Discrete Applied Mathematics 380 (2026) 538–561

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Enumeration of rooted binary perfect phylogenies

Chloe E. Shiff [a,*], Noah A. Rosenberg [b]

[a] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA
[b] Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 18 October 2024
Received in revised form 5 September 2025
Accepted 22 October 2025

Keywords:
Generating functions
Perfect phylogenies
Unlabeled trees

ABSTRACT

Rooted binary perfect phylogenies provide a generalization of rooted binary unlabeled trees. In a rooted binary perfect phylogeny, each leaf is assigned a positive integer value that corresponds in a biological setting to the count of the number of indistinguishable lineages associated with the leaf. For the rooted binary unlabeled trees, these integers equal 1. We enumerate rooted binary perfect phylogenies with $n \geq 1$ leaves and sample size s, $s \geq n$: the rooted binary unlabeled trees with n leaves in which a sample of size $s \geq n$ lineages is distributed across the n leaves. (1) First, we recursively enumerate rooted binary perfect phylogenies with sample size s, summing over all possible n, $1 \leq n \leq s$. We obtain an equation for the generating function, showing that asymptotically, the number of rooted binary perfect phylogenies with sample size s grows with $\approx 0.3519(3.2599)^s s^{-3/2}$, faster than the rooted binary unlabeled trees, which grow with $\approx 0.3188(2.4833)^s s^{-3/2}$. (2) Next, we recursively enumerate rooted binary perfect phylogenies with a specific number of leaves n and sample size $s \geq n$. We report closed-form counts of the rooted binary perfect phylogenies with sample size $s \geq n$ and $n = 2$, 3, and 4 leaves. We provide a recurrence for the generating function describing, for each number of leaves n, the number of rooted binary perfect phylogenies with n leaves as the sample size s increases. We also obtain an equation satisfied by the bivariate generating function counting rooted binary perfect phylogenies with n leaves and sample size s, as well as an asymptotic normal distribution for the number of leaves in a randomly chosen perfect phylogeny with sample size s. (3) We find a generating function for the number of rooted binary perfect phylogenies with the n-leaf caterpillar shape, growing with s. We also find a generating function and exact count $\lfloor 2^s/3 \rfloor$ for the number of rooted binary perfect phylogenies with sample size s and any caterpillar tree shape. A bivariate generating function counting rooted binary perfect phylogenies with n leaves, sample size s, and a caterpillar shape produces an asymptotic normal distribution for the number of leaves in a randomly chosen caterpillar perfect phylogeny with sample size s. (4) Finally, we provide initial results recursively enumerating rooted binary perfect phylogenies with any specific unlabeled tree shape and sample size s. The enumerations further characterize the rooted binary perfect phylogenies, which include the rooted binary unlabeled trees, and which can provide a set of structures useful for various biological contexts.

* Corresponding author.
E-mail addresses: cshiff@stanford.edu (C.E. Shiff), noahr@stanford.edu (N.A. Rosenberg).

https://doi.org/10.1016/j.dam.2025.10.053
0166-218X/

1. Introduction

Rooted binary unlabeled tree structures are classic objects of combinatorics and graph theory [5,8]. In evolutionary biology, rooted binary unlabeled trees are used to describe the possible relationships that a set of unlabeled organisms can possess, so that they arise in inferences about features of speciation histories [7,17].

The rooted binary unlabeled trees can be enumerated recursively. Denoting by u_n the number of rooted binary unlabeled trees with n leaves, for $n \geq 2$, the recursion is

$$u_n = \begin{cases} \sum_{i=1}^{n-1} \frac{1}{2} u_{n-i} u_i, & \text{odd } n \geq 3, \\ \left(\sum_{i=1}^{n-1} \frac{1}{2} u_{n-i} u_i \right) + \frac{1}{2} u_{n/2}, & \text{even } n \geq 2, \end{cases} \tag{1}$$

with $u_0 = 0$ and $u_1 = 1$. The recursion is obtained by summing over possible numbers of leaves i for the right-hand subtree descended from the root. The factor of $\frac{1}{2}$ arises from the fact that each tree is obtained twice—once with its left and right subtrees transposed. If n is even, for $i = \frac{n}{2}$, the recurrence counts the $\binom{u_{n/2}}{2}$ trees with distinct subtrees and the $u_{n/2}$ trees with identical subtrees. The u_n, $n \geq 1$, follow the Wedderburn–Etherington sequence (OEIS A001190), with initial terms 1, 1, 1, 2, 3, 6, 11, 23, 46, 98; they have implicitly defined generating function

$$U(z) = \frac{1}{2} U(z)^2 + \frac{1}{2} U(z^2) + z. \tag{2}$$

The convergence radius for $U(z)$ is approximately 0.4027. The asymptotic growth of u_n approximately follows $0.3188(2.4833)^n n^{-3/2}$ [3,11,15].

Rooted binary *perfect phylogenies* can be viewed as generalizing the rooted binary unlabeled trees. A rooted binary perfect phylogeny is a rooted binary tree in which each leaf is associated with a positive integer [16]. Each integer can be regarded as a multiplicity for the biological entity associated with a leaf—for example, the number of times that a specific DNA sequence is seen in a sample of sequences that are not necessarily distinct. A rooted binary perfect phylogeny has a number of leaves n and a sample size s that represents the sum of the multiplicities at the leaves. Rooted binary perfect phylogenies are a special case of rooted *multifurcating* perfect phylogenies—perfect phylogenies in which internal nodes possess two or more immediate descendants [16]. The rooted binary unlabeled trees correspond to rooted binary perfect phylogenies in the case that $s = n$; the leaf multiplicities of the perfect phylogenies all equal 1 in this equivalence.

In evolutionary biology, perfect phylogenies can sometimes be used as representations of the relationships of genetic sequences [9,16]. The topology of a perfect phylogeny encodes ancestral relationships in a set of sequences that have not experienced recombination, and in which each mutation has occurred only once. *Rooted* perfect phylogenies have one vertex designated as the root, representing a sequence from which all other sequences in the perfect phylogeny descend.

Palacios et al. [16] have recently developed the enumerative combinatorics of rooted perfect phylogenies, focusing on enumerations of various classes of binary trees that are *compatible* with a given rooted perfect phylogeny. Our focus here is different: we enumerate the possible rooted binary perfect phylogenies themselves.

After introducing definitions in Section 2, in Section 3, we recursively enumerate the rooted binary perfect phylogenies with sample size s, considering all possible values of the number of leaves n; we provide the asymptotic approximation of this quantity as s increases. In Section 4, we recursively enumerate rooted binary perfect phylogenies with a specific number of leaves n and sample size s. We provide a recursive equation to compute, for each n, the generating function for the sequence of the number of rooted binary perfect phylogenies with the number of leaves n fixed and the sample size s growing. We analyze the asymptotic growth with s of the number of rooted binary perfect phylogenies with a fixed number of leaves n, and we study the distribution of n in a randomly chosen perfect phylogeny with sample size s. In Section 5, we recursively enumerate the rooted binary perfect phylogenies with sample size s for a caterpillar tree shape. We obtain, for each small n, a closed-form expression for the number of perfect phylogenies for any s. We also provide generating functions for the numbers of perfect phylogenies with n-leaf caterpillar shapes, as well as a generating function and closed form for the total number of perfect phylogenies with sample size s across all caterpillar shapes. We analyze the asymptotic growth with s of the number of rooted binary perfect phylogenies with a caterpillar shape and n leaves, and we study the distribution of n in a randomly chosen perfect phylogeny with sample size s and a caterpillar shape. Section 6 then considers arbitrary tree shapes, obtaining a recurrence for the number of perfect phylogenies with sample size s for any specific tree shape.

2. Preliminaries

2.1. Definitions

We restrict attention to rooted perfect phylogenies that are *binary*, with each internal node possessing exactly two child nodes. Henceforth, the perfect phylogenies that we consider are understood to be rooted and binary, and we sometimes

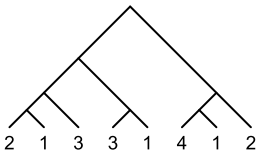

Fig. 1. A perfect phylogeny with sample size $s = 17$ and $n = 8$ leaves. The numbers at the leaves represent leaf multiplicities.

omit these descriptors. Like the rooted binary unlabeled trees, we consider perfect phylogenies to be *non-plane* trees, so that the left–right order in which child nodes are depicted is ignored.

We denote the *number of leaves* in a perfect phylogeny by n. Each leaf is associated with a positive integer, its *multiplicity*—representing in biological applications of perfect phylogenies the number of copies of a biological sequence seen in a sample of sequences. We refer to the sum of the multiplicities in a perfect phylogeny as its *sample size*.

For fixed sample size s, it is convenient to allow an *empty* perfect phylogeny, though we exclude this empty perfect phylogeny from our enumerations. We do include the *trivial* perfect phylogeny with sample size s, namely a perfect phylogeny that consists only of a single leaf of multiplicity s.

Fig. 1 displays an example perfect phylogeny with $n = 8$ leaves. The leaf multiplicities are 2, 1, 3, 3, 1, 4, 1, and 2, for a total sample size $s = 17$.

2.2. Lattices of perfect phylogenies

In the work of Palacios et al. [16], all possible rooted binary trees (with all leaf multiplicities equal to 1) that can be "collapsed" into a specific perfect phylogeny (with leaf multiplicities possibly greater than 1) are enumerated. To facilitate the enumerations, Palacios et al. [16] defined a partial order on rooted binary perfect phylogenies with fixed sample size s, inducing a lattice structure for those rooted binary perfect phylogenies. The lattice structure defines the sense in which trees can be "collapsed."

In particular, recall that a *cherry node* in a rooted tree is an internal node with precisely two descendant leaves. Consider binary perfect phylogenies A and B. In the partial order, A *refines* B if by collapsing cherries of A, B can be produced—where *collapsing* a cherry involves replacing the cherry node by a leaf node with multiplicity equal to the sum of the multiplicities of the leaves previously descended from the cherry node. Trivially, a perfect phylogeny refines itself. A and B are *comparable* if A refines B or B refines A.

Considering all binary perfect phylogenies with sample size s, the partial order of Palacios et al. [16] produces a lattice. Fig. 2 depicts the lattice for the case of $s = 5$. Moving left to right, a path is drawn between pairs (A, B), with A to the left of B, if and only if A refines B. The trivial perfect phylogeny of sample size s is refined by all perfect phylogenies of sample size s and is the maximal element of the lattice. The empty perfect phylogeny refines all perfect phylogenies of sample size s and is the minimal element.

Palacios et al. [16] focused on using the lattice to enumerate rooted binary trees associated with a rooted binary perfect phylogeny. However, the lattice formulation provides a convenient structure for working with perfect phylogenies themselves.

2.3. Description of the enumeration problems

We enumerate several sets of objects. First, we consider the set of (non-empty) rooted binary perfect phylogenies with fixed sample size $s \geq 1$. Denote the size of this set by b_s. Next, we enumerate the rooted binary perfect phylogenies with fixed sample size s and fixed number of leaves n, $1 \leq n \leq s$. Denote the size of this set by $b_{s,n}$; we have $b_s = \sum_{n=1}^{s} b_{s,n}$. We enumerate the rooted binary perfect phylogenies with sample size s and a caterpillar topology with n leaves, where $s \geq n \geq 2$, denoting this quantity $g_{s,n}$; we also enumerate the rooted binary phylogenies with sample size s and *any* caterpillar topology, denoting this quantity $g_s = \sum_{n=2}^{s} g_{s,n}$. Finally, we enumerate the rooted binary perfect phylogenies with sample size s and a specified unlabeled topology T, denoting this quantity $N_{s,T}$.

Note that we have already described $b_{n,n}$ (alternatively, $b_{s,s}$), the number of rooted binary perfect phylogenies with sample size equal to the number of leaves, in Eq. (1). A rooted binary perfect phylogeny with $s = n$ is simply a rooted binary unlabeled tree; each leaf multiplicity in the perfect phylogeny is 1, so that the rooted binary unlabeled trees correspond to the rooted binary perfect phylogenies in which all leaf multiplicities equal 1. Hence, $b_{n,n} = u_n$.

3. Rooted binary perfect phylogenies with sample size s

3.1. Enumeration

To enumerate all rooted binary perfect phylogenies with a fixed sample size $s \geq 1$, we first note that for each s, the trivial perfect phylogeny is permissible. If a perfect phylogeny is not trivial, then each of the two child nodes of the root

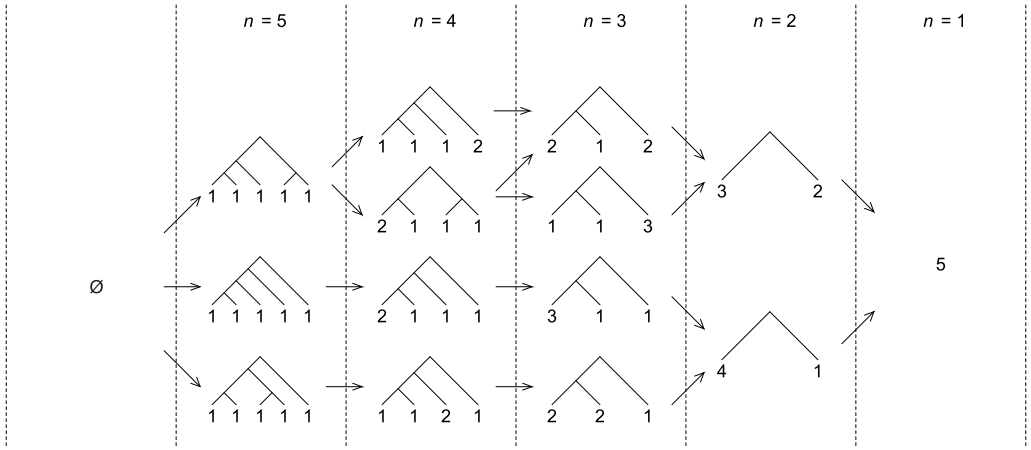

Fig. 2. The lattice of rooted binary perfect phylogenies for sample size $s = 5$. Each column is labeled by its associated number of leaves n.

is itself the root of a perfect phylogeny. In other words, for $s \geq 2$, the perfect phylogeny can be decomposed into two perfect phylogenies, one with sample size i, $1 \leq i \leq s-1$, and the other with sample size $s-i$.

