Sparsification of Large Ultrametric Matrices: Insights into the Microbial Tree of Life *

Evan D. Gorman† and Manuel E. Lladser†

Abstract. Ultrametric matrices have a rich structure that is not apparent from their definition. Notably, the subclass of strictly ultrametric matrices are covariance matrices of certain weighted rooted binary trees. In applications, these matrices can be large and dense, making them difficult to store and handle. In this manuscript, we exploit the underlying tree structure of these matrices to sparsify them via a similarity transformation based on Haar-like wavelets. We show that, with overwhelmingly high probability, only an asymptotically negligible fraction of the off-diagonal entries in random but large strictly ultrametric matrices remain non-zero after the transformation; and develop a fast algorithm to compress such matrices directly from their tree representation. We also identify the subclass of matrices diagonalized by the wavelets and supply a sufficient condition to approximate the spectrum of strictly ultrametric matrices outside this subclass. Our methods give computational access to a covariance model of the microbiologists' Tree of Life, which was previously inaccessible due to its size, and motivate defining a new but wavelet-based phylogenetic β-diversity metric. Applying this metric to a metagenomic dataset demonstrates that it can provide novel insight into noisy high-dimensional samples and localize speciation events that may be most important in determining relationships between environmental factors and microbial composition.

Key words. double principal coordinate analysis, Haar-like wavelets, sparsification, phylogenetic covariance matrix, strictly ultrametric matrix, Tree of life, UniFrac

MSC codes. 05C05, 15A18, 42C40, 65F55, 92C70

1. Introduction.

Ultrametric matrices appear across many domains of mathematics and science. They comprise an important class of matrices called inverse-M matrices [11] and are a key object of study in potential theory and Markov Chains [12]. In scientific applications, ultrametric matrices act as covariance models in phylogenetic comparative analysis [43], network inference [27] and energy models in statistical physics [8]. Further hinting at the pervasiveness of ultrametric matrices in modern data science, recent work has shown that the matrix of normalized Euclidean distances between points in some random subsets of $\mathbb{R}^d$ converge in probability to an ultrametric matrix as the dimension d tends to infinity [54, 55].

In many applications, the underlying ultrametric matrix can be massive, potentially too large to store in computer memory. This raises many challenges in the analysis and application of such matrices. However, if a sparse representation of a matrix can be found, many otherwise impossible tasks become computationally feasible. Examples of the latter include solving linear equations, matrix factorizations, eigenvalue decompositions, and principal component analysis (PCA).

In this paper we focus on the subclass of strictly ultrametric matrices, which is provided in the following definition.

*Submitted to the editors DATE.

Funding: This work was partially funded by the NSF grant No. 1836914.

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309, The United States (Corresponding author e-mail: manuel.lladser@colorado.edu)

In what remains of this manuscript, $n \geq 1$ denotes an integer. Define $[n] := \{1, \ldots, n\}$. In addition, vectors and sometimes functions are represented as column ones. The transpose of a vector or matrix A is denoted A'.

Definition 1.1 ([50]). *A matrix $S \in \mathbb{R}^{n\times n}$ is ultrametric if it is symmetric with non-negative entries and $S(i,j) \geq \min\{S(i,k), S(k,j)\}$ for all $i,j,k \in [n]$. If in addition $S(i,i) > \max\{S(i,t) : t \neq i\}$, for all $i \in [n]$, S is called strictly ultrametric. For $n = 1$, the last inequality is replaced with $S(i,i) > 0$.*

Ultrametric matrices have rich properties that are not made evident by their definition [11]. In particular, if S is strictly ultrametric then it is positive definite (hence invertible), S^{-1} is strictly diagonally dominant with non-positive off-diagonal entries, and $S(i,j) = 0$ if and only if $S^{-1}(i,j) = 0$. These properties were initially proved using probabilistic methods [37]. An alternative proof is based on an equivalence between strictly ultrametric matrices and a subclass of binary trees [40]. The key ingredient for this equivalence is that for $n > 1$, if $S \in \mathbb{R}^{n\times n}$ is symmetric with non-negative entries then it is strictly ultrametric if and only if there exists a permutation matrix P and strictly ultrametric matrices A and B such that

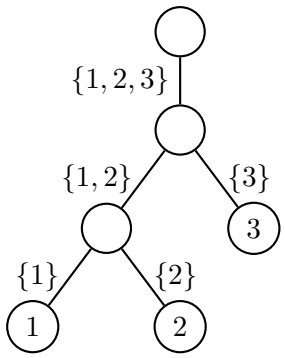
$$P\big(S - \min(S)\,\mathbf{1}\mathbf{1}'\big)P' = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \tag{1.1}$$

where $\min(S)$ is the smallest entry in S, and $\mathbf{1} \in \mathbb{R}^n$ is the column vector of ones [40, Proposition 2.1]. Since A and B are of the same kind as S, this process may be applied recursively and the matrix S encoded as a weighted rooted binary tree with special characteristics. Here we adopt a slightly different encoding to the one in [40], which is more suitable for our purposes. The reader unfamiliar with the jargon and notation of trees may skip ahead to Section 1.2 and come back to make better sense of the construction below.

Let S be a strictly ultrametric matrix of dimensions $n \times n$. We can represent S as a rooted binary tree with $2n$ vertices (of which half are leaves) and hence $(2n-1)$ edges, satisfying the following definition.

Definition 1.2. *An out-rooted bifurcating tree (ORB-tree) with n leaves is a weighted rooted tree with the following properties: each vertex has degree 1 or 3; its leaf set is $[n]$ and excludes the root, which has degree 1; each edge is labeled by the subset of leaves that descend from it; and the length $\ell(e)$ of each edge e is non-negative but $\ell(e) > 0$ when e connects a leave with its parent.*

The representation of a strictly ultrametric matrix S as an ORB-tree may be obtained as follows. The only edge emanating from the root is labeled as $[n]$ and defined to have length $\min(S)$. The only child of the root has two children. One child descends from an edge labeled by the rows (or columns) of S associated with the matrix A before applying the permutation matrix P in (1.1). This edge has length $\min(A)$. Likewise, the other child descends from an edge labeled by the rows associated with the matrix B and has length $\min(B)$. Since A and B are strictly ultrametric, just of smaller dimensions, the tree may be grown recursively



e	$\ell(e)$	δ_e
$\{1,2,3\}$	2	$(1,1,1)'$
$\{1,2\}$	0	$(1,1,0)'$
$\{1\}$	3	$(1,0,0)'$
$\{2\}$	1	$(0,1,0)'$
$\{3\}$	2	$(0,0,1)'$

Figure 1: **ORB-tree and strictly ultrametric matrix correspondence.** The matrix encoding of the tree is $\begin{pmatrix} 5 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 4 \end{pmatrix} = 2\,\delta_{\{1,2,3\}}\delta'_{\{1,2,3\}} + 0\,\delta_{\{1,2\}}\delta'_{\{1,2\}} + 3\,\delta_{\{1\}}\delta'_{\{1\}} + \delta_{\{2\}}\delta'_{\{2\}} + 2\,\delta_{\{3\}}\delta'_{\{3\}}$.

from any descendent of the root that is not associated with a strictly ultrametric matrix of dimensions 1×1. The latter represent edges that parent a leave in the ORB-tree. These edges must have a strictly positive length because 1×1 strictly ultrametric matrices are strictly positive real numbers. To fix ideas see Figure 1.

For each edge e in the ORB-tree define δ_e as the vector of dimension n with entries $\delta_e(i) = 1$ for $i \in e$ and $\delta_e(i) = 0$ for $i \in [n] \setminus e$. It follows from [40, Theorem 2.2] that

$$S = \sum_{e \in E} \ell(e)\, \delta_e\, \delta'_e. \tag{1.2}$$

Conversely, starting with any ORB-tree with edge set E and length function $\ell : E \to [0, \infty)$, the matrix $\sum_{e \in E} \ell(e)\, \delta_e\, \delta'_e$ is strictly ultrametric. This representation of strictly ultrametric matrices as ORB-trees is therefore one-to-one. In fact, for $i, j \in [n]$, the entries of the matrix S associated with an ORB-tree can be computed by direct inspection of tree using that

$$S(i,j) = \sum_{e \in [i \wedge j, \circ]} \ell(e), \tag{1.3}$$

where the $\circ$ denotes the root of the tree. We may say therefore that the entries of a strictly ultrametric matrix are indexed by the leaves of its associated ORB-tree.

We call a matrix with entries such as (1.3) the covariance matrix of the ORB-tree. This terminology is borrowed from the ecology literature where matrices like this are commonly referred to as a tree-structured or phylogenetic covariance matrices. In this setting, the leaves represent organisms, and the matrix entries denote a trait's covariance between pairs of organisms [9]. (The term of cophenetic matrix or cophenetic distance has also been used occasionally in the hierarchical clustering literature [45].)

A key tool in understanding and analyzing non-stationary and noisy continuous signals are wavelets: localized, wave-like functions. Traditional wavelets are defined only in Euclidean spaces and have been remarkably successful in identifying multiscale structures in signals and producing sparse representations of the same. [36]

The Haar wavelet is among the oldest and involves averaging a signal locally at different time or space scales [20]. Recently, the authors of [18] extended it past continuous signals introducing the Haar-like wavelet. This new wavelet is designed for the multiscale analysis of discrete datasets equipped with a partition tree—a hierarchical structure that clusters the data into smaller subsets recursively. Due to the organization of such datasets into different tree levels (i.e., scales) and clusters (i.e., localizations), Haar-like wavelets may identify meaningful structures in data that may be impossible to distinguish otherwise—especially in noisy high dimensional datasets.

Strictly ultrametric matrices can be fully dense; i.e all of their entries be non-zero. Nevertheless, due to the identity in equation (1.3), their entries contain much redundancy, suggesting they may be amenable to some form of compression.

This paper exploits the equivalence between strictly ultrametric matrices and ORB-trees to sparsify and hence compress the former via a change of basis. This basis is composed of the so-called Haar-like wavelets of the associated ORB-trees. The sparsification achieved by these wavelets can be substantial in large, strictly ultrametric matrices, giving computational access to matrices previously inaccessible due to their size. This can be of great value in extensive phylogenetic studies due to the interpretation of these matrices as covariance matrices of phylogenetic trees. It may also find practical applications in the context of double principal coordinate analysis, a metric of phylogenetic diversity among microbial environments.

1.1. Paper organization. In Section 2 we introduce the Haar-like basis from [18] and give a geometric interpretation of its action on ORB-trees. Then, Section 3 presents the conditions under which the Haar-like basis can be used to sparsify large, strictly ultrametric matrices. We show that the basis can substantially sparsify most large random ORB-tree's covariance matrices. We also present an algorithm for directly computing the sparsified matrix from the ORB-tree of the original matrix; in particular, without having to pre-compute the strictly ultrametric matrix from the tree. Following in Section 4, we detail the case in which the Haar-like basis diagonalizes (i.e., fully sparsifies) a strictly ultrametric matrix and provide examples of well-known tree topologies. And, in Section 5, we show that the conditions necessary for diagonalization can be relaxed and the Haar-like basis used to estimate eigenvalues of ORB-tree's covariance matrices.

Finally, in Section 6, we apply our methods to a covariance model of the 97% Greengenes tree, a standard representative phylogeny of microbiologists' Tree of Life. The sparsification opens the door for otherwise impossible tasks related to this model, such as computing the spectrum or inverse of its covariance matrix—standard tasks in phylogenetic comparative methods. We also introduce a new wavelet-based phylogenetic (β-diversity) metric corresponding to a multiscale analysis of organism abundances in microbial environments. This novel metric gives remarkably similar results to other well-known metrics on a previously studied dataset; however, it can also determine the speciation events responsible for the observed microbial compositions and quantify their respective importance.

1.2. Paper notation and terminology. For real-vectors $x = (x_i)_{1\leq i\leq k}$ and $y = (y_i)_{1\leq i\leq k}$ of dimension k, we define $\langle x, y\rangle := x'y = \sum_{i=1}^{k} x_i y_i$ and $\|x\|_2 := \sqrt{\langle x, x\rangle}$. Also, $[\![\cdot]\!]$ denotes the indicator function of the proposition within the parentheses.

In our context, trees are finite undirected connected graphs without cycles.

