Predicting Mutations in SARS-CoV-2 with a Deep Learning Model

Annabelle Hsieh | hsiehannabelle2024@gmail.com
SHINE Lab
La Canada High School, Class of 2024
USC Viterbi Department of Electrical Engineering, SHINE 2023

Introduction

With the recent pandemic, SARS-CoV-2 has been an important topic of research. Due to its rapid mutation rate and contagious nature, it is extremely difficult to create a vaccine that can work effectively against the virus and continue to be used for a long period of time. In my lab, my teammate and I used a pre-made deep learning model to predict mutations in SARS-CoV-2 and tested its accuracy and learning loss with different sequence lengths.

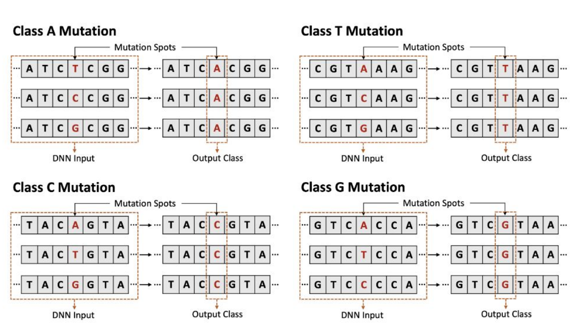

Figure 1. An example of how the deep learning model works as well as its objectives.

Objective & Impact of Professor's Research

The Cyber-Physical Systems lab focuses on using machine learning algorithms and mathematical models to analyze complex networks and process the rules and patterns that define the relationships in those systems. Networks play an important role in much of our everyday life with some of the most prominent examples being the Internet, our social network (i.e. relationships with other people), and our biological network. Understanding these systems can help lead to a greater and more efficient understanding about how these networks work and what they will do in the future.

Acknowledgements

I would like to thank Professor Bogdan for giving me this research opportunity as well as my mentors Xiongye Xiao and Qi Cao for guiding me through my project. Additionally, I would like to thank Marcus Gutierrez (my Center Mentor), Vela Benedicto (my teammate), and ChatGPT (a large language model) for supporting me through this research process.

Research & Learning Process

To complete this project, there were lots of outside information and skills that we needed to know beforehand.

1. Data collection

Mutation Name	Position	Type	Sequence Cut	Is Same
C241T	241	C -> T	GCCGATCATCAGCACATCTAGGTTT'C'GTCCGGGTGTGACCGAAAGGTAAGA	TRUE
A405G	405	A -> G	TTATCAGAGGCACGTCAACATCTTA'A'AGATGGCACTTGTGGCTTAGTAGAA	TRUE
T670G	670	T -> G	GTAATAAAGGAGCTGGTGGCCATAG'T'TACGGCGCCGATCTAAAGTCATTTG	TRUE
C2790T	2790	C -> T	CAAGGTTACAAGAGTGTGAATATCA'C'TTTTGAACTTGATGAAAGGATTGAT	TRUE

Figure 2a. (above) A picture of the csv file I coded.

Figure 2b. (to the left) The company we collected data from.

2. Understanding network science and deep learning

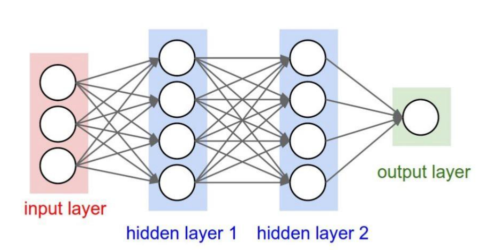

Figure 3a. (above) A visual of how deep learning works.

