FGBERT: Function-Driven Pre-trained Gene Language Model for Metagenomics

Anonymous submission

Abstract

Metagenomic data, comprising mixed multi-species genomes, are prevalent in diverse environments like oceans and soils, significantly impacting human health and ecological functions. However, current research relies on K-mer, which limits the capture of structurally and functionally relevant gene contexts. Moreover, these approaches struggle with encoding biologically meaningful genes and fail to address the One-to-Many and Many-to-One relationships inherent in metagenomic data. To overcome these challenges, we introduce FGBERT, a novel metagenomic pre-trained model that employs a protein-based gene representation as a context-aware and structure-relevant tokenizer. FGBERT incorporates Masked Gene Modeling (MGM) to enhance the understanding of inter-gene contextual relationships and Triplet Enhanced Metagenomic Contrastive Learning (TMC) to elucidate gene sequence-function relationships. Pre-trained on over 100 million metagenomic sequences, FGBERT demonstrates superior performance on metagenomic datasets at four levels, spanning gene, functional, bacterial, and environmental levels and ranging from 1k to 213k input sequences. Case studies of ATP Synthase and Gene Operons highlight FGBERT's capability for functional recognition and its biological relevance in metagenomic research.

Introduction

Metagenomics, the study of mixed genomes of microbial communities in the environment (e.g. gut microbiomes or soil ecosystems) (Mande, Mohammed, and Ghosh 2012; Mathieu et al. 2022), has revealed the critical role in fundamental biological processes like enzyme synthesis, gene expression regulation, and immune function (Pavlopoulos et al. 2023). This deepened understanding highlights the need to accurately interpret the intricate genetic information contained within these diverse communities. Consequently, deciphering the complex sequences of multiple species in metagenomics is vital for unravelling life's mechanisms and advancing biotechnology (Albertsen 2023; Lin et al. 2023).

Unlike traditional genomics focused on single species, metagenomics involves genetic material directly from environmental samples, posing significant challenges due to sample diversity and species abundance (Lu et al. 2022). As shown in Fig. 1, the typical challenges in metagenomics are the presence of One-to-Many (OTM) and Many-to-One (MTO) problems. The OTM problem indicates that a single gene can exhibit various functions in different genomic contexts, underscoring the significance of inter-gene interactions in function regulation (Yang et al. 2021). For example, ATP synthase displays distinct functionalities in diverse organisms such as bacteria, plants, and humans (Fig. 1a). Conversely, the MTO problem implies that different genes can share the same function, emphasizing expression commonality (Al-Shayeb et al. 2022). For example, the CRISPR immune mechanism involves various proteins like Cpf1, Cas1, and Cas13, each contributing to the same defensive function (Fig. 1b) (Yang et al. 2021; Al-Shayeb et al. 2022; Hu et al. 2022).

Recently, various computational methods have emerged for genomic and metagenomic data analysis. However, these methods still face challenges when analyzing metagenomic data. Firstly, the **Semantic Tokenizer** problem. Most machine learning-based methods (Hoarfrost et al. 2022; Liang et al. 2020; Miller, Stern, and Burstein 2022), taking K-mer counts, frequencies or embeddings as input features, providing efficient alternatives to traditional homology searches against reference genome databases. However, K-mer features often have limited representation ability and fail to capture global information. Secondly, the **Function-Driven Modeling** problem. Although recent Transformer models excel in modeling complex DNA contexts through long-range dependencies, they are predominantly tailored for single-species genomic analysis (Wolf et al. 2020; Zhou et al. 2023; Dalla-Torre et al. 2023) and do not adequately address OTM and MTO challenges. This limitation impedes their ability to accurately model the intricate relationships between genes, their function across diverse genomic environments, and their connections among sequences with similar functions. Thirdly, the **Low Generalization** problem. Models like MetaTransformer (Wichmann et al. 2023) and ViBE (Gwak and Rho 2022), designed for specific tasks such as read classification and virus category prediction, fail to grasp the broader biological complexities of metagenomic data, limiting their generalization across diverse metagenomic tasks.

To address these challenges, we propose FGBERT, a novel metagenomic pre-trained model designed to encode contextually-aware and functionally relevant representations of metagenomic sequences. First, to solve the problem of encoding gene sequences with biological meaning, we propose a protein-based gene representation as a context-aware tokenizer, allowing for a flexible token vocabulary for longer

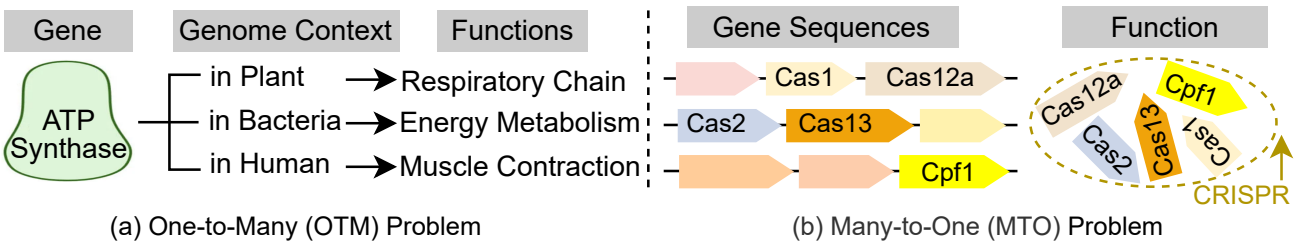


Figure 1: **Motivaion.** Two types of complex relationships between gene sequences and functions in metagenomics. **One-to-Many** problem means that the same gene may display different functions based on the genomic context; for example, ATP synthase works differently in plants, heterotrophic bacteria, and humans. **Many-to-One** problem shows that multiple genes may perform the same function; for instance, different genes from different bacteria, e.g., Cpf1, Cas1, etc., produce the same resistance function within the immune system CRISPR.

metagenomic sequences. This strategy leverages the inherent protein functional and structural information encoded in metagenomic data (Pavlopoulos et al. 2023), overcoming the limitations of K-mer methods and maintaining functional consistency despite potential mutations (D'Onofrio and Abel 2014). Second, we propose two pre-training tasks for function-driven modeling: Masked Gene Modeling (MGM) and Triplet Enhanced Metagenomic Contrastive Learning (TMC) to enhance the co-representation learning of metagenomic gene sequences and functions. Thirdly, FGBERT is pre-trained on over 100 million sequences, showcasing robust performance across diverse datasets spanning gene, functional, bacterial, and environmental levels.

