BMC Bioinformatics

RESEARCH

Open Access

TaxaPLN: a taxonomy-aware augmentation strategy for microbiome-trait classification including metadata

Alexandre Chaussard[1*], Anna Bonnet[1], Sylvain Le Corff[1] and Harry Sokol[2,3]

*Correspondence:
Alexandre Chaussard
alexandre.chaussard@sorbonne-universite.fr
[1]Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Sorbonne Université, Université Paris Cité, CNRS, F-75005 Paris, France
[2]Centre de Recherche Saint-Antoine, CRSA, AP-HP, Sorbonne Université, INSERM UMRS-938, F-75012 Paris, France
[3]Gut, Liver & Microbiome Research, FHU, Paris, France

Abstract

Background The gut microbiome plays a crucial role in human health, making it a cornerstone of modern biomedical research. To study its structure and dynamics, machine learning models are increasingly used to identify key microbial patterns associated with disease and environmental factors, but their performance is often limited by the intrinsic complexity of microbiome data and the small size of available cohorts. In this context, data augmentation has emerged as a promising strategy to overcome these challenges by generating artificial microbiome profiles.

Results We introduce TaxaPLN, a data augmentation method based on PLN-Tree generative models, which leverages the taxonomy and a data-driven sampler to generate realistic synthetic microbiome compositions. Additionally, we propose a conditional extension based on feature-wise linear modulation, enabling covariate-aware generation. Experiments on diverse curated microbiome datasets show that TaxaPLN preserves ecological properties and generally improves or maintains predictive performances, outperforming state-of-the-art baselines on most tasks. Furthermore, the conditional variant of TaxaPLN establishes a new benchmark for metadata-aware microbiome augmentation.

Conclusion TaxaPLN provides a model-based framework for augmenting microbiome datasets while preserving their ecological and clinical relevance. By integrating taxonomic structure and host metadata, it enhances predictive modeling across diverse real-world settings. To facilitate reproducible and scalable microbiome analysis using our method, TaxaPLN is released as an open-source Python package available on PyPI (`plntree`), with MIT-licensed source code hosted at https://github.com/AlexandreChaussard/PLNTree-package.

Keywords Data augmentation, Microbiology, Generative model, Variational inference

Background

The human gut microbiome is a highly complex and diverse ecosystem composed of bacteria, archaea, fungi, and viruses living in our gastrointestinal tract. Its composition is determined via high-throughput sequencing technologies [1], and varies depending on numerous factors such as host genetics, diet, environmental exposure, lifestyle

habits, and health condition [2]. Despite the unique signature of each individual's microbiome, growing evidence suggests that microbial composition can serve as a biomarker for various diseases, including inflammatory bowel disease (IBD), type 1 diabetes (T1D), colorectal cancer (CRC), or psychiatric disorders such as schizophrenia [3–6]. These insights have led to the integration of gut microbiome data into machine learning models for disease classification, prognosis, biomarker discovery, and bacterial interaction network inference [7, 8]. However, leveraging microbiome data for predictive modeling remains challenging due to host-specific variability, high dimensionality, sparsity, and compositional constraints which raise challenges for model generalization [9]. Additionally, thousands of microbial species inhabit the gut microbiome, often distributed in vastly different proportions across individuals. This inherent complexity aggravates the risk of overfitting in machine learning models, particularly when the cohort size is limited, yielding poor generalization.

To mitigate these challenges, data augmentation has emerged as a promising approach to enrich training sets with synthetic samples [10]. Adapting traditional techniques such as synthetic minority oversampling (SMOTE) [11], Mixup [12], and generative adversarial networks (GAN), recent methods have focused on microbiome-specific augmentation, notably Compositional Cutmix and Aitchison Mixup [13] which tackle the compositional nature of microbiome data, or TADA [14], PhyloMix [15] and PhylaGAN [16] which incorporate taxonomic and phylogenetic information in microbiome generation. However, none of these augmentation methods explicitly support the integration of clinical metadata into the generation process, preventing the association of relevant covariates with synthetic microbiome samples (e.g., age, BMI, sex). This limitation ultimately restricts their use in predictive analyses that rely on both microbiome and metadata information, which is a common setting in clinical applications.

Following these advances, we introduce TaxaPLN, a novel data augmentation strategy tailored to microbiome data that aims at improving predictive analysis. In contrast with prior approaches, TaxaPLN is a variational autoencoder (VAE) based on the Poisson Log-Normal Tree (PLN-Tree) generative model [17]. By incorporating the microbial taxonomy as a structured prior on taxa distribution, TaxaPLN generates biologically faithful synthetic microbiome profiles using a post-hoc variational mixture of posteriors [18] (VAMP) sampler. Additionally, TaxaPLN incorporates clinical metadata using a conditional deep architecture based on Feature-wise Linear Modulation (FiLM) [19], enabling the generation of microbiome samples that are associated with patient-level metadata, thus allowing covariate-aware data augmentation in microbiome studies.

We conduct a comprehensive evaluation of TaxaPLN on nine high-quality curated datasets from the *curatedMetagenomicData* database [20]. First, we evaluate the fidelity of TaxaPLN-generated samples regarding the original cohorts using standard ecology diversity metrics [21], asserting valid downstream analysis. We then assess the practical utility of TaxaPLN in supervised learning tasks against several baseline augmentation methods. Our benchmark includes four widely used classifiers in microbiome research: Logistic Regression, Random Forest, XGBoost, and deep neural networks. Overall, we find that TaxaPLN preserves key biological properties of the cohorts while enhancing predictive performance, with notable gains on non-linear classifiers against baseline methods. In particular, TaxaPLN outperforms the standard PLN baseline, showing the interest of the taxonomy in microbiome modeling. Furthermore, performance increases

in covariate-aware data augmentation establish a first baseline in metadata-aware microbiome predictive analysis.

Methods

Framework

Consider a supervised learning setting with n microbiome samples. Let $\mathrm{X} \in \mathbb{R}_{+}^{n \times K_L}$ the abundance matrix for the K_L taxa observed at the finest taxonomic level L, such that the i-th row $\mathrm{X}_i \in \mathbb{R}_{+}^{K_L}$ indicates the abundance of each taxon in the i-th sample. Each sample is paired with a binary or multiclass label $\mathrm{Y} \in \{0, \ldots, M\}^n$ describing the host trait of interest (e.g. disease status), and when available, the abundance can be joined with exogenous covariates $\mathrm{C} \in \mathbb{R}^{n \times d}$ (e.g. age, body mass index, sex) to be used as predictors.

Upon sequencing microbial data, genome-resolved features are aggregated along a curated taxonomic hierarchy up to a level L of interest, typically the *genus* in 16 S rRNA or the *species* in shotgun sequencing, offering an increasing resolution on the microbiome features as L grows. Thus, the obtained microbial counts are related through the taxonomy to form taxa-abundance data, establishing hierarchical compositional relationships along the taxonomy, as illustrated in Fig. 1 and further detailed in Supplementary Material–Section A. Although sequencing pipelines output integer count for each taxon, the abundance is often normalized to assert the compositionality property of microbial samples [9], from which counts can be recovered through multinomial sampling.

