Genomic Surveillance of COVID-19 Variants with Language Models and Machine Learning

Sargun Nagpal[1¶], Ridam Pal[1¶], Ashima[1&], Ananya Tyagi[1&], Sadhana Tripathi[1&], Aditya Nagori[1], Saad Ahmad[1], Hara Prasad Mishra[1], Rintu Kutum[1], Tavpritesh Sethi[1,2*]

1. Indraprastha Institute of Information Technology Delhi, India
2. All India Institute of Medical Sciences, New Delhi, India

*tavpriteshsethi@iiitd.ac.in

[¶]Contributed Equally, [&]Contributed Equally

Abstract.

The global efforts to control COVID-19 are threatened by the rapid emergence of novel variants that may display undesirable characteristics such as immune escape or increased pathogenicity. The current approaches to genomic surveillance do not allow early prediction of emerging variations. Here, we derive Dimensions of Concern (DoC) in the latent space of SARS-CoV-2 mutations and demonstrate their potential to provide a lead time for predicting the increase of new cases in 9 countries across the globe. We learned unsupervised word embeddings from 3,09,060 spike protein coding sequences deposited on GISAID database until April, 2021. We discovered that “blips" in the latent dimensions of embeddings are associated with mutations. We modeled the temporal occurrence of blips and their relationships with the number of new cases in the following months for these countries. Certain dimensions demonstrated a consistent leading relationship between the occurrence of blips and the number of new cases in the following months, thus labeled as potential Dimensions of Concern, DoCs. We validated the predictive importance of DoCs by performing Random Forest-based feature selection and modeling in a temporally split training, validation, testing regime. Twelve dimensions achieved statistical significance and achieved an R-squared of 37% for prediction of number of new cases in the following month. Biological exploration of DoCs revealed that dimensions 3 and 12 captures 3-mers CGG, ACG and CAC that are associated with known variants L452R, K417T and Q677H respectively. Learning and tracking DoCs is extensible to related challenges such as pandemic preparedness, immune escape, pathogenicity modeling and antimicrobial resistance.

Introduction.

COVID-19 is reported to have claimed 3.46 million lives as of May 24, 2021 [1]. A large number of these deaths are attributed to unexpected surges in infections caused by new strains of SARS-CoV-2, prompting health organizations such as CDC to declare these as variants of concern

[2]. B.1.1.7 lineage of the UK origin is circulating in 118 countries and has now been established as the dominant strain in the USA [3], Ireland, Cambodia and is widely spread in Germany, Italy, and Sweden [4]. The B.1.617 lineage is also spreading rapidly in India, UK, Netherland, Australia, Italy, and New Zealand [4]. Such emergence of new variants can seriously undermine the efficacy of global vaccination efforts or cause multiple reinfection waves as new strains escape antibodies. This underscores the need for early detection and prediction of emerging variants. In addition to increased transmissibility, such variants may acquire an increased pathogenic potential as evidenced by the rapid clinical progression and overwhelmed health systems in the devastating second wave of pandemic in India. Mitigating such events in future will require stronger surveillance systems in place.

Several initiatives have focused on providing high quality tracking information for the strains and lineages as these emerge [5]. However, early prediction of emerging variants through genomic signals remains an open challenge. Although domain-based, expert-reasoning approaches are the mainstay of our understanding, these often yield retrospective understanding rather than proactive predictions. On the other hand, black-box machine learning approaches are likely to be biased by underlying data characteristics and do not explain the biological basis of predictions. Here we propose a black-box deep learning based language model along with empirical experiments to demonstrate explainable genomic signals for tracking and predicting the spread of COVID-19 across countries.

Our approach is rooted in language models that have recently shown promise for capturing biological insights from DNA sequences [6]. Word embeddings represent the latent space of a corpus of text [7]. Reading the genome to learn word embeddings for advancing biological understanding is a relatively recent area of research. Standard statistical approaches such as principal components analysis use some form of matrix factorization to capture the latent space of data. However, these are limited by the requirement of linear factorization and are not inherently suited for sequential data that can be represented as words. Word embeddings on the other hand, can be trained with textual inputs and can learn highly nonlinear and contextual relationships. Codons with their tri-nucleotide translations represent a natural basis for word representations and have been utilized in the past for learning embedding models for modeling various outcomes such as mutation susceptibility [8] and gene sequence correlations [9]. More recently, Hie et. al [10] used machine learning along with word embedding techniques for predicting the mutations which might lead to viral escape. The semantics and grammar of the amino acids corresponding to antigenic change and viral fitness was learnt for predicting the viral escape. In this paper, we focus on the semantics of viral DNA sequences to derive *Dimensions of Concern (DoCs)* and demonstrate their causal potential for increasing epidemiological spread patterns across 9 countries. Our approach is extensible to global threats in pathogen surveillance such as emerging infections, pandemics and antibiotic resistance.

Methods.

Datasets.

The dataset was downloaded from GISAID EpiCoV [11] (April 8, 2021 release). 6,34,669 genome sequences with high nucleotide completeness, coverage, complete temporal information and presence of less than 5% non-identified nucleotide bases (N) were downloaded. The sequences included 63 countries including India, United Kingdom, USA, Australia, New Zealand, Germany, Russia, Italy, France, Mexico, Canada, China, Japan, Pakistan, Bangladesh, Iran, Iraq, the continent of South America and Africa. Duplicate samples were removed and whole genome sequences were parsed using CoV-Seq [12] to extract nucleotide sequences corresponding to each of the 12 CDS. Accession ids that did not cover 12 coding regions were discarded yielding 3,09,060 genome sequences. Spike region sequences were extracted for further modeling (Supplementary figure 1). Countrywise COVID-19 data for new cases was taken from a publicly available repository maintained by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

Unsupervised Word Embeddings.

