

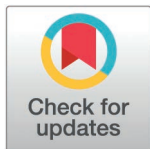
METHODS

Functional group classification using consensus clustering

Pablo Ubilla Pavez^{1,2}, Andrea Paz³, Daniel S. Maynard^{2*}

1 INRIA, Montpellier, France, **2** Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, **3** Département de Sciences Biologiques, Faculté des Arts des Sciences, Université de Montréal, Montréal, Canada

* daniel.maynard@ucl.ac.uk



OPEN ACCESS

Citation: Ubilla Pavez P, Paz A, Maynard DS (2026) Functional group classification using consensus clustering. PLoS Comput Biol 22(5): e1014278. <https://doi.org/10.1371/journal.pcbi.1014278>

Editor: Fuhai Li, Washington University in Saint Louis, UNITED STATES OF AMERICA

Received: October 13, 2025

Accepted: April 27, 2026

Published: May 13, 2026

Copyright: © 2026 Ubilla Pavez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The code is available at https://github.com/pabloubilla/tree_clustering/ and is archived via Zenodo with corresponding DOI:10.5281/zenodo.19222113. The data have been archived via Figshare with corresponding DOI:10.5522/04/31833526.

Funding: This work was partly funded by the Swiss National Science Foundation, Ambizione

Abstract

Functional diversity is a fundamental aspect of community structure and composition, reflecting diversity and redundancy in ecological niches, functional roles, and environmental responses among species within a community. Despite its growing importance for quantifying ecosystem-level biodiversity, existing functional diversity metrics remain difficult to calculate and interpret, hindering their adoption and application beyond the scientific realm. One potential solution to this problem is to categorize species into functional groups based on their traits, which provides a simple, intuitive categorization of functional diversity that allows for the application of traditional species-based metrics. The functional-group approach, however, has several challenges that have limited its adoption, namely, the difficulty in identifying robust functional clusters, which can vary substantially due to trait variability, measurement error, and trait correlation. Here, to address these challenges, we present a multi-step consensus clustering method that integrates trait uncertainty and correlation into the classification of species into functional groups. Our approach proceeds in four main steps: (1) (re)sample trait data from an underlying distribution or with measurement error, (2) fit a Gaussian Mixture Model (to account for correlation) to each resample, (3) build a consensus matrix quantifying how often species pairs are grouped together across the noisy trait sample, and (4) apply traditional hierarchical clustering to this matrix and select the final groups. As a case study of this approach, we apply this method to a global dataset of 47,828 tree species using 18 traits, identifying 42 functional groups with distinct trait patterns and varying degrees of stability. We show how the resulting groups reflect underlying ecological trade-offs and phylogenetic structure, and we demonstrate how traditional diversity metrics (richness and Simpson's Index) can be applied to these functional groups to provide intuitive measures of functional group richness and functional redundancy. Collectively, this framework presents a scalable, interpretable approach for quantifying functional groups that embraces trait correlation and trait uncertainty, allowing for repeatable and intuitive

grant #PZ00P3_193612 to D.S.M, which provided salary and computational supplies; and by ANID, Beca de Magister en el Extranjero (2023-73230729) to P.U.P., which funded his research degree at UCL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

quantification of functional biodiversity that can aid its adoption in biodiversity assessments by conservation and restoration organisations.

Author summary

Biodiversity is commonly quantified by the number of distinct species within a given ecosystem. In contrast, functional diversity uses traits—such as tree height or leaf area—to quantify how species’ ecological strategies vary. Quantifying functional diversity poses challenges: traits are often correlated, and many species have incomplete trait data. To address this, we introduce a method that generates multiple plausible trait profiles for each species. For each scenario, species are clustered based on trait similarity, and we assess how frequently species are grouped together across scenarios. Species that consistently cluster together are likely to fill similar ecological roles, and use these consensus patterns to form “functional groups” of species that share similar trait profiles. Applying this method to a global dataset of over 47,000 tree species, we identified 42 functional groups, varying in robustness, and we show how these groups relate to underlying trait patterns. Finally, we demonstrate how these groups can be used to measure functional diversity in a clear and practical way, helping conservation and restoration efforts track and protect biodiversity more effectively.

Introduction

Quantifying biodiversity is an important task because it helps us define conservation targets, evaluate its change over time in different regions, and understand how it relates to different environmental and human pressures [1–4]. The most common approach to tracking biodiversity is by species richness, which is directly related to phylogeny, and thus the genetic difference between organisms [5–7]. While phylogeny offers important insights into biodiversity, functional diversity presents another valuable perspective by focusing on traits rather than just species [8]. This approach is particularly valuable because traits are closely linked to ecological strategies [9], niches [10,11], and species’ adaptability to environmental changes [12,13]. As we enter an era dominated by big data, the accumulation of extensive trait measurements from various species among different taxonomic groups has led to promising approaches in using these traits as features in predictive models [14]. For example, we can predict how an organism would respond to climate change [15,16], survive in particular habitats [17] or coexist with other species [9,18].

Early efforts to quantify functional diversity primarily focused on functional groups—i.e., sets of species that share traits or ecological roles—which provided tractable means to estimate functional richness via traditional diversity metrics [19,20]. These groupings were often assumed to capture meaningful variation in species’ contributions to ecosystems and provided a convenient way to apply categorical

diversity metrics [21]. However, functional group delineation was criticized for its subjectivity [22]: group boundaries were typically determined by researcher discretion and thus varied across studies, hindering efforts to compare across systems or regions. Moreover, groupings can obscure the continuous nature of functional traits, limiting the resolution of trait-based ecological inference at smaller spatial or temporal scales [23]. In response, the field has increasingly embraced continuous trait-based approaches, where species are positioned in a multi-dimensional trait space to capture fine-scale variation [24,25]. These methods allow richer, quantitative representations of functional composition and facilitate metrics such as functional dispersion and trait divergence.

Yet continuous trait-based metrics bring their own challenges. As trait dimensionality increases, data become complex and interpretation becomes more difficult [26,27]. Ecologists must choose which traits to include, how to weight them, which of the dozens of trait metrics to use, and how to manage missing data and trait uncertainty—all of which can affect the results and inference, hindering comparisons across studies [25]. More problematically, many trait-based metrics are difficult to interpret for non-specialists [22], constraining their use primarily to academic research, and preventing their adoption in conservation frameworks that rely on simpler, threshold-based decision tools. As such, most biodiversity frameworks largely ignore functional diversity [28–30], reflecting the difficulty in obtaining simple, repeatable, and interpretable metrics that hold across studies.

A return to functional group-based approaches offers several compelling advantages, particularly for broad-scale or policy-oriented biodiversity assessments. Grouping species by shared ecological roles, life-history strategies, or resource-use traits provides a natural means of abstraction that simplifies trait data without necessarily discarding its ecological meaning [21,22,31]. Functional groups are more intuitively interpretable and lend themselves to categorical metrics, decision thresholds, and standardized reporting formats that are already used in conservation planning [30]. Moreover, compared to continuous trait metrics, functional classifications can be more robust to missing data, trait inclusion (or exclusion), and taxonomic inconsistencies [32–35], allowing for broader taxonomic and geographic coverage in global biodiversity datasets. From a practical perspective, functional groups are particularly appealing because biodiversity metrics designed for quantifying species diversity can be directly applied to these functional groups [36,37]. In addition, comprehensive lists of functional groups could be constructed for the use of non-experts, helping with evidence-based assessments of functional diversity.

In order to apply functional group approaches, we need to develop robust methods for classifying species to ensure that trait groupings are informed by underlying traits, and not simply expert knowledge. A clustering algorithm finds patterns in the data by grouping observations in the most fitting way possible based on certain assumptions [38,39]. As clustering is an unsupervised task, there is no agreed-upon correct way of knowing which algorithm would suit a particular dataset [40]. When choosing to use clustering algorithms, it is important that the data is clearly understood and what the reason to cluster is, so that the chosen algorithm aligns with the research question. There are various clustering algorithms with different assumptions. Flat clustering algorithms, also known as partition clustering, group species without creating inter-group structures, while hierarchical clustering algorithms create a hierarchy of groups, allowing for group joining or separation [41]. Clustering algorithms can also be parametric, assuming parameters to model cluster distribution, or non-parametric, identifying clusters based solely on data we observe [42]. In trait ecology, these assumptions impact the ecological interpretation of groups. A flat parametric algorithm suggests each functional role lies somewhere in trait-space and species are driven to these functional roles. A density-based approach suggests species with similar ecological functions evolve their traits systematically, occupying a feasible trait-space for that function.

Regardless of the clustering method, there are two key challenges when clustering functional data: (1) the ubiquity of missing data and data uncertainty, and (2) the curse of dimensionality and correlation. First, most clustering algorithms require a complete trait dataset, which leads to the use of imputation methods for missing trait data via methods such as BHPMF [43] and MissForest [44]. Regardless of the method used for imputation, there is uncertainty introduced to the dataset, as the real value remains unknown and the imputed values have intrinsic imputation error (which is often not reported). Second, because traits are often highly correlated with each other, the addition or removal of specific traits

can lead to divergent clustering patterns, compromising some distance metrics and performance. To address the issue, researchers commonly use techniques like PCA [45,46] or recently popularized autoencoders [47,48] to reduce dimensionality and provide orthogonal trait inputs. These methods can be particularly useful in high-dimensional settings (e.g., audio or image classification), but, by favoring traits that capture significant variation across the dataset, they may erase patterns or ignore unique traits that are important for identifying particular groups.

Here, we address these issues by presenting a novel multi-step clustering approach for calculating functional groups, which addresses the combined challenges of trait uncertainty and trait correlation (or trait selection). First, to address uncertainty, our approach relies on resampling the data (by adding measurement error or by sampling from a known distribution reflecting trait variation), where each resample is interpreted as a possible realisation of the underlying distribution. Second, for each resample, we fit a Gaussian Mixture Model (GMM), which addresses high-dimensionality and collinearity by modelling trait covariation, and we identify the number of components that minimizes the Bayesian Information Criterion (Clustering ensemble). This two-step approach results in a clustering ensemble, which we then synthesize using consensus clustering by computing how often each pair of species was linked together (Consensus clustering). These proportions are then treated as a distance metric between species, making it suitable for the use of hierarchical clustering as a final step in the analysis workflow to obtain a single overall group assignment.

To illustrate this approach, we provide a case-study of 47,828 tree species across the globe, classifying them into functional groups using 18 traits related to leaf, crown, root, seed, stem, and wood physiology and morphology. We show that this approach not only identifies the stable functional groups but also provides insight into the underlying functional traits driving these patterns. By comparing this functional group metric to existing species diversity metrics, we illustrate how functional groups provide an intuitive alternative to more complex functional diversity metrics, allowing us to apply species-based metrics (e.g., Simpson's Index) to functional group classification. Collectively, the approach developed here presents an adaptable framework that can be applied to other taxonomic groups and geographic regions, making it a promising tool for future biodiversity assessments and ecological studies that require simple, repeatable, and robust measures of functional diversity.

Materials and methods

Overview

Motivated by the original application of consensus clustering [49], our methodology comprises a multi-step clustering approach that addresses the challenge of uncertainty and correlation between traits (Fig 1). In summary, we resample our dataset S times, generating a clustering ensemble, where, for each resample, we fit a GMM with the total number of components (functional groups) that minimizes BIC (Clustering ensemble), such that every species is assigned to a group which can vary within iterations. This generates a clustering ensemble that is synthesized using consensus clustering by computing how often each pair of species was linked together (Consensus clustering). These proportions are then treated as a distance metric between species, making it suitable for the use of hierarchical clustering. This allows for a flexible clustering scheme, where the number of groups can be decided based on the particular application.

Throughout, we denote matrices with bold and capital letters (e.g., $\mathbf{A}$), vectors with bold and lowercase letters (e.g., $\mathbf{a}$), and scalars and indices with non-bold letters (e.g., a, i, j). The cardinality of a set S , representing the number of elements in the set, is denoted by $|S|$. We otherwise define any notation that is specific to this methodology.

Clustering ensemble

The first step in this method addresses the fact that trait data are noisy, whether due to intraspecific or ontogenetic variation, or to measurement error or imputation noise. We address this uncertainty by adopting an ensemble approach to clustering, where we resample the data in the presence of uncertainty, and fit a Gaussian mixture model to each resample.