Contributions. In this work, we identify three key challenges in metagenomic analysis. To address these issues, we propose FGBERT. To the best of our knowledge, this is the first metagenomic pre-trained model encoding context-aware and function-relevant representations of metagenomic sequences. To summarize: (1) We introduce a new idea of protein-based gene representations to learn biologically meaningful tokenization of long sequences. (2) We propose MGM to model inter-gene relationships and TMC to learn complex relationships between gene sequences and functions. (3) We conduct extensive experiments across various downstream tasks, spanning gene, functional, bacterial, and environmental levels with input sizes from 1k to 213k sequences. FGBERT achieves SOTA performance.

Related Works

Research on Metagenomics. Traditional alignment-based methods like MetaPhlAn5 (Segata et al. 2012) aim to match similarities between query sequences and known reference genomes and are common for taxonomic profiling. Advancements in deep learning have led to new methods like CNN-MGP (Al-Ajlan and El Allali 2019) and DeepVirFinder (Ren et al. 2020), which use CNNs for gene and viral classifications with one-hot encoding. K-mer tokenization (Fiannaca et al. 2018), employed in approaches like MDL4Microbiome (Lee and Rho 2022), is a standard for DNA sequence characterization. Additionally, Virtifier (Miao et al. 2022) maps a nucleotide sequence using a codon dictionary combined with LSTM to predict viral genes. DeepMicrobes (Liang et al. 2020) employs a self-attention mechanism, while DeepTE (Yan, Bombarely, and Li 2020) uses K-mer inputs with CNNs for element classification, and Genomic-nlp (Miller, Stern, and Burstein 2022) applies word2vec for gene function analysis. MetaTransformer (Wichmann et al. 2023) uses K-mer embedding for species classification with Transformer. For pre-training models, LookingGlass (Hoarfrost et al. 2022) uses a three-layer LSTM model for functional prediction in short DNA reads. ViBE (Gwak and Rho 2022) employs a K-mer token-based BERT model pre-trained with Masked Language Modeling for virus identification.

Pre-Training on Genomics. The BERT model, effective in DNA sequence characterization, is limited by the Transformer architecture's computational burden. LOGO (Yang et al. 2022) addresses this by cutting off long sequences into 1-2kb sub-sequences. Enformer (Avsec et al. 2021) combines extended convolution with Transformers for long human genomic data. GenSLMs (Zvyagin et al. 2022) introduce hierarchical language models for whole-genome modelling. DNABERT (Ji et al. 2021), the first pre-trained model on the human genome that focuses on extracting efficient genomic representations. DNABERT2 (Zhou et al. 2023), its successor, uses Byte Pair Encoding on multi-species genomic data. NT (Dalla-Torre et al. 2023) is trained on nucleotide sequences from humans and other species and evaluated on 18 genome prediction tasks. HyenaDNA (Nguyen et al. 2023) presents a long-range genomic model based on single-nucleotide polymorphisms on human reference genomes.

Methods

In this section, we provide a detailed description of the proposed pre-training model FGBERT, which contains the MGM and TMC components as depicted in Fig. 2.

Notation

Given a dataset of metagenomic long sequences $\{\mathcal{X}_i\}_{i=1}^{m}$, we simplify each $\mathcal{X}_i$ into a set of shorter gene sequences $\{x_i\}_{i=1}^{n_i}$ using ultrasonic fragmentation and assembly techniques (Kusters et al. 1993), where n_i represents the variable number of gene sequences derived from each $\mathcal{X}_i$. Each gene sequence x_i is tokenized in a vector $g_i \in \mathbb{R}^d$, where

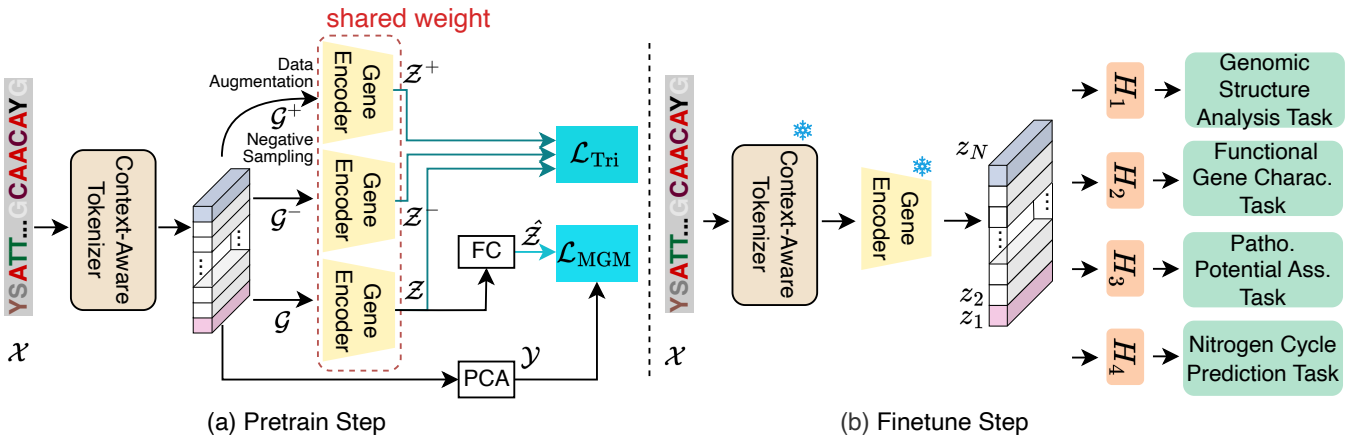


Figure 2: Overview of FGBERT. A metagenomic sequence $\mathcal{X}$ is converted into ordered protein-based gene representations $\mathcal{G}$ via a Context-Aware Tokenizer. Next, we pre-train a Gene Encoder with $\mathcal{L}_{\text{MGM}}$, 15% of these tokens are masked to predict labels $\mathcal{Y}$. Meanwhile, we introduce $\mathcal{L}_{\text{Tri}}$ to distinguish gene sequences. The data augmentation and negative sampling modules generate positive $\mathcal{G}^+$ and negative samples $\mathcal{G}^-$, respectively. Finally, after fine-tuning, FGBERT can handle various downstream tasks.

