TAXOMFM: A TAXONOMY-AWARE FOUNDATION MODEL FOR THE HUMAN GUT MICROBIOME

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing gut microbiome foundation models typically use flat taxon representations, limiting their ability to exploit known taxonomic hierarchies. We introduce TaxoMFM, a microbiome foundation model that integrates taxonomic structure into self-supervised learning. TaxoMFM jointly learns abundance patterns across taxonomic ranks and relative proportions within groups through taxonomy-aware encoding and hierarchical compositional reconstruction. We further construct TaxoGut-HPP, comprising 100,806 gut microbiome profiles from 231 source cohorts and covering 40 classification and 8 regression tasks. TaxoMFM achieves competitive performance across disease, gut health, host trait, and liver ultrasound phenotype prediction. In cross-cohort zero-shot transfer, where target cohorts are excluded from both pretraining and downstream adaptation, TaxoMFM achieves AUROCs of 0.838, 0.736, and 0.722 for colorectal cancer, type 2 diabetes, and inflammatory bowel disease, respectively. Further analyses show that TaxoMFM recovers community changes across taxonomic scales, while preserving community information and producing consistent representations under database updates and species-vocabulary shifts. Our Code is available on GitHub.

1 INTRODUCTION

Foundation models learn general representations from large unlabeled datasets and transfer them to a wide range of downstream tasks (Bommasani et al., 2021). This approach is well suited to gut microbiome data. Gut microbiome profiles are high-dimensional and sparse (Oh & Zhang, 2020). Recent gut microbiome foundation models combine multiple public cohorts for self-supervised pretraining and transfer the learned representations to disease prediction tasks (Pope et al., 2025; Medearis et al., 2026; Karwowska et al., 2026). These studies show the potential of pretraining for microbiome modeling across tasks.

However, microbial signals associated with diseases and host phenotypes can occur at several taxonomic scales. Changes may involve a single species, an entire genus, or a family. For example, a genus can increase in total relative abundance, or maintain the same total while its species proportions change. The taxonomic hierarchy shows how abundant each group is in the community and how its members contribute to that abundance. Current gut microbiome foundation models generally use flat taxon representations that do not explicitly encode these relationships. Previous statistical and machine learning studies have shown that taxonomic or phylogenetic information can help characterize community differences and support downstream prediction (Silverman et al., 2017; Reiman et al., 2020; Jiang et al., 2025). We ask how self-supervised pretraining can use this hierarchy to preserve both abundance distributions across taxonomic scales and composition within lineages.

Existing evaluations largely focus on case–control classification in public disease cohorts. Gut symptoms, such as abdominal pain and bloating, and continuous physiological traits receive less attention. This narrow scope leaves it unclear whether the learned representations capture variation in gut health and host physiology beyond disease status.

To address the modeling limitation, we introduce TaxoMFM, a foundation model that explicitly incorporates the taxonomic hierarchy into both encoding and reconstruction. Its encoder uses relation-biased self-attention to jointly model species, genus, family, and order. Its decoder predicts child proportions within each parent and reconstructs abundances along the taxonomy, ensuring that child

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.

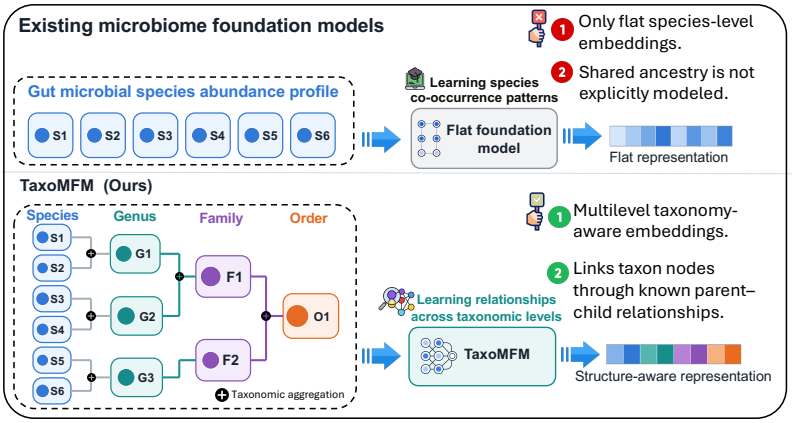


Figure 1: Conceptual comparison of flat and taxonomy-aware microbiome foundation models. Both approaches use the same species-level abundance profile. Existing models represent species as a flat set, whereas TaxoMFM aggregates their abundances along known taxonomic lineages and jointly models species, genus, family, and order nodes to learn a structure-aware representation.

abundances sum to their parent's abundance. Pretraining combines multiscale composition reconstruction with hierarchical composition learning, allowing the model to learn both group abundances and the relative proportions of their members. Together, these components provide a way to learn representations from compositional data with a known hierarchy. Fig. 1 contrasts this approach with flat species-level pretraining.

To broaden the data and evaluation coverage, we construct TaxoGut-HPP, comprising 100,806 gut microbiome profiles from 231 source cohorts. It covers 40 classification and 8 regression tasks spanning disease, gut health, host traits, and continuous clinical phenotypes. This coverage extends evaluation beyond disease classification to gut symptoms and quantitative physiological measurements. Table 1 compares its scale and phenotype coverage with existing pretraining resources. Using TaxoGut-HPP, we evaluate TaxoMFM through downstream prediction, cross-cohort transfer, ablations, and analyses of multiscale recovery and consistency across taxonomic vocabularies.

Our main contributions are as follows:

1. We introduce TaxoMFM, combining taxonomy-aware encoding and hierarchical compositional decoding to learn transferable microbiome representations through joint reconstruction of multiscale abundances and within-lineage proportions.
2. We construct TaxoGut-HPP with 100,806 profiles from 231 source cohorts, expanding pretraining data and evaluation across disease, gut health, host traits, and continuous clinical phenotypes.
3. TaxoMFM outperforms existing foundation models and conventional machine learning methods across a broad range of downstream tasks. Ablations and structural analyses assess the contributions of taxonomy and hierarchical composition learning.

Table 1: Comparison of pretraining and downstream evaluation datasets used by human gut microbiome foundation models.

Model	Size	Cohorts	Downstream task counts			Downstream task coverage			
			Total	Classification	Regression	Disease	Gut health	Host traits	Clinical phenotype
Pope et al. (Pope et al., 2025)	18,480	1	3	3	0	✓	×	✓	×
BiomeGPT (Medearis et al., 2026)	13,524	86	27	27	0	✓	×	×	×
GUT-FORMer (Karwowska et al., 2026)	21,024	93	27	25	2	✓	×	✓	×
MBEmbed (Zahavi et al., 2025)	85,364	3	7	4	3	✓	×	✓	✓
TaxoMFM (ours)	**100,806**	**231**	**48**	**40**	**8**	✓	✓	✓	✓

2 Related Work

2.1 Microbiome foundation models

Recent studies have applied self-supervised pretraining to microbiome profiles. Pope et al. (2025) use an ELECTRA-style objective to detect replaced microbial tokens. MGM (Zhang et al., 2026) applies causal language modeling to genus-level profiles, while Waypoint (Treloar et al., 2026) studies causal pretraining with both genus- and species-level tokenization. For the human gut microbiome, BiomeGPT (Medearis et al., 2026) models species abundance profiles, GUT-FORMer (Karwowska et al., 2026) combines taxonomic and functional profiles, and MBEmbed (Zahavi et al., 2025) learns sample representations from large gut cohorts. These models generally treat each taxon as a separate input unit and learn microbial relationships mainly from co-occurrence. They do not explicitly model the known hierarchy among species, genera, families, and orders during pretraining.

2.2 Taxonomy-aware microbiome learning

DeepMicro (Oh & Zhang, 2020) used autoencoders to learn low-dimensional representations from microbial abundance profiles. Other approaches explicitly incorporated taxonomic or phylogenetic relationships. Ph-CNN (Fioravanti et al., 2018) used phylogenetic distances to define convolutional neighborhoods, while PopPhy-CNN (Reiman et al., 2020) arranged abundances at leaf and ancestral nodes into a tree-based matrix for convolution. TaxoNN (Sharma et al., 2020) grouped OTUs by phylum and combined features learned per group. More recently, DeepPhylo (Wang et al., 2024) combined phylogenetic distance embeddings with abundance information for unsupervised analysis and supervised prediction. PhyloGCNE (Dong et al., 2026) represented microbiome samples as phylogenetic graphs with leaf and ancestral nodes, using edge features to guide message passing. In contrast, TaxoMFM uses relation-biased attention to exchange information across taxonomic ranks. Its hierarchical decoder predicts child proportions within each parent, ensuring that reconstructed child abundances sum to their parent's abundance. Pretraining jointly matches multiscale abundance profiles and within-parent compositions to learn reusable sample representations.

3 TaxoMFM

TaxoMFM consists of a taxonomy-aware encoder and a hierarchical compositional decoder. The encoder uses known taxonomic relations to learn a structure-aware representation of each microbiome sample. During pretraining, species and taxonomic subtrees in the encoder input are masked, and the decoder provides the training signal for the encoder by reconstructing complete abundance profiles through the known hierarchy. After pretraining, the decoder is discarded and the encoder is used for downstream tasks. Fig. 2 summarizes the full framework.

