Explainability across classical and deep learning models in gut microbiome contexts

Stefan-Tudor Lupsan[1]
Supervisor(s): Thomas Abeel[1]**, Bianca Cosma**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 18, 2026

Name of the student: Stefan-Tudor Lupsan
Final project course: CSE3000 Research Project
Thesis committee: Thomas Abeel, Bianca Cosma, Zekeriya Erkin

Abstract

Machine learning models can separate diseased from healthy patients using gut microbiome data, but the explainability techniques used to justify their predictions were developed for ordinary tabular, image, and text data, not for the compositional, high-dimensional, and sparse abundance tables that microbiome studies produce. We test how far these techniques can be trusted in this setting through a comparative study, applying multiple explainability techniques to multiple models reasoning over three levels of representation: classical models on the raw abundance table, the same models on a foundational model's latent space, and end-to-end deep models. The task is distinguishing inflammatory bowel disease from controls on the IBDMDB dataset. We find three failure modes that a single-model analysis would hide: the way we measure or implement an evaluation metric can influence the metric as much as switching the model under test; accuracy should be read separately from biologically correct reasoning; and an encoder's prior can be a stronger influence on the result than the model built on top of it. These findings recommend that explanations of microbiome models should be considered across multiple ways of measuring evaluation metrics, separated from accuracy, with the representation taken into consideration. Otherwise, any single explanation can display substantial artefacts inherent to its implementation or way of measurement.

1 Introduction

The human gut microbiome consists of trillions of microorganisms inhabiting the digestive tract, where they support digestion and influence neurological and psychiatric processes through what has been labelled the microbiota–gut–brain axis [1]. Disruptions of this ecosystem are implicated in a wide spectrum of conditions, ranging from inflammatory bowel disease (IBD) [2] to metabolic and neurodegenerative disorders.

Machine learning is a tool through which we can seek to approach data which presents complex, multivariate relationships, but microbiome data poses distinctive challenges. Abundance tables produced by 16S rRNA or shotgun metagenomic sequencing are high-dimensional, compositional [3], sparse, noisy, and typically come from cohorts that are small enough to be heavily affected by the curse of dimensionality [4]. Classical models such as random forests and elastic nets remain competitive in this regime because they tolerate sparsity and are quite easily analysed and understood [4]. More recently, foundation models for microbial communities, notably the Microbial General Model (MGM) [5], along with task-agnostic models such as TabPFN [6], have broadened the deep learning toolkit of this domain. However, with new models come new explainability challenges, and explainability for these models has only narrowly been explored [5] in microbiome contexts.

These advances complicate interpretation: when a classifier reasons over a learned latent space rather than over unencoded data, its predictions are no longer naturally attributable to specific taxa. Explainable AI (XAI) offers a range of techniques, from feature importance, SHAP values, integrated gradients, attention-based attributions, to evaluation frameworks for assessing explanation quality [7], [8]. Most of these techniques, however, were developed and validated on non-compositional tabular, image, or text data [9], [10], and their behaviour on microbiome inputs, particularly on models that operate over learned representations rather than raw taxa, has not been systematically compared. This raises two concrete questions: how useful are standard XAI techniques on compositional microbiome data, and what can they actually tell us about how these models reach their decisions?

In this paper we address those questions through a comparative analysis across three levels of representation. We apply common explainability techniques to (i) classical models: Elastic Net, Random Forest, and XGBoost trained on the unencoded abundance table; (ii) the same classical models trained on the latent space produced by the MGM encoder; and (iii) end-to-end deep models, MGM and TabPFN, with a classification head. All experiments use the metagenomics data from the Inflammatory Bowel Disease Multi-omics Database (IBDMDB) [11] and are trained to distinguish between IBD and control patients. The explainability techniques we employ are: Shapley values, incremental insertion/deletion, and integrated gradients. Our contribution is twofold: a side-by-side comparison of XAI behaviour across these representation levels, and an assessment of whether feature attributions remain biologically interpretable once a model reasons over a learned latent space rather than unencoded taxa, namely models in categories (ii) and (iii).

The remainder of this paper is organised as follows. Section 2 describes the IBDMDB dataset, the three representation levels we compare, the models used and the explainability techniques applied. Section 3 reports the resulting attributions and their agreement across models. Section 4 reflects on the reproducibility and ethical considerations of working with clinical microbiome data. Section 5 interprets the findings and situates them against prior work on microbiome XAI, and Section 6 concludes and outlines directions for future work.

2 Methods

We evaluated explainability on models predicting inflammatory bowel disease from gut-microbiome metagenomics, across three different levels of analysis:

- Classical models (Elastic Net, Random Forest, and gradient-boosted trees – XGBoost) trained on the unencoded abundance tables
- The same classical models trained on the latent space produced by the MGM encoder
- End-to-end deep models, MGM and TabPFN, with a classification head.

We performed this at the genus level, as MGM accepts only genus-level input. The analysis pipeline comprised three stages: dataset preprocessing, model training under

participant-grouped cross-validation, and computation of explainability metrics. The subsections below describe each stage and distinct component of this pipeline, including the dataset.

2.1 Dataset and preprocessing

We used the genus-level taxonomic profiles [12] released as part of the Integrative Human Microbiome Project IBD multi-omics study (IBDMDB) [11], restricted to the metagenomics dataset. This dataset follows 130 participants, each having samples taken from them at up to 24 points in time. We chose IBD as the disease because it has been strongly linked to markers in the gut microbiome [2] and because there is an existing medical consensus on leading bacterial markers associated with IBD [13], [14], [15], [16], [17]. This gives us a reference against which to sanity-check the outputs of the explainability metrics computed on the models. Among those, we picked IBDMDB, part of the iHMP/HMP2. We collapsed the three-class diagnosis (Crohn's disease, ulcerative colitis, non-IBD) to a binary IBD vs. non-IBD label, because the binary task yields larger per-class samples, strengthening the claims each model and its explanations make. We considered losing the UC vs. CD distinction less important than the larger cohort and simplified task, as the explainability questions we address concern model behaviour rather than the separation of the two IBD subtypes. The resulting cohort contained 1,638 samples from 130 participants, having 578 bacterial species as features, part of 187 distinct genera. IBD is the majority class, accounting for roughly 73% of the samples against 27% non-IBD; this imbalance is what the balanced class weights and probability calibration used by several of the models below are meant to counter.

Microbiome abundance tables are sparse, high-dimensional, and compositional: counts share a fixed sum imposed by the sequencing platform, so the value of any one feature is constrained by the others [3]. We applied two preprocessing methods in response. First, we removed species present in fewer than 5% of samples, leaving 164 features in species-level analysis, and 77 genera. Second, we transformed the filtered abundance matrix with the centred log-ratio (CLR) to reduce the burden of the compositionality of the data.

2.2 Cross-validation

The IBDMDB cohort contains multiple samples per participant. Treating samples as independent would let a model memorise participant-level signatures and inflate test-set metrics. We therefore evaluated all classifiers with a stratified, participant-grouped K-fold scheme ($K = 10$, shuffled, seeded at zero) that assigns each participant entirely to a single fold while approximately balancing the IBD/non-IBD ratio across folds, so the held-out samples in every fold come from participants the model has never seen during training.

We evaluated the models with nested grouped cross-validation: each of the ten outer folds wrapped a three-fold grouped grid search that selected hyperparameters by log-loss, refit the best configuration on the outer training set, and scored the outer held-out samples. We used three inner folds rather than ten to keep each inner split large enough to stay class-balanced under participant grouping. We additionally refit each model on the full dataset for completeness, but this full-data model is not the source of any reported explanation: we computed every reported attribution and faithfulness curve out-of-fold, as described in Section 2.3.

