Yang *et al. Intell. Robot.* **2025**, *5*(3), 541-61
DOI: 10.20517/ir.2025.28

Intelligence & Robotics

Research Article

Open Access

MicroGraphBERT: soil microbial gene sequence classification via fusing taxonomic hierarchies and DNABERT-based contextual embeddings

Han Yang[1], Di Wang[1], Wenjie Pan[2], Chaoying Jiang[2], Weichang Gao[3], Xiaoji Luo[1], Zugui Tu[4]

[1]School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China.
[2]China National Tobacco Corporation Guizhou Branch, Guiyang 550004, Guizhou, China.
[3]Guizhou Tobacco Science Research Institute, Guiyang 550081, Guizhou, China.
[4]College of Tobacco Science, Guizhou University, Guizhou 550025, Guizhou, China.

Correspondence to: Prof. Di Wang, School of Information Science and Engineering, Chongqing Jiaotong University, No. 66 Xuefu Avenue, Nan'an District, Chongqing 400074, China. E-mail: diwang@cqjtu.edu.cn; ORCID: 0000-0001-9679-7592

How to cite this article: Yang, H.; Wang, D.; Pan, W.; Jiang, C.; Gao, W.; Luo, X.; Tu, Z. MicroGraphBERT: soil microbial gene sequence classification via fusing taxonomic hierarchies and DNABERT-based contextual embeddings. *Intell. Robot.* **2025**, *5*(3), 541-61. http://dx.doi.org/10.20517/ir.2025.28

Received: 19 Mar 2025 **First Decision:** 7 May 2025 **Revised:** 7 Jun 2025 **Accepted:** 16 Jun 2025 **Published:** 27 Jun 2025

Academic Editor: Guoxian Yu **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Soil microbial communities are crucial for essential ecosystem functions such as nitrogen cycling and organic matter decomposition. However, accurately classifying their gene sequences remains challenging due to overlooked taxonomic hierarchies, environmental variability, and insufficient structural dynamics. Current methods predominantly focus on intra-sequence nucleotide features while neglecting the community's hierarchical taxonomy. To address these gaps, we analyzed soil samples collected from the loess regions of Guizhou and investigated dynamic changes in microbial community composition across plant growth stages. We propose MicroGraphBERT, a deep learning framework synergizing DNABERT's context-aware embeddings with taxonomy-aware priors via graph attention network to enable joint modeling of sequence and ecological features for microbial classification. Trained on high-throughput sequencing data from the Guizhou loess regions, MicroGraphBERT integrates nucleotide-level contextual semantics from DNABERT and cross-species relational learning with graph attention network to capture both sequence features and taxonomic hierarchies. This approach identifies complex microbial patterns under varying soil conditions, achieving a classification accuracy of 98.72%. Our work advances precision microbiome analytics by providing a scalable solution for soil health monitoring, intelligent fertilizer optimization, and sustainable agroecosystem management.

Keywords: Soil microbiome, taxonomic hierarchies, DNABERT, graph attention network, high-throughput sequencing

1. INTRODUCTION

1.1. Literature review

Soil microorganisms, including *bacteria*, *fungi*, *actinomycetes*, and *protozoa*, are fundamental drivers of soil health, fertility, and structural stability through their roles in organic matter decomposition and biogeochemical cycling[1,2]. Their metabolic diversity enables critical ecosystem functions: nitrogen fixation by rhizobia, phosphorus solubilization by mycorrhizal fungi, and sulfur oxidation by chemolithotrophs collectively enhance nutrient availability and soil organic carbon accumulation[3,4]. Specific functional guilds further illustrate this mechanistic diversity. For instance, *potassium-solubilizing bacteria* secrete organic acids to dissolve silicate minerals, thereby increasing plant-available potassium[5]; *actinomycetes* produce thiopeptide antibiotics that suppress fungal pathogens in the rhizosphere[6]; and *Trichoderma species* enhance plant resilience by synthesizing cell wall-degrading enzymes and inducing systemic resistance[7].

Importantly, the composition and activity of these microbial communities exhibit spatiotemporal variability shaped by environmental gradients and plant phenology[8]. For example, *Pseudomonas* populations dominate during legume flowering stages due to root exudate chemotaxis, while *methanotrophs* proliferate in waterlogged rice paddies. Such environment-community associations provide prior ecological knowledge that can constrain sequence classification models. By integrating taxonomic signatures of known habitat preferences, models improve accuracy when classifying sequences from novel or highly divergent taxa[9]. This context-aware approach is particularly critical in metagenomics, where horizontal gene transfer and convergent evolution blur phylogenetic boundaries. However, accurately classifying microbial sequences in complex soil ecosystems remains challenging due to fragmented analytical frameworks that dissociate sequence semantics from ecological context.

As depicted in Figure 1, conventional approaches fragment sequence analysis and ecological context. The Naive Bayes classifier[10] exemplified feature-driven approaches by constructing bag-of-words models from k-mer frequency statistics under the biologically implausible assumption of nucleotide positional independence. Although the random forest algorithm[11] marginally improved robustness through multi-dimensional feature integration, such frameworks systematically failed to generalize beyond curated datasets. Phymm[12], for instance, depended on predefined species-specific Hidden Markov Models (HMMs), rendering it ineffective for unannotated taxa, while the RDP classifier[13] exhibited inaccuracies in detecting horizontal gene transfer due to its reliance on conserved ribosomal RNA regions. Even hybrid tools such as SPHINX[14], which combined spectral clustering with k-mer and contig features, could not resolve contextual dependencies inherent in microbial community sequences.

Beyond feature engineering, sequence similarity-based approaches emerged as an alternative paradigm, yet they too grappled with fundamental trade-offs between accuracy and scalability. Global alignment algorithms[15] used dynamic programming for global matches but were limited to short sequences; local alignment tools[16] employed BLOSUM62 matrices to identify conserved motifs, though memory constraints hindered scalability in metagenomics. Heuristic methods such as BLAST[17] accelerated alignment via seed extension but exhibited elevated misclassification rates for low-similarity sequences. Multisequence aligners[18] built phylogenetic trees iteratively, yet indel sensitivity compromised reliability. Collectively, both feature-driven and alignment-dependent methods struggled to reconcile biological context with computational tractability, highlighting the need for prior knowledge-integrated approaches.

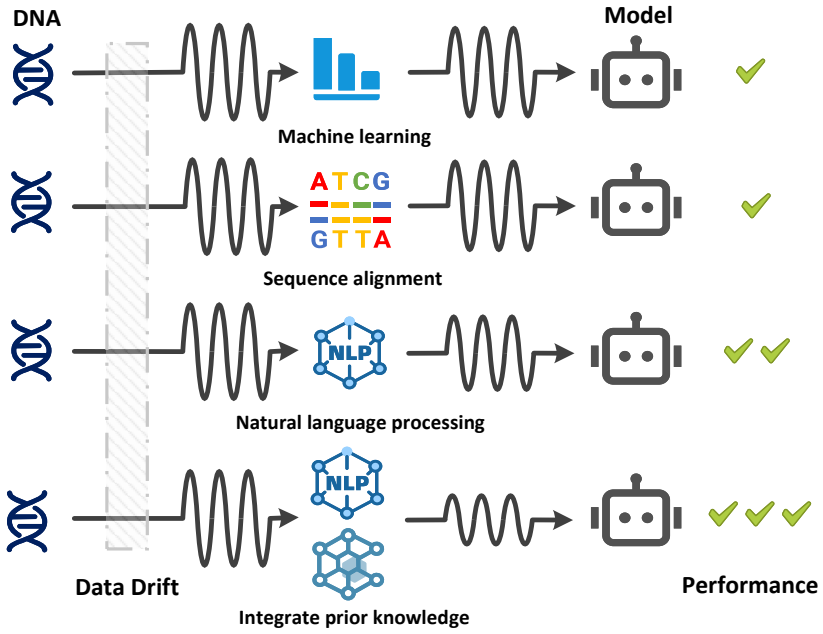

Figure 1. The motivation of microGraphBERT.

Currently, gene sequence classification often depends on features such as sequence patterns, local and global characteristics, context, and word embeddings[19–21]. Transformer architectures advance gene sequence analysis by integrating local sequence patterns and global structural characteristics through self-attention mechanisms, while preserving contextual semantics beyond static word embeddings. This enables unified modeling of nucleotide interactions across scales, overcoming fragmentation in traditional feature-based methods for complex microbial ecosystems. However, real-world microbial communities, influenced by environmental factors and growth stages, can significantly alter soil composition[22]. Thus, developing models that incorporate prior knowledge of microbial communities is crucial for enhancing robustness and accuracy in complex ecosystems. For the reader's convenience, a complete list of the mathematical notations used throughout this paper is provided in Table 1.