3.1 Taxonomy-aware encoder

TaxoMFM takes a species-level abundance profile together with the taxonomic lineage of each species, which is typically available in existing microbiome datasets. Using these lineages, species assigned to the same ancestor are aggregated as

$$x_j^{(\ell)} = \sum_{i:\,\mathrm{ancestor}_\ell(i)=j} x_i^{(\mathrm{species})}, \qquad \ell \in \{\mathrm{genus}, \mathrm{family}, \mathrm{order}\}. \tag{1}$$

Here, $\mathrm{ancestor}_\ell(i)$ denotes species i's ancestor at rank ℓ, so higher-rank abundances are computed from the species input. For each sample, the detected species and their ancestors form a sample-specific active tree. The encoder takes the taxa and abundances from all four levels in this tree as input. For each active node v, x_v and ℓ_v denote its abundance and taxonomic rank, respectively. During pretraining, s_v records whether the node is observed, masked, or partially observed; these states are defined in Section 3.2. We represent the node as

$$\mathbf{h}_v = \mathbf{E}_{\mathrm{taxon}}[v] + \mathbf{E}_{\mathrm{rank}}[\ell_v] + \mathrm{MLP}(x_v) + \mathbf{E}_{\mathrm{state}}[s_v]. \tag{2}$$

Here, $\mathbf{E}_{\mathrm{taxon}}[v]$ is a learned embedding for taxon v itself, while $\mathbf{E}_{\mathrm{rank}}[\ell_v]$ indicates whether it is a species, genus, family, or order node. The MLP maps its scalar abundance x_v into the hidden space. The state embedding is used only for masked pretraining.

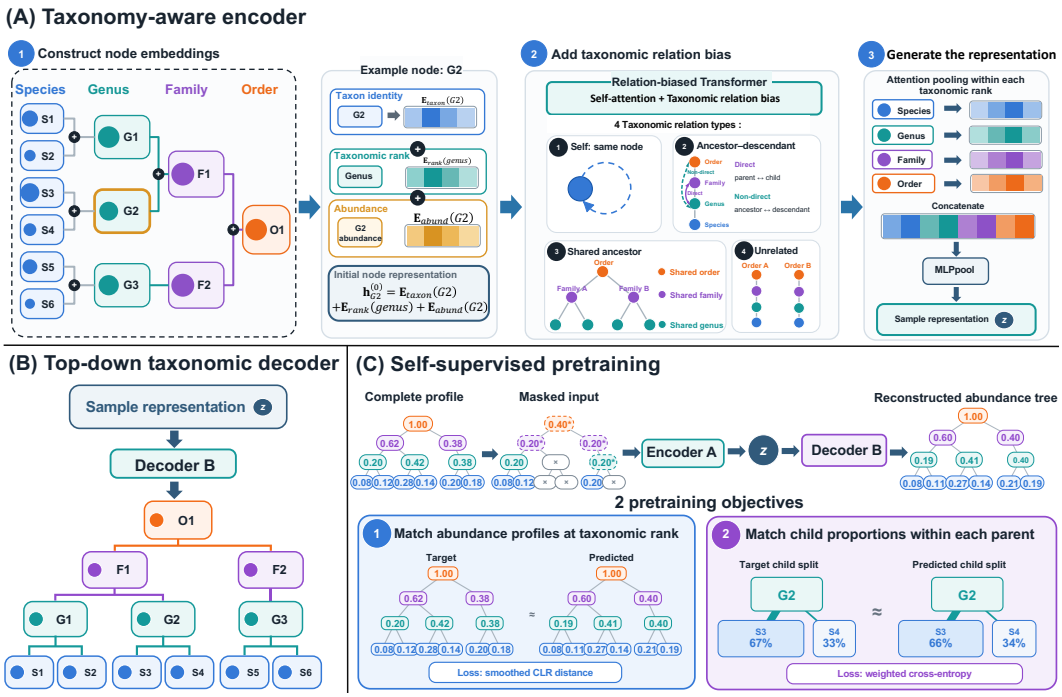


Figure 2: TaxoMFM architecture and self-supervised pretraining. (A) The encoder combines taxon identity, taxonomic rank, and abundance embeddings with relation-biased self-attention and rank-wise pooling to generate the sample representation z. (B) The decoder predicts child proportions, reconstructing abundance profiles top-down through the taxonomy. (C) Pretraining masks species or subtrees and optimizes full-tree reconstruction and within-parent child-proportion matching.

The encoder uses the taxonomy to guide information exchange between active nodes. For every ordered pair (u, v), the taxonomy defines a relation type $r_{uv} \in \mathcal{R}_{\text{rel}}$. We add a learned relation bias to the standard attention score in each head:

$$a_{uv}^{(m)} = \text{Attn}^{(m)}(\mathbf{h}_u, \mathbf{h}_v) + b_{r_{uv}}^{(m)}. \tag{3}$$

The relation type is determined by the positions of u and v in the taxonomy. It distinguishes self, parent, child, ancestor, and descendant pairs. For nodes on different branches, it records whether their lowest common ancestor is a genus, family, or order. Pairs that meet only at the virtual root are treated as unrelated. These relation types are illustrated in Fig. 2A. Each attention head learns a separate bias for every relation type, allowing known taxonomic proximity to guide information exchange in addition to abundance co-occurrence.

After the final encoder layer, each active node has a contextual representation $\mathbf{h}_v$. We first summarize the nodes separately at each taxonomic level:

$$\mathbf{p}^{(r)} = \text{AttnPool}_r\left(\{\mathbf{h}_v\}_{v \in \mathcal{V}_r}\right), \qquad r \in \{\text{species}, \text{genus}, \text{family}, \text{order}\},$$

where $\mathcal{V}_r$ contains the active nodes at level r. This gives one summary vector for each level. We concatenate the four summaries and map them to the sample representation:

$$\mathbf{z} = \text{MLP}_{\text{pool}}\left(\text{Concat}[\mathbf{p}^{(\text{species})}, \mathbf{p}^{(\text{genus})}, \mathbf{p}^{(\text{family})}, \mathbf{p}^{(\text{order})}]\right). \tag{4}$$

$\mathbf{z}$ feeds the decoder during pretraining and the prediction head downstream.

3.2 Hierarchical Compositional Decoding for Pretraining

During pretraining, we mask individual species or complete taxonomic subtrees in the sample-specific active tree. Each node is assigned one of the three states used in Equation 2: observed

when its complete abundance is available, masked when its abundance is hidden, and partially observed when its subtree contains both visible and masked taxa, so its input abundance contains only the visible part. The encoder maps this masked tree to the sample representation $\mathbf{z}$. The decoder treats the total community abundance as one and decomposes it along the taxonomy. It first predicts the proportion assigned to each order, and then recursively predicts the proportions assigned to child families, genera, and species. Reconstructing the complete profiles in this way requires $\mathbf{z}$ to encode both the total abundance of each lineage and the relative composition among its children.

The decoder follows the compositional factorization defined by a fixed taxonomic vocabulary shared across samples. For a parent p with abundance $x_p > 0$ and full child set $\mathcal{C}(p)$ in this vocabulary, the relative abundance of child $c \in \mathcal{C}(p)$ is

$$q_{c|p} = \frac{x_c}{x_p}, \qquad x_c = x_p q_{c|p}. \tag{5}$$

The parent abundance x_p describes the total abundance of a lineage, while $q_{c|p}$ describes its composition among the children. Together, they separate changes in total lineage abundance from changes within the lineage. Given the sample representation $\mathbf{z}$, a shared decoder predicts the conditional composition by scoring each parent–child pair and normalizing across the children:

$$\begin{aligned} s_{pc} &= g_{\text{dec}}\left(\mathbf{z}, \mathbf{E}_{\text{taxon}}[p], \mathbf{E}_{\text{taxon}}[c], \mathbf{E}_{\text{edge}}[p, c]\right), \\ \hat{q}_{c|p} &= \frac{\exp(s_{pc})}{\sum_{k \in \mathcal{C}(p)} \exp(s_{pk})}. \end{aligned} \tag{6}$$

The taxon embeddings identify the parent and child, and the edge embedding identifies their taxonomic-level transition. The same scoring function is shared across all levels. Starting from the virtual root, the decoder converts the predicted conditional compositions into node abundances:

$$\hat{x}_{\text{root}} = 1, \qquad \hat{x}_c = \hat{x}_p \hat{q}_{c|p}. \tag{7}$$

Because the predicted proportions of a parent's children sum to one, their reconstructed abundances also sum to $\hat{x}_p$. The decoder therefore produces consistent profiles across all four levels without an additional hierarchy penalty. We use the hierarchical decoder to pretrain the encoder with two objectives: multiscale composition reconstruction and hierarchical composition learning.

Multiscale composition reconstruction. The decoder reconstructs the complete community profile over the full vocabulary at each taxonomic level. For each rank r, we compare the predicted profile $\hat{\mathbf{x}}^{(r)}$ with the corresponding uncorrupted target $\mathbf{x}^{(r)}$, whose entries are zero for undetected taxa. Because microbial abundances are compositional, we use a smoothed centered log-ratio (CLR) distance (Aitchison, 1986):

$$\mathcal{L}_{\text{rec}} = \sum_{r \in \{\text{species,genus,family,order}\}} \alpha_r D_{\text{CLR}}\left(\mathbf{x}^{(r)}, \hat{\mathbf{x}}^{(r)}\right). \tag{8}$$

Here, D_{CLR} is the mean squared distance between the smoothed CLR transformations of two profiles, and α_r balances the contributions of different taxonomic levels. This objective requires $\mathbf{z}$ to preserve how community abundance is distributed across lineages at multiple biological scales.