For $i = s-i$, we count the $\binom{b_{s/2}}{2} = b_{s/2}(b_{s/2}-1)/2$ perfect phylogenies with distinct perfect phylogenies in the two children of the root and the $b_{s/2}$ perfect phylogenies with identical perfect phylogenies in the subtrees: $\binom{b_{s/2}}{2} + b_{s/2} = \frac{1}{2}(b_{s/2}^2 + b_{s/2})$. We obtain the following result.

Proposition 1. *The number b_s of rooted binary perfect phylogenies with sample size $s \geq 2$ satisfies*

$$b_s = \begin{cases} 1 + \sum_{i=1}^{s-1} \frac{1}{2} b_{s-i} b_i, & \text{odd } s \geq 3, \\ 1 + \left(\sum_{i=1}^{s-1} \frac{1}{2} b_{s-i} b_i \right) + \frac{1}{2} b_{s/2}, & \text{even } s \geq 2, \end{cases} \tag{3}$$

with $b_0 = 0$ and $b_1 = 1$.

The recursion has the same form as Eq. (1), adding a $+1$ term for the trivial perfect phylogeny. The first terms of the sequence appear in Table 1, along with the Wedderburn–Etherington numbers of rooted binary unlabeled trees. The number of rooted binary perfect phylogenies b_s with sample size s appears to grow substantially faster than $b_{s,s}$, the number of rooted binary unlabeled trees with sample size s and multiplicity 1 assigned to each leaf.

3.2. Generating function

To analyze the asymptotic growth of the rooted binary perfect phylogenies with sample size s as $s \to \infty$, we rewrite Eq. (3) in the form

$$b_s = \frac{1}{2}\left(\sum_{i=1}^{s-1} b_{s-i} b_i \right) + \frac{1}{2} b_{s/2} + 1, \quad s \geq 1, \tag{4}$$

with base case $b_0 = 0$ and $b_s = 0$ if s is not a positive integer.

Denote by $B(z)$ the generating function for the rooted binary perfect phylogenies with sample size s, $B(z) = \sum_{s=0}^{\infty} b_s z^s$. To obtain the generating function for the b_s, we multiply Eq. (4) by z^s and sum from $s = 1$ to ∞, obtaining

$$B(z) = \sum_{s=1}^{\infty} \left(\frac{1}{2} \sum_{i=1}^{s-1} b_{s-i} b_i z^s \right) + \sum_{s=1}^{\infty} \frac{1}{2} b_{s/2} z^s + \sum_{s=1}^{\infty} z^s.$$

We simplify by noting $\sum_{s=1}^{\infty} (\frac{1}{2} \sum_{i=1}^{s-1} b_{s-i} b_i z^s) = \frac{1}{2} B(z)^2$, $\sum_{s=1}^{\infty} \frac{1}{2} b_{s/2} z^s = \frac{1}{2} \sum_{s=1}^{\infty} b_s z^{2s} = \frac{1}{2} B(z^2)$, and $\sum_{s=1}^{\infty} z^s = \frac{z}{1-z}$. We have therefore demonstrated the following proposition.

Proposition 2. *The generating function $B(z)$ for the number b_s of rooted binary perfect phylogenies with sample size s satisfies*

$$B(z) = \frac{1}{2} B(z)^2 + \frac{1}{2} B(z^2) + \frac{z}{1-z}. \tag{5}$$

Table 1
The number of rooted binary perfect phylogenies b_s with sample size s (Eq. (3), OEIS A113822), and the number of rooted binary unlabeled trees $b_{s,s} = u_s$ (Eq. (1), OEIS A001190).

s	1	2	3	4	5	6	7	8	9	10	11	12	13	14
b_s	1	2	3	7	14	35	85	226	600	1658	4622	13 141	37 699	109 419
$b_{s,s}$	1	1	1	2	3	6	11	23	46	98	207	451	983	2179

3.3. Asymptotics

Now that we have obtained an equation satisfied by the generating function for the coefficients b_s, we find an asymptotic approximation for the growth of the b_s.

Recall that the generating function $U(z)$ for the number of rooted binary unlabeled trees in Eq. (2) has a radius of convergence $\rho \approx 0.4027$. The rooted binary perfect phylogenies with sample size s include the rooted binary unlabeled trees with s leaves, so that $b_s \geq u_s$, and indeed $b_s > u_s$ for $s \geq 2$. Hence, we have $B(z) > U(z)$ for all z with $0 < z < \rho$. Labeling the radius of convergence of $B(z)$ by β, it follows that $0 \leq \beta \leq \rho < 1$. In addition, because $z^2 < z$ for $0 < z < 1$ and $B(z)$ is monotonically increasing with z for $z > 0$, $B(z^2) < B(z)$ for $0 < z < 1$, so that if $B(z)$ converges at z, $0 < z < 1$, then it also converges at z^2.

To obtain the asymptotic approximation, we first prove a lemma about the relationship of the b_s to the Catalan numbers.

Lemma 3. *Each b_s for $s \geq 1$ is bounded above by the Catalan number $C_s = \frac{1}{s+1}\binom{2s}{s}$.*

Proof. To prove $b_s \leq C_s$ for all $s \geq 1$, we recall the recursion for the Catalan numbers, $C_s = \sum_{k=0}^{s-1} C_k C_{s-1-k} = \sum_{k=1}^{s} C_{k-1} C_{s-k}$ with $C_0 = 1$ [5, p. 26]. We first prove inductively that $\frac{1}{2} b_{s+1} \leq C_s$ for all $s \geq 0$.

For the base case $s = 0$, we have $\frac{1}{2} = \frac{1}{2} b_1 \leq C_0 = 1$; for $s = 1$, we have $1 = \frac{1}{2} b_2 \leq C_1 = 1$. For the inductive step, suppose $\frac{1}{2} b_{s+1} \leq C_s$ for all $s < N$, $N \geq 2$. By the recursion in Eq. (4) and the inductive assumption,

$$\begin{aligned}\frac{1}{2} b_{N+1} &= \left(\sum_{i=2}^{N} \frac{1}{2} b_{N+1-i} \frac{1}{2} b_i\right) + \frac{1}{4} b_N b_1 + \frac{1}{4} b_{(N+1)/2} + \frac{1}{2} \\ &\leq \left(\sum_{i=2}^{N} C_{N-i} C_{i-1}\right) + \frac{1}{4} b_N b_1 + \frac{1}{4} b_{(N+1)/2} + \frac{1}{2}.\end{aligned}$$

By the inductive assumption, noting $b_1 = C_0 = 1$, $\frac{1}{4} b_N b_1 \leq \frac{1}{2} C_{N-1} C_0$. Also by the inductive assumption, $\frac{1}{4} b_{(N+1)/2} + \frac{1}{2} \leq \frac{1}{2} C_{(N-1)/2} + \frac{1}{2}$. The Catalan numbers are strictly monotonically increasing for $N \geq 1$, so that $\frac{1}{2} C_{(N-1)/2} + \frac{1}{2} \leq \frac{1}{2} C_{N-1} = \frac{1}{2} C_{N-1} C_0$ for $N \geq 2$.

We then have $\frac{1}{4} b_N b_1 + \frac{1}{4} b_{(N+1)/2} + \frac{1}{2} \leq C_{N-1} C_0$ and $\frac{1}{2} b_{N+1} \leq \sum_{i=1}^{N} C_{N-i} C_{i-1} = C_N$, and the induction is complete.

To complete the proof that $b_s \leq C_s$ for $s \geq 1$, we proceed again by induction. We note that the result holds in the base case $s = 1$ ($1 = b_1 \leq C_1 = 1$) and $s = 2$ ($2 = b_2 \leq C_2 = 2$), and suppose that it holds for all $s \leq N$, $N \geq 2$. Then

$$\begin{aligned}b_{N+1} &= \left(\sum_{i=2}^{N} \frac{1}{2} b_{N+1-i} b_i\right) + \frac{1}{2} b_N b_1 + \frac{1}{2} b_{(N+1)/2} + 1 \\ &\leq \left(\sum_{i=2}^{N} C_{N+1-i} C_{i-1}\right) + \frac{1}{2} b_N b_1 + \frac{1}{2} b_{(N+1)/2} + 1.\end{aligned}$$

We have $\frac{1}{2} b_N b_1 \leq \frac{1}{2} C_N C_0$ by the inductive hypothesis, and $\frac{1}{2} b_{(N+1)/2} + 1 \leq C_{(N-1)/2} + 1 \leq C_N = C_N C_0$ by the earlier $\frac{1}{2} b_{s+1} \leq C_s$ and the strict monotonicity of the C_N for $N \geq 2$. Then $\frac{1}{2} b_N b_1 + \frac{1}{2} b_{(N+1)/2} + 1 \leq \frac{3}{2} C_N C_0$ and $b_{N+1} \leq \left(\sum_{i=2}^{N} C_{N+1-i} C_{i-1}\right) + \frac{3}{2} C_N C_0 = \left(\sum_{i=1}^{N} C_{N+1-i} C_{i-1}\right) + \frac{1}{2} C_N C_0 < \sum_{i=1}^{N+1} C_{N+1-i} C_{i-1} = C_{N+1}$. □

Corollary 4. *The radius of convergence β for $B(z)$ is positive, and in particular, $\frac{1}{4} \leq \beta \leq \rho$.*

Proof. We have seen that $\beta \leq \rho \approx 0.4027$ because the rooted binary perfect phylogenies include the rooted binary unlabeled trees, whose generating function has radius of convergence ρ.

For the lower bound, because the Catalan generating function $C(z) = (1 - \sqrt{1 - 4z})/(2z)$ has radius of convergence $\frac{1}{4}$, it follows from Lemma 3 that $B(z) \leq C(z)$ for $0 < z < \frac{1}{4}$, so that the generating function $B(z)$ for the smaller sequence $\{b_n\}$ has radius of convergence $\beta \geq \frac{1}{4}$. □

Theorem 5. *The number b_s of rooted binary perfect phylogenies with sample size s has asymptotic growth*

$$b_s \sim [\gamma/(2\sqrt{\pi})](1/r)^s s^{-3/2} \approx \frac{0.3519(3.2599)^s}{s^{3/2}}, \tag{6}$$

where $\gamma \approx 1.2476$ and $r \approx 0.3068$ are constants.

Because $B(z)$ is written in terms of $B(z^2)$ in Proposition 2, the proof of Theorem 5 relies on methods for generating functions defined implicitly. We use the smooth implicit-function schema in Theorem VII.3 from [8, pp. 467–468]. According to this theorem, we begin with an implicitly defined generating function $y(z) = \sum_{n=0}^{\infty} y_n z^n$ that takes the form $y(z) = H\big(z, y(z)\big)$. Suppose that $y(z)$ is analytic at 0, $y_0 = 0$, and $y_n \geq 0$. Suppose also that

i. $H(z, w) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h_{m,n} z^m w^n$ is analytic in a neighborhood of $(z, w) = (0, 0)$.
ii. $H(z, w)$ has coefficients $h_{m,n} \geq 0$ with $h_{0,0} = 0$, $h_{0,1} \neq 1$, and $h_{m,n} > 0$ for some (m, n) with $n \geq 2$.
iii. There exists some point $(z, w) = (r, s)$ in the analytic portion of the domain around $(0, 0)$, such that $H(r, s) = s$ and $H_w(r, s) = 1$.

Then $[z^n]y(z)$ grows with $[\gamma/(2\sqrt{\pi})](1/r)^n n^{-3/2}$, where $\gamma = \sqrt{2rH_z(r, s)/H_{ww}(r, s)}$.

Proof. We verify that $B(z)$ belongs to the smooth implicit-function schema. Eq. (5) gives the implicitly defined generating function. Write $H(z, w) = \frac{1}{2}w^2 + \frac{1}{2}B(z^2) + \frac{z}{1-z} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} h_{m,n} z^m w^n$. We prove $H(z, w)$ satisfies the required conditions.

i. We show $H(z, w)$ is analytic in a neighborhood of $(0, 0)$. First note that $w^2/2$ is entire. Next, for β the radius of convergence of $B(z)$, $\frac{1}{2}B(z^2)$ is analytic for $|z| < \sqrt{\beta}$. Finally, $\frac{z}{1-z}$ is analytic for $z \neq 1$. Hence, noting $\beta < 1$, $H(z, w)$ is analytic for $|w| < \infty$ and $|z| < \sqrt{\beta}$.
ii. For the conditions on $h_{m,n}$, we examine the expansion of $H(z, w)$, and observe $h_{0,0} = 0$, $h_{0,1} = 0 \neq 1$, and $h_{0,2} = \frac{1}{2} > 0$. Each $h_{m,n}$ satisfies $h_{m,n} \geq 0$ for $m, n \geq 0$, as $B(z^2)$ and $z/(1-z)$ have nonnegative coefficients.
iii. We show there exists a solution to the characteristic system

$$H_w(z, w) = 1,\ H(z, w) = w.$$

We first note that $H_w(z, w) = w$, so $H_w(r, s) = 1$ is satisfied for $s = 1$. Thus, we need only show that there exists r with $|r| < \sqrt{\beta}$ such that

$$H(r, 1) = \frac{1}{2} + \frac{1}{2}B(r^2) + \frac{r}{1-r} = 1. \tag{7}$$

Restricting our attention to the positive, real line, we note that:

1. $H(z, 1)$ is a monotonically increasing function for real $z > 0$, as it is a sum of power series with nonnegative coefficients.
2. $H(0, 1) = \frac{1}{2}$, as neither $B(z^2)$ nor $\frac{z}{1-z}$ has a constant term.
3. $H(\frac{1}{3}, 1) > 1$, as $\frac{1}{3}/(1 - \frac{1}{3}) = \frac{1}{2}$ and $B\big((\frac{1}{3})^2\big) > 0$ (because $B(z^2)$ has all non-negative coefficients and at least one positive term; e.g. the coefficient of z^2 is 1). Note also that $\frac{1}{3} < \frac{1}{2} \leq \sqrt{\beta}$ by Corollary 4.

We conclude that there exists some r, $0 \leq r < \frac{1}{3} < \sqrt{\beta}$ such that $H(r, 1) = 1$.

We have therefore shown that $B(z)$ belongs to the smooth implicit-function schema. The smooth implicit-function schema tells us that the same r that solves the characteristic system is indeed the radius of convergence of $B(z)$. With $s = 1$, we solve Eq. (7) for r numerically. We approximate $B(r^2)$ using the terms in Table 1: $B(r^2) \approx \sum_{i=1}^{14} b_i (r^2)^i = 1r^2 + 2r^4 + 3r^6 + 7r^8 + \cdots + 109419r^{28}$. Numerically solving for the positive, real root, we obtain $r \approx 0.306760104888$.

