LLMs Can Learn the Language of the Microbiome

Neythen J. Treloar [1] **Saif Ur-Rehman** [1] **Jenny Yang** [1]

Abstract

We explore the application of large language models (LLMs) to microbiome data, a domain that remains underexplored despite the rise of self-supervised learning in biology. We introduce Atlas, a large-scale pretraining dataset comprising over 539,000 data points from MGnify, spanning multiple DNA sequencing modalities including amplicon, assembly, and whole-metagenome data. Using Atlas, we train the Waypoint family of models, GPT-style causal language models trained to understand microbiomes. To enable standardized evaluation, we present Compass, a benchmark of eight downstream microbiome prediction tasks. We show that our pretrained Waypoint models outperform classical methods and prior foundation models, with gains driven by both dataset scale and representation choices. Our results establish pretrained LLMs as a strong and practical approach for microbiome prediction tasks.

1. Introduction

The human microbiome is a complex and highly variable ecosystem that plays a critical role in health, disease, and environmental biology. Advances in high-throughput DNA sequencing have enabled large-scale profiling of microbial communities across diverse contexts, producing datasets spanning multiple sequencing modalities. While these modalities provide complementary views of microbial composition and function, their heterogeneity poses challenges for unified modelling.

Recent progress in machine learning has been driven by large language models (LLMs) trained via self-supervision, with successful applications in genomics (Ji et al., 2021; Nguyen et al., 2023; Dalla-Torre et al., 2025; Brixi et al., 2025; Munsamy et al., 2026), proteomics (Hayes et al., 2025; Lee et al., 2023) , and whole-genome modelling (Wiatrak et al., 2025; Avsec et al., 2026). Applying this paradigm to microbiome data introduces a new modelling domain, centred on heterogeneous microbial community profiles derived from diverse sequencing modalities and typically scarce labelled data. Early work has begun to model taxonomic profiles as sequences for language model pretraining (Zhang et al., 2026; Pope et al., 2025; Medearis et al., 2026), but questions remain around the integration of varying DNA sequencing modalities, scaling behaviour and standardised evaluation.

We address these challenges by introducing Atlas, a large-scale microbiome pretraining corpus of over 539,000 samples from MGnify (Richardson et al., 2022), spanning diverse DNA sequencing modalities including amplicon, shotgun metagenomics, and metagenome-assembled genomes. We propose a tokenisation strategy that enables consistent representation across modalities and improves robustness to varying taxonomic resolution. We further present Compass, a benchmark suite of eight tasks covering biome classification, drug–microbiome interactions, drug degradation, and infant gut development. Finally, we pretrain the Waypoint family of models and show that pretraining yields consistent improvements across tasks, and our Waypoint models achieve state-of-the-art performance.

2. Results

2.1. Atlas: the Microbiome Pretraining Dataset

We assembled a large-scale pretraining corpus by systematically scraping taxonomic abundance data from the MGnify database (Richardson et al., 2022) across all available pipeline versions (v1.0–v5.0) and four sequencing modalities: amplicon 16S rRNA, whole-genome shotgun metagenomic, metagenomic assembly, and metatranscriptomic (see Appendix for details). The resulting raw collection spans over 4,100 unique studies, representing one of the largest aggregations of publicly available microbiome data to date. Before tokenisation we quality filter the dataset, following the procedure in (Zhang et al., 2026), by removing any taxa with relative abundance lower than 0.0001 and subsequently removing any samples with less than 10 taxa. We then tokenise the sequences of microbes at the genus level, applying a fallback strategy: when a genus-level assignment is not available, we use the most specific higher-rank classification

[1]Outpost Bio. Correspondence to: Neythen J. Treloar <neythen@outpost.bio>, Jenny Yang <jenny@outpost.bio>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

available. The final dataset comprises 539,308 microbiome datapoints, drawn from a broad range of environments including marine, freshwater, terrestrial, host-associated, and engineered ecosystems, reflecting the ecological diversity of the underlying MGnify resource.

2.2. The Waypoint Series of Models: Scaling Microbiome Foundation Models

We pretrained a series of GPT-2–style causal language models ranging from 6M to 170M parameters (not including token and positional embeddings). See Appendix Table 3 for architectural parameters. We also included an alternate 6M model matching the Microbial General Model (MGM) architecture (Zhang et al., 2026) (6M-MGM) enabling us to directly compare our dataset and tokenisation strategy to the existing state of the art microbiome foundation model, and a 85M-gpt2-small architecture corresponding to the GPT-2 small model (Radford et al., 2019). All models share the same tokeniser, context length (512), and pretraining procedure, so that differences in downstream performance can be attributed to model capacity alone.

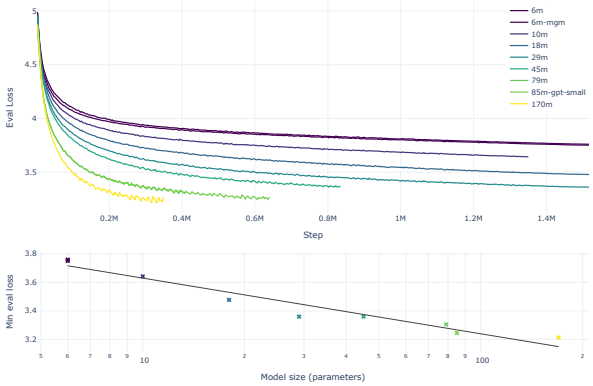

Figure 1. Top: pretraining evaluation loss curves for models with 6 - 170 million parameters. Larger models converge to lower evaluation loss. Some models exited training early because early stopping was used with a patience of 10. Bottom: the minimum evaluation loss decreases linearly with logarithmic model size.