Hierarchical composition learning. Matching the abundance profiles at each taxonomic level does not explicitly emphasize the composition within individual lineages. We therefore supervise the predicted child distribution within every parent:

$$\mathcal{L}_{\text{hier}} = -\sum_{p \in \mathcal{P}: x_p > 0} \omega_p \sum_{c \in \mathcal{C}(p)} q_{c|p} \log \hat{q}_{c|p}, \tag{9}$$

Here, $\mathcal{P}$ contains the virtual root and all order, family, and genus parents. We set ω_p proportional to $\sqrt{x_p}$ and normalize the weights within each parent rank. This gives more weight to abundant lineages without allowing them to dominate the loss. By directly supervising each parent–child transition, this objective captures compositional rearrangements among related taxa, including species replacement within a genus. The full objective is

$$\mathcal{L} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{hier}} \mathcal{L}_{\text{hier}}. \tag{10}$$

Multiscale reconstruction preserves the community profile across taxonomic levels, whereas hierarchical composition learning explicitly preserves the local composition within each lineage. Both losses are backpropagated through the hierarchical decoder to train the encoder. After pretraining, the decoder is discarded, and downstream tasks use the taxonomy-aware encoder and the resulting sample representation $\mathbf{z}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Experiments use the full TaxoGut-HPP dataset, which combines a large pretraining corpus with broad downstream coverage. Standard and taxonomy-aware baselines do not use corpus-level pretraining and are trained separately for each downstream task. All experiments use the same task definitions, participant-level splits, and evaluation metrics. Appendix A.2 summarizes all downstream tasks, including their types, sample sizes, and definitions. In all result tables, bold and underlined values indicate the best and second-best results, respectively.

4.2 COMPARATIVE EXPERIMENTS

Table 2: Representative classification performance. Binary tasks use AUROC and abdominal pain frequency uses macro-AUROC (↑). The baselines use the locked TaxoGut-HPP tasks and splits.

Model	Disease					Gut health (HPP)						
	Colorectal cancer	Type 2 diabetes	IBD	Liver cirrhosis	NAFLD	Current abdominal pain	Abdominal pain frequency	Current bloating	GI sensitivity	Irritable bowel syndrome	*H. pylori* infection/history	Gastro-oesophageal reflux disease
Standard baselines												
Elastic Net	0.826	0.791	0.868	0.811	0.807	0.593	0.572	0.637	0.632	0.582	0.628	0.651
Random Forest	0.839	0.917	0.853	0.836	0.819	0.607	0.583	0.691	0.646	0.547	0.667	0.663
DeepMicro	0.803	0.763	0.865	0.816	0.832	0.602	0.574	0.658	0.627	0.562	0.628	0.646
Taxonomy-aware models												
Ph-CNN	0.844	0.817	0.873	0.855	0.840	0.619	0.561	0.649	0.624	0.578	0.623	0.619
PopPhy-CNN	0.840	0.826	0.877	0.847	0.838	0.630	0.578	0.673	0.642	0.585	0.651	0.667
MDeep	0.851	0.844	0.901	0.910	0.903	0.682	0.587	0.750	0.664	0.599	0.684	0.703
DeepPhylo	0.833	0.811	0.858	0.829	0.841	0.622	0.588	0.689	0.670	0.624	0.611	0.655
DeepBiome	0.836	0.808	0.867	0.845	0.833	0.617	0.578	0.657	0.669	0.654	0.601	0.642
PhyloGCNE	0.846	0.801	0.883	0.871	0.877	0.637	0.593	0.741	0.679	0.577	0.620	0.688
Phylo-Spec	0.831	0.839	0.860	0.822	0.823	0.609	0.548	0.647	0.653	0.609	0.690	0.675
MIOSTONE	0.855	0.834	0.902	0.860	0.878	0.653	0.589	0.724	0.690	0.611	0.698	0.693
General foundation models												
MGM	0.898	0.864	0.922	0.961	0.947	0.707	0.632	0.752	0.669	0.690	0.752	0.712
Waypoint-45M	0.861	0.872	0.891	0.972	0.935	0.696	0.619	0.741	0.698	0.675	0.711	0.732
Gut foundation models												
Pope et al.	0.913	0.903	0.937	0.939	0.963	0.711	0.641	0.751	0.724	0.688	0.768	0.692
BiomeGPT	0.933	0.917	0.947	0.984	0.945	0.732	0.648	0.764	0.748	0.728	0.773	0.716
GUT-FORMer	0.928	0.898	0.966	0.971	0.952	0.741	0.652	0.789	0.734	0.736	0.797	0.728
MBEmbed	0.909	0.911	0.928	0.944	0.940	0.721	0.619	0.738	0.740	0.695	0.747	0.749
TaxoMFM (ours)	**0.953**	**0.934**	**0.985**	**0.996**	**0.986**	**0.776**	**0.665**	**0.803**	**0.794**	**0.745**	**0.813**	**0.785**

Table 3: Representative regression results. In each paired result, the first value is Spearman ρ and the second is R^2 (↑). The baselines use the locked TaxoGut-HPP tasks and splits.

Model	Host traits			Liver ultrasound phenotypes		
	Age	BMI	VAT mass	Attenuation coefficient	Speed of sound	Velocity
Standard baselines						
Elastic Net	0.132 / 0.038	0.148 / 0.022	0.214 / 0.048	0.136 / 0.038	0.113 / 0.036	0.111 / 0.026
Random Forest	0.117 / 0.031	0.191 / 0.045	0.246 / 0.075	0.128 / 0.040	0.193 / 0.042	0.136 / 0.024
DeepMicro	0.280 / 0.066	0.221 / 0.065	0.326 / 0.108	0.162 / 0.061	0.128 / 0.033	0.135 / 0.037
Taxonomy-aware models						
Ph-CNN-Reg	0.178 / 0.049	0.134 / 0.048	0.217 / 0.050	0.139 / 0.043	0.146 / 0.027	0.110 / 0.025
PopPhy-CNN-Reg	0.184 / 0.028	0.213 / 0.043	0.238 / 0.052	0.119 / 0.026	0.164 / 0.036	0.125 / 0.031
MDeep	0.304 / 0.105	0.246 / 0.059	0.298 / 0.113	0.254 / 0.092	0.132 / 0.046	0.189 / 0.051
DeepPhylo	0.212 / 0.046	0.201 / 0.064	0.235 / 0.069	0.145 / 0.039	0.229 / 0.074	0.112 / 0.036
DeepBiome	0.288 / 0.094	0.282 / 0.087	0.249 / 0.088	0.153 / 0.047	0.146 / 0.048	0.171 / 0.043
PhyloGCNE	0.101 / 0.044	0.118 / 0.029	0.244 / 0.083	0.133 / 0.040	0.155 / 0.056	0.118 / 0.032
Phylo-Spec	0.142 / 0.052	0.109 / 0.043	0.207 / 0.073	0.149 / 0.058	0.187 / 0.065	0.115 / 0.026
MIOSTONE	0.251 / 0.085	0.239 / 0.049	0.324 / 0.125	0.210 / 0.081	0.219 / 0.074	0.164 / 0.034
General foundation models						
MGM	0.304 / 0.109	0.272 / 0.085	0.315 / 0.127	0.293 / 0.094	0.271 / 0.089	0.209 / 0.054
Waypoint-45M	0.328 / 0.121	0.263 / 0.083	0.331 / 0.129	0.308 / 0.106	0.286 / 0.102	0.239 / 0.063
Gut foundation models						
Pope et al.	0.322 / 0.122	0.286 / 0.082	0.391 / 0.183	0.259 / 0.084	0.263 / 0.094	0.258 / 0.077
BiomeGPT	0.346 / 0.138	0.322 / 0.108	0.358 / 0.151	0.329 / 0.128	0.322 / 0.128	0.231 / 0.058
GUT-FORMer	0.348 / 0.159	0.308 / 0.091	0.336 / 0.165	0.303 / 0.101	0.258 / 0.088	0.194 / 0.038
MBEmbed	0.366 / 0.134	0.311 / 0.096	0.378 / 0.144	0.318 / 0.103	0.301 / 0.105	0.247 / 0.069
TaxoMFM (ours)	**0.382 / 0.173**	**0.337 / 0.141**	**0.414 / 0.202**	**0.347 / 0.136**	**0.345 / 0.142**	**0.287 / 0.101**

Table 2 and Table 3 reveal a consistent but task-dependent pattern. On public disease tasks, several foundation models already approach high performance, so the gains of TaxoMFM are generally modest. The separation is clearer on gut-health outcomes, especially gastrointestinal sensitivity and gastro-oesophageal reflux disease, where the fine-grained clinical labels are more challenging for the baselines. Regression remains difficult for all methods, as reflected by the moderate Spearman ρ and R^2 values, but TaxoMFM improves consistently across host traits and liver ultrasound phenotypes. Overall, the benefit is most visible on clinically detailed and continuous phenotypes rather than on the more saturated disease classification tasks.