To compute the constant γ, we use $H_z(r, s) = rB'(r^2) + 1/(1-r)^2$ and $H_{ww}(r, s) = 1$. We approximate $B'(r^2)$ by the terms in Table 1:

$$B'(r^2) \approx \sum_{i=1}^{14} i b_i (r^2)^{i-1} = (1 \cdot 1)r^0 + (2 \cdot 2)r^2 + (3 \cdot 3)r^4 + (4 \cdot 7)r^6 + \cdots + (14 \cdot 109419)r^{26}. \tag{8}$$

We obtain $B'(r^2) \approx 1.4871$. Then

$$\gamma = \sqrt{2r\left[rB'(r^2) + \frac{1}{(1-r)^2}\right]} \approx 1.2476,$$

from which $b_s \sim [\gamma/(2\sqrt{\pi})](1/r)^s s^{-3/2} \approx 0.3519(3.2599)^s/s^{3/2}$. □

Fig. 3 plots the logarithm of the exact number of rooted binary perfect phylogenies b_s from Eq. (3) alongside the logarithm of the asymptotic growth from Theorem 5. We can observe, for example, that the asymptotic approximation $0.3519(3.2599)^{60}/60^{3/2}$ gives 4.6930×10^{27}; the exact value is 4, 753, 678, 474, 171, 125, 902, 623, 929, 051.

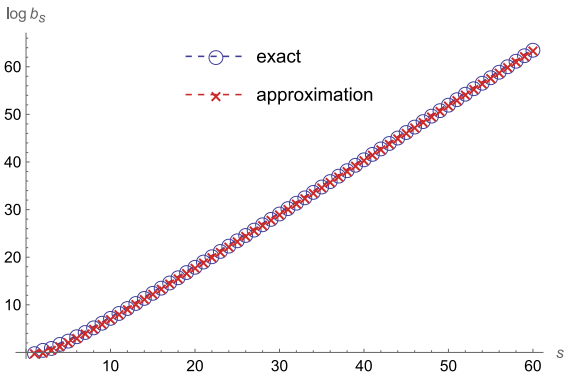

Fig. 3. The number b_s of perfect phylogenies with sample size s. Exact values are computed from Eq. (3). The asymptotic approximation is computed from Eq. (6).

4. Rooted binary perfect phylogenies with sample size s and n leaves

Having enumerated all rooted binary perfect phylogenies with sample size s, we now decompose the enumeration across perfect phylogenies with different numbers of leaves. A (non-empty) perfect phylogeny with sample size s must possess a number of leaves in $[1, s]$. In the case that the number of leaves n is equal to s, the rooted binary perfect phylogenies are simply rooted binary unlabeled trees, as each leaf has multiplicity 1.

4.1. Enumeration

We generalize to consider all pairs (s, n) with $1 \leq n \leq s$. Let $b_{s,n}$ be the number of rooted binary perfect phylogenies with sample size s and n leaves, where $b_{s,n} = 0$ if $s < n$, or $s \notin \mathbb{N}$, or $n \notin \mathbb{N}$.

Proposition 6. *The number $b_{s,n}$ of rooted binary perfect phylogenies with sample size s and n leaves, $1 \leq n \leq s$, satisfies*
(i) $b_{s,1} = 1$ for all $s \geq 1$.
(ii) For (s, n) with $s \geq n \geq 2$,

$$b_{s,n} = \begin{cases} \displaystyle\sum_{j=1}^{n-1} \sum_{i=j}^{s-n+j} \tfrac{1}{2} b_{s-i,n-j}\, b_{i,j}, & s \text{ odd or } n \text{ odd}, \\ \left(\displaystyle\sum_{j=1}^{n-1} \sum_{i=j}^{s-n+j} \tfrac{1}{2} b_{s-i,n-j}\, b_{i,j} \right) + \tfrac{1}{2} b_{s/2,n/2}, & s \text{ even and } n \text{ even}. \end{cases} \tag{9}$$

In Eq. (9), the index i counts the sample size assigned to the right subtree and the index j counts its number of leaves. The left subtree then has sample size $s - i$, with $n - j$ leaves.

We observe that in the case $s = n$, in which a perfect phylogeny has all leaf multiplicities equal to 1, the recursion recovers Eq. (1). In this case, we must have $i = j$. Because $s = n$, the cases become cases for n odd or n even. Because $n = s$ and $j = i$, we obtain for $n \geq 3$ odd:

$$b_{n,n} = \sum_{j=1}^{n-1} \frac{1}{2} b_{n-j,n-j} b_{j,j} = \sum_{j=1}^{n-1} \frac{1}{2} u_{n-j}\, u_j = u_n.$$

The case for even $n \geq 2$ reduces to

$$b_{n,n} = \left(\sum_{j=1}^{n-1} \frac{1}{2} b_{n-j,n-j}\, b_{j,j} \right) + \frac{1}{2} b_{n/2,n/2} = \left(\sum_{j=1}^{n-1} \frac{1}{2} u_{n-j}\, u_j \right) + \frac{1}{2} u_{n/2} = u_n.$$

Proof. We count rooted binary perfect phylogenies with sample size s and n leaves by considering all partitions of the sample and leaves into left and right subtrees. We index the sample size of the right subtree by i and the number of leaves of the right subtree by j.

The right subtree has sample size $i \geq j$. Because the left subtree has $n - j$ leaves, it has sample size at least $n - j$, so that the right subtree has sample size at most $i \leq s - (n - j)$.

Table 2
The number $b_{s,n}$ of rooted binary perfect phylogenies with sample size s and n leaves. Entries are obtained using Eq. (9); the "total" is $b_s = \sum_{n=1}^{s} b_{s,n}$. The total follows OEIS sequence A113822; $b_{s,3}$ follows A002620. The main diagonal and its subdiagonal follow A001190 and A085748. For completeness, $b_{12,12} = 451$, $b_{13,12} = 3264$, $b_{13,13} = 983$, $b_{14,12} = 15\,886$, $b_{14,13} = 7777$, and $b_{14,14} = 2179$.

Sample size (s)	Number of leaves (n)											Total
	11	10	9	8	7	6	5	4	3	2	1	
1											1	1
2										1	1	2
3									1	1	1	3
4								2	2	2	1	7
5							3	4	4	2	1	14
6						6	9	10	6	3	1	35
7					11	20	24	17	9	3	1	85
8				23	46	61	49	30	12	4	1	226
9			46	106	152	138	93	44	16	4	1	600
10		98	248	386	387	290	157	66	20	5	1	1658
11	207	582	974	1072	878	535	253	90	25	5	1	4622
12	1376	2473	2951	2633	1774	939	383	124	30	6	1	13 141
13	6262	8061	7763	5727	3340	1534	562	160	36	6	1	37 699
14	21 899	22 657	18 119	11 551	5881	2420	792	208	42	7	1	109 419

Given the number of leaves j for the right subtree, $1 \le j \le n-1$, and sample size $i, j \le i \le s-n+j$, if $j \neq \frac{n}{2}$ or $i \neq \frac{s}{2}$ or both, then we count $b_{s-i,n-j}\, b_{i,j}$ perfect phylogenies: $b_{i,j}$ for the right subtree and $b_{s-i,n-j}$ for the left. Each distinct perfect phylogeny is obtained twice, the second time with the left and right subtrees transposed.

If the sample size s and number of leaves n are both even, then we count both the $\binom{b_{s/2,n/2}}{2}$ perfect phylogenies with two distinct subtrees of sample size $\frac{s}{2}$ and $\frac{n}{2}$ leaves as well as the $b_{s/2,n/2}$ perfect phylogenies with two identical subtrees, as in Eqs. (3) and (9). □

Fig. 4 shows an example of the enumeration, considering all possible rooted binary perfect phylogenies with $(s, n) = (8, 6)$. Table 2 gives the values of $b_{s,n}$ for small (s, n), illustrating that the sums $\sum_{n=1}^{s} b_{s,n}$ agree with the values obtained for b_s via Proposition 1.

For fixed small n, $b_{s,n}$ can be stated in closed form. First, $b_{s,1} = 1$ for $s \ge 1$. We obtain a sequence of corollaries of Proposition 6.

Corollary 7. *For $s \ge 2$, the number $b_{s,2}$ of rooted binary perfect phylogenies with $n = 2$ leaves is*

$$b_{s,2} = \left\lfloor \frac{s}{2} \right\rfloor. \tag{10}$$

Proof. From Proposition 6, we have

$$b_{s,2} = \begin{cases} \displaystyle\sum_{i=1}^{s-1} \tfrac{1}{2} b_{s-i,1} b_{i,1}, & \text{odd } s \ge 3, \\ \left(\displaystyle\sum_{i=1}^{s-1} \tfrac{1}{2} b_{s-i,1} b_{i,1} \right) + \tfrac{1}{2} b_{s/2,1}, & \text{even } s \ge 2. \end{cases}$$

Because $b_{s,1} = 1$ for all $s \ge 1$, we obtain $b_{s,2} = \frac{s-1}{2}$ for odd $s \ge 3$, and $b_{s,2} = \frac{s}{2}$ for even $s \ge 2$. Summarizing the odd and even cases in one expression, $b_{s,2} = \lfloor \frac{s}{2} \rfloor$. □

Corollary 8. *For $s \ge 3$, the number $b_{s,3}$ of rooted binary perfect phylogenies with $n = 3$ leaves is*

$$b_{s,3} = \left\lfloor \frac{s-1}{2} \right\rfloor \left\lceil \frac{s-1}{2} \right\rceil. \tag{11}$$

Proof. Using Proposition 6,

$$b_{s,3} = \sum_{j=1}^{2} \sum_{i=j}^{s-3+j} \frac{1}{2} b_{s-i,3-j} b_{i,j} = \sum_{i=2}^{s-1} b_{s-i,1} b_{i,2},$$

noting $\sum_{i=1}^{s-2} \frac{1}{2} b_{s-i,2} b_{i,1} = \sum_{i=2}^{s-1} \frac{1}{2} b_{s-i,1} b_{i,2}$ by an exchange of $s - i$ and i.

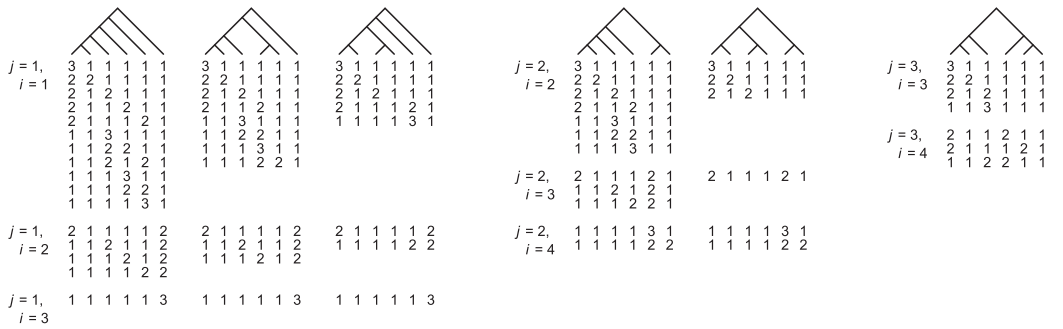

Fig. 4. The enumeration of all $b_{8,6} = 61$ rooted binary perfect phylogenies with sample size $s = 8$ and $n = 6$ leaves. The number of leaves in the right subtree is indicated by j, and i indicates the sample size for the right subtree.

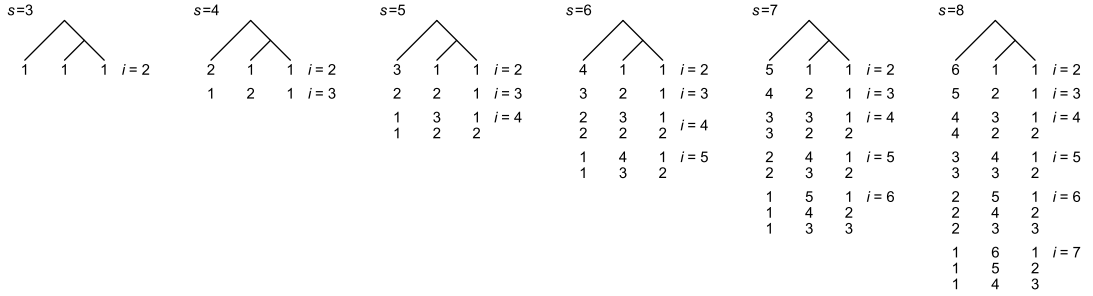

Fig. 5. The $b_{s,3}$ rooted binary perfect phylogenies with $n = 3$ leaves, for each s from 3 to 8, as obtained by Proposition 6. The value of i indicates the sample size for the right subtree.

Using Corollary 7, $b_{s,3} = \sum_{i=2}^{s-1} \lfloor \frac{i}{2} \rfloor$. The summation yields

$$b_{s,3} = \begin{cases} 2\left(1 + 2 + \cdots + \frac{s-3}{2}\right) + \frac{s-1}{2} = \left(\frac{s-1}{2}\right)\left(\frac{s-1}{2}\right), & \text{odd } s \geq 3, \\ 2\left(1 + 2 + \cdots + \frac{s-2}{2}\right) = \left(\frac{s-2}{2}\right)\left(\frac{s}{2}\right), & \text{even } s \geq 4. \end{cases}$$

We can summarize both cases in the single expression $b_{s,3} = \lfloor \frac{s-1}{2} \rfloor \lceil \frac{s-1}{2} \rceil$. □

For small s, the rooted binary perfect phylogenies with $n = 3$ leaves appear in Fig. 5.

