machine learning & knowledge extraction

MDPI

Article

Machine Learning Monte Carlo Approaches and Statistical Physics Notions to Characterize Bacterial Species in Human Microbiota

Michele Bellingeri [1,2,*], Leonardo Mancabelli [3,4], Christian Milani [4,5], Gabriele Andrea Lugli [4,5], Roberto Alfieri [1,2], Massimiliano Turchetto [1,2], Marco Ventura [4,5] and Davide Cassi [1,2]

1 Dipartimento di Scienze Matematiche, Fisiche e Informatiche, University of Parma, Via G.P. Usberti, 7/a, 43124 Parma, Italy; roberto.alfieri@unipr.it (R.A.); massimiliano.turchetto@unipr.it (M.T.); davide.cassi@unipr.it (D.C.)
2 Istituto Nazione di Fisica Nucleare, INFN, Gruppo Collegato di Parma, 43124 Parma, Italy
3 Department of Medicine and Surgery, University of Parma, 43124 Parma, Italy; leonardo.mancabelli@unipr.it
4 Interdepartmental Research Centre "Microbiome Research Hub", University of Parma, 43124 Parma, Italy; christian.milani@unipr.it (C.M.); gabrieleandrea.lugli@unipr.it (G.A.L.); marco.ventura@unipr.it (M.V.)
5 Laboratory of Probiogenomics, Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, 43124 Parma, Italy
* Correspondence: michele.bellingeri@unipr.it

Abstract: Recent studies have shown correlations between the microbiota's composition and various health conditions. Machine learning (ML) techniques are essential for analyzing complex biological data, particularly in microbiome research. ML methods help analyze large datasets to uncover microbiota patterns and understand how these patterns affect human health. This study introduces a novel approach combining statistical physics with the Monte Carlo (MC) methods to characterize bacterial species in the human microbiota. We assess the significance of bacterial species in different age groups by using notions of statistical distances to evaluate species prevalence and abundance across age groups and employing MC simulations based on statistical mechanics principles. Our findings show that the microbiota composition experiences a significant transition from early childhood to adulthood. Species such as *Bifidobacterium breve* and *Veillonella parvula* decrease with age, while others like *Agathobaculum butyriciproducens* and *Eubacterium rectale* increase. Additionally, low-prevalence species may hold significant importance in characterizing age groups. Finally, we propose an overall species ranking by integrating the methods proposed here in a multicriteria classification strategy. Our research provides a comprehensive tool for microbiota analysis using statistical notions, ML techniques, and MC simulations.

Keywords: Monte Carlo simulation; machine learning; human microbiota; statistical physics; microcanonical ensemble; canonical ensemble; database learning

Citation: Bellingeri, M.; Mancabelli, L.; Milani, C.; Lugli, G.A.; Alfieri, R.; Turchetto, M.; Ventura, M.; Cassi, D. Machine Learning Monte Carlo Approaches and Statistical Physics Notions to Characterize Bacterial Species in Human Microbiota. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2375–2399. https://doi.org/10.3390/make6040117

Academic Editor: Dominik Heider

Received: 1 August 2024
Revised: 14 October 2024
Accepted: 15 October 2024
Published: 18 October 2024

1. Introduction

Machine learning (ML) techniques are fundamental in analyzing extensive, complex biological data from different areas of biological science [1]. ML is a subset of artificial intelligence that allows computers to learn from data without being explicitly programmed. These advanced computational methods are precious in microbiome research, enabling the integration and interpretation of vast datasets to uncover intricate patterns and relationships within the microbiota. The microbiota, the community of microorganisms that colonizes the human body, is considered to affect a wide range of physiological processes, from immunity to digestion, and plays a crucial role in determining human health [2–4]. An ever-increasing number of studies have highlighted the possible correlation between the abundance, variability, and richness of bacterial species belonging to the human microbiota and many health disorders and/or diseases [3,5–7]. In this context, investigating in depth the human gut microbiota composition and its correlation with the different host intrinsic

parameters, such as age, gender, health condition, and lifestyle, could be crucial for better understanding the main factors that might impact individuals' health status [6,7].

The study of the human microbiota was made possible by developing modern next-generation sequencing techniques, which allow the precise and in-depth identification of the microbial populations that inhabit the human body [8,9]. However, these modern techniques, such as 16S rRNA gene profiling and shotgun metagenomic sequencing, require sophisticated and advanced technological approaches capable of analyzing large and complex databases and identifying subtle patterns within them. Despite these advancements, there currently needs to be a unified statistical methodology for analyzing metagenomic data, leading to significant variability in the approaches used. In this context, the development of MLs has provided powerful tools to address this challenge, allowing us to extract useful information from large microbiota databases and develop models to predict the abundance of bacterial species in response to various factors.

In this study, we evaluate a new approach by integrating statistical physics notions with the Monte Carlo (MC) method to characterize bacterial species in the human microbiota. Specifically, in addition to classic statistical classification strategies such as species prevalence and abundance, we introduce and utilize different statistical distance notions to assess how much the average occurrence of a species in the age group samples deviates from the general average across groups. Then, we use MCs, which are computational algorithms based on repeated random sampling, to obtain numerical results [10], widely used in various fields of biology [11,12], ecology [13], and physics [14]. MCs are random numerical experiments on a computer where we can observe the outcomes of these experiments, and they are instrumental when dealing with complex systems with high uncertainty or randomness, such as bacterial species sampling. While MC methods are not strictly ML, they are frequently employed as tools within ML algorithms. MC can support ML for parameter estimation in large datasets, generate synthetic data to augment a dataset, or make a random distribution of the empirical data [15]. Our research adopts the MC to generate synthetic data to analyze the prevalence of bacterial species in the human microbiota. We perform MCs whose rationale is based on the concepts of the microcanonical and canonical ensembles in statistical mechanics [16]. MCs furnish random scenarios where the occurrences of species are varied while keeping key parameters fixed. MCs make synthetic data that allow us to find the statistical significance of bacterial species for age groups by comparing their empirical prevalence with that of the numerical experiment. Last, since we evaluate the importance of bacterial species in the age groups with different classification strategies, we propose a further analysis to obtain an overall species ranking by evaluating the results of the different strategies together.

In summary, this manuscript introduces an innovative methodology integrating ML, MC simulations, and statistical physics to characterize bacterial species in the human microbiota. This approach provides a novel statistical approach to analyzing complex microbiota datasets, allowing for more comprehensive insights into microbial populations. By incorporating statistical physics principles, such as microcanonical and canonical ensembles, we enhance the understanding of bacterial species' significance across different age groups, offering a deeper dimension of analysis beyond traditional methods. Additionally, we develop a multicriteria classification strategy to rank bacterial species based on their importance within age groups, considering both low- and high-prevalence species. The methods proposed in this study provide valuable insights into the relationship between microbiota composition and human health, offering a robust framework for advancing microbiome research.

The article is organized as follows: Section 2 describes the state of the art of the microbiota bacterial species research; Section 3 defines the methodology used in this manuscript; Section 4 shows and discusses the results; and finally, Section 5 summarizes and carries out the research conclusions.

2. State of the Art

The human microbiota has gained increasing attention in recent years due to its profound impact on health and disease [6,7]. Many studies have explored the link between microbiota composition and various health conditions, from gastrointestinal disorders to metabolic and immune-related diseases [3,17,18]. These studies leverage next-generation sequencing (NGS) technologies such as 16S rRNA gene sequencing and shotgun metagenomics to analyze the taxonomic composition and functional capacity of microbial communities [8,9].

However, the challenge of analyzing large and complex microbiome datasets has led researchers to adopt advanced computational tools. Traditional statistical methods, though useful, often struggle with the high dimensionality and sparsity inherent in microbiome data. As a result, ML methods have become indispensable for uncovering patterns within these data. ML approaches, including supervised and unsupervised learning algorithms, have been applied to identify microbial biomarkers, classify microbial communities, and predict disease states based on microbiome profiles [6,7].

Despite their success, these approaches often focus on species abundance and prevalence while overlooking the potential role of low-prevalence species in shaping microbial community structure. Moreover, there is limited consensus on the best statistical methodologies for integrating complex datasets, particularly when considering multiple factors such as age, health conditions, and geographic location [19].

Recent advancements in MC methods have provided new avenues for addressing these challenges. MC methods are particularly well-suited for simulating random processes in complex systems and have been applied in fields ranging from ecology to statistical mechanics [20]. While MC methods are not traditionally considered ML, they can be integrated within ML frameworks to support parameter estimation and synthetic data generation. Combining ML techniques with MC simulations, this hybrid approach could overcome some of the limitations of traditional statistical methods in microbiome research by generating robust predictions based on synthetic datasets and exploring the importance of low-prevalence species.

This study builds upon the existing literature by introducing a novel application of MC simulations in microbiome research. It provides new insights into the role of statistical distances in classifying bacterial species and offers a comprehensive approach for integrating ML and MC methods.

3. Methods

The Methods section is organized as follows: Section 3.1 introduces the microbiota bacterial species database and provides the related references. Section 3.2 defines and explains the statistical indicators we propose to characterize bacterial species. Section 3.3 introduces ML techniques and then defines MC simulations. Finally, we explain how we utilize microcanonical and canonical ensembles from statistical physics to conduct MC simulations for a statistical analysis to identify and characterize bacterial species.

3.1. Database

The data used in this study were obtained from publicly available datasets regarding the human gut microbiota across different life stages. These comprehensive datasets included samples from various regions worldwide, providing a diverse and representative overview of the global human gut microbiota [21]. In detail, this study included a total of 5896 sequenced fecal samples collected from 71 public bioprojects across 34 different countries. The datasets included sequenced fecal samples collected from healthy individuals ranging from birth to over 100 years old, with a robust statistical representation of all the different age groups. The collected samples, as reported in the previous manuscript [21], were used to assess the microbiota composition at the species level through the METAnnotatorX2 software (http://probiogenomics.unipr.it/cmu/ accessed on 29 June 2021) following the standard filtering parameters reported in the manual with *Homo sapiens* reads removal [22].

Moreover, the samples included in the analysis were categorized into four age groups, that is, G1 (0–4 years), G2 (5–17 years), G3 (18–64 years), and G4 (65 years and older), following the guidelines provided by the World Health Organization (WHO) [23].

3.2. Statistical Analyses

3.2.1. Average Occurrences

The columns of the matrix database W represent bacterial species; the rows of the database W are the human fecal samples. The element w_{ij} of W indicates the relative abundance of species j in sample i. Figure 1A depicts a simple example of the database W used here.

A

Groups	*W*	S1	S2	S3	S4	S5	S6
G1	C1	0	0.5	0	0	0	0.5
G1	C2	0.1	0	0.3	0	0	0.6
G1	C3	0	0	0.45	0	0	0.55
G2	C4	0	0	0.4	0	0.6	0
G2	C5	0	0	0.2	0	0.8	0
G3	C6	0	0	0	0.2	0.8	0
G3	C7	0	0	0.2	0.1	0.7	0
G3	C8	0	0	0.05	0.55	0.4	0
G4	C9	0	0.25	0	0.75	0	0
G4	C10	0	1	0	0	0	0

B

Groups	*A*	S1	S2	S3	S4	S5	S6
G1	C1	0	1	0	0	0	1
G1	C2	1	0	1	0	0	1
G1	C3	0	0	1	0	0	1
G2	C4	0	0	1	0	1	0
G2	C5	0	0	1	0	1	0
G3	C6	0	0	0	1	1	0
G3	C7	0	0	1	1	1	0
G3	C8	0	0	1	1	1	0
G4	C9	0	1	0	1	0	0
G4	C10	0	1	0	0	0	0

Figure 1. Bacterial species database example. Rows are fecal samples (C1, C2,. . .,C3). Columns are bacterial species (S1, S2,. . .,S6). Rows/samples are split into four groups by age (G1, G2, G3, G4). (**A**) Matrix database W, in which each cell indicates the relative abundance of bacterial species (column) in the sample (row). (**B**) Matrix database A, in which each cell indicates the presence/absence of a bacterial species (column) in the sample (row).

The average occurrence of the species j among all samples is:

$$\overline{m_j} = \frac{1}{N} \sum_{i=1}^{N} w_{ij} \tag{1}$$

where N is the total number of samples; we can call $\overline{m_j}$ the average weighted occurrence. $\overline{m_j}$ estimates how much of a bacterial species is present among all samples. The average weighted occurrence $\overline{m_j}$ is commonly called the 'relative average abundance' of bacterial species [21,24].

The samples are divided by age into four groups. We compute $\overline{m_j^y}$ as the average occurrence of species j within group y and call it the average weighted occurrence (AWO).