In what remains of this manuscript, T denotes an ORB-tree with n leaves and branch length function $\ell : E \to \mathbb{R}$. We denote the vertex and edge set of T as V and E, respectively. The root of T is denoted as $\circ$. The set of internal nodes of T is denoted as I, whereas its set of leaves is denoted as L. By definition, $\circ \in I$ and I and L partition V. From the definition of ORB-tree it also follows that $|L| = |I| = n$, hence $|V| = 2n$. Note that $|E| = |V| - 1$ because T is a tree. We define $|T| := |V|$. We use this later notation when we want to emphasize a direct relationship with the ORB-tree.

For $i, j \in V$, a path of length l between i and j is a sequence $v_0, \ldots, v_l \in V$ such that $v_0 = i$, $v_l = j$, and $\{v_k, v_{k+1}\} \in E$ for $0 \leq k < l$. Unless otherwise stated, we write $[i, j]$ to denote the set of edges in the shortest path in T between i and j. This path is unique because T has no cycles. The depth of i, denoted $\mathrm{depth}(i)$, is defined as $|[i, \circ]|$ i.e. the number of edges that connect i with the root. We say that i is an ancestor of j, or alternatively j is a descendent of i, when $i \in [\circ, j]$. In particular, every node is an ancestor and a descendant from itself. Further, $(i \wedge j)$ denotes the so-called least-common ancestor to i and j. This is the $v \in V$ that minimizes $|[v, \circ]|$, among all the nodes that are ancestors to both i and j.

We define

$$\ell(i, j) := \sum_{e \in [i,j]} \ell(e).$$

In addition, for $J \subset L$ and $i \in V$, define $\ell(J, i)$ as the column vector of dimension $|J|$ with entries $\ell(j, i)$, for $j \in J$. $\ell(i, J)$ is the transpose of $\ell(J, i)$.

For each $i \in V$, $T(i)$ denotes the subtree of T rooted at i. In particular, the vertex set of $T(i)$ is the subset of nodes in T that descend from i, and its edge set is the subset of edges that connect two descendants of i. $L(i)$ denotes the leaf set of $T(i)$. Likewise, for each $e = \{i, j\} \in E$, if i is closest to the root than j, $T(e)$ and $L(e)$ denote $T(i)$ and $L(i)$, respectively.

2. Haar-like basis of ORB-trees.

In this section we specialize the concept of Haar-like basis given in [18] to our setting of ORB-trees. The key new result in this section shows that the Haar-like basis of an ORB-tree interacts nicely with its covariance matrix (i.e. the strictly ultrametric matrix associated with the tree). This is somewhat unexpected because the covariance matrix is determined by the topology of the tree and its branch length function, whereas the basis is solely determined by the tree's topology.

To construct the Haar-like wavelets, it is convenient to represent the nodes in $I \setminus \{o\}$ momentarily as binary strings. With this convention, the (only) child of the root is ε—the so-called empty string. Further, the children of each node $v \in I \setminus \{o\}$ are $v0$ (i.e. the string v with the character zero appended at the end) and $v1$ (i.e. v with the character one appended at the end).

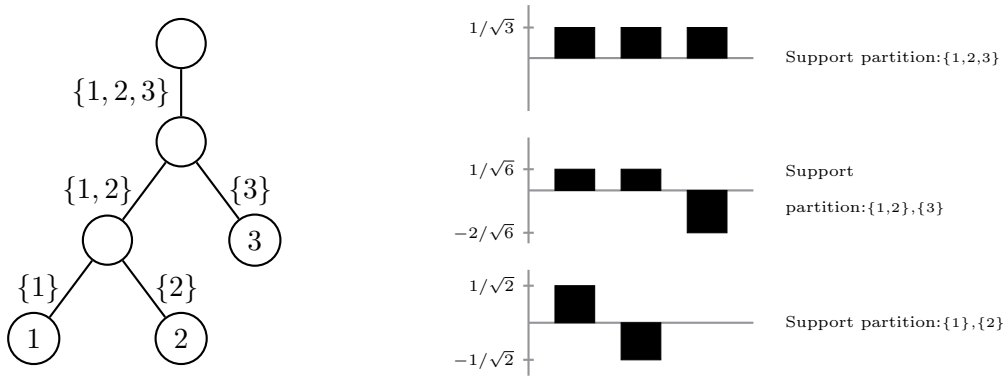


Figure 2: **Visualization of the Haar-like Wavelet basis associated with an ORB-tree.** Left: ORB-tree with leaves $1, 2, 3$. Edges are labeled by the subsets of leaves that descend from them. Right: Haar-like basis associated with the ORB-tree on the left.

Definition 2.1 (Specialization from [18]). *The Haar-like basis associated with T is the set of transformations $\{\varphi_v\}_{v\in I}$ defined as follows:*

$$\varphi_o(i) := \frac{1}{\sqrt{|L|}}, \text{ for all } i \in L;$$

and, for each $v \in I$ with $v \neq o$:

$$\varphi_v(i) := \begin{cases} +\sqrt{\frac{|L(v1)|}{|L(v0)|\cdot|L(v)|}} & , i \in L(v0); \\ -\sqrt{\frac{|L(v0)|}{|L(v1)|\cdot|L(v)|}} & , i \in L(v1); \\ 0 & , \text{otherwise.} \end{cases}$$

The Haar-like matrix associated with T is the matrix Φ with columns φ_v, $v \in I$.

To fix ideas see Figure 2.

The terminology of basis in Definition 2.1 is justified by the fact that

$$\text{if } u, v \in I \text{ then } \langle \varphi_u, \varphi_v \rangle = [\![u = v]\!]. \tag{2.1}$$

In particular, $\{\varphi_v\}_{v\in I}$ is an orthonormal basis of $\mathbb{R}^{|L|}$. (See the Appendix for a self-contained justification of the orthonormality of the Haar-like basis.) Note that the Haar-like matrix Φ has its rows indexed by L and its columns indexed by I. Hence, since $|L| = |I|$, Φ is a square matrix, an orthonormal one.

Clearly, for each $v \in I$, φ_v has $L(v)$ as its support. Further, because φ_o is constant, the orthogonality property implies for $v \neq o$ that $\sum_{i\in L} \varphi_v(i) = 0$. These two properties are essential for our arguments onwards.

The following definition is useful to understand the relationship between the Haar-like

basis of an ORB-tree and its associated covariance matrix.

Definition 2.2. *The trace branch length of T is the function $\ell^* : E \to [0, \infty)$ defined as $\ell^*(e) := |L(e)|\,\ell(e)$, for each $e \in E$.*

Theorem 2.3. *If $v \in I$ then $S\,\varphi_v = diag(\ell^*(L, v))\,\varphi_v$.*

Proof. Consider $v \in I$ and $j \in L$. If $j \notin L(v)$ then $(i \wedge j) = (v \wedge j)$, hence

$$(S\,\varphi_v)(j) = \sum_{i \in L(v)} \ell(i \wedge j, \circ)\,\varphi_v(i) = \ell(v \wedge j, \circ) \sum_{i \in L(v)} \varphi_v(i).$$

But, if $v = \circ$ then $\ell(v \wedge j, \circ) = 0$. Instead, if $v \neq \circ$ then $\sum_{i \in L(v)} \varphi_v(i) = 0$. In either case: $(S\,\varphi_v)(j) = 0$. This shows the lemma for $j \notin L(v)$ because the entry associated with j in $\mathrm{diag}(\ell^*(L, v))\,\varphi_v$ is $\ell^*(v, j) \cdot \varphi_v(j)$, and the support of φ_v is $L(v)$.

Next suppose that $j \in L(v)$. Then

$$\begin{aligned}
(S\,\varphi_v)(j) &= \sum_{i \in L(v)} \sum_{e \in [i \wedge j, o]} \ell(e)\,\varphi_v(i) \\
&= \sum_{i \in L(v)} \sum_{e \in [i \wedge j, v]} \ell(e)\,\varphi_v(i) + \sum_{i \in L(v)} \varphi_v(i) \cdot \sum_{e \in [v, o]} \ell(e) \\
&= \sum_{i \in L(v)} \varphi_v(i) \sum_{e \in [i \wedge j, v]} \ell(e),
\end{aligned}$$

where for the last identity we have used that $\sum_{i \in L(v)} \varphi_v(i) = 0$ if $v \neq \circ$, and $\sum_{e \in [v,o]} \ell(e) = 0$ if $v = \circ$. But note that if $i \in L(v)$ is such that $(i \wedge j) = v$ then $\sum_{e \in [i \wedge j, v]} \ell(e) = 0$. Instead, if $(i \wedge j) \neq v$ then $\varphi_v(i) = \varphi_v(j)$. As a result

$$\begin{aligned}
(S\,\varphi_v)(j) &= \varphi_v(j) \sum_{i \in L(v):\, i \wedge j \neq v} \sum_{e \in [i \wedge j, v]} \ell(e) \\
&= \varphi_v(j) \sum_{e \in [j, v]} \sum_{i \in L(v):\, e \in [i \wedge j, v]} \ell(e) \\
&= \varphi_v(j) \sum_{e \in [j, v]} \ell(e)\,|L(e)| \\
&= \varphi_v(j)\,\ell^*(j, v),
\end{aligned}$$

which shows the result. ∎

It follows from the theorem that for each $u, v \in I$:

$$(\Phi' S \Phi)(u, v) = \varphi_u' S \varphi_v = \sum_{i \in L(v) \cap L(u)} \varphi_u(i)\,\varphi_v(i)\,\ell^*(i, v). \tag{2.2}$$

The importance of the diagonal of $\Phi' S \Phi$ in the discussion ahead, motivates to define for $v \in I$ the quantities

$$\lambda_v := (\Phi' S \Phi)(v, v) = \sum_{i \in L(v)} \varphi_v^2(i)\, \ell^*(i, v). \tag{2.3}$$

For $v \in I$, because φ_v has $L(v)$ as its support and $\|\varphi_v\|_2 = 1$, λ_v is a weighted average of the trace branch length between each leaf in $L(v)$ and v. In particular, since $L(u) \supset L(v)$ when u is an ancestor of v, the closer the internal node v is to the root, the more terms are averaged. (This emulates the averaging at different scales that the standard Haar wavelet transform does to a continuous signal.) Furthermore, since $\ell^*(e) = \ell(e) > 0$ when e joins a leave with its parent, $\lambda_v > 0$.

On the other hand, because $(\Phi' S \Phi)(u, v) = 0$ when $u, v \in I$ are such that $L(u) \cap L(v) = \emptyset$, the identity in (2.2) suggests that the Haar-like matrix can be used to sparsify the covariance matrix of the ORB-tree. The following result is critical to assess how effective this sparsification is in practice.

Lemma 2.4. *For all $u, v \in V$, $L(u) \cap L(v) \neq \emptyset$ if and only if u is an ancestor of v or vice versa.*

Proof. If u is an ancestor of v then $L(v) \subset L(u)$; in particular, $L(u) \cap L(v) = L(v) \neq \emptyset$. The same conclusion applies if v is an ancestor of u. Conversely, suppose that $L(u) \cap L(v) \neq \emptyset$. Without loss of generality assume that $u \neq v$. From the hypothesis, there is $w \in L$ that descends from both u and v. But, since there is a unique path from w to $\circ$, u and v must be both in this path; in particular, either u is an ancestor of v or vice versa. ∎

3. Sparisfication of Covariance Matrices of ORB-trees.

In this section we quantify how much of the covariance matrix of an ORB-tree can be sparsified by its Haar-like matrix. To state our main result we require the following definitions.

Definition 3.1. *The average subtree size of T is the quantity, $avg(T) := \frac{1}{|T|} \sum_{v \in V} |T(v)|$.*

Definition 3.2. *The internal and external path lengths of T are the quantities defined as $IPL(T) := \sum_{v \in I} depth(v)$ and $EPL(T) := \sum_{v \in L} depth(v)$, respectively [46]. The total path length of T is the quantity $TPL(T) := IPL(T) + EPL(T)$.*

We note the relationship:

$$\mathrm{avg}(T) = 1 + \frac{\mathrm{TPL}(T)}{|T|}, \tag{3.1}$$

because

$$\mathrm{TPL}(T) = \sum_{v\in V}\sum_{u\in V\setminus\{\circ\}} [\![v \in T(u)]\!] = \sum_{u\in V\setminus\{\circ\}} |T(u)| = \left\{\sum_{u\in V} |T(u)|\right\} - |T|.$$

Definition 3.3. *The interior of T is the tree $\mathring{T}$ obtained by trimming the leaves of T.*

Clearly, $\mathrm{IPL}(T) = \mathrm{TPL}(\mathring{T})$.

