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; *rintuk@iiitd.ac.in

[¶] Contributed Equally, [&] Contributed Equally

Keywords: SARS-CoV-2, COVID-19, Machine Learning, Language Modeling, Natural Language Processing.

Abstract

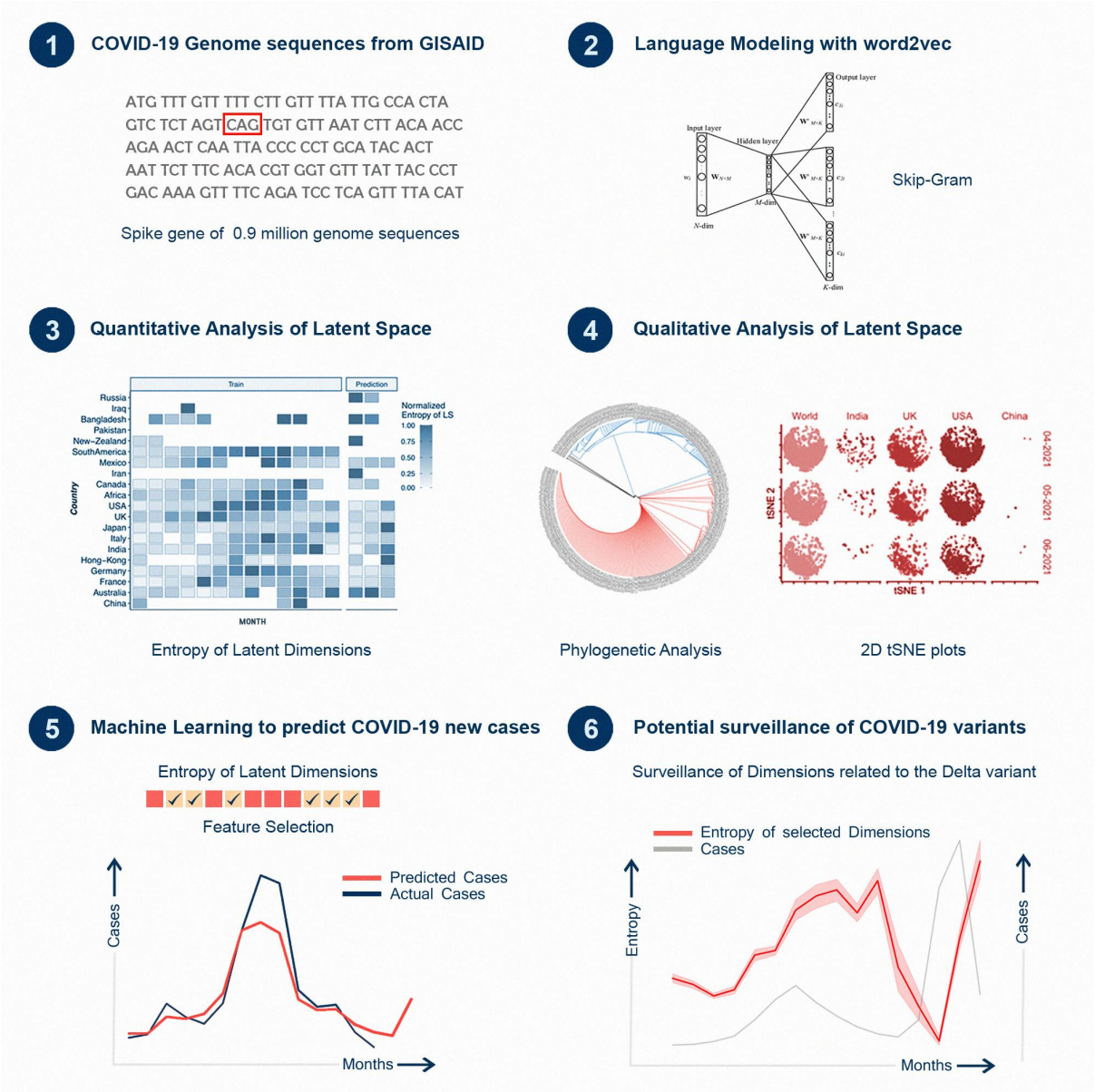
The global efforts to control COVID-19 are threatened by the rapid emergence of novel SARS-CoV-2 variants that may display undesirable characteristics such as immune escape or increased pathogenicity. Early prediction of emerging strains could be vital to pandemic preparedness but remains an open challenge. Here, we developed *Strainflow*, to learn the latent dimensions of 0.9 million high-quality SARS-CoV-2 genome sequences, and used machine learning algorithms to predict upcoming caseloads of SARS-CoV-2. In our *Strainflow* model, SARS-CoV-2 genome sequences were treated as documents, and codons as words to learn unsupervised codon embeddings (latent dimensions). We discovered that codon-level changes lead to a change in the entropy of the latent dimensions. We used a machine learning algorithm to find the most relevant latent dimensions called *Dimensions of Concern* (DoCs) of SARS-CoV-2 spike genes, and demonstrate their potential to provide a lead time for predicting new caseloads in several countries. The DoCs capture codons associated with global Variants of Concern (VOCs) and Variants of Interest (VOIs), and may be surveilled to predict country-specific emergence and spread of SARS-CoV-2 variants.

Highlights

- We developed a genomic surveillance model for SARS-CoV-2 genome sequences, *Strainflow,* where sequences were treated as documents with words (codons) to learn the codon context of 0.9 million spike genes using the skip-gram algorithm.
- Time series analysis of the information content (Entropy) of the latent dimensions learned by *Strainflow* shows a leading relationship with the monthly COVID-19 cases for seven countries (e.g., USA, Japan, India, and others).
- Machine Learning modeling of the entropy of the latent dimensions helped us to develop an epidemiological early warning system for the COVID-19 caseloads.

- The top codons associated with the most relevant latent dimensions (DoCs) were linked to SARS-CoV-2 variants, and these DoCs may be used as a surrogate to track the country-specific spread of the variants.

Graphical abstract



Introduction

COVID-19 is reported to have claimed 4.37 million lives as of Aug 17, 2021 (WHO, 2021). A large number of these deaths are attributed to unexpected surges in infections caused by new strains of SARS-CoV-2, prompting international health organizations such as the CDC and WHO to declare these as variants of concerns (Khan et al., 2020). Such emergence of new variants can seriously undermine the efficacy of global vaccination programs, the need for future booster doses, or cause multiple reinfection waves as new strains escape previously developed antibodies. Several initiatives have focused on providing high-quality tracking information for the strains and lineages as these emerge (Hadfield et al., 2018). 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 insight rather than proactive predictions. On the other hand, machine learning approaches are likely to be biased by underlying data characteristics and do not explain the biological basis of predictions.