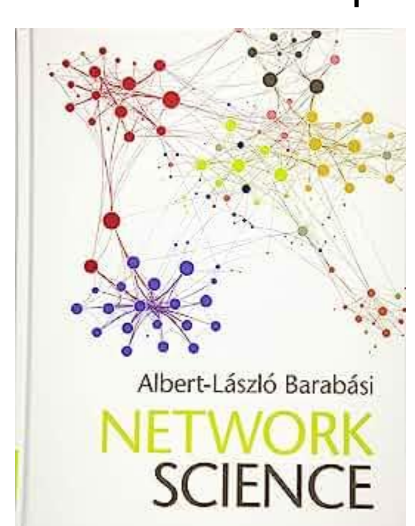

Figure 3b. (to the right) The textbook we read to learn the basics of network science.

3. Learning how to use PyTorch + PyMOL

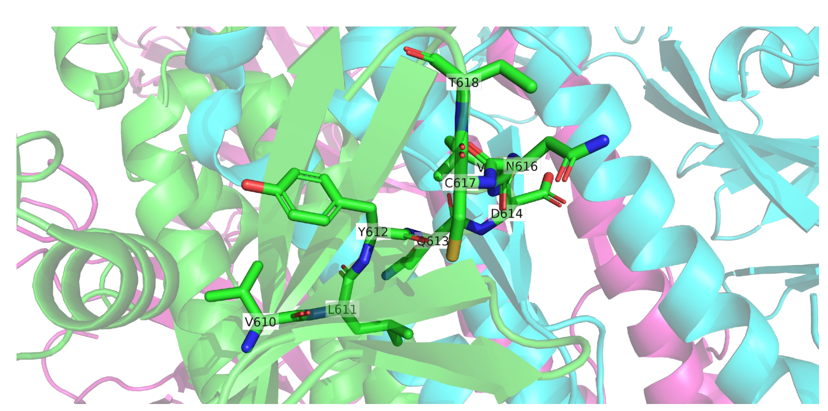

Figure 4. A model of the nine predicted amino acids within the D614G mutation.

4. Understanding and adapting the code

Figure 5. An image of the code written to produce a confusion matrix that depicts the model's accuracy.

```
from sklearn.metrics import confusion_matrix

actual = [y[0].cpu() for y in y_true_save]
predicted = [y[0].cpu() for y in y_pred_save]

cm = confusion_matrix(actual, predicted)

print(cm)

[[ 945  165  134   24]
 [ 125 1218   23  101]
 [  96   43 1160  146]
 [  23  128  112  841]]
```

Methods & Results

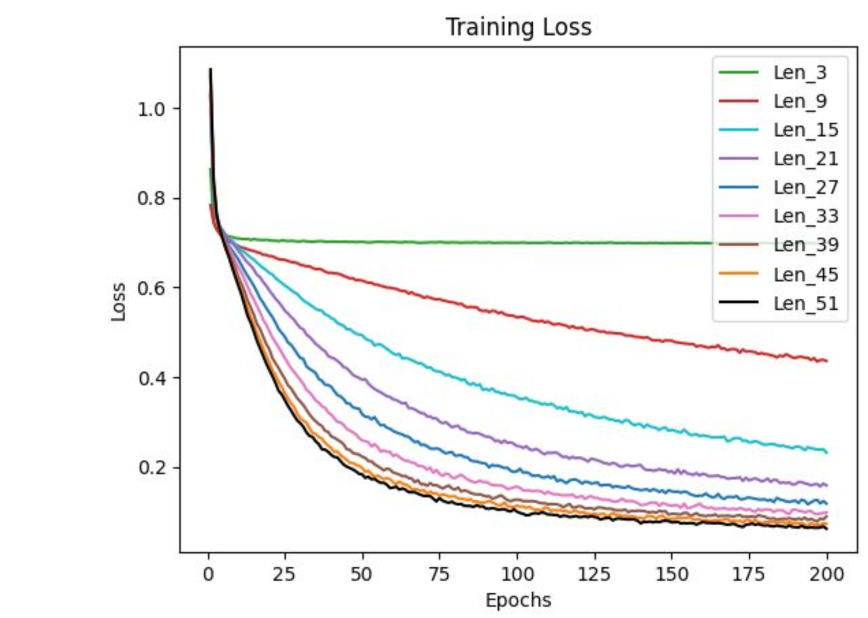

Figure 6a. This graph depicts the training loss values for each respective sequence length. As seen in the graph, as the sequence length increases, the training loss for the deep learning model exponentially decreases to a greater extent.