As mentioned earlier, the identity in (2.2) guarantees that some entries of $\Phi' S \Phi$ vanish. The following result estimates the least number of such entries. Our lower bound is independent of the branch lengths and depends—only—on the tree topology.

Theorem 3.4. *If ζ denotes the fraction of vanishing entries in $\Phi' S \Phi$ then*

$$\zeta \geq 1 + \frac{1}{|L|} - 2\frac{avg(\mathring{T})}{|\mathring{T}|} = 1 - \frac{1}{|L|} - 2\frac{TPL(\mathring{T})}{|\mathring{T}|^2}.$$

Proof. Recall that $|I| = |L|$ and, for $v \in I$, the support of φ_v is $L(v)$. Hence, from the identity in (2.2), $(\Phi' S \Phi)(u, v) = 0$ when $u, v \in I$ and $L(u) \cap L(v) = \emptyset$. As a result, using that $L(u) \neq \emptyset$ when $u \in I$, and Lemma 2.4, we obtain that

$$\begin{aligned}
|I|^2\zeta &\geq |I|^2 - |\{(u,v) \in I \times I \text{ such that } L(u) \cap L(v) \neq \emptyset\}| \\
&= |I|^2 - |I| \\
&\qquad - 2|\{(u,v) \in I \times I \text{ such that } v \neq u \text{ descends from } u \text{ and } L(u) \cap L(v) \neq \emptyset\}| \\
&= |I|^2 - |I| - 2\sum_{u\in I}\left(|\mathring{T}(u)| - 1\right) \\
&= |I|^2 + |I| - 2\sum_{u\in I}|\mathring{T}(u)|,
\end{aligned}$$

Since $|I| = |L| = |\mathring{T}| = |T|/2$, $|L|^2\zeta \geq |L|^2 + |L| - |\mathring{T}| \cdot \mathrm{avg}(\mathring{T})$, which shows the inequality in the theorem. The alternative lower-bound for ζ follows by applying the identity in equation (3.1) to $\mathring{T}$, completing the proof of the theorem. ∎

It follows from the first lemma in [46, Section 6.4] that for an ORB-Tree T, $\mathrm{EPL}(T) - \mathrm{IPL}(T) = 2|I| - 1$, which together with the previous theorem let us conclude the following asymptotic result.

Corollary 3.5. *If either $avg(\mathring{T}) \ll |\mathring{T}|$, $TPL(\mathring{T}) \ll |\mathring{T}|^2$, $IPL(T) \ll |I|^2$, or $EPL(T) \ll |L|^2$ as $|T| \to \infty$, then $\zeta = 1 - o(1)$.*

In other words, if T grows so that either of the asymptotic inequalities in the above corollary applies, then an asymptotically negligible fraction of the off-diagonal entries in $\Phi' S \Phi$ will be non-zero.

The last asymptotic condition in the corollary (i.e., that $\mathrm{EPL}(T) \ll |L|^2$) is of relevance in phylogenetic studies. In that context, the external path length of a tree is called its Sackin's index [3, 10, 29]. This index is used as a measure of the imbalance of phylogenetic trees. In particular, since phylogenetic trees are neither too balanced nor too imbalanced [2, 4]; Haar-like bases should be rather effective in sparsifying covariance matrices of phylogenetic trees in practice. We come back to this point in Section 6.

3.1. Sparsification of covariance matrices of maximally balanced ORB-trees. In our context, the following definition gives the most balanced topology among the ORB-trees.

Definition 3.6 (Perfect Binary Trees). *A perfect binary tree is an ORB-tree in which all leaves have the same depth.*

To fix ideas see Figure 3.

Let T be a perfect binary tree of height $(h+1)$. In particular, $|L| = 2^h$ and $|V| = 2^{h+1}$. At level $k \geq 1$, T contains 2^{k-1} nodes, each of which is the root of a perfect binary tree of height $(h-k)$. Since a perfect binary tree of height h contains $(2^{h+1}-1)$ nodes, and the interior of a perfect binary tree of height $(h+1)$ is a perfect binary tree of height h:

$$\mathrm{avg}(\mathring{T}) = \frac{2^h + \sum_{k=1}^{h} 2^{k-1} \cdot (2^{h-k+1} - 1)}{2 \cdot 2^{h-1}} = h + 2^{-h} \ll |\mathring{T}|.$$

In particular, due to Theorem 3.5, we can conclude that the Haar-like matrix can be used to asymptotically annihilate (via a similarity transformation) the off-diagonal entries of the covariance matrix of a perfect binary tree as its height tends to infinity.

3.2. Sparsification of covariance matrices of maximally imbalanced ORB-trees. The following definition provides what we may regard as the most unbalanced topology among the ORB-trees.

Definition 3.7 (Binary Caterpillar Trees). *The binary caterpillar tree of height $h \geq 1$ is the ORB-tree with vertices $\circ, 1, \ldots, h, 1', \ldots, (h-1)'$ and edges of the form $\{\circ, 1\}$, $\{i, i+1\}$, for $i = 1, \ldots, (h-1)$, and $\{j, j'\}$ for $j = 1, \ldots, (h-1)$.*

To fix ideas see Figure 4.

Let T be a binary Caterpillar tree of height h. In particular, $|V| = 2h$, $|L| = h$, and each node in level $k \geq 1$ is the root of a snake binary subtree of height $(h-k)$. Therefore, each internal node has as children one leaf node and one internal node that is the root of a binary

caterpillar subtree, and the subgraph of internal nodes is a path. As a result:

$$\operatorname{avg}(\mathring{T}) = \frac{2\sum_{k=1}^{h-1}(h-k)}{2(h-1)} = \frac{h}{2} + 1 \sim \frac{|\mathring{T}|}{2}.$$

Hence, the lower-bound provided by Theorem 3.4 is trivial, and we cannot guarantee that the Haar-like matrix associated with a sizeable binary caterpillar tree annihilates its off-diagonal entries in any significant way.

3.3. Sparsification of covariance matrices of large random ORB-trees. Perfect binary trees and caterpillar trees are opposite extremes of how balanced (or imbalanced) ORB-trees can be. It is therefore unclear how much sparsification the Haar-like matrix of a large but generic ORB-tree can induce on its covariance matrix. To address this issue we consider a natural ensemble of random ORB-trees

In what follows, $\mathbb{T}$ denotes a uniformly at random ORB-tree with $|I|$ internal nodes. Such trees may be generated using the Catalan distribution [46, Section 6.7]. This probability model produces full binary trees with a given number of internal nodes; which we may turn into an ORB-tree by appending their root to a new one.

Let $\mathbb{S}$ denote the covariance matrix of $\mathbb{T}$, and ζ the number of zeroes in the random matrix $\Phi'\mathbb{S}\Phi$, where Φ is the Haar-like matrix associated with $\mathbb{T}$. It turns out that the mean and variance of the internal path length of $\mathbb{T}$ are given by

$$\mathbb{E}\big(\mathrm{IPL}(\mathbb{T})\big) \sim \sqrt{\pi}|I|^{3/2}; \tag{3.2}$$

$$\mathbb{V}\big(\mathrm{IPL}(\mathbb{T})\big) \sim \left(\frac{10}{3} - \pi\right)|I|^3. \tag{3.3}$$

The identity in equation (3.2) follows from [15, Proposition VII.3.]. The identity in (3.3) may be regarded a refinement of [15, Note VII.12].

As the following result implies, the Haar-like basis of most large ORB-trees should be highly effective in sparsifying their covariance matrix.

> Corollary 3.8. *If $\mathbb{T}$ is a uniformly at random ORB-tree with $|I|$ internal nodes then $\zeta = 1 - o(1)$ with overwhelmingly high probability, as $|I| \to \infty$.*

Proof. Let $t > 0$. Let μ and σ^2 denote the mean and variance of $\mathrm{IPL}(\mathbb{T})$, respectively. Due to Cantelli's inequality (a one sided version of the well-known Chebyshev's inequality): $\mathbb{P}\big(\mathrm{IPL}(\mathbb{T}) \geq \mu + t\sigma\big) \leq (1+t^2)^{-1}$. But $(\mu + t\sigma) = \Omega\big(t|I|^{3/2}\big)$ because of equations (3.2)-(3.3). In particular, there is a constant $c > 0$ such that

$$\mathbb{P}\left(\frac{\mathrm{IPL}(\mathbb{T})}{|I|^2} \leq \frac{ct}{\sqrt{|I|}}\right) \geq \frac{t^2}{1+t^2}.$$

So, if $t \to \infty$ so that $t = o(\sqrt{|I|})$ then $\frac{\mathrm{IPL}(\mathbb{T})}{|I|^2} = o(1)$ with a probability converging to one as $|I| \to \infty$. The result now follows from Corollary 3.5. ∎

3.4. Fast Sparsification Algorithm. A non-trivial challenge to storing and manipulating large strictly ultrametric matrices is that they are almost always fully dense in applications. Further, in the context of phylogenetic covariance matrices, the ORB-trees associated with such matrices are formed in advance. This allows us to sparsify these matrices without computing them, or even storing them in computer memory. It also allows us to anticipate which entries may remain nonzero after sparsification. In fact, due to equation (2.2) and Lemma 2.4, all that is required to sparsify these matrices from their ORB-tree is to precompute the leaves that descend from each internal node (i.e., the sets $L(v)$, with $v \in I$) and the trace branch length between them (Definition 2.2). This can be achieved with two postorder traversals of the ORB-tree. We convey these ideas in the following pseudo-code (Algorithm 3.1), which is fully coded and available on GitHub.

Algorithm 3.1 Phylogenetic covariance matrix sparsification

Input. ORB-tree T with covariance matrix S
Output. Only possibly non-zero entries in $\Phi' S \Phi$
for $v \in I$ in postorder traversal of T **do**
  for $i \in L$ **do**
    if $v = \circ$ **then**
      $\varphi_o(i) \leftarrow \frac{1}{\sqrt{|L|}}$
    else if $i \in L(v0)$ **then**
      $\varphi_v(i) \leftarrow + \sqrt{\frac{|L(v1)|}{|L(v0)| \cdot |L(v)|}}$
      $\ell^*(v,i) \leftarrow \ell^*(i,v0) + |L(v0)| \cdot \ell(v0,v)$
    else if $i \in L(v1)$ **then**
      $\varphi_v(i) \leftarrow - \sqrt{\frac{|L(v0)|}{|L(v1)| \cdot |L(v)|}}$
      $\ell^*(v,i) \leftarrow \ell^*(i,v1) + |L(v1)| \cdot \ell(v1,v)$
    else
      $\varphi_v(i) \leftarrow 0$
    end if
  end for
end for
for $v \in I$ in postorder traversal of T **do**
  while parent$(v) \neq \emptyset$ **do**
    $u \leftarrow$ parent(v)
    $M(u,v) \leftarrow \sum_{i \in L(v) \cap L(u)} \varphi_u(i)\, \varphi_v(i)\, \ell^*(v,i)$
  end while
end for
return $M(u,v)$ for $u, v \in I$ such that $L(u) \cap L(v) \neq \emptyset$

4. Spectrum of Covariance Matrices of Trace-balanced ORB-trees. While Theorem 3.4 guarantees that some entries in $\Phi' S \Phi$ vanish—regardless of branch lengths, additional constraints on the latter can lead to further sparsification. The next definition identifies the class of ORB-trees whose Haar-like basis fully sparsifies (i.e, diagonalizes) their associated strictly ultrametric matrix.

Definition 4.1. *T is called trace-balanced at a node v when, for all $i, j \in L(v)$, $\ell^*(i, v) = \ell^*(j, v)$. T is called trace-balanced when it is trace-balanced at each $v \in I \setminus \{\circ\}$.*

We note that a tree is always trace-balanced at a leaf. Also, if an ORB-tree is trace-balanced at the child of its root then it is also trace-balanced at the root. This is why the definition of trance-balanced trees only considers nodes in $I \setminus \{o\}$. See Figures 3-4 for depictions of trace-balanced trees.