In formula:

$$\overline{m_j^y} = \frac{1}{N_y} \sum_{i \in y} w_{ij} \tag{2}$$

where N_y is the total number of samples of the group y. $\overline{m_j^y}$ estimates the amount of a bacterial species present among the group samples. This furnishes a first and simple estimate of the bacterial species' presence within a given age range of the subjects examined.

Second, we convert the species abundances into simple occurrences (presence/absence). We call this new database occurrences matrix A. For this, A is a binary matrix in which the elements of the species/columns are 0 (no occurrence) and 1 (occurrence) (Figure 1B). The element a_{ij} of the matrix A is 1 if species j occurs in sample i and 0 otherwise.

The average species occurrence among all samples in matrix A is:

$$\overline{u_j} = \frac{1}{N} \sum_{i=1}^{N} a_{ij} \tag{3}$$

We can call $\overline{u_j}$ the average binary occurrence (ABO), whereas it is usually called the 'prevalence' of the bacterial species [21,24].

Then, we compute $\overline{\mu_j^y}$, the average binary occurrence of species j within group y. For each column j, we divide the total number of 1s occurring within group y by the total number of samples. $\overline{\mu_j^y}$ represents the average occurrence among the sample of species j within group y.

In formula:

$$\overline{\mu_j^y} = \frac{1}{N_y} \sum_{i \in y} a_{ij} \tag{4}$$

where N_y is the total number of samples of group y, since this average is computed considering the simple presence–absence of species occurrence in the sample. $\overline{u_j}$ and $\overline{\mu_j^y}$ estimate how frequent it is to find a particular bacterial species among samples without considering the relative abundance of the species in the sample.

3.2.2. Relative Distances

We compute the distance between the observed (Obs) and the expected (Exp) species occurrence within each group. We define this distance as the relative deviation of the average species occurrence within the group from the average species occurrence among groups (among all samples). This is represented by the following ratio:

$$d = \frac{Obs - Exp}{Exp} \tag{5}$$

where Obs indicates the average species occurrence within a group, and Exp is the average species occurrence among all samples (among groups).

The average weighted occurrence among all samples $\overline{m}_j$ represents the expected occurrence. The average weighted occurrence of species j within group $\overline{m_j^y}$ represents the observed occurrence.

The weighted distance becomes:

$${}_w d_j^y = \frac{Obs - Exp}{Exp} = \frac{\overline{m_j^y} - \overline{m}_j}{\overline{m}_j} \tag{6}$$

The greater the distance d_j^y, the greater the difference between the observed and the expected species occurrence within group y. From now on, ${}_w d_j^y$ is the relative weighted distance (RWD). We can see this distance as the relative deviation of the relative aver-

age abundance within the group (observed occurrence) concerning the relative average abundance across all samples (expected occurrence).

We can translate Equation (6) using the binary occurrence. The average binary occurrence among samples $\overline{\mu}_j$ represents the expected occurrence. The average binary occurrence of species j within group $\overline{\mu_j^y}$ represents the observed occurrence. We call this the relative binary distance (RBD), and it is computed as follows:

$$d_j^y = \frac{Obs - Exp}{Exp} = \frac{\overline{\mu_j^y} - \overline{\mu}_j}{\overline{\mu}_j} \tag{7}$$

This statistical distance is the relative deviation of the species prevalence within the group (observed occurrence) concerning the prevalence across all groups (expected occurrence).

Let us consider how the RBD and RWD distances evaluate specific cases of species occurrence.

We start with the case of a bacterial species occurring only within a group.

The average binary occurrence of species j is:

$$\overline{u_j} = \frac{1}{N}\sum_{i=1}^{N} a_{ij} = \frac{\sum_{i=1}^{N} a_{ij}}{N} \tag{8}$$

And the average binary occurrence of species j within group y is:

$$\overline{\mu_j^y} = \frac{1}{N_y}\sum_{i\in y} a_{ij} = \frac{\sum_{i\in y} a_{ij}}{N_y} \tag{9}$$

Substituting Equations (8) and (9) in Equation (7) of the relative binary distance, we obtain:

$$d_j^y = \frac{\frac{\sum_{i\in y} a_{ij}}{N_y} - \frac{\sum_{i=1}^{N} a_{ij}}{N}}{\frac{\sum_{i=1}^{N} a_{ij}}{N}} \tag{10}$$

Reversing the denominator:

$$= \left(\frac{\sum_{i\in y} a_{ij}}{N_y} - \frac{\sum_{i=1}^{N} a_{ij}}{N}\right)\cdot\frac{N}{\sum_{i=1}^{N} a_{ij}}$$

$$= \frac{\sum_{i\in y} a_{ij}}{N_y}\cdot\frac{N}{\sum_{i=1}^{N} a_{ij}} - \frac{\sum_{i=1}^{N} a_{ij}}{N}\cdot\frac{N}{\sum_{i=1}^{N} a_{ij}}$$

$$= \frac{\sum_{i\in y} a_{ij}}{N_y}\cdot\frac{N}{\sum_{i=1}^{N} a_{ij}} - 1 \tag{11}$$

In the case species j occurs only in group y, we have the equivalence between the prevalence within the group and across all groups:

$$\sum_{i\in y} a_{ij} = \sum_{i=1}^{N} a_{ij}$$

And:

$$\sum_{i\in y} a_{ij}$$

For this, Equation (11) results in:

$$= \frac{\sum_{i\in y} a_{ij}}{N_y}\cdot\frac{N}{\sum_{i=1}^{N} a_{ij}} - 1 = \frac{N}{N_y} - 1 \tag{12}$$

Equation (12) indicates that if a bacterial species occurs only within group y, the value of the RBD is only determined by the $\frac{N}{N_y}$ ratio, that is, the ratio between the total number of samples and the number of samples in group y. The last outcome implies that species with different prevalences that occur only within a group will return the same RBD value. In this case, species with different occurrences have the same distance, and the RBD is not able to discriminate their importance in characterizing the group. This implies that Equation (12) also provides the RBD maximum value. Figure 1 depicts an example of RBD computation for species that occur only within a group. From Figure 1B, we calculate the RBD values for species S1 and S6 for G1; we compute d_1^1 and d_6^1.

Using Equation (6) for species S1, we have:

$$d_1^1 = \frac{\overline{\mu_1^1} - \overline{\mu}_1}{\overline{\mu}_1} = \frac{\frac{1}{3} - \frac{1}{10}}{\frac{1}{10}} = 2.\overline{3}$$

And for species S6:

$$d_6^1 = \frac{\overline{\mu_6^1} - \overline{\mu}_6}{\overline{\mu}_6} = \frac{\frac{3}{3} - \frac{3}{10}}{\frac{3}{10}} = 2.\overline{3}$$

Even though the occurrences are different, and species S6 is more frequent than S1 in group 1, the RBD value is the same.

Let us consider the case of a bacterial species that does not occur within a group. In this case, $\overline{\mu_j^y} = 0$ and Equation (6) results in:

$$d_j^y = \frac{\overline{\mu_j^y} - \overline{\mu}_j}{\overline{\mu}_j} = \frac{0 - \overline{\mu}_j}{\overline{\mu}_j} = -1 \tag{13}$$

This result implies that all the species not occurring within a group will return the RBD value $d_j^y = -1$ and that $d_j^y = -1$ is also the minimum value. RBD is constrained in the interval [−1,1]. It is easy to show that the maximum and the minimum values derived in Equations (12) and (13) for RBD are also the limits of the weighted counterpart RWD.

We can introduce a second-order hierarchy to rank the ties to solve the RWD and RBD problem of presenting ties in species ranking when species occur only within a group. For example, in the case of ties, we can rank the species presenting the same RWD and RBD values using measures of their average occurrence within the group; for example, we can use the relative abundance $\overline{m_j^y}$ and the prevalence $\overline{\mu_j^y}$ of the species in the group as a second-order criterion to rank ties. Therefore, species are first ranked according to their relative distance, thus assessing their statistical distance from the overall average occurrence, and then according to their average occurrence within the group. We can choose the ranking strategy to solve ties with the rationale we prefer for the problem. For example, suppose we want to prioritize the number of times a species appears, i.e., its prevalence. In this case, we can use the average binary occurrence $\overline{\mu_j^y}$ as a second-order criterion to rank ties. If we want to prioritize the abundance of a species in the sample, i.e., the relative species abundance, we can adopt its average weighted occurrence $\overline{m_j^y}$ as a second-order criterion to rank ties. The selection of the second-order ranking criterion should be guided by the rationale that aligns most closely with the objectives of the analysis. In this research, we rank ties using the species binary occurrence. In Figure 2, we furnish an example of a second-order rank methodology for solving ties.

A

Groups	S1	S2	S3
G1	0.1	0.5	0
G1	0.1	0	0.3
G1	0.1	0	0.45
G2	0	0	0
G2	0	0	0
G3	0	0	0
G3	0	0	0
G3	0	0	0

B

Rank	RWD	RWD ($\overline{m_j^y}$)	RWD ($\overline{\mu_j^y}$)
1	**S1** (1.67)	**S3** (1.67;0.25)	**S3** (1.67;1)
2	**S2** (1.67)	**S2** (1.67;0.17)	**S1** (1.67;0.67)
3	**S3** (1.67)	**S1** (1.67;0.1)	**S2** (1.67;0.33)

Figure 2. Simple bacterial species database example showing ties in the relative distance ranking strategies. Rows are fecal samples (C1, C2,. . .,C3). Columns are bacterial species (S1, S2,. . .,S6). Rows/samples are split into four groups by age (G1, G2, G3, G4). (**A**) Species database in which all three species occur only within G1. (**B**) RWD strategy ranking and values. (2nd column) As we can see, the RWD value is the same for S1, S2, and S3. (3rd column) When using the weighted average occurrence $\overline{m_j^y}$ as a second-order criterion, the ranking becomes S3, S2, and S1. (4th column) When adopting the binary average occurrence $\overline{\mu_j^y}$ as a second-order criterion, the ranking becomes S3, S1, S2. We give the RWD values and the second-order ranking values within brackets.

3.2.3. Inside–Outside Distances

Then, we used a second type of distance by computing the difference between the species occurrence inside the group and the species occurrence outside the group, that is:

$$ {}_w\Delta_j^y = \overline{m_j^y} - \overline{m_j^{\sim y}} \tag{14}$$

where $\overline{m_j^y}$ is the average weighted occurrence of species j within group y and $\overline{m_j^{\sim y}}$ is the average weighted occurrence of species j outside group y. Since ${}_w\Delta_j^y$ is the difference between the inside and outside average species occurrence, we refer to ${}_w\Delta_j^y$ as the inside–outside weighted distance (IOWD). In other terms, IOWD is the difference between the relative species abundance within the group and the relative species abundance outside the group.

We can modify Equation (14) using the species binary occurrence and defining the inside–outside binary distance (IOBD):

$$\Delta_j^y = \overline{\mu_j^y} - \overline{\mu_j^{\sim y}} \tag{15}$$

here, $\overline{\mu_j^y}$ is the average binary occurrence of species j within group y, and $\overline{\mu_j^{\sim y}}$ is the average binary occurrence of species j outside group y. IOBD is the difference between the species prevalence within the group and the species prevalence outside the group.

We can compute the range limits for the outside–inside distances. Let us take Equation (14), giving the IOWD. The maximum value occurs, satisfying the following three conditions: (i) species j occurs only within a group y ($\overline{m_j^{\sim y}} = 0$, species j does not occur outside the group y), (ii) species j occurs in all the samples of group y, and (iii) species j abundances equal 1 (i.e., $w_{ij} = 1$, meaning that j is the only bacterial species in the sample i). These three conditions lead to $\overline{m_j^y} = 1$, and the maximum IOWD becomes ${}_w\Delta_j^y = \overline{m_j^y} = 1$.

At the opposite end, the minimum value ${}_w\Delta_j^y$ occurs when (i) species j does not occur in the group y ($\overline{m_j^y} = 0$), (ii) species j occurs in all the samples outside group y, and (iii) species j abundances outside group y equal 1. These three conditions lead to $\overline{m_j^{\sim y}} = 1$, and the minimum IOWD becomes ${}_w\Delta_j^y = -1$. It is easy to show that the minimum and the maximum values for IOBD in Equation (15) are the same as those derived above for the IOWD. The minimum and the maximum values for Equation (14) require $w_{ij} = 1 = a_{ij}$, thus demonstrating that Equations (14) and (15) return the same range limits [1,−1]. The distances computed using Equations (14) and (15) do not present the ties problem in ranking, as we find for RBD and RWD. The species producing IOWD and IOBD values corresponding to the closed interval [1,−1] range limits have the same occurrences among samples. This means that they present identical j columns in the bacterial species database, so having the same IOWD and IOBD is a proper way to evaluate their importance.