Given microbial samples X with corresponding traits labels Y, the objective of data augmentation is to generate $\tilde{n}$ synthetic microbiomes $\tilde{\mathrm{X}} \in \mathbb{R}^{\tilde{n} \times K_L}$ and labels $\tilde{\mathrm{Y}} \in \{0, \ldots, M\}^{\tilde{n}}$ such that $\tilde{n} + n = \beta n$, where $\beta \geqslant 1$ is the augmentation ratio. Then, defining the augmented training dataset $\{\mathrm{X}, \tilde{\mathrm{X}}\} \in \mathbb{R}^{(n+\tilde{n}) \times K_L}$ with labels $\{\mathrm{Y}, \tilde{\mathrm{Y}}\} \in \{0, \ldots, M\}^{n+\tilde{n}}$, we aim to improve the performances of a classifier by using

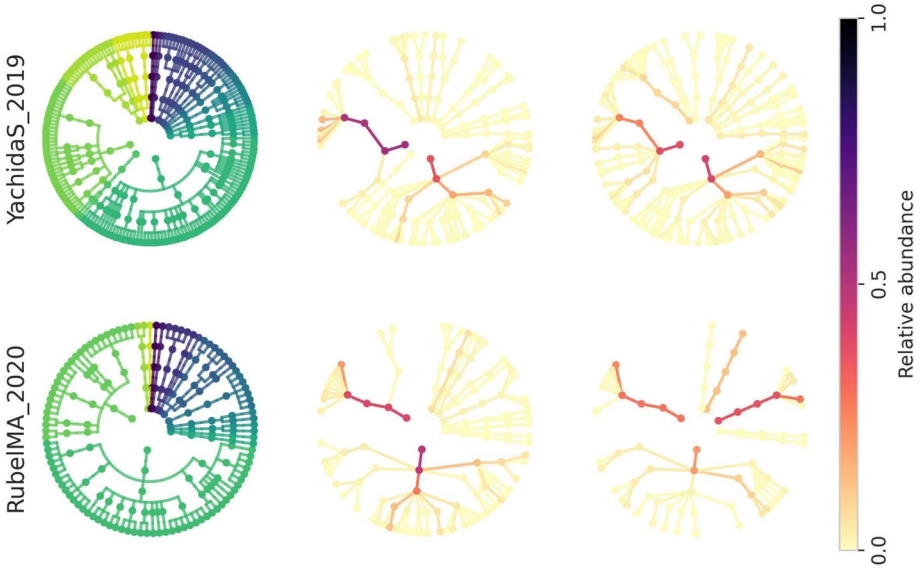

Fig. 1 Taxonomy and taxa-abundance sample illustrations from two studies of the *curatedMetagenomicData* database [20] with prevalence filtering at 15%. First column displays the taxonomic tree used in our experiments for the considered dataset, spanning from the *class* to the *species* level. Second and third columns show taxa-abundance profile for a control and a disease-associated sample respectively. In the taxonomy, colors indicate taxa belonging to the same *class*, coloring in abundance profiles represent relative abundances

the augmented training set over the raw data, evaluating on real microbiomes only. Similarly, conditional data augmentation involves generating covariates $\tilde{\mathrm{C}} \in \mathbb{R}^{\tilde{n} \times d}$, yielding the augmented training set $\{\mathrm{X}, \tilde{\mathrm{X}}, \mathrm{C}, \tilde{\mathrm{C}}\}$ with labels $\{\mathrm{Y}, \tilde{\mathrm{Y}}\}$.

TaxaPLN generative models

Poisson Log-Normal (PLN) models have become a standard approach for modeling count data in ecological studies thanks to their probabilistic framework and ability to capture overdispersion [22]. As specialized variational auto-encoders (VAEs) adapted to abundance data, they offer interpretable latent representations and stable training procedures. These properties make PLN-based models preferable to GANs in clinical applications, as microbiome datasets are often limited in size, while GANs typically require large sample sizes and offer less stability and interpretability.

Building on recent developments that incorporate hierarchical dependencies into PLN models, TaxaPLN extends the Poisson Log-Normal Tree (PLN-Tree) framework [17] to enable metadata-aware microbiome generation while leveraging taxonomic structures. At its core, TaxaPLN preserves the hierarchical probabilistic prior of PLN-Tree, and uses the taxonomy as a guide: latent microbial abundances are propagated through the taxonomic levels via an unconstrained Gaussian Markov chain, and observed counts are generated through parent-sum constrained PLN blocks within taxonomic clades, preserving the compositional nature of taxa-abundance microbiome data. Formally, given covariates C, the microbial counts X and their latent representations Z are modeled such that $\{(\mathrm{Z}_i, \mathrm{X}_i, \mathrm{C}_i)\}_{1 \leqslant i \leqslant n}$ are independent and distributed according to the following hierarchical process.

- The latent Gaussian Markov process

$$\begin{aligned} & \mathrm{Z}_i^1 \sim \mathcal{N}\left(\mathrm{C}_i \mathrm{B}, \boldsymbol{\Sigma}_1\right) , \\ \forall \ell < L, \quad & \mathrm{Z}_i^{\ell+1} \mid \mathrm{Z}_i^\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(\mathrm{Z}_i^\ell, \mathrm{C}_i), \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}(\mathrm{Z}_i^\ell, \mathrm{C}_i)\right) , \end{aligned} \tag{1}$$

is such that $\mathrm{Z}_i^\ell \in \mathbb{R}^{K_\ell}$ where K_ℓ is the number of taxa at the level ℓ of the taxonomy, $\mathrm{B} \in \mathbb{R}^{d \times K_1}$ are the metadata weights at the first taxonomic level, and $\boldsymbol{\mu}_{\boldsymbol{\theta}_\ell}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_\ell}$ are functions parameterized by $\boldsymbol{\theta}_\ell$ outputting respectively the mean and covariance matrix of a Gaussian distribution of dimension K_ℓ, the number of taxa at level ℓ of the taxonomy.

- Conditionally on the latent process, the count emission process is defined as

$$\begin{aligned} & \mathrm{X}_{ik}^1 \mid \mathrm{Z}_{ik}^1 \sim \mathcal{P}(\exp\left(\mathrm{Z}_{ik}^1\right)) , \\ \forall \ell < L, k \leqslant K_\ell, \quad & \check{\mathrm{X}}_{ik}^\ell \mid \mathrm{X}_{ik}^\ell, \check{\mathrm{Z}}_{ik}^\ell \sim \mathcal{M}\left(\mathrm{X}_{ik}^\ell, \sigma(\check{\mathrm{Z}}_{ik}^\ell)\right) , \end{aligned} \tag{2}$$

such that $\sigma(x) = \{\exp\left(x_k\right) / \sum_{j=1}^d \exp\left(x_j\right)\}_{1 \leqslant k \leqslant d}$ is the softmax transform, $\mathcal{M}\left(N, p\right)$ is the multinomial distribution with total count N and probabilities p, and for any random variable V, $\check{\mathrm{V}}_{ik}^\ell$ refers to the vector of children variables of node k at layer ℓ for sample i along the taxonomy.

In particular, TaxaPLN inherits the identifiability guarantees established by Chaussard et al. [17], ensuring that the model's parameters and latent representations admit meaningful interpretation, which is a crucial feature in clinical applications where biological insight and reproducibility are essential.

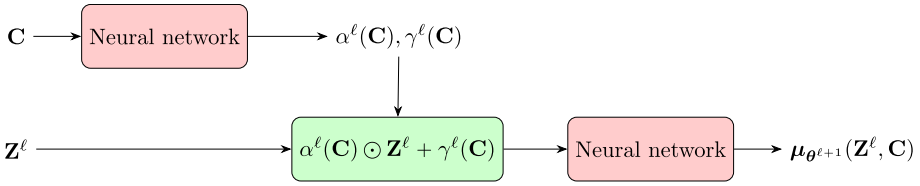

Fig. 2 Feature-wise linear modulation architecture applied to the latent mean parameter from the prior distribution of TaxaPLN at level $\ell < L$. A similar architecture can be derived for the covariance parameters

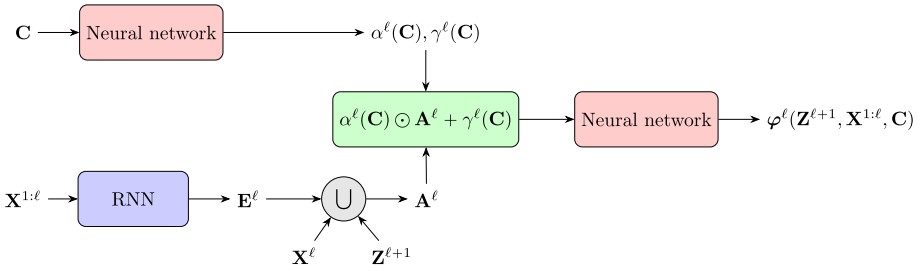

Fig. 3 Feature-wise Linear Modulation applied to the residual amortized backward architecture of TaxaPLN, illustrated for the variational parameters $\boldsymbol{\varphi}^\ell$ at layer $\ell \leqslant L$. The amortizing Recurrent Neural Network is denoted by RNN, while the symbol "U" indicates a concatenation of entries. We denote the variable $\mathrm{X}^{1:\ell} = (\mathrm{X}^1, \ldots, \mathrm{X}^\ell)$ and E^ℓ the last output of the recurrent network after inputting the sequence of microbial compositions at each taxonomic level $\mathrm{X}^{1:\ell}$