The following two results show the relevance of the above definition in terms of the eigenvalues of the ultrametric matrix associated with an ORB-tree.

Lemma 4.2. *If $v \in I$ then φ_v is an eigenvector of S if and only if T is trace-balanced at v, in which case the eigenvalue associated with φ_v is $\ell^*(i, v)$, for any $i \in L(v)$.*

Proof. Fix $v \in I$. Due to Theorem 2.3, $(S\varphi_v)(i) = \ell^*(i, v)\, \varphi_v(i)$, for each $i \in L$. This shows the lemma because $\varphi_v(i) > 0$ if and only if $i \in L(v)$. ∎

Since the covariance matrix S has dimensions $|L| \times |L|$, but $|I| = |L|$ because T is an ORB-tree, the following result is immediate from the previous lemma.

Corollary 4.3. *The Haar-like basis of T diagonalizes its covariance matrix if and only if T is trace-balanced. In this case, the spectrum of S is*

$$\sigma(S) = \bigcup_{v \in I} \{\ell^*(v, i) \text{ for any } i \in L(v)\},$$

and the multiplicity of $\ell^*(v, i)$ is $\big|\{u \in I : \ell^*(v, i) = \ell^*(u, j), \text{ for some } j \in L(u)\}\big|$.

Due to Corollary 4.3, we can assert the following.

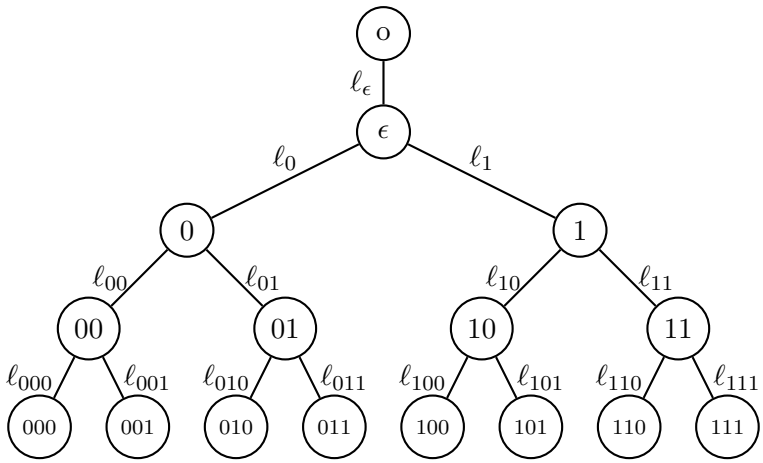


Figure 3: **Visualization of a perfect binary tree of height 4.** Such tree is trace-balanced if and only if $\ell_\alpha = \ell_\beta$ for each pair of binary strings α and β of the same length. If all these lengths are strictly positive, the eigenvalues of its covariance matrix are ℓ_{000} (multiplicity 4), $\ell_{000} + 2\,\ell_{00}$ (multiplicity 2), $\ell_{000} + 2\,\ell_{00} + 4\,\ell_0$ (multiplicity 1), and $\ell_{000} + 2\,\ell_{00} + 4\,\ell_0 + 8\,\ell_\epsilon$ (multiplicity 1). Otherwise, some multiplicities need to be added up.

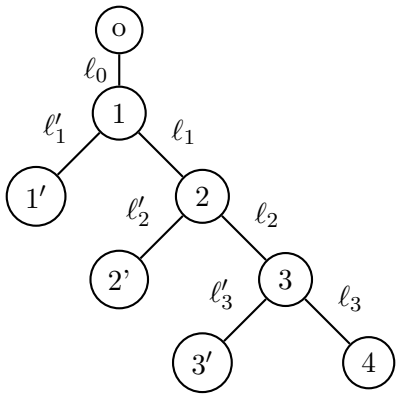


Figure 4: **Visualization of a binary caterpillar tree of height 4.** This tree is trace-balanced if and only if $\ell'_3 = \ell_3$, $\ell'_2 = \ell_3 + 2\,\ell_2$, and $\ell'_1 = \ell_3 + 2\,\ell_2 + 3\,\ell_1$. In such case, if $\ell_3, \ell_2, \ell_1, \ell_0 > 0$ then its covariance matrix spectrum is $\{\ell_3, \ell_3 + 2\,\ell_2, \ell_3 + 2\,\ell_2 + 3\,\ell_1, \ell_3 + 2\,\ell_2 + 3\,\ell_1 + 4\,\ell_0\}$, and each eigenvalue is simple.

Corollary 4.4. *The perfect binary tree of height h is trace-balanced if and only if it has constant branch lengths at each level. In this case, if ℓ_j denotes the common length of the edges that connect a node at depth j with another at depth $(j+1)$, the spectrum of the associated covariance matrix is*

$$\sigma(S) = \left\{ \sum_{k=j}^{h-1} 2^{h-k} \ell_k, \ \textit{with } j = 0, \ldots, h-1 \right\}.$$

Furthermore, the multiplicity of the eigenvalue $\lambda = \sum_{k=j}^{h-1} 2^{h-k} \ell_k$ *is* $\max\left\{1, \sum_{j \in \Lambda} 2^{j-1}\right\}$, *where*

$$\Lambda := \left\{ j' \in \{0, \ldots, h-1\} \textit{ such that } \sum_{k=j'}^{h-1} 2^{h-k} \ell_k = \sum_{k=j}^{h-1} 2^{h-k} \ell_k \right\}.$$

Next, consider the binary caterpillar tree from Definition 3.7. In particular, its internal and leaf set are $I = \{\circ, 1, \ldots, h-1\}$ and $L = \{1', \ldots, (h-1)', h\}$, respectively. Let ℓ_0 denote the branch length of $\{\circ, 1\}$, ℓ_i the length of $\{i, i+1\}$ for $i = 1, \ldots, (h-1)$, and ℓ'_j the branch length of $\{j, j'\}$ for $j = 1, \ldots, (h-1)$. Due to Corollary 4.3, we have the following result.

Corollary 4.5. *The Caterpillar tree of height h is trace-balanced if and only if $\ell'_j = \sum_{k=j}^{h-1} (h-k) \cdot \ell_k$, for $j = 1, \ldots, h-1$. In this case, the eigenvalues of the associated covariance matrix S are as follows, repeated according to their multiplicity: $\ell'_0 \geq \ell'_1 \geq \cdots \geq \ell'_{h-1}$, where $\ell'_0 := \sum_{k=0}^{h-1} (h-k) \cdot \ell_k$.*

We finish this section with a definition and result that characterizes the possible spectrums of covariance matrices of trace-balanced trees. Because the result's proof is constructive, it can be used to form strictly ultrametric matrices with the desired spectrum and multiplicities.

Definition 4.6. *In a tree T, a function $f : I \to [0, \infty)$ is called decreasing when, for all distinct $u, v \in I$, if u is an ancestor of v then $f(u) \geq f(v)$. In addition, f is called strictly positive at the fringe when $f(u) > 0$ whenever u is a parent of a leaf.*

Corollary 4.7. *In a trace-balanced ORB-tree T the function $v \longrightarrow \ell^*(v, i)$, with $v \in I$ and any $i \in L(v)$, is decreasing, and strictly positive at the fringe. Conversely, given any ORB-tree topology T and decreasing function $f : V \to [0, \infty)$ that is strictly positive at the fringe, there is a branch length function $\ell : E \to [0, \infty)$ such that $\sigma(S) = f(I)$. Furthermore, the multiplicity of $\lambda \in \sigma(S)$ is $|f^{-1}(\{\lambda\})|$.*

Proof. From the definitions of trace-balanced and ORB-tree, it is immediate that the transformation $v \in I \longrightarrow \ell^*(v, i)$, with $i \in L(v)$, is well-defined and strictly positive at the fringe of T. Also, f is decreasing because if u is an ancestor of v then, for each $i \in L(v)$: $\ell^*(u, i) = \ell^*(u, v) + \ell^*(v, i) \geq \ell^*(v, i)$. This shows the first statement in the corollary.

For the second statement consider an ORB-tree topology $T = (V, E)$ and function $f : I \to [0, \infty)$ that is both decreasing and strictly positive at the fringe. Due to Corollary 4.3, it suffices to show that there is a branch length function $\ell : E \to [0, +\infty)$ such that $f(v) = \ell^*(v, i)$, for all $i \in L(v)$. To do so, let $e = \{u, v\} \in E$ be so $\text{depth}(u) < \text{depth}(v)$. Define

$$\ell(e) := \begin{cases} f(u) & , v \in L; \\ \frac{f(u) - f(v)}{|L(e)|} & , v \in I. \end{cases} \tag{4.1}$$

Observe that if $v \in L$ then $|L(e)| = 1$ so $f(u) = \ell^*(e)$. Further, $\ell(e) > 0$ because f is strictly positive at the fringe of T. Instead, if $v \in I$ then $f(u) = f(v) + \ell(e) \cdot |L(e)| = f(v) + \ell^*(e)$, and $\ell(e) \geq 0$ because f is decreasing. In particular, if we extend the domain of f to all

of V defining $f(v) := 0$ for $v \in L$ then, for all $e = \{u, v\} \in E$ such that $\text{depth}(u) < \text{depth}(v)$: $f(u) = f(v) + \ell^*(e)$. From this, a simple inductive argument on the difference $d := \text{depth}(v) - \text{depth}(u) > 0$ shows that $f(u) - f(v) = \ell^*(u, v)$; implying that $f(u) = \ell^*(u, i)$, for all $i \in L(u)$, as claimed. ■

5. Spectrum Approximation in Roughly Trace-balanced ORB-trees. We know that the Haar-like wavelet associated with an internal node of an ORB-tree is an eigenvector of its covariance matrix if and only if the node is trace-balanced (Lemma 4.2). On the other hand, the Haar-like matrix of an ORB-tree can sometimes sparsify its covariance matrix significantly (Theorem 3.4). Together, these two facts suggest that the diagonal entry in $\Phi' S \Phi$ associated with an "approximately" trace-balanced internal node should be near the spectrum of S. In this section, we formalize this intuition quantifying what it is required for an internal node to be approximately trance-balanced.

In what follows, for a given function $x : L \to \mathbb{R}$ and non-empty $J \subset L$, we define the average value and variance of x over J as the quantities

$$\text{avg}(x; J) := \frac{1}{|J|} \sum_{j \in J} x(j);$$

$$\text{var}(x; J) := \frac{1}{|J|} \sum_{j \in J} \big(x(j) - \text{avg}(x; J)\big)^2.$$

In addition, for each $v \in V$, let $\text{parent}(v)$ denote the parent of node v in T. Define

$$\rho_v := \frac{|L(v)|}{|L\big(\text{parent}(v)\big)|}.$$

The following result aids in formalizing the intuition mentioned earlier.

Lemma 5.1. *If A is a symmetric matrix of dimensions $n \times n$ then, for all $\lambda \in \mathbb{R}$:*

$$\textit{distance}\big(\lambda, \sigma(A)\big) \leq \min_{x \in \mathbb{R}^n :\, \|x\|_2 = 1} \|(A - \lambda)x\|_2.$$

Proof. Since A is symmetric, all its eigenvalues are real and $\mathbb{R}^n$ has an orthonormal basis of eigenvectors $v_1, \ldots, v_n$. Say $Av_i = \lambda_i v_i$. In particular, if $\lambda \in R$ and $x \in \mathbb{R}^n$ then

$$(A - \lambda)x = \sum_{i=1}^{n} (\lambda_i - \lambda)\langle x, v_i \rangle v_i.$$

So, if $\|x\|_2 = 1$ then

$$\begin{aligned} \|(A - \lambda)x\|_2 &= \sqrt{\sum_{i=1}^{n} (\lambda_i - \lambda)^2 \langle x, v_i \rangle^2} \\ &\geq \min_{\lambda' \in \sigma(A)} |\lambda' - \lambda| \cdot \sqrt{\sum_{i=1}^{n} \langle x, v_i \rangle^2} = \text{distance}\big(\lambda, \sigma(A)\big), \end{aligned}$$

which shows the lemma. ■

Next, we provide a sufficient condition for λ_v, with $v \in I$, to be a good approximation of an eigenvalue of S. We also quantify explicitly the cosine between $S\varphi_v$ and $\lambda_v\varphi_v$ to assess how close φ_v is to be an eigenvector of S.

