LEVERAGING LARGE LANGUAGE MODELS FOR METAGENOMIC ANALYSIS

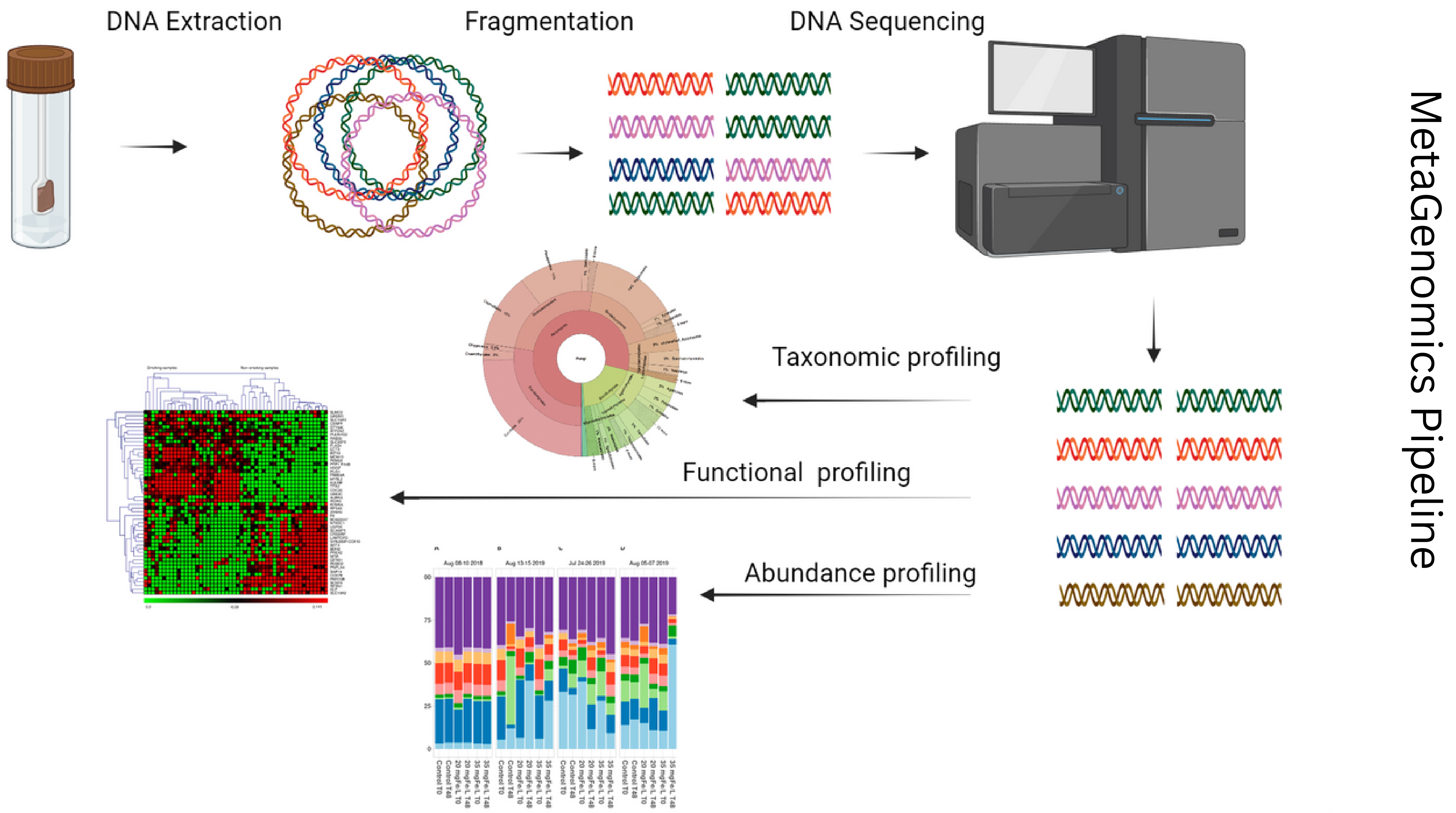
M.S. Refahi , B.A. Sokhansanj and G.L. Rosen

Drexel University

IEEE SPMB 2023

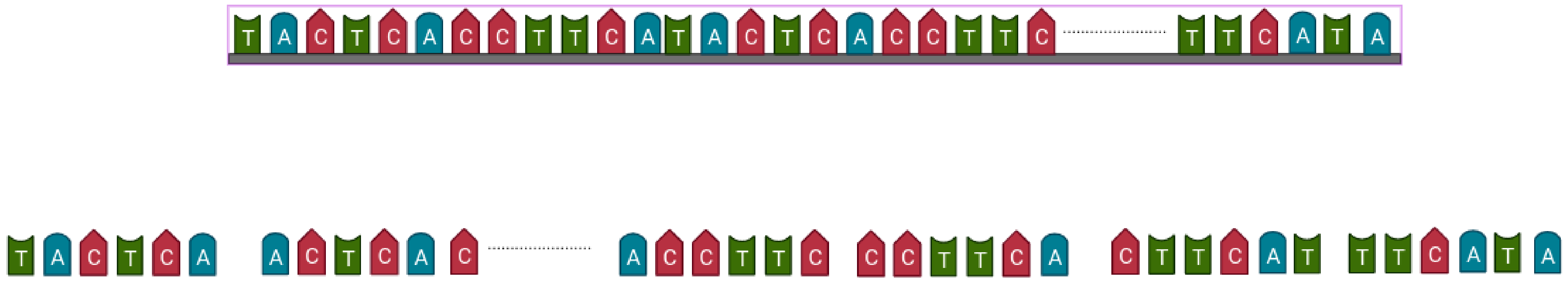
Introduction

- Metagenomics: The study of genetic material directly extracted from environmental or clinical samples through sequencing methods.
- Metagenomics enables the identification of diverse microbial species within a given environment.
- Understanding richness and abundance provides insights into ecosystem health and functioning.
- Analysis of metagenomic data identifies functional genes and pathways in microbial genomes.



Introduction

- Researchers harness Natural Language Processing (NLP) to train machine-learning models, generating vector representations from word sequences.
- This breakthrough in representation learning holds promise for analyzing biological sequence data.
- The first step in any natural language-based model is tokenization of sentences. Each token represents a meaningful unit of the text.
- In biological sequences, each sequence is analogous to a sentence, and DNA sequences are composed of individual nucleotides. Effective extraction of meaningful sub-words from biological sequences using k-mer tokenization.



