0. In particular, if $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_\kappa)$, we have

$$M\mathbf{1} = A^{-t}\Lambda A^t \mathbf{1} = A^{-t}\Lambda e^1 = \lambda_1 A^{-t} e^1 = \lambda_1 D_\pi^{-1} A e^1 = \lambda_1 D_\pi^{-1} u^1 = \lambda_1 \mathbf{1}.$$

Thus, M has constant row sum. Requiring sum of rows equal 1 on a square matrix is equivalent to saying that $\mathbf{1}$ is an eigenvector of eigenvalue 1, so the last claim also follows easily. ∎

Example 3.5 (Tamura–Nei model, TN93). Tamura and Nei presented in [37] a continuous-time model based on the observed changes in human mitochondrial DNA. They proposed an arbitrary stationary distribution π and observed that probabilities of transitions (changes within purines or within pyrimidines) and transversions (changes between purines and pyrimidines) depend on the frequencies of the obtained nucleotide and on a single parameter for transversions and two for transitions. This is a time-reversible model whose transition matrices have the form

$$M = \begin{pmatrix} *_1 & \pi_2 c & \pi_3 b & \pi_4 b \\ \pi_1 c & *_2 & \pi_3 b & \pi_4 b \\ \pi_1 b & \pi_2 b & *_3 & \pi_4 d \\ \pi_1 b & \pi_2 b & \pi_3 d & *_4 \end{pmatrix}, \tag{3.2}$$

where $*_i$ is chosen so that each row sums to 1. Here we identified Σ with the set of nucleotides $\{A, G, C, T\}$, in this order. The matrix M^t has the following π-orthogonal basis B of eigenvectors:

$$B = \left\{ u^1 = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}, u^2 = \begin{pmatrix} \pi_1\pi_{34} \\ \pi_2\pi_{34} \\ -\pi_3\pi_{12} \\ -\pi_4\pi_{12} \end{pmatrix}, u^3 = \frac{1}{\pi_{34}} \begin{pmatrix} 0 \\ 0\pi_3\pi_4 \\ -\pi_3\pi_4 \end{pmatrix}, u^4 = \frac{1}{\pi_{12}} \begin{pmatrix} \pi_1\pi_2 \\ -\pi_1\pi_2 \\ 0 \\ 0 \end{pmatrix} \right\},$$

where $\pi_{12} = \pi_1 + \pi_2$ and $\pi_{34} = \pi_3 + \pi_4$. If the columns of A are the vectors of B, then

$$M^t = A \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_3 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix} A^{-1}, \tag{3.3}$$

where $\lambda_1 = 1$, $\lambda_2 = \lambda_1 - b$, $\lambda_3 = \lambda_1 - \pi_{12}b - \pi_{34}d$, $\lambda_4 = \lambda_1 - \pi_{34}b - \pi_{12}c$ are the eigenvalues of M. The entries of M can be written in terms of the eigenvalues using that

$$b = \lambda_1 - \lambda_2, \quad c = \frac{\pi_{12}\lambda_1 + \pi_{34}\lambda_2 - \lambda_4}{\pi_{12}}, \quad d = \frac{\pi_{34}\lambda_1 + \pi_{12}\lambda_2 - \lambda_3}{\pi_{34}},$$

$$*_1 = \frac{\pi_1}{\pi_{12}}\left(\pi_{12}\lambda_1 + \pi_{34}\lambda_2 + \frac{\pi_2}{\pi_1}\lambda_4\right), \quad *_2 = \frac{\pi_2}{\pi_{12}}\left(\pi_{12}\lambda_1 + \pi_{34}\lambda_2 + \frac{\pi_1}{\pi_2}\lambda_4\right),$$

$$*_3 = \frac{\pi_3}{\pi_{34}}\left(\pi_{34}\lambda_1 + \pi_{12}\lambda_2 + \frac{\pi_4}{\pi_2}\lambda_2\right), \quad *_4 = \frac{\pi_4}{\pi_{34}}\left(\pi_{34}\lambda_1 + \pi_{12}\lambda_2 + \frac{\pi_3}{\pi_4}\lambda_3\right).$$

A matrix M satisfying (3.3) is an *algebraic* TN93 matrix (and we do not necessarily assume $\lambda_1 = 1$). If we impose $c = d$, then we have an HKY85 matrix [18], and if we impose $b = c = d$, we obtain the Equal-Input model (EI) [15], [7]. If we adopt the extra assumption that the stationary distribution is uniform, $\pi = (1/4, 1/4, 1/4, 1/4)$, we recover the RY3.3c model of [39]. Hence, TN93 is a B-time-reversible model and so are its submodels HKY85, EI, and RY3.3c.

Note that

$$\langle u^1, u^1\rangle_\pi = 1, \quad \langle u^2, u^2\rangle_\pi = \pi_{12}\pi_{34}, \quad \langle u^3, u^3\rangle_\pi = \frac{\pi_3\pi_4}{\pi_{34}}, \quad \langle u^4, u^4\rangle_\pi = \frac{\pi_1\pi_2}{\pi_{12}}.$$

The following table gives the π-inner product among the vectors in the standard basis and the basis B:

(3.4)

$\langle\cdot,\cdot\rangle_\pi$	u^1	u^2	u^3	u^4
e^1	1	π_{34}	0	π_2/π_{12}
e^2	1	π_{34}	0	$-\pi_1/\pi_{12}$
e^3	1	$-\pi_{12}$	π_4/π_{34}	0
e^4	1	$-\pi_{12}$	$-\pi_3/\pi_{34}$	0

For any B-time-reversible model, as the standard basis is also a π-orthogonal basis, we have that $\langle u^i, e^j\rangle_\pi$ is the jth coordinate of u^i (in the standard basis) divided by π_j (because $\langle e^j, e^j\rangle_\pi = 1/\pi_j$).

Remark 3.6. Note that there are B-time-reversible models that are not *multiplicatively closed*. This important property has been argued to be needed for consistency of phylogenetic inference; see [34]. For instance, regarding the models in the previous example, while TN93 is multiplicatively closed, its submodel HKY85 is not (the product of two HKY85 matrices with the same stationary distribution π is not necessarily an HKY85 matrix).

Example 3.7. The well-known Kimura models with 2 or 3 parameters [23], [24] and the Jukes–Cantor model [21] are also instances of ATR models. All these models have uniform stationary distribution π and are B-time-reversible models with

$$B = \left\{ u^1 = \frac{1}{4}\begin{pmatrix}1\\1\\1\\1\end{pmatrix}, u^2 = \frac{1}{4}\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix}, u^3 = \frac{1}{4}\begin{pmatrix}1\\-1\\1\\-1\end{pmatrix}, u^4 = \frac{1}{4}\begin{pmatrix}1\\-1\\-1\\1\end{pmatrix} \right\}. \tag{3.5}$$

Working with coordinates in this basis simplifies the parametrization map, as already noted by [14]. For these models this technique is also known as a discrete Fourier transform or Hadamard transform; see [19], for instance.

3.1. Phylogenetic algebraic varieties. As above, let π be a fixed positive stationary distribution and let $B = \{u^1 = \pi, \ldots, u^\kappa\}$ be a π-orthogonal eigenbasis. We call A the matrix of change of basis from B to the standard basis $e^1, \ldots, e^\kappa$.

Remark 3.8 (rerooting). If we have a B-time-reversible model on a phylogenetic tree T, one can chose any node of T to play the role of the root and use the same transition matrices to describe the Markov process on T. Indeed, let us prove that we can change the root from node r to an adjacent node s without changing transition matrices. Let e_0 be the edge from r to s. Expression (2.1) can be written as

$$p_{\boldsymbol{i}} = \pi_{i_r} M^{e_0}_{i_r,i_s} \prod_{e\in E(T), e\neq e_0} M^e_{i_{p(e)},i_{c(e)}}.$$

As M^{e_0} is π-time-reversible, $\pi_{i_r} M^{e_0}_{i_r,i_s} = \pi_{i_s} M^{e_0}_{i_s,i_r}$, so expression (2.1) is still valid if we root the tree at s.

From now on we do not specify the placement of a root (we conveniently place a root at any node if necessary). However, the distribution π at the root node r in expression (2.1) is necessary: we cannot incorporate this distribution into one of the transition matrices on the edges adjacent to r (as is usually done in the general Markov model or for G-equivariant models) because by doing so we would get a new matrix not belonging to the ATR model.

As we are considering a B-time-reversible model, we have $\pi^r = \pi$ and the entries of π^r are not free parameters anymore in the expression (2.3). Thus, by extending to the complex numbers field, the map ϕ_T is defined on the parameter set

$$\mathcal{P}^{\mathbb{C}} = \{(M^e)_{e\in E(T)} \mid M^e \in \mathcal{M}^1_{\mathbb{C}}, (M^e)^t = AD^eA^{-1}\},$$

where each D^e is a diagonal matrix whose first entry is 1, and the map ϕ_T is

$$\begin{array}{rcl} \mathcal{P}^{\mathbb{C}} & \xrightarrow{\phi_T} & \bigotimes^n W \\ (M^e)_{e\in E(T)} & \mapsto & \displaystyle\sum_{i_1,\dots,i_n} p^T_{i_1\dots i_n} e^{i_1}\otimes\cdots\otimes e^{i_n}\,, \end{array}$$

where $W = \langle e^1,\dots,e^\kappa\rangle_{\mathbb{C}}$ is a $\mathbb{C}$-vector space.

DEFINITION 3.9. *The* phylogenetic variety *of a tree T evolving under a B-time-reversible model is the Zariski closure $\mathcal{V}_T$ of* $\operatorname{Im}\phi_T$ *in the tensor space $\otimes^n W$. We denote by $\mathcal{I}_T \subset \mathbb{C}[p_{1\dots1},\dots,p_{\kappa\dots\kappa}]$ the ideal of this algebraic variety. Its elements are called* phylogenetic invariants.

The main goal of this work is to give phylogenetic invariants for ATR models. As ATR models are submodels of the general Markov model, phylogenetic invariants for the general Markov model are also in $\mathcal{I}_T$. Thus, Theorem 2.2 holds and the $(\kappa+1)\times(\kappa+1)$ minors of those flattening matrices are phylogenetic invariants.

