A Foundation Model for Metagenomic Sequences

Ollie Liu[◇⋆] **Sami Jaghouar**[♠] **Johannes Hagemann**[♠] **Jeff Kaufman**[♡] **Willie Neiswanger**[◇⋆]

[◇] University of Southern California [♠] Prime Intellect [♡] Nucleic Acid Observatory

Abstract

We pretrain a 7-billion-parameter autoregressive transformer language model, which we refer to as a *metagenomic foundation model (MGFM)*, on a novel corpus of diverse metagenomic DNA and RNA sequences comprising over 1.5 trillion base pairs. This dataset is sourced from a large collection of human wastewater samples, processed and sequenced using deep metagenomic (next-generation) sequencing methods. Unlike genomic models that focus on individual genomes or curated sets of specific species, the aim of MGFM is to capture the full distribution of genomic information present within this wastewater. We carry out byte-pair encoding (BPE) tokenization on our dataset, tailored for metagenomic sequences, and then pretrain our model. In this paper, we first detail the pretraining dataset, tokenization strategy, and model architecture, highlighting the considerations and design choices that enable the effective modeling of metagenomic data. We then show results of pretraining this model on our metagenomic dataset, providing details about our losses, system metrics, and training stability over the course of pretraining. Finally, we demonstrate the model's capabilities through empirical results on an initial set of genomic benchmark and out-of-distribution detection tasks, showcasing its potential for various metagenomic applications.

1 Introduction

The development of large language models trained on internet-scale text datasets has revolutionized natural language processing, finding increasingly broad applications across numerous domains. In recent years, this modeling technology has been adapted to genomic sequences—e.g., DNA or RNA strands that carry genetic information—leveraging the wealth of data generated by advances in genome sequencing over the past few decades [14, 21, 9, 36, 12]. These genomic language models aim to harness modeling power for tasks such as genome classification, phenotype prediction, gene network inference, human genome analysis, and biological design for medical and therapeutic applications. To date, most of these models have been trained on either the human genome or carefully curated collections of genomes from selected species [7, 1].

Parallel to these developments, there has been significant work on large-scale health monitoring driven largely by widespread public health crises, such as the COVID-19 pandemic [26, 23]. One notable example of this is the genomic monitoring of *wastewater*, which involves sequencing material from samples of municipal sewage [11, 8]. Wastewater contains a complex mix of organic materials generated from human activities and, when collected across multiple time points and locations, can reveal valuable information about the microbiome at a societal scale [2, 16]. Consequently, there have been various efforts to collect wastewater and sequence *metagenomic information*, i.e., information about the diverse collections of organisms and organic material present in these samples [19, 17, 18]. A key motivation for much of this work is the potential to track the prevalence of human pathogens, effectively creating an early warning system for pandemics. Multiple ongoing initiatives are collecting

Corresponding authors: me@ollieliu.com, neiswang@usc.edu

Foundation Models for Science Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

vast amounts of metagenomic information to monitor genomic trends, estimate the prevalence of sequences of interest, and detect new or emerging potential pathogens [8, 15, 16].

These wastewater metagenomic sequencing efforts present two significant opportunities. First, they provide a novel and rich source of metagenomic data, rivaling the scale of datasets used to pretrain large language models (i.e., trillions of nucleic acid base pairs), encompassing highly diverse genomic information across the broad human-adjacent microbiome [4, 30]. This metagenomic data often exhibits unique distributional characteristics in terms of genomic sequence length, heterogeneity, and composition/type of organisms, distinguishing it from previous genome modeling datasets. Second, this data opens up a new domain area for downstream applications of foundation models trained on this information. Such models could be fine-tuned for various tasks crucial to pathogen monitoring, including tracking frequencies, trends, and growth of different sequence types; representation learning for sequenced metagenomic reads; sequence alignment, error-correction, and infilling; and human pathogen detection and taxonomic classification [8].

In this paper, we take an initial step toward developing a foundation model for metagenomic data by pretraining a model on a large, new dataset sequenced from wastewater. This metagenomic dataset, which has never before been used for model training, provides a unique resource for modeling the broad distribution of sequences present in the human microbiome. Specifically, we pretrain a 7-billion-parameter autoregressive transformer model, which we refer to as a *Metagenomic Foundation Model (MGFM)*, on a diverse corpus of DNA and RNA sequences comprising over 1.5 trillion base pairs sourced from wastewater samples, which were processed and sequenced using deep metagenomic (next-generation) sequencing [3, 8]. The MGFM adopts a decoder-style language model architecture, similar to those found in the GPT and Llama families of models [22, 31], which we describe and motivate in more detail in Sec. 3.3. This choice allows us to take advantage of the broad (and rapidly growing) ecosystem of techniques and infrastructure focused on this class of models.

In the following sections, we first describe our metagenomic dataset and detail the BPE tokenization strategy used to process the sequence data. We then provide comprehensive details of our MGFM model architecture and of the pretraining process on our dataset. Subsequently, we demonstrate the model's performance over the course of pretraining and on metagenomic test data. Additionally, we demonstrate that our pretrained MGFM achieves reasonable scores on standard genomic evaluation tasks—designed to evaluate models trained on human and animal genomes—highlighting its generalization capabilities. As an initial demonstration of downstream application potential, we construct a novel detection benchmark and show that MGFM performs well on this out-of-distribution detection task. We hope our paper serves as an initial step toward a foundation model for metagenomic data, which in the future can be fine-tuned to aid in public health applications such as pathogen monitoring and early detection of emerging health threats.