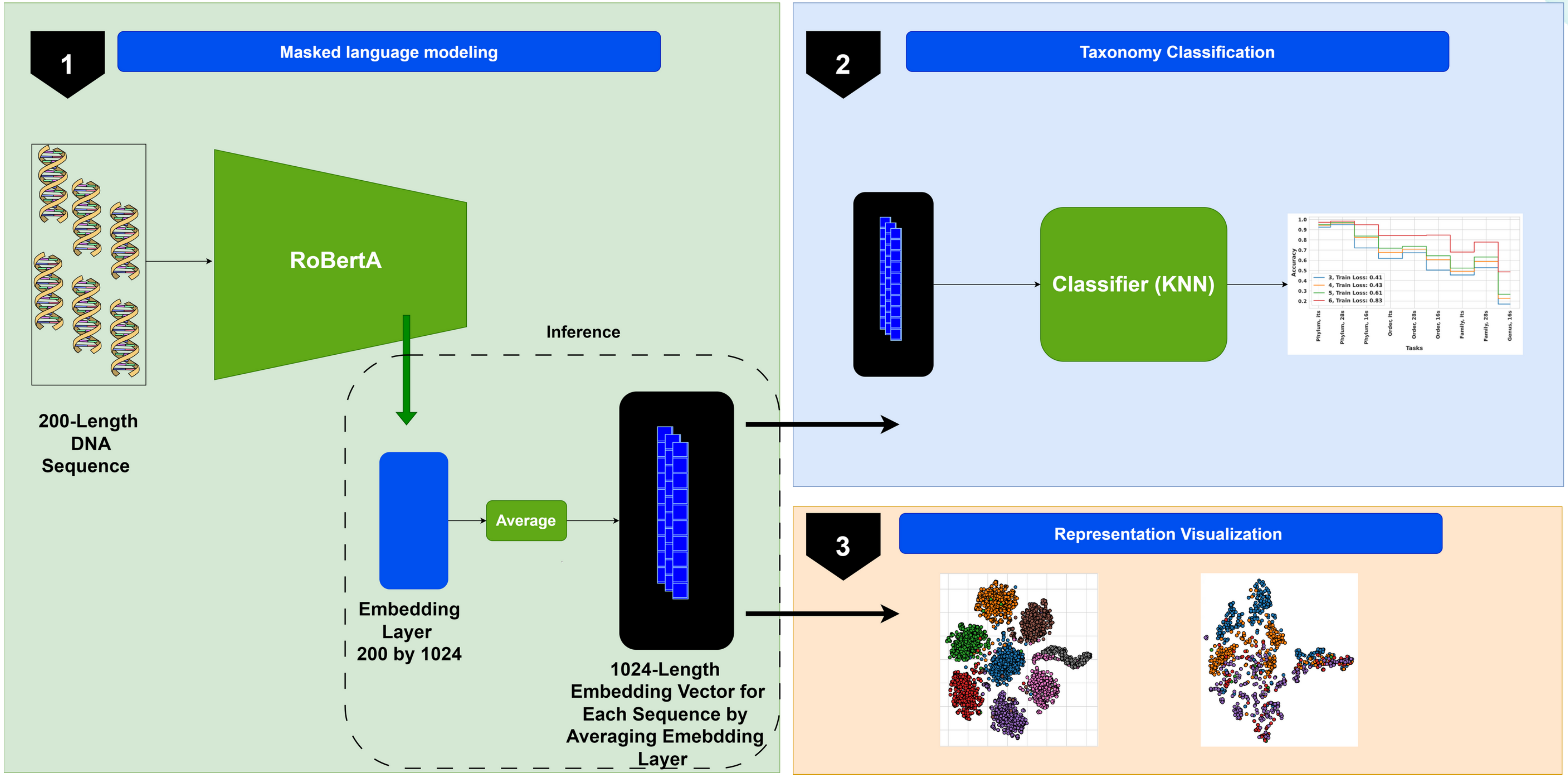
6mer Tokenization

Introduction

- Language modeling is a specific task in NLP where the model tries to predict a token based on surrounding tokens.
- Various techniques, including Word2Vec, LSTM, GloVe, and Transformers, are employed to obtain representation embedding vectors for biological k-mer sequences based on Language Modeling.
- RoBERTa, a version of the BERT model, exhibits superiority in its transformer architecture. Notably, RoBERTa focuses more on masked token prediction compared to BERT.
- To capture intricate patterns in metagenomics, we hyperparameter-tuned the RoBERTa model using metagenomic data.

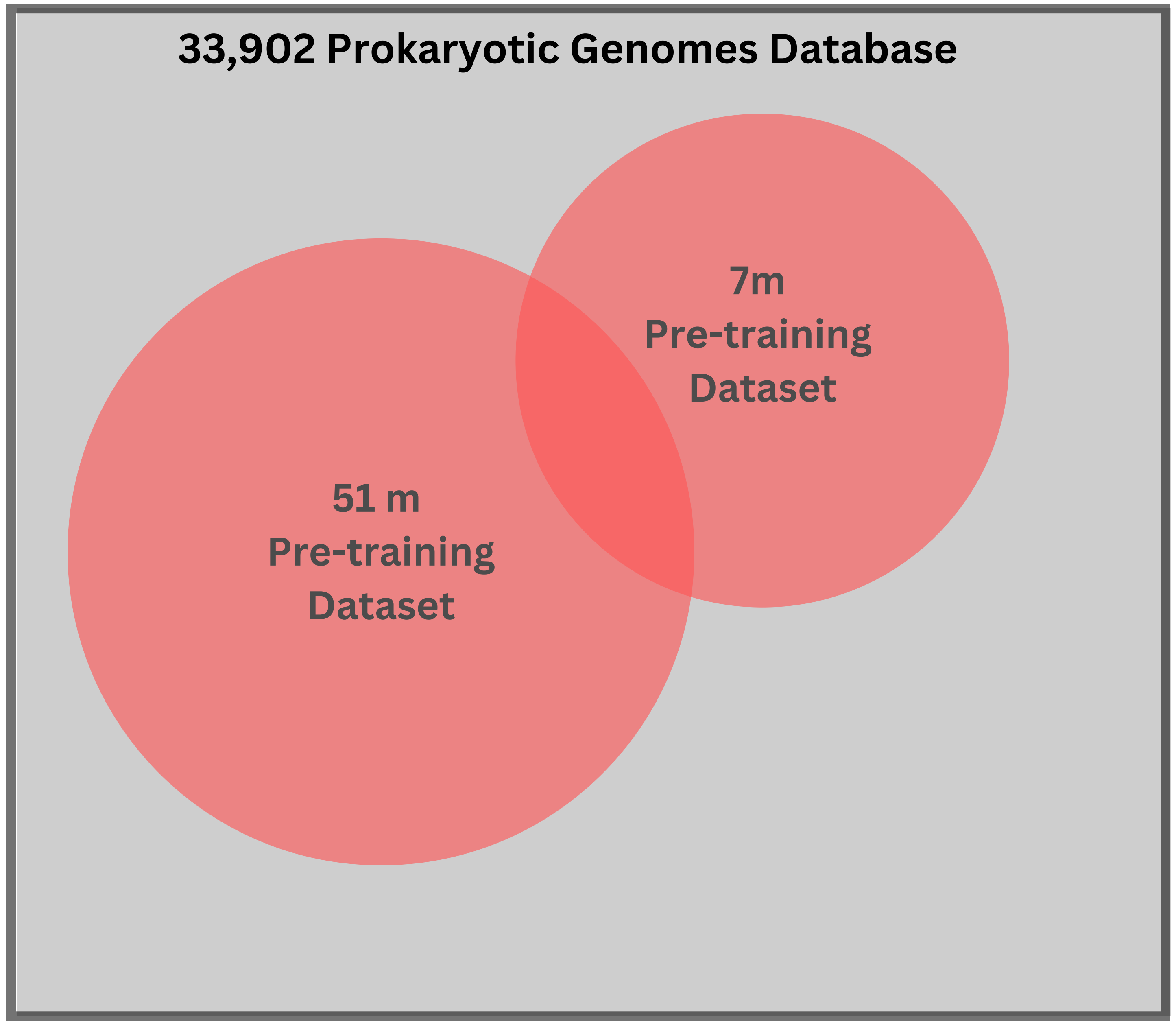
Masked Language Modeling 6-mer Representation

