RETHINKING FOUNDATION MODELS FOR MICROBIOME PHENOTYPE PREDICTION: WHAT ACTUALLY MATTERS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models for the microbiome increasingly rely on large-scale pretraining and Transformer encoders, but it remains unclear whether these choices address the main difficulties of phenotype prediction from sparse abundance profiles. We revisit this assumption through a controlled study of inflammatory bowel disease (IBD) prediction that disentangles pretraining, architectural inductive bias, representation complexity, and cohort shift. A pretrained 34M-parameter MGM encoder and a 1.1M-parameter embedding model without any pretraining reach comparable held-out accuracy (92.8% versus 92.3%), indicating that neither scale nor pretraining alone determines downstream utility. Permutation-invariant models are similarly strong: DeepSets and a minimal learned-embedding model reach 91.9% and 92.3% in the same structural comparison. Under leakage-robust grouped cross-validation, however, classical models remain highly competitive: HistGradientBoosting reaches 0.979 AUROC, compared with 0.957 for the learned embedding, and a simple presence-vector MLP reaches 0.969. The decisive difficulty appears under cohort shift. When models trained on one cohort are evaluated on independent cohorts, performance drops sharply; on the balanced external cohort, AUROC ranges from 0.621 for the embedding model to 0.668 for Random Forest, far below the 0.974 in-family result. Cohort-probe analyses show that study identity is highly decodable from every representation (cohort accuracy 0.93–0.97 versus disease accuracy 0.76–0.82), and remains so after balancing disease composition, indicating that learned representations are dominated by cohort-associated structure. Crucially, adding a single additional cohort to the training set raises disease identification on an independent cohort from 32% to 49%, whereas gradient-reversal debiasing and model scaling do not. These results indicate that cross-cohort generalization in microbiome phenotype prediction is governed not by model capacity, pretraining, or explicit debiasing, but by the cohort coverage of the training distribution.

1 INTRODUCTION

Foundation models have transformed domains in which large-scale pretraining produces representations that transfer across tasks and datasets. This success has motivated a rapidly growing line of microbiome foundation models, including MGM (Zhang et al., 2026), BiomeGPT (Medearis et al., 2026), and Waypoint (Treloar et al., 2026), which pretrain Transformer encoders on collections ranging from thousands to hundreds of thousands of microbial profiles. The underlying premise is appealing: microbial communities contain recurring structure, and sufficiently broad pretraining may capture reusable representations of that structure.

Microbiome phenotype prediction, however, differs in several important respects from modalities in which foundation modeling has been most successful. A sample is naturally represented by a sparse set or abundance vector of taxa rather than a sequence with an intrinsic order (Gloor et al., 2017; Weiss et al., 2017). Available phenotype labels are limited, and profiles collected by different studies can differ substantially (Gibbons et al., 2018; Vujkovic-Cvijin et al., 2020). These properties raise a basic question that is easy to obscure when model scale, pretraining, tokenization, and down-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

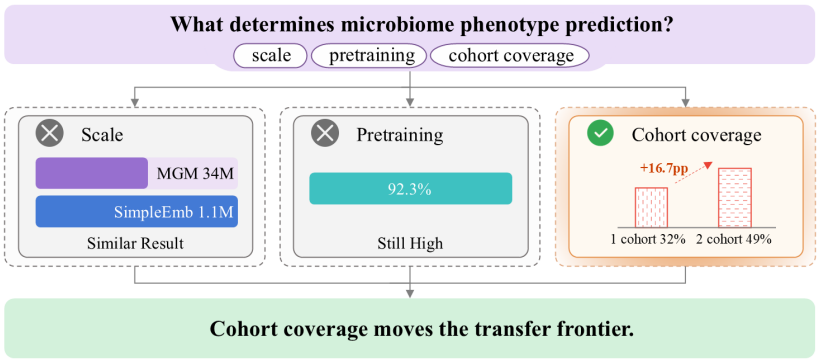

Figure 1: **Overview.** (A) Which factor determines microbiome phenotype prediction: scale, pretraining, or cohort coverage? (B) Within-cohort, pretrained MGM (34M) and SimpleEmb (1.1M, no pretraining) are statistically indistinguishable. (C) Cross-cohort, adding one training cohort raises disease identification from 32% to 49%. Cohort coverage moves the transfer frontier.

stream evaluation change simultaneously: *what actually makes a microbiome representation useful for phenotype prediction?*

We study this question using IBD prediction as a focused test bed. Rather than proposing another large pretrained model, we use a deliberately minimal embedding model as a diagnostic probe and organize the study around three questions. **First, does foundation-model pretraining provide a useful downstream representation?** We compare a pretrained MGM encoder with classical abundance-based models, a task-trained Transformer, DeepSets, and minimal embedding models. **Second, what inductive bias and representation complexity are needed?** We separate sequential from permutation-invariant architectures and compare raw abundance, taxon presence, random embeddings, and learned embeddings. **Third, what ultimately limits generalization?** We move from a held-out split to exact-input-grouped cross-validation and then to independent cohorts. Figure 1 summarizes the three questions and their answers.

Our study makes three contributions. (1) We provide a controlled decomposition of microbiome representation learning across *pretraining, inductive bias, and representation complexity*, showing that these factors should not be conflated when evaluating microbiome foundation models, and that neither scale nor pretraining alone determines downstream utility. (2) Using cohort probes, we show that learned representations are dominated by cohort-associated structure rather than disease structure, and that this persists after balancing disease composition. (3) We show that cross-cohort generalization is governed primarily by the *cohort coverage* of the training distribution: adding one extra cohort improves transfer to an independent cohort by 16.7 percentage points, whereas gradient-reversal debiasing and model scaling do not. Together, the results argue for a shift in emphasis from scaling microbiome encoders toward expanding training-cohort coverage.

2 Related Work

Microbiome foundation models. MGM (Zhang et al., 2026) pretrains a Transformer with next-genus prediction on 263,302 samples and uses rank-based tokenization. BiomeGPT (Medearis et al., 2026) combines species and abundance-bin embeddings with masked-abundance pretraining on 13,300 human gut metagenomes. Waypoint (Treloar et al., 2026) further scales Transformer-based microbiome modeling to more than 539k samples.

Task-specific microbiome prediction. Classical machine-learning pipelines remain strong for microbiome phenotype prediction. SIAMCAT (Wirbel et al., 2021) emphasizes rigorous preprocessing and cross-study evaluation, while DeepMicro (Oh & Zhang, 2020) studies unsupervised representation learning followed by conventional classifiers. Multi-study benchmarks show both transferable and non-specific disease signals (Pasolli et al., 2016; Duvallet et al., 2017); batch-aware methods can improve integration but do not remove the need for held-out-cohort testing (Ling et al.,

2022; Austin et al., 2025). Recent benchmarking therefore motivates direct comparison between foundation-model representations and strong classical baselines (Mu et al., 2026).

Inductive bias and interpretation. DeepSets (Zaheer et al., 2017) provides a general permutation-invariant architecture for set-valued inputs, while FT-Transformer (Gorishniy et al., 2021) adapts self-attention to tabular prediction. We use these model families to test whether an abundance-ranked taxon list benefits from sequential modeling. Unlike additive attribution summaries such as SHAP (Lundberg & Lee, 2017), we use leave-one-out (LOO) perturbation: removing a taxon yields sample-specific evidence that can be checked by deletion tests and literature-curated directions.

3 Experimental Setting

3.1 Data and prediction task

The primary benchmark, `clean_2538`, contains 659 training and 167 test samples, with 55.7% disease prevalence. Across the dataset, 366 genera are observed within a vocabulary of 1,226 taxa. Profiles are sparse and long-tailed: 251 genera occur in fewer than 5% of samples, whereas only 42 occur in more than half. We additionally use a larger `merged_all` collection (3,350 train / 838 test) and independent Qiita cohorts for transfer evaluation. Cross-dataset evaluations remove identified sample overlap between source training and target test sets. This design reflects the marked population and temporal heterogeneity documented in large human-microbiome and IBD cohorts (Human Microbiome Project Consortium, 2012; Gevers et al., 2014; Halfvarson et al., 2017; Lloyd-Price et al., 2019; Franzosa et al., 2019).

The task is binary IBD prediction from genus-level microbiome profiles. We distinguish three evaluation regimes: (i) the original held-out split; (ii) 3 repeats of 5-fold StratifiedGroupKFold, grouping exact genus-sequence hashes to prevent identical inputs from appearing in train and validation folds; and (iii) evaluation on independent cohorts. The latter is essential because strong within-cohort performance need not imply robustness to study-specific variation.

3.2 Model families

We compare four families. **Classical abundance models** use the raw 1,226-dimensional vector (logistic regression, SVM, Random Forest, HistGradientBoosting, Extra Trees, XGBoost, MLP). **Pretrained microbiome representations** use the MGM encoder plus a downstream classifier, and we also transfer BiomeGPT at genus level with per-fold masked pretraining. **Task-trained structural baselines** are FT-Transformer and DeepSets. **Minimal representations** are taxon presence vectors and SimpleEmb, which maps each genus to a learned embedding, applies masked mean pooling, and classifies with an MLP.