4.3 Ablation study

We compare six variants. **Species-only** removes the genus, family, and order nodes and uses a flat species input. **Without relation bias** keeps all four taxonomic ranks but removes the relation-dependent attention bias. **Random taxonomy** preserves the rank sizes and branching pattern but shuffles the parent–child assignments. **Without hierarchical composition learning** removes the direct supervision of child distributions within each parent. **Without multiscale reconstruction** removes the reconstruction objective at the four taxonomic ranks. Finally, **without HPP pretraining** retains the full model and both objectives but excludes HPP samples from pretraining.

Table 4: Ablation study of TaxoMFM. Classification tasks report AUROC (↑), and liver phenotype regression reports Spearman's ρ / R^2 (↑).

Model variant	Disease classification			Gut health classification			Liver phenotype regression		
	Colorectal cancer	IBD	Type 2 diabetes	Current abdominal pain	Current bloating	Irritable bowel syndrome	Attenuation coefficient	Speed of sound	Velocity
Structural ablations									
Species-only (flat)	0.914	0.943	0.909	0.736	0.758	0.723	0.313/0.113	0.320/0.125	0.226/0.060
w/o relation bias	0.934	0.964	0.926	<u>0.771</u>	<u>0.782</u>	0.715	0.327/0.123	0.331/0.130	<u>0.273</u>/<u>0.088</u>
Random taxonomy	0.906	0.933	0.880	0.686	0.694	0.635	0.223/0.084	0.206/0.071	0.158/0.035
Pretraining objective ablations									
w/o hierarchical composition learning	0.927	0.973	0.919	0.764	0.763	0.719	<u>0.331</u>/<u>0.129</u>	<u>0.339</u>/<u>0.134</u>	0.271/0.084
w/o multiscale reconstruction	0.920	0.966	0.914	0.748	0.773	<u>0.733</u>	0.320/0.107	0.332/0.132	0.252/0.074
Pretraining data ablation									
w/o HPP pretraining	<u>0.946</u>	<u>0.980</u>	<u>0.928</u>	0.751	0.766	0.722	0.316/0.105	0.317/0.109	0.233/0.068
TaxoMFM (full)	**0.953**	**0.985**	**0.934**	**0.776**	**0.803**	**0.745**	**0.347 / 0.136**	**0.345 / 0.142**	**0.287 / 0.101**

The ablation results show that TaxoMFM benefits jointly from taxonomic structure, pretraining objectives, and data coverage. Replacing the true taxonomy with a random hierarchy that preserves rank sizes and branching patterns caused the largest performance drop. The mean AUROC decreased from 0.957 to 0.906 on the three disease tasks and from 0.775 to 0.672 on the three gut-health tasks. The mean Spearman's ρ on the three liver phenotypes decreased from 0.326 to 0.196. The random hierarchy also performed worse than the species-only model, which achieved corresponding group means of 0.922, 0.739, and 0.286. This result shows that the improvement does not come from simply adding higher-rank nodes. It depends on the real taxonomic relationships among microorganisms. Removing the relation bias reduced the mean performance of the three task groups by 0.016, 0.019, and 0.016, respectively, showing an additional benefit from explicitly modeling taxonomic relations. The two pretraining objectives also provided complementary supervision. Removing multiscale reconstruction reduced the mean performance by 0.024, 0.023, and 0.025 across the three task groups, while removing hierarchical composition learning resulted in decreases of 0.018, 0.026, and 0.013. Finally, excluding HPP from pretraining reduced the mean AUROC on public disease tasks by only 0.006, but reduced the mean AUROC on gut-health tasks by 0.028 and the mean Spearman's ρ on liver phenotypes by 0.038. These results show that HPP primarily expands the model's pretraining coverage of gut health and continuous clinical phenotypes.

4.4 Zero-shot cross-cohort transfer

We evaluate target-cohort zero-shot transfer on three external disease cohorts: colorectal cancer from Yang et al. (2021), type 2 diabetes from MetaCardis (2020), and IBD from HMP/IBDMDB. For each task, the target cohort is excluded from both pretraining and downstream adaptation. Models are trained using only labeled source cohorts and are then evaluated directly on the unseen target cohort

using AUROC. As shown in Table 5, TaxoMFM achieves the best performance on all three tasks, with AUROCs of 0.838, 0.736, and 0.722, respectively. Compared with the strongest baseline for each task, TaxoMFM improves AUROC by 0.075 on colorectal cancer, 0.039 on type 2 diabetes, and 0.029 on IBD. These results show that taxonomy-aware pretraining improves cross-cohort transfer across cancer, metabolic disease, and inflammatory disease.

4.5 Analysis of taxonomy-aware pretraining

The previous experiments show that taxonomic structure improves downstream performance. We next examine what the model learns from this structure through three targeted analyses.

Distinguishing changes across taxonomic scales. We first test whether TaxoMFM can recover community changes at different taxonomic scales from incomplete observations. We use only samples excluded from model pretraining. For each sample, we select a genus containing multiple detected species and construct three controlled changes: changing a single species, changing the total abundance of the genus, or redistributing abundance among its species while keeping the genus total fixed. For each original–perturbed pair, we mask the same species within the selected genus and reconstruct both complete communities. We then compare the reconstructed difference with the known perturbation. Parent error measures the recovery of genus-level abundance changes, conditional error measures the recovery of within-genus compositional changes, and localization error measures whether the change is assigned to the correct genus. Combined error, our primary metric, is the sum of the parent and conditional errors.

Table 5: Zero-shot performance using AUROC (↑).

Model	CRC Yang et al. (2021) $n = 194$	T2D MetaCardis (2020) $n = 783$	IBD HMP/ IBDMDB $n = 130$
Random Forest	0.619	0.634	0.631
MIOSTONE	0.643	0.562	<u>0.693</u>
PhyloGCNE	0.610	0.602	0.567
MGM	0.744	0.670	0.664
BiomeGPT	<u>0.763</u>	0.683	0.679
Species-only	0.759	<u>0.697</u>	0.688
TaxoMFM (ours)	**0.838**	**0.736**	**0.722**

Table 6: Multiscale compositional recovery under controlled perturbations (↓).

Model variant	Parent	Conditional	Localization	Combined
Species-only (flat)	0.736	<u>0.994</u>	<u>0.489</u>	<u>1.730</u>
Random taxonomy	0.823	1.414	0.621	2.237
w/o hierarchical composition learning	<u>0.722</u>	1.228	0.525	1.950
TaxoMFM (full)	**0.642**	**0.956**	**0.431**	**1.598**

As shown in Table 6, TaxoMFM achieves the lowest error on all four metrics. Its combined error is 1.598, representing reductions of 7.6%, 28.6%, and 18.1% relative to species-only, random taxonomy, and the model without hierarchical composition learning, respectively. The lower parent and conditional errors indicate that TaxoMFM better recovers both lineage-level abundance changes and within-lineage compositional shifts. Performance declines when the taxonomy is randomized or hierarchical composition learning is removed, indicating that both components contribute to multiscale compositional recovery.

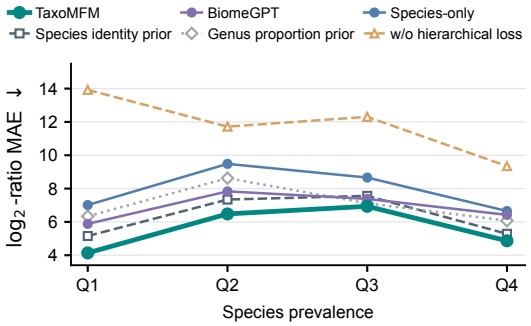
Conditional recovery of long-tail species. We evaluated conditional abundance recovery on independent external samples. Species were divided into four groups according to their prevalence in the pretraining data, from Q1 (rarest) to Q4 (most prevalent). For each instance, we masked the abundances of approximately 20% of detected species, including the target and at least one additional species. We retained species identities and constructed model inputs from the renormalized visible composition. The prediction target was the target species' abundance divided by the total visible abundance. We measured recovery using species-averaged absolute $\log_2$-ratio error, with lower values indicating better performance.

Table 7 and Fig. 3 show TaxoMFM has the lowest error across all four prevalence groups. Its errors are 29.6%, 17.3%, 6.1%, and 24.2% lower than the BiomeGPT values in Q1–Q4, respectively, with the largest differences in Q1 and Q4. In Q1, TaxoMFM's error is 4.139, compared with 5.153 for the species identity prior and 6.337 for the genus proportion prior. Baseline rankings vary across groups: the species identity prior performs best among baselines in Q1, Q2, and Q4, whereas the genus proportion prior performs best in Q3.

Table 7: Long-tail recovery from Q1 (rarest) to Q4 (commonest).

Method	Q1	Q2	Q3	Q4
Species identity prior	5.153	7.342	7.563	5.284
Genus proportion prior	6.337	8.624	7.118	6.074
Species-only	7.012	9.483	8.661	6.653
Random taxonomy	23.126	19.847	21.203	13.014
w/o hierarchical loss	13.924	11.718	12.306	9.347
BiomeGPT	5.881	7.832	7.384	6.422
TaxoMFM	**4.139**	**6.479**	**6.934**	**4.866**

Figure 3: Recovery trends across prevalence groups.



