LEARNING THE LANGUAGE OF THE MICROBIOME WITH TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised pretraining has become central to biological machine learning, yet microbiome data remains comparatively underexplored in terms of both modeling approaches and evaluation frameworks. To address this gap, we present Atlas, a pretraining dataset of 539,308 microbiome datapoints from the MGnify database. Using Atlas, we train the Waypoint family of microbiome foundation models: a series of GPT-2 style causal language models ranging from 6M to 170M parameters. We also introduce Compass, a curated benchmark of eight predictive tasks spanning biome classification, drug-microbiome interactions, drug degradation, and infant gut development. Using this benchmark, we compare the performance of Waypoint models against classical baselines and the existing Microbial General Model (MGM) foundation model. Our results show that pretraining consistently improves downstream task performance, that both dataset scale and tokenization strategy affect model quality, that pretraining supports favorable scaling behavior and that representations learned during pretraining transfer between microbiome domains. Furthermore, pretrained transformer models outperform classical methods on Compass tasks with roughly 10,000 or more training examples, a number that is attainable for modern microbiome studies. Finally, we demonstrate that the Waypoint models achieve state-of-the-art performance among publicly available microbiome foundation models. Overall, our work highlights the importance of large-scale self-supervised pretraining and tokenisation choices in this domain and establishes Atlas, Compass, and the Waypoint models as valuable resources for the research community in this emerging field.

1 INTRODUCTION

Microbial communities have emerged as a powerful variable in human health, disease, and environmental biology. Advances in high-throughput sequencing have made it possible to profile microbial communities at scale, generating large repositories of compositional abundance data across diverse ecological contexts. Yet translating this wealth of data into actionable biological insight remains a significant challenge, owing to the high dimensionality, sparsity, and compositional nature of microbiome profiles, as well as the heterogeneity in sequencing and taxonomic assignment pipelines.

Much recent progress in machine learning has been driven by the use of self-supervised pretraining to learn meaningful representations, especially in language modelling. These techniques have also been applied to biological data, giving rise to DNA language models Ji et al. (2021); Nguyen et al. (2023); Dalla-Torre et al. (2025); Brixi et al. (2025); Munsamy et al. (2026), protein language models Hayes et al. (2025); Lee et al. (2023), and whole-genome models Wiatrak et al. (2025); Avsec et al. (2026). This naturally raises the question of whether the architectures and training paradigms that succeed in natural language and molecular biology transfer effectively to microbiomes and this is an emerging frontier Zhang et al. (2026); Pope et al. (2025); Medearis et al. (2026). The foundation model paradigm is particularly attractive for microbiome research, where labelled datasets are typically small and task-specific, but large collections of unlabelled compositional profiles are publicly available. Several recent efforts have begun to treat taxonomic abundance profiles as sequences amenable to language model pretraining Zhang et al. (2026); Pope et al. (2025); Medearis et al. (2026). However, these approaches differ in their tokenisation strategies, pretraining corpora, and architectural choices, and systematic evaluation of how these design decisions affect downstream performance has been limited. In particular, the effect of model scale on microbiome representation

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

learning, a question extensively studied in language Kaplan et al. (2020); Hoffmann et al. (2022) and protein Cheng et al. (2024) modelling, remains poorly understood. Equally, the field lacks standardised benchmarks for comparing models across the range of biologically meaningful prediction tasks that microbiome data can support.

In this paper, we take first steps toward addressing these gaps. We assemble and release a large-scale pretraining corpus, named Atlas, of 539,308 microbiome datapoints from the MGnify database Richardson et al. (2022). We introduce a fallback tokenisation strategy that preserves taxonomic information when genus-level assignments are unavailable and show that it, together with the larger pretraining corpus, improves performance relative to MGM. We also present Compass, a curated benchmark of eight predictive tasks spanning biome classification, drug-microbiome interactions, drug degradation, and infant gut development, designed to provide a comprehensive and reproducible evaluation framework for microbiome foundation models. Finally, we pretrain the Waypoint models, a series of GPT-2 style causal language models ranging from 6M to 170M parameters, and evaluate each on Compass, alongside classical baselines and the existing MGM foundation model Zhang et al. (2026). Our results show that pretraining consistently improves downstream performance, supports favourable model scaling and produces representations that transfer across microbiome domains. We also show that task-specific dataset size is associated with the relative performance of Waypoint and classical baselines and that Waypoint models achieve state-of-the-art performance among microbiome foundation models, with dataset scale and tokenisation strategy emerging as important drivers of model quality.

2 Data, Models, and Evaluation

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 three sequencing modalities: amplicon 16S rRNA, whole-genome shotgun metagenomic and metagenomic assembly (Appendix A.1). 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 a procedure used previously in related work 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. Following quality filtering, the dataset comprises 539,308 microbiome datapoints, partitioned 90/10 into 485,377 pretraining and 53,931 held-out benchmarking samples, 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.

Previous approaches have tokenised microbiomes at the genus level Zhang et al. (2026); Pope et al. (2025), but other work has argued for the advantages of tokenising at the species level Medearis et al. (2026). Tokenising at the species level can improve taxonomic precision; however, for many Amplicon sequence variants (ASVs) lower-level taxonomic assignments cannot be made. This is particularly relevant for amplicon sequencing, where the 16S rRNA marker often lacks the resolution needed to assign an ASV at the species level. To address this limitation, we introduce a fallback tokenisation approach in which a target level of taxonomic specificity is defined (Appendix A.1). Classifications more specific than the target rank are collapsed to that rank, while taxa that cannot be assigned at the target rank are represented using the most specific higher-rank classification available. To determine whether to tokenise at the genus or species level, we analysed the dataset using both tokenisation strategies. We found minimal difference in the distribution of sequence lengths and the number of total tokens between the two approaches (Table S1, Figure S1) and so we decided to move forward with genus level tokenisation with fallback.

2.2 The Waypoint series of models: scaling microbiome foundation models

To investigate the effect of model scale on microbiome representation learning, we pretrained a series of GPT-2–style causal language models ranging from 6M to 170M parameters (not including token and positional embeddings) See Table S2 for architectural parameters and Appendix A.2 for further details. Models were scaled by jointly increasing the hidden dimension and the number of

transformer layers, while holding the per-head dimension fixed at 64 throughout (i.e. the number of attention heads grows in proportion to the hidden dimension). The resulting series spans roughly one order of magnitude in parameter count (6M-170M). We also included an alternate 6M model matching the MGM Zhang et al. (2026) architecture (6M-MGM) enabling us to understand the performance gain from using a larger pretraining dataset and fallback tokenisation, and an 85M-gpt2-small architecture corresponding to the GPT-2 small model Radford et al. (2019). All models share the same tokeniser, maximum context length (512), and pretraining procedure, so that differences in downstream performance can be attributed to model capacity alone.

2.3 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 (Appendix A.3). 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 (Table S3). The MGnify dataset Richardson et al. (2022) contributes two biome classification tasks: task 1 requires predicting the environmental origin of a microbiome sample at up to five levels of a biological ontology hierarchy (e.g., Environmental $\rightarrow$ Aquatic $\rightarrow$ Marine), while task 2 restricts this to classification of data points from gut microbiomes. The Shi et al. dataset, derived from a 16S rRNA amplicon study of drug–microbiome interactions in stool derived communities Shi et al. (2025), contributes three tasks: task 3 requires classifying the sample from which a drug perturbed community originated from; task 4 is a binary classification of whether a sample received a drug or was a control; and task 5 requires predicting the ATC drug class that was applied to a community from the resulting composition. The Mastrorilli et al. dataset Mastrorilli et al. (2026) contributes task 6, a regression task predicting drug degradation rate from community composition and drug identity. Finally, the Roswall et al. infant microbiome dataset Roswall et al. (2021) contributes task 7, predicting the age of an infant from a gut microbiome sample, and task 8, predicting delivery mode (vaginal vs. cesarean section). Tasks 1–5 and 7–8 are classification tasks scored by macro-averaged F1 across classes, averaged across target columns where multiple targets are present (tasks 1 and 2). This scoring scheme weights classes equally regardless of prevalence, reflecting the benchmark's aim to assess generalisation across the full label space rather than rewarding majority-class prediction. Task 6 is a regression task scored by the coefficient of determination (R^2), clipped to be greater or equal to zero to prevent large negative values from skewing the overall score. 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 the mean of these four averages was taken as the final dataset score (Appendix A.3).

