Mathematical Foundations of Beta Diversity: Why Common Metrics Fail in Microbiome Analysis

Zihan Zhu[1], Yiqian Zhang[1, 2, 3], Wenhao Li[4], Michael Greenacre[5], Satabdi Saha[4], Yushu Shi[6], Liangliang Zhang[1, 7*]

[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, 44106, OH, USA.
[2]Division of Biostatistics, College of Public Health, Ohio State University, 1841 Neil Ave, Columbus, 43210, OH, USA.
[3]Department of Statistics, College of Arts and Sciences, The Ohio State University, 1958 Neil Ave., Columbus, 43210, OH, USA.
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 7007 Bertner Ave, Houston, 77030, TX, USA.
[5]Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.
[6]Weill Cornell Medicine, Cornell University, 1300 York Ave, New York, 10065, NY, USA .
[7]Case Comprehensive Cancer Center, 2103 Cornell Road, Cleveland, 44106, OH, USA .

*Corresponding author(s). E-mail(s): lxz716@case.edu;

Abstract

Background
In microbiome studies, beta diversity quantifies dissimilarity between samples and is often visualized using ordination techniques. It enables researchers to characterize ecological structure, compare microbial communities, assess environmental or host-driven heterogeneity, and track longitudinal shifts over time. Although many diversity indices were originally developed with practical goals in mind, they lack a unified framework to ensure theoretical rigor and validity. This gap makes it challenging for researchers to evaluate and select appropriate beta diversity measures for microbiome analyses, potentially leading to biased analyses and invalid conclusions.

Results
To bridge the persistent knowledge gaps, we systematically evaluate the commonly used beta diversity measures according to key mathematical properties, including whether they are true metrics, conform to Euclidean geometry, and satisfy conditional negative definiteness. We show that their violations can compromise downstream analyses such as PCoA, PERMANOVA, and kernel-based tests. In addition, drawing on mathematical consensus, we introduce a novel four-category classification of beta diversity measures: scale difference, difference scale, Hamming difference, and distribution difference. Complementing this framework, we build diagnostic tools for assessing Euclidean validity and develop remedial strategies that correct problematic dissimilarity matrices while preserving ordination structures. We demonstrate the effectiveness of these solutions using real-world microbiome datasets.

Conclusions
These results establish a unified framework for evaluating beta diversity in microbiome research, supported by an R package, interactive Shiny app, and step-by-step tutorials. The framework provides a clear roadmap for selecting and refining dissimilarity metrics, paving the way for future methodological advances.

Keywords: Beta Diversity, Dissimilarity Measures, Euclidean Property, Microbiome Analysis

1 Background

The origins of ecology as a science trace back to the application of experimental and mathematical approaches to understanding organism–environment interactions, community structure and succession, and population dynamics [1]. Ecology has emerged as a scientific discipline since the mid-19th century, shaped largely by early work in botany and zoology. By studying the distribution and behavior of plants and animals, scientists have come to recognize the intricate relationships between organisms and their environments. For instance, plant distributions have been found to depend on climate, soil, and other environmental factors, while animal populations are influenced by food availability, predation, and species interactions. Although microorganisms were historically understudied due to their invisibility to the naked eye, difficulty in cultivation, and vast unknown diversity, the advent of high-throughput sequencing technologies has enabled comprehensive profiling of microbial communities and thus catalyzed rapid advances in microbiome and microbial ecology research [2, 3].

Microbiome research differs fundamentally from traditional human omics studies, which focus on molecular data (e.g. genomics, transcriptomics, or proteomics) derived from a single organism. In contrast, microbiome studies investigate a complex, multi-species community of microorganisms inhabiting diverse environments, including the human body and beyond. These microbial communities exhibit intricate ecological interactions—such as mutualism, competition, and antagonism—that influence both

the microbial composition and host physiology [4]. To quantitatively capture the structure and variability of these communities, microbiome researchers rely on diversity indices rooted in ecological theory [5], while also developing new metrics tailored to the unique features of microbiome data, such as phylogenetic relatedness and compositionality. As such, ecological principles serve as a foundational lens for understanding microbial ecosystems [6].

Among these principles, understanding how biological communities vary across space and time has long been a central theme in ecology. The alpha–beta–gamma framework, introduced by Whittaker [7], provides a structured approach to partitioning biodiversity within and between habitats. By linking plant population distributions and community characteristics to continuous environmental gradients, his work reframed community ecology as a study of compositional change along ecological axes. This foundational perspective paved the way for modern diversity metrics, particularly beta diversity, which captures the turnover or differentiation among communities across varying environments or conditions [8, 9]. Extending this concept to microbiome research, beta diversity quantifies pairwise dissimilarity between samples and enables comparison of ecological variation due to environmental, spatial, or host-related factors. A simple bibliometric study shows that the proportion of microbiome studies with beta diversity included in title, abstract or key words keeps increasing in last two decades, as detailed in Supplementary Section 1.

To support these analyses, a wide array of beta diversity measures has been developed, each reflecting distinct ecological assumptions-whether based on presence–absence, abundance, or phylogeny [10]. Since the 1970s, beta diversity has undergone significant refinement—moving from early quantitative indices like Jaccard [11] and Sørensen [12], and their subsequent use or promotion by MacArthur [13], Cody [14], and Wilson & Shmida [15], to more sophisticated approaches that incorporate phylogenetic (Faith, 1992 [16]) and functional trait information (Petchey & Gaston, 2002 [17]). In the modern microbiome era, advances in high-throughput sequencing have made beta diversity a central tool for analyzing microbial community variation, with new metrics like UniFrac (Lozupone & Knight, 2005 [18]) designed to capture phylogenetic and compositional structures across environmental and host-related gradients. Their integration into software such as `vegan`, `phyloseq`, and `betapart` has facilitated broad application across ecological fields. Despite these advances, microbiome data pose unique analytical challenges due to its compositional nature, prevalence of zeros, and complex phylogenetic structure, all of which complicate dissimilarity estimation. Consequently, the choice of beta diversity measures can profoundly impact ordination results, statistical inferences, and biological conclusions [18, 19].

A critical issue arises when beta diversity measures do not satisfy key mathematical properties of a metric. In such cases, the resulting dissimilarities may fail to represent the true spatial relationships between samples. This non-metric nature can introduce distortions to downstream analyses such as ordination, clustering, or kernel-based models, ultimately affecting the interpretation of community dissimilarity patterns. To illustrate this issue, we present a simple example highlighting the impact of non-metric dissimilarities on principal coordinates analysis (PCoA) visualizations. We construct

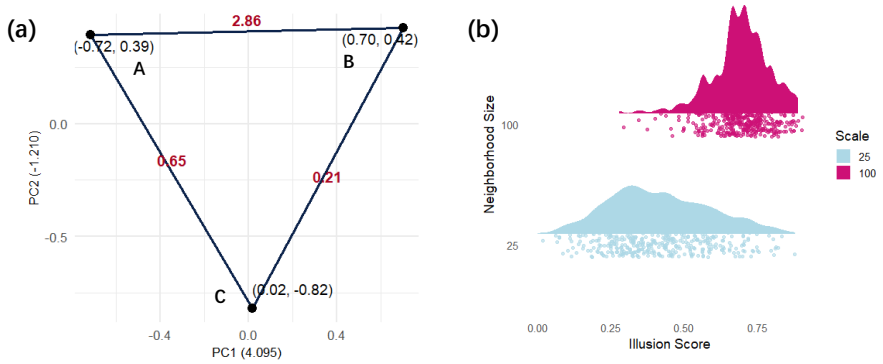

Fig. 1: Illustration of geometric and statistical distortions due to non-Euclidean dissimilarities. (a) A toy example illustrating the principal coordinates analysis (PCoA) of three samples. (b) Raincloud plot of the illusion score for Bray-Curtis dissimilarity applied to Data (1) (introduced in Section 4.1).

a Bray–Curtis dissimilarity matrix among samples for a problematic dataset (to be introduced in Section 4.1). We then perform eigen-decomposition of the dissimilarity matrix and present the results in Figure 1. We select three samples and plot their embeddings in the two-dimensional principal coordinate. As shown in Figure 1 (a), the first eigenvalue is 4.095 while the second is -1.21, indicating that the beta diversity matrix is not positive definite. Although samples A and B are substantially more dissimilar than the others based on the bray-curtis distance of original data (2.86 versus 0.65 and 0.21), their projected distance in the PCoA plot appears nearly equal to that of other sample pairs—highlighting the spatial distortion introduced during ordination. This visual artifact is termed the illusion of similarity (IOS). We quantify IOS as the fraction of neighbors in the PCoA space that are not true neighbors in the original space. Figure 1 (b) displays the distribution of IOS score for each sample under different neighborhood size. In ideal cases, IOS score is 0, indicating no illusion of similarity. However, the mean IOS score of this data can reach as high as 0.69, underscoring the extent of distortion introduced by PCoA when the underlying beta diversity measure violates metric properties.

The above example illustrates a broad and serious issue in microbiome research. Despite growing use of beta diversity in microbiome studies [20], many studies lack a unified framework to assess the theoretical validity of these measures. Without a rigorous foundation and evaluation framework, it becomes difficult for researchers to assess or select appropriate beta diversity metrics—raising the risk of analytical distortions,

biased results, and misleading biological conclusions. To bridge this persistent knowledge gaps, we systematically review and evaluate the commonly used beta diversity measures according to key mathematical properties, including whether they are true metrics, conform to Euclidean geometry, and satisfy conditional negative definiteness. We show that their violations can compromise downstream analyses such as PCoA, PERMANOVA, and kernel-based methods. To guide robust and interpretable analyses, we introduce a unified classification framework, develop diagnostic tools to assess distortion, and provide practical recommendations for selecting or adjusting dissimilarity measures. We demonstrate the effectiveness of these solutions using real-world microbiome datasets.The resulting framework is implemented in an R package with an interactive Shiny application and step-by-step tutorials to support broad adoption.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive overview of commonly used dissimilarities, detailing their definitions, ecological motivations, and computational formulations. In Section 3, we highlight three key mathematical properties that ideal dissimilarities should satisfy and propose a classification system that groups existing measures according to shared theoretical features. To address the practical challenge of verifying these properties, we also introduce a diagnostic framework implemented in R, alongside recommended remedies for commonly observed violations. In Section 4, we demonstrate the utility of these tools through applications to multiple real-world microbiome datasets. Finally, Section 5 summarizes our findings and discusses broader implications for beta diversity analysis in microbial ecology.

2 Method

We begin by introducing the concept of dissimilarity metrics, which serve as the foundation for quantifying beta diversity in microbial community analysis. Let $d(i, j)$ denote a generic dissimilarity between microbial communities i and j, computed from input profiles such as species abundance, relative abundance, or presence–absence data. These pairwise values form the entries of a dissimilarity matrix $D \in \mathbb{R}^{N \times N}$, where each element satisfies

$$D_{ij} = d(i, j), \quad D_{ii} = 0.$$

Such matrices serve as inputs for a variety of downstream analyses, including ordination techniques (e.g., Principal Coordinates Analysis), statistical tests (e.g., PERMANOVA, MiRKAT), and community structure visualizations, which are used to reveal underlying ecological gradients.

In what follows, we detail the analytical principles and methodological choices that inform our evaluation and application of beta diversity measures. We first survey commonly used dissimilarities and their mathematical definitions. We then present a theoretical framework for evaluating key geometric and statistical properties—such as the metric condition, Euclidean embeddability, and conditional negative definiteness—which underpin many ecological interpretations. Next, we assess how deviations

from these properties impact downstream analyses, and introduce diagnostic and correction tools to mitigate such effects. Together, these methods provide a comprehensive and geometry-aware foundation for evaluating beta diversity in microbiome studies.

2.1 An Overview of Commonly Used Dissimilarity Measures

Let's first give a clear distinction: dissimilarity measures are like rulers—they provide the mathematical tools to measure the distances between microbial communities. Beta diversity, in contrast, is the landscape these rulers help us explore: it captures how community composition varies across environments or samples. For clarity and consistency in presenting mathematical formulas for each dissimilarity measure, we first define the notation and symbols used throughout. Let N denote the number of samples and K the number of observed taxa. For sample i, let a_{ik} represent the absolute abundance (sequence read counts) of taxon k, with corresponding relative abundance $r_{ik} = a_{ik}/a_{i\cdot}$, where $a_{i\cdot} = \sum_{k=1}^{K} a_{ik}$. Let $p_{ik} = I(a_{ik} > 0)$ denote presence–absence indicators. As follows, we present dissimilarity measures in roughly chronological order and categorize them into two groups based on disciplinary origin: those developed primarily by biological researchers and those arising from quantitative fields such as mathematics and statistics. We begin with the biologically motivated measures. Additionally, following both prior literature and our own validation, we use the term distance exclusively for measures that satisfy all metric properties. All others are referred to as dissimilarities, irrespective of their original designation.

2.1.1 Jaccard Distance

The Jaccard index, introduced by the Swiss botanist Paul Jaccard in 1901 [11], is one of the earliest measures of similarity and diversity between two sets of species. Originally, it considered the presence-absence information of the elements in two finite sets A and B. The Jaccard similarity index is mathematically defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

Here $|\cdot|$ is the cardinality of a set. Hence, $|A \cap B|$ represents the number of elements present in both A and B and $|A \cup B|$ represents the number of elements present in A or B. The Jaccard index yields a value between 0 and 1, where 0 denotes disjoint sets and 1 denotes identical sets. The corresponding Jaccard distance, often used as a dissimilarity measure, is given by:

$$d_J(A, B) = 1 - J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

This definition yields a proper distance-like metric that increases as the sets become less similar and satisfies the properties of a metric on the space of all finite sets [21–23].

For presence–absence data, the Jaccard distance between samples i and j is defined as

$$d_J(i,j) = 1 - \frac{\sum_k p_{ik}p_{jk}}{\sum_k I(p_{ik} + p_{jk} > 0)}.$$

The Jaccard index extends to non-binary data by representing sets as non-negative vectors, denoting degrees of membership. In this setting, set operations are replaced with element-wise minima and maxima. For abundance data, the generalized (non-binary) Jaccard similarity index [24] is defined as

$$J_{\text{non-binary}}(i,j) = \frac{\sum_k \min(a_{ik}, a_{jk})}{\sum_k \max(a_{ik}, a_{jk})}.$$

This extension accommodates continuous or abundance data, allowing for a more nuanced comparison of communities or feature sets. The non-binary Jaccard index collapses to the binary case when all values are restricted to 0 or 1. After more than a century, Jaccard-based measures remain indispensable in a broad range of applications, capturing essential notions of similarity and difference with a conceptual elegance that has proven timeless. It continues to find widespread application in fields such as ecology, information retrieval, and machine learning [25].