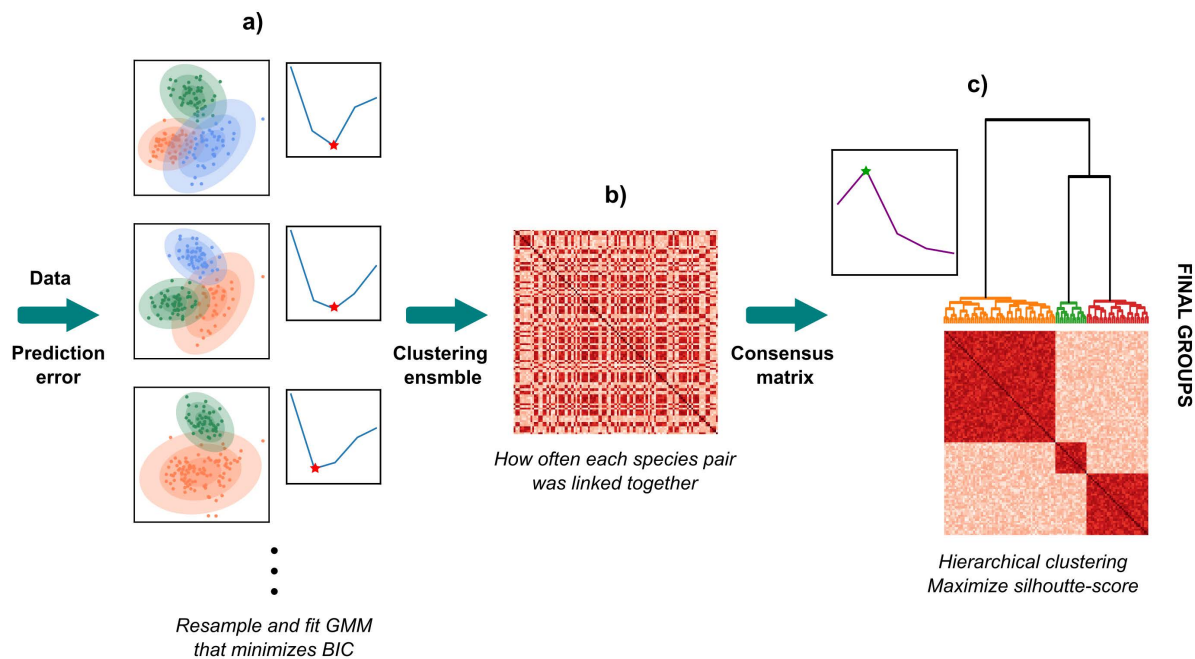


Fig 1. Conceptual diagram for consensus clustering with data uncertainty. **a)** The input of the model is the data and prediction error to account for data uncertainty. These data and prediction error are used to perform resamples that are clustered using GMM by optimizing the number of groups with BIC. **b)** These resamples yield a clustering ensemble, where for each pair of species it is possible to compute the proportion of times they were linked together. This builds a consensus matrix. **c)** This consensus matrix is used as a distance matrix to perform hierarchical clustering to optimize the silhouette score (due to the absence of a likelihood for BIC-based selection) and obtain the final group assignments.

<https://doi.org/10.1371/journal.pcbi.1014278.g001>

Specifically, let $\mathbf{X}$ denote a matrix in which each row corresponds to a species, and each column represents a trait. Formally, $\mathbf{X}$ has N rows (species) and M columns (traits). Let us consider $\mathcal{N}$ as the set containing all species, and $\mathcal{M}$ as the set containing all traits. We are not considering the case where some traits could be categorical, as this further complicates the clustering problem [50].

The clustering problem itself consists of finding a partition of $\mathcal{N}$, denoted as $\mathcal{P}$, containing a collection of sets that do not share any element with each other and that would jointly account for every species in $\mathcal{N}$. We denote G as the number of sets from the partition, which is equivalent to the number of groups. Also, denote $\mathcal{C}_g$ as the set g in the partition. The first challenge of this problem is that we do not know the value of G , as this is part of what we are trying to answer.

The second challenge of our clustering problem is that our matrix of observations $\mathbf{X}$ comes with uncertainty. This happens because some entries of the matrix were imputed using predictive methods. To address this uncertainty, we are going to consider the input of the clustering problem to be a random variable $\mathbf{W}$, which we can resample S times. Then, let $\mathbf{W}^{(s)}$ represent the s -th instance for clustering, where these instances are independently and identically distributed. Each resampling iteration yields a distinct partition $\mathcal{P}^{(s)}$, which may come with a different total number of groups $G^{(s)}$.

In this section, we will first detail how to model the random variable $\mathbf{W}$, then detail how to fit a GMM for each resample of this variable, and finally how to choose the best value of G for each resample.

Resampling from error distribution. To account for trait variation and prediction errors in our data, we use a resampling approach that is repeated in order to obtain a robust cluster. Think of W_{ij} as a random variable for the value of trait j for species i , with the following distribution:

$$W_{ij} = X_{ij} + \epsilon_{ij} \quad \epsilon_{ij} \sim E_{ij} \quad \forall i \in \{1, \dots, N\}, \quad j \in \{1, \dots, M\} \quad (1)$$

Here, E_{ij} denotes the error distribution for species i and trait j . For example, this can be specified non-parametrically by sampling from empirical prediction errors, or parametrically using a normal distribution centered at zero with the observed standard deviation. The matrix $\mathbf{W}$ is then a random matrix, where each entry represents a random variable reflecting uncertainty in trait values.

Fitting a Gaussian mixture model

We adopt to use Gaussian Mixture Models for each resample because of the intrinsic ability of a GMM to account for correlation between traits, and its ability to allow for different correlation structures for each group [51]. Moreover, GMMs provide a flexible probabilistic framework for clustering that can capture a wide range of cluster shapes through mixtures of Gaussian components. Compared to methods such as k-means or PAM, which rely on distance-based partitioning, GMMs allow clusters to differ in covariance structure and orientation. This flexibility is particularly useful for ecological trait data. In particular, the data used in our experiments exhibit a clear correlation structure, and the traits are approximately normally distributed (S1 and S2 Figs), making GMM a well-suited choice in this case. Nevertheless, depending on the ecological application and the expected importance of trait covariation, alternative clustering methods may be more appropriate.

For the effects of this section, consider $\mathbf{W} = \mathbf{w}$ to be any realization of the random variable that serves as input for the clustering algorithm. Note that in cases with uncertainty in the data, fitting a GMM without resampling likely leads to over-fitting (S10 Fig).

In a GMM, each group g would be defined by a Multivariate Normal distribution with mean μ_g , covariance matrix Σ_g , and weight π_g , that determines the probability of a species belonging to group g . We will refer to the complete set of parameters as: $\theta^{(g)} = \{\mu_g, \Sigma_g, \pi_g\}_{g=1}^G$. These parameters are latent variables of the model (unobserved) that we are trying to recover by maximizing log-likelihood.

To estimate the parameters $\theta^{(g)}$, we maximize the log-likelihood using the EM-algorithm [52], which involves two steps:

E-step: Calculate the posterior probabilities τ_{ig} that i belongs to g , given the current parameter estimates. Using Bayes and the law of total probability:

$$\begin{aligned}\tau_{ig} &= P(i \in \mathcal{C}_g | \mathbf{W}_i = \mathbf{w}_i; \theta^{(g)}) \quad \forall i \in \mathcal{N} \quad \forall g \in \{1, \dots, G\} \\ &= \frac{P(\mathbf{W}_i = \mathbf{w}_i | i \in \mathcal{C}_g; \theta^{(g)}) \cdot P(i \in \mathcal{C}_g; \theta^{(g)})}{P(\mathbf{W}_i; \theta^{(g)})} \\ &= \frac{\phi(\mathbf{W}_i = \mathbf{w}_i; \mu_g, \Sigma_g) \cdot \pi_g}{\sum_{h=1}^G \phi(\mathbf{W}_i = \mathbf{w}_i; \mu_h, \Sigma_h) \cdot \pi_h}\end{aligned}\quad (2)$$

where ϕ is the Multivariate Normal density function.

M-step: Update the parameters μ_g , Σ_g , and π_g by maximizing the expected log-likelihood (3) using the posterior probabilities τ_{ig} from the E-step (2).

$$l(\theta^{(g)}; \mathbf{W} = \mathbf{w}) = \sum_{i=1}^N \ln \left(\sum_{g=1}^G \pi_g \phi(\mathbf{W}_i = \mathbf{w}_i; \mu_g, \Sigma_g) \right) \quad \forall g \in \{1, \dots, G\} \quad (3)$$

The algorithm alternates between these steps until the parameter estimates converge. After fitting the model, group membership for each species can be interpreted either as a probability distribution, where $P(i \in \mathcal{C}_g) = \tau_{ig}$, or by assigning each species to the group with the highest posterior probability, $i \in \mathcal{C}_g | g = \arg \max_g \tau_{ig}$.

Choosing the optimal number of groups G is challenging since increasing G naturally improves the log-likelihood by adding more parameters. Thus, the goal is to find the smallest G that still adequately fits the data.

An approach to choose the value of G is to use the Bayesian Information Criterion (BIC), as we are working directly with a probability distribution. This metric balances the trade-off between the log-likelihood and the number of parameters. This metric has been shown to estimate the correct true number of groups in multiple situations under certain simulated conditions and is usually recommended against other metrics like the Akaike Information Criterion [53]. Note that the number of parameters increases linearly as we fit more groups. Defining $|\theta_{\text{unique}}^{(G)}|$ as the number of parameters for a GMM with G groups (without considering repeated parameters), the BIC value for a GMM with G components is:

$$|\theta_{\text{unique}}^{(G)}| = \underbrace{(G-1)}_{\text{Number of weights}} + \underbrace{(G \cdot M)}_{\text{Number of means}} + \underbrace{(G \cdot M \cdot (M+1) \cdot 2^{-1})}_{\text{Number of covariances}} \quad (4)$$

$$BIC_G = -2 \cdot l(\theta^{(G)}; \mathbf{X}) + |\theta_{\text{unique}}^{(G)}| \cdot \ln(N) \quad (5)$$

Using this metric, we can try different values of G and choose the GMM that minimizes the BIC score. Then, we can define $G^* = \arg \min_G BIC_G$. Note that other metrics used for choosing the number of clusters, such as the silhouette-score or the elbow method, do not penalize model complexity, can be affected by correlation, and ignore the probabilistic nature of the model [53–55].

Consensus clustering

Through resampling, we obtained a collection of partitions $\mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \dots, \mathcal{P}^{(S)}$, each of which is the best clustering result for a particular resample of our data. This is useful, as we generated a distribution of clusters that accounts for uncertainty. However, this process introduces a new challenge: determining a singular assignment for each species from the multitude of individual clusterings. To address this issue, we need to decide a singular assignment for species i based on how often it was linked to the other species $i' \neq i$.

The idea of consensus clustering was first introduced for visualizing gene expression [49]. The main purpose of this algorithm is to find consensus between multiple cluster algorithms or multiple realizations of the same algorithm. While the original framework proposes bootstrapping as a resampling method, we saw the potential to incorporate prediction error as a resampling mechanism to address uncertainty. Then, we construct a consensus matrix $\mathbf{C}$ that summarizes the clustering ensemble previously obtained. In this matrix, component $C_{ij'}$ shows how often species i was linked to species i' throughout the S iterations of resampling.

$$C_{ij'} = \sum_{s=1}^S \frac{\sum_{g=1}^{G^{(s)}} \mathbb{I}\{(i \& i') \in C_g^{(s)}\}}{S} \quad i \neq i' \quad (6)$$

$$C_{ii} = 1$$

Note that the clustering ensemble and the construction of the consensus matrix could be done simultaneously in an efficient way as described in Algorithm 1.

In this section, we will go through how to build a clustering structure with the consensus matrix, and how to ultimately choose the number of groups with this structure.

Algorithm 1 Consensus Clustering with Trait Resampling and GMM random assignments

```

1: Input: Original matrix  $\mathbf{X} \in \mathbb{R}^{N \times M}$ , error distributions  $E$ , number of samples  $S$ , number of random assignments  $A$ 
2: Output: Final matrix  $\mathbf{C}$  containing average cluster co-occurrences
3: Initialize  $\mathbf{C}$  as an  $N \times N$  zero matrix
4: for  $s=1$  to  $S$  do
5:   Sample  $\mathbf{W}^{(s)}$  as described in (1)
6:   Fit GMM to  $\mathbf{W}^{(s)}$  to obtain probability matrix  $\boldsymbol{\tau}^{(s)}$  as described in (2)
7:   for  $a=1$  to  $A$  do
8:     Sample partition  $\mathcal{P}^{(s,a)}$  from  $\boldsymbol{\tau}^{(s)}$ 
9:     Initialize  $\mathbf{C}^{(s,a)}$  as an  $N \times N$  zero matrix
10:    for each pair  $(i, i')$  do
11:       $\mathbf{C}_{i,i'}^{(s,a)} \leftarrow 1$  if  $\exists \mathcal{C} \in \mathcal{P}^{(s,a)} \mid (i, i') \in \mathcal{C}$ , otherwise 0
12:    end for
13:     $\mathbf{C} \leftarrow \mathbf{C} + \mathbf{C}^{(s,a)}$ 
14:  end for
15: end for
16:  $\mathbf{C} \leftarrow \frac{1}{S \cdot A} \mathbf{C}$ 

```

Hierarchical clustering

Hierarchical clustering, particularly linkage-based methods, is well-suited for our clustering framework because it requires only a distance metric rather than specific observations. This is ideal for our case, as the consensus matrix $\mathbf{C}$ can be interpreted as a distance matrix. By defining $\mathbf{D} = \mathbf{I} - \mathbf{C}$, we treat species with high consensus as closer to each other for clustering purposes, making hierarchical clustering a natural choice for this step.

