Briefings in Bioinformatics, 2025, **26(2)**, bbaf149
https://doi.org/10.1093/bib/bbaf149
Problem Solving Protocol

FGeneBERT: function-driven pre-trained gene language model for metagenomics

Chenrui Duan[1,2], Zelin Zang[3], Yongjie Xu[1,2], Hang He[4], Siyuan Li[1,2], Zihan Liu[1,2], Zhen Lei[3,5,6], Ju-Sheng Zheng[4,*], Stan Z. Li[2,*]

[1]College of Computer Science and Technology, Zhejiang University, No. 866, Yuhangtang Road, 310058 Zhejiang, P. R. China
[2]School of Engineering, Westlake University, No. 600 Dunyu Road, 310030 Zhejiang, P. R. China
[3]Centre for Artificial Intelligence and Robotics (CAIR), HKISI-CAS Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong 310000, China
[4]School of Medicine and School of Life Sciences, Westlake University, No. 600 Dunyu Road, 310030 Zhejiang, P. R. China
[5]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[6]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China

[*]Corresponding authors. Ju-Sheng Zheng, School of Medicine and School of Life Sciences, Westlake University, No. 600 Dunyu Road, 310030 Zhejiang, P. R. China. E-mail: zhengjusheng@westlake.edu.cn; Stan Z. Li, School of Engineering, Westlake University, No. 600 Dunyu Road, 310030 Zhejiang, P. R. China. E-mail: stan.zq.li@westlake.edu.cn

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 FGeneBERT, a novel metagenomic pre-trained model that employs a protein-based gene representation as a context-aware and structure-relevant tokenizer. FGeneBERT incorporates masked gene modeling to enhance the understanding of inter-gene contextual relationships and triplet enhanced metagenomic contrastive learning to elucidate gene sequence–function relationships. Pre-trained on over 100 million metagenomic sequences, FGeneBERT demonstrates superior performance on metagenomic datasets at four levels, spanning gene, functional, bacterial, and environmental levels and ranging from 1 to 213 k input sequences. Case studies of ATP synthase and gene operons highlight FGeneBERT's capability for functional recognition and its biological relevance in metagenomic research.

Keywords: metagenomics; DNA; pre-trained language model; transformer

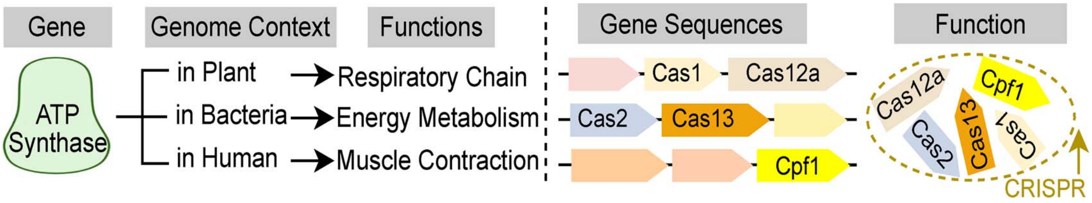
Introduction

Metagenomics, the study of mixed genomes of microbial communities in the environment (e.g. gut microbiomes or soil ecosystems) [1–3], has revealed the critical role in fundamental biological processes like enzyme synthesis, gene expression regulation, and immune function [4–6]. 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 unraveling life's mechanisms [7, 8], and advancing biotechnology [9, 10], with important applications in biodiversity conservation and epidemiology [11, 12].

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 [13, 14]. 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 [15]. 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 [16]. For example, the CRISPR (CRISPR-Cas systems are essential adaptive immune components in microorganisms that defend against mobile genetic elements and form the basis of advanced genome engineering technologies. Unlike the commonly used DNA-targeting Cas9 or Cas12 systems, Cas13 is an RNA-guided, programmable RNA-targeting system that enables gene manipulation at the transcriptional level.) immune mechanism [16, 17] involves various proteins like Cpf1, Cas1, and Cas13, each contributing to the same defensive function (Fig. 1b) [18, 19].

Recently, various computational methods have emerged for genomic and metagenomic data analysis. Traditional alignment-based methods like MetaPhlAn5 [20] 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 [21] and DeepVirFinder [22], which use CNNs for gene and viral classifications with one-hot encoding. K-mer tokenization [23], employed in approaches

Received: December 25, 2024. **Revised:** February 22, 2025. **Accepted:** March 14, 2025



(a) One-to-Many (OTM) Problem

(b) Many-to-One (MTO) Problem

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.

like MDL4Microbiome [24], is a standard for DNA sequence characterization. Additionally, Virtifier [25] maps a nucleotide sequence using a codon dictionary combined with LSTM to predict viral genes. DeepMicrobes [26] employs a self-attention mechanism, while DeepTE [27] uses K-mer inputs with CNNs for element classification, and Genomic-nlp [28] applies word2vec for gene function analysis. MetaTransformer [29] uses K-mer embedding for species classification with Transformer. For pre-training models, LookingGlass [30] uses a three-layer LSTM model for functional prediction in short DNA reads. ViBE [31] employs a K-mer token-based BERT model pre-trained with Masked Language Modeling for virus identification. The BERT model, effective in DNA sequence characterization, is limited by the Transformer architecture's computational burden. LOGO [32] addresses this by cutting off long sequences into 1-2kb sub-sequences. Enformer [33] combines extended convolution with Transformers for long human genomic data. GenSLMs [34] introduce hierarchical language models for whole-genome modeling. DNABERT [35], the first pre-trained model on the human genome that focuses on extracting efficient genomic representations. DNABERT2 [36], its successor, uses Byte Pair Encoding on multi-species genomic data. NT [37] is trained on nucleotide sequences from humans and other species and evaluated on 18 genome prediction tasks. HyenaDNA [38] presents a long-range genomic model based on single-nucleotide polymorphisms on human reference genomes.