Corollary 9. *For $s \geq 4$, the number $b_{s,4}$ of rooted binary perfect phylogenies with $n = 4$ leaves is*

$$b_{s,4} = \begin{cases} \frac{(s-1)(s-3)(5s-1)}{48}, & \textit{odd } s \geq 5, \\ \frac{s(s-2)(5s-11)}{48} + \frac{1}{2}\lfloor \frac{s}{4} \rfloor, & \textit{even } s \geq 4. \end{cases} \tag{12}$$

Proof. Using Eq. (9), we see that:

$$b_{s,4} = \begin{cases} \displaystyle\sum_{j=1}^{3} \sum_{i=j}^{s-4+j} \frac{1}{2} b_{s-i,4-j}\, b_{i,j}, & \text{odd } s \geq 5, \\ \left(\displaystyle\sum_{j=1}^{3} \sum_{i=j}^{s-4+j} \frac{1}{2} b_{s-i,4-j}\, b_{i,j} \right) + \frac{1}{2} b_{s/2,2}, & \text{even } s \geq 4. \end{cases}$$

The outer sum considers values of 1, 2, and 3 for the sample size j assigned to the right subtree. Assigning $j = 3$ gives the same inner sum as $j = 1$, as $\sum_{i=3}^{s-1} \frac{1}{2} b_{s-i,1} b_{i,3} = \sum_{i=1}^{s-3} \frac{1}{2} b_{s-i,3} b_{i,1}$ by an exchange of i and $s - i$. Noting that $b_{s,1} = 1$ for $s \geq 1$, the problem becomes:

$$b_{s,4} = \begin{cases} \displaystyle\sum_{i=1}^{s-3} b_{s-i,3} + \sum_{i=2}^{s-2} \frac{1}{2} b_{s-i,2} b_{i,2}, & \text{odd } s \geq 5, \\ \left(\displaystyle\sum_{i=1}^{s-3} b_{s-i,3} + \sum_{i=2}^{s-2} \frac{1}{2} b_{s-i,2} b_{i,2} \right) + \frac{1}{2} b_{s/2,2}, & \text{even } s \geq 4. \end{cases}$$

For s odd, $s-3$ is even and $s-2$ is odd. We use Corollaries 7 and 8 to obtain the summands, resolving floor and ceiling functions separately for odd and even quantities. The sums are completed using $\sum_{k=1}^{n} k = n(n+1)/2$ and $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$. Then

$$\begin{aligned}
\sum_{i=1}^{s-3} b_{s-i,3} &= \sum_{i=1}^{s-3} \left\lfloor \frac{s-i-1}{2} \right\rfloor \left\lceil \frac{s-i-1}{2} \right\rceil \\
&= \sum_{k=1}^{\frac{s-3}{2}} \left\lfloor \frac{s-(2k-1)-1}{2} \right\rfloor \left\lceil \frac{s-(2k-1)-1}{2} \right\rceil + \left\lfloor \frac{s-2k-1}{2} \right\rfloor \left\lceil \frac{s-2k-1}{2} \right\rceil \\
&= \sum_{k=1}^{\frac{s-3}{2}} \left(\frac{s-2k-1}{2}\right)\left(\frac{s-2k+1}{2}\right) + \left(\frac{s-2k-1}{2}\right)^2 \\
&= (s-1)(s-3)(2s-1)/24. \qquad (13)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=2}^{s-2} b_{s-i,2}\, b_{i,2} &= \sum_{i=2}^{s-2} \left\lfloor \frac{s-i}{2} \right\rfloor \left\lfloor \frac{i}{2} \right\rfloor \\
&= \sum_{k=1}^{\frac{s-3}{2}} \left\lfloor \frac{s-(2k+1)}{2} \right\rfloor \left\lfloor \frac{2k+1}{2} \right\rfloor + \left\lfloor \frac{s-(2k)}{2} \right\rfloor \left\lfloor \frac{2k}{2} \right\rfloor \\
&= \sum_{k=1}^{\frac{s-3}{2}} \left(\frac{s-2k-1}{2}\right)k + \left(\frac{s-2k-1}{2}\right)k \\
&= (s+1)(s-1)(s-3)/24. \qquad (14)
\end{aligned}$$

Summing Eq. (13) and half of Eq. (14), we obtain the result in Eq. (12) for odd s.

The even case is similar, except that now $s-3$ is odd and $s-2$ is even.

$$\begin{aligned}
\sum_{i=1}^{s-3} b_{s-i,3} &= \sum_{i=1}^{s-3} \left\lfloor \frac{s-i-1}{2} \right\rfloor \left\lceil \frac{s-i-1}{2} \right\rceil \\
&= \left\lfloor \frac{s-(s-3)-1}{2} \right\rfloor \left\lceil \frac{s-(s-3)-1}{2} \right\rceil \\
&\quad + \sum_{k=1}^{\frac{s-4}{2}} \left\lfloor \frac{s-(2k-1)-1}{2} \right\rfloor \left\lceil \frac{s-(2k-1)-1}{2} \right\rceil + \left\lfloor \frac{s-2k-1}{2} \right\rfloor \left\lceil \frac{s-2k-1}{2} \right\rceil \\
&= 1 + \sum_{k=1}^{\frac{s-4}{2}} \left(\frac{s-2k}{2}\right)^2 + \left(\frac{s-2k-2}{2}\right)\left(\frac{s-2k}{2}\right) \\
&= s(s-2)(2s-5)/24. \qquad (15)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=2}^{s-2} b_{s-i,2}\, b_{i,2} &= \sum_{i=2}^{s-2} \left\lfloor \frac{s-i}{2} \right\rfloor \left\lfloor \frac{i}{2} \right\rfloor \\
&= \left\lfloor \frac{s-(s-2)}{2} \right\rfloor \left\lfloor \frac{s-2}{2} \right\rfloor + \sum_{k=1}^{\frac{s-4}{2}} \left\lfloor \frac{s-(2k+1)}{2} \right\rfloor \left\lfloor \frac{2k+1}{2} \right\rfloor + \left\lfloor \frac{s-(2k)}{2} \right\rfloor \left\lfloor \frac{2k}{2} \right\rfloor \\
&= \frac{s-2}{2} + \sum_{k=1}^{\frac{s-4}{2}} \left(\frac{s-2k-2}{2}\right)k + \left(\frac{s-2k}{2}\right)k \\
&= s(s-1)(s-2)/24. \qquad (16)
\end{aligned}$$

We obtain Eq. (12) for even s by summing Eq. (15), half of Eq. (16), and $\frac{1}{2} b_{s/2,2} = \frac{1}{2}\lfloor \frac{s}{4} \rfloor$. □

4.2. Generating function

Beyond the closed-form expressions for the cases of $n = 2$, 3, and 4, for a specific value of n more generally, we can obtain a generating function for the sequence $b_{s,n}$ describing the number of rooted binary perfect phylogenies with fixed n and increasing values of the sample size s.

Denote by $B_n(z)$ the generating function describing the sequence of values for the number of rooted binary perfect phylogenies with n leaves and increasing sample size s:

$$B_n(z) = \sum_{s=1}^{\infty} b_{s,n} z^s.$$

Recall that $b_{s,n} > 0$ only for integers (s, n) with $s \geq n \geq 1$. We set $B_n(z) = 0$ for non-integer n.

Proposition 10. *The generating function $B_n(z)$ for the number $b_{s,n}$ of rooted binary perfect phylogenies with sample size s and n leaves satisfies*
(i) $B_1(z) = \frac{z}{1-z}$.
(ii) For $n \geq 2$,

$$B_n(z) = \left[\sum_{j=1}^{n-1} \frac{1}{2} B_{n-j}(z)\, B_j(z)\right] + \frac{1}{2} B_{n/2}(z^2). \tag{17}$$

Proof. (i) For $n = 1$, for all $s \geq 1$, $b_{s,1} = 1$. Hence, $B_1(z) = \frac{z}{1-z}$.

(ii) For $n \geq 2$, we use the recursive Eq. (9),

$$\begin{aligned} B_n(z) &= \sum_{s=n}^{\infty} \left(\sum_{j=1}^{n-1} \sum_{i=j}^{s-n+j} \frac{1}{2} b_{s-i,n-j} b_{i,j} \right) z^s + \sum_{s=n}^{\infty} \frac{1}{2} b_{s/2,n/2} z^s \\ &= \sum_{s=n}^{\infty} \left(\sum_{j=1}^{n-1} \sum_{i=j}^{s-n+j} \frac{1}{2} b_{s-i,n-j} z^{s-i} b_{i,j} z^i \right) + \sum_{s=n}^{\infty} \frac{1}{2} b_{s/2,n/2} z^s. \end{aligned}$$

We adjust the summation limits, noting that $b_{s,n} = 0$ for $s < n$:

$$\begin{aligned} B_n(z) &= \sum_{s=2}^{\infty} \left(\sum_{j=1}^{n-1} \sum_{i=1}^{s-1} \frac{1}{2} b_{s-i,n-j} z^{s-i} b_{i,j} z^i \right) + \sum_{s=2}^{\infty} \frac{1}{2} b_{s/2,n/2} z^s \\ &= \left[\sum_{j=1}^{n-1} \frac{1}{2} B_{n-j}(z)\, B_j(z)\right] + \frac{1}{2} B_{n/2}(z^2), \end{aligned}$$

completing the proof. □

Iterating from the base case $B_1(z) = \frac{z}{1-z}$, we can write an explicit form for $B_n(z)$ for a fixed n. Using Proposition 10, we have

$$\begin{aligned} B_2(z) &= \frac{1}{2} B_1(z)^2 + \frac{1}{2} B_1(z^2) \\ &= \frac{1}{2} \frac{z^2}{(1-z)^2} + \frac{1}{2} \frac{z^2}{1-z^2} \\ &= \frac{z^2}{(1-z)^2(1+z)}. \end{aligned} \tag{18}$$

Next, using $B_2(z)$ from Eq. (18), we have

$$\begin{aligned} B_3(z) &= \frac{1}{2} B_1(z)\, B_2(z) + \frac{1}{2} B_2(z)\, B_1(z) \\ &= \frac{z^3}{(1-z)^3(1+z)}. \end{aligned} \tag{19}$$

For $B_4(z)$, we use Eq. (19):

$$\begin{aligned} B_4(z) &= \frac{1}{2}B_1(z)B_3(z) + \frac{1}{2}B_2^2(z) + \frac{1}{2}B_3(z)B_1(z) + \frac{1}{2}B_2(z^2) \\ &= \frac{z^4}{(1-z)^4(1+z)} + \frac{1}{2}\frac{z^4}{(1-z)^4(1+z)^2} + \frac{1}{2}\frac{z^4}{(1-z^2)^2(1+z^2)} \\ &= \frac{z^4(2+2z^2+z^3)}{(1-z)^4(1+z)^2(1+z^2)}. \end{aligned} \tag{20}$$

We can continue iteratively to get generating functions $B_n(z)$ for larger n.

4.3. Asymptotics

We next study the asymptotics of $b_{s,n}$ for fixed n, as $s \to \infty$.

Proposition 11. *As $s \to \infty$, for fixed $n \geq 1$, the number $b_{s,n}$ of rooted binary perfect phylogenies with sample size s and n leaves has asymptotic growth $b_{s,n} \sim [C_{n-1}s^{n-1}]/[2^{n-1}(n-1)!]$, where $C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1}$ is a Catalan number.*

First, we need a lemma concerning the generating function $B_n(z)$. The lemma uses the concept of a Δ-domain, which, informally, describes a certain region around a point that is enclosed by a ball centered at the point [8, p. 389]; a function that is analytic in a ball around the point (excluding at the point itself) is analytic in a Δ-domain at the point.

Lemma 12. *Fix $n \geq 1$. $B_n(z) \sim a_n/(1-z)^n$ as $z \to 1$ in a Δ-domain in the neighborhood of $z = 1$, for a constant a_n.*

Proof. From Proposition 10, we see that $B_1(z) = \frac{z}{1-z}$ and that $a_1 = 1$. By Eq. (18), $a_2 = \frac{1}{2}$, by Eq. (19), $a_3 = \frac{1}{2}$, and by Eq. (20), $a_4 = \frac{5}{8}$.

The recursive form of $B_n(z)$ in Proposition 10, together with $B_1(z) = \frac{z}{1-z}$, indicates that each $B_n(z)$ has a pole at $z = 1$ and potentially also at other roots of $z^k = 1$. $B_n(z)$ is otherwise analytic, and hence, it is analytic in a Δ-domain around $z = 1$.

We continue by induction, supposing that for each k, $1 \leq k \leq n$, $B_k(z) \sim a_k/(1-z)^k$ as $z \to 1$ in a Δ-domain around $z = 1$, for a constant a_k. We seek to show that $B_{n+1}(z) \sim a_{n+1}/(1-z)^{n+1}$ as $z \to 1$ in a Δ-domain. Starting from Eq. (17),

$$(1-z)^{n+1}B_{n+1}(z) = \left[\sum_{j=1}^{n}\frac{1}{2}B_{n+1-j}(z)(1-z)^{n+1-j}B_j(z)(1-z)^j\right] + (1-z)^{n+1}\left[\frac{1}{2}B_{(n+1)/2}(z^2)\right].$$

Taking $z \to 1$ and applying the inductive hypothesis,

$$\lim_{z\to 1}(1-z)^{n+1}B_{n+1}(z) = \sum_{j=1}^{n}\frac{1}{2}a_{n+1-j}a_j,$$

noting that the term $(1-z)^{n+1}\frac{1}{2}B_{(n+1)/2}(z^2)$ is zero for n even, and for n odd, it has limit $(1-z)^{n+1}\frac{1}{2}a_{(n+1)/2}/(1-z^2)^{(n+1)/2} \to 0$ as $z \to 1$. We have proven the lemma, with $a_n = \sum_{j=1}^{n-1}\frac{1}{2}a_{n-j}a_j$ for $n \geq 2$. □

Proof of Proposition 11. By Lemma 12, writing $A(z) = \sum_{n=1}^{\infty}a_nz^n$, we have

$$A(z) = z + \sum_{n=2}^{\infty}\left(\sum_{j=1}^{n-1}\frac{1}{2}a_{n-j}a_j\right)z^n = z + \frac{1}{2}A(z)^2,$$

from which $A(z) = 1 - \sqrt{1-2z}$. This generating function is the exponential generating function for the rooted labeled binary trees [17, p. 17], with $a_n = (2n-3)!!/n! = C_{n-1}/2^{n-1}$. Note that this expression recovers $a_2 = \frac{1}{2}$, $a_3 = \frac{1}{2}$, $a_4 = \frac{5}{8}$, in accord with Eqs. (18)–(20).

We have obtained $B_n(z) \sim C_{n-1}/[2^{n-1}(1-z)^n]$ as $z \to 1$ in a Δ-domain at $z = 1$. Corollary 2.16 from [5] states that if $a(z)$ is analytic in a delta-domain Δ such that $a(z) \sim D(1-z/z_0)^{-n}$ for $z \to z_0$, $z \in \Delta$, where D is a constant and n is a complex number that is not a non-positive integer, then as $s \to \infty$, $[z^s]\,a(z) \sim Ds^{n-1}z_0^{-s}/\Gamma(n)$.