Methodology



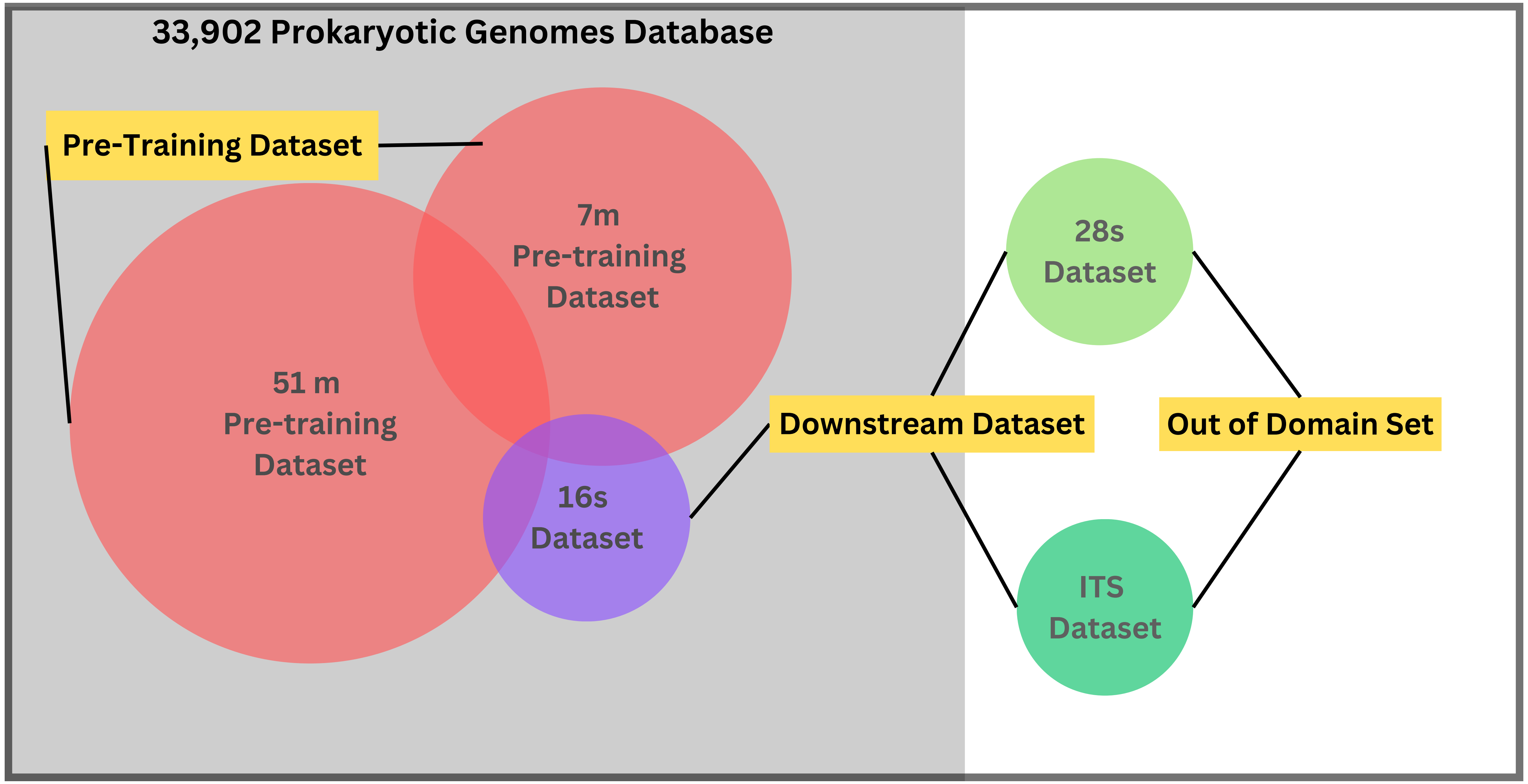
Pre-Train Datasets

- **Pretraining-Dataset:** Comprehensive dataset of **33,902 prokaryotic genomes** from NCBI, average length of 3.4 Mb.
  Prokaryotic genomes, unlike eukaryotes, exhibit compact structures and high gene density.
  Two subsets: 7 million genomic fragments (200bp) and 51 million fragments, maintaining diversity.



Downstream Datasets

- **Bacterial-16s** DairyDB Dataset: comprises 10,612 full-length 16S rRNA genes from microbial species in dairy products. The 16S rRNA gene serves as a marker for taxonomic and phylogenetic analyses. The dataset includes 80,227 fragments (200bp) representing 42 phyla, 197 orders, and 1069 genera
- **Fungi-ITS Dataset:** From Fungi RefSeq ITS project: 15,551 sequences from non-gene coding region. Extracted 50,068 fragments (200bp) showcasing diversity across 6 phyla, 235 orders, and 516 families.
- **Fungi-28s Dataset:** Curated 28s rRNA genes from mothur project: 8,506 unique sequences. Extracted 42,766 fragments (200bp) covering sequences from 8 phyla, 105 orders, and 293 families.



