Assessing LLMs to Improve the Prediction of COVID-19 Status

Carmen Truong c5truong@ucsd.edu
Kathleen Nguyen kan019@ucsd.edu
Lorenzo Ramos j8ramos@ucsd.edu
Sean Chan swchan@ucsd.edu
Rob Knight rknight@health.ucsd.edu
Kalen Cantrell kcantrel@ucsd.edu

Abstract

In this study, we assess the performance of four large language models (LLMs)—DNABERT, DNABERT-2, GROVER and AAM—in predicting COVID-19 status from microbiome data. Given the increasing recognition of the microbiome's role in health outcomes, we focus on how the pretraining of these models impacts their predictive capabilities. These four models were chosen for their distinct pre-training strategies: DNABERT and GROVER were trained on the human genome, DNABERT-2 incorporated multi-species genomes, and AAM was trained on 16s ribosomal RNA (rRNA) sequencing data. We assessed each model's performance by using embeddings extracted from fecal and hospital-derived 16s data labeled with COVID-19 status. For our evaluation metrics, we used AUROC and AUPRC to benchmark.

Code: https://github.com/ramosrenzo/COVID-LLM

1 Introduction 3
2 Methods 6
3 Results 9
4 Discussion 9
5 Conclusion 9
References 10
Appendices A1
B Contributions A1

1 Introduction

With the vast amount of data available today, there are plenty of opportunities to harness it for global progress, from generating personalized recommendations to improving communication across languages. This is made possible with the power of Large Language Models (LLMs). A large language model is a type of machine learning model that is trained on large sets of data to learn patterns and relationships among forms of written content via deep neural networks (Toloka AI 2199). Originally developed for natural language processing (NLP), LLMs have since expanded into a wide range of sectors, including healthcare.

The COVID-19 pandemic highlighted the importance of data in shaping public health responses and accelerating medical research. The virus was first detected in December 2019 in Wuhan, China when patients experienced symptoms of an atypical pneumonia-like illness from an unknown cause (Centers for Disease Control and Prevention 2024). It quickly spread worldwide, causing unprecedented levels of sickness and death. The need for real-time analysis of the virus's impact and the development of treatments created vast amounts of healthcare data. AI models, including LLMs, played a significant role in analyzing this data during the pandemic, especially in processing scientific literature, summarizing research findings, and tracking developments in real time (Farhat et al. 2023). While LLMs weren't directly involved in developing predictive models, they assisted medical professionals and researchers by extracting relevant insights from data and providing accessible information to both fields. As a result, LLMs contributed to vaccine development and medical research. Not only were LLMs used to process scientific literature and summarize research findings, they were also involved in predicting COVID-19 status. These models were applied to various types of data, such as text-based descriptions, genomic sequences, and even audio recordings. Text-based LLMs, such as BioBERT and PubMedBERT, were used to analyze clinical records, extract medical information, and identify patterns linked to COVID-19 diagnosis and patient outcomes. Genomic LLMs, such as GenSLMs, were used to classify and cluster different COVID-19 genome sequences by distinguishing between variants (Zvyagin et al. 2023). Additionally, LLMs were employed to analyze speech and audio data, detecting COVID-19-specific vocal biomarkers in coughs or speech patterns (Anibal et al. 2024). Through these applications, LLMs provided insight and supported the development of predictive models for COVID-19, making a contribution to pandemic response efforts.

Although the Public Health Emergency has ended, COVID-19 continues to affect people globally. The virus remains highly mutative, with new variants likely to emerge, which presents the ongoing challenges of tracking and managing its spread (Markov et al. 2023). However, the many lessons learned from the pandemic continue to drive progress in combating COVID-19 and future health crises. The ability of Large Language Models to track and offer insights into COVID-19 data not only improves our response to the virus, but they also serve as a test case for how LLMs can transform healthcare. By improving models and processing more datasets, LLMs can assist in extracting information from medical content, supporting public health communication, and aiding in the development of predictive models for future health crises. Continuing to improve the capabilities of Large Language Models will not only strengthen the ability to predict and manage COVID-19, but also pre-

pare for the broader application of LLMs. LLMs are not just a tool for responding to current issues, but as a means to shape a better, data-driven future in healthcare and beyond.

Our project leverages the power of Large Language Models, with a focus on pre-trained genomic transformers, to improve the current state of predictive models for COVID-19 diagnostics. We will compare the diverse approaches of Random Forest Classifiers, DNABERT, DNABERT-2, GROVER, and Attention All Microbes (AAM) to determine which method most effectively integrates Large Language Models into microbiome-based COVID-19 predictions. The objective is to improve diagnostic classification by enhancing both prediction accuracy and feature selection.