1.2. Contribution

Due to the limitations in current microbial community analysis methods, MicroGraphBERT was proposed in this study, which leverages the prior knowledge of microbial community composition and function to efficiently identify soil microorganisms. Initially, the DNABERT model was fine-tuned using microbial gene data from the loess regions of Guizhou, allowing it to adapt to specific gene sequence features[23]. Subsequently, embedding features of the gene sequences were extracted from the fine-tuned DNABERT model, capturing nuanced contextual information within the sequences. Next, based on comprehensive statistical analysis, a hierarchical classification network was constructed by establishing taxonomic relationships among gene sequences to reflect the complex structure of microbial communities. Thereafter, the features of this network were updated using graph attention network (GAT)[24] to extract characteristics of microbial community structure. Finally, the classification results were output based on the integrated features. The main contributions of this paper in comparison with the extant literature are summarized as follows:

(1) This study presents the first multimodal fusion framework that jointly models DNA sequence semantic features and microbial community structures. Contextual semantic features of nucleotide sequences are captured

Table 1. Symbol table for microGraphBERT framework

Symbol	Definition
S	Raw gene sequences (e.g., 16S rRNA sequences)
$\mathcal{H}$	Taxonomic hierarchy tree
$\mathcal{Y}$	Numeric labels
$\hat{S}$	Filtered ASV table
$\mathcal{G}$	Hierarchical graph
$\tilde{S}$	Balanced dataset (SMOTE + undersampling)
DNABERT_θ	Pre-trained DNABERT-2 model
$\mathbf{X}_{\text{DNA}}$	BPE-tokenized DNA sequence matrix
$\mathbf{H}_{\text{DNA}}$	Contextual embeddings
$\mathbf{F}_{\text{DNA}}$	Global pooling features
GAT_ϕ	Hierarchical graph attention network
h_i^l	Node i feature at layer l
h_i^{l+1}	Updated node i feature
α_{ij}^m	Attention weight (head m, nodes i-j)
$\mathbf{W}^m$	Weight matrix (head m)
a	Learnable attention vector
$\sigma(\cdot)$	LeakyReLU activation
$\|\cdot\|$	Concatenation
$\mathbf{F}_{\mathcal{G}}$	Aggregated graph features
$\mathbf{F}_{\text{align}}$	Aligned DNA-graph features
$\mathbf{F}_{\text{fuse}}$	Concatenated fused features
$\hat{y}$	Predicted probability distribution
L	Total loss including Cross-entropy loss and L2 regularization
AdamW	Optimizer

through fine-tuning of a pre-trained DNABERT, while cross-species topological associations across taxonomic hierarchies are dynamically learned by a taxonomy-aware GAT. The limitations of conventional single-modal analyses are overcome by synergistic integration of natural language processing for sequence interpretation and graph neural networks (GNNs) for association modeling. Functional gene annotation and multi-level taxonomic feature extraction are jointly optimized, establishing a new methodological paradigm for computational biology.

(2) This study develops a hierarchical GNN architecture informed by taxonomic priors, where multi-level evolutionary topologies spanning from domain to species are explicitly embedded to construct a hierarchical graph adjacency matrix that encodes microbial phylogenetic constraints. During feature propagation, a multi-head graph attention mechanism is employed, guided by the adjacency matrix to mediate node-wise information interactions while dynamically learning association weights across taxonomic hierarchies. This framework achieves unified modeling of gene sequence semantic features and taxonomic graph structures through systematic integration.

2. METHODS

2.1. Data collection

The data was collected from specific loess zones in the Guizhou ecological region, which is characterized by high microbial diversity and favorable ecological conditions, making it well-suited for tobacco cultivation [25]. Loess provides optimal conditions for tobacco growth due to its favorable physical and chemical properties, including a moderate pH, high organic matter content, and strong water and nutrient retention. These characteristics make the loess regions in Guizhou an ideal setting for exploring the intricate relationship between soil microorganisms and tobacco growth.

To further investigate this relationship, soil samples were collected during nine plant growth stages within the

Figure 2. Tobacco soil sampling illustration. (A) Panoramic view of the tobacco field; (B) Individual tobacco plant; (C) Soil sampling in the tobacco field; (D) Soil sample processing.

same plot, including bud formation, vigorous growth, lower leaf maturity, middle leaf maturity, upper leaf maturity, senescence, and two additional stages. At each stage, both rhizosphere and bulk soil (BS) samples were collected, with six samples per stage, resulting in a total of 114 samples. Additionally, soil samples were collected after tobacco leaf harvest to serve as a control group for comparison with samples from other growth stages. The schematic for tobacco soil sampling is shown in Figure 2.

2.2. Data exploration

After sample processing, amplicon sequence variant (ASV) data representing unique species information was obtained for each sample. To investigate soil microbial characteristics, a systematic analysis of microbial community dynamics across the plant growth cycle was performed. Rhizosphere soil (RS) and BS were designated in this study, with numerical labels indicating days post-planting. Comparative analysis of these two soil types across five ecological parameters yielded five figure sets that elucidate their compositional differences and functional relationships.

Based on species annotation, the top 10 phyla by relative abundance were selected, with others categorized as "Others". Figure 3 shows significant differences in microbial community composition between rhizosphere and BS were observed. Specifically, the relative abundance of *Firmicutes* increased while *Acidobacteriota* decreased in RS. This shift is likely due to root exudates rich in sugars and organic acids during the plant growth stages between days 45 and 85. These compounds promote *Firmicutes* growth while inhibiting *Acidobacteriota*. Overall, bacterial compositions in both soil types changed significantly across the tobacco growth cycle, with rhizosphere RS exhibiting more pronounced shifts, highlighting the regulatory role of root exudates.

The heatmap generated based on the abundance information of the top 35 genera, as shown in Figure 4, revealed significant differences in microbial communities between RS and BS. The analysis showed that the relative abundances of *Firmicutes*, *Proteobacteria*, *Actinobacteriota*, *Bacteroidota*, and *Cyanobacteria* were higher in RS than in BS. These bacterial groups can utilize organic substances secreted by plant roots, such as carbohydrates and amino acids, as nutrient sources, thereby gaining a competitive advantage in the rhizosphere environment. The increased abundance of Cyanobacteria in some rhizosphere samples may be related to favorable microenvironmental conditions such as light and moisture, suggesting the presence of photosynthetic activity.

In contrast, the microbial community in BS exhibited higher stability, with their higher relative abundances of *Acidobacteriota*, *Gemmatimonadota*, and *Chloroflexi*. This is likely due to their ability to adapt to nutrient-poor environments or low organic matter content. The smaller variations in the microbial community of BS reflect the relative stability of its environmental conditions and the reduced influence of plant roots.

The petal diagram, as illustrated in Figure 5, analyzed the quantity of microbial feature sequences in RS and BS across different growth stages.

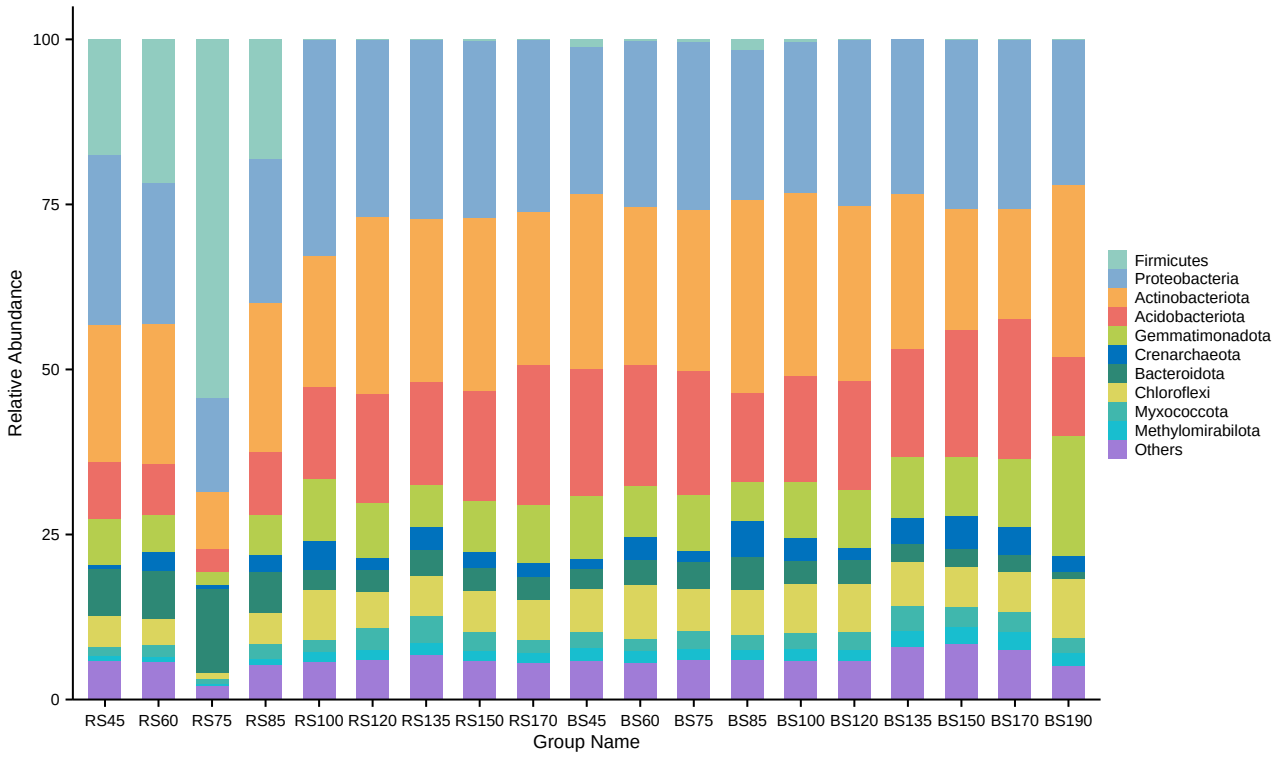

Figure 3. The bar chart of species abundance at phylum-level.

