Abstract

K-mer-based feature extraction is a widely used technique for microbial taxonomy classification, offering insights into the relationships between DNA sequences and taxonomic labels. The 16S rRNA gene database used in this study contains nucleotide sequences of varying lengths (90 to 564 bases) and was filtered to include only records with complete taxonomic annotations down to the species level. The specific challenge addressed here is identifying the optimal k-mer size for predicting species-level classification with high accuracy. A neural network model was developed to test our hypothesis. Here we show that by systematically evaluating k-mer values maximize prediction accuracy across taxonomic levels, with larger k-mer performing better with strong correlation between the two. The model highlights the trade-offs between k-mer size, computational efficiency, and classification performance, demonstrating that certain k-mer values outperform others in predicting taxonomic labels. These findings advance our understanding of feature optimization in microbial taxonomy classification and set the stage for future genome-based research in this field.

Summary of previous findings

Previous studies have explored microbial taxonomy classification using k-mer analysis and machine learning. Early approaches, such as k-mer frequency analysis with support vector machines (SVMs) and random forest classifiers, demonstrated moderate accuracy but were limited by scalability and challenges with high-dimensional data. Deep learning models have also been applied in this domain, such as Convolutional Neural Networks (CNNs) for classifying DNA sequences based on k-mers. Although these models exhibited superior accuracy compared to traditional methods, their computational requirements and the difficulty in optimizing hyperparameters limited their applicability in large-scale studies. Zhao et al. (2024) introduced DSNetax, a deep learning species annotation method based on a deep-shallow parallel framework, which achieved promising results but faced challenges with training complexity and resource requirements. Asgari et al. (2018) used a k-mer-based representation to predict environments and host phenotypes from 16S rRNA sequencing data, demonstrating the potential of k-mer analysis for taxonomic tasks. However, their approach focused on phenotype prediction rather than detailed taxonomic classification.

In contrast, this study employs a sequential neural network rather than more complex architectures like deep neural networks or CNNs. This choice offers a balance between computational feasibility and predictive performance, particularly when working with large-scale datasets and limited resources. By focusing on optimizing k-mer sizes and utilizing a streamlined neural network design, we address some of the scalability and resource challenges highlighted in previous research.

Results

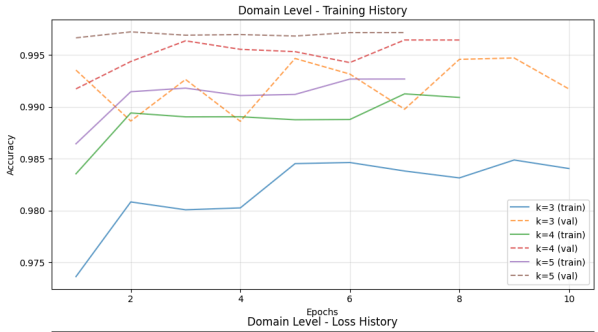
This study examined the effect of k-mer sizes (3, 4, and 5) on neural network performance for classifying microbial species across seven taxonomic levels: domain, phylum, class, order,

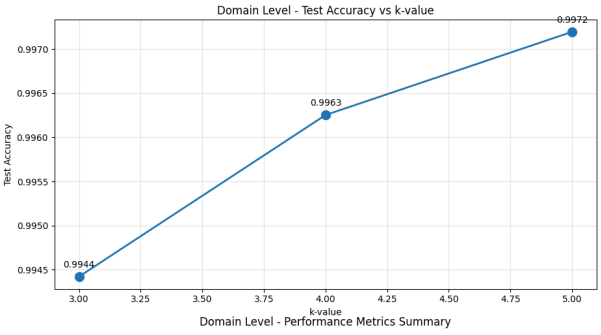
family, genus, and species. Neural networks were trained for each k-mer and taxonomic level, with evaluation based on accuracy and loss metrics. The analysis used 25% of the trimmed dataset (2.21 GB), containing nucleotide sequences with complete species-level annotations. Sparse categorical cross-entropy loss was employed, effectively handling multi-class classification tasks with multiple taxonomic labels.