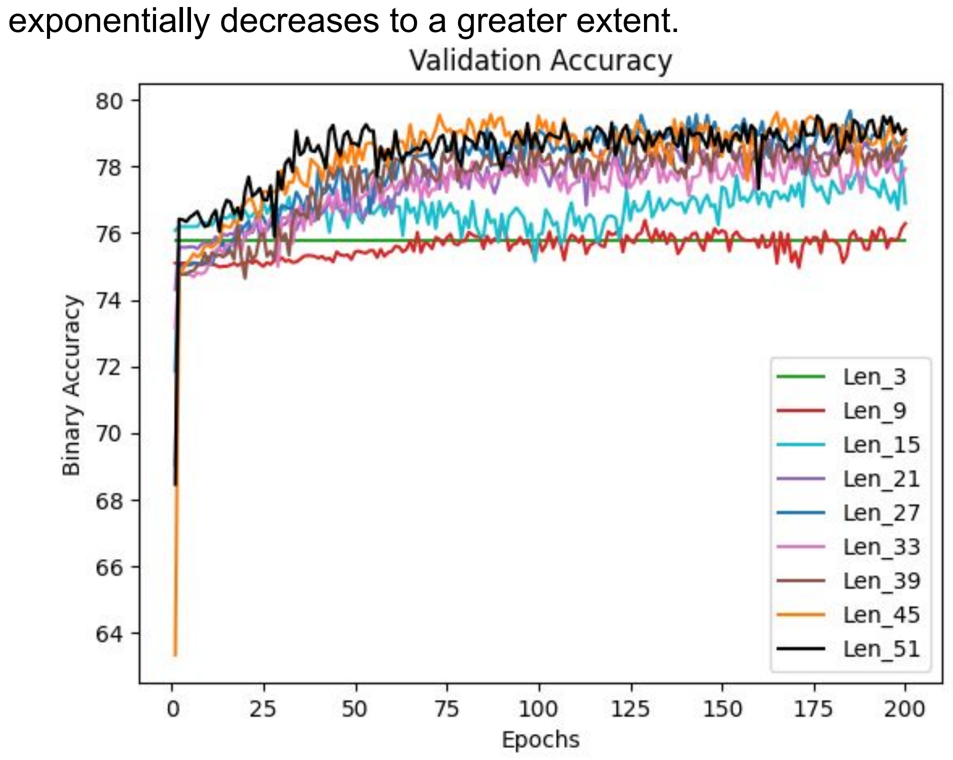

Figure 6b. This graph illustrates the impact of sequence length on binary accuracy for the deep learning model. After the sequence length of 21, there seems to be an overall limit to the impact of sequence length on binary accuracy. However, sequence lengths of 27 and 51 appear to perform best.