For a sample containing taxa $S = \{g_1, \ldots, g_k\}$, SimpleEmb computes

$$\mathbf{h}(S) = \frac{1}{k} \sum_{g \in S} \mathbf{e}_g, \qquad \hat{y} = f_\theta(\mathbf{h}(S)), \tag{1}$$

where $\mathbf{e}_g$ is a trainable genus embedding and f_θ is an MLP. This weak architecture tests whether learned taxon representations are needed and exposes an explicit representation space. Unless otherwise noted, we use AdamW (lr 10^{-3}, weight decay 10^{-4}), cosine annealing, batch size 32, and class-weighted loss.

3.3 Representation and attribution analyses

We vary embedding dimension from 16 to 768 and compare learned embeddings with raw/presence inputs, frozen random embeddings, and linear versus nonlinear heads. For sample-specific interpretation, LOO attribution for genus g is the prediction change after removing g from the profile. We assess attribution stability across folds, deletion of top-ranked versus random genera, association with genus prevalence and differential abundance, and agreement with literature-curated IBD directions where taxa overlap. Finally, we test whether providing LOO evidence to Qwen2-7B-Instruct

Table 1: **IBD classification performance on `clean_2538` (659 train / 167 test).** All methods evaluated on an identical split. SimpleEmb reports 5-seed ensemble mean ± std. † = this work.

Method	ACC	AUC	Prec.	Rec.	F1	Params
Lower bound						
Majority class	0.5569	0.5000	0.5569	1.0000	0.7154	0
Classical ML (raw 1226-dim abundance)						
Logistic Regression	0.8802	0.9241	0.9195	0.8602	0.8889	2.5K
Linear SVM	0.8563	0.8641	0.9367	0.7957	0.8605	2.5K
Random Forest	0.899±0.007	0.961±0.002	0.917±0.008	0.901±0.013	0.909±0.007	∼200 trees
XGBoost	0.922±0.000	0.968±0.000	0.935±0.000	0.925±0.000	0.930±0.000	∼200 trees
MLP (sklearn)	0.889±0.011	0.959±0.002	0.935±0.005	0.860±0.023	0.896±0.012	330K
MGM Transformer (34M params, pretrained)						
MGM pretrained + MLP	0.9281	0.9763	–	–	–	34M
Task-trained structural baselines						
FT-Transformer	0.923±0.020	0.959±0.006	0.935±0.023	0.927±0.020	0.931±0.018	–
DeepSets	0.919±0.010	0.981±0.004	0.961±0.016	0.890±0.014	0.924±0.009	–
SimpleEmb (this work)						
SimpleEmb + Linear	0.910±0.004	0.968±0.003	0.928±0.005	0.910±0.009	0.919±0.004	0.93M
SimpleEmb†	0.923±0.007	0.974±0.003	0.951±0.008	0.910±0.013	0.930±0.007	1.14M

improves the faithfulness of natural-language explanations relative to prompting from the raw genus list alone.

4 Does Pretraining or Scale Determine Downstream Utility?

4.1 Held-out evaluation

On the original held-out split, SimpleEmb reaches 0.923 accuracy and 0.974 AUROC, similar to XGBoost (0.922 accuracy; 0.968 AUROC) while using no external pretraining. The pretrained 34M-parameter MGM encoder followed by an MLP reaches 0.928 accuracy and 0.976 AUROC (Table 1), statistically indistinguishable from SimpleEmb despite using 30× more parameters and large-scale pretraining. This parity—rather than a clear winner—is the first indication that neither scale nor pretraining alone determines downstream utility.

4.2 Scale and pretraining are not the deciding factors

To isolate architecture from pretraining, we train a Transformer directly on the target task. FT-Transformer, without microbiome pretraining, reaches 0.923 accuracy, close to the pretrained MGM (0.928) and to the permutation-invariant models (Table 1). DeepSets and SimpleEmb, both permutation-invariant, reach 0.919 and 0.923 accuracy without pretraining. The spread among the task-trained and pretrained architectures on the original split is small relative to the degradation observed under distribution shift (Section 6). More importantly, the grouped-CV comparison does not support a consistent advantage for SimpleEmb over strong conventional alternatives. Pairwise McNemar tests show no significant difference between SimpleEmb and HistGradientBoosting, MGM+MLP, Extra Trees, logistic regression, or SVM after Benjamini–Hochberg correction (all $q_{\mathrm{BH}} > 0.05$; Table 10). Significant differences are observed only against the weaker KNN and BernoulliNB baselines. Thus, the relevant result is not that one architecture wins within cohort, but that substantially different model classes converge to a similar performance regime while cross-cohort transfer remains much more fragile.

Table 2: Leakage-robust grouped cross-validation (15 folds). Values are mean ± standard deviation.

Method	Accuracy	AUROC	AUPRC
BiomeGPT architecture transfer	.937±.021	.981±.010	.978±.025
HistGradientBoosting	.925±.025	.979±.010	.981±.010
SIAMCAT (LASSO)	.922±.015	.975±.009	.976±.016
Random Forest	.916±.023	.971±.013	.970±.019
SimpleEmb + MLP	.914±.020	.957±.014	.952±.032
Raw MLP	.872±.031	.945±.024	.950±.030
MGM + MLP	.853±.028	.929±.026	.924±.048
DeepMicro	.846±.025	.917±.027	.911±.040

Table 3: **Representation and head ablation under grouped-CV (15 folds).** Presence = multi-hot token presence vector; Embedding64 = trainable $E{=}64$ embedding + masked mean pool.

Model	**Params**	**ACC**	**AUROC**	**AUPRC**
Presence + Linear	2,454	0.876 ± 0.037	0.949 ± 0.019	0.957 ± 0.018
Presence + MLP	315,138	**0.912** ± 0.025	**0.969** ± 0.012	**0.973** ± 0.012
Embedding64 + Linear	78,594	0.857 ± 0.036	0.916 ± 0.029	0.909 ± 0.049
Embedding64 + MLP (ours)	96,130	0.906 ± 0.021	0.956 ± 0.016	0.954 ± 0.031

5 What Representation Complexity Is Actually Needed?

5.1 Grouped validation changes the ranking

The leakage-robust grouped evaluation gives a more cautious picture (Table 2). A genus-level transfer of the BiomeGPT architecture is numerically strongest (0.981 AUROC), but HistGradientBoosting (0.979), SIAMCAT (0.975), and several tree ensembles are similarly strong. SimpleEmb reaches 0.957 AUROC and does not outperform the strongest classical baselines. The result is important precisely because it is negative: sophisticated learned representations are not required to obtain strong within-cohort discrimination on these data.

5.2 Presence is nearly sufficient for prediction

The most revealing representation ablation compares a multi-hot taxon-presence vector with a learned embedding under the same classifier family. Presence+MLP reaches 0.969 AUROC, whereas a 64-dimensional learned embedding reaches 0.956. Linear heads are weaker for both representations (0.949 and 0.916 AUROC, respectively). The stronger performance of Presence+MLP indicates that learned embeddings are not required for strong within-cohort prediction; much of the gain over linear models can instead be obtained through nonlinear decision boundaries.

This observation changes the role of SimpleEmb in the paper. Its value is not that an embedding is required for high classification accuracy. Instead, the embedding provides a compact continuous space in which patients and taxa can be compared and perturbed. Predictive performance and representational utility should therefore be evaluated separately; Section 7 develops this representational utility through retrieval, clustering, and perturbation analyses.

5.3 Embedding dimension and data efficiency

Embedding capacity saturates quickly (Table 4). AUROC rises from 0.924 at dimension 16 to 0.956 at 64, but increasing the embedding dimension twelve-fold from 64 to 768 yields essentially no further improvement (0.956 vs. 0.957). The same saturation is visible in accuracy, which changes only from 0.906 to 0.914. Thus, once a modest representation capacity is reached, additional parameters provide little predictive benefit. Nested group-CV learning curves (Figure 2) show AUROC 0.956 at the full training fraction and 0.900 with only 20% of the data—about 400× smaller than the MGM pretraining corpus. In this setting, scale and data are not the scarce resources; the bottleneck is elsewhere.

Table 4: **Embedding dimension ablation under grouped 3×5-fold CV.** Only the embedding dimension changes; all other settings are fixed.

