Supporting Information

for

Quantifying Mutational Response to Track the Evolution of SARS-CoV-2

Spike Variants: Introducing a Statistical-Mechanics-Guided Machine Learning Method

Satyam Sangeet†, Raju Sarkar†, Saswat K. Mohanty† and Susmita Roy†∗

†Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, West Bengal 741246, India

∗Corresponding Author

Email: susmita.roy@iiserkol.ac.in

This file includes:

• Schematic representation of the workflow
• Month-wise entropic variation of B.1 and Beta (B.1.351) variant in South Africa.
• Mutational Response Function of B.1 and Beta variant in South Africa.
• Month-wise entropic variation of B.1.1.28 and Gamma (P.1) variant in Brazil.
• Mutational Response Function of B.1.1.28 and Gamma variant in Brazil.
• Month-wise entropic variation of Kappa (B.1.617.1) variant in India.
• Mutational Response Function of B.1, Kappa (B.1.617.1) and Delta (B.1.617.2) variant in India.
• Month-wise entropic variation of B.1 and Delta variant (protein sequence)
• Month-wise entropic variation of B.1.1 and BA.1 variant (genomic sequence)
• Month-wise entropic variation of B.1.1 and BA.1 variant (protein sequence)
• Mutational Response Function of B.1.1 and BA.1 variant (genomic sequence)
• Mutational Response Function of B.1.1 and BA.1 variant (protein sequence)
• Feature Selection for model training
• Comparison of Training Loss and Accuracy between original and upgraded model.
• Hyperparameter tuning of the upgraded model. Training loss and accuracy of the model with varying number of hidden layers and neurons in each layer.
• Training Loss and Accuracy of the upgraded model for different activation function and learning rate.
• Receiver Operating Curve for original and upgraded model.
• Model metrics showing Accuracy, Precision, Recall and F1-Score of the upgraded model.
Figure S1: Schematic representation of the workflow. The process starts with the genomic data procurement from GISAID followed by Multiple Sequence Alignment (MSA). The entropy is calculated
for the aligned sequence in a month wise manner. On the other hand, protein sequence data is procured from NCBI followed by calculating the four features for the collected sequences. Once the first stage is completed, the next step is to build the feedforward classification model and training the model on the curated dataset. The model prediction is then analyzed for the possible amino acids into which the predicted residue could mutate into.
**Figure S2. Mutational position and pattern of Beta Variant.** Key mutations in the spike protein of Beta variant, first emerged in South Africa, highlighted in yellow (upper panel). The graph shows the time-dependent entropy pattern between the Beta variant and its ancestor B.1 variant from South Africa (lower panel). The entropic fluctuation in the NTD (21596-22472 nt, blue box) and RBD (22515-23183 nt, orange box) domain of the spike protein is very relevant when the transition happens from B.1. to Beta variant.

