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. Early prediction of emerging strains could be vital to pandemic preparedness but remains an open challenge. Here, we derive Dimensions of Concern (DoCs) 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. We modeled viral DNA sequences as documents with codons treated as words to learn unsupervised word embeddings. We discovered that "blips" in latent dimensions of the learned embeddings were associated with mutations. Latent dimensions which harbored blips that consistently preceded and were predictive of new caseloads were analyzed further as *Dimensions of Concern*, *DoCs*. The DOCs captured CGG, CTG, AGG, AGT, GAC and, CAC codons associated with major global VoCs L452R, R190S, and D1118H, thus validating our approach biologically. Tracking these DOCs can provide a practical approach to predict country-specific emergence and spread of viral strains for genomic surveillance and is extensible to related challenges such as immune escape, pathogenicity modeling, and antimicrobial resistance.

COVID-19 is reported to have claimed 3.46 million lives as of May 24, 2021 [1]. Many of these deaths are attributed to unexpected surges in infections caused by new strains of SARS-CoV-2, prompting health organizations such as the CDC to declare these as variants of concern (VoCs) [2]. Such emergence of new variants can seriously undermine the efficacy of global vaccination programs or cause multiple reinfection waves as new strains may escape antibodies. As a result, several initiatives have focused on providing high-quality tracking information for the strains and lineages as these emerge [3]. 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, which 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 model 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 [4]. Word embeddings represent the latent space of a corpus of text [5] and can capture highly nonlinear and contextual relationships. Codons with their tri-nucleotide translations represent a biological basis for word representations. They have been utilized to learn embeddings for modeling various outcomes such as mutation susceptibility [6] and gene sequence correlations [7]. Our empirical experiments with the learned embeddings uncovered explainable genomic signals and predicted new caseloads across nine countries.

Reading the genome to learn word embeddings for advancing biological understanding is a relatively recent area of research. Recently, Hie et al. [8] used machine learning and word embedding techniques to model the semantics and grammar of amino acids corresponding to antigenic change to predict the mutations that might lead to viral escape. This paper focuses on the semantics of viral DNA sequences to derive *Dimensions of Concern (DoCs)* and demonstrate their causal potential for increasing epidemiological spread patterns across nine countries. Our approach is extensible to global threats in pathogen surveillance such as emerging infections, pandemics, and antibiotic resistance.

Following the linguistic idea that the meaning of a word is characterized by the company it keeps, we leveraged a word2vec language model to learn the latent representations of codons in spike protein sequences of SARS-COV-2. Sequence-level embeddings were then obtained from the codon embeddings and investigated for the presence of genomically meaningful characteristics. The phylogenetic tree (obtained from the embeddings) for the United Kingdom (Figure 1(a)) shows two clear temporally split clusters for 2020 and 2021 sequences, which may be indicative of different strains in these time periods. The temporality of the collected sequences was found to be preserved in the two clusters, even though the model was trained only on genome sequences. Figures 1(b,c) demonstrate that geospatial information is preserved in the sequence embeddings. Some countries form distinct clusters, while others are dispersed across many clusters. This behavior was expected because viral strains spread from one country to another, and our samples had different collection dates for each country. This analysis suggests that our language model captures the temporal emergence of strains and geographic information in a country-specific manner.

