
Taxonomic Tokenisation and Self-Supervised Pretraining for Transferable Microbiome Representations

Neythen J Treloar, Saif Ur-Rehman & Jenny Yang

Outpost Bio

{neythen, saif, jenny}@outpost.bio

Abstract

Taxonomic abundance profiles provide a compact view of microbial communities, but heterogeneous and partially resolved taxonomies make learning transferable representations challenging. We investigate whether taxonomic tokenisation and self-supervised pretraining can address this problem. We assemble Atlas, a corpus of 539,308 microbiome profiles from diverse ecological contexts, and introduce a fallback tokenisation scheme that retains the most specific available taxonomic assignment when genus-level labels are unavailable. Using Atlas, we pretrain the Waypoint family of models, GPT-style causal language models trained to understand microbiomes. We evaluate their learned representations on Compass, a benchmark spanning eight environmental and gut-microbiome prediction tasks. We show that pretraining consistently improves downstream performance and enables favourable scaling with model capacity. Further, models pretrained without human-associated or gut samples still perform well on human gut tasks, demonstrating transfer learning across ecological domains. We find that incomplete vocabulary coverage across independently processed datasets can limit downstream performance, identifying taxonomic tokenisation as an important remaining limitation. Together, these results show that self-supervised pretraining can learn transferable microbiome representations while motivating more robust tokenisation methods for microbiome data.

1 Introduction

Microbial communities are ecosystems whose composition is linked to human health, environmental processes, and biotechnological function. High-throughput sequencing now yields microbiome profiles at substantial scale, creating an opportunity to learn reusable representations from large datasets. Yet this remains challenging: taxonomic abundance profiles are high-dimensional, sparse, compositional, and taxonomic assignments vary in resolution and are affected by differences in sequencing and bioinformatics pipelines. Self-supervised pretraining from large unlabelled corpora has produced transferable representations across genomic, protein, and whole-genome modelling tasks [1–9]. Recent microbiome foundation models have similarly treated taxonomic abundance profiles as sequences amenable to language-model pretraining [10–12]. However, the representation of a microbiome as a discrete sequence is not straightforward. Existing approaches differ in their taxonomic resolution, tokenisation strategy, pretraining corpus, and architecture [10–12], and the role of model scale remains comparatively underexplored in microbiome representation learning, despite its importance in language and protein modelling [13–15]. Here, we apply taxonomic tokenisation and self-supervised pretraining to learn transferable microbiome representations. We assemble Atlas, a corpus of 539,308 taxonomic abundance profiles from the MGnify resource [16], spanning diverse ecological contexts. We introduce a fallback tokenisation scheme that retains

the most specific available taxonomic assignment when genus-level labels are unavailable, rather than discarding partially resolved lineages. We pretrain the Waypoint family of causal transformer models and evaluate them with Compass, a benchmark of eight environmental and gut-microbiome prediction tasks. Our results show that pretraining consistently improves downstream performance and enables favourable scaling with model capacity. Token-coverage differences between independently processed datasets reveal a practical limitation of our taxonomic tokenisation scheme. Finally, models pretrained without gut or human-associated samples perform well on human-gut tasks, indicating transfer learning between ecological domains. Together, these findings support self-supervised pretraining as a route to learning general microbiome representations, while motivating more robust tokenisation methods.

2 Results

Pretraining improves downstream performance and enables model scaling. We assembled a large-scale pretraining corpus by systematically scraping taxonomic abundance data from the MGnify database [16] (see Appendix for details). We order microbes according to their z-score normalised relative abundances in descending order, then tokenise the resulting 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 available. The final Atlas dataset comprises 539,308 microbiomes, drawn from a broad range of environments including marine, freshwater, terrestrial, host-associated, and engineered ecosystems.

We pretrained the Waypoint family of causal transformer models ranging from 6M to 170M parameters on Atlas (see Appendix Table 3 for architectural parameters). We also included an alternate 6M model matching the Microbial General Model (MGM) architecture [10] (6M-MGM) enabling us to directly compare our dataset and tokenisation strategy to the existing state of the art microbiome foundation model. MGM was 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. We also included an 85M-gpt2-small architecture corresponding to the GPT-2 small model [17]. All models share the same tokeniser, context length (512), and pretraining procedure, so that differences in downstream performance can be attributed to model capacity. Pretraining evaluation loss curves for all nine models are shown in Figure 1A, larger models achieve consistently lower evaluation loss. Microbiome embeddings were generated from the 170M model by mean pooling over output token embeddings, T-SNE projections (Figure 1B) of these embeddings show clear structure with respect to sample source locations (for more t-SNE and PCA plots see Figure S1).