Here, we propose ***Strainflow***, a hybrid architecture of machine learning and language modeling, along with empirical experiments to demonstrate explainable genomic signals for tracking and predicting the spread of SARS-CoV-2 across countries. Strainflow is rooted in language models for generating sequence embeddings that have recently shown promise for capturing biological insights from DNA sequences. Typically, in language models, word embeddings represent the latent space (dimensions) of a corpus of text (Mikolov et al., 2013) and can capture highly nonlinear and contextual relationships. Codons (tri-nucleotides, 3-mers) 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 (Yilmaz, 2020) and gene sequence correlations (Wu et al., 2021). Recently, Hie et. al (Hie et al., 2021) used machine learning along with word embedding techniques to model the semantics and grammar of amino acids corresponding to antigenic change to predict the mutations which might lead to 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 7 countries. Our approach, Strainflow, is extensible to global threats in pathogen surveillance such as emerging infections, pandemics, and antibiotic resistance.

Results

Genomic sequence-based language modeling captures emerging diversity in the SARS-CoV-2 spike gene.

Following the linguistic idea that the context of a word is characterized by its neighborhood, we leveraged a word2vec language model and developed "*Strainflow*" to learn the latent representations of codons in spike gene sequences of SARS-CoV-2. Strainflow is trained on 0.3 million high-quality genome sequences from March 1, 2020 to March 31, 2021, collected from the GISAID database (**Fig 1A**). To understand the latent space representation (LR) learned by Strainflow, we performed qualitative analysis using tSNE for both global and local (country) levels on a monthly basis. The global tSNE highlights dynamic emerging patterns derived from LR of spike genes of SARS-CoV-2 (**Fig 1B**) from September, 2020 to March, 2021, along with specific geographic locations (country-level) such as India, UK, USA, and China.

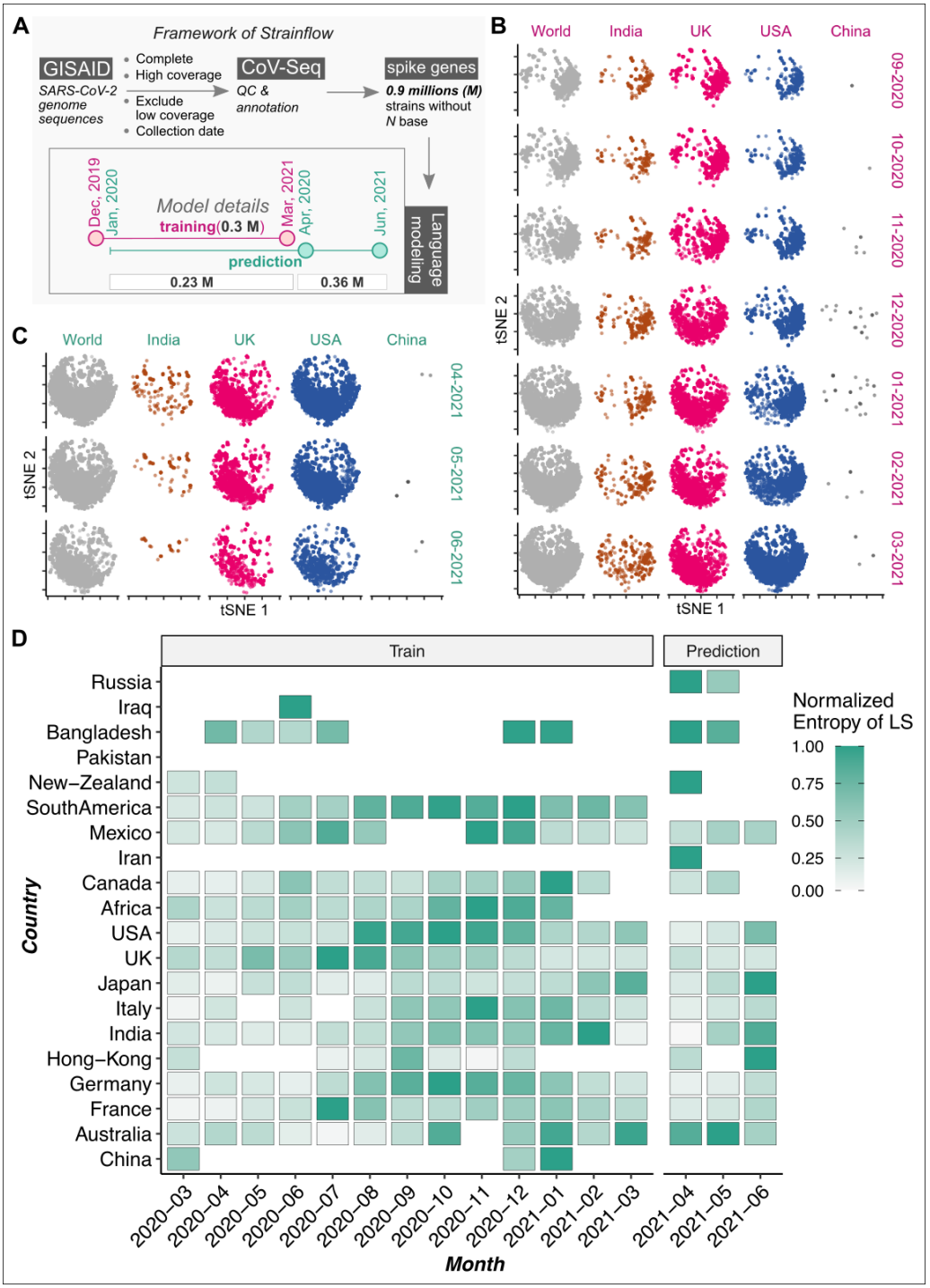


Figure 1: Latent space of spike gene derived using Strainflow preserves spatiotemporal information of SARS-CoV-2 spread. *(A) The implementation framework of Strainflow (details described in the method section) (**B**) tSNE plot showing distinct spatio-temporal relationship based on the latent space learned from the spike gene of 0.308 million SARS-CoV-2 genomes collected till 31 March 2021 (world), India, UK, USA, and China. (**C**) Embeddings estimated or predicted from the Strainflow model for 0.36 million SARS-CoV-2 spike genes from the month of April, 2021 to June, 2021. (**D**) Heatmap showing the normalized entropy for 20 countries from March, 2020 to June, 2021. For Pakistan, we couldn't estimate the fast sample entropy due to fewer number of strain sequences (<20).*