3 EXPERIMENTS AND RESULTS

3.1 LARGER MODELS ACHIEVE LOWER PRETRAINING LOSS

Pretraining evaluation loss curves for all nine models are shown in Figure 1A. All models converge from a shared initial loss of $\sim$5.0, with larger models achieving consistently lower evaluation loss throughout training. The two 6M-parameter models (6M and 6M-MGM) produce almost indistinguishable loss curves, demonstrating that the difference in attention head count between the two architectures has a minimal effect on pretraining performance. Scaling up model capacity yields consistent improvements, with eval loss decreasing monotonically with model size (Figure 1B). 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 early 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. Microbiome embeddings were generated from the 170M model by mean pooling over output token embeddings. In both PCA (Figure 1C) and t-SNE (Figure 1D) projections of these embeddings we see clear structure with respect to the second level of the biome hierarchy (for all other levels see Figure S2).

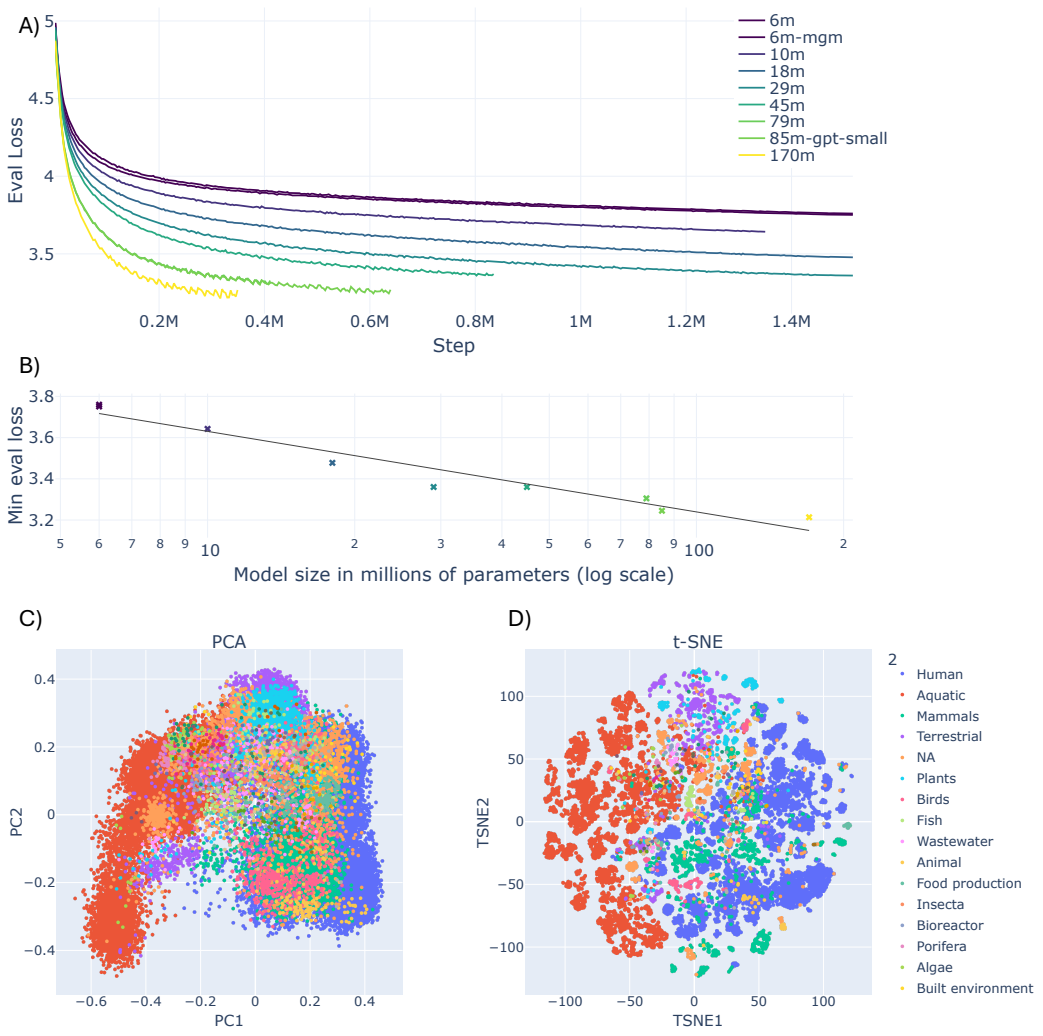

Figure 1: A) Pretraining evaluation loss curves for models with 6 - 170 million parameters. Larger models converge to lower evaluation loss. Some models exited training early because early stopping was used with a patience of 10. B) The minimum evaluation loss decreases linearly with logarithmic model size. C, D) PCA and t-SNE plots of mean pooled model embeddings coloured according to the second level of biome hierarchy.

3.2 Pretraining improves downstream performance

We evaluated our pretrained Waypoint models against classical baselines and the current state-of-the-art in microbiome foundation models (Appendix A.4 for details). As classical baselines we included simple multi layer perceptron (MLP), linear models, random forest models and XGBoost operating on bag-of-taxa abundance matrices. Class-weighted losses were used for classification tasks for all models, and simple hyperparameter optimisation was performed for each baseline. 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 between 100% and 59.8% across datasets (Table S4). This variation in vocabulary overlap directly motivated evaluating the classical baselines both with and without filtering of taxa absent from the pretraining vocabulary (the filtered versions denoted "no UNK"), where the latter removes all out of vocabulary taxa to enable a fairer comparison to the Waypoint models operating under a fixed token set. Note that even in the no UNK setting the baseline models still have an information advantage as they are given the precise relative abundances of each taxon, whereas the Waypoint models just see rank

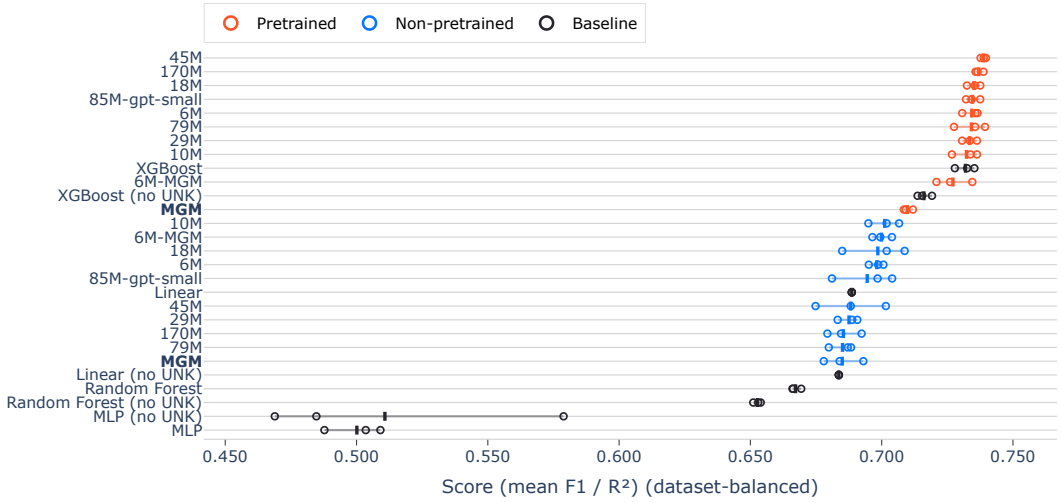

Figure 2: Overall Compass scores for all models, taken as the dataset balanced average of the scores across the 8 benchmark tasks, where classification tasks (1-5 and 7-8) are scored using macro averaged F1 and the regression task (6) is scored using R^2. Solid lines indicate mean score and the hollow circles individual replicates. Pretrained models consistently outperform their non-pretrained counterparts; most also outperform the strongest classical baselines. 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 more direct comparison.

