Modelling and Analyzing Variants of Concern (VOCs) in SARS-CoV-2 Genome Sequences

Sara El-Shawa[1,3] , Sheen Thusoo[1] , Elham Dolatabodi[1] , Hassaan Maan[1,2] , Deval Pandya[1] , Graham W. Taylor[1,3] and Bo Wang[1,2]

Vector Institute[1] University of Toronto[2] University of Guelph[3]

Background

COVID-19, caused by the novel SARS-CoV-2 coronavirus, has been prevalent since 2019 and continues to spread globally. As mutations accumulate through global transmission of COVID-19, VOCs are becoming more pervasive. It is critical to understand and analyze these mutations because they are more transmissible and could be immune to treatments (vaccine). This project aims to use Neural Networks to identify and analyze patterns in COVID-19 viral genome sequences that enable surveillance and characterization of VOCs.

Genome Sequence Alignment and Adjacency Matrix

The data used was a subset of GISAID COVID-19 viral genome sequences [1] that were isolated from infected individuals around the world. For this project, 1000 sequences were sampled from VOCs and normal COVID-19 genomes each (total of 2000). A multiple-sequence alignment was performed on these genome sequences using MAFFT [2]. Using the 'ape' R package [4] and the Kimura-80 model of nucleotide substitution, the DNA distance of these 1815 sequences was computed. Using the DNA Distance, an adjacency matrix (1815 x 1815) was computed. This adjacency matrix was then used to developed an undirected graph. Edges were added if the distance of two sequences exceeded a threshold value.

Visualization of Low Dimensional Embeddings

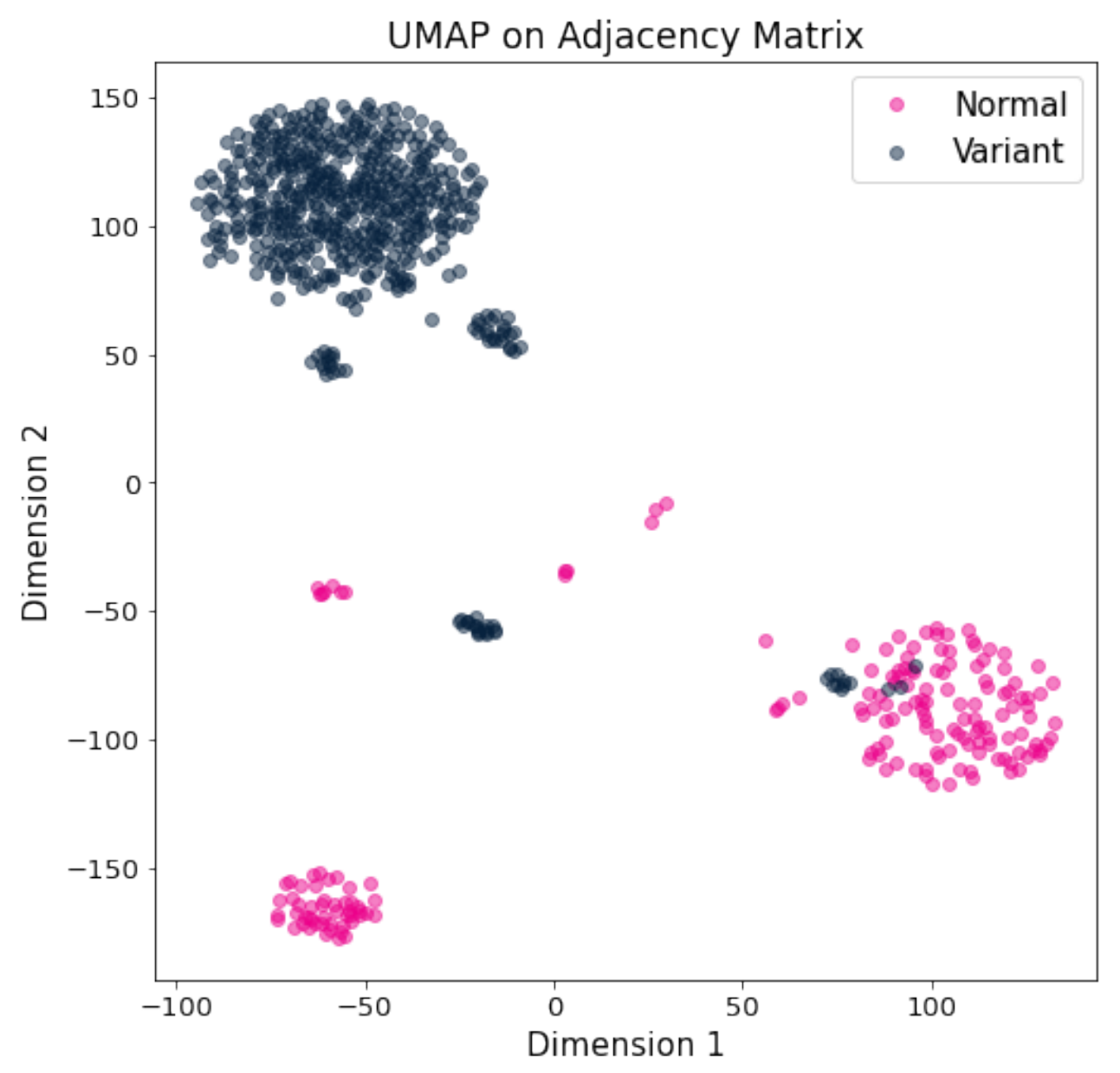

Figure 1: Projection of COVID-19 genome sequences onto 2D space using UMAP [3].

