

# METAGENE-1: Metagenomic Foundation Model for Pandemic Monitoring

Ollie Liu<sup>1</sup>, Sami Jaghouar<sup>2</sup>, Johannes Hagemann<sup>2</sup>, Shangshang Wang<sup>1</sup>,  
Jason Wiemels<sup>1</sup>, Jeff Kaufman<sup>3</sup>, and Willie Neiswanger<sup>1</sup>

<sup>1</sup>University of Southern California, <sup>2</sup>Prime Intellect, <sup>3</sup>Nucleic Acid Observatory

We pretrain METAGENE-1, a 7-billion-parameter autoregressive transformer model, which we refer to as a *metagenomic foundation model*, 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 METAGENE-1 is to capture the full distribution of genomic information present within this wastewater, to aid in tasks relevant to pandemic monitoring and pathogen detection. 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 performance of METAGENE-1, which achieves state-of-the-art results on a set of genomic benchmarks and new evaluations focused on human-pathogen detection and genomic sequence embedding, showcasing its potential for public health applications in pandemic monitoring, biosurveillance, and early detection of emerging health threats.

 [Website: metagene.ai](http://metagene.ai)

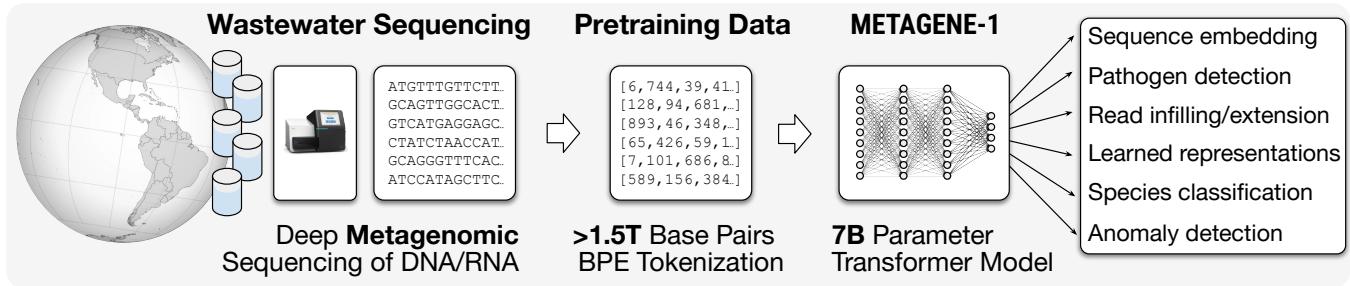
 [Model Weights: huggingface.co/metagene-ai](https://huggingface.co/metagene-ai)

 [Code Repository: github.com/metagene-ai](https://github.com/metagene-ai)

## 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 (Ji et al., 2021, Nguyen et al., 2024b, Dalla-Torre et al., 2023, Zhou et al., 2023, Fishman et al., 2023). These large genomic 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 human genomes or on curated collections of genomes from selected species (Consens et al., 2023, Benegas et al., 2024).

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 (Salomon et al., 2021, Reinhart et al., 2021). One notable example of this is the genomic monitoring of wastewater, which involves sequencing material from samples of municipal sewage (Farkas et al., 2020, Consortium, 2021). 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 (Bogler et al.,



**Figure 1: Overview of METAGENE-1 and applications.** Wastewater samples are collected and undergo deep metagenomic sequencing to generate DNA and RNA sequences totaling over 1.5 trillion base pairs. These sequences are tokenized using byte-pair encoding (BPE) to create the pretraining dataset. The data is used to train METAGENE-1, a 7B-parameter transformer model that enables a wide range of metagenomic analysis and monitoring applications.

2020, Levy et al., 2023). 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 (Medema et al., 2020, Mao et al., 2020, McClary-Gutierrez et al., 2021). 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 vast amounts of metagenomic information to monitor genomic trends, estimate the prevalence of sequences of interest, and detect new or emerging potential pathogens (Consortium, 2021, Keshaviah et al., 2021, Levy et al., 2023).

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 (Breitwieser et al., 2019, Tisza and Buck, 2021). 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 and embedding for sequenced metagenomic reads; sequence alignment, error-correction, and infilling; and human pathogen detection and taxonomic classification (Consortium, 2021).

In this paper, we take an initial step toward developing a metagenomic foundation model 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 METAGENE-1, 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 (Bragg and Tyson, 2014, Consortium, 2021). This dataset, comprising short uncurated sequences from tens of thousands of species, allows METAGENE-1 to excel at representing the complexities of microbial and viral diversity, providing unique advantages in biosurveillance applications. METAGENE-1 adopts a decoder-style language model architecture, similar to those found in the GPT and Llama families of models (Radford et al., 2019, Touvron et al., 2023), 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. An overview of METAGENE-1 data, model architecture, and applications is shown in Figure 1.

In the following sections, we first describe our metagenomic dataset and detail the tokenization strategy used to process the sequence data. We then provide comprehensive details of the METAGENE-1 model architecture and of the pretraining process on our dataset. Subsequently, we develop, and demonstrate our model’s performance, on pathogen detection and metagenomic embedding benchmarks. METAGENE-1 achieves state-of-the-art performance on these and other standard genomic evaluation tasks—designed to evaluate models trained on human and animal genomes—highlighting its generalization capabilities. As an initial demonstration of the downstream application potential, we construct an anomaly detection scenario, and show that METAGENE-1 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