2.3 Explainability techniques

We applied three explainability techniques to our models: Shapley values, integrated gradients and incremental insertion/deletion. Not every technique applies to every model; the precise explainer and its settings vary by model and are stated in each model's subsection below and in Table 2.

We computed every explainability metric out-of-fold: within each of the ten cross-validation folds described in Section 2.2. A local method produced a samples-by-features matrix per fold, which we collapsed to a per-fold global ranking by taking the mean of the absolute attributions across samples. We then averaged the ten per-fold rankings into a single ranking. This averaged ranking is the one we reported, and it fed the insertion/deletion evaluation.

Shapley values (SHAP). We used whichever SHAP estimator is exact or efficient for each model class: closed-form explainers for the tree ensembles and the linear model, and coalition-sampling estimators for the deep and latent-space models. We always explained the positive (IBD) class. We implemented two different ways of measuring SHAP. The *imputation* variant leaves the trained model untouched and represents the absent feature by a neutral background value, the feature's mean, which due to CLR is 0. The MGM models could not be queried in this way, due to compute restrictions. Instead, we filled this imputation role with integrated gradients for the end-to-end classifier. In the cases where MGM just acts as an encoder, integrated gradients propagate the classifier models' SHAP values from the embedding space back to the original taxa. The *native* variant instead removes the feature from the model's view entirely and re-evaluates it on the reduced feature set, which is only meaningful for models that can be queried on an arbitrary feature subset.

Integrated gradients (IG). For the MGM-based models, which are differentiable, we attributed at the token-embedding layer, took the all-padding sequence as the baseline, and aggregated the resulting per-token attributions to a per-genus score. This metric attributes the positive-class probability directly for the end-to-end MGM classifier, and serves as the second stage of a composed attribution for the frozen-encoder latent models, where we propagate SHAP values back to the input genera through the encoder.

Incremental insertion and deletion. Finally, we evaluated how faithful a given ranking is to the model that produced it with an incremental insertion/deletion sweep. This metric tracks the area under the receiver operating characteristic curve (AUROC) as the top features by their SHAP values are added or removed from the model's view. A faithful ranking yields a steep insertion curve and a steep deletion drop, as it would mean the model actually relies on the reported high-SHAP features to make its decisions. We used two modes for representing a removed feature: a *mask* mode, which replaces it with a neutral baseline value and is available for

Table 1: Curated reference panel of IBD-associated genera used for the biological-plausibility (recall@k) check, with the direction of the reported association and its supporting studies.

Genus	Reported IBD signal	Source studies
Depleted in IBD		
Faecalibacterium	Depleted	[13], [14], [15]
Roseburia	Depleted	[14], [15], [16]
Coprococcus	Depleted	[15], [17]
Lachnospira	Depleted	[15], [16]
Eubacterium	Depleted	[14], [16]
Alistipes	Depleted	[14], [17]
Odoribacter	Depleted	[15], [17]
Gemmiger	Depleted	[15], [17]
Fusicatenibacter	Depleted	[15], [17]
Lawsonibacter	Depleted	[15], [17]
Enriched in IBD		
Escherichia	Enriched	[13], [14], [17]
Klebsiella	Enriched	[14], [17]
Veillonella	Enriched	[13], [15], [17]
Haemophilus	Enriched	[13], [15], [17]

every model; and a *refit* mode, which drops it outright and rebuilds the model on the retained subset, mirroring the native SHAP variant and only possible for TabPFN and MGM. We reported the AUROC-versus-k curves for each model and preprocessing condition.

Biological plausibility (recall@k). The techniques above ask whether an explanation is faithful to the model that produced it; they do not ask whether it has any biological basis. To assess the latter we compared each model's averaged genus ranking against a curated reference panel of genera with an established IBD association (Table 1), scoring the overlap with recall@k: the fraction of the panel genera present in a model's top-k ranked genera. Because recall@k will eventually converge to one for every model, signal lies in how early the curve rises; we therefore summarised each curve by the area between it and the diagonal expected of a random ranking (expected recall@$k = k/n$ over n features), further referred to as *recovery score*. Note that this metric does not distinguish between genera that are expected to be enriched and depleted in IBD.

2.4 Plausibility reference panel

The panel (Table 1) is a purposive, not a systematic, selection of five studies, so it carries our own selection bias and makes no claim to comprehensiveness. Our filter was requiring each genus to appear in at least two of the five studies. We chose studies that were not based on the training dataset, IBDMDB.

We do not read the panel as a definitive list of IBD-causal bacteria. We instead take the union of the genera these studies report as differentially abundant in IBD, with the following filters: (i) required presence in at least two of the studies; (ii) have its direction, observed enrichment or depletion, stay consistent in the studies; (iii) be present in the training dataset.

2.5 Models

We trained five model families spanning the three levels of analysis introduced above: three classical tabular classifiers (elastic-net logistic regression, Random Forest, and XGBoost), the same three classifiers refitted on the MGM latent space, and two deep models (a fine-tuned MGM transformer and TabPFN). The classical trio represents the models commonly reported as competitive baselines on tabular microbiome classification [4], chosen to span fundamentally different ways of combining features. We were not aiming for state-of-the-art accuracy, but for adequate baselines whose response to explainability tests we could analyse.

We trained every model with a fixed random seed of zero and, unless stated otherwise, selected all tunable hyperparameters inside the nested grouped cross-validation described in subsection 2.2. The remainder of this subsection describes each model's caveats and the hyperparameter grid over which we tuned it.

Elastic Net. An ℓ_1/ℓ_2-regularised logistic regression (the `saga` solver, balanced class weights, 20,000 maximum iterations, and a tolerance of 10^{-4}), fitted inside a pipeline that first standardises every feature to zero mean and unit variance. As the only linear model in our set it is the one for which SHAP attributions are exact rather than approximate. We jointly tuned the inverse-regularisation strength $C \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and the L1/L2 mixing ratio $\ell_1 \in \{0, 0.05, 0.1, 0.3, 0.5\}$, a $7 \times 5 = 35$-point grid.

Random Forest. A tree ensemble, where we fixed balanced class weights to counter the IBD/non-IBD imbalance and left the remaining scikit-learn defaults untouched, tuning the number of estimators ($\{300, 500\}$), the maximum tree depth ($\{\text{None}, 20\}$), and the minimum samples per leaf ($\{1, 5\}$) inside the inner cross-validation.

XGBoost. A gradient-boosted tree ensemble using the histogram tree method and the log-loss training objective. We compensated for class imbalance with per-sample balanced weights (computed from the training labels at fit time so the minority class receives a gradient signal proportional to its prevalence). We tuned the number of estimators ($\{50, 100, 200, 500\}$), the maximum tree depth ($\{2, 3, 5\}$), and the learning rate ($\{0.001, 0.003, 0.01, 0.03, 0.05, 0.1\}$).

TabPFN. A pretrained tabular foundation model that performs in-context learning, so it exposes no tunable hyperparameters and we skipped the inner grid search for it altogether. We ran it with `balance_probabilities` enabled, a softmax temperature of 0.9, and a cached fit mode that reuses the encoded context across calls.

MGM. MGM is a microbiome-specific GPT-2 model, entered in two distinct roles. In the *frozen-encoder* role we used the pretrained transformer purely as a feature extractor: we froze its weights, tokenised abundance profiles with MicroTokenizer and `MicroCorpus` (maximum sequence length 512), and pooled the final transformer layer into a per-sample embedding. We then fitted one of the three classical classifiers on those embeddings, reusing the exact grids listed above (with the pipeline step prefix `clf__`). In the *end-to-end*

Table 2: Explainability techniques applied to each model, by representation level. ✓ marks a technique the model supports and that we computed; a dash marks one that does not apply. SHAP_{imp} and SHAP_{nat} are the imputation and native Shapley variants; ID_{mask} and ID_{refit} are the masking and refitting insertion/deletion modes.