Remark 3.10 (degree two nodes). Assume that T has a degree two node s and let M^{e_1} and M^{e_2} be transition matrices at the edges incident to it. Then the image by ϕ_T of these parameters coincides with the one obtained by deleting node s, joining e_1 and e_2 in a new edge e_0, and considering the matrix $M^{e_0} = M^{e_1}M^{e_2}$ at e_0. Therefore, adding or removing degree two nodes in a tree will not affect the map ϕ_T (when we add a degree two node on an edge and split it into two edges, we can trivially put the identity matrix at one of these edges).

The map ϕ_T parametrizes a dense subset of $\mathcal{V}_T$. According to the result of [10] and its generalization in [1], if T has no nodes of degree two, the fibers of ϕ_T are finite. Therefore the dimension of $\mathcal{V}_T$ coincides with the dimension of the space of parameters, which is $(\kappa-1)|E(T)|$ in this case.

A special point in $\operatorname{Im}\phi_T$ is the image of identity transition matrices. This is called the *no-evolution point* in [6] and is denoted by $p^0 = \phi_T(\{Id\}_{e\in E(T)})$. This point has a special relevance: in biological applications, transition matrices should not be far from identity (because it is difficult to obtain reliable data evolving on a tree with transition matrices far from the identity), so the points in $\mathcal{V}_T$ of most interest (biologically speaking) are those close to p^0.

In terms of probabilities it is easy to see that the image of ϕ_T lies in the hyperplane

$$H = \left\{ p \in \otimes^n W \mid \sum_{\boldsymbol{i}} p_{\boldsymbol{i}} = 1 \right\}, \tag{3.6}$$

so we also have $\mathcal{V}_T \subset H$. Thus, $\sum_{\boldsymbol{i}} p_{\boldsymbol{i}} - 1$ is a (trivial) phylogenetic invariant.

We make the previous map homogeneous by extending it to square matrices that diagonalize through A, without imposing $\lambda_1 = 1$. Let $C\mathcal{V}_T$ be the cone over $\mathcal{V}_T$; then $C\mathcal{V}_T$ is the Zariski closure of the following homogeneous map:

$$\begin{array}{ccc} \tilde{\mathcal{P}}^{\mathbb{C}} & \xrightarrow{\psi_T} & \bigotimes^n W \\ (M^e)_{e\in E(T)} & \mapsto & \sum_{i_1,\dots,i_n} p^T_{i_1\dots i_n} e^{i_1}\otimes\cdots\otimes e^{i_n}\,, \end{array}$$

where $\tilde{\mathcal{P}}^{\mathbb{C}} = \{(M^e)_{e\in E(T)} \mid (M^e)^t = AD^eA^{-1}\}$ and $p^T_{i_1,\dots,i_n}$ is defined by the same expression as (2.2). Actually, $\mathcal{V}_T = C\mathcal{V}_T \cap H$ (see also [3], [5]). Indeed, if the row sum of a matrix M^e is λ_1^e (see Lemma 3.4), then M^e is the product of λ_1^e by a Markov matrix $\tilde{M}^e$ whose rows sum to one. Hence in (2.1) we have $p^T_{\mathbf{i}} = (\prod_{e\in E(T)}\lambda_1^e)\pi_{i_r}\prod_{e\in E(T)}\tilde{M}^e_{i_{p(e),c(e)}}$, and if $s := \prod_{e\in E(T)}\lambda_1^e$, we have $\psi_T(\{M^e\}) = s\phi_T(\{\tilde{M}^e\})$. We extend the definition of a B-time-reversible model on a phylogenetic tree (Definition 3.3) in order to allow all matrices M^e to be in $\tilde{\mathcal{P}}^{\mathbb{C}}$.

We could do a change of coordinates in the parameter space $\tilde{\mathcal{P}}^{\mathbb{C}}$: instead of dealing with the entries of M^e we could deal directly with its eigenvalues $\lambda_1^e,\dots,\lambda_\kappa^e$. Thus, we could also express $p^T_{i_1,\dots,i_n}$ in terms of the eigenvalues of M^e's. This change in the parameter space and the analogous change of coordinates in the target space will be studied in the next section.

4. New coordinates for phylogenetic varieties of ATR models. From now on, the π-inner product will be simply denoted by $\langle\,,\,\rangle$. This inner product was introduced on $W = \mathbb{R}^n$ but can be extended naturally to any tensor power $\otimes^n W = W\otimes\cdots\otimes W$ as

$$\langle w_1\otimes\cdots\otimes w_n, v_1\otimes\cdots\otimes v_n\rangle = \langle w_1,v_1\rangle\langle w_2,v_2\rangle\dots\langle w_n,v_n\rangle.$$

Actually, it can also be extended to the complex number field by taking the conjugate of the second component in the inner product. However, we do not introduce this notation because in all inner products we will use, the second component will be a vector with real coordinates. Thus, we can think of $\langle w,v\rangle$ with the definition we have already introduced and take w in $\otimes^n\mathbb{C}^\kappa$ and $v\in\otimes^n\mathbb{R}^\kappa$.

Let $B = \{u^1 = \pi,\dots,u^\kappa\}$ be a π-orthogonal eigenbasis. Then the basis

$$B_n = \{u^{i_1}\otimes u^{i_2}\otimes\cdots\otimes u^{i_n} \mid i_j\in\Sigma\}$$

is a π-orthogonal basis of $\otimes^n W$. To simplify notation we call

$$u^{\boldsymbol{i}} = u^{i_1}\otimes\cdots\otimes u^{i_n} \text{ and } e^{\boldsymbol{i}} = e^{i_1}\otimes\cdots\otimes e^{i_n}$$

if $\boldsymbol{i} = (i_1,\dots,i_n)\in\Sigma^n$. Let A be the $\kappa\times\kappa$ matrix of change of basis from B to the standard basis as in the previous section.

If $p\in\otimes^n W$ and $p_{i_1\dots i_n}$ are its coordinates in the standard basis, then its coordinates in the basis B_n shall be denoted by $\bar{p}$ and are obtained as $\bar{p} = (A^{-1}\otimes\cdots\otimes A^{-1})\,p$, where $\otimes$ denotes the Kronecker product of matrices. That is, for $\boldsymbol{i} = (i_1,\dots,i_n)$, the $\boldsymbol{i}$-coordinate of the tensor p in the basis B_n, $\bar{p}_{i_1\dots i_n}$, is the $i_1\dots i_n$ entry of the vector $\bar{p}$. Since B_n is a π-orthogonal basis, this coordinate $\bar{p}_{i_1\dots i_n}$ can also be computed as

$$\bar{p}_{i_1\dots i_n} = \frac{\langle p, u^{\boldsymbol{i}}\rangle}{\langle u^{\boldsymbol{i}}, u^{\boldsymbol{i}}\rangle}\,. \tag{4.1}$$

Reparametrization. Let T be a tree evolving under a B-time-reversible model. If we change coordinates on the parameter space $\widetilde{\mathcal{P}}^{\mathbb{C}}$ so that (transposes of) transition matrices are diagonalized (and hence written in the basis B) and use coordinates $\bar{p}$ in the target space of ψ_T, we have a much simpler parametrization of $C\mathcal{V}_T$:

$$\begin{array}{ccc} \prod_{e\in E(T)} \mathbb{C}^\kappa & \xrightarrow{\varphi_T} & \bigotimes^n W \\ (\Lambda^e)_{e\in E(T)} & \mapsto & \sum_{i_1,\dots,i_n} \bar{p}^T_{i_1\dots i_n} u^{i_1}\otimes\cdots\otimes u^{i_n}, \end{array}$$

where $\Lambda^e = \operatorname{diag}(\lambda^e_1,\dots,\lambda^e_\kappa)$ is the diagonal matrix formed by the eigenvalues of M^e, $e\in E(T)$. We denote the Zariski closure of the image of φ_T by CV_T.

DEFINITION 4.1. *We denote by I_T the ideal of CV_T in $\mathbb{C}[\bar{p}_{1\dots1},\dots,\bar{p}_{\kappa\dots\kappa}]$. A polynomial that belongs to all I_T for $T\in\mathcal{T}_n$ is called a* model invariant *(as it holds for any tree evolving under the B-time-reversible model). A polynomial that belongs to some I_T for $T\in\mathcal{T}_n$ but that does not belong to $I_{T'}$ for some $T'\in\mathcal{T}_n$ is called a* topology invariant*; see* [30, *Chapter* 8].

From the computational point of view, if we want to work with any π, we can work with polynomials with coefficients in the field of fractions $\mathbb{C}(\pi_1,\dots,\pi_\kappa)$. Our computations in `Macaulay2` [17] follow this approach.

Using these new coordinates $\bar{p}$ in the basis B_n, the following result relating marginalization and new coordinates will be useful.

LEMMA 4.2 (marginalization). *For any $p\in\otimes^n W$ define the marginalization $p^+\in\oplus^{n-1}W$ of p over the last component as*

$$p^+_{i_1\dots i_{n-1}} = \sum_{j\in\Sigma} p_{i_1\dots i_{n-1}j}. \tag{4.2}$$

Then, for a π-orthogonal basis B_n as above, we have

$$\bar{p}_{i_1\dots i_{n-1}1} = \overline{p^+}_{i_1\dots i_{n-1}}. \tag{4.3}$$

Furthermore, if T_n is an n-leaf tree and T_{n-1} is the tree obtained from T_n by deleting the pendant edge leading to leaf n, then for any $p\in\operatorname{Im}(\phi_{T_n})$ we have that $p^+\in\operatorname{Im}(\phi_{T_{n-1}})$.

Proof. Note that the (i,j)-entry of A^{-1} is the ith coordinate of e^j in the basis B, which is $\frac{\langle e^j,u^i\rangle}{\langle u^i,u^i\rangle}$. As $\langle u^1,u^1\rangle = 1$ and $\langle e^i,u^1\rangle = 1$ for any $i=1,\dots,\kappa$, the first row of A^{-1} is $\mathbf{1}^t$. Thus, if $p\in\otimes^n W$, the slice of $\bar{p}$ with last component indexed by 1 is

$$(A^{-1}\otimes\overset{n-1}{\cdots}\otimes A^{-1}\otimes\mathbf{1}^t)\,p,$$

which is equal to $(A^{-1}\otimes\overset{n-1}{\cdots}\otimes A^{-1})\,p^+$. Thus, $\bar{p}_{i_1\dots i_{n-1}1} = \overline{p^+}_{i_1\dots i_{n-1}}$. The last claim is well known and follows directly from [40, Proposition 5.52]. ∎

Markov action. The following action of $GL_\kappa(\mathbb{C})^n$ on tensors in $\otimes^n W$,

$$(N_1,\dots,N_n)\cdot p = (N_1\otimes\dots\otimes N_n)p,$$

is called the *Markov action.* We can restrict this action to diagonal matrices so that we have an action of an $n\kappa$-dimensional torus $\mathbb{T} = (\mathbb{C}^*)^\kappa\times\overset{n}{\cdots}\times(\mathbb{C}^*)^\kappa$. The torus $\mathbb{T}$ acts on tensors with coordinates $\bar{p}$ as follows: if $(D^1,\dots,D^n)$ is in $\mathbb{T}$ and $D^i = \operatorname{diag}(d^i_1,\dots,d^i_\kappa)$, then $(D^1,\dots,D^n)\cdot\bar{p}$ has coordinates $d^1_{i_1}\dots d^n_{i_n}\bar{p}_{i_1\dots i_n}$.