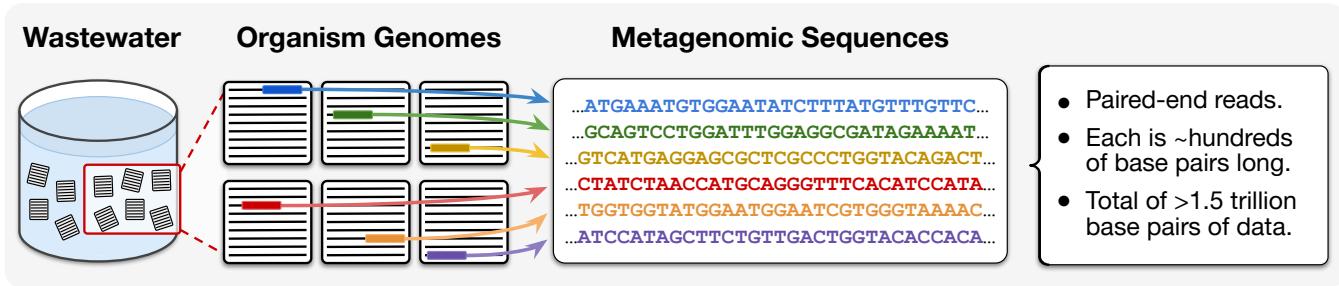
Language models trained on genomic sequences have been an area of active research, with many aiming to train on long DNA sequences from specific species, gained from publicly available sources. For instance, models such as DNABERT (Ji et al., 2021), HyenaDNA (Nguyen et al., 2024b), GROVER (Sanabria et al., 2024), and Caduceus (Schiff et al., 2024) 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. For 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 (Zhou et al., 2023), NucleotideTransformer (Dalla-Torre et al., 2023), GENA-LM (Fishman et al., 2023), SpliceBERT (Chen et al., 2023), and DNAGPT (Zhang et al., 2023) 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 (Zhou et al., 2023, Fishman et al., 2023, Sanabria et al., 2024, Zhou et al., 2024).

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 and GPT model families, which we further motivate in Sec. 3.3.

## 3. METAGENE-1: Metagenomic Foundation Model

We pretrain a 7-billion-parameter autoregressive transformer language model, referred to as METAGENE-1, 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.



**Figure 2: Overview of the metagenomic data collection and sequencing pipeline for model pretraining.** The process begins with the collection of wastewater (left), which contains genomic fragments from a diverse collection (e.g., tens of thousands) of constituent organisms (center). These samples are processed via high-throughput metagenomic sequencing to produce millions of paired-end reads (right), each consisting of hundreds of base pairs. The complete dataset comprises over 1.5 trillion base pairs of metagenomic sequences used for model pretraining.

### 3.1. Metagenomic Dataset

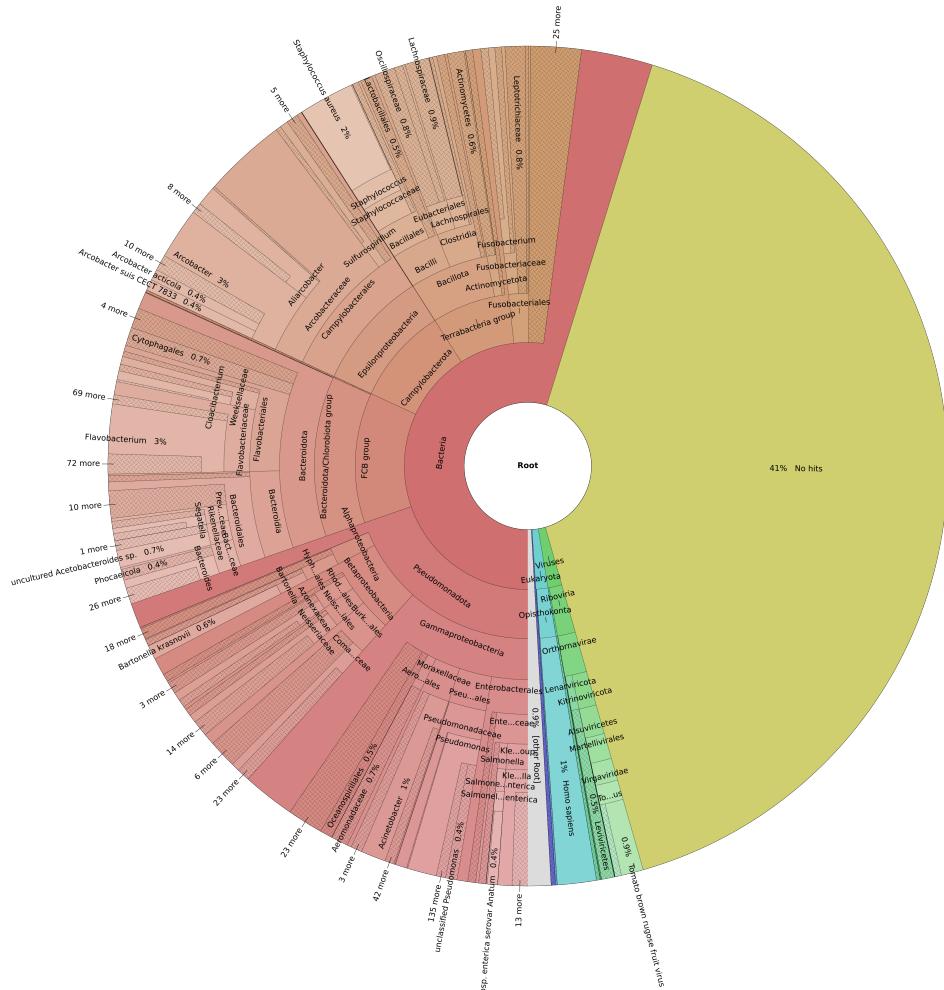
One of the goals of our metagenomic foundation model is to train on a genomic dataset that captures the immense diversity of the microbiome surrounding humans. To achieve this, we leverage a newly collected metagenomic dataset—never before used in model training—comprising material from a broad range of organisms, including bacteria, viruses, cells from human and other eukaryotes, 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 sequence models, which often focus on curated collections of specific (known) 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<sup>1</sup>. Paired-end reads can offer 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 large genomic 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. We illustrate this metagenomic data collection and sequencing pipeline in Figure 2.

This metagenomic sequence corpus was collected over a six-month period by the Nucleic Acid Observatory (NAO) (Consortium, 2021) in collaboration with partners (Marc Johnson and Clayton Rushford at the University of Missouri<sup>2</sup> and Jason Rothman in Katrine Whiteson's lab<sup>3</sup> at the University of California, Irvine). Samples of wastewater were sourced from multiple locations across the United States, in particular from cities in California and Missouri. After wastewater samples were collected, the material was filtered and nucleic acids extracted (Rothman et al., 2021, Robinson et al., 2022) before undergoing metagenomic sequencing. In full, the metagenomic dataset for pretraining comprises over 1.5 trillion base pairs. Our hope is that 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 METAGENE-1.

