THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Quantifying Mutational Response to Track the Evolution of SARS-CoV-2 Spike Variants: Introducing a Statistical-Mechanics-Guided Machine Learning Method

Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

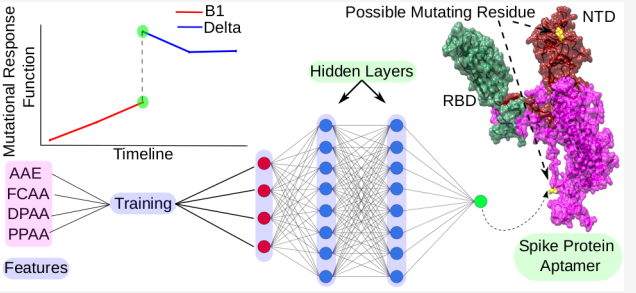
Satyam Sangeet, Raju Sarkar, Saswat K. Mohanty, and Susmita Roy*

Cite This: https://doi.org/10.1021/acs.jpcb.2c04574

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The emergence of SARS-CoV-2 and its variants that critically affect global public health requires characterization of mutations and their evolutionary pattern from specific Variants of Interest (VOIs) to Variants of Concern (VOCs). Leveraging the concept of equilibrium statistical mechanics, we introduce a new responsive quantity defined as "Mutational Response Function (MRF)" aptly quantifying domain-wise average entropy-fluctuation in the spike glycoprotein sequence of SARS-CoV-2 based on its evolutionary database. As the evolution transits from a specific variant to VOC, we find that the evolutionary crossover is accompanied by a dramatic change in MRF, upholding the characteristic of a dynamic phase transition. With this entropic information, we have developed an ancestral-based machine learning method that helps predict future domain-specific mutations. The feedforward binary classification model pinpoints possible residues prone to future mutations that have implications for enhanced fusogenicity and pathogenicity of the virus. We believe such MRF analyses followed by a statistical mechanics augmented ML approach could help track different evolutionary stages of such species and identify a critical evolutionary transition that is alarming.



INTRODUCTION

Severe Acute Respiratory Syndrome (SARS) has recently re-emerged in China in December 2019.[1] The rapid spread of the virus results in the selection of SARS-CoV-2 variants with varying mutations in the Receptor Binding Domain (RBD) and N-Terminal domain (NTD)[2−4] predominantly among its other structural domains (Figure 1). Consequently, the virus left its fresh footprints over its evolutionary trajectory, intervened by several Variants of Interest (VOIs) and Variants of Concerns (VOCs). This trajectory, although short-term, may carry a critical characteristic evolutionary signal; therefore, a thorough statistical analysis of this short-term trajectory is of utmost necessity at the current time.

The transition from one class of variant to the other involves mutational fluctuation both at the genomic level and amino acidic level. To unveil the detailed nature of the fluctuation during such transition, we have borrowed fundamental concepts concerning the fluctuations of systems described by statistical mechanics.

One of the most fundamental and general results described by statistical mechanics is deriving the relation between the spontaneous fluctuations and the response to external fields of physical observables concerning a system.[5] Although the fluctuation relations receive more attention by the phenomena involved far from equilibrium, nonetheless, a generic fluctuation−dissipation relation of statistical mechanics is rather relevant as a general concept, regardless of the Hamiltonian or equilibrium behavior of the system.[6]

For example, using equilibrium fluctuation of thermodynamic observables, we connect the heat capacity (as the response to an energy perturbation) to the energy fluctuations and entropy fluctuation, isothermal compressibility to fluctuation in volume, and number density.[7] On the other hand, using fluctuations in nonequilibrium statistical mechanics, one can derive celebrated Einstein's relation between diffusion and mobility.[8] These response functions often exhibit a sudden change when the system approaches the transition point where it is about to change its phase.[9,10] Systems undergoing such phase/state transitions often appear in the natural and social sciences.[11,12] Therefore, a lot of effort has been put into identifying an appropriate responsive order parameter for the

Received: June 30, 2022
Revised: September 16, 2022

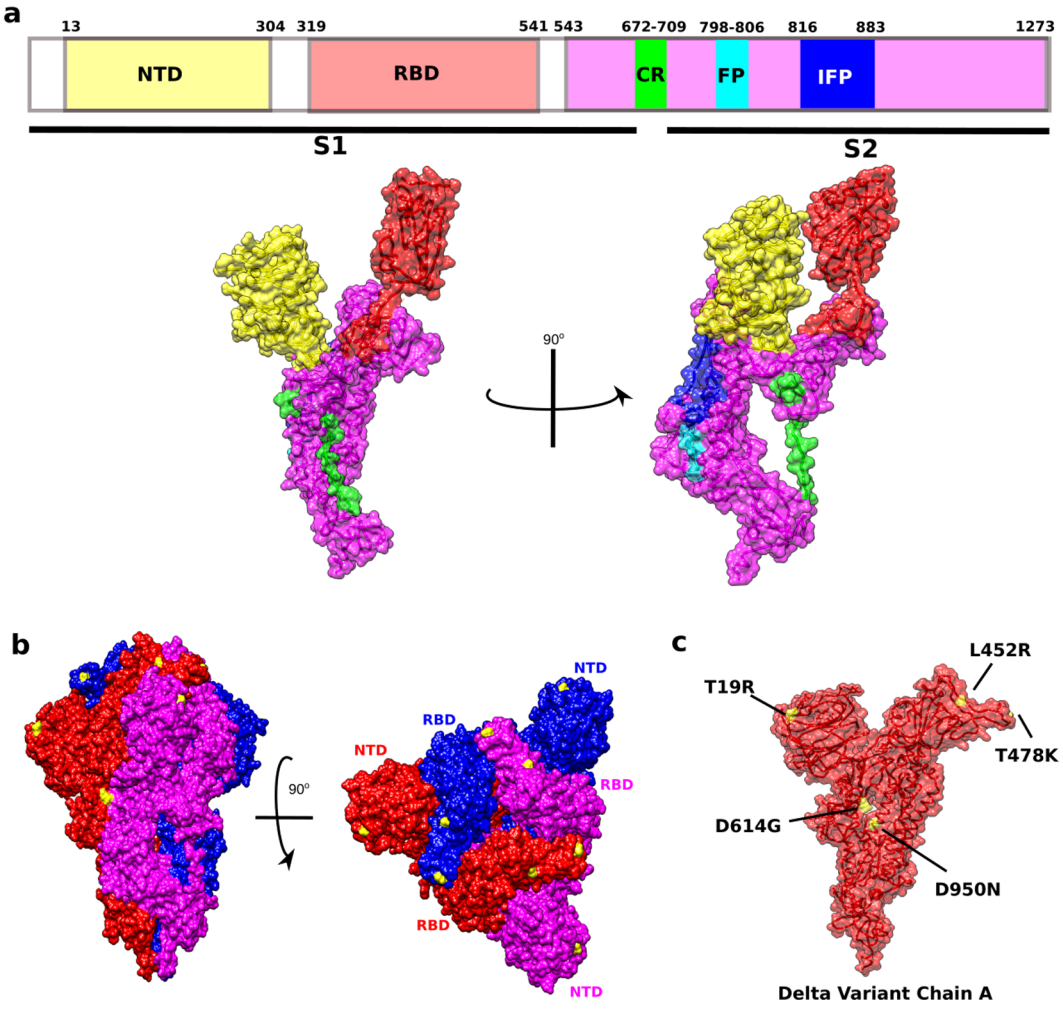


Figure 1. Illustration of the spike surface glycoprotein sequence of SARS-CoV-2 and its specific structural domains. (a) Surface glycoprotein sequence of SARS-CoV-2 shows major regions and their positions (upper panel). Yellow color corresponds to the N-Terminal Domain ranging from position 13 to 304. Red color corresponds to the Receptor Binding Domain ranging from position 319 to 541. Green color corresponds to the S1/S2 Cleavage Region (CR). Cyan color represents the Fusion Peptide, and blue color represents the Internal Fusion Peptide (IFP). Purple color corresponds to the S2 domain of the glycoprotein sequence. Lower panel corresponds to the 3D representation of a single chain of spike glycoprotein. The colored regions correspond to the upper panel representation (b) 3-dimensional representation of the spike protein of SARS-CoV-2. The figure shows the trimeric chain with Chain A (red), Chain B (blue), and Chain C (magenta) in the closed state. (c) 3-Dimensional representation of the single spike chain depicting the key mutations of the Delta variant (B.1.617.2) highlighted in yellow color.