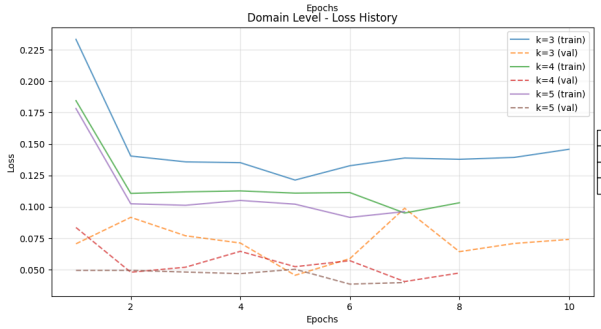
An initial baseline approach was to use the Needleman-Wunsch algorithm for pairwise global alignments to compare new sequences against multiple sequence alignments (MSAs) or individual gene sequences within families. While it provides precise percent identity calculations and could help identify the closest sequence matches, the computational demands were prohibitive. For just 5% of the dataset, approximately 782 billion alignments would have been required, making this method infeasible within the project's scope and resources.

The neural network performed well across taxonomic levels, with accuracy improving as k-mer size increased. At higher levels like domain and phylum, accuracies exceeded 90%, with k=5 consistently performing best. Accuracy declined at deeper levels due to more distinct classes, though k=5 maintained the highest performance. Larger k-mer sizes improved results overall but showed diminishing returns at deeper ranks, highlighting the trade-off between accuracy and computational demands.

Alternative approaches were considered but not implemented due to computational limitations. Expanding k-mer sizes to 11 or higher was proposed to improve accuracy but proved impractical due to exponential feature space growth and memory demands. Alternatives like random forests or support vector machines were also explored but lacked scalability for high-dimensional datasets. Neural networks were chosen for their efficiency in handling complex data and strong classification performance, enhanced by dropout and early stopping to prevent overfitting.



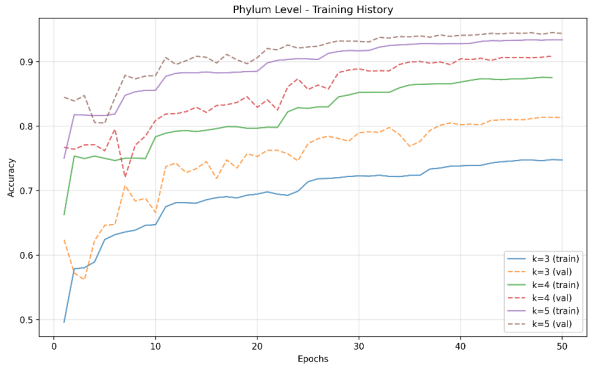


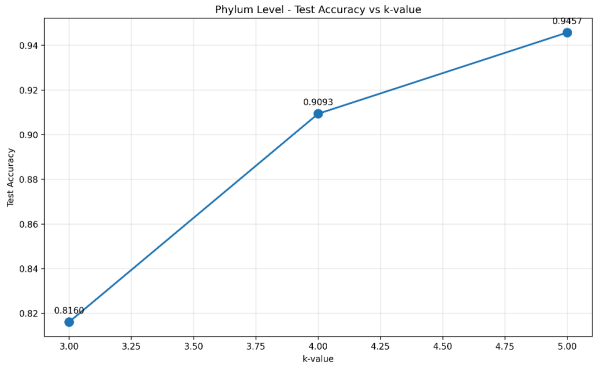


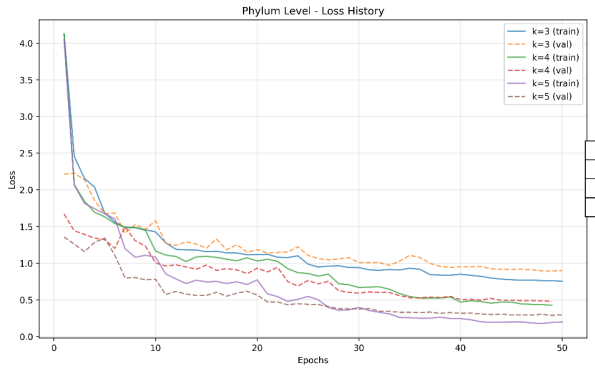
Domain Level - Performance Metrics Summary

k-value	Test Acc	Test Loss	Classes	Features
3	0.9944	0.0471	2	64
4	0.9963	0.0533	2	256
5	0.9972	0.0505	2	1024