<sup>1</sup>Where RNA sequences are first converted into DNA via reverse transcription. <sup>2</sup><https://bondlsc.missouri.edu/person/marc-johnson>. <sup>3</sup><https://jasonrothman.weebly.com/>

We show an estimate of the metagenomic composition of this pretraining dataset in Figure 3, using the *Kraken 2* (Wood et al., 2019) sequence classification software (see Figure 7 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.



**Figure 3:** Metagenomic composition of the METAGENE-1 pretraining dataset, estimated via *Kraken 2* (Wood et al., 2019) sequence classification, and visualized via *Krona* (Ondov et al., 2011). See Figure 7 for a more-detailed view.

### 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 genomic models.

BPE offers several advantages for our model. 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.

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 ~370 billion tokens ( $\approx 1.69$  trillion base pairs) for pretraining. We give a table showing full tokenizer details, including a list of all special tokens, in Appendix B.

### 3.3. METAGENE-1 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 (Radford et al., 2019, Touvron et al., 2023). 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 METAGENE-1 stands in contrast to some of the alternative approaches explored in recent genomic models, which include BERT-style bidirectional encoders (Ji et al., 2021, Zhou et al., 2023, 2024) or non-attention based architectures (Nguyen et al., 2024b,a). Our decision to use this particular model architecture was driven by the following motivations:

- 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.
- 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.
- 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 METAGENE-1. 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

Model Details	METAGENE-1
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	$6 \times 10^{-4}$
$\beta_1, \beta_2$	0.9, 0.95

**Table 1:** METAGENE-1 architecture details.

below. We use an attention mask which prevents attention between the distinct packed sequence reads. METAGENE-1 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 METAGENE-1

### 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) (Chowdhery et al., 2022) 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.

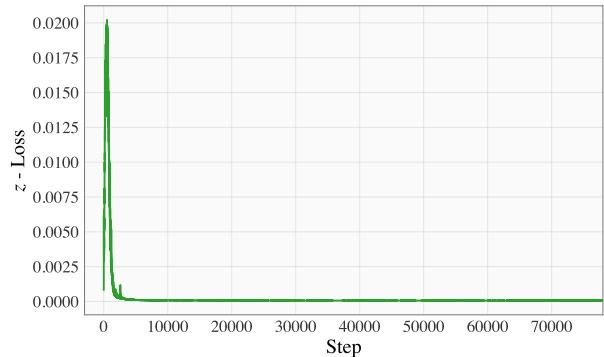
For training, we use a global batch size of 30,720, a sequence length of 512, and a micro-batch size of 48. We observe this combination to offer the best trade-off between high MFU and reduced memory usage; it also allows us to shard the optimizer state and gradients within a single node. Further tests 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.

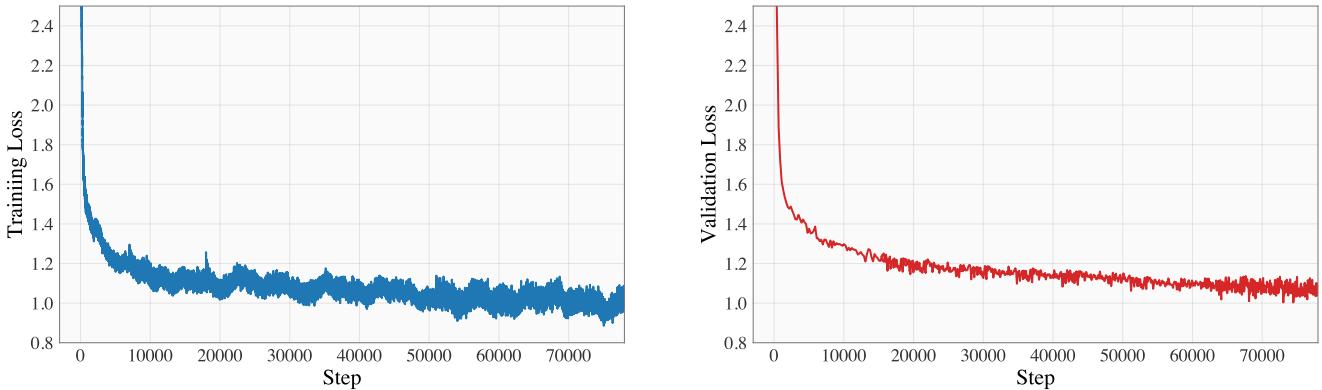
### 4.2. Stability

Foundation model pretraining is prone to suffer from training instability, which can be more pronounced when scaling models to billions of parameters (Wortsman et al., 2023). 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 (Team, 2024).

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.



**Figure 4:** We show z-loss during pretraining, which aids and gives an indicator of stability.



**Figure 5:** METAGENE-1 loss curves during pretraining. We show training loss (left), and validation loss on a held out metagenomic sample (right).

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 (Team, 2024) to preserve the Llama architecture and leverage optimized inference pipelines.

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 the training. Fortunately, no stability issues were observed, and these 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*<sup>4</sup> (Dao, 2023) which avoids materializing the full mask, which would have been inefficient.

#### 4.4. Continual Pretraining

After the initial stage of pretraining is complete, we carry out a second stage of pretraining which constitutes about 9% of our total number of pretraining tokens. In this second stage of training, we extend our dataset to a broader distribution of genomic sequences relative to our original metagenomic distribution, and we follow practices for continual learning, such as annealing the learning rate both to enact a *warmup* period (*i.e.*, a linear ramp up to account for the shifted data distribution), and a *cooldown* period (*i.e.*, a ramp down of the learning rate at the end of training for improved performance (Hägele et al., 2024)).

The modified training distribution aims to allow for us to maintain performance on metagenomic tasks, such as metagenomic embedding and classification, while also achieving improved performance on a broader set of genomic tasks (*i.e.*, tasks involving non-metagenomic data). For this, we sample sequences from the dataset provided by Zhou et al. (2023), which includes genomic sequences from known organisms—both from human genomes and a curated selection of genomes from multiple species (*e.g.*, fungi, mammalian, invertebrate, bacteria)—and shuffle it into our metagenomic reads at a 1:8 ratio.