development of early warning signals. An early warning signal may also involve an increase in variance and correlation time of that sensitive order parameter as the system approaches the transition point. A response function uniquely provides a quantitative measure of the response of a system to an external perturbation. In equilibrium statistical mechanics, such response functions are defined in terms of mean square fluctuations in the conjugate system properties. Near a phase transition or/and instability point, these response functions exhibit dramatic changes that essentially signal a change in the state of the system, for example, a critical temperature in gas−liquid or magnetic systems. Here, we follow the concept of probing phase/state transition from equilibrium statistical mechanics and attempted to capture mutational state transition by quantifying the mean value of mutational entropy and its mean square fluctuation for each domain, as described in the Methods. It is worth mentioning here that mutational fluctuations are always present in an evolving sequence due to the stochastic nature of the elementary mutation process (similar to a biomolecular reaction in a cellular system). These fluctuations exert more profound effects and can induce new functional activities in small-scale systems (like mutation occurs only in a few sites compared to the whole genomic/protein sequence but render a functional change), which may not be predicted or well captured by any mean-field/mean value description.

The viral mutation in the real world is a stochastic phenomenon.[13] Such stochasticity is always associated with any mutational event as a random/chance event.[14,15] Although the mutational transition is a complex nonequilibrium process, in a given VOI or VOC, certain mutations are still observed to survive in a time-invariant manner within a limited span and can be considered as a unique mutational steady state. When the number of such random mutational events uprises, it may be captured by a relevant fluctuation relation near a dramatic mutational transition which governs a mutational state to evolve from an old to a new mutational steady state. Information entropy has long been utilized as a parameter to

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

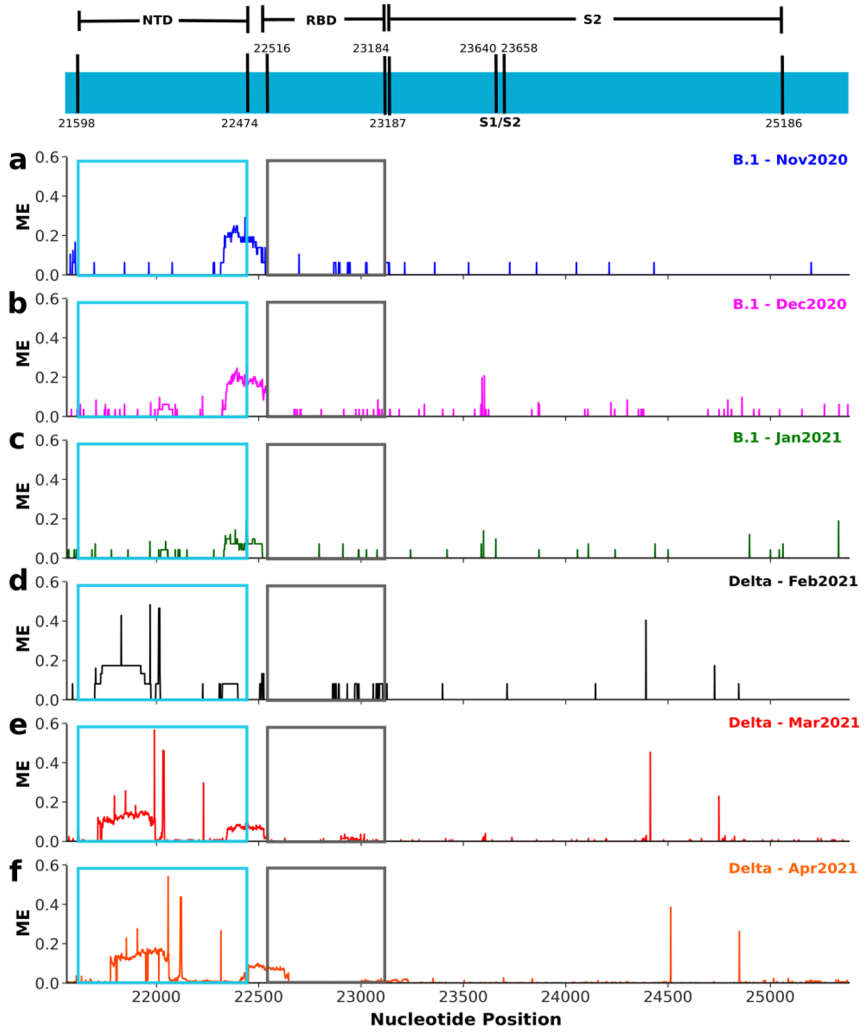


Figure 2. Mutational Entropy (ME) variation of each genomic position as time evolves. The figure shows the entropy change at the genomic level with time evolution. Upper panel corresponds to the genomic sequence of SARS-CoV-2 with distinct domains. For the B.1 variant (Ancestral variant), the mutational entropy of the NTD region (cyan box) gradually decreases with minor changes occurring in the RBD region (dark gray box) from November 2020 until January 2021 (a–c). On the other hand, the entropy change obtained in the Delta variant (B.1.617.2) starts to decrease in the RBD region (d–f) from February 2021 to April 2021.

analyze epidemics by evaluating the time of maximum diffusion of an epidemic or pandemic,[16] analyzing phylogenetically informative genetic regions, helping design primers for PCR amplification,[17] and studying genetic mutations by analyzing the entropy spectrum of the genomes.[18] Thus, we surmise that information entropy and its relevant fluctuation may collectively capture such random mutational effect during the transition from one mutational state of a variant to a new one.[19]

The surge of mutational events, with ten mutations occurring in the surface glycoprotein of the Alpha variant (B.1.1.7[20,21]) to 37 mutations occurring in the surface glycoprotein of the Omicron variant (B.1.1.529[22]), raises an important question: Is it possible to predict future mutations? How accurately can we predict these mutations? Predicting the mutational pattern of a species may provide insights into the mutation process and future activity of the species and may help design potential drugs targeting the viral species.[23,24] Machine Learning techniques offer assistance for analyzing the available mutational data with several models utilizing different methods to predict the mutations in different scenarios, such as Long–Short-Term Memory (LSTM) models[25] and Neural Networks and Rough Set theory[26] to predict mutations in viral species, Statistical Relational Learning for generating resistant mutations in HIV reverse transcriptase inhibitors,[27] Deep Convolutional Neural Network[28] to classify and predict mutation from histopathology images of lung cancer, Variational Autoencoders (VAE) to examine the heterogeneity and fluctuation of chromatin structure,[29] Interactive Interface to analyze biomedical and clinical data sets,[30] Supervised Machine Learning Model of coarse-grained molecular dynamic force fields,[31] and Multilayer Perceptron classifier[32] to predict mutations in the influenza virus. As a new approach, information entropy can be utilized for data compression and transmission and to calculate the target-class imbalances in binary classification models of machine learning.