The inside–outside distances can adequately evaluate the case of ranking ties shown above for the relative distances RBD and RWD. As we did above for RBD and RWD, we computed IOBD for species S1 and S6 for G1 in Figure 1; we computed Δ_1^1 and Δ_6^1.

Using Equation (15) for species S1, we have:

$$\Delta_1^1 = \overline{\mu_1^1} - \overline{\mu_1^{\sim 1}} = \frac{1}{3} - \frac{0}{7} = 0.\overline{3}$$

And for species S6:

$$\Delta_6^1 = \overline{\mu_6^1} - \overline{\mu_6^{\sim 1}} = \frac{3}{3} - \frac{0}{7} = 1$$

The result $\Delta_6^1 > \Delta_1^1$ demonstrates that the inside–outside distance can discriminate species that occur only within a group but with different occurrences. This property may be necessary when a bacterial species database presents many species occurring only in one group.

3.3. Monte Carlo Numerical Simulations

3.3.1. Machine Learning

ML allows computers to learn from data without being explicitly programmed [25]. Computers learn from huge amounts of different data, from numbers to pictures, and create new algorithms on their own that can identify patterns in data, make predictions, and improve their performance over time [26]. ML presents various applications, such as image and speech recognition, predicting proteins and molecule interactions, self-driving cars, analyzing epidemiological data for identifying risk factors, and medical diagnosis [25]. MC methods are not strictly ML, but they are frequently employed as tools within ML algorithms. MC can support ML in generating synthetic data [15] and dropout training in deep neural networks, support adaptive algorithms [27], or make a random distribution of empirical data [28]. Our research adopts the MC method to generate synthetic data to analyze the prevalence of bacterial species in the human microbiota. Microcanonical and canonical MC simulations create random scenarios where the occurrences of species are varied while keeping key parameters fixed. These synthetic data allow for comparisons between empirical (observed) and expected occurrences, providing a basis for assessing the statistical significance of bacterial species across different age groups.

3.3.2. Microcanonical Simulation

We first perform a microcanonical Monte Carlo (MM) numerical simulation. The MM simulation keeps the total number of elements (occupied sites) fixed in every random assignment of occurrences. The word microcanonical arises from the microcanonical ensemble in statistical mechanics [16], and it was extended to non-thermodynamical problems, such as percolation theory [14]. In percolation theory, the microcanonical approach to percolation focuses on the behavior of individual sites within the lattice [14]. It is based

on the idea of considering all the microstates (i.e., the possible configurations) of a system with the same total number of occupied sites and assigning the same probability to each of them. In other words, the microcanonical ensemble assumes that the only information known about the system is the total number of occupied sites.

In detail, the microcanonical approach consists of randomizing the columns of matrix A (Figure 1B) by permuting each species binary occurrence column. The microcanonical randomization preserves the total number of 1s and 0s in the column, thus fixing the total number of binary occurrences. We iterate the process 10^6 times.

3.3.3. Canonical Simulation

Then, we perform a canonical Monte Carlo (CM) numerical simulation. We fix the probability of having a species occurrence in every random assignment of occurrences. The word canonical, too, arises from the canonical ensemble in statistical mechanics [16], and it was extended to percolation theory [14]. Consider a lattice with a finite number of sites where each site can be occupied or empty. The canonical approach assigns an occupation probability of p to each site; $1 - p$ is the probability of having an empty site. For these reasons, unlike the microcanonical approach, the canonical approach preserves only the probability of occupied sites. In the canonical approach, the total number of occupied sites can vary between simulations [14].

In our canonical MC simulation, we compute the average occurrence among samples for each species j. To do this, we divide the total number of 1s by the total number of samples. This computation returns the average binary occurrence (or prevalence) in Equation (3). $\overline{\mu}_j$ represents the probability p of finding species i among samples. The probability $1 - p$ represents the probability of not finding species i among samples. Using probability p, we can sort the occurrences from a binomial distribution. We assign 1 with probability p and 0 with probability $1 - p$ in each element a_{ij} of the randomized matrix. The canonical-like randomization preserves the average number of occurrences $\overline{\mu}_j$ (at least for a higher number of iterations). We iterate the process 10^6 times.

3.3.4. Monte Carlo Statistical Analyses

To evaluate the significance of the MC outcomes, we follow this scheme. First, we compute $\overline{\rho_j^y}$, which is the average occurrence among samples of finding the species j within group y of the randomized matrix. Last, to evaluate the probability of having the observed average occurrence by chance, we count how many times $\overline{\rho_j^y} > \overline{\mu_j^y}$ and divide this value by the total number of iterations (M). We obtain a p-value indicating the probability of having, by chance, a higher species occurrence within the group.

Therefore, we can compute p_j^y, which indicates the significance of observing species j in group y by chance, as follows:

$$p_j^y = \frac{1}{M} \sum_M \delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})} \tag{16}$$

where δ is the Kronecker delta function for which $\delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})} = \begin{cases} 1 \; if \; \overline{\rho_j^y} > \overline{\mu_j^y} \\ 0 \; if \; \overline{\rho_j^y} \leq \overline{\mu_j^y} \end{cases}$, and M is the total number of MC simulations (10^6). To perform a very large set of MC simulations, it is essential to obtain accurate predictions and produce reliable statistical results. Having a high number of simulations allows for accurately estimating the simulated occurrence of species with low empirical prevalence that, however, characterize a certain age group.

In the case of ties, which are species presenting the same p-value, we rank these ties according to the prevalence of the species. We performed the numerical MC simulations and the statistical analyses using the software R version 4.3.1, with packages *MASS* and *openxlsx*. The MC simulations were coded in parallel using the R programming language with *doParallel* and *foreach* modules and executions iterated 1 million times took approximately

60 h on 64 cores and 200 GB RAM. We performed the numerical simulations using the High Performance Computing (HPC) cluster of the University of Parma and the CINECA supercomputer Galileo100.

Table 1 lists the bacterial species classification strategies used in this research with formulas and meanings.

Table 1. List of the bacterial species ranking strategies.

Strategy	Acronym	Formula		Meaning
Average weighted occurrence	AWO	$\overline{m_j^y} = \frac{1}{N_y}\sum_{i\in y} w_{ij}$	w_{ij} of W indicates the relative abundance of species j in sample i; N_y is the total number of samples of group y.	Weighted abundance of a species in a group.
Average binary occurrence	ABO	$\overline{\mu_j^y} = \frac{1}{N_y}\sum_{i\in y} a_{ij}$	a_{ij} of A indicates the presence of species j in sample i; N_y is the total number of samples of group y.	Binary abundance of a species in a group, commonly called 'species prevalence'.
Relative weighted distance	RWD	${}_w d_j^y = \frac{\overline{m_j^y} - \overline{m}_j}{\overline{m}_j}$	$\overline{m_j^y}$ is the average weighted occurrence within group y; $\overline{m_j}$ is the average weighted occurrence among all samples.	Relative deviation of the average weighted abundance of a species in the group from the overall mean.
Relative binary distance	RBD	$d_j^y = \frac{\overline{\mu_j^y} - \overline{u}_j}{\overline{u}_j}$	$\overline{\mu_j^y}$ is the average binary occurrence within group y; $\overline{u_j}$ is the average binary occurrence among all samples.	Relative deviation of the average binary abundance of a species in the group from the overall mean.
Inside–outside weighted distance	IOWD	${}_w\Delta_j^y = \overline{m_j^y} - \overline{m_j^{\sim y}}$	$\overline{m_j^y}$ is the average weighted occurrence of species j within group y; $\overline{m_j^{\sim y}}$ is the average weighted occurrence of species j outside group y.	Difference between the average weighted abundance of a species within and outside a group.
Inside–outside binary distance	IOBD	$\Delta_j^y = \overline{\mu_j^y} - \overline{\mu_j^{\sim y}}$	$\overline{\mu_j^y}$ is the average occurrence of species j within group y; $\overline{\mu_j^{\sim y}}$ is the average occurrence of species j outside group y.	Difference between the average binary abundance of a species within and outside a group.
Microcanonical Monte Carlo	MM	$p_j^y = \frac{1}{M}\sum_M \delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})}$	$\overline{\rho_j^y}$ average within group y of the randomized matrix, $\overline{\mu_j^y}$ is the average binary occurrence within group y, δ is the Kronecker delta function for which $\delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})} = \begin{cases} 1 \; if \; \overline{\rho_j^y} > \overline{\mu_j^y} \\ 0 \; if \; \overline{\rho_j^y} \leq \overline{\mu_j^y} \end{cases}$, and M the total number of simulations.	Evaluates the probability to have a species within a group by permuting its binary occurrence.
Canonical Monte Carlo	CM	$p_j^y = \frac{1}{M}\sum_M \delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})}$	$\overline{\rho_j^y}$ average within group y of the randomized matrix, $\overline{\mu_j^y}$ is the average binary occurrence within group y, δ is the Kronecker delta function for which $\delta_{(\overline{\rho_j^y}, \overline{\mu_j^y})} = \begin{cases} 1 \; if \; \overline{\rho_j^y} > \overline{\mu_j^y} \\ 0 \; if \; \overline{\rho_j^y} \leq \overline{\mu_j^y} \end{cases}$, and M the total number of simulations.	Evaluates the probability to have a species within a group by sorting the binary occurrence at random.

4. Results and Discussion

4.1. Average Occurrence

In Figure 3, in the top row, we plot the frequency distribution of the binary occurrence of species (prevalence). The species prevalence is highly skewed with a long right tail, considering all samples together (Figure 3A, chart 'All') and the prevalence within the groups (Figure 3B). The highly skewed distribution with a long right tail indicates that the database presents few species occurring in most of the samples, and most of the bacterial species show minor occurrences, i.e., most of the species are rare. Table 2 shows the ten most common species (greater prevalence) with their relative binary occurrence values. As we can see, for groups G2, G3, and G4, most of the species with the highest prevalence occur in more than 80% of the sample. Only for G1 do the most common species show

a prevalence lower than 0.75. Figure 3A depicts the scatterplots of the average weighted occurrence ($\overline{m_j}$) vs. average binary occurrence ($\overline{u_j}$). We find a positive correlation between them for both the average among all samples (Figure 3A, chart 'All') and within the groups (Figure 3C, charts G1, G2, G3, G4). Computing the Pearson correlation coefficient r [29] to quantify the correlation between $\overline{m_j}$ and $\overline{u_j}$, we obtain the values $r = 0.667$ for all the samples and $r = \{0.674, 0.640, 0.627, 0.674\}$ for each group, respectively. The Pearson correlation outcomes indicate a positive correlation between the variables; when $\overline{m_j}$ increases, $\overline{u_j}$ also increases. Despite the good correlation, there are some less correlated points, showing how the prevalence of bacterial species is not always correlated with their relative species abundance (or weighted occurrence). This discrepancy highlights the complex nature of microbial ecosystems, where a highly prevalent taxon within a population does not necessarily dominate in abundance. Previous research has demonstrated that the gut microbiota undergoes significant taxonomic and functional shifts influenced by various factors, including age, diet, and health status [30,31].

Table 2. Ten most common species for each group with its prevalence (relative binary occurrence).

G1		G2		G3		G4	
Bifidobacterium longum	0.73	*Bacteroides unknown_species*	0.98	*Blautia unknown_species*	0.99	*Blautia unknown_species*	0.98
Escherichia coli	0.65	*Blautia unknown_species*	0.98	*Ruminococcus unknown_species*	0.98	*Ruminococcus unknown_species*	0.98
Blautia unknown_species	0.61	*Ruminococcus unknown_species*	0.97	*Clostridium unknown_species*	0.97	*Eubacterium unknown_species*	0.96
Clostridium unknown_species	0.61	*Clostridium unknown_species*	0.96	*Eubacterium unknown_species*	0.97	*Clostridium unknown_species*	0.96
Bacteroides unknown_species	0.61	*Bacteroides uniformis*	0.94	*Roseburia unknown_species*	0.96	*Faecalibacterium unknown_species*	0.95
Ruminococcus unknown_species	0.58	*Eubacterium unknown_species*	0.94	*Faecalibacterium prausnitzii*	0.96	*Roseburia unknown_species*	0.94
Bacteroides uniformis	0.57	*Roseburia unknown_species*	0.93	*Faecalibacterium unknown_species*	0.96	*Faecalibacterium prausnitzii*	0.94
Blautia wexlerae	0.57	*Faecalibacterium unknown_species*	0.92	*Eubacterium rectale*	0.93	*Enterocloster unknown_species*	0.9
Flavonifractor plautii	0.54	*Faecalibacterium prausnitzii*	0.92	*Bacteroides unknown_species*	0.93	*Bacteroides uniformis*	0.88
Ruminococcus gnavus	0.51	*Blautia wexlerae*	0.91	*Enterocloster unknown_species*	0.91	*Bacteroides unknown_species*	0.88

We draw the scatterplots of the species occurrences across all samples vs. species occurrences within groups for both binary (prevalence, Figure 3D) and weighted (relative abundance, Figure 3D) occurrences. The scatterplots in Figure 3D allow us to visually identify which species are prevalent in the group compared to their presence across all groups. The points above the bisector line indicate species with an average occurrence within a group higher than the average occurrence among groups (among all samples). On the contrary, points below the bisector line indicate species that occur less within the group than when considering all samples. Points on the bisector lines indicate similar average species occurrence within and across groups.