2.1.2 Kulczyński Dissimilarity

The Kulczyński dissimilarity is named after Stanisław Kulczyński, a Polish botanist and phytosociologist. In 1928, Kulczyński introduced this measure to quantify the similarity between ecological communities, originally using species presence-absence data [26]. His work focused on understanding the relationships among vegetation samples, and the Kulczyński dissimilarity has since become a valuable tool in studies of ecology, community structure, and biodiversity assessment.

The original binary version based on presence-absence data is given by:

$$S_{ij}^{(\text{binary})} = \frac{1}{2}\left(\frac{\sum_k p_{ik}p_{jk}}{\sum_k p_{ik}} + \frac{\sum_k p_{ik}p_{jk}}{\sum_k p_{jk}}\right).$$

This formula reflects the original form of the measure, which accounts purely for species occurrences without considering their abundances. The definition was later extended to handle species abundance, making it a more nuanced measure of dissimilarity. This extension is especially useful when in contexts where the proportional overlap of species abundances between samples is of interest.

For abundance data, the Kulczyński similarity index between two samples i and j is defined as [27–29]:

$$S_{ij} = \frac{1}{2}\left(\frac{\sum_k \min(a_{ik},\, a_{jk})}{a_{i\cdot}} + \frac{\sum_k \min(a_{ik},\, a_{jk})}{a_{j\cdot}}\right).$$

The term $\sum_k \min(a_{ik},\, a_{jk})$ captures the shared abundance of each species between the two samples, highlighting overlap in species composition. The denominators

$a_{i\cdot} = \sum_k a_{ik}$ and $a_{j\cdot} = \sum_k a_{jk}$ normalize the shared abundances relative to the overall abundances. The Kulczyński dissimilarity is then calculated by subtracting the similarity from one [30]:

$$D_{ij} = 1 - S_{ij} = \frac{1}{2}\left(\frac{\sum_k |a_{ik} - a_{jk}|}{a_{i\cdot}} + \frac{\sum_k |a_{ik} - a_{jk}|}{a_{j\cdot}}\right).$$

By averaging the normalized shared abundances and subtracting from one, the Kulczyński dissimilarity balances contributions from both samples and provides a measure of dissimilarity that does not overly penalize species with small distribution ranges. It is also considered a suitable alternative to the Jaccard distance for analyzing groups of species with overlapping distributions, as species with small distribution areas can be grouped alongside those with larger ones when there is a subset relationship [31–33]. This makes the Kulczyński distance a valuable tool in ecological studies aiming to assess biodiversity patterns and community structures across different environments or temporal scales.

2.1.3 Bray Curtis Dissimilarity

The Bray–Curtis dissimilarity was first introduced by Bray and Curtis [34] in 1957. They developed this measure to interpret complex ecological data related to forest communities. Since then, the Bray–Curtis dissimilarity has become one of the most well-known tools to quantify compositional differences among ecosystems. It assesses the difference between two communities via the abundances of various species present in each community. It's modern form is mathematically defined as

$$\mathrm{BC}_{ij} = \frac{\sum_k |a_{ik} - a_{jk}|}{\sum_k (a_{ik} + a_{jk})}.$$

This formulation quantifies the dissimilarity by summing the absolute differences in the abundances of species and normalizing it by the total abundances of both communities. In practice, it is common to use relative abundances for this calculation, as defined in Supplementary Equation (1). Notably, the Bray–Curtis dissimilarity is also mathematically related to the quantitative Sørensen similarity index [12, 35] QS_{ij}, as both emphasize shared species abundance while ignoring joint absences. The formula can be expressed as

$$\mathrm{BC}_{ij} = 1 - \mathrm{QS}_{ij} = 1 - \frac{2\sum_k \min(a_{ik},\ a_{jk})}{\sum_k (a_{ik} + a_{jk})}.$$

However, it is important to note that the Bray–Curtis dissimilarity cannot be considered as a true distance because it does not satisfy the triangle inequality [36]. Furthermore, as [37] pointed out, Bray-Curtis dissimilarity between two communities often comes from the difference in "size" (or sequencing depth) rather than true composition. These limitations mean that while the Bray–Curtis dissimilarity is highly effective for comparing community compositions, it may not be suitable for analysis that rely on the properties of distance, such as certain ordination techniques and distance-based clustering algorithms.

2.1.4 UniFrac Dissimilarity

Unlike traditional ecological diversity metrics designed for plants and animals, UniFrac dissimilarity was specifically developed for microbiome research [38]. By incorporating phylogenetic relationships among microbial taxa, it refines community similarity assessment beyond what abundance- or presence-based measures offer. Both unweighted and weighted versions of UniFrac will be introduced in the following paragraphs.

The unweighted UniFrac distance is developed by Lozupone and Knight [18] in 2005. It quantifies the dissimilarity between microbial communities by leveraging phylogenetic information, focusing exclusively on the prevalence of taxa [18]. Mathematically, it is defined as the fraction of the lengths of unshared branches for two communities in a phylogenetic tree, i.e.,

$$d_{\text{Unweighted UniFrac}}(i,j) = \frac{\sum_{k=1}^{\bar{K}} b_k |p_{ik} - p_{jk}|}{\sum_{k=1}^{\bar{K}} b_k I(p_{ik} + p_{jk} > 0)},$$

where $\bar{K}$ denotes the number of branches. The unweighted UniFrac dissimilarity accounts for evolutionary relationships among taxa, offering a robust approach to capture community differences beyond simple taxonomic comparisons. Considering the presence or absence of taxa, it emphasizes rare or unique evolutionary lineages, which makes it particularly valuable in ecological studies where such lineages might reflect critical environmental adaptations or evolutionary pressures [18, 39, 40]. A graph illustration on calculating the unweighted UniFrac distance between two samples is provided in Figure 2.

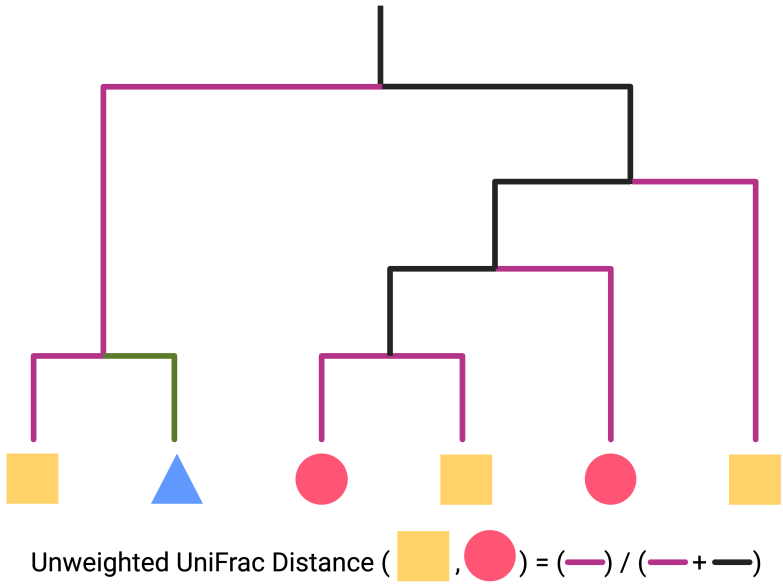

Fig. 2: **Unweighted UniFrac between two samples 'Circle' and 'Square'.** The blue triangle represents a taxon that does not belong to either of the two samples. The unshared branch lengths, representing phylogenetic branches unique to each community (highlighted in purple), are summed and divided by the sum of the unshared branch lengths and shared branch lengths (highlighted in black). This distance quantifies the phylogenetic dissimilarity between the two communities by comparing unique evolutionary lineages to the overall tree structure.

The Weighted UniFrac, introduced by Lozupone et al. [41] in 2007, extends the unweighted UniFrac distance by incorporating the relative abundances of taxa within microbial communities. Unlike the unweighted version, weighted UniFrac quantifies the dissimilarity between communities not only based on the phylogenetic distances of taxa but also by weighting these distances according to their relative abundances [41, 42]. Mathematically, it is defined as [41]:

$$d_{\text{Weighted UniFrac}}(i,j) = \frac{\sum_{k=1}^{\bar{K}} b_k |a_{ik} - a_{jk}|}{\sum_{k=1}^{\bar{K}} b_k |a_{ik} + a_{jk}|}$$

The numerator accounts for the weighted differences in abundance along each branch, while the denominator normalizes this difference by the total branch length of the phylogenetic tree. By combining evolutionary relationships and relative abundances, Weighted UniFrac offers a more nuanced measure of community dissimilarity, capturing both taxonomic composition and ecological dominance. As a result, it particularly useful in studies where abundance disparities between communities are important, such as those involving environmental gradients or host-associated [38, 41].

2.1.5 Minkowski Distance

Minkowski distances, introduced by Hermann Minkowski in the early 20th century [43], form a family of metrics that unify various notions of distance under a single mathematical framework. For points $x = (x_1, x_2, \ldots, x_K)$ and $y = (y_1, y_2, \ldots, y_K)$ in a K-dimensional space, the Minkowski distance is defined as

$$d_p(x, y) = \left(\sum_{i=1}^{K} |x_i - y_i|^p \right)^{1/p},$$

where $p \geq 1$. For microbiome data, the Minkowski distance between community i and j is

$$d_p(i, j) = \left(\sum_{k=1}^{K} |a_{ik} - a_{jk}|^p \right)^{1/p}.$$

The Minkowski distances qualify as true distance ensured by the Minkowski inequality [44]. By adjusting the parameter p, one can seamlessly transition between different distance measures, each characterized by distinct geometric and analytical properties.

When $p = 2$, the Minkowski distance becomes a special case known as the Euclidean distance

$$d_2(x, y) = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2},$$

whose roots trace back to Euclid's Elements [45]. The Euclidean distance represents the shortest straight-line path between two points, making it intuitively appealing and widely used. A key property of Euclidean distance is its rotational invariance: rotations and translations of the coordinate system do not alter the measured distances. This invariance is crucial in many applications, ensuring that the distance captures intrinsic relationships rather than coordinate-based artifacts. As a result, Euclidean distance serves as the foundation for numerous statistical and machine learning techniques, including principal component analysis [46, 47], clustering methods like K-means [48, 49], and regularized linear regression [50, 51]. Its squared form also appears naturally in optimization problems, where minimizing squared Euclidean distances often simplifies to solvable analytical solutions [52].

When $p = 1$, the Minkowski distance becomes a special case known as the Manhattan (or taxicab) distance

$$d_1(x, y) = \sum_{i=1}^{K} |x_i - y_i|.$$

It is so named because it reflects distances one might travel along a grid of perpendicular city streets [53]. Unlike the Euclidean distance, the Manhattan distance does not directly measure "straight-line" displacement; instead, it sums absolute differences across each dimension. This shift changes the geometry of the space, producing diamond-shaped level sets rather than circular ones, as illustrated in Figure 3. Such

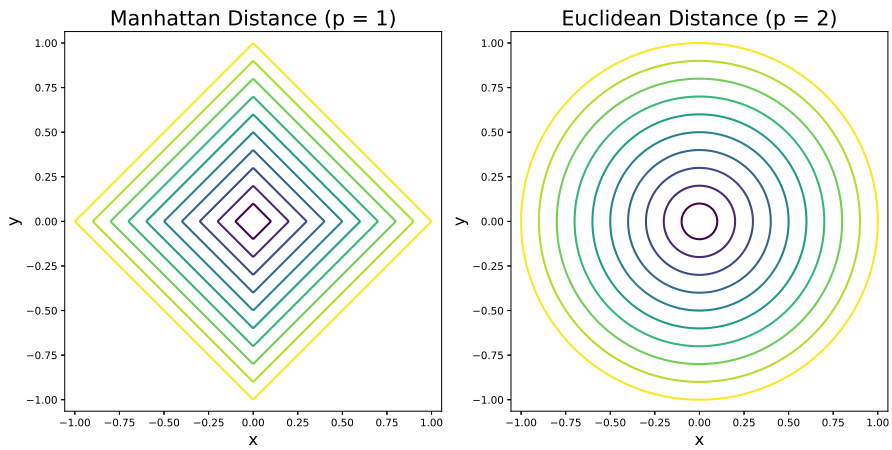

Fig. 3: Comparison of level sets for Manhattan distance ($p = 1$) and Euclidean distance ($p = 2$), illustrating the diamond-shaped geometry of Manhattan distance and the circular geometry of Euclidean distance.

geometry can be advantageous in contexts involving high-dimensional data, where the Euclidean distance may become less interpretable or dominated by large outliers. The Manhattan distance also tends to preserve meaningful differences in cases where numerous features are zero-inflated, a common occurrence in microbiome data and species abundance matrices [27]. Its sensitivity to coordinate axes and additive nature sometimes makes it the preferred measure in fields like text mining and genomics, where absolute differences in counts or frequencies are more meaningful than squared deviations [27].

In ecology, conservation science, and microbiome research, selecting a suitable Minkowski distance helps quantify compositional differences among communities and detect gradients of biodiversity or microbial assemblages [54]. In machine learning, the choice of p influences the shape of clusters and the performance of distance-based classifiers, potentially improving outcomes when data exhibit particular distributional or structural characteristics [52].

2.1.6 Chi-square Distance

The chi-square distance is a statistical measure that quantifies the dissimilarity between two distributions, especially in histogram comparison or compositional data. Its origins trace back to the pioneering work of British mathematician Karl Pearson in 1900 [55]. Pearson introduced the chi-square statistic as part of his development of the goodness-of-fit test, a method designed to evaluate how well observed data align with a theoretical distribution, especially within categorical data and contingency tables [55]. While his initial focus was on hypothesis testing against theoretical distribution, the chi-square statistic was later adapted as a measure for comparing empirical distributions directly, without referring to any theoretical distribution [56]. This adaptation

expanded its utility across various fields, including image processing, pattern recognition, machine learning, and biological sciences, where comparing distributions of categorical data is essential [57].