In the present study, we analyze both genetic and protein sequence in different variants of SARS-CoV-2 and introduce a mutational response function (MRF) to quantify the entropic fluctuation during the transition from one variant to another. The fluctuation parameter captures the transition pattern both at the genomic and protein level and exhibits a clear transition

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX–XXX

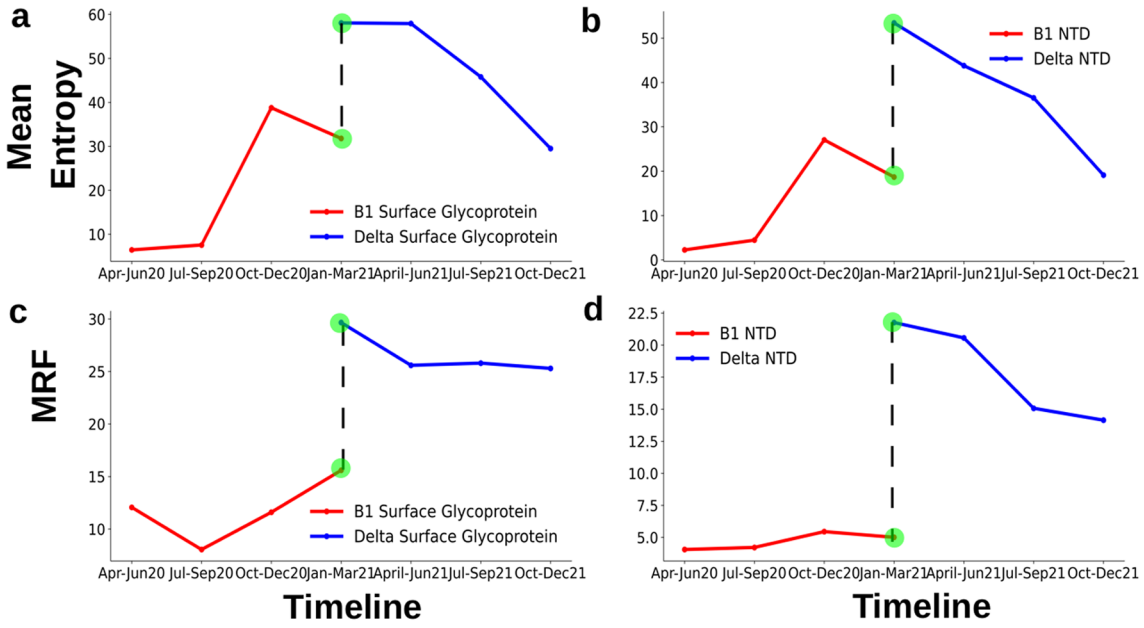


Figure 3. Mean Entropy and Mutational Response Function (MRF) for the Surface Glycoprotein and N-Terminal Domain of the B.1 and Delta variants (genomic sequence). (a) Mean entropy of surface glycoprotein for the B.1 (red) and Delta variants (blue) shows the point of emergence of the Delta variant corresponding to January−March 2021. (b) Mean entropy (domain-wise) for the NTD domain of the B.1 (red) and Delta variants (blue). The mean entropy for the NTD domain corresponds to a higher amplitude variation compared to the RBD domain (see Figure S8). (c) MRF capturing the mutational phase transition from the B.1 variant (red) to the Delta variant (blue) with a drastic jump in MRF values at the crossover point of the two variants. (d) MRF for the NTD domain for both variants showing a similar pattern specifically contributing to the transitional shift.

from one variant to another. In the study of phase transitions, the significance of such critical fluctuations that abound in the vicinity of a critical point has often been highlighted. The current investigation also attempts to establish a cause−mutation relationship as we have quantified the randomness for the spike surface glycoprotein of SARS-CoV-2 and we can compare the parent−daughter sequences along the branch of the phylogenetic tree to quantify the occurrence and nonoccurrence of the mutations (see Figure S1 for the process overflow). Moreover, we classify the occurrence and non-occurrence of the mutations as unity and zero, respectively, and explore the domain of classification problems.

RESULTS

Entropy Variation with Time and Mutational Response Function. The entropy pattern observed in the genomic sequence of the virus shows the pattern change with the progression of time (Figure 2). The entropy of the B.1 variant slowly increases over time, especially in the region 22000 to 22500, which corresponds to the N-Terminal Domain (NTD) region of the SARS-CoV-2. This pattern holds true considering the fact that the Delta variant (B.1.617.2) buds off from the B.1 variant. Thus, the entropic amplitude will slowly increase in the B.1 variant over time. On the other hand, the entropic variation of the Delta variant slowly goes down as the variant starts to acclimatize with the host population. A similar pattern is observed in the Delta strain from February 2021 (first detected in India) to April 2021. This pattern of entropic behavior lays out a framework to predict the evolution of a variant and whether the variant has the potential to fall in the category of VOIs or VOCs. The study reveals a significant result of the sudden drop of entropy variation in the NTD region of the B.1 variant. As the Delta variant was first detected in February 2021 in India, we expect to observe an increasing entropic variation in the B.1 variant until the month of January 2021. However, the entropy variation of the B.1 variant suddenly drops in January 2021, suggesting that the Delta variant was already in circulation in the host population from January 2021. This observation becomes more critical as it leaves a characteristic to understand when to expect the possible emergence of a new variant. A similar pattern is observed for the Beta variant (Figure S2), the Gamma variant (Figure S4) and the Kappa variant (Figure S6) suggesting a common transitional behavior throughout different variants.

Mean Entropy and Mutational Response Function (MRF) Distinguishing Mutational States and the Emergence of VOC. At a genomic level, the mean variation and fluctuation of information entropy may indicate the emergence of a new variant. When we measure the mean entropy of the surface glycoprotein of the B.1 variant (Figure 3a, red) it shows a sudden increment after September 2020 until December 2020. The sudden drop of mean entropy of the B.1 variant in January 2021 corresponds to the emergence of the Delta variant as is evident from Figure 3a blue curve. This drastic mean entropic change indeed suggests that the sudden increment of the entropic variation depicts the emergence of a new variant. As soon as a new variant emerges, the entropic fluctuation of the ancestral variant again decreases rapidly. N-Terminal Domain (Figure 3b) shows a higher entropic variation as compared to the Receptor Binding Domain (RBD) (Figure S9a). This phenomenon is important to notice, as the key region of mutation accumulation is the NTD for majority of VOCs (except Omicron; see Figure S13) suggesting the plausibility of the NTD domain showing a mutational fluctuation. Our analysis indicates the importance

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

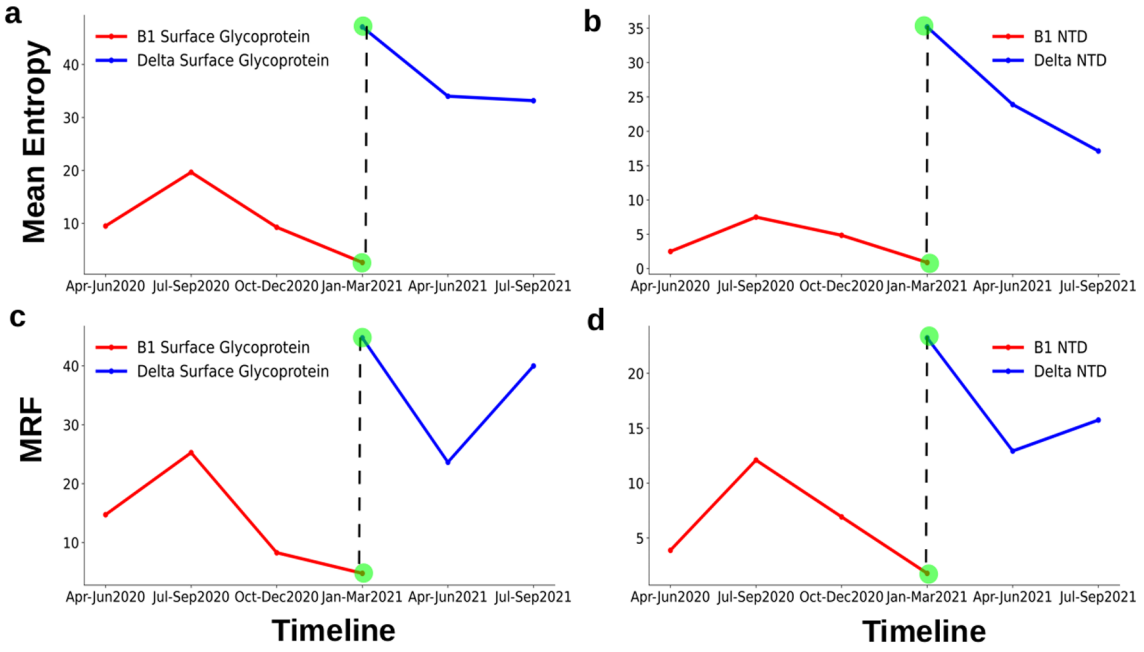


Figure 4. Mean Entropy and Mutational Response Function (MRF) for the Surface Glycoprotein and N-Terminal Domain (NTD) of the B.1 and Delta variants (protein sequence). (a) Mean entropy of B.1 (red) and Delta variant (blue) showing a clear jump when the Delta variant comes into existence. (b) Mean entropy (domain-wise) for the NTD domain of the B.1 (red) and Delta variants (blue). The mean entropy for the NTD domain corresponds to a higher amplitude variation, showing a similar trend obtained for genomic sequence (Figure 2b). Moreover, NTD has a higher entropy compared to the RBD domain (see Figure S10). (c) MRF capturing the mutational phase transition from the B.1 variant (red) to the Delta variant (blue) with a drastic jump in MRF values at the crossover point of the two variants. (d) MRF for the NTD domain for both variants, showing a similar pattern specifically contributing to the transitional shift. The same signature is obtained for the transitional pattern for the B.1 to the Delta variant if one uses either protein or genomic sequence.