Additionally, a key novelty of TaxaPLN is the integration of clinical metadata in the dynamic, enabling covariate-aware microbiome generation through feature-wise linear modulation (FiLM) [19]. For a given sample i and taxonomic level ℓ, the FiLM module takes the metadata vector $\mathrm{C}_i \in \mathbb{R}^d$ as input to two neural networks, α^ℓ, γ^ℓ, that return two vectors with dimensionality matching the regular input A_i in the original architectures of Chaussard et al. [17]. The FiLM block then applies a feature-wise affine transformation

$$\mathrm{A}_i^{\mathrm{FiLM}} = \alpha^\ell(\mathrm{C}_i) \odot \mathrm{A}_i + \gamma^\ell(\mathrm{C}_i) \,, \tag{3}$$

where $\odot$ is the Hadamard product. Fig. 2 depicts the FiLM insertions into the mean parameter of the TaxaPLN prior dynamic. Empirically, FiLM has shown similar benefits as attention mechanisms while being much less parameter-heavy, which can be critical in microbiome studies where n is often small compared to the amount of taxa.

Parameter inference for TaxaPLN models is carried out within the conditional variational framework described in [17]. This approach relies on the optimization of a surrogate objective known as the evidence lower bound (ELBO), which entails a variational proxy $q_{\boldsymbol{\varphi}}(\mathrm{Z} \mid \mathrm{X}, \mathrm{C})$ of the model's posterior $p_{\boldsymbol{\theta}}(\mathrm{Z} \mid \mathrm{X}, \mathrm{C})$. For TaxaPLN models, we extend the specialized residual backward amortized variational approximation of Chaussard et al. [17] to encode the taxonomic structure of microbial data while including metadata through the FiLM module. An illustration of the encoding parameters is provided in Fig. 3. During training, model fitness to the original data is then monitored by the ELBO maximization, where higher values indicate a better fit.

Generating microbiomes with TaxaPLN

Motivated by non-parametric augmentation methods that condition synthetic sample generation on observed data, TaxaPLN employs a Variational Mixture of Posteriors

(VAMP) prior [18] as a post-training sampler. The VAMP prior defines a data-driven latent distribution by forming a uniform mixture of the variational posteriors using microbiome samples from the training set. Formally, if the variational approximation of the posterior is $q_{\boldsymbol{\varphi}}(\mathrm{Z}|\mathrm{X}, \mathrm{C})$, parameterized by $\boldsymbol{\varphi}$, the VAMP distribution is given by

$$\bar{p}_{\boldsymbol{\varphi}}(\mathrm{Z}) = \frac{1}{n}\sum_{i=1}^{n} q_{\boldsymbol{\varphi}}(\mathrm{Z} \mid \mathrm{X}_i, \mathrm{C}_i) \ . \tag{4}$$

This methodology enables the generation of synthetic samples that are generally more representative of the training data than samples from the prior [18, 23], thereby emphasizing the biological integrity characterizing the cohort in microbiome analysis. Following the VAMP prior definition, the generative procedure consists in drawing uniformly a microbiome sample from the training set, then encoding it using the variational sampler to obtain its latent representation, before decoding the latent representation into a new sample using the emission distribution. To assert the practical benefit of our approach, Supplementary Material–Section C provides a generative benchmark of TaxaPLN against the traditional PLN-Tree prior. Results notably highlight the faithfulness of TaxaPLN generations to the original microbiome cohorts, aligning with previous findings in other areas of research [18, 23].

Metagenomic datasets

The *curatedMetagenomicData* package is an extensive collection of human microbiome shotgun metagenomic data from different body sites that has been standardized and curated from 93 published studies [20], thus facilitating comparative metagenomics research. Focusing on stool samples, it includes over 15,000 taxa-abundance samples in relative form sequenced with MetaPhlAn3 and annotated up to the species level, as well as rich annotations such as the age category, the sex, the geographical location, and the health status of the host. This vast database integrates studies covering a variety of conditions, including inflammatory bowel disease (IBD), type 1 diabetes (T1D), adenoma, schizophrenia, colorectal cancer (CRC), soil-transmitted helminth infection (STH), and other microbiome-related disorders.

Using this package, we conduct a comprehensive evaluation of TaxaPLN augmentation on 9 studies spanning several conditions and displaying varying amount of samples and imbalance. To mitigate overfitting and reduce noise from spurious taxa which are critical concerns in low-sample-size microbial studies, we perform prevalence filtering at the *species* level, retaining only those taxa present in at least 15% of the samples. The datasets composition and properties are provided in Table 1.

The resulting set of taxa annotations enable to determine the datasets' associated taxonomy which hierarchically organize the taxa-abundance samples. An illustration on two example datasets is provided in Fig. 1. Additionally, to extend the applicability of count-based models to relative-form microbial data, we convert normalized proportions into count data by sampling from a multinomial distribution with a total count of 100,000 and probabilities defined by the relative abundances of each sample. The large total count ensures minimal information loss during this transformation. Finally, when covariates are available for a given study, we collect the age, sex, and body mass index

Table 1 Curated metagenomics datasets considered in our experiments, extracted from Pasolli et al. [20] with prevalence filtering at 15%. Diagnosis: diagnosis task, $\mathbf{n}$: number of samples, $\mathbf{p}$: number of featured taxa, $\mathbf{n_d}$: number of samples in the positive diagnosis group, $\mathbf{n_c}$: number of control samples, n_d/n: imbalance ratio towards the positive class

Study ID	$\mathbf{n}$	$\mathbf{p}$	Diagnosis	$\mathbf{n_d}$	$\mathbf{n_c}$	$\mathbf{n_d/n}$	References
WirbelJ_2018	125	189	CRC	60	65	48%	[24]
KosticAD_2015	120	104	T1D	31	89	26%	[4]
RubelMA_2020	175	103	STH	89	86	51%	[25]
ZhuF_2020	171	155	Schizophrenia	90	81	52%	[6]
ZellerG_2014	103	216	Adenoma	42	61	41%	[5]
YachidaS_2019	318	177	Adenoma	67	251	21%	[26]
YuJ_2015	128	198	CRC	74	54	58%	[27]
NielsenHB_2014	396	174	IBD	148	248	37%	[28]
HMP_2019_ibdmdb	1627	103	IBD	1201	426	74%	[3]

(BMI) of the patients for Section–Conditional Augmentation experiments, with feature-specific preprocessings detailed in Supplementary Material–Section F.1.

Evaluation tasks

Before evaluating TaxaPLN on data augmentation tasks, we first assess the realism of the synthetic microbiome data it generates following the diversity benchmark methodology of Chaussard et al. [17]. The goal is to assert whether samples generated by TaxaPLN preserve the biological signature of microbial communities, as quantified by the α-diversity and β-diversity metrics [21]. Resemblance in α-diversity is evaluated using the Shannon and Simpson indices, comparing synthetic and real cohorts via Mann–Whitney U tests and Kolmogorov–Smirnov (KS) divergences. To assess β-diversity similarity, we employ Principal Coordinates Analysis (PCoA) on Aitchison distance and Bray–Curtis dissimilarity, two metrics commonly used in shotgun metagenomics studies.