The central part of the diagram represents the microbial feature sequences shared among all stages, while the petals depict the sequences unique to each stage. The diagram shows a gradual decline in the rhizosphere feature sequences from RS45 to RS100, followed by an increase between RS120 and RS170. In contrast, BS maintained higher sequence numbers throughout the growth stages, with a significant rise during RS120 to RS170. This highlights the dynamic response of rhizosphere microbes to plant growth stages, in contrast to the relatively stable BS microbial community.

Phylogenetic trees constructed via multiple sequence alignment of soil samples from different plant growth stages, as shown in Figure 6, revealed that*Proteobacteria*, *Firmicutes*, and *Bacteroidota* were dominant, with particularly high abundances of *Firmicutes* and *Bacteroidota* in the RS75 sample. This suggests active microbial roles in nitrogen cycling, organic matter decomposition, plant-rhizosphere symbiosis, and complex organic matter breakdown during this stage.

UPGMA-based clustering analysis, as shown in Figure 7, revealed high similarity in microbial composition among RS samples, while BS exhibited stable microbial community structures across different growth stages.

These results suggest that rhizosphere microbial dynamics are closely linked to plant growth stages, likely due to selective pressures from root exudates and shifts in rhizosphere microenvironmental conditions. In contrast, the stability of BS microbial communities is primarily regulated by inherent soil physicochemical properties such as pH, organic matter content, moisture, and temperature. In contrast, the stability of BS microbial communities is primarily regulated by inherent soil physicochemical properties such as pH, organic matter content, moisture, and temperature.

The analysis results indicate that soil microbial communities exhibit significant dynamic changes during the plant growth cycle, driven by a combination of plant root activities, soil environmental conditions, and mi-

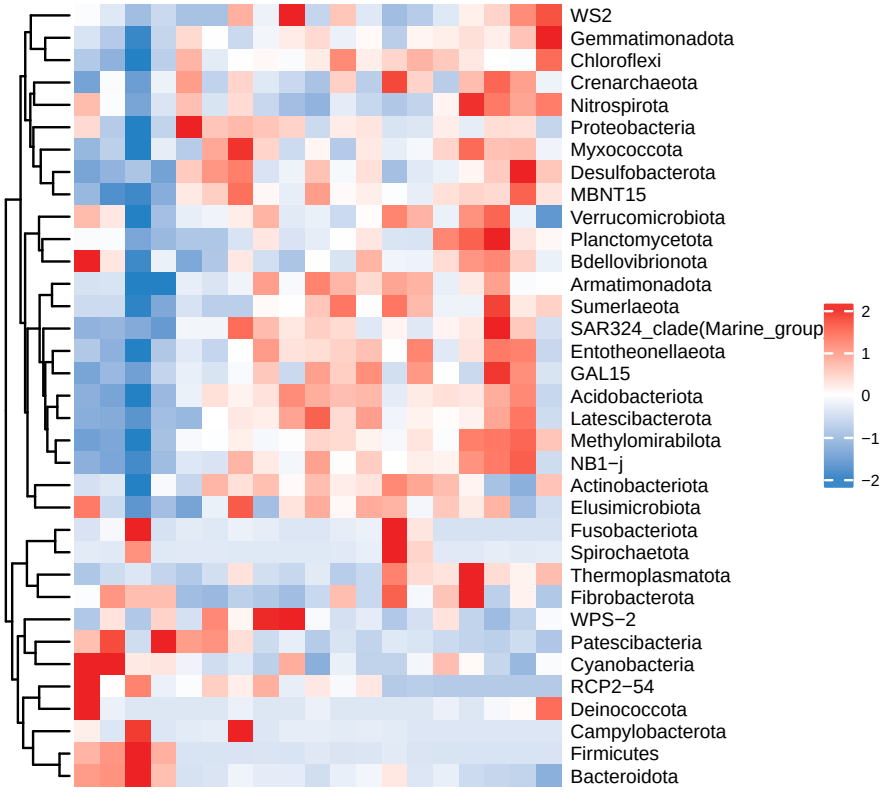

Figure 4. The cluster heatmap of species abundance at phylum-level.

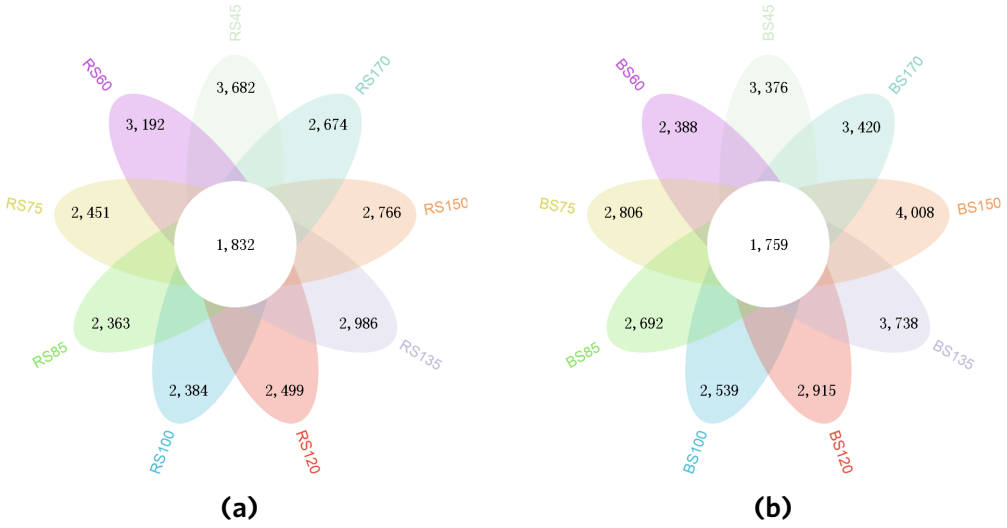

Figure 5. The petal diagram of soil at different growth stages. (A) RS; (B) BS. RS: Rhizosphere soil; BS: bulk soil.

crobial ecological functions. These changes present challenges for microbial classification and identification, including uneven data distribution, difficulties in classifying low-abundance microorganisms, and interference from environmental factors.

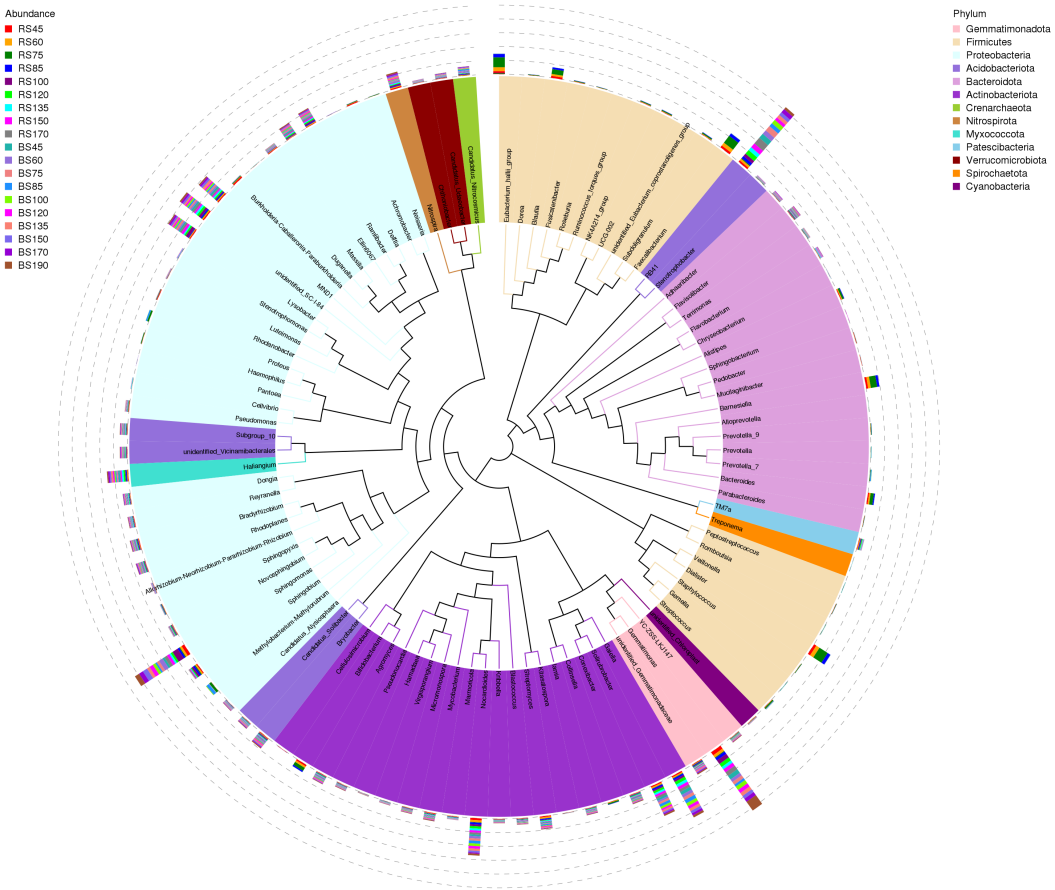

Figure 6. The phylogenetic tree of genera at different growth stages.