Despite these advancements, current approaches exhibit three critical limitations when analyzing metagenomic data. Firstly, the Semantic Tokenizer problem. Most machine learning-based methods [26, 28, 30], 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 [36, 37, 39]. Some models have also been developed for long-range DNA-related tasks, demonstrating promising performance in long-range species classification tasks [40, 41], although the practical applications of this problem remain poorly defined [42]. However, these tasks 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 persists in metagenomic models, with recent work [43] proposing optimized strategies to enhance model adaptability. Models like MetaTransformer [29] and ViBE [31], 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 FGeneBERT (Function-Driven Pre-trained Gene Language Model for Metagenomics), 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 metagenomic sequences. This strategy leverages the inherent protein functional and structural information encoded in metagenomic data [44], overcoming the limitations of K-mer methods and maintaining functional consistency despite potential mutations [45]. 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, FGeneBERT is pre-trained on over 100 million sequences, showcasing robust performance across diverse datasets spanning gene, functional, bacterial, and environmental levels.

In this work, we identify three key challenges in metagenomic analysis. To address these issues, we propose FGeneBERT. 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: (i) We introduce a new idea of protein-based gene representations to learn biologically meaningful tokenization of long sequences. (ii) We propose MGM to model inter-gene relationships and TMC to learn complex relationships between gene sequences and functions. (iii) We conduct extensive experiments across various downstream tasks, spanning gene, functional, bacterial, and environmental levels with input sizes from 1 to 213 k sequences. FGeneBERT achieves the state-of-the-art performance.

Methods

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

Datasets

We pre-train and evaluate FGeneBERT on distinct datasets ranging from thousands to hundreds of thousands of sequences. For pre-training, we utilize the MGnify database [updated February 2023 (https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2023_02/)], comprising 2 973 257 435 protein sequences aggregated from 30 FASTA files of diverse microbial communities. Table 1 details microbial genomes across

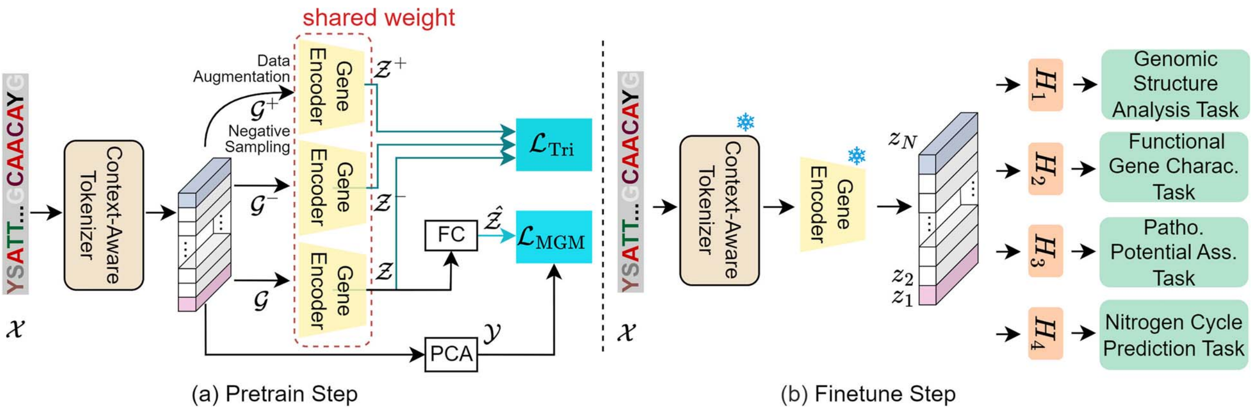


Figure 2. Overview of FGeneBERT. A metagenomic sequence $\mathcal{X}$ is converted into ordered protein-based gene representations $\mathcal{G}$ via a Context-Aware Tokenizer. Next, we pre-train a Gene Encoder with $\mathcal{L}_{\mathrm{MGM}}$, 15% of these tokens are masked to predict labels $\mathcal{Y}$.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

Table 1. Microbial species and genome numbers in each environment in the MGnify dataset. Columns denote environmental catalogs, microbial species, and assembled genomes, respectively

Catalogues	Species	Genomes
Human gut	4744	289 232
Human oral	452	1225
Cow rumen	2729	5578
Marine	1496	1504
Pig gut	1376	3972
Zebrafish feacel	79	101
Non-model fish gut	172	196

environments (e.g. Human gut: 289 232 genomes, 4744 species; marine: 1504 genomes, 1496 species), distinct from its protein sequences used for pre-training, with cataloged microbial species provided. For downstream evaluation, we test FGeneBERT across four task levels—(i) gene structure analysis, (ii) functional gene prediction, (iii) pathogenicity potential assessment, and (iv) nitrogen cycle prediction—using datasets split into training, validation, and test sets (8:1:1) with five-fold cross-validation. As detailed in Table 2: E-K12 [46] (4315 operons) for gene structure; CARD [47] (1966 sequences) for functional prediction across 269 AMR families (CARD-A), 37 drug classes (CARD-D), and 7 resistance mechanisms (CARD-R); VFDB core [48] (8945 sequences, 15 VF categories) for virulence factors; ENZYME [49] (5761 sequences, 7 categories) for enzyme functions with unique EC numbers [50]; PATRIC core [51] (5000 sequences, 110 classes) for pathogenicity; and NCycDB [52] (213 501 sequences, 68 gene families) for nitrogen cycle processes (see Supplementary section *Overview of Downstream Tasks*).