Following the biological validation, we assess the utility of TaxaPLN augmentation in two supervised learning scenarios: vanilla and conditional. The vanilla setting evaluates TaxaPLN's ability to enhance classification performance through the generation of synthetic microbiome profiles without any exogenous input, relying solely on microbial composition for host trait prediction. In the conditional setting, TaxaPLN is trained using both microbial data and covariates via the FiLM architecture defined in Section–TaxaPLN Generative Models, enabling to generate synthetic microbiome samples conditioned on clinical metadata. This setting aims to improve classification performance when both microbiome composition and covariates are used as features. In both scenarios, the impact of data augmentation is quantified using the Area Under the Precision–Recall Curve (AUPRC), a metric well-suited for balanced and imbalanced datasets [29], as commonly found in Table 1.

Augmentation settings

TaxaPLN augmentation is evaluated on four machine learning models commonly applied in microbial-trait association: Logistic Regression (LR), Multi-Layer Perceptron (MLP), Gradient Boosting classifier (XGBoost), and Random Forest (RF). For all classifiers, we employ the default implementations from `scikit-learn` [30] with parameters detailed in Supplementary Material–Table B3. For each study in Table 1, we adopt a 5-fold cross-validation scheme repeated 25 times with different random seeds to ensure robust performance estimates. Within each five splits of a cross-validation

iteration, we fit a label-specific TaxaPLN model on the training data using the hyperparameters detailed in Section–Practical Analysis of TaxaPLN. For the taxonomy ablation study, we additionally fit two label-specific PLN models, implemented via the `pyPLNmodels` package [31]. The learned label-specific models are then used to augment the training data with a fixed augmentation ratio $\beta = 2$. For all augmentation strategies, microbiome data are first augmented in raw count form and then preprocessed with the centered log-ratio (CLR) transformation [32] prior to classifier training, mitigating compositional constraints and facilitating the use of parametric models (see Supplementary Material–Section E). Predictions are made on each test split of the cross-validation procedure, allowing the computation of one AUPRC value per split. For each iteration of the cross-validation procedure, we average the AUPRC scores across the five test splits, yielding one AUPRC value per iteration. Repeating the cross-validation while changing seed yields 25 AUPRC evaluations per dataset, which serve as the benchmark metric for comparing augmentation strategies. Additionally, Supplementary Material–Section E presents alternative results using proportion data instead of CLR-transformed, assessing the robustness of TaxaPLN to different preprocessing choices.

Baseline methods

TaxaPLN is benchmarked against several augmentation methods, each generating an equivalent number of synthetic microbiome samples for evaluation, both for biological metrics analysis and downstream classification performances. To assess the contribution of the taxonomy-informed prior, we first compare TaxaPLN to a PLN baseline using the VAMP sampler described by Equation (4), with the parameterization introduced by Chiquet et al. [22]. In addition, we compare TaxaPLN to two widely used augmentation strategies: Vanilla Mixup, which generates synthetic samples by forming uniform convex combinations of microbiome compositions [12], and SMOTE with $k = 5$ neighbours, which performs local uniform convex interpolation based on Euclidean distance [11]. We also benchmark TaxaPLN against state-of-the-art microbiome-specific methods, including Compositional CutMix [13], which replaces subparts of one sample with those from another before renormalizing to preserve compositional constraints, and PhyloMix [15], a biologically guided variant of Mixup and CutMix that leverages the taxonomy and phylogeny to inform the partitioning strategy. Since phylogenetic trees are not consistently defined across all taxa, our experiments with PhyloMix are restricted to taxonomic structures available in *curatedMetagenomicData*. Importantly, as none of these mixup-based strategies support the inclusion of covariates, the conditional augmentation experiments focus solely on comparing TaxaPLN and PLN, each paired with its corresponding conditional VAMP sampler. All experiments are conducted using the original implementations provided by the respective authors, with default hyperparameters.

Results

TaxaPLN generates realistic microbiomes

TaxaPLN models, along with PLN baselines, are trained on the metagenomics datasets listed in Table 1 using the 5-fold cross-validation strategy described in the Augmentation Settings section, enabling the generation of synthetic microbiome samples conditioned on class labels. To evaluate generation quality, we select the models trained during the

first cross-validation iteration and sample synthetic microbiome profiles using TaxaPLN and the other baseline augmentation methods considered in our benchmark.

First, we assess the similarity between the generated and original microbiomes using α-diversity metrics, specifically the Shannon and Simpson indices. To evaluate whether the distributions of diversity indices differ significantly, we conduct Mann–Whitney U tests and provide empirical Kolmogorov-Smirnov divergence evaluations. Fig. 4 presents results for a representative study, while complete benchmarks are provided in Supplementary Material–Section D.1.

Our findings indicate that synthetic microbiomes generated by TaxaPLN preserve the Shannon and Simpson α-diversity signatures of the considered cohorts. Similarly, PhyloMix and Compositional Cutmix generate microbiome profiles whose α-diversity metrics are generally not significantly different from those of the original cohort, suggesting that key biological signatures are preserved. Conversely, PLN-generated cohorts consistently differ across all datasets, underscoring the importance of incorporating taxonomic structures in microbiome modeling. Besides, SMOTE and Vanilla Mixup generations yield significant α-diversity shifts across all studies, being also consistently rejected. These results indicate that both PLN, SMOTE, and Vanilla Mixup distort biological signatures of microbiome data, potentially impairing downstream analysis.

Additionally, we assess β-diversity alignments between generated and original microbiome samples using both the Aitchison distance and Bray–Curtis dissimilarity through Principal Coordinates Analysis (PCoA). As shown in Fig. 5, TaxaPLN preserves the β -diversity patterns of the original cohort, whereas other methods often introduce noticeable shifts. In particular, TaxaPLN-generated samples closely align with real samples in the Bray–Curtis projected space, indicating more accurate preservation of the original cohort's ecological structure. In the Aitchison geometry, TaxaPLN further demonstrates the ability to explore the space of microbiome compositions while generalizing the geometric structures observed in the original data, contrary to other methods which tend to distort patterns. Together, these results suggest that TaxaPLN achieves a balance between exploration and ecological fidelity, enabling meaningful augmentation of microbiome cohorts.

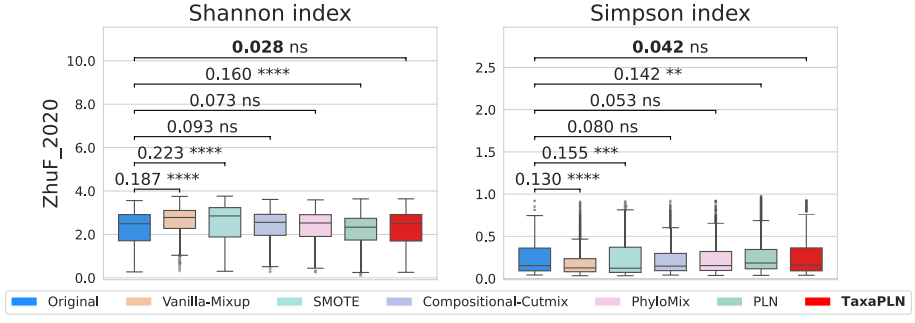

Fig. 4 Shannon and Simpson indices distributions of schizophrenic-labeled microbiomes from *ZhuF_2020*. Each method generates 500 samples. Mann–Whitney U tests are performed to assess the statistical significance of distribution differences between generated and original samples for each method. Significance *P*-values thresholds are denoted by: $^{****}P \leqslant 0.0001$; $^{***}P \leqslant 0.001$; $^{**}P \leqslant 0.01$; $^{*}P \leqslant 0.05$; ns: $P > 0.05$. Kolmogorov-Smirnov divergence between original samples and generated microbiomes is provided with bold value indicating the minimal distance