E	Params	ACC	Macro-F1	AUROC	AUPRC
16	24,994	0.855 ± 0.027	0.854 ± 0.027	0.924 ± 0.020	0.918 ± 0.041
32	48,706	0.879 ± 0.033	0.878 ± 0.034	0.941 ± 0.019	0.944 ± 0.022
64	96,130	0.906 ± 0.021	0.905 ± 0.021	0.956 ± 0.016	0.954 ± 0.031
128	190,978	0.908 ± 0.024	0.907 ± 0.024	0.957 ± 0.015	0.956 ± 0.026
256	380,674	0.909 ± 0.020	0.909 ± 0.020	0.956 ± 0.015	0.952 ± 0.037
512	**760,066**	**0.914** ± 0.020	**0.913** ± 0.019	**0.957** ± 0.014	0.952 ± 0.032
768	1,139,458	**0.914** ± 0.016	**0.913** ± 0.016	0.957 ± 0.015	0.952 ± 0.035

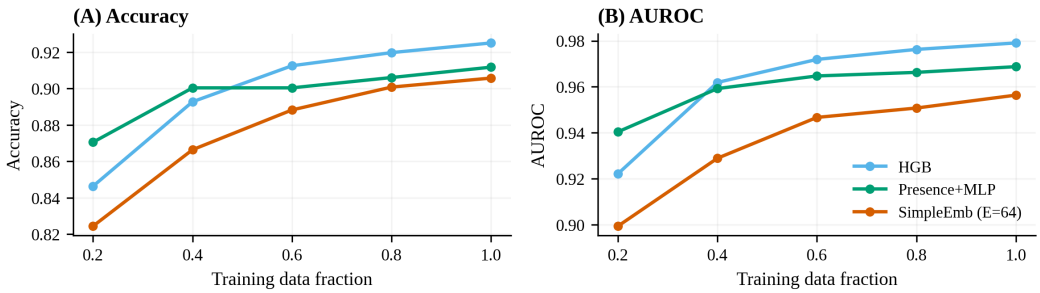

Figure 2: **Data efficiency (nested group-CV).** (A) Accuracy and (B) AUROC of SimpleEmb (E=64) as a function of training data fraction.

6 Cross-Cohort Shift Is the Dominant Bottleneck

6.1 Cross-Dataset Transfer

Within-dataset random splits remain strong: clean→clean reaches 0.982 AUROC and merged→merged 0.855. Transfer is asymmetric (Table 5). Training on the smaller clean set and testing on the decontaminated merged set yields 0.773 AUROC and severe sensitivity collapse, whereas training on the larger merged collection and testing on clean reaches 0.943 AUROC. The larger and more diverse source therefore improves transfer, but does not establish cohort invariance.

6.2 Independent Cohorts Expose the Real Failure Mode

The strongest evidence comes from leave-one-source-out evaluation (Table 6). Models trained on study 2538 degrade substantially on independent cohorts. On the balanced 10317 cohort, SimpleEmb reaches only 0.621 AUROC, HistGradientBoosting 0.615, and Random Forest 0.668, compared with approximately 0.974 AUROC for the in-family holdout. On disease-only cohorts, disease identification rates are likewise low and variable. Importantly, the learned embedding is not the most robust model; Random Forest is consistently among the most robust across the reported external cohorts.

6.3 Cohort-Associated Structure in Learned Representations

To characterize what the models encode beyond within-cohort discrimination, we probe four representations—raw abundance, SimpleEmb, the pretrained MGM encoder, and a random-initialized MGM encoder of the same architecture—with logistic-regression classifiers that predict either disease status or study identity (2538 versus 10317, the two cohorts that contain healthy controls). All probes use 5-fold stratified cross-validation over 1,402 samples after removing six profiles with zero observed abundance.

Study identity is highly decodable from every representation (Table 7; Figure 3): cohort-probe accuracy ranges from 0.93 to 0.97, whereas disease-probe accuracy is lower (0.76–0.82). Because

Table 5: **Within-dataset and cross-dataset transfer evaluation.** Cross-dataset transfer test sets are decontaminated of training sample overlap.

Train → Test	n_{train}	n_{test}	**ACC**	**AUC**	**Rec./Spec.**
Within-dataset (random split)					
clean → clean	659	167	0.921±0.009	0.975±0.002	0.910 / 0.935
merged → merged	3,350	838	0.804±0.003	0.856±0.001	0.703 / 0.911
Cross-dataset transfer (decontaminated)					
clean → merged*	659	733	0.629±0.089	0.722±0.118	0.325 / 0.935
merged → clean†	**3,350**	**167**	**0.909**±0.007	**0.950**±0.004	**0.912 / 0.905**

Table 6: Independent-cohort evaluation after training exclusively on study 2538. Cohort 10317 contains healthy controls and is reported by AUROC; the remaining cohorts are disease-only and report disease identification rate. SimpleEmb is trained end-to-end over all 826 study-2538 profiles and evaluated with a logistic-regression probe on the encoder representation; HGB and Random Forest operate directly on the raw abundance vector.

Cohort	SimpleEmb	HGB	Random Forest
10317 (AUROC)	0.621±0.004	0.615	0.668±0.009
1939	0.322±0.014	0.503	0.712±0.012
11484	0.413±0.027	0.552	0.827±0.024
10283	0.085±0.006	0.134	0.132±0.013
1629	0.126±0.009	0.183	0.275

the two cohorts differ in disease prevalence, a high cohort probe could reflect disease composition rather than study-specific technical variation. We therefore repeat the cohort probe after downsampling each cohort to equal disease and healthy counts (2538: 393/393; 10317: 191/191), averaging over five independent draws. Cohort decodability is essentially unchanged (Table 7), indicating that the representations encode study-associated structure that is not explained by disease composition alone. Interestingly, the pretrained MGM encoder shows slightly *lower* cohort decodability than a random-initialized encoder of the same architecture (0.929 versus 0.964), indicating that pretraining mildly reduces—rather than introduces—cohort-specific structure, yet a large cohort-dominance gap remains in every representation.

6.4 Cohort coverage, not debiasing, improves cross-cohort generalization

The cohort-probe result suggests a natural intervention: if representations are dominated by cohort structure, then suppressing that structure should improve transfer. We test this in two ways. First, we train with a gradient-reversal (domain-adversarial) objective (Ganin et al., 2016) that explicitly penalizes cohort decodability; second, we simply expand the training set with an additional cohort that contains healthy controls (10317). We evaluate disease identification on independent disease-only cohorts.

Gradient-reversal debiasing does not improve generalization: disease identification changes by less than ±0.01 on all four independent cohorts. Unlike DEBIAS-M-style preprocessing-bias correction (Austin et al., 2025), removing cohort signal at the representation level does not recover cross-cohort generalization. In contrast, adding the 10317 cohort to the training set improves disease identification on the independent 1939 cohort from 0.322 to 0.489 (a 16.7 percentage-point gain; mean over five seeds, standard deviation 0.014), with smaller gains on 11484 (0.40 to 0.43) and no change on 10283 or 1629. The gain is stable across all five seeds. This asymmetry is informative: the benefit concentrates on cohorts whose baseline disease probability is in an ambiguous middle range (0.3–0.5), and not on cohorts where the model is already confident (disease probability ≈0.15). This suggests that some failures may reflect shifted decision boundaries that additional cohort coverage can partly correct, whereas others may arise from missing or non-transferable signal.

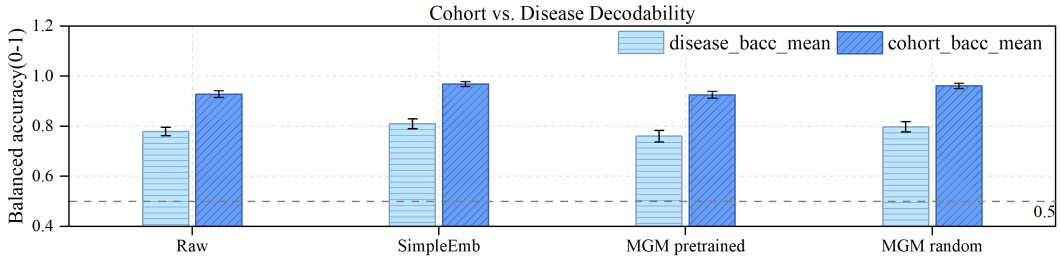

Figure 3: **Balanced accuracy (5-fold CV, 5 seeds, mean±sd) of a logistic-regression probe predicting study identity vs disease; dashed=chance(0.5). In every representation, cohort decodability (0.93-0.97) substantially exceeds disease (0.76-0.82), and the two rise together — no disentanglement for free.**

Table 7: Cohort-probe analysis. A logistic regression predicts study identity (2538 vs 10317) from each representation, reported as accuracy / balanced accuracy under 5-fold stratified cross-validation. The balanced setting downsamples each cohort to equal disease/healthy counts (five draws, averaged) to remove disease-composition confounds.

Representation	Unbalanced	Balanced
Raw abundance	0.933±0.012 / 0.929±0.014	0.929 / 0.914
SimpleEmb	0.970±0.010 / 0.970±0.010	0.973 / 0.970
MGM (pretrained)	0.930±0.013 / 0.926±0.014	0.930 / 0.922
MGM (random init)	0.964±0.010 / 0.962±0.011	0.966 / 0.958

These results substantially change the scientific conclusion. Within a cohort, many model classes can extract enough signal for high discrimination, and increasing architecture complexity produces limited gains. Across cohorts, the signal learned by all models is fragile. The central challenge is not to fit a more expressive function, nor to explicitly remove cohort signal, but to increase the cohort coverage of the training distribution so that learned decision boundaries remain valid under study-specific variation. Consistent with the probes, the representation geometry is organized by cohort rather than disease: cohort separation in cosine space is consistently, if modestly, larger than disease separation (Figure 4), a kNN decoder predicts cohort identity more accurately than disease (Figure 5), and cohort-colored UMAP projections separate by study rather than disease (Figure 6).