A low dimensional representation for the genome sequences, was learned using the word2vec model. The sequence of 3-mers was considered as words for training the word2vec model [7], implemented in Gensim [13]. The word2vec model was trained using the skip gram algorithm, with fixed window size of twenty and vector size of thirty six. For generating a consensus embedding for a particular strain, genomic sequences were represented by taking the mean of each 3-mer occurring in the sequence dimension wise. The mean was calculated by summing across all the k-mers over each dimension and then dividing it by the total number of 3-mers present in the sequence. The document embeddings of the extracted genome sequences were visualized in the Tensorboard projector [14] for understanding the clustering of variants across time.

For selecting the vector size of the word2vec model, we calculated PIP (Pairwise Inner Product) loss [15]. PIP loss is a metric used for calculating the dissimilarity between two word embeddings. For the embedding matrix of strains (E), PIP matrix is defined as the dot product the embedding matrix to its transpose ($E.E^{T}$). The PIP loss between two embeddings is defined as the norm of difference between their PIP matrices.

$$||PIP(E1) - PIP(E2)|| = \left\|E_1E_1^T - E_2E_2^T\right\| = \sqrt{\sum_{i,j}((v_i^{(1)}, v_j^{(1)}) - (v_i^{(2)}, v_j^{(2)}))^2}$$

Various word2vec models were trained on the dataset with different vector sizes varying in multiples of three. Based on the PIP loss calculations, we found out that word embeddings with 36 dimensions showed a differential dent in the curve (change in straight line) due to which we selected this to be the dimension of word embeddings (Supplement figure 2).

Hierarchical Clustering using Cosine Distances.

Hierarchical clustering was performed on the word2vec embeddings generated using the pairwise cosine distances. Hierarchical clustering was done with "complete method", taking other parameters as default, implemented using the R package 'hclust' [16]. Since the total number of strains were large, we sampled 25 random strains from each of the 16 countries and calculated the pairwise cosine distances among them to generate a hierarchical clustering dendrogram. Phylogenetic trees of 400 random spike proteins sequences were plotted for all countries present in the dataset.

Latent Space Dimensions.

Each dimension of the mean embeddings for a country was analysed for the presence of temporal anomalies. Countries having a minimum of 20 samples in any given month were selected and the same number of records (minimum samples in any given month for that country) were sampled without replacement from each month. These records were used to define control limits of ±1 standard deviation from the mean value for each dimension, and all values in the full dataset outside those limits were categorized as 'Blip' points. Blip counts in each month were normalized by calculating the number of blips per sample collected in a month for each dimension (normalized blips). The embedding dimensions were then compared in terms of the total normalized blips for each country to observe the significant dimensions and dis(similarity) in trends among different countries. Cumulative counts of normalized blips were analyzed to understand the temporal accumulation of blips in each dimension.

Cross correlations analysis.

Cross correlation between monthly blip dimensions and new cases was performed. Since both the caseload and blip time series exhibited exponential increase over time, these were transformed on a log scale. Successive differences of the data points were taken in order to make the time series stationary, a necessary assumption for the cross-correlation tests. Finally, on the log transformed and differenced series, cross-correlation analysis between blip dimensions at time t+h and new cases at t was performed, where h = 0, ±1, ±2, ±3...±7 was taken. Line plots and heatmaps were plotted between the processed data to explore the relationships further.

Dimensions of Concern (DoCs).

Dimensions of Concern(DoCs) were defined as those dimensions which were statistically significant, leading the case count in cross-correlation analysis and predictive of new case count in the respective countries. Country wise monthly total new cases data was taken at the end of each month. Total new cases data for each month was merged to monthly blip dimensions data from March, 2020 to March, 2021. We used Boruta [17], a wrapper algorithm around randomforest to select the important blip dimension for the prediction of subsequent 1 month's new cases. R software version 3.6 and Boruta package was used to fit the importance model. Features versus z-score based importance measure was plotted to visualize DoCs.

Model development & evaluation.
A random forest model using 1000 trees was trained using next month's new cases as outcome variable and *DoCs* as predictors. Temporal splitting was performed to validate models on future data. Training was performed upon data from March, 2020 to December, 2021 and the fitted model was tested upon data from January, 2021 to March, 2021.

Biological significance of DoCs.
Top ten 3-mers having the highest absolute weights (contribution) were identified corresponding to each of the 36 dimensions in order to map the latent dimensions of the sequence embeddings to specific mutations. Literature based screening of SARS-CoV-2 lineages and associated variants linked to the spike protein was conducted [18][19] and a list of codons associated with VOIs and VOCs was curated [20][21][22]. Supplementary figure 4 shows the intersection of the k-mers extracted using the data-driven approach and manually curated list.

Results.