<sup>4</sup>Named function `flash_attn_varlen_func` in the *FlashAttention-2* Python package.

## 5. Empirical Results

### 5.1. Pretraining Performance

As an initial analysis of METAGENE-1, in Figure 5, 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. Pathogen Detection Benchmark

Our initial experiments evaluate METAGENE-1’s reliability in detecting human pathogens. To this end, we construct four datasets with binary labels, aiming to classify human pathogens versus non-pathogens. These datasets are constructed from four distinct sequencing deliveries, which are excluded from our pretraining data. For each delivery, we extract two sets of sequencing reads: pathogen and non-pathogen. Pathogen reads are defined as a subset of sequencing reads meeting two criteria: (1) Kraken 2 (Wood et al., 2019)<sup>5</sup> identifies at least one hit on a  $k$ -mer associated with a human-infecting virus, and (2) the read aligns with a human-infecting virus genome in GenBank<sup>6</sup>. The sub-tasks in this pathogen detection benchmark represent different deliveries, which vary by collection location, sequencing pipeline, date, or a combination of these factors. Each dataset contains 1,600 training samples and 2,000 test samples. We intentionally use a small training set to mimic real-world scenarios where rare human pathogens are expensive to identify.

	DNABERT-2	DNABERT-S	NT-2.5b-Multi	NT-2.5b-1000g	METAGENE-1
<b>PATHOGEN-DETECT (AVG.)</b>	87.92	87.02	82.43	79.02	<b>92.96</b>
<b>PATHOGEN-DETECT-1</b>	86.73	85.43	83.80	77.52	<b>92.14</b>
<b>PATHOGEN-DETECT-2</b>	86.90	85.23	83.53	80.38	<b>90.91</b>
<b>PATHOGEN-DETECT-3</b>	88.30	89.01	82.48	79.83	<b>93.70</b>
<b>PATHOGEN-DETECT-4</b>	89.77	88.41	79.91	78.37	<b>95.10</b>

**Table 2:** Results on the Pathogen Detection benchmark. The metric used for all evaluations is MCC. The header row reports macro-averaged performance metrics. See Section 5.2 for details.

We evaluate the performance of METAGENE-1 and other genomic foundation models on the pathogen detection datasets, measured using the Matthews correlation coefficient (MCC). All models were trained with a consistent set of hyperparameters: DNABERT (Zhou et al., 2024) variants undergo full-model fine-tuning, while Nucleotide Transformer (NT) (Dalla-Torre et al., 2023) variants and METAGENE-1 are fine-tuned using low-rank adapters (LoRA) (Hu et al., 2021). For sequence-level classification, we use the built-in pooler for DNABERT and NT models provided in HuggingFace Transformers (Wolf, 2019), and use mean-pooled representations for METAGENE-1. Additional experimental details can be found in Appendix C.1.

As shown in Table 2, METAGENE-1 consistently outperforms all other models across the Pathogen Detection benchmark, with gains ranging from approximately 3 to 17 MCC points over the strongest competing models. These results highlight METAGENE-1’s strong performance in pathogen detection tasks, particularly in scenarios with diverse sequencing conditions or delivery pipelines.

<sup>5</sup>We use the 2024-06 Standard Database for identification. <sup>6</sup>We use the 2024-06 GenBank release available at <https://www.ncbi.nlm.nih.gov/genbank/>.

	DNABERT-2	DNABERT-S	NT-2.5b-Multi	NT-2.5b-1000g	METAGENE-1
<b>HUMAN-VIRUS (AVG.)</b>	0.564	0.570	0.675	0.710	<b>0.775</b>
<b>HUMAN-VIRUS-1</b>	0.594	0.605	0.671	0.721	<b>0.828</b>
<b>HUMAN-VIRUS-2</b>	0.507	0.510	0.652	0.624	<b>0.742</b>
<b>HUMAN-VIRUS-3</b>	0.606	0.612	0.758	0.740	<b>0.835</b>
<b>HUMAN-VIRUS-4</b>	0.550	0.551	0.620	<b>0.755</b>	0.697
<b>HMPD (AVG.)</b>	0.397	0.403	0.449	0.451	<b>0.465</b>
<b>HMPD-SINGLE</b>	0.292	0.293	0.285	0.292	<b>0.297</b>
<b>HMPD-DISEASE</b>	0.480	0.486	0.498	0.489	<b>0.542</b>
<b>HMPD-SEX</b>	0.366	0.367	0.487	0.476	<b>0.495</b>
<b>HMPD-SOURCE</b>	0.451	0.465	0.523	<b>0.545</b>	0.526
<b>HVR (AVG.)</b>	0.479	0.479	0.546	0.524	<b>0.550</b>
<b>HVR-p2p</b>	0.548	0.550	0.559	<b>0.650</b>	0.466
<b>HVR-s2s-align</b>	0.243	0.241	0.266	<b>0.293</b>	0.267
<b>HVR-s2s-small</b>	0.373	0.372	0.357	0.371	<b>0.467</b>
<b>HVR-s2s-tiny</b>	0.753	0.753	1.000	0.782	<b>1.000</b>
<b>HMPR (AVG.)</b>	0.347	0.351	0.348	0.403	<b>0.476</b>
<b>HMPR-p2p</b>	0.566	<b>0.580</b>	0.471	0.543	0.479
<b>HMPR-s2s-align</b>	0.127	0.129	0.144	<b>0.219</b>	0.140
<b>HMPR-s2s-small</b>	0.419	0.421	0.443	<b>0.459</b>	0.432
<b>HMPR-s2s-tiny</b>	0.274	0.274	0.332	0.391	<b>0.855</b>
<b>GLOBAL AVERAGE</b>	0.475	0.479	0.525	0.545	<b>0.590</b>

**Table 3:** Results on the Genomic Embedding (Gene-MTEB) benchmark. See Section 5.3 for details.

### 5.3. Genomic Embedding Benchmark

Next, we assess METAGENE-1’s ability to generate high-quality representations in a zero-shot manner. These representations are crucial for lightweight development of predictive models using a frozen foundation model (Devlin, 2018, Karpukhin et al., 2020, *inter alia*). They enhance interpretability by enabling sparse autoencoders to produce semantically meaningful encodings (Bricken et al., 2023, Gao et al., 2024). Additionally, they are vital for anomaly detection methods that rely on them for effective modeling (Yang et al., 2024). Drawing inspiration from MTEB (Muennighoff et al., 2022), we introduce a large-scale genomics embedding benchmark, termed Gene-MTEB, to advance the development of robust genomics representations.