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 [53], 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 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 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 [54] 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 [55]. 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 [56]. 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 [57, 58]. 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 [44], 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.

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

Table 2. Description of experimental datasets. #Seq. and #Class means the number of sequences and categories in each dataset

Task	Dataset	Description	#Seq.	#Class
Gene structure prediction (Gene level)	E-K12	Gene operons	4315	1379
Functional prediction (Functional level)	CARD-A	AMR family	1966	269
	CARD-D	Drug class	1966	37
	CARD-R	Resistance mech.	1966	7
	VFDB	Virulence fact.	8945	15
	ENZYME	Enzyme func.	5761	7
Pathogenicity prediction (Bacteria level)	PATRIC	Pathogenic genes	5000	110
Nitrogen cycle prediction (Environmental level)	NCycDB	Cycling genes	213 501	68

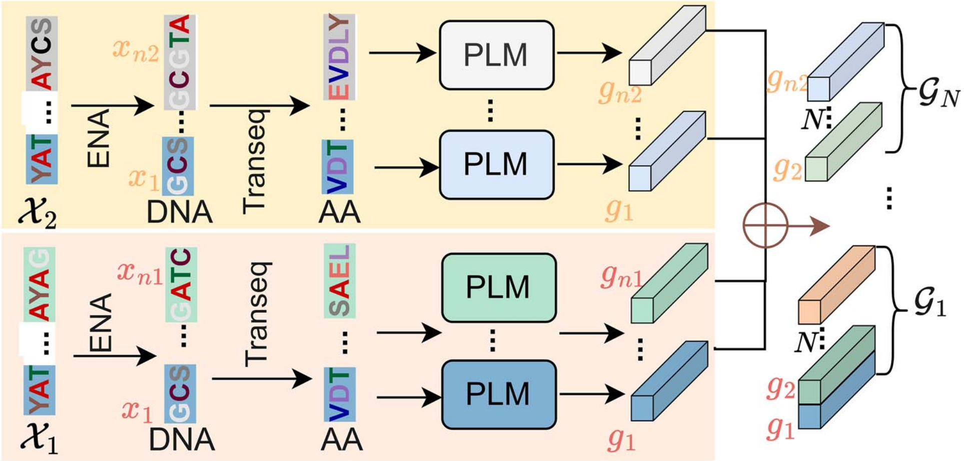


Figure 3. Framework of Context-Aware Tokenizer. Gene sequences $\{x_i\}_{i=1}^{n_i}$ extracted from metagenomic sequence $\mathcal{X}_i$ are translated into amino acid sequences and encoded into ESM-2 representations as $\{g_i\}_{i=1}^{n_i}$. These representations are then concatenated sequentially every N representation to form gene groups $\mathcal{G}$.

unmasked genome context $\mathcal{G}_{/M}$:

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

where M denotes the index set of the masked gene.

In addition, considering genetic polymorphism [59, 60], 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, $\mathcal{L}_{\mathrm{MGM}}$, which incorporates feature reconstruction and probability prediction:

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

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 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 sharing the same EC number cluster closely in feature space, while those with distinct EC numbers are more separated [61], with recent advancements [62] optimizing feature extraction for improved clustering. 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:

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

where $d(\cdot,\cdot)$ is a distance function (e.g. Euclidean distance) defined on the embedding space. We adopt the SupCon-Hard loss [63] to incorporate multiple positive and negative samples per anchor, enhancing robustness by mining difficult samples, with recent work [64] extending its application to biological contexts. 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 learning 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 work [65] established that balancing triplet triviality and hardness is key to effective contrastive learning, while recent studies [66, 67] further refine this by introducing advanced sampling strategies and efficiency improvements. 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 $\mathcal{L}_{\mathrm{Tri}}(x_i, \{x_{j(i)}\}_{j=1}^{N_j}, \{x_{k(i)}\}_{k=1}^{N_k})$ is defined:

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

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

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

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:

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{MGM}} + \lambda\mathcal{L}_{\mathrm{Tri}}, \tag{6}$$

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

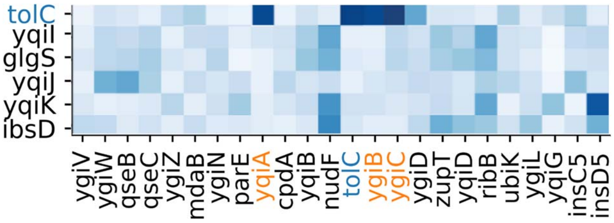


Figure 4. Visualization of attention. The attention value is from the first head of the last (19th) attention layer. Darker shading indicates higher attention weight.

Table 3. Classification results on CARD-R

Class name	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 and discussion

Experiments 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, enhancing our understanding of gene regulation mechanisms. We utilize the *Escherichia coli* K12 RegulonDB dataset (E-K12) [46], a comprehensive resource of operon annotations, with further details provided in Supplementary Table S2.

Results analysis. The attention heatmap in Fig. 4 shows that gene operon tolC has high attention weight with the operons *ygiB* and *ygiC* and *yqiA*. This suggests a significant interaction among these operons, indicating the presence of a shared genetic operon tolC-ygib. This inference finds support in biological research, which associates these operons with the DUF1190 domain-containing protein YgiB [68].

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 [69, 70], necessitating the use of deep learning methods to rapidly and accurately detect ARG presence in metagenomic data. We utilize the CARD dataset [47], which provides a structured classification framework for resistance profiles, with our model performing three-category classification across its gene annotations.

