Exploring Set-Aggregated Genome Embeddings for Microbiome Abundance Prediction

Younhun Kim [1 2] **Georg K. Gerber** [1 2] **Travis E. Gibson** [1 2]

Abstract

Microbiome functions are encoded within the genes of the community-wide metagenome. A natural question is whether properties of a microbial community can be predicted just from knowing the raw DNA sequences of its members. In this work, we employ set-aggregated genome embeddings (SAGE) to predict community-level abundance profiles, exploiting the few-shot learning capabilities of genomic language models (GLMs). We benchmark this approach to show improved generalization on novel genomes compared to classical bioinformatics approaches. Model ablation shows that community-level latent representations directly result in improved performance. Lastly, we demonstrate the benefits of intermediate transformations between latent representations and demonstrate the differences between GLM embedding choices.

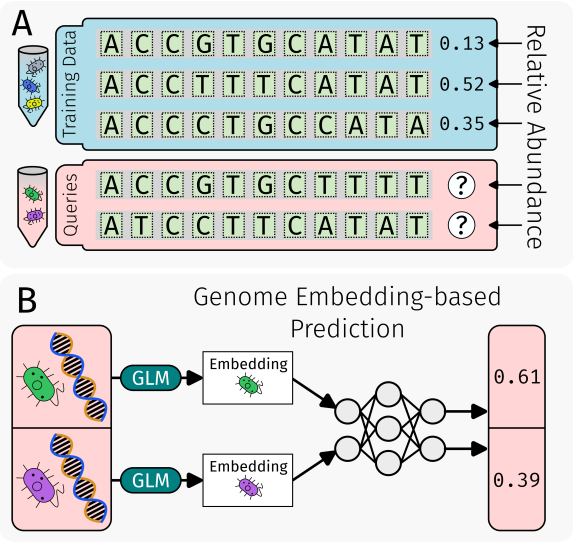

Figure 1. Overview of (a) the predictive task and (b) the deep-learning approach. The primary goal is to use genomic sequences directly to produce abundance predictions per microbiome sample.

1. Introduction

Microbiomes are intensely studied for their wide array of impactful applications, notably in the area of human health and medicine (Gilbert et al., 2018; Fan & Pedersen, 2021). A big central question shared by all major applications is: *can we predict properties of never-before-seen microbiomes?* That is, given knowledge of what taxa are present, can we make predictions about a target desired property? This is spurred by a big hurdle in microbiome science, namely that it is difficult to isolate & culture arbitrary bacterial strains at scale. Thus, it is infeasible to understand all novel microbiomes by growing or isolating each individual member (Lagier et al., 2018).

Thanks to advancements in molecular sequencing, it is now commonplace to perform DNA sequencing of microbiomes, producing reads from the "pool" of genomes (*metagenome*) present in any sample. Very few algorithms utilize genomic sequence information end-to-end. Recent line of research has showed that genomic language models (GLMs) such as Evo (Nguyen et al., 2024) and DNABERT (Ji et al., 2021) can be feature-extracted for microbiome tasks. The two prototypical tasks which have appeared in literature are environmental source prediction and host phenotype prediction (Ludwig et al., 2025; Yoo & Rosen, 2025).

However, many present-day questions in the microbiome field involve predictions on a per-species or per-strain basis, one for each found in the input community. Examples include bacterial interactions (Gibson et al., 2025), perturbational response (Ng et al., 2019), and engraftment prediction after fecal microbiota transplant (Smillie et al., 2018). It is also unclear whether the community-level pooling representations, as seen in Set-Attention (Lee et al., 2019), is helpful for encoding microbiomes, if it can be improved, or would benefit from different choices of GLMs.

To understand the above questions, we investigate the task of *relative abundance prediction* directly from genomic sequences (Figure 1a). Here, we assume that the raw sequencing reads from each microbiome experiment have been

[1]Brigham and Women's Hospital, Boston, MA, USA [2]Harvard Medical School, Boston, MA, USA. Correspondence to: Travis E. Gibson <tegibson@bwh.harvard.edu>.

pre-processed into clusters with representative sequences ("taxa"). Our datasets come with a "ground truth" table of each dataset; each (i, j) cell is the observed abundance of taxa j in microbiome sample i, normalized so each row sums to 1. The target objective is to predict the abundance table (each (i, j) cell) only using information about which taxa j are found per sample (the support of each row) and the representative gene sequence(s) of each j. The predictive models we study are neural networks which take the nucleotide sequence embeddings as input (Figure 1b), with a particular focus on a family of models that use permutation-invariant pooling on to latently represent each taxa and community.