At the core of the investigation and comparative framework is the application of Random Forest Classifiers, a machine learning algorithm. When Random Forest was applied to classify microbes associated with COVID-19, it resulted in a high predictive accuracy among each sample of nares, stool, forehead, and floor inside the hospital (Marotz et al. 2020). Beyond the original study and its use of Random Forest Classifiers, we want to utilize Large Language Models. Contemporary solutions have increasingly relied on LLMs, which benefit from extensive pre-training on genomic data, offering its insight in analyzing biological information.

Models like DNABERT and GROVER both adapted the transformer architecture from Bidirectional Encoder Representations from Transformers (BERT). In addition to BERT's abilities, DNABERT and GROVER have been specifically designed to interpret biological sequences. DNABERT, for instance, focuses on DNA sequence data and is effective at predicting disease-associated genetic variants. GROVER is optimized to process both DNA and RNA sequences, allowing for the simultaneous analysis of multiple sequence types.

Beyond these models, our study also incorporates DNABERT-2 and AAM into our comparative framework. DNABERT-2 is a successor of DNABERT as it refines the original architecture and training process of DNABERT, leading to an improved contextual understanding of k-mers. Its superior tokenization and representation capabilities make it a promising candidate for detecting subtle genetic markers associated with COVID-19. By leveraging DNABERT-2's improved performance, we expect to capture more nuanced genomic features that may correlate with disease status. Unlike DNABERT-based models that generate embeddings at the sequence level, Attention All Microbes is specifically designed to derive sample-level embeddings from microbiome data. It employs advanced attention mechanisms to aggregate and denoise data from entire microbial communities. This approach captures global microbial interactions and community structures, which has the potential of revealing characteristic features that are indicative of COVID-19 status. AAM's focus on sample-level data offers a complementary perspective to the sequence-based embeddings used in DNABERT and DNABERT-2.

1.1 Literature Review

Over the course of the pandemic, the world experienced millions of cases and deaths, prompting the development of vaccines and treatments aimed at improving the conditions

of COVID-19. In response to the evolving crisis, accurate results of COVID-19 cases had become essential for healthcare systems to effectively prevent and control the disease (Patil, Mollaei and Barati Farimani 2023). With the help of machine learning, computational biology has been able to make advancements and reveal the potential of utilizing microbiome data to predict health outcomes (Bao et al. 2024), including infectious diseases such as COVID-19. The microbiome — the community of microorganisms residing in the human body — has been present in a range of diseases, with growing evidence suggesting its role in influencing immune responses and disease severity (Yeoh et al. 2021). A significant area of research has been the use of machine learning techniques to analyze microbiome data in the context of predicting their COVID-19 status — whether an individual has tested positive or negative for the virus.

Among classical machine learning methods, Random Forest algorithms have demonstrated strong performance in classification tasks involving microbiome data (Hernández Medina et al. 2022). Marotz, et al. (2020) applied Random Forest classifiers to predict COVID-19 status using microbiome profiles obtained from 16S rRNA gene amplicon sequencing of various sample types, including nares, stool, skin (from the forehead), and hospital floor (Marotz et al. 2020). The authors used a 20-time repeated, stratified 5-fold cross-validation to optimize hyperparameters and evaluate the model, identifying key Amplicon Sequence Variants (ASVs) linked to COVID status. The model then achieved an AUROC of 0.89 for nares samples, 0.82 for stool, and 0.79 for forehead skin, demonstrating high prediction accuracy despite the imbalance of data. To assess model performance, AUPRC values of 0.76, 0.72, and 0.7 were calculated for each sample type, reflecting strong classification capabilities. This study highlighted Random Forest as a robust method for microbiome-based COVID-19 classification, offering both high accuracy and the ability to identify microbial markers that could inform diagnostic strategies.

While Random Forest models have shown promising results, there is also an increasing interest in the application of Large Language Models in healthcare. LLMs can handle unstructured biological data, such as genomic and microbiome sequences. Models like DNABERT, DNABERT-2, GROVER, and AAM are effective in modeling such data, leveraging transformer architectures to capture relationships within biological sequences and microbial communities.