d is the token dimension. Each sequence is associated with a reduced-dimensional representation $y_i \in \mathbb{R}^{100}$. Suppose a gene group $\mathcal{G} = \{g_i\}_{i=1}^{N}$ is formed by concatenating N gene vectors sequentially. During the pre-training phase, each gene token $g_i \in \mathcal{G}$ is processed by the context-aware genome language encoder $\mathcal{F}(\cdot)$, generating the knowledge representations z_i, where $z_i = \mathcal{F}(g_i)$. These representations are then transformed by a fully connected layer into $\hat{z}_i$, defined as $\hat{z}_i = \mathcal{H}(z_i)$, where $\mathcal{H}(\cdot)$ represents a multi-classification head. In addition, we incorporate contrastive learning into the methodology. For each gene x_i, we introduce a data augmentation module to generate positive samples $x_{j(i)}$ and a hard negative sampling strategy for constructing negative samples $x_{k(i)}$.

Context-aware Masked Gene Modeling (MGM) for One-to-Many problem

Context-Aware Tokenizer. To develop a biologically meaningful tokenization of long sequences, we design a context-aware tokenizer utilizing the Protein Language Model (PLM), such as ESM-2 (Lin et al. 2022) with 15 billion parameters, integrating biological prior knowledge. As illustrated in Fig. 3, this tokenizer framework begins by extracting DNA gene sequences $\{x_i\}_{i=1}^{n_i}$ from a metagenomic sequence $\mathcal{X}_i$ utilizing the European Nucleotide Archive (ENA) software (Gruenstaeudl 2020). This conversion enhances the flexibility of analyzing longer metagenomic sequences.

Secondly, these DNA sequences x_i are translated into Amino Acid (AA) sequences using the Transeq software (McWilliam et al. 2013). This translation addresses the issue of degenerate DNA codes, where certain non-'ATCG' symbols like 'Y' or 'S' can represent multiple nucleotides (e.g., 'Y' can be 'C' or 'G'). By translating to AA sequences, we eliminate this redundancy, as different DNA sequences can map to the same AA sequence, ensuring consistency in representing biologically equivalent sequences (Lawson et al. 2004; Jain et al. 2017). Thirdly, these AA sequences are transformed into 1280D normalized ESM-2 representations, with an additional gene orientation vector, resulting in 1281D gene representations $\{g_i\}_{i=1}^{n_i}$. The utility of ESM-2 lies in its ability to reveal gene relationships and functional information inherent in metagenomic data (Pavlopoulos et al. 2023), preserving important intra-gene contextually-aware information. Finally, these representations every N representations are concatenated sequentially to form gene groups $\mathcal{G}$, which serve as the basis for subsequent modeling tasks.

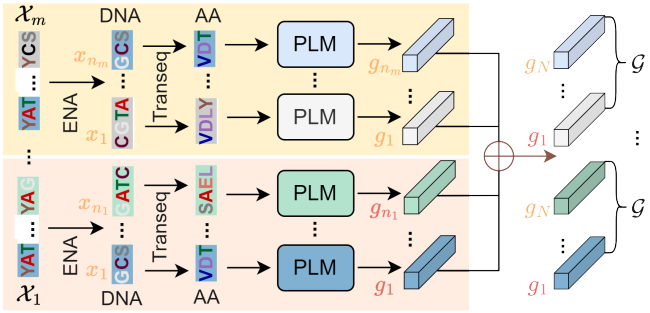


Figure 3: Framework of our Context-Aware Tokenizer.

Masked Gene Modeling. We propose the MGM to enhance the model's understanding of the relationships between genes within metagenomic sequences and their function regulations across diverse genomic environments (OTM problem). During pre-training, each gene token is masked with a 15% probability and predicted based on its unmasked genome context $\mathcal{G}_{/M}$:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{g\sim\mathcal{G}}\mathbb{E}_M \sum_{i\in M} -\log p(g_i|\mathcal{G}_{/M}), \qquad (1)$$

where M denotes the index set of the masked gene.

In addition, considering genetic polymorphism (Pastinen,

Ge, and Hudson 2006; Zhang et al. 2021), the MGM component focuses on detecting multiple genes that could coexist at a single genomic site, denoted as $\hat{z}_i = [\hat{z}_1, \hat{z}_2, \hat{z}_3, \hat{z}_4]$. This requires the model not only to predict individual genes but also to identify various combinations of genes occurring within the same site. Thus, we enhance the model with a comprehensive loss function, $\mathcal{L}_{\text{MGM}}$, which incorporates feature reconstruction and probability prediction:

$$\mathcal{L}_{\text{MGM}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{y_i^T \hat{z}_i}{\|y_i\| \cdot \|\hat{z}_i\|}\right)^{\gamma} + \frac{\alpha}{N}\sum_{i=1}^{N}||\tilde{z}_i - \tilde{y}_i||_2^2, \quad (2)$$

where γ is a reconstruction loss with the scaled cosine error, and α is a weighting factor to balance the importance of the two loss functions. The first item, Feature Reconstruction Loss (FRL), quantifies the distance between the model prediction $\hat{z}_i$ and its corresponding label y_i. The second item, Probability Prediction Loss (PPL), evaluates the discrepancy between the predicted embedding probability $\tilde{z}_i = \frac{e^{\hat{z}_i}}{\sum_{j=1}^{C} e^{\hat{z}_j}}$ and the true category probability $\tilde{y}_i = \frac{e^{y_i}}{\sum_{j=1}^{C} e^{y_j}}$, both processed via the softmax function. C denotes the number of gene combinations and is set to 4.