2. Datasets

We used two datasets to train and evaluate our models. The first is the American Gut Project (AGP) (McDonald et al., 2018), a 16S-V4 amplicon sequencing dataset processed via the Human Microbiome Compendium (Turnbaugh et al., 2007) using DADA2 (Callahan et al., 2016). This yields taxa-specific gene sequences referred to as Amplicon Sequence Variants (ASVs). The second is the whole-metagenome sequencing (WMS) dataset from the MetaPhlAn4 publication (Blanco-Míguez et al., 2023), where taxa are Species-level Genome Bins (SGBs). ASVs are *de novo* sequences; SGBs are reference-database clusters defined by marker genes. After quality screening (Appendix A), the 16S dataset contained 3,851 healthy adult samples and the WMS dataset contained 6,015 healthy adult samples. We train separate models per dataset.

2.1. Train-Test Split

We benchmark using two main types of splitting methodologies. The first is a simple, uniformly random train-test split ("random split") where samples are globally shuffled and split at a 4:1 train/test ratio. Second, to test generalization to unseen microbiome compositions, we constructed an out-of-distribution test split ("JS-maximizing split"). To be precise, we computed pairwise Jensen-Shannon (JS) divergence between samples and applied PCoA (Gower, 1966) to the resulting distance matrix. The first principal coordinate was used to split samples at a 4:1 train/test ratio, placing more compositionally distant samples in the test set.

3. Predictive Models

Here, we define methods sorted into two categories: (1) distance-aware, sequence-naive (methods can only use distances computed by bioinformatics sequence comparison), and (2) sequence-aware embedding methods. #2 is the main focus of our paper; #1 serves as a baseline.

3.1. Nearest-Neighbor Baseline Methods

We assume knowledge of a pairwise distance metric δ between taxa, spanning both the training-set and test-set taxa. For 16S, δ is derived from % sequence identity via Mothur (Schloss et al., 2009); for WMS, δ is the phylogenetic tree distance from PhyloPhlAn3 (Asnicar et al., 2020). Each outputs a relative abundance probability vector:

1. *Uniform*: Assigns $1/N$ relative abundance to each taxon.
2. *Global Average*: maps each query taxon to its nearest neighbor in the training set; weights by geometric-mean abundance across all training samples.
3. *kNN Average*: same nearest-neighbor mapping of taxa, but weights by geometric-mean abundance within the k most similar training samples by Jaccard similarity.

These methods were previously benchmarked in (Asher & Bashan, 2024) and shown to outperform sequence and distance-naive neural network. Distance-aware neural networks were not tested.

3.2. Phylogenetic Embedding Methods

We include distance-aware, sequence-naive neural network methods by first constructing UMAP embeddings (McInnes et al., 2018) from δ, sweeping output dimensions $d \in \{20, 40, 60, 80, 100\}$. Each taxon is represented as a $1 \times d$ vector and passed into a deep-learning model (Section 4) specific to the choice of d.

3.3. Genomic Sequence Embeddings

This family of sequence-aware neural networks is the main focus of our paper. We compare three pre-trained genomic language models (GLMs): Evo (Nguyen et al., 2024), Evo 2 (Brixi et al., 2025), and DNABERT-S (Zhou et al., 2024). These models were selected to represent the current landscape of microbe-focused GLM and restricted to those explicitly tested on microbial genome prediction, as opposed to GLMs primarily trained & solely evaluated on human sequences. For Evo/Evo 2, we extract intermediate layer representations ($d = 4096$); for DNABERT-S we use the final layer ($d = 768$). Weights were not fine-tuned. Each 16S ASV is embedded as a single d-dimensional vector; each WMS SGB is embedded across its g PhyloPhlAn marker genes, producing a $g \times d$ matrix. For WMS, we apply incremental PCA (Ross et al., 2008) across the entire reference database gene set (independent from any train-test split) to reduce dimensionality and manage memory constraints.