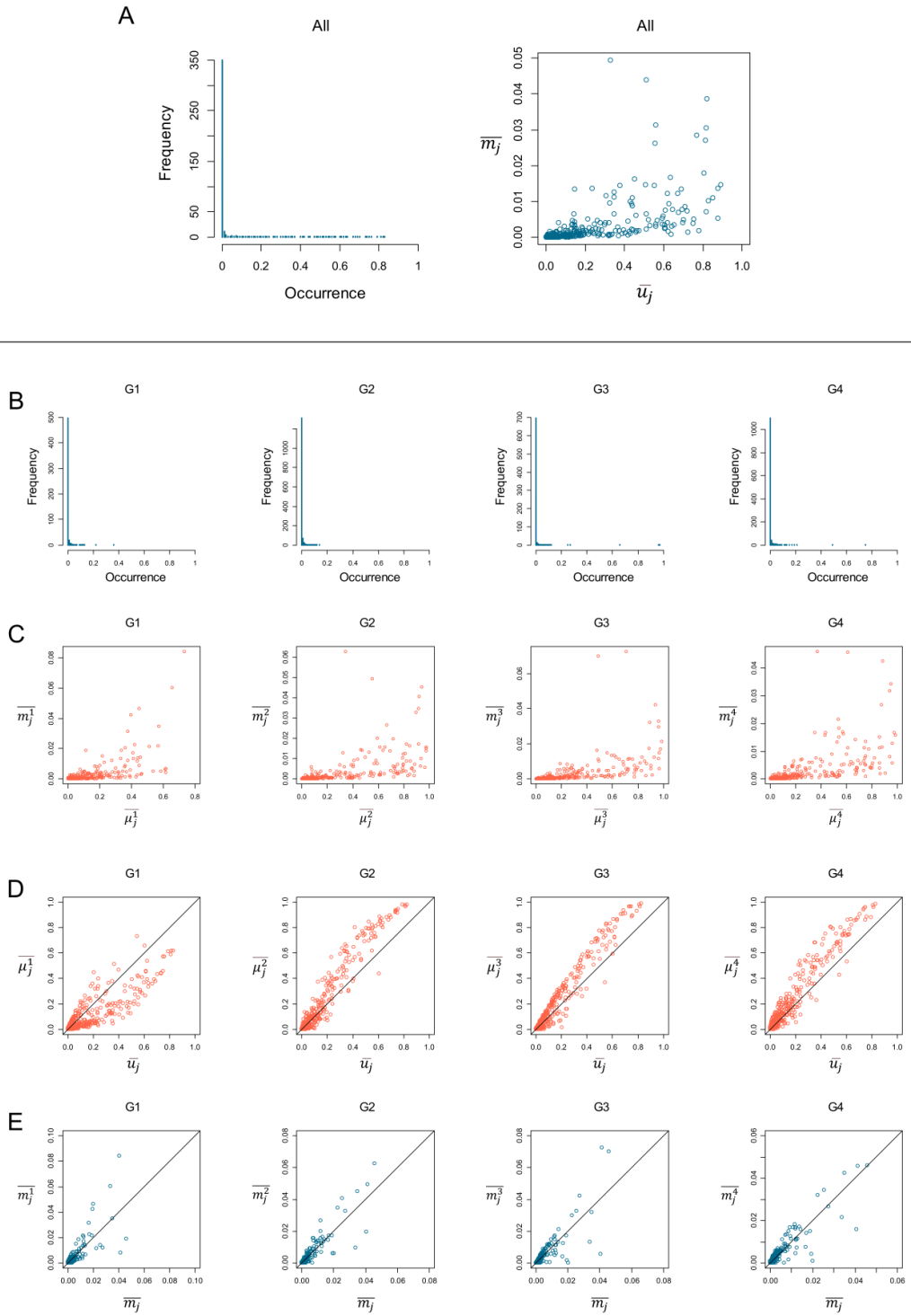

Figure 3. (**A**) Left panel: bacterial species binary occurrence frequency distributions for all samples (All); x-axis: species binary occurrence; y-axis: frequency of the occurrence value. The x-axis is normalized by the total number of samples for each plot; in this way, the occurrence may range from 0 (no occurrence) to 1 (the species occurs in all the samples). Right panel: scatterplots of the average weighted occurrence (m_j) vs. average binary occurrence (u_j) for all samples. (**B**) Bacterial species binary occurrence frequency distributions for age groups (G1, G2, G3, G4). (**C**) Scatterplots of the average weighted occurrence (m_j) vs. average binary occurrence (u_j) for age groups (G1, G2, G3, G4). (**D**) Scatterplots of the average binary occurrence within group ($\overline{\mu_j^y}$) vs. the average binary occurrence ($\overline{u_j}$) for age groups (G1, G2, G3, G4). (**E**) Scatterplots of the average weighted occurrence within group ($\overline{m_j^y}$) vs. the average weighted occurrence ($\overline{m_j}$) for age groups (G1, G2, G3, G4).

4.2. Statistical Distance Notions

Figure 4 compares the group species prevalence ($\overline{\mu_j^y}$) for the 50 species with the highest $\overline{\mu_j^y}$ and the value $\overline{\mu_j^y}$ for the species with the highest statistical distance values, RBD and IOBD. We find a significant difference between the $\overline{\mu_j^y}$ values for the 50 species ranked by the average binary occurrence (ABO) (Figure 4 black line, top row) and ranked by the relative binary distance (RBD) (Figure 4 red line, top row). The $\overline{\mu_j^y}$ values for the first 50 species ranked by ABO are above 0.6 for all groups, whereas the $\overline{\mu_j^y}$ values for the first 50 species ranked by RBD are very low (<0.1). This result indicates that species with a higher RBD in the group may exhibit a low prevalence in the same group. The high RBD exhibited by species with a very low average occurrence in the group would suggest that species with a very low prevalence within the group may be highly characteristic of the group.

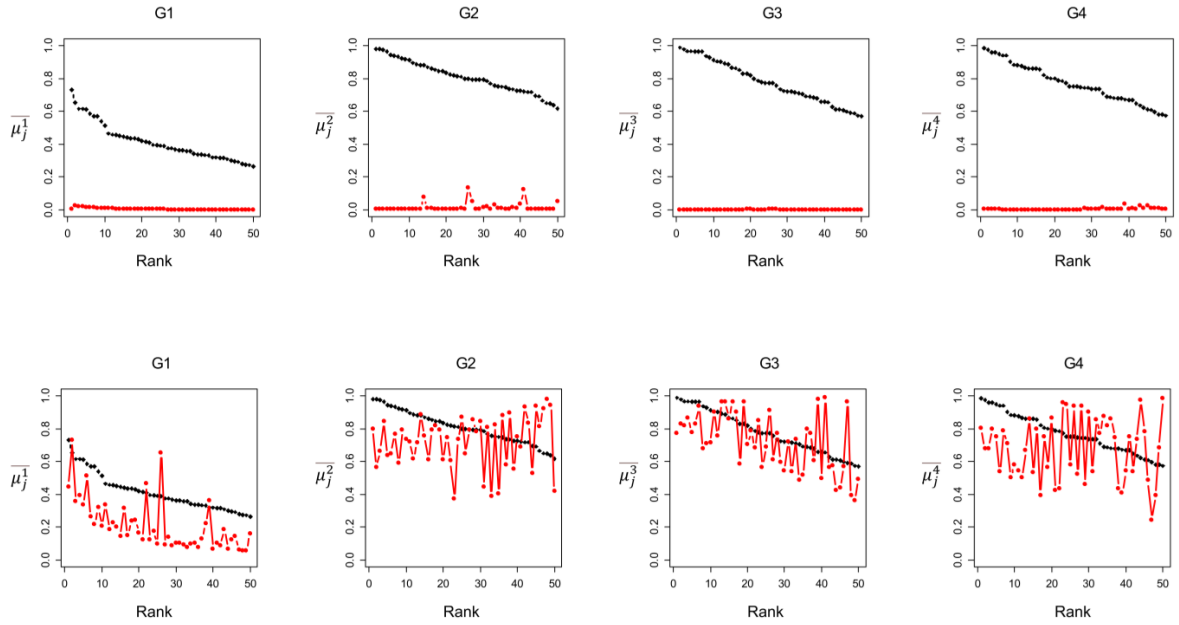

Figure 4. Comparison between species average binary occurrence for each group ($\overline{\mu_j^y}$) and the notions of statistical distance. Top row: first 50 species ranked by average binary occurrence (ABO) (black line) and relative binary distance (RBD) (red line); bottom row: first 50 species ranked by average binary occurrence (ABO) (black line) and inside–outside binary distance (IOBD) (red line).

Nevertheless, as noted in the 'Relative Distances' section of the Methods, when a bacterial species is exclusively found within a group, the RBD value is calculated as the ratio of the total number of samples to the group's sample size (Equation (12)). Consequently, species exclusive to a single group will exhibit identical RBD values regardless of their varying prevalence within that group. To elucidate this issue, we determined the number of group-exclusive species among the top 50 species ranked by both RBD and RWD, revealing counts of 50, 5, 50, and 27 species for groups G1, G2, G3, and G4, respectively. This result unveils that many species with the highest RWD and RBD ranking are exclusively present within one group (for groups G1 and G3, all the first 50 species are exclusive to those groups). The RWD and RBD return rank ties for species that occur only in one group; therefore, RWD and RBD cannot discriminate the relative importance of these species. Table 3 lists the twenty species of highest rank for the G1 for each classification strategy. As we can see, the RBD and RWD rankings differ from the other strategies, confirming the peculiar outcomes provided by these classification strategies. These latest results indicate how the statistical distances RWD and RBD may present problems in selecting the most characteristic species of a group in our database.

Table 3. Twenty species of highest rank for the G1 for each ranking strategy.