7 What Can an Explicit Embedding Still Provide?

Although SimpleEmb is not the strongest classifier under every protocol, its explicit representation supports analyses that are less natural with tree ensembles or a raw presence vector. Patient-level kNN retrieval in the learned space reaches 97.0% accuracy on the reported test split, and disease samples show lower intra-class cosine similarity than healthy samples (0.54 versus 0.64), consistent with greater heterogeneity in the learned disease representation. These analyses should be interpreted as properties of the learned target-task space, not as evidence that the embedding recovers a universal biological taxonomy. Indeed, literature-defined functional genus groups do not show elevated intra-group cosine similarity relative to random groups (0.0005 vs. 0.0046). The embedding therefore appears to organize taxa according to features useful for the supervised IBD task rather than reproducing known functional groupings, reinforcing its role as a task-specific analytical representation.

7.1 Perturbation-based evidence

LOO attribution measures how much the model prediction changes when an observed genus is removed. Across five folds, global importance rankings have mean pairwise Spearman correlation $\rho = 0.72$, while the top-50 Jaccard overlap is only 0.07 (Table 8). The deletion test is more directly informative: removing the top-50 LOO-ranked genera reduces AUROC by 0.039, compared with

Table 8: **LOO attribution reliability across 5 folds.**

Metric	Value
Spearman ρ (mean across fold pairs)	0.72 ($p<0.001$)
Top-50 Jaccard (mean)	0.07
Deletion test (top-50 AUC drop)	0.039 (vs 0.007 random)
Permutation control (mean $\|$LOO$\|$)	real 0.0096 vs. shuffled 0.0705
Literature direction agreement	5/6 (83.3%)
Prevalence-importance correlation	r=0.05 ($p>0.05$)

0.007 for random deletion. Importance is essentially uncorrelated with prevalence ($r = 0.05$), and among the six literature-curated IBD genera present in the vocabulary (e.g., Faecalibacterium and Roseburia (Sokol et al., 2008; Machiels et al., 2014)), five have the reported direction of association (Table 8). These results support LOO as predictive evidence, while not implying biological causality.

The relationship with differential abundance is also weak. Across all genera the Spearman correlation is -0.23 ($p < 0.001$), but this weak negative association vanishes at higher prevalence (0.11, $p = 0.269$ for prevalence ≥ 50). Thus, the classifier's perturbation sensitivity is not reducible to a ranking of the largest marginal abundance shifts. Whether this reflects reproducible multivariate biology or cohort-specific interactions remains unresolved by the present data and is precisely why external robustness is important.

7.2 Evidence-grounded language-model interpretation

We use the LLM only after prediction, as an interface for verbalizing model evidence. Supplying the classifier output and top LOO evidence substantially reduces hallucinated genera, raises prediction consistency, and improves direction accuracy compared with prompting from the raw genus list alone.

This is intentionally not evidence that an LLM should perform microbiome diagnosis: direct few-shot prompting performs far worse than the trained classifier. The useful separation is between *prediction* and *communication*: a specialized model produces the prediction and evidence, and the LLM is constrained to explain it. We treat this as a downstream demonstration rather than the main contribution.

8 Discussion and Conclusion

Our results do not support either "small models beat foundation models" or the reverse. Within cohort, pretrained MGM, task-trained Transformers, permutation-invariant models, and simple classical baselines achieve similar performance. Under grouped validation, a presence-vector MLP is already highly competitive, indicating that neither pretraining nor representation complexity alone determines downstream utility.

SimpleEmb should therefore be viewed as a diagnostic representation rather than a universally superior classifier. Its main value is analytical: it provides a compact continuous space for retrieval and perturbation, while supporting sample-specific evidence that can be verbalized by a language model.

The dominant difficulty instead emerges under cohort shift. Study identity is more readily decoded than disease across all tested representations, gradient-reversal debiasing does not improve transfer, and adding an additional training cohort improves disease identification on one independent cohort by 16.7 percentage points. These results suggest that broader cohort coverage may be more important than further model scaling or explicit invariance objectives. However, our experiments are limited primarily to IBD at genus-level resolution, and the cohort-coverage gain is demonstrated with one added training cohort; whether this trend generalizes across diseases, taxonomic resolutions, and larger numbers of cohorts remains open.

Reproducibility Statement

Architecture and hyperparameters are described in Section 4; evaluation protocols in Sections 4–5; the cohort-probe and cohort-coverage analyses are described in Section 5. Model checkpoints and reproduction scripts are included in the anonymized repository.

AI Use Statement

In this work, we used generative AI tools to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, and to implement methods. We have not used generative AI tools for other tasks requiring disclosure. We reviewed all AI-assisted work and take responsibility for the final content, including text, claims, or artifacts produced with the aid of generative AI.

References

George I. Austin et al. Processing-bias correction with DEBIAS-M improves cross-study generalization of microbiome-based prediction models. *Nature Microbiology*, 10:897–911, 2025.

Claire Duvallet et al. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8:1784, 2017.

Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, 4(2):293–305, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Dirk Gevers et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host & Microbe*, 15(3):382–392, 2014.

Sean M. Gibbons et al. Correcting for batch effects in case-control microbiome studies. *PLoS Computational Biology*, 14(4):e1006102, 2018.

Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jonas Halfvarson et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2:17004, 2017.

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207–214, 2012.

Wodan Ling et al. Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, 13:5418, 2022.

Jason Lloyd-Price et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569:655–662, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kathleen Machiels, Marie Joossens, João Sabino, Vicky De Preter, Ingrid Arijs, Venessa Eeckhaut, Vera Ballet, Karolien Claes, Filip Van Immerseel, Kristin Verbeke, et al. A decrease of the butyrate-producing species roseburia hominis and faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut*, 63(8):1275–1283, 2014.

Nicholas A Medearis, Siyao Zhu, and Ali R Zomorrodi. Biomegpt: A foundation model for the human gut microbiome. *bioRxiv*, 2026. doi: 10.64898/2026.01.05.697599.

Jin Mu, Zheng-Zheng Tang, and Guanhua Chen. Systematic benchmarking of foundation models and classical baselines for microbiome-based disease prediction. *Research Square*, 2026. doi: 10.21203/rs.3.rs-8912605/v1.

Min Oh and Liqing Zhang. Deepmicro: deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10, 2020. doi: 10.1038/s41598-020-63159-5.

Edoardo Pasolli et al. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Computational Biology*, 12(7):e1004977, 2016.

Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of crohn disease patients. *Proceedings of the National Academy of Sciences*, 105(43):16731–16736, 2008.

Neythen J. Treloar, Saif Ur-Rehman, and Jenny Yang. Learning the language of the microbiome with transformers. *bioRxiv*, 2026. doi: 10.64898/2026.05.02.722381.

Ivan Vujkovic-Cvijin et al. Host variables confound gut microbiota studies of human disease. *Nature*, 587:448–454, 2020.

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, 2017.

Jakob Wirbel, Konrad Zych, Morgan Essex, Nicolai Karcher, Ece Kartal, Guillem Salazar, Peer Bork, Shinichi Sunagawa, and Georg Zeller. Microbiome meta-analysis and cross-disease comparison enabled by the siamcat machine learning toolbox. *Genome Biology*, 22, 2021. doi: 10.1186/s13059-021-02306-1.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haohong Zhang, Yuli Zhang, Zixin Kang, Jiayun Xiong, Ronghua Yang, and Kang Ning. Mgm as a large-scale pretrained foundation model for microbiome analyses in diverse contexts. *Advanced Science*, 2026. doi: 10.1002/advs.202513333.

Table 9: **Ablation: contribution of each architectural component.** All variants use the identical MLP architecture. Δ shows improvement over variant A.