- DNABERT and DNABERT-2 are both transformer-based models, but they differ in the data they were trained on. DNABERT was primarily trained on single-species genomes, while DNABERT-2 was trained on multi-species genomes. Despite this difference, both models generate sequence-level embeddings that capture patterns within microbial DNA sequences. DNABERT excels in understanding DNA and RNA, whereas DNABERT-2 improves accuracy with more nuanced embeddings. When applied to microbiome data, these models can enhance the predictive performance of COVID-19 status classification by identifying specific sequence features associated with infection.
- GROVER is a transformer-based model that was trained on the human genome to understand and generate biological sequences. It captures patterns in these sequences, which could help identify microbial dynamics linked to COVID-19 outcomes. By

modeling biological sequence relationships, GROVER can enhance prediction models for disease status by improving the understanding of microbial features associated with what progresses the infection.

- Attention All Microbes uses attention-based mechanisms to generate sample-level embeddings by aggregating microbiome data, reducing noise, and capturing global microbial interactions. This approach focuses on entire microbial communities, which can provide more accurate predictive features for COVID-19 status.

1.2 Data Description

We utilized sequencing data and biome tables from the QIITA database (Study ID: 13092) (Gonzalez et al. 2018). The complete dataset comprises 972 samples collected from hospitalized ICU patients with COVID-19, healthcare providers, and hospital surfaces before, during, and after admission. None of the healthcare providers tested positive for COVID-19. SARS-CoV-2 was assessed using RT-qPCR and microbial communities were identified by 16S rRNA gene amplicon sequencing. We used amplicon sequence variants(ASVs) of 150 base pairs. The dataset was filtered to include only samples labeled as "not detected" or "positive" for their COVID-19 status. Additionally, we focused on four sample environments: nares (n=89), stool (n=44), forehead (n=84), and inside floor (n=120). After filtering, the final dataset had a total of 337 samples which we further divided into an 80:20 training and test split for each of the four environments.

2 Methods

2.1 AAM

The Attention All Microbes (AAM) model is an attention-based neural network designed to analyze microbial sequencing data. It better captures the contextual relationship between different parts of a DNA sequence by using attention mechanisms to capture complex patterns within microbial communities. Compared to other models, AAM outputs a sample-level embedding instead of a sequence-level embedding.

The sample-level embeddings help to reduce the influence of sequencing noise and sample variability. This "denoising" effect ensures that the latent representation reflects genuine microbial signatures, which can be crucial when relating microbiome profiles to COVID-19 outcomes. These embeddings are subsequently used as input features for downstream machine learning models. By combining the latent features extracted by AAM with clinical and demographic data, we aim to enhance the predictive performance of COVID-19 status models. We used a baseline keras model that was developed by the creator of AAM, Kalen Cantrell, and trained it on 80% of the data and made predictions on the remaining 20%. Using these embeddings, we hope to be able to predict and classify the COVID status of individuals based on their microbial data.