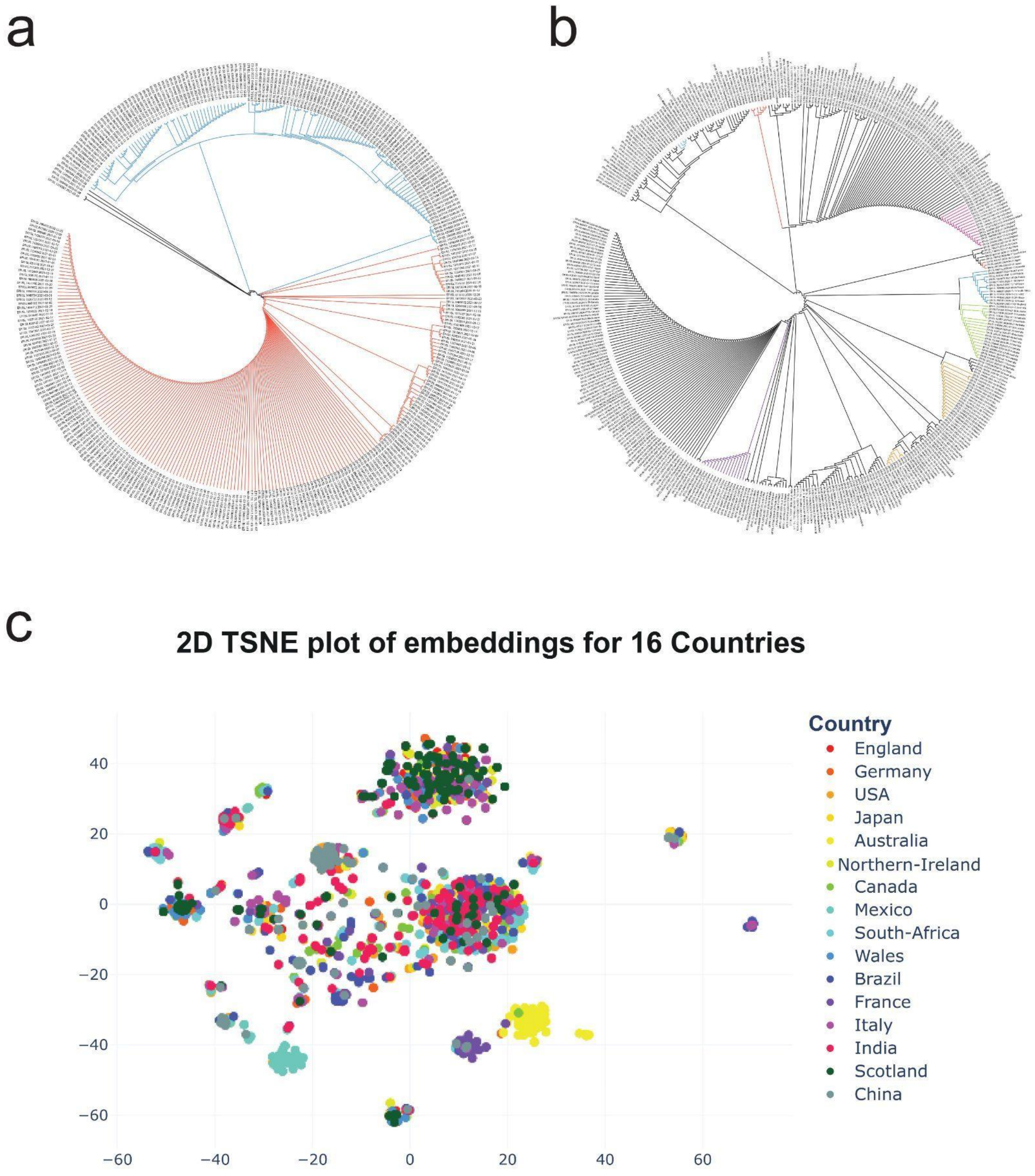

Figure 1: Latent space preserves spatiotemporal information of COVID-19 spread. *Phylogenetic trees are constructed using cosine similarities between 400 randomly sampled sequence embeddings. (**a**) Dendrogram for strains from the U.K.: Cluster 1 (blue) contains strains from the period Oct 2020 - Dec 2020, while Cluster 2 (orange) contains strains collected between Jan 2021 - Mar 2021. (**b**) Dendrogram for 16 countries across the globe: Chinese, Australian, Mexican, and England strains form tight clusters (marked in green, purple, and magenta), while strains from Italy, France, Brazil, Japan, Canada, USA, Scotland, and India are dispersed with other countries. (**c**) T-SNE plot for visualization of high-dimensional latent embeddings (200 randomly sampled sequences from each country). Distinct clusters for China, Australia, and Mexico were observed.*