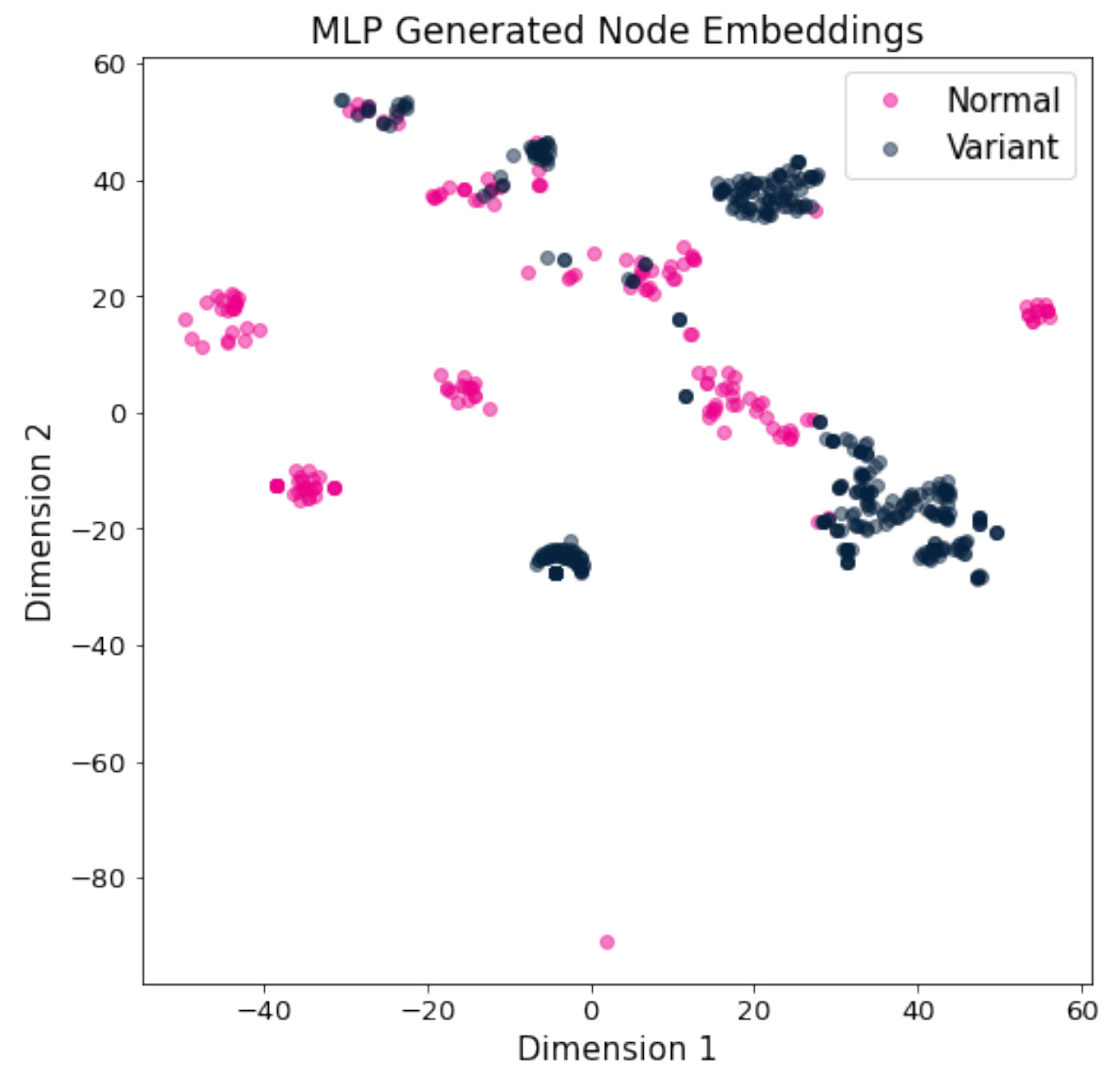

Figure 2: Node embedding developed by MLP model through Pipeline B.