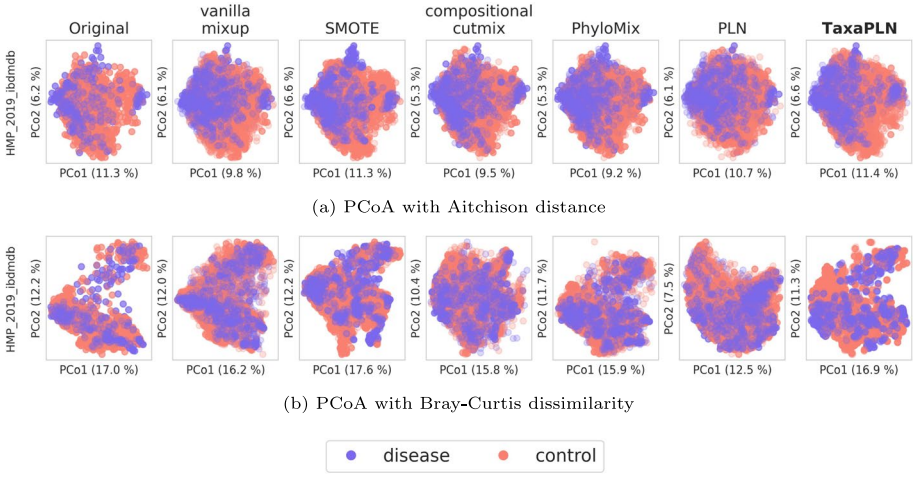

Fig. 5 Aitchison and Bray–Curtis β-diversity PCoA plots of augmented datasets compared to the original cohort of *HMP_2019_ibdmdb*. Each method augments the training fold with $\beta = 2$. Results on all datasets are reported in Additional File 1–Section D.2

Data augmentation

Vanilla augmentation

TaxaPLN augmentation is evaluated on 9 microbiome-trait classification tasks drawn from studies summarized in Table 1. These datasets span a diverse range of microbiome-associated conditions, including inflammatory bowel disease, schizophrenia, type 1 diabetes, adenoma, parasitic infections, and colorectal cancer, thus providing a comprehensive benchmark for assessing our method across varied biomedical contexts. Performance is measured using a 5-fold cross-validation scheme repeated 25 times, as detailed in Section–Augmentation Settings. The resulting AUPRC scores are reported in Fig. 6, along with their 95% confidence intervals. To assess statistical significance, we conduct one-tailed paired t-tests comparing the mean AUPRC of TaxaPLN with that of each baseline augmentation strategy. Relative AUPRC gains to no-augmentation baseline are also provided in Fig. 7.

The augmentation results indicate that TaxaPLN generally enhances or preserves the predictive performance of various classifiers. Specifically, TaxaPLN achieves maximum test AUPRC improvements of 4.1%, 2.6%, 1.6%, and 0.9% when used with MLP, XGBoost, Random Forests, and Logistic Regression classifiers respectively. In comparison, the second-best augmentation method, PhyloMix, achieves maximum gains of 2.1%, 2.0%, 1.0%, and 1.3% with the same classifiers. On average across studies, TaxaPLN ranks first among all augmentation methods when paired with the MLP and XGBoost classifiers, yielding respective mean test AUPRC gains of 0.74% and 0.83% across studies. It also ranks third with Random Forest, close behind Vanilla Mixup and SMOTE, with an average gain of 0.82%, but ranks fifth with Logistic Regression, showing an average gain of 0.02%. As observed for all methods, relative gains are generally modest, since most studies already achieve AUPRC scores around 80% or higher likely due to the high quality and curation of the *curatedMetagenomicData* microbiome database and relative simplicity of the diagnoses tasks [20].

For non-linear models, TaxaPLN leads to performance degradation in only 2 out of 27 experiments, with a maximum AUPRC drop of -2.0%. In contrast, PhyloMix, Compositional Cutmix, SMOTE, and Vanilla Mixup lead to performance drops in 8/27, 10/27,

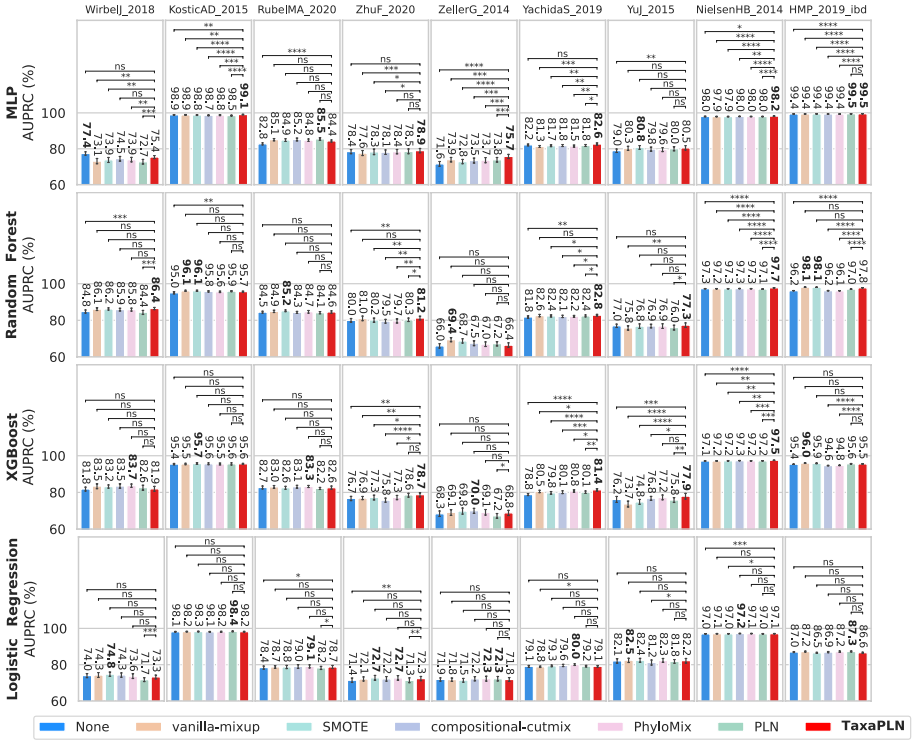

Fig. 6 Vanilla augmentation performances on the supervised learning tasks. The evaluation was performed on real microbiome datasets from Table 1 sequenced at the *species* level. TaxaPLN augmentation is evaluated with five classifiers and compared against five baseline methods according to the Augmentation Settings protocol. TaxaPLN augmentation generally preserves or enhance predictive performances in varying dimensionality and data availability contexts. Performance comparison between TaxaPLN and other methods is assessed using one-tailed paired *t*-tests with significance: ****P-value $\leqslant 0.0001$; ***P-value $\leqslant 0.001$; **P-value $\leqslant 0.01$; *P-value $\leqslant 0.05$; ns: P-value >0.05

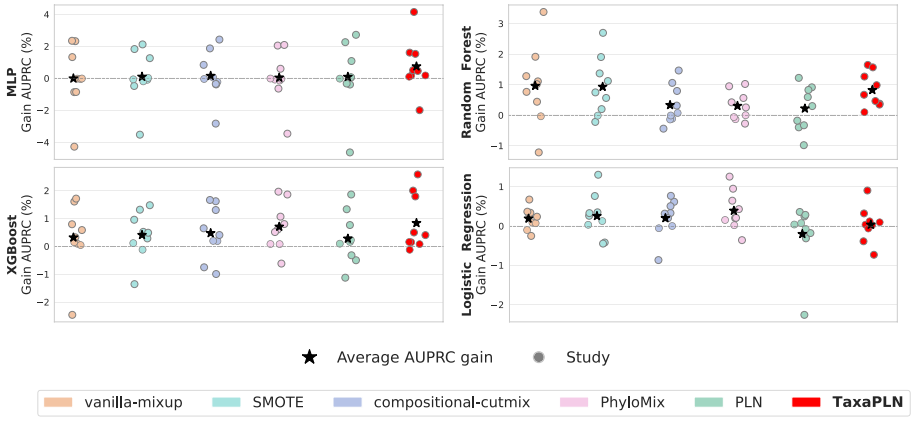

Fig. 7 Average performance gains (ΔAUPRC) of vanilla augmentation methods on supervised learning tasks relative to the no-augmentation baseline. Each point shows the difference between the average AUPRC with and without augmentation for a given study in Table 1. The star symbols denote the mean AUPRC gain across studies for each method. On average, TaxaPLN achieves the highest improvement and generally enhances or maintains performance, particularly when paired with non-linear classifiers

8/27, and 5/27 experiments, with respective maximum declines of -3.5%, -2.8%, -3.5%, and -4.3%. Assessing the significance of TaxaPLN gains over the second best method, PhyloMix, one-tailed paired *t*-tests show that TaxaPLN achieves significantly better performance on 6/9 studies with MLP, 4/9 with Random Forest, and 4/9 with XGBoost. These results suggest that TaxaPLN performs at least as well as PhyloMix with non-linear classifiers, and offers significant improvements in roughly half of the tested scenarios.