To investigate the information content in the latent dimensions or space (LD or LS) of the *Strainflow* model, we performed qualitative and quantitative analysis on 0.9 million SARS-CoV-2 spike genes collected from December, 2019 to June, 2021. Qualitative analysis was performed by performing dimensionality reduction with a fast tSNE method called Flt-SNE (Linderman et al., 2019). We compared the 2D t-SNE plot of the world with four countries (India, UK, USA, China) from September, 2020 to June, 2021, which clearly highlights the dynamic changes in the spike genes across countries in different months (**Fig 1B** and **Fig 1C**). Additionally, the quantitative analysis of the latent space was performed by calculating the fast sample entropy (Pan et al., 2011) for each LD. To compare the monthly entropy of the LS per country, mean entropy was calculated and normalized across the months for each country. We observed the highest entropy (information content) for India, UK, USA and China in the months of February-2021, July-2020, August-2020, and January-2021 respectively (**Fig 1D**). Interestingly, we observed high entropy for 4 months from August, 2020 to November, 2020 in the USA (**Fig 1D**). This highlights that the spike protein latent space representation learned by the Strainflow model could be used as a proxy to capture the spatiotemporal entropy or diversity in the emerging SARS-CoV-2 strains across different countries.

Preservation of spatiotemporal information for SARS-CoV-2 spread depicted through phylogenetic analysis.

Sequence-level embeddings were obtained from the codon embeddings and investigated for the presence of genomically meaningful characteristics. The phylogenetic tree derived from the embeddings for the United Kingdom (**Fig 2A**) 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, although the model was trained only on genome sequences.

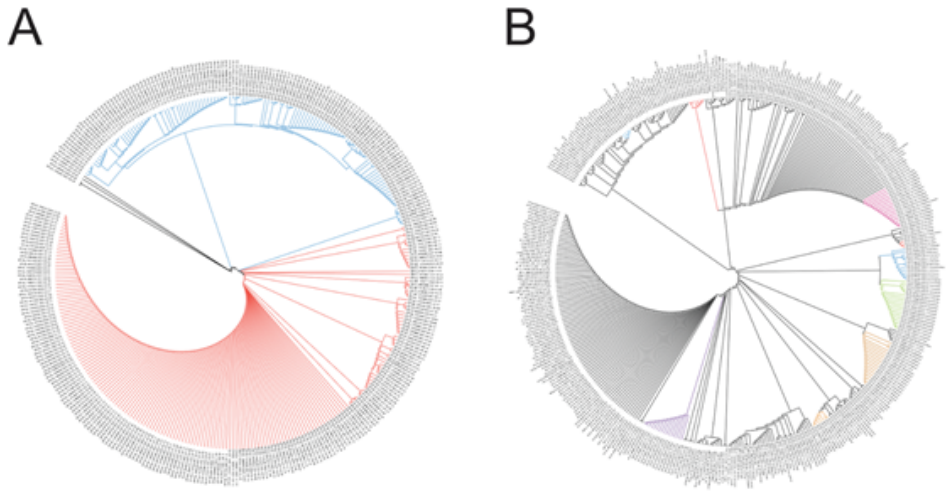


Figure 2: Phylogenetic trees constructed using cosine similarities between 400 randomly sampled sequence embeddings. *(**A**) Dendrogram for strains from the U.K.: Cluster 1 (blue) contains strains from the period Oct 2020 - Dec 2020, while Cluster 2 (orange) contains strains collected between Jan 2021 - 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.

The phylogenetic tree with globally collected sequences (**Fig 2B**) demonstrates that geospatial information is also preserved in the sequence embeddings. Strains from countries like China, Australia, Mexico, and England were found in distinct clusters, while those from Italy, France, Brazil, Japan, Canada, USA, Scotland, and India were found to be dispersed with other countries. This behavior is expected because viral strains spread from one country to another, and our samples had different collection dates for each country. Overall, Strainflow captures the temporal emergence of strains and geographic information in a country-specific manner.

Entropy in the latent space dimensions captures variability in the spike gene.
Analysis of the sequence embeddings for different time periods revealed that the value of each latent dimension changes with time. We attributed these changes to mutations 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 in. A mutation results in a codon that is known to be found in a different context. This alters the semantic sense of the sequence, and this modification of meaning is reflected as a change in values across different latent dimensions. The changed values cause a change in the degree of disorder or the entropy of the latent space.

To compare different geographical regions, the sum of sample entropy was computed for each latent dimension across all the months. This revealed that certain geospatial regions such as France and Germany (**Fig 3A**) and USA and Canada (**Fig 3B**) have similar total entropies across the latent dimensions, indicating that these regions have been accumulating similar changes.

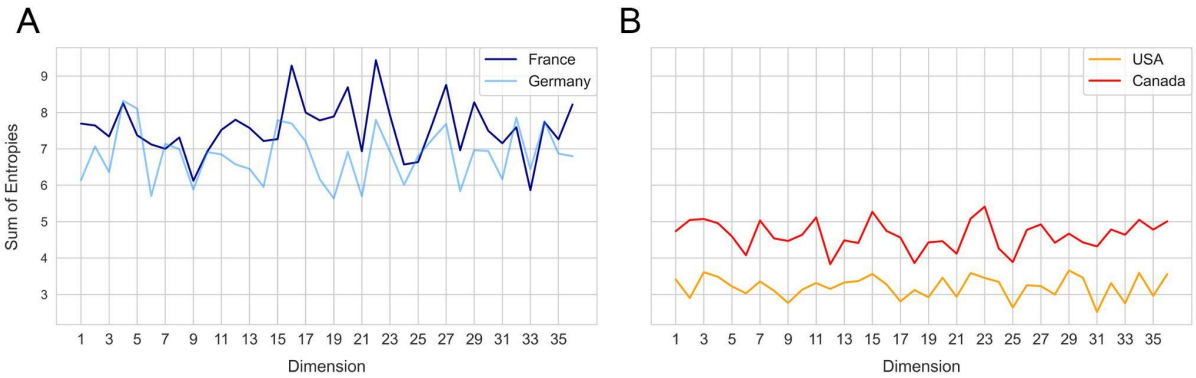


Figure 3: Sum of sample entropy for each latent dimension for different countries. *Country pairs (**A**) - France and Germany, (**B**) USA and Canada show a similar distribution of total sample entropy across dimensions, while each pair differs from the other.*

Entropy dimensions are predictive of new COVID-19 caseloads.