Variant	ACC	AUC	Prec.	Rec.	F1	ΔACC
A: Raw→MLP	0.889±0.011	0.959±0.002	0.935±0.005	0.860±0.023	0.896±0.012	—
B: RandomEmb→MLP	0.932±0.006	0.978±0.005	0.962±0.005	0.914±0.010	0.937±0.006	+4.31%
C: SimpleEmb→Linear	0.910±0.004	0.968±0.003	0.928±0.005	0.910±0.009	0.919±0.004	+2.16%
D: SimpleEmb→MLP	0.923±0.007	0.974±0.003	0.951±0.008	0.910±0.013	0.930±0.007	+3.47%

A Supplementary Tables and Figures

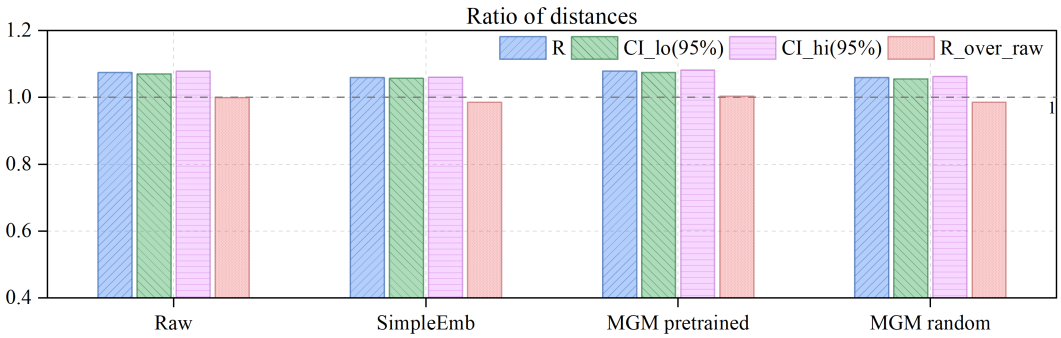

Figure 4: **Cohort separation is consistently but only modestly larger than disease separation in cosine geometry.**

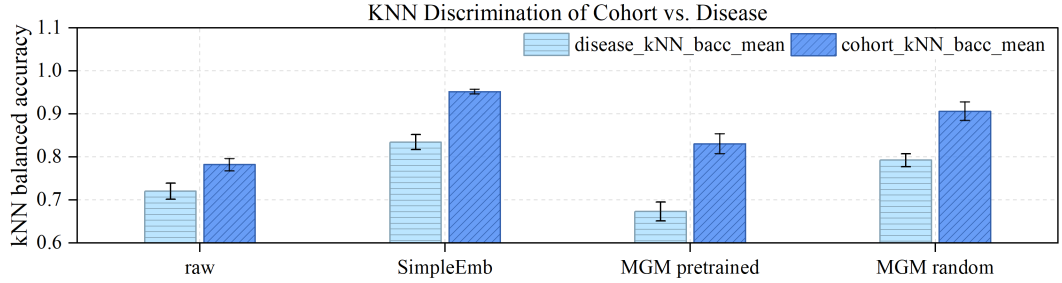

Figure 5: k**NN (**k=5**, cosine) balanced accuracy of cohort-versus-disease decoding.**

B Limitations

The study has limitations. Only IBD is evaluated, so conclusions may not generalize to other phenotypes. Independent-cohort performance remains poor, so the model is not yet a clinically transferable diagnostic. LOO attribution measures predictive sensitivity, not causal effect. Literature validation covers only 6 of 20 curated IBD-associated genera. The LLM benchmark uses automatic faithfulness criteria without expert validation. Finally, the cohort-coverage gain is shown on one added cohort (10317) and one target (1939); how far it scales remains open.

Code and Data Availability

Code for training, evaluation, and analysis is available at an anonymized repository: `https://anonymous.4open.science/r/iclr-D07F/`. The datasets are assembled from publicly available Qiita studies (2538, 10283, 1939, 10317, 11484, 1629) together with the American Gut

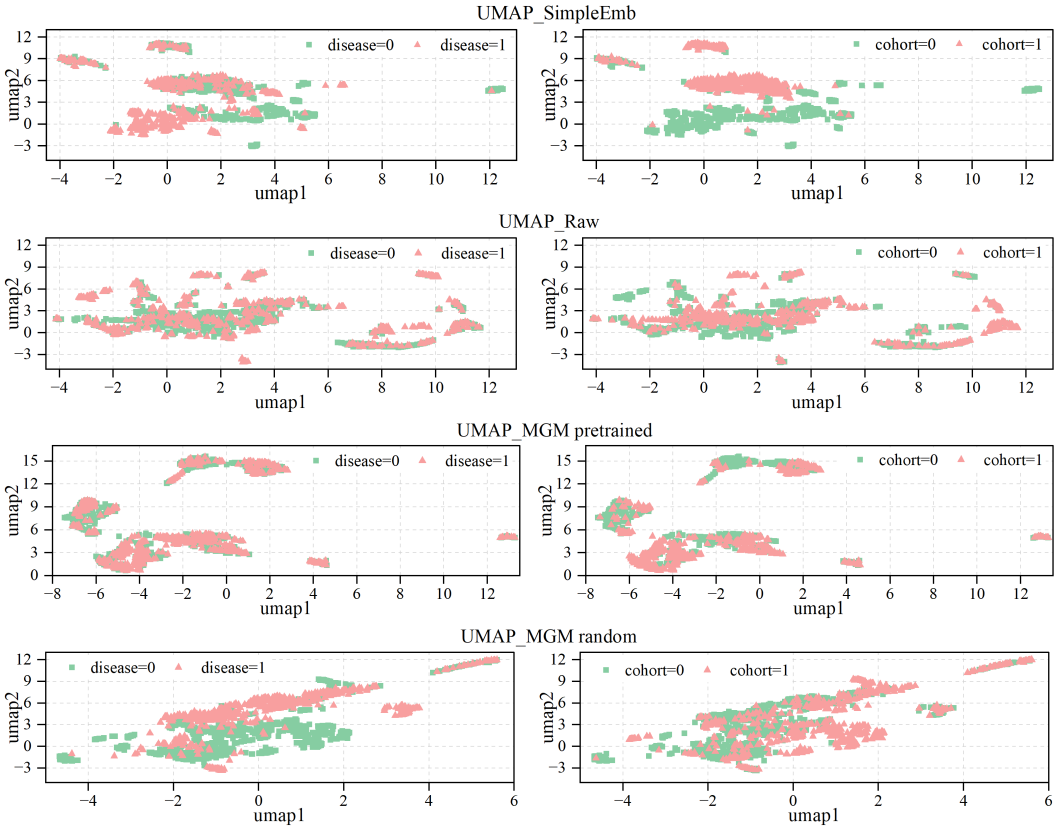

Figure 6: **Cohort-colored UMAP**

Table 10: **Pairwise McNemar tests (SimpleEmb vs. baselines on grouped-CV predictions).** n_{disc} = discordant pairs; B / C = pairs where only SimpleEmb / only the baseline is correct. q_{BH} = Benjamini–Hochberg FDR-corrected p-value. Bold = significant at $q < 0.05$.

Baseline	n_{disc}	B	C	q_{BH}
KNN	329	213	116	1.2×10^{-6}
BernoulliNB	375	216	159	1.5×10^{-2}
MGM + Logistic Regression	286	161	125	1.1×10^{-1}
MGM + Random Forest	406	186	220	2.4×10^{-1}
Logistic Regression	197	88	109	2.8×10^{-1}
RBF-SVM	228	125	103	2.8×10^{-1}
HistGradientBoosting	139	75	64	5.9×10^{-1}
Linear SVM	278	133	145	6.8×10^{-1}
MGM + MLP	299	154	145	7.7×10^{-1}
MLP	235	115	120	8.7×10^{-1}
Extra Trees	156	79	77	9.4×10^{-1}

Project (AGP) and Fecal Transplantation Program (FTP) cohorts; processed genus-abundance vectors and labels are available from the authors upon request.

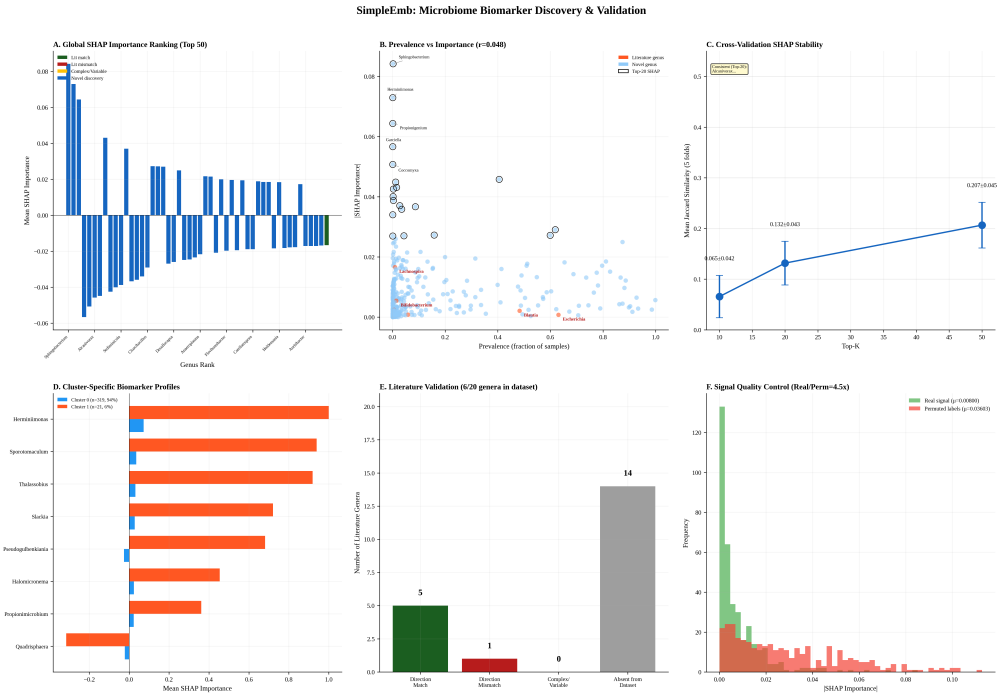

Figure 7: **Comprehensive LOO attribution analysis.** (A) Global importance ranking (top-50 genera, color-coded by literature agreement). (B) Prevalence vs. importance. (C) Cross-validation stability (Jaccard). (D) Cluster-specific biomarker profiles. (E) Literature direction validation. (F) Signal quality: real vs. permuted-label importance.