Representation	Model	SHAP_{imp}	SHAP_{nat}	IG	ID_{mask}	ID_{refit}
Unencoded abundance	Elastic Net	✓	–	–	✓	–
	Random Forest	✓	–	–	✓	–
	XGBoost	✓	–	–	✓	–
MGM latent space	Elastic Net	✓[a]	✓	✓[a]	✓	✓
	Random Forest	✓[a]	✓	✓[a]	✓	✓
	XGBoost	✓[a]	✓	✓[a]	✓	✓
End-to-end deep	MGM	–	✓	✓	✓	✓
	TabPFN	✓	✓	–	✓	✓

[a] For the frozen-encoder latent models the SHAP_{imp} and IG columns denote a single composed attribution: imputation SHAP on the encoder bottleneck dimensions, propagated to genera by integrated gradients through the encoder. They are not two independent computations.

role we fine-tuned the same backbone with a fresh sequence-classification head for 20 epochs at a learning rate of 10^{-3}, batch size 8, weight decay 10^{-3}, 100 warmup steps, and a $90/10$ train/validation split, with early stopping on validation loss (patience 5) and class imbalance handled through weighted cross-entropy. We did not grid-search the fine-tuning hyperparameters as an exhaustive search was computationally out of reach. In both roles MGM uses raw (non-CLR, unfiltered) genus-level abundances aligned to its bundled ∼5,700-genus vocabulary, with cohort genera missing from the vocabulary dropped.

2.6 Software environments

The pipeline ran across two separate Python environments, split because MGM's and TabPFN's dependency requirements are completely incompatible.

Main workspace (Python 3.14.5). The three classical model families, TabPFN, every explainability computation applied to them, and all figure generation ran in a Python 3.14.5 environment. Its core scientific stack was NumPy 2.4.4, pandas 3.0.2, SciPy 1.17.1, scikit-learn 1.8.0, and Matplotlib 3.10.9. Gradient-boosted trees used XGBoost 3.2.0; Shapley attribution used SHAP 0.51.0, backed by Numba 0.65.1 and llvmlite 0.47.0; and the foundation-model classifier used TabPFN 8.0.3 with `tabpfn-extensions` 0.4.1 for its interpretability routines, on top of PyTorch 2.12.0.

MGM environment (Python 3.11.15). We isolated MGM in a second environment on Python 3.11.15, because `microformer-mgm` 0.5.8 transitively pins a 2023-era stack: PyTorch 2.0.1, NumPy 1.24.3, and transformers 4.33.3, which exists only for Python <3.12 and which cannot be installed alongside the other environment's NumPy 2.x. Alongside those, this environment used `tokenizers` 0.13.3, `accelerate` 0.24.1, `pytorch-lightning` 2.0.6, pandas 2.0.3, scikit-learn 1.3.1, and SciPy 1.15.3. Its attribution stack was: SHAP 0.49.1 (the last release compatible with NumPy 1.x), `captum` 0.8.0 for integrated gradients, and `shapiq` 1.4.1.

Table 3: Out-of-fold discrimination and calibration, pooled over the held-out folds of the nested cross-validation ($n = 1{,}638$; IBD prevalence 73.8%). AUROC is the area under the receiver operating characteristic curve and PR-AUC the area under the precision–recall curve. Bold marks the best value per column, excluding the baseline.

Model class	Model	AUROC	PR-AUC	Log loss
Classical	Elastic Net	**0.666**	0.594	0.640
	Random Forest	0.664	**0.608**	**0.563**
	XGBoost	0.631	0.574	0.625
Classical over MGM's latent space	Elastic Net	0.639	0.580	0.656
	Random Forest	0.653	0.605	0.566
	XGBoost	0.637	0.595	0.600
End-to-end deep	MGM	0.576	0.545	1.211
	TabPFN	0.605	0.571	0.749
Baseline	Stratified dummy	0.496	0.499	6.189

3 Results

We selected the findings reported in this section from broader comparisons of the models and explainability techniques, which can be found in their entirety in the appendix (Figures 4, 5, 6, 7, 8). Table 3 reports the out-of-fold classification performance of every model and serves as a reference for the analyses that follow.

3.1 Switching the method of computing insertion/deletion shifts the result as much as switching the model does

In our pursuit to review prominent XAI techniques in the context of compositional microbiome data, we wanted to validate whether different methods of computing and implementing the same metric, in this case insertion/deletion curves, would arrive at the same, or nearly identical, results. Since a substantial difference would imply that the results are influenced by the implementation of the metric as much as they are influenced by the model, such a test would help evaluate these

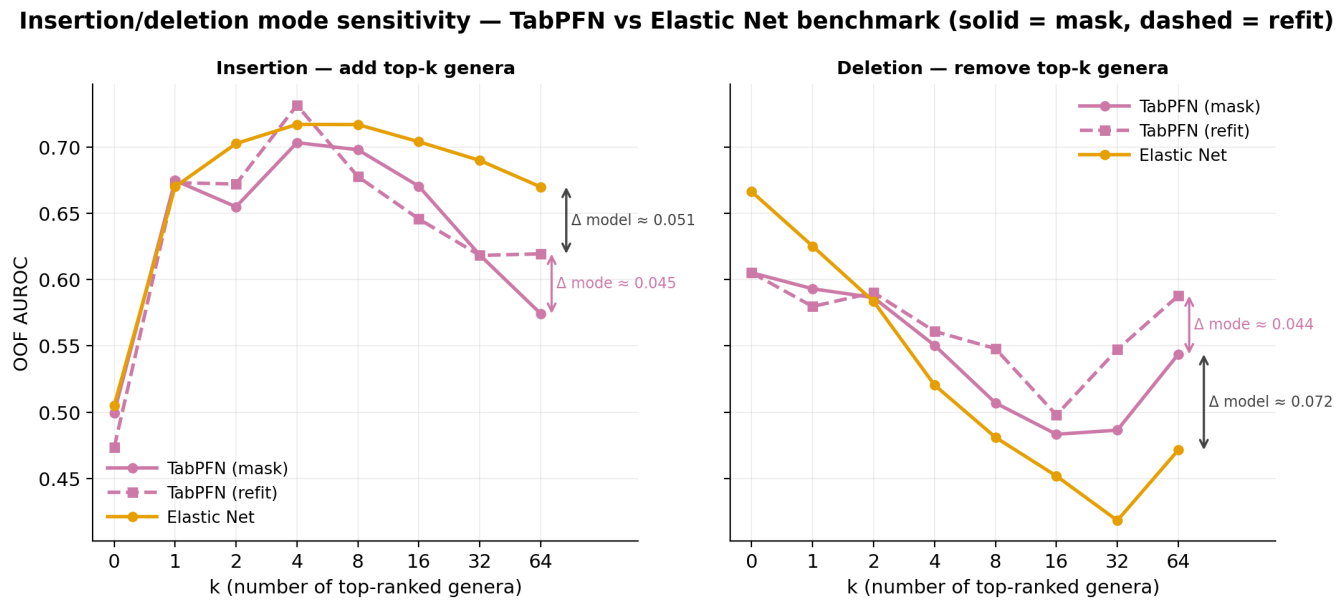

Figure 1: Insertion (panel A) and deletion (panel B) out-of-fold AUROC as a function of the number of top-ranked genera k reintroduced or removed, for TabPFN under both insertion/deletion methods and for Elastic Net as a benchmark. Colour encodes the model (orange: Elastic Net; purple: TabPFN) and line style encodes the method (solid with circles: $\mathrm{ID}_{\mathrm{mask}}$; dashed with squares: $\mathrm{ID}_{\mathrm{refit}}$); Elastic Net admits only the mask mode.