To systematically evaluate these pretrained models we introduce Compass, a curated benchmark of eight predictive tasks (see Appendix for details). 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 [16], drug-microbiome interactions [? 18], and infant gut microbiomes [19]. We compare the Waypoint models against non pretrained versions, classical methods and MGM. Pretraining consistently improves Compass scores for all model sizes (Figure 1C). Our 6M-MGM model also outperforms MGM, showing the value of our larger pretraining corpus and fallback taxonomic tokenisation scheme. The per-task breakdown (Appendix Figure S2) reveals that the benefit of pretraining is consistent across tasks. Figures providing additional complementary metrics for the classification tasks can be found in the appendix (Figures S4, S5, S6). Pretraining is also a prerequisite for favourable model scaling (Figure 1D), with pretrained models tending to perform better at higher parameter counts, in contrast to the opposite trend for non-pretrained models.

Labelled dataset size and token coverage shape performance XGBoost trained on the complete taxonomic abundance matrices performs competitively with our pretrained models (Figure 1C). Unlike the transformer models, this baseline has access both to taxa outside the pretraining vocabulary and to their exact relative abundances; the transformers instead receive rank-ordered sequences drawn from a vocabulary fixed during pretraining. Vocabulary coverage varies substantially between datasets (Table 2). Total token coverage is complete for the MGnify tasks, but falls to 77.4% for the Shi et al. tasks [?], 93.5% for the Mastorilli et al. task [18], and 59.8% for the Roswall et al. tasks [19]. To isolate the effect of vocabulary coverage, we additionally evaluate “no UNK”

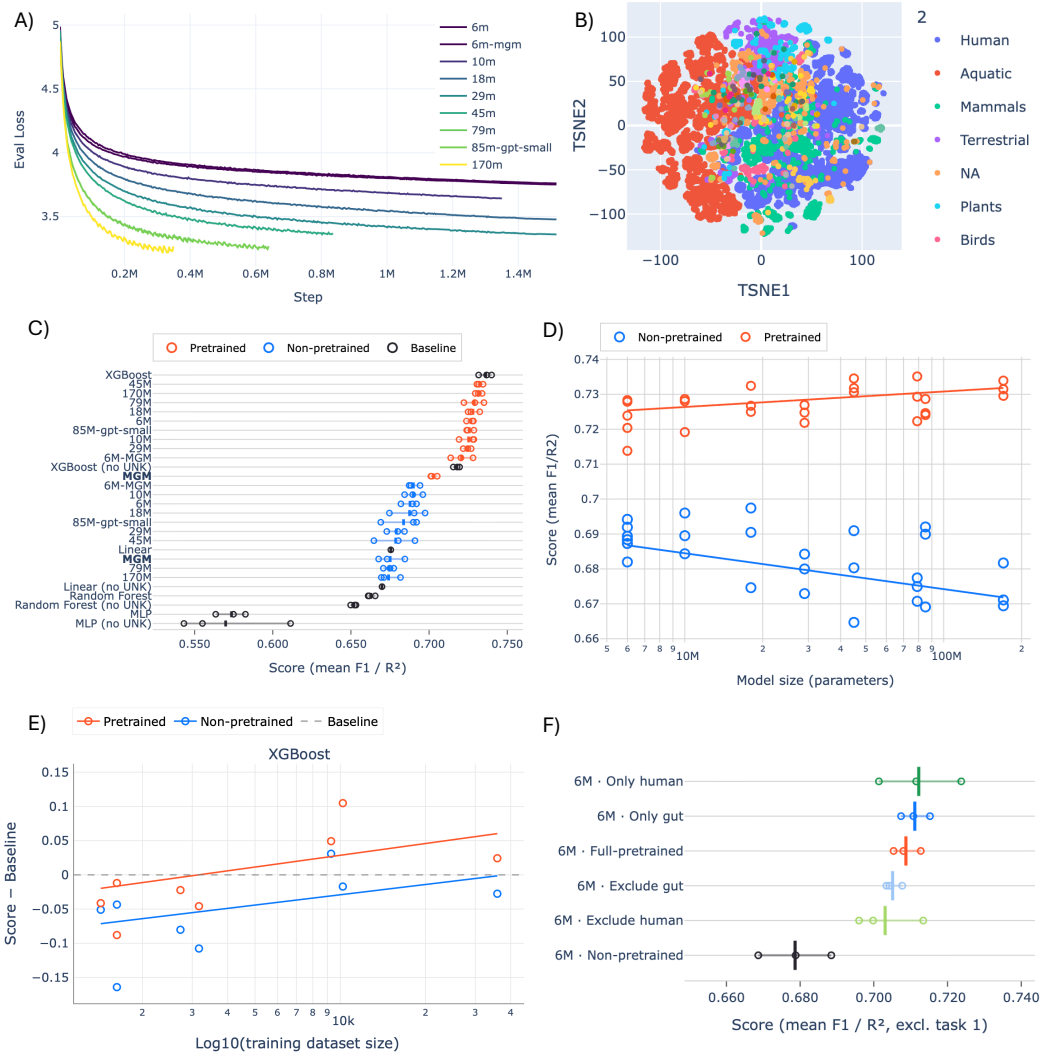


Figure 1: **A)** Pretraining evaluation loss curves for models with 6 - 170 million parameters. Larger models converge to lower evaluation loss. Some models exited training early because early stopping was used with a patience of 10. **B)** t-SNE plots of mean pooled model embeddings coloured according to sample location. **C)** Aggregate Compass performance for pretrained and randomly initialised Waypoint models, MGM, and classical abundance-based baselines. **D)** Aggregate Compass performance as a function of model parameter count. **E)** Difference in mean Compass task score between the 45M Waypoint model and XGBoost as a function of the number of labelled training examples. Positive values indicate that Waypoint outperforms XGBoost. **F)** Aggregate Compass performance of 6M Waypoint models pretrained on size-controlled subsets of Atlas. Horizontal lines indicate means and hollow circles show three fine-tuning replicates.