For our application, we employ the Ward linkage method due to its ability to create compact, homogeneous clusters while minimizing variance [40], making it widely adopted in ecological studies for phylogenetics and species distribution [49,56,57]. It forms clusters by merging pairs that result in the smallest increase in within-cluster variance at each step using an agglomerative approach. In doing so, it implicitly results in clusters that are compact and approximately spherical in multivariate trait space, which is consistent with ecological niche theory, which holds that species occupy bounded, unimodal regions in multidimensional niche and trait space [58,33]. Nevertheless, although we adopt Ward's method due to its simplicity and ecological relevance, we stress that any hierarchical clustering linkage method could be used in place, depending on the desired application and resolution of small vs. large clusters, and the expected shape of the ecologically relevant clusters in multivariate trait space.

Choosing the number of groups for hierarchical clustering. As with the GMM procedure, we once again face the challenge of choosing the number of groups to obtain a single consensus cluster. We address this using a dendrogram generated by the Ward linkage, which is a tree-like diagram that illustrates the arrangement of clusters formed at various levels. We denote the number of groups as K to distinguish it from those in the GMM. Unlike GMM, which is a flat clustering algorithm, hierarchical clustering allows flexibility to adjust K by cutting the dendrogram at different levels [59].

To choose the best value for K , we use the silhouette score, which measures how well a species aligns with its cluster versus other clusters. This score is effective in identifying the true number of clusters across various applications [54]. In contrast to the previous resampling step, which used a likelihood-based criterion such as BIC, this approach relies on a distance matrix and therefore does not permit the use of likelihood-based methods. The silhouette score $s_K(i)$ for species i in a clustering with K groups is defined as:

$$s_K(i) = \frac{b_K(i) - a_K(i)}{\max\{a_K(i), b_K(i)\}} \quad \forall i \in \mathcal{N} \quad (7)$$

$$a_K(i) = \frac{1}{|C_{K(i)}| - 1} \sum_{i' \in C_{K(i)}, i' \neq i} (1 - C_{ii'})$$

$$b_K(i) = \min_{C_{K(i)} \neq C_{K'}} \frac{1}{|C_{K'}|} \sum_{i' \in C_{K'}} (1 - C_{ii'})$$

Where intuitively: $a_K(i)$ is the average distance from species i to all other species in the same cluster, and $b_K(i)$ is the average distance from species i to all the species in the nearest next cluster. Then K^* is the value that maximizes $\sum_{i=1}^N s_K(i)$.

By maximizing the silhouette score, we obtain a final partition of $\mathcal{N}$, denoted as $\mathcal{P}_{K^*}$, which serves as the outcome of the clustering framework. Note again that this partition is hierarchical, not flat like the one used for GMM, allowing flexibility to adjust the number of groups ($K < K^*$ or $K > K^*$) within the same clustering structure.

It is important to note that the optimum in terms of silhouette-score comes from a hierarchical clustering perspective that is different from the resampling approach. The optimal number of clusters chosen at the last step does not need to match the mean of optimums obtained during the resampling step.

Case study: Functional groups of trees

Traits are particularly important for trees, as they govern water, nutrient, and light economies within individuals [1,60], which underpin key life-history trade-offs that lead to trait selection across different environments [61]. Trees have unique traits related to woody structure, size, and longevity [62,63], which sets them apart from herbaceous, non-dendritic plant species, whose analysis is often restricted to leaf-related traits [64,65]. The unique traits of trees can provide insight into broader ecological functions of trees, such as a forest's carbon storage capacity [66,67] and its resilience to climate change [15,68,69]. Such analyses are indispensable for predicting the impacts of environmental shifts and planning effective conservation strategies.

In this case study, we apply our consensus clustering approach in the presence of trait uncertainty, as some of the trait measurements have been imputed using machine learning methods [70]. We address two main questions: (1) how many groups there are, and (2) what these groups are. Question (1) is addressed from an optimization perspective, where we maximize metrics that are related to how good a clustering result is. Question (2) is addressed from a perspective of trait uncertainty, where each resample of the data yields a different possible clustering.

Data. The dataset for this case study is based on observed species-level values and imputed values from a study analyzing global tree trait relationships [70]. Imputed values were generated using machine learning methods based on the phylogenetic distances between species, which introduces the uncertainty we will address in the clustering problem. The data consists of $N=47,828$ species of trees and $M=18$ traits (for details see Table 1, S1 and S2 Figs). This data accounts for around 70% of all tree species, across different taxa and geographic regions [71].

Prior to clustering, all traits were log-transformed, as is typical for functional trait analysis in plants [72]. This is for two reasons: first, ecologically meaningful differences in traits are typically proportional rather than absolute, as strategies are differentiated by orders of magnitude rather than additive differences. In addition, because most ecological traits are strongly right-skewed, the log transformation stabilizes the variance and ensures approximate normality, which satisfies the assumptions of Gaussian processes.

Error data. In the case of trees, the error distribution for trait j was computed by collecting all empirical errors available for that trait (S3 Fig). Maynard et al. produced species-level mean estimates of 18 traits, along with the corresponding error (observed - predicted) for each species that had one or more observed trait values. By collecting these trait-specific

Table 1. Trait statistics. Each trait specifies the unit of measurement, the mean and standard deviation for natural-log transformed values, the mean, standard deviation, minimum, and maximum values for the original scale, the number of error samples, and the Mean Absolute Error (MAE) on the standardised log scale.

Trait name	ln()		Original Scale				N. Errors	MAE (log)
	Mean	Std. Dev.	Mean	Std. Dev.	Min	Max		
Bark thickness (mm)	1.21	0.78	4.64	4.76	0.35	42.70	1,085	0.405
Crown diameter (m)	2.39	0.49	12.38	7.36	2.84	40.00	—	—
Crown height (m)	2.42	0.42	12.29	5.39	3.50	41.30	—	—
Leaf K per mass ($\frac{mg}{g}$)	2.14	0.42	9.29	3.92	1.04	27.43	1,511	0.242
Leaf N per mass ($\frac{mg}{g}$)	2.91	0.30	19.23	5.64	4.72	43.50	4,538	0.099
Leaf P per mass ($\frac{mg}{g}$)	-0.02	0.43	1.08	0.47	0.16	3.90	3,018	0.199
Leaf Vcmax per dry mass ($\frac{\mu mol}{g \cdot s}$)	-1.17	0.37	0.33	0.12	0.08	0.92	527	0.261
Leaf area (cm ²)	7.79	1.48	5,693.45	8,526.82	6.64	108,949.93	699	0.793
Leaf density ($\frac{g}{cm^3}$)	-1.03	0.26	0.37	0.08	0.04	0.74	1,488	0.160
Leaf thickness (mm)	-1.51	0.31	0.23	0.09	0.06	1.21	1,847	0.153
Root depth (m)	0.95	0.88	4.19	6.51	0.46	60.00	—	—
Seed dry mass (mg)	4.05	2.42	522.19	1,638.89	0.09	41,832.44	5,204	0.863
Specific leaf area ($\frac{cm^2}{g}$)	2.45	0.38	12.37	4.52	1.53	38.46	4,627	0.143
Stem conduit diameter (μm)	3.36	0.32	30.43	10.29	8.92	72.70	252	0.189
Stem diameter (m)	-0.61	0.55	0.64	0.45	0.15	6.48	—	—
Stomatal conductance ($\frac{mmol}{m^2 \cdot s}$)	5.03	0.46	170.90	85.51	25.42	715.19	866	0.334
Tree height (m)	3.23	0.41	27.44	11.56	6.26	119.80	—	—
Wood density ($\frac{g_{dry}}{cm^3_{wet}}$)	-0.55	0.23	0.59	0.13	0.18	1.16	7,224	0.103

<https://doi.org/10.1371/journal.pcbi.1014278.t001>

errors, we obtain an empirical posterior error distribution for each trait, nested within taxonomic groups (i.e., one set for angiosperms, one set for gymnosperms).

Specifically, let $\mathcal{I}_j^{(obs)}$ and $\mathcal{I}_j^{(pred)}$ be the sets of observed and predicted values for trait j , respectively. We also differentiate between a set of angiosperms $\mathcal{A}$ and a set of gymnosperms $\mathcal{G}$, as these taxonomic groups are significantly different and some methodologies to measure their traits could vary [73]. We specifically have $|\mathcal{A}| = 47,263$ angiosperms species and $|\mathcal{G}| = 565$ gymnosperm species. Then for traits that we know some observed values, we can sample errors using the following error sets:

$$\mathcal{E}_{\mathcal{A}j} = \{X_{ij}^{(pred)} - X_{ij}^{(obs)}, \forall i \in (\mathcal{I}_j^{(obs)} \cap \mathcal{A})\} \quad E_{\mathcal{A}j} = \text{Uniform}(\mathcal{E}_{\mathcal{A}j}) \quad \forall j : |\mathcal{I}_j^{(obs)}| > 0 \quad (8)$$

$$\mathcal{E}_{\mathcal{G}j} = \{X_{ij}^{(pred)} - X_{ij}^{(obs)}, \forall i \in (\mathcal{I}_j^{(obs)} \cap \mathcal{G})\} \quad E_{\mathcal{G}j} = \text{Uniform}(\mathcal{E}_{\mathcal{G}j}) \quad \forall j : |\mathcal{I}_j^{(obs)}| > 0$$

Note that for some traits we do not have error samples (marked empty in Table 1). This happens because these traits were predicted using quantile random forest [70], therefore the observed and predicted values are not comparable for these traits. For this case, we assume that the error distribution comes from the union of all sets of errors that we do have.

$$\mathcal{E}_{\mathcal{A}j'} = \bigcup_{j \in \mathcal{M}^{(obs)}} \mathcal{E}_{\mathcal{A}j} \quad \forall j' : |\mathcal{I}_{j'}^{(obs)}| = 0 \quad (9)$$

$$\mathcal{E}_{Gj'} = \bigcup_{j \in \mathcal{M}^{(obs)}} \mathcal{E}_{Gj} \quad \forall j' : |T_{j'}^{(obs)}| = 0$$

This configuration of errors is well-suited for this particular case study, but it is recommended to adopt an error distribution that suits the particular application. In our case study, for example, we do not consider correlations between errors, in part because the sparsity of the original data set means that the covariance matrix cannot be robustly calculated apart from a subset of species [70]. We thus assume uncorrelated errors in our approach, with the prediction errors being sampled directly from the posterior empirical error distribution for each trait. In other settings, however, more complex error distributions could be incorporated, such as with Bayesian extension of probabilistic matrix factorization (BHPMF), which has been used to impute missing data in other plant datasets [43]. Such decisions should be based on the data availability and the question of interest.

Model configuration. For GMM fitting, we consider $S=50$ resamples of the data, as this yields a reasonable standard error (~ 0.07) given the computational limitations. As with any bootstrap or resampling procedure, one should generally take the maximum number of resamples possible, given computational resources and time availability. For each resample, we look for the optimum between $G=10\text{--}60$. For the hierarchical clustering step we test values $K=2\text{--}120$. These ranges were chosen to reflect the values observed in our experiments, where the optimal solutions consistently fell within the mid-range, indicating that at least a local optimum had been identified.

To test different configurations of our framework, we varied two components: (1) feature scaling, and (2) group assignment. For (1) feature scaling, we compare (1.i) the original log-scale values (as described in Data), and (1.ii) a robust scaling approach, where each trait column was centered by its median and scaled by its interquartile range—that is, the difference between the 75th and 25th percentiles. For (2) group assignment (see Fitting a Gaussian Mixture Model), we evaluated (2.i) a sampling-based strategy, where group membership for each species was drawn from the probabilities τ_{ig} , using $A=10$ assignments per realization of $\mathbf{W}$, and (2.ii) a deterministic strategy, where each species was assigned to its most probable group by taking the maximum of τ_{ig} . Among the resulting four configurations, we selected the one that maximizes the final framework metric: the silhouette score.

Analysis. Using the final clustering partition $\mathcal{P}_{K^*}$ derived from the global tree dataset, we conducted a series of downstream analyses to visualize the resulting clusters, evaluate their stability, examine trait composition and variation within and among clusters, and compare functional biodiversity metrics based on this clustering approach with those obtained from existing methods.

