USC Viterbi
School of Engineering

Prediction of SARS-CoV-2 Mutations Through Deep Learning Algorithms and Z Descriptor Properties

Vela Benedicto | vtbened@gmail.com
Crescenta Valley High School, Class of 2025
USC Viterbi Department of Electrical Engineering | Cyber Physical Systems, SHINE 2023

SHINE
Summer High School Intensive in Next-Generation Engineering

Introduction

SARS-CoV-2 has been a detrimental virus that has devastated the world since March 2020, becoming an important topic for research in the past coming years. Due to its rapid ability to mutate and generate despite existing vaccines, it is critical in being able to detect future variant mutations in order to devise the best solution for public safety.

Objective & Impact of Professor's Research

Professor Bogdan works in the cyber physical systems lab, an interdisciplinary field that combines fields such as biology and chemistry with artificial intelligence (AI). His research also involves applying deep learning techniques such as neural networks. Within our lab, we focused on predicting the next base using neural networks, also taking into consideration their z descriptor properties (affected by the R-group).

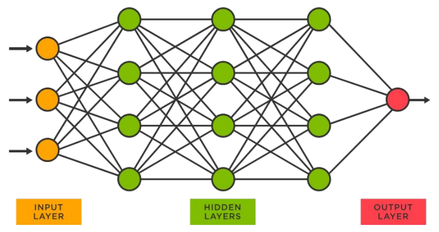


Fig 1. Neural network visual. Example of how neural networks are modeled and run through deep learning.

Methodology

1. Obtain the SARS-CoV-2 genome.
2. Using VSCode, scale the z descriptor values from values of 0-1.
3. Assign A, G, C, and T bases to numerical values of 0-3 respectively.
4. Within the prediction sequence, for every three bases, map the corresponding amino acid and concatenate with z descriptor values onto an array.
5. Using pytorch, develop a neural network and run 27 bases for 300 epochs, graphing using matplot.
6. Use pymol for additional visual representation to map out areas of prediction and the effect on the protein structure.

Data & Learning Process

In order to gain a better understanding of neural networks and deep learning algorithms using python, we first read existing literature within the research field, including consulting books, courses, and videos as seen below.

Fig 2. Collection of materials read. Gisaid, a genomic database on influenza viruses, was used to obtain reference sequences.

Among the first trials we came across was being able to slice the correct base at its mutation site. Using VSCode, we were able to generate a CSV file after extracting bases from a Gisaid fasta file. The same code was used to extract the target sequence (accounting for 25 bases before and after the mutation site) later on when applying z descriptor properties.

Using Pymol, an open source visualization tool for biological macromolecules, we are able to model the affected areas of mutations and predict their protein structure.

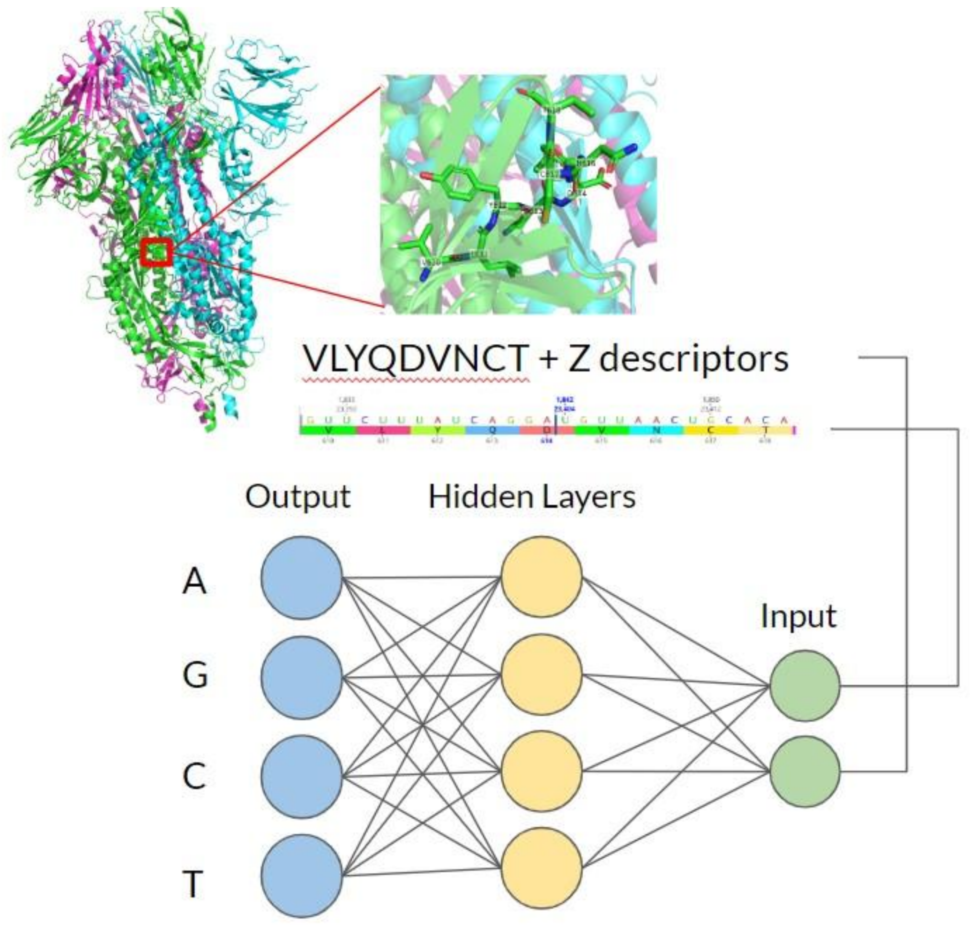
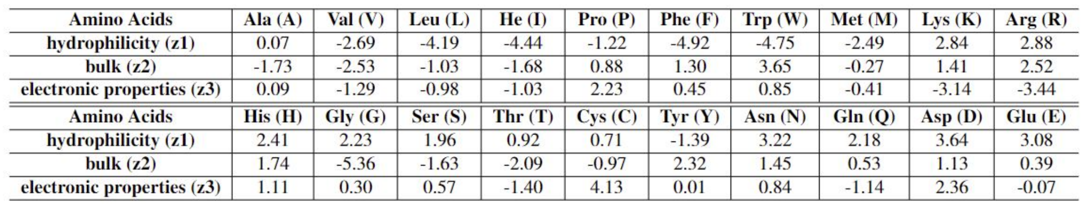