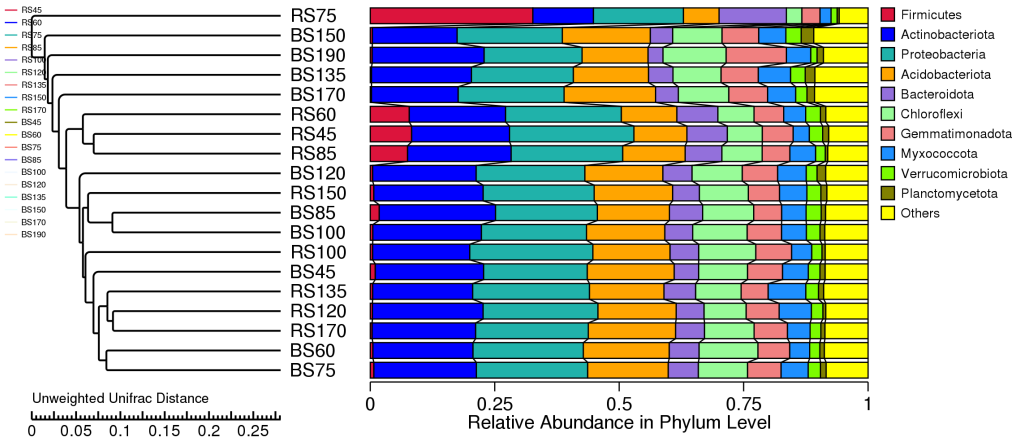

Figure 7. The UPGMA clustering tree.

2.3. Microbial genomic sequence dataset

The dataset, derived from soil samples, includes 126,078 high-quality gene sequences classified into 486 taxonomic categories. It comprises three key components: gene sequences, microbial hierarchical classification relationships, and numerical labels. The hierarchical relationships depict microbial evolution from domain to

Table 2. Dataset format

Sequence	Classification	Label
TGGGGAATATTGCGCAGG...	k__Bacteria;p__Basidiomycota...	0
TTTCCGTAGGTGAACCAG...	k__Fungi;p__Chytridiomycota...	1
GCGAGAAACCTTAGCACT...	k__Archaea;p__Crenarchaeota...	2
...	...	...

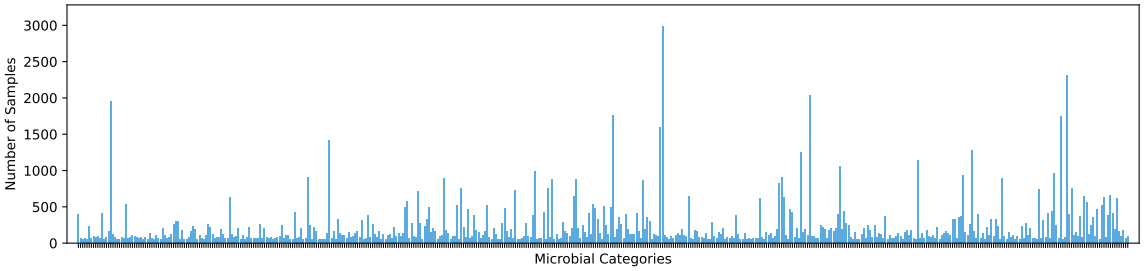

Figure 8. Classification distribution in the microbial dataset.

Table 3. Distribution of unique categories and total instances across different taxonomic ranks in the dataset

Ranks	Unique categories	Total instances
Phylum	43	106,251
Class	102	99,231
Order	183	86,909
Family	229	66,999
Genus	122	23,449
Custom taxonomy	486	126,078

species, while numerical labels denote different categories. Data format details are in Table 2. The dataset was randomly divided into training (70%), validation (10%), and testing sets (20%). Due to its origin from actual soil samples, the dataset exhibits significant regional characteristics and microbial composition diversity, resulting in data imbalance. As shown in Figure 8, the sample sizes of certain microbial categories are far higher than others, which manifests as a pronounced class imbalance problem.

To address the imbalance shown in Figure 8, the SMOTE [26] method was applied. It combines over-sampling of minority classes and under-sampling of majority classes to generate synthetic samples and achieve balance. This method ensures that both majority and minority classes are adequately represented during training, thereby enhancing model generalization and performance.

Table 3 provides a detailed breakdown of the quantity of gene sequences and the corresponding counts of classifications across various hierarchical levels. This information is crucial for understanding the distribution and diversity of gene sequences within the dataset, offering insights into the complexity and structure of the microbial community under investigation.

2.4. The proposed framework

This study has developed a high-precision microbial sequence classification model, named MicroGraphBERT, to accurately distinguish between different microbial categories. Existing methods primarily rely on gene sequences, which are susceptible to various factors that reduce classification accuracy. To address this issue, a new framework has been proposed that integrates gene sequence features with the hierarchical classification

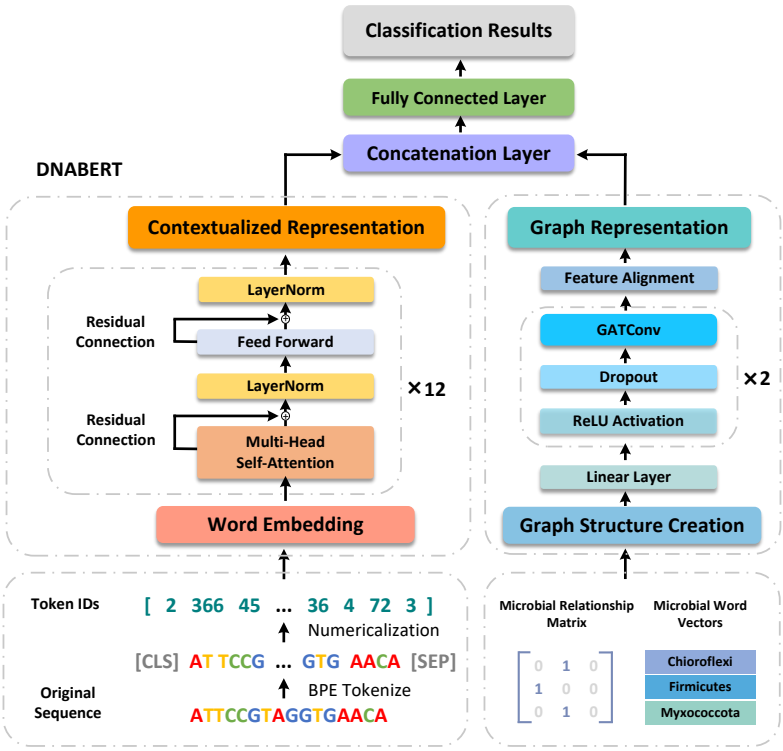

Figure 9. The diagram illustrates the architecture of MicroGraphBERT.

network of microbial communities to enhance model performance, as shown in Figure 9.

MicroGraphBERT Soil Microbial Classification Framework integrates gene sequence and microbial community structure features through a multimodal fusion mechanism to achieve complementary representations. The pre-trained DNABERT-2 model performs contextual semantic encoding on raw gene sequences, leveraging a multi-layer Transformer architecture to capture long-range dependencies in nucleotide sequences and generate global feature representations. This process involves tokenization, embedding, and pooling operations to extract sequence-level biological information.

Concurrently, microbial community structure features are modeled using a hierarchical GNN, where a taxonomy-based graph structure is constructed, and node features are iteratively propagated via the GAT to aggregate graph-level global representations.