**Figure S3. Mutational Response Function (MRF) of B.1 and Beta variant averaged over five months.** The response function clearly captures the key fluctuation regions: The MRF of B.1 variant (upper panel) shows a distinct variation in the NTD (Blue box) and RBD (Orange box) domain of spike glycoprotein as compared to the MRF of Beta variant (lower panel). The MRF captures the entropic fluctuations being localized in the NTD and RBD domain as the virus evolves from an ancestor to a Variant of Interest.
Figure S4. Mutational position and pattern of Gamma variant. Key mutations in the spike glycoprotein of Gamma (P.1) variant first emerged in Brazil, highlighted in yellow (upper panel). Chain A is highlighted in red color whereas chain B and chain C are highlighted in dark grey color. The graph shows the entropic variation in the spike glycoprotein in a time dependent manner for B.1.1.28 variant and Gamma (P.1) variant (lower panel). The entropic variation in B.1.1.28 variant increases very slowly as highlighted in the NTD (blue box) and RBD (orange box) domain, but not significant enough to make a decisive change. On the other hand, the entropic variation in gamma variant decreases very rapidly in the NTD and RBD domain suggesting the less immediate concern with gamma variant being a dominant version of SARS-CoV-2.
Figure S5. Mutational Response Function (MRF) of B.1.1.28 and Gamma variant averaged over five months. The MRF of gamma variant (lower panel) shows the distinct region of NTD having a higher fluctuation as compared to other regions with key positions in RBD domain showing a MRF of around 0.2. On the other hand, the MRF of B.1.1.28 (upper panel) remains almost constant suggesting a clear fluctuation throughout the spike glycoprotein.
Figure S6. Mutational position and pattern of Kappa variant. Key mutations in the surface glycoprotein of Kappa (B.1.617.1) variant, first identified in India, highlighted in yellow (upper panel). Chain A of surface glycoprotein is highlighted in red color whereas chain B and chain C are highlighted in dark grey color. The graph shows the entropic variations of B.1 and Kappa variant in a time-dependent manner (lower panel). The entropic variation of NTD region (blue box) in B.1 variant in month of January 2021 decreased drastically as compared to previous months, suggesting that although the Kappa variant was identified in the month of February 2021 in India along with Delta variant, but it was already in circulation in the host population from December 2020 to January 2021. On the other hand, the entropic variation of Kappa variant shows clear fluctuations in the NTD region suggesting a prominent change occurring in NTD domain. The RBD domain (orange box) shows a similar pattern in both B.1 variant and Kappa variant.
Figure S7. Mutational position and pattern of Delta variant. Key mutations in the surface glycoprotein of Delta (B.1.617.2) variant, first identified in India, highlighted in yellow (upper panel). Chain A of surface glycoprotein is highlighted in red color whereas chain B and chain C are highlighted in dark grey color. The graph shows the entropic variations of B.1 and Delta variant in a time-dependent manner (lower panel). The entropic variation of NTD region (blue box) in B.1 variant in month of January 2021 decreased drastically as compared to previous months, suggesting that although the Delta variant was identified in the month of February 2021 in India along with Kappa and B.1.617.3 variant, but it was already in circulation in the host population from December 2020 to January 2021. On the other hand, the entropic variation of Delta variant shows clear fluctuations in the NTD region suggesting a prominent change occurring in NTD domain. The RBD domain (orange box) shows a similar pattern in both B.1 variant and Delta variant.
Figure S8. Mutational Response Function (MRF) of B.1, Kappa (B.1.617.1) and Delta (B.1.617.2) variant. The MRF clearly captures the transition of the entropic fluctuation from (a) ancestor (B.1) variant to (b) Variant of Concern (Kappa) to the (c) Variant of Interest (Delta). As we go from B.1 to Delta, the fluctuation becomes very localized with more pronounced fluctuation response in the NTD domain (blue box) of Delta variant as compared to a constant fluctuation in the B.1 variant. This behavior of the virus suggests a key role of fluctuation in the NTD domain and how it influences the infectivity and spread of the virus in the general population. Throughout the transition from B.1 to Kappa to Delta variant, the MRF
of RBD domain (orange box) decreases suggesting an important role of NTD fluctuation in the transmissibility of the virus.