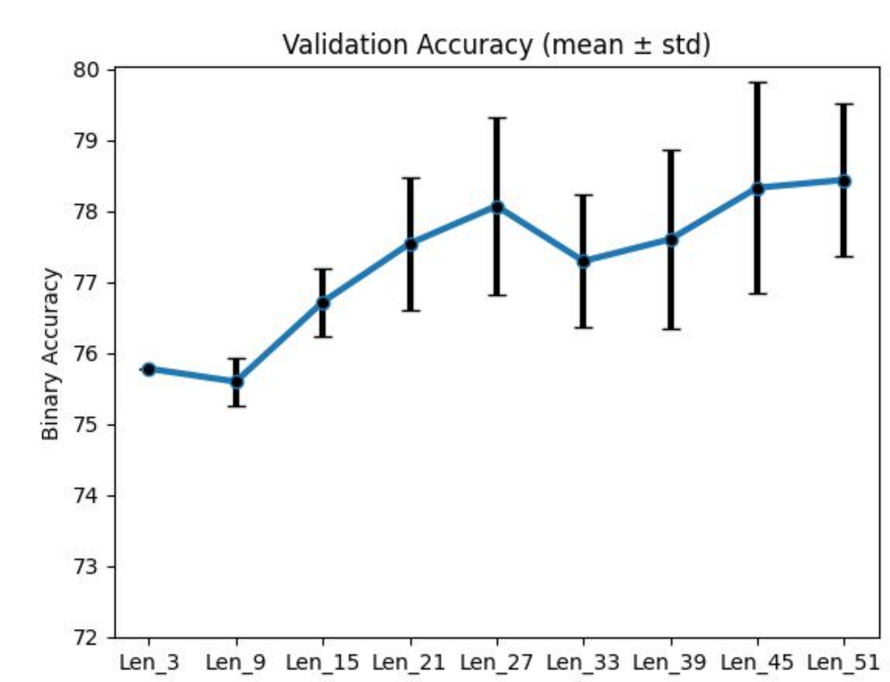

Figure 6c. The chart above shows the mean binary accuracy for each sequence length as well as the standard deviation. Based on the figure, sequence lengths of 27 and 51 appear to have the best binary accuracy with the least standard deviation.

Results Analysis

Overall, the model performed best when cut into sequences with a length of 27 bases. As see in the confusion matrix below, the model was correct more than 70% of the time for all bases and reached over 80% accuracy for bases C and T.

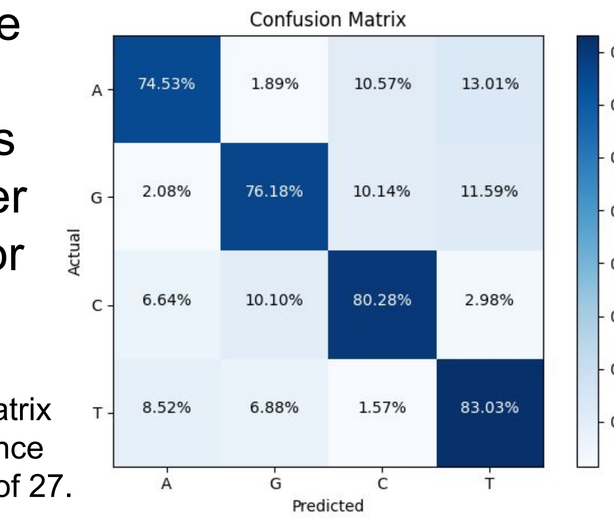

Figure 7. A confusion matrix of the model's performance with a sequence length of 27.

Next Steps for You & Advice to Future SHINE participants

In the future, I would like to look more into the application of network science and deep learning, especially in relation to mental health as it is a subject that I feel is very relevant today. When I initially started, the work seemed overwhelming and almost impossible to do given my little experience in research. However, with time, I definitely picked up many important skills from my mentors. To future SHINE students, I would recommend not being afraid to ask questions and to come in with the mentality that you are here to learn, so it is ok if you cannot exactly match the work the Ph.D. students are doing.

Citations

"Network Science by Albert-László Barabási." *BarabásiLab*, 2023, networksciencebook.com/chapter/1. Accessed 19 July 2023.
"GISAID - Gisaid.org." *Gisaid.org*, GISAID, 2023, gisaid.org/. Accessed 19 July 2023.
Jain, Siddharth, et al. *Generator Based Approach to Analyze Mutations in Genomic Datasets*. Vol. 11, no. 1, 26 Oct. 2021, www.nature.com/articles/s41598-021-00609-8, https://doi.org/10.1038/s41598-021-00609-8. Accessed 19 July 2023.
Xiao, Xiongye, et al. *Deciphering the Generating Rules and Functionalities of Complex Networks*. Vol. 11, no. 1, 25 Nov. 2021, www.nature.com/articles/s41598-021-02203-4, https://doi.org/10.1038/s41598-021-02203-4. Accessed 19 July 2023.