2.3. Larger Scale Waypoint Models Show Improved Pretraining Performance

We pretrained all models with an autoregressive next token prediction objective (see Appendix for details). Pretraining evaluation loss curves for all nine models are shown in Figure 1. All models converge from a shared initial loss of $\sim$5.0, with larger models achieving consistently lower eval loss throughout training. Scaling up model capacity yields consistent improvements, with evaluation loss decreasing monotonically with model size. Interestingly the 79m and 85m-gpt-small models show very similar loss curves, with the difference in final loss being primarily due to the 79m model hitting the early stopping condition and exiting training, despite being slightly different in size and architecture. Larger models also converge in fewer steps, while smaller models continue to improve slowly, suggesting that the smaller models have not fully converged and further gains remain available with extended training.

2.4. Compass: The Microbiome Benchmark

To systematically evaluate these pretrained models we introduce Compass, a curated benchmark of eight predictive tasks spanning diverse gut and environmental microbiome tasks (see Appendix for details). The benchmark is designed to evaluate the ability of models to extract biologically meaningful information from compositional microbiome data. For now we focus primarily on gut microbiome problems, but this can be expanded upon in the future. The eight tasks are drawn from four independent datasets (Appendix Table 4) covering environmental origin (Richardson et al., 2022), drug–microbiome interactions (Shi et al., 2024; Mastrorilli et al., 2026), and infant gut microbiomes (Roswall et al., 2021).

2.5. Pretrained LLMs Outperform Non-pretrained LLMs and Baseline Models

We evaluated our pretrained Waypoint LLMs against classical baselines, including logistic regression (classification), ridge regression (regression), and random forest models. The sequencing data in each dataset used in the benchmark was processed differently by different authors, meaning that coverage of downstream taxa by the pretraining vocabulary varied substantially across datasets (Appendix Table 2). Coverage was complete for MGnify-biomes (Tasks 1–2), remained high for Mastrorilli et al. (Task 6), but was lower for Shi et al. (Tasks 3–5) and substantially reduced for Roswall et al. (Tasks 7–8). This variation in vocabulary overlap directly motivated evaluating the classical baselines both with and without filtering of taxa absent from the pretraining vocabulary (denoted "no unk"), where the latter removes all out-of-vocabulary taxa to isolate the effect of vocabulary mismatch and enable a fairer comparison to models operating under a fixed token set. We also compared our model against MGM (Zhang et al., 2026), an existing microbiome foundation model pretrained on 260K MGnify samples using a causal language modelling objective. Like our model, MGM represents each sample as a sequence of genus-level tokens; however, it does not support fallback to higher taxonomic ranks when genus labels are unavailable. MGM uses the same underlying GPT-2 architecture as our 6M-MGM model, providing a direct assessment of the effects of our

larger pretraining corpus and fallback tokenisation strategy. We attempted to include comparisons with two additional microbiome foundation models (Pope et al., 2025; Medearis et al., 2026), but to our knowledge these have not been made publically available. For each LLM, we additionally evaluated non-pretrained variants by reinitialising weights prior to benchmarking.

Figure 2 shows mean score across all eight benchmark tasks for each model and baseline, with replicates shown as individual points. Pretraining consistently and substantially improves performance over both non-pretrained counterparts and the classical baselines, and all pretrained models outperform the original MGM model, which itself outperforms all non-pretrained transformers. This demonstrates that without pretraining the relatively small per-task datasets are not sufficient to train these architectures. Additionally, the 6M-MGM architecture outperforms the original MGM despite sharing the same architecture, indicating that the larger pretraining dataset and the more informative input sequences enabled by fallback tokenisation contribute to improved performance.

The per-task breakdown (Appendix Figure 5) reveals that the benefit of pretraining is consistent across tasks, demonstrating the effectiveness of self supervised pretraining to learn general representations of microbiomes. From the per task breakdown we also see that the (no unk) variants of the baseline models clearly outperform all LLMs for Task 4 and Task 7, demonstrating that for some of the small per task datasets we see no clear advantage from using LLMs. Figures providing additional complementary metrics for the classification tasks can be found in the appendix (Appendix Figures 7, 8, 9).

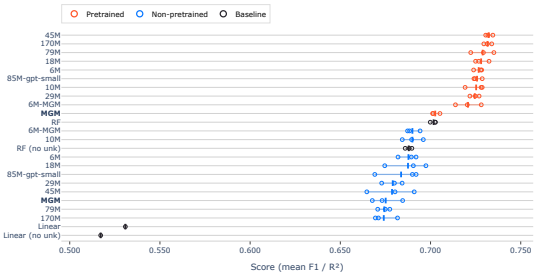

Figure 2. Overall benchmark scores for all models, taken as the average of the scores across the 8 benchmark tasks, where classification tasks (1-5 and 7-8) are scored using macro averaged F1 and the regression task (6) is scored using R^2.

2.6. LLMs Outperform Baselines at Obtainable Dataset Sizes

To understand how model performance scales with the amount of available training data, we plot the difference in the benchmark score of the best performing Waypoint LLM (45M) compared to the RF (no unk) baseline against benchmark training dataset size (Figure 3). We see that transformer models increasingly outperform the baseline as the number of training examples grows. Notably, from approximately 10k training examples onwards, our Waypoint LLM consistently achieves higher mean scores than the baseline. The pretrained Waypoint LLM outperforms the non-pretrained model across the full range of dataset sizes, suggesting that pretraining provides a consistent increase in performance within the range of dataset sizes in the benchmark. Notably, for small datasets (below 1k examples), Waypoint LLM underperforms the baseline, highlighting that a minimum quantity of training data is required before LLMs become competitive. These results suggest that pretrained LLMs are a practical choice for the tasks studied here whenever training datasets of at least 10k examples can be assembled.