Visualization

To visualize the clustering results, we use the following dimensionality reduction methods:

PCA [46]: Linear dimensionality reduction by creating new, uncorrelated variables ordered by variance.

t-SNE [74]: Uses a probabilistic approach to match a lower-dimensional distribution to the original.

UMAP [75]: Matches graph structures between high and low dimensions capturing non-linear relations.

It is important to acknowledge that dimensionality reduction methods may result in the loss of some information; however, they are useful to allow human interpretability of the data patterns.

Statistics

For each cluster $\mathcal{C}_k \in \mathcal{P}_{K^*}$, we computed aggregate statistics, such as the mean and standard deviation of each trait across species $i \in \mathcal{C}_k$. Additionally, we analyzed the cluster sizes $|\mathcal{C}_k|$ and examine the dendrogram that produces the partition.

For species level stability we explored how similar a species is to other species in its own cluster, and how similar it is to all species overall. This is expressed as:

$$stab_i = \frac{\sum_{i': i' \in C_{k(i)}, i' \neq i} C_{ii'}}{\sum_{i' \neq i} C_{ii'}}$$

We evaluated clustering robustness by comparing our consensus-based approach to repeated Gaussian Mixture Model (GMM) fits under resampling. On a reduced dataset, we generated multiple resampled subsets and computed pairwise Adjusted Rand Index (ARI) scores between resulting clusterings. We then applied our consensus procedure to groups of resampled runs and compared the stability of the resulting consensus clusterings.

When comparing group-based functional diversity metrics to traditional metrics, we use generalized additive models (GAMs) to account for spatial autocorrelation. Specifically, using the *mgcv* package in R, we fit GAMs using a B-spline basis for each predictor ($k=7$, second-order penalty with first-order derivative), with a spherical smooth of geographic coordinates to account for spatial autocorrelation. Models were estimated using restricted maximum likelihood (REML) with an elevated smoothing parameter ($\gamma = 1.4$) to guard against overfitting (e.g., `gam(y ~ s(x, bs='bs', k=7, m=c(2, 1)) + s(Longitude, Latitude, bs="sos"))`). To estimate residual correlation after accounting for spatial autocorrelations, we fit separate GAMs for the predictor and outcome variables, retaining only the geographic spline term. We then calculated the Pearson correlation between the residuals of both models, providing a measure of association between x and y after removing spatially structured variation in both variables.

Software and hardware

The analysis was primarily done in *Python* 3.12.4, using *scikit-learn* for Gaussian Mixture Models [76], *SciPy* for hierarchical clustering [77], *pandas* for data manipulation [78], *matplotlib* [79], and *seaborn* [80] for visualizations. Computational experiments were conducted on a compute node equipped with 2× Intel Xeon Gold 6426Y (Sapphire Rapids) CPUs (32 physical cores total) and 512 GB of RAM; and initial model testing and downstream analyses were conducted on an AMD Ryzen Threadripper PRO 5995WX workstation (64 physical cores) with 1024 GB DDR4 RAM. The primary experiment (the most computationally demanding and corresponding to Fig 2) required approximately 7.5 hours using the described hardware configuration.

Code is available at https://github.com/pabloubilla/tree_clustering/ and is archived via Zenodo with corresponding DOI:10.5281/zenodo.19222113. The data have been archived via Figshare with corresponding DOI:10.5522/04/31833526.

Results

Group selection through consensus clustering

For each individual data resampling, we obtain a corresponding BIC curve that identifies the optimal number of groups (G^*) at different values (Fig 2). The mean BIC curves exhibit similar patterns for both scaled and non-scaled resamples, although the BIC values are generally lower for the non-scaled resamples. However, since the BIC values come from different distributions, this does not imply that not scaling yields a better fit.

When comparing the four combinations of scaled or not scaled, and argmax or random assignment, we find they display similar patterns by quickly increasing the score until around 30 groups, then reaching a plateau with slow decreases after reaching their respective optima (Fig 2). The highest silhouette score is achieved with 42 groups for the scaled data at 0.20 using the argmax assignment, which was subsequently used for the final group assignment through hierarchical clustering.

Note that a value of 0.20 represents the optimal result given the level of uncertainty in the data; lower uncertainty would be expected to yield higher consensus and correspondingly higher silhouette scores. If all traits had zero error, the consensus would necessarily be 1.0, and if trait noise was substantially greater, the silhouette score would necessarily be

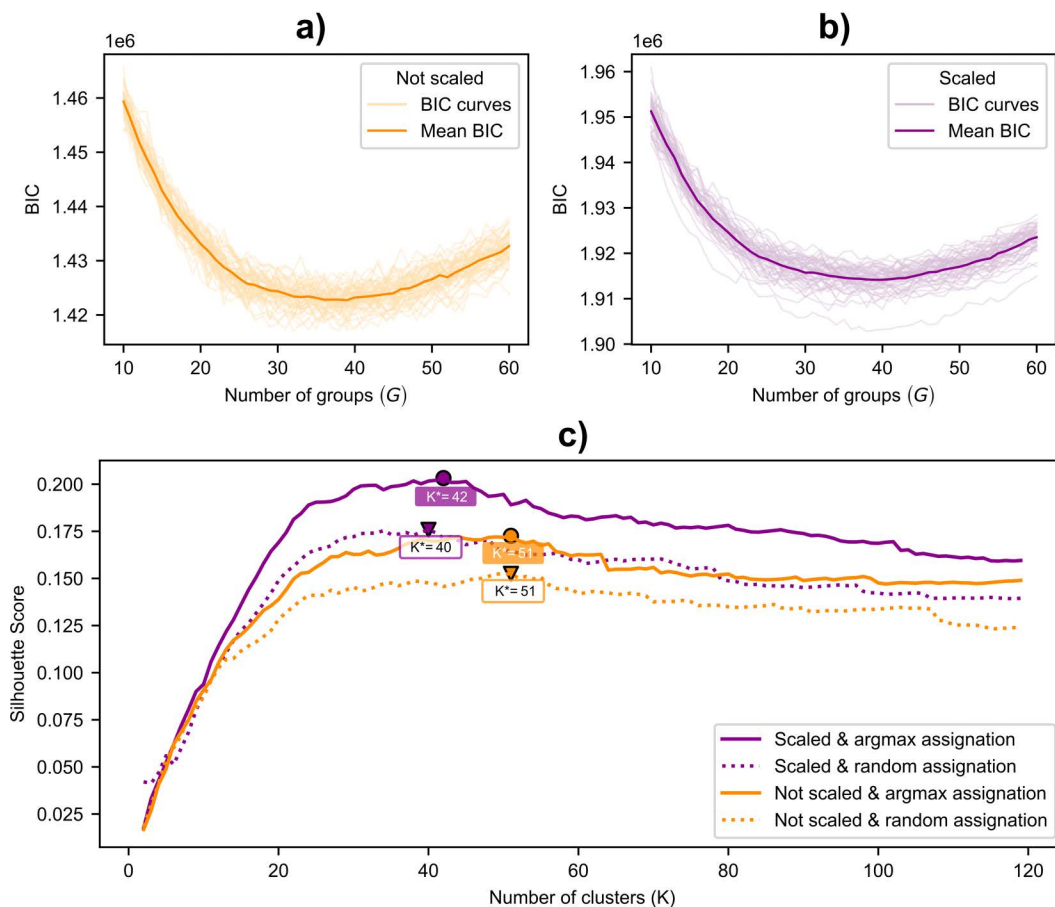


Fig 2. Optimization for number of groups. All BIC curves and the resulting average curve, shown for **a)** unscaled and **b)** scaled data. **c)** Silhouette score for different aggregations of the consensus matrix. A higher value of K means going lower in the dendrogram, hence splitting one group. Note that BIC is optimized independently for each resample in (a) and (b), and therefore partitions with the same number of clusters may differ substantially. In contrast, the Silhouette score is optimized at the aggregated level across all resamples. As a result, the optimal number of clusters identified by these two criteria will not necessarily coincide.

<https://doi.org/10.1371/journal.pcbi.1014278.g002>

lower. Thus, the silhouette score can also be used to indicate the extent to which trait error (or noise) weakens cohesion across the groups.

Using a reduced dataset of 1,365 species, we find that the stability of the clustering results is markedly improved following the consensus step (S4 Fig).

Functional groups results and visualization

When grouping species using hierarchical clustering into 42 groups, we observe a wide range of group sizes and varying levels of consensus within the groups (Fig 3). The consensus values are approximately normally distributed, with a mean of 0.39 and a standard deviation of 0.17. Ideally, average consensus values should be near 1, with dissimilarities (off-diagonal values) close to 0, forming a distinct red diagonal in the consensus matrix. We observe a similar result when looking at the species-level stability (S14 Fig). Our results show most clusters tend not to pair with others, except the last six clusters which show uncertain assignments and have average consensus below 0.2 (S1 Table).

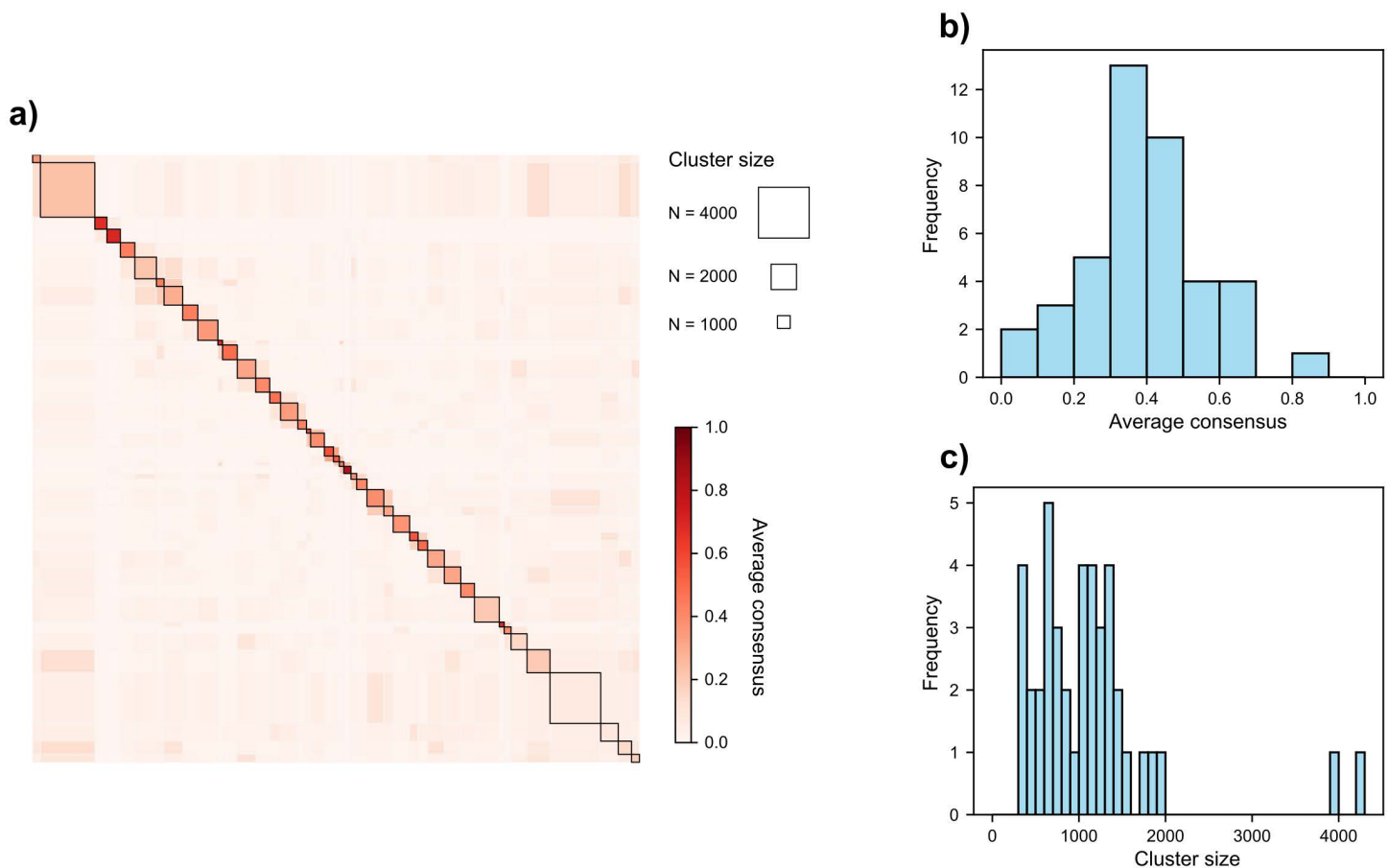


Fig 3. Summary of functional groups results. (a) The consensus matrix displays the average consensus for all species pairs within clusters along the diagonal, while off-diagonal values represent the average consensus between species from different clusters (row vs. column). Square sizes are proportional to their respective cluster size. (b) Distribution of average consensus for the 42 observed groups, ranging from 0 (no consensus) to 1 (maximum consensus), divided into bins of width 0.1. (c) Size distribution of the 42 groups, where cluster size refers to the number of species per group, with bins of width 100.