variants of each baseline after removing taxa that cannot be represented by the transformer tokeniser. These vocabulary-matched variants consistently underperform their full-information counterparts, demonstrating that the excluded taxa contain predictive information. The loss of performance associated with incomplete vocabulary coverage is therefore likely to limit the absolute performance attainable by the current tokenisation scheme and identifies it as an important target for future work. We note that even the no-UNK comparisons remain conservative because they retain exact abundance values, whereas Waypoint receives only abundance-derived rank order.

Comparing Waypoint-45M model with XGBoost (the full information variant) shows a positive relationship between the number of training examples and Waypoint’s relative score (Figure 1C–D). Below approximately 3,000 examples, the pretrained model underperforms XGBoost. But for the datasets in Compass with around 10,000 examples and above this is reversed. This crossover should not be interpreted as a universal threshold, but it roughly demonstrates the data regime in which the transformer becomes competitive with strong baseline models, even under a representational disadvantage.

Pretrained representations transfer across microbiome domains Compass is dominated by human-gut tasks, whereas Atlas spans host-associated, marine, freshwater, terrestrial, and engineered microbiomes. We therefore tested whether the benefit of pretraining depends on exposure to the target ecological domain. We pretrained 6M models using the full corpus, only human-associated samples, only gut samples, all samples excluding human-associated microbiomes, or all samples excluding gut microbiomes. Because these datasets vary in size we downsample each to 110k datapoints. Further, to evaluate the models on performance in the gut regime we remove task 1 from the benchmark for this analysis, as it is the only task that involves data from other domains. We find that every pretrained model outperforms the non-pretrained model, including models that never encounter gut or human-associated samples during pretraining (Figure 1F). Models pretrained without the target domain therefore retain most of the downstream benefit of pretraining, indicating that the learned representations capture ecological structure that transfers across microbiome domains. For non-sized matched scores, scores with task 1 included and per task scores see Figures S7, S8, S9

3 Discussion

Our results show that self-supervised pretraining on taxonomic abundance profiles improves downstream performance, enables model scaling and learns representations that transfer across ecological domains. Together, these findings suggest that pretraining captures statistical and ecological regularities shared across microbial communities rather than merely memorising domain-specific taxonomic associations. We demonstrate that our pretrained transformers are not uniformly preferable to classical models and show their relative advantage depends on the amount of task specific data. The consistent decline of baseline methods in the no-UNK setting indicates that gains can be made by generalising the representation of taxa across datasets. Fallback tokenisation addresses one source of information loss by retaining the most specific available taxonomic assignment when genus-level resolution is unavailable. It does not, however, solve vocabulary mismatch between datasets processed with different sequencing and bioinformatics pipelines. Future models could represent taxonomic lineages compositionally, using rank-specific or hierarchical tokens that permit unseen taxa to share information with known ancestors. Combining these representations with explicit continuous or discretised abundance information may also recover signal discarded by rank-only sequences.

Two main limitations qualify our conclusions. Compass contains only eight tasks from four datasets and is focused on human-gut microbiomes, so it does not yet establish transfer across the full range of environments represented in Atlas. Similarly, the observed crossover at approximately 10,000 labelled examples is descriptive of the tasks studied here, and is not guaranteed to hold in general.

Despite these limitations, the results establish a clear role for self-supervised pretraining in microbiome representation learning, it improves downstream performance and supports transfer between ecological domains. While tokenisation and abundance representation have been identified as central opportunities for the next generation of microbiome foundation models.