Results analysis. FGeneBERT's performance on CARD-A is significant, as shown in Table 4. 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 [71]. To mitigate this issue, we adjust the data sampling strategy to increase the frequency of fewer samples, improving the model's prediction accuracy. Table 3 demonstrates FGeneBERT's high prediction accuracy for the CARD-R category, exhibiting superior classification results for both majority and minority classes, with over 85% accuracy. Supplementary Table S8 reveals FGeneBERT has better performance for all 269 AMR Gene Family categories, with 100% classification accuracy for majority

Table 4. Macro F1 (% ↑) and Weighted F1 (% ↑) on eight downstream tasks. This includes 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 with boldface. The second highest results are highlighted 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**

categories like CTX, ADC, and CMY, as well as for minor ones like AXC, CRH, and 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 [72]. We employ the VFDB core dataset [48], a detailed repository of VF characteristics from well-characterized bacterial pathogens, to predict these critical microbial elements. Results analysis. Our model achieves the SOTA results on VFDB, as reported in Table 4. 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. We leverage the ENZYME core dataset [49], a comprehensive repository of enzyme nomenclature, to facilitate these predictions. Results analysis. Our experimental results demonstrate FGeneBERT'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 threats. We use PATRIC core dataset [51], which has 5000 pathogenic bacterial sequences across 110 classes. Results analysis. Table 4 shows FGeneBERT'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, FGeneBERT 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. We utilize the NCycDB dataset [52], a comprehensive resource spanning key N cycle gene families and processes, to facilitate this analysis. Results analysis. Table 4 presents FGeneBERT'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, FGeneBERT 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 Model efficiency analysis.

Ablation study

Ablation study on the performance of MGM and TMC

We perform the ablation study to investigate the effectiveness of our proposed components. Table 5 compares the performance of FGeneBERT without MGM and TMC modules across four datasets. Specifically, MGM is designed to address the OTM task, while TMC is tailored to handle the MTO task. The decrease in M.F1 score after removing the MGM and TMC, respectively, highlights their roles in enhancing model performance across different tasks. While TMC contributes to performance, it is evident that MGM has

Table 5. Ablation study of M.F1 on CARD-D. *w/o.* MGM denotes FGeneBERT without MGM, *w/o.* Triplet denotes FGeneBERT without TMC

Method	Operons	ARG prediction		
	E-K12	CARD-A	CARD-D	CARD-R
FGeneBERT	**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

Bold values highlight the best (maximum) results in each column. Negative values indicate how much lower the corresponding result is compared to the best result.

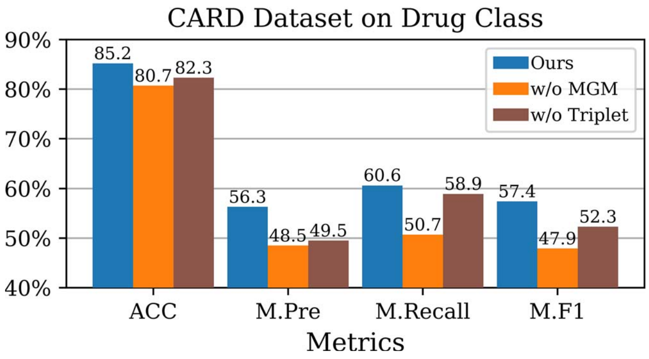


Figure 5. Ablation studies of our proposed modules on four downstream tasks.

a more substantial impact. Fig. 5 illustrates four more metrics on the CARD dataset.

Ablation study on visualization results of MGM

We conduct a visualization experiment to validate the effectiveness of MGM in the OTM scenario. ATP synthases can exhibit different functions in different organisms to adapt to their respective environmental conditions, even though the basic functions are the same [73].

We collect 1177 ATP synthase sequences from UniProt [74] and color them according to six taxonomies, i.e. Hydrolases (35.6%, dark blue), Isomerases (3.6%, orange), Ligases (35.7%, green), Lyases (10.7%, red), Oxidoreduct (8.9%, light purple) and Transferases (5.3%, brown) (numbers in parentheses indicate category ratios). Figure 6(a) shows the clustering results of ATPase protein embeddings without genome contextual analysis, indicating that more dispersed clustering results of genes in different genome contexts. Figure 6(b) presents the clustering results of ATPase embeddings after our designed protein-based gene embedding, which shows genes belonging to the same category (the same color) apparently cluster together. Concretely, Isomerases (orange) and Transferases (brown), which account for the smallest percentage, cluster together, whereas, in the left plot, these two are scattered. This demonstrates that our proposed MGM can resolve the One-to-Many problem between sequences and functions.

Ablation study on clustering results of TMC

We further explore the impact of integrating TMC into our model on gene operon prediction in the Many-to-One (MTO) scenario. Initially, we evaluate FGeneBERT without the TMC module. Afterward, we add the TMC module, and K-means are performed using the embeddings from the network's intermediate layers. We compute the normalized mutual information (NMI), adjusted Rand index (ARI), and Silhouette coefficient index to measure the clustering quality. The results in Table 6 show significant improvements in ARI and Silhouette coefficient with TMC, highlighting its effectiveness in enhancing clustering performance across various metrics.

Table 6. Clustering results of TMC for many-to-one problem

Method	NMI	ARI	Silhouette coefficient
+Tokenizer	0.44	0.34	0.7
+MGM	0.66	0.51	0.72
+TMC	**0.72**	**0.59**	**0.75**

Bold values indicate the best performance in each column.

Ablation study on context-aware tokenizer

We replace the FGeneBERT's tokenizer with ESM2 and BPE representations and assessed performance across downstream tasks. Table 7 shows that FGeneBERT outperforms both methods in capturing complex gene sequence-function relationships. Although DNABERT2 compresses sequences effectively with BPE, our context-aware tokenizer demonstrates superior accuracy in metagenomic data analysis. For complete results, see Supplementary Table S6.