<https://doi.org/10.1371/journal.pcbi.1014278.g003>

The cluster size distribution is left-skewed, with a mean size of 1,139 species and a standard deviation of 790. The majority of clusters contain close to 1,000 species, while a few larger clusters deviate significantly from this norm (Fig 3 and S1 Table). The size distribution is somewhat inconsistent, displaying gaps, which might be due to the relatively low number of clusters and the large range of potential cluster sizes. Notably, there is no cluster smaller than 300 species, with the minimum size being 324 species.

We are able to visualize the functional groups using t-SNE (Fig 4, interactive version in S1 Appendix), where we generally observe a clear separation between groups. Group 2 (blue square), being the largest, shows compactness at the centre of the distribution, whereas group 39 (red x), being the second largest, is spread across the entire distribution. Group 23 (black square) stands out by being tightly clustered, fully corresponding to all Pinales species (S8 and S9 Figs). Group 3 (teal cross) and 4 (magenta x) are also clearly distinct, both encompassing mostly Myrtales species. Other dimensionality reductions do not show as clear a separation between clusters, but some are reasonably identified regardless of the reduction methodology (S5 and S6 Figs). Note that this visualization is only intended as a qualitative representation of cluster structure.

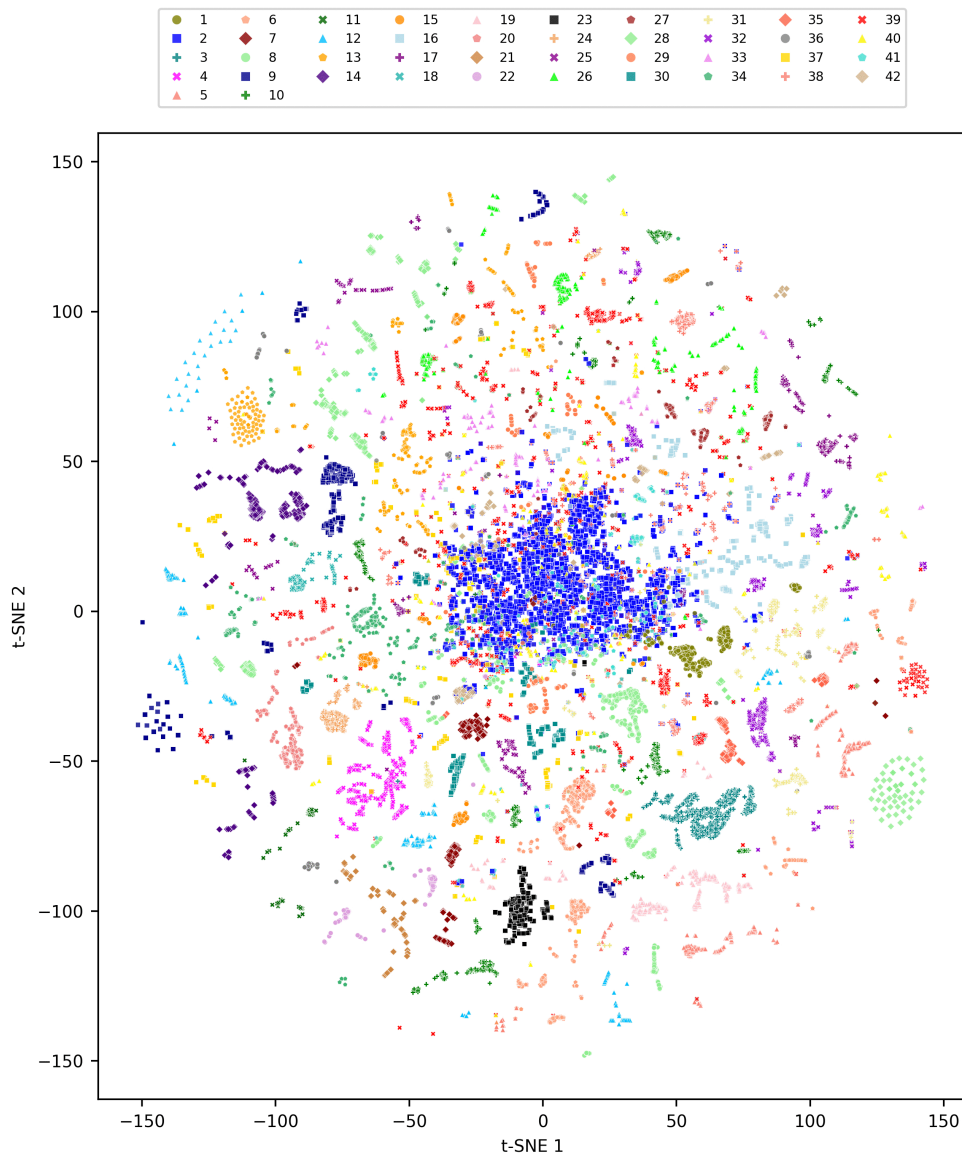


Fig 4. t-SNE visualization of clustering results. Species are visualized in a two-dimensional functional space using t-Distributed Stochastic Neighbor Embedding (t-SNE), a probabilistic technique that reduces high-dimensional data (originally 18 dimensions) to two dimensions while preserving the local structure and similarity relationships of the data points. Different colors and markers represent distinct groups, chosen to maximize visual distinction among clusters based on their distribution. The distances in this 2D space are probabilistic, reflecting the likelihood of similarity between species, and approximate the original high-dimensional probability distribution. Note that t-SNE is intended for visualisation of the shape and location of clusters, but it does not have the same interpretation as other dimensionality reduction techniques such as PCA (see [S5](#) and [S6 Figs](#) for alternate visualisations). Please also see the interactive version of this figure in [S1 Appendix](#), or available from the GitHub repository.

<https://doi.org/10.1371/journal.pcbi.1014278.g004>

The majority of the clusters are comprised of species that exhibit functionally unique combinations of traits, often aligning with well known taxonomic structures (e.g., within genus or family). However, we confidently identified one large, taxonomically broad group consisting of 4,291 species with non-distinct or ‘average’ traits, comprising taxa from across many families and orders ([Fig 4](#); see interactive version in [S1 Appendix](#)). Other taxonomically dispersed groups, regardless of their size, show low consensus values ([S1 Table](#)). Whether this large dispersed group represents ecological generalists,

reflects convergent evolution of common trait combinations, or captures species not readily assigned to distinct clusters is a compelling question for future work [81–84].

Group trait patterns

When exploring the trait patterns underpinning the groups, we see that they are not determined from just one or two traits, but rather all traits appear to be relevant in the identification of different functional groups (Fig 5, S2 Table). The lowest standard deviation across groups is for stem diameter at 70% of the original variation, while the highest is for leaf density at 99%. If a trait was not relevant, we would expect that there is no pattern to the assignment across that particular trait, meaning the standard deviation across groups would be much closer to 0. These results reinforce the idea that group 2 acts as a generalist group, with average values across all traits and low within-group variation, while groups 38–42 also show average trait values, but with much higher variation across species (S7 Fig).

Functional groups to quantify functional diversity

The value of a functional group approach lies in the fact that functional groups provide a natural analogue of species, allowing the entire suite of species-based diversity metrics to be applied to functional groups. At the simplest level, functional group richness (the number of unique groups in a community) provides an intuitive metric for quantifying functional diversity that doesn't require complex metrics or calculations. On the other hand, Simpson's Index—a widely used measure of species-based diversity—can likewise be applied to functional groups, in which case it provides a direct measure of 'functional redundancy', i.e., the probability that any two randomly selected species in a community belong to the same functional group. Indeed, any species-based metric can be applied directly to functional group richness, providing relatively simple and intuitive measures of richness, redundancy, divergence, and evenness.

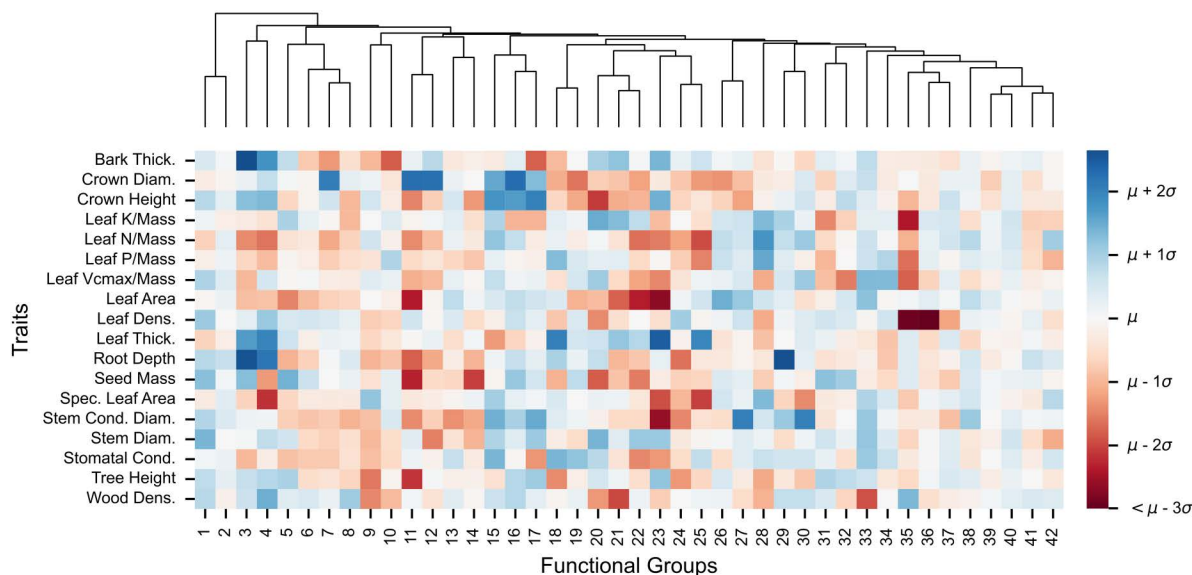


Fig 5. Centroids of 42 functional groups. Mean values were calculated for each group across their species after normalizing the original dataset for each trait. The colour gradient represents these normalized values and is consistent across all traits, with values clipped at 3 standard deviations from the mean to enhance visualization contrast. A dendrogram, constructed using the Ward-linkage method, is included to illustrate the relative consensus distances among the groups.

<https://doi.org/10.1371/journal.pcbi.1014278.g005>

To illustrate how such metrics can be applied to functional groups, we use a recent global dataset of tree community composition [12], which contains presence-absence of around 37,000 tree species at 200-km resolution across the globe. We used our tree-species functional group classification to assign each tree species to a functional group. We then calculated functional redundancy ('FRedund', Simpson's Index) and functional group richness ('FGR', number of groups) across all species contained within each pixel. We then compared these group-derived metrics with three traditional diversity metrics reported in Paz et al.: species richness (number of unique species), mean pairwise distance (Rao's quadratic entropy, a measure of functional clustering), and functional richness (convex hull volume across all traits) (see Fig 6).