Triplet Metagenomic Contrastive Framework (TMC) for Many-to-One problem

Contrastive Learning. The MGM component enhances the model's ability to learn contextual relationships between genes, helping to alleviate the OTM problem present in metagenomic data. However, when faced with the MTO problem, the model's ability to describe the relationships between sequences with the same function remains weak. For instance, when annotating Enzyme Commission (EC) numbers for partial metagenomic sequences, we observe that sequences with the same EC number are close to each other in feature space, while sequences with different EC numbers are farther apart (Jumper et al. 2021). Therefore, we introduce a contrastive learning technique to capture the functional relationships between gene classes, enabling different genes with similar functions to cluster together and further optimize model training. Generally speaking, the objective of contrastive learning is to learn an embedding function $\mathcal{F}$ such that the distance between positive pairs is smaller than the distance between negative pairs:

$$d(\mathcal{F}(x_a), \mathcal{F}(x_p)) < d(\mathcal{F}(x_a), \mathcal{F}(x_n)), \quad (3)$$

where $d(\cdot, \cdot)$ is a distance function (e.g., Euclidean distance) defined on the embedding space. We adopt SupCon-Hard loss (Khosla et al. 2020) to consider multiple positive and negative samples for each anchor, encouraging the model to mine difficult samples and enhancing its robustness. Additionally, data augmentation and negative sampling modules are included to create positive and negative samples, further improving the model's capacity to recognize commonalities among gene classes.

Positives Sampling. The strategy for sampling triplets is crucial to learn a well-organized embedding space. For each gene group $\mathcal{G}$, as an anchor gene x_i within a gene batch I, a mutation strategy is proposed to augment orphan sequences (i.e., functions associated with individual sequences) to generate a large number of pairs of positive samples $x_{j(i)} \in \mathcal{G}_i^+$, where $\mathcal{G}_i^+$ is the set of positive samples for anchor x_i. Specifically, 10 random mutations are performed for each gene sequence, with mutation ratios randomly generated according to a standard normal distribution. The number of mutations is calculated based on sequence lengths. This process aims to generate new sequences that are functionally similar to the original sequence but sequentially different, providing additional training data to improve the predictive power and accuracy of orphan EC numbers.

Hard Negatives Sampling. Previous studies (Hermans, Beyer, and Leibe 2017) have demonstrated that a critical component of successful contrastive learning is balancing the triviality and hardness of the sampled triplets. For the negative sample pair $x_{k(i)} \in \mathcal{G}_i^-$, where $\mathcal{G}_i^-$ represents the set of negative samples for the anchor x_i, this balance is particularly important. We determine the centers for each functional group by averaging the embeddings of all sequences within that group. Subsequently, we compute the Euclidean distances $d(\cdot)$ based on these centers. For negative sample selection, we choose samples that are similar to the anchor in latent space but from different clusters, thus increasing the learning difficulty compared to random selection. The triplet loss $\mathcal{L}_{\text{Tri}}(x_i, \{x_{j(i)}\}_{j=1}^{N_j}, \{x_{k(i)}\}_{k=1}^{N_k})$ is defined:

$$\mathcal{L}_{\text{Tri}} = -\sum_{i \in I} \log(1/|\mathcal{G}_i^+| \sum_{j \in \mathcal{G}_i^+} \exp(S_{z_i, z_{j(i)}}/\tau)/\mathcal{G}_i). \quad (4)$$

For all negative samples in $\mathcal{G}_i^-$, the probability of selecting each negative sample x_k for the anchor x_i as follows:

$$\mathcal{G}_i = \sum_{x_{j(i)}} \exp(S_{z_i, z_{j(i)}}/\tau) + \sum_{x_{k(i)}} p_{x_k} \exp(S_{z_i, z_{k(i)}}/\tau), \quad (5)$$

where τ is the temperature hyper-parameter, and S is the similarity function, typically cosine similarity. The term p_{x_k} represents the probability of selecting the negative sample x_k for the anchor x_i, calculated as $p_{x_k} = w_{x_k} / \sum_{x_m \in \mathcal{G}_i^-} w_{x_m}$ with $w_{x_k} = \frac{1}{d(x_i, x_k)}$.

Finally, MGM and TMC constitute a unified pre-training framework with a total loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MGM}} + \lambda \mathcal{L}_{\text{Tri}}, \quad (6)$$

where λ is a hyper-parameter tuning the influence between two loss functions.

Experiments

We pre-train FGBERT on a large amount of metagenomic data and comprehensively assess its generalization on different datasets ranging from thousands to hundreds of thousands of sequences, as detailed in Tab. 1. Our model is tested across four task levels: (1) Gene Structure Analysis, (2) Functional Gene Prediction, (3) Pathogenicity Potential Assessment, and (4) Nitrogen Cycle Prediction. More detailed descriptions of downstream tasks can be found in the Overview of Downstream Tasks section of the Appendix.

Table 1: Description of Experimental Datasets.

Task	Dataset	Description	#Seq.	#Class
Gene Structure Prediction **(Gene Level)**	E-K12	Gene Operons	4,315	1,379
Functional Prediction **(Functional Level)**	CARD-A	AMR Family	1,966	269
	CARD-D	Drug Class	1,966	37
	CARD-R	Resistance Mech.	1,966	7
	VFDB	Virulence Fact.	8,945	15
	ENZYME	Enzyme Func.	5,761	7
Pathogenicity Prediction **(Bacteria Level)**	PATRIC	Pathogenic Genes	5,000	110
Nitrogen Cycle Prediction **(Environmental Level)**	NCycDB	Cycling Genes	213,501	68

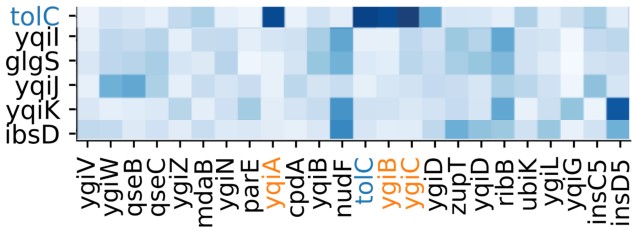


Figure 4: Visualization of Attention.

We use the MGnify database (updated February 2023), which comprises 2,973,257,435 protein sequences from various microbial communities, detailed in the MGnify dataset section of the Appendix.

Results on Four Level Downstream Tasks

Level 1 Task A: Gene Structure Analysis → Gene Operons Prediction. This task is to identify the transcription factor binding sites that are strongly correlated with operon regulation in the gene regulatory network, which helps us to understand the mechanism and network of gene regulation. The dataset used is the E. coli K12 RegulonDB dataset (E-K12) (Salgado et al. 2018), which contains 4315 operons. Detailed information is listed in Appendix Tab.9.