References

- [1] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 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>.
- [2] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution, 2023. URL <https://arxiv.org/abs/2306.15794>.
- [3] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. 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>.
- [4] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. 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>.
- [5] Geraldene Munsamy, Gavin Ayres, Carla Greco, Keith Kam, Gus Minto-Cowcher, John St John, Tanggis Bohnuud, Matthew Bakalar, William Chow, Robert Pecoraro, Marcello der Torossian Torres, Aaron Kollasch, Marcus Leung, Hassan Sirelkhatim, Francesco Farina, Connor McGinnis, Srijani Sridhar, Daniel Anderson, Francesco Oteri, Ali Taghibakhshi, Jeremie Dona, Tyler Shimko, Cedric Stenbeeke, Alexandros Papadopoulos, Malcolm Krolick, Fabian Spöndlin, Purba Gupta, Sandeep Kumar, Anne Bara, Jared Wilbur, Noelia Ferruz, Timur Rvachov, Fangping Wang, Hanqun Cao, Hyun-Su Lee, Japan Mehta, Raphael Chaleil, Valerio Pereno, Sid Potti, Chris Emerson, Roy Tal Dew, Kevin K Yang, Eric Nguyen, Neha Tadimet, Jillian F. Banfield, Alicia Frame, Emma Bolton, David Ruau, Rory Kelleher, Anthony Costa, Kimberley Powell, Cesar de la Fuente-Nunez, Glen-Oliver Gowers, Oliver Vince, Jonathan Finn, and Philipp Lorenz. 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>.
- [6] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. 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>.
- [7] Hansol Lee, Songyeon Lee, Ingo Lee, and Hojung Nam. 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>.
- [8] Maciej Wiatrak, Ramon Viñas Torné, Maria Ntemourtsidou, Adam Dinan, David C. Abelson, Divya Arora, Maria Brbić, Aaron Weimann, and R. Andres Floto. 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>.

- [9] Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, Raina Thomas, Vincent Dutordoir, Matteo Perino, Soham De, Alexander Karollus, Adam Gayoso, Toby Sargeant, Anne Mottram, Lai Hong Wong, Pavol Drotár, Adam Kosiorek, Andrew Senior, Richard Tanburn, Taylor Applebaum, Souradeep Basu, Demis Hassabis, and Pushmeet Kohli. 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>.
- [10] 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*, n/a(n/a):e13333, 2026. doi: <https://doi.org/10.1002/adv.202513333>. URL <https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adv.202513333>.
- [11] 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):1–24, 05 2025. doi: 10.1371/journal.pcbi.1011353. URL <https://doi.org/10.1371/journal.pcbi.1011353>.
- [12] Nicholas A. Medearis, Siyao Zhu, and Ali R. Zomorodi. 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>.
- [13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [14] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL <https://arxiv.org/abs/2203.15556>.
- [15] Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models, 2024. URL <https://arxiv.org/abs/2411.02142>.
- [16] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, Tom Curtis, Alejandra Escobar-Zepeda, Tatiana A Gurbich, Varsha Kale, Anton Korobeynikov, Shriya Raj, Alexander B Rogers, Ekaterina Sakharova, Santiago Sanchez, Darren J Wilkinson, and Robert D Finn. 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>.
- [17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.
- [18] Eleonora Mastrorilli, Pamela Herd, Federico E. Rey, Andrew L. Goodman, and Michael Zimmermann. 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>.
- [19] Josefine Roswall, Lisa M Olsson, Petia Kovatcheva-Datchary, Staffan Nilsson, Valentina Tremaroli, Marie-Christine Simon, Pia Kiilerich, Rozita Akrami, Manuela Krämer, Mathias Uhlén, Anders Gummesson, Karsten Kristiansen, Jovanna Dahlgren, and Fredrik Bäckhed. 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.

- [20] Handuo Shi, Daniel P. Newton, Taylor H. Nguyen, Sylvie Estrela, Juan Sanchez, Michael Tu, Po-Yi Ho, Qinglin Zeng, Brian C. DeFelice, Justin L. Sonnenburg, and Kerwyn Casey Huang. Nutrient competition predicts gut microbiome restructuring under drug perturbations. *Cell*, 188(24):6971–6986.e14, Nov 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.10.038. URL <https://doi.org/10.1016/j.cell.2025.10.038>.

A Appendix

A.1 Pretraining Dataset

Data Acquisition Taxonomic abundance data were downloaded from the MGnify public database [16] 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 S3) 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 [10] 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

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: Pretraining data coverage for each benchmark dataset, measured in two ways: unique token coverage, the percentage of distinct tokens in the dataset that appear in the pretraining data, and total token coverage, the percentage of all tokens in the dataset that appear in the pretraining data.

A.3 Benchmark Construction

Datasets. The benchmark draws on four publicly available microbiome datasets. The MGnify dataset [16] 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 [?] 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 [18] 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 [19] contains longitudinal 16S rRNA profiles of infant gut communities with associated delivery mode and timepoint metadata.

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 $n_{\text{microbiomes}} \times n_{\text{drugs}}$ 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 [16]; 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 S4) captures pairwise discriminability by averaging AUC values across all class pairs, macro precision–recall (Appendix Figure S5) summarizes the balance between precision and recall across classes without being dominated by label frequency, balanced accuracy (Appendix Figure S6) 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.

Benchmark score aggregating. To prevent datasets represented by multiple benchmark tasks from disproportionately influencing aggregate performance, we used dataset-balanced averaging. For each model and replicate, task scores were first averaged within each source dataset and these dataset-level means were then averaged across datasets. Specifically, for dataset (d), containing task set (T_d), we calculated

$$s_{m,d} = \frac{1}{|T_d|} \sum_{t \in T_d} s_{m,t}$$

and defined the dataset-balanced benchmark score as

$$S_m = \frac{1}{D} \sum_{d=1}^D s_{m,d}$$

where ($s_{m,t}$) is the score of model (m) on task (t) and ($s_{m,d}$) is the score of model (m) on dataset (d), and ($D=4$) is the number of source datasets. The four dataset groups comprised tasks 1–2, tasks 3–5, task 6, and tasks 7–8. This gives each source dataset equal weight, irrespective of how many prediction tasks it contributes.