We apply this corollary with $B_n(z)$ in the role of $a(z)$, $z_0 = 1$, and $D = C_{n-1}/2^{n-1}$. Noting that $\Gamma(n) = (n-1)!$, we obtain

$$b_{s,n} \sim \frac{C_{n-1}s^{n-1}}{2^{n-1}(n-1)!}. \quad \square \tag{21}$$

4.4. Bivariate generating function

As s grows large, we show that the distribution of the number of leaves n in a rooted binary perfect phylogeny selected uniformly at random among those with sample size s follows an asymptotic normal distribution. To obtain this result, we first obtain a bivariate generating function. Let

$$B(z,u)=\sum_{s=1}^{\infty}\sum_{n=1}^{\infty}b_{s,n}u^n z^s=\sum_{n=1}^{\infty}B_n(z)\,u^n. \tag{22}$$

Then by Proposition 10,

$$\begin{aligned}B(z,u)&=B_1(z)\,u^1+\sum_{n=2}^{\infty}\left[\left(\sum_{j=1}^{n-1}\frac{1}{2}B_{n-j}(z)B_j(z)\right)+\frac{1}{2}B_{n/2}(z^2)\right]u^n\\&=\frac{z}{1-z}u+\sum_{n=2}^{\infty}\left[\left(\sum_{j=1}^{n-1}\frac{1}{2}B_{n-j}(z)\,u^{n-j}\,B_j(z)\,u^j\right)+\frac{1}{2}B_n(z^2)\,u^{2n}\right]\\&=\frac{uz}{1-z}+\frac{1}{2}\left[\sum_{n=1}^{\infty}B_n(z)\,u^n\right]\left[\sum_{n=1}^{\infty}B_n(z)\,u^n\right]+\sum_{n=1}^{\infty}\frac{1}{2}B_{n/2}(z^2)\,u^n\\&=\frac{uz}{1-z}+\frac{1}{2}B(z,u)^2+\frac{1}{2}B(z^2,u^2).\end{aligned} \tag{23}$$

4.5. Asymptotic normal distribution for the number of leaves

We now proceed to the asymptotic normal distribution.

Theorem 13. *Let X_s denote the random variable describing the number of leaves of a perfect phylogeny chosen at random with sample size s. Then as s grows large,*
(i) $\mathbb{E}[X_s]\sim\mu s$, with $\mu\approx 0.7326$,
(ii) $\mathrm{Var}[X_s]\sim\sigma^2 s$, with $\sigma^2\approx 0.2325$,
(iii) The distribution of X_s has a limiting normal distribution

$$\frac{X_s-\mathbb{E}[X_s]}{\sqrt{\mathrm{Var}[X_s]}}\to N(0,1).$$

Proof. We use the combinatorial central limit theorem in Theorem 2.23 in [5]. Suppose X_s is a sequence of random variables with

$$\mathbb{E}[u^{X_s}]=\frac{[z^s]\,y(z,u)}{[z^s]\,y(z,1)},$$

where $y(z,u)$ is a power series that is the solution of $y=F(z,y,u)$. Suppose also that the following conditions hold for F:

1. $F(z,y,u)=\sum_{s=0}^{\infty}\sum_{m=0}^{\infty}F_{s,m}(u)\,z^s y^m$ is analytic around $(0,0,0)$.
2. The coefficients $F_{s,m}(1)$ are real and non-negative;
3. There exists (z_0,y_0) such that $y_0=F(z_0,y_0,1)$ and $1=F_y(z_0,y_0,1)$, with $F_z(z_0,y_0,1)\neq 0$ and $F_{yy}(z_0,y_0,1)\neq 0$.

Then (i) $\mathbb{E}[X_s]\sim\mu s$, where $\mu=F_u(z_0,y_0,1)/[z_0F_z(z_0,y_0,1)]$.
(ii) $\mathrm{Var}[X_s]\sim\sigma^2 s$, where

$$\sigma^2=\mu+\mu^2+\frac{1}{z_0F_z^3F_{yy}}\left[F_z^2(F_{yy}F_{uu}-F_{yu}^2)-2F_zF_u(F_{yy}F_{zu})+F_u^2(F_{yy}F_{zz}-F_{yz}F_{yu})\right], \tag{24}$$

and all derivatives of F are evaluated at $(z_0,y_0,1)$.

(iii) If $\sigma^2>0$, then $(X_s-\mathbb{E}[X_s])/\sqrt{\mathrm{Var}[X_s]}$ has a limiting standard normal distribution $N(0,1)$.

To apply the combinatorial central limit theorem, denote by X_s the number of leaves of a perfect phylogeny with sample size s, selected uniformly at random. In terms of generating functions $B(z)$ and $B(z,u)$,

$$\mathbb{E}[X_s]=\frac{\sum_{n=1}^{s}nb_{s,n}}{b_s}=\frac{\sum_{n=1}^{s}nb_{s,n}}{[z^s]\,B(z)}. \tag{25}$$

In the same manner, because the bivariate generating function has $[z^s]\,B(z,u)=\sum_{n=1}^{s}b_{s,n}u^n$,

$$\mathbb{E}[u^{X_s}]=\frac{\sum_{n=1}^{s}b_{s,n}u^n}{b_s}=\frac{[z^s]\,B(z,u)}{[z^s]\,B(z)}. \tag{26}$$

We let $B(z,u) = F\big(z, B(z,u), u\big)$, so that in Eq. (23),

$$F(z,y,u) = \frac{uz}{1-z} + \frac{1}{2}y^2 + \frac{1}{2}B(z^2,u^2).$$

We treat $B(z^2,u^2)$ as a known function in z and u, because it is analytic in the region with $|z| < \sqrt{r}$ and $|u| < 1$, where $r \approx 0.3068$ is the constant that specifies the region of convergence for $B(z)$.

We now verify the conditions. For condition 1, we can write $F(z,y,u) = \sum_{s=0}^{\infty}\sum_{m=0}^{\infty} F_{s,m}(u)z^s y^m$ by setting $F_{0,2}(u) = \frac{1}{2}$, and for all $m \neq 2$, $F_{0,m}(u) = 0$. For $s \geq 1$, we set

$$F_{s,0}(u) = \begin{cases} u, & \text{odd } s \geq 1, \\ u + \frac{1}{2}\sum_{n=1}^{s/2} b_{s/2,n}u^{2n}, & \text{even } s \geq 2. \end{cases} \tag{27}$$

For $s \geq 1$ and $m \geq 1$, $F_{s,m}(u) = 0$. We see that $F(z,y,u)$ is analytic around $(0,0,0)$ as $\frac{1}{2}y^2$ is entire, $\frac{uz}{1-z}$ is analytic everywhere except at $z = 1$, and $B(z^2,u^2)$ is analytic for $|z| < \sqrt{r}$ and $|u| < 1$.

For condition 2, we have defined the coefficients $F_{s,m}(u)$ in verifying condition 1; all are real and non-negative in Eq. (27) at $u = 1$ because $b_{s,n} \geq 0$ for all s, n.

For condition 3, we note that $F(z,y,1) = \frac{z}{1-z} + \frac{1}{2}y^2 + \frac{1}{2}B(z^2) = H(z,y)$ and $F_y(z,y,1) = H_y(z,y)$ where $H(z,y)$ is the function defined in the proof of Theorem 5 (with variable y here in place of w previously). In that proof, we found the solution $(z,y) = (z_0,y_0) = (r,1)$ that satisfies $F(z,y,1) = y$ and $F_y(z,y,1) = 1$: $z_0 = r \approx 0.306760104888$ and $y_0 = 1$, where r is the radius of convergence of $B(z)$. We also have $F_z(z,y,u) = u/(1-z)^2 + \frac{1}{2}\frac{\partial}{\partial z}B(z^2,u^2)$, so that $F_z(z_0,y_0,1) = F_z(r,1,1) = 1/(1-r)^2 + rB'(r^2) \neq 0$, because all coefficients for $B(z)$ are non-negative so that $B'(r^2)$ is non-negative, and $1/(1-r)^2$ is positive. Finally, $F_{yy}(z_0,y_0,1) = F_{yy}(r,1,1) = 1 \neq 0$.

With the conditions of the theorem verified, we approximate μ and σ^2. Because $|u| < 1$ and $|z| < r < 1$ in the region where $B(z,u)$ converges, we can approximate $B(z^2,u^2)$ by its first terms:

$$\begin{aligned} B(z^2,u^2) \approx \frac{1}{2}\big[&(u^2)z^2 + (u^2+u^4)z^4 + (u^2+u^4+u^6)z^6 + (u^2+2u^4+2u^6+2u^8)z^8 \\ &+ (u^2+2u^4+4u^6+4u^8+3u^{10})z^{10} + (u^2+3u^4+6u^6+10u^8+9u^{10}+6u^{12})z^{12}\big], \end{aligned} \tag{28}$$

making use of terms $b_{s,n}$ that represent the coefficients of $u^n z^s$ from Table 2.

We approximate $B(z^2) \approx \sum_{i=1}^{14} b_i (z^2)^i = 1z^2 + 2z^4 + 3z^6 + 7z^8 + \cdots + 109419z^{28}$ with the first 14 terms as in the proof of Theorem 5. We then compute approximations to the derivatives:

$$\begin{aligned} F_z(z,y,u) &\approx \frac{u}{(1-z)^2} + \frac{1}{2}B'(z^2)2z, \\ F_{zz}(z,y,u) &\approx \frac{2u}{(1-z)^3} + \frac{1}{2}[B''(z^2)4z^2 + 2B'(z^2)], \\ F_u(z,y,u) &\approx \frac{z}{1-z} + \frac{1}{2}[(2u)z^2 + (2u+4u^3)z^4 + (2u+4u^3+6u^5)z^6 \\ &\quad + (2u+8u^3+12u^5+16u^7)z^8 + (2u+8u^3+24u^5+32u^7+30u^9)z^{10} \\ &\quad + (2u+12u^3+36u^5+80u^7+90u^9+72u^{11})z^{12}], \\ F_{uu}(z,y,u) &\approx \frac{1}{2}[2z^2 + (2+12u^2)z^4 + (2+12u^2+30u^4)z^6 + (2+24u^2+60u^4+112u^6)z^8 \\ &\quad + (2+24u^2+120u^4+224u^6+270u^8)z^{10} \\ &\quad + (2+36u^2+180u^4+560u^6+810u^8+792u^{10})z^{12}], \\ F_{zu}(z,y,u) &\approx \frac{1}{(1-z)^2} + \frac{1}{2}[(4u)z + (2u+4u^3)4z^3 + (2u+4u^3+6u^5)6z^5 \\ &\quad + (2u+8u^3+12u^5+16u^7)8z^7 + (2u+8u^3+24u^5+32u^7+30u^9)10z^9 \\ &\quad + (2u+12u^3+36u^5+80u^7+90u^9+72u^{11})12z^{11}], \\ F_{yz}(z,y,u) &= 0, \\ F_{yu}(z,y,u) &= 0, \\ F_{yy}(z,y,u) &= 1. \end{aligned}$$

We then approximate the derivatives of F at $z = z_0 = r$, $y = y_0 = 1$, and $u = 1$, where $B'(r^2)$ is approximated as in Eq. (8) in the proof of Theorem 5, and $B''(r^2) \approx \sum_{i=1}^{14} i(i-1)b_i(r^2)^{i-2}$.

We obtain $F_z(r,1,1) \approx 2.5370$, $F_{zz}(r,1,1) \approx 8.7686$, $F_u(r,1,1) \approx 0.5701$, $F_{uu}(r,1,1) \approx 0.1854$, and $F_{zu}(r,1,1) \approx 3.1929$. Then $\mu = F_u(r,1,1)/[rF_z(r,1,1)] \approx 0.7326$, and simplifying Eq. (24) with $F_{yz} = 0$, $F_{yu} = 0$, and $F_{yy} = 1$, we obtain $\sigma^2 = \mu + \mu^2 + (F_z^2F_{uu} - 2F_zF_uF_{zu} + F_u^2F_{zz})/(z_0F_z^3) \approx 0.2325$. □

We note a minor technicality. The statement of Theorem 2.23 of [5] includes an additional condition, namely $F(0, y, u) \equiv 0$, which does not hold in our scenario, as $F(0, y, u) = \frac{1}{2}y^2 + B(0, u^2) = \frac{1}{2}y^2 \not\equiv 0$. However, this condition is used only to guarantee the existence of a solution $y = y(z, u)$ to $y = F(z, y, u)$ with non-negative Taylor coefficients (see the beginning of the proof of Remark 2.20 of [5]). In our setting, this existence is guaranteed, as we defined $F(z, y, u)$ to be the implicit generating function $B(z, u) = F\big(z, B(z, u), u\big)$, where $B(z, u) = \sum_{s=1}^{\infty}\sum_{n=1}^{\infty} b_{s,n} z^s u^n$ and the coefficients $b_{s,n}$ are counts that are necessarily nonnegative for all s, n.

Theorem 2.23 of [5] also states the condition $F(z, 0, u) \not\equiv 0$. It is not clear where this condition is required for obtaining the conclusions of the theorem, but in any case, in our situation, it is straightforward to verify, as $F(z, 0, u) = \frac{uz}{1-z} + \frac{1}{2}B^2(z^2, u^2) \not\equiv 0$.

Computing from Table 2 the sequence of values μ_s/s describing the mean number of leaves in perfect phylogenies of size s, we obtain 1, $\frac{3}{4}$, $\frac{2}{3}$, $\frac{19}{28}$, $\frac{24}{35}$, $\frac{73}{105}$, $\frac{419}{595}$, $\frac{641}{904}$, $\frac{107}{150}$, $\frac{11869}{16580}$, $\frac{18253}{25421}$, $\frac{113467}{157692}$, and $\frac{353277}{490087}$ for $s = 1, 2, \ldots, 12, 13$. The numerical value at $s = 13$ is approximately 0.7208, close to the limiting value of approximately 0.7326. The sequence of approximate values of σ_s^2/s gives 0, 0.125, 0.222, 0.265, 0.278, 0.281, 0.268, 0.265, 0.257, 0.254, 0.250, 0.248, and 0.246 for $s = 1, 2, \ldots, 12, 13$, nearing the limit of approximately 0.2325.

5. Rooted binary perfect phylogenies with a caterpillar shape

We have counted perfect phylogenies with sample size s, and with sample size s and number of leaves n. Each perfect phylogeny has an associated unlabeled tree shape; we now count the perfect phylogenies with sample size s and a caterpillar shape (with n leaves).

5.1. Enumeration

A caterpillar tree with $n \geq 2$ leaves has exactly 1 cherry node. In other words, for $n \geq 3$, a caterpillar tree is constructed by adjoining a caterpillar tree with $n-1$ leaves and a single-leaf tree to a shared root. Denote by $g_{s,n}$ the number of caterpillar rooted binary phylogenies with sample size s and $n \geq 2$ leaves. We set $g_{s,n} = 0$ if $s < n$ (or $s \notin \mathbb{N}$, or $n \notin \mathbb{N}$).