For this benchmark, we curate eight classification tasks (Human-Virus-1-4, MHPD-single, HMPD-disease, HMPD-source, HMPD-sex), and eight clustering tasks (HVR-p2p, HVR-s2s-align, HVR-s2s-small, HVR-s2s-tiny, HMPR-p2p, HMPR-s2s-align, HMPR-s2s-small, HMPR-s2s-tiny). Datasets for these tasks are sourced from the Human Microbiome Project (Peterson et al., 2009), and held-out portions of our metagenomic dataset. Details and access to all benchmark datasets are provided on the project HuggingFace page. All classification tasks carry out logistic regression on top of embeddings and all clustering tasks carry out mini-batch  $k$ -means. Embeddings for all models are accessed via mean pooling on the last hidden state.

Results on Gene-MTEB are shown in Table 3. Here, *accuracy* is shown for classification and *V-measure* for clustering tasks. We find that METAGENE-1 shows strong embedding performance across the board, and in particular for Human-Virus datasets, scoring over 6 points above all other models. Continual training with representation learning objectives, such as contrastive losses, could further enhance its embedding quality beyond its current LM-based pretraining.

	CNN	HyenaDNA	DNABERT	NT-2.5B-Multi	DNABERT-2	METAGENE-1
<b>TF-MOUSE (AVG.)</b>	45.3	51.0	57.7	67.0	68.0	<b>71.4</b>
0	31.1	35.6	42.3	<b>63.3</b>	56.8	61.5
1	59.7	80.5	79.1	83.8	<b>84.8</b>	83.7
2	63.2	65.3	69.9	71.5	79.3	<b>83.0</b>
3	45.5	54.2	55.4	69.4	66.5	<b>82.2</b>
4	27.2	19.2	42.0	47.1	<b>52.7</b>	46.6
<b>TF-HUMAN (AVG.)</b>	50.7	56.0	64.4	62.6	<b>70.1</b>	68.3
0	54.0	62.3	68.0	66.6	<b>72.0</b>	68.9
1	63.2	67.9	70.9	66.6	<b>76.1</b>	70.8
2	45.2	46.9	60.5	58.7	<b>66.5</b>	65.9
3	29.8	41.8	53.0	51.7	<b>58.5</b>	58.1
4	61.5	61.2	69.8	69.3	77.4	<b>77.9</b>
<b>EMP (AVG.)</b>	37.6	44.9	49.5	58.1	56.0	<b>66.0</b>
H3	61.5	67.2	74.2	78.8	78.3	<b>80.2</b>
H3K14AC	29.7	32.0	42.1	56.2	52.6	<b>64.9</b>
H3K36ME3	38.6	48.3	48.5	62.0	56.9	<b>66.7</b>
H3K4ME1	26.1	35.8	43.0	55.3	50.5	<b>55.3</b>
H3K4ME2	25.8	25.8	31.3	36.5	31.1	<b>51.2</b>
H3K4ME3	20.5	23.1	28.9	40.3	36.3	<b>58.5</b>
H3K79ME3	46.3	54.1	60.1	64.7	67.4	<b>73.0</b>
H3K9AC	40.0	50.8	50.5	56.0	55.6	<b>65.5</b>
H4	62.3	73.7	78.3	81.7	80.7	<b>82.7</b>
H4AC	25.5	38.4	38.6	49.1	50.4	<b>61.7</b>
<b>PD (AVG.)</b>	77.1	35.0	84.6	<b>88.1</b>	84.2	82.3
ALL	75.8	47.4	90.4	<b>91.0</b>	86.8	86.0
No-TATA	85.1	52.2	93.6	94.0	<b>94.3</b>	93.7
TATA	70.3	5.3	69.8	<b>79.4</b>	71.6	67.4
<b>CPD (AVG.)</b>	62.5	48.4	<b>73.0</b>	71.6	70.5	69.9
ALL	58.1	37.0	<b>70.9</b>	70.3	69.4	66.4
No-TATA	60.1	35.4	69.8	<b>71.6</b>	68.0	68.3
TATA	69.3	72.9	<b>78.2</b>	73.0	74.2	75.1
<b>SSD</b>	76.8	72.7	84.1	<b>89.3</b>	85.0	87.8
<b>COVID</b>	22.2	23.3	62.2	<b>73.0</b>	71.9	72.5
<b>GLOBAL WIN %</b>	0.0	0.0	7.1	21.4	25.0	<b>46.4</b>

**Table 4:** Results on the Genome Understanding Evaluation (GUE) benchmark. Non-METAGENE-1 results are adapted from Zhou et al. (2023). The metric used for all evaluations is MCC, except for the COVID task, which uses F1 score. The header rows report macro-averaged performance metrics. The final row shows *Global Win %*, i.e., the percentage of tasks in which a given method achieves top score under the associated metric.