2 Related Work

Language models trained on genomic sequences have been an area of active research, with many aiming to train on long DNA sequences from specific species, gained from publicly available sources. For instance, models such as DNABERT [14], HyenaDNA [21], GROVER [27], and Caduceus [28] are examples primarily trained on long sequences of *human DNA*. These models typically use encoder-based architectures or decoder-only non-transformer architectures, aiming to handle long sequence lengths. In terms of tokenization, these initial human-focused genome models have commonly employed either k-mer tokenization (with fixed values like k=3) or single-nucleobase tokenization.

Recently, the scope of genomic models has expanded to include multi-species datasets, with models like DNABERT-2 [36], NucleotideTransformer [9], GENA-LM [12], SpliceBERT [5], and DNAGPT [35] being trained on a mix of human genome data and manually curated sets from other species (for example, mixes of species from a taxonomic class, such as collections of mammals). Some of these models have also explored alternative tokenization strategies, such as byte-pair encoding, learned for their particular genomic distributions [36, 12, 27, 37].

Our metagenomic foundation model differs from these prior works in a few important ways. First, our pretraining dataset comprises shorter metagenomic sequences (arising from metagenomic next-generation/massively-parallel sequencing methods) performed on samples of human wastewater collected across many locations; these samples contain potentially tens-of-thousands of species across a wide range of taxonomic ranks, and capture a representative distribution of the full human-adjacent microbiome. This includes both recognized species and many unknown or unclassified sequences

(see Sec. 3.1). Another distinction is the model architecture: we use a decoder-only transformer model, akin to the Llama/GPT families, which we further motivate in Sec. 3.3.

3 Metagenomic Foundation Model (MGFM)

We pretrain a 7-billion-parameter autoregressive transformer language model, which we refer to as a *metagenomic foundation model (MGFM)*, on a novel corpus of diverse metagenomic DNA and RNA sequences comprising over 1.5 trillion base pairs. This dataset is sourced from a diverse set of human wastewater samples, which were processed and sequenced using deep metagenomic (next-generation) sequencing methods. Before training, we carry out byte-pair encoding (BPE) tokenization on our dataset, tailored for these nucleic acid sequences. The following sections provide detailed descriptions of the pretraining dataset, tokenization strategy, and model architecture, highlighting the considerations and design choices that enable the effective modeling of metagenomic data.

3.1 Metagenomic Dataset

One of the goals of our metagenomic foundation model is to train on a genomic dataset that captures the immense diversity of the microbiome surrounding humans. To achieve this, we leverage a newly collected metagenomic dataset—never before used in model training—comprising a broad range of organisms, including bacteria, human cells, human-infecting pathogens, and a diverse array of other species, which was collected via *metagenomic sequencing of human wastewater* (i.e., municipal influent). This approach contrasts with prior genomic language models, which often focus on specific species or genomic types. By incorporating DNA and RNA sequences collected from wastewater, we aim to model the complexity of microbial and viral interactions in human-associated environments.

The dataset was generated using deep metagenomic sequencing, specifically leveraging Illumina sequencing technology, commonly referred to as next-generation sequencing (NGS) or high-throughput sequencing, in which billions of nucleic acid fragments are simultaneously sequenced in a massively parallel manner.

This method produces paired-end reads, where each read consists of two contiguous sequences of base pairs from opposite ends of a DNA or RNA fragment. Paired-end reads canoffer advantages in accuracy and alignment over single-end reads,particularly for complex metagenomic samples. Notably, the nature of metagenomic NGS results in much shorter reads compared to datasets used in many previous genomic language models. In our dataset, most reads range from 100 to 300 base pairs in length (after adapter removal and quality trimming), which introduces unique challenges for modeling, but also provides a rich diversity and large set of biological information.

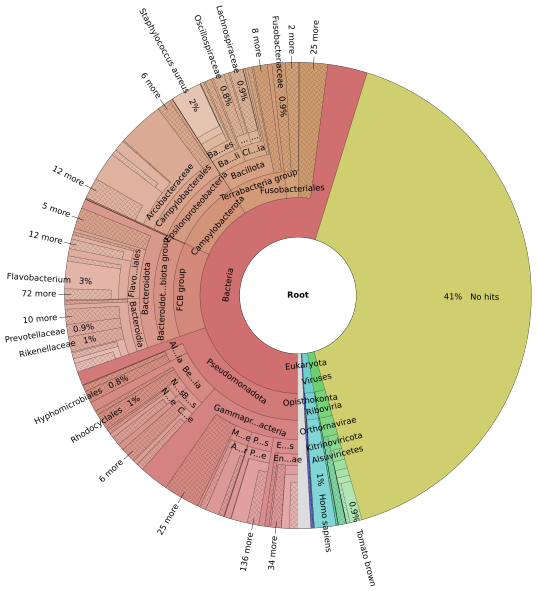

Figure 1: Metagenomic composition of the MGFM pretraining dataset, estimated via *Kraken 2* [32] sequence classification. See Fig. 5 for a more-detailed view.