We then attempted to decipher the relationship between monthly sample entropy and monthly new COVID-19 cases in different countries. Detrended cross-correlation coefficient was calculated at different lag values, which revealed that entropy dimensions have a leading relationship with new cases (**Fig 4A, 4B**). This suggests that the genome sequence data in a given month can be used to predict new cases in subsequent months. A lead period of two months was chosen and Boruta algorithm was employed to assign feature importance scores to different dimensions, which revealed that dimension 32 is the most significant predictor of new cases **(Fig 4C)**. Significant dimensions from Boruta analysis were termed Dimensions of Concern (DoCs). Random forest based regression modeling on the DoCs achieved a total R-squared of 73% on the test set. The predicted cases were found to be highly correlated with the actual cases (**Table 1**), which suggests that our model can indicate the directional change of cases for different countries. Further, the predicted relative change in cases between successive months was found to be correlated to the actual relative changes (**Table S1**), which suggests that our model can also indicate the magnitude of change that we expect to observe in the cases.

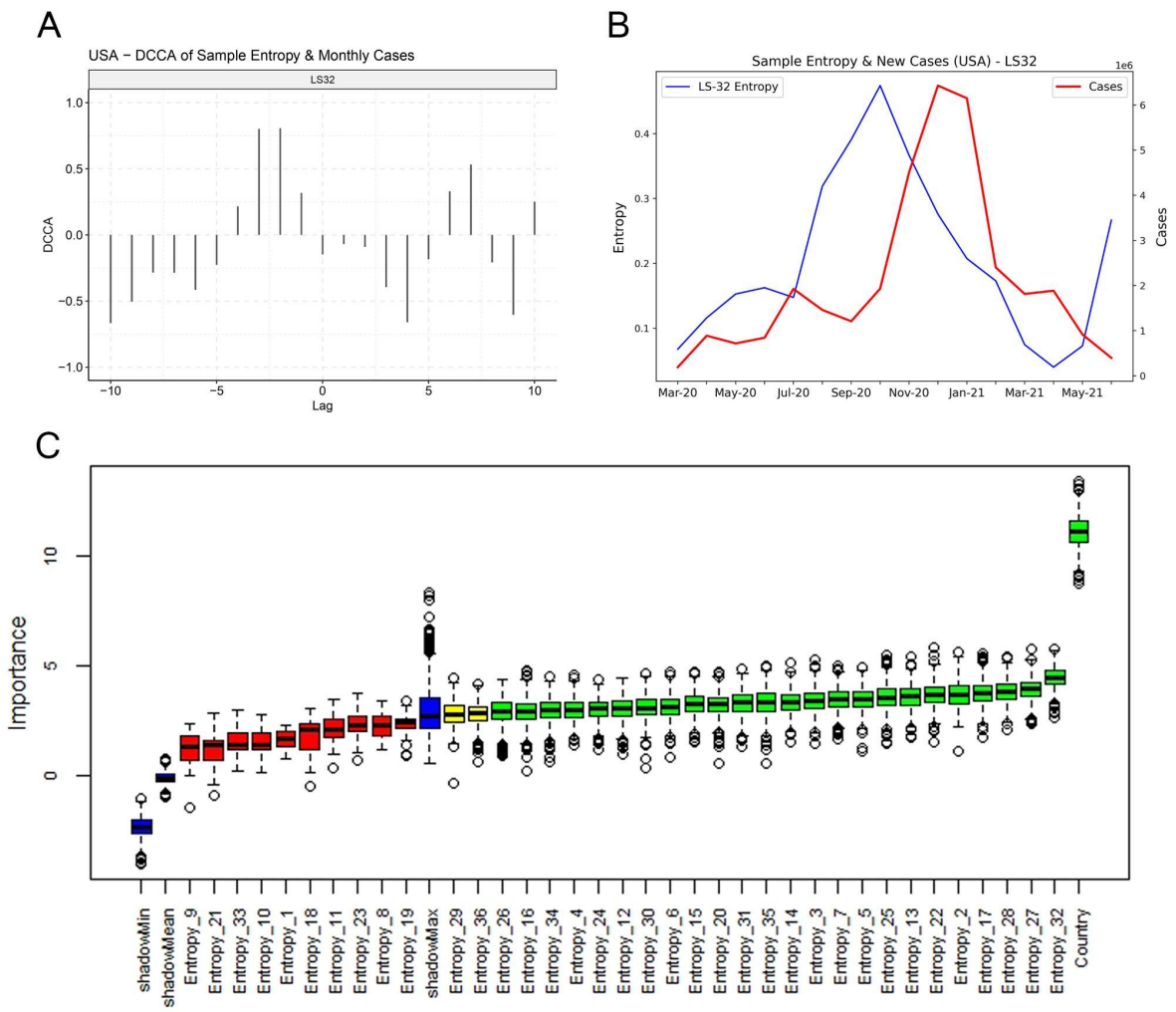


Figure 4: **Relationship of the entropy of latent space dimensions with COVID-19 caseloads.** (*A) Detrended Cross-correlation coefficient values for different lags between Entropy dimension 32 and new cases for USA. High values are observed for a lead of 1 and 2*

*months. (**B**) Line plot for Sample Entropy dimension 32 and monthly new cases for USA, indicating that the entropy in dimension 32 has a leading relationship with the cases. (**C**) Feature importance scores from the Boruta algorithm for predicting cases in the month following the next month.*

Country	Pearson Correlation	p value	Spearman's Correlation	p value
USA	0.97	8.41×10^{-9}	0.94	0.00
India	0.91	6.13×10^{-6}	0.97	0.00
Germany	0.91	6.78×10^{-6}	0.87	7.57×10^{-6}
France	0.86	7.35×10^{-5}	0.97	0.00
England	0.82	2.89×10^{-4}	0.66	1.22×10^{-2}
Japan	0.71	4.38×10^{-3}	0.63	1.92×10^{-2}
Brazil	0.48	8.61×10^{-2}	0.45	1.12×10^{-1}

Table 1: **Pearson and Spearman's Correlation coefficients between predicted and actual cases in different countries.**

Our model can be therefore used to predict the COVID-19 caseloads in several countries. Both USA (**Fig 5A**) and Japan (**Fig 5C**) show an increase in sample entropy in all the DoCs across the time period April - June 2021, concurrent with the respective spreads in these countries. Our model predicts new caseloads with a two-month lead time, which strongly predicts a spike in new cases both in USA (**Fig 5B**) and Japan (**Fig 5D**) in the months of July and August, 2021. Therefore our model may be used as an epidemiological early warning system to predict new caseloads.