Maintaining representations across taxonomic vocabularies. Different taxonomic database versions can rename, split, or omit the same species or SGB. We therefore constructed paired MetaPhlAn 4 current and MetaPhlAn 3 legacy views from profiles in the pretraining dataset. The legacy view retained identical species names and 33 curated MetaPhlAn 3-to-4 mappings, whereas TaxoMFM routed unmatched abundance to genus, family, or order nodes shared across the vocabularies. Because each pair came from the same sample, only the vocabulary changed.

Table 8: Cross-vocabulary consistency under a MetaPhlAn 3-to-4 shift.

Model	Legacy abundance retained (%) ↑	Cosine separation ↑	Bidirectional R@1 (%) ↑	Bidirectional R@5 (%) ↑
Structural controls				
TaxoMFM leaf-only	45.79	0.309	67.69	84.94
Species-only (flat)	45.79	0.370	82.42	84.52
Random taxonomy	**81.44**	0.045	1.23	3.51
External baselines				
BiomeGPT	45.79	0.078	83.11	83.99
Waypoint-45M	64.30	0.116	81.06	83.54
GUT-FORMer	41.05	0.050	79.53	88.84
MBEmbed	33.14	0.290	75.32	89.18
DeepMicro	41.05	0.104	85.73	90.79
TaxoMFM (full)	**81.44**	**0.379**	**86.64**	**96.06**

Figure 4: Cross-vocabulary transfer performance across models, with darker cells indicating higher within-metric rank.



We report retained legacy abundance (the mean fraction of original abundance preserved through species matching or ancestor routing), cosine separation, and bidirectional Recall@1 and Recall@5. Detailed metric definitions are provided in Appendix C.4. As shown in Table 8 and Fig. 4, TaxoMFM retained 81.44% of the abundance and achieved a cosine separation of 0.379, a Recall@1 of 86.64%, and a Recall@5 of 96.06%. Disabling ancestor routing reduced retained abundance to 45.79% and Recall@1 to 67.69%. A random taxonomy retained the same abundance but achieved only 1.23% Recall@1. Shared ancestors therefore preserve information across vocabularies, while the correct taxonomic structure maintains sample-specific representations.

5 Conclusion

We introduced TaxoMFM, a microbiome foundation model that incorporates taxonomic hierarchies and relationships across species, genus, family, and order. Experiments on TaxoGut-HPP show improvements in disease classification, gut health assessment, continuous phenotype prediction, and cross-cohort zero-shot transfer. TaxoMFM also recovers community changes across taxonomic scales and masked species abundances while adapting to vocabulary shifts through shared taxonomic ancestors. Together, these results establish a framework for learning biologically informed microbiome representations that transfer across tasks, cohorts, and taxonomic vocabularies.

AI USE STATEMENT

We used large language models (LLMs) to assist with dataset collection, manuscript polishing, and code review. The authors take responsibility for the final content of this work.

REFERENCES

John Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall, 1986.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258.

Daniel Chang, Vinod K Gupta, Benjamin Hur, Sergio Cobo-López, Kevin Y Cunningham, Nam Soo Han, Insuk Lee, Vanessa L Kronzer, Levi M Teigen, Lioudmila V Karnatovskaia, et al. Gut microbiome wellness index 2 enhances health status prediction from gut microbiome taxonomic profiles. *Nature communications*, 15(1):7447, 2024.

Biao Dong, Bin Wang, Jiongjin Chen, Xiaomin Xu, and Zhenjiang Zech Xu. Adaptive graph learning of microbial phylogeny enables accurate and interpretable microbiome-based host phenotype prediction. *Applied and Environmental Microbiology*, 2026. doi: 10.1128/aem.00788-26.

Diego Fioravanti, Ylenia Giarratano, Valerio Maggio, Claudio Agostinelli, Marco Chierici, Giuseppe Jurman, and Cesare Furlanello. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics*, 19:49, 2018. doi: 10.1186/s12859-018-2033-5.

Sofia K Forslund, Rima Chakaroun, Maria Zimmermann-Kogadeeva, Lajos Markó, Judith Aron-Wisnewsky, Trine Nielsen, Lucas Moitinho-Silva, Thomas S B Schmidt, Gwen Falony, Sara Vieira-Silva, Solia Adriouch, Renato J Alves, Karen Assmann, Jean-Philippe Bastard, Till Birkner, Robert Caesar, Julien Chilloux, Luis Pedro Coelho, Leopold Fezeu, Nathalie Galleron, Gerard Helft, Richard Isnard, Boyang Ji, Michael Kuhn, Emmanuelle Le Chatelier, Antonis Myridakis, Lisa Olsson, Nicolas Pons, Edi Prifti, Benoit Quinquis, Hugo Roume, Joe-Elie Salem, Nataliya Sokolovska, Valentina Tremaroli, Mireia Valles-Colomer, Christian Lewinter, Nadja B Søndertoft, Helle Krogh Pedersen, Tue H Hansen, MetaCardis Consortium*, Jens Peter Gøtze, Lars Køber, Henrik Vestergaard, Torben Hansen, Jean-Daniel Zucker, Serge Hercberg, Jean-Michel Oppert, Ivica Letunic, Jens Nielsen, Fredrik Bäckhed, S Dusko Ehrlich, Marc-Emmanuel Dumas, Jeroen Raes, Oluf Pedersen, Karine Clément, Michael Stumvoll, and Peer Bork. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature*, 600(7889):500–505, 2021. doi: 10.1038/s41586-021-04177-9.

Vinod K Gupta, Minsuk Kim, Utpal Bakshi, Kevin Y Cunningham, John M Davis III, Konstantinos N Lazaridis, Heidi Nelson, Nicholas Chia, and Jaeyun Sung. A predictive index for health status using species-level gut microbiome profiling. *Nature communications*, 11(1):4635, 2020.

Yifan Jiang, Matthew Aton, Qiyun Zhu, and Yang Young Lu. Modeling microbiome-trait associations with taxonomy-adaptive neural networks. *Microbiome*, 13(1), 2025. doi: 10.1186/s40168-025-02080-3.

Zuzanna Karwowska, Marcin Możejko, Wojciech Nowak, Anastasia Romanchenko, Ewa Szczurek, and Tomasz Kosciolek. A universal taxonomic and functional human gut microbiome model for disease classification and phenotype discovery. *bioRxiv*, 2026. doi: 10.64898/2026.04.30.721924.

Jason Lloyd-Price, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, Nadim J Ajami, Kevin S Bonham, Colin J Brislawn, David Casero, Holly Courtney, Antonio Gonzalez, Thomas G Graeber, A Brantley Hall, Kathleen Lake, Carol J Landers, Himel Mallick, Damian R Plichta, Mahadev Prasad, Gholamali Rahnavard, Jenny Sauk, Dmitry Shungin, Yoshiki Vázquez-Baeza, Richard A White, IBDMDB Investigators, Jonathan Braun, Lee A Denson, Janet K Jansson, Rob Knight, Subra Kugathasan, Dermot P B McGovern, Joseph F Petrosino, Thaddeus S Stappenbeck, Harland S Winter, Clary B Clish, Eric A Franzosa, Hera Vlamakis, Ramnik J Xavier, and Curtis Huttenhower. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662, 2019. doi: 10.1038/s41586-019-1237-9.

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.

Min Oh and Liqing Zhang. DeepMicro: deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10:6026, 2020. doi: 10.1038/s41598-020-63159-5.

Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. Accessible, curated metagenomic data through experimenthub. *Nature methods*, 14(11):1023–1024, 2017.

Quintin Pope, Rohan Varma, Christine Tataru, Maude M. David, and Xiaoli Fern. Learning a deep language model for microbiomes: The power of large scale unlabeled microbiome data. *PLOS Computational Biology*, 21(5):e1011353, 2025. doi: 10.1371/journal.pcbi.1011353.

Lee Reicher, Smadar Shilo, Anastasia Godneva, Guy Lutsker, Liron Zahavi, Saar Shoer, David Krongauz, Michal Rein, Sarah Kohn, Tomer Segev, et al. Deep phenotyping of health–disease continuum in the human phenotype project. *Nature Medicine*, 31(9):3191–3203, 2025.

Derek Reiman, Ahmed A. Metwally, Jun Sun, and Yang Dai. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2993–3001, 2020. doi: 10.1109/JBHI.2020.2993761.

Daphna Rothschild, Sigal Leviatan, Ariel Hanemann, Yossi Cohen, Omer Weissbrod, and Eran Segal. An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PloS one*, 17(3):e0265756, 2022.

Divya Sharma, Andrew D. Paterson, and Wei Xu. TaxoNN: ensemble of neural networks on stratified microbiome data for disease prediction. *Bioinformatics*, 36(17):4544–4550, 2020. doi: 10.1093/bioinformatics/btaa542.

Justin D. Silverman, Alex D. Washburne, Sayan Mukherjee, and Lawrence A. David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6:e21887, 2017. doi: 10.7554/eLife.21887.

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.

Bin Wang, Yulong Shen, Jingyan Fang, and Xiaoquan Su. DeepPhylo: phylogeny-aware microbial embeddings enhanced predictive accuracy in human microbiome data analysis. *Advanced Science*, 11(45):2404277, 2024. doi: 10.1002/advs.202404277.

Yongzhi Yang, Lutao Du, Debing Shi, Cheng Kong, Jianqiang Liu, Guang Liu, Xinxiang Li, and Yanlei Ma. Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications*, 12(1):6757, 2021. doi: 10.1038/s41467-021-27112-y.