Results Analysis. The attention heatmap in Fig. 4 shows that gene operon tolC has high attention weights with the operons ygiB, ygiC, and yqiA. This suggests a significant interaction among these operons, indicating the presence of a shared genetic operon tolC-ygib. Biological studies (Karp et al. 2019) support this inference by demonstrating that these operons are associated with the DUF1190 structural domain-containing protein YgiB.

Level 2 Task B: Functional Gene Prediction → Antimicrobial Resistance Genes (ARG) Prediction . This task is crucial for understanding ARGs and facilitates the identification of resistance mechanisms. However, existing methods suffer from high false-positive rates and category bias (Jian et al. 2021; Arnold, Huang, and Hanage 2022), necessitating the use of deep learning methods to rapidly and accurately detect ARG presence in metagenomic data. The CARD dataset (Jia et al. 2016) categorizes each gene into one of 269 AMR Gene Families (CARD-A), 37 Drug Classes (CARD-D), or 7 Resistance Mechanisms (CARD-R). Our model performs a three-category classification for each gene sequence in CARD, respectively.

Table 2: Classification Results on CARD-R.

ClassName	Total Count	Correct Num	Correct Ratio
Antibiotic Inactivation	252	238	94.44%
Antibiotic Target Alteration	70	59	84.29%
Antibiotic Target Protection	28	27	96.43%
Antibiotic Efflux	25	18	72.00%
Antibiotic Target Replacement	14	12	85.72%

Results Analysis. FGBERT's performance on CARD-A is significant, as shown in Tab. 3. This category's broad range (269 classifications) creates a long-tail distribution, necessitating an understanding of the biological properties of gene sequences for accurate annotation (McArthur et al. 2013). To mitigate this issue, we adjust the data sampling strategy to increase the frequency of fewer samples, improving the model's prediction accuracy. Tab. 2 demonstrates FGBERT's high prediction accuracy for the CARD-R category, exhibiting superior classification results for both majority and minority classes, with over 85% accuracy. Appendix Tab.15 reveals FGBERT has better performance for all 269 AMR Gene Family categories, with 100% classification accuracy for majority categories like CTX, ADC, CMY, as well as for minor ones like AXC, CRH, KLUC.

Level 2 Task C: Functional Gene Prediction → Virulence Factors (VF) Prediction. This task is to detect microbial elements like bacterial toxins, which enhance pathogen infectivity and exacerbate antimicrobial resistance. Existing methods for metagenomic analysis, particularly those co-predicting ARGs and VFs, are inadequate and suffer from threshold sensitivity issues (Yang et al. 2016). VFDB dataset (Chen et al. 2005) includes the major VFs of the most characterized bacterial pathogens, detailing their structural features, functions, and mechanisms. We use VFDB core dataset, including 8945 VF sequences and 15 VF categories.

Results Analysis. Our model achieves the SOTA results on VFDB, as reported in Tab. 3. It significantly outperforms the genomic pre-trained model; for instance, M.F1 and W.F1 scores improve by 30% and 9.5%, respectively, compared to DNABERT2. This highlights the limitations of directly applying the genomic pre-trained model to metagenomic data for precise functional annotation. Conversely, ESM-2, as a PLM, excels by leveraging intrinsic protein information in metagenomic data, highlighting its effectiveness.

Level 2 Task D: Functional Gene Prediction → Enzyme Function Prediction . This task is critical for understanding metabolism and disease mechanisms in organisms. While traditional methods rely on time-consuming and labor-intensive biochemical experiments, advanced technologies can offer efficient and accurate predictions for large-scale genomic data. ENZYME (Bairoch 2000) is a repository of information related to the nomenclature of enzymes. We organize 5761 data, 7 categories of ENZYME core dataset, each enzyme has a unique EC number (Nomenclature 1992).

Results Analysis. Our experimental results demonstrate

Table 3: Macro F1 and Weighted F1 on eight downstream tasks: Gene Operon Prediction on E-K12, ARG Prediction on three CARD categories, Virulence Factors Classification on VFDB, Enzyme Function Annotation on ENZYME, Microbial Pathogens Detection on PATRIC, Nitrogen Cycle Processes Prediction on NCycDB. RF denotes Random Forest, and VT represents Vanilla Transformer. The highest results are highlighted in boldface, and the second with underline.

Method	Operons		ARG Prediction						Virus		Enzyme		Pathogen		N-Cycle	
	E-K12		CARD-A		CARD-D		CARD-R		VFDB		ENZYME		PATRIC		NCycDB	
	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1	M.F1	W.F1
RF	20.2	34.8	22.4	35.3	36.1	49.0	47.8	57.6	22.4	38.5	33.6	41.2	25.3	29.8	67.0	71.7
SVM	38.6	45.2	27.6	40.5	33.6	47.2	43.3	66.2	28.0	41.4	31.3	43.6	26.6	31.2	66.9	70.3
KNN	39.9	41.0	36.9	54.4	36.4	51.3	36.2	63.5	27.3	47.1	31.4	42.9	11.0	27.4	68.8	73.2
LSTM	40.4	42.5	47.1	60.3	39.1	62.3	47.5	84.2	36.7	66.3	42.8	51.0	41.3	49.7	71.9	81.2
BiLSTM	38.2	43.8	47.4	61.9	43.5	58.1	58.9	80.3	46.1	72.1	38.7	50.2	43.3	48.5	82.0	88.4
VT	43.3	47.8	57.1	70.0	49.8	68.1	55.7	86.4	58.0	81.0	68.2	75.8	49.8	57.3	84.5	90.7
HyenaDNA	42.4	47.1	50.9	68.2	53.6	78.1	66.2	88.1	61.0	70.4	79.6	83.6	51.1	57.6	92.4	96.0
ESM-2	38.2	42.5	57.2	71.4	56.0	82.1	68.2	90.0	60.7	84.4	92.5	96.7	56.0	67.5	95.8	96.1
NT	45.1	44.8	58.5	72.0	56.2	80.2	68.0	90.3	58.3	71.6	74.1	76.7	46.1	61.9	75.1	86.5
DNABERT2	51.7	52.4	65.2	79.8	51.5	78.7	61.2	88.6	58.2	82.3	85.4	85.2	52.9	60.6	88.6	95.7
Ours	**61.8**	**65.4**	**78.6**	**90.1**	**57.4**	**85.2**	**69.4**	**91.4**	**75.7**	**90.2**	**99.1**	**98.8**	**99.3**	**99.0**	**99.5**	**99.2**

FGBERT's superior performance on the ENZYME dataset. It outperforms ESM-2, the second-highest method, by approximately 6.62% in M.F1 and 2.09% in W.F1, demonstrating its ability to discern distinct enzyme function characteristics. This observation highlights that our model not only captures gene-protein contextual relationships but also effectively models the relationships between sequences and functions within metagenomic data.