If $\bar{p}=\varphi_T((\Lambda^e)_{e\in E(T)})$, then $(D^1,\dots,D^n)\cdot\bar{p}=\varphi_T((\tilde{\Lambda}^e)_{e\in E(T)})$, where $\tilde{\Lambda}^{e_i}=D^i\Lambda^{e_i}$ if e_i is the pendant edge to leaf l_i, and $\tilde{\Lambda}^e=\Lambda^e$ otherwise. Hence we have that the Zariski closure of $\mathbb{T}\cdot CV_T$ is again CV_T, that is, CV_T is invariant by the action of $\mathbb{T}$.

Remark 4.3. Let $\bar{p}=\varphi_T(\{\Lambda^e\}_e)$. If $\bar{q}\in\mathrm{Im}\varphi_T$ has the same parameters as $\bar{p}$ except the matrices Λ^{e_i} on the pendant edges are replaced by identity matrices, then

$$\bar{p}=(\Lambda^{e_1},\dots,\Lambda^{e_n})\cdot\bar{q},\quad \text{i.e.,}\ \ \bar{p}_{i_1\dots i_n}=\lambda_{i_1}^{e_1}\dots\lambda_{i_n}^{e_n}\bar{q}_{i_1\dots i_n}. \tag{4.4}$$

Throughout the paper $\bar{q}$ will denote the image by φ_T of a set of parameters with identity matrices at the pendant edges.

4.1. Star trees evolving under ATR models. In the following lemma we prove that if T is a star tree, then CV_T is a toric variety.

LEMMA 4.4. *Let T be the star tree with n leaves and let it evolve under a B-time-reversible model. Then CV_T is a toric variety (not necessarily normal), φ_T is a monomial parametrization, and $I_T\subset\mathbb{C}[\bar{p}_{i_1\dots i_n}\mid i_j\in\Sigma]$ is generated by binomials. Moreover the no-evolution point p^0 is a nonsingular point of $C\mathcal{V}_T$ and $\mathcal{V}_T$.*

Proof. By Remark 4.3, we know that CV_T is the closure of the orbit of $p^0=\varphi_T(\{Id\})$ under the action of $\mathbb{T}$. This implies that CV_T (and hence $C\mathcal{V}_T$) is a toric variety and p^0 is a nonsingular point of $C\mathcal{V}_T$ and $\mathcal{V}_T$. Again from Remark 4.3 we have that the parametrization φ_T is monomial on the eigenvalues of Λ^i, given that the coordinates of p^0 in the basis B_n are expressions in terms of π only. From this, we obtain that the ideal I_T can be generated by binomials (see [13]). ∎

For G-equivariant models it was already known that no-evolution points are nonsingular points of star trees [6, Corollary 3.9]. The proof of [6, Theorem 5.4] shows that p^0 is a nonsingular point on *any* tree evolving under a G-equivariant model.

4.2. Gluing trees. We recall here a procedure to glue trees and substitution parameters that was introduced in [3].

Gluing trees. Let T_1 and T_2 be two phylogenetic trees with leaf sets $\{l_1,\dots,l_m,s_1\}$ and $\{s_2,l_{m+1},\dots,l_n\}$, respectively. We call T' the tree with leaf set $\{l_1,\dots,l_n\}$ obtained by identifying s_1 and s_2 in a node s. We then call $T=T_1{*}T_2$ the tree obtained by deleting this node s and replacing the two edges e_1,e_2 incident to it by a single edge e_0; see Figure 1. We call $\alpha=\{l_1,\dots,l_m\}$ and $\beta=\{l_{m+1},\dots,l_n\}$, so that T has leaf set $L=\alpha\cup\beta$.

Gluing parameters. If T_1 and T_2 evolve under the B-time-reversible model with matrices $(M^e)_{e\in E(T_1)\cup E(T_2)}$, then we define transition matrices at the edges of $T=T_1*T_2$ as follows: if $e\neq e_0$, then $e\in E(T_i)$ for some $i=1,2$ and we assign to e the transition matrix as in T_i; if $e=e_0$, we let $M^e=M^{e_1}M^{e_2}$.

According to Remark 3.10, we can indistinguishably use the tree $T=T_1{*}T_2$ with the set of parameters just described, or the tree T' with the degree two node s and parameters $(M^e)_{e\in E(T_1)\cup E(T_2)}$.

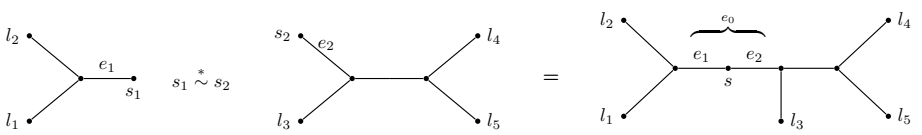

FIG. 1. *Gluing of a tripod and a quartet. The result is a tree with* 5 *leaves. In this case* $m=2,n=5,\alpha=\{l_1,l_2\}$, *and* $\beta=\{l_3,l_4,l_5\}$.

The following lemma is obvious if we think about distributions, but for the sake of completeness we prove it in section SM1 of the supplementary material for any set of parameters.

LEMMA 4.5. *Let T be obtained by gluing two trees T_1 and T_2 as above, let $p^{T_i} \in \operatorname{Im}\psi_{T_i}$, $i=1,2$, and let p^T be the tensor obtained by gluing parameters on T_1 and T_2. Let $\boldsymbol{i}=(\boldsymbol{i}_\alpha,\boldsymbol{i}_\beta)$ be a collection of states at the leaves of T. Then for any state $k\in\Sigma$ at node $s=s_1\sim s_2$ we have*

$$p^T_{\boldsymbol{i}_\alpha,k,\boldsymbol{i}_\beta}=\frac{1}{\pi_k}p^{T_1}_{\boldsymbol{i}_\alpha,k}p^{T_2}_{k,\boldsymbol{i}_\beta}.$$

The following theorem expresses a tensor evolving on $T=T_1*T_2$ under a B-time-reversible model in terms of tensors evolving on T_i. Here $\operatorname{Flat}(\bar{p})$ is as in Definition 2.1 exchanging coordinates in the standard basis by coordinates in B_n.

THEOREM 4.6. *Let T_1 and T_2 be two trees evolving under a B-time-reversible model, and let $T=T_1*T_2$ be the tree obtained by gluing T_1 and T_2 as above. Let $p^{T_i}\in\operatorname{Im}\psi_{T_i}$, $i=1,2$, and let p^T be the tensor obtained by gluing parameters on T_1 and T_2. Then, in coordinates in the basis B_n, we have*

$$\bar{p}_{i_1\dots i_n}=\sum_{j\in\Sigma}\langle u^j,u^j\rangle\,\bar{p}^{T_1}_{i_1\dots i_m j}\bar{p}^{T_2}_{j i_{m+1}\dots i_n} \tag{4.5}$$

for any $i_1,\dots,i_n\in\Sigma$. If B is a π-orthonormal eigenbasis, then the expression becomes

$$\bar{p}_{i_1\dots i_n}=\sum_{j\in\Sigma}\bar{p}^{T_1}_{i_1\dots i_m j}\bar{p}^{T_2}_{j i_{m+1}\dots i_n} \tag{4.6}$$

and we have $\operatorname{Flat}_{1\dots m|m+1\dots n}(\bar{p})=\operatorname{Flat}_{1\dots m|s_1}(\bar{p}^{T_1})\operatorname{Flat}_{s_2|m+1\dots n}(\bar{p}^{T_2})$.

As an immediate consequence of the last statement we recover Theorem 2.2 for ATR models. We proceed to prove the theorem.

Proof. Let $\boldsymbol{i}=(i_1,\dots,i_n)=(\boldsymbol{i}_\alpha,\boldsymbol{i}_\beta)$ be a collection of states at the leaves of T. We start by expressing $\langle p,u^{\boldsymbol{i}}\rangle$ in terms of scalar products of the corresponding tensors on the subtrees T_1 and T_2. We have

$$\langle p,u^{\boldsymbol{i}}\rangle=\left\langle\sum_{\boldsymbol{j}=(\boldsymbol{j}_\alpha,\boldsymbol{j}_\beta)}p_{\boldsymbol{j}}e^{\boldsymbol{j}},u^{\boldsymbol{i}}\right\rangle=\sum_{\boldsymbol{j}_\alpha,\boldsymbol{j}_\beta}p_{\boldsymbol{j}_\alpha,\boldsymbol{j}_\beta}\langle e^{\boldsymbol{j}_\alpha},u^{\boldsymbol{i}_\alpha}\rangle\langle e^{\boldsymbol{j}_\beta},u^{\boldsymbol{i}_\beta}\rangle.$$

By (2.2), $p_{\boldsymbol{j}_\alpha,\boldsymbol{j}_\beta}=\sum_{j_s\in\Sigma}p_{\boldsymbol{j}_\alpha,j_s,\boldsymbol{j}_\beta}$, where $s=s_1\sim s_2$, and by Lemma 4.5 we get

$$p_{\boldsymbol{j}_\alpha,\boldsymbol{j}_\beta}=\sum_{j_s\in\Sigma}\frac{p^{T_1}_{\boldsymbol{j}_\alpha,j_s}}{\pi_{j_s}}p^{T_2}_{j_s,\boldsymbol{j}_\beta}=\sum_{j_r,j_s\in\Sigma}\frac{p^{T_1}_{\boldsymbol{j}_\alpha,j_s}}{\pi_{j_s}}\delta_{j_r,j_s}p^{T_2}_{j_r,\boldsymbol{j}_\beta},$$

where $\delta_{i,j}$ is the Kronecker delta. Hence,

$$\langle p,u^{\boldsymbol{i}}\rangle=\sum_{\boldsymbol{j}_\alpha,j_s}\frac{p^{T_1}_{\boldsymbol{j}_\alpha,j_s}}{\pi_{j_s}}\langle e^{\boldsymbol{j}_\alpha},u^{\boldsymbol{i}_\alpha}\rangle\left(\sum_{j_r}\delta_{j_s,j_r}\sum_{\boldsymbol{j}_\beta}p^{T_2}_{j_r,\boldsymbol{j}_\beta}\langle e^{\boldsymbol{j}_\beta},u^{\boldsymbol{i}_\beta}\rangle\right).$$