With Logistic Regression, TaxaPLN augmentation yields limited improvements, with statistically significant gains over the no-augmentation baseline observed in only 3 out of 9 datasets. Although other augmentation methods can further improve linear classification by up to 1.3%, Fig. 7 shows that all augmentation strategies generally offer much more modest benefits in this setting compared to non-linear classifiers. For TaxaPLN, this behavior can be interpreted through the PCoA plots in Aitchison space, which correspond to principal component analysis on CLR-transformed microbial abundances. As shown in Fig. 5, TaxaPLN augmentation preserves local structures in the CLR space, introducing fine-scale diversity that linear models are typically unable to exploit. While this limits the effectiveness of TaxaPLN in linear settings, the observed behavior remains consistent with current microbiome analysis practices, which predominantly rely on non-linear classifiers such as Random Forest, XGBoost, or deep neural networks.

Additionally, we observe that TaxaPLN generally performs better than PLN, with significantly better results in 18 out of the 36 experiments. Taken together with the biological conservation demonstrated by TaxaPLN relative to PLN, these findings highlight the benefit of incorporating taxonomic structure to generate synthetic microbiome data that are both biologically relevant and effective for data augmentation, as previously observed with PhyloMix.

Metadata-aware augmentation

Incorporating host-related clinical metadata into microbiome-trait classification allows to investigate how environmental and physiological factors influence microbial communities [33, 34], while potentially enhancing predictive performance. Unlike non-parametric augmentation methods in the baseline, TaxaPLN augmentation provides a probabilistic framework that can naturally accommodate exogenous information via conditional sampling, as detailed in Section–TaxaPLN Generative Models.

We evaluate TaxaPLN's conditional augmentation on 7 out of the 9 datasets from Table 1 that all include clinical information on patients' age, sex, and body mass index (BMI). Prior to model training, these covariates are preprocessed according to Supplementary Material–Section F.1. Performance evaluation follows the 5-fold cross-validation protocol Section–Augmentation Settings. AUPRC performances are reported in Fig. 8. The comparative baseline is limited to PLN, as no other augmentation strategy explicitly supports the incorporation of covariates into microbiome generation to the best of our knowledge. Moreover, the set of benchmarked classifiers is restricted to Random Forest and XGBoost, as these models naturally accommodate the heterogeneous data types resulting from the concatenation of covariates and microbiome compositions.

The results show that conditional TaxaPLN augmentation consistently enhances or preserves the predictive performance of both Random Forest and XGBoost classifiers. Specifically, TaxaPLN achieves maximum test AUPRC improvements of 2.2% and

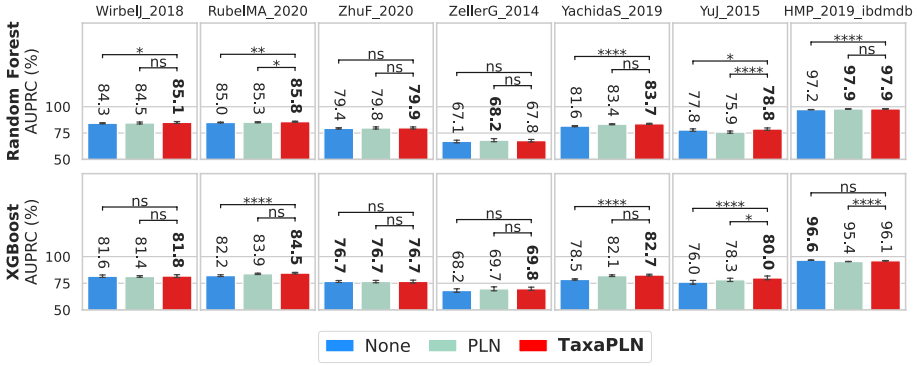

Fig. 8 Conditional augmentation performances on the supervised learning tasks using covariates from Supplementary Material–Section F.1. Performance comparison is assessed using one-tailed paired *t*-tests with significance: ****P-value $\leqslant 0.0001$; ***P-value $\leqslant 0.001$; **P-value $\leqslant 0.01$; *P-value $\leqslant 0.05$; ns: P-value >0.05

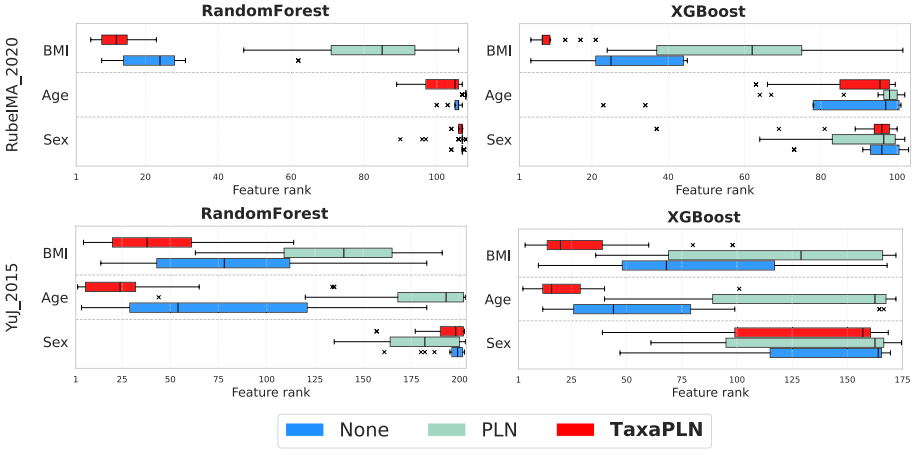

Fig. 9 Metadata ranking across features for covariate-aware microbiome classification, based on feature importance from Random Forest and XGBoost using Mean Decrease in Impurity (MDI). For the 25 cross-validation iterations, we compute the average impurity-based feature importance across folds and report the ranking of metadata features under TaxaPLN augmentation, PLN augmentation, and no augmentation. A lower rank indicates higher model-attributed importance for diagnosis. Full results are reported in Additional File 1–Section F.3

4.1% on Random Forest and XGBoost respectively, with corresponding average gains of 0.9% and 1.7%. In comparison, the PLN baseline yields average AUPRC improvements of 0.4% on Random Forest and 1.1% on XGBoost, with more frequent negative impact than TaxaPLN. Statistical analysis using one-tailed paired *t*-tests indicates that TaxaPLN significantly outperforms the baseline in over half of the experiments, and in four cases for PLN. A generative benchmark comparing synthetic microbiomes produced by TaxaPLN and PLN in Supplementary Material–Section F.2 further compliments these findings by showing that TaxaPLN preserves biological statistics of the data while PLN exhibits significant biological shifts.

To assess how augmentation affects the role of clinical covariates in microbiome–trait classification, we evaluate the importance of microbiome and covariate features using the mean decrease in impurity (MDI) from fitted tree-based predictors in Fig. 9. Across studies and classifiers, TaxaPLN augmentation generally increases the relative importance attributed to clinical metadata compared to the no-augmentation baseline, whereas PLN tends to reduce their importance. Although the magnitude of these

effects varies across datasets and classifiers, the BMI consistently emerges as one of the most informative clinical information, while the importance of age and sex varies across studies. Notably, we observe that age contributes comparatively more to colorectal cancer diagnosis in the *YuJ_2015* cohort than to STH parasitic infection diagnosis in *RubelMA_2020*, which is consistent with prior medical evidence in both studies [25, 27]. These results suggest that TaxaPLN augmentation enhances the relative importance of clinically relevant metadata for microbiome–trait classification tasks, whereas PLN tends to reduce their relative importance in the fitted tree-based models.