The chi-square distance can be computed using the formulation from the `vegdist` *R* package [58]:

$$\chi_{ij} = \sqrt{\sum_{k=1}^{K} \frac{\left(\frac{a_{ik}}{a_{i\cdot}} - \frac{a_{jk}}{a_{j\cdot}}\right)^2}{\bar{a}_{\cdot k}}} = \sqrt{\sum_{k=1}^{K} \frac{(r_{ik} - r_{jk})^2}{\bar{a}_{\cdot k}}},$$

where $\bar{a}_{\cdot k} = a_{\cdot k}/N$ and $a_{\cdot k} = \sum_i a_{ik}$. The transformation $r_{ik}/\sqrt{\bar{a}_{\cdot k}}$ is usually referred as chi-square transformation. It aggregates the values across all communities for each species k and takes average according to the number of communities. This normalization term acts as a weighting factor to adjust for the relative importance of each species based on its contribution to the total abundance.

The chi-square distance has become an invaluable tool in various applications. In microbiome research, it is employed to compare the composition of microbial communities by measuring differences in the abundance of microbial taxa between samples [27]. It enables researchers to detect significant variations in microbial populations, especially rare species that might be critical in understanding disease mechanisms or ecological dynamics [27].

2.1.7 Hellinger Distance

Hellinger distance, established in the context of probability theory by Ernst Hellinger in 1909 [59], measures the dissimilarity between two probability distributions in a way that is both bounded and symmetric [60]. Unlike traditional measures that might magnify large discrepancies or diminish the importance of rare components, Hellinger distance applies a square-root transformation to the probability values, ensuring that both common and rare categories contribute meaningfully to the assessment. For two discrete probability distributions $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$, the Hellinger distance is defined as

$$d_{Hellinger}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2}.$$

This formulation constrains the distance to lie within the interval $[0, 1]$, with 0 indicating identical distributions and 1 representing maximum divergence.

Hellinger distance has found broad utility in ecology, microbiome research, and other biological and environmental sciences [61], where the underlying data often represent relative abundances rather than absolute counts,

$$d_{Hellinger}(i, j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{K} (\sqrt{r_{ik}} - \sqrt{r_{jk}})^2}.$$

By incorporating a square-root transformation, the measure downweights dominant taxa and alleviates the influence of very small proportions, improving robustness against sampling errors and compositional biases. In statistical inference and machine learning, it provides stable and interpretable comparisons between distributions, facilitates clustering and classification tasks, and can enhance performance in non-parametric estimation [62–64]. Its versatile nature and sound mathematical properties make Hellinger distance a valuable tool for capturing patterns and differences in complex, probability-based data sets [65].

2.1.8 Mahalanobis Distance

Mahalanobis distance, introduced by Prasanta Chandra Mahalanobis in 1936 [66], provides a multivariate measure of dissimilarity that accounts for correlations among variables. Unlike Euclidean distance, which treats each dimension independently, Mahalanobis distance adjusts for the covariance structure of the variables, giving more balanced, scale-invariant assessments of how similar or different points are in a multidimensional space. For two vectors $x = (x_1, x_2, \ldots, x_K)$ and $y = (y_1, y_2, \ldots, y_K)$, and an variance–covariance matrix Σ [67], the Mahalanobis distance is defined as

$$d_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}.$$

This formulation not only re-scales features according to their variances, but also reduces redundancy by down-weighting directions of high correlation. Its use spans fields such as ecology, where it helps discern ecological niches by comparing species assemblages under correlated environmental gradients [27], and in pattern recognition, machine learning, and outlier detection, where capturing the "shape" and internal structure of data is critical. Mahalanobis distance's ability to incorporate covariance information makes it a powerful tool for multivariate analyses, guiding the identification of clusters, discriminating among groups, and improving classification accuracy in complex, high-dimensional data sets.

In microbiome analyses, each dimension commonly represents an OTU (Operational Taxonomic Unit), and the matrix Σ reflects the variance and covariance of these OTUs across samples,

$$d_M(i, j) = \sqrt{\sum_{k=1}^{K} \sum_{l=1}^{K} (a_{ik} - a_{jk})(\Sigma^{-1})_{k,l}(a_{il} - a_{jl})},$$

where $(\Sigma^{-1})_{k,l}$ is the (k, l)-th element of Σ^{-1}. Notably, as R packages vegan [68] and phyloseq [69] do, we replace the true covariance matrix Σ with the sample covariance matrix S. It is derived from the abundance patterns of OTUs rather than their representative sequences, while Pavoine et al. 2004 [70] interprets $\Sigma_{k,l}$ as the distance between species k and l. By incorporating these abundance-based covariance structures, Mahalanobis distance can capture the correlated nature of microbial communities more accurately. Literature on multivariate analyses in microbial ecology [27, 71] underscores the importance of considering abundance-based

covariance patterns, as they help in revealing meaningful ecological interactions and community-level structures.

2.1.9 Aitchinson Distance

Aitchison distance, introduced by John Aitchison in 1982 [72] in the context of compositional data analysis, provides a meaningful notion of distance that respects the simplex structure inherent to compositional vectors. Traditional distances defined on the Euclidean space, such as the Minkowski distances, are not well-suited for compositional data. This is because compositional data are characterized by nonnegative components that sum to a constant. Moreover, the unit-sum constraint and differences in absolute scale can further distort the distances when these traditional measures are applied. The Aitchison distance circumvents these issues by mapping compositions into a log-ratio space, where standard Euclidean geometry can be applied. The Aitchison distance is defined as

$$d_A(i,j) = \left(\sum_{k=1}^{K} \left(\log \frac{r_{ik}}{g(i)} - \log \frac{r_{jk}}{g(j)} \right)^2 \right)^{\frac{1}{2}},$$

where

$$g(i) = \left(\prod_{k=1}^{K} r_{ik} \right)^{\frac{1}{K}}.$$

This formulation treats ratios of parts, rather than their absolute differences, ensuring that the analysis focuses on relative information—crucial in fields such as geochemistry or microbiome research, where relative abundances carry more interpretative significance than absolute counts [73, 74].

In practice, compositional data are often characterized by the presence of zeros, the occurrence of outliers, and heavy-tailed distributions. To address such complexities, robust versions of the Aitchison distance have been developed [75]. These variants incorporate robust estimators of center and scatter, or employ zero-replacement techniques that maintain the compositional nature of the data while diminishing the influence of problematic values. By reducing sensitivity to extreme observations and properly handling zeros, robust Aitchison distances enhance stability and reliability in real-world applications, improving interpretations of ecological niches, dietary patterns, environmental pollution gradients, and microbial community structures.

2.2 Theoretical Properties of an Ideal Dissimilarity Matrix

To provide a mathematical justification and theoretical framework for existing beta diversity indices introduced above, we outline three fundamental mathematical properties that are often desired for an ideal dissimilarity matrix in microbial community analysis: the metric property, the Euclidean property, and conditional negative definiteness (CND). These properties form a hierarchy in terms of restrictiveness and analytical utility. While many indices satisfy the basic conditions of a metric, fewer are Euclidean or conditionally negative definite. Although the Euclidean property

and CND are mathematically equivalent when applied to squared dissimilarities, they emphasize different perspectives: the Euclidean property supports geometric interpretations and ordination techniques such as principal coordinates analysis, while CND enables the construction of valid kernels for statistical learning and regression in Hilbert spaces. Discussing all three properties clarifies the theoretical capabilities and limitations of different dissimilarity measures and informs their appropriate use in ecological inference and downstream analysis.

2.2.1 Metric Property

A dissimilarity matrix $D \in \mathbb{R}^{N \times N}$ satisfies the metric property if its entries D_{ij} fulfill the following four conditions for all $i, j, k \in \{1, \ldots, N\}$:

1. Non-negativity:
$$D_{ij} \geq 0$$
2. Identity of indiscernible (reflexivity):
$$D_{ij} = 0 \iff i = j$$
3. Symmetry:
$$D_{ij} = D_{ji}$$
4. Triangle inequality:
$$D_{ik} \leq D_{ij} + D_{jk}$$

This set of conditions follows the definition of a metric in the sense of Gower [76]. These properties ensure that the dissimilarity measure functions in accordance with the principles of a mathematical distance. In microbial ecology, this means the index adheres to intuitive criteria such as identity (zero dissimilarity for identical communities), symmetry between community comparisons, and consistency when evaluating direct versus indirect differences. However, some widely used indices, such as Bray–Curtis dissimilarity, do not satisfy all four properties. In particular, they may violate the triangle inequality for certain sample pairs, and are therefore considered semi-metrics [53]. Although semi-metric dissimilarities remain meaningful and interpretable, their failure to satisfy the triangle inequality can limit their applicability in distance-based statistical methods that assume metric properties, such as metric multidimensional scaling (mMDS) [77].

Ecological interpretations of the triangle inequality arise naturally from fundamental principles of community assembly. A central hypothesis is the individualistic hypothesis, originally proposed by Whittaker [78], which posits that species respond independently to underlying environmental gradients, such as changes in soil pH (e.g., from 5 to 8), altitude (e.g., from 1000 to 3000 meters), or host diet (e.g., from high-fiber to high-fat intake). This hypothesis implies that changes in community composition occur gradually along such gradients, forming ecological continua. In this context, beta diversity—as a measure of dissimilarity between communities—should reflect this gradualism: if community B lies between A and C along a gradient, its compositional dissimilarity to both A and C should be more than A and C are to each other. This logic implies that the dissimilarity between A and C is bounded by the sum of dissimilarities

involving B—a property exactly formalized by the triangle inequality. Further support comes from the idea of ecological continuity, especially in long-standing, undisturbed habitats. Odum [79] suggested that such continuity promotes gradual shifts in species composition over time or space, reinforcing the notion that community transitions are typically smooth rather than abrupt breaks or gaps in between. Thus, the triangle inequality aligns not merely with a mathematical structure, but with ecological expectations about how communities respond to continuous environmental variation.

In microbial ecology, this idea translates into hypotheses about how microbiomes vary across gradients of host phenotype, tissue niche, or environmental exposure. For example, if a set of gut microbiomes vary along a diet-based continuum, then samples from intermediate diets should bridge the dissimilarities between more extreme ones. Consequently, observing violations of the triangle inequality in beta diversity matrices may signal ecological discontinuity, hidden host stratification, or methodological artifacts, making it an important diagnostic tool for both theoretical and practical purposes.

2.2.2 Euclidean Property

A dissimilarity matrix D is said to have the Euclidean property if there exists an embedding of the N communities into a Euclidean space $\mathbb{R}^m$, for some $1 \leq m \leq N-1$, such that the squared dissimilarities correspond exactly to squared Euclidean distances between embedded points. That is, there exists a set of points $\{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^m$ such that

$$D_{ij}^2 = \|x_i - x_j\|^2, \quad \forall i, j \in \{1, \ldots, N\}.$$

The concept of Euclidean embedding has deep mathematical roots. In 1841, Cayley introduced the Cayley–Menger determinant, providing early criteria for embedding point configurations in Euclidean space. Menger later formalized the necessary and sufficient conditions for a finite metric space to admit such an embedding [80]. A dissimilarity matrix that fulfills this criterion is said to have the Euclidean property, meaning that its pairwise values can be faithfully represented as distances in a Euclidean space without geometric distortion.

This property is particularly crucial for ordination methods in beta diversity analysis. When a dissimilarity matrix is Euclidean, techniques such as Principal Coordinates Analysis (PCoA) can accurately map the samples into a low-dimensional Euclidean space, preserving the dissimilarity structure as faithfully as possible. In this setting, the visualized distances between points directly correspond to the original dissimilarities. This allows researchers to intuitively interpret ecological gradients and clustering patterns in the data. In contrast, if the matrix lacks the Euclidean property, the embedding may introduce distortions, leading to misleading visual representations of community relationships [76, 81]. Thus, the Euclidean property is essential not only for mathematical consistency but also for ensuring meaningful ecological interpretation in ordination and visualization.

2.2.3 Conditional Negative Definiteness

In mathematics, a symmetric function[1] $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be conditionally negative definite (CND), if for any finite set of points $\{x_1, x_2, \ldots, x_N\} \subseteq \mathcal{X}$ and any real coefficients $\{c_1, \ldots, c_N\}$ satisfying $\sum_{i=1}^{N} c_i = 0$, the following inequality holds:

$$\sum_{i,j=1}^{N} c_i c_j f(x_i, x_j) \leq 0.$$

In beta diversity analysis, we often consider a squared dissimilarity matrix $D^{(2)} = [D_{ij}^2]$. It is said to be conditionally negative definite if

$$\mathbf{c}^\top D^{(2)} \mathbf{c} \leq 0$$

for all real vectors $\mathbf{c} = (c_1, \ldots, c_N)^\top$ with $\sum_{i=1}^{N} c_i = 0$.

The CND property plays a central role in ecological data analysis, particularly due to its connection with ordination. Many widely used dissimilarity measures, such as Rao's quadratic entropy [82] and variance-based dissimilarities [83], are derived from conditionally negative definite functions. When the squared dissimilarities satisfy the CND condition, the corresponding matrix can be embedded into a Euclidean space without introducing curvature distortions.

Several ordination techniques, including Principal Coordinates Analysis (PCoA) and Double Principal Coordinates Analysis (DPCoA) [70], explicitly require this property. Notably, even when the original dissimilarity matrix D is not CND, its squared version $D^{(2)}$ may be—this holds, for example, the Manhattan distance. This distinction is particularly relevant when evaluating whether a dissimilarity measure supports faithful Euclidean embedding. We further explore this connection in the next subsection.

2.2.4 Relationship among Metric, Euclidean and CND Properties

The relationship among the metric property, the Euclidean property, and conditional negative definiteness (CND) has been studied extensively, beginning with Menger's characterization theorem [84]. This theorem provides necessary and sufficient conditions for a semi-metric space to be isometrically embeddable in a Euclidean space, based on the positive semi-definiteness (PSD) of the associated squared distance matrix. This result was later generalized by Gower [76], who formalized the equivalence between the Euclidean property and the CND property for dissimilarities. Specifically, a dissimilarity measure is Euclidean if and only if its squared dissimilarity matrix is conditionally negative definite. Thus, the Euclidean and CND properties are mathematically equivalent. Furthermore, Blumenthal [85, 86] proved that if a semi-metric satisfies the Euclidean property, it must also satisfy the triangle inequality, and therefore is a metric. As most dissimilarity measures are at least semi-metrics, this result implies that every Euclidean dissimilarity is also a metric. However, the converse does

[1] A binary function is symmetric if $f(x, y) = f(y, x)$.

not hold: not all metric dissimilarities are Euclidean. That is, some metric spaces cannot be isometrically embedded into any Euclidean space [87].