Level 3 Task E: Pathogenicity Potential Assessment → Genome Pathogens Prediction. This task assesses the pathogenic potential of pathogens to cope with the public health risks caused by newly emerging pathogens. Accurate deep-learning algorithms are key for the precise identification of pathogens, improving the ability to respond to drug resistance threat. We use PATRIC core dataset (Gillespie et al. 2011), which has 5000 pathogenic bacterial sequences across 110 classes.

Results Analysis. Tab. 3 shows FGBERT's classification of pathogenic bacteria species within PATRIC, demonstrating superior performance over baselines by recognizing crucial genera features. The PATRIC dataset presents a significant challenge due to its large number of categories and sparse data. Baselines generally underperform because they require more data to discern the subtle differences between numerous categories. In contrast, FGBERT stands out with M.F1 and W.F1 scores of 99.27% and 99.03%, respectively. This robust performance indicates its advanced learning capability, making it well-suited for high-dimensional classification tasks and highlighting the benefits of using protein-based gene representations for enhanced functional annotation accuracy.

Level 4 Task F: Nitrogen Cycle Prediction → Nitrogen (N) Cycling Process Prediction. This task focuses on the functional genes related to the N cycle, linking them to environmental and ecological processes. NCycDB (Tu et al. 2019) contains 68 gene (sub)families and covers 8 N cycle processes with 213,501 representative sequences at 100% identity cutoffs, each involving a specific gene family.

Results Analysis. Tab. 3 presents FGBERT's classification results on NCycDB, suggesting its ability to recognize key features of particular N cycle processes and improve gene family classification by recognizing domains. Although baselines show improved performance on NCycDB compared to PATRIC, due to a larger amount of data per category aiding in discrimination among diverse categories, FGBERT still leads with macro F1 (M.F1) and weighted F1 (W.F1) scores of 99.49% and 99.22%, respectively. However, pre-trained baselines require more time and memory for tokenizing large datasets, as analyzed in Sec. Model Efficiency.

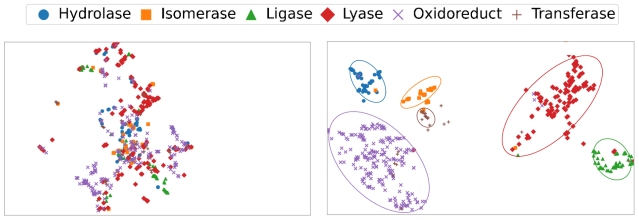
Ablation Study

Ablation Study on the Performance and Visualization of MGM and TMC. We conducted an ablation study to assess the contributions of the MGM and TMC modules to our model's performance. As shown in Tab. 5 and Appendix Fig.7, removing either MGM or TMC resulted in a decrease in M.F1 scores across four datasets, underscoring the importance of these components. Notably, MGM had a more pronounced impact on overall performance, while TMC still provided significant improvements, particularly in clustering metrics like ARI and Silhouette Coefficient.

To further validate MGM, we performed a visualization experiment in the One-to-Many scenario using 1177 ATP synthase sequences from UniProt (uni 2023). Sequences were categorized into six taxonomies, revealing that without genome contextual analysis, the embeddings produced dispersed clusters. In contrast, MGM's integration led to more cohesive clustering, especially in categories with lower representation, effectively addressing the One-to-Many problem (Fig. 5). We examined TMC's impact on gene operon prediction by comparing models with and without this module. The integration of TMC showed marked improvements in clustering performance across various metrics (NMI, ARI, Silhouette Coefficient), as seen in Appendix Tab.12.

Table 4: The Scalability of FGBERT.

Model	Pre-training Data	# Layers	Hidden Dim	# Heads	Operons (500/1000/1500 ep)	CARD-A (500/1000/1500 ep)	CARD-D (500/1000/1500 ep)	CARD-R (500/1000/1500 ep)
FGBERT-T	50M	10	640	5	50.9 / 61.5 / 63.5	74.6 / 88.6 / 90.7	72.4 / 83.7 / 86.4	77.3 / 90.1 / 91.1
FGBERT-S	100M	19	1280	10	55.1 / 65.4 / 65.9	80.4 / 90.1 / 91.2	76.9 / 85.2 / 87.5	81.4 / 91.4 / 93.0
FGBERT-B	150M	25	2560	25	57.1 / 66.7 / 67.6	82.7 / 91.1 / 93.2	81 / 87.6 / 90.0	83.7 / 93.4 / 94.8



(a) Input protein embeddings (b) Contextualized gene embeddings

Figure 5: T-SNE Visualization of Different Embeddings for ATP Synthases. Each dot denotes a sequence and is colored according to different functions.

Table 5: Ablation Study of M.F1 on CARD-D.

Method	Operons	ARG Prediction		
	E-K12	CARD-A	CARD-D	CARD-R
FGBERT	**61.8**	**78.7**	**57.4**	**69.4**
w/o. MGM	-8.1	-6.7	-10.5	-6.7
w/o. Triplet	-7.4	-5.4	-5.8	-3.4

Ablation Study on the Context-Aware Tokenizer. Lastly, we evaluated the effect of substituting the context-aware tokenizer with ESM2 and BPE representations as shown in Tab. 6. FGBERT outperformed both alternatives in capturing gene sequence-function relationships, particularly in metagenomic data analysis. Although BPE offers effective sequence compression, our tokenizer demonstrated superior accuracy, as detailed in Appendix Tab.13.

Model Efficiency Analysis

We analyze the time complexity and memory efficiency of our tokenizer compared to four genomic pre-trained methods on six datasets in Fig. 6. Our tokenizer demonstrates superior efficiency, achieving a significant reduction in both time and memory usage. Notably, on NCyc dataset (brown) with 213,501 sequences, ours reduces processing time by 31.05% and memory usage by 94.33% compared to DNABERT2. For the CARD dataset (orange) with 1,966 sequences, time and memory usage decreased by 61.70% and 58.53%. Although HyenaDNA uses less memory on Operons, CARD, VFDB, and ENZYME datasets, it underperforms ours in time cost and overall performance.