Latent space preserves spatiotemporal information of COVID-19 spread. The plot in Figure 1(a) is for a randomly selected subsample of the strains from the UK, having collection dates from March, 2020 - March, 2021. We observed 2 major clusters in the strains. The first cluster contains strains having collection date majorly from the range October, 2020 to December, 2020 (colored in Blue). The other cluster contains strains majorly from January, 2021 to March, 2021. The two different clusters for 2020 strains and 2021 strains may be indicative of the mutation of the strains. The dendrogram in Figure 1(b) shows hierarchical clustering based upon the latent space embeddings of nucleotide sequence of the spike protein. Clusters of strains belonging majorly to a particular country are highlighted in various colours. It is observed that some of the Chinese, Australian, Mexican and England strains form tight clusters (marked in green, purple and Magenta). Strains from Italy, France, Brazil, Japan, Canada, USA, Scotland and India are dispersed with strains from different countries and are not very tightly bound in large clusters. This behaviour is expected because the strains from different countries are selected having various different collection dates. The t-SNE plot (Figure 1(c)) suggests that there have been clusters of nucleotide sequences of spike proteins specific to a geographic location based on embedding features. Separate clusters of nucleotide sequences for some countries; Mexico, Australia and China have been significantly captured by latent space of sequence embeddings.

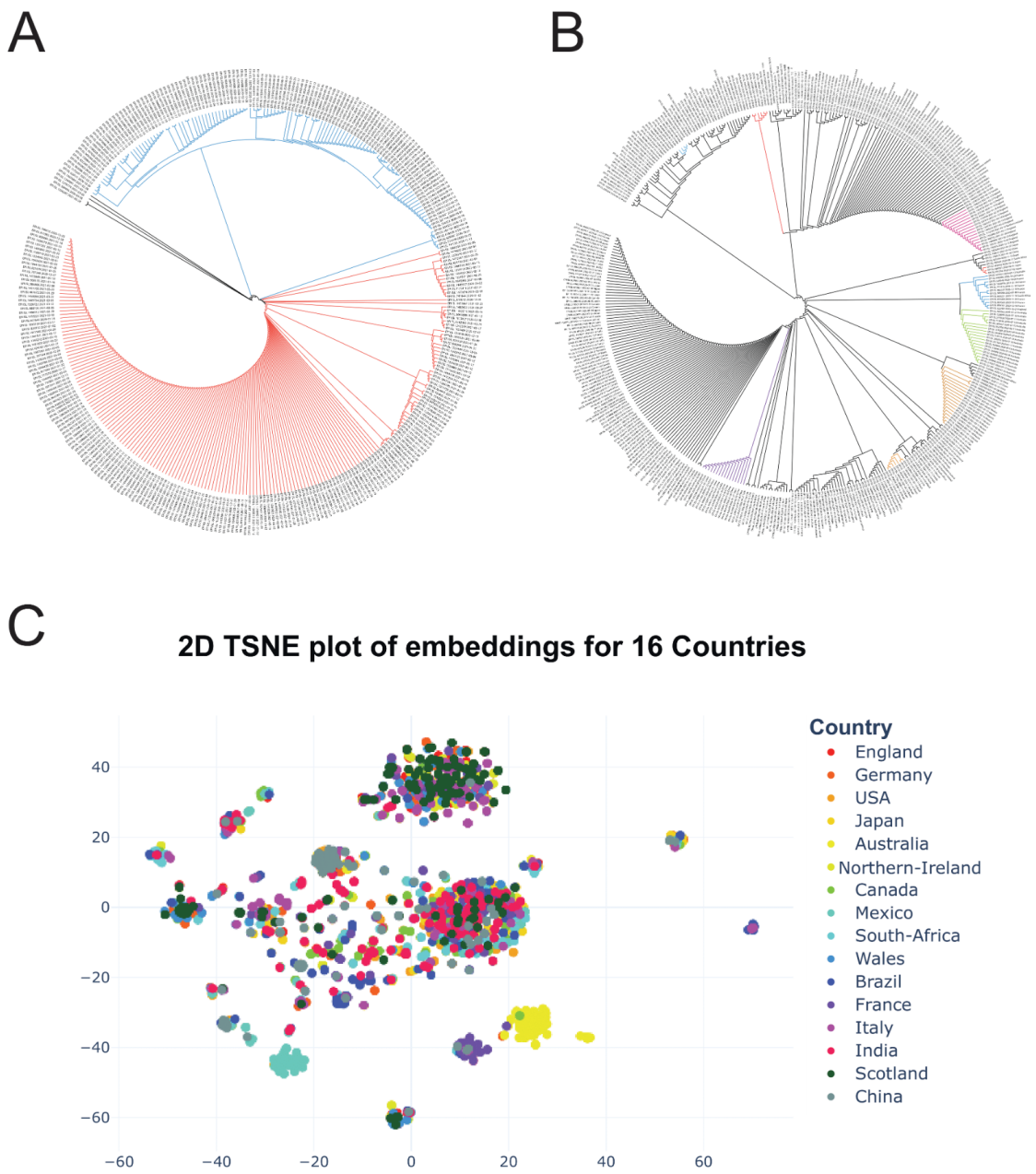

Figure 1: *Phylogenetic tree constructed using cosine similarities between sequence embeddings. The model captured temporal emergence of strains and geographic information in country-specific manner, e.g. United Kingdom (A) and globally in 16 countries(B) respectively. 400 randomly sampled sequences from the complete data have been used to construct the phylogenetic tree. (C) shows a t-SNE plot for clusters formed on the basis of embedding features. 200 randomly sampled sequences from each country have been used to construct the t-SNE plot.*

Latent space dimensions capture mutations in spike protein. The 36-dimensional mean embedding vectors for the countries showed sudden changes ("blips") in values. We hypothesized that these blips were caused by changes in the spike protein of the genome sequence. This hypothesis was confirmed in simulation experiments in synthetic datasets (Supplementary Information 1). Similar blip patterns were observed for geospatial regions such as England and Scotland within the United Kingdom (Figure 2(a)) whereas other countries such as Brazil had different patterns (Figure 2(b)).