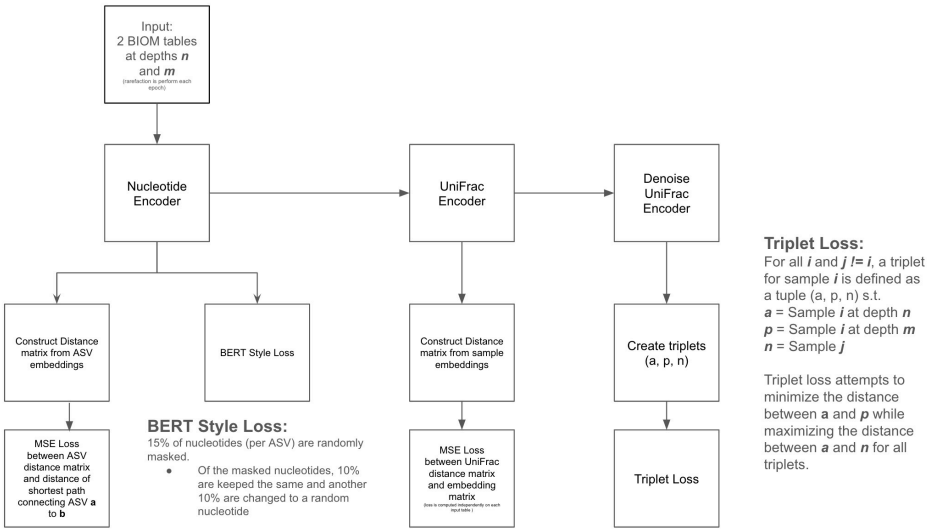

Figure 1: Architecture of AAM

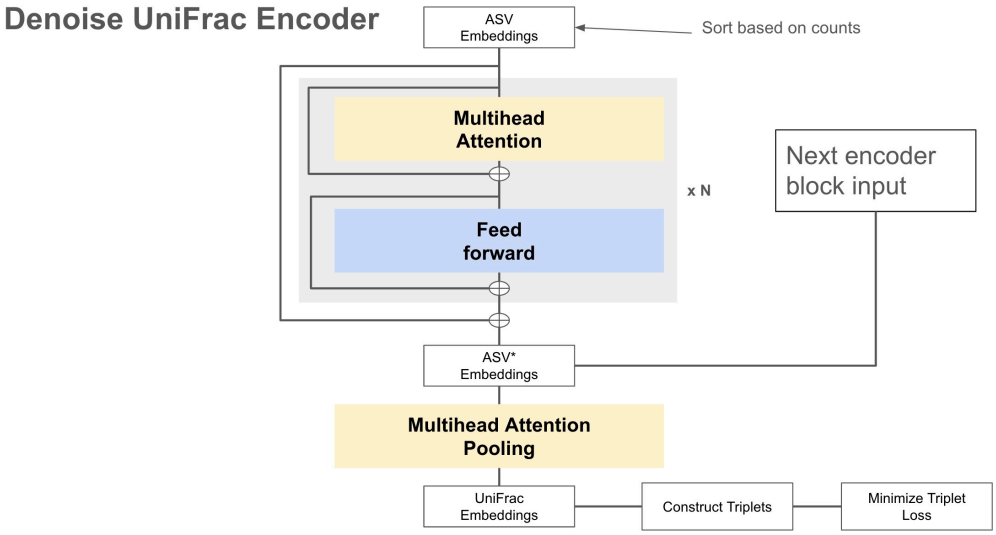

Figure 2: Architecture of Denoise UniFrac Encoder

2.2 DNABERT

DNABERT builds on Bidirectional Encoder Representations from Transformers (BERT) by adapting the transformer architecture for DNA sequences. The use of BERT relates to natural language processing tasks. For instance, BERT is used for sentiment analysis and text summarization. DNABERT, on the other hand, relates more to medical use, particularly bioinformatics. The model is used to find important patterns in DNA sequences and analyze the relationship within its context. Due to their application differences, there is a

significant contrast between the two models that are relevant to our experiment of working with microbiome data:

- Training data: BERT was trained on textual data. DNABERT was specifically trained on the human genome.
- Tokenization: BERT tokenizes text into either words or sub-words. DNABERT tokenizes DNA sequences with k-mer representation, with each different k leading to a different tokenization of a DNA sequence.

We leveraged the pre-trained DNABERT model available on HuggingFace for k-mer 6. DNA sequences were individually inputted in the model by sample and we extracted the hidden states as output. Then, the hidden states were mean-pooled to acquire final model embeddings at the sequence-level.

2.3 DNABERT-2

DNABERT-2 improves on its predecessor, DNABERT, by addressing sequence length and training limitations as well as increasing the scope of the data. (Zhou et al. 2023) The model achieves higher performance than the original in six out of seven different tasks which include epigenetic marks prediction, transcription factor prediction on both human and mouse genome, covid variants classification, promoter detection and splice site prediction. There are three major changes between the two models that are relevant to our experiments:

- Training data scope: DNABERT-2 is trained on multi-species genomes in addition to the human genome.
- Tokenization: DNABERT-2 replaces k-mer tokenization with Byte Pair Encoding (BPE) to prevent information leakage.
- Embedding type: DNABERT-2 uses Attention with Linear Biases (ALiBi) instead of positional embeddings like DNABERT. Positional embeddings introduce a 512 base pair limitation, but by using ALiBi instead, this limitation is eliminated.

We leveraged the pre-trained DNABERT-2 model available on HuggingFace. DNA sequences were individually inputted in the model by sample, and we extracted the hidden states as output. Then, the hidden states were mean-pooled to acquire final model embeddings at the sequence-level.

2.4 GROVER

GROVER is a foundation language model that adapted the transformer encoder BERT architecture (Sanabria et al. 2024). Unlike DNABERT and DNABERT-2, which were pre-trained for classification tasks, GROVER was built for general genome modeling and can be fine-tuned for other various tasks such as CTCF motif binding, promoter classification, etc. Also, in addition to BPE-generated vocabulary, GROVER incorporates five special tokens commonly used in transformer-based language models.

- Five Special Token Representations:
  - CLS - Classification token
  - PAD - Ensures uniform sequence length during batching
  - UNK - Represents unknown tokens outside vocabulary
  - SEP - Used to indicate end of sequence
  - MASK - Masked tokens
- Training data: GROVER was exclusively trained on the human genome (hg19).
- Tokenization: Like DNABERT-2, GROVER also employs BPE, which constructs a vocabulary optimized for genome sequences. BPE helps mitigate frequency imbalance regarding genomic k-mers and allows for a more flexible and informative representation of DNA sequences.
- Training objective: GROVER was trained for masked token prediction, but could easily be fine-tuned.