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. 7. Our tokenizer demonstrates superior efficiency, achieving a significant reduction in time and memory usage. Notably, on the 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 1966 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 Table 8.

Sensitivity analysis

Our sensitivity analysis indicates that FGeneBERT can be optimized effectively using small batch size b without larger performance degradation shown in 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. Next, we analyze another hyper-parameter of balance ratio α on CARD dataset. FGeneBERT obtains good performance for different values of α, demonstrating its robustness and insensitivity to this hyper-parameter. We set α to be 0.4.

Model configurations

All our experiments are performed on four NVIDIA V100 GPUs and the PyTorch framework. The encoder of FGeneBERT is initialized with Roberta [75]. For pre-training, the parameters are set as follows: batch size of 1000, 19 encoder layers, 10 attention heads, an embedding dimension of 1280, and relative position encoding. During the pre-training stage, the model is trained for 500 epochs using the AdamW optimizer [76] with a weight decay of 0.02. The learning rate starts at 1e-5, warms up to 1e-4 over

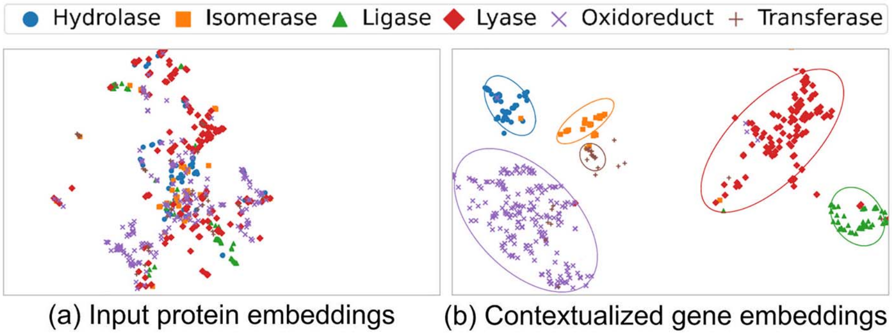


Figure 6. T-SNE visualization of different embeddings for ATP synthases. Each dot denotes a sequence and is grouped according to different functions.

Table 7. The ablation study using protein-based gene representation as a context-aware tokenizer

Dataset	Method	Acc	M.Pre	M.Recall	M.F1	W.Pre	W.Recall	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	<u>0.59</u>	<u>0.58</u>	<u>0.55</u>	<u>0.54</u>	<u>0.67</u>	<u>0.54</u>	<u>0.55</u>
CARD-A	FGBERT	**0.91**	**0.77**	**0.8**	**0.78**	**0.9**	**0.91**	**0.9**
	ESM2	0.82	0.74	0.76	0.73	0.87	0.82	0.81
	BPE	<u>0.84</u>	<u>0.75</u>	<u>0.77</u>	<u>0.75</u>	<u>0.88</u>	<u>0.86</u>	<u>0.86</u>

Bold values indicate the best performance, and underlined values indicate the second-best performance.

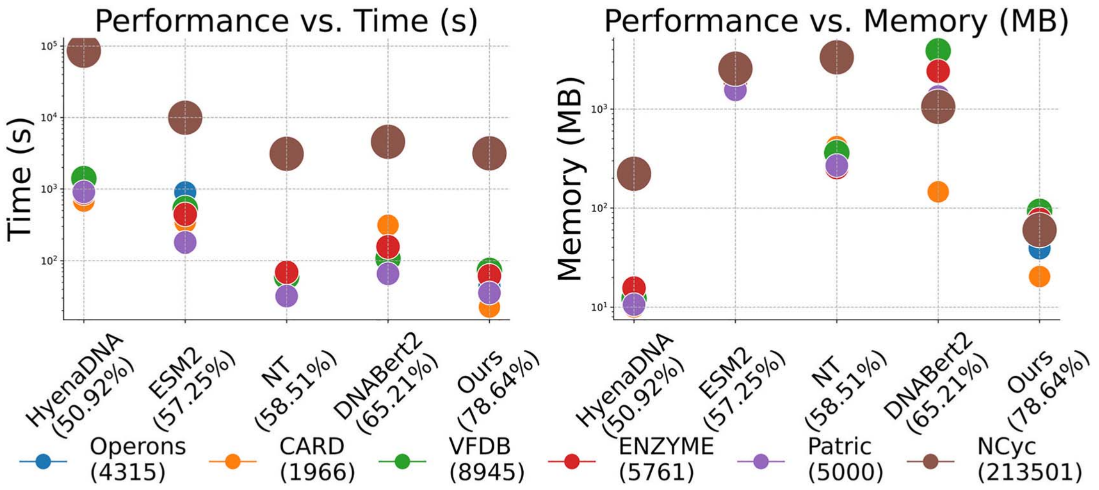


Figure 7. Comparative analysis on tokenization efficiency: time(s) vs. memory (MB). Each point denotes a specific dataset, with the size indicating its scale.

Table 8. Time complexity and memory efficiency of pre-training and scalability of FGBERT

Model	P[a]	# L[b]	H[a]	# H[b]	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]P and H represent the pre-training data and hidden dimension, respectively. [b]# L, and # H represent the number of Layers and Heads, respectively.

the first 5000 steps, and then decreases back to 1e-5 following a cosine decay schedule. The complete model consists of 954.73 million parameters and requires 2.55 billion FLOPs, as detailed in Supplementary Table S4.