Figure S9. Response of Receptor Binding Domain (RBD) of B.1 and Delta variant with time. (a) Mean entropic response of RBD of B.1 (red) and Delta (blue) Surface Glycoprotein sequence with progressing time. For both the variants we didn’t capture much difference in the entropic fluctuation of RBD domain (b) MRF response of RBD of B.1 (red) and Delta (blue) variant. MRF also captures a similar magnitude of fluctuation of RBD domain for both variants.
Figure S10. Mean Entropy and Mutational Response Function of Receptor Binding Domain (RBD) of B.1 and Delta variant (protein sequence) (a) Mean entropy of RBD of B.1 (red) and Delta variant (blue). As can be seen that not much of a mutational jump is observed when the transition from ancestor to a VOC occurs. (b) MRF for B.1 (red) and Delta variant (blue) calculated by taking 3 months’ time average. It can be seen that the mutational jump is not significant for the RBD.
Figure S11. The graph shows the entropic variations of protein sequence of B.1 and Delta variant in a time-dependent manner. The entropic variation of NTD region (green shade) in B.1 variant in month of January 2021 decreased drastically as compared to previous months, suggesting that although the Delta variant was identified in the month of February 2021 in India along with Kappa and B.1.617.3 variant, but it was already in circulation in the host population from December 2020 to January 2021. On the other hand, the entropic variation of Delta variant shows clear fluctuations in the NTD region suggesting a prominent change occurring in NTD domain. The RBD domain (yellow shade) in Delta variant shows very less entropic variation which decreases further as the time goes by.
Figure S12. Mutational Entropy of B.1.1 variant and BA.1 variant (Omicron) genomic sequence. The graph represents the entropic variation in a time dependent manner for both the variants. As can be seen, the entropic variation of N-Terminal Domain (NTD) (cyan shade) and Receptor Binding Domain (green shade) for B.1.1 variant starts to increase from Feb 2021 onwards with a higher variation in the month of July 2021 and August 2021. Subsequently, in BA.1 variant, the RBD domain shows a slightly higher variation than the NTD. This can be accounted by the fact that a higher number of mutations takes place in RBD (15 mutations) of BA.1 variant than in NTD (11 mutations).
Figure S13. Mutational entropy of B.1.1 and BA.1 variant (protein sequence). The graph represents the entropic variation of B.1.1 and BA.1 variant in a time dependent manner. As can be seen, the N-Terminal Domain (NTD) (green shade) for B.1.1 variant shows a higher entropic variation until it drops in the month of December 2021. On the other hand, in BA.1 variant, the entropic variation in Receptor Binding Domain (RBD) shows a higher profile than NTD.
Figure S14. Mean Entropy and Mutational Response Function for Surface Glycoprotein and N-Terminal Domain of B.1.1 and BA.1 variant (genomic sequence). (a) The mean entropy of surface glycoprotein of B.1.1 (red) and BA.1 variant (blue) shows the mutational jump for BA1 variant from B.1.1 variant corresponding to August 2021. (b) Mean entropy for the N-Terminal Domain of B.1.1 (red) and BA.1 variant (blue). NTD also resembles a similar jump from ancestor to VOC. (c) MRF captures the mutational transition from B.1.1 (red) to BA.1 variant (blue) corresponding to the transitional point of two variants. (d) MRF values for both the variants show a similar pattern contributing to the transitional shift.
Figure S15. Mean entropy and Mutational Response Function for surface glycoprotein and N-Terminal Domain of B.1.1 and BA.1 variant (protein sequence). (a) Mean entropy of B.1.1 variant (red) and BA.1 variant (blue) shows a mutational transition corresponding to November-December 2021. The transition month doesn’t necessarily correspond to the emergent month of the variant as in the case of protein sequences, very less data is being submitted in NCBI. (b) Mean entropy of NTD of B.1.1 (red) and BA.1 variant (blue) shows a mutational jump between the two variants. (c) MRF of B.1.1 (red) and BA.1 variant (blue) captures the transition event for the two variants with a mutational jump at the transition point. (d) MRF for the NTD of both the variants depict a transition event with a higher mutational jump.
Figure S16. Mean entropy and Mutational Response Function (MRF) of Receptor Binding Domain (RBD) of B.1.1 and BA.1 variant (genomic sequence). (a) Mean entropy of RBD of B.1.1 (red) and BA.1 variant (blue) representing the mutational jump from B.1.1 to BA.1. (b) MRF of B.1.1 (red) and BA.1 variant (blue) by taking 3 months’ time average. The inherent value of the MRF for RBD of BA.1 starts at a higher value than the Delta variant (Figure S9).
Figure S17. Mean entropy and Mutational Response Function (MRF) for Receptor Binding Domain (RBD) of B.1.1 and BA.1 variant (protein sequence). (a) Mean entropy of B.1.1 (red) and BA.1 variant (blue) shows a clear mutational jump from ancestor to VOC. (b) MRF of B.1.1 (red) and BA.1 variant (blue) capturing the mutational jump between the two variants depicts a clear transitioning point between the two variants.
Omicron Variant

The entropy pattern observed in the genomic sequence of the virus shows the pattern change with the progression of time (Figure . S12, Figure . S13). The entropy of the B.1.1 variant slowly increases over time, especially in the region 21000 to 23500. This erratic behavior of the virus could be attributed due to the large number of mutations in the NTD and RBD domain of the surface glycoprotein. On the other hand, the entropic variation of the Omicron variant maintains a higher entropic variation in the RBD domain (Figure . S12, BA.1 variant green box). This behavior is more pronounced in the case of the protein sequence (Figure . S13). Initially the entropic variation of B.1.1 variant sustains its high value. But we observe a sudden drop in the variation in the month of December 2021. If we look at the mutational entropy of BA.1 variant in the month of December 2021 (Figure . S13) it corresponds to a sudden higher mutational entropy suggesting the dominant circulation of the omicron variant in the general population. The study reveals a significant result of the sustained entropy variation in the RBD region of the BA.1 variant which is different as compared to the other VOCs. Moreover, the mean entropy also captures the sudden transition of the BA.1 variant corresponding to the June-August 2021 period. This transition captures a key phenomenon suggesting that even though the omicron variant was detected at a later stage, it was already in circulation in the host population. MRF captures a similar trend for B.1.1 and BA.1 variant for both the protein (Figure . S15) and genomic sequence (Figure . S14). Furthermore, the MRF captures a clear tipping point for the RBD domain when the virus transitions from B.1.1 to BA.1 variant (Figure . S17) suggesting the pronounced role of RBD in the viral transmission of Omicron variant.