XAI technique-classes in the aforementioned context.

We hypothesised that the differences would not be substantial, and tested both insertion/deletion implementations of Section 2.3 on TabPFN: the mask mode, ranked by its imputation SHAP values, and the refit mode, ranked by its native SHAP values. Our showcase of the results (Figure 1) also includes the Elastic Net model as a point of reference, both because it had the strongest performance on this metric and because its SHAP values are exact, as it is a logistic regression.

We observed in the insertion sweep (Figure 1, panel A) that at a single inserted genus the two implementations were indistinguishable (insertion AUROC 0.675 and 0.673), but they separated as the input filled in: by the full top-64 genera the mask mode reached 0.574 against the refit mode's 0.620, a gap of 0.045 AUROC. That method-induced gap was almost as large as the 0.051 AUROC separating the better TabPFN mode from Elastic Net's insertion AUROC at the same k (0.670). Switching the measurement method thus impacted the faithfulness comparison by as much as switching the model under test.

The deletion sweep reproduced this mode sensitivity and, more pointedly, reversed which mode appeared faithful. Deletion removes the top-k genera from the full input, so if the SHAP ranking of features is faithful to the model, their removal should drive AUROC down; at the full top-64 genera Elastic Net fell furthest, to 0.471, while TabPFN's mask and refit modes held at 0.544 and 0.588 respectively (Figure 1, panel B). The refit method, the more faithful-looking of the two on insertion, was therefore the less faithful-looking on deletion.

The method choice did not merely display a difference in results, while keeping one method strictly superior to the other, indicating there are better and worse approaches of calculating this metric for TabPFN in this context. Instead, it completely inverted the ordering of which mode reads as more faithful depending on the sweep direction.

We did not observe the same pattern in the two approaches of computing feature importance for MGM, as the native approach outperformed the two-step IG and SHAP approach across all metrics. The latter did not produce any useful or meaningful results, but this is expanded on in the appendix.

3.2 Accuracy did not predict biological plausibility

We also wanted to explore if the models relied on the correct markers, not just that they score well on accuracy, as a model's ability to separate IBD from controls is a necessary, but not sufficient condition for models to be considered trustworthy [7]. Among others, making sure the model relies on the correct biological markers is essential [7].

We achieved this by measuring how its explanation recovered the established IBD biomarkers, scoring each model's averaged genus ranking against the curated panel (Table 1) with the recall@k and recovery score metrics defined in Section 2.3.

Elastic Net, having the strongest accuracy (Figure 1, full comparison in Table 3), produced the least plausible ranking in the entire set: its recall@k held at 0.14 through the top 25 genera and reached only 0.43 by $k = 50$. Random Forest and XGBoost, the next two on discrimination, ranked the panel essentially at chance (recovery score -0.00 and 0.02). Across the raw-abundance models, stronger discrimination thus came with flat or below-chance recovery of the known biomarkers.

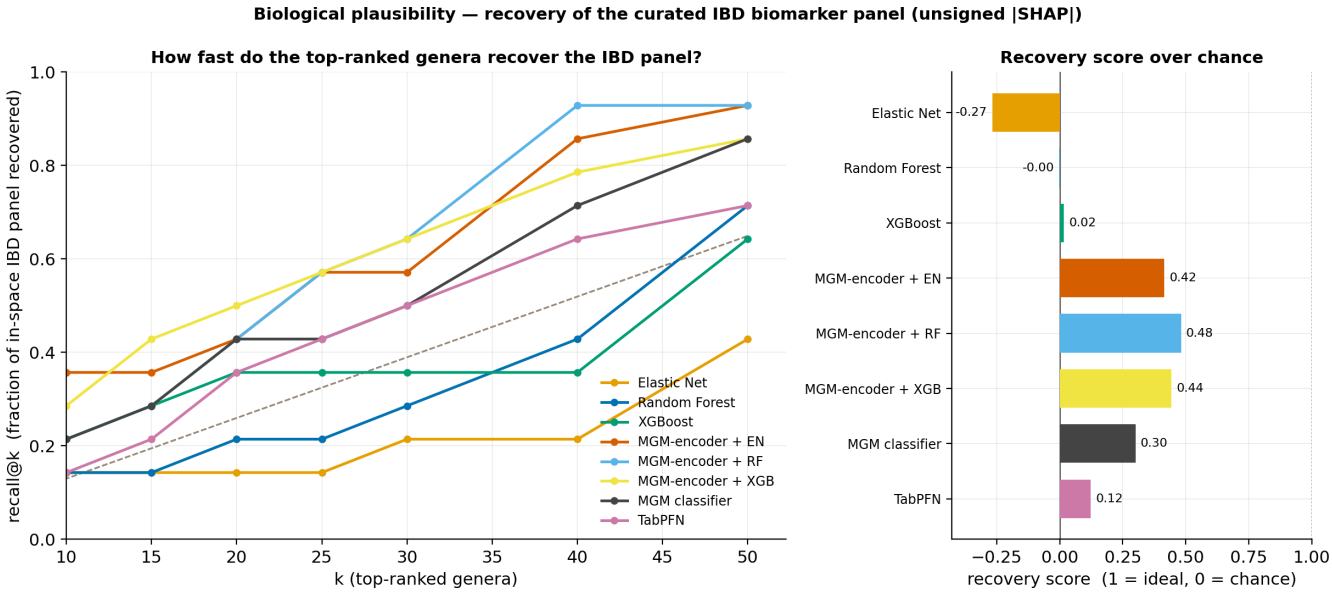

Figure 2: Biological plausibility of each model's unsigned |SHAP| genus ranking against the curated IBD biomarker panel of Table 1, using the recall@k and recovery score metrics defined in Section 2.3. Left: recall@k per model (solid lines) against the random-ranking baseline (faint dashed grey, expected recall@$k = k/n$); a curve bowing above the baseline is front-loaded with panel biomarkers. Right: the recovery score per model. Rankings use unsigned |SHAP|, so recovery reflects enrichment and depletion signals equally.

The rankings that did concentrate the panel near the top came from the models that discriminated least well. Every MGM-based ranking, the three classical heads fitted on the MGM latent space and the end-to-end MGM classifier, cleared chance by a wide margin (recovery score 0.30 to 0.48), recovering half the panel within the top 15 genera, where the raw-abundance models were still at a single hit (Figure 2). The highest recovery score (0.48) came from a classical head on the MGM latent space, while the end-to-end classifier, the weakest discriminator we evaluated (AUROC 0.576), still cleared chance comfortably at 0.30. Looking at the best and worst performing models, recovery score varied from -0.27 for Elastic Net up to 0.48 for the strongest MGM-encoder head, across models whose AUROC differed by only 0.09.

That the strongest classifier was the least plausible bears directly on how these attributions should be used. Because the panel is not ground truth (Section 2.4), a low recovery score does not on its own convict a model of learning an incorrect signal; but the spread we observed, roughly three-quarters of a recovery score unit, from -0.27 for Elastic Net up to 0.48 for the strongest MGM-encoder head, across models whose AUROC differed by only 0.09, shows that plausibility and discrimination answer different questions and have to be reported separately. A model can be the cohort's best discriminator while ranking canonical biomarkers below chance, the behaviour expected of a classifier leaning on cohort-specific or compositional structure orthogonal to the curated panel.