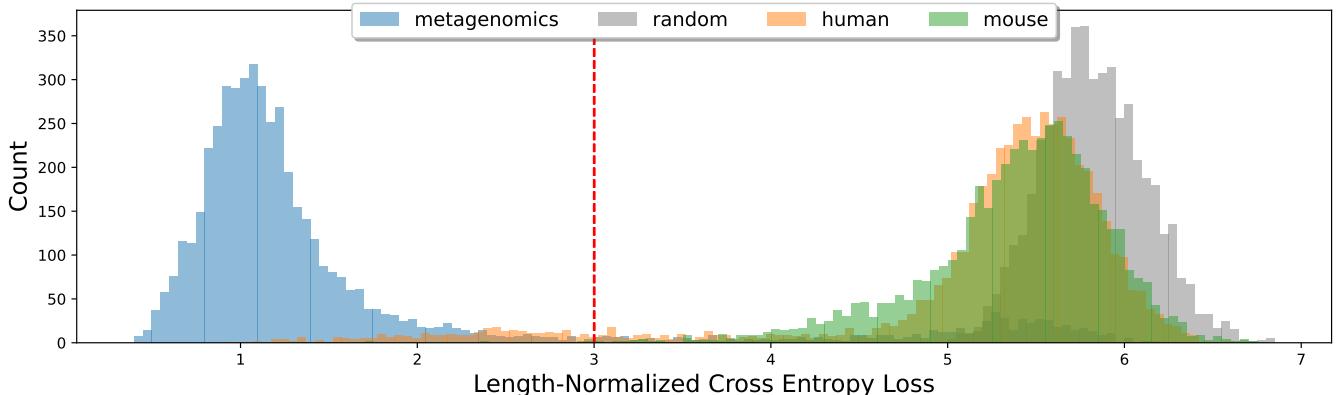
#### 5.4. Genome Understanding Evaluation Benchmark

We now investigate the viability of METAGENE-1 as a general-purpose foundation model. Importantly, we aim to assess its performance on nucleotide sequences sampled from a diverse array of species. One such example is long-sequence full-animal-genome datasets. In many prior genomic sequence models' pretraining datasets, this type of genomic data is found in abundance (Dalla-Torre et al., 2023, Ji et al., 2021, Nguyen et al., 2024b,

Zhou et al., 2023). As a pilot study, we perform fine-tuning experiments on the Genome Understanding Evaluation (GUE) benchmark (Zhou et al., 2023), which comprises 28 sequence-level classification tasks curated from this type of genomics data.

Following Section 5.2, we fine-tune low-rank adapters (LoRA) (Hu et al., 2021) 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 perform a grid search over linearly spaced learning rates from  $1e-4$  to  $1e-3$  and select LoRA modules from query-value and query-key-value-dense combinations. We fix all other hyperparameters and select the best configuration based on validation performances. Additional details on training hyperparameters can be found in Appendix C.2. Following the metrics selected in Zhou et al., we report Matthews correlation coefficient (MCC) on all but the COVID task, which instead uses the F1 score.

In Table 4, we present METAGENE-1’s performance on the GUE benchmark. Our findings show that METAGENE-1 outperforms or remains competitive with state-of-the-art foundation models specializing in multi-species genomics prediction, achieving a top score on 13 out of 28 GUE subtasks (compared with DNABERT-2, the second highest scoring model, that achieves a top score on 7 out of 28 subtasks). Notably, METAGENE-1 excels in Epigenetic Marks Prediction (EMP) tasks but shows room for improvement in (Core) Promoter Detection (PD/CPD). We attribute this to limitations in the pre-training data mixture, and believe that a more tailored pre-training dataset could potentially enhance METAGENE-1’s performance in this area.



**Figure 6:** Distribution of the length-normalized cross entropy loss across all datasets, given by METAGENE-1.

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 5:** OOD detection performance between metagenomics sequences and other data sources.

### 5.5. Anomaly Detection from Wastewater

Our final experiment aims to show the feasibility of METAGENE-1 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 sample 5000 sequences from, respectively, our metagenomics pretraining data, the mouse and human genomes from the GUE dataset, as well as *uniform 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 5. Note that this anomaly detection study is performed using a checkpoint of METAGENE-1 that has only been pretrained on metagenomic data (*i.e.*, without second-stage training).

Figure 6 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 5 reports numerical results of our OOD detection tests. METAGENE-1 achieves strong performance for separating metagenomics sequences from other data sources.

## 6. Safety Considerations

Metagenomic foundation models like METAGENE-1 demonstrate improved capabilities on tasks that can aid in biosurveillance, genomic anomaly detection, and pandemic monitoring. While still relatively small in scale compared with many modern language models, METAGENE-1 shows state-of-the-art results on benchmarks and enables potential downstream uses. However, these capabilities merit careful consideration of safety and must be balanced against potential risks. This category of genomic model—and especially, future larger variants of it—could pose risks to human health and safety by enabling harmful applications, such as the design of novel pathogenic DNA sequences or synthetic genetic materials. These potential abuses were considered when deciding to open source METAGENE-1. The final decision was based on weighing the beneficial applications, such as pandemic preparedness, against the potential for misuse. Based on our safety considerations, which we outline below, we believe that the current iteration of METAGENE-1 poses minimal risk, and its release is justified by its significant positive potential. However, we also recognize and discuss the need for careful safety considerations before open sourcing increasingly capable models of this type.

**Relation to other open source genomic models.** METAGENE-1 is a genomic foundation model that builds upon a lineage of similar open-source efforts, such as NucleotideTransformer (Dalla-Torre et al., 2023), DNABERT (Ji et al., 2021), HyenaDNA (Nguyen et al., 2024b), Evo (Nguyen et al., 2024a), and more. At 7 billion parameters, METAGENE-1 matches the largest of these existing models. The key distinction of METAGENE-1 lies in the model’s training data: a highly diverse set of metagenomic sequences derived from wastewater, with a focus on the human microbiome. This dataset, comprising short uncurated sequences from tens of thousands of species, allows METAGENE-1 to excel at representing the complexities of microbial and viral diversity in metagenomic samples, providing unique advantages in biosurveillance applications. Similar to other genomic foundation models, and unlike large language models, these models alone do not possess significant reasoning or control capabilities (given that complex control instructions cannot easily be provided via input context, which is restricted to genomic sequences).

**Tailored for detection, not design.** METAGENE-1 was specifically designed for anomaly detection in metagenomic data, not for complex genomic design tasks. The training data, model architecture, and task design are geared toward detecting and classifying anomalies in short sequences of a few hundred base pairs. Notably, all metagenomic data used in pretraining METAGENE-1 consist exclusively of sequences ranging from 100 to 300 base pairs. Unlike large genomic models focused on longer sequence generation, METAGENE-1’s capabilities are tailored to analyzing these short metagenomic reads. Its architectural constraints, including a

maximum context length of 512 tokens, further limit its applicability to sequence design tasks. These design decisions ensure that the model’s primary utility lies in detecting pathogens and monitoring biosurveillance trends, rather than enabling misuse in synthetic biology.