Feature Selection

Feature 1 - Pair Prediction of Amino Acids

This feature relies on the fact that different amino acids occur in different pairs in a protein sequence and a mutation in one amino acid can possibly change the amino acid pair resulting in a different pair frequency. Based on the occurrence of these variable pairs for a particular amino acid, we can calculate the actual frequency and the predicted frequency of the pairs. The surface glycoprotein sequence of the Delta variant (Accession ID: QYJ09734.1) is 1271 residue long which consists of 88 Asparagine (N) and 62 Aspartic Acid (D). The frequency of “ND” pair occurring together is:

\[
\text{Frequency of "ND" Pair} = \frac{88}{1271} \times \frac{62}{1270} = 4.29
\]  

(i.e., there will be occurrence of 4 “ND” pairs in the sequence which corresponds to the predicted frequency. The actual frequency of the “ND” pair is also 4. Thus, the difference between its actual and predicted frequencies becomes zero. In a similar fashion, each amino acid pair can be compared with its actual and predicted frequencies. As point mutation affects only a single amino acid, which is connected to two of its neighboring residues except for the terminal one, we assign each amino acid with the sum of difference between its actual and predicted frequencies.

Feature 2 - Distribution Probability of Amino Acids

This feature is based upon the occupancy of sub-populations and partitions. In a protein sequence, certain amino acids cluster together in specific regions rather than homogeneously spreading throughout the length of the protein. Thus, any change/mutation in a particular amino acid will result in the change of how that amino acid is distributed in a protein sequence. The position of each type of amino acid in a protein sequence can be viewed as a particular distribution whose probability can be calculated as [3];
Distribution Probability = \frac{r!}{(q_0!Xq_1!Xq_2!...Xq_n!)} \frac{r!}{(r_1!Xr_2!Xr_3!...Xr_n!)}X^{n-r} (Equation S2)

Where, “r” is the number of a particular amino acid, “n” is the number of partitions, “r_n” is the number of amino acids in the nth partition and “q_n” is the number of partitions with the same number of amino acids.

This distribution probability can be alluded to in statistical mechanics, which arranges the rudimentary particles in the energy state. We calculate the predicted and actual distribution of an amino acid in a protein sequence. We take the ratio of predicted to actual distribution and then take the natural log of the value obtained. This value then can be assigned as the second feature. The magnitude of the value obtained after taking the ratio is very high. So, in order to reduce the magnitude of the value and bring it down to the range of values for other features, we take the natural log.

Feature 3 - Future Composition of Amino Acids

This feature relies on the translation probabilities between the RNA codons and the translated amino acids. This feature incorporates the mutational probability of an amino acid i.e., the probability with which a particular amino acid mutates into another one. This feature can be acknowledged by the fact that an amino acid is coded by a codon. Any mutation in one of the nucleotide bases results in a different codon which translates into a different amino acid. The mutational probabilities of amino acids are presented in Table. For example, Methionine (M) is coded by the codon “AUG”. The mutation at first position of “AUG” can mutate “AUG” into “CUG”, “GUG” and “UUG” which codes for Leucine, Valine and Leucine respectively. In a similar manner, mutation at other two positions could lead Methionine to mutate into different amino acids. Thus, for Methionine we have a final mutational probability relation which looks like:

\text{Mutational Probability of Methionine (M)} = \frac{1}{9}R + \frac{3}{9}I + \frac{2}{9}L + \frac{1}{9}K + \frac{1}{9}T + \frac{1}{9}V \quad (Equation S3)

To calculate the future composition of Methionine in a sequence, we plug in the respective amino acid count in the above equation. After calculating the future composition of all the 20 amino acids, we calculate the future contribution of each amino acid to the protein sequence. We also calculate the actual contribution of each amino acid to the protein sequence and take the ratio to get the final value which is assigned as the third feature.

Fourth Feature - Amino Acid Entropy

This feature allows for capturing the effect of mutation in terms of entropy of individual amino acid residues. The rationale behind selecting this feature is conservation of residue position throughout evolution. If mutation is not occurring in the protein sequence through several generations, then the amino acid at the particular position will remain conserved and thus the entropy of the residue will be zero. But any mutation incorporation would change the amino acid pattern resulting in an increasing entropy value.