Further analysis of the sequence embeddings for different periods revealed sudden changes ("blips") in values in each dimension. We attributed these blips to changes in the spike protein of the genome sequence. This hypothesis was based on the fact that the embedding of a word (codon) represents the context it occurs. A mutation results in a codon that is known to be found in a different context. This alters the semantic sense of the original sequence, and this modification of meaning is reflected as a change in values (blips) across different latent dimensions. This hypothesis was validated in simulation experiments in synthetic datasets (Supplementary Information).

Monthly normalized blip counts were computed for each dimension (blip dimensions). A comparison of total normalized blip counts for different locations revealed that geospatially proximate regions have similar blip patterns across the latent dimensions (Figure 2(a)), indicating that these regions have been accumulating similar changes. Conversely, some regions that show dissimilar patterns suggest that they have accumulated unlike changes(Figure 2(b)).

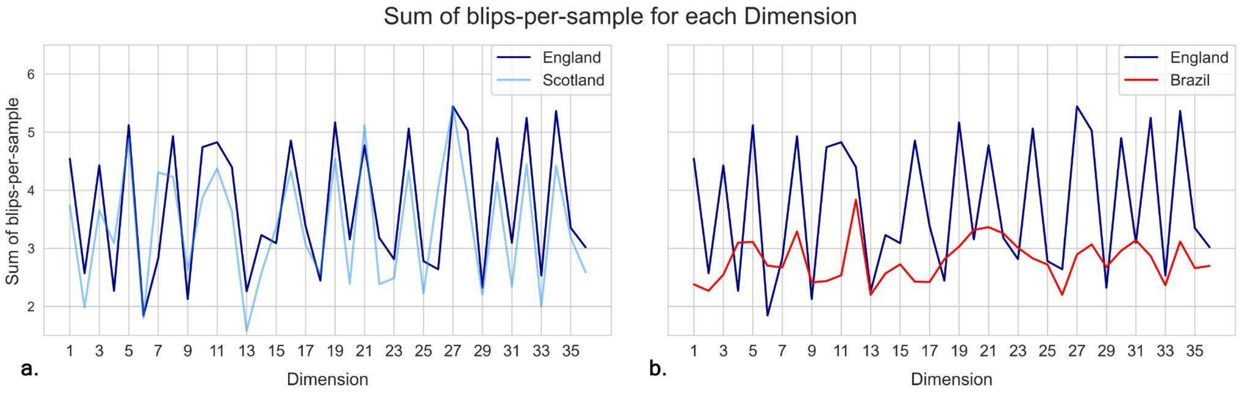

Figure 2: Latent space dimensions capture mutations in the spike protein.
*Countries in the U.K. (**a**) - England and Scotland show a similar distribution of total normalized blips across dimensions, while (**b**) England and Brazil have dissimilar distributions.*

We then attempted to decipher the relationship between blip dimensions and monthly new COVID-19 cases in different countries. Cross-correlation between the two revealed that blip dimensions have a leading relationship with new cases (Figure 3(a)-(d)). This suggests that the genome sequence data in a given month can be used to predict new cases in subsequent months. Boruta algorithm was employed to assign feature importance scores to different dimensions, which revealed that dimension 3 is the most significant predictor of new cases (Figure 3(e)). Statistically significant dimensions from Boruta analysis were termed Dimensions of Concern (DoCs). Random forest-based regression modeling on the DoCs achieved an R-squared of 37% on the test set, out of which 3% variance was explained by the blip dimensions.