of NTD in modulating the change of the viral protein. The localized mutations of NTD, such as recurrent deletion regions (RDRs[33]), might play an essential role in modulating the viral mechanics; for example, deletions at position 69/70 (Δ69/70) might allosterically change the protein conformation.[34]

Again, the entropic fluctuation between variants is efficiently captured by Mutation Response Function (MRF, a statistically relevant response function capturing mutational phase transition under external perturbation, as defined in the method part, eq 7) (Figure 3c). This response function is calculated considering a steady-state condition under the interval of three months period achieved by a constant, small influx of domain mutation. Moreover, MRF also captures the similar response pattern of the NTD (Figure 3d) of both the variants again pointing to the importance of NTD in modulating the fluctuation. The role of NTD in modulating the viral mechanics becomes clearer, as the virus evolves from the B.1 to the Kappa to the Delta variant (Figure S8). Furthermore, a similar signature is obtained when performing the same examination on protein sequences of the B.1 and Delta variants (Figure 4). The mean entropy of the surface glycoprotein sequence (Figure 4a) and NTD (Figure 4b) captures the mutational jump from one variant to other. MRF also captures a similar transition between the B.1 and the Delta variant for their surface glycoprotein (Figure 4c) and NTD (Figure 4d) suggesting the role of NTD in modulating the viral dynamics for the Delta variant. MRF captures a similar trend for the Beta variant (Figure S3) and the Gamma variant (Figure S5). Here, we find the discontinuity in the mutational response function (mean square fluctuation of mutational entropy, Figure 3c,d) because the mean value of the domain entropy shows a discontinuous pattern (similar to the first order phase transition according to the Ehrenfest classification) (Figure 3a,b). It amplifies, in the form of mutational response, when we calculate the mean square fluctuation of that domain entropy. As the evolution transits from a specific variant to VOC, the biological mutations in a particular domain collectively bear that signature of discontinuity. We have observed this signature across several variants (as per the availability of mutational data), including Omicron. Thus, tracking the genomic fluctuation using information entropy and quantifying the mutational response by MRF calculation may provide a way to better understand/predict the emergence of a new variant or a new mutational phase. These outcomes make all the more sense because the Delta variant was first identified in February 2021, but the genomic level response became evident in January 2021 (Figure 3). The lag of 1 month before capturing the variant in the patient samples, we assume, might have given the Delta variant an added advantage to propagate in the host population.

Toward Developing a Statistical Mechanics Guided Machine Learning Model. With the rapid mutations being transpired in the spike glycoprotein of SARS-CoV-2, randomness seems to play a crucial role in engineering the mutations of the virus because randomness insinuates the maximum probability with which a spike surface glycoprotein would occur with a rapidly changing environment. Since both MRF and mean entropy bolster the significance of entropy-based information and capture the evolution of SARS-CoV-2 during mutational phase transition between the B.1 and Delta variants, we rely on such entropic details to build a Machine Learning (ML) model to further predict possible mutational positions in

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

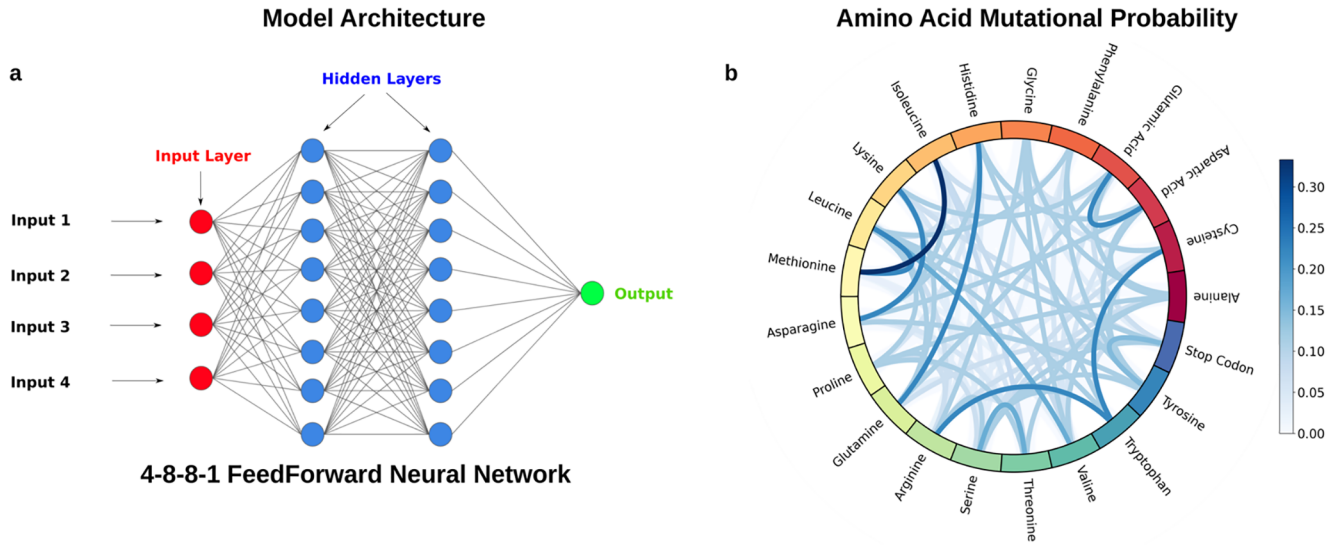


Figure 5. Feedforward model architecture and mutational probability of amino acid residue. (a) Model architecture used for mutation prediction. The model used for the current study has a 4−8−8−1 multilayer perceptron architecture. The input layer has four neurons (red) corresponding to four features, two hidden layers each with eight neurons (blue), and a single output for classification (green). The input layers correspond to the four features with a "tanh" activation function for the first three layers and a sigmoid activation function for the output layer. (b) Circos plot representing the mutational probability of respective amino acids. The mutational probability is derived from the RNA Codon table. The color bar represents the magnitude of the probability of mutations.

the surface glycoprotein sequence. As the current model (Figure 5a) focuses on the primary sequence of the protein, the predictions were structured in mainly two ways: (i) predicting the possible mutational position in the spike surface glycoprotein sequence and (ii) the would-be mutated amino acid residue at the predicted position.

Feature Selection. Feature selection is an essential aspect of developing an ML model. In the current study, four features were taken to quantify the protein sequences for the model to learn. These are as follows: (i) The first feature focused on the pair-predictability of the amino acids in the protein sequence. Pair occurrence of the amino acids is a vital aspect to consider while examining the mutational occurrence as a point mutation can change the number of pairs appearing together (see Supplementary Methods). (ii) The second feature focused on the distribution probability of the amino acids in a protein sequence. The distribution pattern of amino acids is an important aspect to consider because this property quantifies why amino acids are clustered together rather than homogeneously spreading throughout the sequence length. Thus, any change in the amino acid position will change the way the distribution naturally appears (see Supplementary Methods). (iii) The third feature focused on the current and future composition of the amino acids. Understanding the composition of amino acids in the protein sequence and how this composition is affected due to mutations may provide us with the probability of a certain amino acid mutating into another particular one. This quantification is compiled based on the RNA codon table mutational probability (Figure 5b) (see Supplementary Methods). (iv) The fourth feature focused on the entropy of the amino acid residues. As a unique feature, entropy provided us with the mutational pattern in the genomic and protein sequences of SARS-CoV-2 (see Supplementary Methods). Feature assignment of the protein sequence (Table S1) was followed by hyperparameter optimization in order to find the optimal model architecture. After comparing different parameters (Figure S18), and training the model with different activation functions (Figure S19a,b) and learning rates (Figure S19c,d), 4−8−8−1 feedforward neural network architecture was found to be the optimal framework (Figure 5a). Table S1 shows the assignment of each feature to the residues of the surface glycoprotein of delta sequence (PubMed Accession ID: QYJ09734.1). The target is determined by comparing two delta sequences (QYJ09734 and UCL70994).