Drexel UNIVERSITY

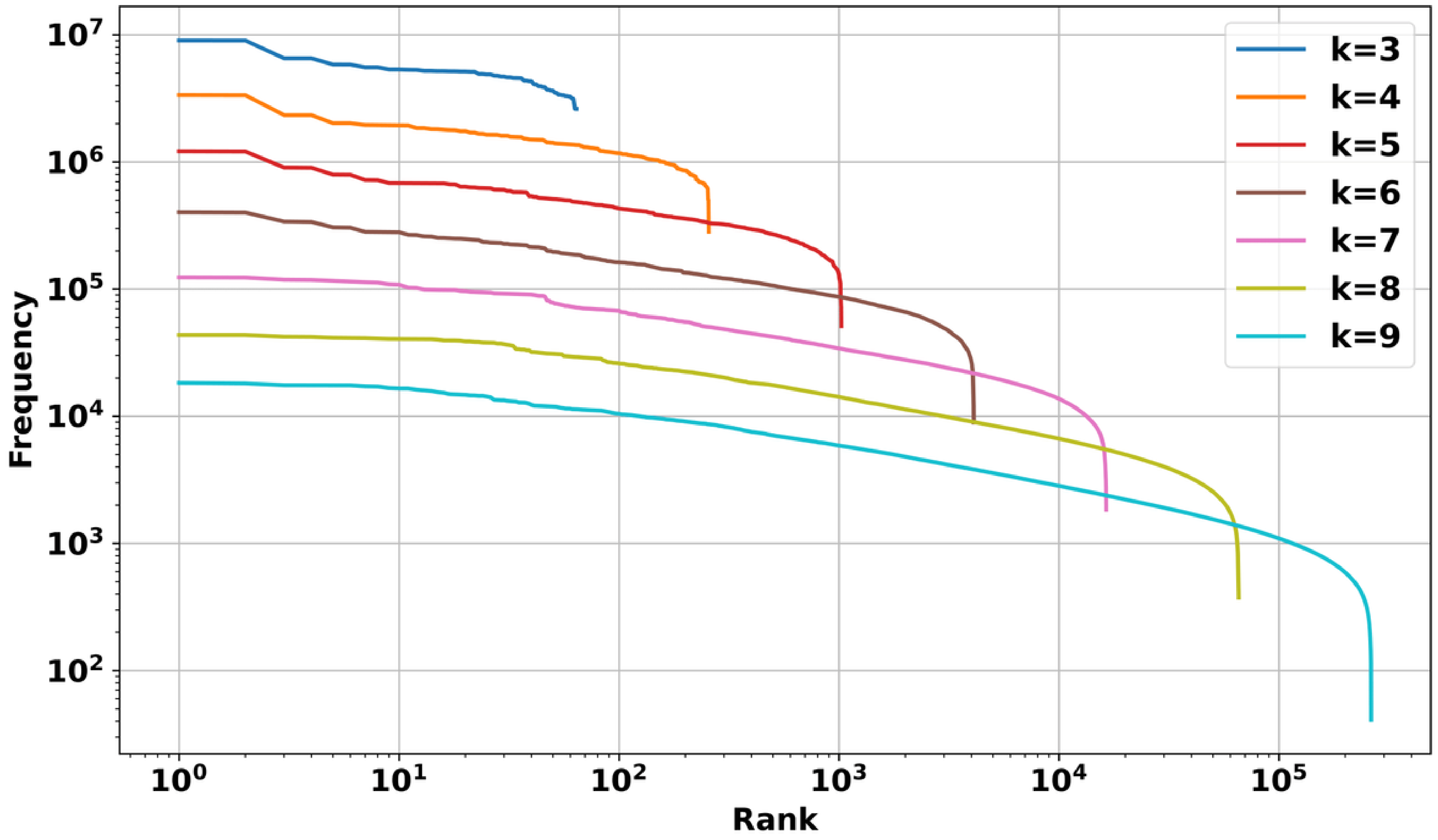
DNA Language Model Training

- Customized RoBERTa model tailored for analyzing prokaryotic genomes. Main architecture and transformers from RoBERTa, with modifications for optimized performance in genomics sequences.
- Input length selected as 200bp, considering the short-read nature of popular Illumina sequencing.
- Parameter selection strategy based on two metrics: **pre-training loss** and **downstream task accuracy** (taxonomy classification).
- Due to the time and cost involved in training large language models, a subset of parameters is selected for experimentation.
- Iterative testing involves swapping one parameter at a time to assess its impact on both downstream task accuracy and pre-training loss.
- Parameters chosen for evaluation include k-mer size, positional embedding type, maximum positional embedding size, model size, dataset size, and embedding size.
- Parameters such as the number of attention heads and forward layers investigated for optimization(Model Size).

Drexel UNIVERSITY

Zipf's Law of the pre-training dataset of ~34k genomes for Different k-mer Lengths

The Zipf's Law plot demonstrates a natural increase as the number of kmer decreases

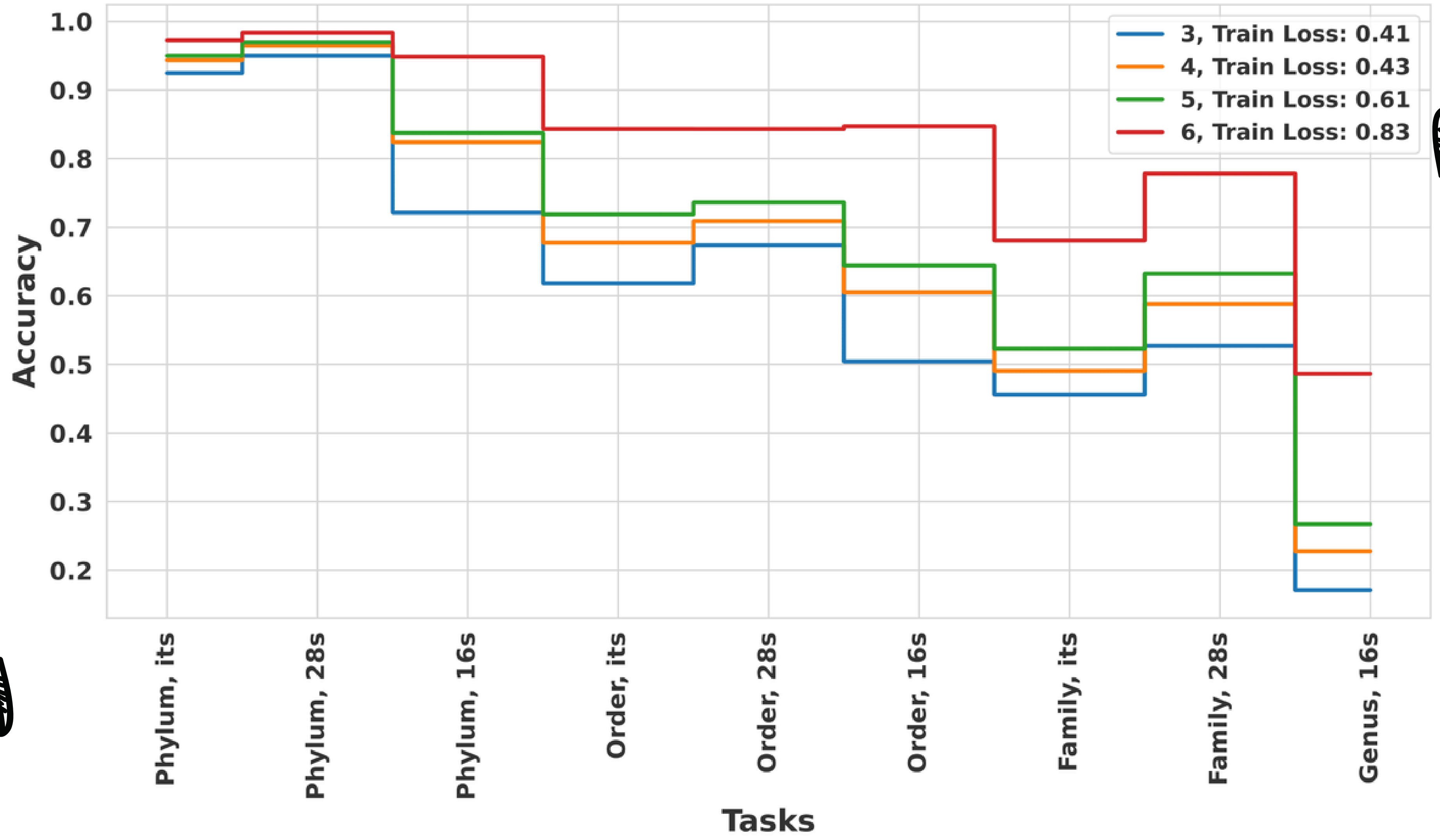


Zipf's Law plots can inform optimal vocabulary choices by highlighting a balance between high-frequency words, which could convey insignificant meaning in certain cases, and low-frequency (rare) words, potentially posing training challenges.