Now we observe that

$$\delta_{i,j}=\frac{1}{\pi_i}\pi_i(e^i)^te^j=\langle\pi_ie^i,e^j\rangle=\left\langle\pi_ie^i,\sum_{k\in\Sigma}\frac{\langle e^j,u^k\rangle}{\langle u^k,u^k\rangle}u^k\right\rangle=\sum_{k\in\Sigma}\frac{\pi_i\langle e^i,u^k\rangle\langle u^k,e^j\rangle}{\langle u^k,u^k\rangle}.$$

Therefore, using this expression we obtain

$$\langle p, u^{\boldsymbol{i}}\rangle = \sum_{\boldsymbol{j}_\alpha, j_s} \frac{p^{T_1}_{\boldsymbol{j}_\alpha, j_s}}{\pi_{j_s}} \langle e^{\boldsymbol{j}_\alpha}, u^{\boldsymbol{i}_\alpha}\rangle \left(\sum_{j_r} \sum_{k\in\Sigma} \frac{\pi_{j_s}\langle e^{j_s}, u^k\rangle \langle u^k, e^{j_r}\rangle}{\langle u^k, u^k\rangle} \sum_{\boldsymbol{j}_\beta} p^{T_2}_{j_r, \boldsymbol{j}_\beta} \langle e^{\boldsymbol{j}_\beta}, u^{\boldsymbol{i}_\beta}\rangle \right)$$
$$= \sum_{\boldsymbol{j}_\alpha, j_s} \frac{p^{T_1}_{\boldsymbol{j}_\alpha, j_s}}{\pi_{j_s}} \langle e^{\boldsymbol{j}_\alpha}, u^{\boldsymbol{i}_\alpha}\rangle \left(\sum_{k\in\Sigma} \frac{\pi_{j_s}\langle e^{j_s}, u^k\rangle}{\langle u^k, u^k\rangle} \sum_{j_r, \boldsymbol{j}_\beta} p^{T_2}_{j_r, \boldsymbol{j}_\beta} \langle u^k, e^{j_r}\rangle \langle e^{\boldsymbol{j}_\beta}, u^{\boldsymbol{i}_\beta}\rangle \right).$$

The last sum in the previous expression is $\langle p^{T_2}, u^k \otimes u^{\boldsymbol{i}_\beta}\rangle$ so we get

$$\langle p, u^{\boldsymbol{i}}\rangle = \sum_{k\in\Sigma} \frac{1}{\langle u^k, u^k\rangle} \sum_{\boldsymbol{j}_\alpha, j_s} p^{T_1}_{\boldsymbol{j}_\alpha, j_s} \langle e^{\boldsymbol{j}_\alpha}, u^{\boldsymbol{i}_\alpha}\rangle \langle e^{j_s}, u^k\rangle \langle p^{T_2}, u^k \otimes u^{\boldsymbol{i}_\beta}\rangle .$$

The proof finishes by observing that $\langle p^{T_1}, u^{\boldsymbol{i}_\alpha} \otimes u^k\rangle = \sum_{\boldsymbol{j}_\alpha, j_s} p^{T_1}_{\boldsymbol{j}_\alpha, j_s} \langle e^{\boldsymbol{j}_\alpha}, u^{\boldsymbol{i}_\alpha}\rangle \langle e^{j_s}, u^k\rangle$ and dividing $\langle p, u^{\boldsymbol{i}}\rangle$ by $\langle u^{\boldsymbol{i}}, u^{\boldsymbol{i}}\rangle$:

$$\bar{p}_{\boldsymbol{i}} = \tfrac{\langle p, u^{\boldsymbol{i}}\rangle}{\langle u^{\boldsymbol{i}}, u^{\boldsymbol{i}}\rangle} = \textstyle\sum_k \langle u^k, u^k\rangle \tfrac{\langle p^{T_1}, u^{\boldsymbol{i}_\alpha} \otimes u^k\rangle}{\langle u^{\boldsymbol{i}_\alpha}, u^{\boldsymbol{i}_\alpha}\rangle \langle u^k, u^k\rangle} \tfrac{\langle p^{T_2}, u^k \otimes u^{\boldsymbol{i}_\beta}\rangle}{\langle u^k, u^k\rangle \langle u^{\boldsymbol{i}_\beta}, u^{\boldsymbol{i}_\beta}\rangle} = \textstyle\sum_{k\in\Sigma} \langle u^k, u^k\rangle \bar{p}^{T_1}_{\boldsymbol{i}_\alpha k} \bar{p}^{T_2}_{k \boldsymbol{i}_\beta}.$$

The last claim in the statement of the theorem is straightforward. ∎

The above theorem is also valid when one of the two subtrees, say T_2, is a tree with two leaves and a single edge. In this case this operation is equivalent to multiplying $\bar{p}^{T_1}$ by a diagonal matrix and so it is equivalent to the Markov action on one leaf.

Remark 4.7. In general, the gluing procedure is not equivalent to a toric fiber product as described in [33]. Indeed, as we will see in section 5.3 for the TN93 model, the parametrization for quartets obtained by gluing two tripods (which have a monomial parametrization) is not monomial, and hence the corresponding ideal is not the toric fiber product of both toric ideals. However, for the toric models Kimura 3-parameter and its submodels studied in [32], the gluing procedure is equivalent to a toric fiber product if we use Fourier coordinates.

Thus, with Theorem 4.6 we recover a well-known result of Evans and Speed [14].

COROLLARY 4.8 (see [14, 32]). *On a tree evolving under the Kimura 3-parameter model or one of its submodels (Kimura 2-parameter and Jukes–Cantor), the Fourier coordinates of a tensor $p \in \operatorname{Im}\psi_T$ have a monomial expression in terms of the eigenvalues of the transition matrices. Moreover, the ideal of the corresponding phylogenetic variety V_T is generated by binomials.*

Proof. For the Kimura 3-parameter model and its submodels, π is the uniform distribution. The Fourier basis in (3.5) is a π-orthonormal basis that diagonalizes all transition matrices in these models. If we biject the set $\Sigma = \{1, 2, 3, 4\}$ with the additive group $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ by identifying $1 = (0, 0), 2 = (0, 1), 3 = (1, 0), 4 = (1, 1)$, and denote by $\boxplus$ the sum in G, then it is easy to see that for a tripod tree we have (see also [32])

$$\bar{p}_{i_1 i_2 i_3} = \begin{cases} \lambda^1_{i_1} \lambda^2_{i_2} \lambda^3_{i_3} & \text{if } i_1 \boxplus i_2 \boxplus i_3 = 0; \\ 0 & \text{otherwise.} \end{cases}$$

Then the corollary follows easily by Theorem 4.6 and induction. Indeed, if T is an n-leaf tree, we consider a cherry on it and view T as $T = T_1 * T_2$, where T_2 is a tripod tree. The induction hypothesis is that for any $m < n$, $\bar{p}_{i_1 \dots i_m} = 0$ if $i_1 \boxplus \cdots \boxplus i_m \neq 0$

and $\bar{p}_{i_1\dots i_m}$ has a monomial expression in the eigenvalues if $i_1 \boxplus \cdots \boxplus i_m = 0$. By Theorem 4.6 we have

$$\bar{p}_{i_1\dots i_n} = \sum_{j\in\Sigma} \bar{p}^{T_1}_{i_1\dots i_{n-2}j}\bar{p}^{T_2}_{ji_{n-1}i_n}$$

so that, by the induction hypothesis, the only (possibly) nonzero summand is for $j = i_{n-1} \boxplus i_n$, which implies $i_1 \boxplus \cdots \boxplus i_n = 0$ and gives a monomial expression. ∎

Theorem 4.6 gives an inductive procedure to build the parametrization map of a phylogenetic tree evolving under a B-time-reversible model. Although this gluing procedure mimics the one described in [3] and [12], it is not obvious that this leads to the expression of the ideal of these phylogenetic varieties in terms of subtrees as was done in these two papers. We pose the following question.

QUESTION 4.9. *Can the ideal of the phylogenetic variety evolving on a B-time-reversible model be described in terms of the ideals of flattenings at its interior nodes and edges as in* [3], [12]*?*

We do not deal with this general question in this paper, but for the TN93 model on quartets we construct from tripods and edge flattenings a local complete intersection that describes the variety on an open set (see section 5).

5. Invariants for trees evolving under the TN93 model. In this section we showcase how the framework developed previously can be used to find phylogenetic invariants for the Tamura–Nei model TN93 introduced in Example 3.5. The stationary distribution π is fixed (and we will assume it to be generic when convenient) and the π-orthogonal basis B is specified in Example 3.5. We start by computing phylogenetic invariants for trees with three leaves, then we use the marginalization and tree gluing to find phylogenetic invariants for quartets and n-leaf trees.