Conclusion

In this paper, 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, FGeneBERT for metagenomic data. Our experiments and visualizations demonstrate the superior performance of our model. For the future, it remains to be explored how to incorporate multi-omics

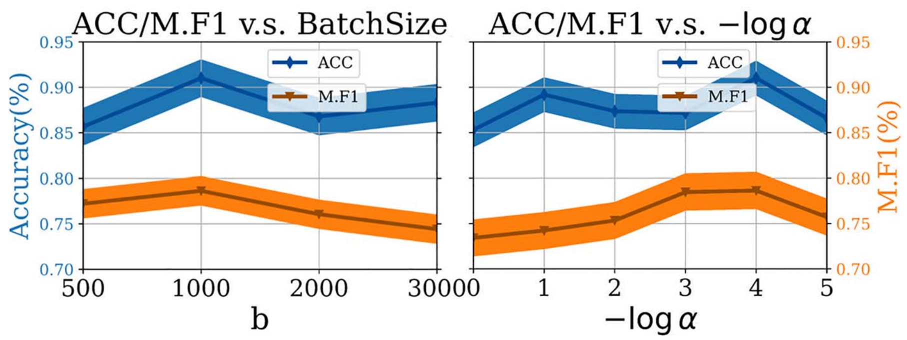


Figure 8. Sensitivity w.r.t. hyper-parameters α, b of CARD dataset on AMR gene family.

data, such as metabolomics, into our metagenomic pre-trained model.

Key Points

- Metagenomic analysis faces three key challenges: the lack of context-aware representations, limited capture of inter-gene relationships, and inefficient modeling of gene-function associations in long sequences.
- We introduce FGeneBERT, the first metagenomic pre-trained model that encodes biologically meaningful, context-aware, and function-relevant representations of metagenomic data.
- FGeneBERT employs protein-based gene representations for efficient tokenization of long sequences and uses MGM and TMC modules to model inter-gene relationships and gene-function associations, respectively.
- Comprehensive experiments show that FGeneBERT consistently achieves SOTA performance in metagenomic analysis.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: No competing interest is declared.

Funding

This work was supported by National Science and Technology Major Project (No. 2022ZD0115101), National Natural Science Foundation of China Project (No. U21A20427), Project (No. WU2022A009) from the Center of Synthetic Biology and Integrated Bioengineering of Westlake University, Project (No. WU2023C019) from the Westlake University Industries of the Future Research Funding. We thank the AI Station of Westlake University for the support of GPUs. This work was supported by the InnoHK program and Ant Group through CAAI-Ant Research Fund.

Data availability

All data processed in this study were obtained exclusively from public sources and are available at https://github.com/jane-pyc/OpenMeta/datasets. Source codes are freely available at https://github.com/jane-pyc/OpenMeta/models/FGBERT.

References

1. Mathieu A, Leclercq M, Sanabria M. *et al.* Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. *Front Microbiol* 2022;**13**:811495. https://doi.org/10.3389/fmicb.2022.811495
2. De D, Nayak T, Das G. *et al.* Metagenomics and bioinformatics in microbial ecology: current status and beyond. In: Thatoi H, Pradhan SK, Kumar U. (Eds.), *Applications of Metagenomics*. Amsterdam, Netherlands: Elsevier, 2024, pp. 359–85. https://doi.org/10.1016/B978-0-323-98394-5.00009-2.
3. Han J, Zhang H, Ning K. Techniques for learning and transferring knowledge for microbiome-based classification and prediction: review and assessment. *Brief Bioinform* 2025;**26**:bbaf015.
4. Duan CR, Zang Z, Li S. *et al.* Phylogen: language model-enhanced phylogenetic inference via graph structure generation. *Adv Neural Inform Process Syst* 2024;**37**:131676–703.
5. Ariaeenejad S, Gharechahi J, Shahraki MF. *et al.* Precision enzyme discovery through targeted mining of metagenomic data. *Na Products Bioprospect* 2024;**14**:7. https://doi.org/10.1007/s13659-023-00426-8
6. Teukam YGN, Zipoli F, Laino T. *et al.* Integrating genetic algorithms and language models for enhanced enzyme design. *Brief Bioinform* 2025;**26**:bbae675. https://doi.org/10.1093/bib/bbae675
7. Albertsen M. Long-read metagenomics paves the way toward a complete microbial tree of life. *Nat Methods* 2023;**20**:30–1. https://doi.org/10.1038/s41592-022-01726-6
8. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
9. Liu S, Moon CD, Zheng N. *et al.* Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome* 2022;**10**:76. https://doi.org/10.1186/s40168-022-01272-5
10. Aplakidou E, Vergoulidis N, Chasapi M. *et al.* Visualizing metagenomic and metatranscriptomic data: a comprehensive review. *Comput Struct Biotechnol J* 2024;**23**:2011–33. https://doi.org/10.1016/j.csbj.2024.04.060
11. Sarumi OA, Heider D. Large language models and their applications in bioinformatics. *Comput Struct Biotechnol J* 2024;**23**:3498–505. https://doi.org/10.1016/j.csbj.2024.09.031
12. Duan CR, Zang Z, Li S. *et al.* Phylogen: language model-enhanced phylogenetic inference via graph structure generation. *Adv Neural Inform Process Syst* 2025;**37**:131676–703.
13. Zhang Z, Duan C, Lin T. *et al.* Gvfom: a novel external force for active contour based image segmentation. *Inform Sci* 2020;**506**:1–18.
14. Hongyuan L, Diaz DJ, Czarnecki NJ. *et al.* Machine learning-aided engineering of hydrolases for pet depolymerization. *Nature* 2022;**604**:662–7. https://doi.org/10.1038/s41586-022-04599-z