In the following result we use the notation: $\neg 0 = 1$ and $\neg 1 = 0$.

Theorem 5.2. *If $v \in I$ then $\lambda_v = \rho_{v1} \cdot \overline{\ell^*\big(L(v0), v\big)} + \rho_{v0} \cdot \overline{\ell^*\big(L(v1), v\big)}$, and*

$$\begin{aligned} &\mathit{distance}\big(\lambda_v, \sigma(S)\big) \\ &\quad \le \|(S-\lambda_v)\varphi_v\|_2 \\ &\quad = \sqrt{\rho_{v0} \cdot \rho_{v1} \cdot \Big\{\overline{\ell^*(L(v1), v)} - \overline{\ell^*(L(v0), v)}\Big\}^2 + \sum_{\alpha \in \{0,1\}} \rho_{v\alpha} \cdot \mathit{var}\big(\ell^*(L, v); L(v\neg\alpha)\big)}. \end{aligned}$$

Furthermore,

$$\cos(S\varphi_v, \lambda_v\varphi_v) = \frac{1}{\sqrt{1 + \Big\{\frac{\|(S-\lambda_v)\varphi_v\|_2}{\lambda_v}\Big\}^2}}.$$

Proof. Fix $v \in I$. To make the λ_v more explicit, observe that if $x : L \to \mathbb{R}$ is a function (or vector) then

$$\sum_{i \in L} \varphi_v^2(i) \cdot x(i) = \rho_{v1} \cdot \mathrm{avg}\big(x; L(v0)\big) + \rho_{v0} \cdot \mathrm{avg}\big(x; L(v1)\big). \tag{5.1}$$

In particular, due to Theorem 2.3:

$$\lambda_v = \varphi_v' S\varphi_v = \sum_{i \in L(v)} \varphi_v^2(i) \cdot \ell^*(v, i) = \rho_{v1} \cdot \overline{\ell^*\big(L(v0), v\big)} + \rho_{v0} \cdot \overline{\ell^*\big(L(v1), v\big)}, \tag{5.2}$$

which shows the first identity in the theorem.

On the other hand, Lemma 5.1 implies that

$$\mathrm{distance}(\lambda_v, \sigma(S)) \le \|(S-\lambda_v)\varphi_v\|_2.$$

But, from Theorem 2.3, we also have for $i \in L$ that $(S\varphi_v - \lambda_v\varphi_v)(i) = \varphi_v(i) \cdot \big(\ell^*(v, i) - \lambda_v\big)$. As a result

$$\begin{aligned} \|(S-\lambda_v)\varphi_v\|_2^2 &= \sum_{i \in L(v)} \varphi_v^2(i) \cdot \big(\ell^*(v, i) - \lambda_v\big)^2 \\ &= \frac{\rho_{v1}}{|L(v0)|} \sum_{i \in L(v0)} (\ell^*(i, v) - \lambda_v)^2 + \frac{\rho_{v0}}{|L(v1)|} \sum_{i \in L(v1)} (\ell^*(i, v) - \lambda_v)^2, \end{aligned} \tag{5.3}$$

where for the last identity we have used the equation (5.1). To complete the proof of the theorem note that $(\rho_{v0}+\rho_{v1})=1$. In particular, from the identity in equation (5.2), we may rewrite

$$\begin{aligned}\sum_{i\in L(v0)}(\ell^*(i,v)-\lambda_v)^2 &= \sum_{i\in L(v0)}\Big(\rho_{v0}\Big\{\overline{\ell^*(L(v1),v)}-\overline{\ell^*(L(v0),v)}\Big\}+\ell^*(i,v)-\overline{\ell^*\big(L(v0),v\big)}\Big)^2\\ &= |L(v0)|\,\rho_{v0}^2\left\{\overline{\ell^*(L(v1),v)}-\overline{\ell^*(L(v0),v)}\right\}^2\\ &\qquad+\sum_{i\in L(v0)}\Big(\ell^*(i,v)-\overline{\ell^*\big(L(v0),v\big)}\Big)^2.\end{aligned}$$

Namely

$$\frac{1}{|L(v0)|}\sum_{i\in L(v0)}(\ell^*(i,v)-\lambda_v)^2=\rho_{v0}^2\Big\{\overline{\ell^*(L(v1),v)}-\overline{\ell^*(L(v0),v)}\Big\}^2+\mathrm{var}\big(\ell^*(L,v);L(v0)\big).$$

Similarly,

$$\frac{1}{|L(v1)|}\sum_{i\in L(v1)}(\ell^*(i,v)-\lambda_v)^2=\rho_{v1}^2\Big\{\overline{\ell^*(L(v1),v)}-\overline{\ell^*(L(v0),v)}\Big\}^2+\mathrm{var}\big(\ell^*(L,v);L(v1)\big).$$

The second identity in the theorem is now a direct consequence of (5.3) and the last two identities.

Finally, again due to Theorem 2.3, we find that

$$\cos(S\varphi_v,\lambda_v\varphi_v)=\frac{\varphi_v'S\varphi_v}{\|S\varphi_v\|_2}=\frac{\lambda_v}{\sqrt{\varphi_v'\mathrm{diag}(\ell^*(L,v))^2\varphi_v}}.$$

But, similarly as we argued before

$$\varphi_v'\mathrm{diag}(\ell^*(L,v))^2\varphi_v=\sum_{i\in L}\varphi_v^2(i)\ell^*(i,v)^2=\lambda_v^2+\sum_{i\in L}\varphi_v^2(i)(\ell^*(i,v)-\lambda_v)^2=\lambda_v^2+\|(S-\lambda_v)\varphi_v\|_2^2,$$

which implies that

$$\cos(S\varphi_v,\lambda_v\varphi_v)=\frac{1}{\sqrt{1+\frac{\sum_{i\in L}\varphi_v^2(i)(\ell^*(i,v)-\lambda_v)^2}{\lambda_v^2}}}.$$

The third identity in the theorem follows now from equation (5.3). ■

It follows from the theorem that $S\varphi=\lambda_v\varphi_v$ if and only if $\overline{\ell^*(L(v1),v)}=\overline{\ell^*(L(v0),v)}$ and $\mathrm{var}\big(\ell^*(L,v);L(v0)\big)=\mathrm{var}\big(\ell^*(L,v);L(v1)\big)=0$. But these conditions are precisely equivalent to having the ORB-tree trace-balanced at v. Lemma 4.2 may be therefore regarded a corollary of Theorem 5.2.

We also emphasize that the first upper-bound for distance$\big(\lambda_v,\sigma(S)\big)$ may be computed efficiently using Theorem 2.3. Nevertheless, its alternative expression gives a way to quantify how approximately trace-balanced an ORB-tree is.

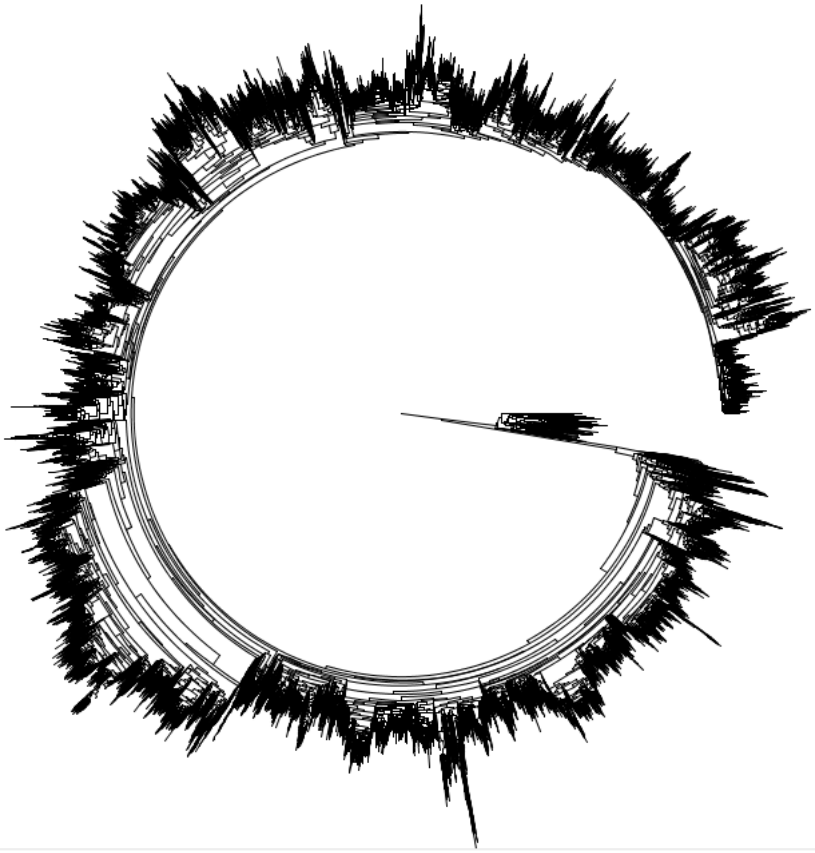


Figure 5: **Circular layout and basic statistics of the 97% Greengenes tree.** The tree has 99,322 leaves, 198,642 edges, and height (i.e. maximal leaf depth) 107. The average branch length is 1.42×10^{-2} units, with lengths varying between 1.5×10^{-4} and 1.0.

6. New Insights into the Microbial Tree of Life. Many methods in microbiology rely on a phylogenetic tree relating microorganisms. At the microbial level, however, the notions of genus or species are ill-defined because microorganisms do not interbreed. So microbes' taxonomy and phylogeny are often based on the so-called 16S ribosomal RNA (16S rRNA) gene. This gene is present in all known single cell organisms and can therefore be used as a phylogenetic marker. An operational taxonomic unit (OTU) is a cluster of these markers defined by some least level of DNA sequence similarity among its (highly) conserved regions.

Greengenes is a standarized database based on the 16S rRNA marker. It has been a standard reference in microbial studies, particularly metagenomics, and is the default option in QIITA [19]—a widely used open-source management platform for microbial analyses. Greengenes phylogenetic trees are built using FastTree [42] and their associated taxonomies are assigned using tax2tree [38]. Trees are typically stored in the newick format [14], which encodes their topology and branch lengths. Trees can be visualized from their newick format using software such as FigTree [1] or ETE Toolkit [23].

Figure 5 displays the Greengenes tree when OTUs are thresholded at a 97% sequence similarity—the average similarity of macro-organisms' DNA in the same species. The tree represents the inferred evolutionary history of modern day microorganisms from common ancestors. Its root is at the center of the circular layout, and each OTU is associated with a single leaf in the tree and vice versa. Branch lengths are a proxy of evolutionary time such as the estimated expected number of mutations per nucleotide site [22], and interior nodes (also called splits) are inferred speciation events that have led to the present-day microorganisms in the database.

A fundamental problem in microbiology is to link environmental factors (such as acidity, light, nutrients, salinity, temperature, etc) with microbial composition. A valuable tool for this has been the concept of β-diversity (i.e., a measure of differences between microbial composition across different environments). Early approaches to β-diversity include the Bray-Curtis dissimilarity [6] or the Jaccard distance [25], which ignored the evolutionary relationships between microorganisms found in different environments. Nonetheless, one would expect microbes with a shared evolutionary history to similarly thrive or struggle in similar environments. Phylogenetic informed metrics were introduced precisely to convey this idea. These metrics require a phylogenetic tree relating the microorganisms observed in samples from all the environments under study. Among other more recent phylogentic trees such as SILVA [44] and WoL [53], Greengenes has been a common choice of representative phylogeny. We base our application on the latter—though our discussion applies to any phylogeny.

Double Principal Coordinate Analysis (DPCoA) [41] is a phylogenetically informed β-diversity metric between pairs of microbial environments, which provides similar insights [17] to other more recent though more widely used distances such as unweighted and weighted UniFrac [33].