Table 11: **Calibration metrics and error analysis on the held-out test set** (n=167)**.**

Metric	**Value**
ECE (Expected Calibration Error)	0.0496
Brier Score	0.0555
Accuracy	0.9281
AUC	0.9802
Sensitivity	0.9247
Specificity	0.9324
Error breakdown ($12/167 = 7.2\%$)	
False positives (Healthy→IBD)	5 (conf. 0.71–0.86)
False negatives (IBD→Healthy)	7 (conf. 0.007–0.056)

Table 12: **Few-shot LLM evaluation on IBD diagnosis** (n=100)**.** Qwen2-7B-Instruct is given k examples of (genus list, diagnosis) pairs.

Method	**ACC**	**F1**	**Uncertain**	**Correct/Total**
LLM (k=0, zero-shot)	0.420	0.296	0	42/100
LLM (k=1, 1-shot)	0.596	0.594	1	59/99
LLM (k=3, 3-shot)	0.550	0.549	0	55/100
LLM (k=5, 5-shot)	0.620	0.618	0	62/100
SimpleEmb (trained)	0.920	—	0	92/100

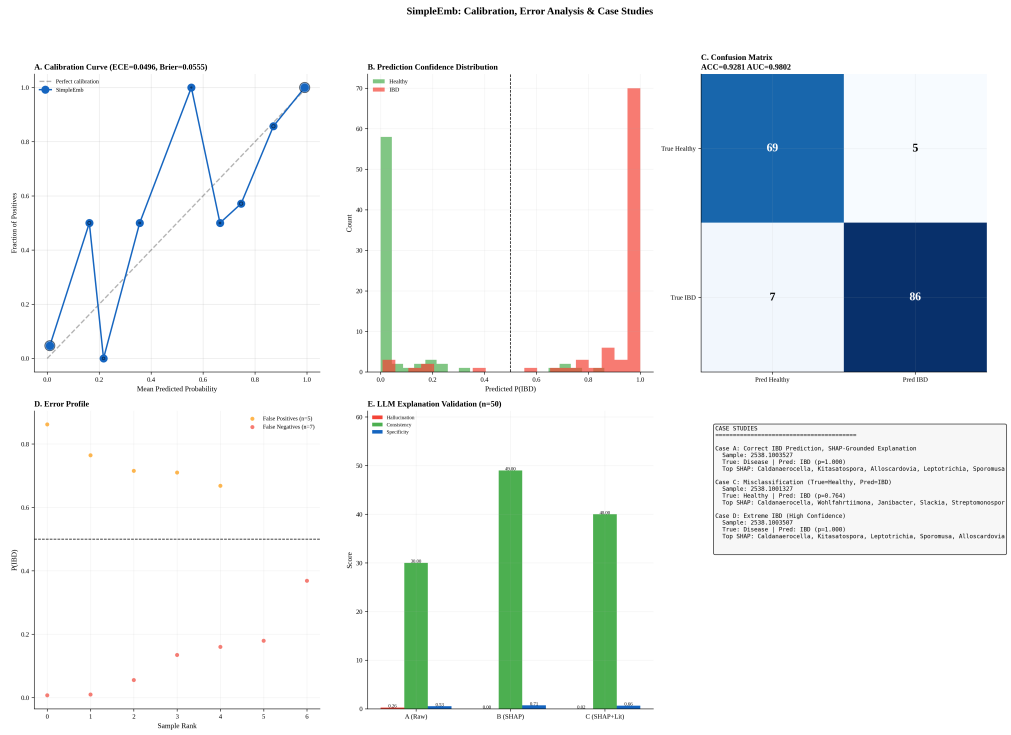

Figure 8: **Calibration, error analysis, and case studies.** (A) Calibration curve (ECE=0.05, Brier=0.056). (B) Prediction confidence distribution. (C) Confusion matrix. (D) Error profile. (E) LLM explanation validation summary. (F) Case study overview.

Table 13: **Stratified LLM explanation quality (LOO+LLM variant, 167 samples).**

Stratum	n	Hallucination	Consistency	Direction
Healthy samples	74	0.42	0.97	0.92
IBD samples	93	0.31	0.99	0.95
Correct predictions	155	0.37	0.98	0.94
Wrong predictions	12	0.17	1.00	0.91

Table 14: **Comparison of microbiome foundation models.** SimpleEmb achieves competitive performance with 30× fewer parameters and no pretraining, while providing per-sample LOO attribution and LLM-based interpretation.

Model	Architecture	Granularity	Pretraining	Interpretability	Weights
MGM	8L Transformer (34M)	Genus	Next-genus, 263k	Attention + LOO	Public
BiomeGPT	Transformer + dual emb	Species	Masked abundance, 13.3k	CLS attention	Unavailable
Waypoint	Transformer (6–170M)	Genus	539k+	Not emphasized	Unavailable
SimpleEmb	SimpleEmb + MLP (1.1M)	Genus	None	LOO + LLM	Public

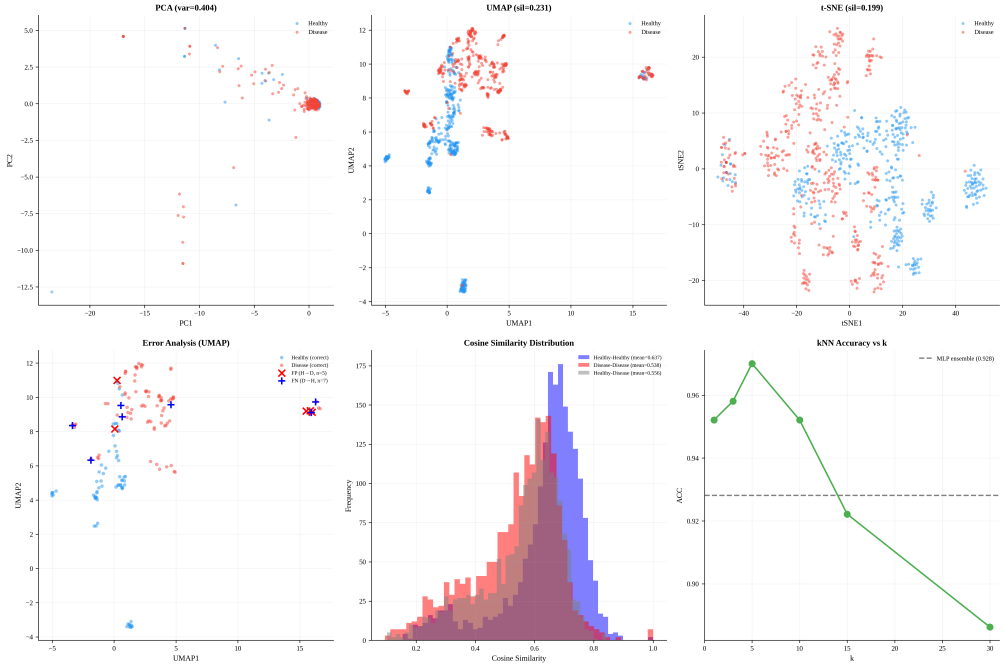

Figure 9: **Representation space analysis.** (A) PCA 2D projection. (B) UMAP projection (cosine metric). (C) t-SNE projection. (D) Prediction error analysis. (E) Intra-class cosine similarity. (F) kNN classification accuracy.

Table 15: Dataset summary statistics.

Property	**clean_2538**	**merged_all**
Training samples	659	3,350
Test samples	167	838
Disease (IBD) fraction	55.7%	50.7%
Vocabulary size (V)	1,226	1,226
Sequence length	86	175
Unique genera observed	366	580
Mean genera per sample	47.8 ± 16.2	—

Table 16: **Nested grouped-CV learning curves (AUROC).** The training fraction is sampled from only the outer training groups.