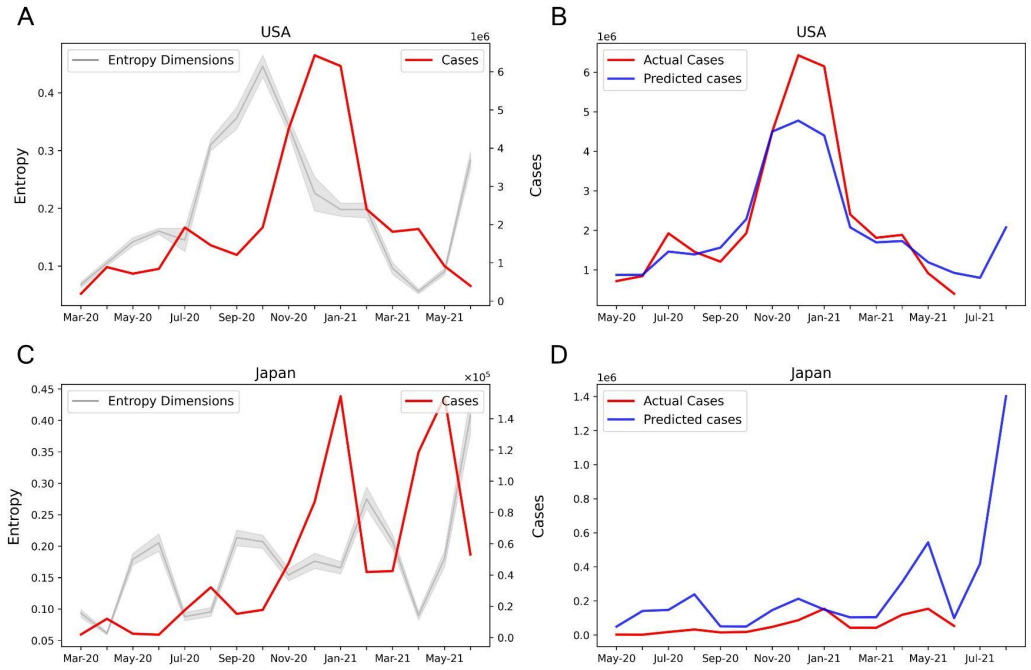


Figure 5: Prediction of new COVID-19 cases with Sample Entropy values of the latent

dimensions. *(**A**) Line plot showing the Entropy values of DoCs and new COVID-19 cases for the USA. (**B**) Actual and predicted cases based on the entropy values of DoCs for the USA. The model predicts a rise in cases for July and August 2021. (**C**) Entropy of DoCs and new cases for Japan. (**D**) Actual and predicted cases for Japan. A spike in cases is predicted for July and August 2021.*

Codons associated with Dimensions of Concerns (DoCs) could be linked to SARS-CoV-2 variants.

We further assessed the potential link of DoCs with SARS-COV-2 by extracting the top 10 contributing codons and their associated weights for each dimension (**Table S2**). The intuition behind this idea is that the codons with high weights in a given dimension, when mutated in the viral sequence, are likely to cause a significant change in the entropy of the associated dimensions. Therefore each DoC can be linked to codons, which can further be mapped to Variants of concern (VOCs) and Variants of Interest (VOIs) (**Table S3**). Despite the fact that our model cannot directly capture the SARS-CoV-2 variants, it was observed that dimension-32 captures the CTG, CGG codons (ranks 5 and 8 respectively), known to occur in the T19R variant. Similarly, dimension 3 captures three codons (ACG, CGG, CAC) that are associated with multiple variants such as K417T, L452R, and D1118H, causing increased infectivity, pathogenicity, and spread. Dimension 30 captures codons CAT and CAC associated with Δ69 and D1118H respectively which are linked to B.1.1.7 lineage.

Codon weights (**Table S2**) of a given DoC provide an opportunity to associate with specific Variants of concern (VOCs) and Variants of Interest (VOIs); and to predict emerging SARS-CoV-2 variants and take preventive measures in advance. Distinct dimensions capture country-specific changes and may be surveilled to monitor the spread of the pandemic. This approach was back-validated with several real-world examples. For instance, dimension 32 captures the codons CGG (R) and CAC (R), which are found in B.1.429 lineage (L425R mutation). Dimension 3 captures CGG which is seen in L452R (associated with lineage B.1.617.1), which was first observed in India in December 2020 and was found to have increased infectivity and transmissibility.

Temporal tracking of entropy in Dimensions of Concern (DoCs) may be used as a surrogate to track the spread of various SARS-CoV-2 variants.

To investigate the potential of DoCs to track the spread of SARS-CoV-2, we used the codon level information of the SARS-CoV-2 delta variant for the spike gene and extracted the weights of these codons specific to the DoCs. We selected DoCs (dimension 3, 4, 12, 13, 15, 16, 25, 28, 30, 32) with high absolute weights for the codons related to the delta variants (**Fig 6A**). The entropy of the DoCs was contrasted with the caseloads in England (**Fig 6B**), India (**Fig 6C**), and USA (**Fig 6D**). Overall, the temporal tracking of the entropy in Dimensions of Concern (DoCs) may be used as a surrogate to track the spread of various SAR-CoV-2 variants.