ABO	AWO	RBD	RWD	IOBD	IOWD	MM	CM
Bifidobacterium longum	*Bifidobacterium longum*	*Methylobacterium unknown_species*	*Microbacterium oleivorans*	*Bifidobacterium breve*	*Bifidobacterium longum*	*Bifidobacterium longum*	*Bifidobacterium longum*
Escherichia coli	*Escherichia coli*	*Cutibacterium avidum*	*Neisseria meningitidis*	*Bifidobacterium longum*	*Escherichia coli*	*Escherichia coli*	*Escherichia coli*
Blautia unknown_species	*Bifidobacterium breve*	*Vibrio harveyi*	*Rhizobium daejeonense*	*Erysipelatoclostridium ramosum*	*Bifidobacterium breve*	*Ruminococcus gnavus*	*Ruminococcus gnavus*
Clostridium unknown_species	*Bifidobacterium bifidum*	*Actinomyces urogenitalis*	*Rubrobacter unknown_species*	*Bifidobacterium bifidum*	*Bifidobacterium bifidum*	*Bifidobacterium unknown_species*	*Bifidobacterium unknown_species*
Bacteroides unknown_species	*Bacteroides uniformis*	*Staphylococcus hominis*	*Scandinavium goeteborgense*	*Veillonella parvula*	*Bacteroides fragilis*	*Bifidobacterium breve*	*Bifidobacterium breve*
Ruminococcus unknown_species	*Bacteroides fragilis*	*Nocardia nova*	*Serratia nematodiphila*	*Ruminococcus gnavus*	*Veillonella parvula*	*Bifidobacterium bifidum*	*Bifidobacterium bifidum*
Bacteroides uniformis	*Phocaeicola dorei*	*Acinetobacter lwoffii*	*Acidovorax oryzae*	*Veillonella unknown_species*	*Ruminococcus gnavus*	*Bifidobacterium pseudocatenulatum*	*Erysipelatoclostridium ramosum*
Blautia wexlerae	*Blautia wexlerae*	*Streptococcus peroris*	*Cloacibacterium normanense*	*Enterococcus faecalis*	*Enterococcus faecalis*	*Erysipelatoclostridium ramosum*	*Eggerthella lenta*
Flavonifractor plautii	*Ruminococcus gnavus*	*Azoarcus communis*	*Frigoribacterium unknown_species*	*Clostridium innocuum*	*Bifidobacterium pseudocatenulatum*	*Eggerthella lenta*	*Veillonella parvula*
Ruminococcus gnavus	*Bifidobacterium pseudocatenulatum*	*Acidovorax oryzae*	*Gleimia unknown_species*	*Veillonella atypica*	*Phocaeicola dorei*	*Veillonella parvula*	*Clostridium innocuum*
Bifidobacterium unknown_species	*Prevotella copri*	*Mycolicibacterium elephantis*	*Herbaspirillum huttiense*	*Eggerthella lenta*	*Parabacteroides distasonis*	*Clostridium innocuum*	*Enterocloster bolteae*
Eubacterium unknown_species	*Veillonella parvula*	*Serratia liquefaciens*	*Afipia broomeae*	*Klebsiella michiganensis*	*Erysipelatoclostridium ramosum*	*Enterocloster bolteae*	*Veillonella unknown_species*
Phocaeicola vulgatus	*Parabacteroides distasonis*	*Micromonospora endophytica*	*Aggregatibacter kilianii*	*Hungatella effluvii*	*Klebsiella pneumoniae*	*Veillonella unknown_species*	*Coprococcus phoceensis*
Bacteroides thetaiotaomicron	*Enterococcus faecalis*	*Myxococcus xanthus*	*Agreia unknown_species*	*Haemophilus unknown_species*	*Staphylococcus epidermidis*	*Streptococcus unknown_species*	*Haemophilus parainfluenzae*
Bifidobacterium breve	*Phocaeicola vulgatus*	*Micrococcus yunnanensis*	*Lysobacter enzymogenes*	*Lactobacillus rhamnosus*	*Bifidobacterium dentium*	*Coprococcus phoceensis*	*Hungatella effluvii*
Faecalibacterium unknown_species	*Faecalibacterium unknown_species*	*Ralstonia pickettii*	*Mannheimia unknown_species*	*Enterocloster bolteae*	*Enterobacter hormaechei*	*Haemophilus parainfluenzae*	*Intestinibacter bartlettii*
Roseburia unknown_species	*Anaerostipes hadrus*	*Metakosakonia unknown_species*	*Massilia unknown_species*	*Veillonella infantium*	*Blautia wexlerae*	*Hungatella effluvii*	*Enterococcus faecalis*
Faecalibacterium prausnitzii	*Collinsella aerofaciens*	*Neisseria flavescens*	*Achromobacter insuavis*	*Haemophilus parainfluenzae*	*Haemophilus haemolyticus*	*Intestinibacter bartlettii*	*Veillonella atypica*
Enterocloster unknown_species	*Bifidobacterium adolescentis*	*Streptomyces albidochromogenes*	*Alicycliphilus denitrificans*	*Coprococcus phoceensis*	*Haemophilus parainfluenzae*	*Enterococcus faecalis*	*Haemophilus unknown_species*
Phocaeicola dorei	*Eubacterium rectale*	*Cutibacterium unknown_species*	*Micrococcus luteus*	*Sellimonas intestinalis*	*Veillonella atypica*	*Veillonella atypica*	*Phocaeicola sartorii*

Then, we find a reduced difference between the $\overline{\mu_j^y}$ values for the 50 species ranked by prevalence (average binary occurrence (ABO), Figure 4 black line, bottom row) and ranked by the inside–outside binary distance (IOBD) (Figure 4 red line, bottom row). However, we observe that the $\overline{\mu_j^y}$ variability for species ranked by IOBD is very high. Some bacterial species have a high occurrence, while others close in rank show a much lower occurrence. This result indicates that low-prevalence species within the group may instead have a large statistical distance between the average occurrence inside and outside the group, meaning that they are much more prevalent in the group compared to their occurrence in the other groups. These species may be good candidates to characterize the group. Figure 4 shows an interesting pattern for G1. G1 shows the highest difference between the $\overline{\mu_j^1}$ values of the species with the highest occurrence and the $\overline{\mu_j^1}$ values of the species with the largest IOBD. Differently from the other groups, G1 is characterized by a set of bacterial species that occur preferentially in G1, that is, bacterial species that show a higher difference between their prevalence in individuals of young age and their prevalence in older ages. In detail, the ten species ranked by IOBD mainly belong to six different

genera, i.e., *Bifidobacterium*, *Clostridium*, *Enterococcus*, *Erysipelatoclostridium*, *Ruminococcus*, and *Veillonella*, which are typical of the infant microbiota and consistent with previous results obtained from the pooled analysis of these datasets [21]. In particular, the highest-ranked species are *Bifidobacterium bifidum*, *Bifidobacterium breve*, and *Bifidobacterium longum*, which are widely recognized as primary colonizers of the infant gut, confirming the validity of the statistical approach used.

The IOWD and IOBD show some advantages concerning the relative distances RWD and RBD. We compute the number of species that occur only within a group in the first 50 species ranked by IOWD and IOBD, discovering that no species occur only within a group for both ranking strategies. Further, we outline that the IOBD and IOWD do not create ties when species occur only in one group. If a species j occurs only in group y, the average occurrence outside group y is $\overline{\mu_j^{\sim y}} = 0$. Consequently, Equation (15) becomes $\Delta_j^y = \overline{\mu_j^y}$, indicating that the statistical distance IOBD is the average occurrence of species j within group y. The same reasoning can be applied to IOWD computed in Equation (14), and the IOWD value for species occurring only in one group becomes ${}_w\Delta_j^y = \overline{m_j^y}$. These results ensure no ties for species with different occurrence values in the IOBD and IOWD species rank.

4.3. Monte Carlo Simulations

Figure 5 depicts the scatterplots for each group of the species binary occurrence vs. the MC binary occurrence outcomes for both microcanonical and canonical approaches. The x-axis (*Sim*) represents the MC simulation outcomes of the species occurrences, and the y-axis represents the empirical/observed (*Obs*) occurrences of the species (prevalence). The bisector line indicates the perfect agreement between the species' empirical and simulated occurrences. Agreement between two measurements refers to the degree of concordance between them and can be evaluated with different statistical points of view [32,33]. Here, agreement refers to the difference between the empirical and MC simulated occurrences, and, therefore, the bisector indicates no difference between the empirical and simulated occurrences. Points above the bisector line are bacterial species with empirical occurrences higher than the simulated ones; however, species below the bisector line present simulated occurrences higher than empirical ones. G1 presents many of the most prevalent species below the bisector line concerning the other groups. Differently, other groups (G2, G3, and G4) present many of the more prevalent species above the bisector line.

Figure 6A outlines the 20 bacterial species with the highest difference between the observed occurrence (*Obs*) and the simulated occurrence (*Sim*) in G1 by green points. These species are the most distant above the bisector line, indicating their empirical occurrence in G1 is higher than in the simulated one. The higher observed occurrence indicates that these bacterial species appear much more frequently in G1 than expected by chance, indicating their high prevalence in the microbiota of individuals in G1. We call these taxa 'the characterizing species' of G1. On the contrary, black points are the twenty bacterial species with the lowest difference between the observed occurrence (*Obs*) and the simulated occurrence (*Sim*) in group 1 (G1). They are 'rare species', lying with the highest distance below the bisector line, unveiling that their simulated occurrence in G1 is higher than the empirical one.

Figure 6B shows that the rare bacterial species with low occurrence in G1 (black points) lie distant above the bisector line in groups G2, G3, and G4, indicating that these bacterial species present a higher observed occurrence than expected by chance in G2, G3, and G4.

Group 1 is the set of younger individuals. The age of individuals increases from G1 to G4. The results of MCs tell us from a numerical–statistical perspective that when age increases, there is a clear transition of the species composition in the microbiota of individuals. In G2, the characterizing species that in the G1 plot lie above the bisector line (Figure 6) are superimposed on the bisector line, indicating that their occurrences do not deviate from what is expected by chance. Then, in G3 and G4, the characterizing species

are clearly below the bisector line, showing that their empirical occurrence in the group is lower than expected by chance. This pattern demonstrates that the transition emerges passing from G1 to G3 and that G2 represents the transition age group.

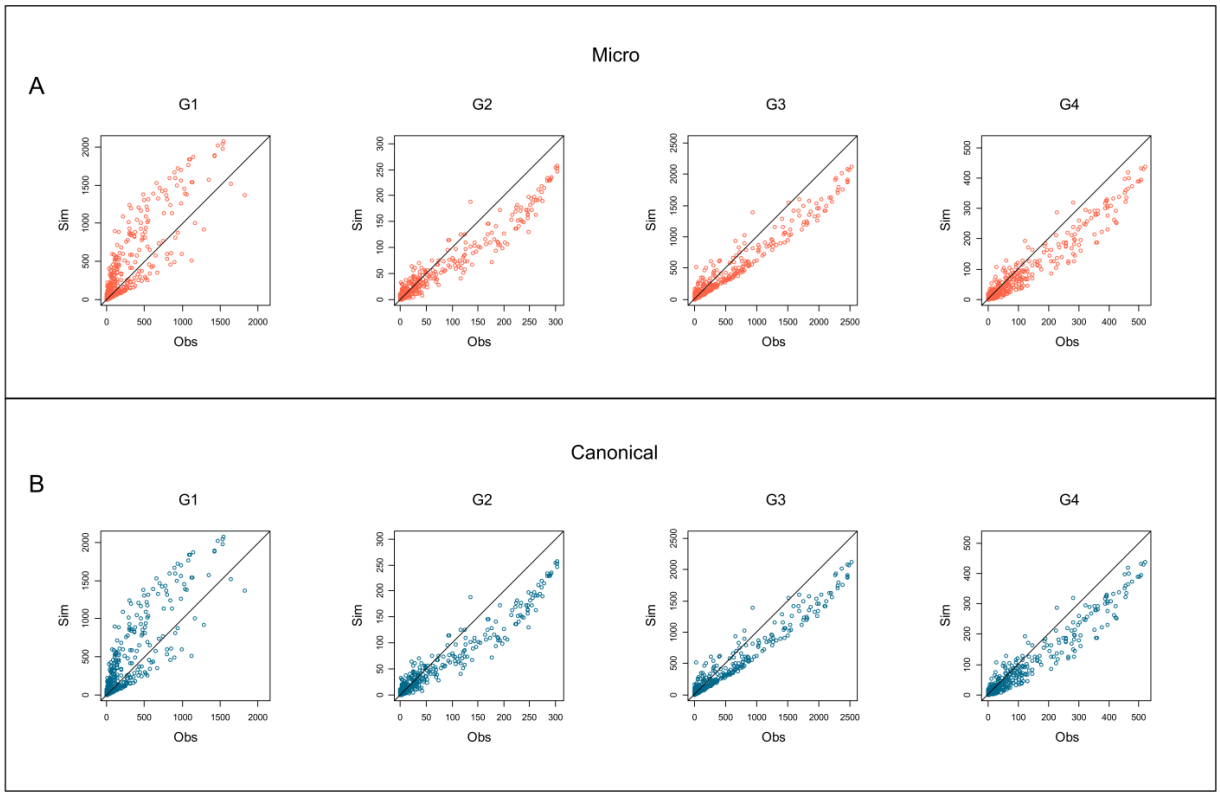

Figure 5. Scatterplots of the species binary occurrence (presence/absence in the sample) for each group. *Y*-axis (*Sim*) represents the Monte Carlo (MC) simulation outcomes of the species occurrences; *x*-axis (*Obs*) represents the empirical (observed) occurrences of the species. (**A**) Microcanonical MC simulations; (**B**) canonical MC simulations. Columns depict the scatterplots for the four groups by age (G1, G2, G3, G4). The bisector line indicates the complete agreement between the empirical and the simulated occurrences, that is, the bisector indicates no difference between the empirical and simulated occurrences. The bacterial species simulated occurrences (*Sim*) are the average over the entire set of MC simulations.