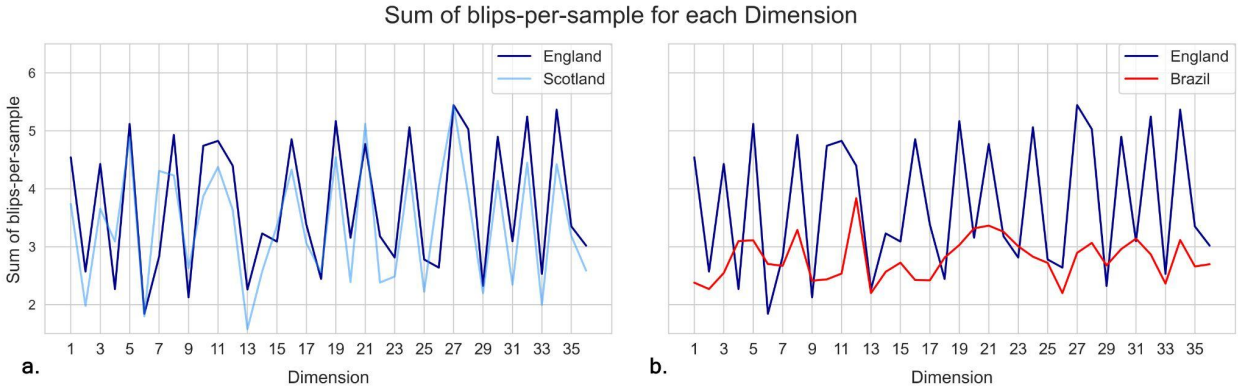

Figure 2: Sum of blips per sample in each dimension for different countries. *Countries in UK (a) - England and Scotland show similar distribution of blip counts per sample across dimensions, whereas when (b) England was compared with Brazil, it didn't show any similarity in distribution.*

Deciphering the relationship between blip dimensions and new cases.

First, cross-correlation between blip dimensions and monthly new cases time-series showed blip dimensions were leading the new cases as shown in Figure 3(a). The higher correlation at negative lags showed that blip dimensions are associated with the future months' new cases. The heatmap in Figure 3(c) and 3(d) for Brazil and England showed an overall leading relationship of blip dimensions with new cases. Therefore, spike protein mutations in a given month can derive new cases in subsequent months.

Secondly, we identified the significant blip dimensions associated with the subsequent 1 month's new cases. We have used a Random forest based feature selection algorithm (Boruta), which showed that dimension-3 was the most significant predictor of next month's new cases across countries (Figure 3(e)). The random forest model on the selected features achieved a total R-squared of 37% on the testing set. Out of which 3% variance was explained by Blips dimensions.