We aligned the final sequences using CLUSTALW (https://www.genome.jp/tools-bin/clustalw) and then calculated the entropy of each position for the sequence as:

\[ H_i = - \sum p_i \log(p_i) \quad (Equation S4) \]

The values obtained were then assigned as the fourth feature to the model.
Hyperparameters Tuning

We selected four major hyperparameters: Number of Hidden layers, Number of Neurons in each hidden layer, Learning rate and Activation Function. After training the model for different combinations of hyperparameters we found the optimum values to be: 2 hidden layers, 8 neurons in each hidden layer, learning rate of 0.001 and hyperbolic-tangent (TanH) activation function for the first three layers to make for an optimum model. The activation function for the final layer was determined to be a sigmoid activation function as the final output is a binary classification.

**Figure S18. Hyperparameter tuning of the model with variable number of hidden layers and neurons.**

The graph shows the training accuracy and loss when the hyperparameters were tuned for the upgraded model. The difference in training loss was not significant for 2, 3 and 4 hidden layers. Moreover, increasing the number of neurons didn’t account for any further significant difference to the model training. Similar behavior was observed with training accuracy. Thus, the final model architecture for the upgraded model was fixed with 4 inputs corresponding to 4 features, 2 hidden layers each consisting of 8 neurons and a single output for classification i.e., a 4-8-8-1 feedforward backpropagating model.
Figure S19. Training loss and accuracy for the model with different activation function for the hidden layers and variable learning rate. Sigmoid activation function was fixed for the output layer as the process is a binary classification. The model was trained with ReLU and TanH activation function for the hidden layers. There was no significant difference in the training loss of the model with ReLU and TanH activation function (a) but the training accuracy showed a better trend with TanH activation function (b). Thus, TanH function was selected as the activation function. The training loss (c) and accuracy (d) of the model showed a better behavior with a learning rate of 0.001. We didn’t go beyond a learning rate of 0.0001 to avoid the issue of getting stuck in a plateau region and model taking large time to converge.
**Figure S20. Training Loss and Accuracy of the original and upgraded model.** The original model with 3 features reached a training loss of around 0.4 with 250 epochs but the upgraded model with 4 features reached a training loss of nearly 0 with 250 epochs suggesting a better learning by the model with 4 features. The training accuracy of the original model reached a maximum of around 83% with 250 epochs iteration whereas the training accuracy of the upgraded model reached nearly 99% with 250 epochs suggesting a better accuracy of the model.

**Figure S21. Receiver Operating Curve for the original and upgraded model.** The original model had an AUC of 0.532 whereas the AUC for the upgraded model was 0.996 suggesting a better performance of the upgraded model as compared to the original model.
Figure S22. Model metrics for trained model. Graph showing the Accuracy, Precision, Recall and F1-Score of the upgraded model trained on 117 sequences. The model achieves a good accuracy of around 99% with precision being almost 99.7% suggesting a precise prediction. The model showed a score of nearly 99% for the recall suggesting that the model correctly labels the true positives in delta variant case. The F1-score of the model captures the true balance between precision and recall usually when there is an uneven class distribution. The F1-score of the model reaches a score of nearly 99.2% suggesting a better model metric and a better balance between the precision and recall.
Table S1: Numeric representation of QYJ09734.1, Delta variant sequence.

| Residue Position | Amino Acid | Input Features | Target |
|------------------|------------|----------------|--------|
|                  |            | I              | II     | III    | IV     |       |
| 1                | M          | 1              | 1.0986 | 10.80  | 0      | 0      |
| -                | -          | -              | -      | -      | -      | -      |
| 77               | K          | 2              | 1.6178 | 1.0842 | 0.6789 | 1      |
| 78               | R          | 3              | 1.6178 | 1.1320 | 0      | 0      |
| 79               | F          | 1              | 0      | 0.96990| 0      | 0      |
| -                | -          | -              | -      | -      | -      | -      |
| -                | -          | -              | -      | -      | -      | -      |
| -                | -          | -              | -      | -      | -      | -      |
| 710              | I          | 4              | 0.7368 | 1.2380 | 0      | 0      |
| 711              | A          | 4              | 0.7695 | 1.1194 | 0      | 0      |
References

1. Noa, C. E. F., Harunari, P. E., de Oliveira, M. J. & Fiore, C. E. Entropy production as a tool for characterizing nonequilibrium phase transitions. *Phys. Rev. E* **100**, 12104 (2019)

2. Mishin, Y. Thermodynamic theory of equilibrium fluctuations. *Ann. Phys. (N. Y)*. **363**, 48–97 (2015).

3. Feller, W. An introduction to probability theory and its application 3rd Edition Vol. 1 (Wiley, New York, 1968)

4. Katah, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acid Res* **30**(14), 3059-3066 (2002)