Liron Zahavi, Zachary Levine, and Eran Segal. Self-supervised representation learning for microbiome improves downstream prediction in data-limited settings and cross-cohort generalizability. In *ICML Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

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*, pp. e13333, 2026. doi: 10.1002/advs.202513333.

A Downstream tasks

A.1 Data sources and sample composition

TaxoGut-HPP integrates public microbial abundance data from curatedMetagenomicData (cMD), GMHI, and GMWI2 with data from the HPP and DayTwo cohorts. The dataset covers 231 source cohorts and contains 100,806 abundance profiles.

Table A1 summarizes cohort and sample counts for each resource, with one abundance profile counted as one sample.

Table A1: Data sources and sample composition of TaxoGut-HPP.

Data resource	**Cohorts**	**Samples**
cMD (Pasolli et al., 2017)	80	12,351
GMHI main (Gupta et al., 2020)	94	4,344
GMHI validation (Gupta et al., 2020)	1	782
GMWI2 main (Chang et al., 2024)	49	7,129
GMWI2 test (Chang et al., 2024)	5	1,005
HPP (Reicher et al., 2025)	1	9,044
DayTwo (Rothschild et al., 2022)	1	66,151
Total	**231**	**100,806**

A.2 Downstream task definitions

Table A2: Downstream task definitions.

Task	**Type**	**N**	**Definition**
Public health-status and disease tasks			
Colorectal cancer	Binary	17,265	Malignant tumors of the colon or rectum.
Type 2 diabetes	Binary	17,298	Metabolic disease involving insulin resistance and elevated blood glucose.
IBD	Binary	16,369	Inflammatory bowel disease, including Crohn's disease and ulcerative colitis.
Liver cirrhosis	Binary	15,948	Liver scarring and structural changes caused by chronic injury.
NAFLD	Binary	15,882	Nonalcoholic fatty liver disease: liver fat accumulation unrelated to excessive alcohol use.
Healthy vs. diseased	Binary	25,240	Healthy or non-healthy status recorded in the source data.
Crohn's disease	Binary	16,188	Chronic inflammatory bowel disease affecting different parts of the digestive tract.
Ulcerative colitis	Binary	16,142	Chronic inflammatory bowel disease affecting the lining of the colon and rectum.
Multiple sclerosis	Binary	16,185	Immune-mediated demyelinating disease of the central nervous system.
Rheumatoid arthritis	Binary	16,136	Autoimmune disease causing chronic joint inflammation.
Ankylosing spondylitis	Binary	16,004	Chronic inflammatory disease primarily affecting the spine and sacroiliac joints.
Graves' disease	Binary	15,896	Autoimmune disease causing an overactive thyroid.
Behçet's disease	Binary	15,816	Systemic inflammatory disease affecting the mouth, genitals, eyes, and blood vessels.

Continued on next page

Table A2 (continued)

Task	Type	N	Definition
Myalgic encephalomyelitis / chronic fatigue syndrome	Binary	15,846	Chronic illness characterized by persistent fatigue and worsening symptoms after exertion.
Colorectal adenoma	Binary	16,038	Adenomatous lesions of the colon or rectum with potential for malignant progression.
Pancreatic cancer	Binary	15,839	Malignant tumors of the pancreas.
Coronary artery disease	Binary	15,881	Disease involving narrowed or blocked coronary arteries and reduced blood supply to the heart.
Atherosclerotic cardiovascular disease	Binary	16,376	Cardiovascular disease caused by atherosclerotic plaque.
Impaired glucose tolerance	Binary	15,960	Elevated blood glucose after a glucose challenge, below the diabetes range.
Melanoma	Binary	15,926	Malignant tumors arising from melanocytes.
Parkinson's disease	Binary	16,286	Neurodegenerative disease primarily affecting movement.
Schizophrenia	Binary	15,886	Psychiatric disorder affecting thought, perception, and behavior.
Soil-transmitted helminths	Binary	15,810	Infection with parasitic worms transmitted through contaminated soil.
Clostridioides difficile infection	Binary	15,824	Intestinal infection caused by Clostridioides difficile.
Acute diarrhea	Binary	15,835	Short-term onset of frequent loose or watery stools.
Premature birth	Binary	15,826	Preterm-birth phenotype recorded in the source data.
HPP gastrointestinal questionnaire: classification			
Current abdominal pain	Binary	6,642	Self-reported presence of current abdominal pain.
Abdominal pain frequency	Ordinal (7)	6,662	Frequency of abdominal pain or discomfort over the past three months, from never to daily.
Current bloating	Binary	6,539	Self-reported presence of current abdominal bloating or distension.
GI sensitivity	Binary	5,564	Gastrointestinal sensitivity, assessed by self-reported sensitive stomach.
Soft-stool frequency	Ordinal (5)	6,661	Frequency of soft stools over the past three months, from never to always.
Hard-stool frequency	Ordinal (5)	6,661	Frequency of hard stools over the past three months, from never to always.
Nausea in the past three months	Binary	6,352	Whether nausea caused discomfort over the past three months.
HPP medical-history tasks			
Irritable bowel syndrome	Binary	3,174	History of irritable bowel syndrome, involving recurrent abdominal pain and altered bowel habits.
H. pylori infection/history	Binary	3,174	History of Helicobacter pylori infection or associated gastritis.
Gastro-oesophageal reflux disease	Binary	3,174	History of reflux of stomach contents causing symptoms or complications.
Anorectal disease	Binary	3,174	History of haemorrhoids, anal fissures, perianal abscesses, or fistulas.
Gallstone / cholecystectomy	Binary	3,174	History of gallstones or surgical removal of the gallbladder.
Fatty liver history	Binary	3,174	History of fatty liver, including nonalcoholic fatty liver disease.

Continued on next page

Table A2 (continued)

Task	Type	N	Definition
Peptic ulcer disease	Binary	3,174	History of peptic ulcers, typically in the stomach or duodenum.
Host-trait regression			
Age	Regression	9,044	Age in years.
BMI	Regression	9,044	Body mass index: weight divided by squared height (kg/m^2).
VAT mass	Regression	9,044	Mass of visceral adipose tissue.
HPP liver-ultrasound regression			
Attenuation coefficient	Regression	5,346	Liver ultrasound attenuation coefficient (dB/cm/MHz).
Speed of sound	Regression	5,350	Speed of ultrasound propagation through liver tissue (m/s).
Velocity	Regression	5,350	Median shear-wave propagation velocity in liver tissue (m/s).
HPP gastrointestinal questionnaire: regression			
Bowel-movement satisfaction	Regression	6,641	Self-rated bowel-movement satisfaction on a 0–10 scale; higher scores indicate greater satisfaction.
Daily bowel-movement count	Regression	6,636	Average number of bowel movements per day.

B ADDITIONAL DOWNSTREAM RESULTS

The main classification table reports 12 representative tasks, and Tables A3–A8 provide 28 additional classification tasks. The main regression table reports six representative tasks, and Table A8 provides two additional HPP regression tasks. Together, these tables cover 40 classification and eight regression tasks. Binary classification tasks are evaluated using AUROC. For the three ordinal multiclass outcomes, abdominal-pain frequency, soft-stool frequency, and hard-stool frequency, we report macro-averaged one-vs-rest AUROC. Regression tasks are evaluated using Spearman's ρ and R^2.

Table A3: Additional AUROC results for general and immune-related diseases (part I).

Model	Healthy vs. diseased	Crohn's disease	Ulcerative colitis	Multiple sclerosis	Rheumatoid arthritis
Standard baselines					
Elastic Net	0.760	0.872	0.849	0.869	0.834
Random Forest	0.782	0.908	0.872	0.901	0.859
DeepMicro	0.809	0.873	0.868	0.908	0.861
ExtraTrees	0.773	0.902	0.882	0.895	0.864
Taxonomy-aware models					
Ph-CNN	0.799	0.896	0.884	0.910	0.868
PopPhy-CNN	0.811	0.907	0.899	0.923	0.883
MDeep	0.846	0.933	0.793	0.943	0.910
DeepPhylo	0.802	0.901	0.878	0.918	0.877
DeepBiome	0.816	0.914	0.903	0.951	0.891
PhyloGCNE	0.822	0.919	0.910	0.931	0.903
Phylo-Spec	0.784	0.878	0.847	0.883	0.859
MIOSTONE	0.838	0.883	0.927	0.934	0.936
General foundation models					
MGM	0.858	0.948	0.934	0.947	0.931
Waypoint-45M	0.842	0.936	0.936	0.950	0.933
Gut foundation models					
Pope et al.	0.851	0.953	0.945	0.951	0.938
BiomeGPT	0.874	<u>0.964</u>	0.951	<u>0.963</u>	<u>0.959</u>
GUT-FORMer	<u>0.879</u>	0.955	<u>0.959</u>	0.959	0.958
MBEmbed	0.869	0.947	0.944	0.949	0.946
TaxoMFM (ours)	**0.891**	**0.979**	**0.967**	**0.976**	**0.971**

Table A4: Additional AUROC results for general and immune-related diseases (part II).