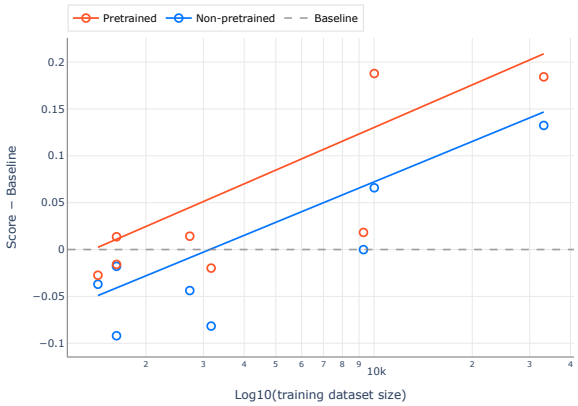

Figure 3. The difference in mean score of transformer models compared to the mean score of the random forest baseline, against the size of the training dataset for each task in the benchmark. At 10k datapoints and above, all pretrained transformers are better than the baseline. Solid lines are ordinary least squares lines of best fit.

2.7. Pretraining Enables Model Scaling

We next investigate how benchmark performance varies as a function of model size, and in particular whether pretraining alters scaling behavior. Figure 4 shows benchmark scores for the three replicates for both the pretrained and non-pretrained LLMs. A clear divergence emerges between pretrained and non-pretrained models as model size increases.

Pretrained models exhibit a modest but consistent improvement in performance with increasing parameter count. In contrast, non-pretrained models show a degradation in performance as model size grows, with larger models failing to translate additional capacity into improved task performance. This difference suggests that pretraining is critical for effectively utilizing increased model capacity. Without pretraining, larger models may overfit or fail to generalize, leading to negative scaling with increased size. Conversely, pretraining provides a strong initialization that allows larger models to leverage their additional representational capacity, resulting in steady performance gains. This finding has practical implications for model selection: investing in larger architectures is only justified when pretraining is also employed. Overall, these results demonstrate that pretraining on microbiomes fundamentally changes the scaling properties of LLMs, enabling positive performance scaling with model size and unlocking the benefits of increased capacity that are otherwise inaccessible.

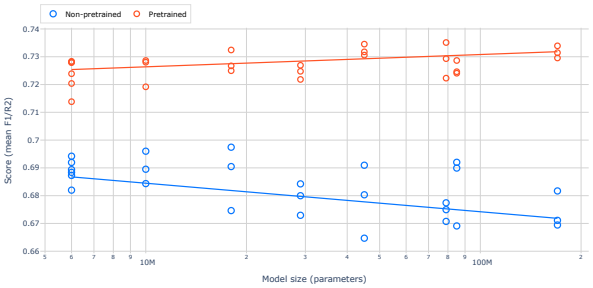

Figure 4. Benchmark performance as a function of model size for pretrained and non-pretrained transformer models. Each point represents the score from an individual training replicate, solid lines are ordinary least squares lines of best fit.

3. Discussion

In this work, we introduced Atlas, Compass, and the Waypoint model family: respectively, a large scale microbiome pretraining dataset, a benchmark spanning eight predictive tasks across gut and environmental microbiomes, and the new state of the art in microbiome foundation models. Atlas integrates data from multiple DNA sequencing modalities, including amplicon, shotgun metagenomics, and metagenome-assembled genomes, enabling unified modelling across heterogeneous microbial community profiles. We systematically evaluate pretrained and non-pretrained LLMs across a range of dataset and model scales. Our results demonstrate three key findings: (i) self-supervised pretraining on large, heterogeneous microbiome datasets consistently improves downstream performance, (ii) pretrained LLMs outperform classical baselines at realistic dataset sizes, and (iii) pretraining is necessary to realise positive model scaling.

We demonstrate that pretraining consistently improves performance on microbiome tasks, providing a practical path to improved performance without additional wetlab data generation. This is particularly important in microbiome settings, where labelled data are often scarce and costly to obtain. We show that a model with identical architecture to the original MGM model (Zhang et al., 2026), when pretrained on Atlas, achieves superior performance, demonstrating that dataset scale and tokenisation are key independent drivers of model quality. From a practical perspective, pretrained LLMs outperform random forest baselines at around 10k labelled samples, a regime already attainable in many studies and increasingly feasible with large-scale cohorts and high-throughput screening (Mastrorilli et al., 2026). We further show that pretraining is necessary to unlock the benefits of model scaling: without it, larger models degrade in performance, whereas pretrained models exhibit positive scaling behaviour.

This study has several limitations. First, Compass focuses on gut and environmental microbiomes and does not capture the full diversity of microbiome applications, including other body sites such as oral, skin, and respiratory niches; expanding coverage would improve its utility as a general benchmark. Second, baseline comparisons are limited to random forest and linear models. While additional methods (e.g., gradient boosting or alternative deep learning approaches) could provide a broader comparison, our goal is to characterise scaling behaviour in pretrained transformer models rather than exhaustively survey microbiome machine learning methods, and the chosen baselines are sufficient for this purpose. We were also unable to include two recent microbiome foundation models due to model unavailability (Pope et al., 2025; Medearis et al., 2026), limiting direct comparison to similar methods. More broadly, our approach relies on taxonomic labels for tokenisation, leading to incomplete token coverage across datasets and reflecting inconsistencies between sequencing pipelines and reference databases. This issue is amplified with a dataset that includes different DNA sequencing modalities, where outputs vary in terms of resolution and annotation. Addressing this will require more generalisable tokenisation strategies that better integrate heterogeneous sequencing outputs.