ordered lists of taxa. Second, we compared against MGM, an existing microbiome foundation model pretrained on 260K samples from MGnify using a causal language modelling objective Zhang et al. (2026). Our 6M-MGM model uses the same architectural parameters as MGM and thus provides a direct assessment of the effect of our larger pretraining corpus and fallback tokenisation scheme with the prior state of the art. We also attempted to include comparisons with two additional microbiome foundation models Pope et al. (2025); Medearis et al. (2026), but to our knowledge these models have not been made publicly available. For each transformer architecture, we additionally evaluated non-pretrained variants by reinitialising weights prior to benchmarking.

Figure 2 shows the overall Compass score for each model and baseline, with replicates shown as individual points. In terms of the overall benchmark score the MLP baseline is not competitive, the Linear and Random Forest models are competitive with our non-pretrained transformers and the XGBoost models are competitive with our pretrained transformers. We see that all of our pretrained Waypoint models outperform XGBoost apart from 6M-MGM. Further, besides MLP, which performs poorly across the board, for every baseline model the no UNK variant underperforms the variant that can see all the tokens, this highlights that having full knowledge of the taxa is important for good performance and implies that the token mismatch between pretraining and benchmarking is likely having a deleterious effect on the performance of the Waypoint models and MGM. We also see that pretraining consistently and substantially improves performance over non-pretrained counterparts, and all pretrained Waypoint models outperform the MGM model, which itself outperforms all non-pretrained transformers. Additionally, 6M-MGM 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 effect of the improved tokenisation scheme is further supported by comparing the non-pretrained MGM and 6M-MGM models, where the only difference is the tokenisation scheme; the superior performance of 6M-MGM in this setting highlights the importance of tokenisation as a key modelling decision.

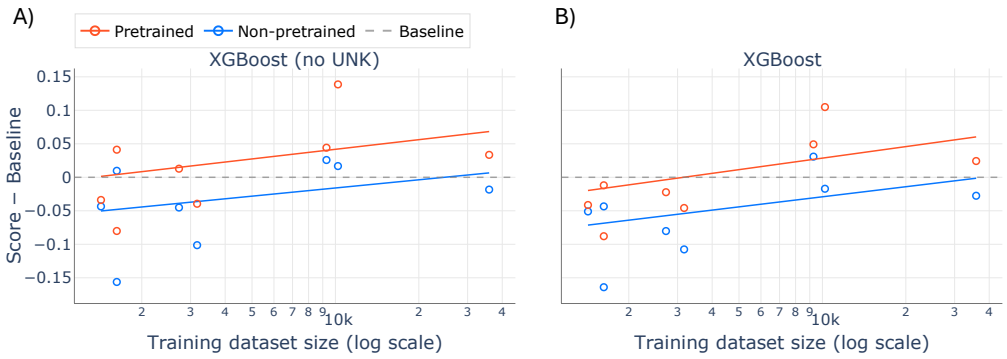

Figure 3: The difference in mean score of transformer models compared to the mean score of XGBoost (no UNK) (A) and XGBoost (B) baselines, against the size of the training dataset for each task in the benchmark. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens are unknown to the transformer models and here we plot both the XGBoost model with these tokens removed (the no UNK variant) and the standard XGBoost variant with all tokens. On Compass tasks with approximately 10,000 or more training examples, all pretrained transformers score above both XGBoost baselines. Solid lines are ordinary least squares lines of best fit.

The per-task breakdown (Figure S3) reveals that the benefit of pretraining is consistent across all tasks. This demonstrates the effectiveness of self supervised pretraining to learn general representations of microbiomes that improve performance across a wide variety of downstream tasks. From the per task breakdown we also see that the "no UNK" variant of the XGBoost clearly outperforms all transformers for task 4 and task 7, demonstrating that for some of the small per task datasets we see no clear advantage for using transformers over baseline models even when vocabulary is kept consistent. Supplementary figures provide additional complementary metrics for the classification tasks. One-vs-one macro ROC AUC (Figure S4), macro precision–recall (Figure S5) and balanced accuracy (Figure S6).

3.3 Waypoint becomes more competitive as dataset size increases

To understand how model performance scales with the amount of available task specific training data, we plot the difference in the benchmark score of the best performing Waypoint model (45M) compared to the XGBoost baselines against benchmark dataset size. Figure 3 shows the difference in mean score (averaged over the three independent runs) between Waypoint and XGBoost as a function of training set size for each task for both the no UNK (Figure 3A) and standard (Figure 3B) XGBoost variants. We observe that Waypoint's relative performance is higher on the larger Compass tasks. The pretrained 45M model achieves higher mean scores than both XGBoost baselines on tasks with approximately 10,000 or more training examples. The pretrained model outperforms the non-pretrained model across the full range of dataset sizes examined, showing that pretraining provides a stable benefit within the range of dataset sizes in the benchmark. Conversely, on tasks with fewer than 4,000 examples, the Waypoint model underperforms the no-UNK baseline, implying that a minimum quantity of training data is required before transformer-based approaches become competitive. These results suggest that Waypoint models are a practical choice for the tasks studied here whenever training datasets of at least 10k examples can be assembled.

3.4 Pretraining changes downstream model-scaling behavior

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 transformer models. 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,

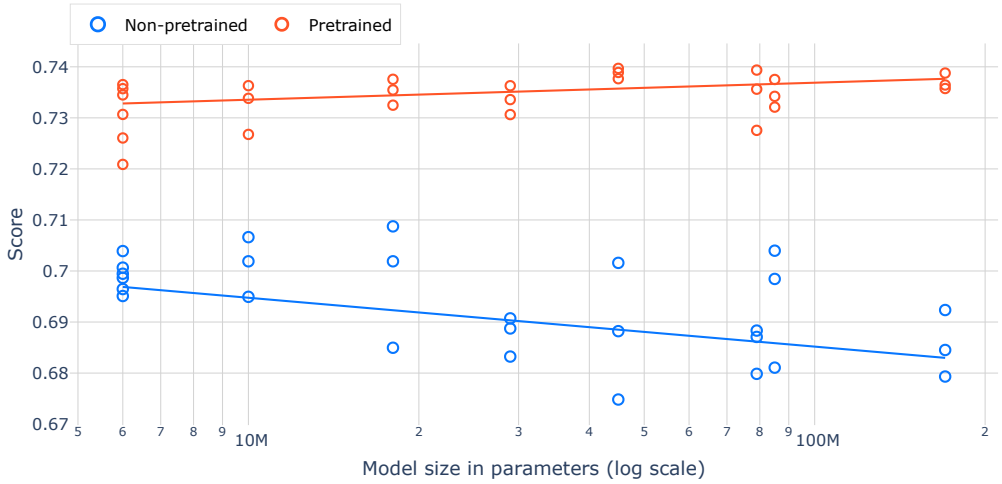

Figure 4: Benchmark performance as a function of model size for pretrained and non-pretrained transformer models. Each point represents the score from an individual benchmarking replicate, solid lines are ordinary least squares lines of best fit. Pretrained models show consistent improvements with increasing parameter count, while non-pretrained models exhibit a decline in performance at larger scales, highlighting the importance of pretraining for effective model scaling.