Drexel UNIVERSITY

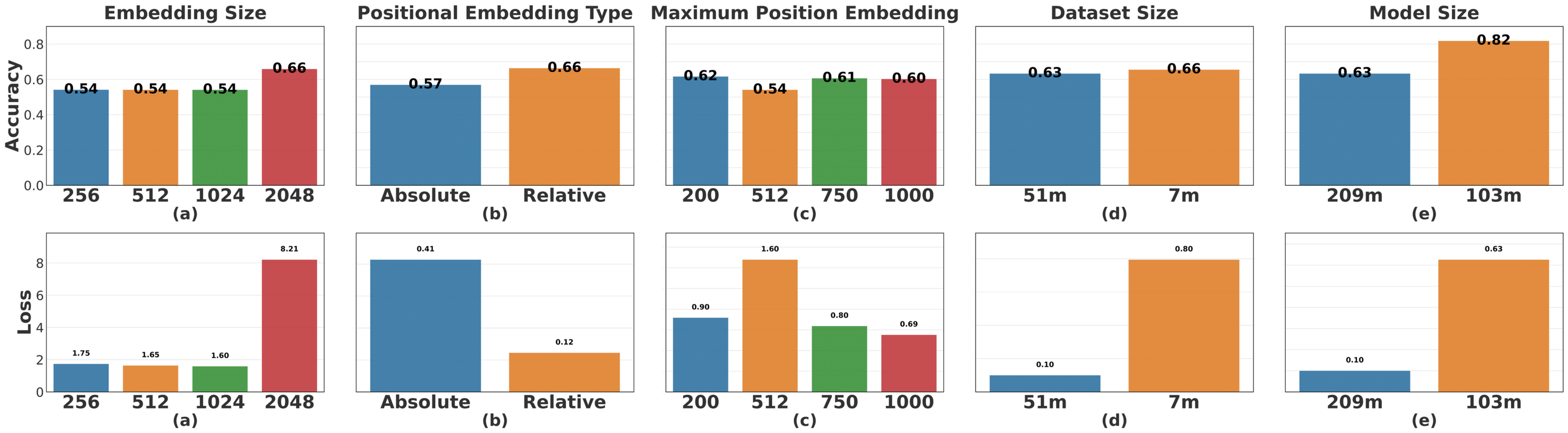
Comparison of Model Accuracies for Different Tasks with Respect to K-mer Size

We have 9 tasks across three datasets ITS, 16s, 28s and are calculating accuracy for various taxonomy levels within each dataset.



In kmer analysis, as vocabulary grows, training loss may rise. This highlights the need for alternative metrics for better comparisons.

Comparison of Average Accuracy and Training Loss for Different Parameters



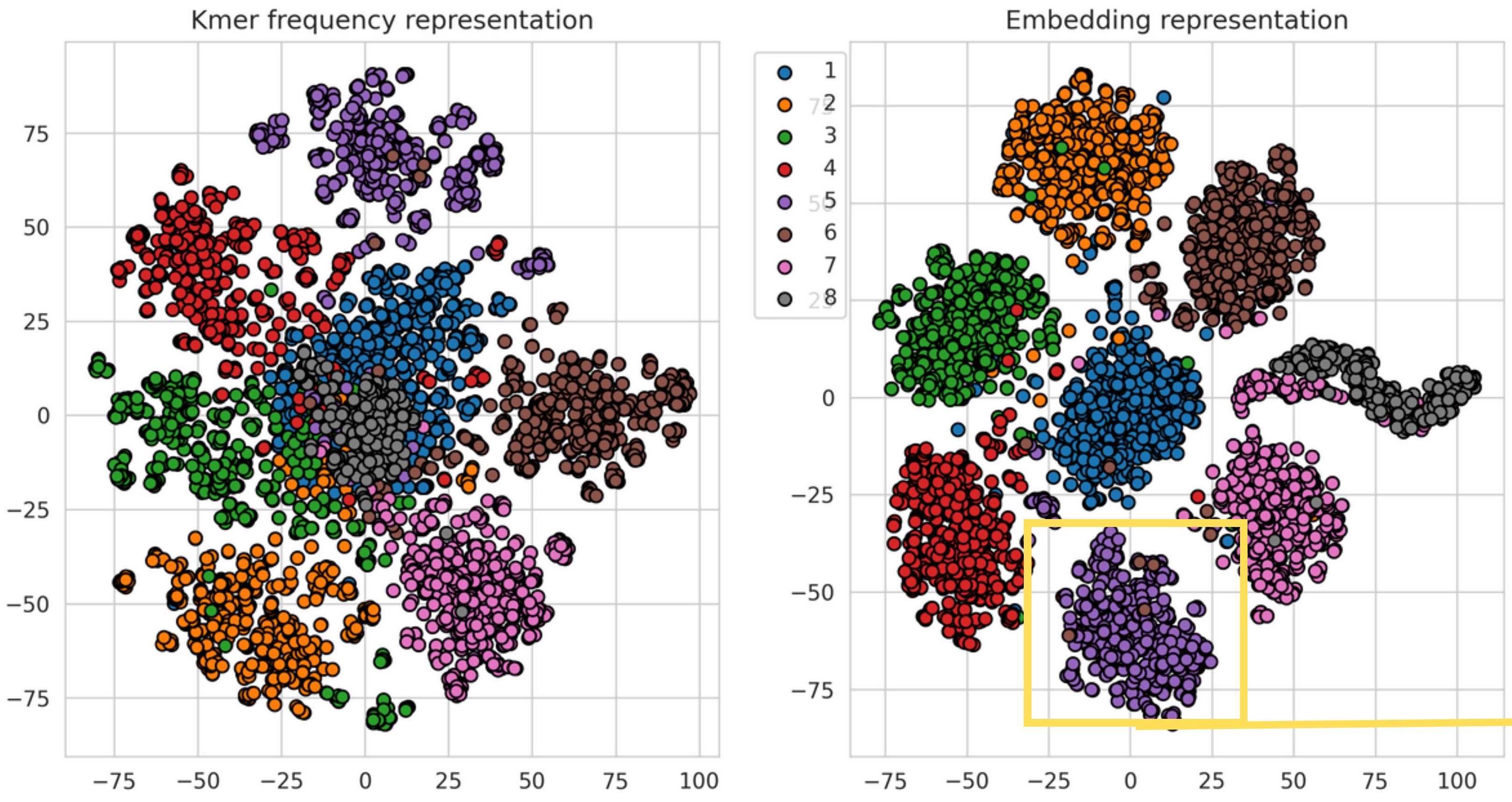
An example emphasizes that high accuracy doesn't ensure parameter effectiveness. A low pretraining loss suggests inadequate model training.

- Default value for max_position_embeddings in BERT and RoBERTa: 512
- Crucial to note its role with positional embedding type in position encoding
- Consider tuning based on sequence length for optimal model performance.