5.1. Invariants for tripods. Let T be a star tree with three leaves l_1, l_2, l_3 and three edges e_1, e_2, e_3 (we call it the *tripod*). The joint probability tensor for the point of no-evolution $p^0 = \psi_T(Id, Id, Id)$ has the following coordinates in the standard basis:

$$p^0_{i_1i_2i_3} = \sum_{i_r\in\Sigma} \pi_{i_r} Id_{i_r,i_1} Id_{i_r,i_2} Id_{i_r,i_3} = \begin{cases} \pi_{i_r} & \text{if } i_1 = i_2 = i_3 = i_r; \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

LEMMA 5.1. *Consider the evolutionary model TN93 on the tripod T. The coordinates $\bar{p}^0$ in the basis B_n for the no-evolution point p^0 are*

$$\bar{p}^0_{111} = 1, \qquad \bar{p}^0_{222} = \frac{\pi_{34}-\pi_{12}}{\pi^2_{12}\pi^2_{34}}, \qquad \bar{p}^0_{333} = \frac{\pi^2_4-\pi^2_3}{\pi^2_3\pi^2_4}, \qquad \bar{p}^0_{444} = \frac{\pi^2_2-\pi^2_1}{\pi^2_1\pi^2_2},$$
$$\bar{p}^0_{122} = \bar{p}^0_{212} = \bar{p}^0_{221} = \frac{1}{\pi_{12}\pi_{34}}, \qquad \bar{p}^0_{133} = \bar{p}^0_{313} = \bar{p}^0_{331} = \frac{\pi_{34}}{\pi_3\pi_4},$$
$$\bar{p}^0_{144} = \bar{p}^0_{414} = \bar{p}^0_{441} = \frac{\pi_{12}}{\pi_1\pi_2}, \qquad \bar{p}^0_{233} = \bar{p}^0_{323} = \bar{p}^0_{332} = \frac{-1}{\pi_3\pi_4},$$
$$\bar{p}^0_{244} = \bar{p}^0_{424} = \bar{p}^0_{442} = \frac{1}{\pi_1\pi_2}, \qquad \textit{and} \qquad \bar{p}^0_{i_1i_2i_3} = 0 \textit{ otherwise.}$$

Remark 5.2. For a distribution π such that $\pi_{12} \neq \pi_{34}$, $\pi_1 \neq \pi_2$, and $\pi_3 \neq \pi_4$, there are exactly 45 entries of $\bar{p}^0_{i_1i_2i_3}$ that vanish: when $\{i_1, i_2, i_3\}$ contains either a unique 3 or a unique 4 or when $i_j = 2$ for some j and $i_k = 1$ for $k \neq j$. These are the genericity conditions we will consider for the distribution π from now on.

Proof of Lemma 5.1. From (4.1) and (5.1) we have the expression

$$\bar{p}^0_{i_1i_2i_3} = \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} \sum_{j\in\Sigma} \pi_j \langle e^j, u^{i_1}\rangle \langle e^j, u^{i_2}\rangle \langle e^j, u^{i_3}\rangle,$$

which is invariant after reordering i_1, i_2, i_3. First we prove that $\bar{p}^0_{i_1i_2i_3} = 0$ for the cases mentioned in the previous remark.

Case I: $\{i_1, i_2, i_3\}$ contains only one 3. Without loss of generality assume that $i_1 = 3$ and $i_2, i_3 \neq 3$. From (3.4) we have

$$\begin{aligned}
\bar{p}^0_{3i_2i_3} &= \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} \sum_{j\in\Sigma} \pi_j \langle e^j, u^3\rangle \langle e^j, u^{i_2}\rangle \langle e^j, u^{i_3}\rangle \\
&= \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} \left(\pi_3 \langle e^3, u^3\rangle \langle e^3, u^{i_2}\rangle \langle e^3, u^{i_3}\rangle + \pi_4 \langle e^4, u^3\rangle \langle e^4, u^{i_2}\rangle \langle e^4, u^{i_3}\rangle\right) \\
&= \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} \left(\frac{\pi_3\pi_4}{\pi_{34}} \langle e^3, u^{i_2}\rangle \langle e^3, u^{i_3}\rangle - \frac{\pi_3\pi_4}{\pi_{34}} \langle e^4, u^{i_2}\rangle \langle e^4, u^{i_3}\rangle\right) = 0,
\end{aligned}$$

where the last equality holds because the last two rows of (3.4) for $u^i \neq u^3$ are equal.

Case II: $\{i_1, i_2, i_3\}$ contains only one 4. The proof is analogous to the previous case by noting that the first two rows of (3.4) are equal for $u^i \neq u^4$.

Case III: $i_j = 2$ for some j and $i_k = 1$ for $k \neq j$: Assume, without loss of generality, that $i_1 = i_2 = 1$ and $i_3 = 2$. We have

$$\bar{p}^0_{112} = \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} \sum_{j\in\Sigma} \pi_j \langle e^j, u^1\rangle \langle e^j, u^1\rangle \langle e^j, u^2\rangle = \frac{1}{\langle u^{\mathbf{i}}, u^{\mathbf{i}}\rangle} (\pi_{12}\pi_{34} - \pi_{12}\pi_{34}) = 0.$$

The remaining $64 - 45 = 19$ coordinates $\bar{p}^0_{i_1i_2i_3}$ are nonzero for generic π and can be easily computed in a similar fashion. ∎

By (4.4), for any point $p \in \mathrm{Im}\varphi_T$ the coordinates $\bar{p}_{i_1i_2i_3}$ with i_1, i_2, and i_3 satisfying any of the cases of Remark 5.2 vanish. Thus, CV_T is contained in a linear space $\mathcal{L}_3 \subset \otimes^3 W$ of dimension 19 defined by the 45 linear equations, and V_T is contained in the linear space $\mathcal{L}_3 \cap H$ of dimension 18, where H is the space defined in (3.6).

We now provide a set of generators for I_T and a set of polynomials defining a complete intersection that cuts out $CV_T \subset \mathcal{L}_3$ in an open set.

PROPOSITION 5.3. *Let T be a tripod that evolves under the TN*93 *model. For generic π, I_T is a binomial ideal minimally generated by* 45 *linear monomials,* 9 *quadratic binomials,* 29 *cubic binomials, and* 3 *quintic binomials. If we consider $CV_T \subset \mathcal{L}_3$, then the variety X_T defined by the* 9 *polynomials*

$$\begin{gathered}
\bar{p}_{222}\bar{p}_{441} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{221}\bar{p}_{442}, \qquad \bar{p}_{222}\bar{p}_{414} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{212}\bar{p}_{424}, \\
\bar{p}_{222}\bar{p}_{144} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{122}\bar{p}_{244}, \qquad \bar{p}_{332}\bar{p}_{441} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{331}\bar{p}_{442}, \\
\bar{p}_{323}\bar{p}_{414} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{313}\bar{p}_{424}, \qquad \bar{p}_{233}\bar{p}_{144} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{133}\bar{p}_{244}, \\
\bar{p}_{144}\bar{p}_{414}\bar{p}_{441} - \frac{\pi_1\pi_2\pi_{12}}{(\pi_1-\pi_2)^2}\bar{p}_{111}\bar{p}^2_{444}, \qquad \bar{p}_{133}\bar{p}_{313}\bar{p}_{331} - \frac{\pi_3\pi_4\pi_{34}}{(\pi_3-\pi_4)^2}\bar{p}_{111}\bar{p}^2_{333}, \\
\bar{p}_{332}\bar{p}_{323}\bar{p}_{233} - \frac{\pi_3\pi_4\pi^2_{12}}{(\pi_3-\pi_4)^2(\pi_{12}-\pi_{34})}\bar{p}_{222}\bar{p}^2_{333}
\end{gathered}$$

is a complete intersection that cuts out CV_T in the open set $\bar{p}_{iii} \neq 0$ for $i = 1, 2, 3, 4$. Furthermore, CV_T is an irreducible component of X_T.

Proof. From Lemma 4.4 it follows that φ_T is a monomial map and I_T is generated by binomials. Lemma 5.1 and (4.4) guarantee that the 45 linear polynomials $\bar{p}_{i_1 i_2 i_3}$ are in I_T for the indices $\{i_1, i_2, i_3\}$ satisfying the conditions of Remark 5.2. `Macaulay2` computations provide the remaining generators and guarantee that they are a minimal set of generators. The comprehensive list is included in section SM2.

To study CV_T, we use tools from toric geometry. We refer the reader to [31, Chapter 4] for an introduction to toric ideals. Consider a rescaling $\tilde{p}$ of $\bar{p}$ such that it is a monic monomial parametrization. The vanishing ideal of V_T is the toric ideal generated by the columns of the matrix C in (5.2),

$$C = \left(\begin{array}{cccc|cccc|cccc|cccc|ccc} 1&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&1&1&0&0&0&0&0&0&1&0&0&1&0&0\\ 0&0&1&0&0&0&0&0&1&1&0&0&0&0&1&1&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&0&1&1&0&0&0&0&1&1\\ \hline 1&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0\\ 0&1&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&1&0\\ 0&0&1&0&0&0&0&1&0&1&0&0&0&1&0&1&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&1&0&1&0&0&0&1&0&1\\ \hline 1&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0\\ 0&1&0&0&1&1&0&0&0&0&0&0&0&0&0&1&0&0&1\\ 0&0&1&0&0&0&0&1&1&0&0&0&0&1&1&0&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&1&1&0&0&0&0&1&1&0 \end{array}\right). \tag{5.2}$$

The columns of the matrix D in (5.3) are generators of $\ker(C)$, and hence the 9 polynomials defined in the statement are the generators of the lattice basis ideal J of C when considering the aforementioned parametrization $\tilde{p}$.

$$D = \left(\begin{array}{ccc|ccc|ccc} 0&0&0&0&0&0&-1&-1&0\\ 1&1&1&0&0&0&0&0&-1\\ 0&0&0&0&0&0&0&-2&-2\\ 0&0&0&0&0&0&-2&0&0\\ \hline 0&0&-1&0&0&0&0&0&0\\ 0&-1&0&0&0&0&0&0&0\\ -1&0&0&0&0&0&0&0&0\\ \hline 0&0&0&0&0&-1&0&1&0\\ 0&0&0&0&-1&0&0&1&0\\ 0&0&0&-1&0&0&0&1&0\\ \hline 0&0&1&0&0&1&1&0&0\\ 0&1&0&0&1&0&1&0&0\\ 1&0&0&1&0&0&1&0&0\\ \hline 0&0&0&0&0&1&0&0&1\\ 0&0&0&0&1&0&0&0&1\\ 0&0&0&1&0&0&0&0&1\\ \hline 0&0&-1&0&0&-1&0&0&0\\ 0&-1&0&0&-1&0&0&0&0\\ -1&0&0&-1&0&0&0&0&0 \end{array}\right). \tag{5.3}$$

Theorem 2.1 in [8] (see also Theorem 2.7 in [27]) guarantees that this lattice basis ideal is a complete intersection if every mixed submatrix of D has at least as many columns as rows. Recall that a matrix is mixed if in every column there is at least one positive and one negative entry. From the structure of D it is intuitive to see

that every mixed submatrix is either a square matrix or has more rows than columns; however, we provide a code that verifies that every submatrix of D with fewer rows than columns is not mixed. The code is included in the CORA repository. By [13, Corollary 2.5] $I_T = J : (\prod_{ijk} \tilde{p}_{ijk})^\infty$, where $J : (\prod_{ijk} \tilde{p}_{ijk})^\infty$ denotes the saturation of J by the ideal generated by the product of all variables, and by [20, Corollary 2.1], I_T is a minimal prime of J, hence CV_T is an irreducible component of X_T. ∎

Remark 5.4. For a nongeneric distribution π, there might be more invariants arising from the vanishing of $\bar{p}^0_{iii}$, with $i \in \{2,3,4\}$. For instance, when $\pi_{12} = \pi_{34}$, $\pi_1 = \pi_2$, and $\pi_3 = \pi_4$, then I_T is minimally generated by 9 quadrics, 24 cubics, 3 quintics, and the additional linear invariants $\bar{p}_{222} = \bar{p}_{333} = \bar{p}_{444} = 0$.