Mathematically, target assignment can be represented according to eq 1:

$$\text{target} = \begin{cases} 0, \text{ if } aa_i^1 = aa_i^2 \\ 1, \text{ if } aa_i^1 \neq aa_i^2 \end{cases} \quad (1)$$

Here, aa_i^1 corresponds to an amino acid at the ith position of first sequence and aa_i^2 corresponds to the amino acid at the ith position of the second sequence.

An important aspect of model development is determining whether the model can capture the relationship between the input features and the target output. To resolve this, we compared the predicted with the actual mutational position and classified the predicted mutational position as positives, false positives, negatives and false negatives. With this classification, we quantified the model performance in terms of model accuracy, precision, recall, and F1 score which can be described according to eqs 2,3,4, and eq 5:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (2)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (3)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (4)$$

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

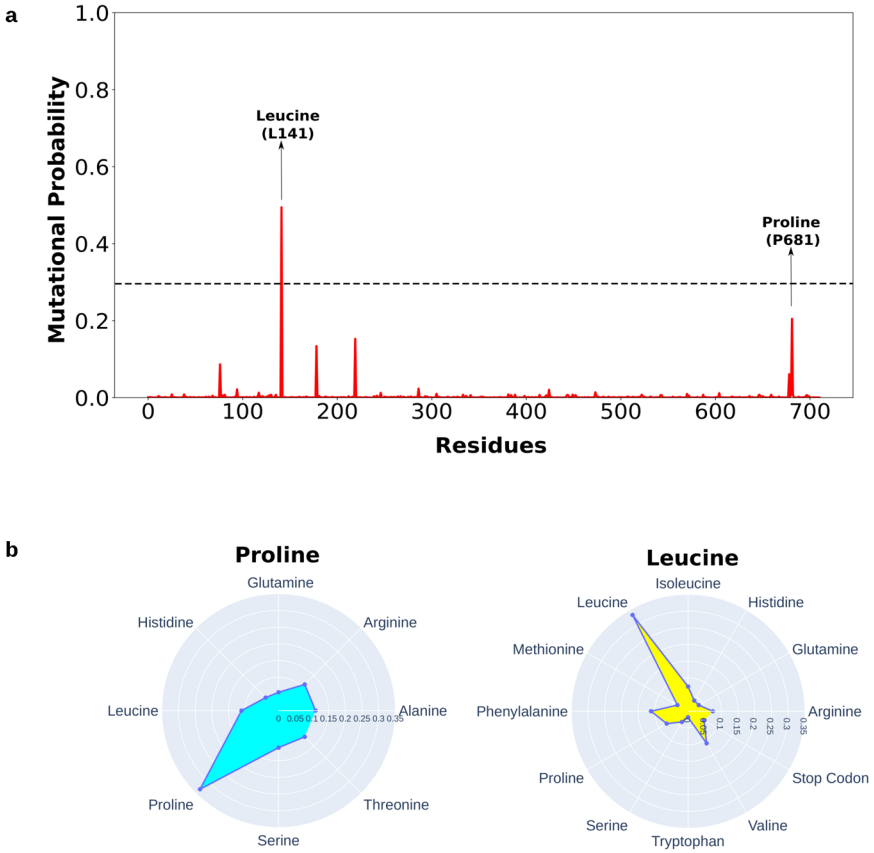


Figure 6. Mutation prediction from 4 to 8−8−1 feedforward model. (a) The model predicts five possible mutational positions in the delta sequence, out of which two major positions are 141 and 681. The cutoff value of 0.3 corresponds to the probability of mutating in future sequences. Our model predicted that two important residues are leucine and proline. (b) Radar plot of the predicted amino acids. The plot represents the probability of mutating the respective amino acids into other amino acids. The color of the radar plot corresponds to the categories to which these residues belong: yellow representing amino acid with nonpolar aliphatic side chain and cyan representing amino acid with polar uncharged side chain.

$$\text{F1 score} = \frac{2 \times \text{recall} \times \text{precision}}{(\text{recall} + \text{precision})} \times 100 \quad (5)$$

Here, TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. Figure 6a shows the mutation prediction in the Delta variant sequence (PubMed Accession ID: UAP60488). The model predicts five possible positions that have the potential to mutate in the upcoming generations of the Delta variant. The dotted line corresponding to a value of 0.3 is the cutoff value for the mutation, suggesting those residues with a mutational probability of more than 0.3 have a higher chance of occurring in the future generation. Residues having a mutational probability less than the cutoff value have a lower chance of occurring in future generations. The would-be mutated residues were confirmed from the RNA codon table. Moreover, the model (labeled as upgraded model, see Figure S20) exhibits an optimistic accuracy of 98.5% (see Figure S22) with a precision score of 99% (see Figure S22). The F1-score of 99.25% (see Figure S22) corresponds to a good balance between precision and recall. We further trained the model with the first three features (labeled as the original model, see Figure S20), excluding the fourth feature (amino acid entropy) to examine model performance differences. The original model exhibited a decrease in the performance, achieving a training loss and accuracy of approximately 48% and 83% (Figure S20, red curve). On the other hand, the upgraded model shows a better model performance (Figure S20, blue curve). The difference in the model performances was further confirmed by calculating the Receiver Operating Curve (ROC) (Figure S21). ROC curve is a graphical plot that represents the indicative capacity of a binary classifier framework as its segregation limit is changed. The upgraded model (Figure S21, red curve) shows a higher Area Under the Curve (AUC) value of 0.996, whereas the original model (Figure S21, black curve) shows an AUC of 0.532. This analysis shows the impact of selecting a relevant model feature in increasing the model accuracy.

We confirm the possible mutational positions and found that leucine (L, position 141) and proline (P, position 681) are predicted as the possible residues that can mutate. The prediction from the model can be supported by the fact that in the Delta variant the mutation at position 681 is an important one causing proline to mutate into arginine (P681R) resulting in an increased fusogenicity (the proficiency of the viral species to fuse with the host membrane) and pathogenicity of the Delta variant.[35]

DISCUSSION

We examine the genomic sequences of different VOCs of SARS-CoV-2 to understand the spread and transition of the variants from one form to another. During such transition, we

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

quantify the mutational response of different variants by calculating the mutational entropy fluctuations introducing a new parameter, MRF, for different surface glycoprotein domains accounting for their genetic and protein sequences. MRF indeed distinguishes the crossover point between any two drastically different mutational states/phases as the viral sequence evolves. In the later part of this study, we also attempt to build a feedforward neural network model including such information entropy as one of the essential features, aiming to predict the possible mutational residues in the protein sequence of SARS-CoV-2. Such an investigation of over 30 000 base pair long genomic sequences of SARS-CoV-2[36] and its variants is important to gain an insight into the mutational pattern and the viral spread. Moreover, SARS-CoV-2 appears to show an exceptionally high recurrence of recombination emerging because of the absence of proofread mechanism and diversity.[37,38] Therefore, genomic level information on different domains of surface glycoprotein also helps to examine the transmission pattern of the virus to understand the spread of the disease and the efficacy of administrated vaccines and drugs.[39,40]

On one hand, as the incorporation of different mutations makes the dormant space of spike genes sequentially more disordered, the emergence of such disorderedness has become a useful monitoring parameter in our ML model to characterize the viral mutational trajectory as the virus evolves from one variant to another. We analyze the spike sequences submitted to GISIAD database (gisaid.org) for different VOCs. After filtering out the ambiguous, redundant, and incomplete sequences, genomic entropy was calculated in a month-wise manner to analyze the transmission dynamics. The mutational entropy (Figure 2) depicts a clear domain-wise trajectory of the B.1 and Delta variants. This trajectory is more pronounced if we track the spread of the B.1 and Delta variants for a longer duration (see Figure S7), suggesting a clear tipping point when a new variant emerges (January 2021 in the case of the Delta variant). A similar pattern is also observed for other VOCs (see Figure S2, S4, and S6), suggesting a typical behavioral pattern of viral transmission.