To further explore model scalability, we train three FGBERT variants—FGBERT-T, FGBERT-S, and FGBERT-B—differing in layers, hidden dimensions, and attention heads. Each variant is trained on 50 million, 100 million, and 150 million sequences to assess the impact on time and memory during pre-training, as shown in Tab. 4.

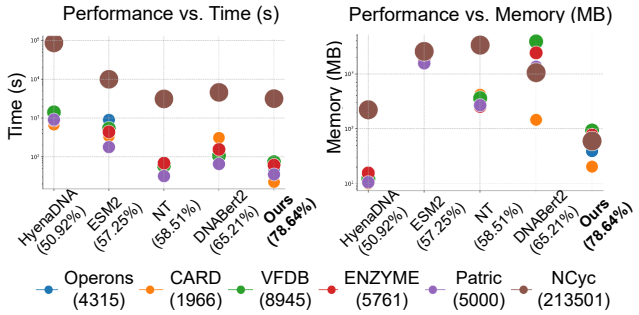


Figure 6: Comparative Analysis on Tokenization Efficiency: Time(s) vs. Memory (MB). Each point denotes a specific dataset, with the size indicating its scale.

Table 6: Ablation study on protein-based gene representation as a context-aware tokenizer.

	Method	Acc	M.Pre	M.Re	M.F1	W.Pre	W.Re	W.F1
E-K12	FGBERT	**0.68**	**0.68**	**0.61**	**0.61**	**0.77**	**0.67**	**0.65**
	ESM2	0.54	0.51	0.48	0.49	0.61	0.51	0.52
	BPE	0.59	0.58	0.55	0.54	0.67	0.54	0.55
CARD-A	FGBERT	**0.91**	**0.77**	**0.80**	**0.78**	**0.90**	**0.91**	**0.90**
	ESM2	0.82	0.74	0.76	0.73	0.87	0.82	0.81
	BPE	0.84	0.75	0.77	0.75	0.88	0.86	0.86

Sensitivity Analysis

Our sensitivity analysis indicates that FGBERT can be optimized effectively using a small batch size b without larger performance degradation (see Appendix Fig.8). This is important for resource-constrained scenarios, highlighting our model maintains good performance even with limited data. We choose a batch size of 1000. Additionally, we examined the impact of the balance ratio α on the CARD dataset. FGBERT consistently performed well across different α values, underscoring its robustness and insensitivity to this hyperparameter. We set α to 0.4.

Conclusion

In this work, we propose a new idea of protein-based gene representation, preserving essential biological characteristics within each gene sequence. With the new context-aware tokenizer, we propose MGM, a gene group-level pre-training task, designed to learn the interactions between genes. Additionally, we develop TMC, a contrastive learning module to generate multiple positive and negative samples to distinguish the gene sequences. MGM and TMC constitute a joint pre-training model, FGBERT for metagenomic data. For the future, it remains to be explored how to incorporate multi-omics data, such as metabolomics, into our metagenomic pre-trained model.

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (partial)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)
- Does this paper make theoretical contributions? (no)

If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes/partial/no)
- All novel claims are stated formally (e.g., in theorem statements). (yes/partial/no)
- Proofs of all novel claims are included. (yes/partial/no)
- Proof sketches or intuitions are given for complex and/or novel results. (yes/partial/no)
- Appropriate citations to theoretical tools used are given. (yes/partial/no)
- All theoretical claims are demonstrated empirically to hold. (yes/partial/no/NA)
- All experimental code used to eliminate or disprove claims is included. (yes/no/NA)

Does this paper rely on one or more datasets? (yes)
If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (yes/)
- All novel datasets introduced in this paper are included in a data appendix. (yes)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)
- Does this paper include computational experiments? (yes)

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (no).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (no)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (no)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes/no)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes)

References

2023. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531.

Al-Ajlan, A.; and El Allali, A. 2019. CNN-MGP: convolutional neural networks for metagenomics gene prediction. *Interdisciplinary Sciences: Computational Life Sciences*, 11: 628–635.

Al-Shayeb, B.; Skopintsev, P.; Soczek, K. M.; Stahl, E. C.; Li, Z.; Groover, E.; Smock, D.; Eggers, A. R.; Pausch, P.; Cress, B. F.; et al. 2022. Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell*, 185(24): 4574–4586.

Albertsen, M. 2023. Long-read metagenomics paves the way toward a complete microbial tree of life. *Nature Methods*, 20(1): 30–31.

Arnold, B. J.; Huang, I.-T.; and Hanage, W. P. 2022. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4): 206–218.

Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; and Kelley, D. R. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10): 1196–1203.

Bairoch, A. 2000. The ENZYME database in 2000. *Nucleic acids research*, 28(1): 304–305.

Chen, L.; Yang, J.; Yu, J.; Yao, Z.; Sun, L.; Shen, Y.; and Jin, Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic acids research*, 33(suppl_1): D325–D328.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023–01.

D'Onofrio, D. J.; and Abel, D. L. 2014. Redundancy of the genetic code enables translational pausing. *Frontiers in genetics*, 5: 140.

Fiannaca, A.; La Paglia, L.; La Rosa, M.; Lo Bosco, G.; Renda, G.; Rizzo, R.; Gaglio, S.; and Urso, A. 2018. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC bioinformatics*, 19: 61–76.

Gillespie, J. J.; Wattam, A. R.; Cammer, S. A.; Gabbard, J. L.; Shukla, M. P.; Dalay, O.; Driscoll, T.; Hix, D.; Mane, S. P.; Mao, C.; et al. 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity*, 79(11): 4286–4298.

Gruenstaeudl, M. 2020. annonex2embl: automatic preparation of annotated DNA sequences for bulk submissions to ENA. *Bioinformatics*, 36(12): 3841–3848.

Gwak, H.-J.; and Rho, M. 2022. ViBE: a hierarchical BERT model to identify eukaryotic viruses using metagenome sequencing data. *Briefings in Bioinformatics*, 23(4): bbac204.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hoarfrost, A.; Aptekmann, A.; Farfañuk, G.; and Bromberg, Y. 2022. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature communications*, 13(1): 2606.