Let T be the ORB-tree associated with a phylogenetic tree (e.g. the 97% Greengenes tree), and S the covariance matrix of T. In the context of phylogenetic informed metrics, environments are represented as probability mass functions over the OTUs (i.e. leaves). We denote those functions with lower-case letters such as a and b, and interpret them as probability models over L. In particular, $a : L \to [0, +\infty)$ satisfies that $\sum_{x\in L} a(x) = 1$ and, for each $e \in E$, $a(e) = \sum_{x\in e} a(x)$. With this convention, the DPCoA distance between two environments a and b is defined as [41, 17]:

$$d(a,b) := \left\{ \sum_{e\in E} \ell(e) \left(a(e) - b(e)\right)^2 \right\}^{1/2} = \sqrt{(a-b)'S(a-b)}. \tag{6.1}$$

In particular, since S is positive definite, DPCoA corresponds to a Mahalanobis distance [35]; implying that $d(\cdot,\cdot)$ is a metric—in the mathematical sense—in $\mathbb{R}^{|L|}$.

The weighted and unweighted UniFrac distances are instead defined as follows [33]:

$$d_w(a,b) := \sum_{e\in E} \ell(e)\, |a(e) - b(e)|;$$
$$d_u(a,b) := \frac{\sum_{e\in E} \ell(e) \left| [\![a(e) > 0]\!] - [\![b(e) > 0]\!] \right|}{\sum_{e\in E} \ell(e)}.$$

Both versions of UniFrac are known to satisfy the triangular inequality [34]. DPCoA is also more robust to unbiased noise but more sensitive to outliers than UniFrac [17].

Regardless of the metric of choice, the standard approach to linking environmental factors with microbial composition goes roughly as follows [31]. First, environmental samples are collected, and each environment is represented by its OTU composition on the leaves of the phylogeny of reference. Then, the pairwise distance matrix between the environments is computed, and the environments are embedded into a low-dimensional Euclidean space using standard techniques such as multidimensional scaling (MDS) [5]. Despite the noisy and high-dimensional nature of microbial datasets [47, 32, 21], this approach has been remarkably reliable for the ordination [28] of microbial environments in as little as 1-2 dimensions, and for correlating environmental factors with microorganisms. However, this approach does not usually explain correlations, which need to be justified by other means.

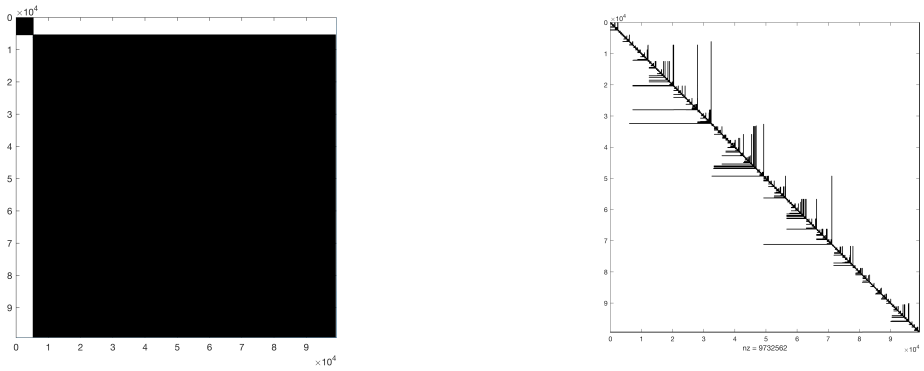
In what remains of this section, we apply our methods to the Greengenes phylogeny. First, we demonstrate significant sparsification of the associated covariance matrix after applying the Haar-like wavelet transform. Then, we motivate a new wavelet-based phylogenetic β-diversity metric corresponding to a multiscale analysis of the phylogenetic tree. Finally, applying the new metric to a previously studied dataset shows that the wavelet-based metric can give novel insights into the relationship between environmental factors and OTU composition.

6.1. Greengenes Phylogenetic Covariance Matrix Sparsification. The 97% Greengenes tree has about 100,000 leaves. We can think of it as an ORB-tree by adding an external root $\circ$ and connecting it to the original root with a branch of length 0. (Alternatively, we could think of the Greengenes tree as two ORB-trees with their roots merged.) We denote the resulting ORB-tree as T.

The identity in equation (1.3) implies that the covariance matrix S of T is a 2×2 block diagonal matrix, with each block corresponding to an ORB-subtree. Nevertheless, approximately 94% of the almost 10 billion entries in S are non-zero because one of the ORB-subtrees (corresponding to the Archaea domain) is much smaller than the other—see Figure 6(a). This makes storing the covariance matrix of T challenging. Further, basic computational tasks such as finding the spectrum and inverting S for parameter estimation in phylogenetic comparative methods [26, 16] is infeasible because this large matrix is almost fully dense. We may use, however, the Haar-like matrix Φ associated with T to sparsify S. From Theorem 3.4, we can guarantee that $\zeta \geq 0.9989$, i.e. at least 99.89% of the entries in the similar matrix $\Phi' S \Phi$ vanish. This significant compression of the matrix S can be appreciated in Figure 6(b).

We implemented Algorithm 3.1 using the sparse matrix packages from SciPy [51] to compute $\Phi' S \Phi$. As proof-of-principle we used this compressed representation of S to compute its largest 500 eigenvalues to machine precision using SciPy's implementation of the Lanczos algorithm. As seen in Figure 7, the eigenvalues of S decay rapidly. In fact, we found that $\lambda_1(S) \sim 1.27 \times 10^5$, $\lambda_2(S) \sim 4.75 \times 10^3$, and $\mathrm{trace}(S) \sim 1.65 \times 10^5$, so the top and top-two eigenvalues account for approximately 77% and 80% of the trace of S, respectively.

As seen in Figure 7 also, the sorted diagonal entries in $\Phi' S \Phi$ (i.e. the quantities λ_v, with $v \in I$) approximate with ample accuracy the spectrum of S. For instance, $\max_{v \in I} \lambda_v$ underestimates $\lambda_1(S)$ with only about a 0.06% relative error. Anticipating this overall accuracy from T alone remains an open problem as neither our mathematical results, particularly



(a) (b)

Figure 6: **Heatmaps of matrices associated with the 97% Greengenes.** Black (white) pixels denote non-zero (vanishing) entries. (a) Phylogenetic covariance matrix S of the 97% Greengenes tree. S has dimensions $\sim 10^5 \times 10^5$. (b) Sparsified matrix $\Phi' S \Phi$.

Theorem 5.2, nor more general ones such as the Gershgorin's circle theorem, Sylvester's determinant theorem, and bounds found in [49, 7, 24] have been able to explain it.

6.2. A Wavelet Based Phylogenetic β-diversity Metric. Let T be the ORB-tree associated with a phylogenetic tree. Recall that φ_v, with $v \in I$, is supported on $L(v)$, and together these functions form an orthonormal base of $\mathbb{R}^{|L|}$. In particular, just as wavelets are traditionally used to localize signals at different scales, we may use the Haar-like basis of T to localize environmental OTU distributions on subsets of leaves defined by splits in the tree. This is particularly appealing from a biological standpoint. Indeed, the opposite signs of φ_v on the leaves of the left and right subtrees dangling from v may be interpreted as a speciation event that conferred more fitness to present-day microorganisms descending from one of the subtrees than the other. We propose the following definition to convey these features into a phylogenetic β-diversity metric.

Recall that $\lambda_v = (\Phi' S \Phi)(v, v) > 0$, for each $v \in I$. Further, for a given environment a (i.e., OTU distribution over L), $\Phi' a$ is the projection of a onto the Haar-like basis of the reference tree.

Definition 6.1. *The Haar-like distance between two environments a and b is the quantity*

$$d_h(a, b) := \sqrt{\sum_{v \in I} \lambda_v \, \Delta_v^2}, \text{ where } \Delta = (\Delta_v)_{v \in I} := \Phi'(a - b).$$

The specifics of this distance can be motivated as follows. On one hand, the terms Δ_v^2, with $v \in I$, convey the idea that d_h regards two environments similar (different) when their OTU compositions project similarly (differently) onto the Haar-like basis of the reference tree.

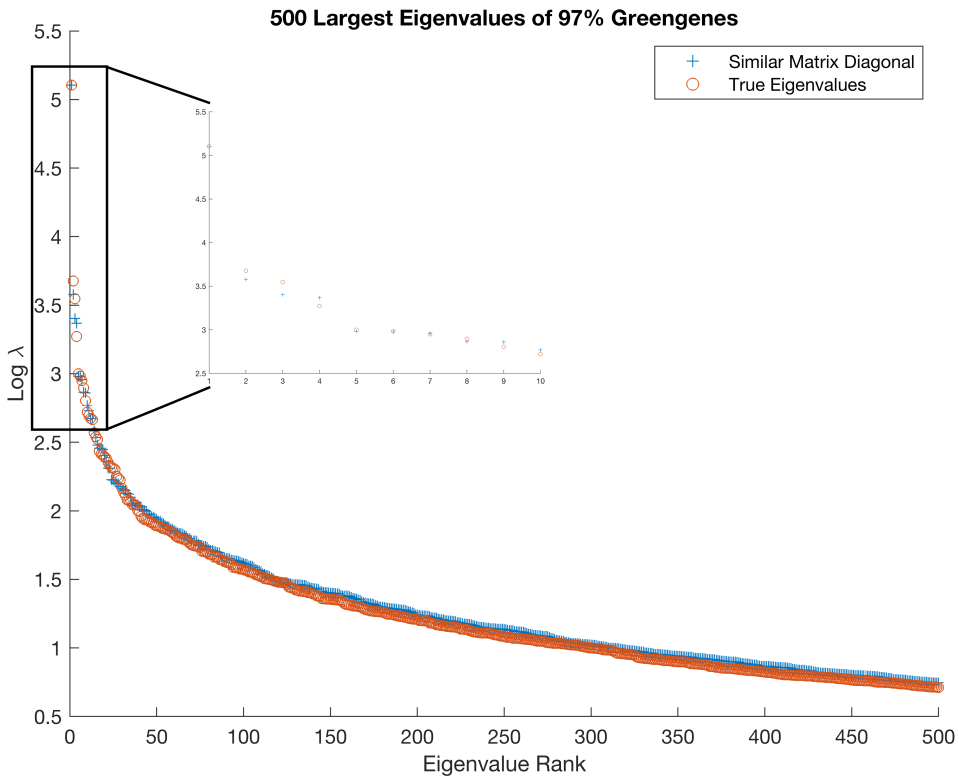


Figure 7: **Spectrum decay of 97% Greengenes tree covariance matrix and corresponding approximation using Haar-like wavelets.** Only the 500 most dominant eigenvalues of S are plotted as a function of their rank. Logarithms are in base-10.

On the other hand, the weights λ_v, with $v \in I$, are motivated by the success of DPCoA in various biological investigations. To explain this, consider the matrices $D := \mathrm{diag}(\lambda_v : v \in I)$ and $E := \Phi' S \Phi - D$. Observe that $d_h(a,b) = \sqrt{\Delta' D \Delta}$; in particular, d_h is a metric in $\mathbb{R}^{|L|}$ because D is positive definite, and $d(a,b) = \sqrt{\Delta' D \Delta + \Delta' E \Delta}$. In large phylogenetic trees, however, we expect E to be mostly filled with zeroes due to Corollary 3.8—which suggests considering d_h as an alternative metric to DPCoA.