5.2. Invariants for n-leaf trees arising from tripods. In this section we focus on phylogenetic invariants of trees with n leaves evolving under TN93 that can be obtained from invariants of the tripod either by gluing or by marginalization. Let T be a tree with n leaves, $n > 3$, and consider a tensor $p = \varphi_T(\{\Lambda^e\}_{e \in E(T)})$. We can always obtain p by gluing a tensor p^{T_1} on a tripod T_1 to a tensor p^{T_2} on an $(n-1)$-leaf tree T_2. Indeed, let l_1 and l_2 be two leaves in T forming a cherry and consider the interior edge e_0 adjacent to the cherry; then we split this edge into two edges e_1 and e_2 as in Figure 1 and call T_1 the tripod tree formed by the cherry and e_1 and T_2 the tree $T \setminus T_1$. Theorem 4.6 ensures that $p = p^{T_1} * p^{T_2}$, where $p^{T_1} = \varphi_{T_1}(\{\Lambda^e\}_{e \in E(T_1)})$ and $p^{T_2} = \varphi_{T_2}(\{\Lambda^e\}_{e \in E(T_2)})$, with $\Lambda^{e_1} = \Lambda^{e_0}$, $\Lambda^{e_2} = Id$, where e_0, e_1, e_2 are the edges involved in the gluing as denoted in Figure 1, and the remaining matrices Λ^e for $e \in E(T_i)$ coincide with Λ^e for the corresponding $e \in E(T)$.

PROPOSITION 5.5. *The linear equations of the form $\bar{p}_{i_1 \dots i_n} = 0$ hold for any phylogenetic n-leaf tree T evolving under a TN93 model when*

(i) *exactly one i_k is equal to 3 or 4, or*

(ii) *exactly one i_k is equal to 2 and the rest are equal to 1.*

Proof. Case $n = 3$ is proven by Lemma 5.1. If T is a tree with n leaves and $n > 3$, we use as induction hypothesis that the equations hold for trees with fewer than n leaves. Since $n > 3$, T has a cherry that does not contain the distinguished element. Let T_1 be the tripod formed from this cherry and consider the decomposition $T = T_1 * T_2$, where T_2 is an $(n-1)$-leaf tree. Since there will be a single distinguished element in T_2, we can assume without loss of generality that it is leaf l_n and that the cherry considered in T_1 has leaves l_1 and l_2 (reordering indices if necessary).

In case (i), consider $l \in \{3,4\}$ and $i_k \neq l$ for $1 \leq k \leq n-1$. By the induction hypothesis, $\bar{p}^{T_2}_{ji_3 \dots i_{n-1} l} = 0$ for $j \neq l$ and $\bar{p}^{T_1}_{i_1 i_2 l} = 0$. Therefore,

$$\bar{p}_{i_1 \dots i_{n-1} l} = \sum_{j \in \Sigma} \langle u^j, u^j \rangle \bar{p}^{T_1}_{i_1 i_2 j} \bar{p}^{T_2}_{j i_3 \dots i_{n-1} l} = \langle u^l, u^l \rangle \bar{p}^{T_1}_{i_1 i_2 l} \bar{p}^{T_2}_{l i_3 \dots i_{n-1} l} = 0.$$

In case (ii) note that $\bar{p}^{T_i}_{1 \dots 1 j} = 0$ for any $j \neq 1$ by the induction hypothesis and case (i). Hence, we get $\bar{p}_{1 \dots 12} = \sum_{j \in \Sigma} \langle u^j, u^j \rangle \bar{p}^{T_1}_{11j} \bar{p}^{T_2}_{j1 \dots 12} = 0$. ∎

We pose the following question.

QUESTION 5.6. *Do the equations in Proposition 5.5 determine a system of generators for the set of linear model invariants for n-leaf trees?*

In the case of quartets, this is proven in Proposition 5.11. This problem is related to the space of phylogenetic mixtures and model selection (see [5]), and we expect to address it for n-leaf trees in a forthcoming paper.

By decomposing a tree into a tripod and the remaining subtree as in the proof of Proposition 5.5, we can obtain other linear invariants for $n \geq 4$ based on certain leaf configurations of the tree.

LEMMA 5.7. *Let T be a tree evolving under a TN93 model. If nodes l_j and l_k form a cherry, coordinates $\bar{p}_{i_1 \dots i_n}$, where $\{i_j, i_k\} = \{3,4\}$, vanish for any $p \in \operatorname{Im} \varphi_T$.*

Proof. Without loss of generality we can assume that l_1 and l_2 form a cherry as above (such that $i_1 = 3$ and $i_2 = 4$), and we can view T as gluing a tripod tree T_1 and an $(n-1)$-leaf tree T_2. Then

$$\bar{p}_{34 i_3 \dots i_n} = \sum_{j \in \Sigma} \langle u^j, u^j \rangle \, \bar{p}^{T_1}_{34j} \bar{p}^{T_1}_{j i_3 \dots i_n},$$

which is zero by Remark 5.2. ∎

When both 3 and 4 appear exactly once, $\bar{p}_{i_1 \dots i_{n-2} 34}$ is a model invariant, as proved in Proposition 5.5. As we will see in Proposition 5.11, if both 3 and 4 appear twice, $\bar{p}_{i_1 \dots i_4}$ yield topology invariants for quartets. A natural question that remains to be addressed for larger trees is the following.

QUESTION 5.8. *Are equations $\bar{p}_{i_1 \dots i_n} = 0$ topology invariants for n-leaf trees if both 3 and 4 appear at least twice?*

Next we go beyond linear invariants. Any phylogenetic invariant of the tripod can easily be extended to a model invariant for trees with n leaves.

LEMMA 5.9. *Let $f(\{\bar{p}_{ijk}\})$ be a phylogenetic invariant of the tripod. Then, for trees with $n \geq 3$ leaves evolving under TN93, $f(\{\bar{p}_{ijk1\dots1}\})$ is a model invariant.*

Proof. Let $f \in \mathbb{C}[\bar{p}_{ijk} \mid i,j,k \in \Sigma]$ be a phylogenetic invariant for the tripod and let $\tilde{f} \in \mathbb{C}[\bar{p}_{ijk1\dots1} \mid i,j,k \in \Sigma]$ be the extension of f via $\bar{p}_{ijk} \mapsto \bar{p}_{ijk1\dots1}$. The fact that $\tilde{f}$ is a model invariant for n-leaf trees follows directly from Lemma 4.2. Indeed, let p be a tensor in $\operatorname{Im}\psi_T$ for any tree T. By marginalizing $p \in \otimes^n W$ over the last $n-3$ components, we have a tensor $p^+ \in \operatorname{Im} \psi_{T_3}$ on the tripod T_3. In coordinates in the basis B_n we obtain $\tilde{f}(\{\bar{p}_{ijk1\dots1}\}) = f(\{\overline{p^+}_{ijk}\})$ and it vanishes because f is a model invariant for tripods. ∎

Example 5.10. The tripod invariants in Proposition 5.3 can be extended to model invariants for quartets as follows:

$$\bar{p}_{2221}\bar{p}_{4411} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{2211}\bar{p}_{4421}, \quad \bar{p}_{2221}\bar{p}_{4141} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{2121}\bar{p}_{4241},$$
$$\bar{p}_{2221}\bar{p}_{1441} - \frac{\pi_{34}-\pi_{12}}{\pi_{34}}\bar{p}_{1221}\bar{p}_{2441}, \quad \bar{p}_{3321}\bar{p}_{4411} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{3311}\bar{p}_{4421},$$
$$\bar{p}_{3231}\bar{p}_{4141} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{3131}\bar{p}_{4241}, \quad \bar{p}_{2331}\bar{p}_{1441} + \frac{\pi_{12}}{\pi_{34}}\bar{p}_{1331}\bar{p}_{2441},$$
$$\bar{p}_{1441}\bar{p}_{4141}\bar{p}_{441} - \frac{\pi_1\pi_2\pi_{12}}{(\pi_1-\pi_2)^2}\bar{p}_{1111}\bar{p}^2_{4441}, \quad \bar{p}_{1331}\bar{p}_{3131}\bar{p}_{331} - \frac{\pi_3\pi_4\pi_{34}}{(\pi_3-\pi_4)^2}\bar{p}_{1111}\bar{p}^2_{3331},$$
$$\bar{p}_{3321}\bar{p}_{3231}\bar{p}_{2331} - \frac{\pi_3\pi_4\pi^2_{12}}{(\pi_3-\pi_4)^2(\pi_{12}-\pi_{34})}\bar{p}_{2221}\bar{p}^2_{3331}.$$

5.3. Invariants for quartets. We now turn our attention to quartets. We call $l_i l_j | l_k l_m$ the trivalent tree with four leaves l_i, l_j, l_k, l_m whose interior edge separates leaves l_i, l_j from l_k, l_m (that is, l_i, l_j and l_k, l_m are cherries in this tree). We will focus on the tree $l_1 l_2 | l_3 l_4$, but all the results in this section are analogous for the two

other tree topologies, $l_1l_3|l_2l_4$ and $l_1l_4|l_2l_3$. By $\varphi_T(\Lambda^1,\dots,\Lambda^4,\Lambda)$ we mean that Λ^i is assigned to the edge incident to leaf l_i for $i=1,\dots,4$, and Λ is assigned to the interior edge.