3.3 Explanation agreement is governed by the representation, not the model

We wanted to investigate whether the representation of the data had a substantial effect on the feature importance ranking produced, so whether the same models ranked features differently when trained on the latent representation generated by the MGM encoder, as opposed to the raw data, and along which axis (model or representation) greater variance lies. This is relevant because the signals the models learn and the rankings they produce should fundamentally be a feature of the data, not of the representation of that data; because greater variance generated by the difference in representation would mean the encoder introduces substantial noise and bias into the data.

We tested this by measuring the rank-biased overlap (RBO, $p = 0.9$) between every pair of models' unsigned |SHAP| genus rankings (Figure 3). The design admits a controlled comparison, as we fitted the three classical heads, Elastic Net, Random Forest, and XGBoost, both on the raw abundance table and on the MGM latent space. We hypothesised that the difference between models would be much greater than the one between representations.

The comparison came out the opposite way to the assumption above. Holding the algorithm fixed and swapping the representation collapsed agreement: a model agreed with its own re-encoded twin at RBO 0.39 (Elastic Net), 0.38 (Random Forest), and 0.37 (XGBoost). Holding the representation fixed and swapping the algorithm preserved it: the three raw-abundance heads agreed with one another at RBO 0.64 to 0.75, and the three MGM-encoder heads agreed even more tightly, at 0.75 to 0.81. The representation thus accounted for roughly twice the ranking agreement that the choice of model

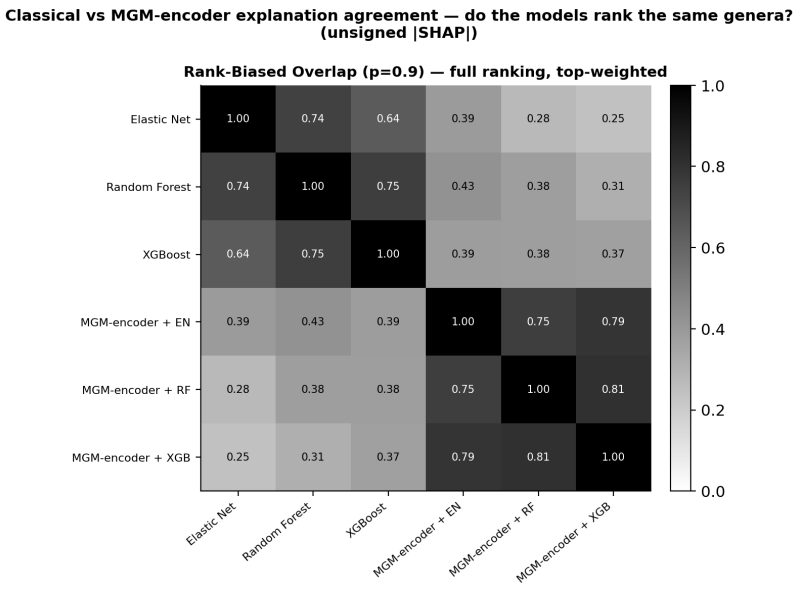

Figure 3: Cross-model agreement of the unsigned |SHAP| genus rankings, measured as rank-biased overlap (RBO, $p = 0.9$, top-weighted over the full ranking) between every pair of models. Lighter cells denote stronger agreement. The three classical heads appear twice, once on raw abundance (Elastic Net, Random Forest, XGBoost) and once on the MGM latent space (MGM-encoder + EN/RF/XGB), so the same algorithm can be compared across the two representations.

did.

When also taking into consideration the aforementioned plausibility result (Section 3.2), a conclusion arises. The MGM-encoder rankings were the most biologically plausible, recovering half the IBD panel within the top 15 genera, yet the least likely to agree with any model outside their own representation. This leads us to believe that the MGM encoder grants more weight to certain genera in its embeddings, steering the classifiers built on top towards those taxa, so that their attributions reflect the encoder's fixed prior rather than what each individual classifier learned. This pattern is in accordance with MGM's design, as MGM was built so that this contextualised representation, not the downstream estimator, carries the microbial signal [5]. The same dominance of the learned embedding over the classifier built on top of it has been reported for other deep microbiome representation learners [18]. Read against the plausibility result of Section 3.2, this fixed prior is not arbitrary: it concentrates our panel's IBD genera near the top, so the tight agreement among the latent-space models is agreement on a biologically reasonable representation, guided by MGM's preferred genera.

4 Responsible research

Ethics. We used only the public, de-identified IBDMDB release [11], [12], collected under the original study's consent and approval; we added no new data and ran no patient contact. The models here are for studying explainability, not for clinical use, and our own results caution against reading their feature rankings as causal biology (Section 3.2).

Reproducibility. Every reported number comes from a fixed seed and a participant-grouped cross-validation, computed out-of-fold (Section 2.2). The dataset is public and versioned, the code is also public [19], and we pin the exact package versions for both environments (Section 2.6). Due to the high compute costs for some of the metrics, the findings might not be trivially reproducible, but all of the required information is provided.

AI Use. Artificial Intelligence in the form of large language models (Claude 4.7/4.8) was used in different facets of the creation of this work, including: searching for relevant publications in different areas and filtering them by relevance, debugging code, generating docstring and documentation for files and functions in the codebase, rephrasing and proofreading, as well as coding the figure generation scripts.

5 Discussion

We asked how useful standard XAI techniques are on compositional microbiome data, and what they can tell us about how these models decide. Each finding answers part of that, and all three carry the same caveat: the answer depends on a choice that is easy to leave implicit, the measurement method and implementation, the metric reported, or the representation the model reads from.

The model and method of computing the metric make the same amount of difference in the insertion/deletion curve, leading us to believe that neither is a sufficient explanation on its own, both introduce substantial noise and method-specific bias. This implementation dependence is not a quirk of our setup but a caveat of the underlying attribution methods.

The cause is the way these methods, mask and refit, approach representing absent features. Since there is not a fixed way of doing so in the SHAP process [10], it is left up to choice. Integrated gradients have the same problem, requiring

an explicit baseline that its axioms do not determine [9]. The wider evaluation literature already reports that perturbation-based faithfulness metrics give inconsistent verdicts under exactly these kinds of changes in measurement method [8]. The order inversion we observe is a concrete realisation of that instability, which is why a read from a single implementation cannot be taken at face value for TabPFN on this data, and why the techniques are not self-validating in this context.

Looking at the plausibility result, the strongest classifier being the least plausible bears directly on how these attributions should be used. Because the panel is not ground truth (Section 2.4), a low recovery score does not on its own convict a model of learning an incorrect signal. However, the fact that the most plausible and least plausible models do not greatly differ in classification performance shows that plausibility and discrimination answer different questions and have to be reported separately. A model can be the cohort's best classifier while ranking canonical biomarkers below chance, with its behaviour leaning on cohort-specific signals orthogonal to the curated panel. The plausibility result narrows what these techniques can tell us: a faithful explanation of an accurate model is not necessarily a biologically correct one [7], because the compositional structure means an apparent per-taxon signal need not correspond to any change in that taxon [3]. Accuracy is therefore not evidence that a model uses the right biology.