These results demonstrate how functional group diversity complements and reflects key elements of traditional metrics, while also providing novel insight into the functional structure of natural communities. After accounting for spatial autocorrelation, Functional Group Richness (FGR) exhibits a saturating positive relationship with traditional species richness (Pearson $r=0.58$, $p<0.001$; Fig 6a), which flattens after ~500 species, and likewise shows a strong linear correlation with functional richness ($r=0.74$, $p<0.001$). Conversely, FGR has weak residual correlation with Rao's Q ($r=0.15$, $p<0.001$;

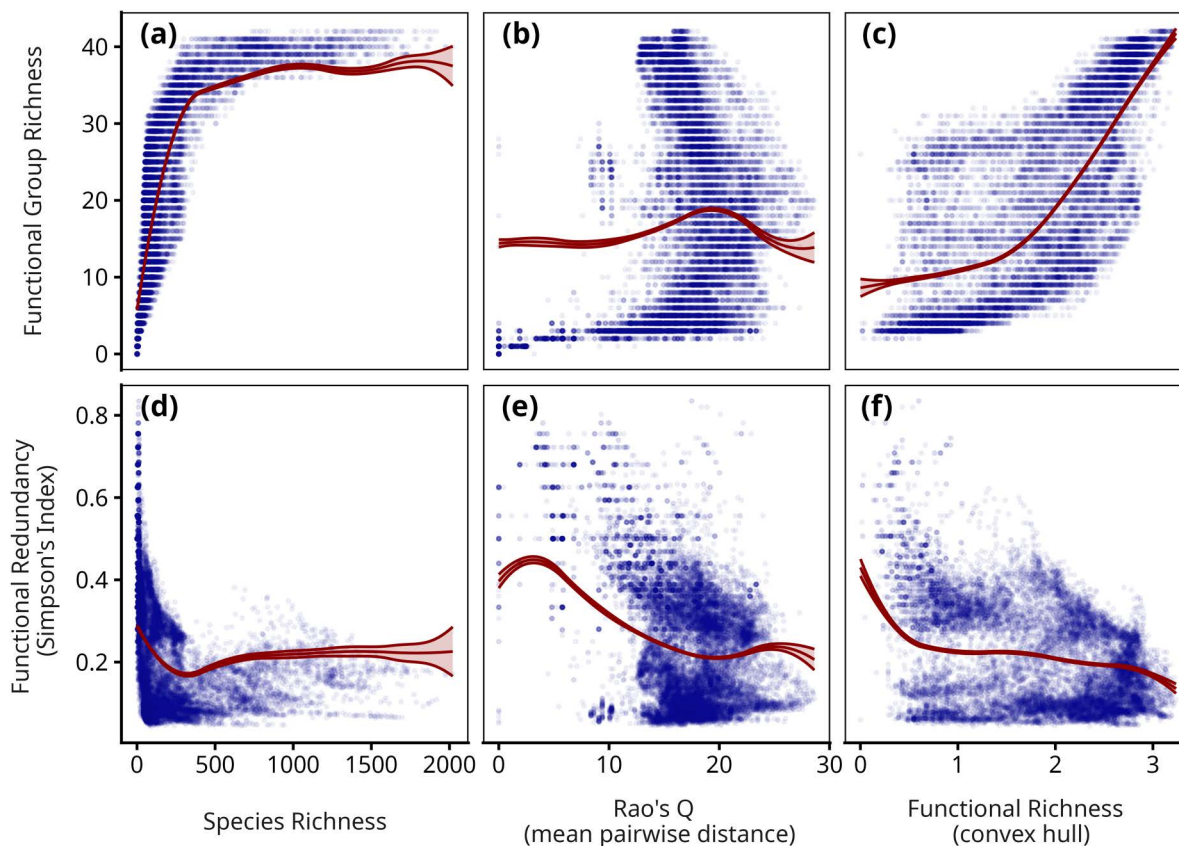


Fig 6. Functional group diversity metrics compared to traditional measures of species richness and functional diversity, calculated for each of $n=16,048$ forested pixels across the globe (with more than 3 clusters), as reported in Paz et al. [12]. The points show the raw observations, with the lines denoting the fitted trends from generalised additive models, adjusted for spatial autocorrelation (see Methods), along with 95% confidence intervals. (a-c) Functional group richness (the number of unique groups) is highly correlated with species richness at low levels, but saturates after approximately 500 species in the system. Functional redundancy is weakly correlated with Rao's Q, which reflects spatial clustering in trait space, but, compared to functional redundancy exhibits a stronger (albeit still relatively weak) correlation with functional richness. (d-f) Functional redundancy (Simpson's Index applied to groups) is inversely correlated with species richness, revealing significant variation in redundancy at high levels of richness, but is only weakly correlated with Rao's Q and Functional richness, capturing complementary aspects of both. These results illustrate how traditional diversity metrics can be applied to functional groups, yielding intuitive and simple metrics for quantifying functional composition.

<https://doi.org/10.1371/journal.pcbi.1014278.g006>

[Fig 6b](#)), highlighting that this metric is largely independent of distance-based similarity. Functional Redundancy, on the other hand, aims to capture overlap between species, and accordingly exhibits a moderate negative correlation with Rao's Q ($r = -0.40$, $p < 0.001$; [Fig 6e](#)), along with a negligible correlation with species richness ($r = -0.04$, $p < 0.001$) and a slight negative relationship with functional richness ($r = -0.21$, $p < 0.001$; [Fig 6f](#)). Collectively, these results illustrate how group-based diversity metrics can provide complementary insight into functional structure by integrating different aspects of species richness, functional clustering, and functional breadth (also see [S15 Fig](#) for spatial projections). Moreover, the simplicity of group-based calculations helps to overcome some of the computational limitations of existing functional diversity metrics (e.g., the inability to calculate convex hull volume in communities where there are fewer species than traits) while also being based on well-known, interpretable species-diversity metrics.

Discussion

Here, we present a clustering framework that enables the identification of functional groups while explicitly accounting for trait uncertainty and trait correlation. By generating multiple scenarios of a trait dataset and clustering each independently, we capture the variability inherent in imputed trait values or in the presence of intraspecific variation. By aggregating these results into a consensus matrix and applying hierarchical clustering, we are able to identify robust functional groups and their corresponding level of certainty (consensus). In doing this, this approach increases agreement across bootstrap samples and reduces variation relative to a single clustering approach ([S4 Fig](#)). When applied to a global dataset of tree species, this approach identified 42 functional groups, most of which exhibited moderate to high consensus. We also show how these groups can be interpreted in terms of their average trait composition and linked to other biodiversity metrics across space.

In terms of our methodology, our main novelty relied on the use of consensus clustering with resamples of the prediction error and the optimization of groups in each step of consensus clustering. Consensus clustering has been used before, but mainly to reach consensus between different clustering methods, or by using bootstrapping in the original dataset [[49,85](#)]. To our knowledge, these techniques have not been used to solve the problem of trait uncertainty. Other methods have been proposed to solve this issue, like incorporating trait uncertainty into a Mixture Model [[86–88](#)], but the complexity quickly scales, and it becomes practically infeasible when we also have uncertainty in the number of groups. Our model provides a novel, scalable solution to addressing the issues of trait uncertainty and trait collinearity, especially suited for datasets constructed using imputed traits, as is the case in nearly all large trait databases.

In the case study, our consensus clustering approach resulted in the identification of 42 functional groups of tree species worldwide, of which 36 demonstrated an average consensus greater than 20%, indicating a good level of agreement, and 9 of these groups had a consensus exceeding 50%, reflecting high agreement. The groups with high consensus exhibited complex trait patterns, suggesting that all traits play a relevant role at some partition level. Conversely, the 6 groups with less than 20% consensus underscore the significance of accounting for uncertainty and highlight the need for further data collection. Our findings also reveal that while functional groups share some overlap with taxonomic groups, the observed differences in phylogenetic distance and consensus underscore the importance of distinguishing between taxonomic and functional classifications. While we only consider global trait patterns, a clear extension is to rerun these methods on local scales to identify functional (sub)groups within specific regions. For example, using only species from tropical rainforests or boreal forests could reveal detailed functional patterns, overlooked at a global scale, due to the distinct structures of resource competition in each region [[9,66,89,90](#)].

The key benefit of a functional group approach is that it allows us to directly apply biodiversity metrics constructed for taxonomic data, such as species richness, Simpson index, and Shannon-Wiener index [[36,37,91](#)]. In addition, the complete consensus matrix allows us to apply continuous diversity metrics designed for phylogenetic-like dendrograms, such as Faith's phylogenetic diversity [[7](#)]. In doing this, the method and framework proposed here are intended to overcome some of the challenges in adopting functional diversity metrics in conservation and restoration, by providing simple and

intuitive metrics that do not rely on having access to the original trait database or expert knowledge of the dozens of different functional diversity metrics. As a simple example, when doing species selection for reforestation, practitioners could directly use the list of tree functional groups to select focal species to ensure that their target composition contains a diversity of functional roles and a low level of functional redundancy. Similarly, current efforts to quantify and track biodiversity are largely limited to measures of species richness, but these species-based metrics can easily be converted into functional diversity metrics by applying functional group labels to common diversity calculations. Because a list of functional groups is simple to share and access, such approaches are repeatable and transparent, facilitating adoption and comparison of results across taxa and study sites.

An important distinction between our approach and traditional functional group classification is that our method is unsupervised, as opposed to the original approach of assigning groups based on expert knowledge [20]. Thus, there is no ground truth for the resulting group assignments, and while this method provides a data-driven approach that minimizes human bias, it also means that the resulting clusters may not reflect the ecological question of interest. Indeed, we used Ward's linkage method in our case study to identify compact spherical clusters, in line with expectations from niche theory [58,33]; other linkage methods that likewise favor compact groups, such as average linkage, would be expected to yield similar results to Ward's (S13 Fig). However, alternate approaches may work best in other ecological settings. For example, if one expects ecologically relevant clusters to track environmental gradients, then single linkage might be better suited due to its ability to chain together clusters [92].

Similarly, we use BIC-based cluster-size selection when a likelihood is available (e.g., for individual cluster selection across the resamples, Fig 1a), and otherwise we use silhouette scores in the absence of a likelihood (e.g., for the consensus matrix, Fig 1c). In practice, any cluster-size selection method is valid, and these criteria can and will yield different results and cluster assignments, such that the user is encouraged to compare approaches and tailor these decisions to their application and question of interest. For example, here, by comparing the functional groups to underlying phylogenetic structure or trait distribution (e.g., Fig 5), we show how the resulting groups can be linked to known ecological roles and species labels, facilitating group interpretation and aiding in downstream analysis. Future steps to validate these groups could include incorporating known categories, such as the presence of species in different biomes [93,94], functional roles labelled by experts [69,95], temperature and precipitation niches [96,97], and environmental tolerances [83,98].

As with any unsupervised clustering method, the choice of clustering algorithm is an important decision. For building a clustering ensemble, we used Gaussian mixture models for their ability to handle multivariate trait correlation, which makes them less sensitive to trait collinearity or trait omission. However, such models still assume a specific cluster distribution that can be hard to validate in practice [53]. An alternative approach would be a non-parametric, density-based method like HDBSCAN [99], which may perform better when functional groups are complex and non-spherical, but its sensitivity to observation distance can lead to unstable clustering results (S11 and S12 Figs). Furthermore, it is not clear that a density-based cluster would reflect what we understand as a functional group, since these methods allow for species very far apart in functional space to be classified as the same group if they are connected by other species (e.g., a horseshoe-shaped pattern where all species are in the same group). Thus, future work could explore density-based perturbations to improve HDBSCAN's stability [100], and comparison of such approaches to distance-based methods, which could aid in the identification of novel groups and ecological similarities across species. On the other hand, a key benefit of the hierarchical consensus clustering approach used here is that it naturally allows for further subdivisions, such that these groups could be further partitioned into smaller subgroups and compared to deeper taxonomic levels like family and genus. The method presented here is thus intended as a general and flexible framework for assigning functional groups, which can be tailored to the specific question of interest.

Conclusion

Collectively, our study presents a novel framework for functional group classification that addresses trait uncertainty and trait correlation. By applying this method to a case study of global tree traits, we show how this approach can be used to

identify ecologically meaningful clusters and the traits underpinning cluster assignment. By applying traditional species-based diversity metrics to functional groups, we further illustrate how functional groups can be used to generate interpretable, intuitive functional diversity metrics that allow for scalable and repeatable comparisons across study systems. In doing so, this approach provides a scalable method for incorporating functional biodiversity into restoration or conservation activities, and for quantifying and tracking changes in biodiversity through time.

Supporting information

S1 Table. Detail of group size and group average consensus for each of the functional groups as displayed in Fig 3.

(PDF)

S2 Table. Detailed trait statistics for cluster centroids after standardizing each trait as shown in Fig 5. Mean values near 0 and standard deviation near 1 indicates that it recreates the original distribution.

(PDF)

S1 Appendix. Interactive t-SNE. Interactive version of Fig 4 with trait axes and species names. Trait loadings on each axis were calculated using multivariate regression, with the figure displaying only those traits with partial correlation $r > 0.3$. (HTML)

S1 Fig. Correlation matrix for all 18 traits. Correlation matrix for all 18 traits (after natural log), with a color scale where blue denotes positive correlations and red signifies negative correlations. Most traits exhibit significant correlations, highlighting the necessity of accounting for these relationships in clustering and distance analyses.

(PDF)

S2 Fig. Trait distribution across the dataset. Trait distribution for the complete dataset of 18 traits (after natural log transformation), showing spikes due to repeated imputation. The blend of continuous and discrete distributions underscores the issue of trait uncertainty in clustering algorithms.

(PDF)

S3 Fig. Error distributions for traits. Distribution of errors across traits for angiosperms (a) and gymnosperms (b). Errors are calculated by subtracting log predicted values from log observed values for species with available data. The x-axis represents logged trait error values, and the y-axis represents frequency.

(PNG)

S4 Fig. Adjusted Rand Index Distribution. We evaluate clustering consistency on a subsample of 1,365 species (including all gymnosperms to ensure structural heterogeneity) using two approaches. In the first approach (blue), a Gaussian Mixture Model (GMM) is applied independently to each of 500 resampled datasets, producing 500 clustering results. The Adjusted Rand Index (ARI) is then computed for all pairwise comparisons among these clusterings, reflecting the variability of the method across resamples. In the second approach (orange), the 500 resampled clusterings are partitioned into 10 groups of 50. Within each group, a consensus clustering is derived using the full consensus framework, yielding 10 consensus clusterings. The ARI is then computed for all pairwise comparisons among these consensus results, providing a measure of agreement after aggregation.

