Title: Language model analysis of SARS-CoV-2 Spike from UK B.1.1.7 and South African V501.V2 lineages

Authors: Brian Hie[1,2], Bryan Bryson[2,3]*, Bonnie Berger[1,4]*

Affiliations: [1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [2]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA; [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. *Correspondence: B. Bryson (bryand@mit.edu), B. Berger (bab@mit.edu).

Summary: We used a neural language model, part of a framework called constrained semantic change search (CSCS) that we developed for viral escape prediction, to predict semantic change of full-length Spike sequences from the B.1.1.7 lineage (originally observed in the United Kingdom) and the V501.V2 lineage (originally observed in South Africa) compared to a background set of 911 Spike sequences from SARS-CoV-2 surveillance and to PT188-EM, a laboratory-generated mutant that reduces sensitivity to polyclonal neutralizing sera. The model predicts high semantic change for both B.1.1.7 and V501.V2 compared to the background set of surveilled sequences, but lower semantic change than that of PT188-EM. Analyzing the effect of individual mutations within these new mutant lineages, our model highlights high semantic change associated with a P681H substitution to the furin cleavage site, a A701V substitution increasingly observed in Malaysia and Singapore, and insertions and deletions to the N-terminal domain, suggesting these and similar variants should be monitored for escape. Our model and analysis can be applied readily to future Spike mutants from surveillance efforts.

Results

The recent emergence of SARS-CoV-2 lineages B.1.1.7 (Rambaut et al., 2021) and V501.V2 (Pond et al., 2021; Tegally et al., 2020) has generated concern over the effect of these mutations on potential escape from human immunity to previous versions of SARS-CoV-2 (Andreano et al., 2021; Garry, 2021; Greaney et al., 2020, 2021), particularly since these lineages contain mutations to the Spike surface protein that is targeted by antibodies (Greaney et al., 2020). In previous work (Hie et al., 2021), we developed a computational model of viral evolution, based on a machine learning algorithm called a *language model*, that learns patterns of antigenicity and viral fitness from sequence data alone, which we used to predict escape mutations to SARS-CoV-2 Spike as well as influenza A hemagglutinin and HIV-1 envelope glycoprotein with higher accuracy than previous machine learning approaches to escape prediction. In particular, we provided evidence that the "semantic change" learned by the language model could approximate antigenic change of viral surface proteins, since high semantic change correlates with greater functional change, which, along with viral fitness, contributes to viral escape (Hie et al., 2021).

We therefore used our language model to quantify the semantic change of B.1.1.7 and V501.V2 Spike sequences relative to wildtype Spike (originally observed in Wuhan, China); see **Methods** for more details. For comparison, semantic change with respect to wildtype was also computed across a genetic background composed of 911 unique, surveilled Spike sequences. As a positive control, we also quantified the semantic change of a laboratory-generated mutant sequence, PT188-EM (Andreano et al., 2021). Importantly, while our CSCS model considers both predicted semantic change and predicted viral fitness ("grammaticality"), we largely focus

on semantic change in this analysis because all viral sequences that we analyze are guaranteed to be infectious and are therefore "grammatical."

B.1.1.7, a lineage originally observed in the United Kingdom with a total of eight mutations (H69del/V70del, Y145del, N501Y, A570D, P681H, T716I, S982A, and D118H), had a semantic change percentile rank of 98.8, with 11 background sequences that have higher predicted semantic change. For V501.V2, a lineage originally observed in South Africa with a total of six mutations (D80A, D215G, K417N, E484K, N501Y, and A701V), the model predicts a semantic change with a percentile rank of 99.5 and 5 background sequences with higher predicted semantic change. The semantic change of PT188-EM with a total of three mutations (F140del, E484K, and insertion $_{248a}$KTRNKSTSRRE$_{248k}$) has a percentile rank of 99.9 with a single background sequence that had higher predicted semantic change.

Consistent with all three of these strains preserving infectivity, the grammaticality scores of these three sequences were within the range observed among the background set. V501.V2 had the highest grammaticality (19.5 percentile rank) followed by B.1.1.7 (9.7 percentile rank) and PT188-EM (4.8 percentile rank).

Our model predictions indicate potentially altered antigenicity of all three sequences-of-interest (**Figure 1**). Our model's predictions are consistent with a number of results from recent phylogenetic and *in vitro* laboratory analyses of B.1.1.7 and V501.V2. First, we predict antigenic drift associated with both lineages, a finding consistent with phylogenetic analysis (Pond et al., 2021; Rambaut et al., 2021). Interestingly, we predict less semantic change-based escape potential of B.1.1.7 (98.8 percentile rank of semantic change) compared to V501.V2 (99.5 percentile rank), consistent with deep mutational scan-based escape maps of yeast-displayed receptor binding domain (RBD) also show that B.1.1.7 RBD mutants have less escape potential

than V501.V2 with respect to polyclonal sera (Greaney et al., 2021). We also predict less escape potential among the two naturally circulating variants compared to the PT188-EM, which was artificially selected for polyclonal escape potential *in vitro* (Andreano et al., 2021).

We can also estimate the individual contributions of mutations to semantic change for each of B.1.1.7, V501.V2, and PT188-EM. We quantified the semantic change relative to wildtype Spike for the sequences made by each single mutation, which are summarized in **Tables 1-3**. For B.1.1.7, the mutation resulting in the highest semantic change was P681H, a substitution in the furin cleavage site and is part of a four-residue insertion unique to SARS-CoV-2. This region has been implicated in increased transmission from previous coronaviruses (Hoffmann et al., 2020; Peacock et al., 2020) (**Table 1**). The semantic change associated with P681H alone reaches a percentile rank of 48.7 compared to the background sequences.