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 appear harder to optimize and 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, as scaling a randomly initialised model may even be detrimental. Scaling results for each task are shown in Figure S7, which shows that performance as a function of model size and the effect of pretraining can be variable between tasks, even for those derived from the same underlying datasets, so broader experiments are needed to determine how generally this scaling pattern holds.

3.5 Pretrained representations generalise between microbiome domains

The Atlas corpus spans a broad range of ecological contexts, from host-associated to marine, freshwater, terrestrial, and engineered environments. The Compass benchmark, in contrast, is dominated by gut tasks. This raises a natural question: do the gains from pretraining derive specifically from in-domain (gut or human) samples, or do learned representations transfer across biomes? To answer this, we pretrained the 6M Waypoint model on five subsets of Atlas, defined from the biome lineage metadata (Appendix A.5 for details): the complete corpus (*Full-pretrained*); only human-associated samples (*Only human*); only gut samples (*Only gut*); all samples except human-associated (*Exclude human*); and all samples except gut (*Exclude gut*). Any datapoints with missing labels such that their membership in a set was ambiguous were removed to prevent unintentional contamination of the data. Every subset model used the identical architecture, tokeniser, and pretraining and fine-tuning procedure as the standard 6M model, and each was evaluated on the Compass benchmark across three replicates. Because the subsets have varying numbers of data points and necessarily contain less data than the full dataset (Table S5) we also pretrained five additional models in which each dataset was randomly subsampled to 110k datapoints to control for the effect of dataset size on model performance. For all subset models 90% of the corresponding pretraining data subset was used for training and the remaining 10% for validation. Additionally, we plotted the aggregate benchmark score excluding task 1, this is because task 1 is the only task in the benchmark which is

not focused on the human gut microbiome and so removing this allows us to more clearly assess the effect of different pretraining datasets within this domain.

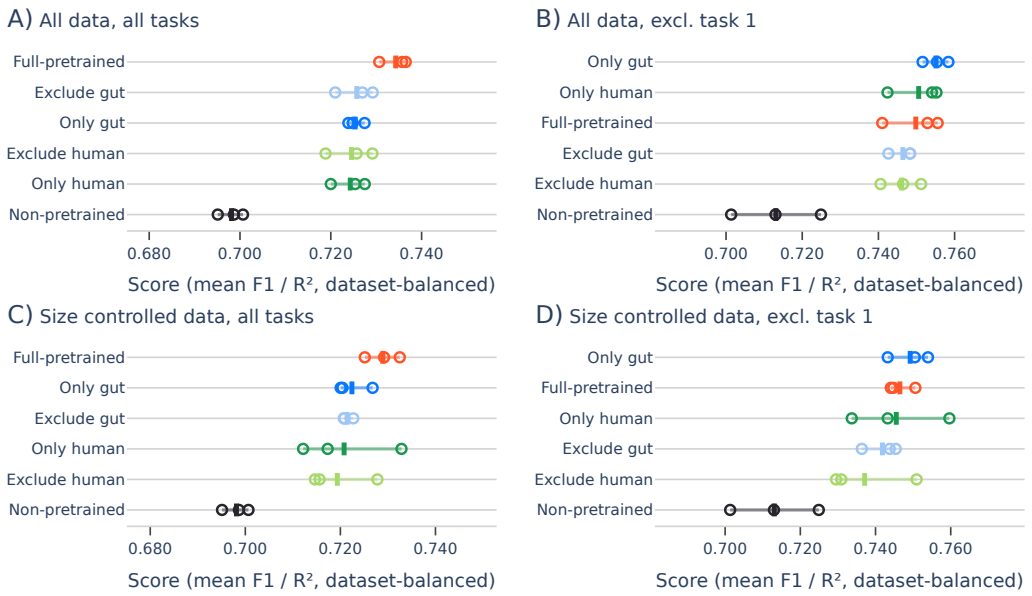

Figure 5: Pretraining transfers across microbiome domains. Compass scores for Waypoint-6M models pretrained on the complete Atlas corpus, domain-restricted subsets (*Only human* and *Only gut*), or complementary subsets (*Exclude human* and *Exclude gut*), compared with a non-pretrained model. (A–B) Models pretrained using all available samples in each subset; (C–D) models pretrained after each subset was randomly downsampled to 110,000 samples to control for corpus size. Panels A and C report scores across all eight Compass tasks, whereas panels B and D exclude the non-gut biome-classification task (task 1). Points show three independent benchmarking replicates and vertical ticks indicate their means. Models pretrained without human-associated or gut samples retain most of the benefit of pretraining, indicating transfer across ecological domains.

The mean benchmark score for each pretraining dataset without subsampling is shown in Figure 5A. Every pretrained condition substantially outperforms the non-pretrained model, with the model trained on the full corpus performing the best. Strikingly, the four subset models cluster tightly together and are essentially indistinguishable from one another. The higher scores of the *Exclude gut* and *Exclude human* models relative to the non-pretrained model demonstrates that transfer learning between microbiome domains is effective, providing evidence that representations learned through autoregressive pretraining transfer across ecological contexts. In all four aggregate comparisons in Figure 5, the Exclude models score above the non-pretrained model. Removing task 1 from the benchmark (Figure 5 B) notably, but expectedly, lowers the relative scores of *Full-pretrained* and *Exclude gut* such that the gut and human focused datasets are now at the top, however we still see clear transfer learning in comparison to the non-pretrained model. The results from the size controlled pretraining datasets (Figure 5 C, D) show very similar trends, with the expected decrease in score across all pretrained conditions due to lower dataset sizes (the non-pretrained model is unchanged as it uses no pretraining data).

The per-task breakdowns (Figure S8, Figure S9) confirm that this pattern is consistent across the benchmark. For most tasks, all pretrained subsets outperform the non-pretrained model, and the in-domain-only and out-of-domain subsets overlap heavily, with differences between subsets generally falling within the replicate-to-replicate variation.

The results in this section demonstrate that pretraining transfers across ecological contexts: models that never saw gut samples, or never saw human-associated samples, still recover the large majority of the benefit of pretraining on a benchmark dominated by human gut tasks, outperforming the non-pretrained model. The representations a Waypoint model learns from a microbial community remain useful well outside the ecological context it was trained on. This is an encouraging result

for microbiome modelling in general, where public data is abundant for a handful of well-studied environments and sparse for most others: it suggests that models for under-sampled biomes need not wait for domain-specific data at scale, and can instead inherit most of their representational quality from whatever broad corpus is available.

4 Discussion and Limitations

In this work, we introduced Atlas, a large-scale microbiome pretraining corpus; Compass, a benchmark of eight predictive tasks; and Waypoint, a family of GPT-2-style microbiome foundation models. Across Compass, self-supervised pretraining consistently improved downstream performance relative to random initialisation. Our 6M-MGM model also outperformed the architecturally matched MGM model Zhang et al. (2026), suggesting that the larger pretraining corpus and fallback tokenisation improve representation quality. Within Compass, pretrained Waypoint models outperform both XGBoost baselines on tasks with approximately 10,000 or more labelled examples, suggesting that transformer-based approaches can be competitive in large cohort and screening studies; for example, the drug-metabolism dataset of Mastrorilli et al. (2026) used in task 6 of Compass. Pretraining also changed model-scaling behaviour: performance generally improved with model size after pretraining but declined without it. Finally, models pretrained without human-associated or gut samples retained most of the benefit of pretraining on predominantly human-gut tasks, with performance similar to domain-matched models after controlling for corpus size. Together, these findings show that corpus construction, tokenisation, and pretraining are central to learning scalable and transferable representations of microbial communities.