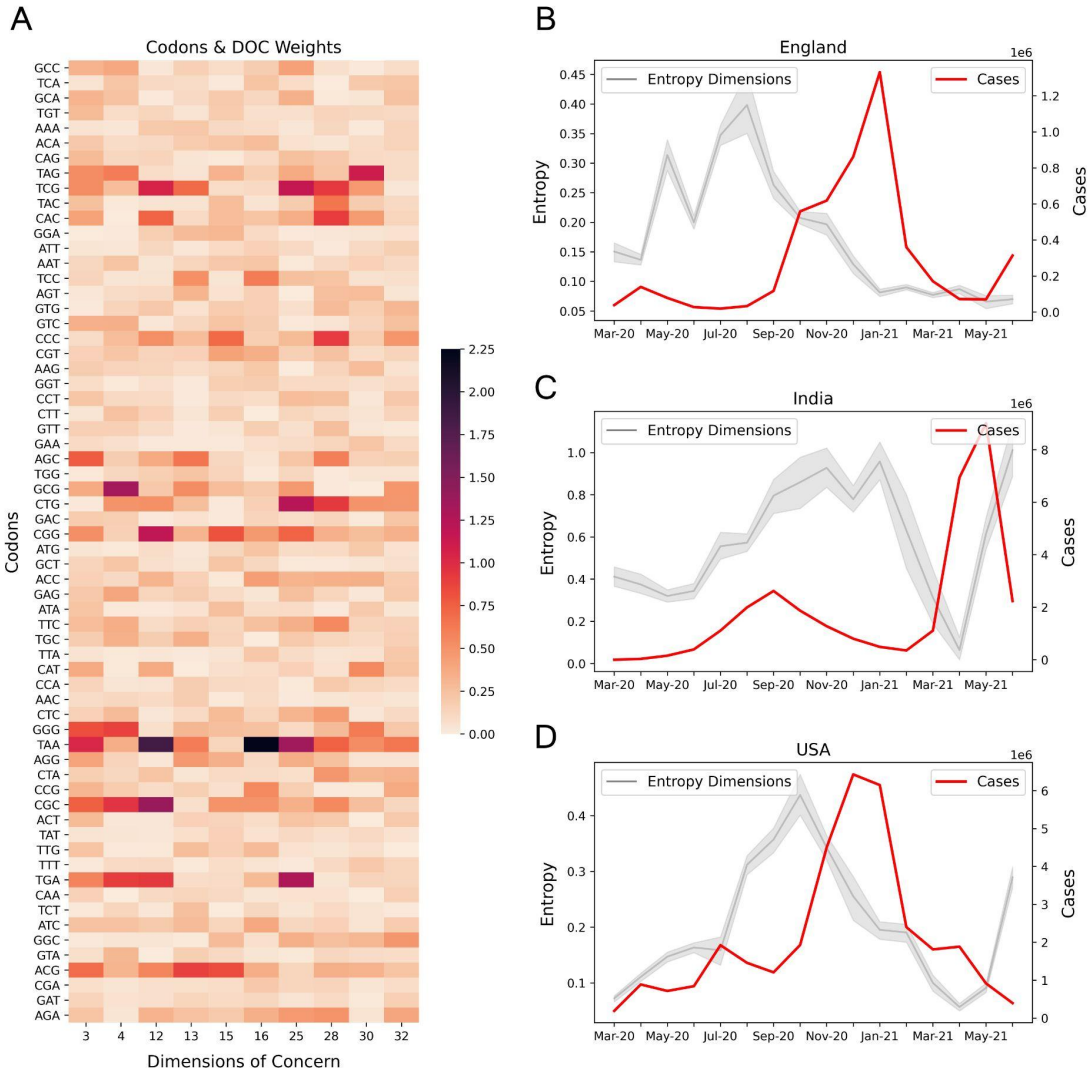


Fig 6. Potential association of codons observed in SARS-CoV-2 Delta variant (lineage B.1.617.2) with their corresponding DoCs, and the trend of caseloads with the entropy of the DoCs. *(**A**) Heat map showing the absolute latent space weights of the codons in the DoCs used for modeling. Line plots showing the entropies of DoCs and cases in countries, (**B**) England, (**C**) India, and (**D**) USA. The entropies of the DoCs show an increasing trend in the months April - June 2021 for India and USA, indicating a possible surge in the delta variant in these countries.*

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 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 (Mugnai et al., 2020) 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 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. The pipeline 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 (Shishir et al., 2021) is important to highlight the mechanistic understanding of mutations. SARS-CoV-2 seems to show a particularly high frequency of recombinations arising due to the absence of a proof-reading mechanism and sequence diversity, which calls for urgency in studying its transmission pattern (Mandal et al., 2021)(Rouchka et al., 2020). 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 (Srivastava et al., 2021)(Li et al., 2020).

Entropy is a measure of the disorder of a system. We hypothesized that mutations increase the chaotic dynamics in the latent space of spike genes. To calculate entropy, we used the accelerated versions of the Approximate Entropy and Sample Entropy algorithms, called Fast Approximate Entropy and Fast Sample Entropy (Pan et al., 2011). Both algorithms aim to quantify how often different patterns of data are found in a time series. Fast Approximate Entropy, however, is a biased statistic and depends on the length of the series. Since we could have different counts of genome sequences collected each month, we preferred Sample Entropy, which is independent of the length of the series. Entropy values were calculated for each latent dimension in each month. Thereafter, Detrended Cross-Correlation Analysis (DCCA) (Podobnik and Stanley, 2007) was performed between the entropy dimensions and the new cases. DCCA is a modification of the standard cross-correlation analysis for finding relationships between non-stationary time series. High cross-correlation for different lead periods revealed that the entropy values in a given month could be used to predict the new cases in different countries in subsequent months. Different countries had different lead times at which the highest cross-correlation was observed between the entropy dimensions and the cases, ranging from 1-6 months. Overall, a lead time of two months was chosen to model the new cases.

A similar analysis was done with daily values of entropy and new cases. Entropy was calculated in rolling windows, and cross-correlation analysis was performed between entropy and new cases at different lead periods. Although the cross-correlation values were found to be significant, the values were low and ranged between -0.1 to 0.1. Therefore, we decided to use the monthly entropy values for the modeling exercise.

We also experimented with “blips” as a feature to predict the new cases. Blips are sudden changes in the values of the latent dimensions. These changes may be caused by a mutation, which changes the words (codons) in a given genome sequence. This hypothesis was validated in simulation experiments in synthetic datasets. Each dimension of the spike gene

embeddings for a country was analyzed 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. Similar to entropy, blips were found to have a leading relationship with the cases. However, regression modeling results with sample entropy were found to be better.

To predict new COVID-19 cases, a Random Forest regression model was trained on the monthly entropy data. With sample entropy, the R-squared value achieved was 73%, while with approximate entropy, the value was only 10%. Therefore the model trained on sample entropy was selected. The predictions from the model were found to be highly correlated with the actual cases, indicating that our model can be used for preemptive warning signals for the rise in cases in different countries. Further, the actual and the predicted difference in the number of cases in consecutive months was found to be correlated, which suggests that the relative change in the cases in consecutive months predicted by our model is linked to the relative change in the number of cases. Overall, we recommend that our model be used to predict dangerous trends and not the actual number of cases. Further, the mapping from latent dimensions to Variants of Concern (VOCs) and Variants of Interest (VOIs) may help us track the country-specific spread of different variants.

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 (Wang et al., 2020). The machine learning models can also be trained on genomic sequences for COVID-19 classification (Arslan, 2021). 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. 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 events in the future will require stronger surveillance systems in place. Our study offers a promising solution in this direction and lays the foundation for proactive genomic surveillance of COVID-19.

Limitations of our study

Our approach of codon (3-mers) embeddings does not indicate the position where the codon change may have happened in the spike gene. This is because low-dimensional embeddings do not preserve the positional encoding of words. However, we are investigating advanced approaches such as complex-valued word embeddings with positional encodings (Wang et al., 2019) and transformer models such as BERT (Wolf et al., 2020)(Lee et al., 2020) to overcome our current limitation. 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.

Further, our Strainflow model is trained only on the spike gene of the viral genome, which does not represent the complete variation spectrum of the virus. To mitigate this shortcoming, we are currently developing a genome-level Strainflow model (version 2.0) for SARS-CoV-2.