We have mentioned before that while traditional phylogenetic metrics (in conjunction with embedding techniques) have been remarkably successful at correlating microbial composition with environmental factors, these correlations cannot usually be explained from the metrics alone. The wavelet nature of the Haar-like distance has, however, the potential to explain said correlations. Indeed, the biological interpretation of the Haar-like basis conveyed by their sign flip suggests that if $\lambda_v \Delta_v^2$ is comparatively large (small) for some $v \in I$, then the speciation event associated with v has a significant (little) influence differentiating the OTU distributions between two environments a and b. (There may be discrepancies between

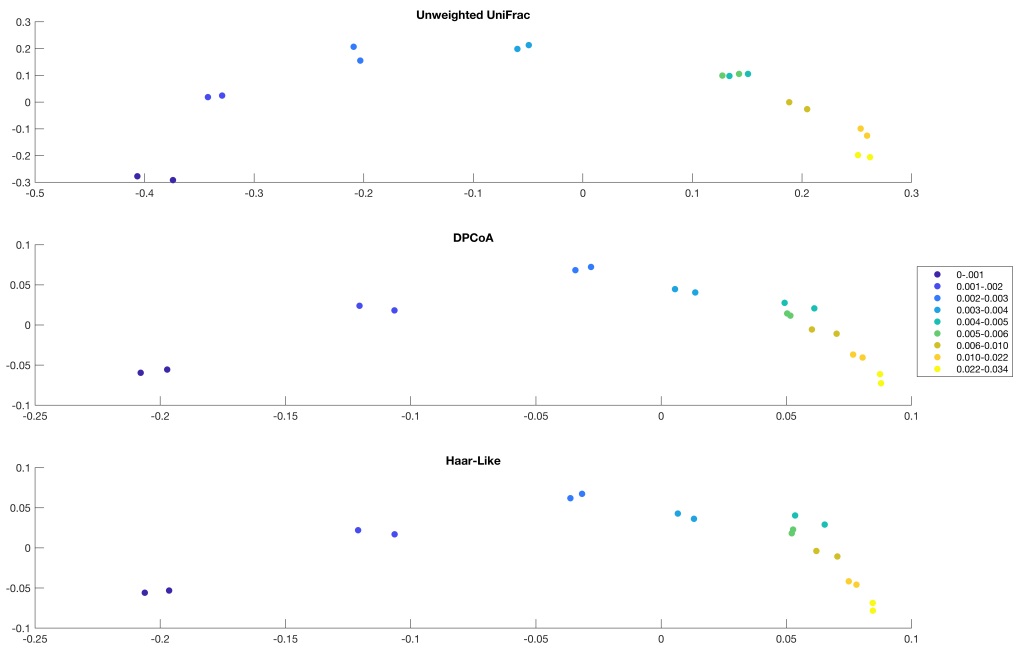


Figure 8: **2-D MDS embeddings of samples from Guerrero Negro w.r.t. different metrics.** The embeddings are based on unweighted UniFrac (top), DPCoA (middle), and the Haar-like distance (bottom). Depth varies from 0-0.034 meters.

taxonomy and splits in a tree. In particular, while the aforementioned correlations may be explained by a phylogeny, they are not necessarily explained by a taxonomic classification.)

6.3. Haar-like Distances of the Guerrero Negro microbial mat. A microbial mat is a biofilm of layered groups of microorganisms with coupled biochemistries. Their rich biodiversity, combined with the environmental gradients of light, oxygen, etc., offer an ideal setting to test phylogenetic β-diversity metrics.

To demonstrate the insight gained from Haar-like distance we applied it to a 16S rRNA data set of 18 soil samples (obtained from QIITA [19]) at different depths of the Guerrero Negro microbial mat [30], located in Baja California Sur, Mexico. We used the 97% Greengenes as the reference phylogeny.

Earlier work [30] based on unweighted UniFrac showed a gradient of microbial composition in the mat with respect to depth—see top plot in Figure 8. (For a discussion regarding the "horseshoe" shape in the plot see [30, 39, 13].) As seen on the bottom plot of the same figure, we can practically reproduce this gradient using the Haar-like distance instead. Furthermore, as seen on the bottom two plots, the DPCoA and Haar-like distance produce nearly indistinguishable embeddings.

While the three phylogenetic β-diversity metrics imply that soil depth drives a measur-

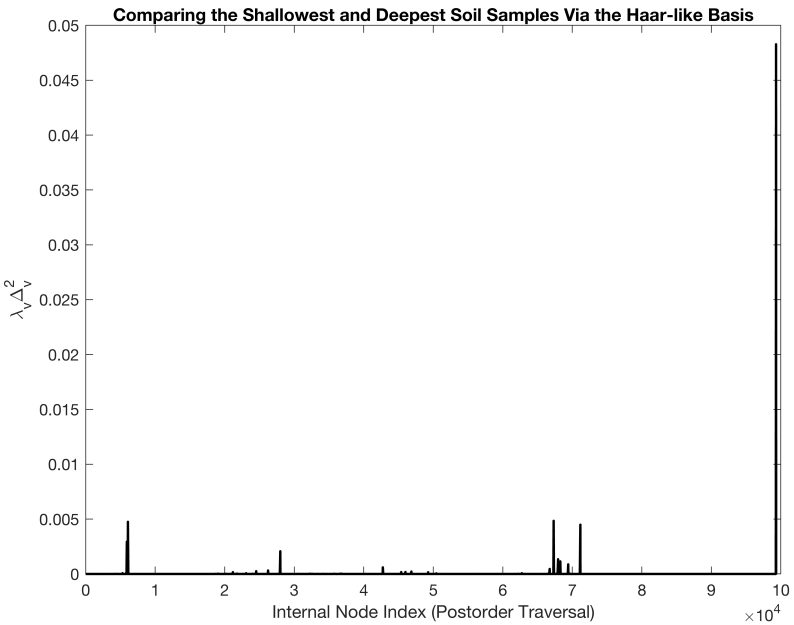


Figure 9: **Plot of $\lambda_v \, \Delta_v^2$, with $v \in I$, to measure the Haar-like distance between the shallowest and deepest sample in the Guerrero Negro dataset.** The average non-zero value of $\lambda_v \, \Delta_v^2$ is $\sim 2.09 \times 10^{-5}$. The standard deviation of these values is $\sim 7.55 \times 10^{-4}$.

able change in OTU composition, we can go a step further with the Haar-like distance and determine which splits are responsible for this trend and quantify their importance. We demonstrate this by comparing the two extremes in the dataset: let a and b be the OTU compositions of the shallowest and deepest environment, respectively. Define $\Delta = \Phi'(b - a)$. Following the logic described in Section 6.2, we computed $v \in I \longrightarrow \lambda_v \, \Delta_v^2$, indexing interior nodes according to a postorder traversal of the 97% Greengenes tree. As seen in Figure 9, the 3 largest values are statistically significant. These are associated with the Haar-like wavelets φ_{99311} ($\lambda_v \Delta_v^2$-value $\sim 4.84 \times 10^{-2}$), φ_{67317} ($\lambda_v \Delta_v^2$-value $\sim 4.84 \times 10^{-3}$), and φ_{6079} ($\lambda_v \Delta_v^2$-value $\sim 4.75 \times 10^{-3}$). These correspond to splits at depths 10, 34, and 18 of the 97% Greengenes tree, respectively.

Notably, the split associated with φ_{99311} corresponds to the largest $\lambda_v \Delta_v^2$-value. According to the associated taxonomic classification, the (say) left descendants of this split associated correspond to the phylum level classification of Cyanobacteria. This is consistent with the conclusion in [30], which correlated Cyanobacteria abundance changes with soil depth and explained the correlation by their ability to photosynthesize.

The other two wavelets provide novel insight into other important OTU composition differences driving the observed soil depth gradient in the Guerrero Negro dataset. Indeed, while the descendants of the split associated with φ_{67317} do not exhaust a taxonomic classification, all leaves under that split are classified as Anaerolineae. This differentiation between the

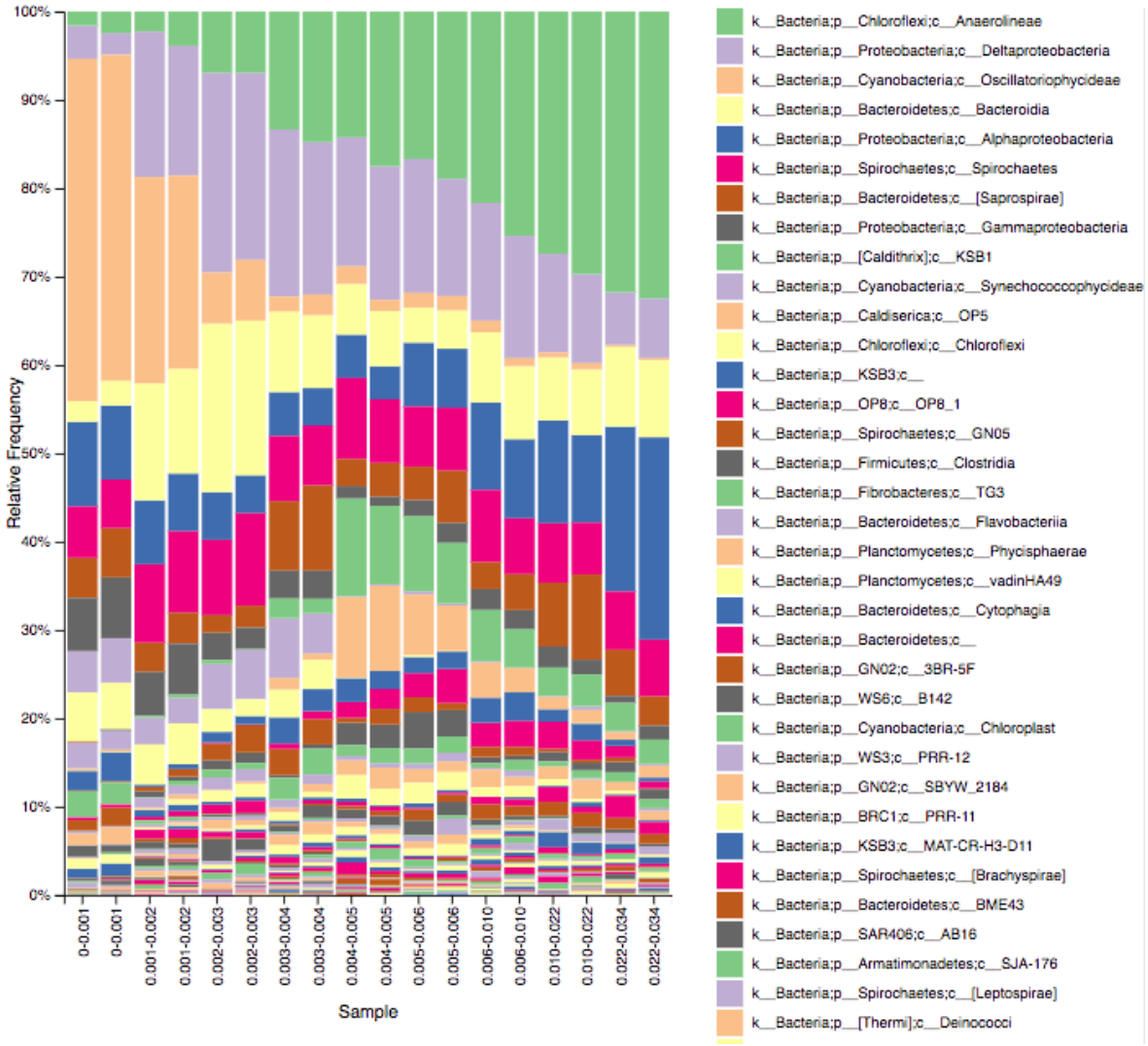


Figure 10: **OTU compositions plotted at class level taxonomy in QIITA.** Only the 35 most abundant OTU class names are displayed in the legend.

shallowest and deepest sample may be due to Anaerolineae's role as a anaerobic digester [52]. This claim is reinforced by the top green bars in the OTU composition plot in Figure 10, which show a rapid increase of Anaerolineae with depth.

Finally, the split associated with φ_{6079} subdivides the Cyanobacteria phylum into further classes, including Oscillatoriophycideae which, according to Figure 10, is the third most abundant class of the Guerrero Negro dataset. The relevance of this split to differentiate shallow from deep samples may be explained by Oscillatoriophycideae's photoautotrophic capability [48]. Again, this is supported by Figure 10, which shows a sharp decrease in Oscillatoriophycideae with respect to soil depth.

Our analysis of the Guerrero Negro mat shows that the Haar-like distance may be a valid alternative to other more common phylogenetic β-diversity metrics, primarily because it provides a systematic method for detecting statistically significant speciation events (and corresponding levels of OTU classification) that can link OTU composition with environmental factor gradients.

Appendix A. Orthonormality of Haar-like bases. The statement that the Haar-like basis $\{\varphi_v\}_{v\in I}$ associated with an ORB-tree is orthonormal is based on the concept of multiresolution analysis of Euclidean spaces in [18]. Here we justify this fact by first principles.