Train fraction	**Embedding64+MLP (ours)**
20%	0.900 ± 0.027
40%	0.929 ± 0.022
60%	0.947 ± 0.015
80%	0.951 ± 0.017
100%	0.956 ± 0.016

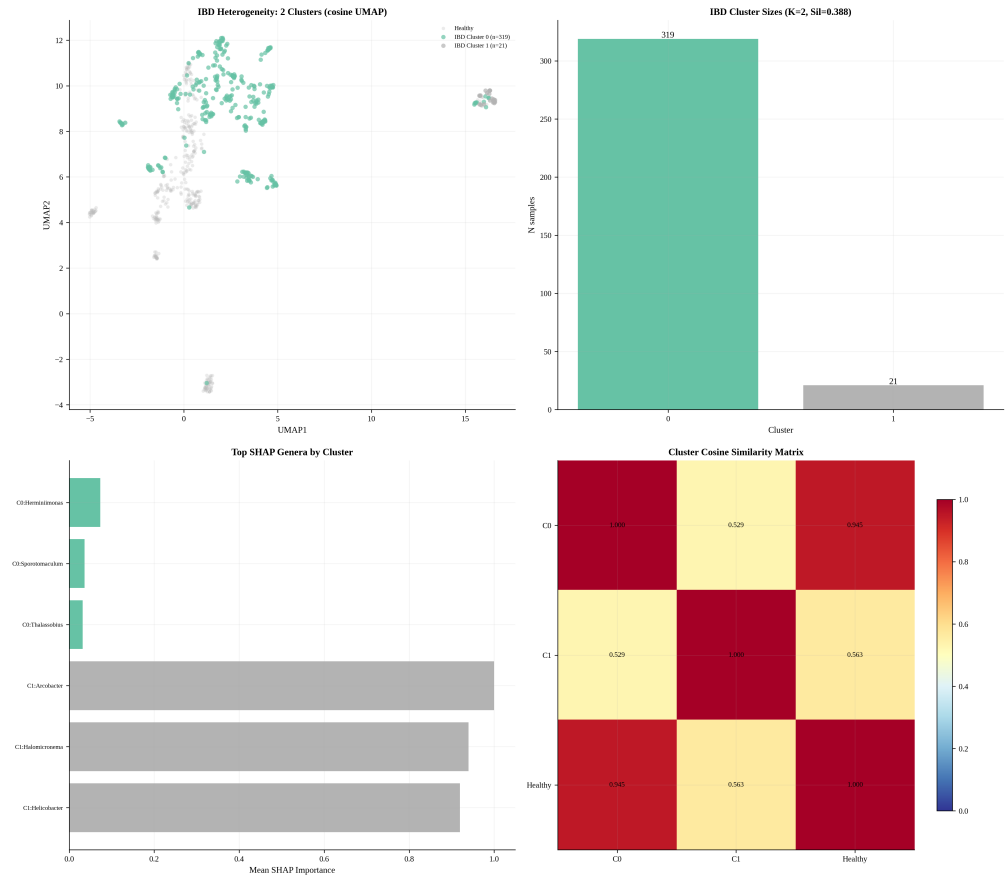

Figure 10: **IBD heterogeneity analysis.** (A) UMAP with IBD clusters colored separately from healthy samples. (B) Cluster sizes with silhouette score. (C) Top LOO-attributed genera per cluster. (D) Cluster cosine similarity matrix.

Table 17: Embedding space quality metrics.

Metric	**Value**
kNN (k=5, cosine) test ACC	97.0%
PCA 2D explained variance	40.4%
IBD intra-class cosine similarity	0.54
Healthy intra-class cosine similarity	0.64
Cross-class cosine similarity	0.56
Functional-group intra-group cosine	0.0005 (vs. 0.0046 random)

Table 18: **LOO attribution vs. differential abundance (Spearman ρ).**

Prevalence Strata	ρ	p-value
All genera (n=365)	-0.23	<0.001
Prevalence $\geq$10 (n=186)	0.01	0.901
Prevalence $\geq$50 (n=108)	0.11	0.269
Prevalence $\geq$200 (n=66)	0.16	0.188

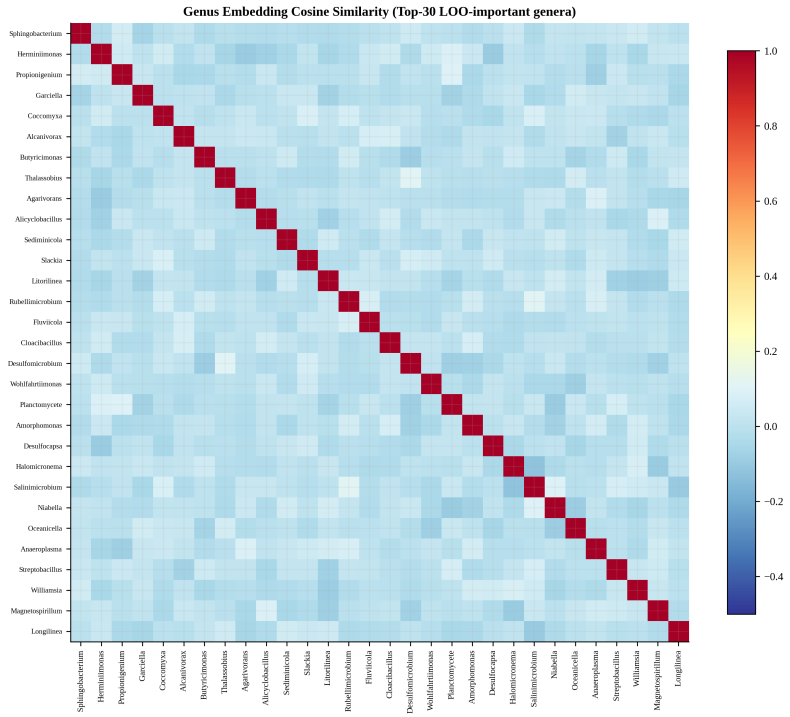

Figure 11: **Genus embedding similarity heatmap.** Cosine similarity among the top-30 LOO-important genera in the learned embedding space.

Table 19: **Out-of-fold case analysis (grouped split, seed 42).** Values are disease probabilities from three models on the same held-out samples. Extreme low-richness samples expose failure modes of all models.

Case	Truth	Active tokens	SimpleEmb	HGB	MGM
SimpleEmb correct; HGB wrong	Healthy	86	H (0.003)	D (0.883)	D (0.887)
HGB correct; SimpleEmb wrong	Disease	86	H (0.104)	D (0.997)	H (0.175)
SimpleEmb correct; MGM wrong	Disease	4	D (1.000)	D (0.749)	H (0.398)
High-conf. false positive	Healthy	1	D (1.000)	D (0.824)	D (1.000)
High-conf. false negative	Disease	2	H (0.000)	D (0.749)	D (0.889)
All models correct	Disease	23	D (1.000)	D (1.000)	D (1.000)

Table 20: **LLM explanation validation (all 167 test samples, Qwen2-7B-Instruct).** Bold indicates best.

Metric	Raw LLM	LOO+LLM	LOO+Lit+LLM
Hallucinated genera (avg/response)	0.72	**0.36**	0.35
Input genera mentioned (avg)	5.47	9.02	**9.05**
LOO consistency (top-15)	0.70	0.99	**1.00**
Direction accuracy	0.27	**0.94**	0.93
Prediction consistency (%)	63%	**98%**	98%

Table 21: **Module-transfer ablation on the clean-2538 holdout** (5 seeds, mean±std; bold = per-column best).

Variant	Accuracy	AUROC	AUPRC
A0: SimpleEmb (baseline)	0.9246±0.0072	0.9753±0.0046	0.9829±0.0028
A1: +rank-bin (MGM)	0.9377±0.0029	0.9774±0.0029	0.9826±0.0027
A2: +abundance-bin	0.9293±0.0116	0.9749±0.0045	0.9797±0.0036
A3: gated pooling	0.9162±0.0066	0.9664±0.0045	0.9749±0.0060
A4: +masked-abundance pretraining	0.9126±0.0081	0.9734±0.0041	0.9790±0.0032
A6: +token dropout (DeepMicro)	**0.9389±0.0045**	**0.9822±0.0022**	**0.9851±0.0022**
A8: linear head	0.9114±0.0045	0.9681±0.0025	0.9783±0.0015

Table 22: **Retention test under the 15-fold leakage-robust grouped-CV protocol** (3 seeds × StratifiedGroupKFold(5)).

Candidate	AUROC wins	CI-sig. folds	ΔAUPRC	Verdict
A1 rank-bin	9/15	2/15	+0.002	rejected
A6 token dropout	**15/15**	8/15 (rule: ≥10)	+0.020	not retained; follow-up

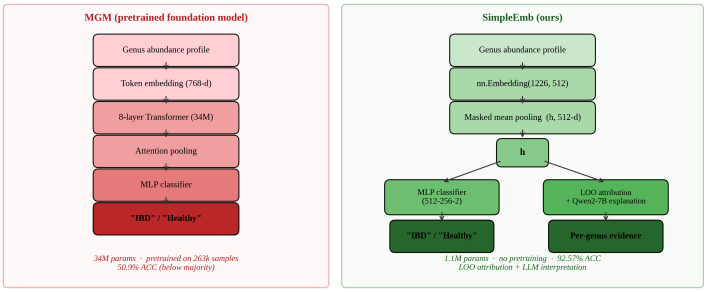

Figure 12: **Architecture comparison.** SimpleEmb (right) maps each present genus to a learned embedding, summarizes by masked mean pooling, and predicts IBD; the same embedding space powers LOO attribution, verbalized by a frozen Qwen2-7B. MGM (left) is a 34M-parameter pretrained Transformer.