This metagenomic sequence corpus was collected over a six-month period by the Nucleic Acid Observatory (NAO) [8] in collaboration with partners (Marc Johnson and Clayton Rushford at the University of Missouri and Jason Rothman in Katrine Whiteson's lab at the University of California, Irvine). Samples of wastewater were sourced from various locations across the United States, namely from cities in California, Missouri, and Massachusetts. After wastewater samples were collected, the material was filtered and nucleic acids extracted [25, 24] before undergoing metagenomic sequencing. In full, the metagenomic dataset for pretraining comprises over *1.5 trillion base pairs*. Our hope is that

Where RNA sequences are first converted into DNA via reverse transcription. [2]https://bondlsc.missouri.edu/person/marc-johnson. [3]https://jasonrothman.weebly.com/

this careful sampling and processing approach yields a clean dataset for sequence modeling, which captures a wide array of genomic content, offering a strong foundation for the training of MGFM.

We show an estimate of the metagenomic composition of this pretraining set in Fig. 1, using the *Kraken 2* [32] sequence classification software (see Fig. 5 for a more-detailed view). At the highest level, this visualization shows that 55% of reads are hits for bacteria, 2% of reads are eukaryotes (predominantly *Homo sapiens*), 2% of reads are viruses, and 41% of reads have *no hits* and are unclassified or of unknown origin.

3.2 Tokenization

In developing our metagenomic foundation model, we sought a tokenization strategy that would enable high-accuracy sequence modeling, accommodate novel nucleic acid sequences, and align with best practices in modern large language models. We opted for byte-pair encoding (BPE) as our tokenization method, as it satisfies these criteria, and drawing inspiration from its successful application in recent language models.

BPE offers several advantages for our MGFM. Unlike fixed-length k-mer tokenization, it allows for flexible token sizes, which is beneficial for capturing varying levels of genomic information, and can allow the model to adapt to different sequence patterns and structures. Moreover, BPE's ability to tokenize novel sequences is particularly valuable for modeling diverse metagenomic sequences containing unknown, varied, and possibly novel organisms. The method also has the potential to capture semantic information within a vocabulary of tokens, which can lead to more nuanced representations of genomic data.

Model Details	**MGFM**
Architecture	Llama-2-7B
Embedding Size	4096
Intermediate Size	11008
Number of Attention Heads	32
Number of Hidden Layers	32
Vocabulary Size	1024
Sequence Length	512
Normalization	RMSNorm
Regularization	z-loss
Position Embedding	Rotary
Bias	None
Warmup Steps	2000
Batch Size	30720
Weight Decay	0.1
Learning Rate Schedule	Cosine Decay
Initial Learning Rate	6×10^{-4}
β_1, β_2	$0.9, 0.95$

Table 1: MGFM architecture details.

To implement this strategy, we first trained a BPE tokenizer on a uniformly-at-random sampled subset of our pretraining dataset, comprising 2 billion base pairs. After analyzing the distribution of token sizes and considering training efficiency, we settled on a vocabulary size of 1,024 unique tokens. This vocabulary size strikes a balance between capturing sufficient genomic complexity, maintaining sufficiently long sequence lengths (based on the distribution of token sizes), and allowing for computational efficiency. Following this tokenizer training, we applied this BPE tokenizer to our entire pretraining dataset, effectively preparing it for model ingestion and training, yielding a set of over 300 billion tokens for pretraining. We give a table showing full tokenizer details, including a list of all special tokens, in Appendix B.

3.3 MGFM Architecture

For our metagenomic foundation model, we pretrain a 7-billion-parameter autoregressive language model, using a standard dense transformer architecture, similar to the architecture used in popular language models such as the GPT and Llama model families [22, 31]. Specifically, we implement a decoder-only style transformer with a causal language modeling objective, where the model aims to predict the next token in a sequence based on the previous tokens.

This architecture choice for MGFM stands in contrast to some of the alternative approaches explored in recent genomic models, which include BERT-style bidirectional encoders [14, 36, 37] or non-attention based architectures [21, 20]. Our decision to use this particular model architecture was driven by the following motivations:

1. *Ecosystem*: By aligning with this widely-adopted architecture, we can take advantage of the growing ecosystem of techniques and associated implementations developed for autoregressive decoder-only transformer models. This extends to both pretraining optimizations and downstream applications in fine-tuning and inference.
2. *Infrastructure*: Given our large dataset size, this architecture allows us to leverage scalable pretraining infrastructure specifically designed for distributed training of this model type. This infrastructure has demonstrated success in recent language models, enabling efficient training on massive datasets.

3. *Data characteristics*: The nature of our metagenomic sequence data, which primarily consists of short sequences, does not necessitate architectures designed for extremely long context lengths. This makes the transformer a suitable and efficient choice for our use case.

We next describe some of the specific configuration details of MGFM. First, the model operates with a context length of 512 tokens, which is sufficient for all of the metagenomic sequences in our pretraining dataset. For efficiency, we pack shorter sequences within this context window, a process detailed in Section 4.3 below. We use an attention mask which prevents attention between the distinct packed sequence reads. MGFM consists of 32 layers and 32 attention heads, with an embedding size of 4096 and a hidden layer size of 11008. We employ root mean square layer normalization throughout the model, with a normalization epsilon of 1e-5. These configurations result in a model with approximately 7 billion parameters in total. All architecture details are summarized in Table 1.

4 Pretraining MGFM

4.1 Training Infrastructure

Our model is trained on four nodes, each equipped with 8 H100 SXM5 GPUs interconnected via Ethernet with 40 GB/s bandwidth. This interconnect bandwidth poses a significant performance bottleneck, as it is an order of magnitude slower than NVIDIA's InfiniBand and faster Ethernet interconnects. Despite this limitation, we were able to achieve 40% model FLOPS utilization (MFU) [6] by employing a hybrid sharding strategy. Specifically, we use PyTorch's `HYBRID_SHARD_ZERO2` strategy implemented in its Fully Sharded Data Parallel (FSDP) utilities. This design choice provides the benefit of model and optimizer state sharding within each node, while practicing standard data parallelism across nodes to reduce the inter-node communication overhead. In practice, it only requires an all-reduce operation on the gradient buckets during the optimizer step.