Proposition 14. *The number $g_{s,n}$ of rooted binary perfect phylogenies with caterpillar shape, sample size s, and n leaves, $2 \leq n \leq s$, satisfies*
(i) $g_{s,2} = \lfloor \frac{s}{2} \rfloor$ for all $s \geq 2$.
(ii) For (s, n) with $s \geq n \geq 3$,

$$g_{s,n} = \sum_{i=n-1}^{s-1} g_{i,n-1} = \sum_{i_1=n-1}^{s-1} \sum_{i_2=n-2}^{i_1-1} \cdots \sum_{i_{n-3}=3}^{i_{n-4}-1} \sum_{i_{n-2}=2}^{i_{n-3}-1} \left\lfloor \frac{i_{n-2}}{2} \right\rfloor. \tag{29}$$

Proof. (i) Recognizing that the only tree shape with $n = 2$ leaves is the 2-leaf caterpillar tree, we see that we already proved this result in Corollary 7.

(ii) For $n \geq 3$, the left subtree of a caterpillar of size n is a caterpillar of size $n-1$. We assign sample size i to the left subtree, $n-1 \leq i \leq s-1$, and $s-i$ to the leaf in the right subtree:

$$g_{s,n} = \sum_{i=n-1}^{s-1} g_{i,n-1} b_{s-i,1} = \sum_{i=n-1}^{s-1} g_{i,n-1}. \tag{30}$$

Proceeding iteratively, we have

$$g_{s,n} = \sum_{i_1=n-1}^{s-1} g_{i_1,n-1} = \sum_{i_1=n-1}^{s-1} \sum_{i_2=n-2}^{i_1-1} g_{i_2,n-2} = \cdots = \sum_{i_1=n-1}^{s-1} \sum_{i_2=n-2}^{i_1-1} \cdots \sum_{i_{n-3}=3}^{i_{n-4}-1} \sum_{i_{n-2}=2}^{i_{n-3}-1} g_{i_{n-2},2}.$$

We apply the base case of $n = 2$ to complete the proof. □

We can apply Proposition 14 with specific small values of n, completing the sum in Eq. (29). The case of $n = 3$ was obtained in Corollary 8, and we will write its solution in a different form. We proceed via calculations similar to those performed in obtaining Corollaries 7–9.

We use an approach that avoids summations that include floor and ceiling functions, as appeared in the proofs of Corollaries 8 and 9. Separating the $n = 2$ result $g_{s,2} = \lfloor \frac{s}{2} \rfloor$ (Corollary 7) into cases for odd and even s, we can increase n incrementally, observing from Eq. (30) that for fixed n, $g_{s,n}$ as a function of s can be written with odd and even cases, each consisting of a polynomial of degree $n-1$ in s. It is convenient to instead define the cases in terms of odd and even $s-n$. In particular, for $s-n$ even, we define $f^e_{s,n}$ to be the polynomial describing the number of caterpillar perfect phylogenies with sample size s and n leaves. For $s-n$ odd, we define $f^o_{s,n}$ as the corresponding polynomial for the number of caterpillar perfect phylogenies with sample size s and n leaves. Note that both polynomials are functions that can be calculated for all (s, n) with $s \geq n \geq 2$; however, each represents the number of caterpillar perfect phylogenies only in its associated case. With these definitions, the number of caterpillar perfect phylogenies $g_{s,n}$ can be written in a form that is convenient for computation, containing only a single floor function.

Proposition 15. *For (s, n) with $n \geq 2$ and $s \geq n$, the number of caterpillar perfect phylogenies with n leaves and sample size s is*

$$g_{s,n} = \lfloor f^e_{s,n} \rfloor = \left\lfloor \left(\sum_{i=n-1}^{s-1} f^e_{i,n-1} \right) - \frac{s-n}{2^{n-1}} \right\rfloor. \tag{31}$$

The proposition relies on a lemma.

Lemma 16. *For (s, n) with $n \geq 2$ and $s \geq n$,*

$$f^e_{s,n} - f^o_{s,n} = \frac{1}{2^{n-1}}. \tag{32}$$

Proof. We proceed by induction on n. For the base case, $n = 2$, by Corollary 7, we have $g_{s,2} = \lfloor \frac{s}{2} \rfloor$. Thus for s even, $g_{s,2} = \frac{s}{2}$, and for s odd, $g_{s,2} = \frac{s}{2} - \frac{1}{2}$. In other words, we have $f^e_{s,2} = \frac{s}{2}$ and $f^o_{s,2} = \frac{s-1}{2}$. It follows that $f^e_{s,2} - f^o_{s,2} = \frac{1}{2}$.

For the inductive step, suppose for $n \geq 3$ that $f^e_{s,n-1} - f^o_{s,n-1} = 1/2^{n-2}$. If $s-(n-1)$ is even, then $f^e_{s,n-1}$ is an integer, with $g_{s,n-1} = f^e_{s,n-1} = \lfloor f^e_{s,n-1} \rfloor$. If instead $s-(n-1)$ is odd, then $f^o_{s,n-1} = g_{s,n-1}$ is an integer, and by the inductive hypothesis, $\lfloor f^e_{s,n-1} \rfloor = f^e_{s,n-1} - 1/2^{n-2} = f^o_{s,n-1}$.

By Eq. (30) and the inductive hypothesis, using $\mathbb{1}_{\{x\}} = 1$ if x holds and $\mathbb{1}_{\{x\}} = 0$ otherwise,

$$\begin{aligned} g_{s,n} &= \sum_{i=n-1}^{s-1} g_{i,n-1} = \sum_{i=n-1}^{s-1} \lfloor f^e_{i,n-1} \rfloor \\ &= \left(\sum_{i=n-1}^{s-1} f^e_{i,n-1} \right) - \frac{1}{2^{n-2}} \sum_{i=n-1}^{s-1} \mathbb{1}_{\{i-(n-1) \text{ is odd}\}}. \end{aligned}$$

We then use that

$$\sum_{i=n-1}^{s-1} \mathbb{1}_{\{i-(n-1) \text{ is odd}\}} = \sum_{i=0}^{s-n} \mathbb{1}_{\{i \text{ is odd}\}} = \begin{cases} \frac{s-n}{2}, & s-n \text{ even}, \\ \frac{s-(n-1)}{2}, & s-n \text{ odd}. \end{cases}$$

We then obtain expressions for $g_{s,n}$ in the case of even $s-n$ and odd $s-n$. Because $g_{s,n} = f^e_{s,n}$ for even $s-n$ and $g_{s,n} = f^o_{s,n}$ for odd $s-n$, we have

$$f^e_{s,n} = \left(\sum_{i=n-1}^{s-1} f^e_{i,n-1} \right) - \frac{s-n}{2^{n-1}}, \tag{33}$$

$$f^o_{s,n} = \left(\sum_{i=n-1}^{s-1} f^e_{i,n-1} \right) - \frac{s-(n-1)}{2^{n-1}}. \tag{34}$$

Now we see that $f^e_{s,n} - f^o_{s,n} = 1/2^{n-1}$, completing the induction. □

Proof of Proposition 15. From Lemma 16, for each $n \geq 2$, $f^e_{s,n}$ exceeds $f^o_{s,n}$ by a quantity that is less than 1. Hence, for odd $s-n$, for which $g_{s,n} = f^o_{s,n}$ and $f^o_{s,n}$ is an integer, $\lfloor f^e_{s,n} \rfloor = f^o_{s,n} = g_{s,n}$. For even $s-n$, $g_{s,n} = f^e_{s,n}$ and $f^e_{s,n}$ is an integer, so that $\lfloor f^e_{s,n} \rfloor = f^e_{s,n} = g_{s,n}$. We conclude in both odd and even cases that $g_{s,n} = \lfloor f^e_{s,n} \rfloor$, with $f^e_{s,n}$ specified by Eq. (33). □

We can then compute $g_{s,n}$ for the smallest n by iteratively summing polynomials to calculate $f^e_{s,n}$ in Eq. (33), taking the floor of the output. We present the first several functions $g_{s,n}$.

$$g_{s,2} = \left\lfloor \frac{s}{2} \right\rfloor \tag{35}$$

$$g_{s,3} = \left\lfloor \frac{(s-1)^2}{4} \right\rfloor \tag{36}$$

$$g_{s,4} = \left\lfloor \frac{s(s-2)(2s-5)}{24} \right\rfloor \tag{37}$$

$$g_{s,5} = \left\lfloor \frac{(s-1)(s-3)(s^2-4s+1)}{48} \right\rfloor \tag{38}$$

$$g_{s,6} = \left\lfloor \frac{s(s-2)(s-4)(2s^2-13s+16)}{480} \right\rfloor \tag{39}$$

Table 3
The number $g_{s,n}$ of rooted binary perfect phylogenies with sample size s and a caterpillar shape with n leaves. Entries are obtained using Proposition 14; the "total" is $g_s = \sum_{n=2}^{s} g_{s,n}$. The total follows A000975 (with the index shifted so that term s in A000975 is g_{s+1}).

Sample size (s)	Number of leaves (n)												Total
	13	12	11	10	9	8	7	6	5	4	3	2	
2												1	1
3											1	1	2
4										1	2	2	5
5									1	3	4	2	10
6								1	4	7	6	3	21
7							1	5	11	13	9	3	42
8						1	6	16	24	22	12	4	85
9					1	7	22	40	46	34	16	4	170
10				1	8	29	62	86	80	50	20	5	341
11			1	9	37	91	148	166	130	70	25	5	682
12		1	10	46	128	239	314	296	200	95	30	6	1365
13	1	11	56	174	367	553	610	496	295	125	36	6	2730

$$g_{s,7} = \left\lfloor \frac{(s-1)(s-3)^2(s-5)(2s^2-12s+1)}{2880} \right\rfloor \tag{40}$$

$$g_{s,8} = \left\lfloor \frac{s(s-2)(s-4)(s-6)(4s^3-50s^2+176s-151)}{40320} \right\rfloor \tag{41}$$

$$g_{s,9} = \left\lfloor \frac{(s-1)(s-3)(s-5)(s-7)(s^4-16s^3+78s^2-112s+3)}{80640} \right\rfloor. \tag{42}$$

The values of $g_{s,n}$ for $2 \le n \le s \le 13$ appear in Table 3.

5.2. Generating function

We next obtain a generating function for the number of perfect phylogenies with the fixed caterpillar shape with n leaves, as s increases.

Proposition 17. *The generating function $G_n(z)$ for the number $g_{s,n}$ of rooted binary perfect phylogenies with sample size $s \ge n$ and the caterpillar topology with $n \ge 2$ leaves satisfies*

$$G_n(z) = \frac{z^n}{(1-z)^n(1+z)}. \tag{43}$$

Proof. We proceed by induction. We obtained the result for $n = 2$ in Eq. (18), as the caterpillar is the only shape with 2 leaves. Suppose the generating function for the number of rooted binary perfect phylogenies with sample size s and the n-leaf caterpillar follows Eq. (43). We apply Eq. (30) to obtain the generating function associated with the caterpillar with $n+1$ leaves (propagating $b_{s-i,1} = 1$ through the calculation for clarity):

$$\begin{aligned} G_{n+1}(z) &= \sum_{s=1}^{\infty} g_{s,n+1} z^s \\ &= \sum_{s=1}^{\infty} \left(\sum_{i=n}^{s-1} g_{i,n} b_{s-i,1} \right) z^s \\ &= \sum_{s=1}^{\infty} \sum_{i=n}^{s-1} (g_{i,n} z^i)(b_{s-i,1} z^{s-i}). \end{aligned}$$

Because $g_{s,n} = 0$ for $1 \le s \le n-1$, we add additional zeros and simplify a convolution:

$$\begin{aligned} G_{n+1}(z) &= \sum_{s=1}^{\infty} \sum_{i=1}^{s-1} (g_{i,n} z^i)(b_{s-i,1} z^{s-i}) \\ &= G_n(z) B_1(z) \\ &= \frac{z^n}{(1-z)^n(1+z)} \frac{z}{1-z}. \end{aligned}$$

The induction is complete. □

Note that g_s, the total number of caterpillar perfect phylogenies with sample size s, allowing all possible values of n, $2 \leq n \leq s$, follows the Lichtenberg sequence, OEIS sequence A000975 (the index is shifted, so that if A000975 is denoted $\{a_s\}$, then $a_s = g_{s+1}$). We verify this equivalence by showing an identity of generating functions. Denote the generating function for the number of caterpillars with sample size s, considering all possible numbers of leaves, by $G(z) = \sum_{s=2}^{\infty} g_s z^s$.

We have that

$$g_s = \sum_{n=2}^{s} g_{s,n} \tag{44}$$

and $G_n(z) = \sum_{s=n}^{\infty} g_{s,n} z^s$, and we add zeros to obtain

$$G(z) = \sum_{s=2}^{\infty} g_s z^s = \sum_{s=2}^{\infty}\left(\sum_{n=2}^{\infty} g_{s,n}\right) z^s = \sum_{n=2}^{\infty}\left(\sum_{s=2}^{\infty} g_{s,n} z^s\right) = \sum_{n=2}^{\infty} G_n(z).$$

By Proposition 17, we then have

$$\begin{aligned} G(z) &= \sum_{n=2}^{\infty} \frac{z^n}{(1-z)^n(1+z)} \\ &= \frac{1}{1+z}\left(\frac{1}{1-\frac{z}{1-z}} - \frac{z}{1-z} - 1\right) \\ &= \frac{z^2}{(1+z)(1-z)(1-2z)}, \end{aligned} \tag{45}$$

where the summation requires $|z| < \frac{1}{2}$. The Lichtenberg sequence [12,13] has generating function $z/[(1+z)(1-z)(1-2z)]$, differing only in missing a factor of z, so that its term a_s accords with our g_{s+1}. Using the exact form for the Lichtenberg sequence [12], we have

$$g_s = \left\lfloor \frac{2^s}{3} \right\rfloor. \tag{46}$$

Note that if we were to consider the 1-leaf perfect phylogeny a caterpillar and to allow a trivial perfect phylogeny with $s = 0$, then we would obtain a sequence $\{g'_s\}$ for the total number of perfect phylogenies with sample size s and $n \geq 1$ leaves; for all $s \geq 0$, $g'_s = g_s + 1$. This sequence, with generating function $G(z) + \frac{1}{1-z}$ to account for the extra perfect phylogeny with 1 leaf (for all $s \geq 1$) and the trivial perfect phylogeny ($s = 0$), accords with A005578, which has generating function $(1 - z - z^2)/[(1+z)(1-z)(1-2z)] = G(z) + \frac{1}{1-z}$.