We used the pre-trained GROVER model available on HuggingFace. Acquiring the final model embeddings at the sequence level is identical to DNABERT-2.

3 Results

4 Discussion

5 Conclusion

References

Anibal, James T, Adam J Landa, Nguyen TT Hang, Miranda J Song, Alec K Peltekian, Ashley Shin, Hannah B Huth, Lindsey A Hazen, Anna S Christou, Jocelyne Rivera et al. 2024. "Omicron detection with large language models and YouTube audio data." *medRxiv*

Bao, Z., Z. Yang, R. Sun et al. 2024. "Predicting host health status through an integrated machine learning framework: insights from healthy gut microbiome aging trajectory." *Scientific Reports* 14, p. 31143. [Link]

Centers for Disease Control and Prevention. 2024. "COVID-19 Timeline." Jul. [Link]

Farhat, F., S. S. Sohail, M. T. Alam, S. Ubaid, Ashhad M. Shakil, and D. Ø. Madsen. 2023. "COVID-19 and beyond: leveraging artificial intelligence for enhanced outbreak control." *Frontiers in Artificial Intelligence* 6, p. 1266560. [Link]

Gonzalez, Antonio, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D Swafford, Stephanie B Orchanian et al. 2018. "Qiita: rapid, web-enabled microbiome meta-analysis." *Nature methods* 15(10): 796–798

Hernández Medina, R., S. Kutuzova, K. N. Nielsen et al. 2022. "Machine learning and deep learning applications in microbiome research." *ISME Communications* 2, p. 98. [Link]

Markov, P. V., M. Ghafari, M. Beer et al. 2023. "The evolution of SARS-CoV-2." *Nature Reviews Microbiology* 21: 361–379. [Link]

Marotz, C., P. Belda-Ferre, F. Ali, P. Das, S. Huang, K. Cantrell, L. Jiang, C. Martino, R. E. Diner, G. Rahman, D. McDonald, G. Armstrong, S. Kodera, S. Donato, G. Eckl-Mensah, N. Gottel, M. C. S. Garcia, L. Y. Chiang, R. A. Salido, J. P. Shaffer, M. Bryant, K. Sanders, G. Humphrey, G. Ackermann, N. Haiminen, K. L. Beck, H. C. Kim, A. P. Carrieri, L. Parida, Y. Vázquez-Baeza, F. J. Torriani, R. Knight, J. A. Gilbert, D. A. Sweeney, and S. M. Allard. 2020. "Microbial context predicts SARS-CoV-2 prevalence in patients and the hospital built environment." *medRxiv [Preprint]*. [Link]

Patil, Saurabh, Parisa Mollaei, and Amir Barati Farimani. 2023. "Forecasting COVID-19 New Cases Using Transformer Deep Learning Model." *medRxiv [Preprint]*. [Link]

Sanabria, Melissa, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. 2024. "DNA language model GROVER learns sequence context in the human genome." *Nature Machine Intelligence* 6(8): 911–923

Toloka AI., "History of LLMs." [Link]

Yeoh, Y. K., T. Zuo, G. C. Lui, F. Zhang, Q. Liu, A. Y. Li, A. C. Chung, C. P. Cheung, E. Y. Tso, K. S. Fung, V. Chan, L. Ling, G. Joynt, D. S. Hui, K. M. Chow, S. S. S. Ng, T. C. Li, R. W. Ng, T. C. Yip, G. L. Wong, F. K. Chan, C. K. Wong, P. K. Chan, and S. C. Ng. 2021. "Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19." *Gut* 70(4): 698–706. [Link]

Zhou, Zhihan, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu.

2023. “Dnabert-2: Efficient foundation model and benchmark for multi-species genome.” *arXiv preprint arXiv:2306.15006*

Zvyagin, Maxim, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma et al. 2023. “GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics.” *The International Journal of High Performance Computing Applications* 37 (6): 683–705

Appendices

A.1 Project Proposal A1

A.1 Project Proposal

PDF

B Contributions

- Sean: set up and ran AAM Model, wrote AAM Methods section, and setup LATEX report.
- Carmen: set up and ran DNABERT model, wrote DNABERT Methods section, wrote Introduction, and wrote Literature Review.
- Lorenzo: set up and ran GROVER model, wrote GROVER Methods section, and wrote Abstract.
- Kathleen: set up and ran DNABERT-2 model, preprocessed data, prepared environments and build scripts, wrote DNABERT-2 Methods section, and wrote Data Description.