Taken together, these results demonstrate that TaxaPLN provides an effective and biologically grounded strategy for metadata-aware data augmentation in microbiome studies. By generating synthetic samples that preserve both microbial structure and covariate associations, TaxaPLN enhances downstream classification performance and has the potential to improve our understanding of the relationships between exogenous covariates and microbiome composition.

Practical analysis of TaxaPLN

Model parametrization Selecting a TaxaPLN architecture involves specifying the parameterization of both the latent dynamics and the variational approximation. In our experiments, we adopt a similar deep architecture to that of Chaussard et al. [17] for PLN-Tree by using a linear latent dynamics and a residual backward variational approximation, with two-layer FiLM modules, as detailed in Supplementary Material–Section B.1. Despite not being tailored to individual datasets, this architecture displays consistent performance gains across varied studies, thus highlighting the flexibility of the TaxaPLN framework. Besides, while a comprehensive hyperparameter tuning study is beyond the scope of this work, we note that appropriate architecture selection could further optimize augmentation performances. In particular, smaller datasets would benefit from lower-capacity models with regularization to prevent overfitting, whereas larger datasets could leverage more expressive architectures to capture fine-grained patterns in microbiome composition. Additionally, the number of taxa relative to the cohort size should also inform the choice of deep architecture, in particular in high-dimensional low-sample-size contexts where overparameterization in deep learning modules can hinder generalization. Ultimately, model parameterization and computational power play a central role in the time efficiency of the augmentation pipeline. In our experiments, training TaxaPLN for 10,000 epochs on an NVIDIA TITAN X (12GB) GPU takes on average 16 min, and up to 36 min for larger datasets. As shown in Supplementary Material–Table B2, training time improves with increased computational power, averaging 11 min on an NVIDIA TITAN RTX (24GB) with the same hyperparameter settings. While this introduces a non-negligible overhead compared to non-parametric methods, the cost remains comparable to that of other deep generative models such as MB-GAN [35]. Moreover, as illustrated in Supplementary Material–Fig. B2, ELBO convergence is typically reached before 10,000 epochs, suggesting that early stopping could substantially reduce runtime while maintaining model fitness to the original data.

Augmentation ratio As observed with mixup-based methods, the number of generated microbiome samples has a measurable impact on downstream performances, depending on both the augmentation strategy and the classifier type. In particular, for a given dataset, optimizing the augmentation ratio β through cross-validated grid-search

can improve downstream performance, albeit at the cost of increased computational overhead. In this study, we aim to identify a reasonable default value for β that generalizes well across microbiome datasets and classifiers by analyzing its impact on predictive performance across the curated datasets from Table 1. Fig. 10 reports the average AUPRC performances across the different microbiome studies for several values of β, grouped by classifier type. Results reveal distinct behaviors for each classifier: tree-based models generally benefit from increased β until reaching a performance plateau, MLPs exhibit a peak followed by a decline, and Logistic Regression remains largely insensitive to the augmentation ratio. Overall, $\beta = 2$ emerges as a reasonable compromise across classifiers and datasets in our experiments to improve classification performances with microbiome data.

Cohort size effect We investigate the impact of training set size on the effectiveness of TaxaPLN augmentation. Specifically, we conduct a control experiment using the *HMP_2019_ibdmdb* dataset ($n = 1627$), where subsets of the training data are randomly selected to represent 5% ($n = 81$) to 30% ($n = 488$) of the full cohort, with stratification on diagnosis labels. This procedure is repeated 100 times with different random seeds to assess variability in performance. For each subset, a TaxaPLN model is trained and used for vanilla data augmentation with $\beta = 2$, followed by classification with different models. Performance results are reported in Fig. 11. We observe that although data augmentation on larger training sets generally yields better performances, the relative improvements brought by TaxaPLN are more pronounced when training data are limited, highlighting the practical relevance of TaxaPLN in low-resource settings. Besides, the performance gains are generally comparable to those obtained with PhyloMix, with TaxaPLN achieving superior average performance in 13 out of 16 scenarios, thus underpinning the interest of TaxaPLN for data augmentation across varying sample size settings.

Discussion

As explored in this study, predictive modeling in microbiome research is often hindered by the intrinsic complexity of microbial data and the limited size of cohorts in clinical applications. By generating synthetic microbiome profiles, data augmentation is a promising approach to mitigate these challenges and enhance model robustness and generalization. While previous methods such as SMOTE [11], Vanilla Mixup [12], Compositional Cutmix [13], or PhyloMix [15] have demonstrated competitive performances, their lack of modeling groundings and inability to incorporate patient-level

Fig. 10 Impact of augmentation ratio on average classification performances with TaxaPLN. Averages are obtained from the performances on all microbiome datasets in Table 1. The augmentation ratio β takes value spanning between 1 (no augmentation) and 5

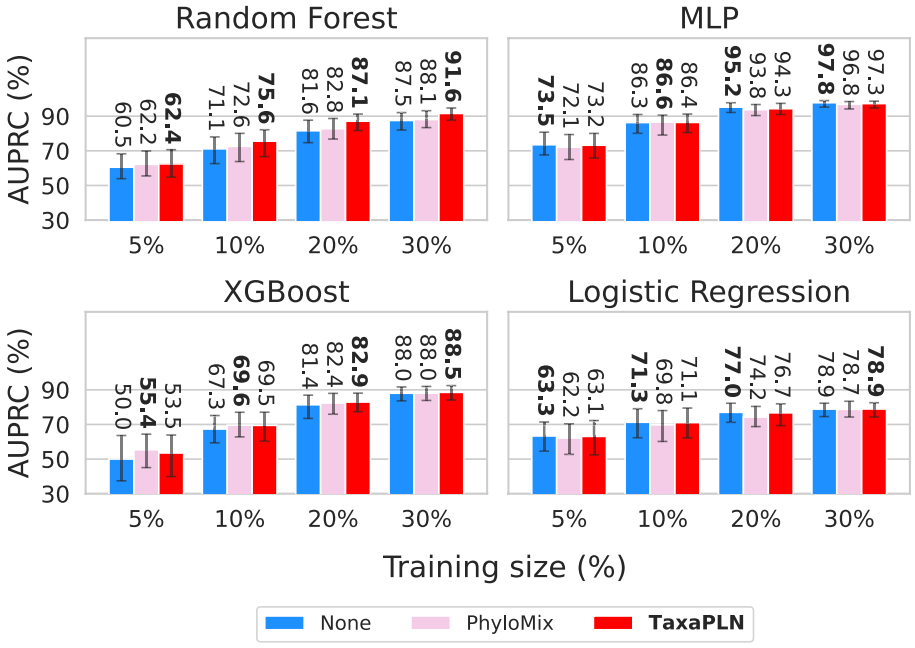

Fig. 11 Effect of augmentation using TaxaPLN and PhyloMix with varying numbers of training samples from the *HMP_2019_ibdmdb* cohort, with augmentation ratio $\beta = 2$. Average performance is computed over 100 hold-out splits at each specified training size, with performance measured through average AUPRC $\pm 95\%$ confidence intervals

metadata ultimately limit their interpretability and applicability in clinical contexts. This study aims at addressing these limitations by introducing TaxaPLN, a model-based data augmentation framework that leverages taxonomic relationships and exogenous metadata to generate synthetic microbiome profiles. Through comprehensive benchmarks on varied curated microbiome datasets [20], we demonstrate that TaxaPLN augmentation not only preserves key ecological statistics of the original cohorts, but also generally improves or maintains performance in classification tasks, notably with non-linear classifiers where it outperforms prior baselines. Moreover, its ability to incorporate clinical metadata in a VAE framework represents, to our knowledge, the first benchmark for data augmentation in covariate-aware microbiome analysis, which is central in clinical research.