Fig 3. Model of the research process. VLYQDYNCT, amino acid abbreviations, are concatenated with the z descriptors. Both are taken as input and fed through a neural network, leading the output to be one of four bases for predictions. PC: Vela Benedicto

Results & Analysis



Amino Acids	Ala (A)	Val (V)	Leu (L)	He (I)	Pro (P)	Phe (F)	Trp (W)	Met (M)	Lys (K)	Arg (R)
hydrophilicity (z1)	0.07	-2.69	-4.19	-4.44	-1.22	-4.92	-4.75	-2.49	2.84	2.88
bulk (z2)	-1.73	-2.53	-1.03	-1.68	0.88	1.30	3.65	-0.27	1.41	2.52
electronic properties (z3)	0.09	-1.29	-0.98	-1.03	2.23	0.45	0.85	-0.41	-3.14	-3.44
Amino Acids	His (H)	Gly (G)	Ser (S)	Thr (T)	Cys (C)	Tyr (Y)	Asn (N)	Gln (Q)	Asp (D)	Glu (E)
hydrophilicity (z1)	2.41	2.23	1.96	0.92	0.71	-1.39	3.22	2.18	3.64	3.08
bulk (z2)	1.74	-5.36	-1.63	-2.09	-0.97	2.32	1.45	0.53	1.13	0.39
electronic properties (z3)	1.11	0.30	0.57	-1.40	4.13	0.01	0.84	-1.14	2.36	-0.07

Table 1: Z Descriptor values for each amino acid. Z descriptor values indicate properties of each amino acid which are influenced by the R-group. PC: Xiongye Xiao