Taken together, our results establish pretraining LLMs as a practical and effective approach for microbiome modelling. Atlas provides a foundational dataset for pretraining models, Compass enables standardised evaluation, and the Waypoint models achieve state-of-the-art performance. As microbiome datasets continue to expand, we expect LLM-based approaches to play an increasingly important role.

Impact Statement

This work advances the application of LLMs to microbiome science by introducing a multimodal pretraining framework that integrates diverse DNA sequencing modalities at scale. By releasing the Atlas dataset, Compass benchmark, and Waypoint models, we aim to lower barriers to entry and enable broader participation in microbiome machine learning, particularly in settings where access to large, well-curated datasets is limited. These resources may accelerate research in areas such as human health, drug–microbiome interactions, and environmental biology.

At the same time, the increasing use of large-scale LLMs in the life sciences raises several considerations. First, models trained on aggregated public datasets may inherit biases in sampling, such as over representation of certain populations, environments, or experimental protocols, potentially limiting generalisability. Second, the use of heterogeneous sequencing modalities introduces challenges in data harmonisation and interpretation, which may affect downstream conclusions if not carefully managed.

Finally, the growing adoption of LLM-based approaches in biology may shift research practices toward more compute-intensive methods, raising questions around accessibility and environmental cost. We hope that by providing open datasets, benchmarks, and models, this work supports transparent, reproducible, and equitable progress in applying machine learning to the life sciences, while encouraging continued discussion of its broader impacts.

Model, Data and Code Availability

The Waypoint model weights, the Atlas dataset and the Compass benchmark are available on Hugging Face at https://huggingface.co/outpost-bio/, and example pre-training and finetuning/benchmarking code is available on GitHub at https://github.com/Outpost-Bio/waypoint. All artifacts are released under the Apache License, Version 2.0. The Hugging Face repositories are gated; access is granted automatically upon request via the Hugging Face interface.

References

Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., Bycroft, C., Nicolaisen, L., Arvaniti, E., Pan, J., Thomas, R., Dutordoir, V., Perino, M., De, S., Karollus, A., Gayoso, A., Sargeant, T., Mottram, A., Wong, L. H., Drotár, P., Kosiorek, A., Senior, A., Tanburn, R., Applebaum, T., Basu, S., Hassabis, D., and Kohli, P. Advancing regulatory variant effect prediction with alphagenome. *Nature*, 649 (8099):1206–1218, Jan 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-10014-0. URL https://doi.org/10.1038/s41586-025-10014-0.

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., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R., Ng, M. Y., Pannu, J., Ré, C., Schmok, J. C., John, J. S., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. L. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, Feb 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02523-z. URL https://doi.org/10.1038/s41592-024-02523-z.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

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, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

Lee, H., Lee, S., Lee, I., and Nam, H. Amp-bert: Prediction of antimicrobial peptide function based on a bert model. *Protein Science*, 32(1):e4529, 2023. doi: https://doi.org/10.1002/pro.4529. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4529.

Mastrorilli, E., Herd, P., Rey, F. E., Goodman, A. L., and Zimmermann, M. Linking interpersonal differences in gut

microbiota composition and drug biotransformation activity. *bioRxiv*, 2026. doi: 10.64898/2026.01.21.700809. URL https://www.biorxiv.org/content/early/2026/01/21/2026.01.21.700809.

Medearis, N. A., Zhu, S., and Zomorrodi, A. R. Biomegpt: A foundation model for the human gut microbiome. *bioRxiv*, 2026. doi: 10.64898/2026.01.05.697599. URL https://www.biorxiv.org/content/early/2026/01/05/2026.01.05.697599.

Munsamy, G., Ayres, G., Greco, C., Kam, K., Minto-Cowcher, G., St John, J., Bohnuud, T., Bakalar, M., Chow, W., Pecoraro, R., der Torossian Torres, M., Kollasch, A., Leung, M., Sirelkhatim, H., Farina, F., McGinnis, C., Sridhar, S., Anderson, D., Oteri, F., Taghibakhshi, A., Dona, J., Shimko, T., Stenbeeke, C., Papadopoulos, A., Krolick, M., Spoendlin, F., Gupta, P., Kumar, S., Bara, A., Wilbur, J., Ferruz, N., Rvachov, T., Wang, F., Cao, H., Lee, H.-S., Mehta, J., Chaleil, R., Pereno, V., Potti, S., Emerson, C., Dew, R. T., Yang, K. K., Nguyen, E., Tadimeti, N., Banfield, J. F., Frame, A., Bolton, E., Ruau, D., Kelleher, R., Costa, A., Powell, K., de la Fuente-Nunez, C., Gowers, G.-O., Vince, O., Finn, J., and Lorenz, P. Designing ai-programmable therapeutics with the eden family of foundation models. *bioRxiv*, 2026. doi: 10.64898/2026.01.12.699009. URL https://www.biorxiv.org/content/early/2026/01/12/2026.01.12.699009.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., and Ré, C. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution, 2023. URL https://arxiv.org/abs/2306.15794.