Figure 1. Domain Level Results



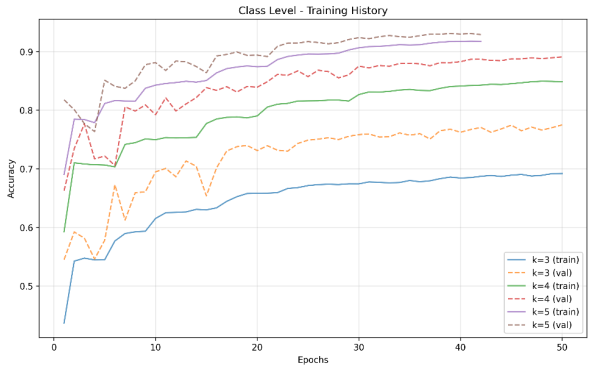


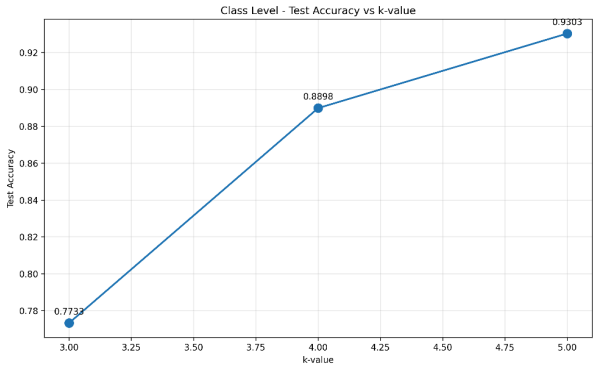


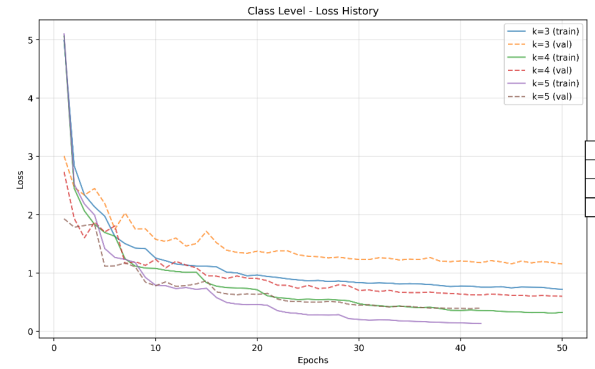
Phylum Level - Performance Metrics Summary

k-value	Test Acc	Test Loss	Classes	Features
3	0.8160	0.8856	117	64
4	0.9093	0.4726	117	256
5	0.9457	0.2832	117	1024

Figure 2. Phylum Level Results



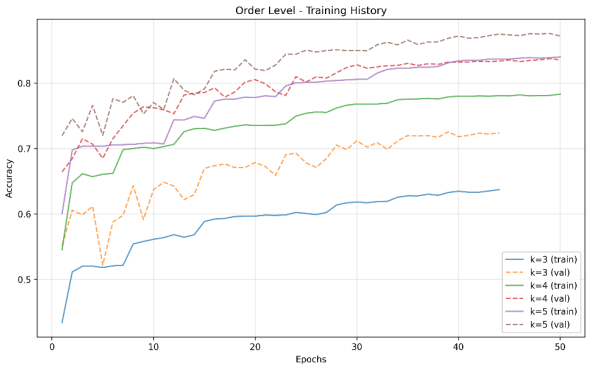


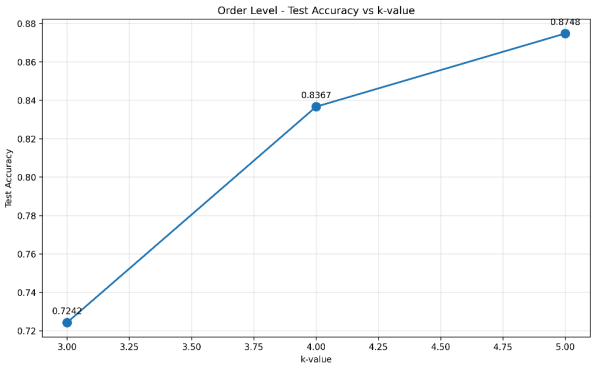


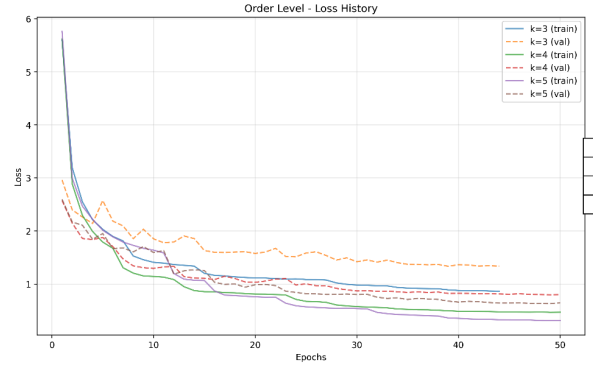
Class Level - Performance Metrics Summary

k-value	Test Acc	Test Loss	Classes	Features
3	0.7733	1.1579	284	64
4	0.8898	0.6013	284	256
5	0.9303	0.3930	284	1024