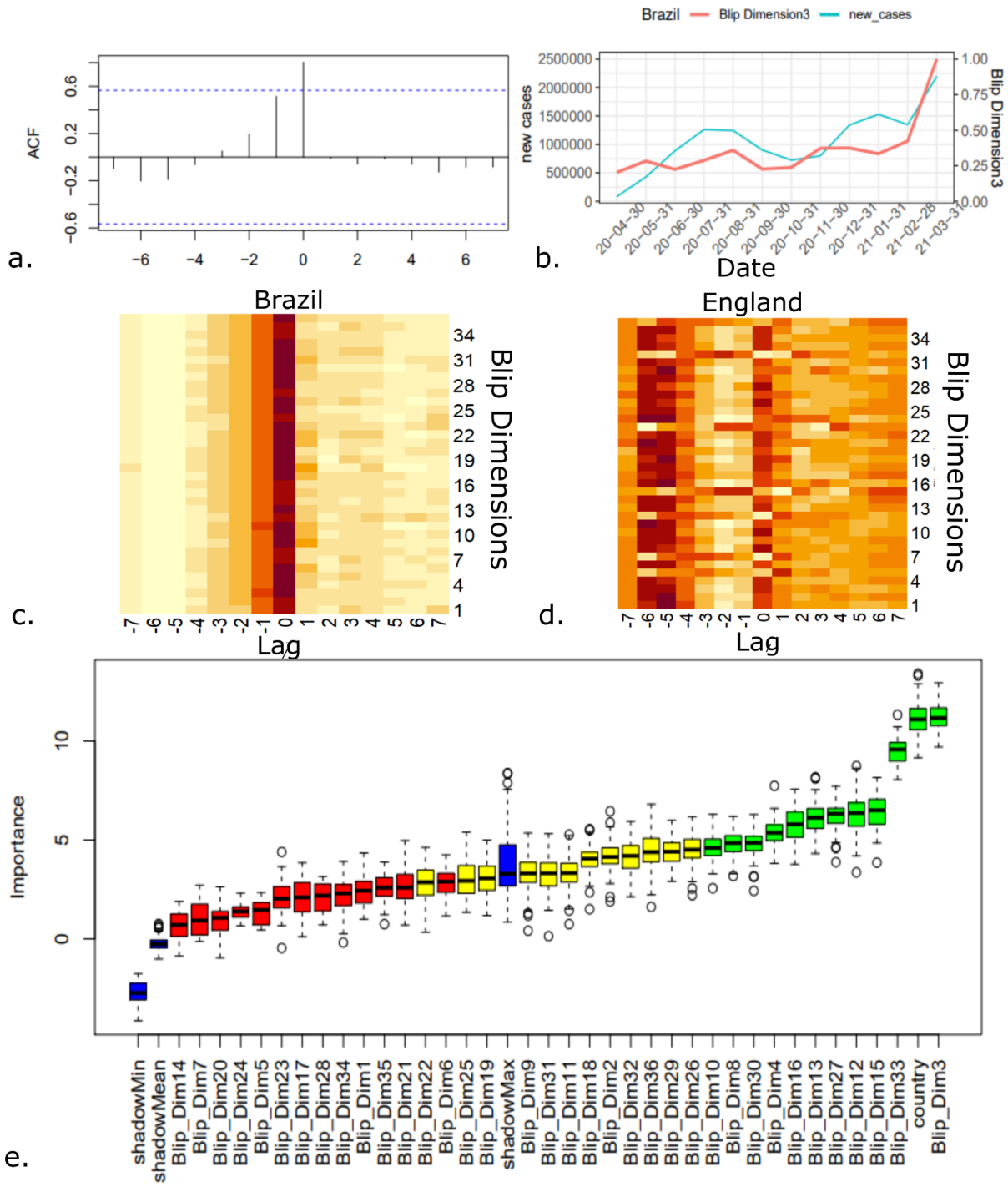

Figure 3: *a) Cross-correlation between Delta log transformed blip dimension 3 and Delta log transformed new cases for Brazil. b) Line plot for blip dimension 3 and monthly new cases for Brazil. c,d) Heatmap of cross-correlation between delta log transformed blip dimensions and*

delta log transformed new cases for Brazil and England (Darker colors represent higher correlations) e) Important features selected through boruta feature selection algorithm.

Dimensions of Concern capturing biologically important K-mer contributions are predictive of new caseloads. Table 1 illustrates the top contributing 3-mers and their associated weights (contribution) for the significant dimensions obtained from Boruta analysis (Dimensions 3, 33, 15, 12 and 30). The biological significance of these dimensions is explored by finding the Variants of concern (VOCs) and Variants of Interest (VOIs) associated with the top contributing codons in these dimensions. It was observed that dimension-12 captured the CGG and CTG codons (ranks 3 and 9 respectively), known to occur in the L452R variant. Similarly, dimension-33 captured six 3-mers (CAC, TGC, GAC, GAG, TGT, ACG) that are associated with multiple variants such as Q677H, W152C, D1118H, Δ156 and K417T, causing increased infectivity and spread. Dimension 30 captures tri-mers CAT and CAC associated with Δ69 and D1118H respectively which are linked to B.1.1.7 lineage.

Geospatial analysis indicated that dimension-33 was associated with UK variants such as Q677H and D1118H belonging to lineage B.1.525 and B.1.1.7 respectively. The decrease in blip counts in this dimension for the period November 2020 to January 2021 as shown in (Figure 4) coincided with the second and third national lockdown, start of vaccination and enforcement of stricter policy for social distancing [23][24].

Distinct dimensions captured country specific changes. For instance, dimensions 12 and 15 captured the 3-mer changes in codons AGC (Serine, S), ACG (Threonine,T) known to be associated with the P.1 lineage variant R190S, K417T [18] in Brazil. This variant was responsible for the enhanced transmissibility rate in Brazil from December, 2020. Both the dimensions in Figure 4 show an increase in the normalized blip counts across the three months. The increase in each month was followed by a rise in the active cases in Brazil in the next month. Dimensions 12 and 3 capture the doubled mutant L452R and dimension 33 captures the mutant Q1071H associated with lineage B.1.617.1 which were first observed in India in December 2020 [19] and were found to have increased infectivity and transmissibility. All these dimensions in Figure 4 also show an increase in the normalized blip counts across the three months for India. Therefore our analysis suggests that temporal tracking of these dimensions may be used as a surrogate to track the spread of the pandemic.

Latent Dimensions from Word Embeddings											
3		**12**		**13**		**15**		**30**		**33**	
3-mers	**Weights**	**3-mers**	**Weights**	**3-mers**	**Weights**	**3-mers**	**Weights**	**3-mers**	**Weights**	**3-mers**	**Weights**
TAA	1.02	TAA	1.85	ACG	0.9	ACG	0.84	TAG	1.11	TAA	6.31
GGG	0.83	CGC	1.38	TCG	0.7	CGG	0.81	GGG	0.62	CGC	2.96
AGC	0.77	CGG	1.18	AGC	0.65	CCC	0.71	CAT	0.55	CAC	2.19
CGC	0.75	TCG	1.05	TAA	0.61	CGC	0.5	TAA	0.53	CGA	2.09

ACG	0.69	TGA	0.93	GCG	0.53	CGT	0.42	CTG	0.5	CCG	2.03
TGA	0.58	CAC	0.73	TCC	0.51	AGG	0.35	TCG	0.46	TGC	1.89
TAG	0.53	ACG	0.57	AGG	0.47	TAG	0.33	CAC	0.45	GAC	1.81
TCG	0.53	CCC	0.52	TGC	0.35	GGA	0.3	ACC	0.35	GAG	1.81
CGG	0.52	CTG	0.51	AGT	0.32	CTC	0.29	ACG	0.33	TGT	1.79
CAC	0.42	AGC	0.37	CGG	0.32	GGC	0.28	CTA	0.3	ACG	1.78