These theoretical relationships allow us to categorize dissimilarity measures into three levels of increasing mathematical adequacy:

1. Semi-metric dissimilarity: satisfies non-negativity, reflexivity, and symmetry, but may violate the triangle inequality.
2. Non-Euclidean metric dissimilarity: satisfies all metric properties, but not Euclidean.
3. Euclidean dissimilarity: both metric and Euclidean (i.e., squared dissimilarities are CND).

In practice, dissimilarity measures that are both metric and Euclidean are generally preferred, as they ensure compatibility with distance-based ordination methods and enable faithful geometric representation of ecological relationships. Therefore, understanding and verifying these properties is critical for selecting appropriate measures in downstream beta diversity analyses.

2.2.5 Consequences of Violating Theoretical Conditions

Many commonly used methods—like PCoA, PERMANOVA, and MiRKAT—assume the dissimilarity is Euclidean. But this assumption is often not checked in practice, which can compromise the mathematical validity and reliability of the results. To illustrate the impact of such violations, we present three examples where unmet conditions lead to untrustworthy inferences. Additionally, we introduce diagnostic metrics to quantify the severity of these violations.

PCoA.

The centered Gram matrix used in PCoA is given by

$$G = -\frac{1}{2} J D^{(2)} J, \quad \text{where } J = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top.$$

If D is Euclidean, then $D^{(2)}$ is CND, and the Gram matrix G is positive semi-definite (PSD). In this case, all eigenvalues of G, denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$, represent non-negative contributions to total variation, and their relative magnitudes describe the proportion of variance captured by each principal coordinate:

$$\text{Proportion}_i = \frac{\lambda_i}{\sum_{i=1}^N \lambda_i}.$$

However, when the dissimilarity is non-Euclidean, G may contain negative eigenvalues, indicating curvature or non-Euclidean structure in the original space. To quantify this, we define the fraction of negative inertia (FNI) as:

$$\text{FNI} = \frac{\sum_{i=1}^N |\lambda_i| \cdot I(\lambda_i < 0)}{\sum_{i=1}^N |\lambda_i|}.$$

FNI measures the proportion of variation that lies outside of Euclidean space and cannot be faithfully captured by PCoA. In such cases, the standard interpretation of eigenvalues as variance contributions becomes invalid, since the resulting axes no longer represent orthogonal directions in flat Euclidean geometry.

PERMANOVA.

PERMANOVA evaluates the association between dissimilarity and sample groupings via a pseudo-F statistic:

$$F = \frac{\text{MS(model)}}{\text{MS(residual)}},$$

where the mean squares are defined using the Gram matrix G and the projection (hat) matrix H of the metadata design matrix:

$$\text{MS(model)} = \frac{1}{d-1}\text{tr}(HGH), \quad \text{MS(residual)} = \frac{1}{N-d}\text{tr}((I-H)G(I-H)).$$

Assuming $G = Q\Lambda Q^\top$ is the eigen-decomposition of the Gram matrix, we can express:

$$\text{MS(model)} = \frac{1}{d-1}\sum_{i=1}^{N} \lambda_i \left\|(HQ)_{\cdot,i}\right\|^2,$$

where $(HQ)_{\cdot,i}$ denotes the projection of the i-th principal coordinate onto the metadata space. When $\lambda_i < 0$, the contribution to MS(model) becomes negative, violating the interpretation of variance decomposition and leading to potential misinterpretation of the pseudo-F statistic. To address this, we propose a modified pseudo-F statistic that considers only positive eigencomponents:

$$\hat{F} = \frac{N-d}{d-1} \cdot \frac{\sum_{i=1}^{N} |\lambda_i| \left\|(HQ)_{\cdot,i}\right\|^2 I(\lambda_i > 0)}{\sum_{i=1}^{N} |\lambda_i| \left\|(HQ)_{\cdot,i}\right\|^2},$$

and define the corresponding coefficient of determination as:

$$\hat{R}^2 = \frac{\sum_{i=1}^{N} |\lambda_i| \left\|(HQ)_{\cdot,i}\right\|^2 I(\lambda_i > 0)}{\sum_{i=1}^{N} |\lambda_i| \left\|(HQ)_{\cdot,i}\right\|^2}.$$

MiRKAT.

MiRKAT tests the association between a continuous or categorical outcome Y and a dissimilarity matrix via a kernel-based score statistic. Let H be the projection matrix for confounders and $R_Y = (I-H)Y$ be the residuals. The test statistic is:

$$T = R_Y^\top G R_Y = \text{tr}(R_Y^\top Q\Lambda Q^\top R_Y),$$

which quantifies the squared association between the outcome residuals and the principal coordinates, weighted by their corresponding eigenvalues. If the dissimilarity

is non-Euclidean and some $\lambda_i < 0$, the contribution of those axes becomes negative, undermining the interpretation of T as a variance-based measure of association. MiRKAT uses permutation testing to evaluate significance under the null. Given residuals R_Y^{perm}, the null distribution of the test statistic is approximated by

$$T_{\text{perm}} = (R_Y^{\text{perm}})^\top G R_Y^{\text{perm}},$$

and the p-value is computed as:

$$p = \frac{\#\{T_{\text{perm}} > T\}}{\#\text{ permutations}}.$$

Though this procedure remains valid computationally, the interpretability of T as a variance-based statistic is compromised when G is not PSD.

Summary.

These examples demonstrate that when the dissimilarity matrix is not Euclidean, the core assumptions behind commonly used analysis methods can break down. Negative eigenvalues introduce interpretational issues in ordination, hypothesis testing, and variance partitioning. To ensure valid and interpretable results, the use of Euclidean dissimilarities—where the squared dissimilarity matrix is conditionally negative definite—is strongly recommended in downstream beta diversity analysis.

2.3 Classification of Dissimilarities

A wide range of dissimilarity measures have been developed to quantify structural differences between microbial communities, each grounded in different mathematical spaces and constructions. However, these measures vary greatly in their theoretical rigor—some are proper metrics, some are Euclidean, and others lack formal geometric properties altogether. While specific dissimilarities such as Bray–Curtis [36] and UniFrac [38, 88] have been studied in isolation, a unified geometric framework is still missing. Here, we summarize the mathematical underpinnings of four common types of dissimilarity measures, based on the metric space in which the community profiles are represented: Euclidean space, Hamming space, the compositional simplex, and the space of probability distributions.

Euclidean space. Abundance profiles are often modeled as vectors in $\mathbb{R}^K$, where each coordinate corresponds to a taxon's count or transformed count. A general class of dissimilarities in this space is based on the Minkowski distance [89]:

$$\|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{k=1}^{K} |x_k - y_k|^p \right)^{1/p}.$$

This formulation unifies several widely used distances. For example, $p = 2$ yields the Euclidean distance, while $p = 1$ gives the Manhattan distance. These metrics are compatible with a variety of multivariate techniques—including PCA, PCoA,

MANOVA—due to their well-defined geometric properties such as translation invariance, symmetry, and rotational consistency [76, 90–92]. In particular, Euclidean distances preserve variance decomposition and linear structure, making them especially suitable for ordination and hypothesis testing methods discussed in Section 2.2.

Hamming space. Presence-absence profiles are naturally embedded in the binary space $\mathbb{H}^K = \{0,1\}^K$, where each position indicates whether a taxon is detected [93]. The Hamming distance between two such vectors $\mathbf{p}$ and $\mathbf{q}$ is defined as:

$$d_{\mathbb{H}}(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{K} |p_k - q_k|,$$

which counts the number of mismatched taxa [93, 94]. Binary similarity measures such as the Jaccard index or unweighted UniFrac are derived from inner products and overlaps in this space.

Probability space and f-divergence. When relative abundances are viewed as discrete probability distributions $P = (p_1, \ldots, p_K)$, dissimilarity can be quantified using f-divergences [95–97], defined as:

$$D_f(P\|Q) = \sum_{k=1}^{K} q_k f\left(\frac{p_k}{q_k}\right),$$

where f is a convex function with $f(1) = 0$. Notable examples include the Jensen-Shannon distance [98]

$$f_{JS}(x) = x \log x - (x+1) \log \frac{x+1}{2},$$

and the Hellinger distance [59, 99]

$$f_{\text{Hellinger}}(x) = (\sqrt{x} - 1)^2,$$

both of which are symmetric and well-suited for measuring shifts in community composition. These divergences provide a statistical interpretation of dissimilarity as information loss or distributional discrepancy.

Compositional space. This space is presented last because it is fundamentally different from the other spaces: distances cannot be directly defined within the simplex due to its constrained geometry. Micorbial relative abundances are compositional by nature, lying in the simplex:

$$S^K = \left\{ \mathbf{x} \in \mathbb{R}^K \mid x_k > 0, \sum_{k=1}^{K} x_k = 1 \right\}.$$

Direct application of Euclidean distance in this constrained space can lead to misleading interpretations [72]. To address this, log-ratio transformations such as the centered log-ratio (CLR) are applied:

$$\mathrm{CLR}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \ldots, \log \frac{x_K}{g(\mathbf{x})} \right),$$

where $g(\mathbf{x})$ is the geometric mean of the components. The resulting Aitchison distance in the transformed space accounts for the relative scale and compositional nature of microbial communities.

This geometric classification offers a unified view of how dissimilarities map onto different mathematical spaces and provides insights into clustering behavior, ordination suitability, and metric validity. However, it does not explicitly trace the construction of these dissimilarities, from the transformation of single element to the pairwise operations (such as subtraction) performed in geometric spaces. For example, dissimilarities arising from square-root, log, or ratio transformations may share similar analytical behavior despite originating from different geometric spaces.

To address these gaps, we propose a categorization system that classifies dissimilarity measures into four types based on their scaling procedures, data transformations, and the geometric spaces in which distances are computed. This classification helps clarify which dissimilarities are proper metrics, which are Euclidean, and which may pose challenges in downstream analyses such as ordination or hypothesis testing. Table 1 summarizes four major categories. The discussion on the categorizing commonly used dissimilarity measures is detailed in Supplementary Section 2.

Category	Space	Class	Dissimilarity	Metric?	Euclidean?
scale difference	$\mathbb{R}^n$ or S^n	$D_1(\mathbf{x},\mathbf{y}) = \lVert \psi_1(\mathbf{x}) - \psi_1(\mathbf{y}) \rVert_p$	Euclidean	Yes	Yes
			Manhattan	Yes	No
			Mahalanobis	Yes	Yes
			Chord	Yes	Yes
			Aitchison	Yes	Yes
			Robust Aitchison	Yes	Yes
			Chi-square	Yes	Yes
difference scale	$\mathbb{R}^n$ or S^n	$D_2(\mathbf{x},\mathbf{y}) = \lVert \psi_2(\mathbf{x} - \mathbf{y}) \rVert_p$	Bray-Curtis	No	No
			Kulczyński	Unclear	Unclear
			Weighted UniFrac	Likely	No
presence difference	$\mathbb{H}^n$	$D_3(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x}\cdot\mathbf{y}}{\psi_H(\mathbf{x},\mathbf{y})}$	Jaccard	Yes	No
			Unweighted UniFrac	Yes	No
distribution difference	Distributions	$D_f(P\lVert Q) = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ$	Hellinger	Yes	Yes
			Jensen-Shannon	Yes	Unclear

Table 1: Classification of commonly used dissimilarity measures. Measures Category 1 operate in Euclidean space (abundance data) or compositional simplex (relative abundance data) and compute the L_p norm after applying a normalization function ψ_1 to each data point individually before calculating the difference. Measures in Category 2 also operate in Euclidean space or compositional simplex but differ by first computing the difference between data points and then applying a normalization function ψ_2. Measures in Category 3 are Defined in the Hamming space $\mathbb{H}^n$ and usually assess dissimilarity based on presence-absence data. Measures in Category 4 measures the dissimilarity between probability mass functions, which can also be applied to relative abundance data.

Category 1 (scale difference): Transform-then-Distance (Euclidean embeddings). Measures in this group apply a normalization function $\psi_1(\cdot)$ to each data point individually (e.g., abundance or relative abundance), and then compute the L_p distance between the transformed vectors:

$$D_1(\mathbf{x}, \mathbf{y}) = \|\psi_1(\mathbf{x}) - \psi_1(\mathbf{y})\|_p.$$

Common examples include Euclidean, Mahalanobis, and Aitchison distances. Notably, when $\psi_1(\mathbf{x})$ maps into a Euclidean space (e.g., CLR-transformed compositions), the resulting dissimilarities naturally preserve Euclidean geometry, ensuring compatibility with ordination methods like PCoA and hypothesis tests like PERMANOVA.

Category 2 (difference scale): Normalize-after-Difference (non-Euclidean forms). This group computes element-wise differences between communities before normalization:

$$D_2(\mathbf{x}, \mathbf{y}) = \|\psi_2(\mathbf{x} - \mathbf{y})\|_p.$$

Examples include Bray–Curtis and Weighted UniFrac. These measures often lack a clear geometric embedding and may violate the triangle inequality or the Euclidean property. As a result, distortions can arise in PCoA plots, and negative eigenvalues in the Gram matrix are common. We provide numerical evidence in Section 4.

Category 3 (presence difference): Presence–absence dissimilarities (Hamming space). Presence–absence profiles lie in the binary space $\mathbb{H}^K = \{0, 1\}^K$, where dissimilarity is derived from shared and mismatched entries. Measures like Jaccard and unweighted UniFrac can be written as:

$$D_3(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\psi_H(\mathbf{x}, \mathbf{y})},$$

where the denominator counts the number of taxa present in either community. Due to the discrete geometry of Hamming space, these measures typically lack Euclidean embeddability and should be carefully evaluated before downstream analysis.

Category 4 (distribution difference): Divergence-based dissimilarities (probability space). Relative abundances can also be viewed as probability mass functions. In this view, f-divergences provide a general framework:

$$D_f(P\|Q) = \sum_{k=1}^{K} q_k f\left(\frac{p_k}{q_k}\right),$$

where f is a convex function such as the Hellinger or Jensen–Shannon form. These divergences are intuitive and interpretable as average log-ratio deviations, but not all satisfy the triangle inequality or support Euclidean embedding.

2.4 Diagnostics and Remedial Corrections

As discussed in Sections 2.1 and 2.2, many widely used dissimilarity measures in microbial ecology lack the Euclidean property, which may lead to negative eigenvalues in the associated Gram matrix and distort downstream analyses such as ordination or hypothesis testing. While structural classifications (Section 2.3) provide useful heuristics, rigorously establishing whether a given dissimilarity is metric or Euclidean remains theoretically nontrivial and often depends on the characteristics of the data [76, 100, 101]. To address this challenge, we develop diagnostic tools for empirically assessing these properties and propose practical remedies for cases where non-Euclidean dissimilarities must be used. The implementations, along with reproducible examples, are available at https://github.com/bioscinema/Betadiag.