Figure 3. Class Level Results



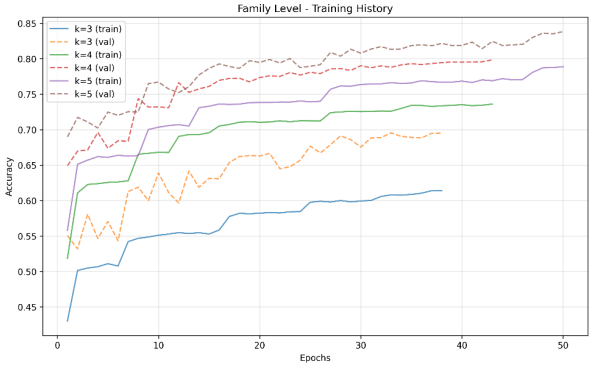


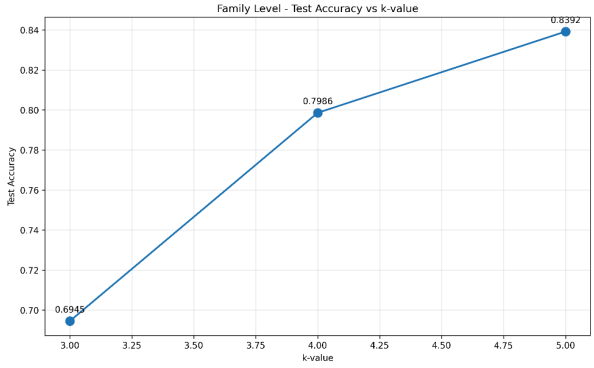


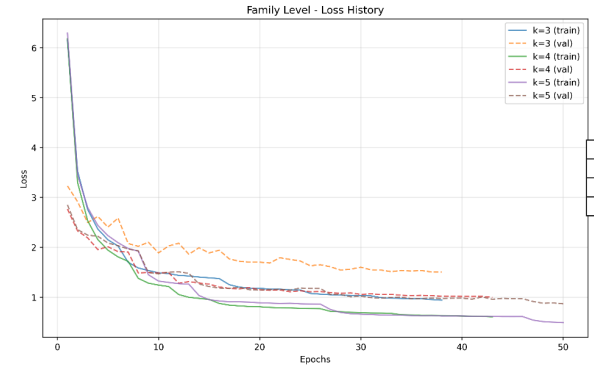
Order Level - Performance Metrics Summary

k-value	Test Acc	Test Loss	Classes	Features
3	0.7242	1.3430	791	64
4	0.8367	0.8007	791	256
5	0.8748	0.6398	791	1024