For training, we use a global batch size of 30,720, a sequence length of 512, and a micro-batch size of 48. We observe this combination to offer the best trade-off between high MFU and reduced memory usage; it also allows us to shard the optimizer state and gradients within a single node. Further tests on fewer nodes yield MFU values of 0.51 and 0.47 for 1-node and 2-node setups, respectively. These results suggest that interconnect bandwidth was the main bottleneck in our training environment.

Node failure. During training, we experienced three node failures, one GPU failure, one network failure, and one disk failure. All failures required us to restart the training from the latest checkpoint.

4.2 Stability

Foundation model pretraining is prone to suffer from training instability, which can be more pronounced when scaling models to billions of parameters [33]. Such instabilities often arise during the middle or late stages of training, and are often characterized by a sudden spike in loss and/or other divergent behaviors. Failure to identify these problems can result in considerable wasted compute resources. Additionally, the characteristics of the input data have been shown to influence training stability, as highlighted by recent work in large multimodal language models [29].

Given that we scaled directly from sub-billion parameters to a 7 billion parameter model, and that training on metagenomic sequences is less studied compared to natural language, we anticipated a relatively high risk of encountering stability issues. To mitigate such risks, we followed best practices from Wortsman et al. and implemented a variant of the *z-loss*, referred to as *max-z-loss*, introduced by Yang et al. with a coefficient of 2e-4. We opted against the recommendation of QK-layer normalization [29] to preserve the Llama architecture and leverage optimized inference pipelines.

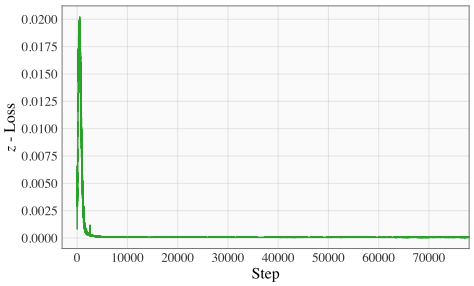

Figure 2: We show z-loss during pretraining, which aids and gives an indicator of stability.

During training, we monitored the norms of the language model head, the query, key, and value outputs, as well as the gradient norms. Wortsman et al. empirically shows that a significant increase in any of these metrics may signify potential instability, allowing us to intervene early by restarting

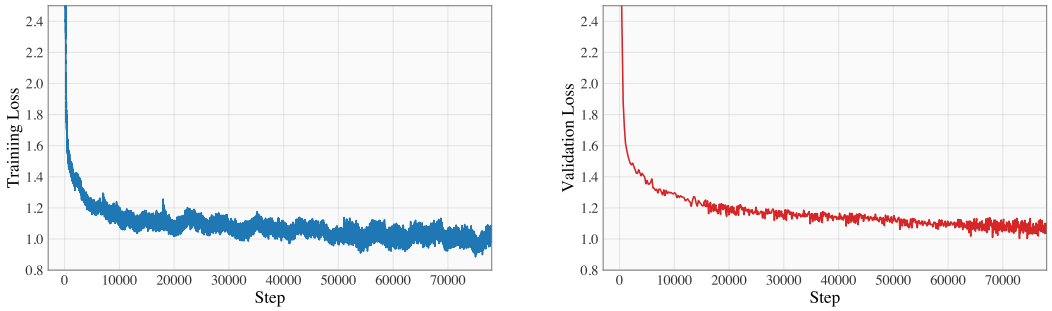

Figure 3: MGFM loss curves during pretraining. We show training loss (left), and validation loss on a held out metagenomic sample (right).

the training. Fortunately, no stability issues were observed, and the monitored metrics remained consistent throughout the training process.

4.3 Context Stuffing

A significant portion of our dataset contains sequences with fewer tokens than our model's context length. To optimize compute efficiency and avoid wasting resources on padding tokens, we pack the sequence dimension with multiple samples, where applicable. We modify the attention mask to ensure that tokens from different samples cannot attend to one another. This is implemented using the variable length function in *FlashAttention-2* [10] which avoids materializing the full mask, which would have been inefficient.

5 Empirical Results

5.1 Pretraining Performance

As an initial analysis of MGFM, in Figure 3, we show two loss curves generated over the course of pretraining. On the left, we show the training loss over one epoch of our 1.5-trillion-base-pair pretraining dataset. On the right, we show the validation loss, computed on a held-out portion of our metagenomic dataset. In the training curve we note that there are slight systematic oscillations over the course of training, which occur due to pseudo-random data shuffling (implemented for efficiency reasons); however, these do not appear in our validation loss curve.

5.2 Fine-Tuning Performance on Out-of-Distribution Data

We now investigate the viability of MGFM as a general-purpose foundation model. Importantly, we aim to show reasonable performance on nucleotide sequences sampled from out-of-domain distributions. One such example is long-sequence full-animal-genome datasets. In many prior genomic language models' pretraining datasets, this type of genomic data is found in abundance [9, 14, 21, 36]. As a pilot study, we perform fine-tuning experiments on the GUE benchmark [36], which comprises 28 sequence-level classification tasks curated from this type of genomics data.