As discussed in Section 2.2, a dissimilarity matrix D satisfies the metric property if, for any three distinct samples i, j, k, the triangle inequality holds:

$$D_{ik} \leq D_{ij} + D_{jk}.$$

We implement an `R` function to exhaustively verify this condition across all triplets in the dataset. In addition, we define two sample-level diagnostic scores that quantify the extent to which each sample contributes to collinearity or nonlinearity in the dissimilarity structure:

$$\text{CollinearityScore}(i) = \frac{1}{\binom{N}{3}} \sum_{j,k:\, i \neq j \neq k} I(D_{ik} = D_{ij} + D_{jk}), \tag{1}$$

$$\text{NonlinearityScore}(i) = \frac{1}{\binom{N}{3}} \sum_{j,k:\, i \neq j \neq k} I(D_{ik} > D_{ij} + D_{jk}). \tag{2}$$

A high collinearity score indicates that the dissimilarity involving sample i is exactly additive with respect to other samples, suggesting potential redundancy. In contrast, a high nonlinearity score implies that triangle inequality is violated in configurations involving i, suggesting that its information may be poorly captured in methods assuming a Euclidean embedding.

To assess the Euclidean nature of D, we construct its centered Gram matrix G and perform eigen-decomposition of G and compute the FNI introduced in Section 2.2.5. A nonzero FNI indicates that D is not Euclidean, and the associated Gram matrix is not positive semi-definite. In such cases, linear dimensionality reduction techniques such as PCoA will introduce geometric distortions, as negative eigenvalues lack interpretable variance decomposition. These diagnostic tools offer a computationally efficient and interpretable framework for assessing the geometric adequacy of dissimilarity matrices, informing the choice of ordination methods and guiding the application of correction procedures when necessary.

When non-Euclidean structure is detected, a natural solution is to select an alternative dissimilarity measure that satisfies Euclidean properties. However, in cases where a non-Euclidean dissimilarity is preferred due to ecological interpretability or established usage, several remedial techniques can be employed to mitigate its limitations in downstream analysis.

One widely used class of remedies involves directly adjusting the dissimilarity matrix to restore metric properties. The `ape` R package [102] provides two such corrections, namely the Cailliez [103] and Lingoes [104] methods, which aim to remove negative eigenvalues from the centered Gram matrix. The Cailliez method identifies the smallest constant $c \geq 0$ such that adding c uniformly to all off-diagonal entries of the dissimilarity matrix ensures that the resulting Gram matrix becomes positive semi-definite. This approach follows the analytical solution proposed by Gower and Legendre [76] and implemented in `ape`. In contrast, the Lingoes method applies a transformation to the squared dissimilarity matrix, adding a constant sufficient to render the matrix conditionally negative definite. Specifically, it transforms the original distances via:

$$D'_{ij} = \sqrt{D_{ij}^2 + c}, \quad \text{where } c = |\lambda_{\min}|,$$

and $\lambda_{\min}$ denotes the most negative eigenvalue of the original Gram matrix. This adjustment typically introduces less geometric distortion compared to the Cailliez correction, as it preserves relative relationships among samples more faithfully. Importantly, while both corrections ensure that the modified dissimilarity matrix satisfies the metric condition, they do not guarantee that the resulting distances are strictly Euclidean. That is, negative curvature may still persist in the embedded space, and downstream analyses may continue to be affected by residual non-Euclidean structure.

Notably, many widely used microbiome analysis tools—including the `ordinate` function in the `phyloseq` package [69]—do not apply these corrections by default. As a result, it remains the user's responsibility to evaluate and, if needed, preprocess dissimilarity matrices prior to applying ordination methods that assume Euclidean geometry. Similar limitations apply to other analysis pipelines, underscoring the practical relevance of explicit diagnostics and corrections.

When direct modification of the dissimilarity matrix fails to ensure Euclidean structure, an alternative strategy is to adjust the corresponding Gram matrix G to enforce positive semidefiniteness (PSD). We introduce two matrix-level remedies: Higham's nearest PSD approximation [105] and Tikhonov regularization [106].

Higham's method aims to find the closest PSD matrix G^+ to a given symmetric but indefinite Gram matrix G, under the Frobenius norm. The approach begins with the eigen-decomposition $G = Q\Lambda Q^\top$, where Λ contains the eigenvalues and Q the corresponding eigenvectors. Negative eigenvalues in Λ are replaced with zeros to form Λ^+, and the corrected matrix is reconstructed as $G^+ = Q\Lambda^+ Q^\top$. Iterative projections are applied to refine the approximation, ensuring symmetry and numerical stability. This procedure is widely used in multidimensional scaling and kernel methods, as it guarantees the minimal Frobenius perturbation while restoring the PSD property.

Tikhonov regularization provides an alternative strategy by adding a small ridge term to the diagonal of G, yielding the adjusted matrix $G_\alpha = G + \alpha I$, where $\alpha > 0$ is a regularization parameter and I is the identity matrix. This uniform shift raises all eigenvalues by α, rendering G_α PSD when α exceeds the absolute value of the most negative eigenvalue. Compared to Higham's method, Tikhonov regularization minimally distorts the eigenspectrum and preserves the relative contributions of all eigencomponents. However, it modifies the eigenvectors and may introduce rotational distortions in the embedded space.

The choice between these two remedies depends on the analytical goals. Higham's method strictly enforces non-negativity of eigenvalues and preserves the original eigenvectors, making it suitable when eigenspace fidelity is essential (e.g., in kernel PCA or spectral clustering). However, by truncating all negative eigenvalues, this approach may lead to a loss of information carried by small but potentially meaningful components [27]. In contrast, Tikhonov regularization maintains the variance distribution across components and avoids abrupt truncation, making it preferable in applications that prioritize smooth eigenvalue decay and numerical conditioning. In practice, we recommend evaluating both methods when correcting non-PSD Gram matrices, and selecting based on their impact on percent variance explained (PVE) and stability in downstream analysis.

Beyond direct corrections to the dissimilarity or Gram matrix, recent methods have been developed to accommodate non-Euclidean structure without enforcing strict Euclidean constraints. Deng et al. [107] proposed Non-Euclidean Multidimensional Scaling (Neuc-MDS), which extends classical MDS by incorporating negative eigenvalues into a generalized bilinear form, allowing sample configurations to be embedded in non-metric spaces while minimizing the STRESS criterion. Gisbrecht and Schleif [108] introduced a scalable strategy for transforming non-metric dissimilarities into positive-definite kernel matrices. Their approach combines Nystroem approximation, double centering, and eigenvalue correction, offering linear computational complexity and suitability for large-scale microbiome datasets.

Together, these strategies provide a flexible toolkit for addressing non-Euclidean dissimilarities in microbial community analysis. Depending on the analytical objective—whether prioritizing ecological interpretability, geometric fidelity, or computational scalability—researchers can select from a range of diagnostic and corrective methods to ensure valid and robust downstream inference.

3 Results

3.1 Datasets and Analysis Strategy

Building on the theoretical framework in previous sections, we now turn to empirical evaluations across multiple real-world microbiome datasets. Our objective is to examine how often violations of metric and Euclidean properties in commonly used beta diversity measures can be found, and evaluate whether remedial transformations can mitigate such issues. To achieve this objective, we develop a structured evaluation pipeline that quantifies the prevalence of property violations and evaluates their impact on downstream analyses.

1. **Distance Matrix Construction.** For each dataset, we compute pairwise dissimilarities using a diverse panel of beta diversity measures, including presence–absence-based (e.g., Jaccard), abundance-based (e.g., Bray–Curtis), phylogenetically informed (e.g., UniFrac), and compositional (e.g., Aitchison) distances.
2. **Assessment of Metric Violations.** We assess the triangle inequality by evaluating all sample triplets to identify violations and quantify their severity using the

number of violations, collinearity, and nonlinearity scores, which provide insight into metric distortions and structural dependencies.

3. **PCoA and Euclidean Properties.** We perform Principal Coordinates Analysis (PCoA) and compute the Fraction of Negative Inertia (FNI), which quantifies the proportion of variance attributed to negative eigenvalues. An FNI of zero indicates that the dissimilarity matrix is fully embeddable in a Euclidean space, while higher values reflect increasing deviation from Euclidean geometry. This provides a diagnostic of how severely a given beta diversity measure violates the Euclidean property and potentially distorts the resulting ordination.
4. **Association via PERMANOVA.** We apply PERMANOVA [109] to assess how well each dissimilarity measure captures compositional variation associated with key covariates. Performance is evaluated using the pseudo-F statistic and R^2, which quantify the strength and proportion of group-level separation, respectively. Higher values of pseudo-F and R^2 indicate stronger covariate-associated structure, reflecting the informativeness of a dissimilarity measure in detecting biologically meaningful patterns.
5. **Association via MiRKAT.** We apply the MiRKAT framework [19, 110, 111], which uses dissimilarity-induced kernels to test associations with continuous or categorical outcomes. Performance is evaluated using the R^2 statistic, as the permutation-based nature of MiRKAT renders p-values independent of the probability distribution of test statistic and less informative about effect size [112, 113]. Given the typically weak statistical signals in microbiome data, we focus on R^2 as a measure of explanatory power.
6. **Remedial Corrections.** To address non-metric and non-Euclidean distortions, we apply Higham's matrix correction and Tikhonov regularization. We then re-run PERMANOVA and MiRKAT using the corrected matrices to evaluate improvements in pseudo-F and R^2.

We apply the structured analysis pipeline to six diverse microbiome datasets to evaluate the behavior of different beta diversity measures across varied biological contexts.

(1) *Cardiometabolic Dataset [114]:* This dataset from the MetaCardis consortium profiles gut microbiome and metabolites in 1,241 middle-aged Europeans across healthy, dysmetabolic, and ischemic heart disease (IHD) groups. Fecal, serum, and urine samples were collected alongside extensive clinical phenotyping to identify cardiometabolic signatures while adjusting for confounders such as polypharmacy and metabolic disorders.

(2) *COVID-Gut Microbiome Dataset [115]:* This dataset examines gut microbiome alterations associated with acute SARS-CoV-2 infection and post-recovery states. Conducted at Rutgers University, it includes 60 adults (20 COVID-19 positive, 20 recovered, and 20 healthy controls), with stool samples analyzed by 16S rRNA sequencing. The study identifies sustained depletion of commensal taxa (e.g., Bacteroidaceae, Ruminococcaceae) post-infection, exacerbated by antibiotic use.

(3) *IBD Dataset [116, 117]:* This UC-focused dataset integrates metagenomic, proteomic, and metabolomic profiles to investigate microbial protease activity in

inflammatory bowel disease. It includes 40 UC patients from UCSD and a validation cohort of 210 individuals (73 UC, 117 Crohn's disease, 20 healthy controls). The study identifies overabundant Bacteroides vulgatus proteases linked to disease severity.

(4) *Pig Gut Microbiome Dataset [118]:* This animal model investigates the impact of long-term fructose and resistant starch intake on gut microbiota and metabolic profiles in Göttingen minipigs. Juvenile pigs were fed high-risk or lower-risk diets over five months, with fecal, plasma, and urine samples analyzed by metagenomics and metabolomics.

(5) *ORIGINS Oral Microbiome Dataset [119, 120]:* Part of the ORIGINS study, this dataset evaluates associations between subgingival microbial communities and prediabetes risk. The dataset includes 300 adults aged 20 to 55, with a mean age of 34 years and 77% female participants, from a cohort in the Service Employees International Union 1199. Plaque samples from 1,188 sites were analyzed for 11 periodontitis-related species, alongside clinical periodontal assessments.

(6) *HIV-Gut Microbiome Dataset [121]:* This dataset explores the relationship between gut microbiome composition, metabolic disease, and systemic inflammation in people living with HIV (PLWH) and high-risk men who have sex with men (HR-MSM). The dataset includes stool and plasma samples from 113 men, categorized as HIV-negative, untreated HIV-positive, and ART-treated HIV-positive individuals, with further stratification based on lipodystrophy (LD) status. Gut microbiota composition was analyzed using 16S rRNA sequencing, and plasma metabolomics and immune markers were assessed to identify metabolic disease predictors.

3.2 Results

Figure 4 summarizes the metric properties of commonly used beta diversity measures across the six datasets. Each dissimilarity is evaluated for adherence to the triangle inequality, with blue indicating satisfaction of the metric condition and red indicating violations. As expected from theoretical considerations (Table 1), Euclidean, Manhattan, Mahalanobis, Aitchison, and related L_p-based dissimilarities consistently satisfy the metric property across all datasets. Notably, Mahalanobis distance could not be computed for the Cardiometabolic dataset (Dataset (1)) due to singularity in the covariance matrix—a common occurrence when the number of features exceeds the number of samples, as is typical in microbiome data. In contrast, several widely used measures—most notably Bray–Curtis and both versions of weighted UniFrac—frequently violate the triangle inequality, indicating non-metric behavior in specific datasets.

Bray–Curtis, although widely adopted in microbiome research, fails to satisfy the triangle inequality and is often referred to as a semi-metric [122, 123]. It operates on absolute abundance differences without accounting for compositional constraints or phylogenetic structure, making it susceptible to distortion in high-dimensional microbial data. Our findings confirm its non-metric behavior across all datasets, reinforcing concerns about its use in distance-based analyses. In contrast, Jaccard dissimilarity—a presence-absence–based metric [21, 22, 124]—consistently satisfies the triangle inequality.