Table 1: *Showing the top contributing 3-mers for the significant dimensions obtained from Boruta analysis. The biological significance of these dimensions is found by finding the Variants of concern(VOCs) and Variants of Interest(VOIs) associated with the top codons in each dimension.*

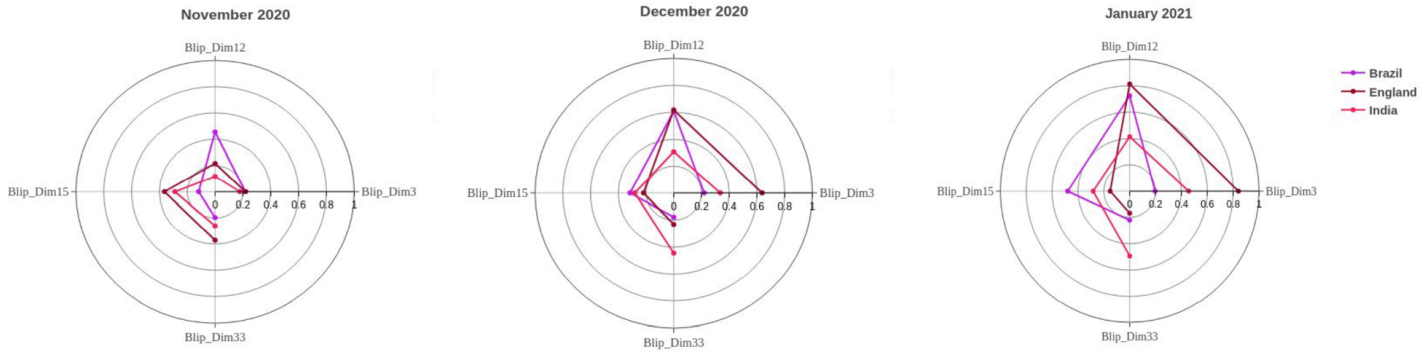

Figure 4: *Radar plots depicting the normalized blip counts for selected dimensions in the months of November 2020, December 2020 and January 2021, for Brazil, England and India. Both Brazil and India show an increase in normalized blip counts in Dimension 12 (associated with the L452R variant) across the three months, whereas England shows a decrease in Dimension 33 (associated with the D1118H variant), concurrent with the respective spreads in these countries.*

Discussion.

We have implemented an approach for analyzing the emerging strains based on the latent space of spike proteins coding nucleotide sequences. We chose the nucleotide sequences instead of proteins in order to capture and track the variations that may not have immediate functional consequences. Our approach has two main underlying tenets: (i) long range interactions are known to modulate the functional [25] interaction between receptor binding domain and ACE2 receptors, hence may be captured in the NLP models that capture 3-mer changes and context, and (ii) latent space dimensions may be differentially correlated with indicators of spread, thus

providing a data-driven handle for tracking and predicting variants of concern. The pipeline takes advantage of temporal changes in semantic similarity between mutating sequences. Preservation of phylogenetic structure based upon the similarity matrix obtained using the embedding validated its potential.SARS-CoV-2 is a single stranded positive RNA virus with approximately 30,000 nucleotide bases arranged in 12 coding regions [26]. Analyzing the dynamic patterns of 30,000 base pair long sequences while shedding light on underlying correlations is important to shed light on mechanistic understanding of mutations. SARS-CoV-2 seems to show a particularly high frequency of recombinations arising due to the absence of proof-reading mechanism and sequence diversity calling its urgency in studying its transmission pattern [27].

Therefore predicting mutations in the spike protein, which binds to ACE2 receptors can help us estimate the spread of disease, and the efficacy of therapeutic treatments and vaccines [28][29].
An empirical analysis of our embedding dimensions led us to the understanding of mutations that are likely to cause blips in each of latent dimensions. We carried out controlled substitutions, insertions and deletions were simulated to create new sequences keeping a sequence which had no blips in any dimension as a reference. For substitution, we substituted 3-mers in the reference sequence with each possible 3-mer at each position, except the first and the last, enabling us to study the effect of substituting each kmer at each position. Similarly, studies on insertion and deletion of 3-mers were carried out. In addition, we tried random insertion and deletion of nucleotide bases in the sequence to cause frameshifts.

Our analysis showed that certain 3-mers have a strong weight in multiple dimensions and these 3-mers when substituted, inserted or deleted from the sequence cause blips in multiple dimensions. Other 3-mers have strong weights in fewer dimensions and blips can be observed in those associated dimensions when a mutation involving those 3-mers takes place. Frameshifts cause blips in multiple dimensions as a large number of 3-mers in the sequence get modified. To interpret our embeddings, we identified the top 10 3-mers for each dimension that have the highest weights for those dimensions. At a given time step, we look at the number of dimensions for which blips are observed. If the number of such dimensions is high, it could either be because of a frameshift in the sequence or a mutation involving the 3-mers having high weights in multiple dimensions, that we have identified. If the number of latent dimensions having blips is low, we look at those specific dimensions where blips occur and find the top 3-mers associated with those dimensions. In our studies each genomic sequence is represented in 36 dimensional latent vector space and each dimension is capturing the 3-mers at nucleotide level which are showing blips that are driven by substitution, insertion and deletion of 3-mers. The top ten 3-mers for each dimension are filtered on the basis of weights and mapped with variants of interest and concern.