While TaxaPLN shows notable empirical performance, our study offers room for future developments. For instance, we adopted a general-purpose architecture and fixed augmentation ratio for consistency across our experiments. However, performance would benefit from dataset-specific hyperparameter tuning, particularly in low-data regimes. Besides, generation quality is sensitive to the fit of the variational approximation, which can lead to limited diversity when model capacity is not tailored to data availability constraints. Finally, although TaxaPLN augmentation works best when paired with current best-performing microbiome classifiers such as Random Forest, it shows limited impact on linear models.

As such, this study opens several promising directions for future work. The generative framework of TaxaPLN offers a broad design space with numerous variants yet to be explored. For instance, implementing β-VAE regularization [36] or early stopping strategies could encourage further exploration by regularizing the variational approximation, namely in low-data regimes. Besides, leveraging the latent space could open

opportunities for advanced sampling strategies, such as latent mixup or Riemannian geodesic interpolation [23], which could also enhance exploration while improving augmentation performances. As a model-based approach, TaxaPLN also offers tools for investigating the biology of the gut microbiome. Exploring structured parameterizations in TaxaPLN, namely sparse or block-diagonal precision matrices, could yield interpretable insights into microbial community structures, revealing interactions that guide synthetic data generation. Similarly, linear FiLM networks could be used to understand the impact of specific covariates on microbial compositions. Finally, automatically tuning the augmentation ratio to maximize downstream performance could further enhance the practical impact of augmentation strategies. Thus, exploring data-driven methods to identify the optimal number of synthetic samples would yield substantial interest for practitioners.

Conclusion

This study presents TaxaPLN, a model-based data augmentation framework for microbiome analysis. By integrating taxonomic relationships and clinical metadata into a generative model, TaxaPLN produces biologically realistic synthetic profiles that enhance predictive modeling across a broad range of real-world classification tasks, including both vanilla and metadata-aware microbiome augmentation scenarios. In particular, its applicability to limited sample size settings and capacity to integrate patient-level metadata align well with the constraints of clinical microbiome studies. To support reproducibility and facilitate its use, we release TaxaPLN as a user-friendly open-source Python package available via PyPI (`plntree`), with full source code and experiments hosted on GitHub.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06312-z.

Supplementary Material.

Acknowledgements
The authors would like to thank the anonymous reviewers and associate editor for their valuable feedback which have contributed to improve the quality of the manuscript.

Author contributions
AC proposed the methodological development under the supervision of SLC and AB while HS supervised the metagenomics aspects. AC developed the code, carried out the experiments, and wrote the first draft of the manuscript. Reviewing and editing have been performed by all authors.

Funding
This work was supported by the Institute of Computing and Data Sciences (ISCD) from Sorbonne Université that funded AC's PhD thesis.

Data availability
The datasets supporting the conclusions of this article are available on the ***curatedMetagenomicData*** database at https://doi.org/doi:10.18129/B9.bioc.curatedMetagenomicData. Pretrained models are available on Zenodo at https://doi.org/10.5281/zenodo.15736784, while the Python source code for our method is accessible on GitHub under MIT license at https://github.com/AlexandreChaussard/PLNTree-package, along with comprehensive installation instructions and tutorials to facilitate reproducibility.

Declarations

Ethics approval and consent to participate
Microbiome data used in this study originate from the publicly available *curatedMetagenomicData* database, which aggregates datasets approved by the respective institutional review boards. No additional ethics approval was required for our work.

Consent for publication
Not applicable.

Conflict of interest
The authors declare no conflict of interest.

Received: 31 July 2025 / Accepted: 27 October 2025
Published online: 28 November 2025

References

1. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.
2. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24(4):392–400.
3. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655–62.
4. Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen AM, et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 2015;17(2):260–73.
5. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10(11):766.
6. Zhu F, Ju Y, Wang W, Wang Q, Guo R, Ma Q, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun. 2020;11(1):1612.
7. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Front Microbiol. 2021;12:634511.
8. Qian W, Stanley KG, Aziz Z, Aziz U, Siciliano SD. SPLANG—a synthetic Poisson-Lognormal-based abundance and network generative model for microbial interaction inference algorithms. Sci Rep. 2024;14(1):25099.
9. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
10. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array. 2022;16:100258.
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
12. Zhang H, Cissé M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. In: ICLR. OpenReview.net; 2018. Vancouver, Canada.
13. Gordon-Rodríguez E, Quinn TP, Cunningham JP. Data Augmentation for Compositional Data: Advancing Predictive Models of the Microbiome. In: NeurIPS; 2022. Los Angeles, USA.
14. Sayyari E, Kawas B, Mirarab S. TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification. Bioinformatics. 2019;35(14):i31–40.
15. Jiang Y, Liao D, Zhu Q, Lu YY. PhyloMix: enhancing microbiome-trait association prediction through phylogeny-mixing augmentation. Bioinformatics. 2025;41(2):btaf014.
16. Sharma D, Lou W, Xu W. phylaGAN: data augmentation through conditional GANs and autoencoders for improving disease prediction accuracy using microbiome data. Bioinformatics. 2024;40(4):btae161.
17. Chaussard A, Bonnet A, Gassiat E, Le Corff S. Tree-based variational inference for Poisson log-normal models. Stat Comput. 2025;35(5):1–35.
18. Tomczak JM, Welling M. VAE with a VampPrior. In: AISTATS. vol. 84. PMLR; 2018. p. 1214–1223.
19. Perez E, Strub F, de Vries H, Dumoulin V, Courville AC. FiLM: Visual Reasoning with a General Conditioning Layer. In: AAAI. AAAI Press; 2018. p. 3942–3951.
20. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. Nat Methods. 2017;14(11):1023–4.
21. Bastiaanssen TF, Quinn TP, Loughman A. Bugs as features (part 1): concepts and foundations for the compositional data analysis of the microbiome-gut-brain axis. Nat Mental Health. 2023;1(12):930–8.
22. Chiquet J, Mariadassou M, Robin S. The Poisson-Lognormal model as a versatile framework for the joint analysis of species abundances. Front Ecol Evol. 2021;9:588292.
23. Chadebec C, Thibeau-Sutre E, Burgos N, Allassonnière S. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. IEEE Trans Pattern Anal Mach Intell. 2022;45(3):2879–96.
24. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25(4):679–89.
25. Rubel MA, Abbas A, Taylor LJ, Connell A, Tanes C, Bittinger K, et al. Lifestyle and the presence of helminths is associated with gut microbiome composition in Cameroonians. Genome Biol. 2020;21:1–32.
26. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med. 2019;25(6):968–76.
27. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, et al. Metagenomic analysis of Faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66(1):70–8.
28. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32(8):822–8.
29. Davis J, Goadrich MH. The relationship between Precision-Recall and ROC curves. In: ICML. vol. 148 of ACM International Conference Proceeding Series. ACM; 2006. p. 233–240.
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.

31. Batardiere B, Kwon J, Chiquet J. pyPLNmodels: a Python package to analyze multivariate high-dimensional count data. J Open Sour Softw. 2024;9(104):6969.
32. Aitchison J. The statistical analysis of compositional data. J Roy Stat Soc: Ser B (Methodol). 1982;44(2):139–60.
33. Wu H, Li Y, Jiang Y, Li X, Wang S, Zhao C, et al. Machine learning prediction of obesity-associated gut microbiota: identifying Bifidobacterium Pseudocatenulatum as a potential therapeutic target. Front Microbiol. 2025;15:1488656.
34. Wu H, Lv B, Zhi L, Shao Y, Liu X, Mitteregger M, et al. Microbiome–metabolome dynamics associated with impaired glucose control and responses to lifestyle changes. Nat Med 2025;p. 1–10.
35. Rong R, Jiang S, Xu L, Xiao G, Xie Y, Liu DJ, et al. MB-GAN: microbiome simulation via generative adversarial network. GigaScience. 2021;10(2):giab005.
36. Kumar A, Poole B. On Implicit Regularization in β-VAEs. In: ICML. vol. 119. PMLR; 2020. p. 5480–5490.

Publisher's Note