The agreement result speaks most directly to how these models decide. The MGM-encoder models were the least likely to agree with any other class of models, while having similar rankings with each other. This leads us to believe the MGM encoder grants more weight to certain genera in its embeddings, steering the classifiers built on top towards those taxa, so that their attributions reflect the encoder's fixed prior more than the raw signal in the data. When attributions are governed by the representation rather than the classifier, an explanation of a latent-space model describes the encoder more than the model deployed on top of it. This is the expected behaviour of a foundation model whose value lies in a fixed, pretrained representation [5]. The encoder's rankings were also the most plausible [18], but that plausibility reflects its fixed prior over genera, not what the downstream classifier learned. Furthermore, the encoder's bias towards certain genera could be the reason for their decreased performance compared to the unencoded classical models (Table 3), as they are unable to capture full cohort-specific signals. So once a model reasons over a learned representation, its attributions stay biologically readable while no longer describing that particular model.

5.1 Limitations

Several limitations bound these conclusions. We studied a single disease in a single cohort (IBDMDB), so the specific agreement and plausibility numbers may not transfer to other diseases or sequencing pipelines; the patterns we report are qualitative claims about how the metrics behave, not calibrated constants. The plausibility panel is a curated, non-exhaustive reference rather than ground truth, so low plausibility flags a divergence from current consensus rather than proving a model wrong (Section 2.4). The end-to-end MGM model was fit from a single, pilot-chosen hyperparameter configuration rather than a full search, for reasons of compute, so its absolute performance should be read as an adequate baseline rather than a tuned ceiling. Finally, the refit insertion/deletion mode that exposed the protocol sensitivity was only feasible for the models that can be re-evaluated on an arbitrary feature subset (TabPFN and MGM), so we cannot say how far the same instability extends to the classical models.

6 Conclusions and future work

This work asked how useful standard explainability (XAI) techniques are on compositional microbiome data, and what they can actually tell us about how a model reaches its decisions. Its novelty is not a new attribution method but a side-by-side comparison across three levels of representation, which exposes failure modes that any single-model explanation analysis would hide. We arrive at three conclusions.

First, the explainability metrics are not self-validating on this data: the choice of how an insertion/deletion curve represents an absent feature shifted the faithfulness verdict for TabPFN as much as swapping the model under test, and even inverted which protocol looked more faithful between the two sweeps. Faithfulness on compositional microbiome data therefore has to be reported across more than one protocol, or not trusted at all.

Second, a model being accurate, or even faithfully explained, does not make its explanation biologically correct: in our set the strongest discriminator produced the least plausible ranking, placing canonical IBD biomarkers below chance. Discrimination and biological plausibility answer different questions and must be measured and reported separately; accuracy is not evidence that a model uses the right biology.

Third, once a model reasons over a learned representation, its explanation can describe the representation more than the model, as we found for MGM: swapping the data representation collapsed ranking agreement far more than swapping the algorithm did, roughly twice as much. An attribution computed over a learned latent space therefore characterises the encoder's fixed prior as much as the classifier deployed on top of it.

6.1 Future work

There are several directions in which future work could be focused to extend and strengthen these findings: replicating the comparison across other cohorts, diseases, sequencing pipelines, and curated biomarker panels would establish which patterns are general and which are artefacts of IBDMDB. A second direction is to close the gaps our compute budget left open, with a full hyperparameter search for the end-to-end deep models, an extension of the refit insertion/deletion mode to the classical models, to measure how far the protocol instability we found for TabPFN actually reaches, and with multiple runs per experiment, to be able to test for statistical significance.

References

[1] J. F. Cryan et al., "The Microbiota-Gut-Brain Axis," *Physiological Reviews*, vol. 99, no. 4, pp. 1877–2013, Oct. 2019, ISSN: 0031-9333. DOI: 10.1152/physrev.00018.2018 Accessed: Apr. 23, 2026. [Online]. Available: https://journals.physiology.org/doi/full/10.1152/physrev.00018.2018

[2] J. Ni, G. D. Wu, L. Albenberg, and V. T. Tomov, "Gut microbiota and IBD: Causation or correlation?" *Nature Reviews Gastroenterology & Hepatology*, vol. 14, no. 10, pp. 573–584, Oct. 2017, ISSN: 1759-5053. DOI: 10.1038/nrgastro.2017.88 Accessed: Jun. 9, 2026. [Online]. Available: https://www.nature.com/articles/nrgastro.2017.88

[3] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Frontiers in Microbiology*, vol. 8, Nov. 15, 2017, ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02224 Accessed: May 5, 2026. [Online]. Available: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2017.02224/full

[4] R. Hernández Medina et al., "Machine learning and deep learning applications in microbiome research," *ISME Communications*, vol. 2, no. 1, p. 98, Dec. 1, 2022, ISSN: 2730-6151. DOI: 10.1038/s43705-022-00182-9 Accessed: Apr. 22, 2026. [Online]. Available: https://academic.oup.com/ismecommun/article/7461082

[5] H. Zhang, Y. Zhang, Z. Kang, J. Xiong, R. Yang, and K. Ning, "MGM as a Large-Scale Pretrained Foundation Model for Microbiome Analyses in Diverse Contexts," *Advanced Science*, vol. 13, no. 24, e13333, Apr. 2026, ISSN: 2198-3844. DOI: 10.1002/advs.202513333 PMID: 41580987.

[6] N. Hollmann et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025, ISSN: 1476-4687. DOI: 10.1038/s41586-024-08328-6 Accessed: May 22, 2026. [Online]. Available: https://www.nature.com/articles/s41586-024-08328-6

[7] A. B. Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." arXiv: 1910.10045 [cs], Accessed: Apr. 22, 2026. [Online]. Available: http://arxiv.org/abs/1910.10045 pre-published.

[8] M. Nauta et al., "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," *ACM Comput. Surv.*, vol. 55, 295:1–295:42, 13s Jul. 13, 2023, ISSN: 0360-0300. DOI: 10.1145/3583558 Accessed: Apr. 24, 2026. [Online]. Available: https://dl.acm.org/doi/10.1145/3583558

[9] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks." arXiv: 1703.01365 [cs.LG], Accessed: Jun. 8, 2026. [Online]. Available: http://arxiv.org/abs/1703.01365 pre-published.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA: Curran Associates Inc., Dec. 4, 2017, pp. 4768–4777, ISBN: 978-1-5108-6096-4. Accessed: Jun. 8, 2026. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230

[11] J. Lloyd-Price et al., "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, May 2019, ISSN: 1476-4687. DOI: 10.1038/s41586-019-1237-9 Accessed: Apr. 23, 2026. [Online]. Available: https://www.nature.com/articles/s41586-019-1237-9

[12] *The Inflammatory Bowel Disease Multi'omics Database*.

[13] D. Gevers et al., "The treatment-naive microbiome in new-onset Crohn's disease," *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, Mar. 2014. DOI: 10.1016/j.chom.2014.02.005 [Online]. Available: https://www.cell.com/cell-host-microbe/fulltext/S1931-3128(14)00063-8

[14] E. A. Franzosa et al., "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, no. 2, pp. 293–305, Dec. 10, 2018, ISSN: 2058-5276. DOI: 10.1038/s41564-018-0306-4 Accessed: Jun. 9, 2026. [Online]. Available: https://www.nature.com/articles/s41564-018-0306-4

[15] J. Zheng et al., "Noninvasive, microbiome-based diagnosis of inflammatory bowel disease," *Nature Medicine*, vol. 30, no. 12, pp. 3555–3567, Dec. 2024. DOI: 10.1038/s41591-024-03280-4 [Online]. Available: https://www.nature.com/articles/s41591-024-03280-4

[16] R. Pittayanon et al., "Differences in gut microbiota in patients with vs without inflammatory bowel diseases: A systematic review," *Gastroenterology*, vol. 158, no. 4, 930–946.e1, Mar. 2020. DOI: 10.1053/j.gastro.2019.11.294 [Online]. Available: https://www.gastrojournal.org/article/S0016-5085(19)41893-3/fulltext