15. Yang C, Chowdhury D, Zhang Z. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J* 2021;**19**:6301–14. https://doi.org/10.1016/j.csbj.2021.11.028
16. Al-Shayeb B, Skopintsev P, Soczek KM. *et al.* Diverse virus-encoded crispr-cas systems include streamlined genome editors. *Cell* 2022;**185**:4574–4586.e16. https://doi.org/10.1016/j.cell.2022.10.020
17. Yanping H, Chen Y, Jing X. *et al.* Metagenomic discovery of novel crispr-cas13 systems. *Cell Discov* 2022;**8**:107. https://doi.org/10.1038/s41421-022-00464-5
18. Yang H, Patel DJ. Structures, mechanisms and applications of rna-centric crispr–cas13. *Nat Chem Biol* 2024;**20**:673–88. https://doi.org/10.1038/s41589-024-01593-6
19. Zilberzwige-Tal S, Altae-Tran H, Kannan S. *et al.* Reprogrammable rna-targeting crispr systems evolved from rna toxin-antitoxins. *Cell* 2025. Advance online publication. https://doi.org/10.1016/j.cell.2025.01.034
20. Segata N, Waldron L, Ballarini A. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012;**9**:811–4. https://doi.org/10.1038/nmeth.2066
21. Al-Ajlan A, El Allali A. Cnn-mgp: convolutional neural networks for metagenomics gene prediction. *Interdiscipl Sci: Comput Life Sci* 2019;**11**:628–35. https://doi.org/10.1007/s12539-018-0313-4
22. Ren J, Song K, Deng C. *et al.* Identifying viruses from metagenomic data using deep learning. *Quant Biol* 2020;**8**:64–77. https://doi.org/10.1007/s40484-019-0187-4
23. Fiannaca A, La Paglia L, La Rosa M. *et al.* Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinform* 2018;**19**:61–76. https://doi.org/10.1186/s12859-018-2182-6
24. Lee SJ, Rho M. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Sci Rep* 2022;**12**:824. https://doi.org/10.1038/s41598-022-04773-3
25. Yan Miao F, Liu TH, Liu Y. Virtifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* 2022;**38**:1216–22.
26. Liang Q, Bible PW, Liu Y. *et al.* Deepmicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics Bioinform* 2020;**2**:lqaa009.
27. Yan H, Bombarely A, Li S. Deepte: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* 2020;**36**:4269–75. https://doi.org/10.1093/bioinformatics/btaa519
28. Miller D, Stern A, Burstein D. Deciphering microbial gene function using natural language processing. *Nat Commun* 2022;**13**:5731. https://doi.org/10.1038/s41467-022-33397-4
29. Wichmann A, Buschong E, Müller A. *et al.* Metatransformer: deep metagenomic sequencing read classification using self-attention models. *NAR Genomics Bioinformat* 2023;**5**:lqad082.
30. Hoarfrost A, Aptekmann A, Farfañuk G. *et al.* Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat Commun* 2022;**13**:2606. https://doi.org/10.1038/s41467-022-30070-8
31. Gwak H-J, Rho M. Vibe: a hierarchical bert model to identify eukaryotic viruses using metagenome sequencing data. *Brief Bioinform* 2022;**23**:bbac204.
32. Yang M, Huang L, Huang H. *et al.* Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res* 2022;**50**:e81–1. https://doi.org/10.1093/nar/gkac326
33. Avsec Ž, Agarwal V, Visentin D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 2021;**18**:1196–203. https://doi.org/10.1038/s41592-021-01252-x
34. Zvyagin M, Brace A, Hippe K. *et al.* Genslms: genome-scale language models reveal sars-cov-2 evolutionary dynamics. *Int J High Perform Comput Appl* 2022;**37**:683–705. https://doi.org/10.1177/10943420231201154
35. Ji Y, Zhou Z, Liu H. *et al.* Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* 2021;**37**:2112–20. https://doi.org/10.1093/bioinformatics/btab083
36. Zhou Z, Ji Y, Li W. *et al.* Dnabert-2: efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006. 2023.
37. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* 2024;**21**:1–11. https://doi.org/10.1038/s41592-024-02523-z
38. Nguyen E, Poli M, Faizi M. *et al.* Hyenadna: long-range genomic sequence modeling at single nucleotide resolution. *Adv Neural Inf Process Syst.* 2023;**36**:43177–201.
39. Wolf T, Debut L, Sanh V. *et al.* Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
40. Fishman V, Kuratov Y, Shmelev A. *et al.* Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Res* 2025;**53**:gkae1310.
41. Ma M, Liu G, Cao C. *et al. Hybridna: A Hybrid Transformer-mamba2 Long-Range Dna Language Model* 2025. https://doi.org/10.48550/arXiv.2502.10807
42. Cheng W, Song Z, Yang Z. *et al.* Dnalongbench: a benchmark suite for long-range dna prediction tasks. *bioRxiv* 2025;2025–01. https://doi.org/10.1101/2025.01.06.631595
43. Gündüz HA, Mreches R, Moosbauer J. *et al.* Optimized model architectures for deep learning on genomic data. *Commun Biol* 2024;**7:516**. https://doi.org/10.1038/s42003-024-06161-1
44. Pavlopoulos GA, Baltoumas FA, Liu S. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* 2023;**622**:594–602. https://doi.org/10.1038/s41586-023-06583-7
45. D'Onofrio DJ, Abel DL. Redundancy of the genetic code enables translational pausing. *Front Genet* 2014;**5**:140. https://doi.org/10.3389/fgene.2014.00140
46. Salgado H, Martínez-Flores I, Bustamante VH. *et al.* Using regulondb, the *Escherichia coli* k-12 gene regulatory transcriptional network database. *Curr Protoc Bioinformatics* 2018;**61**:1–32. https://doi.org/10.1002/cpbi.43
47. Jia B, Raphenya AR, Alcock B. *et al.* Card 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;**45**:D566–73.
48. Chen L, Yang J, Jun Y. *et al.* Vfdb: a reference database for bacterial virulence factors. *Nucleic Acids Res* 2005;**33**:D325–8.
49. Bairoch A. The enzyme database in 2000. *Nucleic Acids Res* 2000;**28**:304–5. https://doi.org/10.1093/nar/28.1.304
50. McDonald AG, Tipton KF. Enzyme nomenclature and classification: the state of the art. *The FEBS Journal*. Chichester, UK: Wiley Online Library, 2023;**290**:2214–31. https://doi.org/10.1111/febs.16494
51. Gillespie JJ, Wattam AR, Cammer SA. *et al.* Patric: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* 2011;**79**:4286–98. https://doi.org/10.1128/IAI.00207-11