4. Deep Learning Architecture & Training

Each input taxa is generically a $g \times d$ matrix, representing g genes embedded using a d-dimensional GLM output. Thus,

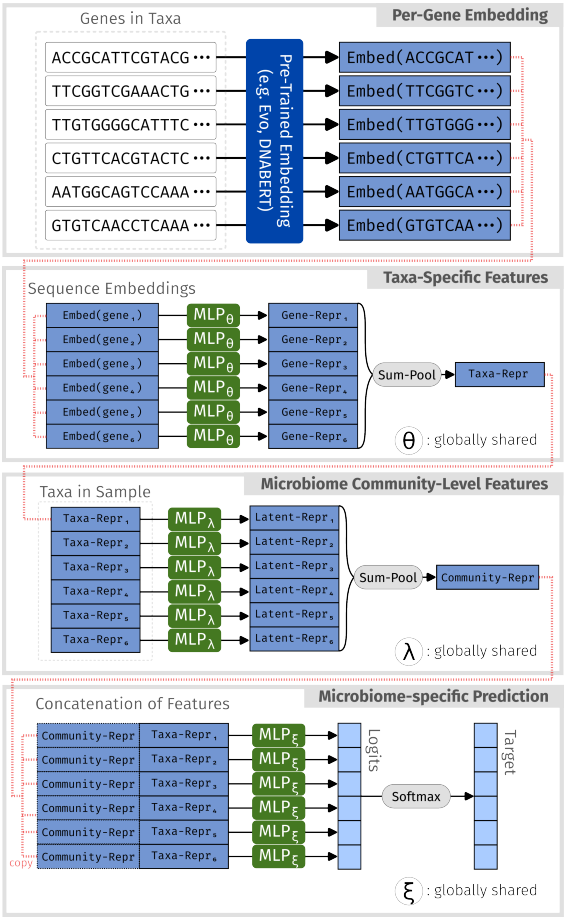

Figure 2. Architectural diagram of the benchmarked SAGE implementation. The architecture is divided into modules implementing successive hierarchies: microbiomes are sets of taxa, taxa are sets of genes, and genes are represented by sequences. Intermediate MLPs are broadcasted across each slice, to serve as latent transformations between represented spaces.

a microbiome sample consisting of t taxa is a (padded and masked) $t \times g \times d$ tensor. To handle these types of input representations, we dubbed the architectural strategy *Set-Aggregated Genomic Embedding* (SAGE), to refer to the pooling of genomic features in each taxa.

The architecture (Figure 2) can be understood in terms of its *modules* (gray boxes). Each "level" of aggregation is done by first transforming each input representation using a multi-layer perceptron (MLP) submodule, which is broadcasted across each of the constituent feature vector slices. Then, the transformed latent representations are sum-pooled. This is a canonical example of a "permutation-equivariant" neural network (Lee et al., 2019), which has the property of being agnostic to the ordering of the microbial taxa, and to the ordering of the genes. The final outputs are taxa-specific logits, which is produced via a third MLP submodule that

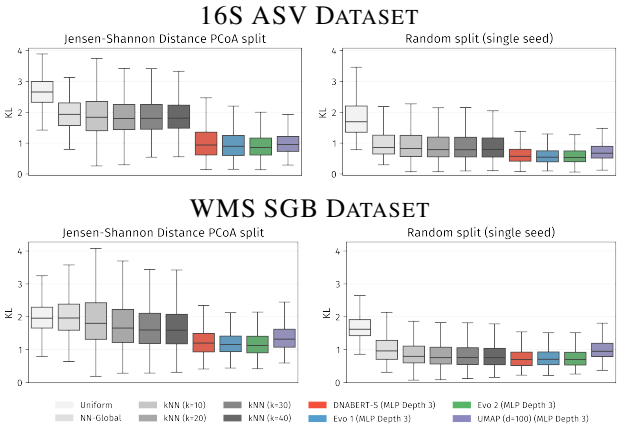

Figure 3. Benchmarking results on the 16S ASV dataset (top row) and the WMS SGB dataset (bottom row). Shown are the test-set median KL loss values for the JS-maximizing split (left) and on a random split (right) for a single seed. Summary statistic across multiple random split seeds can be found in the Appendix.

uses the concatenation of the taxa-specific and community-wide features.