Several limitations of this study should be acknowledged. First, Compass currently focuses exclusively on gut and environmental microbiome tasks; while the pretraining corpus is ecologically diverse, the benchmark itself does not yet reflect the full breadth of microbiome applications, including oral, skin, or respiratory microbiomes. Expanding the benchmark to cover a wider range of body sites and environments is an important direction for future work. Second, we were unable to include comparisons with two recently published microbiome foundation models Pope et al. (2025); Medearis et al. (2026) due to the absence of publicly available models, which limits the scope of our comparative analysis. We encourage the release of model weights alongside future publications to facilitate community benchmarking. Furthermore, our models currently have two obvious limitations which should be addressed in subsequent work; they are not aware of relative abundance values beyond the rank ordering of taxa and they are vulnerable to token mismatch between the pretraining and finetuning datasets. Previous papers have suggested the efficacy of more fine-grained treatment of relative abundance Medearis et al. (2026), although other work has suggested that rank ordering is just as effective Pope et al. (2025). We see that token mismatch is an issue of varying magnitude with all of our benchmarking datasets (Table S4), likely limiting the efficacy of our models, so future work to generalise tokenisation schemes to work across datasets should be done. Additional work should also explore a broader range of model architectures beyond GPT-2 to assess their suitability for microbiome modelling tasks.

Taken together, our results show that self-supervised pretraining on large-scale microbiome sequence data is a powerful and practical strategy for improving performance across diverse downstream tasks and enables foundation models to learn general representations of microbiomes. The release of the Atlas dataset will enable the community to pretrain their own microbiome foundation models. The Compass benchmark provides a standardised and reproducible framework for evaluating future models, and we hope it will serve as a community resource for tracking progress in microbiome machine learning. We have also demonstrated that Waypoint models have achieved state of the art performance and we are releasing the Waypoint-6M, Waypoint-45M and Waypoint-170M foundation models to the community. As microbiome datasets continue to grow in scale and diversity, we anticipate that the benefits of foundation model pretraining will become increasingly pronounced, making this a productive direction for future research.

AI use statement

In this work, we used generative AI tools to draft parts of this research paper and edit this research paper to improve readability. We have not used generative AI tools for formatting references, generating synthetic data sets, to help develop theoretical models or conceptual frameworks, to propose

or refine hypotheses, to design or provide feedback on research methodology or experiments, to implement methods, to assist with translation, to clean and reformat dataset, to support qualitative and thematic data analysis or to interpret results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work. Additionally, we used generative AI tools for creating and editing software code. We have reviewed all AI-assisted work. All code has been verified by at least two authors and is thoroughly unit tested and the final contents of this paper have been proofread by two human reviewers. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Details of the methodology are found in the supplementary. Additionally, the Waypoint model weights, the Atlas dataset and the Compass benchmark are available on Hugging Face at [ANONYMISED], and example pre-training and finetuning/benchmarking code is available on GitHub at [ANONYMISED]. 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.

Author Contributions

[ANONYMISED]

Acknowledgments

[ANONYMISED]

References

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.

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 Taghibakshi, 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.

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.

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.

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.

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.

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.

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.

Hansol Lee, Songyeon Lee, Ingoo 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.

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.

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

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 Spoendlin, 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 Tadimeti, 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.

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.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

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.

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.

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.

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.

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/advs.202513333. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202513333.

A Methods

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 three sequencing modalities: amplicon 16S rRNA, whole-genome shotgun metagenomic and metagenomic assembly. 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 removed, 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%, 485,377) and a held-out benchmarking set (10%, 53,931) 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. 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), taxa within each sample are ordered by their z-scored relative abundances, calculated using statistics from the pretraining corpus. For taxon (t) in sample (i), the z-score is calculated as $z_{i,t} = \frac{x_{i,t}-\mu_t}{\sigma_t}$, where $(x_{i,t})$ is the observed relative abundance of taxon (t) in sample (i), and (μ_t) and (σ_t) are its mean and standard deviation, respectively, across the pretraining corpus. Tokens are then sorted from highest to lowest z-score. The means and standard deviations of each token in the pretraining dataset are serialised with the tokeniser, allowing the same normalisation to be applied to downstream tasks. Sequences are truncated to a maximum length of 510 tokens and then flanked by `<bos>` and `<eos>` tokens (to give a total maximum length of 512 tokens). The vocabulary is initialised during pretraining using all taxa in the pretraining dataset. During downstream benchmarking, unknown taxa 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, leaving 436,840 samples for training and 48,537 for validation; 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

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 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. (2025) 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.

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 infant 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. Supplementary figures report additional complementary metrics; one-vs-one macro ROC AUC (Figure S4) captures pairwise discriminability by averaging AUC values across all class pairs, macro precision–recall (Figure S5) summarizes the balance between precision and recall across classes without being dominated by label frequency, balanced accuracy (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 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 were taken from the repository provided in the paper Zhang et al. (2026).

To benchmark non-pretrained version of the models, weights were simply randomly reinitialised prior to benchmarking.

Baselines Four baselines operating on taxon abundance features were included for comparison. All four share a common feature pipeline.

For the "no UNK" variants taxa that the pretrained model cannot represent are first removed, so that the baselines see exactly the same taxonomic information as the tokenizer-based models and the comparison is not confounded by differences in vocabulary coverage. Each taxon name is tokenized with the pretrained model's tokenizer and dropped if any of its tokens maps to the unknown token; the relative abundances of the surviving taxa in that sample are then renormalised to sum to one.

A bag-of-taxa vector is then constructed for each sample, where each dimension corresponds to a unique taxon and its value is the observed relative abundance (0 if absent). Abundances are then centred-log-ratio (CLR) transformed row-wise with a pseudocount of 10^{-6}, and each taxon dimension is standardised to zero mean and unit variance using training-split statistics. For tasks including drug identity, a one-hot drug encoding is appended after the transform and is itself left untransformed.

For all four models, hyperparameters were selected by grid search on the validation split, scored by cross-entropy for classification and mean squared error for regression, and class imbalance was handled by class-balanced weighting of the loss ($n/(K \cdot n_k)$ for class k). For logistic regression (classification) and ridge regression (regression), the regularisation strength was selected by grid search over $0.01, 0.1, 1, 10$ with a maximum of 1000 solver iterations. For the random forest baseline, the search covered number of trees $(100, 200, 300, 500)$ and maximum depth $(10, 20,$ unlimited$)$, with $\sqrt{d}$ features considered per split. For the XGBoost baseline, maximum tree depth was searched over $4, 6, 8$. Because the feature matrix is very wide (approximately 29,000 taxon columns), the number of boosting rounds and the learning rate were fixed to a coupled default of 200 trees at a learning rate of 0.1 rather than scanned, as a lower learning rate simply requires proportionally more trees. Remaining parameters were fixed at row subsampling 0.8, column subsampling per tree 0.2, minimum child weight 1.0, and L_2 regularisation 1.0, using the histogram tree-construction method. The feedforward neural network baseline is a multilayer perceptron with two hidden layers of 256 and 128 units, each followed by a ReLU activation and dropout, and a linear output head. Dropout rate $(0.1, 0.3)$ and learning rate $(10^{-3}, 3 \times 10^{-4})$ were grid-searched. The network was trained with Adam (weight decay 10^{-5}) at batch size 256 for up to 200 epochs, using weighted cross-entropy for classification and weighted mean squared error for regression. Early stopping used a held-out 10% of the training rows with a patience of 5 epochs, kept separate from the validation split so that the latter is used solely for hyperparameter selection. For multi-target regression a single shared

network with a multi-output head was used, whereas the other three baselines fit one independent model per target.