Figure 7 shows the average binary occurrence in the group ($\overline{\mu_j^1}$) for the first 10 characterizing bacterial species with the highest difference between *Obs* and *Sim* in G1. The average binary occurrence in G1 ($\overline{\mu_j^1}$) is generally high and decreases in the other groups (Figure 7, top row). Three characterizing bacterial species, *Veillonella dispar*, *Enterococcus faecalis*, and *Hydrogenoanaerobacterium* unknown_species, present $\overline{\mu_j^1} < 0.3$, outlining how the MCs unveil that species of lower prevalence can be essential to characterize the microbiota of the age group. Figure 7, bottom row, depicts $\overline{\mu_j^1}$ for the first rare species in G1, that is, species minimizing the difference between *Obs* and *Sim* in G1. $\overline{\mu_j^1}$ of the rare species identified by the MC simulations in G1 are very low, and they quickly increase in the other higher age groups, showing that these rare species in G1 become dominant in the microbiota with age. These results reflected the fluctuation and the adaptation of the intestinal microbiota during the human life span. In fact, species like *Bifidobacterium breve*, *Bifidobacterium longum*, and *Veillonella parvula* are predominant in the infant gut microbiota and decrease as individuals age. Interestingly, *Bifidobacterium longum*, *Ruminococcus gnavus*, and *Clostridium innocuum* decrease less significantly, indicating that these taxa remain present in adults. Conversely, certain bacterial species, such as *Agathobaculum butyriciproducens*, *Eubacterium rectale*, and

Coprococcus, are found to have a low prevalence in infants and increase significantly in adults. This dynamic change aligns with findings in the literature, which indicate that gut microbiota composition evolves with age due to varying physiological stages and dietary habits. Species *Eubacterium rectale* presents a higher prevalence in G1 ($\overline{\mu_j^1} > 0.3$) than many characterizing species (see panels for G1, Figure 7), demonstrating how the simple prevalence of bacterial species may not be a reliable proxy for their importance in characterizing age groups.

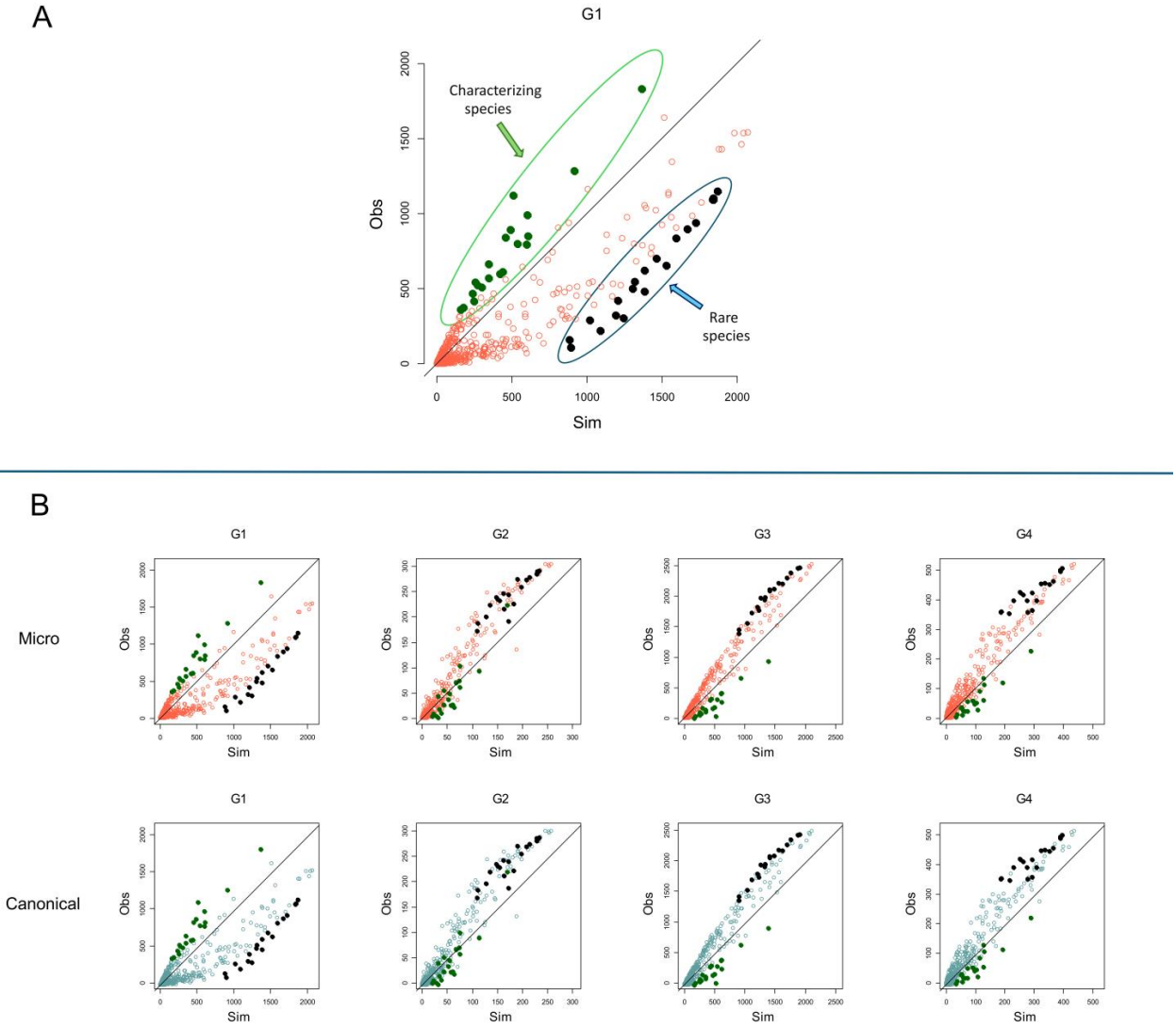

Figure 6. (Panel (**A**)) Scatterplots of the observed species binary occurrence (*Obs*) and the simulated occurrence (*Sim*) in group 1 (G1) for the microcanonical Monte Carlo approach. We outline the points in green for the 20 bacterial species with the highest difference between *Obs* and *Sim*. These 'characterizing species' are the more distant below the bisector line in chart G1, indicating that the observed (empirical) occurrence in group 1 is higher than the simulated one. The points in black are the 20 bacterial species with the lowest difference between the observed occurrence (*Obs*) and the simulated occurrence (*Sim*) in group 1 (G1). These are the 'rare species' that are the more distant above the bisector line in chart G1, indicating that the simulated occurrence in G1 is higher than the empirical one. (Panel (**B**)) Scatterplots of the species binary occurrence for each group, where green points are the 20 bacterial species with the highest difference between *Obs* and *Sim*, and black points are the 20 bacterial species with the lowest difference between *Obs* and *Sim*.

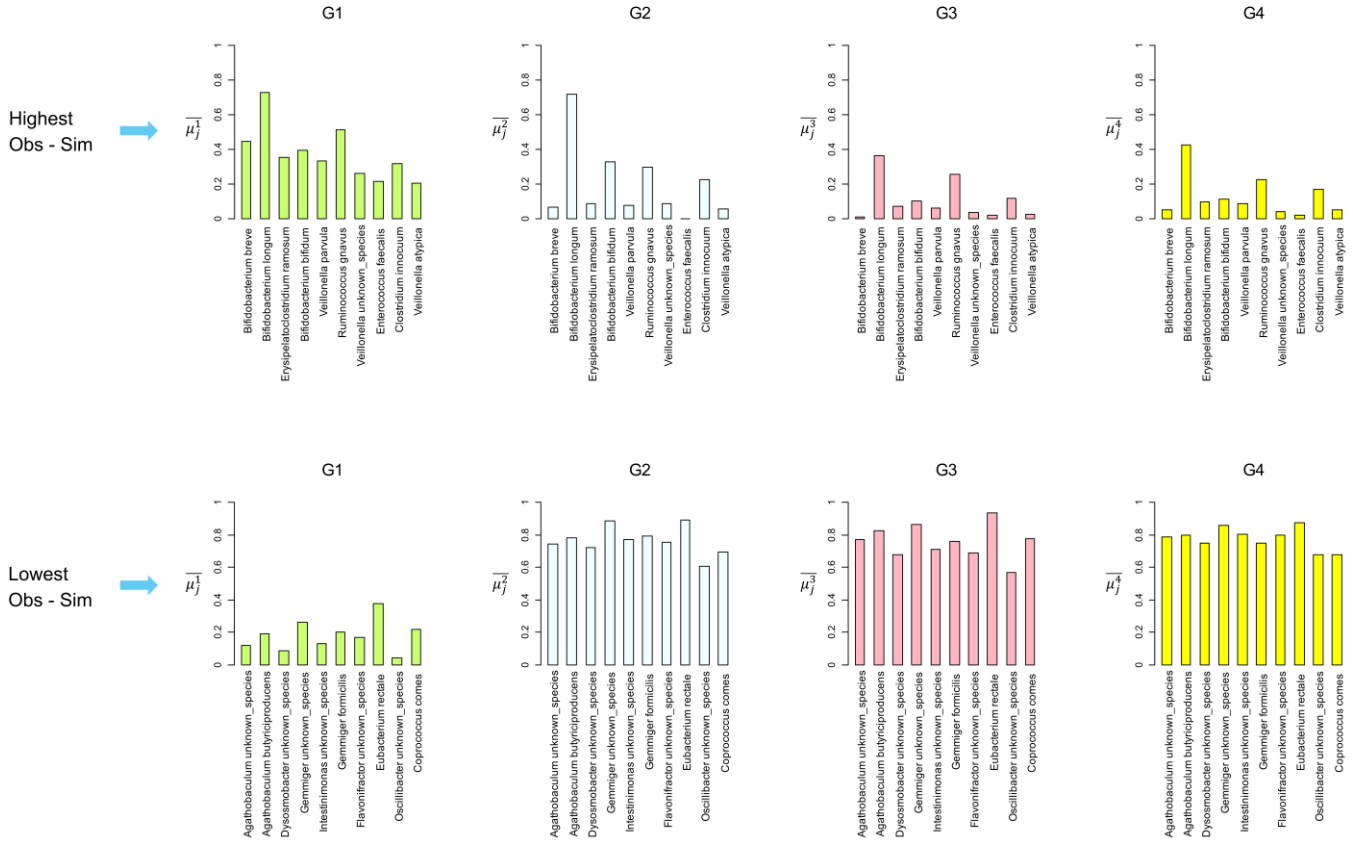

Figure 7. First row: Barplots of the average binary occurrence in the group ($\overline{\mu_j^y}$) for the 10 bacterial species with the highest difference between the observed occurrence (*Obs*) and the simulated occurrence (*Sim*) in group 1 (G1). These are characterizing species lying more distant above the bisector line in Figure 6, indicating that the simulated occurrence in group 1 is lower than the empirical one. Second row: Barplots of the average binary occurrence in the group ($\overline{\mu_j^y}$) for the ten bacterial species with the lowest difference between the observed occurrence (*Obs*) and the simulated occurrence (*Sim*) in group 1 (G1). These are rare species lying more distant below the bisector line in Figure 6, indicating that the simulated occurrence in group 1 is higher than the empirical one.

The results from the two different approaches to the MCs are similar. Figure S1 in the Supplementary Materials file shows the scatterplots of the *p*-values carried out with the microcanonical approach against the *p*-values obtained from the canonical approach for each age group. The *p*-values of the two MC approaches are correlated, demonstrating that the two approaches yield similar significance values for the occurrence of the bacterial species.

Figure 8 illustrates the scatterplots of the group prevalence ($\overline{\mu_j^y}$) vs. the MC simulated *p*-value for the same group (*p*-val) for both microcanonical and canonical approaches. The points lying on the *x*-axis indicate bacterial species with very low and significant *p*-values (*p*-value < 0.05). These are species for which it is highly unlikely to obtain their empirical occurrence by chance with the MCs. In other words, these species preferentially occur within a group if comparing their occurrence in other groups. On the contrary, species with a *p*-value approaching 1 are bacterial species that are likely to present an MC simulation occurrence in the group higher than the empirical occurrence in the same group; that is, they are species that do not preferentially occur in the group. It should be noted that many species with *p*-value ≈ 1 present a very high occurrence in the group. For example, there are many species in G1 with $\overline{\mu_j^1} > 0.5$ and with *p*-value ≈ 1; that is, they occur in more than half of the samples in the group, but it is very likely to obtain these occurrences by chance. Thus, they are not significant for MC simulations. This result suggests that the

simple species prevalence in a group is insufficient to identify the more critical bacterial species for a particular age group.