PROPOSITION 5.11. *The following* 172 *linear equations hold for any quartet* T *evolving under the evolutionary model* $\mathcal{M}=TN93$*:*

(i) $\bar{p}_{i_1i_2i_3i_4}=0$ *if exactly one* i_k *is equal to* 3 *or* 4 *for* $k\in\{1,2,3,4\}$,

(ii) $\bar{p}_{1112}=\bar{p}_{1121}=\bar{p}_{1211}=\bar{p}_{2111}=0$,

and these equations generate all linear model invariants. Moreover, if $T=l_1l_2|l_3l_4$*, then* $\bar{p}_{3434},\bar{p}_{3443},\bar{p}_{4334}$*, and* $\bar{p}_{4343}$ *are linear topology invariants of* T*.*

Proof. As a particular case of Proposition 5.5 we obtain the list of model invariants in (i) and (ii). On the other hand, from Lemma 5.7 we get that $\bar{p}_{3434},\bar{p}_{3443},\bar{p}_{4334}$, and $\bar{p}_{4343}$ vanish when p evolves on $T=l_1l_2|l_3l_4$. It can be easily checked that $\bar{p}_{3434},\bar{p}_{3443},\bar{p}_{4334}$, and $\bar{p}_{4343}$ are generically nonzero for either tree $l_1l_3|l_2l_4$ or $l_1l_4|l_2l_3$, and hence they are topology invariants for T. For example, if $T=l_1l_3|l_2l_4$, let $q=\varphi_T(Id,Id,Id,Id,\Lambda)$ for a diagonal matrix Λ; then

$$\bar{q}_{3434}=\bar{q}_{4343}=\frac{\pi_{12}\pi_{34}}{\pi_1\pi_2\pi_3\pi_4}(\lambda_1-\lambda_2) \tag{5.4}$$

(see Lemma SM2.1 in subsection SM2.2), which is nonzero if $\lambda_1\neq\lambda_2$.

Let $\mathcal{L}_4$ be the linear space defined by the 172 equations in (i) and (ii). As these are linearly independent equations (they involve different coordinates), $\mathcal{L}_4$ has dimension $256-172=84$. In what follows we prove that there is a subset of points in $CV_{l_1l_2|l_3l_4}\cup CV_{l_1l_3|l_2l_4}\cup CV_{l_1l_4|l_2l_3}$ that spans $\mathcal{L}_4$.

For each $j\in\Sigma=\{1,2,3,4\}$, set $D_j=\mathrm{diag}(e^j)$. Consider the 84 4-tuples not appearing as a subindex in the equations of $\mathcal{L}_4$ and let $\mathbf{i}=(i_1,\dots,i_4)$ be any of these 4-tuples.

If $\mathbf{i}$ is different from $(3,4,3,4),(4,3,4,3),(3,4,4,3),(4,3,3,4)$, take $T=l_1l_2|l_3l_4$ and a diagonal matrix Λ, and define the point

$$p^{\mathbf{i}}=\varphi_T(D_{i_1},D_{i_2},D_{i_3},D_{i_4},\Lambda).$$

By (4.4), the unique possibly nonzero coordinate of $p^{\mathbf{i}}$ in the basis B_4 is $i_1\dots i_4$. Using the expressions given in subsection SM2.2, we see that this coordinate is nonzero if Λ has nonzero generic elements in the diagonal. Therefore, we have exactly 80 linearly independent points in this set.

If $\mathbf{i}=(3,4,3,4)$ or $(4,3,4,3)$, consider $T=l_1l_3|l_2l_4$ and the point $p^{\mathbf{i}}=\varphi_T(D_{i_1},D_{i_2},D_{i_3},D_{i_4},\Lambda)$, where $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$ has $\lambda_2\neq\lambda_1$. Note that these are two linearly independent points whose B_4 coordinates are all zero except those indexed by 3434, 4343, respectively, which coincide with expression (5.4).

If $\mathbf{i}=(3,4,4,3)$ or $(4,3,3,4)$, an analogous argument applies by considering points $p^{\mathbf{i}}$ in the variety of $T=l_1l_4|l_2l_3$.

The 80 linearly independent points above together with the four points $p^{(3,4,3,4)}$, $p^{(4,3,4,3)}$, $p^{(3,4,4,3)}$, and $p^{(4,3,3,4)}$ form a set of 84 linearly independent points in $\mathcal{L}_4$ because all have a single nonzero coordinate and all of them are in different positions. ∎

Remark 5.12. From the previous result we get that the 84-dimensional space $\mathcal{L}_4$ defined as the set of tensors where the 172 equations in (i) and (ii) vanish coincides with the linear span of $CV_{l_1l_2|l_3l_4}\cup CV_{l_1l_3|l_2l_4}\cup CV_{l_1l_4|l_2l_3}$. If we add equations $\bar{p}_{3434}=0,\bar{p}_{3443}=0,\bar{p}_{4334}=0$, and $\bar{p}_{4343}=0$. then the zero set $\mathcal{L}_{l_1l_2|l_3l_4}$ of dimension 80 is the

linear span of $CV_{l_1l_2|l_3l_4}$. The table below displays which equations hold for each of the three possible quartet topologies, thus providing their topology invariants.

	$\bar{p}_{3344}=0$	$\bar{p}_{4433}=0$	$\bar{p}_{3434}=0$	$\bar{p}_{4343}=0$	$\bar{p}_{3443}=0$	$\bar{p}_{4334}=0$
$l_1l_2\|l_3l_4$	No	No	Yes	Yes	Yes	Yes
$l_1l_3\|l_2l_4$	Yes	Yes	No	No	Yes	Yes
$l_1l_4\|l_2l_3$	Yes	Yes	Yes	Yes	No	No

LEMMA 5.13. *Let $T=l_1l_2|l_3l_4$ be a quartet and consider the flattening matrix $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$. Then, for any $i,j\in\Sigma$, column (i,j) is a linear combination of columns $(1,1),(1,2),(1,3)$, and $(1,4)$. Moreover, for generic π, there is an open set $\mathcal{U}\subseteq CV_T$ containing the no-evolution point p^0 such that, for any $p\in\mathcal{U}$,* $\operatorname{rank}\left(\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})\right)=4$ *and*

(i) *the submatrix formed by columns $(1,j),(j,1),(2,j),(j,2)$ has rank 1 for $j\in\{3,4\}$,*
(ii) *the submatrix formed by columns $(1,2),(2,1)$ has rank 1,*
(iii) *the submatrix formed by columns $(1,1),(1,2),(2,2)$ has rank 2,*
(iv) *the submatrix formed by columns $(1,1),(1,2),(1,j),(j,j)$ has rank 3 for $j\in\{3,4\}$.*

The submatrices described in the statement are displayed in Tables SM2 to SM5 in section SM3. Note that the generic rank 4 claimed here was already known by a more detailed proof of Theorem 2.2 given in [28].

Proof. Given $\bar{p}=\varphi_T(\Lambda^1,\dots,\Lambda^4,\Lambda)$, consider $\bar{q}=\varphi_T(Id,\dots,Id,\Lambda)$ so that

$$\bar{p}_{i_1i_2i_3i_4}=\lambda^1_{i_1}\lambda^2_{i_2}\lambda^3_{i_3}\lambda^4_{i_4}\bar{q}_{i_1i_2i_3i_4}$$

for any $i_j\in\Sigma$. In standard coordinates $\bar{q}$ is $q=\psi_T(Id,\dots,Id,A^{-t}\Lambda A^t)$, which can be written as the gluing $q^{T_1}*q^{T_2}$, where q^{T_1} evolves on the tripod T_1 with the identity matrix at leaves 1, 2 and matrix $A^{-t}\Lambda A^t$ at the third leaf, and q^{T_2} is the no-evolution point on the tripod T_2. Note that q^{T_1} coincides with the marginalization q^+ of q over leaf l_3. By Lemma 4.2, we have $\bar{q}^{T_1}_{i_1i_2k}=\overline{q^+}_{i_1i_2k}=\bar{q}_{i_1i_21k}$. Therefore using Theorem 4.6 we have

$$\bar{q}_{i_1i_2ij}=\sum_{k\in\Sigma}\langle u^k,u^k\rangle\,\bar{q}^{T_1}_{i_1i_2k}\bar{q}^{T_2}_{kij}=\sum_{k\in\Sigma}\langle u^k,u^k\rangle\,\bar{q}^{T_2}_{kij}\bar{q}_{i_1i_21k}$$

for any $i_1,i_2,i,j\in\Sigma$. In particular, column (i,j) of $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{q})$ is a linear combination of columns $(1,k)$ for $k\in\Sigma$. In Table SM1 we display $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{q})$ to visualize the submatrices in the statement.

Now we have

$$\begin{aligned}\bar{p}_{i_1i_2ij}&=\lambda^1_{i_1}\lambda^2_{i_2}\lambda^3_i\lambda^4_j\bar{q}_{i_1i_2ij}=\sum_{k\in\Sigma}\langle u^k,u^k\rangle\,\lambda^3_i\lambda^4_j\bar{q}^{T_2}_{kij}\lambda^1_{i_1}\lambda^2_{i_2}\bar{q}_{i_1i_21k}\\&=\sum_{k\in\Sigma}\langle u^k,u^k\rangle\,\lambda^3_i\lambda^4_j\bar{q}^{T_2}_{kij}(\lambda^3_1\lambda^4_k)^{-1}\bar{p}_{i_1i_21k},\end{aligned}\tag{5.5}$$

where the last equality holds if $\lambda^3_1\neq 0$ and Λ^4 is invertible. Thus, on an open set of V_T (and hence on the whole variety), column (i,j) of $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ is a linear combination of columns $(1,1)$, $(1,2)$, $(1,3)$, and $(1,4)$. Moreover, the rank of $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ is exactly 4 in an open set containing the no-evolution point, since the 4-minor formed by rows and columns $(1,k)$ does not vanish at p^0.

The nonvanishing coordinates of the no-evolution point q^{T_2} in Lemma 5.1 determine how many nonvanishing summands there are in (5.5), and hence yield the rank of submatrices (i)–(iv). ∎

For the TN93 process on the quartet $T = l_1l_2|l_3l_4$ we have 5 transition matrices with 3 free parameters each, hence $\dim V_T = 15$ and $\dim CV_T = 16$ (see subsection 3.1). By Remark 5.12, CV_T lies in the linear space $\mathcal{L}_{l_1l_2|l_3l_4}$ of dimension 80. Next we provide $\operatorname{codim}(CV_T) = 64$ elements in the ideal of CV_T that define the variety locally at the no-evolution point.

Inspired by the results in [6, Theorem 5.4] that provide local equations for equivariant models, we consider invariants arising from

(a) extending the tripod equations in Proposition 5.3 by adding 1 in the first and third positions, respectively (which are phylogenetic invariants for T by Lemma 5.9), and

(b) rank constraints on $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ as in Lemma 5.13; more precisely, for each of the submatrices of rank r, $r = 1, 2, 3$, in Lemma 5.13, consider a nonvanishing r-minor (namely, one containing only rows and columns of type $(1, j)$ for $j \in \Sigma$), and consider all $(r+1)$-minors containing it (check Tables SM2 to SM5 in section SM3 for the precise description of these minors).