A.5 Biome Subset Pretraining

To test whether the benefits of pretraining are specific to in-domain data or transfer across ecological contexts, we pretrained the 6M model on five subsets of the Atlas corpus. Subsets were defined from the hierarchical biome lineage metadata associated with each sample (e.g. `root: Host-associated: Human: Digestive system: Fecal`), the same field used to construct the biome-classification tasks. We defined the *human* domain as all data points whose lineage falls under `Host-associated: Human`, and the *gut* domain as any data point where the third biome level is equal to `Digestive system`, these can be human or non human. The five conditions were: *Full-pretrained*, the complete Atlas pretraining set; *Only human* and *Only gut*, restricted to the respective in-domain samples; and *Exclude human* and *Exclude gut*, comprising all pretraining samples *outside* the respective domain. Any datapoints with missing labels such that their membership in a set was ambiguous were automatically removed to prevent any unintentional contamination of the data.

Each subset model was pretrained from scratch using the identical architecture, tokeniser, z-score abundance ordering, optimiser settings, context length, and early-stopping criteria as the standard 6M model (see *Pretraining*). The tokeniser vocabulary was held fixed to that of the full corpus across all conditions so that downstream tokenisation and token coverage are identical, isolating the effect of pretraining-data composition. Each pretrained subset model was then fine-tuned and evaluated on all eight Compass tasks following the same downstream procedure as the main experiments, with three independent replicates per task. The non-pretrained reference model is the randomly-initialised 6M model evaluated under the same fine-tuning protocol.

B Supplementary Figures and Tables

Table S1: Tokenised corpus statistics at genus and species resolution across the full Atlas corpus. Sequence length is measured as tokens per sample. Mean, Median, Min and Max refer to sequence length across the dataset.

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

Table S2: Model configurations used in the scaling study. Layers denote the number of transformer blocks, hidden dimension the size of the model's internal representations, and heads the number of attention heads per layer.

Model	Layers	Hidden dimension	Number of 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

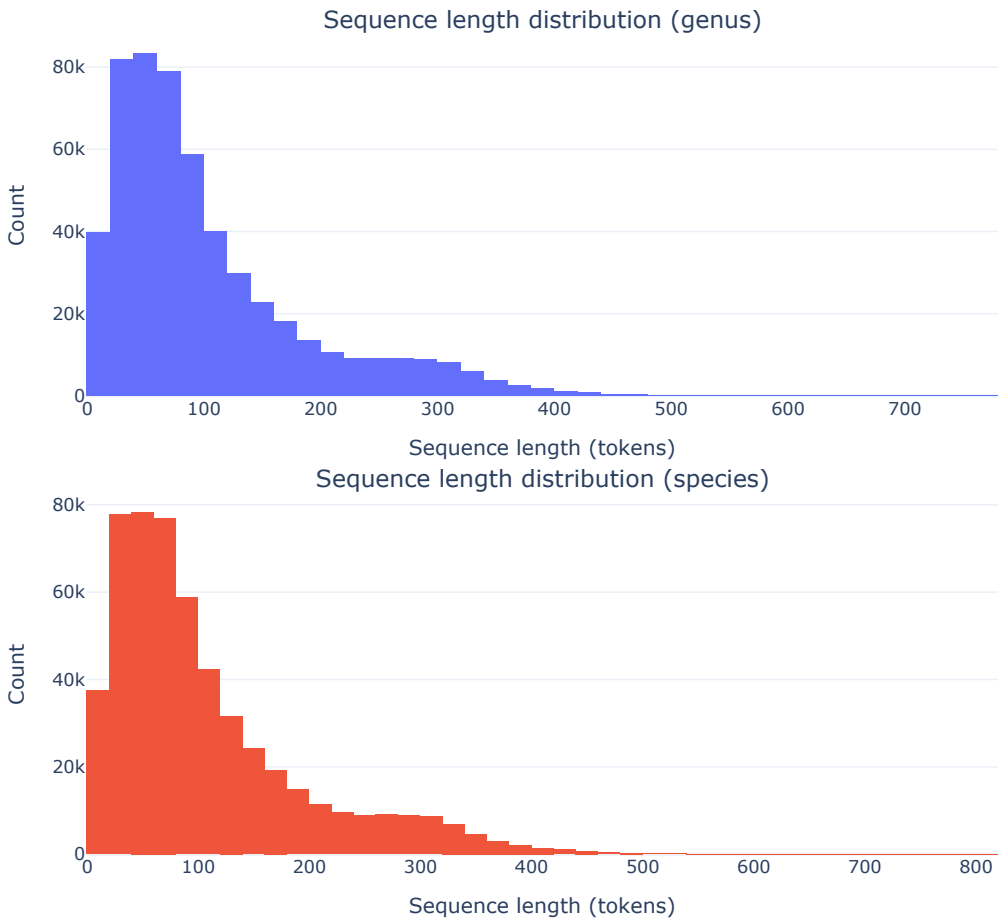

Figure S1: The distribution of sequence lengths when tokenising at the genus (top) and species (bottom) taxonomic levels.

Table S3: 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.

				N samples			
Task	**Name**	**Dataset**	**Type**	**Total**	**Train**	**Val**	**Test**
1	Biome	Richardson et al. (2022)	Class.	41,399	33,121	4,139	4,139
2	Gut biome	Richardson et al. (2022)	Class.	12,538	10,016	1,268	1,254
3	SDC	Shi et al. (2025)	Class.	3,394	2,725	340	329
4	Drug vs. non-drug	Shi et al. (2025)	Class.	3,960	3,168	396	396
5	Drug class (ATC)	Shi et al. (2025)	Class.	1,778	1,425	182	171
6	Drug degradation	Mastrorilli et al. (2026)	Reg.	15,419	9,282	3,084	3,053
7	Infant age	Roswall et al. (2021)	Class.	2,031	1,625	203	203
8	Delivery mode	Roswall et al. (2021)	Class.	2,031	1,625	203	203



Figure S2: A-D) PCA and t-SNE plots of mean pooled model embeddings coloured according to the 1st, 3rd, 4th, 5th levels of biome hierarchy.

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 S4: 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.