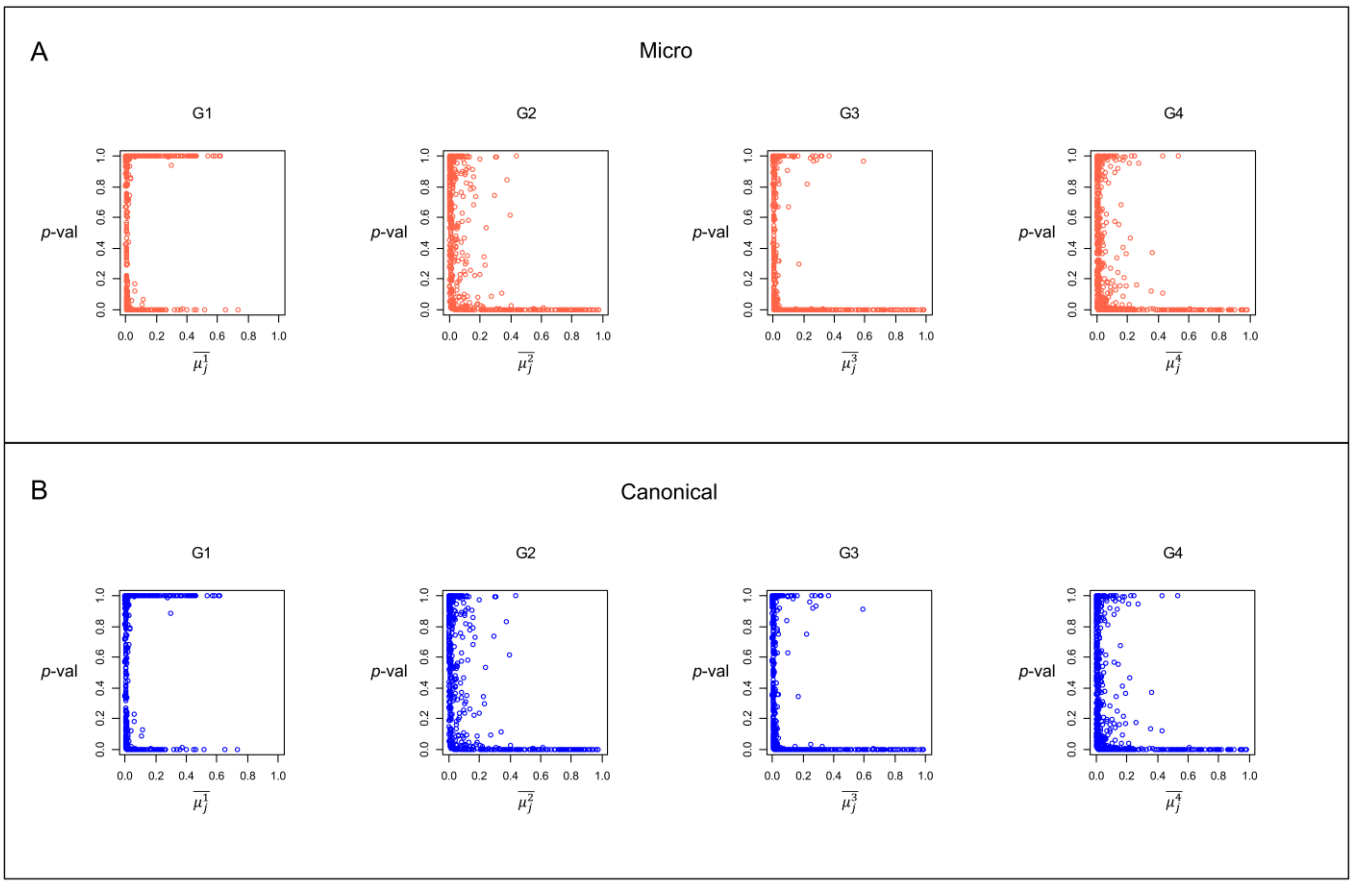

Figure 8. Scatterplots of the species average binary occurrence for each group ($\overline{\mu_j^y}$) vs. the Monte Carlo (MC) *p*-value (*p*-val) to obtain the occurrence by chance. (**A**) Microcanonical MC simulations; (**B**) Canonical MC simulations. Columns depict the scatterplots for the age four groups (G1, G2, G3, G4).

4.4. Species Rank Correlations

We analyze the species rank correlations among classification strategies by computing the number of common bacterial species in the first 50 species for each rank. The set S_i defines the first 50 bacterial species for strategy i, and S_j defines the first 50 bacterial species for strategy j; the cardinality of the intersection between the two species set $|S_i \cap S_j|$ returns the number of common species between the ranks. Table 4 depicts the species rank intersection for the first 50 species ranked by each strategy for each group. Therefore, the element i, j in Table 4 indicates the number of species that the ranks S_i and S_j share, that is, $|S_i \cap S_j| = 10$ indicates that the two ranks share 10 bacterial species in the first 50 ranked species. The relative distances RBD and RWD do not share common species with the other strategies, showing how these two strategies, which give high importance to species, including rare ones, that occur solely within a group, return peculiar species rankings. Most importantly, G1 differs from all other groups, unveiling the lowest overlap between the ranks provided by the MC simulations and the ranks based on species occurrence (ABO and AWO). In G1, the MCs present 13 (MM) and 12 (CM) common species with the highest occurrence species rank (ABO). In G2, G3, and G4, the MCs present more than 48 common species in common with the highest occurrence species rank (ABO). This result has two main implications. On the one hand, it demonstrates how G1 differs from all others in terms of bacterial species composition. Additionally, the MCs identify bacterial species as highly important for group 1, which may not necessarily be of high binary or weighted

occurrence within the group. Therefore, MC simulation approaches may be a valuable tool for gathering additional information on the importance of bacterial species for age groups.

Table 4. Species rank intersection for the first 50 species ranked by each strategy for each group. Element i, j of tables indicates the species overlapping between the rank set of strategy S_i and the rank set of strategy S_j, that is $\left|S_i \cap S_j\right|$. In other words, the element i, j indicates the number of species that S_i and S_j share, that is, $\left|S_i \cap S_j\right| = 10$ means that the two ranks share 10 bacterial species in the first 50 ranked species. Let us consider an example with two hypothetical ranks of 5 species. For example, if $S_i = \{B, E, U, A, T\}$ and $S_i = \{W, E, H, T, M\}$, then the intersection is $S_i \cap S_j = \{E, T\}$, and the cardinality $\left|S_i \cap S_j\right| = 2$. This indicates that the two hypothetical ranks share 2 elements among the top 5 positions. The higher the cardinality of the intersection, the greater the similarity between the two ranks. Bold font indicates column and row titles.

					G1										**G2**			
$S_i \cap S_i$	**ABO**	**AWO**	**RBD**	**RWD**	**IOBD**	**IOWD**	**MM**	**CM**	$S_i \cap S_i$	**ABO**	**AWO**	**RBD**	**RWD**	**IOBD**	**IOWD**	**MM**	**CM**
ABO	50	33	0	0	13	21	13	12	**ABO**	50	30	0	0	37	25	50	50
AWO	0	50	0	0	16	30	16	15	**AWO**	0	50	0	0	29	32	30	30
RBD	0	0	50	5	0	0	0	0	**RBD**	0	0	50	37	0	0	0	0
RWD	0	0	0	50	0	0	0	0	**RWD**	0	0	0	50	0	1	0	0
IOBD	0	0	0	0	50	27	44	45	**IOBD**	0	0	0	0	50	31	37	37
IOWD	0	0	0	0	0	50	27	26	**IOWD**	0	0	0	0	0	50	25	25
MM	0	0	0	0	0	0	50	46	**MM**	0	0	0	0	0	0	50	50
CM	0	0	0	0	0	0	0	50	**CM**	0	0	0	0	0	0	0	50
					G3										**G4**			
$S_i \cap S_i$	**ABO**	**AWO**	**RBD**	**RWD**	**IOBD**	**IOWD**	**MM**	**CM**	$S_i \cap S_i$	**ABO**	**AWO**	**RBD**	**RWD**	**IOBD**	**IOWD**	**MM**	**CM**
ABO	50	35	0	0	38	29	49	48	**ABO**	50	31	0	0	31	24	48	48
AWO	0	50	0	0	27	34	34	34	**AWO**	0	50	0	0	21	28	29	29
RBD	0	0	50	14	0	0	0	0	**RBD**	0	0	50	40	0	0	0	0
RWD	0	0	0	50	0	0	0	0	**RWD**	0	0	0	50	0	0	0	0
IOBD	0	0	0	0	50	32	39	40	**IOBD**	0	0	0	0	50	27	32	32
IOWD	0	0	0	0	0	50	29	29	**IOWD**	0	0	0	0	0	50	25	25
MM	0	0	0	0	0	0	50	49	**MM**	0	0	0	0	0	0	50	50
CM	0	0	0	0	0	0	0	50	**CM**	0	0	0	0	0	0	0	50

4.5. Overall Ranking

In this article, we propose eight classification strategies using different rationales for identifying and ranking the characterizing bacterial species for the different age groups. The different strategies furnish different species ranking. Therefore, we perform a last analysis to obtain an overall species ranking by evaluating the results of the different strategies together. We count the frequency of each bacterial species appearing in the first ten species in each strategy, leading to a comprehensive rank evaluation. The results of this analysis are in Table 5. Tables S1–S4 depict the overall ranking computing for each group.

Table 5 lists the ten species with the highest score according to the overall ranking for each group, with the percentage indicating the number of times the species is ranked within the first ten species for each ranking strategy. For example, 75% indicates that the species is ranked in the first ten species in 75% of the cases, i.e., it figures in six of the eight ranking strategies. For instance, *Bifidobacterium longum*, *B. breve*, and *Ruminococcus gnavus* are prevalent in younger individuals (G1), while species such as *Faecalibacterium prausnitzii* and *Eubacterium rectale* become more significant in older groups (G3 and G4). This transition aligns with the literature, indicating that gut microbiota composition evolves with age due to varying physiological stages and dietary habits.

Table 5. The 10 highest ranking species according to the overall ranking. The percentage beside each species indicates the fraction of occurrences of the species within the first ten species for each ranking strategy. For example, 75% indicates that the species is ranked in the first ten species 75% of the time, i.e., it figures 6 times over the 8 ranking strategies. Bold font indicates column and row titles.

Rank	**G1**		**G2**		**G3**		**G4**	
1	*Bifidobacterium longum*	75%	*Bacteroides unknown_species*	62.5%	*Faecalibacterium prausnitzii*	75%	*Intestinimonas unknown_species*	63%
2	*Bifidobacterium breve*	75%	*Faecalibacterium prausnitzii*	62.5%	*Faecalibacterium unknown_species*	75%	*Faecalibacterium prausnitzii*	63%
3	*Ruminococcus gnavus*	75%	*Faecalibacterium unknown_species*	62.5%	*Eubacterium rectale*	75%	*Faecalibacterium unknown_species*	63%
4	*Escherichia coli*	62.5%	*Ruminococcus unknown_species*	62.5%	*Eubacterium unknown_species*	75%	*Ruminococcus unknown_species*	63%
5	*Bifidobacterium bifidum*	62.5%	*Bacteroides uniformis*	62.5%	*Roseburia unknown_species*	75%	*Bacteroides uniformis*	63%
6	*Veillonella parvula*	62.5%	*Eubacterium rectale*	62.5%	*Roseburia inulinivorans*	75%	*Gemmiger unknown_species*	63%
7	*Enterococcus faecalis*	62.5%	*Phocaeicola vulgatus*	62.5%	*Blautia unknown_species*	63%	*Blautia unknown_species*	50%
8	*Erysipelatoclostridium ramosum*	50%	*Blautia unknown_species*	50%	*Ruminococcus unknown_species*	63%	*Agathobaculum butyriciproducens*	50%
9	*Veillonella atypica*	50%	*Parabacteroides unknown_species*	50%	*Lachnospira unknown_species*	63%	*Eubacterium rectale*	50%
10	*Haemophilus parainfluenzae*	50%	*Gemmiger unknown_species*	50%	*Gemmiger unknown_species*	63%	*Eubacterium unknown_species*	50%

An overall ranking, that is, a multicriteria approach to find and classify important bacterial species for each age group, can be helpful when each criterion to rank species accounts for specific and important information about species occurrence in the sample database. For example, we are interested in evaluating both the binary occurrence and the sample abundance. In that case, we have to consider notions of statistical distance focusing on the species' prevalence and abundance together. On the other hand, when the information for the classification problem we need to solve is specific and exhaustive, using a multicriteria approach is not recommended, as it would confuse the bacterial species classification with information derived from other classification methods based on different rationales.

Finally, using a multicriteria approach can help discover 'eccentric criteria' that provide results utterly different from the results of other criteria. Tables S1–S4 show that RBD and RWD provide a species ranking that differs from all other strategies. This evidence suggests that RBD and RWD may have classification issues with the database under examination.

5. Conclusions

This manuscript proposes and tests different classification strategies to characterize important bacterial species in human microbiota for different age groups. First, in addition to the classic statistical notions of species prevalence and abundance, we introduce different notions of statistical distance to classify important species for each age group. Among the approaches used, RBD and RWD appear less effective, as they return many rank ties for species of different prevalence, and they frequently rank species of a negligible prevalence within the group. The other statistical distance notions IOBD and IOWD return more reliable results. On the one hand, they do not return ties for species of different prevalence; on the other hand, IOBD and IOWD prioritize both low- and high-occurrence species within the group, suggesting that low-prevalence species within the group may hold greater significance to the group's identity. These species may be good candidates to characterize the group.