52. Qichao T, Lin L, Cheng L. *et al.* Ncycdb: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics* 2019;**35**:1040–8.
53. Kusters KA, Pratsinis SE, Thoma SG. *et al.* Ultrasonic fragmentation of agglomerate powders. *Chem Eng Sci* 1993;**48**:4119–27. https://doi.org/10.1016/0009-2509(93)80258-R
54. Lin Z, Akin H, Rao R. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022;**2022**:500902.
55. Gruenstaeudl M. annonex2embl: automatic preparation of annotated dna sequences for bulk submissions to ena. *Bioinformatics* 2020;**36**:3841–8. https://doi.org/10.1093/bioinformatics/btaa209
56. McWilliam H, Li W, Uludag M. *et al.* Analysis tool web services from the embl-ebi. *Nucleic Acids Res* 2013;**41**:W597–600. https://doi.org/10.1093/nar/gkt376
57. Lawson CL, Swigon D, Murakami KS. *et al.* Catabolite activator protein: Dna binding and transcription activation. *Curr Opin Struct Biol* 2004;**14**:10–20. https://doi.org/10.1016/j.sbi.2004.01.012
58. Jain S, Hassanzadeh FF, Schwartz M. *et al.* Duplication-correcting codes for data storage in the dna of living organisms. *IEEE Trans Inform Theory* 2017;**63**:4996–5010. https://doi.org/10.1109/TIT.2017.2688361
59. Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. *Hum Mol Genet* 2006;**15**:R9–R16. https://doi.org/10.1093/hmg/ddl044
60. Zhang Z, Han Z, Yuqing W. *et al.* Metagenomics assembled genome scale analysis revealed the microbial diversity and genetic polymorphism of *Lactiplantibacillus plantarum* in traditional fermented foods of Hainan, China. *Food Res Int* 2021;**150**:110785. https://doi.org/10.1016/j.foodres.2021.110785
61. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
62. Yang Y, Jerger A, Feng S. *et al.* Improved enzyme functional annotation prediction using contrastive learning with structural inference. *Commun Biol* 2024;**7:1690**. https://doi.org/10.1038/s42003-024-07359-z
63. Khosla P, Teterwak P, Wang C. *et al.* Supervised contrastive learning. *Adv Neural Inform Process Syst* 2020;**33**:18661–73.
64. Mikhael PG, Chinn I, Barzilay R. Clipzyme: reaction-conditioned virtual screening of enzymes. In: *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*. Vienna, Austria: JMLR.org; 2024, Article No. 1453, pp. 1–17.
65. Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737. 2017. https://www.sci-hub.ru/10.48550/arXiv.1703.07737
66. Haigen H, Wang X, Zhang Y. *et al.* A comprehensive survey on contrastive learning. *Neurocomputing* 2024;**610**:128645. https://doi.org/10.1016/j.neucom.2024.128645
67. Naeve Z, Mitchell L, Reed C. *et al.* Introducing dynamic token embedding sampling of large language models for improved inference accuracy. *Authorea Preprints* 2024. https://doi.org/10.36227/techrxiv.173014793.37761346/v1
68. Karp PD, Billington R, Caspi R. *et al.* The biocyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2019;**20**:1085–93. https://doi.org/10.1093/bib/bbx085
69. Jian Z, Zeng L, Taojie X. *et al.* Antibiotic resistance genes in bacteria: occurrence, spread, and control. *J Basic Microbiol* 2021;**61**:1049–70. https://doi.org/10.1002/jobm.202100201
70. Arnold BJ, Huang I-T, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* 2022;**20**:206–18. https://doi.org/10.1038/s41579-021-00650-4
71. McArthur AG, Waglechner N, Nizam F. *et al.* The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013;**57**:3348–57. https://doi.org/10.1128/AAC.00419-13
72. Yang Y, Jiang X, Chai B. *et al.* Args-oap: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured arg-database. *Bioinformatics* 2016;**32**:2346–51. https://doi.org/10.1093/bioinformatics/btw136
73. Hong S, Pedersen PL. Atp synthase and the actions of inhibitors utilized to study its roles in human health, disease, and other scientific areas. *Microbiol Mol Biol Rev* 2008;**72**:590–641. https://doi.org/10.1128/MMBR.00016-08
74. Uniprot: The universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31. https://doi.org/10.1093/nar/gkac1052
75. Devlin J, Chang M-W, Lee K. *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics; 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
76. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. 2017. https://doi.org/10.48550/arXiv.1711.05101

Briefings in Bioinformatics, 2025, **26(2)**, bbaf149
https://doi.org/10.1093/bib/bbaf149
Problem Solving Protocol