As a minimal setup, we fine-tune low-rank adapters (LoRA) [13] and a linear classification head that projects average-pooled representations from the last hidden layer to the class logits. This setup is aimed to emulate downstream users with a limited compute budget. For each experiment, we choose the learning rate between 1e-3 and 1e-4 based on the convergence behavior of the training loss. The LoRA parameters are only introduced to the query and value projections for all but the epigenetic marks predictions (EMP) tasks. For the latter, additional adapters are added to the the key and dense layers. All other hyperparameters are fixed across all tasks. We report final test performances after 5 epochs of training. Additional details on training hyperparameters can be found in Appendix C. Following Zhou et al., we report Matthews correlation coefficient (MCC) on all but the COVID task, which uses the F1 score. These results are presented in Table 2.

Named function `flash_attn_varlen_func` in the *FlashAttention-2* Python package.

	CNN	HyenaDNA	DNABERT	NT-2.5B-Multi	DNABERT-2	MGFM
TF-MOUSE	45.3	51.0	57.7	67.0	68.0	65.9
0	31.1	35.6	42.3	63.3	56.8	55.4
1	59.7	80.5	79.1	83.8	84.8	80.0
2	63.2	65.3	69.9	71.5	79.3	78.1
3	45.5	54.2	55.4	69.4	66.5	73.1
4	27.2	19.2	42.0	47.1	52.7	43.0
TF-HUMAN	50.7	56.0	64.4	62.6	70.1	66.3
0	54.0	62.3	68.0	66.6	72.0	67.9
1	63.2	67.9	70.9	66.6	76.1	69.6
2	45.2	46.9	60.5	58.7	66.5	62.6
3	29.8	41.8	53.0	51.7	58.5	54.8
4	61.5	61.2	69.8	69.3	77.4	76.7
EMP	37.6	44.9	49.5	58.1	56.0	53.7
H3	61.5	67.2	74.2	78.8	78.3	75.2
H3K14AC	29.7	32.0	42.1	56.2	52.6	50.3
H3K36ME3	38.6	48.3	48.5	62.0	56.9	54.5
H3K4ME1	26.1	35.8	43.0	55.3	50.5	41.9
H3K4ME2	25.8	25.8	31.3	36.5	31.1	38.8
H3K4ME3	20.5	23.1	28.9	40.3	36.3	37.8
H3K79ME3	46.3	54.1	60.1	64.7	67.4	61.3
H3K9AC	40.0	50.8	50.5	56.0	55.6	52.0
H4	62.3	73.7	78.3	81.7	80.7	78.7
H4AC	25.5	38.4	38.6	49.1	50.4	46.8
PD	77.1	35.0	84.6	88.1	84.2	79.4
ALL	75.8	47.4	90.4	91.0	86.8	82.6
NO-TATA	85.1	52.2	93.6	94.0	94.3	92.6
TATA	70.3	5.3	69.8	79.4	71.6	62.9
CPD	62.5	48.4	73.0	71.6	70.5	67.2
ALL	58.1	37.0	70.9	70.3	69.4	62.7
NO-TATA	60.1	35.4	69.8	71.6	68.0	66.8
TATA	69.3	72.9	78.2	73.0	74.2	72.0
SSD	76.8	72.7	84.1	89.3	85.0	82.2
COVID	22.2	23.3	62.2	73.0	71.9	69.7

Table 2: Fine-tuning results on the GUE benchmark. Non-MGFM results are adapted from Zhou et al.. The metrics used for evaluation is MCC, except for the COVID task, which uses F1 score. The header rows report macro-averaged performance metrics.

We observe MGFM to attain comparable performance to strong baselines reported in Zhou et al.. On the human transcription factor (HUMAN-TF) prediction tasks, MGFM consistently outperforms adapter-tuned NucleotideTransformer models, a family of encoder-style foundation models pretrained on genomics distributions similar to those in the GUE benchmark. On the other hand, MGFM trails behind state-of-the-art performances in select tasks from EMP and promoter detection (PD). While these performances may benefit from a more careful round of hyperparameter tuning, this gap highlights the importance of in-distribution pretraining dataset to downstream applications. As the exploratory analysis in Figure 5 suggests, while our metagenomic sequences offer a rich profile of microorganisms, they lack long/full genome sequences from diverse animal species. An effort to improve the GUE performance of MGFM, via full-model fine-tuning and continual pretraining on a mixture of our metagenomics dataset and public genomics datasets, is currently underway.

Benefits of pretraining. To ablate the benefits of pretraining for downstream performance, we perform LoRA fine-tuning from *randomly initialized* model weights. These experiments are trained

with the same hyperparameters, but we double the number of training epochs (i.e., 10 epochs) to compensate for the lack of pretraining. Additionally, we report the *best* test performance from evaluations after every epoch, in order to provide an accurate lower bound of performance gains from pretraining. We show experiments on the transcription factor prediction tasks from mouse and human genomes, and report the ablation results in Table 3.

	TF-MOUSE					TF-HUMAN				
	0	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**
MGFM w/o PT	31.7	67.7	72.6	50.7	24.6	55.6	57.7	40.5	32.0	51.2
MGFM	55.4	80.0	78.1	73.1	43.0	67.9	69.6	62.6	54.8	76.7
PT Δ	23.7	12.3	5.5	22.4	18.4	12.3	11.9	22.1	22.8	25.5

Table 3: Ablation results on the performance gain from pretraining on our metagenomics dataset. First row: LoRA fine-tuning results for MGFM *without pretraining*. Second row: LoRA fine-tuning results for standard MGFM. Third row: the performance gain from pretraining.