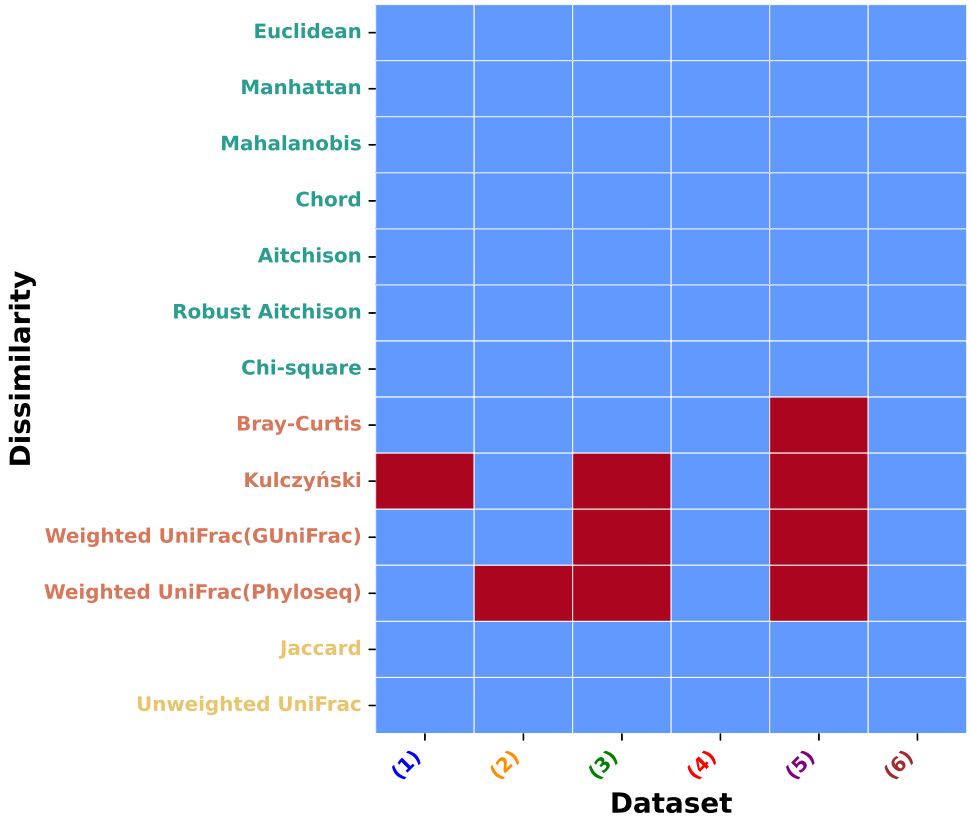

Fig. 4: *Metric property of dissimilarity measures across datasets.* This heatmap indicates whether each dissimilarity measure satisfies the metric property within each dataset. Blue squares denote that the dissimilarity is metric; red squares indicate violations of the metric condition. Dissimilarities are grouped and color-coded by category to facilitate interpretation across types of measures.

Interestingly, the metric property of weighted UniFrac is software-dependent. The definition implemented in `QIIME 2`, which follows the formulation in [38], ensures metric properties by omitting the normalization denominator. In contrast, the versions implemented in `phyloseq` and `GUniFrac` retain the original form with a denominator [41], leading to systematic violations of the triangle inequality. Our empirical results align with the latter: both `phyloseq`- and `GUniFrac`-derived distances consistently fail to meet the metric condition, in contrast to the theoretically valid but less commonly used `QIIME 2` variant.

It is important to note that even dissimilarities that qualify as formal metrics (e.g., Jaccard, unweighted UniFrac) may still yield non-Euclidean distance matrices,

potentially distorting ordination outcomes. Thus, while metric compliance is a necessary prerequisite for distance-based modeling, it does not guarantee faithful Euclidean embedding—a distinction.

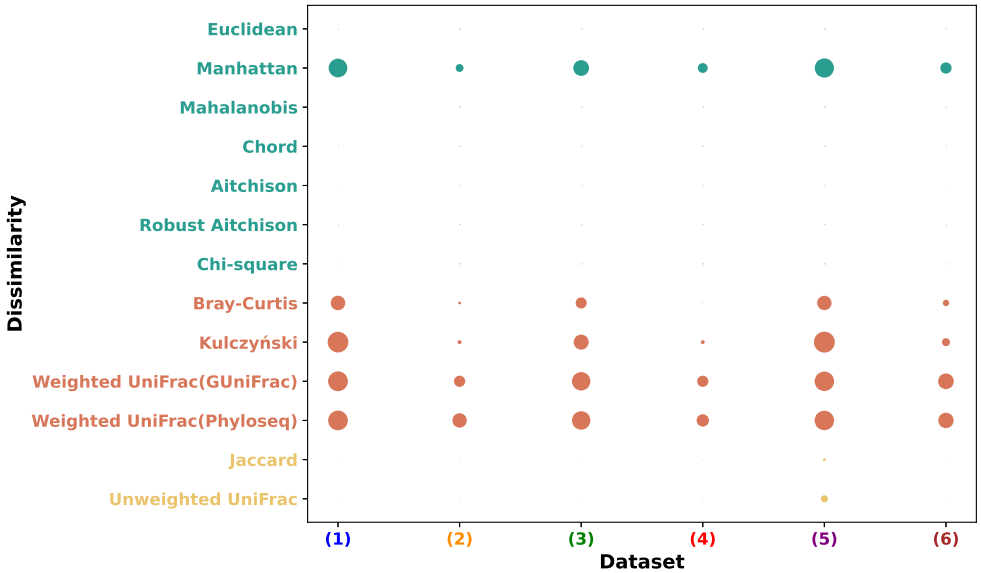

Fig. 5: *Euclidean deviation (FNI) of dissimilarity measures across datasets.* The bubble plot visualizes the Fraction of Negative inertia (FNI) for each dissimilarity measure across datasets. Larger bubbles correspond to higher FNI values, indicating greater deviation from Euclidean geometry. An FNI of zero implies that the dissimilarity is Euclidean for that dataset. Dissimilarity labels are color-coded by category.

Figure 5 visualizes the Euclidean deviation of each dissimilarity measure using the Fraction of Negative Inertia (FNI), defined as the proportion of variance explained by negative eigenvalues in principal coordinates analysis (PCoA). An FNI of zero indicates that the dissimilarity matrix admits an exact Euclidean embedding, whereas higher values signify increasing departures from Euclidean geometry. Combined with the metric diagnostics in Figure 4 and theoretical classifications in Table 1, this analysis highlights the nuanced geometric behavior of commonly used beta diversity measures applied in certain datasets.

As expected, Euclidean distance consistently yields zero FNI across all datasets, as expected from its inner-product–based formulation. This confirms that dissimilarities constructed from L_2 norms preserve Euclidean geometry and can be faithfully represented in principal coordinates analysis (PCoA). In contrast, other L_p-based dissimilarities such as Manhattan (L_1) and Mahalanobis—while valid metrics—often yield nonzero FNI values, indicating imperfect Euclidean embeddability. Manhattan distance, in particular, is prone to geometric distortions in high-dimensional microbiome data due to its tendency to produce collinear sample configurations, and

Mahalanobis becomes sensitive to ill-conditioned or singular covariance structures. These results underscore that only L_2-based dissimilarities (Chord, Aichison, Chi-sq) inherently guarantee a Euclidean embedding.

In contrast, Bray–Curtis, which fails the triangle inequality and is formally a semi-metric [34, 122], consistently produces nonzero FNI. These distortions are known to affect ordination results [125], introducing artifacts in low-dimensional projections that can mask or exaggerate true ecological gradients [126, 127]. Our results reinforce these concerns by demonstrating that Bray–Curtis not only lacks metric validity but also substantially deviates from Euclidean geometry in empirical data.

Interestingly, some dissimilarities that satisfy the triangle inequality—such as Jaccard and unweighted UniFrac [18, 38, 124]—also exhibit moderate FNI. This highlights the important distinction between metric and Euclidean properties: being a metric is necessary but not sufficient for guaranteeing a faithful Euclidean representation. For example, Jaccard, while formally metric [21, 22], can be highly sensitive to rare taxa, while unweighted UniFrac is susceptible to exaggerated influence from deep phylogenetic branches [128]. Both can introduce non-Euclidean artifacts in PCoA, despite satisfying metric axioms.

Weighted UniFrac shows the highest and most variable FNI values across datasets. As discussed in Figure 4, its definition is software-dependent: the original form with a normalization denominator [41], used in `phyloseq` and `GUniFrac`, fails to satisfy metric properties, while the alternative formulation implemented in `QIIME 2` [38] omits the denominator and restores the metric condition. However, even under the latter, the weighted UniFrac remains a tree-based distance and thus inherently non-Euclidean [129, 130]. This explains the persistent presence of negative eigenvalues across all weighted UniFrac variants, emphasizing that correcting for metricity does not guarantee Euclidean behavior, especially for phylogeny-informed distances.

Taken together, these findings underscore the importance of distinguishing between metric validity and Euclidean embeddability when selecting or interpreting dissimilarity measures. While the former ensures theoretical consistency in distance-based modeling, the latter is essential for ordination and kernel-based analyses that rely on inner product geometry.

Figure 6 summarizes the overall structural robustness of each dissimilarity measure using a rose diagram, which aggregates how often a measure satisfies both the metric and Euclidean properties across datasets. A full bar (value of 6) indicates that a dissimilarity consistently behaves as a valid metric and admits a Euclidean embedding, suggesting strong geometric reliability. This integrated view complements the results from Figures 4 and 5, allowing a more intuitive comparison of robustness across dissimilarity categories.

Overall, dissimilarities based on L_2 norms—such as Euclidean, Chord, Aitchison, Robust Aitchison, and Chi-square—demonstrate the highest structural stability, satisfying both the triangle inequality and Euclidean geometry in nearly all datasets. This is consistent with their theoretical formulations in Table 1, which position them in Category 1. These distances are particularly well suited for ordination and kernel-based

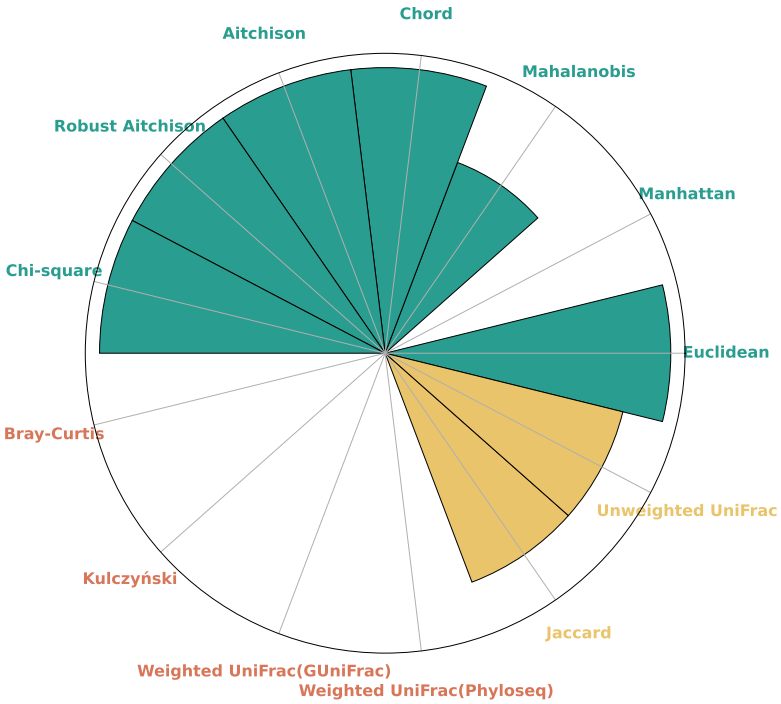

Fig. 6: *Frequency of metric and Euclidean behavior across datasets.* This rose diagram summarizes, for each dissimilarity measure, the number of datasets in which it is both metric and Euclidean (i.e., FNI = 0). A full bar (value of 6) indicates consistent metric and Euclidean behavior across all datasets, suggesting greater structural reliability. Bar colors reflect the dissimilarity category.

inference methods, as they preserve inner-product structure and avoid negative eigenvalues in PCoA. Their robustness also enhances the interpretability of downstream clustering or embedding [125, 126].

In contrast, dissimilarities in Categories 2 and 3 exhibit more variable behavior. Bray–Curtis, although widely used, fails both the metric and Euclidean conditions in all datasets, confirming its semi-metric nature [122, 123]. As noted in prior studies [126, 127], such measures can introduce spurious clustering patterns in compositional datasets, where samples from distinct biological groups may appear artificially close due to projection artifacts. This underscores the risk of using structurally unstable distances in ordination-based microbiome studies. Weighted UniFrac performs particularly poorly in this regard: both the `phyloseq` and `GUniFrac` implementations fail to meet metric or Euclidean criteria in any dataset. Even the theoretically valid `QIIME 2` variant, while satisfying metricity, remains non-Euclidean due to its dependence on phylogenetic tree structure. This highlights a broader limitation of Category 2 dissimilarities, which may incorporate biologically relevant information but at the cost of geometric interpretability. Interestingly, presence–absence–based measures

such as Jaccard and unweighted UniFrac (Category 3) show intermediate robustness. They satisfy the metric condition consistently, but moderate FNI values indicate partial deviation from Euclidean geometry. This reflects a fundamental trade-off: while these measures emphasize rare taxa and phylogenetic breadth, they are more prone to introducing ordination distortions that could obscure true ecological gradients.

Taken together, these findings suggest that the choice of dissimilarity measure should balance biological relevance with geometric stability. While metrics like Bray–Curtis or UniFrac may capture meaningful ecological or phylogenetic signals, their geometric fragility can impair inference in distance-based methods such as PERMANOVA [131] and MiRKAT [19]. Robust dissimilarities—those that maintain both metric and Euclidean properties across datasets—offer greater consistency, reduce the risk of spurious associations, and improve the interpretability of multivariate analyses in microbiome research.

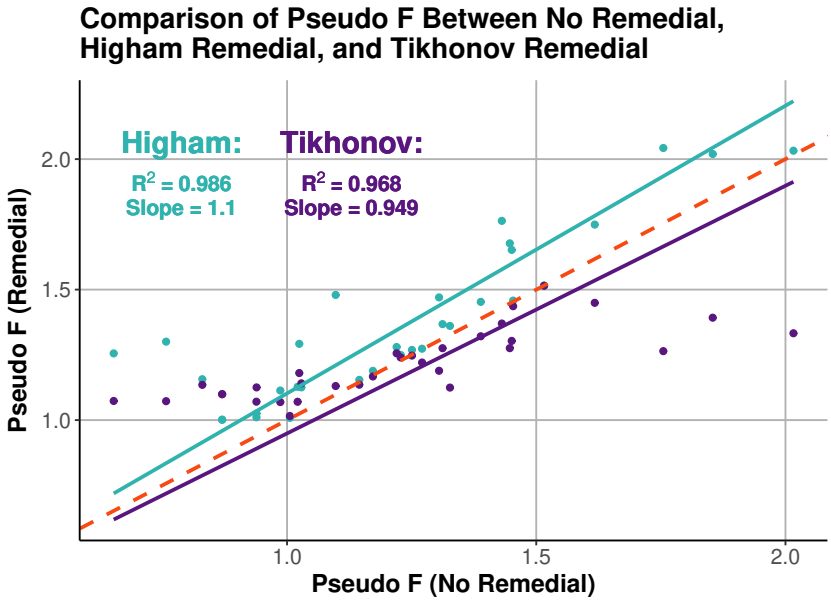

Fig. 7: Comparison of PERMANOVA pseudo-F values before and after remedial methods. The horizontal axis shows pseudo-F values obtained using a non-Euclidean dissimilarity measure ("No Remedial"), while the vertical axis shows the corrected pseudo-F values after applying the Higham (teal) or Tikhonov (purple) remedial methods. A red dashed line indicates the 1:1 reference. Regression lines and statistics are shown for each remedial approach; notably, the Higham method increases the original pseudo-F values by roughly 11%.