The biological significance of our embeddings-based approach can be understood from the known variants and their properties. One-third of the spike region is known to have mutated so far as evidenced from the data deposited in publicly available databases such as GISAID [30]. The classification of variants into three classes i.e. Variant of Interest (VOI), Variant of Concern (VOC) and Variant high consequence (VOHC)[18] provide some insights into the genomic features of these strains. B.1.1.7, B.1.351, P.1, B.1.427, and B.1.429 lineages are associated with VOCs that are capable of evading host immune mechanisms and cause outbreak clusters with increased infectivity, transmissibility and severity. These are known to accumulate a number of mutations before these transform into established lineages. For example, Δ69/70, A570D, D614G, P681H, T716I, S982A, D1118H, K1191N characterize the B.1.1.7 strain, identified first in the UK. Similarly, the P.1 lineage, reported first in Brazil, includes the variants T20N, D138Y, R190S,K417N, K417T, E484K, N501Y, D614G, H665Y. The B.1.526, B.1.526.1, B.1.617, B.1.617.1, B.1.617.2, B.1.617.3 and P.2 lineages are associated with VOIs and are characterized by reduced neutralization post-vaccination and increased transmissibility. The P.2 lineage, first detected in Brazil [18], includes mutations F565L, D614G, V1176F and E484K. The B.1.617 lineage includes the L452R, E484Q and D614G variants, its sub lineages B.1.617.1 (including additional variants T95I, G142D, E154K, P681R and Q1071H) and B.1.617.2 (including additional variants T19R, G142D, Δ156, Δ157, R158G, T478K, P681R, D950N) have shown capability of reduced susceptibility to monoclonal antibody treatment and hence require urgent attention [31]. The L452R mutation, characterized by replacement of CTG (Leu, L) with CGG (Arg, R) at 452nd position of the Receptor Binding Domain (RBD) of the Spike protein and by a T-to-G nucleotide substitution at 22917th genomic position of the spike region, offers a selective significance in both infection and transmission, making it a predominant variant across many countries [28][32].

Latent space dimensional analysis revealed that the codons AGG (Arginine, R) and AGT (Serine, S), associated with R190S mutation [33]], were captured in 5 out of total 36 dimensions. Similarly, CTG (Leu, L) and CGG (Arg, R), associated with L452R, were captured in 18 and 22 dimensions respectively. Capturing these codons adds a special significance to our studies as variants R190S and L452R are known to exhibit antibody neutralizing characteristics resulting in increased infectivity [18][20]. Besides this, the variants crucial to India and UK were also captured by our dimensional analysis, for instance, the codons CAT(Gln, Q) and CAC(His, H), associated with Q1071H of B.1.617.1 lineage, codons AGA(Arg,R) and GGA(Gly,G), associated with R158G of B.1.617.2 lineage, codons ACA (Thr, T) and ATA (Ile, I) linked to T716I of B.1.1.7 lineage were captured in multiple dimensions.

While all the 36 dimensions are able to capture the 3-mers associated with variants of concern and interest, dimensions 3, 12, 15, 30, and 33 were found to be statistically significant from Boruta feature selection for the prediction of new cases in the successive month. These Dimensions of Concern (DoCs) show a strong relationship with the current pandemic trends as

illustrated in the radar plots. Dimensions 3 and 12 captured the codons AGC (Serine, S), CGG (Arg, R), ACG (Thr, T) and CAC (His, H), the CGG and CTG codons are significant as they are associated with the Indian double mutant L452R. The latent space dimensional analysis also predicted the deletions associated with variants. For instance, dimensions 4 and 33 captured GAG(E) associated with the mutation Δ156 belonging to B.1.617.2 (which was recorded in December 2020 in India).

Prediction models can be used for training the genomic sequence for predicting infection severity based on Co-associations between the SNPs of Co-morbid Diseases and COVID-19 [34]. The machine learning models can also be trained on genomic sequences for COVID-19 classification [35]. Our exploratory cross correlation analysis suggested a leading relationship of normalized blip dimension with new COVID-19 cases. This motivated the use of machine learning techniques such as Boruta to select important predictors of future new cases. We have used the Random Forest model for predicting the monthly new cases by taking the Dimensions of Concern as a feature vector. Although the variance explained by our model is low, however, we were able to compute the variability associated with spike protein mutations. So our method showed a potential way to estimate the new cases variability associated with spike protein mutations. Our methods can be incorporated with the epidemic projections model to better predict the epidemic trajectories.

There are few limitations to our study. Although the embeddings indicate 3-mers with high weights, these do not indicate the position where the 3-mer change may have happened in the genome. This is because low-dimensional embeddings do not preserve the positional encoding of words. However, we plan to evaluate advanced approaches such as complex valued word embeddings with positional encodings [36] and transformer models such as BERT [37][38]. The latter are considered expensive and data-hungry models and it will remain to be evaluated if the gain of positional information may be countered by the loss of prediction accuracy for forecasting new cases in the future. However, we believe that the availability of sequences for a wide variety of viral pathogens presents an exciting opportunity to train data-hungry models that may be able to transfer insights across pathogens and yet remain interpretable. Another limitation of our study is the relatively small number of samples that were used to construct the supervised predictive models. This is to eliminate the sampling bias that may arise while building the supervised models as some countries had a disproportionately higher number of samples submitted to GISAID. However, the unsupervised embeddings and temporal cross-correlations were learned upon the full datasets and these presented clear patterns in DoCs and significant cross correlations with caseloads. Nevertheless, it is important for our models to receive at least 30 samples per month for getting reliable prediction results for the future months. This also underscoring the need for a more reliable and agile approach to deposit country-level datasets on repositories such as GISAID. We make an appeal to the countries to facilitate the sharing of such data in order to be prepared for any future waves of the current pandemic and for preventing new

emergence of strains. We believe our study is an instance of the new paradigm of pathogen surveillance using a novel language modeling approach that is potentially scalable to infectious disease surveillance and antimicrobial resistance.