Figure 4. Order Level Results



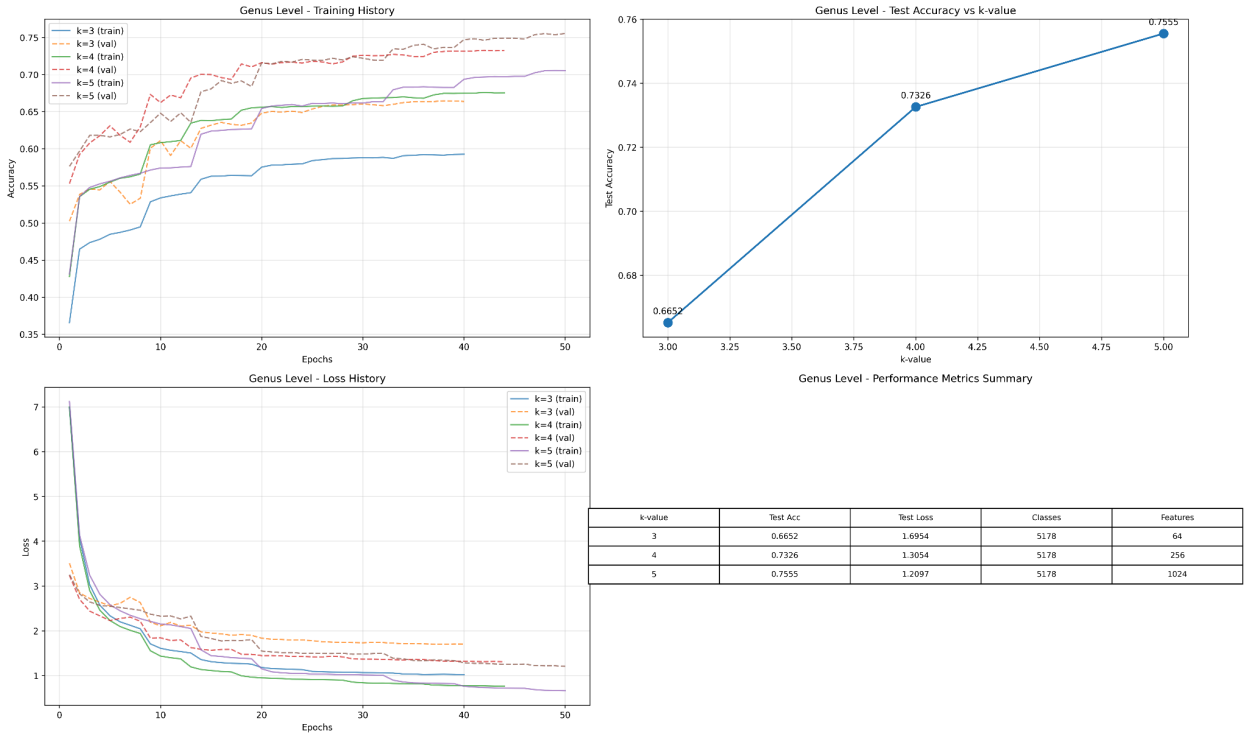




Family Level - Performance Metrics Summary

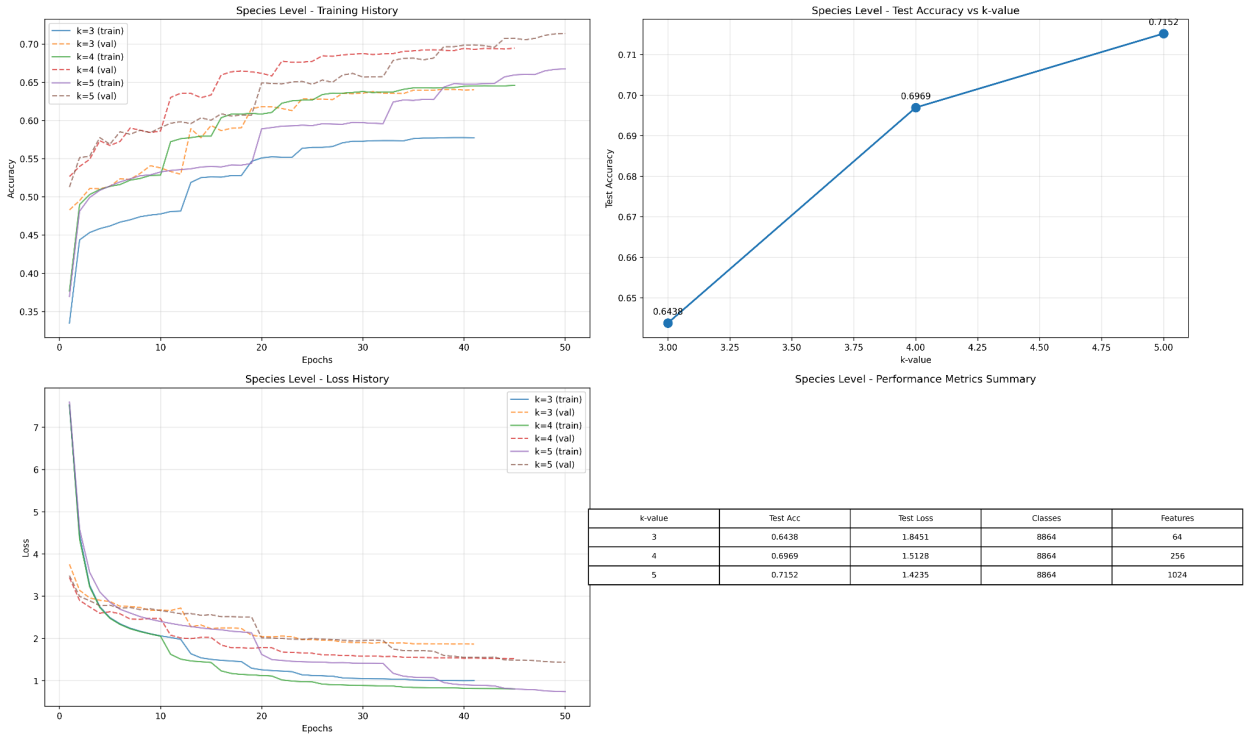
k-value	Test Acc	Test Loss	Classes	Features
3	0.6945	1.5119	1674	64
4	0.7986	0.9965	1674	256
5	0.8392	0.8604	1674	1024