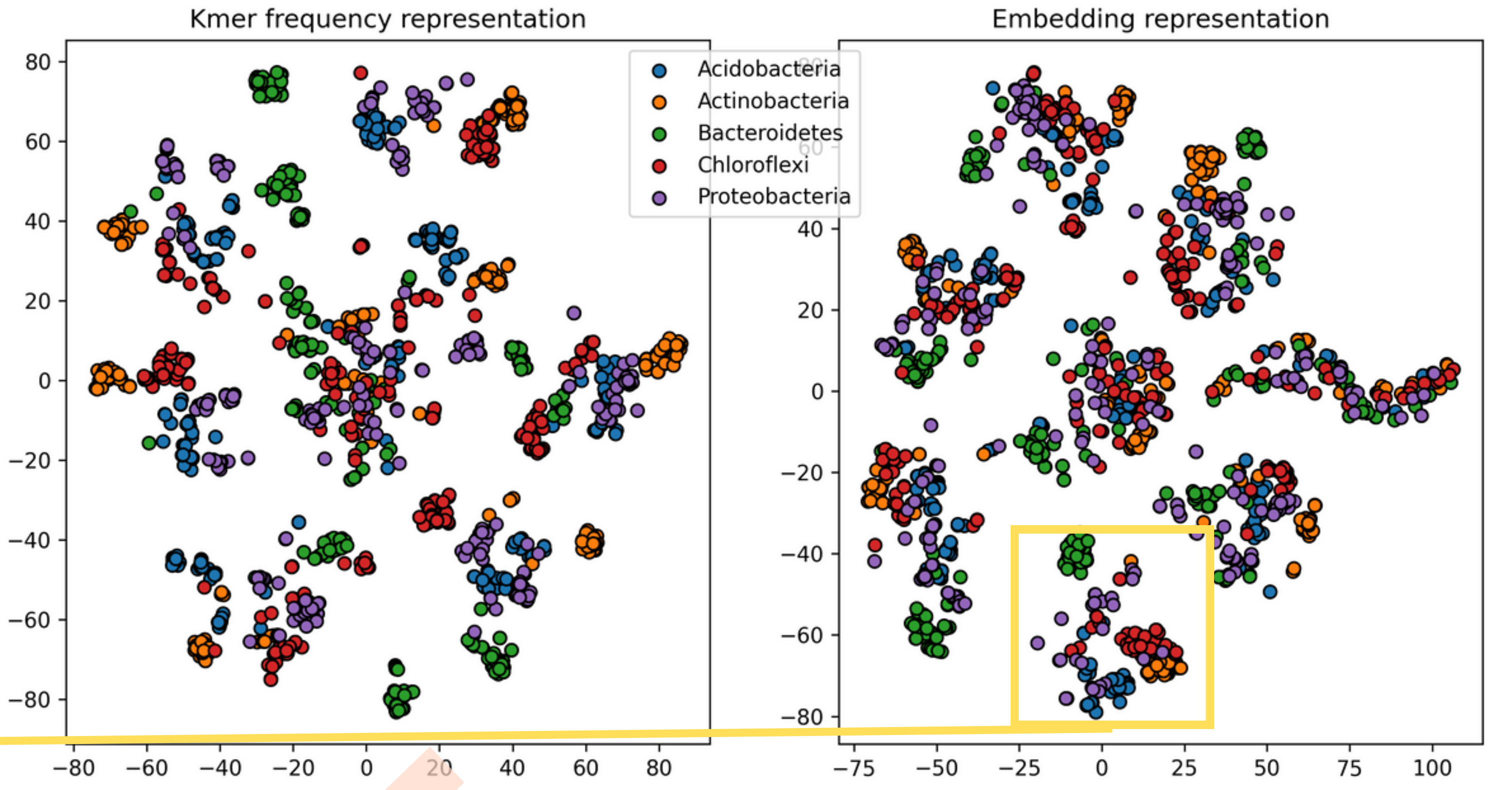
- Study emphasis: Overfitting effects on pre-training tasks
- Highlight: Potential lack of generalization to other datasets in downstream tasks, despite low loss for MLM (Masked Language Model)

Phylum Distribution and Positional Insights of Embeddings(6mer)

Fragments positions in 16s dataset



Top five phyla in 16s dataset



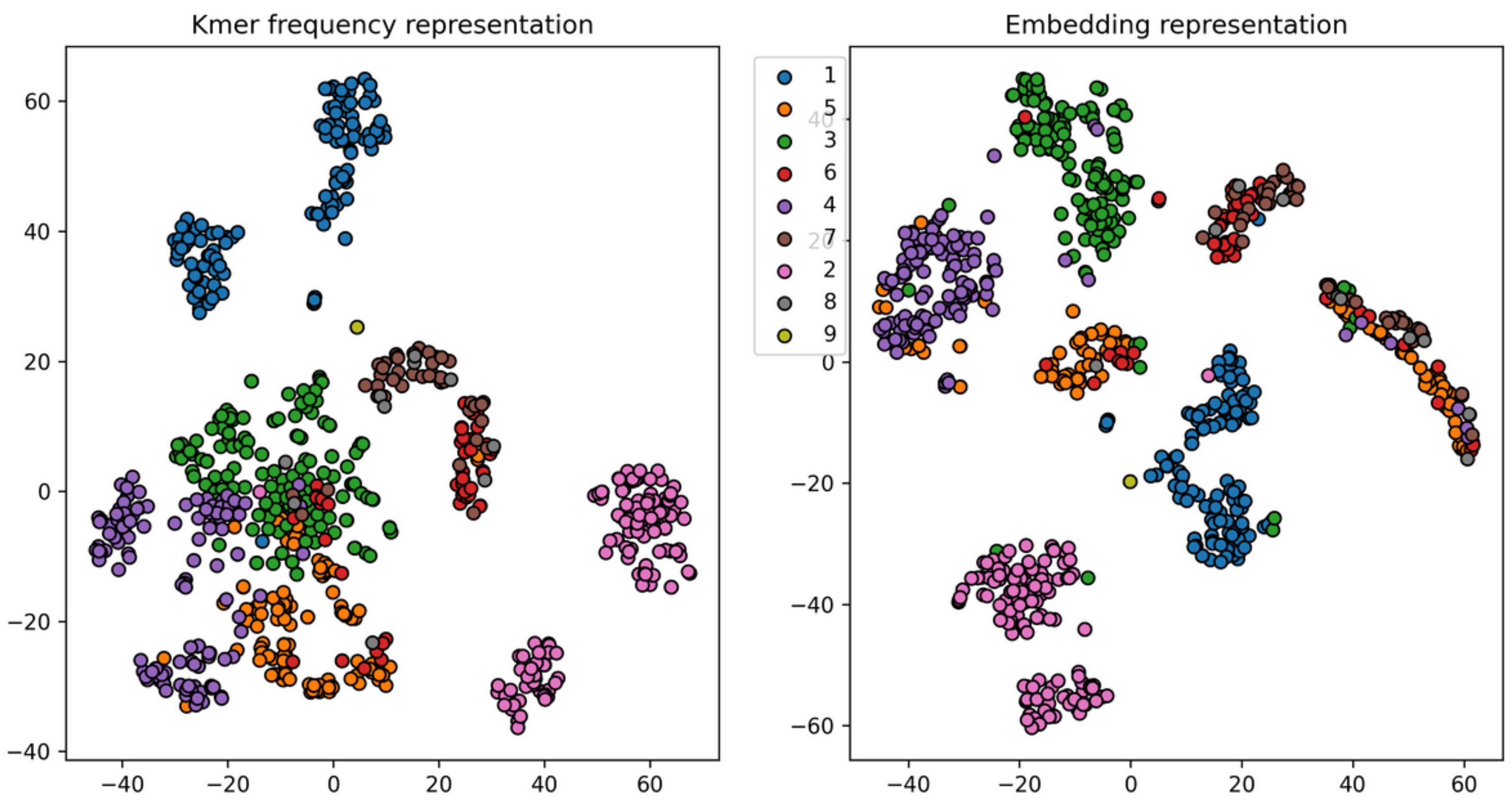
- 8 clusters representing positions of fragments in sequences
- Model excels in accurately classifying local clusters within each positional cluster

- It is highly interesting to utilize such a pre-trained model for positional-based applications, such as detecting variable regions in 16S rRNA genes .