Apart from characterization of mean genomic entropy, the fluctuation of entropy is well quantified with MRF, which acts as an order parameter capturing the mutational phase-transition point; for instance, a clear depiction of viral entropic fluctuation showing a pronounced difference between the B.1 and the Delta variant (Figure 3).

Interestingly, we find that the entropic fluctuation becomes more clustered toward the NTD as the variant evolves from the B.1 to the Kappa to the Delta variant (see Figure S8). This chaotic behavior of NTD seems true as more mutations accumulate in the NTD (Omicron is an exception, see Figures S13 and S14). Nevertheless, we believe that the decreasing chaotic behavior of RBD can be studied by employing time delay dynamics, as is done in gene networks.[41] However, this unique behavior raises an important question of exploring the potential role of NTDs in viral infectivity. In our early work,[42] we have demonstrated a key interchain interaction at the interface of NTD and RBD, bestowing thermodynamic stability to the viral mechanics. NTD has been classified in terms of Recurring Deletion Regions (RDRs)[33] which play an essential role in allowing the virus to escape the immune system. The highly flexible loop of NTD stabilizes the surface exposed tertiary design[43] of the viral protein. These pieces of evidence indicate an important contribution of NTDs in viral infectivity. It may be a possibility that the occurrence of recurrent deletions in specific regions makes the NTD more compact allowing for better interchain interaction between NTD and RBD and allowing for better viral mechanics leading to viral infectivity. It will be important to see how the molecular level information changes for different variants to gain more insight into the potential of NTD in the viral spread and disease severity. Moreover, with the increasing stability of the later viral strains as compared to early, higher entropic variants,[44] it becomes imperative to dissect the role of different spike domains.

To further explore the question of mutational occurrence, we trained a feedforward neural network to predict possible mutational residues in the surface glycoprotein sequence. The major challenge was to quantify the random event of a mutational occurrence, which was achieved by developing four features namely: Amino Acid Pair Predictability (AAPP), Amino Acid Distribution Probability (AADP), Future Composition of Amino Acids (FCAA), and the Amino Acid Residual Entropy (AARE). The trained model was able to achieve an accuracy of 99% and predicted five possible mutational sites in the protein sequence. The reliability of the model comes from the fact that the model predicts two important mutational sites: P681R (proline to arginine) and L141 (Recurrent Deletion Region[33]). Although we trained the model on protein sequences, the prediction model can also be used to train the genomic sequences to predict disease severity[45] or classify COVID-19.[46]

In summary, this statistical mechanical concept of entropy-guided mutational response function (MRF) followed by neural network modeling appears like a powerful approach to capturing and understanding the mutational signature of SARS-CoV-2. However, from such modeling aspects, there are unchartered areas that remain to be addressed to further fine-tune the model because of the following limitations: (i) SARS-CoV-2 does not have a tremendous amount of historic data. Thus, data availability becomes one of the key factors in making the model more robust. (ii) Here, we have explored a new feature (Amino Acid Entropy) to understand and predict new mutation points for SARS-CoV-2.[33,47] Similarly, other features can also be explored that can serve as an early warning signal to predict not only future mutational position but also a timeline when a new mutation will be expected to evolve.

METHODS

Data Procurement. There is the availability of over 2.9 million genomic sequences of SARS-CoV-2 in the GISAID database (https://www.gisaid.org/). To ensure that genomes with only full sequences and less than 5% N content were included in the current investigation, a preliminary data set was created using the options "complete", "high coverage", "low coverage exclusion", and "collection date complete". The genomic sequence data were then downloaded in a month-wise manner for the variants in the current study. The month-wise data were selected based on the emergence of the variant in the particular country.

Multiple Sequence Alignment (MSA). Following the data collection in fasta format, the MAFFT (https://mafft.cbrc.jp/alignment/server/)[48,49] server was used for the execution of multiple sequence alignment. For every month, we obtained an MSA file that was used for further calculation.

Entropy Calculation. In sequence analysis, entropy refers to the measure of character (column) variance across various

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

sequences. Shannon's Entropy,[50] Schneider's Entropy,[51] Shenkin's Entropy,[52] Gerstein's Entropy,[53] and Gap Normalized Entropy[54] are some of the entropy formulas that may be used to forecast the entropy plot by multiple sequence alignment. There are mainly two steps involved in the entropy calculation of an MSA: (i) Performing multiple sequence alignment of the sequences and (ii) calculation of entropy for each column in order to get a consensus entropic value.

For the current study we employed the Shannon entropy to calculate residue wise entropic fluctuation. The entropy of a random variable X with potential outcomes X_1, X_2, ..., X_n that appears with a probability $P(x_1)$, $P(x_2)$, ..., $P(x_n)$ can be defined as

$$H(x) = -\sum p(x_i)\cdot\log(p(x_i)) \tag{6}$$

In the case of MSA of genomic sequence of variants, x = A, T, G, and C, whereas, in the case of protein sequence, x ∈ {S}, where S is a set consisting of the 20 amino acids. The individual probability of respective bases is calculated column-wise resulting in a consensus entropic value for each position of the genomic/protein sequence. The resultant values were then plotted using custom python code, with x-axis corresponding to nucleotide position and the y-axis corresponding to the entropy value. For instance, in the case of the B.1 variant for November 2020, we take all the sequences that are gathered in that month, run them through a multiple sequence alignment procedure, and then use the aligned sequences from that month to determine the probability of a certain site. Now, we did not always discover the same amount of sequence data deposited for each month. In order to calculate probabilities, the data set is standardized. To study how the pattern evolves as time passes (from November 2020 to December 2020 to January 2021), we do this for each month, as illustrated in Figure 2.

Mutational Response Function. To capture the mutational entropy fluctuation in SARS-CoV-2 variant, we lay out a new mathematical framework by introducing a new parameter, Mutational Response Function (MRF). It aptly quantifies the mutational entropy fluctuation of every ith residue as the sequence evolves over a given time. Mathematically, MRF for each ith residue can be defined as,

$$\mathrm{MRF}(i) = \frac{\langle \Delta S(i)^2 \rangle}{\overline{S}(i)} = \frac{\frac{1}{T}\sum_{t=1}^{T}(S(i) - \overline{S})^2}{\overline{S}} \tag{7}$$

where, $S(i)$ is the entropy of the ith nucleotide, $\overline{S}$ is the time-averaged entropy of that ith nucleotide for 3 months ($T = 3$). From here, we have collectively monitored domain-wise MRF by adding this residual MRF involved in any concerned domain (e.g., RBD, NTD, etc.) as shown in Figure 3.

Data Procurement and Preprocessing for Model Development. Delta variant (B.1.617.2) sequences for Neural Network Model development were downloaded from the National Center for Biotechnology Information (NCBI) Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2,[55] last accessed on October 4, 2021) database. The protein sequences were downloaded from two countries: India, because of the origin of the delta variant, and USA for the high number of sequence submissions. The initial preprocessing was done by selecting three filters to download the optimal sequences. The "Pango lineage" filter was set to "B.1.617.2" to obtain only the Delta variant sequences. The "Proteins" filter was set to "Surface Glycoprotein" to obtain only the spike protein sequences. The "Nucleotide Completeness" filter was set to "Complete" to obtain only the complete sequences. After applying these filters, the data sets from both countries were combined to obtain a total of 2436 sequences.