The two feature modalities are aligned through an attention mechanism to eliminate modality-specific distribution discrepancies, followed by concatenation to form fused features. These fused features are then fed into a multi-layer perceptron (MLP) for classification, with cross-entropy loss combined with L2 regularization forming the optimization objective. To enhance model robustness, training employs a hybrid strategy of SMOTE oversampling and random undersampling, while parameter updates are dynamically adjusted using the AdamW optimizer. The computational workflow is presented as pseudo-code in Algorithm 1, where each critical operation is annotated with technical specifications. Corresponding mathematical notations are rigorously documented in Table 1 to ensure symbolic consistency.

Algorithm 1 MicroGraphBERT Soil Microbial Classification Framework

Require:

Raw gene sequences S, taxonomy $\mathcal{H}$, labels $\mathcal{Y}$

Ensure:

Predicted label y and community embeddings

1: **Preprocessing**
- Extract ASV table and filter $\boxtimes \hat{\mathcal{S}}$
- Build graph $\mathcal{G}$ from $\mathcal{H}$
- Balance data $\boxtimes \tilde{\mathcal{S}}$ (SMOTE + undersampling)

2: **DNABERT Feature Extraction** (DNABERT$_\theta$)
- Tokenize (BPE) $\boxtimes \mathbf{X}_{\text{DNA}}$
- Embed $\boxtimes \mathbf{H}_{\text{DNA}}$
- Pool $\boxtimes \mathbf{F}_{\text{DNA}}$

3: **Hierarchical GNN** (GAT$_\phi$)

4: **for** each layer $l \in \{1, \ldots, L\}$ **do**

5: Propagate features:

$$h_i^{l+1} = \|_{m=1}^{M} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^m W^m h_j^l\right) \quad \text{where } \alpha_{ij} = \text{Softmax}(\text{LeakyReLU}(a^T [WH_i || WH_j]))$$

6: **end for**

7: Aggregate graph features $\boxtimes \mathbf{F}_{\mathcal{G}}$

8: **Multimodal Fusion**
- Align $\mathbf{F}_{\text{DNA}}$ and $\mathbf{F}_{\mathcal{G}} \boxtimes \mathbf{F}_{\text{align}}$
- Concatenate $\boxtimes \mathbf{F}_{\text{fuse}}$
- Classify via MLP $\boxtimes \hat{y}$

9: **Training**
- Total loss: $L = \text{CrossEntropy}(\hat{y}, y) + \lambda \|\theta\|_2^2$
- Update parameters (AdamW)

2.5. Fine-tuning DNABERT

DNABERT is a BERT-derived model tailored for genomic sequence data. It uses transformer-based encoders and employs k-mer or byte pair encoding (BPE) strategies to segment sequences. Unlike BERT, sentence embeddings are removed.

In DNABERT, the attention mechanism and multi-head attention are core components. The attention mechanism dynamically focuses on different parts of the input sequence, capturing long-range dependencies, while multi-head attention enables parallel computation of multiple attention heads, enabling the model to capture diverse features across distinct subspaces. This enhances pattern recognition and understanding of biological sequences. These mechanisms are shown as formulas (1), (2) and (3):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$head_i = Attention(XR_i^Q, XR_i^K, XR_i^V) \tag{2}$$

$$MultiHead(Q, K, V) = concat(head_1, head_2, \ldots, head_n) \cdot R \tag{3}$$

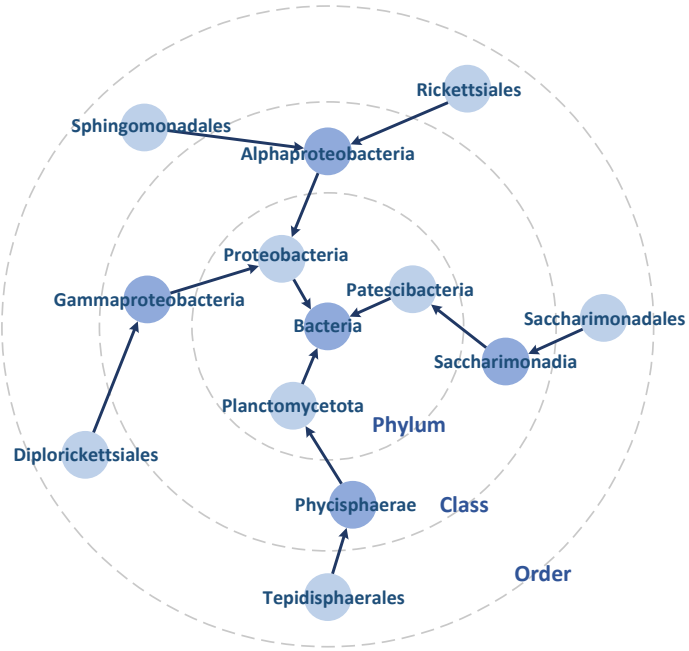

Figure 10. Microbial hierarchical network.

where R_i^Q, R_i^K, R_i^V are learnable parameters corresponding to the query, key, and value matrices of the i-th attention head, respectively. Q represents the information to be retrieved, K is used for matching with the Q, and V contains the actual data content that is weighted and aggregated based on the matching results between the Q and K. The index i indicates the i-th attention head, n represents the total number of attention heads, d_k is the dimension of the key vectors serving as a scaling factor to prevent excessively large values during similarity computation, and R is the final linear transformation weight matrix used to map the concatenated output to the desired dimensions.

This study used DNABERT-2-117M to generate high-quality DNA embeddings for clustering and classifying genes based on their similarities and differences. Trained on a large multi-species genomic dataset, the model was fine-tuned on soil microbiome data to capture unique features. DNABERT-2 employs BPE for word embeddings and incorporates the ALiBi mechanism, which adjusts attention weights through linear bias, improving adaptability to sequences of variable lengths. The model converts gene sequences into 768-dimensional vectors, with the [CLS] token capturing key information for effective classification.

2.6. Extracting hierarchical features with GAT

A soil microbial dataset was analyzed, revealing that microbial communities are influenced by plant root activities, soil conditions, and ecological niches. These factors cause dynamic community changes and challenges in gene sequence identification, such as inconsistent data distribution, difficulty in classifying low-abundance microbes, and environmental interference. To address these issues, a hierarchical graph network representing microbial taxonomy from domain to species was constructed and integrated into the classification model as prior knowledge, as shown in Figure 10. The network defines nodes as microbial taxonomic entities such as phylum, class, and species, representing hierarchical classifications from domain to species. Edges encode taxonomic relationships such as parent-child hierarchy and functional associations.

We processed the graph network using GAT, updating node vectors to effectively reflect soil microbial com-

munity characteristics. It also calculated edge attention coefficients to reveal relationships among microbial classifications. Despite potential data drift caused by environmental factors, the GAT-derived network can be integrated into the classification model as prior knowledge, enhancing model robustness by leveraging both gene sequence similarity and microbial community structure.

Adaptive attention and information transmission within the GAT capture node relationships to extract discriminative features. By dynamically weighting neighbor nodes, it enhances performance in heterogeneous graphs and complex tasks. Widely used in biological networks[27], traffic prediction[28], and other[29], GAT calculates attention using the coefficients of three key formulas, as shown in formulas (4), (5), and (6):

$$e_{ij} = \text{LeakyReLU}(a^T [Wh_i || Wh_j]) \tag{4}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \tag{5}$$

$$h_i^{l+1} = \|_{m=1}^{M} \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^m W^m h_j^l \right) \tag{6}$$

where a is a learnable attention vector, and W is a learnable weight matrix, e_{ij} is the attention score between node i and node j, h_i and h_j are the feature vectors of node i and node j, $||$ is the concatenation operation, $LeakyReLU(\cdot)$ is the activation function used to introduce non-linearity, α_{ij} represents the attention score between node i and node j, $\mathcal{N}(i)$ is the set of neighboring nodes of node i, M is the attention head number, and l is the number of layers of the GAT.