We observe a sizable performance gap between ablation and pretraining results. Our findings indicate that pretraining offers concrete benefits in terms of downstream performance, even if there exists a mismatch in terms of data distributions. Despite this gap, it is worth noting that LoRA fine-tuning from random weights can achieve non-trivial performance, and is often comparable to the CNN baselines. This observation could be attributed to the positive inductive bias introduced from the transformer architecture for sequence modeling.

5.3 Anomaly Detection from Wastewater

Our final experiment aims to show the feasibility of MGFM to detect out-of-distribution (OOD) data at scale, as it serves as a primer for reliable anomaly detection from wastewater samples. In this early study, we respectively sample 5000 sequences from our metagenomics pretraining data, the mouse and human genomes from the GUE dataset, as well as *random* sequences as a control group. All sequences are truncated to 100 base pairs in accordance with the sequence lengths from the GUE dataset. As a baseline, we implement a threshold-based anomaly detector, which classifies samples with length-normalized cross entropy losses below a certain threshold as non-anomalies, and *vice versa*. We select a threshold of 3 based on our observations from the validation curve in Figure 3.

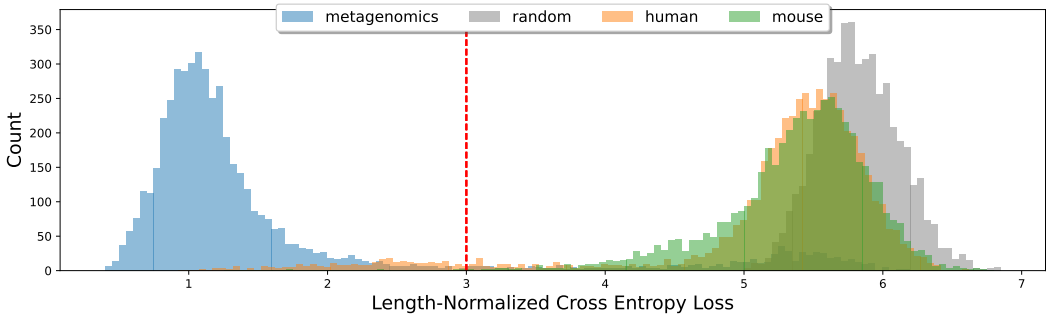

Figure 4: Distribution of the length-normalized cross entropy loss across all datasets.

Figure 4 indicates a clear separation between metagenomics sequences and other data sources. The in-distribution data behaves within our expectation; the human and mouse genomic data both attain a similar mode and spread, and their loss distributions are more similar to that of random sequences, compared to our in-distribution data. Table 4 reports numerical results of our OOD detection tests. MGFM achieves strong performance for separating metagenomics sequences from other data sources.

Group	F1	Loss (Std. Err)	Tokenized Seq Len (Std. Dev)
Metagenomics	-	1.24 (1.31)	24.91 (3.35)
Random	0.91	5.83 (0.29)	27.16 (1.32)
Human	0.94	5.22 (0.22)	27.29 (1.33)
Mouse	0.91	5.38 (0.54)	27.2 (1.34)

Table 4: OOD detection performance between metagenomics sequences and other data sources.

6 Discussion, Limitations, Conclusion

We have reported our current progress on pretraining and evaluating MGFM, the first large-scale foundation model pretrained from metagenomic sequences. We detail our dataset construction, model training, and fine-tuning procedure to facilitate open-science research. We will open-source our datasets, training code, and model checkpoints in the near future.

Our downstream performance over genomics benchmarks suggests the viability of our approach even in the face out-of-domain distributions. The performance gaps with state-of-the-art approaches on these benchmarks indicate that MGFM may benefit from continual pretraining with a diverse mixture of data sources (at least on tasks similar to these genomic benchmarks). We are actively exploring this direction, such as incorporating human reference genomes and multi-species genomic datasets.

In addition, we are actively developing a standardized evaluation suite consisting of classification, embedding, out-of-distribution detection, and pandemic monitoring tasks for metagenomics sequences. We hope our effort can facilitate objective evaluation of MGFM, and we invite both domain experts and the machine learning community to contribute to this research.

7 Acknowledgements

We thank Prime Intellect for their generous Fast Compute Grants and for compute infrastructure support. We also thank Marc Johnson, Clayton Rushford, Jason Rothman, and the team at the Nucleic Acid Observatory for the work on data collection, processing, and helpful analysis. OL thanks the Polymathic AI group and Zhihan Zhou for helpful discussions and feedback.

References

[1] Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. Genomic language models: Opportunities and challenges. *arXiv preprint arXiv:2407.11435*, 2024.

[2] Anne Bogler, Aaron Packman, Alex Furman, Amit Gross, Ariel Kushmaro, Avner Ronen, Christophe Dagot, Colin Hill, Dalit Vaizel-Ohayon, Eberhard Morgenroth, et al. Rethinking wastewater risks and monitoring in light of the covid-19 pandemic. *Nature Sustainability*, 3 (12):981–990, 2020.

[3] Lauren Bragg and Gene W Tyson. Metagenomics using next-generation sequencing. *Environmental microbiology: methods and protocols*, pages 183–201, 2014.

[4] Florian P Breitwieser, Jennifer Lu, and Steven L Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 20(4):1125–1136, 2019.

[5] Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv*, pages 2023–01, 2023.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David

Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

[7] Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. To transformers and beyond: large language models for the genome. *arXiv preprint arXiv:2311.07621*, 2023.