This gives the 64 phylogenetic invariants listed in the statement below.

THEOREM 5.14. *Consider the tree $T = l_1l_2 \mid l_3l_4$ and let it evolve under the TN*93 *model. Then there exist* 64 *equations that cut out the variety CV_T on an open set containing the no-evolution point p^0, arising from*

- *the extension of the* 6 *quadrics and* 3 *cubics in Proposition* 5.3 *with* 1 *in the first leaf,*
- *the extension of the* 6 *quadrics and* 3 *cubics in Proposition* 5.3 *with* 1 *in the third leaf,*
- 2*-minors of columns* $(1,j),(j,1),(2,j),(4,j)$ *of* $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ *for each* $j \in \{3,4\}$ *(*12 *quadrics),*
- 2*-minors of columns* $(1,2),(2,1)$ *of* $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ *(*4 *quadrics),*
- 3*-minors of columns* $(1,1),(1,2),(2,2)$ *of* $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ *(*4 *cubics),*
- 4*-minors of columns* $(1,1),(1,2),(1,j),(j,j)$ *of* $\mathrm{Flat}_{l_1l_2|l_3l_4}(\bar{p})$ *for each* $j \in \{3,4\}$ *(*7 *quartics).*

Proof. `Macaulay2` computations show that the Jacobian of these polynomials at p^0 is indeed 64. Thus, on a Zariski open subset containing p^0, these polynomials define a complete intersection of dimension 16 that coincides with CV_T. ∎

6. Discussion. We have introduced a new approach to working with algebraic time-reversible models that have a given stationary distribution π. We assume that this π can be inferred from data, that is, the given data has reached the equilibrium distribution. We also assumed that this stationary distribution was the same as the one that initiated the process (as it is usual to assume on a time-reversible process). This is an important assumption for our methods to work: if the distribution at the root π^r was supposed to be different from π, the statements of our main results would not hold. It would be interesting to explore a new model that would allow π^r to be parameters as well.

We have illustrated our methods with an insight into the TN93 model. Far from providing an extensive work on this model, we have mainly worked on tripods and quartet trees. We are aware that the tools presented here can allow the extension of this work to trees on any number of leaves, and we aim to develop such work in a

forthcoming project. Another project is exploring the tools we have developed with a view towards model selection by using linear model invariants of different ATR models and studying the space of phylogenetic mixtures (in the sense of [22] and [5]).

Acknowledgment. We would like to thank Jesús Fernández-Sánchez for useful discussions on this topic that led to improvements in the paper.

REFERENCES

[1] E. S. Allman and J. A. Rhodes, *Quartets and parameter recovery for the general Markov model of sequence mutation*, Appl. Math. Res. eXpress, 2004 (2004), pp. 107–131, https://doi.org/10.1155/S1687120004020283.

[2] E. S. Allman and J. A. Rhodes, *Phylogenetic invariants for stationary base composition*, J. Symbolic Comput., 41 (2006), pp. 138–150.

[3] E. S. Allman and J. A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Adv. Appl. Math., 40 (2008), pp. 127–148, https://doi.org/10.1016/j.aam.2006.10.002.

[4] M. Casanellas and J. Fernández-Sánchez, *Relevant phylogenetic invariants of evolutionary models*, J. Math. Pures Appl., 96 (2011), pp. 207–229, https://doi.org/10.1016/j.matpur.2010.11.002.

[5] M. Casanellas, J. Fernández-Sánchez, and A. M. Kedzierska, *The space of phylogenetic mixtures for equivariant models*, Algorithms Mol. Biol., 7 (2012), 33, https://doi.org/10.1186/1748-7188-7-33.

[6] M. Casanellas, J. Fernández-Sánchez, and M. Michałek, *Local equations for equivariant evolutionary models*, Adv. Math., 315 (2017), pp. 285–323, https://doi.org/10.1016/j.aim.2017.05.003.

[7] M. Casanellas and M. Steel, *Phylogenetic mixtures and linear invariants for equal input models*, J. Math. Biol., 74 (2017), pp. 1107–1138.

[8] E. Cattani, R. Curran, and A. Dickenstein, *Complete intersections in toric ideals*, Proc. Amer. Math. Soc., 135 (2007), pp. 329–335, https://doi.org/10.1090/S0002-9939-06-08513-3.

[9] J. A. Cavender and J. Felsenstein, *Invariants of phylogenies in a simple case with discrete states*, J. Classif., 4 (1987), pp. 57–71, https://doi.org/10.1007/BF01890075.

[10] J. T. Chang, *Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency*, Math. Biosci., 137 (1996), pp. 51–73, https://doi.org/10.1016/S0025-5564(96)00075-2.

[11] J. Chifman and L. S. Kubatko, *Quartet inference from SNP data under the coalescent model*, Bioinform., 30 (2014), pp. 3317–3324, https://doi.org/10.1093/bioinformatics/btu530.

[12] J. Draisma and J. Kuttler, *On the ideals of equivariant tree models*, Math. Ann., 344 (2009), pp. 619–644, https://doi.org/10.1007/s00208-008-0320-6.

[13] D. Eisenbud and B. Sturmfels, *Binomial ideals*, Duke Math. J., 84 (1996), pp. 1–45, https://doi.org/10.1215/S0012-7094-96-08401-X.

[14] S. N. Evans and T. P. Speed, *Invariants of some probability models used in phylogenetic inference*, Ann. Statist., 21 (1993), pp. 355–377, https://doi.org/10.1214/aos/1176349030.

[15] J. Felsenstein, *Evolutionary trees from DNA sequences: A maximum likelihood approach*, J. Mol. Evol., 17 (2005), pp. 368–376.

[16] J. Fernández-Sánchez and M. Casanellas, *Invariant versus classical approach when evolution is heterogeneous across sites and lineages*, Syst. Biol., 65 (2016), pp. 280–291.

[17] D. R. Grayson and M. E. Stillman, *Macaulay2, a Software System for Research in Algebraic Geometry*, 2009, https://macaulay2.com/.

[18] M. Hasegawa, H. Kishino, and T.-a. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*, J. Mol. Evol., 22 (1985), pp. 160–174, https://doi.org/10.1007/BF02101694.

[19] M. D. Hendy, D. Penny, and M. A. Steel, *A discrete Fourier analysis for evolutionary trees*, Proc. Natl. Acad. Sci. USA, 91 (1994), pp. 3339–3343, https://doi.org/10.1073/pnas.91.8.3339.

[20] S. Hoçten and J. Shapiro, *Primary decomposition of lattice basis ideals*, J. Symbolic Comput., 29 (2000), pp. 625–639, https://doi.org/10.1006/jsco.1999.0397.

[21] T. Jukes and C. Cantor, *Evolution of protein molecules*, in Mammalian Protein Metabolism, Academic Press, 1969, pp. 21–132.

[22] A. M. Kedzierska, M. Drton, R. Guigó, and M. Casanellas, *SPIn: Model selection for phylogenetic mixtures via linear invariants*, Mol. Biol. Evol., 29 (2012), pp. 929–937, https://doi.org/10.1093/molbev/msr259.

[23] M. Kimura, *A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences*, J. Mol. Evol., 16 (1980), pp. 111–120.

[24] M. Kimura, *Estimation of evolutionary sequences between homologous nucleotide sequences*, Proc. Natl. Acad. Sci. USA, 78 (1981), pp. 454–458.

[25] J. A. Lake, *A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony*, Mol. Biol. Evol., 4 (1987), pp. 167–191, https://doi.org/10.1093/oxfordjournals.molbev.a040433.

[26] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2006, https://doi.org/10.1090/mbk/107.

[27] G. Scheja, O. Scheja, and U. Storch, *On regular sequences of binomials*, Manuscripta Math., 98 (1999), pp. 115–132, https://doi.org/10.1007/s002290050129.

[28] J. Snyman, C. Fox, and D. Bryant, *Parsimony and the rank of a flattening matrix*, J. Math. Biol., 86 (2023), 44.

[29] K. St. John, T. Warnow, B. M. Moret, and L. Vawter, *Performance study of phylogenetic methods: (Unweighted) quartet methods and neighbor-joining*, J. Algorithm., 48 (2003), pp. 173–193, https://doi.org/10.1016/S0196-6774(03)00049-X.

[30] M. Steel, *Phylogeny: Discrete and Random Processes in Evolution*, SIAM, Philadelphia, PA, 2016, https://doi.org/10.1137/1.9781611974485.

[31] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, Mem. Amer. Math. Soc., American Mathematical Society, 1996, https://books.google.de/books?id=J5cVkNIbgXgC.

[32] B. Sturmfels and S. Sullivant, *Toric ideals of phylogenetic invariants*, J. Comput. Biol., 12 (2005), pp. 204–228, https://doi.org/10.1089/cmb.2005.12.204.

[33] S. Sullivant, *Toric fiber products*, J. Algebra, 316 (2007), pp. 560–577, https://doi.org/10.1016/j.jalgebra.2006.10.004.

[34] J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, B. T. Kaine, M. D. Woodhams, and B. R. Holland, *Is the general time-reversible model bad for molecular phylogenetics?*, Syst. Biol., 61 (2012), pp. 1069–1074, https://doi.org/10.1093/sysbio/sys042.

[35] D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version* 4.0b10, Sinauer Associates, Sunderland, MA, 2003.

[36] L. Szekely, M. Steel, and P. Erdos, *Fourier calculus on evolutionary trees*, Adv. Appl. Math., 14 (1993), pp. 200–216, https://doi.org/10.1006/aama.1993.1011.

[37] K. Tamura and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees*, Mol. Biol. Evol., 10 (1993), pp. 512–526, https://doi.org/10.1093/oxfordjournals.molbev.a040023.

[38] S. Tavaré, *Some probabilistic and statistical problems on the analysis of DNA sequences*, Lect. Math. Life Sci., 17 (1986), pp. 57–86.

[39] M. D. Woodhams, J. Fernández-Sánchez, and J. G. Sumner, *A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates*, Syst. Biol., 64 (2015), pp. 638–650, https://doi.org/10.1093/sysbio/syv021.

[40] P. Zwiernik, *Semialgebraic Statistics and Latent Tree Models*, Chapman & Hall, 2019.