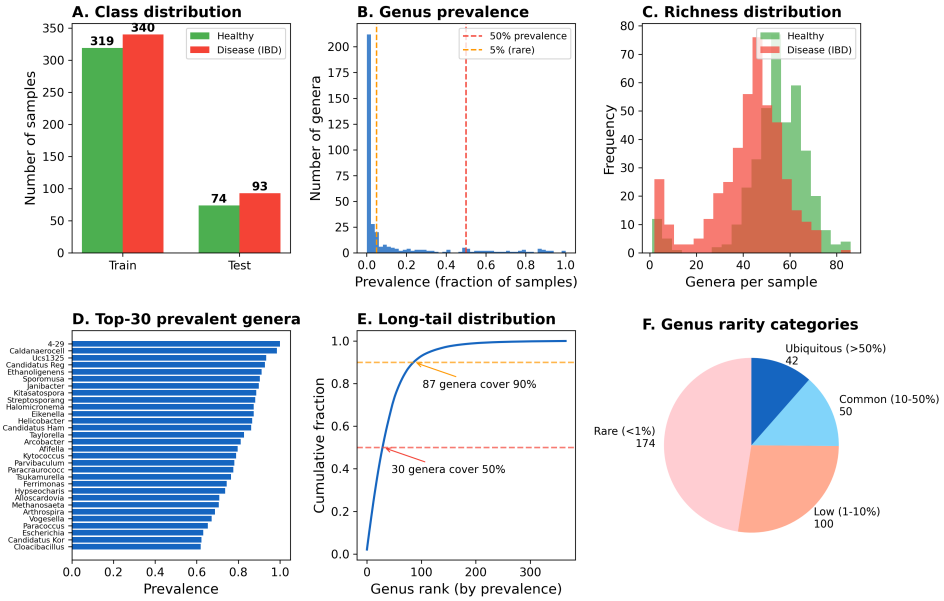

Figure 13: **Dataset statistics for clean_2538.** (A) Class distribution. (B) Genus prevalence distribution. (C) Richness distribution. (D) Top-30 most prevalent genera. (E) Long-tail cumulative distribution.

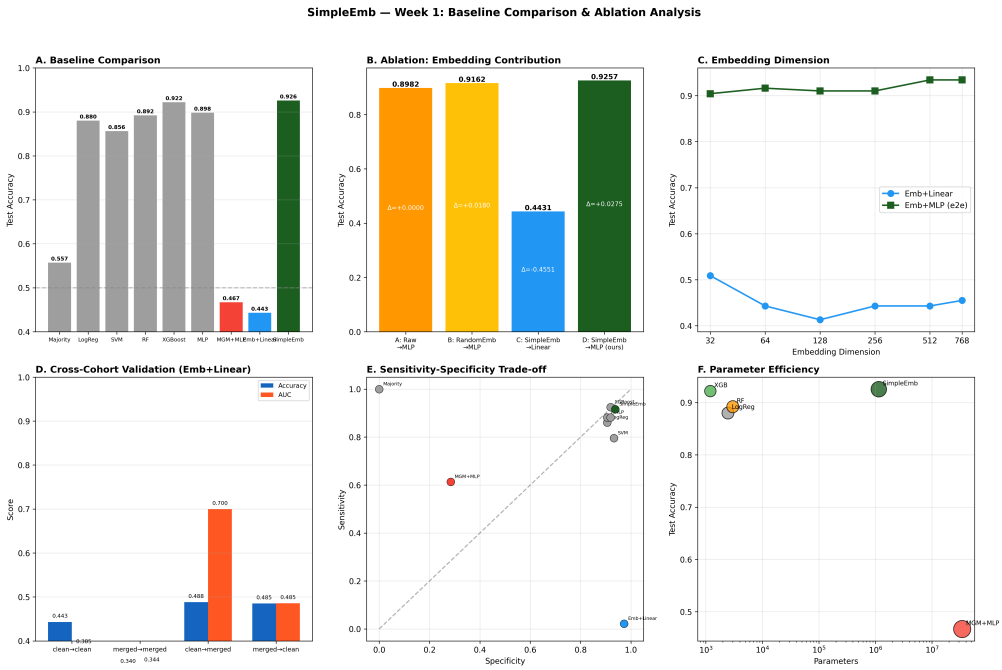

Figure 14: **Comprehensive comparison analysis.** (A) Baseline accuracy comparison. (B) Ablation: embedding contribution. (C) Embedding dimension sweep. (D) Cross-dataset transfer. (E) Sensitivity–specificity trade-off. (F) Parameter efficiency.

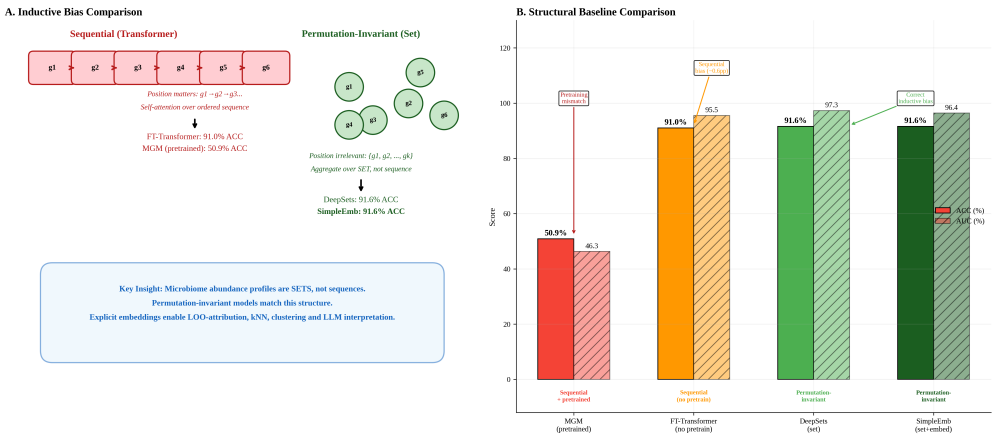

Figure 15: **Inductive bias comparison.** (A) Sequential (Transformer) vs. permutation-invariant (Set) inductive biases. (B) Structural baseline accuracy and AUROC comparison.

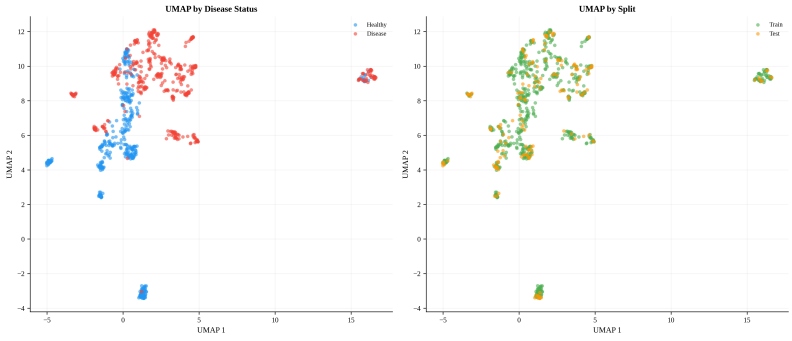

Figure 16: **UMAP visualization of learned embeddings.** Disease and healthy samples in the 2D UMAP projection of the 768-dimensional embedding space.



Figure 17: **Biomarker robustness analysis.** Cross-fold consistency, permutation control, abundance-independence, and cluster validation of LOO-attributed biomarkers.

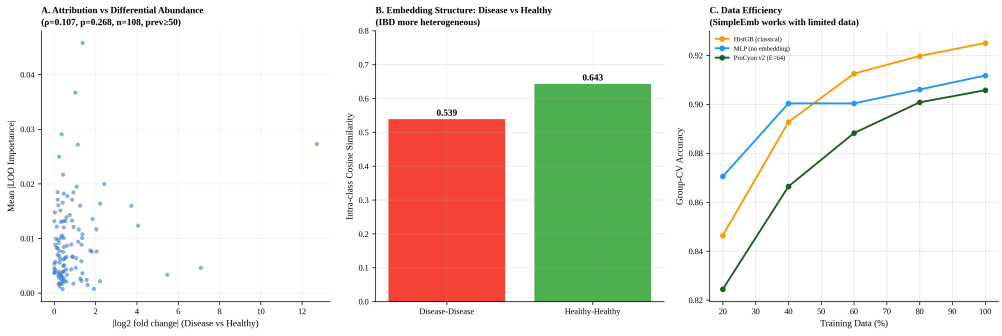

Figure 18: **Attribution biology and embedding structure.** (A) LOO importance vs. $|\log_2$ fold change$|$ scatter (prevalence $\geq$50). (B) Intra-class cosine similarity: disease vs. healthy. (C) Data efficiency comparison.