Model	Ankylosing spondylitis	Graves' disease	Behçet's disease	Myalgic encephalomyelitis / chronic fatigue syndrome
Standard baselines				
Elastic Net	0.892	0.883	0.904	0.862
Random Forest	0.923	0.902	0.947	0.889
DeepMicro	0.907	0.910	0.922	0.902
ExtraTrees	0.917	0.930	0.941	0.877
Taxonomy-aware models				
Ph-CNN	0.921	0.928	0.935	0.904
PopPhy-CNN	0.930	0.941	0.954	0.917
MDeep	0.947	0.951	0.961	0.953
DeepPhylo	0.941	0.936	0.937	0.912
DeepBiome	0.938	0.947	0.949	0.927
PhyloGCNE	0.933	0.949	0.938	0.920
Phylo-Spec	0.902	0.908	0.917	0.886
MIOSTONE	0.945	0.955	0.972	0.932
General foundation models				
MGM	0.964	0.977	0.976	0.939
Waypoint-45M	0.960	0.963	0.966	0.947
Gut foundation models				
Pope et al.	0.975	0.970	0.993	0.938
BiomeGPT	0.983	0.982	0.983	0.960
GUT-FORMer	0.979	0.986	0.991	0.969
MBEmbed	0.970	0.968	0.971	0.944
TaxoMFM (ours)	**0.992**	**0.993**	**0.996**	**0.978**

Table A5: Additional AUROC results for neoplastic and cardiometabolic diseases.

Model	Colorectal adenoma	Pancreatic cancer	Coronary artery disease	Athero-sclerotic cardio-vascular disease	Impaired glucose tolerance	Melanoma
Standard baselines						
Elastic Net	0.833	0.896	0.874	0.903	0.788	0.874
Random Forest	0.868	0.932	0.903	0.934	0.817	0.906
DeepMicro	0.851	0.918	0.882	0.919	0.825	0.895
ExtraTrees	0.873	0.925	0.896	0.928	0.806	0.899
Taxonomy-aware models						
Ph-CNN	0.880	0.939	0.903	0.929	0.823	0.902
PopPhy-CNN	0.892	0.944	0.921	0.947	0.843	0.916
MDeep	0.915	0.952	0.945	0.955	0.864	0.924
DeepPhylo	0.873	0.936	0.914	0.936	0.887	0.913
DeepBiome	0.887	0.948	0.930	0.949	0.856	0.939
PhyloGCNE	0.899	0.940	0.926	0.941	0.849	0.930
Phylo-Spec	0.858	0.922	0.889	0.919	0.814	0.877
MIOSTONE	0.909	0.954	0.938	0.953	0.861	0.933
General foundation models						
MGM	0.928	0.962	0.943	0.976	0.872	0.941
Waypoint-45M	0.919	0.968	0.934	0.962	0.881	0.940
Gut foundation models						
Pope et al.	0.949	0.974	0.952	0.975	0.881	0.948
BiomeGPT	0.931	0.982	0.962	0.988	0.899	0.962
GUT-FORMer	0.940	0.979	0.956	0.979	0.887	0.954
MBEmbed	0.947	0.963	0.944	0.970	0.877	0.968
TaxoMFM (ours)	**0.962**	**0.993**	**0.975**	**0.995**	**0.919**	**0.980**

Table A6: Additional AUROC results for neurological, infectious, and host-related tasks.

Model	Parkinson's disease	Schizophrenia	Soil-transmitted helminths	*Clostridioides difficile* infection	Acute diarrhea	Premature birth
Standard baselines						
Elastic Net	0.898	0.883	0.916	0.869	0.922	0.904
Random Forest	0.924	0.918	0.944	0.899	0.957	0.932
DeepMicro	0.916	0.925	0.930	0.920	0.949	0.941
ExtraTrees	0.919	0.911	0.956	0.883	0.950	0.928
Taxonomy-aware models						
Ph-CNN	0.931	0.928	0.950	0.916	0.959	0.937
PopPhy-CNN	0.941	0.948	0.964	0.934	0.969	0.949
MDeep	0.953	0.959	0.969	0.949	0.973	0.962
DeepPhylo	0.932	0.946	0.947	0.923	0.958	0.941
DeepBiome	0.944	0.951	0.960	0.938	0.965	0.953
PhyloGCNE	0.946	0.941	0.962	0.940	0.978	0.951
Phylo-Spec	0.913	0.910	0.939	0.907	0.931	0.918
MIOSTONE	0.951	0.964	0.966	0.945	0.971	0.959
General foundation models						
MGM	0.971	0.972	0.978	0.955	0.977	0.978
Waypoint-45M	0.965	0.966	0.987	0.951	0.973	0.971
Gut foundation models						
Pope et al.	0.979	0.973	0.985	**0.980**	0.986	0.974
BiomeGPT	0.990	0.985	0.992	0.969	0.989	0.989
GUT-FORMer	0.985	0.979	0.982	0.978	0.993	0.981
MBEmbed	0.973	0.981	0.988	0.959	0.979	0.986
TaxoMFM (ours)	**0.998**	**0.996**	**0.993**	0.978	**0.996**	**0.998**

Table A7: Additional HPP classification results for gastrointestinal outcomes. Soft- and hard-stool frequency use macro-AUROC; the remaining tasks use AUROC.

Model	Soft-stool frequency	Hard-stool frequency	Nausea in the past three months	Anorectal disease	Gallstone / cholecystectomy
Standard baselines					
Elastic Net	0.596	0.587	0.664	0.624	0.669
Random Forest	0.638	0.625	0.698	0.657	0.706
DeepMicro	0.632	0.619	0.728	0.668	0.719
ExtraTrees	0.613	0.621	0.687	0.646	0.698
Taxonomy-aware models					
Ph-CNN	0.639	0.621	0.718	0.668	0.726
PopPhy-CNN	0.651	0.640	0.734	0.687	0.742
MDeep	0.677	0.668	0.754	0.728	0.789
DeepPhylo	0.648	0.637	0.730	0.704	0.745
DeepBiome	0.678	0.667	0.756	0.709	0.768
PhyloGCNE	0.669	0.660	0.768	0.714	0.764
Phylo-Spec	0.631	0.609	0.708	0.660	0.711
MIOSTONE	0.681	0.674	0.781	0.741	0.779
General foundation models					
MGM	0.688	0.677	0.806	0.756	0.819
Waypoint-45M	0.674	0.689	0.813	0.748	0.810
Gut foundation models					
Pope et al.	0.710	0.704	0.841	0.785	0.831
BiomeGPT	0.724	0.706	0.859	0.787	0.849
GUT-FORMer	**0.740**	0.714	0.846	0.796	0.833
MBEmbed	0.694	0.725	0.835	0.778	0.865
TaxoMFM (ours)	0.738	**0.742**	**0.885**	**0.810**	**0.872**

Table A8: Additional HPP classification and regression results. The first two columns report AUROC; the final two report Spearman ρ / R^2 (all ↑).

Model	Fatty liver history	Peptic ulcer disease	Bowel-movement satisfaction	Daily bowel-movement count
Standard baselines				
Elastic Net	0.657	0.629	0.118 / -0.012	0.183 / 0.041
Random Forest	0.741	0.612	0.156 / 0.019	0.258 / 0.096
ExtraTrees	0.726	0.624	0.147 / 0.023	0.246 / 0.101
DeepMicro	0.693	0.655	0.193 / 0.031	0.229 / 0.071
Taxonomy-aware models				
Ph-CNN	0.711	0.642	0.154 / 0.012	0.216 / 0.061
PopPhy-CNN	0.734	0.679	0.181 / 0.026	0.247 / 0.089
MDeep	0.764	0.657	0.211 / 0.042	0.281 / 0.120
DeepPhylo	0.722	0.705	0.169 / 0.021	0.295 / 0.141
DeepBiome	0.751	0.671	0.222 / 0.046	0.268 / 0.109
PhyloGCNE	0.779	0.698	0.105 / 0.004	0.255 / 0.092
Phylo-Spec	0.701	0.633	0.128 / -0.007	0.237 / 0.079
MIOSTONE	0.803	0.691	0.216 / 0.048	0.312 / 0.156
General foundation models				
MGM	0.808	0.703	0.239 / 0.058	0.306 / 0.162
Waypoint-45M	0.792	0.724	0.228 / 0.061	0.324 / 0.168
Gut foundation models				
Pope et al.	0.799	0.741	0.234 / 0.054	0.331 / 0.179
BiomeGPT	0.826	0.716	0.253 / 0.065	0.317 / 0.159
GUT-FORMer	0.839	0.752	0.247 / 0.068	0.341 / 0.193
MBEmbed	0.817	0.761	0.260 / 0.062	0.352 / 0.186
TaxoMFM (ours)	**0.853**	**0.776**	**0.266 / 0.072**	**0.369 / 0.210**

C Detailed experimental settings

Cross-cohort evaluation tests whether disease-related information learned from source cohorts transfers to external populations. Tables 6–8 further examine the information retained in the representations through community perturbations, conditional recovery of rare species, and vocabulary changes. These experiments assess whether models distinguish compositional changes at different taxonomic scales, recover masked abundances from the visible community, and preserve sample distinctions across vocabularies. We describe the data splits, input construction, evaluation procedures, and metrics below.

C.1 Zero-shot cross-cohort transfer (Table 5)

We evaluate cross-cohort transfer for colorectal cancer (CRC), type 2 diabetes (T2D), and inflammatory bowel disease (IBD). Each task comprises labeled source cohorts and one external target test set, to which the trained model is applied directly. A separate pretraining split excludes all 1,107 target test samples from pretraining and downstream training. Table A9 summarizes the task-specific cohort splits.