2.7. Performance evaluation metrics

The performance of multi-class models is evaluated using key metrics such as accuracy, precision, recall, and F1 score. These metrics, derived from a statistical analysis of model predictions, offer insights into performance and reliability. Accuracy reflects the proportion of correctly classified instances across all categories, but may be misleading in imbalanced datasets. Precision and recall provide important supplementary evaluations, while the F1 score, as the harmonic mean of precision and recall, offers a comprehensive assessment. Their calculations are shown in formulas (7), (8), (9) and (10):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1\_Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

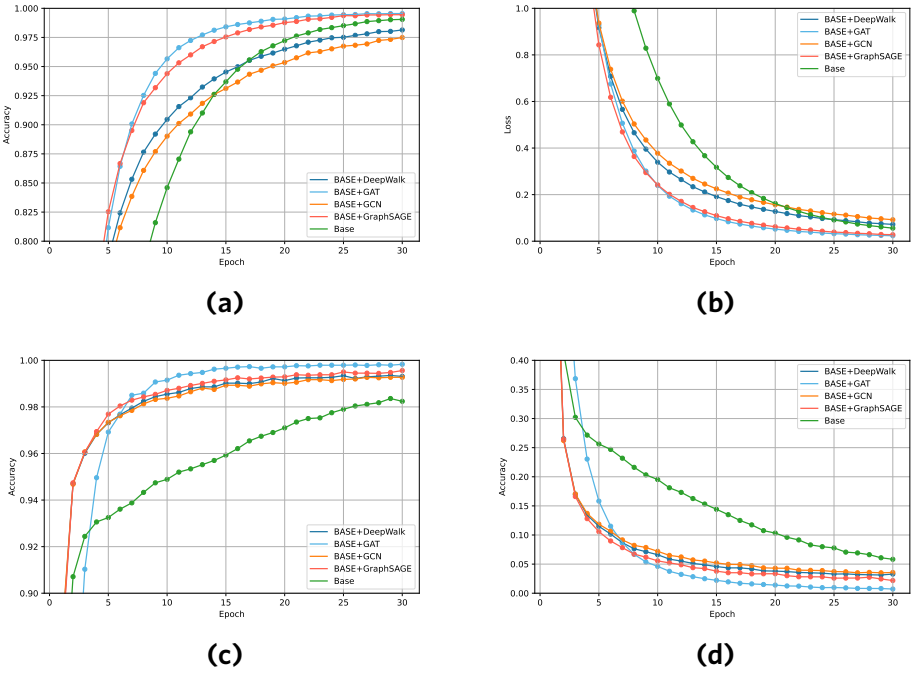

Figure 11. Performance comparison of the ablation study. (A) Training accuracy comparison; (B) Training loss comparison; (C) Validation accuracy comparison; (D) Validation Loss comparison

where TP represents the number of instances correctly predicted as positive, which are indeed positive, FP signifies the number of instances incorrectly predicted as positive, which are actually negative, FN indicates the number of positive instances that the model incorrectly predicted as negative, and TN refers to the number of instances correctly predicted as negative, which are indeed negative.

3. RESULTS AND DISCUSSION

To validate the effectiveness of the MicroGraphBERT framework, a series of experiments was conducted. First, an ablation study was performed to evaluate the contribution of key modules. Subsequently, MicroGraphBERT was compared with the baseline models and other GNN methods. Furthermore, the dynamic evolution of microbial community structures across plant growth stages was analyzed. The results indicated that MicroGraphBERT achieved competitive classification accuracy and effectively captured hierarchical taxonomic relationships.

3.1. Comparative and ablation study results

To systematically evaluate the contribution of key architectural components and validate the robustness of the proposed framework, we designed a comprehensive ablation study. The ablation study compared MicroGraphBERT with other graph-based methods, including GCN, GraphSAGE, and DeepWalk, as well as a baseline model without feature processing. Specifically, DeepWalk captures global structural features via random walks, GraphSAGE aggregates local features from sampled neighbors, and GCN smooths node features through graph convolution with neighboring nodes.The results are presented in Figures 11 and 12.

The analysis of Figure 11 showed that the model with GAT achieved higher accuracy on both the training and validation sets and had the fastest rate of loss reduction. This indicates the strong capability of GAT in learning data features and generalization. GraphSAGE ranked second with good accuracy, suggesting its method of aggregating neighboring node features effectively captures the local structure of microbial communities. In

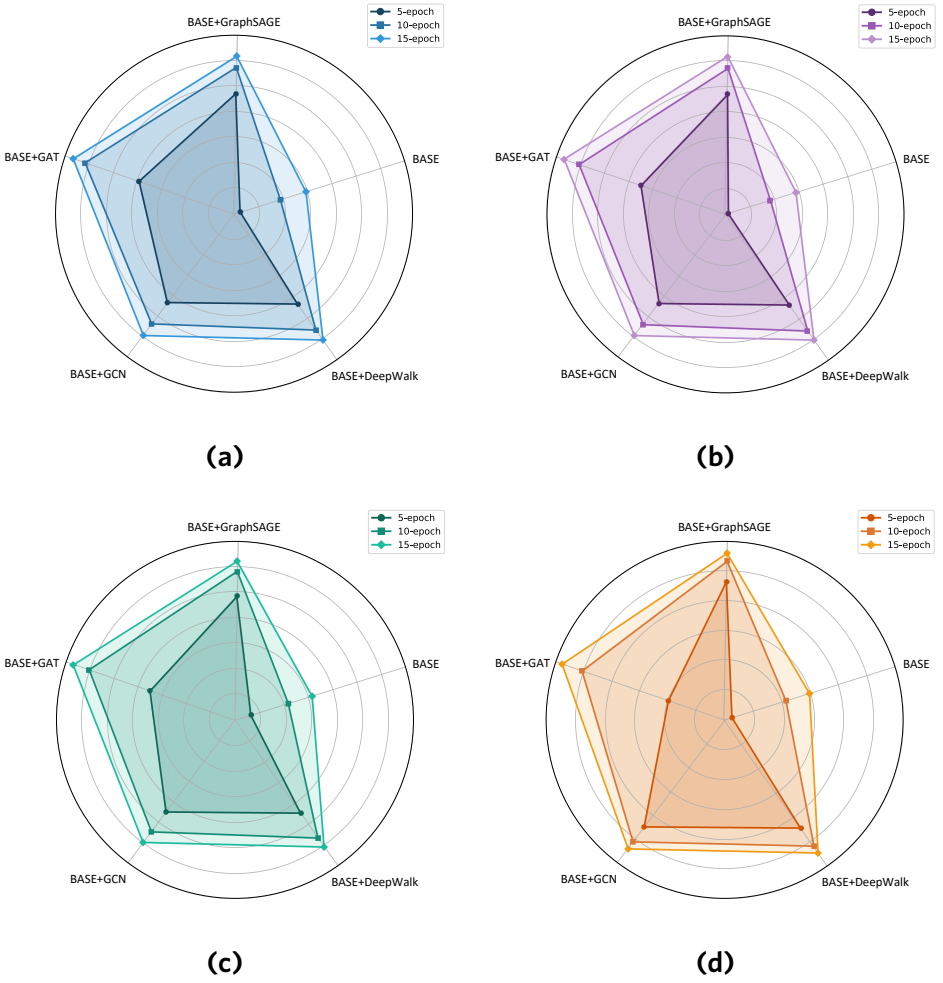

Figure 12. Radar chart of performance metrics across multiple epochs in the ablation study. (A) Accuracy radar chart; (B) Precision radar chart; (C) Recall radar chart; (D) F1 score radar chart.

contrast, the BASE model, without feature processing, showed poorer performance on the validation set, likely due to its lack of specialized handling for microbial community structure features, a limitation that makes it less effective for complex data patterns.

The radar chart in Figure 12 further compares the accuracy, precision, recall, and F1 score of each experiment after 5, 10, and 15 training epochs. It can be seen that the model processed using GAT consistently performed better than the others across all metrics, while the BASE model had the lowest performance. Overall, these results highlight the effectiveness of GAT in handling microbial community structure features.

To further validate the performance of the proposed model, the trained model was evaluated on an independent test set to comprehensively assess its classification capabilities. As shown in Table 4, the GAT model achieved the highest accuracy and F1 score across all metrics, with an accuracy of 98.72%, precision of 98.53%, recall of 98.50%, and an F1 score of 98.64%, outperforming other models.

This result demonstrated the effectiveness of GAT in handling microbial community structure features and its strong generalization ability. In contrast, the BASE model performed the worst; these results highlighted the importance of feature processing in improving model performance.

Table 4. Ablation study results of model performance in microbial classification

Model	Accuracy (%)	Precision (%)	F1 Score (%)	Recall (%)
BASE	96.89	97.17	96.84	96.92
BASE + DeepWalk	97.85	98.13	98.39	98.43
BASE + GraphSAGE	98.36	98.40	98.41	98.45
BASE + GCN	98.14	98.12	98.18	98.21
BASE + GAT	**98.72**	**98.53**	**98.64**	**98.50**

Table 5. Ablation study results of model performance across taxonomic tanks in microbial classification accuracy (%)

Model	Phylum	Class	Order	Family	Genus	Average
BASE	99.17	98.01	97.15	97.87	96.80	97.80
BASE + DeepWalk	99.28	98.17	97.05	97.67	97.16	97.86
BASE + GraphSAGE	99.40	98.08	97.37	98.00	97.21	98.01
BASE + GCN	99.63	98.15	97.28	98.08	97.06	98.04
BASE + GAT	**99.80**	**98.64**	97.26	**98.18**	97.12	**98.20**

To further explore the performance of the models in microbial classification tasks, the dataset was processed into five hierarchical classification tasks - phylum, class, order, family, and genus, thereby facilitating an in-depth analysis of the performance of the models at each classification level. The experimental results are shown in Table 5.

The results indicate that the GAT-based model achieved the highest accuracy, reaching 99.80% at the phylum level, 98.64% at the class level, and 98.18% at the genus level, with an overall average of 98.20%. These figures surpass those of other methods, underscoring the effectiveness of the GAT-based model in microbial classification. However, the classification accuracy at the order level was observed to be lower than the accuracy at the family level in the experiments.