Figure 4. Family Level Results



k-value	Test Acc	Test Loss	Classes	Features
3	0.6652	1.6954	5178	64
4	0.7326	1.3054	5178	256
5	0.7555	1.2097	5178	1024

Figure 5. Genus Level Results



k-value	Test Acc	Test Loss	Classes	Features
3	0.6438	1.8451	8864	64
4	0.6969	1.5128	8864	256
5	0.7152	1.4235	8864	1024

Figure 6. Species Level Results

Conclusion

This study aimed to classify microbial species using a neural network optimized for varying k-mer sizes, evaluating their impact on accuracy across seven taxonomic levels. The findings demonstrated that larger k-mer sizes, particularly k=5, consistently improved accuracy, with the highest performance observed at the domain level (99.72% accuracy) and a gradual decline at deeper taxonomic levels, such as species (71.52%). Despite computational constraints, the analysis provided valuable insights into the trade-offs between feature size and classification performance.

Future work can focus on enhancing neural network models and data representation. One improvement is converting nucleotide sequences (ACTG) to binary bits, reducing computational demands by leveraging their four unique values (0–3). Exploring alternative architectures, like convolutional or recurrent neural networks, may better capture sequence patterns. Another direction is optimizing k-mer storage by mapping similar sequences with minor differences to a single value based on shared classification labels, reducing redundancy and improving consistency. Additionally, structured encoding of sequence variations could group similar rows by classification labels, isolating differences for targeted analysis.

Methods

The classification of microbial species was conducted using a neural network designed to predict taxonomic levels down to the species level. This study focused on evaluating the impact of different k-mer values, ranging from 3 to 5, on the accuracy of taxonomic label predictions across seven levels: domain, phylum, class, order, family, genus, and species.

The neural network architecture consisted of an input layer that accepted nucleotide sequence data of varying lengths, spanning 90 to 564 bases. The hidden layers were carefully designed to optimize feature extraction and classification performance. The first hidden layer comprised neurons with a ReLU activation function, followed by a dropout layer to mitigate overfitting. The second hidden layer reduced the number of neurons to , also using ReLU activation and a dropout layer. Similarly, the third hidden layer employed neurons with ReLU activation and dropout. The output layer was configured to match the number of distinct classes for each taxonomic rank, ranging from 2 classes for domain to 8864 classes for species.

To address computational constraints, the dataset was reduced to 25% of its original size, retaining only records with complete species-level annotations to ensure high-quality input data. Sparse categorical cross-entropy loss was used as the optimization metric, well-suited for multi-class classification. Training employed early stopping, halting after 10 consecutive epochs without improvement, and incorporated mechanisms to revert to the best-performing weights to prevent overfitting and stabilize the model. A total of 21 models were trained, systematically testing three k-mer values across seven taxonomic levels. The experimental design aimed to balance predictive accuracy and computational efficiency, identifying the k-mer size that delivered optimal performance. This method offers a scalable solution for microbial taxonomy classification and provides a solid foundation for future genomic research.

References

Hongyuan Zhao, Suyi Zhang, Hui Qin, Xiaogang Liu, Dongna Ma, Xiao Han, Jian Mao, Shuangping Liu, DSNetax: a deep learning species annotation method based on a deep-shallow parallel framework, *Briefings in Bioinformatics*, Volume 25, Issue 3, May 2024, bbae157, https://doi.org/10.1093/bib/bbae157

Park H, Lim SJ, Cosme J, et al. Investigation of machine learning algorithms for taxonomic classification of marine metagenomes. *Microbiol Spectr*. 2023;11(5):e0523722. doi:10.1128/spectrum.05237-22

Asgari E, Garakani K, McHardy AC, Mofrad MRK. MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples [published correction appears in Bioinformatics. 2019 Mar 15;35(6):1082. doi: 10.1093/bioinformatics/bty652]. *Bioinformatics*. 2018;34(13):i32-i42. doi:10.1093/bioinformatics/bty296

McDonald, D., Jiang, Y., Balaban, M. et al. Greengenes2 unifies microbial data in a single reference tree. Nat Biotechnol 42, 715–718 (2024). https://doi.org/10.1038/s41587-023-01845-1

GreenGenes database release (2024.09). Retrieved from https://ftp.microbio.me/greengenes_release/2024.09/00README