Another limitation of our study is the relatively small number of samples that were used to construct the supervised predictive model for case prediction. As more data becomes available in subsequent months, we can produce more confident case predictions. Nonetheless, 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. However, 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 underscores 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 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.

Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Data		
SARS-CoV-2 genome sequences	GISAID	https://www.gisaid.org/
COVID-19 cases	GitHub	https://github.com/CSSEGISandData/COVID-19
Software and algorithms		
CoV-Seq	GitHub	https://github.com/boxiangliu/covseq
Flt-SNE (version 1.2.1)	GitHub	https://github.com/KlugerLab/Flt-SNE
Anaconda (version 4.10.1)	anaconda.com	https://www.anaconda.com/
Python (version 3.8.5)	python.org	https://www.python.org/downloads/release/python-385/
biopython (version 1.78)	PyPI	https://pypi.org/project/biopython/1.78/
gensim (version 4.0.1)	PyPI	https://pypi.org/project/gensim/4.0.1/
numpy (version 1.19.2)	PyPI	https://pypi.org/project/numpy/1.19.2/
pandas (version 1.1.3)	PyPI	https://pypi.org/project/pandas/1.1.3/
matplotlib (version 3.3.2)	PyPI	https://pypi.org/project/matplotlib/3.3.2/
seaborn (version 0.11.0)	PyPI	https://pypi.org/project/seaborn/0.11.0/
plotly (version 4.14.3)	PyPI	https://pypi.org/project/plotly/4.14.3/
R version (4.1.0)	CRAN	https://cran.r-project.org/
lubridate (version 1.7.0)	CRAN	https://cran.r-project.org/web/packages/lubridate/
dplyr (version 1.0.7)	CRAN	https://cran.r-project.org/web/packages/dplyr/
zoo (version 1.8.9)	CRAN	https://cran.r-project.org/web/packages/zoo/
tseries (version 0.10.48)	CRAN	https://cran.r-project.org/web/packages/tseries/
plyr (version 1.8.6)	CRAN	https://cran.r-project.org/web/packages/plyr/
reshape2 (version 1.4.4)	CRAN	https://cran.r-project.org/web/packages/reshape2/
ggplot2 (version 3.3.5)	CRAN	https://cran.r-project.org/web/packages/ggplot2/
ggpubr (version 0.4.0)	CRAN	https://cran.r-project.org/web/packages/ggpubr/
TSEntropies (version 0.9)	CRAN	https://cran.r-project.org/web/packages/TSEntropies/
DCCA (version 0.1.1)	CRAN	https://cran.r-project.org/web/packages/DCCA/
Boruta (version 7.0.0)	CRAN	https://cran.r-project.org/web/packages/Boruta/
randomForest (version 4.6.14)	CRAN	https://cran.r-project.org/web/packages/randomForest/
ape (version 5.5.0)	CRAN	https://cran.r-project.org/web/packages/ape/
iTOL (version 6.3)	Online	https://itol.embl.de/

Datasets.

Training dataset: The dataset was downloaded from GISAID EpiCoV (Shu and McCauley, 2017) (April 8, 2021 release). 0.63 million 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 (Liu et al., 2020) to extract nucleotide sequences corresponding to each of the 12 CDS. Accession IDs that did not cover 12 coding regions were discarded, yielding 0.31 million high-quality SARS-CoV-2 genome sequences for language modelling. We downloaded country-wise COVID-19 data for new cases from a publicly available repository maintained by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

Test dataset: We downloaded around 0.6 million genome sequences submitted to GISAID from April 2021 to June 2021. We have used 0.36 million sequences to predict their latent space using the *Strainflow* model for the months of April 2021 to June 2021.

Word Embeddings in Strainflow model.

In our Strainflow model, we have adopted a word2vec model (Mikolov et al., 2013). A low dimensional representation for the genome sequences was learned using the word2vec model. Non-overlapping sequences of 3-mers (codons) were considered as words for training the word2vec model, implemented in Gensim (Radim Rehurek, 2010). The word2vec model was trained using the skip-gram algorithm, with a 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 codon 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 codons present in the sequence. For selecting the vector size of the word2vec model, we calculated PIP (Pairwise Inner Product) loss (Yin and Shen, 2018). PIP loss is a metric used for calculating the dissimilarity between two word embedding matrices. For the embedding matrix of strains (E), the PIP matrix is defined as the dot product of the embedding matrix with its transpose ($E.E^{T}$). The PIP loss between two embedding matrices is defined as the norm of the 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 the word embeddings (**Fig. S1**).

Phylogenetic analysis using the latent dimensions of the spike genes

To evaluate the phylogenetic properties based on the latent dimensions of the spike gene, we computed the cosine distances among spike genes of SARS-CoV-2 with the 36 latent dimensions. The pairwise distance was further used for hierarchical clustering using the 'hclust' function in R statistical programming language. This analysis was performed using 400 random sequences of spike genes from 16 countries. The visualization of the phylogenetic tree derived based on the latent dimensions was done using 'iTOL' (Letunic) software (**Fig 2**).

Entropy of the latent dimensions

To quantify the properties of latent dimensions, we have used a well-known information theory based algorithm suitable for time series datasets, called 'Fast Sample Entropy' (Tomčala, 2020). To compute Fast Sample Entropy, we have used the 'FastSampEn' function in the 'TSEntropies' package in R (Tomcala, 2018). Fast Sample Entropy can be computed as follows.

$$FastSampEn(x, m, r) = log(\sum_{i=1}^{N_m} \left| s_{i,m} \right| \mid \sum_{i=1}^{N_{m+1}} \left| s_{i,m+1} \right|),$$

where,

$s_{i,m} = \{\xi \mid (\|y_i - y_\xi\| \leq r, \xi \neq i) \wedge (\xi \notin s_{j,m}, j < i)\}$, $y_i = [x_i, x_{i+1}, \ldots, x_{i+m-1}]$

$s_{i,m}$ is a set of sub-sequences of length m belonging to the i-th neighborhood, and N_m is the number of these neighborhoods.

In our case, 'x' is the latent dimension of the spike genes of the SARS-CoV-2 strains per month for a given country with default values of 'm' and 'r'. Entropy was computed for each latent dimension on a monthly basis for each country. The minimum number of SARS-CoV-2 spike genes was 20. To compare across months, we used average entropy derived from 36 latent dimensions, followed by normalization using all the monthly entropies for a given country (**Fig 1D**).

To compare the entropy of the latent dimensions among countries, we used the total entropy of the country for each dimension and visualized it with line graphs (**Fig 3**).