Alternatively, we could have hierarhically composed Set-Attention blocks (Lee et al., 2019) as seen in (Ludwig et al., 2025; Yoo & Rosen, 2025), instead of sum-pooling. Our ablation analysis (Figure 5) shows that performance saturates on our current architecture, suggesting that additional complexity and capacity may not meaningfully improve generalization on our datasets. However, an architecture like Set-Attention is quite attractive for future work, as the inductive bias better represents interactions between taxa found in dynamical systems models (Gibson et al., 2025; Thompson et al., 2026).

We trained all SAGE models using the Adam optimizer (Kingma & Ba, 2014) on the Kullback-Liebler (KL) divergence loss, initialized with 10^{-4} learning rate and a cosine annealing schedule. Lastly, we did not use the test data in any way during training, such as early stopping, to avoid data leakage. Instead, all models were trained for exactly 80 epochs, which appeared sufficient to reach a local minimum of the training loss in all scenarios.

5. Evaluation Results

5.1. *De Novo* Sequence, Single-Gene Evaluation (16S ASV)

We evaluate test-set KL divergence as our primary error metric (Figure 3, top row; median loss values in Table S1) on the 16S dataset. A quick sanity check shows that the trend from (Asher & Bashan, 2024) is recovered, where kNN outperforms Global Averaging. On the JS-maximizing split, all deep-learning methods consistently outperform all baselines

(p-values in Tables S2,S3). Apparent differences between GLM choices can largely be attributed to embedding sizes and parameter counts (Section 5.3). Varying MLP depth generally does not make a difference, except for random splits on UMAP embeddings (Figure S1(a,b)).

Across random splits, all methods improve with sequence-based models reaching KL ≈ 0.55. Phylogenetic embedding models matched sequence models closely in KL. This is true across all tested UMAP output dimensions d, as model performance increases with d but with diminishing returns and only on randomized splits (Figure S1(c,d)).

5.2. Reference Sequence, Multi-Gene Evaluation (WGS SGB)

The WMS benchmark (Figure 3, bottom row; median loss values in Table S4) showed similar patterns, although there was a smaller overall performance gap between the averaging methods and the neural networks. Still, the sequence-embedding models outperformed baselines in KL (p-values in Tables S5,S6). Adding more marker genes (PhyloPhlAn + MetaPhlAn) did not improve performance (Figure S2), suggesting saturation with the first gene set alone; a full gene ablation analysis is left to future work.

5.3. Model Ablation

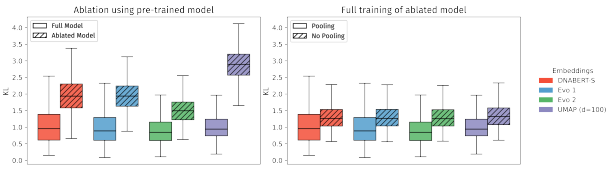

Figure 4. Benchmark of SAGE models for ablating community-level pooling. Models are either (a) given zero vectors for community-level representations during evaluation, or (b) are fully re-trained with community representations removed.

To verify the effectiveness of including community representations, we conducted an ablation study on the community-level pooling module (Figure 4), using the JS-maximizing split 16S dataset. Models were either given zero-vectors for community representations during evaluation (Figure 4, left), or were re-trained with the representations removed outright (Figure 4, right). The ablated intermediary representations resulted in worse performance in either setting, across all GLM embedding types.

Lastly, we explored ablation of MLP_θ, MLP_λ from Figure 5, and performed model capacity ablation by sweeping model sizes. A priori, the MLPs' usefulness is not obvious: a previous work (Yoo & Rosen, 2025) showed that pooling without intermediate transformations results in worse performance, but only marginally and selectively depending on task. Our

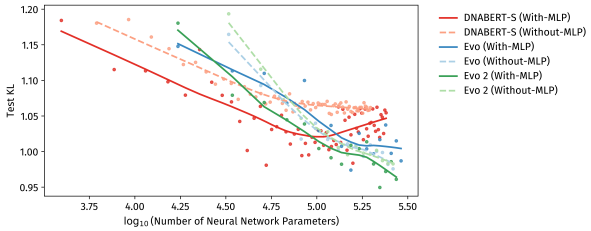

Figure 5. MLP and model capacity ablation analysis. Models are trained with and without the MLP transformations in the taxa-pooling and community-pooling modules, sweeping across parameter counts. Curves are local fits (LOWESS).

analysis found that the "without-MLP" architectures either had poorer (DNABERT-S variants) or approximately equal (Evo, Evo 2) test-loss scaling as parameter count was increased. This analysis also showed that DNABERT-S results in SAGE performance quickly worsening (due to overfitting and not undertraining), whereas Evo 2 results in SAGE performance continuing to improve well past 10^5 parameters.