[17] J. D. Ravichandar et al., "Strain level and comprehensive microbiome analysis in inflammatory bowel disease via multi-technology meta-analysis identifies key bacterial influencers of disease," *Frontiers in Microbiology*, vol. 13, p. 961 020, 2022. DOI: 10.3389/fmicb.2022.961020 [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2022.961020/full

[18] M. Oh and L. Zhang, "DeepMicro: Deep representation learning for disease prediction based on microbiome data," *Scientific Reports*, vol. 10, no. 1, p. 6026, Apr. 7, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-63159-5 Accessed: Apr. 23, 2026. [Online]. Available: https://www.nature.com/articles/s41598-020-63159-5

[19] S. Lupsan, *Paper Codebase*. [Online]. Available: https://gitlab.ewi.tudelft.nl/cse3000/2025-2026-q4/microbiome-foundation-models/stefan-lupsan

[20] D. Rundel, J. Kobialka, C. von Crailsheim, M. Feurer, T. Nagler, and D. Rügamer. "Interpretable Machine Learning for TabPFN," arXiv.org, Accessed: May 31, 2026. [Online]. Available: https://arxiv.org/abs/2403.10923v2

A Appendix

This appendix collects the full-coverage versions of the figures that the Results section reports only in part: the collapsed and full insertion/deletion faithfulness results (Figures 4 and 7); the complete cross-model agreement matrix, including the end-to-end deep models (Figure 8); and the audit of MGM's native attribution against the two-step imputation/SHAP path, which agreed on the ranking but not on faithfulness (Figures 5, 6 and 9).

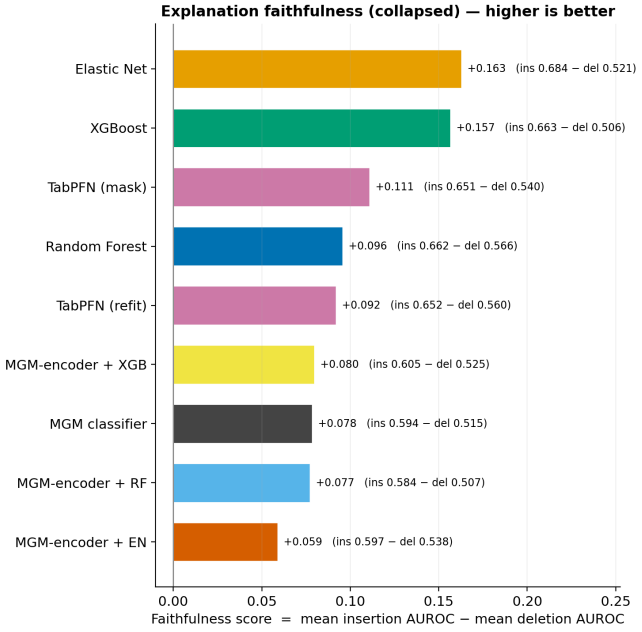

Figure 4: Collapsed explanation-faithfulness score per model, defined as the mean insertion AUROC minus the mean deletion AUROC over the top-k sweep of Figure 7; higher is better, and each bar is annotated with the score and its insertion and deletion components. The raw-abundance Elastic Net and XGBoost rank highest, while the MGM-encoder heads cluster lowest.

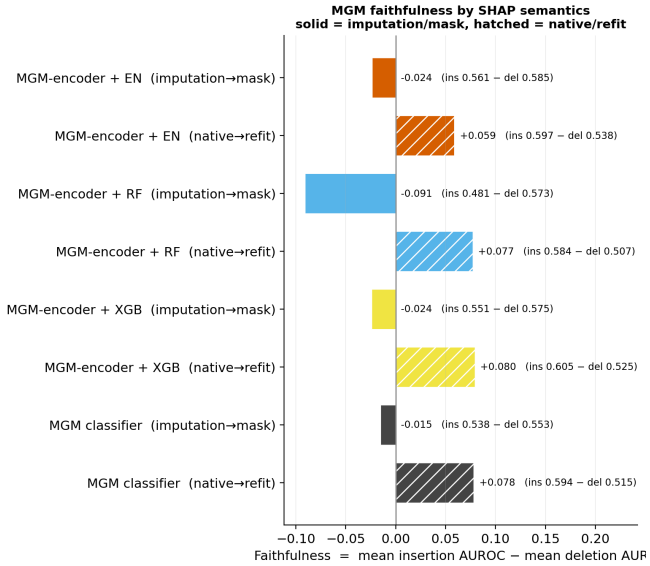

Figure 5: Collapsed faithfulness (mean insertion AUROC minus mean deletion AUROC) for each MGM model under the two SHAP semantics; solid bars are the imputation/mask variant, hatched bars the native/refit variant. Every imputation score is near zero or negative (-0.091 to -0.015) while every native score is positive ($+0.059$ to $+0.080$), so the native attribution was the only one of the two to produce a faithful ranking.

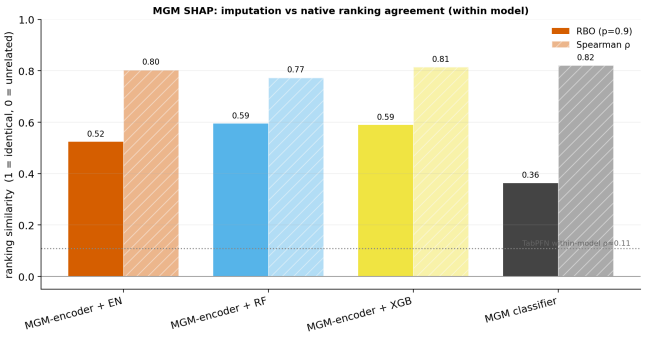

Figure 6: Within-model agreement between the imputation and native SHAP rankings for each MGM model, measured as rank-biased overlap (RBO, $p = 0.9$, solid bars) and Spearman ρ (hatched bars). The dotted line marks the TabPFN within-model mode baseline ($\rho \approx 0.11$) reported by Rundel et al. [20]. Both measures sit far above that baseline (RBO 0.36 to 0.59, ρ 0.77 to 0.82), so, unlike TabPFN, the MGM rankings were stable across the choice of SHAP semantics.