Detrended Cross Correlations Analysis (DCCA)

To investigate the information content (entropy) of the latent dimensions with the new cases observed for COVID-19, we used the Detrended Cross Correlation Analysis (Prass and Pumi, 2019). Here, DCCA captures the long-range cross correlation between time series (entropy of the months and caseloads for a given country). We tested both time series for stationarity using Augmented Dickey-Fuller (ADF) test (Mushtaq, 2011). The ADF test was implemented using the function 'adf.test' available in the 'tseries' (2020a) package in R. Due to the

non-stationary distribution of the estimated entropies and the caseloads for a given country, we used the 'DCCA' R package (2020b). Cross-correlation was calculated between the entropy dimensions at time t+h and new cases at time t, where h = 0, ±1, ±2, ±3 … ±10.

Machine Learning based identification of 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 entropy dimensions data from March, 2020 to June, 2021. We used a regression based machine learning approach called 'Boruta' (Kursa and Rudnicki, 2010), a wrapper algorithm around random forest algorithm (Kursa et al., 2010) to select the most relevant entropy dimensions for the prediction of subsequent two months' new cases. We used the default parameters with the modification of the maximum runs as 1000. We selected the confirmed entropy dimensions as the most relevant features for the prediction of new cases.

Model development and evaluation for prediction of new cases in subsequent months

To predict the new cases in the next to next months, we used a regression based random forest model (Breiman 2001) using the most relevant DoCs based on Boruta. The model training was performed using the data from March, 2020 to April, 2021; and the fitted model was tested on data from May, 2021 to June, 2021. The regression modeling was performed using 1000 decision trees using the 'randomForest' package in R (A. Liaw et al., 2002).

Top codons associated with DoCs

To find the top codons associated with the DoCs, we extracted the absolute weights of each codon for a given DoC. The top 10 codons having the highest absolute weights (contribution) were identified corresponding to each of the DoCs to compare our DoCs with SARS-CoV-2 variants. We collected the SARS-CoV-2 variants and their associated genetic variations at the codon level linked to the spike gene (CDC, 2021); and a list of codons associated with VOIs and VOCs was curated **(Table S4)** (Lopez-Rincon et al.; paola, 2021; Peacock et al., 2021; Srivastava et al., 2021). The curated list is based on the CDC guidelines, and we are consistent with their definition of lineage and variant (CDC, 2021)

Acknowledgements

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, Nishkarsh Saxena, and Anjali for their contribution to the dashboard visualizations.

References

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

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

CDC (2021). SARS-CoV-2 Variant Classifications and Definitions.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

Hie, B., Zhong, E.D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. Science *371*, 284–288.

Khan, M.I., Khan, Z.A., Baig, M.H., Ahmad, I., Farouk, A.-E., Song, Y.G., and Dong, J.-J. (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.

Kursa, M.B., and Rudnicki, W.R. (2010). Feature Selection with theBorutaPackage. Journal of Statistical Software *36*.

Kursa, M.B., Jankowski, A., and Rudnicki, W.R. (2010). Boruta - A System for Feature Selection. Fund. Inform. *101*, 271–285.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240.

Letunic, I. iTOL: Interactive Tree Of Life.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., et al. (2020). The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell *182*, 1284–1294.e9.

Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nat. Methods *16*, 243–245.

Liu, B., Liu, K., Zhang, H., Zhang, L., Bian, Y., and Huang, L. (2020). CoV-Seq, a New Tool for SARS-CoV-2 Genome Analysis and Visualization: Development and Usability Study. J. Med. Internet Res. *22*, e22299.

Lopez-Rincon, A., Perez-Romero, C.A., Tonda, A., Mendoza-Maldonado, L., Claassen, E., Garssen, J., and Kraneveld, A.D. Design of Specific Primer Sets for the Detection of B.1.1.7, B.1.351, P.1, B.1.617.2 and B.1.1.519 Variants of SARS-CoV-2 using Artificial Intelligence.

Mandal, S., Roychowdhury, T., and Bhattacharya, A. (2021). Pattern of genomic variation in SARS-CoV-2 (COVID-19) suggests restricted nonrandom changes: Analysis using Shewhart control charts. J. Biosci. *46*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Mugnai, M.L., Templeton, C., Elber, R., and Thirumalai, D. (2020). Role of Long-range Allosteric Communication in Determining the Stability and Disassembly of SARS-COV-2 in Complex with ACE2. bioRxiv.

Mushtaq, R. (2011). Augmented Dickey Fuller Test. SSRN Electronic Journal.

Pan, Y.H., Wang, Y.H., Liang, S.F., and Lee, K.T. (2011). Fast computation of sample entropy and approximate entropy in biomedicine. Comput. Methods Programs Biomed. *104*.

paola (2021). Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein.

Peacock, T.P., Penrice-Randal, R., Hiscox, J.A., and Barclay, W.S. (2021). SARS-CoV-2 one year on: evidence for ongoing viral adaptation. J. Gen. Virol. *102*.

Podobnik, B., and Stanley, H.E. (2007). Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Non-stationary Time Series.

Prass, T.S., and Pumi, G. (2019). On the behavior of the DFA and DCCA in trend-stationary processes.

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,.

Rouchka, E.C., Chariker, J.H., and Chung, D. (2020). Variant analysis of 1,040 SARS-CoV-2 genomes. PLoS One *15*, e0241535.

Shishir, T.A., Naser, I.B., and Faruque, S.M. (2021). In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLoS One *16*, e0245584.

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. *22*.

Srivastava, S., Banu, S., Singh, P., Sowpati, D.T., and Mishra, R.K. (2021). SARS-CoV-2 genomics: An Indian perspective on sequencing viral variants. J. Biosci. *46*.

Tomcala, J. (2018). Time Series Entropies [R package TSEntropies version 0.9].

Tomčala, J. (2020). New Fast ApEn and SampEn Entropy Algorithms Implementation and Their Application to Supercomputer Power Consumption. Entropy *22*, 863.

Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., and Simonsen, J.G. (2019). Encoding word order in complex embeddings.

Wang, R.Y., Guo, T.Q., Li, L.G., Jiao, J.Y., and Wang, L.Y. (2020). 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).

WHO Coronavirus (COVID-19) Dashboard.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

Wu, F., Yang, R., Zhang, C., and Zhang, L. (2021). A deep learning framework combined with word embedding to identify DNA replication origins. Sci. Rep. *11*, 844.

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

Yin, Z., and Shen, Y. (2018). On the Dimensionality of Word Embedding. Adv. Neural Inf. Process. Syst. *31*.

(2020a). Time Series Analysis and Computational Finance [R package tseries version 0.10-48].

(2020b). Detrended Fluctuation and Detrended Cross-Correlation Analysis [R package DCCA version 0.1.1].