5.3. Asymptotics

We study the asymptotics for $g_{s,n}$, the number of rooted binary caterpillar perfect phylogenies with sample size s and $n \geq 2$ leaves, similarly to our analysis of general perfect phylogenies.

We quickly obtain a result analogous to Lemma 12 directly from the closed form $G_n(z) = \sum_{s=1}^{\infty} g_{s,n} z^s = \sum_{s=n}^{\infty} g_{s,n} z^s = G_n(z) = z^n/[(1-z)^n(1+z)]$ for $n \geq 2$ (Eq. (43)).

Lemma 18. *Fix $n \geq 2$. $G_n(z) \sim h_n/(1-z)^n$ as $z \to 1$ in a Δ-domain in the neighborhood of $z = 1$, for a constant h_n.*

The constant is $h_n = \frac{1}{2}$ for all $n \geq 2$. Applying Corollary 2.16 of [5], we obtain a result similar to Proposition 11.

Proposition 19. *As $s \to \infty$, for fixed $n \geq 2$, the number $g_{s,n}$ of rooted binary perfect phylogenies with sample size $s \geq n$ and the caterpillar topology has asymptotic growth $g_{s,n} \sim s^{n-1}/[2(n-1)!]$.*

5.4. Bivariate generating function

The bivariate generating function for the number of rooted binary perfect phylogenies with a caterpillar topology, with sample size s and $n \geq 2$ leaves, also follows from the closed form $G_n(z)$ (Eq. (43)). Let $G(z,u) = \sum_{n=2}^{\infty} G_n(z)\, u^n$. Then

$$\begin{aligned} G(z,u) &= \frac{1}{1+z}\sum_{n=2}^{\infty} \frac{u^n z^n}{(1-z)^n} \\ &= -\frac{1}{1+z} - \frac{uz}{(1-z)(1+z)} + \frac{1}{1+z}\sum_{n=0}^{\infty}\left(\frac{uz}{1-z}\right)^n \\ &= \frac{u^2 z^2}{(1-z)(1+z)(1-z-uz)}. \end{aligned} \tag{47}$$

5.5. Asymptotic normal distribution for the number of leaves

We obtain an asymptotic normal distribution for the number of leaves. The result is obtained from the bivariate generating function.

Theorem 20. *Let Y_s denote the random variable describing the number of leaves of a caterpillar perfect phylogeny chosen at random with sample size s. Then as s grows large,*
(i) $\mathbb{E}[Y_s] \sim \frac{1}{2}s$.
(ii) $\mathrm{Var}[Y_s] \sim \frac{1}{4}s$.
(iii) The distribution of Y_s has a limiting normal distribution

$$\frac{Y_s - \mathbb{E}[Y_s]}{\sqrt{\mathrm{Var}[Y_s]}} \to N(0, 1).$$

Proof. We refer to Theorem IX.9 in [8] (see also the errata, https://ac.cs.princeton.edu/errata/). Consider a function $F(z, u)$ that is bivariate analytic at $(0, 0)$ and whose expansion has non-negative coefficients. Suppose $F(z, 1)$ is meromorphic in $|z| \le R$ with a pole at $z = \rho$ for a positive $\rho < R$. Suppose that the following also hold:

1. For some $\epsilon > 0$ and $r > \rho$ (and $r \le R$), we can write $F(z, u) = B(z, u)/C(z, u)$ for (z, u) in some domain $D = \{|z| \le r\} \times \{|u - 1| < \epsilon\}$, where $B(z, u)$ and $C(z, u)$ are analytic in D with $B(\rho, 1) \neq 0$.
2. The partial derivatives of C satisfy
$$\frac{\partial C(\rho, 1)}{\partial z} \frac{\partial C(\rho, 1)}{\partial u} \neq 0.$$
3. $\mathbf{v}\left(\frac{\rho(1)}{\rho(u)}\right) \neq 0$, where $\rho(u)$ is the solution to $C\big(\rho(u), u\big) = 0$ and $\rho(1) = \rho$, $\mathbf{m}\big(f(u)\big) = \frac{f'(1)}{f(1)}$, and $\mathbf{v}\big(f(u)\big) = \frac{f''(1)}{f(1)} + \frac{f'(1)}{f(1)} - \left(\frac{f'(1)}{f(1)}\right)^2$ for a function f analytic at 1 with $f(1) \neq 0$.

Then the random variable Y_s with probability generating function $p_s(u) = [z^s]F(z, u)/[z^s]F(z, 1)$, standardized to $(Y_s - \mu_s)/\sigma_s$, converges in distribution to a standard normal random variable with mean 0 and variance 1, where $\mu_s = \mathbf{m}\left(\frac{\rho(1)}{\rho(u)}\right)s + O(1)$ and $\sigma_s^2 = \mathbf{v}\left(\frac{\rho(1)}{\rho(u)}\right)s + O(1)$.

We verify the hypotheses of the theorem for the bivariate generating function $G(z, u)$. First, $G(z, u)$ is analytic in both variables at $(0, 0)$, and its expansion has non-negative coefficients for all (s, n) with $s \ge n \ge 2$ (Eq. (47)).

Next, for a choice of r with $\frac{1}{2} < r < 1$, $G(z, 1)$ is meromorphic for $|z| \le r$, with a pole only at $z = \rho$ for $\rho = \frac{1}{2} < r$. We verify the conditions of the theorem.

1. Write $G(z, u) = B(z, u)/C(z, u)$ in $D = \{|z| \le r\} \times \{|u - 1| < \epsilon\}$ for $B(z, u) = u^2z^2$ and $C(z, u) = (1 - z)(1 + z)(1 - z - uz)$, for small $\epsilon > 0$ and $\frac{1}{2} < r < 1$. Both B and C are analytic with $B(\frac{1}{2}, 1) = \frac{1}{4} \neq 0$.
2. The condition on the partial derivatives is satisfied, with $\rho = \frac{1}{2}$:
$$\frac{\partial C(\rho, 1)}{\partial z} \frac{\partial C(\rho, 1)}{\partial u} = [2(3\rho^2 - \rho - 1)][\rho(\rho - 1)(\rho + 1)] = \frac{9}{16} \neq 0.$$
3. Given u, the location of pole ρ as a function of u is $z = \rho(u) = \frac{1}{1+u}$, so that $\rho(1)/\rho(u) = \frac{1+u}{2}$. Letting $f(u) = \frac{1+u}{2}$, $f(u)$ is analytic at $u = 1$ with $f(1) = 1 \neq 0$. We have $\mathbf{m}\big(f(u)\big) = \frac{1}{2}/1 = \frac{1}{2}$ and $\mathbf{v}\big(f(u)\big) = \frac{0}{1} + \frac{1/2}{1} - (\frac{1/2}{1})^2 = \frac{1}{4} \neq 0$.

We conclude that for the random variable Y_s describing the random number of leaves of a caterpillar perfect phylogeny with sample size s, with probability generating function $p_s(u) = [z^s]G(z, u)/[z^s]G(z, 1)$, $(Y_s - \mu_s)/\sigma_s$ converges in distribution to a standard normal random variable, with $\mu_s = \mathbf{m}\big(f(u)\big)s + O(1) = \frac{1}{2}s + O(1)$ and $\sigma_s^2 = \mathbf{v}\big(f(u)\big)s + O(1) = \frac{1}{4}s + O(1)$. □

Computing from Table 3 the sequence of values μ_s/s describing the mean number of leaves in caterpillars of size s, we obtain $\frac{1}{2}$, $\frac{5}{6}$, $\frac{7}{10}$, $\frac{33}{50}$, $\frac{13}{21}$, $\frac{59}{98}$, $\frac{199}{340}$, $\frac{881}{1530}$, $\frac{967}{1705}$, $\frac{4209}{7502}$, $\frac{1517}{2730}$, and $\frac{6523}{11830}$ for $s = 2$ to 13; the last of these values is approximately 0.551, near the limiting value of $\frac{1}{2}$. Numerical approximations for the corresponding sequence of values σ_s^2/s are 0, 0.083, 0.140, 0.162, 0.193, 0.201, 0.215, 0.219, 0.226, 0.228, 0.231, and 0.232, approaching the limiting value of $\frac{1}{4}$.

6. Rooted binary perfect phylogenies with an arbitrary unlabeled shape

For our last analysis, we generalize the argument we have used for recursively counting perfect phylogenies with a caterpillar tree shape with n leaves (Section 5.1) to an arbitrary tree shape with n leaves, offering some results in the general case of an arbitrary unlabeled tree shape.

Let T be an unlabeled tree shape with $|T|$ leaves. Tree T has left and right subtrees, T_ℓ and T_r, with $|T_\ell|$ and $|T_r|$ leaves. In sequentially decomposing a tree into its left and right subtrees, eventually a single node is reached. Denote by $N_{s,T}$ the number of rooted binary perfect phylogenies with unlabeled tree shape T, where $N_{s,T} = 0$ if $s < |T|$ or s is not an integer.

Proposition 21. *The number $N_{s,T}$ of rooted binary perfect phylogenies with unlabeled tree shape T and sample size $s \geq |T|$ satisfies*
(i) $N_{s,T} = 1$ if T has a single leaf and $s \geq 1$.
(ii) For (s, T) with $s \geq |T| \geq 2$,

$$N_{s,T} = \begin{cases} \sum_{i=|T_\ell|}^{s-|T_r|} N_{i,T_\ell} N_{s-i,T_r}, & T_\ell \neq T_r, \\ \left(\sum_{i=|T_\ell|}^{s-|T_r|} \frac{1}{2} N_{i,T_\ell} N_{s-i,T_r} \right) + \frac{1}{2} N_{s/2,T_\ell}, & T_\ell = T_r. \end{cases} \tag{48}$$

Proof. (i) We have discussed the base case of the single-leaf tree in Proposition 6i. (ii) For the general case, a perfect phylogeny with unlabeled shape T is constructed from perfect phylogenies with unlabeled shapes T_ℓ and T_r. The minimal sample size assigned to T_ℓ is $|T_\ell|$, and the minimal sample size assigned to T_r is $|T_r|$, so that the maximal sample size for T_ℓ is $s - |T_r|$.

If $T_\ell \neq T_r$, then we sum the product of the number of perfect phylogenies for T_ℓ and the number of perfect phylogenies for T_r over all possible values i of the sample size assigned to T_ℓ. Because $i \geq |T_\ell|$ and $s - i \geq |T_r|$, we have $i \leq s - |T_r|$.

If $T_\ell = T_r$ and sample size $i \neq \frac{s}{2}$ is assigned to the left subtree, then a factor of $\frac{1}{2}$ accounts for the fact that each perfect phylogeny traversed is also obtained for sample size $s - i$ assigned to the left subtree. If $T_\ell = T_r$ and $i = \frac{s}{2}$ (for even s), then we count the $\binom{N_{s/2,T_\ell}}{2}$ trees with distinct subtrees and the $N_{s/2,T_\ell}$ trees with identical subtrees. □

Using Proposition 21, we compute $N_{s,T}$ for each unlabeled tree shape with $|T| \leq 8$ and $|T| \leq s \leq 11$, presenting these counts in Tables 4 and 5. For each small value of n, at fixed $s \geq n$, we observe that across shapes T with n leaves, the number of rooted binary perfect phylogenies $N_{s,T}$ tends to be larger for less balanced shapes T and smaller for more balanced shapes.

The proposition can be used to obtain a closed form for the number of rooted binary perfect phylogenies for a fixed shape T as a function of s. For example, suppose T is the 4-leaf symmetric unlabeled shape, with T_ℓ and T_r both corresponding to the 2-leaf caterpillar. Proposition 21 yields, for $s \geq 4$,

$$\begin{aligned} N_{s,T} &= \left(\sum_{i=2}^{s-2} \frac{1}{2} g_{i,2} g_{s-i,2} \right) + \frac{1}{2} g_{s/2,2} \\ &= \left(\sum_{i=2}^{s-2} \frac{1}{2} \left\lfloor \frac{i}{2} \right\rfloor \left\lfloor \frac{s-i}{2} \right\rfloor \right) + \frac{1}{2} \left\lfloor \frac{s}{4} \right\rfloor \mathbb{1}_{\{s \text{ is even}\}} \\ &= \begin{cases} \frac{(s+1)(s-1)(s-3)}{48}, & \text{odd } s \geq 5, \\ \frac{s(s-1)(s-2)}{48} + \frac{1}{2} \lfloor \frac{s}{4} \rfloor, & \text{even } s \geq 4. \end{cases} \end{aligned} \tag{49}$$

Note that the derivation follows the proof of Corollary 9, Eqs. (14) and (16).

Recall that there are only two 4-leaf unlabeled topologies, the symmetric shape and the caterpillar. Adding Eq. (49), counting perfect phylogenies for the symmetric shape, and Eq. (37), for the caterpillar, we obtain Eq. (12), counting all perfect phylogenies with $n = 4$ leaves. In particular, for odd $s \geq 5$, using Lemma 16 and Proposition 15 to remove the floor function,

$$\begin{aligned} \frac{(s+1)(s-1)(s-3)}{48} + \left\lfloor \frac{s(s-2)(2s-5)}{24} \right\rfloor &= \frac{(s+1)(s-1)(s-3)}{48} + \frac{s(s-2)(2s-5)}{24} - \frac{1}{8} \\ &= \frac{(s-1)(s-3)(5s-1)}{48}. \end{aligned}$$

For even $s \geq 4$, by Proposition 15,

$$\begin{aligned} \frac{s(s-1)(s-2)}{48} + \frac{1}{2} \left\lfloor \frac{s}{4} \right\rfloor + \left\lfloor \frac{s(s-2)(2s-5)}{24} \right\rfloor &= \frac{s(s-1)(s-2)}{48} + \frac{1}{2} \left\lfloor \frac{s}{4} \right\rfloor + \frac{s(s-2)(2s-5)}{24} \\ &= \frac{s(s-2)(5s-11)}{48} + \frac{1}{2} \left\lfloor \frac{s}{4} \right\rfloor. \end{aligned}$$

Subtracting Eq. (49), the number of perfect phylogenies for the symmetric shape, from Eq. (37), the corresponding number for the caterpillar, we can quickly demonstrate that for $s \geq 5$, the caterpillar possesses more perfect phylogenies with sample size s. In particular, for odd $s \geq 5$,

$$\frac{s(s-2)(2s-5)}{24} - \frac{1}{8} - \frac{(s+1)(s-1)(s-3)}{48} = \frac{(s-1)^2(s-3)}{16} > 0.$$

Table 4
The number of rooted binary perfect phylogenies $N(T, s)$ that have sample size s and a given unlabeled topology T with n leaves, for small n ($1 \leq n \leq 7$) and s. $N(T, s)$ is calculated according to Proposition 21.