Table A9: Source and target cohort splits for cross-cohort transfer. Positive and negative counts refer to the target test sets.

Task	Source samples	Target study	Target samples	Positive/ negative
CRC	1,517	Yang et al. (2021)	194	97/97
T2D	465	MetaCardis (Forslund et al., 2021)	783	549/234
IBD	318	HMP/IBDMDB (Lloyd-Price et al., 2019)	130	103/27

The public profiles were obtained from curatedMetagenomicData (Pasolli et al., 2017) and GMWI2 (Chang et al., 2024). Table A10 lists the source and target datasets, their resource-specific

study identifiers, and the PubMed identifiers of the corresponding original studies. Splits are organized by participant, with one sample per participant in each task.

Table A10: Resource study IDs and PubMed identifiers for cross-cohort transfer.

Task	Role	Resource	Resource study ID	PMID
CRC	Source	cMD	GuptaA_2019	31719139
CRC	Source	cMD	HanniganGD_2017	30459201
CRC	Source	cMD	ThomasAM_2018b	30936548
CRC	Source	cMD	VogtmannE_2016	27171425
CRC	Source	cMD	WirbelJ_2018	30936547
CRC	Source	cMD	YachidaS_2019	31171880
CRC	Source	GMWI2	Thomas_2019	30936548
CRC	Source	GMWI2	Yang_2020	31971861
CRC	Source	GMWI2	Yu_2015	26408641
CRC	Source	GMWI2	Zeller_2014	25432777
CRC	Target	GMWI2	Yang_2021	34799562
T2D	Source	cMD	KarlssonFH_2013	23719380
T2D	Source	GMWI2	Qin_2012	23023125
T2D	Target	cMD	MetaCardis_2020_a	34880489
IBD	Source	cMD	NielsenHB_2014	24997787
IBD	Target	cMD	HMP_2019_ibdmdb	31142855

C.2 Recovering compositional changes across taxonomic scales (Table 6)

We construct three magnitude-matched compositional perturbations in samples outside pretraining to assess recovery of changes in lineage abundance and within-lineage species proportions. Let x denote species relative abundances and $\mathcal{C}_g$ the species in genus g, with total abundance $P_g = \sum_{s\in\mathcal{C}_g} x_s$. We set the transferred abundance to $\delta = \alpha P_g$, with $\alpha = 0.10$. The three operations increase a single species by δ, proportionally scale all species in the genus to increase its total by δ, or transfer δ between two congeneric species. The first two remove an equal amount proportionally from the community outside the genus; the third preserves the genus total. In each original–perturbed pair, we mask the same detected species other than the donor and recipient, retaining its identity and setting its input abundance to zero. We independently renormalize the visible species abundances in each composition to sum to one. All higher-rank input abundances are then recomputed from the renormalized visible composition. The model reconstructs both complete compositions, and their difference estimates the true perturbation. Comparisons include full TaxoMFM, species-only, random taxonomy, and the variant without hierarchical composition learning.

Let Δ denote the perturbed value minus the original value, with hatted quantities calculated from the reconstructed composition $\hat{x}$. Writing the within-genus proportion as $q_{s|g} = x_s/P_g$, Parent error and Conditional error quantify recovery of changes in genus totals and within-genus proportions, respectively. They are normalized by δ and α:

$$\begin{aligned} E_{\text{parent}} &= \frac{|\Delta\hat{P}_g - \Delta P_g|}{\delta}, \\ E_{\text{conditional}} &= \frac{1}{\alpha}\sum_{s\in\mathcal{C}_g} |\Delta\hat{q}_{s|g} - \Delta q_{s|g}|. \end{aligned} \tag{11}$$

Localization error compares the fractions of predicted and true absolute changes that occur within the target genus. Denoting this fraction for a change vector v by $L_g(v)$ gives

$$L_g(v) = \frac{\sum_{s\in\mathcal{C}_g} |v_s|}{\sum_s |v_s|}, \qquad E_{\text{localization}} = |L_g(\Delta\hat{x}) - L_g(\Delta x)|. \tag{12}$$

Combined error, $E_{\text{combined}} = E_{\text{parent}} + E_{\text{conditional}}$, aggregates errors in recovering genus totals and within-genus proportions. All metrics are averaged across samples and the three perturbations; lower values indicate better recovery.

C.3 Conditional abundance recovery of rare species (Table 7)

We divide eligible species detected in the external evaluation samples into four approximately equal-sized groups according to their prevalence in the pretraining data. Groups Q1–Q4 range from the rarest to the most common species. In each external sample, a detected target species is masked together with other randomly selected species. Each mask covers approximately 20% of detected species and retains at least two visible species. All methods use the same targets and masks. Species identities are retained, masked abundances are set to zero, and visible abundances are renormalized; higher-rank inputs are aggregated from visible species. Let x denote the original abundances, V the visible species, and t the target species. The evaluation target is the log abundance ratio of the target species to the visible community total. The model computes the corresponding prediction from its reconstructed composition $\hat{p}$:

$$\tilde{x}_s = \begin{cases} x_s / \sum_{j \in V} x_j, & s \in V, \\ 0, & s \text{ is masked}, \end{cases} \qquad y_t = \log_2 \frac{x_t}{\sum_{j \in V} x_j}, \qquad \hat{y}_t = \log_2 \frac{\hat{p}_t}{\sum_{j \in V} \hat{p}_j}. \tag{13}$$

Table 7 reports the absolute log-ratio error $e_t = |\hat{y}_t - y_t|$. We first average repeated masks and repeated samples within participants, then average across participants for each species. Species are weighted equally within each prevalence group; lower errors indicate more accurate recovery. Two priors estimated from the training data provide reference predictions. The Species identity prior uses each species' median conditional log abundance ratio. The Genus proportion prior estimates target abundance from its typical within-genus proportion and the abundance of currently visible congeneric species. These comparisons assess the contribution of community context to recovering rare and common species.

C.4 Representation consistency across taxonomic vocabularies (Table 8)

We construct current and legacy vocabulary views for 9,044 HPP samples from the pretraining data to evaluate compositional information retention and representation consistency. Both views originate from the same abundance profile. The current view uses the MetaPhlAn 4 vocabulary; the legacy view uses identical species names and 33 curated MetaPhlAn 3–4 species mappings. TaxoMFM assigns abundance unmatched at the species level to the nearest shared genus, family, or order; abundance without a shared ancestor is removed. Structural controls include leaf-only, which retains only matched species, species-only, and random taxonomy. External baselines encode both views after mapping them to their respective model vocabularies.

Legacy abundance retention is the fraction of original abundance representable after vocabulary matching and ancestor aggregation, averaged across samples. To assess representation consistency, let u_i and v_i denote the L2-normalized current and legacy representations. Cosine separation measures the similarity difference between matched samples and mismatched pairs formed by a cyclic shift. For N samples and shifted index $\pi(i)$,

$$\text{Separation} = \frac{1}{N} \sum_{i=1}^{N} \left[u_i^\top v_i - u_i^\top v_{\pi(i)} \right]. \tag{14}$$

Bidirectional Recall@k queries all samples in one view using the other view and ranks candidates by cosine similarity. It averages the fraction of correct matches retrieved among the top k candidates across both directions:

$$R@k = \frac{100\%}{2N} \sum_{i=1}^{N} \big[\mathbf{1}\{\text{rank}_{\text{current}\to\text{legacy}}(i) \le k\} + \mathbf{1}\{\text{rank}_{\text{legacy}\to\text{current}}(i) \le k\} \big], \qquad k \in \{1, 5\}. \tag{15}$$

Here, $\text{rank}(i)$ is the rank of the correct match for sample i, and $\mathbf{1}$ is the indicator function. Retention measures preserved compositional information, whereas separation and retrieval recall measure sample distinguishability; higher values are better for all metrics.

D COMPUTATIONAL EFFICIENCY AND RESOURCE REQUIREMENTS

We evaluated representation extraction for five models on 1,024 samples using NVIDIA RTX PRO 6000 Blackwell GPUs. All models used existing pretrained checkpoints and BF16 mixed precision. Measurements were repeated five times after warm-up. Table A11 reports steady-state inference performance, with batch size 1 for latency and batch size 8 for throughput and peak GPU memory.

Table A11: Computational efficiency of representation extraction.

Model	Parameters (M)	GFLOPs / sample	Latency (ms) ↓	Throughput (samples/s) ↑	Peak memory (GiB) ↓
MGM	6.83	0.692	3.399	2,113	0.063
Waypoint-45M	45.15	4.844	5.732	1,341	0.237
BiomeGPT	14.73	1.991	4.262	1,809	0.114
GUT-FORMer	1.22	0.483	0.951	8,392	0.029
TaxoMFM	18.25	5.626	0.759	3,852	0.412

As shown in Table A11, TaxoMFM maintains efficient inference while incorporating multilevel taxonomic information. It achieves the lowest single-sample latency of 0.759 ms and a throughput of 3,852 samples/s, outperforming MGM, Waypoint-45M, and BiomeGPT in both measures and trailing only the lighter GUT-FORMer in throughput. Despite requiring more computation and peak memory, TaxoMFM executes efficiently on the GPU, balancing multilevel representation with inference speed.