Pope, Q., Varma, R., Tataru, C., David, M. M., and Fern, X. Learning a deep language model for microbiomes: The power of large scale unlabeled microbiome data. *PLOS Computational Biology*, 21(5):1–24, 05 2025. doi: 10.1371/journal.pcbi.1011353. URL https://doi.org/10.1371/journal.pcbi.1011353.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L., Curtis, T., Escobar-Zepeda, A., Gurbich, T., Kale, V., Korobeynikov, A., Raj, S., Rogers, A., Sakharova, E., Sanchez, S., Wilkinson, D., and Finn, R. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51 (D1):D753–D759, 12 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1080. URL https://doi.org/10.1093/nar/gkac1080.

Roswall, J., Olsson, L. M., Kovatcheva-Datchary, P., Nilsson, S., Tremaroli, V., Simon, M.-C., Kiilerich, P., Akrami, R., Krämer, M., Uhlén, M., Gummesson, A., Kristiansen, K., Dahlgren, J., and Bäckhed, F. Developmental trajectory of the healthy human gut microbiota during the first 5 years of life. *Cell Host Microbe*, 29(5): 765–776.e3, May 2021.

Shi, H., Newton, D. P., Nguyen, T. H., Estrela, S., Sanchez, J., Tu, M., Ho, P.-Y., Zeng, Q., DeFelice, B., Sonnenburg, J., and Huang, K. C. Nutrient competition predicts gut microbiome restructuring under drug perturbations. August 2024.

Wiatrak, M., Viñas Torné, R., Ntemourtsidou, M., Dinan, A., Abelson, D. C., Arora, D., Brbić, M., Weimann, A., and Floto, R. A. A contextualised protein language model reveals the functional syntax of bacterial evolution. *bioRxiv*, 2025. doi: 10.1101/2025.07.20.665723. URL https://www.biorxiv.org/content/early/2025/07/20/2025.07.20.665723.

Zhang, H., Zhang, Y., Kang, Z., Xiong, J., Yang, R., and Ning, K. Mgm as a large-scale pretrained foundation model for microbiome analyses in diverse contexts. *Advanced Science*, n/a(n/a):e13333, 2026. doi: https://doi.org/10.1002/advs.202513333. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202513333.

A. Appendix

A.1. Pretraining Dataset

Data Acquisition Taxonomic abundance data were downloaded from the MGnify public database (Richardson et al., 2022) across all available pipeline versions (v1.0–v5.0) and four sequencing modalities: amplicon 16S rRNA, whole-genome shotgun metagenomic, metagenomic assembly, and metatranscriptomic. Data were obtained in the form of per-study abundance matrices where each column is a sample run accession and each row is a taxon, with values representing relative abundances. All datasets were versioned with DVC to ensure reproducibility of all download and processing steps.

Two quality filters were applied: (i) any taxon with relative abundance below 10^{-4} was set to zero, and (ii) any sample with fewer than 10 taxa remaining after the abundance filter was removed. After filtering, the dataset comprised 539,308 datapoints.

Prior to any tokenisation or model training, the filtered dataset was randomly partitioned into a pretraining set (90%) and a held-out benchmarking set (10%) using a fixed random seed ($s = 42$). The benchmarking partition is used exclusively to supply labelled samples for Tasks 1 and 2 of the benchmark, ensuring no overlap with pretraining data.

Taxonomic Tokenisation Each microbiome sample is represented as an ordered sequence of taxonomic strings. A custom `TaxonomicTokenizer` was built by extracting all genus-level names present in the pretraining data. To decide on whether to tokenise at the genus or species level we analysed the dataset following both tokenisation strategies. We found minimal difference in the distribution of sequence lengths and the number of total tokens between the two approaches (Appendix Table 5, Appendix Figure 6) and so we decided to move forward with genus level tokenisation with fallback. Where a taxon lacked a genus-level classification, the tokeniser falls back to the most specific available rank (family, order, and so on). The vocabulary comprises 14,389 taxonomic tokens plus five special tokens (`<bos>`, `<eos>`, `<pad>`, `<unk>`, `<mask>`), for a total vocabulary size of 14,394.

As in (Zhang et al., 2026) within each sample, taxa are ordered by the z-scored relative abundance computed across the pretraining corpus: for each taxon, t, we compute the z-score of its observed relative abundance using the mean, μ_t, and standard deviation, σ_t, then sort tokens from highest to lowest z-score. The means and standard deviations of each token in the pretraining dataset are serialised with the tokeniser so the same normalisation can be used for downstream tasks. Sequences are truncated to a maximum length of 510 then flanked by `<bos>` and `<eos>` tokens. The vocabulary is initialised during pretraining using all taxa in the pretraining dataset. During downstream benchmarking any unknown tokens are assigned the `<unk>` token and appended to the sequence after the z-score sorted known tokens.

A.2. Pretraining

We pretrain GPT-2-style causal language models with a next-token prediction objective. We explore model scales from 6M to 170M parameters by varying the number of layers and the embedding dimension. All models use HuggingFace `transformers` and were trained using the AdamW optimiser with a linear learning-rate warmup over the first 1,000 steps (learning rate 10^{-3}, weight decay 10^{-3}, batch size 32 per device). Training proceeded for up to 100 epochs with early stopping on validation loss (patience of 10 evaluations at $\sim$half-epoch intervals). A 10% random split of the pretraining set was held out as an in-training validation set for early stopping; this split is distinct from the held-out benchmarking partition. The GPU used and runtime of each pretraining run are shown in Table 1

A.3. Benchmark Construction

Datasets. The benchmark draws on four publicly available microbiome datasets. The MGnify dataset (Richardson et al., 2022) consists of metagenomically profiled samples spanning environments including human gut, skin, respiratory tract, oral cavity, marine, freshwater, soil, and engineered systems; biome labels were parsed from the hierarchical biome lineage metadata field (e.g., `root:Host-associated:Human:Digestive system:Fecal`). The Shi et al. drug–microbiome dataset (Shi et al., 2024) comprises 16S rRNA amplicon profiles of stabilised gut microbial communities exposed to a panel of drugs or control conditions, with drug identity and ATC classification provided as metadata. The Mastrorilli et al. dataset (Mastrorilli et al., 2026) contains paired community composition and drug degradation rate measurements for human gut microbial communities incubated with a panel of compounds. The Roswall et al. infant dataset (Roswall et al., 2021) contains longitudinal 16S rRNA profiles of infant gut communities with associated delivery mode and timepoint metadata.