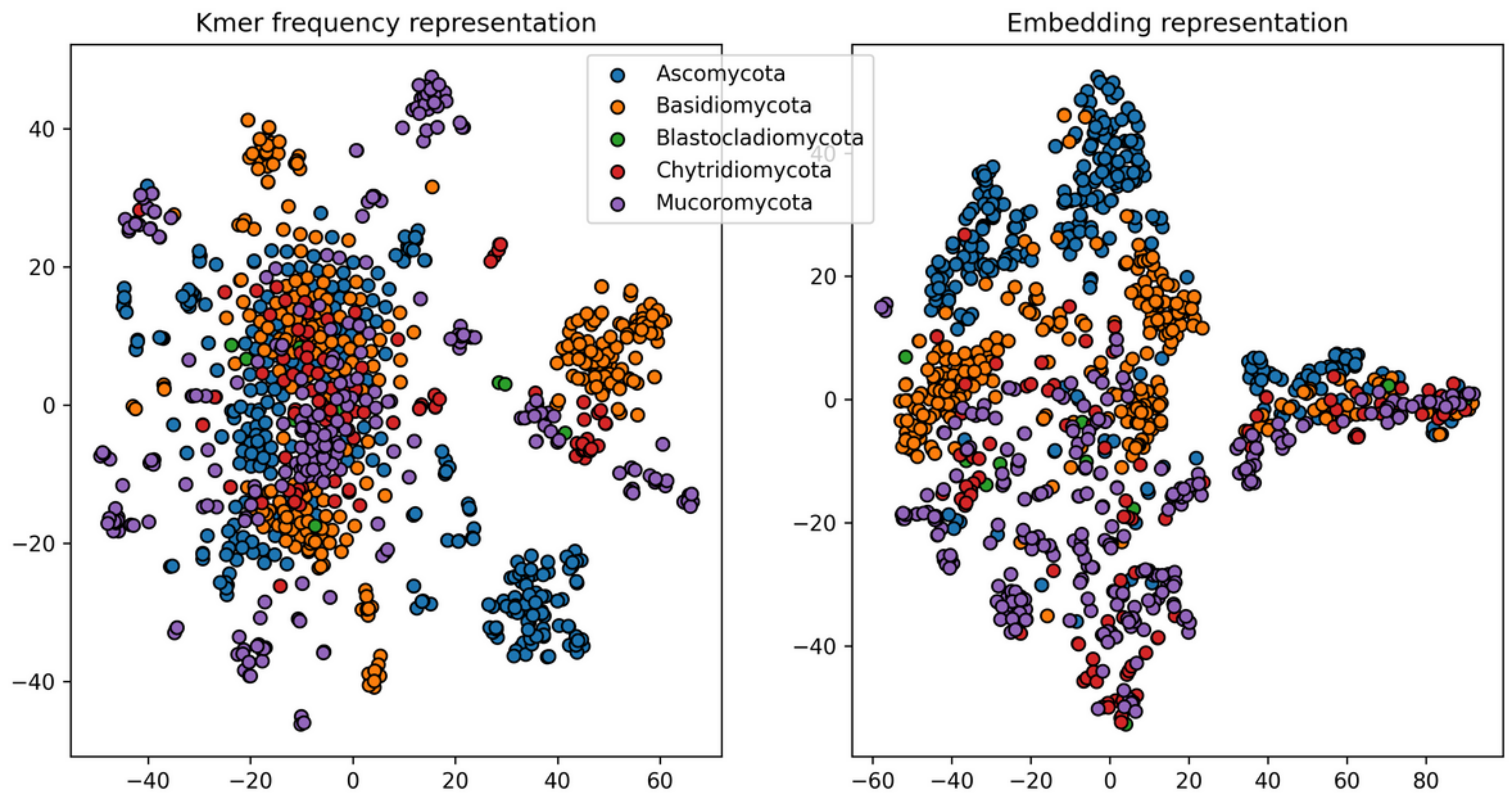
Drexel UNIVERSITY

Phylum Distribution and Positional Insights of Embeddings(6mer)