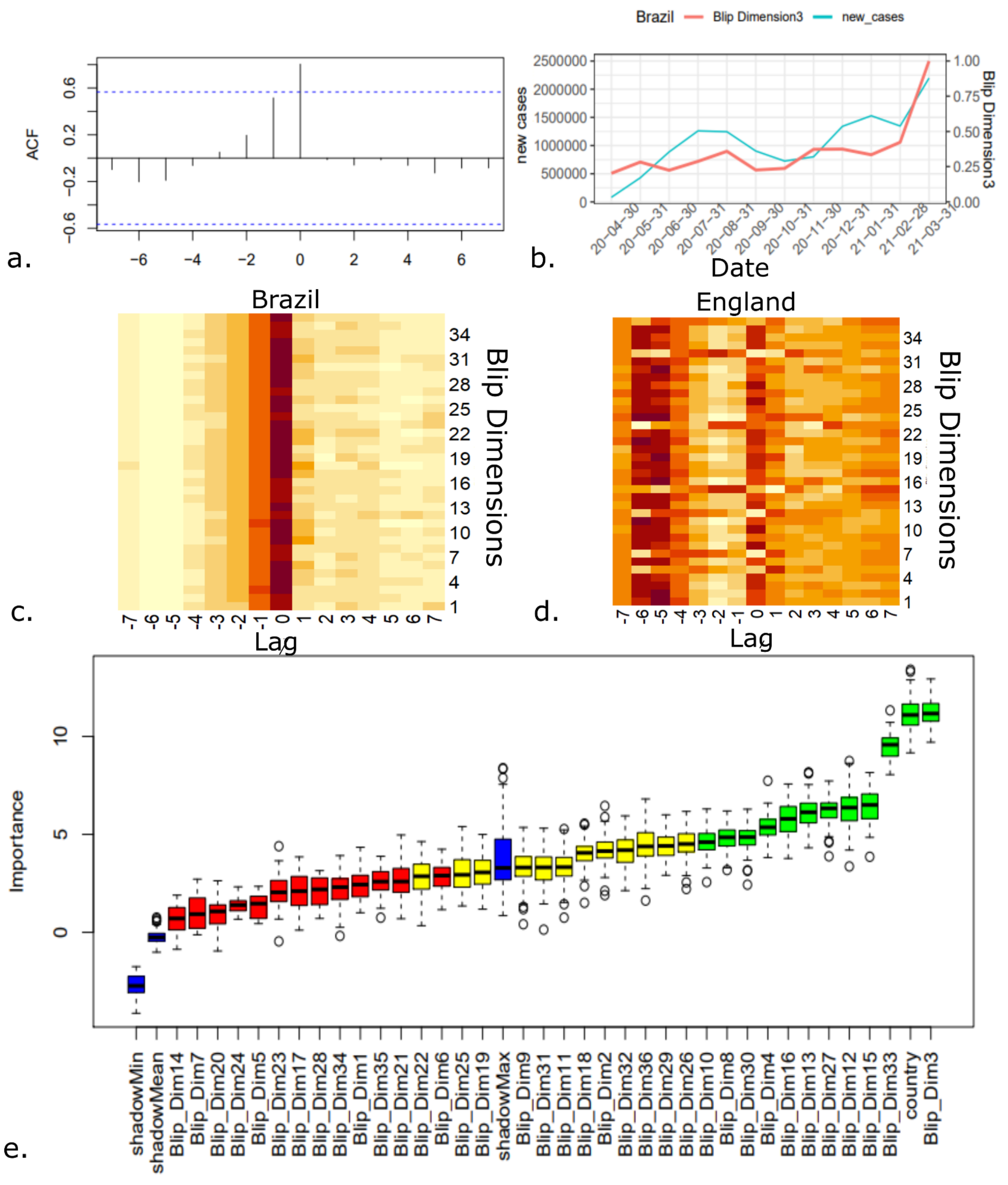

Figure 3: **Blip dimensions are predictive of new COVID-19 caseloads.**
a) *Cross-correlation between Delta log-transformed blip dimension three and Delta log-transformed new cases for Brazil.* ***b)*** *Line plot for blip dimension three and monthly new cases for Brazil.* ***c, d)*** *Heatmaps of cross-correlation between delta log-transformed blip dimensions and delta log-transformed new cases for Brazil and England (Darker colors represent higher correlations). These plots indicate that blip dimensions have an overall*

leading relationship with new cases. ***e)*** *Feature importance scores from the Boruta algorithm for predicting cases in subsequent months.*