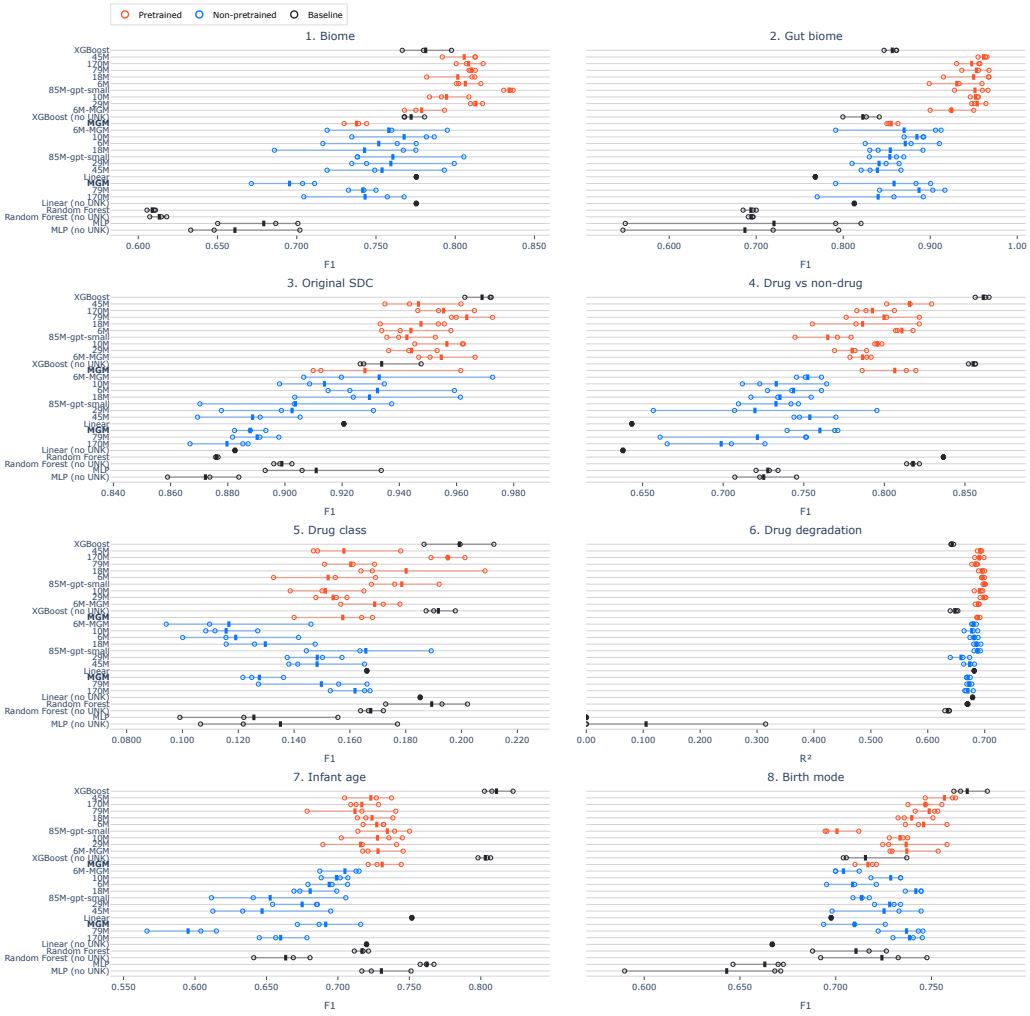

Figure S3: Score per task for all models, classification tasks (1-5 and 7-8) are scored using macro averaged F1 score and the regression task (6) is scored using R^2. Solid lines indicate mean score and the hollow circles individual replicates. Pretrained models consistently outperform non-pretrained models, but for some tasks some baselines outperform 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 S4: 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.

Table S5: Pretraining subsets of Atlas used in the cross-biome transfer experiments. The *human* domain comprises all `Host-associated: Human` samples and the *gut* domain as any data point where the third biome level is equal to `Digestive system`, these can be both human and non human.

Subset	**Pretraining samples**
Full-pretrained	485,377
Only human	153,623
Only gut	112,352
Exclude human	217,393
Exclude gut	258,664
Missing biome metadata	114,361

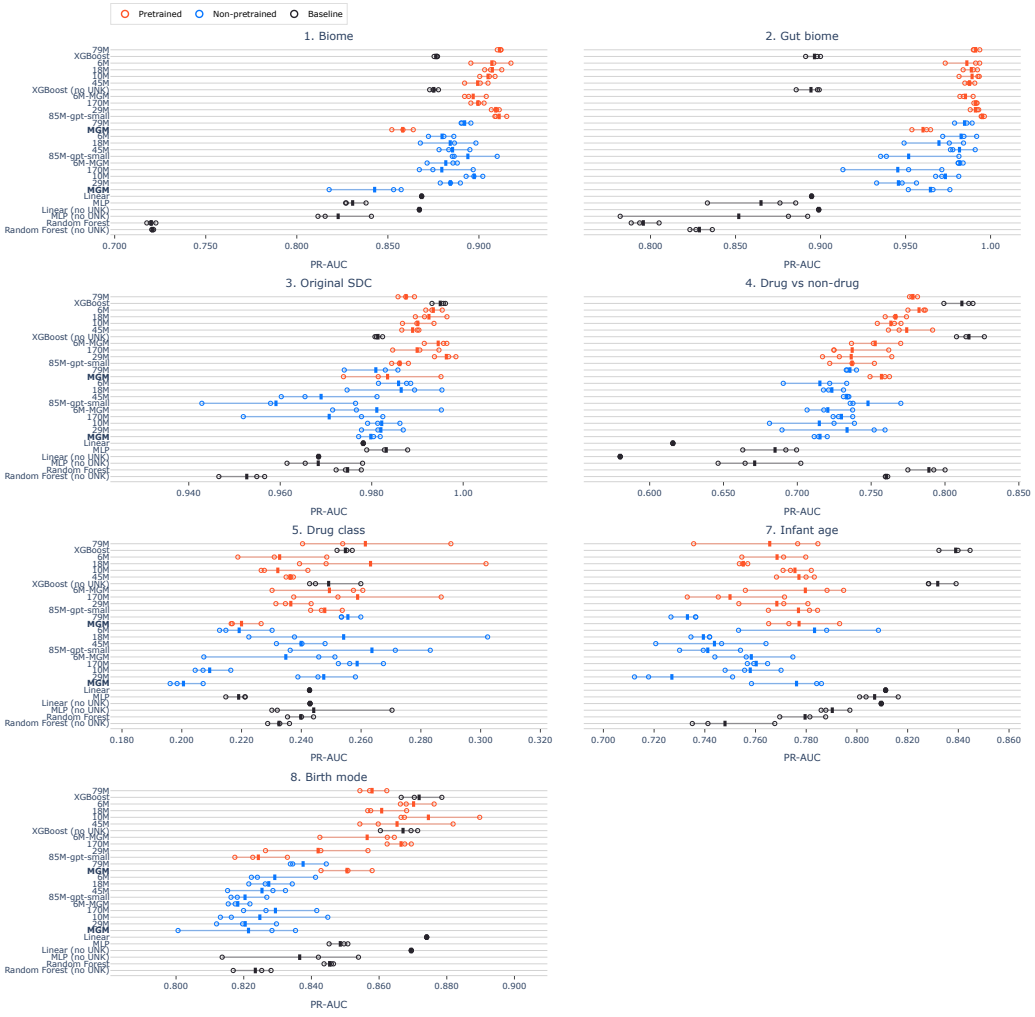

Figure S5: PR AUC per classification task for all models. Pretrained models consistently outperform non-pretrained models, but we see that for some tasks the RF baseline outperforms the transformer models. Because the benchmarking datasets contain taxonomic labels not seen during pretraining, some tokens 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 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.