6. Discussion

Our results support the main hypothesis: genome embedding aggregations results in better generalization to novel sequences unseen during training. Still, the WMS dataset's smaller performance gap was curious to us. We speculate that this is due to the dataset's nature where species-level reference sequences are used, rather than *de novo* variants.

The ablation findings confirm that including extra layers produce more efficiently trainable models. Our dataset is likely too small to extrapolate a full "scaling law" in the manner of (Kaplan et al., 2020; Hoffmann et al., 2022). Rather, this analysis characterizes the sensitivity of performance to model capacity given a limited compute budget (memory capacity) and limited dataset size. While the former is solvable with a larger budget, the latter is not easily fixable. Indeed, microbiome sample collection is difficult to scale, particularly in human studies with a huge diversity of both host and microbiome. Our observed test KL loss of $0.8 \sim 0.9$ may seem high, but the reader should note that we did not use any of the host features in the predictions. With a larger, cleaner, and more well-annotated dataset, it may be possible to push this performance even further.

In future work, we hope to explore other types of inductive biases depending on microbiome-specific tasks. We are especially excited to explore the potential of longer *de novo* long-read sequencing using this approach, which would provide the "best of both worlds" of the single-gene and whole-genome datasets.

References

Asher, E. E. and Bashan, A. Model-free prediction of microbiome compositions. *Microbiome*, 12(1):17, 2024.

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0. *Nature communications*, 11(1):2500, 2020.

Blanco-Míguez, A., Beghini, F., Cumbo, F., McIver, L. J., Thompson, K. N., Zolfo, M., Manghi, P., Dubois, L., Huang, K. D., Thomas, A. M., et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4. *Nature biotechnology*, 41(11):1633–1644, 2023.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., et al. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pp. 2025–02, 2025.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583, 2016.

Fan, Y. and Pedersen, O. Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, 19(1): 55–71, 2021.

Gibson, T. E., Kim, Y., Acharya, S., Kaplan, D. E., DiBenedetto, N., Lavin, R., Berger, B., Allegretti, J. R., Bry, L., and Gerber, G. K. Learning ecosystem-scale dynamics from microbiome data with mdsine2. *Nature Microbiology*, 10(10):2550–2564, 2025.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. Current understanding of the human microbiome. *Nature medicine*, 24(4):392–400, 2018.

Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.-M., Fournier, P.-E., and Raoult, D. Culturing the human microbiota and culturomics. *Nature Reviews Microbiology*, 16(9): 540–550, 2018.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Ludwig, D. W., Guptil, C., Alexander, N. R., Zhalnina, K., Wipf, E. M.-L., Khasanova, A., Barber, N. A., Swingley, W., Walker, D. M., and Phillips, J. L. Setbert: the deep learning platform for contextualized embeddings and explainable predictions from high-throughput sequencing. *Bioinformatics*, 41(7):btaf370, 2025.

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., Aksenov, A. A., Behsaz, B., Brennan, C., Chen, Y., et al. American gut: an open platform for citizen science microbiome research. *Msystems*, 3(3):10–1128, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Ng, K. M., Aranda-Díaz, A., Tropini, C., Frankel, M. R., Van Treuren, W., O'Loughlin, C. T., Merrill, B. D., Yu, F. B., Pruss, K. M., Oliveira, R. A., et al. Recovery of the gut microbiota after antibiotics depends on host diet, community context, and environmental reservoirs. *Cell host & microbe*, 26(5):650–665, 2019.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723): eado9336, 2024.

Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H. Incremental learning for robust visual tracking. *International journal of computer vision*, 77(1):125–141, 2008.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. Introducing

mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

Smillie, C. S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., Hohmann, E. L., Staley, C., Khoruts, A., Sadowsky, M. J., et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell host & microbe*, 23(2):229–240, 2018.

Thompson, J., Connors, B. M., Zavala, V. M., and Venturelli, O. S. Physics-constrained neural ordinary differential equation models to discover and predict microbial community dynamics. *Proceedings of the National Academy of Sciences*, 123(13):e2517661123, 2026.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

Yoo, H. and Rosen, G. Abundance-aware set transformer for microbiome sample embedding. *arXiv preprint arXiv:2508.11075*, 2025.