Let $u, v \in I$. If $u = v$ then

$$\langle \varphi_u, \varphi_v \rangle = \frac{|L(u1)| + |L(u0)|}{|L(u)|} = 1.$$

Instead, there are two possibilities when $u \neq v$. If $L(u) \cap L(v) = \emptyset$ then $\langle \varphi_v, \varphi_u \rangle = 0$ because φ_v and φ_u have disjoint supports. Otherwise, if $L(u) \cap L(v) \neq \emptyset$ then Lemma 2.4 let us assume without any loss of generality that u is an ancestor of v. In particular, $L(v) \subset L(u)$ but also φ_u remains constant over $L(v)$. Therefore, for any given $x \in L(v)$:

$$\begin{aligned}\langle \varphi_u, \varphi_v \rangle &= \varphi_u(x) \cdot \sum_{y \in L(v)} \varphi_v(y) \\ &= \varphi_u(x) \cdot \left\{ \sqrt{\frac{|L(v1)| \cdot |L(v0)|}{|L(v)|}} - \sqrt{\frac{L(v0)| \cdot |L(v1)|}{|L(v)|}} \right\} = 0.\end{aligned}$$

Acknowledgments. This work has been partially funded by the NSF grant No. 1836914.

REFERENCES

[1] A Rambaut, *Figtree v1.3.1. institute of evolutionary biology, university of edinburgh, edinburgh*, 2010.

[2] D. J. Aldous, *Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today*, Statistical Science, (2001), pp. 23–34.

[3] M. G. Blum and O. François, *On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited*, Mathematical Biosciences, 195 (2005), p. 141–153.

[4] M. G. B. Blum and O. François, *Which random processes describe the Tree of Life? A large-scale study of phylogenetic tree imbalance*, Syst. Bio, 55 (2006), pp. 685–691.

[5] I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, 2nd ed., 2005.

[6] J. Bray and J. Curtis, *An ordination of the upland forest communities of southern wisconsin*, Ecologial Monographs, 27 (1957), pp. 325–349.

[7] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numerische Mathematik, 31 (1978), p. 31–48, https://doi.org/10.1007/bf01396012.

[8] D. Capocaccia, M. Cassandro, and P. Picco, *On the existence of thermodynamics for the generalized random energy model*, Journal of Statistical Physics, 46 (1987), p. 493–505, https://doi.org/10.1007/bf01013370.

[9] L. L. Cavalli-Sforza and A. W. Edwards, *Phylogenetic analysis: Models and estimation procedures*, Evolution, 21 (1967), p. 550, https://doi.org/10.2307/2406616.

[10] T. M. Coronado, A. Mir, F. Rosselló, and L. Rotger, *On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index*, BMC Bioinformatics, 21 (2020).

[11] C. Dellacherie, S. Martinez, and S. Martín, *Inverse M-Matrices and Ultrametric Matrices*, vol. 2118 of Lecture Notes in Mathematics, Springer, 2014.

[12] C. Dellacherie, S. Martínez, and J. San Martín, *Ultrametric matrices and induced markov chains*, Advances in Applied Mathematics, 17 (1996), p. 169–183, https://doi.org/10.1006/aama.1996.0009.

[13] P. Diaconis, S. Goel, and S. Holmes, *Horseshoes in multidimensional scaling and local kernel methods*, The Annals of Applied Statistics, 2 (2008), https://doi.org/10.1214/08-aoas165.

[14] J. Felsenstein, J. Archie, W. Day, W. Maddison, C. Meacham, F. Rohlf, and D. Swofford, *The newick tree format*, 1986.

[15] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge Univ. Press, 2013.

[16] R. P. Freckleton, *Fast likelihood calculations for comparative analyses*, Methods in Ecology and Evolution, 3 (2012), p. 940–947, https://doi.org/10.1111/j.2041-210x.2012.00220.x.

[17] J. Fukuyama, P. J. McMurdie, L. Dethlefsen, D. A. Relman, and S. Holmes, *Comparisons of distance methods for combining covariates and abundances in microbiome studies*, Biocomputing, (2012), pp. 213–224.

[18] M. Gavish, B. Nadler, and R. R. Coifman, *Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning*, in ICML, 2010, pp. 367–374, https://icml.cc/Conferences/2010/papers/137.pdf.

[19] A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, J. G. Sanders, J. Shorenstein, H. Holste, S. Petrus, A. Robbins-Pianka, C. J. Brislawn, M. Wang, J. R. Rideout, E. Bolyen, M. J. Dillon, G. Caporaso, P. C. Dorrestein, and R. Knight, *Qiita: rapid, web-enabled microbiome meta-analysis*, Nature Methods, 15 (2018), p. 796–798. https://qiita.ucsd.edu/.

[20] A. Graps, *An introduction to wavelets*, IEEE Computational Science and Engineering, 2 (1995), p. 50–61, https://doi.org/10.1109/99.388960.

[21] J. Hampton and M. E. Lladser, *Estimation of distribution overlap of urn models*, PloS ONE, 7 (2012), p. e42368.

[22] S. Y. Ho and S. Duchêne, *Molecular-clock methods for estimating evolutionary rates and timescales*, Molecular Ecology, 23 (2014), p. 5947–5965, https://doi.org/10.1111/mec.12953.

[23] J. Huerta-Cepas, F. Serra, and P. Bork, *Ete 3: Reconstruction, analysis, and visualization of phylogenomic data*, Molecular Biology and Evolution, 33 (2016), p. 1635–1638, https://doi.org/10.1093/molbev/msw046.

[24] I. C. F. Ipsen and B. Nadler, *Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices*, SIAM Journal on Matrix Analysis and Applications, 31 (2009), p. 40–53, https://doi.org/10.1137/070682745.

[25] P. Jaccard, *The distribution of the flora in the alpine zone.1*, New Phytologist, 11 (1912), pp. 37–50, https://doi.org/https://doi.org/10.1111/j.1469-8137.1912.tb05611.x, https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x, https://arxiv.org/abs/https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x.

[26] D.-C. Jhwueng, *On the covariance of phylogenetic quantitative trait evolution models and their matrix condition*, Communications in Statistics - Simulation and Computation, (2022), p. 1–20, https://doi.org/10.1080/03610918.2022.2037639.

[27] F. Ji, W. Tang, and W. P. Tay, *On the properties of gromov matrices and their applications in network inference*, IEEE Transactions on Signal Processing, 67 (2019), p. 2624–2638, https://doi.org/10.1109/tsp.2019.2908133.

[28] R. H. G. Jongman, *Data analysis in community and landscape ecology*, Cambridge University Press, 1995.

[29] M. C. King and N. A. Rosenberg, *A simple derivation of the mean of the Sackin index of tree balance under the uniform model on rooted binary labeled trees*, Mathematical Biosciences, 342 (2021), p. 108688.

[30] J. Kirk Harris, J. Gregory Caporaso, J. J. Walker, J. R. Spear, N. J. Gold, C. E. Robertson, P. Hugenholtz, J. Goodrich, D. McDonald, D. Knights, and et al., *Phylogenetic stratigraphy in the guerrero negro hypersaline microbial mat*, The ISME Journal, 7 (2012), p. 50–60, https://doi.org/10.1038/ismej.2012.79.

[31] M. E. Lladser and R. Knight, *Mathematical approaches for describing microbial populations: prac-*

tice and theory for extrapolation of rich environments, in The Human Microbiota: How Microbial Communities Affect Health and Disease, D. Fredricks, ed., Wiley-Blackwell, 2013.

[32] R. J. Lladser ME, Gouet R, *Extrapolation of urn models via poissonization: Accurate measurements of the microbial unknown*, PLoS ONE, 6 (2011), p. e21105.

[33] C. Lozupone and R. Knight, *UniFrac: a new phylogenetic method for comparing microbial communities*, Appl Environ Microbiol., 71 (2005), pp. 8228–35, https://doi.org/doi:10.1128/AEM.71.12.8228-8235.2005.

[34] C. Lozupone, M. Lladser, D. Knights, J. Stombaugh, and R. Knight, *UniFrac: an effective distance metric for microbial community comparison*, ISME J., 5 (2011), pp. 169–72.

[35] P. Mahalanobis, *On the generalised distance in statistics*, Proceedings of the National Institute of Sciences of India, 2 (1936), p. 49–55.

[36] S. G. Mallat, *A wavelet tour of signal processing the sparse way*, Elsevier /Academic Press, 2009.

[37] S. Martinez, G. Michon, and J. San Martín, *Inverse of strictly ultrametric matrices are of Stieltjes type*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 98–106.

[38] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz, *An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea*, The ISME Journal, 6 (2011), p. 610–618, https://doi.org/10.1038/ismej.2011.139.

[39] J. T. Morton, L. Toran, A. Edlund, J. L. Metcalf, C. Lauber, and R. Knight, *Uncovering the horseshoe effect in microbial analyses*, mSystems, 2 (2017), https://doi.org/10.1128/msystems.00166-16.

[40] R. Nabben and R. S. Varga, *A linear algebra proof that the inverse of a strictly ultrametric matrix is a strictly diagonally dominant stieltjes matrix*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 107–113.

[41] S. Pavoine, A.-B. Dufour, and D. Chessel, *From dissimilarities among species to dissimilarities among communities: A double principal coordinate analysis*, Journal of Theoretical Biology, 228 (2004), p. 523–537, https://doi.org/10.1016/j.jtbi.2004.02.014.

[42] M. N. Price, P. S. Dehal, and A. P. Arkin, *Fasttree 2 - approximately maximum-likelihood trees for large alignments*, PLoS ONE, 5 (2010), https://doi.org/10.1371/journal.pone.0009490.

[43] E. Purdom, *Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree*, The Annals of Applied Statistics, 5 (2011), https://doi.org/10.1214/10-aoas402.

[44] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, *The silva ribosomal rna gene database project: Improved data processing and web-based tools*, Nucleic Acids Research, 41 (2012), https://doi.org/10.1093/nar/gks1219.

[45] S. Saraçli, N. Doğan, and I. Doğan, *Comparison of hierarchical cluster analysis methods by cophenetic correlation*, Journal of Inequalities and Applications, 2013 (2013), https://doi.org/10.1186/1029-242x-2013-203.

[46] R. Sedgewick and P. Flajolet, *An introduction to the analysis of algorithms*, Addison-Wesley, 2013.

[47] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, *Microbial diversity in the deep sea and the underexplored "rare biosphere"*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 12115–12120, https://doi.org/10.1073/pnas.0605127103.

[48] R. Y. Stanier and G. Cohenbazire, *Phototropic Prokaryotes - Cyanobacteria*, Annual Review of Microbiology, 31 (1977), pp. 225–274, https://doi.org/DOI10.1146/annurev.mi.31.100177.001301.

[49] R. Thompson, *The behavior of eigenvalues and singular values under perturbations of restricted rank*, Linear Algebra and its Applications, 13 (1976), pp. 69–78, https://doi.org/https://doi.org/10.1016/0024-3795(76)90044-6, https://www.sciencedirect.com/science/article/pii/0024379576900446.

[50] R. S. Varga and R. Nabben, *On symmetric ultrametric matrices*, Numerical Linear Algebra, (1993), https://doi.org/10.1515/9783110857658.193.

[51] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *SciPy 1.0: Fun-*

damental Algorithms for Scientific Computing in Python, Nature Methods, 17 (2020), pp. 261–272, https://doi.org/10.1038/s41592-019-0686-2.

[52] Y. Xia, Y. Wang, Y. Wang, F. Y. Chin, and T. Zhang, *Cellular adhesiveness and cellulolytic capacity in anaerolineae revealed by omics-based genome interpretation*, Biotechnology for Biofuels, 9 (2016), https://doi.org/10.1186/s13068-016-0524-z.

[53] Q. Zhu, U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. McDonald, and et al., *Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea*, Nature Communications, 10 (2019), https://doi.org/10.1038/s41467-019-13443-4.

[54] A. P. Zubarev, *On stochastic generation of ultrametrics in high-dimensional Euclidean spaces*, p-Adic Numbers, Ultrametric Analysis, and Applications, 6 (2014), pp. 55–165.

[55] A. P. Zubarev, *On the ultrametric generated by random distribution of points in euclidean spaces of large dimensions with correlated coordinates*, Journal of Classification, 34 (2017), p. 366–383, https://doi.org/10.1007/s00357-017-9236-8.