Once the initial filtered data set was downloaded, it was preprocessed to remove similar sequences. As a majority of the substitution mutations in different SARS-CoV-2 variants occur in the S1 domain, only the S1 domain sequence was kept, resulting in the final sequence length of 711 amino acids. As the current model does not account for deletion/insertion mutation prediction, the data set was again preprocessed to keep only those sequences that have deletion mutations to maintain the sequence length throughout the data set. After the final preprocessing phase, the data set contained 117 sequences.

Model Architecture. We have used the feedforward backpropagation neural network with a model architecture of 4−8−8−1 (Figure 5a). The first layer of the model corresponds to four input features, followed by two hidden layers consisting of eight neurons each and the last layer consists of one neuron which corresponds to the output (target). The activation functions were chosen to be hyperbolic-tangent for the first three layers and sigmoid for the output layer, which corresponds to an output between zero and one (i.e., probability of mutation).

Target Selection for Binary Classification. Phylogenetic analysis was done to understand the evolutionary process of surface glycoproteins. Going through the same branch of phylogenetic tree, we compare the surface glycoprotein sequences and the difference between them indicates the occurrence of mutation which is tagged as "Unity (1)" and no difference indicates nonoccurrence of mutation and is tagged as "Zero (0)".

Predicting the Mutated Amino Acid. Mutation is a random process that is sensitive to different parameters. It becomes very difficult to predict which amino acid a particular residue would mutate to. Thus, we rely again on the RNA codon mutation probability to find the possible residues that can result if a mutation occurs in the predicted position and their respective probabilities of mutation. The higher the probability, the more chances are that the residue at the predicted position would mutate into.

■ ASSOCIATED CONTENT

Data Availability Statement

All data and codes used in the analysis are available from the corresponding author to any researcher for reproducing or extending the analysis under a material transfer agreement with IISER-Kolkata, India.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c04574.

Schematic representation of the workflow, month-wise entropic variation of the B.1 and Beta (B.1.351) variants in South Africa, mutational response function of the B.1 and Beta variants in South Africa, month-wise entropic variation of the B.1.1.28 and Gamma (P.1) variants in Brazil, mutational response function of B.1.1.28 and the Gamma variants in Brazil, month-wise entropic variation of the Kappa (B.1.617.1) variant in India, mutational response function of the B.1, Kappa (B.1.617.1), and Delta (B.1.617.2) variants in India, month-wise entropic

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX

variation of the B.1 and Delta variants (protein sequence), month-wise entropic variation of the B.1.1 and BA.1 variants (genomic sequence), month-wise entropic variation of the B.1.1 and BA.1 variants (protein sequence), mutational response function of the B.1.1 and BA.1 variants (genomic sequence), Mutational Response Function of the B.1.1 and BA.1 variants (protein sequence), feature selection for model training, comparison of training loss and accuracy between original and upgraded models, 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 the original and upgraded models, model metrics showing accuracy, precision, recall and F1-score of the upgraded model, numeric representation of the sequence, and references (PDF)

AUTHOR INFORMATION

Corresponding Author

Susmita Roy − *Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, Kolkata, West Bengal 741246, India;* orcid.org/0000-0001-6411-4347; Email: susmita.roy@iiserkol.ac.in

Authors

Satyam Sangeet − *Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, Kolkata, West Bengal 741246, India*

Raju Sarkar − *Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, Kolkata, West Bengal 741246, India*

Saswat K. Mohanty − *Department of Chemical Sciences, Indian Institute of Science Education and Research Kolkata, Kolkata, West Bengal 741246, India;* orcid.org/0000-0002-1813-589X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.2c04574

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.S., R.S., S.K.M., and S.R. thank the DIRAC IISER-Kolkata for computational support. S.R. acknowledges support from the Department of Biotechnology (DBT) (Grant No. BT/12/IYBA/2019/12) and the Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India (Grant No. SRG/2020/001295).

REFERENCES

(1) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A novel coronavirus from patients with pneumonia in China. *N. Eng. J. Med.* **2020**, *382* (8), 727−733.

(2) Faria, N. R.; Mellan, T. A.; Whittaker, C.; Claro, I. M.; Candido, D. S.; Mishra, S.; Crispim, M. A. E.; Sales, F. C. S.; Hawryluk, I.; McCrone, J. T.; et al. Genomics and Epidemiology of P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **2021**, *372*, 815−821.

(3) Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E. J.; Msomi, N.; et al. Detection of a SARS-CoV-2 Variant of Concern in South Africa. *Nature* **2021**, *592* (7854), 438−443.

(4) Tang, J. W.; Tambyah, P. A.; Hui, D. S. Emergence of new SARS-CoV-2 variant in the UK. *J. Infect* **2021**, *82* (4), e27−e28.

(5) Callen, H. B.*Thermodynamics and an Introduction to the Physical Theories of Equilibrium Thermostatics and Irreversible Thermodynamics*; Wiley: New York, London, and Sydney, 1960.

(6) Einstein, A. Theorie der Opaleszenz von homogenen Flussigkeiten und Flussigkeitsgemischen in der Nahe des kritischen Zustandes. *Ann. Phys.* **1910**, *338* (16), 1275−1298.

(7) Landau, L. D.; Lifshitz, E. M.*Statistical Physics, Part 1*, 3rd ed.; Pergamon Press: 1963.

(8) March, N. H.; Tosi, M. P.*Introduction to Liquid State Physics*; World Scientific: River Edge, NJ, 2002.

(9) Hill, T. L.*Introduction to statistical thermodynamics*; Addison-Wesley: Burlington, MA, 1960.

(10) Morse, P. M.*Thermal physics*; Benjamin, New York: 1969.

(11) Katebi, A.; Kohar, V.; Lu, M. Random Parametric Perturbations of Gene Regulatory Circuit Uncover State Transitions in Cell Cycle. *iScience* **2020**, *23* (6), 101150.

(12) Scheffer, M.*Critical transitions in nature and society*; Princeton University Press: Princeton, NJ, 2009.

(13) Wu, G.; Yan, S. Prediction of Mutations Engineered by Randomness in H5N1 Hemagglutinins of Influenza a Virus. *Amino Acids* **2008**, *35* (2), 365−373.

(14) Everitt, B. S.*Chance rules: an informal guide to probability, risk, and statistics*; Springer, New York, 1999.

(15) Fitch, W. M.; Bush, R. M.; Bender, C. A.; Cox, N. J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 7712−7718.

(16) Lucia, U.; Deisboeck, T. S.; Grisolia, G. Entropy-Based Pandemics Forecasting. *Frontiers in Physics* **2020**, *8*, 00274.

(17) Santos, F. L. S. G.; Paiva Júnior, S. de S. L.; Freitas, A. C. de; Balbino, V. de Q.; Batista, M. V. de A. EntroPhylo: An Entropy-Based Tool to Select Phylogenetic Informative Regions and Primer Design. *Infection, Genetics and Evolution* **2021**, *92*, 104857.

(18) Vopson, M. M.; Robson, S. C. A New Method to Study Genome Mutations Using the Information Entropy. *Physica A: Statistical Mechanics and its Applications* **2021**, *584*, 126383.

(19) Mishin, Y. Thermodynamic theory of equilibrium fluctuations. *Annals of Physics* **2015**, *363*, 48−97.

(20) Davies, N. G.; Abbott, S.; Barnard, R. C.; Jarvis, C. I.; Kucharski, A. J.; Munday, J. D.; Pearson, C. A. B.; Russell, T. W.; Tully, D. C.; Washburne, A. D. Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England. *Science* **2021**, *372* (6538), 3055.

(21) Ostrov, D. A. Structural Consequences of Variation in SARS-CoV-2 B.1.1.7. *Journal of Cellular Immunology* **2021**, *3* (2), 103−108.

(22) Karim, S. S. A.; Karim, Q. A. Omicron SARS-CoV-2 Variant: A New Chapter in the COVID-19 Pandemic. *Lancet* **2021**, *398* (10317), 2126−2128.

(23) Ferguson, N.; Anderson, R. Predicting evolutionary change in influenza A virus. *Nat. Med.* **2002**, *8*, 562−563.

(24) Smith, D. J.; Lapedes, A. S.; De Jong, J. C.; Bestebroer, T. M.; Rimmelzwaan, G. F.; Osterhaus, A. D. M. E.; Fouchier, R. A. M. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **2004**, *305* (5682), 371−376.