The Biological significance of DoCs was assessed by finding the top ten contributing trimers and their associated weights in each of these dimensions (Table 1). These codons were then mapped to Variants of concern (VOCs) and Variants of Interest (VOIs) (Supplementary Table 2). It was observed that codons CGG, CTG, AGG, AGT, GAC, CAC, ACG, and CAT were captured by DoCs associated with VoCs and VoIs such as L452R, R190S, D1118H, K417T, and Δ69. These variants are generally known to exhibit increased infectivity, pathogenicity, and transmissibility.

Dimensions of Concern											
3		**12**		**13**		**15**		**30**		**33**	
trimers	**Weights**	**trimers**	**Weights**	**trimers**	**Weights**	**trimers**	**Weights**	**trimers**	**Weights**	**trimers**	**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: Dimensions of Concern capture biologically significant K-mer contributions. *The table shows the top contributing trimers for the Dimensions of Concern. It provides a way to map the latent space dimensions to specific codons, and by extension, particular Variants of concern (VOCs) and Variants of Interest (VOIs). For example, dimension-12* captures the CGG and CTG codons, known to occur in the L452R variant. Dimension-33 captures the codons ACG, TGT, GAG, GAC, and CAC, associated with multiple variants such as K417T, W152C, Del 156, and D1118H. Dimension 30 captures codons CAT and CAC, associated with Δ69 and D1118H, respectively.

This mapping from latent space blips to different SARS-CoV-2 variants allows for the opportunity to predict emerging strains and take preventive measures in advance. Furthermore, distinct dimensions capture country-specific changes and maybe surveilled to monitor the spread of the pandemic. This hypothesis was confirmed by performing empirical analysis on several DoCs. Dimension-33 captures the codons GAC and CAC, known to occur in the D1118H variant of the U.K. origin. The decrease in blip counts in this dimension for the period November 2020 to January 2021 (Figure 4) coincided with the second and third national lockdown, the start of vaccination, and enforcement of stricter policies for social distancing [9][10]. Dimensions 12 and 15 reflect blips in codons ACG (linked with K471T variant) and AGG (related with R190S variant), respectively, and are known to be associated with the P.1 lineage [11] in Brazil. This variant was responsible for the enhanced transmissibility rate in Brazil from December 2020. Both the dimensions show an increase in the normalized blip counts across the three months (Figure 4). Dimensions 3 and 12 capture the double mutant L452R (associated with lineage B.1.617.1 and B.1.617.2), first observed in India in December 2020 [12], and was found to have increased infectivity and transmissibility. The two dimensions show an increase in the normalized blip counts across the three months for India (Figure 4). Therefore our analysis suggests that temporal tracking of these dimensions may be used as a surrogate to track the spread of the pandemic.