Table 1. GPU and runtime for model pretraining

Model	GPU	Runtime
gpt2-6m-mgm	NVIDIA L40S-48GB	2d 5h 20m
gpt2-6m	NVIDIA L40S-48GB	2d 1h 53m
gpt2-10m	NVIDIA L40S-48GB	2d 12h 20m
gpt2-18m	NVIDIA L40S-48GB	3d 22h 13m
gpt2-29m	NVIDIA L40S-48GB	5d 9h 43m
gpt2-45m	NVIDIA A100-40GB	5d 1h 31m
gpt2-79m	NVIDIA A100-40GB	3d 20h 55m
gpt2-85m-gpt-small	NVIDIA A100-40GB	6d 4h 42m
gpt2-170m	NVIDIA H100-80GB	3d 17h 48m

Dataset	Tasks	Unique Token Coverage	Total Token Coverage
MGnify-biomes	1–2	100% (complete)	100% (complete)
Shi et al.	3–5	73.2% (139/190)	77.4% (53,884/69,659)
Mastrorilli et al.	6	95.9% (231/241)	93.5% (1,576,034/1,686,268)
Roswall et al.	7–8	60.5% (353/583)	59.8% (52,018/87,027)

Table 2. Coverage of unique (the proportion of tokens in the vocabulary that are in the pretraining data) and total tokens (the proportion of total tokens across the dataset that are in the pretraining data) for benchmark datasets.

Data processing. All datasets are processed to a common long format in which each row represents one datapoint, with two list fields storing taxonomic identifiers and corresponding relative abundances in descending order. Taxonomic strings are encoded with standard rank prefixes (e.g., `k__Bacteria; p__Firmicutes; g__Lactobacillus`). ASV-level abundance tables were collapsed to genus level by summing relative abundances within each unique genus-level lineage per sample. For the Mastrorilli et al. dataset, single-rank taxon identifiers were first mapped to full NCBI-derived taxonomic strings prior to genus-level collapse.

Train/validation/test splits. Data are partitioned into train (80%), validation (10%), and test (10%) sets using a fixed random seed (42). Because the Mastrorilli dataset is $n_{microbiomes} \times n_{drugs}$ each microbiome appears many times in the dataset. To prevent the same microbiome appearing in train and testing data all observations for a given microbiome are grouped before splitting. The Mastrorilli et al. dataset uses a 60/20/20 split. All splits are pre-computed and stored with the processed data to ensure reproducibility across model evaluations.

Task definitions. Eight tasks are defined over the four datasets. Tasks 1 and 2 are multi-output classification tasks over the MGnify dataset (Richardson et al., 2022); task 1 predicts all five biome hierarchy levels simultaneously, while task 2 restricts to samples labeled as *Digestive system* at hierarchy level 3 and predicts only the two finer-grained levels. For datapoints with only partial coverage of the biome levels, the levels that are missing are masked so that they do not contribute to the loss. Tasks 3–5 are single-output classification tasks over the Shi et al. dataset: task 3 predicts the sample from which a drug perturbed community originated from, task 4 predicts binary drug/control status, and task 5 predicts the first-level ATC drug class a microbiome was exposed to. Task 6 is a regression task predicting the rate of drug degradation from microbiome composition and a one-hot-encoded drug indicator. Tasks 7 and 8 are single-output classification tasks over the infant dataset, predicting the age of the baby the microbiome was taken from and the delivery mode respectively.

Evaluation metrics. We report macro F1 as the primary summary of performance for classification tasks. This metric computes the F1 score independently for each class and then averages them with equal weight, ensuring that the metric is not over sensitive to performance on majority labels. Appendix figures report additional complementary metrics; one-vs-one macro ROC AUC (Appendix Figure 7) captures pairwise discriminability by averaging AUC values across all class pairs, macro precision–recall (Appendix Figure 8) summarizes the balance between precision and recall across classes without being dominated by label frequency, balanced accuracy (Appendix Figure 9) reports the unweighted mean of per-class recall, reflecting how well each class is recovered on average. For the regression task in the benchmark, performance is evaluated using the coefficient of determination (R^2). Because R^2 can take negative values we clamp it to 0 to maintain consistency

with other metrics reported on a [0, 1] scale.

A.4. Downstream Fine-tuning on the Benchmark

For each benchmark task, a lightweight task-specific head was appended to the GPT-2 backbone. A sequence-level representation was obtained by taking the hidden state of the last non-padding token (i.e. the <eos> position). For Task 6 (drug degradation), which includes drug identity as a feature, a one-hot encoding of the drug compound was concatenated to this before the head. For classification tasks, one independent linear head per target was applied to produce class logits, with cross-entropy loss computed using class-frequency-based weighting to mitigate class imbalance. For the regression task (Task 6), a single linear head per target was used and trained with mean-squared-error loss.