Then, we perform machine learning Monte Carlo simulations based on the physics concepts of the microcanonical and canonical ensembles in statistical mechanics to characterize the bacterial species. MCs furnish important outcomes. First, the MCs tell us from a numerical–statistical perspective that when age increases, there is a clear transition of the species composition in the microbiota of individuals. The transition emerges passing from G1 (0–4 years) to G3 (18–64 years), and G2 (5–17 years) represents the transition age group. MCs demonstrate that the microbiota changes throughout the whole period, from early childhood to adolescence, and stabilizes in adulthood. Some species, such as *Bifidobacterium breve* and *Veillonella parvula*, are predominant in infants but decrease with age, while others, like *Agathobaculum butyriciproducens* and *Eubacterium rectale*, increase. The two MC approaches used yielded similar results, demonstrating the robustness of the findings.

Second, MCs show that low-prevalence species may be statistically significant in characterizing age groups, unveiling how the simple prevalence of a bacterial species in an age group may not be a comprehensive proxy for its importance. Third, MCs are a useful tool for identifying species for age groups by simply computing the bacterial species, maximizing the difference between empirical prevalence and simulated one. These species with a very high difference between empirical and simulated prevalence are very unlikely to occur in the group by chance, and for this, they are highly significant for the age group.

Last, we perform an overall species ranking by evaluating the different classification strategies together. The analyses consider how often the species fall within the first ten species for each ranking strategy. For example, 75% indicates that the species is ranked in the first ten species 75% of the time, i.e., it figures six times over the eight ranking strategies. In this way, we obtain a species classification that considers the abundance of bacterial species in the groups from different statistical points of view. The overall ranking can be viewed as a multicriteria statistical classification strategy that can be helpful when comparing multiple factors or criteria. It is advantageous when the strategies are, in some measure, incommensurable criteria that consider different aspects of the bacterial species occurrence in the samples.

The main limitations of this research relate to the computational complexity of the MC methodology used and the reliance on existing data. First, integrating statistical physics techniques and MC simulations requires significant computational resources, which may limit the scalability of the study to larger datasets. Additionally, the results presented here may be dependent on the representativeness of the available microbiome data, which may change in other geographic regions. Finally, the choice of classification criteria and metrics used could influence the identification of characterizing bacterial species, necessitating further validation to confirm the robustness of the conclusions.

The results presented here can help guide future research in microbial ecology and human health by providing a robust framework for identifying key bacterial species across various contexts, such as the role of the microbiota in health and disease. This approach simplifies and enhances the identification of key bacterial players, improving our ability to analyze and interpret complex microbiome data.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/make6040117/s1, Figure S1: Scatterplots of the Monte Carlo simulated *p*-value.; Table S1: The 10 highest species according to the overall ranking for group G1; Table S2: The 10 highest species according to the overall ranking for group G2; Table S3: The 10 highest species according to the overall ranking for group G3; Table S4: The 10 highest species according to the overall ranking for group G4.

Author Contributions: Conceptualization, M.B., L.M. and D.C.; methodology, M.B., L.M., M.T., R.A. and D.C.; software, M.B., R.A. and M.T.; investigation, M.B., R.A.; data curation, M.B. and L.M.; writing—original draft preparation, M.B., L.M., C.M., G.A.L., M.V. and D.C.; writing—review and editing, M.B. and L.M.; visualization, M.B.; supervision, D.C. and M.V.; project administration, D.C. and M.V.; funding acquisition, M.B., D.C., and M.V.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Ecosister project, funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5—Call for tender No. 3277 of 30/12/2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU Award Number: Project code ECS00000033, Concession Decree No. 1052 of 23/06/2022 adopted by the Italian Ministry. We acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support (project code: code: HP10CZ5SA1). This study was also supported by Fondazione Cariparma as part of the Parma Microbiota project and 'Characterization of the Metabolic Potential of the Human Microbiota in European Populations' project (2023-0555). Part of this research is conducted using the high-performance computing facility of the University of Parma.

Data Availability Statement: No new data were created.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, C.; Jackson, S.A. Machine Learning and Complex Biological Data. *Genome Biol.* **2019**, *20*, 76. [CrossRef] [PubMed]
2. Valdes, A.M.; Walter, J.; Segal, E.; Spector, T.D. Role of the Gut Microbiota in Nutrition and Health. *BMJ* **2018**, *361*, k2179. [CrossRef] [PubMed]
3. Hou, K.; Wu, Z.-X.; Chen, X.-Y.; Wang, J.-Q.; Zhang, D.; Xiao, C.; Zhu, D.; Koya, J.B.; Wei, L.; Li, J.; et al. Microbiota in Health and Diseases. *Signal Transduct. Target. Ther.* **2022**, *7*, 135. [CrossRef] [PubMed]
4. Rooks, M.G.; Garrett, W.S. Gut Microbiota, Metabolites and Host Immunity. *Nat. Rev. Immunol.* **2016**, *16*, 341–352. [CrossRef]
5. Maciel-Fiuza, M.F.; Muller, G.C.; Campos, D.M.S.; do Socorro Silva Costa, P.; Peruzzo, J.; Bonamigo, R.R.; Veit, T.; Vianna, F.S.L. Role of Gut Microbiota in Infectious and Inflammatory Diseases. *Front. Microbiol.* **2023**, *14*, 1098386. [CrossRef]
6. Milani, C.; Ticinesi, A.; Gerritsen, J.; Nouvenne, A.; Andrea Lugli, G.; Mancabelli, L.; Turroni, F.; Duranti, S.; Mangifesta, M.; Viappiani, A.; et al. Gut Microbiota Composition and Clostridium Difficile Infection in Hospitalized Elderly Individuals: A Metagenomic Study. *Sci. Rep.* **2016**, *6*, 25945. [CrossRef]
7. Mancabelli, L.; Milani, C.; Lugli, G.A.; Turroni, F.; Mangifesta, M.; Viappiani, A.; Ticinesi, A.; Nouvenne, A.; Meschi, T.; Van Sinderen, D.; et al. Unveiling the Gut Microbiota Composition and Functionality Associated with Constipation through Metagenomic Analyses. *Sci. Rep.* **2017**, *7*, 9879. [CrossRef]
8. Wensel, C.R.; Pluznick, J.L.; Salzberg, S.L.; Sears, C.L. Next-Generation Sequencing: Insights to Advance Clinical Investigations of the Microbiome. *J. Clin. Investig.* **2022**, *132*, e154944. [CrossRef]
9. Gao, B.; Chi, L.; Zhu, Y.; Shi, X.; Tu, P.; Li, B.; Yin, J.; Gao, N.; Shen, W.; Schnabl, B. An Introduction to next Generation Sequencing Bioinformatic Analysis in Gut Microbiome Studies. *Biomolecules* **2021**, *11*, 530. [CrossRef]
10. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*, 2nd ed.; Springer texts in statistics; Springer New York: New York, NY, USA, 2004; ISBN 9781475741452/1475741456.
11. Manly, B.F.J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; ISBN 9781315273075.
12. Montepietra, D.; Bellingeri, M.; Ross, A.M.; Scotognella, F.; Cassi, D. Modelling Photosystem i as a Complex Interacting Network: Modelling the Photosynthetic System i as Complex Interacting Network. *J. R. Soc. Interface* **2020**, *17*, 20200813. [CrossRef]
13. Soldaat, L.L.; Pannekoek, J.; Verweij, R.J.T.; van Turnhout, C.A.M.; van Strien, A.J. A Monte Carlo Method to Account for Sampling Error in Multi-Species Indicators. *Ecol. Indic.* **2017**, *81*, 340–347. [CrossRef]
14. Newman, M.E.J.; Ziff, R.M. Efficient Monte Carlo Algorithm and High-Precision Results for Percolation. *Phys. Rev. Lett.* **2000**, *85*, 4104–4107. [CrossRef] [PubMed]
15. Nizam, N.I.; Ochoa, M.; Smith, J.T.; Gao, S.; Intes, X. Monte Carlo-Based Data Generation for Efficient Deep Learning Reconstruction of Macroscopic Diffuse Optical Tomography and Topography Applications. *J. Biomed. Opt.* **2022**, *27*, 083016. [CrossRef] [PubMed]
16. Huang, K. *Statistical Mechanics*, 2nd ed.; Wiley India Pvt. Limited: Hoboken, NJ, USA, 2008.
17. O'reilly, C.; Mills, S.; Rea, M.C.; Lavelle, A.; Ghosh, S.; Hill, C.; Ross, R.P. Interplay between Inflammatory Bowel Disease Therapeutics and the Gut Microbiome Reveals Opportunities for Novel Treatment Approaches. *Microbiome Res. Rep.* **2023**, *2*, 35. [CrossRef]
18. Ruiz-Saavedra, S.; Zapico, A.; González, S.; Salazar, N.; de los Reyes-Gavilán, C.G. Role of the Intestinal Microbiota and Diet in the Onset and Progression of Colorectal and Breast Cancers and the Interconnection between Both Types of Tumours. *Microbiome Res. Rep.* **2024**, *3*, 6. [CrossRef]
19. Chen, A.T.; Wu, X.; Ye, G.; Li, W. Editorial: Machine Learning and Deep Learning Applications in Pathogenic Microbiome Research. *Front. Cell Infect. Microbiol.* **2024**, *14*, 1429197. [CrossRef]
20. Jiang, L.; Liu, X.; He, X.; Jin, Y.; Cao, Y.; Zhan, X.; Griffin, C.H.; Gragnoli, C.; Wu, R. A Behavioral Model for Mapping the Genetic Architecture of Gut-Microbiota Networks. *Gut Microbes* **2021**, *13*, 1820847. [CrossRef]

21. Mancabelli, L.; Milani, C.; De Biase, R.; Bocchio, F.; Fontana, F.; Lugli, G.A.; Alessandri, G.; Tarracchini, C.; Viappiani, A.; De Conto, F.; et al. Taxonomic and Metabolic Development of the Human Gut Microbiome across Life Stages: A Worldwide Metagenomic Investigation. *mSystems* **2024**, *9*, e0129423. [CrossRef]
22. Milani, C.; Lugli, G.A.; Fontana, F.; Mancabelli, L.; Alessandri, G.; Longhi, G.; Anzalone, R.; Viappiani, A.; Turroni, F.; van Sinderen, D.; et al. METAnnotatorX2: A Comprehensive Tool for Deep and Shallow Metagenomic Data Set Analyses. *mSystems* **2021**, *6*, e0058321. [CrossRef]
23. Bull, F.C.; Al-Ansari, S.S.; Biddle, S.; Borodulin, K.; Buman, M.P.; Cardon, G.; Carty, C.; Chaput, J.-P.; Chastin, S.; Chou, R.; et al. World Health Organization 2020 Guidelines on Physical Activity and Sedentary Behaviour. *Br. J. Sports Med.* **2020**, *54*, 1451–1462. [CrossRef]
24. Lugli, G.A.; Mancabelli, L.; Milani, C.; Fontana, F.; Tarracchini, C.; Alessandri, G.; van Sinderen, D.; Turroni, F.; Ventura, M. Comprehensive Insights from Composition to Functional Microbe-Based Biodiversity of the Infant Human Gut Microbiota. *NPJ Biofilms Microbiomes* **2023**, *9*, 25. [CrossRef] [PubMed]
25. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
26. Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
27. Rubinstein, R.Y.; Kroese, D.P. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2004; ISBN 978-0-387-20815-5.
28. Barbu, A.; Zhu, S.-C. *Monte Carlo Methods*; Textbook; Springer: Berlin/Heidelberg, Germany, 2020.
29. Pearson, K. VII. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. Lond.* **1895**, *58*, 240–242. [CrossRef]
30. Lozupone, C.A.; Stombaugh, J.I.; Gordon, J.I.; Jansson, J.K.; Knight, R. Diversity, Stability and Resilience of the Human Gut Microbiota. *Nature* **2012**, *489*, 220–230. [CrossRef]
31. Consortium, H.M.P. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* **2012**, *486*, 207–214. [CrossRef]
32. Watson, P.F.; Petrie, A. Method Agreement Analysis: A Review of Correct Methodology. *Theriogenology* **2010**, *73*, 1167–1179. [CrossRef]
33. Ranganathan, P.; Pramesh, C.; Aggarwal, R. Common Pitfalls in Statistical Analysis: Measures of Agreement. *Perspect. Clin. Res.* **2017**, *8*, 187–191. [CrossRef]