For dissimilarity matrices that fail to satisfy Euclidean geometry, we apply two remedial correction methods—Higham projection and Tikhonov regularization—and

compare their effects on PERMANOVA pseudo-F statistics, as shown in Figure 7. The horizontal axis shows pseudo-F values computed using uncorrected, non-Euclidean dissimilarities ("No Remedial"), while the vertical axis shows the corresponding values after correction.

The Higham method strictly enforces positive semidefiniteness (PSD) by removing all negative eigenvalues from the associated Gram matrix. This results in a systematic inflation of pseudo-F values, with a regression slope of 1.10 and $R^2 = 0.986$, indicating an average 10% increase relative to the original statistics. This consistent upward shift suggests that Higham correction enhances apparent group separability, though it may also exaggerate signal strength due to aggressive spectral truncation. In contrast, the Tikhonov method shifts all eigenvalues uniformly by adding a ridge constant, preserving their relative structure while correcting non-Euclidean behavior. This yields more moderate adjustments to the pseudo-F values (slope = 0.949, $R^2 = 0.968$), reflecting improved numerical stability with less risk of over-amplification.

These results highlight a tradeoff between sensitivity and spectral fidelity: Higham correction may increase statistical power but at the cost of altering the geometry more substantially, whereas Tikhonov offers a more conservative correction. The choice of remedy should therefore be guided by the intended downstream application—e.g., signal detection versus geometry preservation.

Complete results for all assessments—including triangle inequality violations, Euclidean deviation with FNI, and the effects of Higham and Tikhonov remedial corrections on PERMANOVA and MiRKAT—are provided in Supplementary Sectgion 3.

4 Discussion

In this study, we established a unified mathematical framework for characterizing beta diversity dissimilarities. By systematically analyzing their mathematical formulations, we categorized widely used dissimilarity measures into four structural classes. We further identified three fundamental mathematical properties that an ideal dissimilarity measure should satisfy and analyzed the relationships between these structural categories and the key theoretical properties. This classification framework facilitates intuitive understanding for researchers and practitioners to select appropriate measures in ecological and microbiome studies , and serves as a foundation for the development of new dissimilarity measures with specific desired properties.

To bridge theory and practice, we developed an R-based diagnostic toolkit to assess whether a given dissimilarity matrix satisfies key geometric properties. Our empirical results, drawn from diverse real-world datasets, demonstrate that Euclidean property—more than metricity—is a critical determinant of downstream statistical validity. Notably, the weighted UniFrac distance, though a biologically appealing measure incorporating both abundance and phylogeny, often violates the Euclidean condition and introduces negative eigenvalues in PCoA, undermining its interpretability in multivariate analyses.

Despite these advances, several limitations warrant discussion. First, the scope of datasets used for evaluation remains limited, and the observed violations of Euclidean

properties may not generalize across all sample types or environmental contexts. Second, while we provide geometric diagnostics and numerical corrections, the ecological consequences of modifying distance matrices (e.g., via Tikhonov or Higham adjustments) remain to be fully understood. Third, our analysis focused on beta diversity in isolation, without incorporating domain-specific biological factors such as trait structure or spatial autocorrelation.

Future work should explore the robustness of dissimilarity measures to biological confounders such as dominant taxa or rare species inflation. Another promising direction lies in examining how zero-inflated abundance distributions affect dissimilarity-based inference. Since microbiome data often exhibit sparsity and heavy-tailed behavior, it is essential to understand how various measures perform under these realistic constraints. For example, [132] introduced the "double-zero problem," emphasizing that shared absences between two communities should not contribute to their similarity. Later, [81] argued that dissimilarity values should remain unchanged when common zeros are added to both communities. While these studies highlight the potential impact of zero inflation on dissimilarity measures, a formal mathematical framework to rigorously characterize this effect, quantify the sensitivity of different measures to zero inflation, and assess the effectiveness of proposed remedies is still lacking. Our work provides such a framework and establishes a solid foundation for further investigation.

Furthermore, the integration of beta diversity concepts into multi-omics analysis represents a compelling extension. Recent advances in cross-omics variation modeling, especially in transcriptomics, proteomics, and metabolomics, call for interpretable and geometrically valid dissimilarities to compare sample profiles across omic layers. Beta diversity measures can serve as dissimilarity-based kernels or embedding layers for integrative models, aiding in sample classification, functional alignment, and pathway concordance detection. However, to achieve this goal, it is essential to design dissimilarity measures that not only preserve biological interpretability but also maintain desirable geometric properties across heterogeneous omics domains. Our work lays the foundation for such developments.

5 Conclusions

Beta diversity and microbiome data analysis exemplify the kind of interdisciplinary integration that unites biologists, ecologists, mathematicians, and statisticians. Some dissimilarity measures were originally developed to address specific biological questions but lack the mathematical rigor needed for broader analytical reliability. Conversely, mathematically sound distance measures have been adopted to quantify beta diversity, yet they often fall short in capturing biological meaning or meeting practical needs in the field. The sheer number of available metrics has led to widespread—sometimes inappropriate—use, with key assumptions often left unexamined. Paradoxically, the familiarity of certain measures may lead researchers to overlook their theoretical limitations and applicability.

Therefore, we present a geometry-aware framework for the evaluation and application of beta diversity dissimilarity measures. By systematically classifying dissimilarities based on their mathematical formulations and examining their theoretical properties, we identify key structural limitations that compromise the validity of downstream ordination and hypothesis testing. Through diagnostic tools and corrective procedures, we offer practical solutions to address non-Euclidean artifacts and improve inference reliability.

Our findings highlight the importance of aligning ecological interpretation with statistical rigor. In particular, we show that Euclidean compatibility is essential for dissimilarity-based multivariate methods to yield meaningful and robust results. This work not only provides actionable guidance for current microbiome studies but also opens new avenues for integrating beta diversity analysis into cross-omics research and systems biology.

Data availability

All data supporting the findings of this study are included within the Article and its Supplementary Information files. Raw data of COVID-Gut Microbiome Dataset are publicly available via QIITA (accession 14812). Raw data of HIV-Gut Microbiome Dataset is accessible via the European Bioinformatics Institute (EBI) under accession number ERP125300 and the Metabolomics Workbench under study ID ST001750, subject to institutional agreements and ethical guidelines. The `Betadiag` software implementing the proposed method—together with the complete scripts and outputs for all real-data analyses—is openly available at: https://github.com/bioscinema/Betadiag. A user-friendly web-based Shiny application for `Betadiag` is also available at: https://yiqianzhang.shinyapps.io/Betadiag/.

Abbreviations

PCoA: Principal Coordinates Analysis
IOS: Illusion of Similarity
CND: Conditional Negative Definiteness
mMDS: Metric Multidimensional Scaling
DPCoA: Double Principal Coordinates Analysis
PSD: Positive Semi-definiteness
FNI: Fraction of Negative Inertia
CLR: Centered Log Ratio
PVE: Percent Variance Explained
Neuc-MDS: Non-Euclidean Multidimensional Scaling

References

[1] Real, L.A., Brown, J.H.: Foundations of Ecology:\Classic Papers with Commentaries. University of Chicago Press, ??? (2012)

[2] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The human microbiome project. Nature **449**(7164), 804–810 (2007)

[3] Wooley, J.C., Godzik, A., Friedberg, I.: A primer on metagenomics. PLoS computational biology **6**(2), 1000667 (2010)

[4] Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J., Relman, D.A.: The application of ecological theory toward an understanding of the human microbiome. Science **336**(6086), 1255–1262 (2012)

[5] Gilbert, J.A., Lynch, S.V.: Community ecology as a framework for human microbiome research. Nature medicine **25**(6), 884–889 (2019)

[6] Zhu, Y.-G., Zhu, D., Rillig, M.C., Yang, Y., Chu, H., Chen, Q.-L., Penuelas, J., Cui, H.-L., Gillings, M.: Ecosystem microbiome science. Mlife **2**(1), 2–10 (2023)

[7] Whittaker, R.H.: Vegetation of the siskiyou mountains, oregon and california. Ecological monographs **30**(3), 279–338 (1960)

[8] Koleff, P., Gaston, K.J., Lennon, J.J.: Measuring beta diversity for presence–absence data. Journal of Animal Ecology **72**(3), 367–382 (2003)

[9] Pane, C., Sorrentino, R., Scotti, R., Molisso, M., Di Matteo, A., Celano, G., Zaccardelli, M.: Alpha and beta-diversity of microbial communities associated to plant disease suppressive functions of on-farm green composts. Agriculture **10**(4), 113 (2020)

[10] Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., Cornell, H.V., Comita, L.S., Davies, K.F., *et al.*: Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology letters **14**(1), 19–28 (2011)

[11] Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat **37**, 547–579 (1901)

[12] Sorensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biologiske skrifter **5**, 1–34 (1948)

[13] MacArthur, R.H.: Patterns of species diversity. Biological reviews **40**(4), 510–533 (1965)

[14] Cody, M.L., Diamond, J.M.: Ecology and Evolution of Communities. Harvard University Press, ??? (1975)

[15] Shmida, A., Wilson, M.V.: Biological determinants of species diversity. Journal of biogeography, 1–20 (1985)

[16] Faith, D.P.: Conservation evaluation and phylogenetic diversity. Biological conservation **61**(1), 1–10 (1992)

[17] Petchey, O.L., Gaston, K.J.: Functional diversity: back to basics and looking forward. Ecology letters **9**(6), 741–758 (2006)

[18] Lozupone, C., Knight, R.: Unifrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology **71**(12), 8228–8235 (2005)

[19] Zhao, N., Chen, J., Carroll, I.M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J.J., Ringel, Y., Li, H., Wu, M.C.: Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. The American Journal of Human Genetics **96**(5), 797–807 (2015)

[20] Gail, M.H., Wan, Y., Shi, J.: Power of microbiome beta-diversity analyses based on standard reference samples. American Journal of Epidemiology **190**(3), 439–447 (2021)

[21] Kosub, S.: A note on the triangle inequality for the jaccard distance. Pattern Recognition Letters **120**, 36–38 (2019)

[22] Lipkus, A.H.: A proof of the triangle inequality for the tanimoto distance. Journal of Mathematical Chemistry **26**(1), 263–265 (1999)

[23] Levandowsky, M., Winter, D.: Distance between sets. Nature **234**(5323), 34–35 (1971)

[24] Ioffe, S.: Improved consistent sampling, weighted minhash and l1 sketching. In: 2010 IEEE International Conference on Data Mining, pp. 246–255 (2010). IEEE

[25] Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**, 1–28 (2015)

[26] Kulczyński, S.: Die Pflanzenassoziationen der Pieninen vol. 3. Imprimerie de l'Université, Lwów (1928)

[27] Legendre, P.: Numerical Ecology. Elsevier, Amsterdam (2012)

[28] Deza, M.-M., Deza, E.: Dictionary of Distances. Elsevier, Amsterdam (2006)

[29] Shi, G.: A comparative study of 39 binary similarity coefficients (1993)

[30] Cheetham, A.H., Hazel, J.E.: Binary (presence-absence) similarity coefficients. Journal of Paleontology, 1130–1136 (1969)

[31] Hausdorf, B., Hennig, C.: Biotic element analysis in biogeography. Systematic Biology **52**(5), 717–723 (2003)

[32] Hennig, C., Hausdorf, B.: Distance-based parametric bootstrap tests for clustering of species ranges. Computational Statistics & Data Analysis **45**(4), 875–895 (2004)

[33] Hennig, C., Hausdorf, B.: A robust distance coefficient between distribution areas incorporating geographic distances. Systematic biology **55**(1), 170–175 (2006)

[34] Bray, J.R., Curtis, J.T.: An ordination of the upland forest communities of southern wisconsin. Ecological monographs **27**(4), 326–349 (1957)

[35] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)

[36] Greenacre, M., Primicerio, R.: Chapter: Measures of distance between samples: Non-euclidean. Universitat Pompeu Fabra Barcelona Department of Economics and Business [Online]. http://www. econ. upf. edu/-michael/stanford/maeb5. pdf (2008)

[37] Greenacre, M.: 'size'and 'shape'in the measurement of multivariate proximity. Methods in Ecology and Evolution **8**(11), 1415–1424 (2017)

[38] Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., Knight, R.: Unifrac: an effective distance metric for microbial community comparison. The ISME journal **5**(2), 169–172 (2011)

[39] Xia, Y., Sun, J.: Beta diversity metrics and ordination. In: Bioinformatic and Statistical Analysis of Microbiome Data: From Raw Sequences to Advanced Modeling with QIIME 2 and R, pp. 335–395. Springer, ??? (2023)

[40] Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D., Li, H.: Associating microbiome composition with environmental covariates using generalized unifrac distances. Bioinformatics **28**(16), 2106–2113 (2012)

[41] Lozupone, C.A., Hamady, M., Kelley, S.T., Knight, R.: Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology **73**(5), 1576–1585 (2007)

[42] Wong, R.G., Wu, J.R., Gloor, G.B.: Expanding the unifrac toolbox. PloS one **11**(9), 0161196 (2016)

[43] Minkowski, H.: Geometrie der Zahlen vol. 1. BG Teubner, Leipzig (1910)

[44] Şuhubi, E.S.: Metric Spaces, pp. 261–356. Springer, Dordrecht (2003). https://doi.org/10.1007/978-94-017-0141-9_5

[45] Heath, T.L.: The Thirteen Books of Euclid's Elements. Dover Publications, Inc, New York (1956)

[46] Elmore, K.L., Richman, M.B.: Euclidean distance as a similarity metric for principal component analysis. Monthly weather review **129**(3), 540–549 (2001)

[47] Korenius, T., Laurikkala, J., Juhola, M.: On principal component analysis, cosine and euclidean measures in information retrieval. Information Sciences **177**(22), 4893–4905 (2007)