Fragments positions in 28s dataset



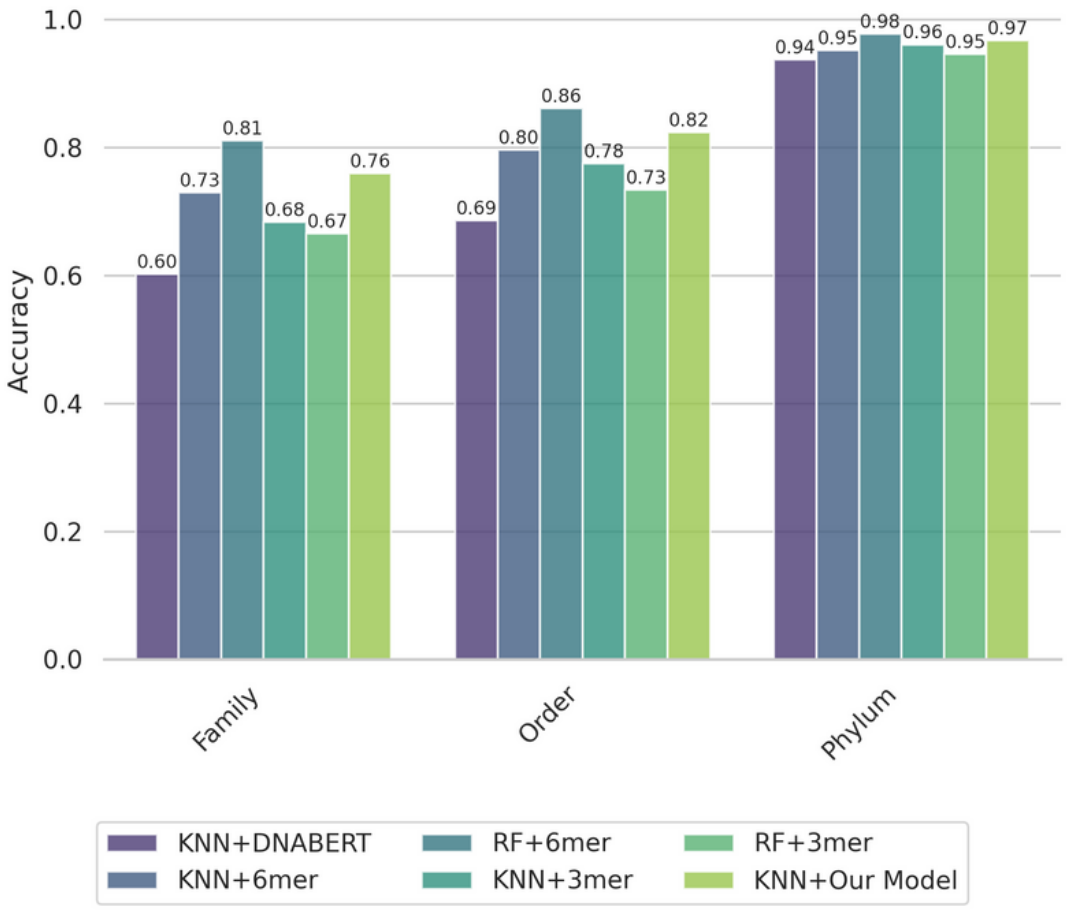
Top five phyla in ITS dataset



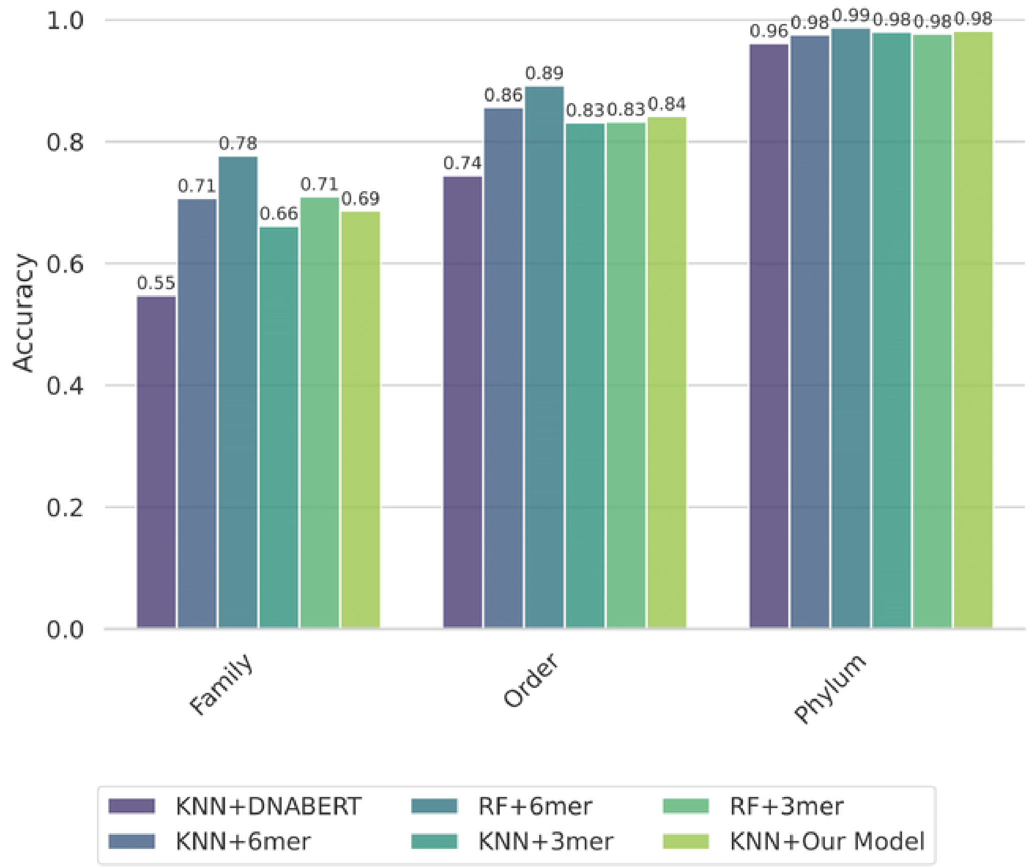
- *Pre-trained model shows potential to detect both positional and taxonomy information*
- *Tested on out-of-domain datasets not used during training*
- *Emphasis on assessing generalizability of results*

Accuracy of Taxonomic Classification

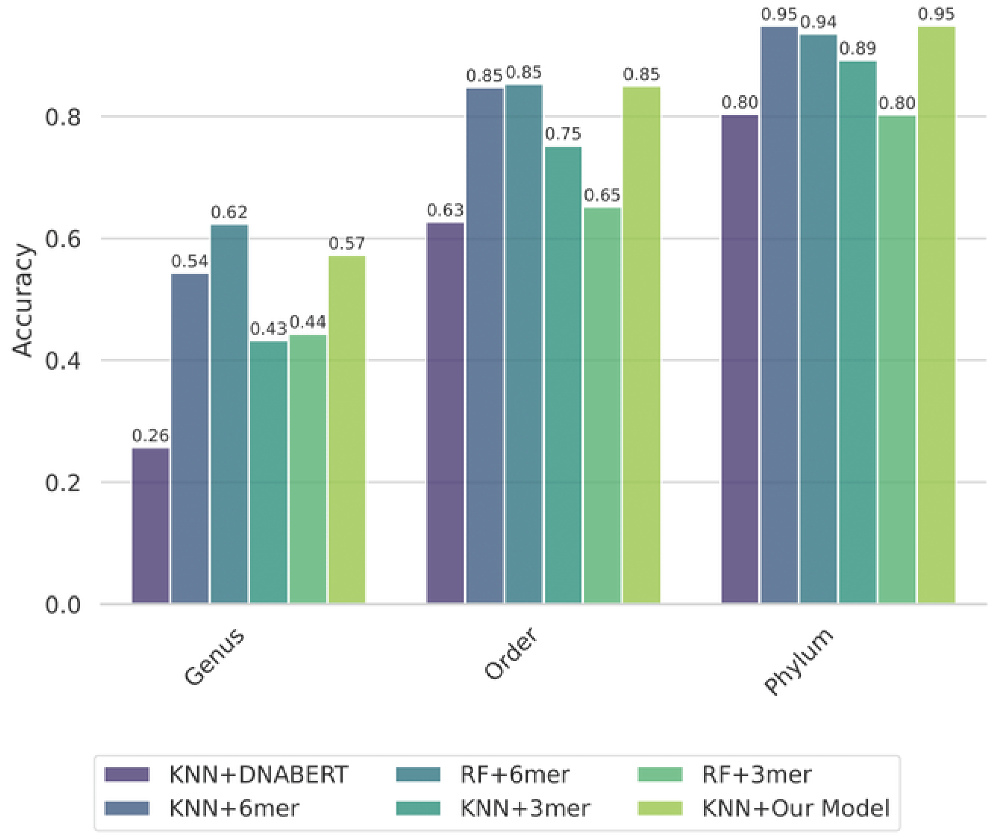
Fungi ITS Dataset



Fungi 28s Dataset



Bacterial 16s Dataset



- K-mer frequency representation combined with Random Forest (RF) yields superior accuracy in most tasks.
- RF prioritizes motif composition over the entire sequence, aligning well with the nature of sequences where taxonomy is more influenced by motif composition than positional information.
- Pretrained model excels in tasks involving local taxonomy class representation, outperforming KNN + k-mer frequency, as confirmed by t-SNE comparison results.
- Our Model Outperforms DNABERT consistently across all tasks, highlighting the challenge of generalizing a model trained on a different domain (e.g., human genomes) to taxonomy-based sequences in prokaryotes.

Drexel
UNIVERSITY

Conclusion

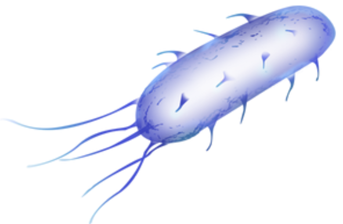
- Demonstrated the effectiveness of our optimized RoBERTa model for taxonomic classification.
- Rigorous performance evaluation across diverse downstream datasets (In-domain and out of domain) and genes.
- Notable observation: Ability to capture positional information, such as variable regions in the 16S dataset.
- The model's applicability for different downstream tasks, like SNP detection, may require higher input data and optimization for longer sequences. Highlighting the need for task-specific fine-tuning to achieve optimal performance.

- Planned use of the pretrained language model for gene function prediction. Suggested exploration for gene function prediction to provide insights into the functional capabilities of different organisms and uncover novel biological pathways.

Drexel
UNIVERSITY

THANK YOU

Any Question ?



M.S. Refahi , B.A. Sokhansanj and G.L. Rosen

This work is supported by the National Science Foundation.
Thanks to Dr. James Brown for his invaluable advice in enhancing our work