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 imbalance. 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 [10]

Baselines Four task-specific baselines were included: linear models, random forests, XGBoost, and a multilayer perceptron (MLP). All baselines used a common bag-of-taxa feature representation in which each dimension corresponds to a unique taxon observed in the training split and contains its relative abundance, or zero when absent. A pseudocount of 10^{-6} was added before applying a centred log-ratio transformation to each sample. Each taxon dimension was then standardised using statistics calculated from the training split. For tasks including drug identity, a one-hot encoding of the compound was appended after transformation and was not standardised. Classification models used class-balanced weighting. Hyperparameters were selected on the validation split using cross-entropy for classification and mean squared error for regression.

For classification, the linear baseline used logistic regression; for regression, it used ridge regression. The regularisation strength C or α was selected from $\{0.01, 0.1, 1, 10\}$, with a maximum of 1,000 solver iterations. For random forests, the number of trees was selected from $\{100, 200, 300, 500\}$ and maximum depth from $\{10, 20, \text{unlimited}\}$, with $\sqrt{d}$ features considered at each split. For XGBoost, maximum depth was selected from $\{4, 6, 8\}$. The number of boosting rounds and learning rate were fixed at 200 and 0.1, respectively; row subsampling was 0.8, column subsampling per tree was 0.2, minimum child weight was 1.0, and L_2 regularisation was 1.0. The histogram tree-construction method was used. The MLP contained hidden layers of 256 and 128 units, each followed by a ReLU activation and dropout. Dropout was selected from $\{0.1, 0.3\}$ and learning rate from $\{10^{-3}, 3 \times 10^{-4}\}$. The MLP was trained with Adam, weight decay 10^{-5} , and batch size 256 for at most 200 epochs. Early stopping used a held-out 10% subset of the training data with a patience of five epochs; this subset was kept separate from the validation set used for hyperparameter selection.

To quantify the effect of vocabulary mismatch, we additionally evaluated “no UNK” versions of the baselines. Taxa that could not be represented by the pretrained tokeniser were removed before feature construction, and the abundances of the remaining taxa were renormalised to sum to one. These baselines therefore receive the same set of taxonomic identities as the transformer models, although they retain access to exact relative abundances rather than abundance-derived rank order.

A.5 Cross-biome pretraining

To determine whether the benefits of pretraining depend on exposure to the target ecological domain, we pretrained the 6M Waypoint model on five subsets of Atlas defined using MGnify biome-lineage metadata. The human domain comprised samples whose lineage fell under `Host-associated: Human`, while the gut domain comprised samples whose third biome level was `Digestive system`, including both human and non-human hosts. The five conditions were the complete Atlas corpus (*Full-pretrained*), only human-associated samples (*Only human*), only gut samples (*Only gut*), all samples excluding human-associated microbiomes (*Exclude human*), and all samples excluding gut microbiomes (*Exclude gut*). Samples with insufficient metadata to determine subset membership were removed to prevent contamination between conditions.

All subset models used the same architecture, tokeniser vocabulary, abundance-ordering procedure, context length, optimisation settings, and early-stopping criteria as the standard 6M model. Holding the vocabulary fixed to that constructed from the complete Atlas corpus ensured that downstream token coverage was identical across conditions. For each subset, 90% of the available profiles were used for pretraining and 10% for validation. Because the unrestricted subsets differed in size, we performed an additional size-controlled experiment in which each condition was randomly subsampled to 110,000 profiles before applying the same training-validation split. Each pretrained model was fine-tuned on all eight Compass tasks using three independent replicates and compared with an identically configured model trained from random initialisation.