Project Pipelines

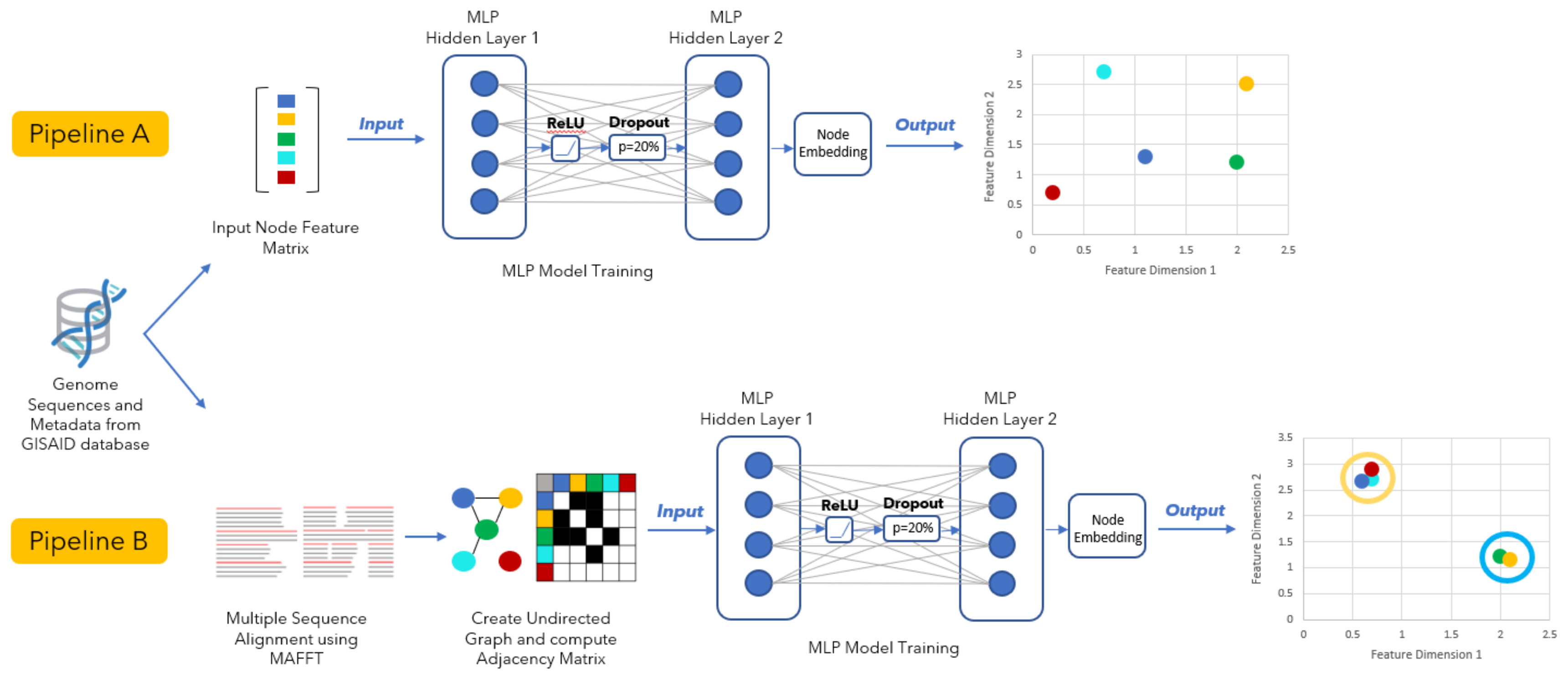

Figure 3: Pipeline A describes the process for inputting a node feature matrix to the model whereas Pipeline B describes that for inputting an adjacency matrix.

MLP Model Training

Our process included two pipelines, each with a distinct input:

- Pipeline A: where the input is the node feature matrix (one-hot encoding representation of 'Country of Origin' and 'Pango Lineage' features for each genome sequence).
- Pipeline B: where the input is the adjacency matrix that was computed via DNA distances between individual genomes.

The rest of the steps for both pipeline are identical (see Figure 3):

- The data was split and masked 70/30 into a training and test set, respectively.
- A Multi-layer Perception (MLP) Network was defined with a ReLU activation and different regularization techniques such as L2 and dropout.
- The target label for each sequence was VOC (1) or Normal (0).

Results on Test Data

Table 1: Performance Metrics on Test Data for Pipelines A and B

Pipeline	Precision	Recall	Accuracy	F1
A	0.50	0.53	0.49	0.51
B	0.97	0.93	0.95	0.94

Conclusion & Next Steps

Preliminary results show the MLP network learns the node embeddings when a distance matrix of the genome sequences is used as the input (Pipeline B). However, the same MLP model fails to learn the embeddings for unseen data when the input data is the node features matrix (Pipeline A).

As this research project is still in its early stages, the following steps will be explored next:

- Build and evaluate MLP network on data from NCBI, which consists of over 660,000 viral genome sequences.
- Calculate genetic distances between COVID-19 genomes using other models such as Jukes and Cantor model.
- Build and evaluate other pipelines using algorithms such as k-nearest neighbors (KNN) on COVID-19 genome sequences.

References

[1] S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017. doi: https://doi.org/10.1002/gch2.1018. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.1018.

[2] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL https://doi.org/10.1093/nar/gkf436.

[3] L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018.

[4] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.