[8] The Nucleic Acid Observatory Consortium. A global nucleic acid observatory for biodefense and planetary health. *arXiv preprint arXiv:2108.02678*, 2021.

[9] 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, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *BioRxiv*, pages 2023–01, 2023.

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

[11] K Farkas, LS Hillary, SK Malham, JE McDonald, and DL Jones. Wastewater and public health: the potential of wastewater surveillance for monitoring covid-19. *Current Opinion in Environmental Science & Health*, 17:14–20, 2020.

[12] Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. Gena-lm: a family of open-source foundational dna language models for long sequences. *bioRxiv*, pages 2023–06, 2023.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[14] 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, 2021.

[15] Aparna Keshaviah, Xindi C Hu, and Marisa Henry. Developing a flexible national wastewater surveillance system for covid-19 and beyond. *Environmental Health Perspectives*, 129(4): 045002, 2021.

[16] Joshua I Levy, Kristian G Andersen, Rob Knight, and Smruthi Karthikeyan. Wastewater surveillance for public health. *Science*, 379(6627):26–27, 2023.

[17] Kang Mao, Kuankuan Zhang, Wei Du, Waqar Ali, Xinbin Feng, and Hua Zhang. The potential of wastewater-based epidemiology as surveillance and early warning of infectious disease outbreaks. *Current Opinion in Environmental Science & Health*, 17:1–7, 2020.

[18] Jill S McClary-Gutierrez, Mia C Mattioli, Perrine Marcenac, Andrea I Silverman, Alexandria B Boehm, Kyle Bibby, Michael Balliet, Daniel Gerrity, John F Griffith, Patricia A Holden, et al. Sars-cov-2 wastewater surveillance for public health action. *Emerging infectious diseases*, 27 (9), 2021.

[19] Gertjan Medema, Frederic Been, Leo Heijnen, and Susan Petterson. Implementation of environmental surveillance for sars-cov-2 virus to support public health decisions: opportunities and challenges. *Current opinion in environmental science & health*, 17:49–71, 2020.

[20] Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. *BioRxiv*, pages 2024–02, 2024.

[21] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[23] Alex Reinhart, Logan Brooks, Maria Jahja, Aaron Rumack, Jingjing Tang, Sumit Agrawal, Wael Al Saeed, Taylor Arnold, Amartya Basu, Jacob Bien, et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

[24] Carolyn A Robinson, Hsin-Yeh Hsieh, Shu-Yu Hsu, Yang Wang, Braxton T Salcedo, Anthony Belenchia, Jessica Klutts, Sally Zemmer, Melissa Reynolds, Elizabeth Semkiw, et al. Defining biological and biophysical properties of sars-cov-2 genetic material in wastewater. *Science of The Total Environment*, 807:150786, 2022.

[25] Jason A Rothman, Theresa B Loveless, Joseph Kapcia III, Eric D Adams, Joshua A Steele, Amity G Zimmer-Faust, Kylie Langlois, David Wanless, Madison Griffith, Lucy Mao, et al. Rna viromics of southern california wastewater and detection of sars-cov-2 single-nucleotide variants. *Applied and environmental microbiology*, 87(23):e01448–21, 2021.

[26] Joshua A Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M Rönn, Marissa B Reitsma, Katherine A Morris, Sarah LaRocca, Tamer H Farag, et al. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51):e2111454118, 2021.

[27] Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, pages 1–13, 2024.

[28] Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

[29] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[30] Michael J Tisza and Christopher B Buck. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences*, 118(23):e2023202118, 2021.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[32] Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13, 2019.

[33] Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities, 2023. URL https://arxiv.org/abs/2309.14322.

[34] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao,

Yupeng Zhang, Zenan Zhou, and Zhiying Wu. Baichuan 2: Open large-scale language models, 2023. URL https://arxiv.org/abs/2309.10305.

[35] Daoan Zhang, Weitong Zhang, Bing He, Jianguo Zhang, Chenchen Qin, and Jianhua Yao. Dnagpt: a generalized pretrained tool for multiple dna sequence analysis tasks. *bioRxiv*, pages 2023–07, 2023.

[36] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[37] Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *ArXiv*, 2024.

Appendix

A Additional Details on the Metagenomic Training Dataset

In Figure 5, we show a visualization of (a relatively small subset of) the composition of metagenomic information contained in our pretraining dataset. This composition is estimated through the *Kraken 2* metagenomic sequence classification software [32], which gives taxonomic hits for reads in our pretraining set (where taxonomic classification is performed using exact k-mer matches) . We show three plots in Figure 5: first, the full pretraining dataset distribution (top); then, as an example subset of this, the distribution of viruses (middle); and finally, as an example subset of this, the distribution of the Steitzviridae family of viruses (bottom).

B Tokenizer Details

Our tokenizer implementation is adapted from `minbpe`. It is trained on a subset of sequences consisting of 2 billion base pairs. These sequences are uniformly sampled from all of the available wastewater sequencing runs from our data sources. Similarly to BPE tokenizers trained on natural language datasets, we treat the beginning of each sequence differently, in our case by prepending a '_' character to the beginning of each read. During pretraining, we postpend a `[BOS]` token to separate each sequence. Our tokenizer consists of the following special tokens: `[PAD]`, `[UNK]`, `[SEP]`, `[BOS]`, `[EOS]`, and `[MASK]` to allow for diverse applications during fine-tuning. In total, it has of a vocabulary size of 1024.