A.6 Supplementary Figures and Tables

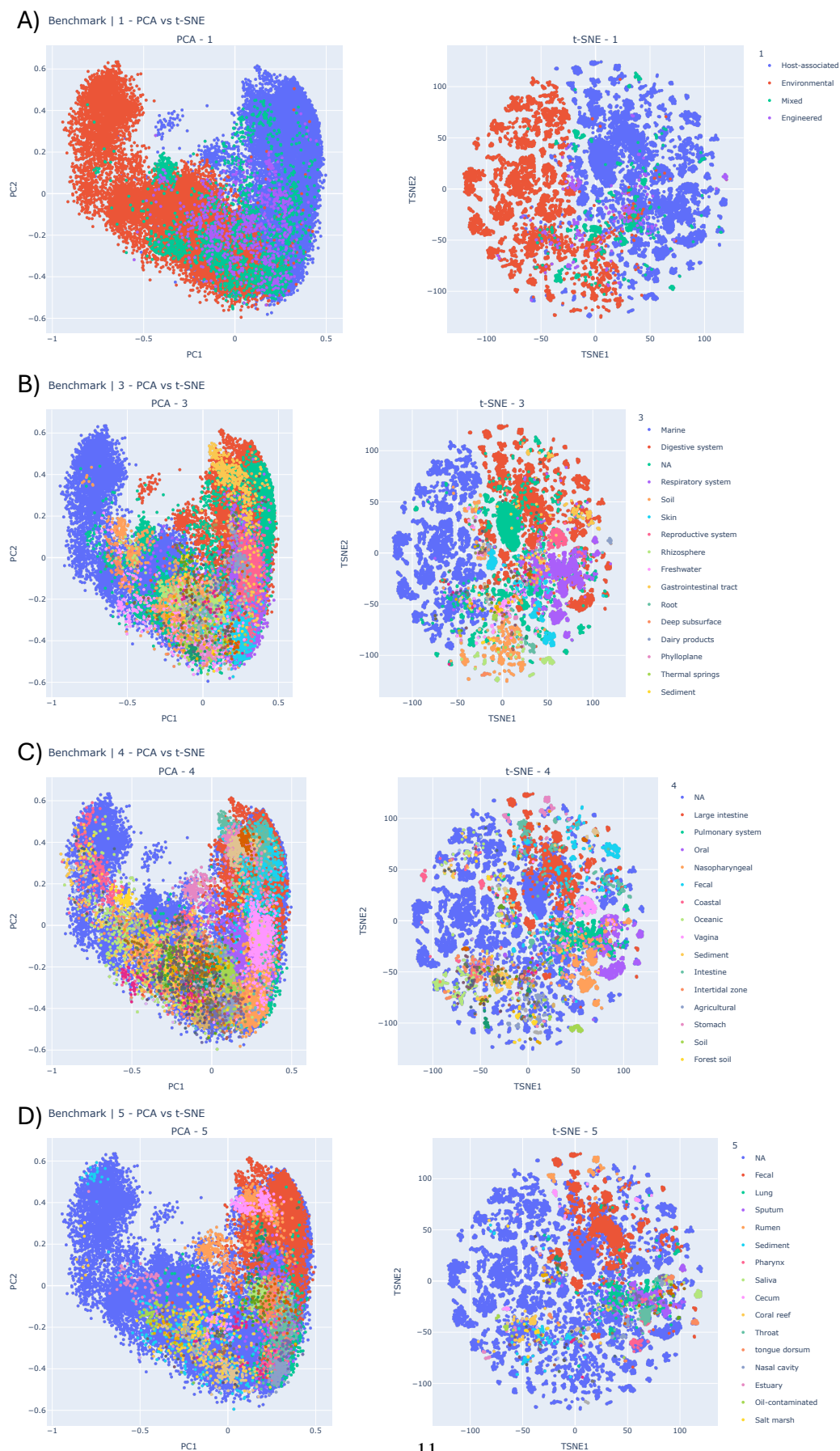


Figure S1: A-D) PCA and t-SNE plots of mean pooled model embeddings coloured according to the 1st, 3rd, 4th, 5th levels of biome hierarchy.

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 (1-5, 7-8) are scored by macro-averaged F_1 ; the regression task (6) by R^2 . Tasks 1-5 and 7-8 use an 80/10/10 train/validation/test split and task 6 a 60/20/20 split grouped by microbiome, both with seed 42. Tasks 2, 3 and 5 are label-filtered subsets of a split assigned over their parent dataset, so their proportions deviate slightly from 80/10/10.

Task	Name	Dataset	Type	N samples			
				Total	Train	Val	Test
1	Biome	[16]	Class.	41,399	33,121	4,139	4,139
2	Gut biome	[16]	Class.	12,538	10,016	1,268	1,254
3	SDC	[20]	Class.	3,394	2,725	340	329
4	Drug vs. non-drug	[20]	Class.	3,960	3,168	396	396
5	Drug class (ATC)	[20]	Class.	1,778	1,425	182	171
6	Drug degradation	[18]	Reg.	15,419	9,282	3,084	3,053
7	Infant age	[19]	Class.	2,031	1,625	203	203
8	Delivery mode	[19]	Class.	2,031	1,625	203	203

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 S2: 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.