Hu, Y.; Chen, Y.; Xu, J.; Wang, X.; Luo, S.; Mao, B.; Zhou, Q.; and Li, W. 2022. Metagenomic discovery of novel CRISPR-Cas13 systems. *Cell Discovery*, 8(1): 107.

Jain, S.; Hassanzadeh, F. F.; Schwartz, M.; and Bruck, J. 2017. Duplication-correcting codes for data storage in the DNA of living organisms. *IEEE Transactions on Information Theory*, 63(8): 4996–5010.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Jia, B.; Raphenya, A. R.; Alcock, B.; Waglechner, N.; Guo, P.; Tsang, K. K.; Lago, B. A.; Dave, B. M.; Pereira, S.; Sharma, A. N.; et al. 2016. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, gkw1004.

Jian, Z.; Zeng, L.; Xu, T.; Sun, S.; Yan, S.; Yang, L.; Huang, Y.; Jia, J.; and Dou, T. 2021. Antibiotic resistance genes in bacteria: Occurrence, spread, and control. *Journal of basic microbiology*, 61(12): 1049–1070.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; et al. 2019. The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in bioinformatics*, 20(4): 1085–1093.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kusters, K. A.; Pratsinis, S. E.; Thoma, S. G.; and Smith, D. M. 1993. Ultrasonic fragmentation of agglomerate powders. *Chemical Engineering Science*, 48(24): 4119–4127.

Lawson, C. L.; Swigon, D.; Murakami, K. S.; Darst, S. A.; Berman, H. M.; and Ebright, R. H. 2004. Catabolite activator protein: DNA binding and transcription activation. *Current opinion in structural biology*, 14(1): 10–20.

Lee, S. J.; and Rho, M. 2022. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Scientific Reports*, 12(1): 824.

Liang, Q.; Bible, P. W.; Liu, Y.; Zou, B.; and Wei, L. 2020. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics*, 2(1): lqaa009.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; et al. 2022. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature*, 604(7907): 662–667.

Mande, S. S.; Mohammed, M. H.; and Ghosh, T. S. 2012. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6): 669–681.

Mathieu, A.; Leclercq, M.; Sanabria, M.; Perin, O.; and Droit, A. 2022. Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. *Frontiers in Microbiology*, 13: 811495.

McArthur, A. G.; Waglechner, N.; Nizam, F.; Yan, A.; Azad, M. A.; Baylay, A. J.; Bhullar, K.; Canova, M. J.; De Pascale, G.; Ejim, L.; et al. 2013. The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7): 3348–3357.

McWilliam, H.; Li, W.; Uludag, M.; Squizzato, S.; Park, Y. M.; Buso, N.; Cowley, A. P.; and Lopez, R. 2013. Analysis tool web services from the EMBL-EBI. *Nucleic acids research*, 41(W1): W597–W600.

Miao, Y.; Liu, F.; Hou, T.; and Liu, Y. 2022. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics*, 38(5): 1216–1222.

Miller, D.; Stern, A.; and Burstein, D. 2022. Deciphering microbial gene function using natural language processing. *Nature Communications*, 13(1): 5731.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Nomenclature, E. 1992. Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes.

Pastinen, T.; Ge, B.; and Hudson, T. J. 2006. Influence of human genome polymorphism on gene expression. *Human molecular genetics*, 15(suppl_1): R9–R16.

Pavlopoulos, G. A.; Baltoumas, F. A.; Liu, S.; Selvitopi, O.; Camargo, A. P.; Nayfach, S.; Azad, A.; Roux, S.; Call, L.; Ivanova, N. N.; et al. 2023. Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983): 594–602.

Ren, J.; Song, K.; Deng, C.; Ahlgren, N. A.; Fuhrman, J. A.; Li, Y.; Xie, X.; Poplin, R.; and Sun, F. 2020. Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, 8: 64–77.

Salgado, H.; Martínez-Flores, I.; Bustamante, V. H.; Alquicira-Hernández, K.; García-Sotelo, J. S.; García-Alonso, D.; and Collado-Vides, J. 2018. Using RegulonDB, the Escherichia coli K-12 gene regulatory transcriptional network database. *Current protocols in bioinformatics*, 61(1): 1–32.

Segata, N.; Waldron, L.; Ballarini, A.; Narasimhan, V.; Jousson, O.; and Huttenhower, C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8): 811–814.

Tu, Q.; Lin, L.; Cheng, L.; Deng, Y.; and He, Z. 2019. NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics*, 35(6): 1040–1048.

Wichmann, A.; Buschong, E.; Müller, A.; Jünger, D.; Hildebrandt, A.; Hankeln, T.; and Schmidt, B. 2023. MetaTransformer: deep metagenomic sequencing read classification using self-attention models. *NAR Genomics and Bioinformatics*, 5(3): lqad082.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Yan, H.; Bombarely, A.; and Li, S. 2020. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics*, 36(15): 4269–4275.

Yang, C.; Chowdhury, D.; Zhang, Z.; Cheung, W. K.; Lu, A.; Bian, Z.; and Zhang, L. 2021. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19: 6301–6314.

Yang, M.; Huang, L.; Huang, H.; Tang, H.; Zhang, N.; Yang, H.; Wu, J.; and Mu, F. 2022. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic acids research*, 50(14): e81–e81.

Yang, Y.; Jiang, X.; Chai, B.; Ma, L.; Li, B.; Zhang, A.; Cole, J. R.; Tiedje, J. M.; and Zhang, T. 2016. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics*, 32(15): 2346–2351.

Zhang, Z.; Han, Z.; Wu, Y.; Jiang, S.; Ma, C.; Zhang, Y.; and Zhang, J. 2021. Metagenomics assembled genome scale analysis revealed the microbial diversity and genetic polymorphism of Lactiplantibacillus plantarum in traditional fermented foods of Hainan, China. *Food Research International*, 150: 110785.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

Zvyagin, M.; Brace, A.; Hippe, K.; Deng, Y.; Zhang, B.; Bohorquez, C. O.; Clyde, A.; Kale, B.; Perez-Rivera, D.; Ma, H.; et al. 2022. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 10943420231201154.