[48] Singh, A., Yadav, A., Rana, A.: K-means with three different distance metrics. International Journal of Computer Applications **67**(10) (2013)

[49] Kapil, S., Chawla, M.: Performance evaluation of k-means clustering algorithm with various distance metrics. In: 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), pp. 1–4 (2016). IEEE

[50] Xu, J., Chi, E., Lange, K.: Generalized linear model regression under distance-to-set penalties. Advances in neural information processing systems **30** (2017)

[51] Krislock, N., Wolkowicz, H.: Euclidean Distance Matrices and Applications. Springer, New York (2012)

[52] Bishop, C.M.: Pattern recognition and machine learning. Springer google schola **2**, 1122–1128 (2006)

[53] Deza, E., Deza, M.M., Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, New York (2009)

[54] Jiang, X., Hu, X., He, T.: Identification of the clustering structure in microbiome data by density clustering on the manhattan distance. Science China Information Sciences **59**, 1–7 (2016)

[55] Pearson, K.: X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **50**(302), 157–175 (1900)

[56] Dodge, Y.: The Concise Encyclopedia of Statistics. Springer, New York (2008)

[57] Liu, X., Wang, D.: A spectral histogram model for texton modeling and texture discrimination. Vision Research **42**(23), 2617–2634 (2002)

[58] Dixon, P.: Vegan, a package of r functions for community ecology. Journal of vegetation science **14**(6), 927–930 (2003)

[59] Hellinger, E.: Neue begründung der theorie quadratischer formen von

unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik **1909**(136), 210–271 (1909)

[60] Zhu, C., Xiao, F.: A belief hellinger distance for d–s evidence theory and its application in pattern recognition. Engineering Applications of Artificial Intelligence **106**, 104452 (2021)

[61] Vázquez-Castellanos, J.F.: Diversity analysis in viral metagenomes. The Human Virome: Methods and Protocols, 203–230 (2018)

[62] Markatou, M., Karlis, D., Ding, Y.: Distance-based statistical inference. Annual Review of Statistics and Its Application **8**(1), 301–327 (2021)

[63] Karagrigoriou, A.: Statistical inference: The minimum distance approach. Taylor & Francis (2012)

[64] Cieslak, D.A., Hoens, T.R., Chawla, N.V., Kegelmeyer, W.P.: Hellinger distance decision trees are robust and skew-insensitive. Data Mining and Knowledge Discovery **24**, 136–158 (2012)

[65] Torgersen, E.: Comparison of Statistical Experiments vol. 36. Cambridge University Press, Cambridge, UK (1991)

[66] Mahalanobis, P.C.: On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-) **80**, 1–7 (2018)

[67] Manly, B.F., Manly, B.: Multivariate statistical methods (1994)

[68] Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M.H.H., Oksanen, M.J., Suggests, M.: The vegan package. Community ecology package **10**(631-637), 719 (2007)

[69] McMurdie, P.J., Holmes, S.: phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. PloS one **8**(4), 61217 (2013)

[70] Pavoine, S., Dufour, A.-B., Chessel, D.: From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. Journal of theoretical biology **228**(4), 523–537 (2004)

[71] Ramette, A.: Multivariate analyses in microbial ecology. FEMS microbiology ecology **62**(2), 142–160 (2007)

[72] Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) **44**(2), 139–160 (1982)

[73] Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: Modeling and Analysis of Compositional Data. John Wiley & Sons, Chichester, UK (2015)

[74] Zhang, Y., Schluter, J., Zhang, L., Cao, X., Jenq, R.R., Feng, H., Haines, J., Zhang, L.: Review and revamp of compositional data transformation: A new framework combining proportion conversion and contrast transformation. Computational and Structural Biotechnology Journal **23**, 4088–4107 (2024)

[75] Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R., Zengler, K.: A novel sparse compositional technique reveals microbial perturbations. MSystems **4**(1), 10–1128 (2019)

[76] Gower, J.C., Legendre, P.: Metric and euclidean properties of dissimilarity coefficients. Journal of classification **3**, 5–48 (1986)

[77] Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer, New York (2007)

[78] Whittaker, R.H.: Evolution and measurement of species diversity. Taxon **21**(2-3), 213–251 (1972)

[79] Odum, E.P.: The strategy of ecosystem development: An understanding of ecological succession provides a basis for resolving man's conflict with nature. science **164**(3877), 262–270 (1969)

[80] Liberti, L., Lavor, C.: Six mathematical gems from the history of distance geometry. International Transactions in Operational Research **23**(3), 897–920 (2016)

[81] Legendre, P., De Cáceres, M.: Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecology Letters **16**(8), 951–963 (2013)

[82] Rao, C.R.: Diversity and dissimilarity coefficients: a unified approach. Theoretical population biology **21**(1), 24–43 (1982)

[83] Anderson, M.J.: Distance-based tests for homogeneity of multivariate dispersions. Biometrics **62**(1), 245–253 (2006)

[84] Menger, K.: Untersuchungen über allgemeine metrik. Mathematische Annalen **100**, 75–163 (1928)

[85] Blumenthal, L.M.: A modern view of distance geometry. Bulletin of the American Mathematical Society **50**(1), 6–16 (1944)

[86] Blumenthal, L.M.: Theory and Applications of Distance Geometry. Oxford University Press, Oxford, UK (1953)

[87] Gower, J.C.: Properties of euclidean and non-euclidean distance matrices. Linear algebra and its applications **67**, 81–97 (1985)

[88] Schloss, P.D.: Evaluating different approaches that test whether microbial

communities have the same structure. The ISME journal **2**(3), 265–275 (2008)

[89] Conway, J.B.: A Course in Functional Analysis vol. 96. Springer, New York (2019)

[90] Seal, H.: Multivariate statistical analysis for biologists. (1964)

[91] DasGupta, A.: Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics. Springer, New York (2011)

[92] Mardia, K.V., Kent, J.T., Taylor, C.C.: Multivariate Analysis. John Wiley & Sons, Hoboken, NJ, USA (2024)

[93] Ash, R.B.: Information Theory. Courier Corporation, Boston, MA (2012)

[94] Richardson, T., Urbanke, R.: Modern Coding Theory. Cambridge university press, Cambridge, UK (2008)

[95] Haroutunian, M., Mkhitaryan, K., Mothe, J.: f-divergence measures for evaluation in community detection. In: Proceedings of the Workshop on Collaborative Technologies and Data Science in Smart City Applications, pp. 1–6 (2018). https://doi.org/10.1145/3289256.3289260

[96] Nishiyama, T.: Generalized bregman and jensen divergences which include some f-divergences. arXiv preprint arXiv:1808.06148 (2018)

[97] Wainwright, M.J.: High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press, Cambridge, UK (2019). https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E

[98] Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991)

[99] Beran, R.: Minimum hellinger distance estimates for parametric models. The Annals of Statistics **5**(3), 445–463 (1977) https://doi.org/10.1214/aos/1176343842

[100] Greenberg, M.J.: Euclidean and non-Euclidean Geometries: Development and History. Macmillan, New York, NY, USA (1993)

[101] De Santis, E., Martino, A., Rizzi, A.: On component-wise dissimilarity measures and metric properties in pattern recognition. PeerJ Computer Science **8**, 1106 (2022)

[102] Paradis, E., Schliep, K.: ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. Bioinformatics **35**(3), 526–528 (2019)

[103] Cailliez, F.: The analytical solution of the additive constant problem. Psychometrika **48**(2), 305–308 (1983)

[104] Lingoes, J.C.: Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika **36**(2), 195–203 (1971)

[105] Higham, N.J.: Computing a nearest symmetric positive semidefinite matrix. Linear Algebra and its Applications **103**, 103–118 (1988) https://doi.org/10.1016/0024-3795(88)90223-6

[106] Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Mathematics Doklady **4**, 1035–1038 (1963)

[107] Deng, C., Gao, J., Lu, K., Luo, F., Sun, H., Xin, C.: Neuc-mds: Non-euclidean multidimensional scaling through bilinear forms. Advances in Neural Information Processing Systems **37**, 121539–121569 (2024)

[108] Gisbrecht, A., Schleif, F.-M.: Metric and non-metric proximity transformations at linear costs. Neurocomputing **167**, 643–657 (2015)

[109] Anderson, M.J.: A new method for non-parametric multivariate analysis of variance. Austral ecology **26**(1), 32–46 (2001)

[110] Wilson, N., Zhao, N., Zhan, X., Koh, H., Fu, W., Chen, J., Li, H., Wu, M.C., Plantinga, A.M.: Mirkat: kernel machine regression-based global association tests for the microbiome. Bioinformatics **37**(11), 1595–1597 (2021)

[111] Jiang, Z., He, M., Chen, J., Zhao, N., Zhan, X.: Mirkat-mc: a distance-based microbiome kernel association test with multi-categorical outcomes. Frontiers in Genetics **13**, 841764 (2022)

[112] Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. John Wiley & Sons, Hoboken, NJ (2010)

[113] Edgington, E., Onghena, P.: Randomization Tests. Chapman and Hall/CRC, Boca Raton, FL, USA (2007)

[114] Fromentin, S., Forslund, S.K., Chechi, K., Aron-Wisnewsky, J., Chakaroun, R., Nielsen, T., Tremaroli, V., Ji, B., Prifti, E., Myridakis, A., *et al.*: Microbiome and metabolome features of the cardiometabolic disease spectrum. Nature medicine **28**(2), 303–314 (2022)

[115] Yin, Y.S., Minacapelli, C.D., Parmar, V., Catalano, C.C., Bhurwal, A., Gupta, K., Rustgi, V.K., Blaser, M.J.: Alterations of the fecal microbiota in relation to acute covid-19 infection and recovery. Molecular biomedicine **3**(1), 36 (2022)

[116] Gonzalez, C.G., Mills, R.H., Zhu, Q., Sauceda, C., Knight, R., Dulai, P.S., Gonzalez, D.J.: Location-specific signatures of crohn's disease at a multi-omics scale.

Microbiome **10**(1), 133 (2022)

[117] Mills, R.H., Dulai, P.S., Vázquez-Baeza, Y., Sauceda, C., Daniel, N., Gerner, R.R., Batachari, L.E., Malfavon, M., Zhu, Q., Weldon, K., *et al.*: Multi-omics analyses of the ulcerative colitis gut microbiome link bacteroides vulgatus proteases with disease severity. Nature microbiology **7**(2), 262–276 (2022)

[118] Curtasu, M.V., Tafintseva, V., Bendiks, Z.A., Marco, M.L., Kohler, A., Xu, Y., Nørskov, N.P., Nygaard Lærke, H., Bach Knudsen, K.E., Hedemann, M.S.: Obesity-related metabolome and gut microbiota profiles of juvenile göttingen minipigs—long-term intake of fructose and resistant starch. Metabolites **10**(11), 456 (2020)

[119] Demmer, R., Jacobs Jr, D., Singh, R., Zuk, A., Rosenbaum, M., Papapanou, P., Desvarieux, M.: Periodontal bacteria and prediabetes prevalence in origins: the oral infections, glucose intolerance, and insulin resistance study. Journal of dental research **94**(9_suppl), 201–211 (2015)

[120] Demmer, R.T., Breskin, A., Rosenbaum, M., Zuk, A., LeDuc, C., Leibel, R., Paster, B., Desvarieux, M., Jacobs Jr, D.R., Papapanou, P.N.: The subgingival microbiome, systemic inflammation and insulin resistance: the oral infections, glucose intolerance and insulin resistance study. Journal of clinical periodontology **44**(3), 255–265 (2017)

[121] Armstrong, A.J., Quinn, K., Li, S.X., Schneider, J.M., Nusbacher, N.M., Doenges, K.A., Fiorillo, S., Marden, T.J., Higgins, J., Reisdorph, N., et al.: Systems analysis of gut microbiome influence on metabolic disease in hiv and high-risk populations. bioRxiv, 2021–03 (2021)

[122] Legendre, P., Legendre, L., et al.: Numerical ecology: developments in environmental modelling. Developments in Environmental Modelling **20**(1) (1998)

[123] Gromov, M., Katz, M., Pansu, P., Semmes, S.: Metric Structures for Riemannian and non-Riemannian Spaces vol. 152. Springer, New York (1999)

[124] Jaccard, P.: Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. **44**, 223–270 (1908)

[125] Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**(3-4), 325–338 (1966)

[126] Pawlowsky-Glahn, V., Egozcue, J.J.: Compositional data and their analysis: an introduction. Geological Society, London, Special Publications **264**(1), 1–10 (2006)

[127] Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M.: Understanding sequencing data as compositions: an outlook and review. Bioinformatics **34**(16),

2870–2878 (2018)

[128] Barwell, L.J., Isaac, N.J., Kunin, W.E.: Measuring β-diversity with species abundance data. Journal of Animal Ecology **84**(4), 1112–1122 (2015)

[129] Gupta, A.: Embedding tree metrics into low dimensional euclidean spaces. In: Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing, pp. 694–700 (1999)

[130] Chen, J., Safro, I.: Algebraic distance on graphs. SIAM Journal on Scientific Computing **33**(6), 3468–3490 (2011)

[131] Anderson, M.J.: Permutational multivariate analysis of variance (permanova). Wiley statsref: statistics reference online, 1–15 (2014)

[132] Legendre, P., Gallagher, E.D.: Ecologically meaningful transformations for ordination of species data. Oecologia **129**(2), 271–280 (2001)

Acknowledgements

Funding

This work was supported by Dr. Liangliang Zhang's junior faculty start-up grant BGT630267, funded by School of Medicine at Case Western Reserve University.

Author information

Authors and Affiliations

Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, 44106, OH, USA
Zihan Zhu, Yiqian Zhang, Liangliang Zhang

Division of Biostatistics, College of Public Health, Ohio State University, 1841 Neil Ave, Columbus, 43210, OH, USA
Yiqian Zhang

Department of Statistics, College of Arts and Sciences, The Ohio State University, 1958 Neil Ave., Columbus, 43210, OH, USA
Yiqian Zhang

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 7007 Bertner Ave, Houston, 77030, TX, USA
Wenhao Li, Satabdi Saha

Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain
Michael Greenacre

Weill Cornell Medicine, Cornell University, 1300 York Ave, New York, 10065, NY, USA
Yushu Shi

Case Comprehensive Cancer Center, 2103 Cornell Road, Cleveland, 44106, OH, USA
Liangliang Zhang

Corresponding author

Correspondence to Liangliang Zhang.

Ethics declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Supplementary Information