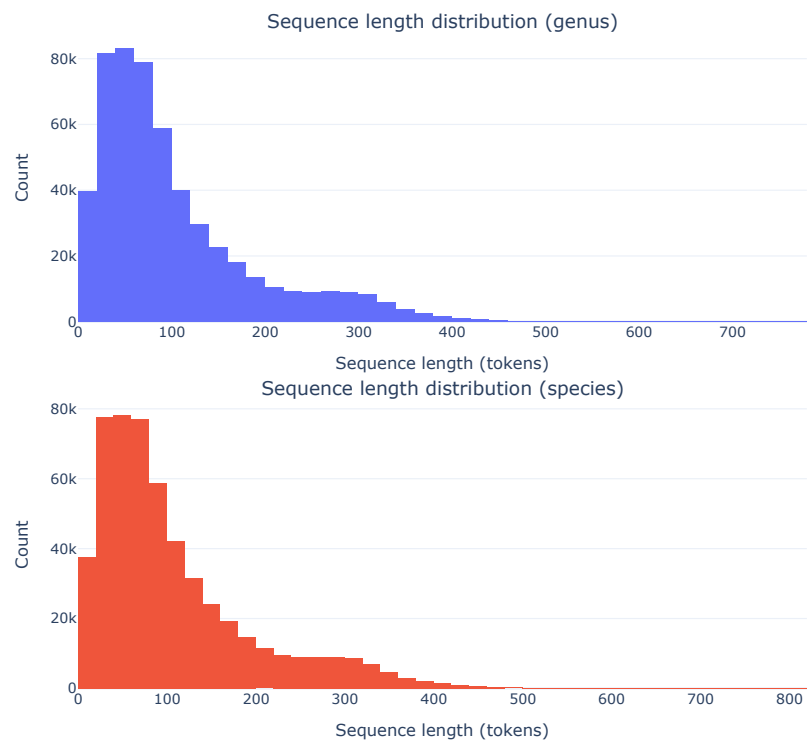


Figure S3: The distribution of sequence lengths when tokenising at the genus (top) and species (bottom) taxonomic levels.

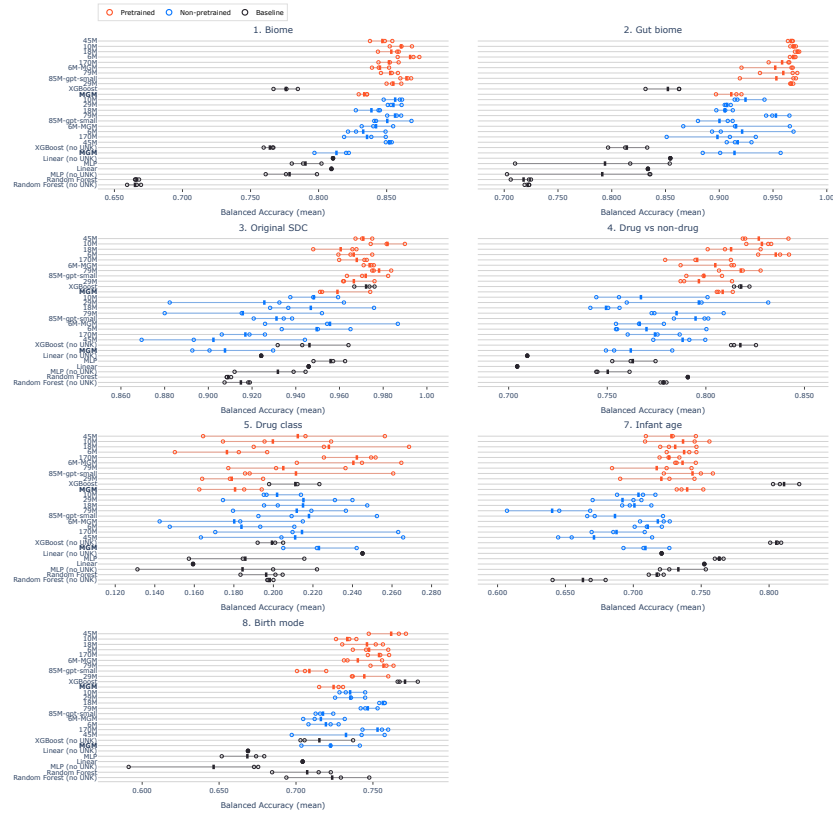


Figure S6: Balanced accuracy per classification task for all models. Pretrained models consistently outperform non-pretrained models, but we see that for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, 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.