**Pros and cons of open source.** Open sourcing a model of this type is a balance between the potential for help and harm. In the case of METAGENE-1, we believe that open source is net positive for research in the area of anomaly detection for pathogen monitoring. We hope that the availability of this model can have a positive impact on facilitating safety research, a prospect that we discuss in Section 7. Nonetheless, we recognize the importance of caution when releasing models in this domain. For future iterations of pathogen-detection models with improved capabilities, we believe strongly in (and we ourselves are committed to) thoroughly evaluating the safety and potential for misuse before an open source release. Larger-scale models, in particular, present additional risks, and we advocate for rigorous safety assessments in determining whether such models should be released publicly. By prioritizing careful oversight and responsible scaling, we aim to mitigate risks while maximizing the benefits of this technology for public health and biosurveillance.

## 7. Discussion, Limitations, Conclusion

We have reported our current progress on pretraining and evaluating METAGENE-1, the first large-scale foundation model pretrained on metagenomic sequences. We detail our dataset construction, model training, and fine-tuning procedure to facilitate open-science research. Additionally, we open-source our training code and model checkpoints.

Our downstream performance on genomic benchmarks indicates the potential of METAGENE-1 as a general-purpose foundation model. Our results also indicate that METAGENE-1 benefits from continual pretraining on a diverse mixture of data sources in addition to metagenomic data (at least for tasks similar to these genomic benchmarks). We are continuing to actively explore this direction, through incorporating additional human reference genomes and multi-species genomic datasets in our metagenomic pretraining data.

**Limitations.** METAGENE-1 is pretrained on a dataset consisting primarily of wastewater metagenomics and multi-species genomic sequences, making it well-suited for downstream tasks within this distribution. However, like many foundation models, it requires additional fine-tuning to achieve optimal performance for specific applications. Additionally, the pretraining data predominantly consist of short metagenomic sequencing reads, limiting the model’s performance to contexts involving shorter metagenomics inputs. This may restrict its effectiveness for tasks involving long-read or full-genome data, where long-sequence models may be necessary (Nguyen et al., 2024a,b).

**Future directions.** There are many potential avenues for future research. An area that we are particularly excited about concerns the *understanding* of genomic foundation models. While a great deal of prior work has studied the mechanistic interpretability of language models (Wang et al., 2022, Hanna et al., 2024, Conmy et al., 2023, Syed et al., 2023), their extensions beyond language and vision have been limited. Future work could systematize approaches to mechanistic interpretability in genomics by leveraging sparse autoencoders (SAEs) (Bricken et al., 2023, Gao et al., 2024, Lieberum et al., 2024) to identify biologically meaningful features, employing attribution methods to trace model predictions to genomic regions (Koo and Ploenzke, 2020, Tseng et al., 2020, Majdandzic et al., 2022), and developing new tools for probing model representations using task-specific datasets (Conneau et al., 2018, Hewitt and Liang, 2019). A better understanding of these models would not only advance their reliability but also help mitigate risks, such as inadvertently generating or propagating harmful genomic sequences.

Finally, 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 METAGENE-1 and future metagenomic models, and we invite both domain experts and the machine learning community to contribute to this research.

**Acknowledgements** We thank Prime Intellect for computing support, and the Nucleic Acid Observatory for metagenomic data resources. In particular, we thank Marc Johnson, Clayton Rushford, and Jason Rothman for metagenomic sequencing support that produced the data for this project. W.N. would like to thank Victor Miller and Mark Schulze for helpful discussions and feedback on this project. O.L. would like to thank Zhihan Zhou for helpful discussions and insights on the GUE benchmark.

## References

- 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.
- 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.
- Lauren Bragg and Gene W Tyson. Metagenomics using next-generation sequencing. *Environmental microbiology: methods and protocols*, pages 183–201, 2014.
- 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.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- 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.
- 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>.
- Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.
- 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.
- The Nucleic Acid Observatory Consortium. A global nucleic acid observatory for biodefense and planetary health. *arXiv preprint arXiv:2108.02678*, 2021.
- 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.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.
- Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- 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.
- 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.
- Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.
- Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *arXiv preprint arXiv:2405.18392*, 2024.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

- 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.
- 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.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.
- 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.
- Peter K Koo and Matt Ploenzke. Interpreting deep neural networks beyond attribution methods: quantifying global importance of genomic features. *BioRxiv*, pages 2020–02, 2020.
- Joshua I Levy, Kristian G Andersen, Rob Knight, and Smruthi Karthikeyan. Wastewater surveillance for public health. *Science*, 379(6627):26–27, 2023.
- Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.
- Antonio Majdandzic, Chandana Rajesh, Ziqi Tang, Shushan Toneyan, Ethan L Labelson, Rohit K Tripathy, and Peter K Koo. Selecting deep neural networks that yield consistent attribution-based interpretations for genomics. In *Machine Learning in Computational Biology*, pages 131–149. PMLR, 2022.
- 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.
- 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.
- 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.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.
- 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, 2024a.
- 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, 2024b.

Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12:1–10, 2011.

Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

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.

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.

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.

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.

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.

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.

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.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

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

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.

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, Wenyan 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, Bin 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>.

Alex Tseng, Avanti Shrikumar, and Anshul Kundaje. Fourier-transform-based attribution priors improve the interpretability and stability of deep learning models for genomics. *Advances in Neural Information Processing Systems*, 33:1913–1923, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

T Wolf. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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

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>.

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>.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, et al. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*, 2024.

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.

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.

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 Pretraining Dataset

In Figure 7, 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 (Wood et al., 2019), 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 7: first, the full pretraining dataset distribution (top); then, an example subset of this showing the distribution of viruses (middle); and finally, an example subset of this showing the distribution of the *Steitzviridae* family of viruses (bottom).