In our preliminary experiments, we have also experimented with a larger vocabulary size of 4096, but this design choice results in many short tokenized sequences that may not be able to provide meaningful learning signal. We have thus decided to move forward with a vocabulary size of 1024 to balance efficiency and downstream performance.

C Fine-Tuning Experiment Details

In Table 5, we show our choices of hyperparameters for fine-tuning experiments.

LoRA modules	query, value$^{\Lambda}$
LoRA rank	8
LoRA α	16
LoRA dropout	0.05
Optimizer	AdamW
Optimizer momentum	β_1, β_2 = 0.9, 0.999
Learning rate	1e-3$^{\Omega}$
LR Scheduler	Linear Warmup + Constant LR
Warmup Steps	50
Weight Decay	0.01
Denominator ϵ	1e-8
Precision	BF16-mixed
Batch size	32
Epochs	5
Epochs (Ablation)	10
Hardware	NVIDIA A100 80GB

Table 5: Hyperparameter settings for fine-tuning experiments. Λ: LoRA is applied to query and value projections for all except for the epigenetic marks prediction tasks, in which case LoRA is applied to the query, key, value, and dense matrices. Ω: we use a learning rate of 1e-4 for tasks HUMAN-TF-0, EPM-H3K36ME3, and EPM-H3K4ME2, as we otherwise observe non-convergent behavior in terms of training loss.

https://github.com/karpathy/minbpe

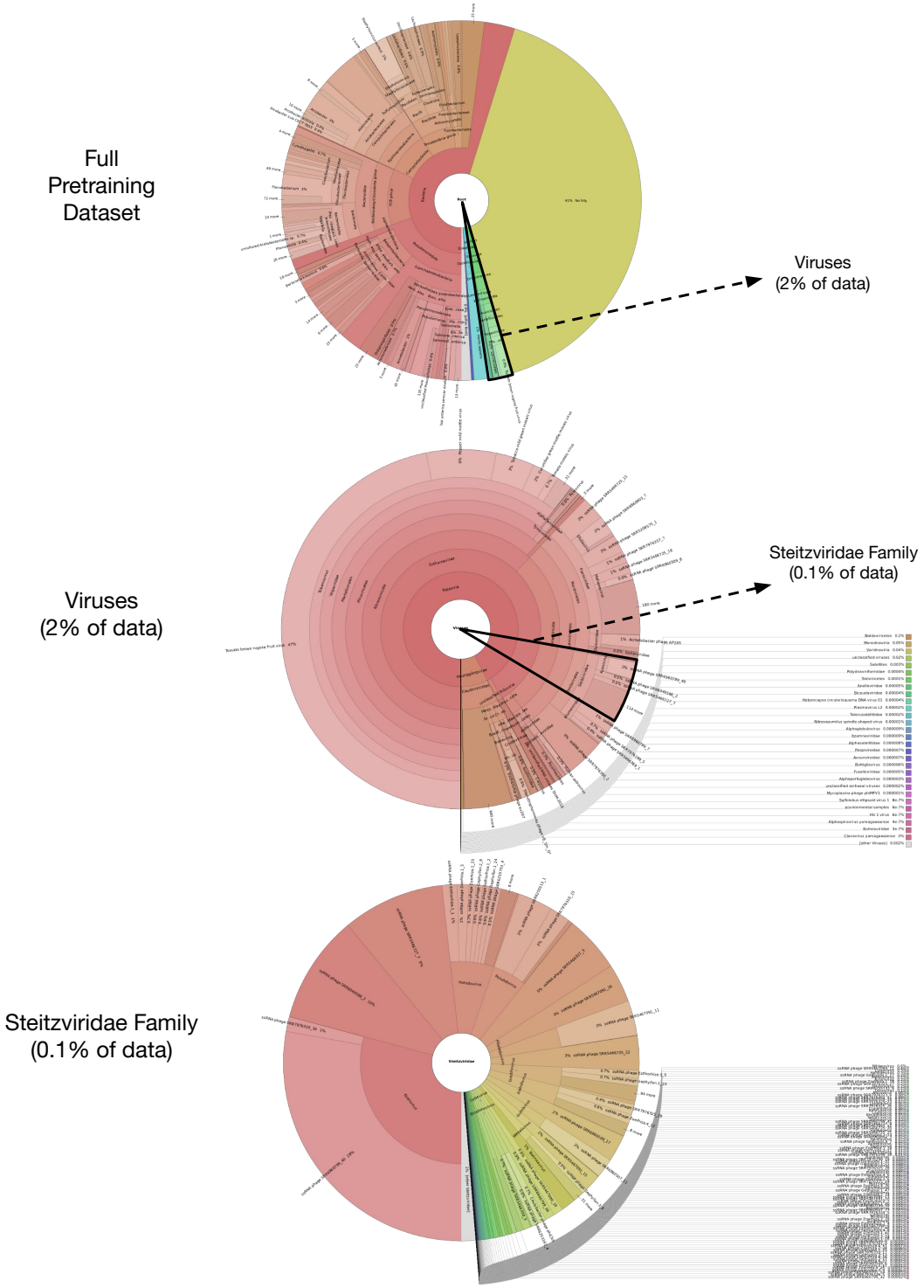

Figure 5: A visualization of the composition of metagenomic information contained in our pretraining dataset, based on *Kraken 2* metagenomic sequence classification hits [32]. We first show the full pretraining dataset distribution (top), and then as an example show the distribution of viruses (middle), and finally the distribution of the Steitzviridae family of viruses (bottom).