To investigate the reasons for this phenomenon, the distribution of classification counts across different taxonomic ranks was visualized using bar charts, as shown in Figure 13. This figure is displayed in five subplots, showcasing the distribution characteristics of microbial samples across different taxonomic levels, from phylum to genus. It reveals marked differences in data distribution, indicating that at the phylum level, there are comparatively limited taxonomic units but each unit comprises a larger number of samples with a broader range of variation. In contrast, at the genus level, the number of taxonomic units expands substantially, yet the range of sample numbers per unit is comparatively smaller.

The observed trend from higher to lower taxonomic levels suggests an inverse relationship between the count of taxonomic units and the sample size per unit. Specifically, higher taxonomic levels, such as phylum and class, are characterized by a smaller number of units but a larger number of samples, which facilitates the acquisition of more stable feature representations and consequently leads to higher classification accuracy. Conversely, lower levels, exemplified by genus, exhibit a larger number of taxonomic ranks with fewer samples, potentially impeding the capacity of the model to discern sufficient features for differentiation and resulting in diminished classification accuracy.

Notably, a pronounced disparity in sample distribution is evident between the order and family levels, despite similar unit counts. This imbalance at the order level contributes to reduced classification accuracy compared to that at the family level, where a more equitable sample distribution enhances the capacity of the model to discern distinct family features, thereby improving overall accuracy.

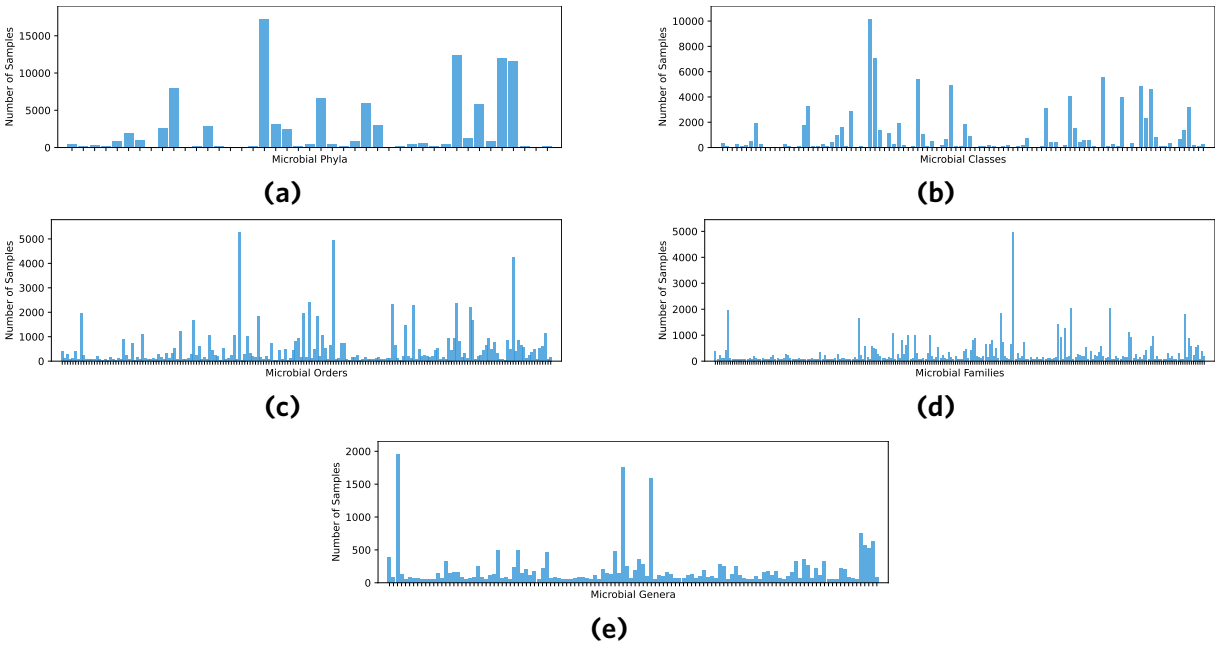

Figure 13. Microbial classification distribution across taxonomic levels. (A) Distribution at phylum level distribution; (B) Distribution at class level; (C) Distribution at order level; (D) Distribution at family level; (E) Distribution at genus level.

3.2. Microbial community structure attention analysis

To elucidate how microbial taxonomic hierarchies interact with graph attention mechanisms for hierarchical classification tasks, we investigated the attention coefficient distribution within the microbial taxonomic network. This analysis aimed to quantitatively validate whether GAT inherently captured hierarchical relationships by adaptively allocating attention weights across taxonomic levels from phylum to genus, which provided interpretability for microbial classification models.

In this study, GAT was used to extract features from the hierarchical taxonomic microbial network. To intuitively illustrate the interaction intensity between microbial nodes and the underlying hierarchical structure, we constructed an attention coefficient figure of the network was constructed based on the attention coefficient matrix generated during feature extraction, as shown in Figure 14. In this figure, nodes represent different microbial taxonomic units, colors denote different microbial phyla, and connection thickness reflects the magnitude of attention coefficients between nodes. The map shows that attention coefficients between nodes increase as the taxonomic hierarchy deepens.

Figure 14 illustrates the distribution of attention coefficients within the microbial taxonomic network across various hierarchical levels. At higher levels such as Phylum and Class, the edges exhibit significantly lower attention coefficients compared to lower levels. This implies that at these levels, the significant differences in features among taxonomic units make them easier to distinguish, which consequently requires less attention. In contrast, at lower levels such as Order and Family, the edges have higher attention coefficients. This indicates that the subtle differences in features among units at these levels necessitate more precise identification and differentiation, which consequently demands more attention resources.

This phenomenon indicates that a layered approach is used to process the microbial taxonomic network. At higher taxonomic levels, significant feature differences are utilized for classification. In contrast, at lower levels, subtle feature nuances require meticulous examination to ensure accurate classification. The distribution of attention coefficients not only exposes the strength of relationships between different levels within the mi-

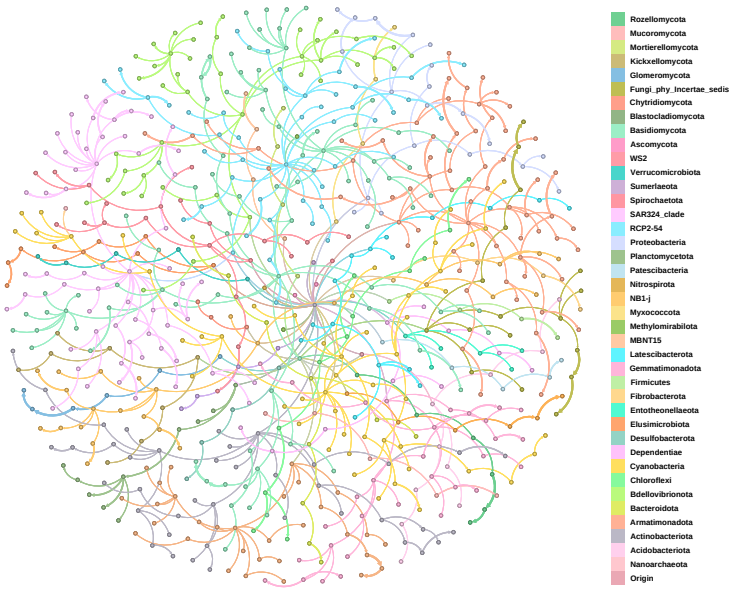

Figure 14. Relationship network graph of microbial classification based on attention coefficients.

Table 6. Comparison of classification accuracy across taxonomic ranks (%)

Model	Phylum	Class	Order	Family	Genus	Custom taxonomy
DNABERT-2	98.73	97.35	95.83	96.89	96.30	94.81
BioSeqBERT-CNN	99.18	98.54	97.13	97.97	96.87	97.86
MicroGraphBERT	**99.80**	**98.64**	**97.26**	**98.18**	**97.12**	**98.72**

crobial taxonomic network but also aids in understanding how attention is allocated across various levels to enhance classification performance. Moreover, the results show that the model can adaptively assign different attention weights to various nodes, which captures hierarchical relationships between nodes more effectively and reflects the intrinsic structure of the microbial taxonomic network accurately.

3.3. Comparison to related works

To evaluate whether integrating hierarchical taxonomic information with contextual sequence semantics improves gene sequence classification, we conducted a comparative analysis of MicroGraphBERT against two BERT-based models DNABERT[30] and BioSeqBERT-CNN[31] that specialize in gene sequence processing. DNABERT modified the BERT model by adjusting the pre-training tasks and redesigning input representations. Specifically, it employs BPE to tokenize DNA sequences into subwords that fit the BERT architecture. During pre-training, DNABERT also adjusts the sequence length and masking strategy to better capture the characteristics of gene sequences. On the other hand, BioSeqBERT-CNN combines BERT with a CNN for gene sequence classification tasks. It first pre-trains a BERT model on a large corpus of biological sequences to obtain contextual embeddings and then uses these embeddings as input to a CNN for classification. In this experiment, the proposed model framework was compared with the two models. As shown in Table 6, the proposed model outperformed DNABERT and BioSeqBERT-CNN in terms of accuracy across six different tasks.