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., Wang, Z., and Liu, H. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *arXiv preprint arXiv:2402.08777*, 10, 2024.

A. Data Processing & Quality Control

A.1. American Gut Project (16S Amplicon)

The AGP dataset sequenced amplicons of a short ∼125bp hypervariable segment (the v4 region) of the 16S gene. We first downloaded the DADA2-processed abundance tables and ASV sequence FASTA files from the Human Microbiome Compendium. To remove outliers which might cause unnecessarily confusing technical bias, we initially performed a 2-step filtering on this dataset. First, to reduce the effect of extremal deviations in molecular amplification and/or sequencing, we removed samples with overall read count in the top 2.5% and bottom 2.5%. Next, to eliminate ASVs and samples with the highest likelihood of having anomalous DADA2 outputs, we removed all ASVs that do not exceed a read count of 10 in at least one sample, and then removed all samples that either have too few ASVs (≤ 50) or too many ASVs (≥ 250). Finally, we removed ASVs which did not map to Bacteria (e.g. Archaeal ASVs) by running Mothur (Schloss et al., 2009) v1.48.5, a software specialized for 16S sequence analysis. After filtering, the result was a collection of 3,851 samples with a collective total of 12,765 ASVs. These samples are largely concentrated in the USA, UK and Canada, although several other regions across the globe are represented as well.

A.2. MetaPhlAn4 study (WGS)

The data used in Blanco-Míguez et al. (2023), itself a compilation of multiple datasets, was requested from the authors of that work. We obtained the abundance table output of MetaPhlAn4 on all samples, as well as the intermediate outputs of PhyloPhlAn3. The latter was required, because there are actually two sets of marker genes per SGB: the marker genes used for SGB abundance estimation (MetaPhlAn4 markers, ∼200 genes per SGB), and the genes used to construct the phylogenetic tree (PhyloPhlAn3 markers, ∼300 genes per SGB). We began with 6,053 samples from healthy adults; we removed the 38 samples which had ≤ 50 SGBs. The resulting dataset consisted of 6015 samples; the dataset is most largely concentrated in the USA, UK and China, though various countries are represented across the globe.

B. Extra Figures and Tables

Method	JS-Maximizing Split (KL)	Random Split (KL)
Uniform	2.6562	1.6868
NN-Global	1.9337	0.8676
kNN ($k = 10$)	1.8382	0.8353
kNN ($k = 20$)	1.7981	0.8089
kNN ($k = 30$)	1.8044	0.7894
kNN ($k = 40$)	1.8097	0.8039
DNABERT-S	0.9374	0.5705
Evo 1	0.8972	0.5444
Evo 2	**0.8578**	**0.5372**
UMAP ($d = 100$)	0.9528	0.6718

Table S1. (16S ASV dataset) Test set KL divergence evaluation of methods. **JS-Maximizing Split**: median test KL for the D_{JS} maximizing split. **Random Split**: median-of-median test KL across ten random seeds. The minimal value in each column is highlighted in bold.

	Uniform	NN-Global	kNN ($k = 10$)	kNN ($k = 20$)	kNN ($k = 30$)	kNN ($k = 40$)	DNABERT-S	Evo 1	Evo 2	UMAP ($d = 100$)
Uniform	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
NN-Global	**0.000***	–	0.168	**0.001***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 10$)	**0.000***	0.168	–	**0.037***	**0.023***	0.053	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 20$)	**0.000***	**0.001***	**0.037***	–	0.891	0.850	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 30$)	**0.000***	**0.000***	**0.023***	0.891	–	0.479	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 40$)	**0.000***	**0.000***	0.053	0.850	0.479	–	**0.000***	**0.000***	**0.000***	**0.000***
DNABERT-S	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	0.290
Evo 1	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.006***
Evo 2	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***
UMAP ($d = 100$)	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	0.290	**0.006***	**0.000***	–

Table S2. (16S ASV dataset) Pairwise Wilcoxon signed-rank, two-tailed p-values (BH-corrected) on per-sample KL, JS-maximizing split. Bold*: $p < 0.05$.