Number of leaves (n)	Topology T	Sample size s										
		1	2	3	4	5	6	7	8	9	10	11
1		1	1	1	1	1	1	1	1	1	1	1
2		-	1	1	2	2	3	3	4	4	5	5
3		-	-	1	2	4	6	9	12	16	20	25
4		-	-	-	1	3	7	13	22	34	50	70
4		-	-	-	1	1	3	4	8	10	16	20
5		-	-	-	-	1	4	11	24	46	80	130
5		-	-	-	-	1	2	5	9	17	27	43
5		-	-	-	-	1	3	8	16	30	50	80
6		-	-	-	-	-	1	5	16	40	86	166
6		-	-	-	-	-	1	3	8	17	34	61
6		-	-	-	-	-	1	4	12	28	58	108
6		-	-	-	-	-	1	4	12	28	58	108
6		-	-	-	-	-	1	2	6	11	23	38
6		-	-	-	-	-	1	2	7	14	31	54
7		-	-	-	-	-	-	1	6	22	62	148
7		-	-	-	-	-	-	1	4	12	29	63
7		-	-	-	-	-	-	1	5	17	45	103
7		-	-	-	-	-	-	1	5	17	45	103
7		-	-	-	-	-	-	1	3	9	20	43
7		-	-	-	-	-	-	1	3	10	24	55
7		-	-	-	-	-	-	1	5	17	45	103
7		-	-	-	-	-	-	1	3	9	20	43
7		-	-	-	-	-	-	1	4	13	32	71
7		-	-	-	-	-	-	1	5	17	45	103
7		-	-	-	-	-	-	1	3	9	20	43

For even $s \geq 4$,

$$\begin{aligned}\frac{s(s-2)(2s-5)}{24} - \frac{s(s-1)(s-2)}{48} - \frac{1}{2}\left\lfloor \frac{s}{4} \right\rfloor &= \frac{s(s-2)(s-3)}{16} - \frac{1}{2}\left\lfloor \frac{s}{4} \right\rfloor \\ &\geq \frac{s(s-2)(s-3)}{16} - \frac{s}{8} \\ &= \frac{s(s-1)(s-4)}{16} \geq 0,\end{aligned}$$

with equality if and only if $s = 4$.

7. Discussion

We have studied the enumerative combinatorics of rooted binary perfect phylogenies. We have provided a recursive formula to enumerate the rooted binary perfect phylogenies with a given sample size s via Eq. (3), and we have provided an asymptotic approximation in Eq. (6). We have also refined the enumeration, counting rooted binary perfect phylogenies for a given sample size s separately for each possible value of the number of leaves n via Eq. (9). We have counted rooted binary perfect phylogenies associated with specific shapes (Eq. (48)), notably a caterpillar shape (Eq. (29)). A summary of results appears in Table 6.

Table 5
The number of rooted binary perfect phylogenies $N(T, s)$ that have sample size s and a given unlabeled topology T with n leaves, for small n ($n = 8$) and s. $N(T, s)$ is calculated according to Proposition 21.

Number of leaves (n)	Topology T	Sample size s										
		1	2	3	4	5	6	7	8	9	10	11
8		-	-	-	-	-	-	-	1	7	29	91
8		-	-	-	-	-	-	-	1	5	17	46
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	4	14	38
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	5	18	50
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	5	18	50
8		-	-	-	-	-	-	-	1	5	18	50
8		-	-	-	-	-	-	-	1	3	10	23
8		-	-	-	-	-	-	-	1	3	11	27
8		-	-	-	-	-	-	-	1	6	23	68
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	5	18	50
8		-	-	-	-	-	-	-	1	3	13	34
8		-	-	-	-	-	-	-	1	4	13	33
8		-	-	-	-	-	-	-	1	1	4	7

Table 6
The main results of the paper. We have variously obtained recursions, generating functions, and asymptotics for the number of rooted binary perfect phylogenies with sample size s: considering all tree shapes, all tree shapes with n leaves, the n-leaf caterpillar shape, all caterpillar shapes, and a single shape that is specified, but that is arbitrary.

Tree shapes	Recursion	Generating function	Asymptotics
All shapes	b_s, Proposition 1	$B(z)$, Proposition 2	Theorem 5
All n-leaf shapes	$b_{s,n}$, Proposition 6	$B_n(z)$, Proposition 10	Proposition 11
n-leaf caterpillar	$g_{s,n}$, Proposition 14	$G_n(z)$, Proposition 17	Proposition 19
All caterpillars	g_s, Eq. (44)	$G(z)$, Eq. (45)	Eq. (46)
Arbitrary shape	$N_{s,T}$, Proposition 21	–	–

The enumerations build on the efforts of Palacios et al. [16] to enumerate the labeled and unlabeled topologies and labeled and unlabeled histories that can be associated with a rooted perfect phylogeny, binary or multifurcating. For rooted binary perfect phylogenies, we provide enumerations that can be employed as a starting point for the enumerations of labeled and unlabeled topologies and labeled and unlabeled histories by Palacios et al. [16].

The recurrence for b_s, the number of rooted binary perfect phylogenies with sample size s (Eq. (3)), is similar to the recurrence for the number of rooted binary unlabeled trees (Eq. (1))—except that it requires the addition of a 1 for the single-leaf perfect phylogeny, whereas the recurrence for the rooted binary unlabeled trees does not include a corresponding possibility. This small difference leads to a large difference in asymptotic growth. Whereas the asymptotic growth of the rooted binary unlabeled trees—the perfect phylogenies with sample size s and s leaves—is approximately $0.3188(2.4833)^s s^{-3/2}$, the growth of the rooted binary perfect phylogenies with sample size s is substantially larger, approximately $0.3519(3.2599)^s s^{-3/2}$ (Eq. (6)).

Some of our results produce known integer sequences. The sequences for b_s (OEIS A113822) and $b_{s+1,s}$ (OEIS A085748) have been reported but little studied; $b_{s+1,s}$ counts rooted binary labeled trees with s leaves in which all leaves except one are labeled "1" and the last leaf is labeled "2"; equivalently, it is the number of rooted binary trees that are unlabeled except that one leaf is given a label. Sequence $b_{s,3}$ follows OEIS A002620 (Corollary 8), the "quarter-squares". The number of ways to place a given sample size across *some* caterpillar shape (Table 3) follows OEIS A000975, a sequence well studied in other contexts.

Interestingly, we observed that across all rooted binary unlabeled trees T with a fixed number of leaves n, the number of rooted binary perfect phylogenies for a fixed s appears to be largest for the caterpillar tree shape. For the case of $n = 4$, we proved this result, showing that for sample size $s \geq 5$, the caterpillar shape has more perfect phylogenies than the symmetric shape (Section 6). Informally, the number of rooted binary perfect phylogenies for fixed s and n, with only the tree shape changing, appears to decrease with increasing tree balance. The symmetry introduced by replacement of an asymmetric internal node by a symmetric internal node decreases the number of perfect phylogenies; it will be informative to systematically examine the number of perfect phylogenies in relation to tree balance indices such as the symmetry nodes index [14].

In accord with the result that caterpillars appear to possess larger numbers of perfect phylogenies, asymptotically as s grows large, whereas the mean number of leaves in a rooted binary perfect phylogeny selected at random grows with approximately $0.7326s$ (Theorem 13), the mean number of leaves in a rooted binary perfect phylogeny with caterpillar shape grows only with $\frac{1}{2}s$ (Theorem 20). The smaller value for the case of caterpillars reflects the fact that a caterpillar possesses only one symmetric node—its cherry—so that many distinct perfect phylogenies can be constructed with sample size s and a fixed small caterpillar size.

Perfect phylogenies have applications in multiple biological settings. They appear in problems concerning DNA sequences descended in a population from an ancestral sequence by a process with little or no genetic recombination [1,2,10, pp. 460–462]; a classic family of "perfect phylogeny problems" seeks to find algorithms for constructing perfect phylogenies from sets of sequences in this context. Recently, perfect phylogenies have also been considered in problems with cell lineages and tumors [6,18]. Our enumerative results assist in characterizing the sizes and combinatorial properties of sets of perfect phylogenies relevant to the various biological applications.

The binary perfect phylogenies are closely related to the rooted *multi-labeled* binary trees [4]. In a rooted multi-labeled binary tree, a shared label can be assigned to multiple leaves. Czabarka et al. [4] report the number $r_{\mathbf{m}} = r_{m_1,m_2,\ldots,m_k}$ of rooted multi-labeled binary trees, each of which is labeled by a given set of "multi-labels" $\{A_1, A_2, \ldots, A_k\}$, where label A_j appears m_j times.

Consider the integers at the leaves of a perfect phylogeny as "labels." A perfect phylogeny with sample size s and n leaves has integer labels $\mathbf{s} = (s_1, s_2, \ldots, s_n)$. The number of unique integer labels that appear in $\mathbf{s}$ is denoted k, and those integer labels appear $\mathbf{m} = (m_{t_1}, m_{t_2}, \ldots, m_{t_k})$ times, where the t_j are the k distinct integer labels represented in $\mathbf{s}$, m_{t_j} represents the number of leaves labeled by integer t_j, $\sum_{j=1}^{k} m_{t_j} = n$, and $\sum_{j=1}^{k} t_j m_{t_j} = s$. For example, the perfect phylogenies with labels $\mathbf{s} = (5, 4, 4, 1, 1)$ and $n = 5$ leaves correspond to the multi-labeled binary trees with 5 leaves and $k = 3$ multi-labels ("1", "4", and "5") and label multiplicities $\mathbf{m} = (m_1, m_4, m_5) = (2, 2, 1)$ (and $(t_1, t_2, t_3) = (1, 4, 5)$). Through the correspondence with multi-labeled trees, perfect phylogenies can potentially also be enumerated by summing enumerations for relevant sets of multi-labeled trees.

From the perspective of the lattice formulation for perfect phylogenies (Fig. 2), we have counted the (non-empty) elements of the lattice, b_s, and $b_{s,n}$, the number of elements that lie $s - n + 1$ "steps" from the minimal element ϕ to the maximal element, a single leaf. However, in describing lattices of binary perfect phylogenies, we have left a number of questions unanswered. In how many ways can the lattice be traversed—via the order relation—between the minimal and maximal perfect phylogenies? How many perfect phylogenies exist with specified features, perhaps concerning numbers of nodes with different numbers of descendants or leaves with specified multiplicities? Applications of the lattice formulation may provide further insights.

Acknowledgments

We thank Michael Fuchs, Bernhard Gittenberger, and Julia Palacios for discussions and for comments on the manuscript. We are very grateful to an anonymous reviewer for suggestions that substantially improved the manuscript. Support was provided by a National Science Foundation Graduate Research Fellowship and by National Institutes of Health grant R01 HG005855 and National Science Foundation grant DMS-2450005.

Data availability

No data were used for the research described in the article.

References

[1] V. Bafna, D. Gusfield, S. Hannenhalli, S. Yooseph, A note on efficient computation of haplotypes via perfect phylogeny, J. Comput. Biol. 11 (2004) 858–866, http://dx.doi.org/10.1089/cmb.2004.11.858.

[2] T.G. Clark, M. De Iorio, R.C. Griffiths, Bayesian logistic regression using a perfect phylogeny, Biostatistics 8 (2006) 32–52, http://dx.doi.org/10.1093/biostatistics/kxj030.

[3] L. Comtet, Advanced Combinatorics, Reidel, Boston, 1974, http://dx.doi.org/10.1007/978-94-010-2196-8.

[4] É. Czabarka, P.L. Erdős, V. Johnson, V. Moulton, Generating functions for multi-labeled trees, Discrete Appl. Math. 161 (2013) 107–117, http://dx.doi.org/10.1016/j.dam.2012.08.010.

[5] M. Drmota, Random Trees, Springer, Vienna, 2009, http://dx.doi.org/10.1007/978-3-211-75357-6.

[6] M. El-Kebir, G. Salas, L. Oesper, B.J. Raphael, Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures, Cell Syst. 3 (2016) 43–53, http://dx.doi.org/10.1016/j.cels.2016.07.004.

[7] J. Felsenstein, Inferring Phylogenies, Sinauer, Sunderland, MA, 2004.

[8] P. Flajolet, R. Sedgewick, Analytic Combinatorics, Cambridge University Press, Cambridge, 2009, http://dx.doi.org/10.1017/cbo9780511801655.

[9] D. Gusfield, Efficient algorithms for inferring evolutionary trees, Networks 21 (1991) 19–28, http://dx.doi.org/10.1002/net.3230210104.

[10] D. Gusfield, ReCombinatorics, MIT Press, Cambridge, 2014.

[11] E.F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. in Appl. Probab. 3 (1971) 44–77, http://dx.doi.org/10.2307/1426329.

[12] A. Hinz, The Lichtenberg sequence, Fibonacci Quart. 55 (2017) 2–12, http://dx.doi.org/10.1080/00150517.2017.12427786.

[13] J. Huang, M. Mickey, J. Xu, The nonassociativity of the double minus operation, J. Integer Seq. 20 (2017) 17.10.3, URL: https://api.semanticscholar.org/CorpusID:20035336.

[14] S. Kersting, M. Fischer, Measuring tree balance using symmetry nodes — a new balance index and its extremal properties, Math. Biosci. 341 (2021) 108690, http://dx.doi.org/10.1016/j.mbs.2021.108690.

[15] B.V. Landau, An asymptotic expansion for the Wedderburn–Etherington sequence, Mathematika 24 (1977) 262–265, http://dx.doi.org/10.1112/s0025579300009177.

[16] J.A. Palacios, A. Bhaskar, F. Disanto, N.A. Rosenberg, Enumeration of binary trees compatible with a perfect phylogeny, J. Math. Biol. 84 (2022) 54, http://dx.doi.org/10.1007/s00285-022-01748-w.

[17] M. Steel, Phylogeny: Discrete and Random Processes in Evolution, Society for Industrial and Applied Mathematics, Philadelphia, 2016, http://dx.doi.org/10.1137/1.9781611974485.

[18] Y. Wu, Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach, Bioinformatics 36 (2020) 742–750, http://dx.doi.org/10.1093/bioinformatics/btz676.