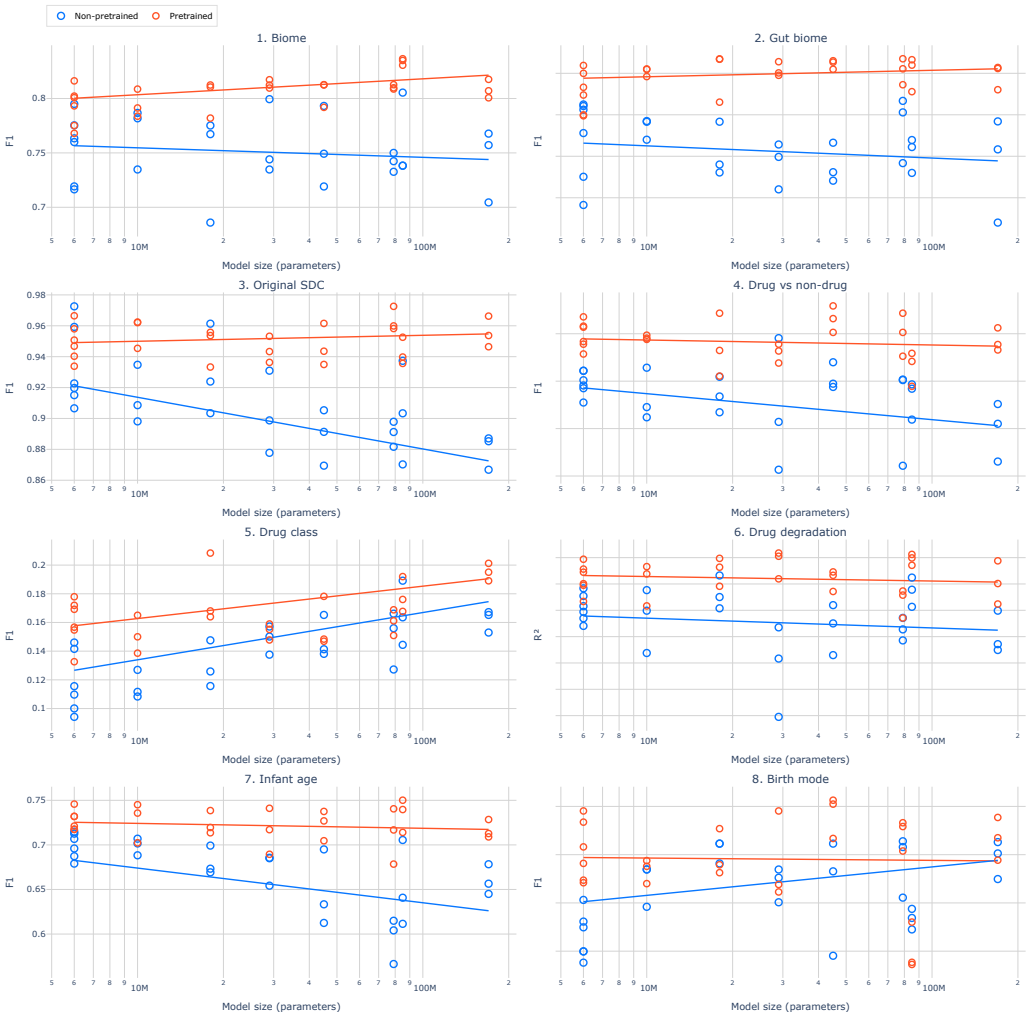

Figure S7: Benchmark performance as a function of model size for pretrained and non-pretrained transformer models for all tasks in the benchmark. Each point represents the score from an individual training replicate, solid lines are ordinary least squares lines of best fit. Scaling behaviour is not uniform across tasks.

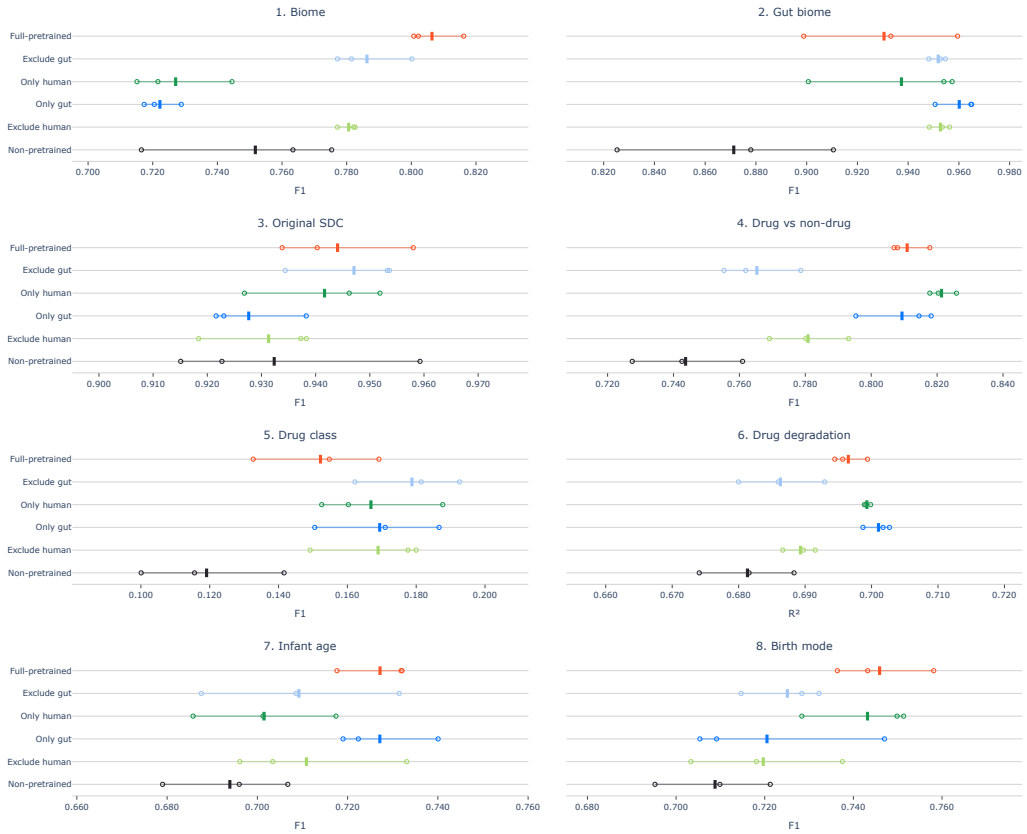

Figure S8: Per-task Compass scores for the 6M model pretrained on different biome subsets of Atlas. Classification tasks (1–5, 7–8) are scored by macro-averaged F_1 and the regression task (6) by R^2. Solid lines indicate the mean and hollow circles individual replicates. The cross-biome transfer observed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

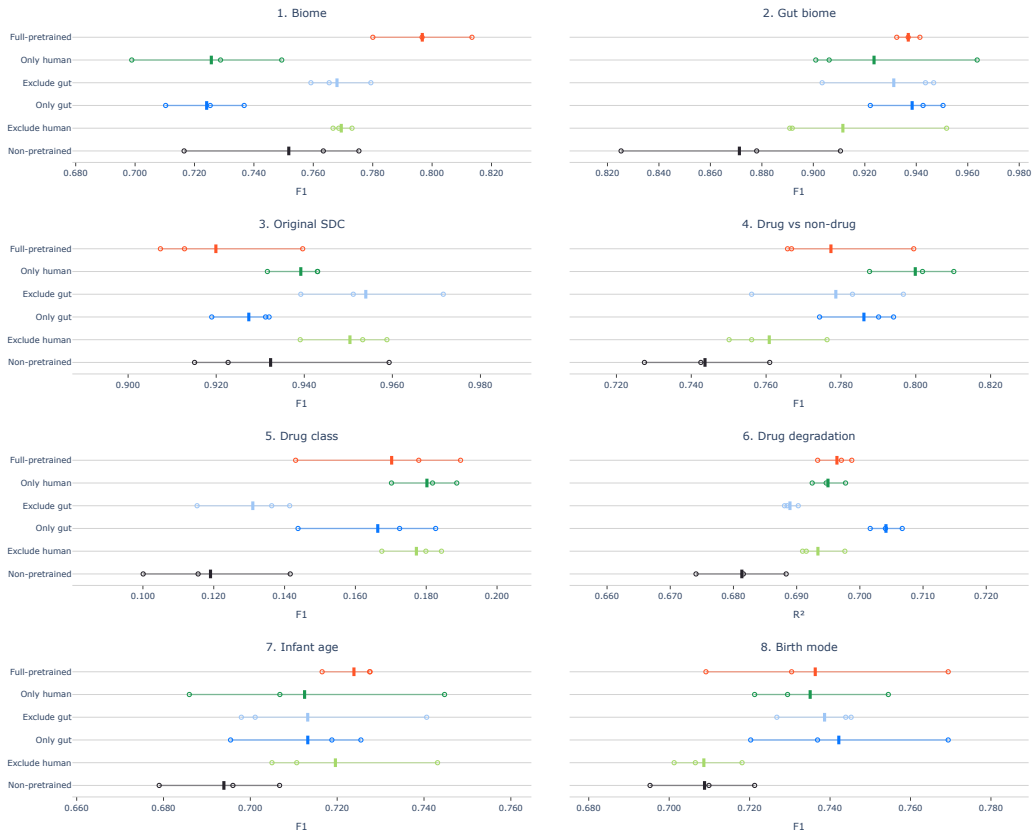

Figure S9: Per-task Compass scores for the 6M model pretrained on different biome subsets of Atlas where all datasets are subsampled down to 110k data points. Classification tasks (1–5, 7–8) are scored by macro-averaged F_1 and the regression task (6) by R^2. Solid lines indicate the mean and hollow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