All model weights (LLM backbone + head) were updated during fine-tuning. Models were trained with the AdamW optimiser (learning rate 3×10^{-5}, weight decay 10^{-3}, batch size 64, linear warmup over 1,000 steps) for up to 300 epochs with early stopping on validation loss (patience of 5 evaluations every 400 steps). For the classification tasks class weights were computed on the training split using the balanced weighting scheme $w_c = N/(K \cdot n_c)$, where N is the total number of training samples, K is the number of classes, and n_c is the number of training samples belonging to class c. These weights were passed to the cross-entropy loss,weighting classes to counteract class inbalance. The best checkpoint (lowest validation loss) was reloaded before test-set evaluation.

Benchmark samples were tokenised using the same `TaxonomicTokenizer` and z-scored relative abundance ordering described above (Section A.1), with token z-score statistics loaded from the file saved alongside each model checkpoint during pretraining. This ensures that the token ordering seen at fine-tuning time is consistent with that used during pretraining.

For the MGM model the weights, tokeniser and token z-score statistics constants were taken from the repository provided in the paper (Zhang et al., 2026)

Baselines Two non-neural baselines were included for comparison. For logistic regression (classification) and ridge regression (regression), a bag-of-taxa feature vector was constructed for each sample, where each dimension corresponds to a unique taxon observed in the training split and its value is the observed relative abundance (0 if absent). The regularisation strength C (or α) was selected by grid search over $\{0.01, 0.1, 1, 10\}$ on the validation set. For the random forest baseline, the same bag-of-taxa matrix was used with hyperparameter search over number of trees ($\{100, 200, 300, 500\}$) and maximum depth ($\{10, 20, \text{unlimited}\}$). Both baselines use class-balanced weighting for classification and, for tasks including drug identity, append a one-hot drug encoding to the feature vector.

A.5. Supplementary Figures and Tables

Table 3. Model configurations used in the scaling study.

Model	Layers	Hidden dim	Heads
6M-MGM	8	256	8
6M	8	256	4
10M	8	320	5
18M	10	384	6
29M	12	448	7
45M	14	512	8
79M	16	640	10
85M-gpt2-small	12	768	12
170M	24	768	12

Table 4. Compass task summary. Classification tasks (task 1-5 and 7-8) are scored by macro-averaged F_1; the regression task (task 6) by R^2.

TASK	NAME	DATASET	TYPE	METRIC
1	BIOME CLASSIFICATION	(RICHARDSON ET AL., 2022)	CLASSIFICATION	F_1-MACRO
2	GUT BIOME CLASSIFICATION	(RICHARDSON ET AL., 2022)	CLASSIFICATION	F_1-MACRO
3	SDC CLASSIFICATION	(SHI ET AL., 2024)	CLASSIFICATION	F_1-MACRO
4	DRUG VS. NON-DRUG	(SHI ET AL., 2024)	CLASSIFICATION	F_1-MACRO
5	DRUG CLASS (ATC)	(SHI ET AL., 2024)	CLASSIFICATION	F_1-MACRO
6	DRUG DEGRADATION RATE	(MASTRORILLI ET AL., 2026)	REGRESSION	R^2
7	INFANT AGE CLASSIFICATION	(ROSWALL ET AL., 2021)	CLASSIFICATION	F_1-MACRO
8	DELIVERY MODE CLASSIFICATION	(ROSWALL ET AL., 2021)	CLASSIFICATION	F_1-MACRO

Table 5. Tokenised corpus statistics at genus and species resolution. Sequence length is measured as tokens per sample.

Lowest Rank	Total tokens	Mean Length	Median Length	Min Length	Max Length
Genus	56,263,516	104.3	76	6	780
Species	58,293,439	108.1	79	11	806

Figure 5. Score per task for all models, classification tasks (1-5 and 7-8) are scored using macro averaged F1 score and the regression task (7) is scored using R^2. Pretrained models consistently outperform non-pretrained models, but for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens are unknown to the transformer models. The baseline methods can use all tokens; when marked as (no unk), the unknown tokens are removed to match the transformer models' vocabulary and enable direct comparison.