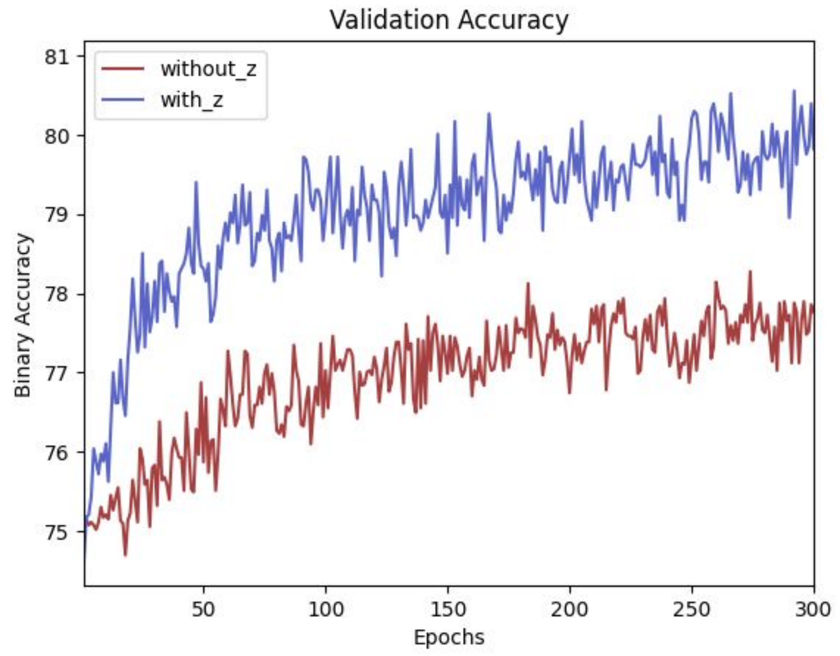


Fig 4. Validation accuracy comparison with and without z descriptor values. During the training process, the model is trained on a training dataset, and its performance is evaluated on a separate dataset called the validation dataset. The validation accuracy measures how well the model performs on the validation dataset and provides an estimate of how well the model will generalize to unseen data. As indicated in blue, there is a higher accuracy when given the z descriptor values, which emphasizes their importance in predicting bases. PC: Vela Benedicto

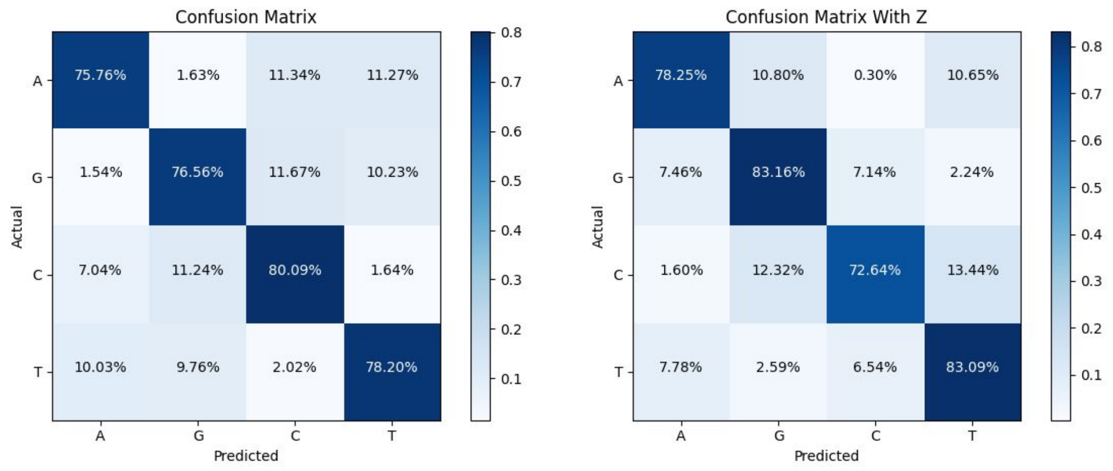


Fig 5. Confusion matrix model with and without z descriptor values. It is used to assess the performance of a classification model, with predictions spanning the x- axis and actual values consisting of the y-axis. *Left*: confusion matrix without z descriptor values, highest accuracy percentage is 80.09, average is 77.65. *Right:* confusion matrix with z descriptor values, highest accuracy is 83.16, average is 79.29. When given the z descriptor values, the model was able to perform significantly better than without the information. PC: Vela Benedicto

Conclusion

Overall, the model was able to display higher accuracy when given the z descriptor properties, which demonstrates their crucial role in determining the base predictions.

Future Work & Advice to Future SHINE participants

Future work involves testing for more mutations of SARS-CoV-2 or developing more efficient deep learning algorithms for testing. Overall, to future SHINE participants, it is best to make the most out of every opportunity and do not hesitate to access any questions to your mentors. I learned a lot from this experience and hope to use this knowledge greatly in conducting future research.

Acknowledgements

I would like to sincerely thank Professor Bogdan for the opportunity to work in his lab. I also deeply appreciate my mentor Xiong Ye Xiao for navigating me through this complex research in addition to Qi Cao for her guidance. Lastly, I am grateful to my lab partner, Annabelle Hsieh, for her support and ECE mentor Marcus Gutierrez for his patience and timeliness.

References

Jain, S., Xiao, X., Bogdan, P., & Bruck, J. (2021, October 26). *Generator Based Approach To Analyze Mutations in Genomic Datasets*. Nature News.https://www.nature.com/articles/s41598-021-00609-8

Xiao, X., Chen, H., & Bogdan, P. (2021, November 25). *Deciphering the Generating Rules and Functionalities of Complex Networks* . Nature News. https://www.nature.com/articles/s41598-021-02203-4