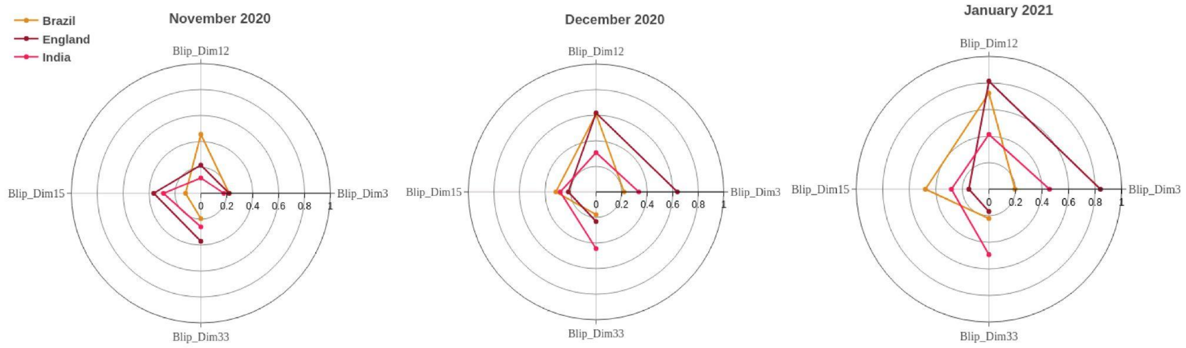

Figure 4: **Temporal tracking of blips in Dimensions of Concern may be used as a surrogate to track the spread of the pandemic.** *The radar plots depict the normalized blip counts for selected dimensions in November 2020, December 2020, and January 2021 for Brazil, England, and India. Both Brazil and India show an increase in normalized blip counts in Dimension 12 (associated with the L452R variant) across the three months, concurrent with the respective spreads in these countries. Conversely, England shows a decrease in Dimension 33 (associated with the D1118H variant), concurrent with the national lockdown and strict social distancing policies.*

Our study exploits the fact that a mutation in a genome sequence changes its ‘meaning’ and, therefore, its latent space representation. We have shown that these changes can be linked to specific strains and could potentially be used to model new COVID-19 cases in several countries across the globe. The latent dimensions may further be employed to predict the clinical consequences of emerging strains. The currently available vaccines are intended for early SARS-CoV-2 strains, but with new emerging variants, immune responses triggered by these vaccines may be weaker and short-lived. As seen in the devastating second wave of the pandemic in India, newer SARS-CoV-2 variants have acquired an increased pathogenic potential resulting in rapid clinical progression and overwhelmed health systems. Mitigating such future events will require more robust surveillance systems in place. Our study offers a promising solution in this direction and lays the foundation for proactive genomic surveillance of COVID-19.

Discussion.

We have implemented an approach for analyzing the emerging strains based on the latent space of spike protein-coding nucleotide sequences. We chose the nucleotide sequences instead of proteins to capture and track the variations that may not have immediate functional consequences. Our approach has two central underlying tenets: (i) long-range interactions are known to modulate the functional [13] interaction between receptor binding domain and ACE2 receptors, hence may be captured in the NLP models that capture trimer 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 and variants of interest. Our approach takes advantage of temporal changes in the semantics of mutating sequences. Preservation of phylogenetic structure based upon the similarity matrix obtained using the embeddings validated that the latent dimensions capture Spatio-temporal information. Analyzing the dynamic patterns and underlying correlations in the 30,000 base pair long sequence of SARS-CoV-2 [14] is vital to highlight the mechanistic understanding of mutations. SARS-CoV-2 seems to show an exceptionally high frequency of recombinations arising due to the absence of proof-reading mechanism and sequence diversity, which calls for urgency in studying its transmission pattern [15][16]. Therefore predicting mutations in the spike protein, which binds to ACE2 receptors, can help us estimate the spread of disease and the efficacy of treatments and vaccines [17][18].

In our study, each genomic sequence is represented in a 36-dimensional latent vector space. Each dimension captures the trimers at the nucleotide level, showing blips driven by substitution, insertion, and deletion of trimers. The top ten trimers for each dimension are filtered based on 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 by the data deposited in publicly available databases such as GISAID [19]. The classification of variants into three classes, i.e., Variants of Interest (VOIs), Variants of Concern (VOCs), and Variants of high consequence (VOHC)[11], 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 several 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 [11], includes mutations F565L, D614G, V1176F, and E484K. The B.1.617 lineage consists of 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 the capability of reduced susceptibility to monoclonal antibody treatment and hence require urgent attention [11]. 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 particular significance in both infection and transmission, making it a predominant variant across many countries [20][21].