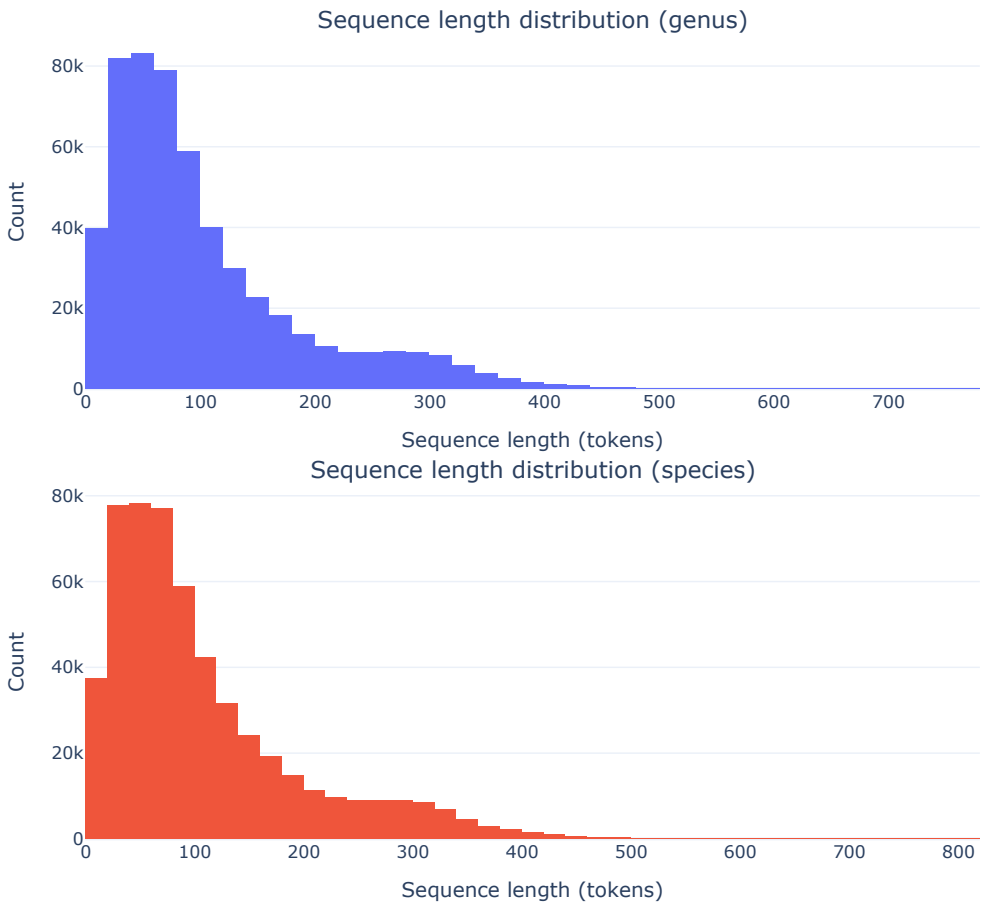

Figure 6. The distribution of sequence lengths when tokenising at the genus (top) and species (bottom) taxonomic levels.

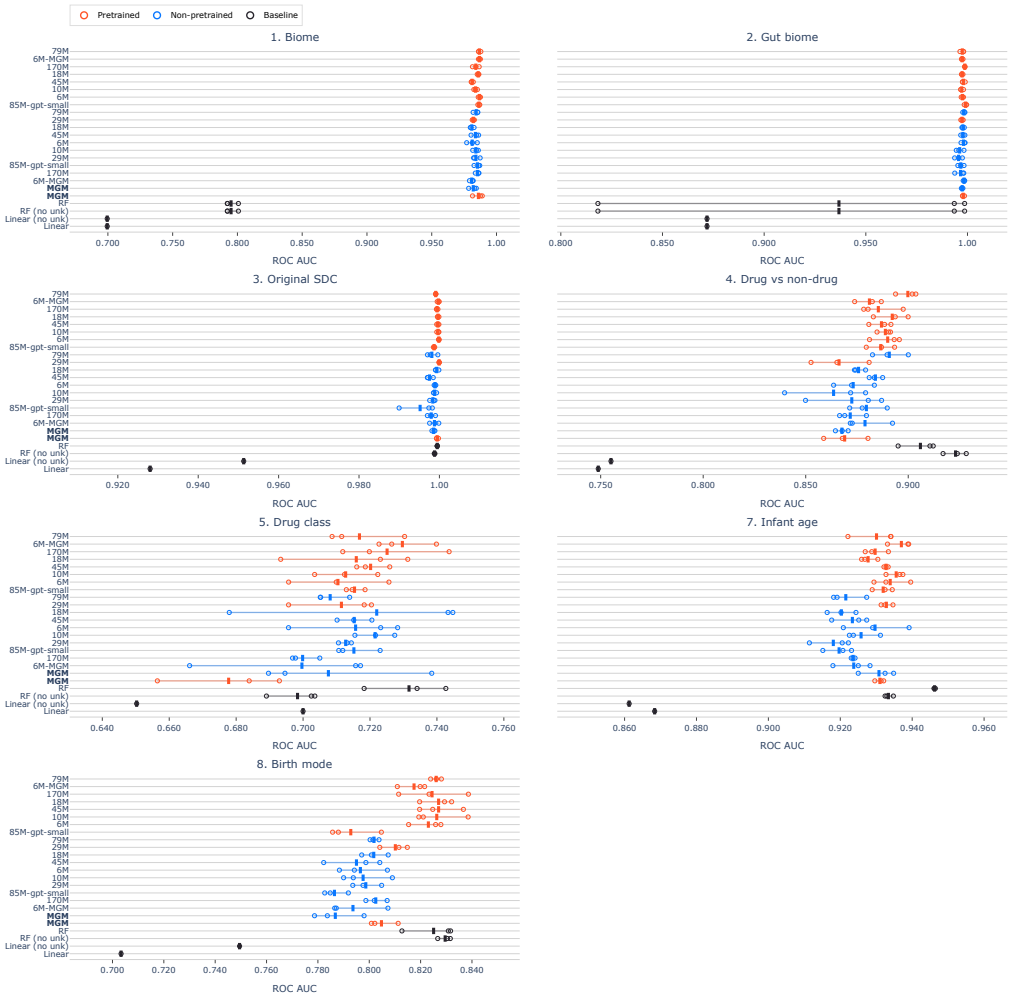

Figure 7. ROC AUC per classification task for all models. Pretrained models consistently outperform non-pretrained models, but we see that for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens are unknown to the transformer models. The baseline methods can use all tokens; when marked as (no unk), the unknown tokens are removed to match the transformer models' vocabulary and enable direct comparison.

Figure 8. PR AUC per classification task for all models. Pretrained models consistently outperform non-pretrained models, but we see that for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens are unknown to the transformer models. The baseline methods can use all tokens; when marked as (no unk), the unknown tokens are removed to match the transformer models' vocabulary and enable direct comparison.

Figure 9. Balanced accuracy per classification task for all models. Pretrained models consistently outperform non-pretrained models, but we see that for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens are unknown to the transformer models. The baseline methods can use all tokens; when marked as (no unk), the unknown tokens are removed to match the transformer models' vocabulary and enable direct comparison.