Acknowledgments.
This work was supported by the Delhi Cluster- Delhi Research Implementation and Innovation (DRIIV) Project supported by the Principal Scientific Advisor Office, Prn.SA/Delhi/ Hub/2018(C) and the Center of Excellence in Healthcare supported by Delhi Knowledge Development Foundation (DKDF) at IIIT-Delhi. We also thank Dr. Chitra Pattabiraman (NIMHANS) for her valuable inputs on viral sequences, and Harleen Kaur and Nishkarsh Saxena for their contribution to the dashboard visualizations.

References.

1 Cumulative confirmed COVID-19 deaths. (https://ourworldindata.org/grapher/cumulative-covid-deaths-region).
2 Khan, M.I. *et al.* (2020) Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. *PLoS One* 15, e0238344
3 Galloway, S.E. *et al.* Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. , *MMWR. Morbidity and Mortality Weekly Report*, 70. (2021) , 95–99
4 Williams, H. *et al.* (2021) Watching Brief: The evolution and impact of COVID-19 variants B.1.1.7, B.1.351, P.1 and B.1.617. *Global Biosecurity* 3,
5 Hadfield, J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123
6 Ng, P. 23-Jan-(2017) , dna2vec: Consistent vector representations of variable-length k-mers.
7 Mikolov, T. *et al.* 16-Jan-(2013) , Efficient Estimation of Word Representations in Vector Space.
8 Yilmaz, A. Assessment of Mutation Susceptibility in DNA Sequences with Word Vectors. , *Journal of Intelligent Systems: Theory and Applications*, 3. (2020) , 1–6
9 Wu, F. *et al.* (2021) A deep learning framework combined with word embedding to identify DNA replication origins. *Sci. Rep.* 11, 844
10 Hie, B. *et al.* (2021) Learning the language of viral evolution and escape. *Science* 371, 284–288
11 Shu, Y. and McCauley, J. (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22,
12 Liu, B. *et al.* (2020) CoV-Seq, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study. *J. Med. Internet Res.* 22, e22299
13 Radim Rehurek, P.S. (2010) , Software Framework for Topic Modelling with Large Corpora. , in *IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS*
14 Papers with Code - Embedding Projector: Interactive Visualization and Interpretation of Embeddings. (https://paperswithcode.com/paper/embedding-projector-interactive-visualization).

15 Yin, Z. and Shen, Y. (2018) On the Dimensionality of Word Embedding. *Adv. Neural Inf. Process. Syst.* 31,
16 hclust function - RDocumentation. (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust).
17 Kursa, M.B. and Rudnicki, W.R. Feature Selection with theBorutaPackage. , *Journal of Statistical Software*, 36. (2010)
18 CDC 06-May-(2021) , SARS-CoV-2 Variant Classifications and Definitions. (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html).
19 SARS-CoV-2 variants of concern as of 6 May 2021. (https://www.ecdc.europa.eu/en/covid-19/variants-concern).
20 SARS-CoV-2 Mutations Complicate COVID-19 Pandemic Response. (https://www.criver.com/eureka/sars-cov-2-mutations-complicate-covid-19-pandemic-response).
21 Das, J.K. *et al.* (2021) Characterizing genomic variants and mutations in SARS-CoV-2 proteins from Indian isolates. *Gene Rep*
22 Peacock, T.P. *et al.* (2021) SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J. Gen. Virol.* 102,
23 clinicaltrialarena 10-Mar-(2021) , Covid-19 cases in the UK down by 92% since January. (https://www.clinicaltrialsarena.com/comment/covid-19-cases-uk-decline/).
24 Davies, N.G. *et al.* (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 372,
25 Mugnai, M.L. *et al.* (2020) Role of Long-range Allosteric Communication in Determining the Stability and Disassembly of SARS-COV-2 in Complex with ACE2. *bioRxiv* DOI: 10.1101/2020.11.30.405340
26 Shishir, T.A. *et al.* (2021) In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. *PLoS One* 16, e0245584
27 Mandal, S. *et al.* (2021) Pattern of genomic variation in SARS-CoV-2 (COVID-19) suggests restricted nonrandom changes: Analysis using Shewhart control charts. *J. Biosci.* 46,
28 Srivastava, S. *et al.* (2021) SARS-CoV-2 genomics: An Indian perspective on sequencing viral variants. *J. Biosci.* 46,
29 Li, Q. *et al.* (2020) The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* 182, 1284–1294.e9
30 Guruprasad, L. HUMAN SARS CoV-2 SPIKE PROTEIN MUTATIONS. .
31 Deng, X. *et al.* (2021) Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* DOI: 10.1016/j.cell.2021.04.025
32 Rouchka, E.C. *et al.* (2020) Variant analysis of 1,040 SARS-CoV-2 genomes. *PLoS One* 15, e0241535
33 paola 11-Jan-(2021) , Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. (https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585).
34 Wang, R.Y. *et al.* 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 International Conference on Computer Science and Network Technology (ICCSNT)*. (2020)
35 Arslan, H. Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. , *Proceedings*, 74. (2021) , 20
36 Wang, B. *et al.* 27-Dec-(2019) , Encoding word order in complex embeddings.
37 Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. , *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. (2020)
38 Lee, J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240