	Uniform	NN-Global	kNN ($k = 10$)	kNN ($k = 20$)	kNN ($k = 30$)	kNN ($k = 40$)	DNABERT-S	Evo 1	Evo 2	UMAP ($d = 100$)
Uniform	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
NN-Global	**0.000***	–	**0.046***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 10$)	**0.000***	**0.046***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 20$)	**0.000***	**0.000***	**0.000***	–	**0.021***	**0.029***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 30$)	**0.000***	**0.000***	**0.000***	**0.021***	–	0.185	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 40$)	**0.000***	**0.000***	**0.000***	**0.029***	0.185	–	**0.000***	**0.000***	**0.000***	**0.000***
DNABERT-S	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***
Evo 1	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	0.986	**0.000***
Evo 2	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	0.986	–	**0.000***
UMAP ($d = 100$)	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–

Table S3. (16S ASV dataset) Pairwise Wilcoxon signed-rank, two-tailed p-values (BH-corrected) on per-sample KL, random split (single seed). Bold*: $p < 0.05$.

	JS-Maximizing Split (KL)	Random Split (KL)
Method		
Uniform	1.9565	1.6196
NN-Global	1.9614	0.9535
kNN ($k = 10$)	1.8006	0.8155
kNN ($k = 20$)	1.6577	0.7746
kNN ($k = 30$)	1.5987	0.7623
kNN ($k = 40$)	1.5917	0.7568
DNABERT-S	1.1993	0.6990
Evo 1	1.1533	0.7125
Evo 2	**1.1272**	**0.6985**
UMAP ($d = 100$)	1.3207	0.9378

Table S4. (WMS SGB dataset) Test set KL divergence evaluation of method. **JS-Maximizing Split**: median test KL for the D_{JS} maximizing split. **Random Split**: median-of-median test KL across ten random seeds. The minimal (best) loss value in each column is highlighted in bold.

	Uniform	NN-Global	kNN ($k = 10$)	kNN ($k = 20$)	kNN ($k = 30$)	kNN ($k = 40$)	DNABERT-S	Evo 1	Evo 2	UMAP ($d = 100$)
Uniform	–	**0.045***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
NN-Global	**0.045***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 10$)	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 20$)	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 30$)	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 40$)	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***
DNABERT-S	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***
Evo 1	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***
Evo 2	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***
UMAP ($d = 100$)	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–

Table S5. (WMS SGB dataset) Pairwise Wilcoxon signed-rank p-values (BH-corrected) on per-sample KL, JS-maximizing split. Bold*: $p < 0.05$.

	Uniform	NN-Global	kNN ($k = 10$)	kNN ($k = 20$)	kNN ($k = 30$)	kNN ($k = 40$)	DNABERT-S	Evo 1	Evo 2	UMAP ($d = 100$)
Uniform	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
NN-Global	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	0.125
kNN ($k = 10$)	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 20$)	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 30$)	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***
kNN ($k = 40$)	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***	**0.000***	**0.000***
DNABERT-S	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.020***	**0.000***
Evo 1	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–	**0.000***	**0.000***
Evo 2	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.020***	**0.000***	–	**0.000***
UMAP ($d = 100$)	**0.000***	0.125	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	–

Table S6. (WMS SGB dataset) Pairwise Wilcoxon signed-rank p-values (BH-corrected) on per-sample KL, random split (single seed). Bold*: $p < 0.05$.