(PNG)

S5 Fig. PCA clustering results. Functional groups identified through Principal Component Analysis (PCA) for dimensionality reduction. PCA transforms the data into principal components that capture the most variance, revealing considerable group overlap and clustering, particularly away from the center of the distribution.

(PNG)

S6 Fig. UMAP clustering results. Functional groups identified using Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. UMAP reduces data complexity by preserving local and global structures, with groups outside the distribution center showing greater dispersion.

(PNG)

S7 Fig. Group-level standard deviation of traits. Standard deviation of log-standardized trait values within each functional group. Higher values indicate greater within-group variation.

(PDF)

S8 Fig. Taxonomic group composition by functional groups. Each taxonomic group composition in terms of functional groups, with proportions displayed for each taxonomic group. Rows sum to 1.

(PDF)

S9 Fig. Functional group composition by taxonomic groups. Composition of functional groups by taxonomic groups, showing the proportion of each functional group belonging to specific taxonomic groups. Rows sum to 1.

(PDF)

S10 Fig. GMM results without resampling. Visualization of clusters in t-SNE space, where the central region suggests overfitting due to repeated observations. The optimal BIC is observed at 800 components.

(PNG)

S11 Fig. HDBSCAN results on original data. Clustering using HDBSCAN with a minimum cluster size of 10 species per group. Black circles indicate noise points. Overfitting issues due to imputed values are evident.

(PNG)

S12 Fig. HDBSCAN results for consensus clustering. t-SNE visualization for HDBSCAN clusters, showing two main clusters and challenges in identifying smaller groups.

(PNG)

S13 Fig. Average linkage clustering results. The t-SNE visualization obtained using the average linkage method yields 34 clusters, compared to 42 clusters produced by the Ward method. This discrepancy highlights the sensitivity of clustering outcomes to the choice of linkage criterion, which should be selected in accordance with the underlying assumptions about cluster structure. Notably, the adjusted Rand index between the two clustering solutions is 0.51, indicating a moderate level of agreement and a substantial overlap in the resulting partitions.

(PNG)

S14 Fig. Species-level stability. Histogram of species stability values computed from the consensus matrix, where stability is defined as the ratio between the average similarity of a species to members of its assigned cluster and its average similarity to all species. Higher values indicate more consistent co-clustering across resamples.

(PNG)

S15 Fig. Global projections of functional group metrics. (a) Functional Richness is the number of functional groups per pixel (200 km nominal diameter hexagonal bins), and (b) Functional Redundancy is Simpson's Index applied to the functional groups, here shown on the log scale for better spatial resolution. Compare to the results of Paz et al. [12].

(PNG)

Author contributions

Conceptualization: Pablo Ubilla Pavez, Andrea Paz, Daniel Maynard.

Data curation: Andrea Paz, Daniel Maynard.

Formal analysis: Pablo Ubilla Pavez, Andrea Paz, Daniel Maynard.

Funding acquisition: Daniel Maynard.

Investigation: Pablo Ubilla Pavez, Daniel Maynard.

Methodology: Pablo Ubilla Pavez, Andrea Paz, Daniel Maynard.

Project administration: Daniel Maynard.

Supervision: Daniel Maynard.

Visualization: Pablo Ubilla Pavez, Daniel Maynard.

Writing – original draft: Pablo Ubilla Pavez, Daniel Maynard.

Writing – review & editing: Pablo Ubilla Pavez, Andrea Paz, Daniel Maynard.

References

1. Díaz S, Zafra-Calvo N, Purvis A, Verburg PH, Obura D, Leadley P, et al. Set ambitious goals for biodiversity and sustainability. *Science*. 2020;370(6515):411–3. <https://doi.org/10.1126/science.abe1530> PMID: 33093100
2. Buckland ST, Studený AC, Magurran AE, Illian JB, Newson SE. The geometric mean of relative abundance indices: a biodiversity measure with a difference. *Ecosphere*. 2011;2(9):art100. <https://doi.org/10.1890/es11-00186.1>
3. Mace GM, Baillie JEM. The 2010 biodiversity indicators: challenges for science and policy. *Conserv Biol*. 2007;21(6):1406–13. <https://doi.org/10.1111/j.1523-1739.2007.00830.x> PMID: 18173464
4. Humphries CJ, Williams PH, Vane-Wright RI. Measuring biodiversity value for conservation. *Annu Rev Ecol Syst*. 1995;26(1):93–111. <https://doi.org/10.1146/annurev.es.26.110195.000521>
5. Pennell MW, FitzJohn RG, Cornwell WK. A simple approach for maximizing the overlap of phylogenetic and comparative data. *Methods Ecol Evol*. 2016;7(6):751–8. <https://doi.org/10.1111/2041-210x.12517>
6. Srivastava DS, Cadotte MW, MacDonald AAM, Marushia RG, Mirotchnick N. Phylogenetic diversity and the functioning of ecosystems. *Ecol Lett*. 2012;15(7):637–48. <https://doi.org/10.1111/j.1461-0248.2012.01795.x> PMID: 22583836
7. Faith DP. Conservation evaluation and phylogenetic diversity. *Biol Conserv*. 1992;61(1):1–10. [https://doi.org/10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3)
8. Freitas JR, Mantovani W. An overview of the applicability of functional diversity in Biological Conservation. *Braz J Biol*. 2018;78(3):517–24. <https://doi.org/10.1590/1519-6984.09416> PMID: 29069162
9. Kraft NJB, Valencia R, Ackerly DD. Functional traits and niche-based tree community assembly in an Amazonian forest. *Science*. 2008;322(5901):580–2. <https://doi.org/10.1126/science.1160662> PMID: 18948539
10. Cadotte MW, Carscadden K, Mirotchnick N. Beyond species: functional diversity and the maintenance of ecological processes and services. *J Appl Ecol*. 2011;48(5):1079–87. <https://doi.org/10.1111/j.1365-2664.2011.02048.x>
11. Kearney M, Simpson SJ, Raubenheimer D, Helmuth B. Modelling the ecological niche from functional traits. *Philos Trans R Soc Lond B Biol Sci*. 2010;365(1557):3469–83. <https://doi.org/10.1098/rstb.2010.0034> PMID: 20921046
12. Paz A, Crowther TW, Maynard DS. Functional and phylogenetic dimensions of tree biodiversity reveal unique geographic patterns. *Glob Ecol Biogeogr*. 2024;33(9). <https://doi.org/10.1111/geb.13877>
13. Sterk M, Gort G, Klimkowska A, van Ruijven J, van Teeffelen AJA, Wamelink GWW. Assess ecosystem resilience: Linking response and effect traits to environmental variability. *Ecol Indic*. 2013;30:21–7. <https://doi.org/10.1016/j.ecolind.2013.02.001>
14. Schneider FD, Fichtmueller D, Gossner MM, Güntsch A, Jochum M, König-Ries B, et al. Towards an ecological trait-data standard. *Methods Ecol Evol*. 2019;10(12):2006–19. <https://doi.org/10.1111/2041-210x.13288>
15. Bussotti F, Pollastrini M, Holland V, Brüggemann W. Functional traits and adaptive capacity of European forests to climate change. *Environ Experim Bot*. 2015;111:91–113. <https://doi.org/10.1016/j.envexpbot.2014.11.006>
16. Heilmeyer H. Functional traits explaining plant responses to past and future climate changes. *Flora*. 2019;254:1–11. <https://doi.org/10.1016/j.flora.2019.04.004>
17. Gross N, Kunstler G, Liancourt P, De Bello F, Suding KN, Lavorel S. Linking individual response to biotic interactions with community structure: a trait-based framework. *Funct Ecol*. 2009;23(6):1167–78. <https://doi.org/10.1111/j.1365-2435.2009.01591.x>
18. Adler PB, Fajardo A, Kleinhesselink AR, Kraft NJB. Trait-based tests of coexistence mechanisms. *Ecol Lett*. 2013;16(10):1294–306. <https://doi.org/10.1111/ele.12157> PMID: 23910482
19. McGrady-Steed J, Morin PJ. Biodiversity, and the dynamics of populations and functional groups. *Ecology*. 2000;81(2):361–73. <https://doi.org/10.2307/177433>
20. Blondel J. Guilds or functional groups: does it matter? *Oikos*. 2003;100(2):223–31. <https://doi.org/10.1034/j.1600-0706.2003.12152.x>

21. Pokorný ML, Sheley RL, Zabinski CA, Engel RE, Svejcar TJ, Borkowski JJ. Plant Functional Group Diversity as a Mechanism for Invasion Resistance. *Restor Ecol*. 2005;13(3):448–59. <https://doi.org/10.1111/j.1526-100x.2005.00056.x>
22. Legras G, Loiseau N, Gaertner J-C. Functional richness: Overview of indices and underlying concepts. *Acta Oecologica*. 2018;87:34–44. <https://doi.org/10.1016/j.actao.2018.02.007>
23. Cadotte MW, Cavender-Bares J, Tilman D, Oakley TH. Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. *PLoS One*. 2009;4(5):e5695. <https://doi.org/10.1371/journal.pone.0005695> PMID: 19479086
24. Laliberté E, Legendre P. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*. 2010;91(1):299–305. <https://doi.org/10.1890/08-2244.1> PMID: 20380219
25. Mouchet MA, Villéger S, Mason NWH, Mouillot D. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Funct Ecol*. 2010;24(4):867–76. <https://doi.org/10.1111/j.1365-2435.2010.01695.x>
26. Legras G, Loiseau N, Gaertner JC, Poggiale JC, Gaertner-Mazouni N. Assessing Functional Diversity: The Influence of the Number of the Functional Traits. *Theor Ecol*. 2020;13(1):117–26. <https://doi.org/10.1007/s12080-019-00433-x>
27. Ohlert T, Kimmel K, Avolio M, Chang C, Forrester E, Gerstner B, et al. Exploring the impact of trait number and type on functional diversity metrics in real-world ecosystems. *PLoS One*. 2022;17(8):e0272791. <https://doi.org/10.1371/journal.pone.0272791> PMID: 36006866
28. Schmeller DS, Weatherdon LV, Loyau A, Bondeau A, Brotons L, Brummitt N, et al. A suite of essential biodiversity variables for detecting critical biodiversity change. *Biol Rev Camb Philos Soc*. 2018;93(1):55–71. <https://doi.org/10.1111/brev.12332> PMID: 28447398
29. Proença V, Martin LJ, Pereira HM, Fernandez M, McRae L, Belnap J, et al. Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. *Biol Conserv*. 2017;213:256–63. <https://doi.org/10.1016/j.biocon.2016.07.014>
30. Burgess ND, Ali N, Bedford J, Bhola N, Brooks S, Cierna A, et al. Global Metrics for Terrestrial Biodiversity. *Ann Rev Environ Resour*. 2024;49(1):673–709. <https://doi.org/10.1146/annurev-environ-121522-045106>
31. Sundstrom SM, Allen CR, Barichiev C. Species, Functional Groups, and Thresholds in Ecological Resilience. *Conserv Biol*. 2012;26(2):305–14. <https://doi.org/10.1111/j.1523-1739.2011.01822.x>
32. Zhu L, Fu B, Zhu H, Wang C, Jiao L, Zhou J. Trait choice profoundly affected the ecological conclusions drawn from functional diversity measures. *Sci Rep*. 2017;7(1):3643. <https://doi.org/10.1038/s41598-017-03812-8> PMID: 28623286
33. Calbi M, Boenisch G, Boulangeat I, Bunker D, Catford JA, Changenet A, et al. A novel framework to generate plant functional groups for ecological modelling. *Ecol Indic*. 2024;166:112370. <https://doi.org/10.1016/j.ecolind.2024.112370>
34. Mouillot D, Villéger S, Parravicini V, Kulbicki M, Arias-González JE, Bender M. Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. *Proc Natl Acad Sci*. 2014;111(38):13757–62. <https://doi.org/10.1073/pnas.1317625111>
35. Villéger S, Mason NWH, Mouillot D. New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology*. 2008;89(8):2290–301. <https://doi.org/10.1890/07-1206.1> PMID: 18724739
36. Skidmore AK, Coops NC, Neinavaz E, Ali A, Schaepman ME, Paganini M, et al. Priority list of biodiversity metrics to observe from space. *Nat Ecol Evol*. 2021;5(7):896–906. <https://doi.org/10.1038/s41559-021-01451-x> PMID: 33986541
37. Keylock CJ. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos*. 2005;109(1):203–7. <https://doi.org/10.1111/j.0030-1299.2005.13735.x>
38. Rokach L, Maimon O. Clustering Methods. In: Maimon O, Rokach L, editors. *Data Mining and Knowledge Discovery Handbook*. Boston (MA): Springer US; 2005. p. 321–52.
39. Williams WT. Principles of Clustering. *Ann Rev Ecol Syst*. 1971;2:303–26.
40. von Luxburg U, Williamson RC, Guyon I. Clustering: Science or Art? In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. JMLR Workshop and Conference Proceedings. 2012. p. 65–79.
41. Madhulatha TS. An overview on clustering methods. *arXiv preprint arXiv:12051117*. 2012.
42. Nugent R, Meila M. An overview of clustering applied to molecular biology. *Statistical methods in molecular biology*. 2010; p. 369–404.
43. Schrod F, Kattge J, Shan H, Fazayeli F, Joswig J, Banerjee A, et al. BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Glob Ecol Biogeogr*. 2015;24(12):1510–21. <https://doi.org/10.1111/geb.12335>
44. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–8. <https://doi.org/10.1093/bioinformatics/btr597> PMID: 22039212
45. Janžeković F, Novak T. PCA—a powerful method for analyze ecological niches. *Principal component analysis—multidisciplinary applications*. 2012. p. 127–42.
46. Kurita T. Principal component analysis (PCA). *Computer vision: a reference guide*. 2019; p. 1–4.
47. Bank D, Koenigstein N, Giryas R. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*. 2023. p. 353–74.
48. Borowiec ML, Dikow RB, Frandsen PB, McKeen A, Valentini G, White AE. Deep learning as a tool for ecology and evolution. *Methods Ecol Evol*. 2022;13(8):1640–60.

49. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach Learn*. 2003;52(1–2):91–118. <https://doi.org/10.1023/a:1023949509487>
50. Morlini I. A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Adv Data Anal Classif*. 2012;6(1):5–28. <https://doi.org/10.1007/s11634-011-0101-z>
51. Reynolds DA, et al. Gaussian mixture models. *Encyclopedia of biometrics*. 2009;741(659–663).
52. Ng SK, Krishnan T, McLachlan GJ. The EM Algorithm. In: Gentle JE, Härdle WK, Mori Y, editors. *Handbook of Computational Statistics: Concepts and Methods*. Berlin, Heidelberg: Springer; 2012. p. 139–72.
53. McLachlan GJ, Rathnayake S. On the number of components in a Gaussian mixture model. *WIREs Data Min Knowl*. 2014;4(5):341–55. <https://doi.org/10.1002/widm.1135>
54. Shahapure KR, Nicholas C. Cluster Quality Analysis Using Silhouette Score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). 2020. p. 747–748.
55. Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conf Ser: Mater Sci Eng*. 2018;336:012017. <https://doi.org/10.1088/1757-899x/336/1/012017>
56. Kreft H, Jetz W. A framework for delineating biogeographical regions based on species distributions. *J Biogeogr*. 2010;37(11):2029–53. <https://doi.org/10.1111/j.1365-2699.2010.02375.x>
57. Chakerian J, Holmes S. Computational Tools for Evaluating Phylogenetic and Hierarchical Clustering Trees. *J Comput Graph Stat*. 2012;21(3):581–99. <https://doi.org/10.1080/10618600.2012.640901> PMID: [32982128](https://pubmed.ncbi.nlm.nih.gov/32982128/)
58. Lu M, Winner K, Jetz W. A unifying framework for quantifying and comparing n-dimensional hypervolumes. *Methods Ecol Evol*. 2021;12(10):1953–68. <https://doi.org/10.1111/2041-210x.13665>
59. Everitt B. *The cambridge dictionary of statistics*. Cambridge, UK: Cambridge University Press. 1998. 5 p.
60. Chave J, Coomes D, Jansen S, Lewis SL, Swenson NG, Zanne AE. Towards a worldwide wood economics spectrum. *Ecol Lett*. 2009;12(4):351–66. <https://doi.org/10.1111/j.1461-0248.2009.01285.x> PMID: [19243406](https://pubmed.ncbi.nlm.nih.gov/19243406/)
61. Maracahipes L, Carlucci MB, Lenza E, Marimon BS, Marimon BH, Guimarães FAG. How to Live in Contrasting Habitats? Acquisitive and Conservative Strategies Emerge at Inter- and Intraspecific Levels in Savanna and Forest Woody Plants. *Perspect Plant Ecol Evol Syst*. 2018;34:17–25. <https://doi.org/10.1016/j.ppees.2018.07.006>
62. Kattge J, Bönsch G, Díaz S, Lavorel S, Prentice IC, Leadley P, et al. TRY Plant Trait Database – Enhanced Coverage and Open Access. *Glob Change Biol*. 2020;26(1):119–88. <https://doi.org/10.1111/gcb.14904>
63. Petit RJ, Hampe A. Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst*. 2006;37(1):187–214. <https://doi.org/10.1146/annurev.ecolsys.37.091305.110215>
64. Joswig JS, Wirth C, Schuman MC, Kattge J, Reu B, Wright IJ, et al. Climatic and soil factors explain the two-dimensional spectrum of global plant trait variation. *Nat Ecol Evol*. 2022;6(1):36–50. <https://doi.org/10.1038/s41559-021-01616-8> PMID: [34949824](https://pubmed.ncbi.nlm.nih.gov/34949824/)
65. Díaz S, Kattge J, Cornelissen JHC, Wright IJ, Lavorel S, Dray S, et al. The Global Spectrum of Plant Form and Function. *Nature*. 2016;529(7585):167–71.
66. Augusto L, Boča A. Tree functional traits, forest biomass, and tree species diversity interact with site properties to drive forest soil carbon. *Nat Commun*. 2022;13(1):1097. <https://doi.org/10.1038/s41467-022-28748-0> PMID: [35233020](https://pubmed.ncbi.nlm.nih.gov/35233020/)
67. Niinemets Ü. The controversy over traits conferring shade-tolerance in trees: ontogenetic changes revisited. *J Ecol*. 2006;94(2):464–70. <https://doi.org/10.1111/j.1365-2745.2006.01093.x>
68. Diaz S, Cabido M. Plant functional types and ecosystem function in relation to global change. *J Veg Sci*. 1997;8(3):463–74. <https://doi.org/10.1111/j.1654-1103.1997.tb00842.x>
69. Valladares F, Niinemets Ü. Shade tolerance, a key plant feature of complex nature and consequences. *Annu Rev Ecol Evol Syst*. 2008;39(1):237–57. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173506>
70. Maynard DS, Bialic-Murphy L, Zohner CM, Averill C, van den Hoogen J, Ma H, et al. Global Relationships in Tree Functional Traits. *Nat Commun*. 2022;13(1):3185.
71. Cazzolla Gatti R, Reich PB, Gamarra JGP, Crowther T, Hui C, Morera A. The Number of Tree Species on Earth. *Proc Natl Acad Sci U S A*. 2022;119(6):e2115329119. <https://doi.org/10.1073/pnas.2115329119>
72. Díaz S, Kattge J, Cornelissen JHC, Wright IJ, Lavorel S, Dray S. The Global Spectrum of Plant Form and Function. *Nature*. 2015;529(7585):1–17. <https://doi.org/10.1038/nature16489>
73. Pérez-Harguindeguy N, Díaz S, Garnier E, Lavorel S, Poorter H, Jaureguiberry P. Corrigendum to: New Handbook for Standardised Measurement of Plant Functional Traits Worldwide. *Aust J Bot*. 2016;64(8):715–6. https://doi.org/10.1071/bt12225_co
74. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(11).
75. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*. 2018.

76. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
77. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–72. <https://doi.org/10.1038/s41592-019-0686-2> PMID: 32015543
78. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editors. *Proceedings of the 9th Python in Science Conference*. 2010. p. 56–61.
79. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9(3):90–5. <https://doi.org/10.1109/mcse.2007.55>
80. Waskom M. seaborn: statistical data visualization. *JOSS*. 2021;6(60):3021. <https://doi.org/10.21105/joss.03021>
81. Marinšek A, Čarni A, Šilc U, Manthey M. What makes a plant species specialist in mixed broad-leaved deciduous forests? *Plant Ecol*. 2015;216(10):1469–79. <https://doi.org/10.1007/s11258-015-0527-z>
82. Chisholm RA, Pacala SW. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proc Natl Acad Sci U S A*. 2010;107(36):15821–5. <https://doi.org/10.1073/pnas.1009387107> PMID: 20733073
83. Boulangeat I, Lavergne S, Van Es J, Garraud L, Thuiller W. Niche breadth, rarity and ecological characteristics within a regional flora spanning large environmental gradients. *J Biogeogr*. 2011;39(1):204–14. <https://doi.org/10.1111/j.1365-2699.2011.02581.x>
84. Kluge J, Kessler M. Influence of niche characteristics and forest type on fern species richness, abundance and plant size along an elevational gradient in Costa Rica. *Plant Ecol*. 2011;212(7):1109–21. <https://doi.org/10.1007/s11258-010-9891-x>
85. Zhang Y, Li T. Extending Consensus Clustering to Explore Multiple Clustering Views. In: *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics; 2011. p. 920–931.
86. Abrevaya J, Donald SG. A GMM Approach for dealing with missing data on regressors. *Rev Econ Stat*. 2017;99(4):657–62. https://doi.org/10.1162/rest_a_00645 PMID: 39267948
87. McCaw ZR, Aschard H, Julienne H. Fitting Gaussian Mixture Models on Incomplete Data. *BMC Bioinform*. 2022;23(1):208. <https://doi.org/10.1186/s12859-022-04740-9>
88. Coleman S, Kirk PDW, Wallace C. Consensus clustering for Bayesian mixture models. *BMC Bioinform*. 2022;23(1):290. <https://doi.org/10.1186/s12859-022-04830-8> PMID: 35864476
89. Messier C, Doucet R, Ruel J-C, Claveau Y, Kelly C, Lechowicz MJ. Functional ecology of advance regeneration in relation to light in boreal forests. *Can J For Res*. 1999;29(6):812–23. <https://doi.org/10.1139/x99-070>
90. Draper FC, Costa FRC, Arellano G, Phillips OL, Duque A, Macía MJ, et al. Amazon tree dominance across forest strata. *Nat Ecol Evol*. 2021;5(6):757–67. <https://doi.org/10.1038/s41559-021-01418-y> PMID: 33795854
91. Whittaker RH. Evolution and measurement of species diversity. *Taxon*. 1972;21(2–3):213–51. <https://doi.org/10.2307/1218190>
92. Legendre P, Legendre L. Chapter 10 - Interpretation of Ecological Structures. In: Legendre P, Legendre L, editors. *Numerical Ecology*. vol. 24 of *Developments in Environmental Modelling*. Elsevier; 2012. p. 521–624.
93. Beech E, Rivers M, Oldfield S, Smith PP. GlobalTreeSearch: The first complete global database of tree species and country distributions. *J Sustain Forest*. 2017;36(5):454–89. <https://doi.org/10.1080/10549811.2017.1310049>
94. Condit R, Ashton PS, Baker P, Bunyavejchewin S, Gunatilleke S, Gunatilleke N, et al. Spatial patterns in the distribution of tropical tree species. *Science*. 2000;288(5470):1414–8. <https://doi.org/10.1126/science.288.5470.1414> PMID: 10827950
95. Rüger N, Berger U, Hubbell SP, Vieilledent G, Condit R. Growth strategies of tropical tree species: disentangling light and size effects. *PLoS One*. 2011;6(9):e25330. <https://doi.org/10.1371/journal.pone.0025330> PMID: 21966498
96. Żywiec M, Muter E, Zielonka T, Delibes M, Calvo G, Fedriani JM. Long-term effect of temperature and precipitation on radial growth in a threatened thermo-Mediterranean tree population. *Trees*. 2017;31(2):491–501. <https://doi.org/10.1007/s00468-016-1472-8>
97. Grossiord C, Sevanto S, Adams HD, Collins AD, Dickman LT, McBranch N, et al. Precipitation, Not Air Temperature, Drives Functional Responses of Trees in Semi-arid Ecosystems. *J Ecol*. 2017;105(1):163–175.
98. Gillner S, Korn S, Hofmann M, Roloff A. Contrasting strategies for tree species to cope with heat and dry conditions at urban sites. *Urban Ecosyst*. 2017;20(4):853–65. <https://doi.org/10.1007/s11252-016-0636-z>
99. McInnes L, Healy J, Astels S, et al. hdbscan: Hierarchical density based clustering. *JOSS*. 2017;2(11):205. <https://doi.org/10.21105/joss.00205>
100. He X. Resampling Methods. *Handbook of Quantile Regression*. Chapman and Hall/CRC. 2017. p. 7–19.