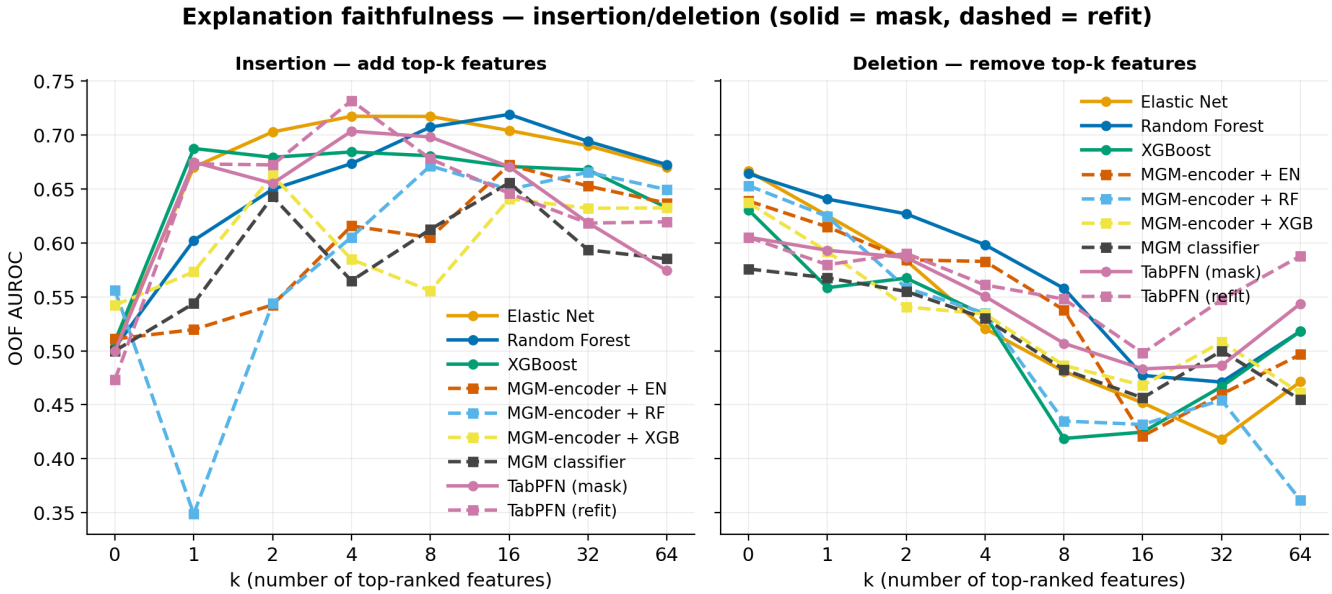

Figure 7: Insertion (left) and deletion (right) out-of-fold AUROC as a function of the number of top-ranked genera k, for all nine model and explanation configurations. Solid lines mask the selected features ($\mathrm{ID}_{\mathrm{mask}}$); dashed lines remove them from context ($\mathrm{ID}_{\mathrm{refit}}$). A steep early rise on insertion indicates that a small set of top-ranked genera carries the prediction; a flat early curve indicates that no compact subset does.

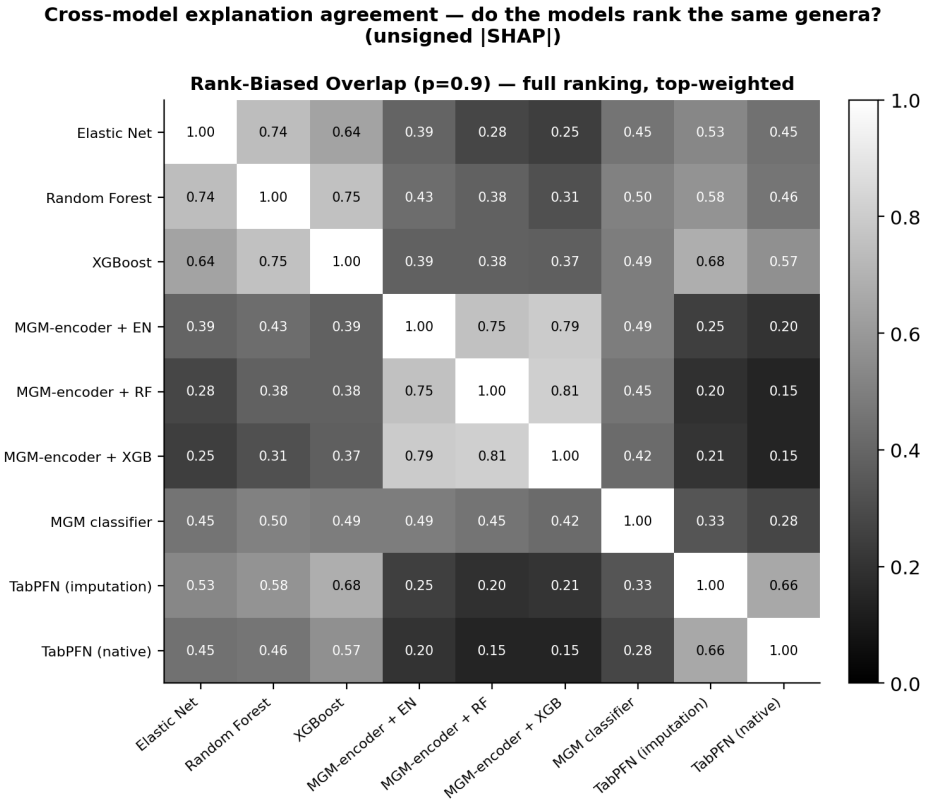

Figure 8: Cross-model agreement of the unsigned |SHAP| genus rankings for every model configuration, measured as rank-biased overlap (RBO, $p = 0.9$, top-weighted over the full ranking) between every pair of models. Lighter cells denote stronger agreement. This is the full counterpart of Figure 3, additionally including the end-to-end deep models (MGM and TabPFN) alongside the classical heads on raw abundance and on the MGM latent space.

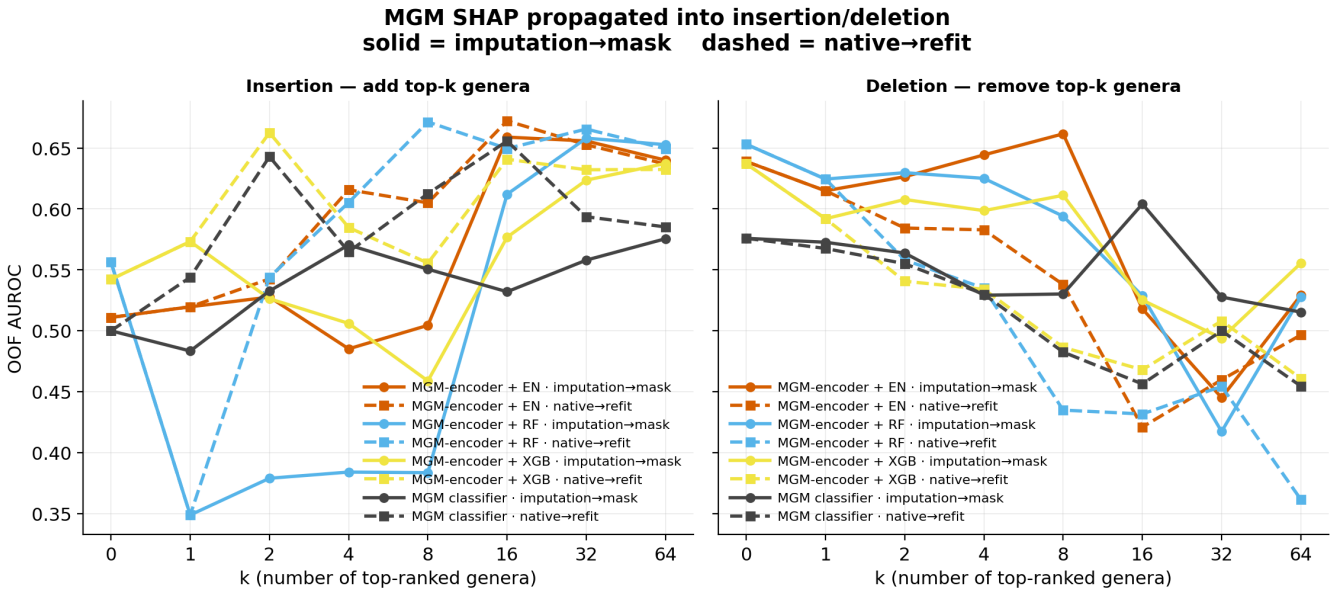

Figure 9: Insertion (left) and deletion (right) out-of-fold AUROC for the four MGM-based models under the two SHAP semantics, against the number of top-ranked genera k (log-spaced 0 to 64). Solid lines propagate imputation SHAP into the mask variant; dashed lines propagate native SHAP into the refit variant. The imputation curves are non-monotonic and fail to rise cleanly on insertion, whereas the native curves behave as a faithful ranking should.