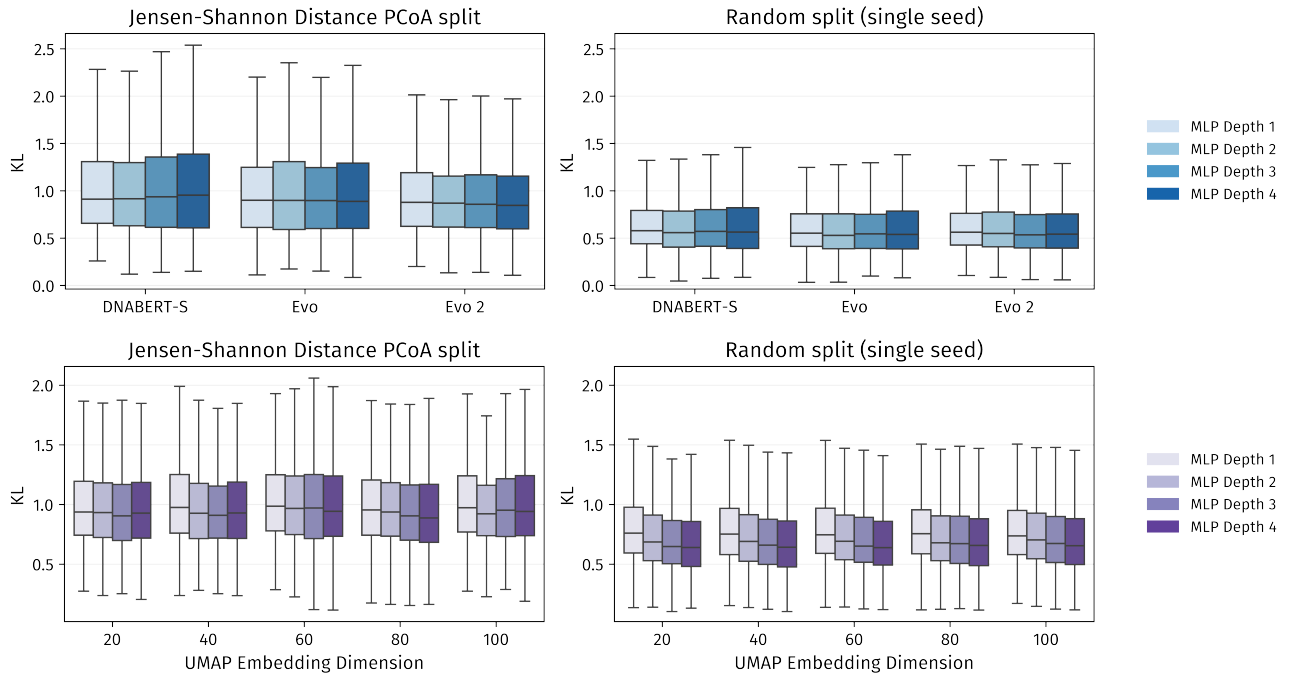

Figure S1. Model ablation comparison, globally varying the MLP capacity by depth. Each hidden layer's width is kept constant (width=32). The sequence embedding-based models are already saturated in performance and do not improve with increased depth. The same is true of UMAP but only for the PCoA split; it consistently improves in performance across all UMAP output dimensions as MLP depth increases.

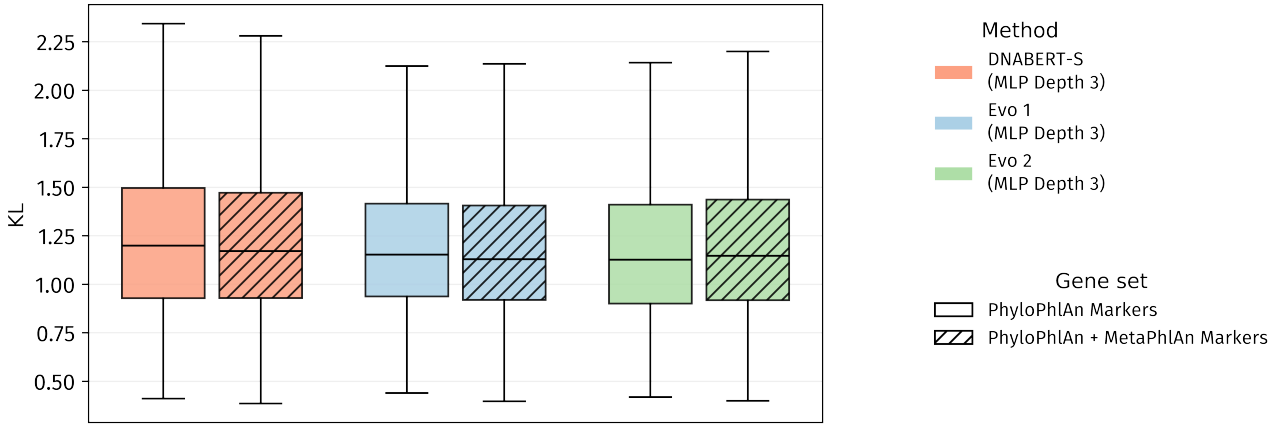

Figure S2. A feature-ablation analysis where we train two families of models: one that uses PhyloPhlAn SGB markers (found in the main text), and another that uses both PhyloPhlAn SGB markers and MetaPhlAn4 SGB markers.