(25) Mohamed, T.; Sayed, S.; Salah, A.; Houssein, E. H. Long Short-Term Memory Neural Networks for RNA Viruses Mutations Prediction. *Mathematical Problems in Engineering* **2021**, *2021*, 1.

(26) Salama, M. A.; Hassanien, A. E.; Mostafa, A. The Prediction of Virus Mutation Using Neural Networks and Rough Set Techniques. *Eurasip Journal on Bioinformatics and Systems Biology* **2016**, *2016* (1), 1−11.

(27) Cilia, E.; Teso, S.; Ammendola, S.; Lenaerts, T.; Passerini, A. Predicting Virus Mutations through Statistical Relational Learning. *BMC Bioinform.* **2014**, *15*, 309.

(28) Coudray, N.; Ocampo, P. S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A. L.; Razavian, N.; Tsirigos, A.

Classification and Mutation Prediction from Non−Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nature Medicine* **2018**, *24* (10), 1559−1567.
(29) Xie, W. J.; Qi, Y.; Zhang, B. Characterizing Chromatin Folding Coordinate and Landscape with Deep Learning. *PLoS Computational Biology* **2020**, *16* (9), e1008262.
(30) Achuthan, S.; Chang, M.; Shah, A.SPIRIT-ML: A Machine Learning Platform for Deriving Knowledge from Biomedical Datasets. In *Data Integration in the Life Sciences. DILS*; Ashish, N., Ambite, J. L., Eds.; Lecture Notes in Computer Science 9162; Springer: Cham, Switzerland, 2015; DOI: 10.1007/978-3-319-21843-4_19.
(31) Wang, J.; Olsson, S.; Wehmeyer, C.; Perez, A.; Charron, N. E.; de Fabritiis, G.; Noe, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5* (5), 755−767.
(32) Yan, S.; Wu, G. Application of Neural Network to Predict Mutations in Proteins from Influenza A Viruses - A Review of Our Approaches with Implication for Predicting Mutations in Coronaviruses. *Journal of Physics: Conference Series* **2020**, *1682*, 012019.
(33) Mccarthy, K. R.; Rennick, L. J.; Nambulli, S.; Robinson-Mccarthy, L. R.; Bain, W. G.; Haidar, G.; Duprex, W. P. Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape. *Science* **2021**, *371* (6534), 1139−1142.
(34) Xie, X.; Liu, Y.; Liu, J.; Zhang, X.; Zou, J.; Fontes-Garfias, C. R.; Xia, H.; Swanson, K. A.; Cutler, M.; Cooper, D.; et al. Neutralization of SARS-CoV-2 Spike 69/70 Deletion, E484K and N501Y Variants by BNT162b2 Vaccine-Elicited Sera. *Nature Medicine* **2021**, *27* (4), 620−621.
(35) Saito, A.; Irie, T.; Suzuki, R.; Maemura, T.; Nasser, H.; Uriu, K.; Kosugi, Y.; Shirakawa, K.; Sadamasu, K.; Kimura, I.; et al. Enhanced Fusogenicity and Pathogenicity of SARS-CoV-2 Delta P681R Mutation. *Nature* **2022**, *602* (7896), 300−306.
(36) Shishir, T. A.; Naser, I. B.; Faruque, S. M. In Silico Comparative Genomics of SARS-CoV-2 to Determine the Source and Diversity of the Pathogen in Bangladesh. *PLoS One* **2021**, *16* (1), e0245584.
(37) Mandal, S.; Roychowdhury, T.; Bhattacharya, A. Pattern of Genomic Variation in SARS-CoV-2 (COVID-19) Suggests Restricted Nonrandom Changes: Analysis Using Shewhart Control Charts. *Journal of Biosciences* **2021**, *46* (1), 11.
(38) Rouchka, E. C.; Chariker, J. H.; Chung, D. Variant Analysis of 1,040 SARS-CoV-2 Genomes. *PLoS One* **2020**, *15* (11), e0241535.
(39) Srivastava, S.; Banu, S.; Singh, P.; Sowpati, D. T.; Mishra, R. K. SARS-CoV-2 Genomics: An Indian Perspective on Sequencing Viral Variants. *Journal of Biosciences* **2021**, *46* (1), 22.
(40) Li, Q.; Wu, J.; Nie, J.; Zhang, L.; Hao, H.; Liu, S.; Zhao, C.; Zhang, Q.; Liu, H.; Nie, L.; et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **2020**, *182* (5), 1284−1294.
(41) Suzuki, Y.; Lu, M.; Ben-Jacob, E.; Onuchic, J. N. Periodic, Quasi-Periodic and Chaotic Dynamics in Simple Gene Elements with Time Delays. *Sci. Rep.* **2016**, *6*, 21037.
(42) Roy, S.; Jaiswar, A.; Sarkar, R. Dynamic Asymmetry Exposes 2019-NCoV Prefusion Spike. *J. Phys. Chem. Lett.* **2020**, *11* (17), 7021−7027.
(43) Klinakis, A.; Cournia, Z.; Rampias, T. N-Terminal Domain Mutations of the Spike Protein Are Structurally Implicated in Epitope Recognition in Emerging SARS-CoV-2 Strains. *Computational and Structural Biotechnology Journal* **2021**, *19*, 5556−5567.
(44) Ghanchi, N. K.; Nasir, A.; Masood, K. I.; Abidi, S. H.; Mahmood, S. F.; Kanji, A.; Razzak, S.; Khan, W.; Shahid, S.; Yameen, M.; et al. Higher Entropy Observed in SARS-CoV-2 Genomes from the First COVID-19 Wave in Pakistan. *PLoS One* **2021**, *16* (8), e0256451.
(45) Wang, R. Y.; Guo, T. Q.; Li, L. G.; Jiao, J. Y.; Wang, L. Y. Predictions of COVID-19 Infection Severity Based on Co-Associations between the SNPs of Co-Morbid Diseases and COVID-19 through Machine Learning of Genetic Data. *2020 IEEE 8th Int. Conf. Comput. Sci. Network Technol.* **2020**, 92−96.
(46) Arslan, H. Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. *Proc. 7th Int. Management Inform. Syst. Conf.* **2021**, *74*, 20.
(47) Plante, J. A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B. A.; Lokugamage, K. G.; Zhang, X.; Muruato, A. E.; Zou, J.; Fontes-Garfias, C. R.; et al. Spike Mutation D614G Alters SARS-CoV-2 Fitness. *Nature* **2021**, *592* (7852), 116−121.
(48) Katoh, K.; Misawa, K.; Kuma, K.-I.; Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30* (14), 3059−3066.
(49) Kuraku, S.; Zmasek, C. M.; Nishimura, O.; Katoh, K. ALeaves Facilitates On-Demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity. *Nucleic acids research* **2013**, *41* (W1), W22.
(50) Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **1948**, *27* (3), 379−423.
(51) Sander, C.; Schneider, R. Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins* **1991**, *9* (1), 56−68.
(52) Shenkin, P. S.; Erman, B.; Mastrandrea, L. D. Information-Theoretical Entropy as a Measure of Sequence Variability. *Proteins* **1991**, *11*, 297−313.
(53) Gerstein, M.; Altman, R. B. Average Core Structures and Variability Measures for Protein Families: Application to the Immunoglobulins. *J. Mol. Biol.* **1995**, *251*, 161−175.
(54) Ramazzotti, M.; Degl'Innocenti, D.; Manao, G.; Ramponi, G. Entropy Calculator: getting the best from your multiple protein alignments. *Ital. J. Biochem* **2004**, *53* (1), 16−22.
(55) Hatcher, E. L.; Zhdanov, S. A.; Bao, Y.; Blinkova, O.; Nawrocki, E. P.; Ostapchuck, Y.; Schaffer, A. A.; Brister, J. R. Virus Variation Resource-Improved Response to Emergent Viral Outbreaks. *Nucleic Acids Res.* **2017**, *45* (D1), D482−D490.

https://doi.org/10.1021/acs.jpcb.2c04574
J. Phys. Chem. B XXXX, XXX, XXX−XXX