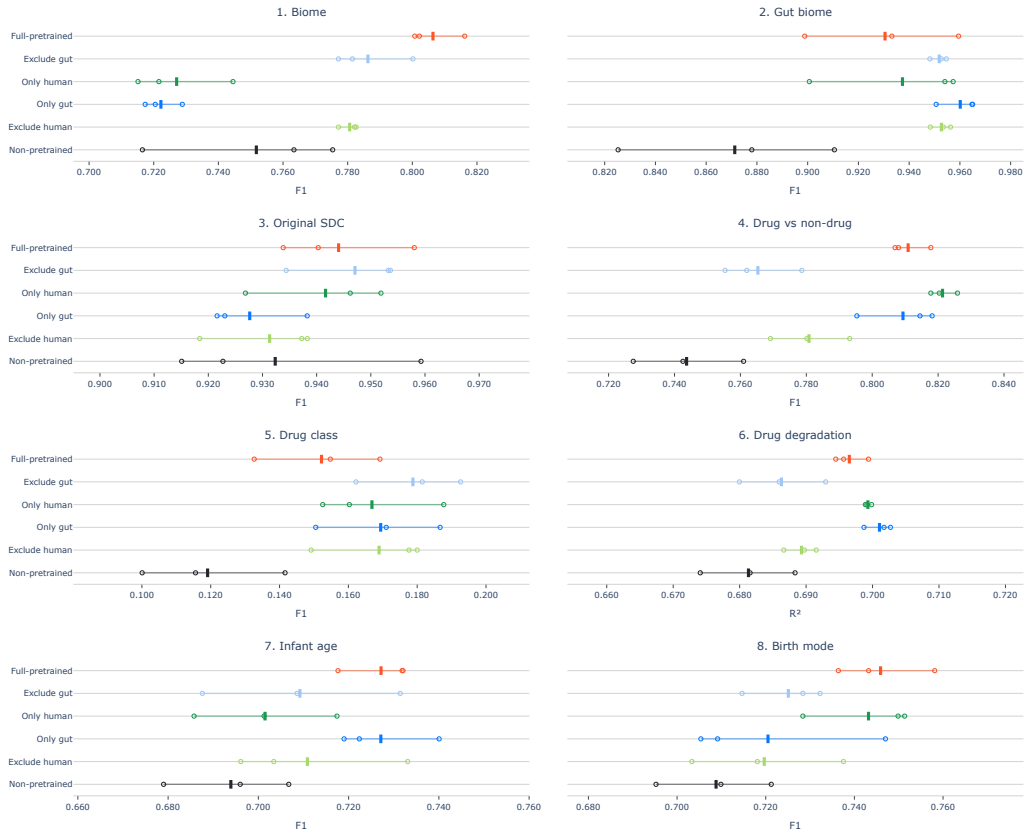


Figure S8: Per-task Compass scores for the 6M model pretrained on different biome subsets of Atlas. Classification tasks (1–5, 7–8) are scored by macro-averaged F_1 and the regression task (6) by R^2 . Solid lines indicate the mean and hollow circles individual replicates. The cross-biome transfer observed in the aggregate holds across individual tasks.

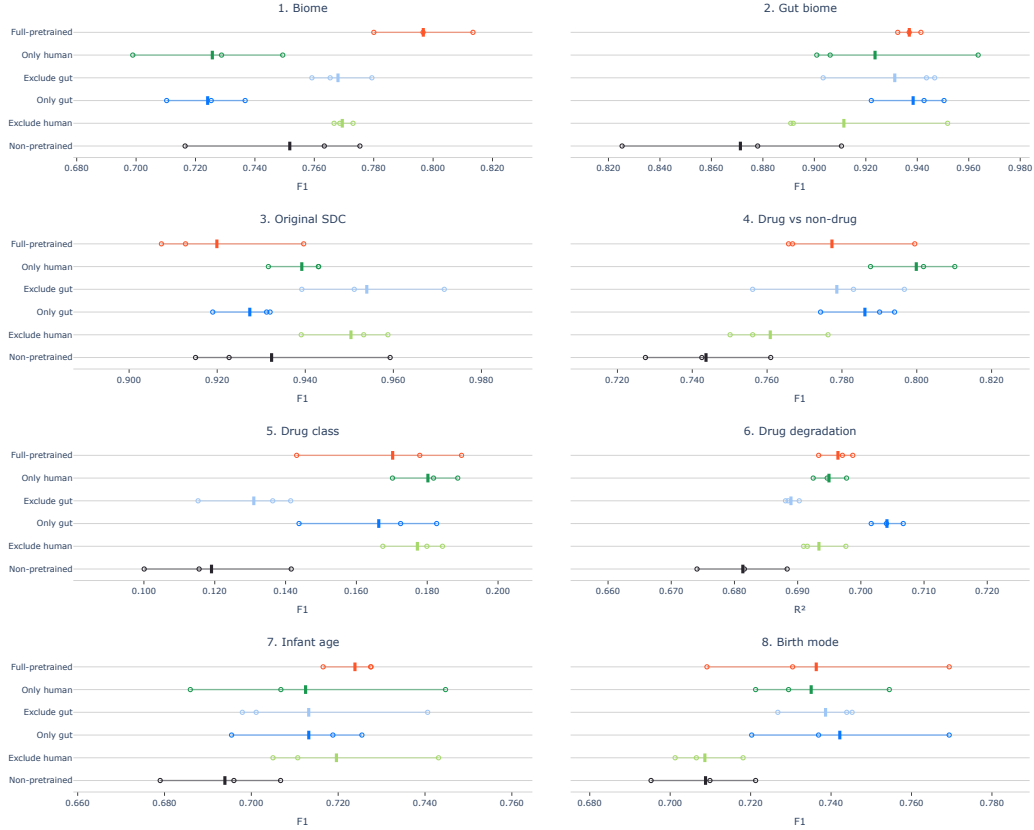


Figure S9: Per-task Compass scores for the 6M model pretrained on different biome subsets of Atlas where all datasets are subsampled down to 110k data points. Classification tasks (1–5, 7–8) are scored by macro-averaged F_1 and the regression task (6) by R^2 . Solid lines indicate the mean and hollow circles individual replicates. The cross-biome transfer observed in the aggregate holds across individual tasks.