Latent space dimensional analysis revealed that the codons AGG (Arginine, R) and AGT (Serine, S), associated with R190S mutation [22], 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. Capturing these codons adds a special significance to our studies as variants R190S and L452R exhibit antibody neutralizing characteristics resulting in increased infectivity [11][23]. Besides this, the variants crucial to India and UK were also captured by our dimensional analysis. For instance, the codons CAT (Gln, Q) 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 can capture the trimers associated with variants of concern and interest, dimensions 3, 12, 13, 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, 12 captured the codons CGG (Arg, R), ACG (Thr, T), and CAC (His, H). Codons CGG and CTG are significant as they are associated with the Indian double mutant L452R. The latent space dimensional analysis also captured the deletions related to variants. For instance, dimensions 4 and 33 captured GAG (E) associated with the mutation Δ156 belonging to B.1.617.2 (recorded in December 2020 in India). Similarly, dimension-13 captured four trimers (ACG, CGG, AGG, AGT) associated with multiple variants such as K417T, L452R, and R190S causing increased infectivity and spread.

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 [24]. The machine learning models can also be trained on genomic sequences for COVID-19 classification [25]. Our exploratory cross-correlation analysis suggested a leading relationship of normalized blip dimension with new COVID-19 cases. This finding motivated using 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 possible way to estimate the new cases variability associated with spike protein mutations. Our methods can be incorporated with the epidemic projections model to predict the epidemic trajectories better.

There are few limitations to our study. Although the embeddings indicate trimers with high weights, these do not tell the position where the trimer 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 [26] and transformer models such as BERT [27][28]. The latter are considered expensive and data-hungry models. 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 used to construct the supervised predictive models. This eliminates 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 complete datasets, and these presented clear patterns in DoCs and significant cross-correlations with caseloads. Nevertheless, our models need to receive at least 30 samples per month for getting reliable prediction results for the coming months. This also underscores the need for a more agile and dependable approach to deposit country-level datasets on repositories such as GISAID. We appeal to the countries to facilitate the ing such data to be prepared for any future waves of the current pandemic and prevent the 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 funded 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, Nishkarsh Saxena, and Anjali 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 Hadfield, J. *et al.* (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123.

4 Ng, P. 23-Jan-(2017) , dna2vec: Consistent vector representations of variable-length k-mers.

5 Mikolov, T. *et al.* 16-Jan-(2013) , Efficient Estimation of Word Representations in Vector Space.

6 Yilmaz, A. Assessment of Mutation Susceptibility in DNA Sequences with Word Vectors. , *Journal of Intelligent Systems: Theory and Applications*, 3. (2020) , 1–6.

7 Wu, F. *et al.* (2021) A deep learning framework combined with word embedding to identify DNA replication origins. *Sci. Rep.* 11, 844.

8 Hie, B. *et al.* (2021) Learning the language of viral evolution and escape. *Science* 371, 284–288.

9 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/).

10 Davies, N.G. *et al.* (2021) Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 372.

11 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).

12 SARS-CoV-2 variants of concern as of 6 May 2021. (https://www.ecdc.europa.eu/en/covid-19/variants-concern).

13 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.

14 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.

15 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.

16 Rouchka, E.C. *et al.* (2020) Variant analysis of 1,040 SARS-CoV-2 genomes. *PLoS One* 15, e0241535.

17 Srivastava, S. *et al.* (2021) SARS-CoV-2 genomics: An Indian perspective on sequencing viral variants. *J. Biosci.* 46.

18 Li, Q. *et al.* (2020) The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* 182, 1284–1294.e9.

19 Guruprasad, L. HUMAN SARS CoV-2 SPIKE PROTEIN MUTATIONS. .

20 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.

21 Deng, X. *et al.* Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *medRxiv* DOI: 10.1101/2021.03.07.21252647.

22 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).

23 SARS-CoV-2 Mutations Complicate COVID-19 Pandemic Response. (https://www.criver.com/eureka/sars-cov-2-mutations-complicate-covid-19-pandemic-response).

24 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).

25 Arslan, H. Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. , *Proceedings*, 74. (2021) , 20.

26 Wang, B. *et al.* 27-Dec-(2019) , Encoding word order in complex embeddings.

27 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).

28 Lee, J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240.