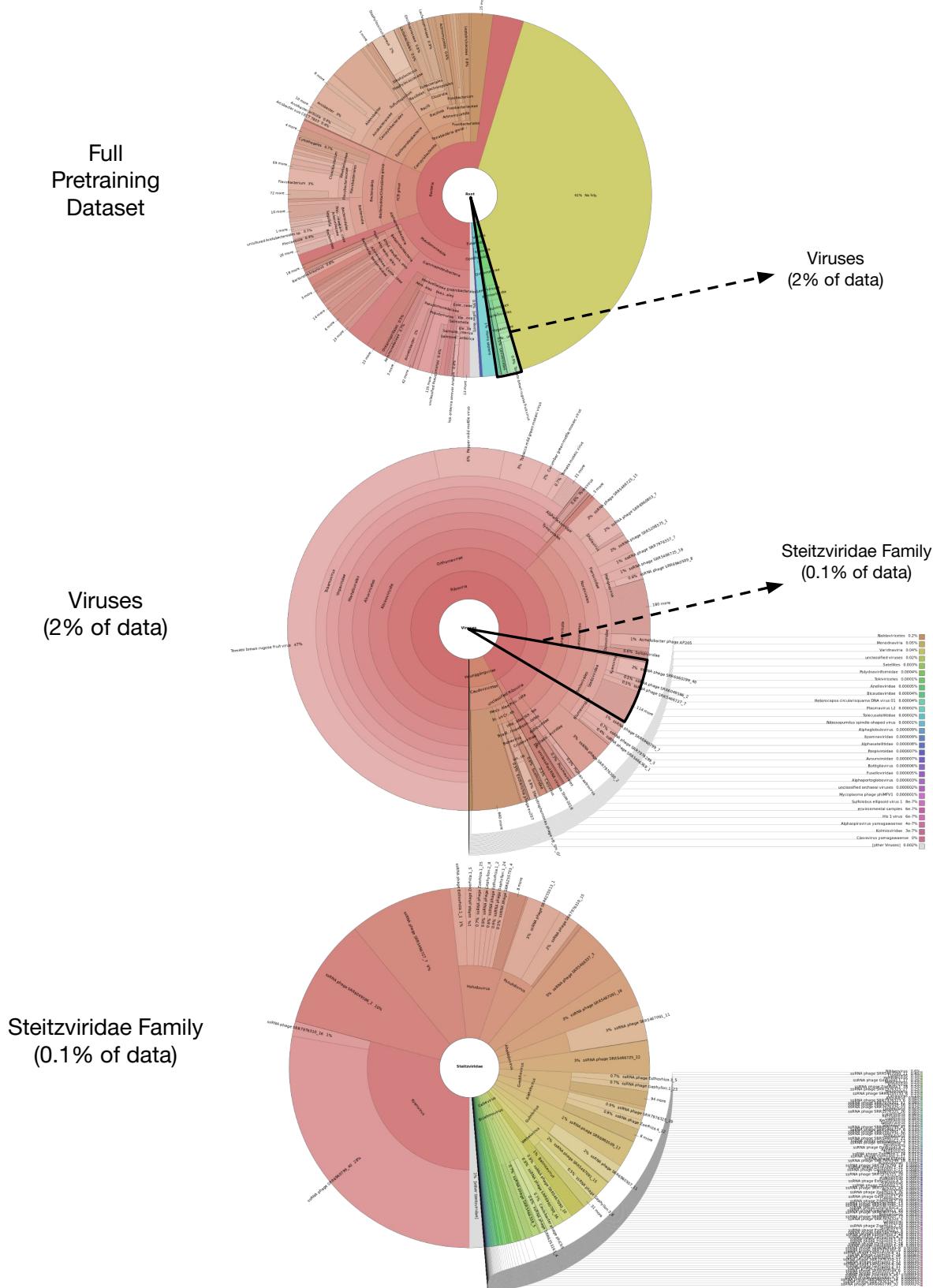
### B. Tokenizer Details

Our tokenizer implementation is adapted from `minbpe`<sup>7</sup>. 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 also experimented with a larger vocabulary size of 4096, but due to length characteristics of our metagenomic data, this design choice results in many short tokenized sequences that may not be able to provide meaningful learning signal. We thus decided to move forward with a vocabulary size of 1024 to balance efficiency and downstream performance.

---

<sup>7</sup><https://github.com/karpathy/minbpe>



**Figure 7:** A visualization of the composition of metagenomic information contained in our pretraining dataset, based on Kraken 2 metagenomic sequence classification hits (Wood et al., 2019). 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).

## C. Additional Experimental Details

### C.1. Additional Details for the Pathogen Detection Benchmark

In Table 6, we show our choices of hyperparameters for fine-tuning experiments.

	DNABERT-★	Full Model
NT-★		LoRA
METAGENE-1		LoRA
LoRA Modules		query, key, value, dense
LoRA Rank		8
LoRA $\alpha$		16
LoRA Dropout		0.1
Optimizer		AdamW
Optimizer Momentum		$\beta_1, \beta_2 = 0.9, 0.999$
Learning Rate		1e-4 <sup>A</sup>
LR Scheduler		Linear Warmup + Constant LR
Warmup Steps		50
Weight Decay		0.01
Denominator $\epsilon$		1e-8
Precision		BF16-mixed
Batch Size		32
Epochs		10
Hardware		NVIDIA A100 80GB

**Table 6:** Hyperparameter settings for the Pathogen Detection fine-tuning experiments. A: for DNABERT-S, we halve the learning rate to 5e-5 as we observe clear oscillation behavior in the training loss.

## C.2. Additional Details for the GUE Benchmark

In Table 7, we show our choices of hyperparameters for fine-tuning experiments.

LoRA Modules	query, key, value, dense <sup>Λ</sup>
LoRA Rank	8
LoRA $\alpha$	16
LoRA Dropout	0.1
Optimizer	AdamW
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Learning Rate	$\{1e-4 \dots 1e-3\}^{\Omega}$
LR Scheduler	Linear Warmup + Constant LR
Warmup Steps	50
Weight Decay	0.01
Denominator $\epsilon$	1e-8
Precision	BF16-mixed
Batch Size	32
Epochs	10
Hardware	NVIDIA A100 80GB

**Table 7:** Hyperparameter settings for the GUE fine-tuning experiments.  $\Lambda$ : LoRA is applied to query-value or query-key-value-dense modules.  $\Omega$ : learning rates are tuned over a equally-spaced grid of  $1e-4, 2e-4, \dots, 1e-3$ . All hyperparameters are selected according to performances on validation sets.