The results demonstrated that MicroGraphBERT achieved the highest classification accuracy across all six tasks, especially in the Phylum and Custom Taxonomy tasks, where it reached accuracies of 99.80% and 98.72%, respectively, which significantly outperformed the other two models.

4. CONCLUSIONS

In this study, soil datasets were meticulously collected from the loess regions of Guizhou, encompassing samples from both rhizosphere and non-rhizosphere soils through nine distinct plant growth stages and two specific post-harvest time points. High-throughput sequencing of these diverse samples generated an extensive volume of gene sequence data, providing a rich foundation for analysis. The analysis revealed that soil microbial communities are significantly influenced by the rhizosphere and exhibit substantial changes across different plant growth stages. This dynamic shift indicates that microbes are actively involved in plant metabolic processes, thus highlighting the temporal variability of microbial composition. Additionally, the presence of numerous low-abundance microbes across various stages suggests a class-imbalanced dataset, which poses significant challenges for accurate classification, particularly for low-abundance species.

To address these challenges, the MicroGraphBERT model was proposed. A hierarchical microbial taxonomic network was extracted from the soil microbiome dataset as prior knowledge, and its features were extracted using GAT. These were fused with contextual embeddings from a fine-tuned DNABERT model, which captures gene sequence semantics. This integration enables the model to consider both genetic and structural features during classification. During training, SMOTE oversampling was used, and 50% of the data was randomly discarded in each epoch to enhance generalizability and robustness. Experiments showed that MicroGraphBERT achieved superior performance in both ablation and comparative studies.

Future work will focus on the incorporation of additional prior knowledge into microbial identification to further explore the underlying mechanisms and interactions within microbial communities, thereby improving classification accuracy and deepening our understanding of plant-soil microbiota dynamics.

DECLARATIONS

Authors' contributions

Conceived and proposed this study: Yang, H.; Wang, D.
Conducted key experiments and collected data: Yang, H.; Luo, X.; Tu, Z.
Wrote and revised the manuscript.: Yang, H.; Wang, D.
Controlled the quality and checked the results of the manuscript: Wang, D.; Pan, W.
Provided data support and experimental environment: Pan, W.; Jiang, C.; Gao, W.

Availability of data and materials

The data will not be publicly shared. Due to the sensitive nature of this research, the underlying data are proprietary to and governed by the access restrictions of the Guizhou Tobacco Science Research Institute. Therefore, they cannot be made publicly available.

Financial support and sponsorship

This work was supported by the Key Program for Science and Technology of China National Tobacco Corporation, grant number 110202102038, and School of Information Science and Engineering, Chongqing Jiaotong University, grant number 2024yjkc003.

Conflicts of interest

Wang, D. is Junior Editorial Board Member of the journal *Intelligence & Robotics*. She was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, or decision-making. Pan, W. and Jiang, C. are employees of China National Tobacco Corporation Guizhou Branch. Gao, W. is an employee of Guizhou Tobacco Science Research Institute. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Coban, O.; De Deyn, G. B.; van der Ploeg, M. Soil microbiota as game-changers in restoration of degraded land. *Science* **2022**, *375*, abe0725. DOI
2. Philippot, L.; Chenu, C.; Kappler, A.; Rillig, M. C.; Fierer, N. The interplay between microbial communities and soil properties. *Nat. Rev. Microbiol.* **2024**, *22*, 226-39. DOI
3. Yang, X.; Cheng, J.; Franks, A. E.; et al. Loss of microbial diversity weakens specific soil functions, but increases soil ecosystem stability. *Soil Biol. Biochem.* **2023**, *177*, 108916. DOI
4. Liao, J.; Dou, Y.; Yang, X. ; An, S. Soil microbial community and their functional genes during grassland restoration. *J. Environ. Manage.* **2023**, *325*, 116488. DOI
5. Wang, Y.; Yan, X.; Su, M.; et al. Isolation of potassium solubilizing bacteria in soil and preparation of liquid bacteria fertilizer from food wastewater. *Biochem. Eng. J.* **2022**, *181*, 108378. DOI
6. Javed, Z.; Tripathi, G. D.; Mishra, M.; Dashora, K. Actinomycetes – The microbial machinery for the organic-cycling, plant growth, and sustainable soil health. *Biocatal. Agric. Biotechnol.* **2021**, *31*, 101893. DOI
7. Bai, B.; Liu, C.; Zhang, C.; et al. *Trichoderma* species from plant and soil: an excellent resource for biosynthesis of terpenoids with versatile bioactivities. *J. Adv. Res.* **2023**, *49*, 81–102. DOI
8. Sokol, N. W.; Slessarev, E.; Marschmann, G. L.; et al. Life and death in the soil microbiome: how ecological processes influence biogeochemistry. *Nat. Rev. Microbiol.* **2022**, *20*, 415–30. DOI
9. Yonatan, Y.; Amit, G.; Friedman, J.; Bashan, A. Complexity–stability trade-off in empirical microbial ecosystems. *Nat. Ecol. Evol.* **2022**, *6*, 693–700. DOI
10. Chandra, B.; Gupta, M. Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. *Expert Syst. Appl.* **2011**, *38*, 1293–8. DOI
11. Deng, H.; Runger, G. Gene selection with guided regularized random forest. *Pattern Recognit.* **2013**, *46*, 3483–9. DOI
12. Brady, A.; Salzberg, S. L. Phymm and phymmBL: metagenomic phylogenetic classification with interpolated markov models. *Nat. Methods* **2009**, *6*, 673–6. DOI
13. Wang, Q.; Garrity, G. M.; Tiedje, J. M.; Cole, J, R. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–7. DOI
14. Haque Mohammed, M.; Ghosh, T. S.; Singh, N. K.; Mande, S. S. SPHINX - an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* **2011**, *27*, 22–30. DOI
15. Díaz, D.; Esteban, F. J.; Hernández, P.; Caballero, J. A.; Dorado, G.; Gálvez, S. Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture. *Parallel Comput.* **2011**, *37*, 244–59. DOI
16. Xia, Z.; Cui, Y.; Zhang, A.; et al. A review of parallel implementations for the Smith–Waterman algorithm. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 1–14. DOI
17. Camacho, C.; Coulouris, G.; Avagyan, V.; et al. BLAST+: architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421. DOI
18. Sievers, F.; Wilm, A.; Dineen, D.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. DOI
19. Le, N. Q. K.; Ho, Q. T.; Nguyen, T. T. D.; Ou, Y. Y. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Brief. Bioinform.* **2021**, *22*, bbab005. DOI
20. Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **2017**, *35*, 128–35. DOI
21. Wang, D.; Yang, S. X. Intelligent feature extraction, data fusion and detection of concrete bridge cracks: current development and challenges. *Intell. Robot.* **2022**, *2*, 391-406. DOI
22. Zhao, S.; Qiu, S.; Xu, X.; Ciampitti, I, A.; Zhang, S.; He, P. Change in straw decomposition rate and soil microbial community composition after straw addition in different long-term fertilization soils. *Appl. Soil Ecol.* **2019**, *138*, 123-33. DOI
23. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R. V. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112-20. DOI
24. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, **2018**. https://openreview.net/forum?id=rJXMpikCZ. (accessed 18 Jun 2025)
25. Gao, W.; Cai, K.; Li, D.; et al. Soil taxonomy and suitability assessment on typical tobacco-planting farmlands in Guizhou, Southwest

China. *SN Appl. Sci.* **2019**, *1*, 877. DOI

26. Fernández, A.; García, S.; Herrera, F.; Chawla, N. V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. DOI
27. Zhang, T.; Wu, Z.; Li, L.; et al. CellGAT: a GAT-based method for constructing a cell communication network integrating multiomics information. *Biomolecules* **2025**, *15*, 342. DOI
28. Kong, X.; Xing, W.; Wei, X.; Bao, P.; Zhang, J.; Lu, W. STGAT: spatial-temporal graph attention networks for traffic flow forecasting. *IEEE Access*, **2020**, *8*, 134363–72. DOI
29. Sellers, T.; Lei, T.; Luo, C.; Jan, G. E.; Ma, J. A node selection algorithm to graph-based multi-waypoint optimization navigation and mapping. *Intell. Robot.* **2022**, *2*, 333-54. DOI
30. Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; Liu, H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *arXiv* **2023**, arXiv:2306.15006. https://arxiv.org/abs/2306.15006. (accessed 18 Jun 2025)
31. Helaly, M. A.; Rady, S.; Aref, M. M. BERT contextual embeddings for taxonomic classification of bacterial DNA sequences. *Expert Syst. Appl.* **2022**, *208*, 117972. DOI