For V501.V2, the variant with the highest individual semantic change is A701V (**Table 2**). A701V is of growing concern in viral surveillance efforts and has reportedly increased to 85% prevalence among sequences reported by the Malaysian Ministry of Health (Abdullah, 2020; Harun, 2020). Our model estimates that the semantic change associated with A701V alone reaches a percentile rank of 50.1 compared to the background sequences.

For PT188-EM, the highest semantic change is due to the 11-residue insertion between Y248 and L249, consistent with experimental data showing rapid fixation of the mutation resulting in complete abrogation of PT188 plasma sample neutralization (**Table 3**). With this insertion alone, semantic change reaches a percentile rank of 99.2. However, we also observed very high semantic change associated with the F140 deletion (percentile rank of 78.7). Interestingly, the lowest semantic change is predicted for E484K (percentile rank of 25.1), which has been shown to reduce sensitivity to polyclonal neutralizing sera (Andreano et al., 2021;

Greaney et al., 2021). The E484K prediction provides some opportunities to learn more about that specific RBD mutation.

Discussion

Our model highlights B.1.1.7 and V501.V2 as outliers in terms of semantic change and therefore potentially altered functional profiles as well, but our model assigns greater semantic change-based escape potential to V501.V2. However, both mutant lineages still do not have as high predicted escape potential as PT188-EM, which has demonstrated resistance to polyclonal neutralization by a number of laboratory sera samples (Andreano et al., 2021). PT188-EM, and other experimentally-verified escape strains obtained in future studies, could be used as a "threshold" sequence to calibrate our model's predictions, enabling rapid identification of surveilled viruses that have high escape potential. While our model does provide an ordering of predicted semantic change, additional experiments are needed to profile the escape potential of all three of these sequences of interest.

Our model highlights insertion $_{248a}$KTRNKSTSRRE$_{248k}$, F140del, A701V, and P681H as mutational variants that have large individual escape potential, all of which are outside of the RBD. These mutations (and similar variants, such as modifications to the N3 or N5 N-terminal domain loops) should continue to be monitored via viral surveillance efforts and further probed by biochemical and *in vivo* experiments. In particular, A701V is of growing concern due to its observation in South Africa and Malaysia (Abdullah, 2020; Harun, 2020; Pond et al., 2021; Tegally et al., 2020). Lastly, and importantly, these computational predictions provide only one perspective into viral evolution and should be considered in the context of alternative epidemiological and experimental data that may provide additional insight into SARS-CoV-2 evolution and escape.

Methods

Data and code availability

Code, scripts for plotting and visualizing, and pretrained models are available at https://github.com/brianhie/viral-mutation/tree/v501. We used the following publicly available datasets for model training:

- *Coronavidae* spike protein sequences from the Virus Pathogen Resource (ViPR) database (https://www.viprbrc.org/brc/home.spg?decorator=corona)
- SARS-CoV-2 Spike protein sequences from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/)
- SARS-CoV-2 Spike and other Betacoronavirus spike protein sequences from GISAID (https://www.gisaid.org/)

Representative mutations for the B.1.1.7 and V501.V2 lineages were obtained from reference (Garry, 2021).

Model training

We used pretrained models where the training procedure was described by Hie et al. (2021). Briefly, *Coronavidae* spike glycoprotein sequences were obtained from the Gene/Protein Search portal of the ViPR database (https://www.viprbrc.org/brc/home.spg?decorator=corona) across the entire *Coronavidae* family. We only included amino acid sequences with "spike" gene products. SARS-CoV-2 Spike sequences were obtained from the Severe acute respiratory syndrome coronavirus 2 datahub at NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/). Betacoronavirus spike sequences from GISAID also used in the analysis by Starr et al. (2020) were obtained from https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS/blob/master/data/alignments/Spike_GISAID/spike_GISAID_aligned.fasta. Across

all coronavirus datasets, we furthermore excluded sequences with a protein sequence length of less than 1,000 amino acid residues. We trained an amino acid residue-level language model on a total of 4,172 unique Spike (and homologous protein) sequences.

Background sequences and semantic change computation

Among sequences in the training corpus, we used all complete, unique SARS-CoV-2 Spike sequences, leading to 911 unique sequences. We computed the semantic change relative to the GISAID wildtype Spike sequence. Semantic change computation is described in detail by Hie et al. (2021); briefly, the output of the final hidden layer was computed for each sequence to obtain a sequence-length-by-embedding dimension matrix, this matrix was averaged across the sequence dimension to obtain fixed-length embeddings for each sequence, and semantic change was computed as the Euclidean distance within this embedding space. Embedding computation code is at https://github.com/brianhie/viral-mutation/blob/a54c297cb4aecc95d03c32c7a2c2c5b556307490/bin/cov.py#L312.

Mutant lineage analysis

Semantic change was also computed for B.1.1.7, V501.V2, and PT188-EM Spike sequences. The semantic change value for each of these sequences was compared to the semantic change values across the background sequences to obtain a percentile rank score. Percentile rank scores are computed at https://github.com/brianhie/viral-mutation/blob/a54c297cb4aecc95d03c32c7a2c2c5b556307490/bin/cov.py#L366. We also computed these percentile rank scores for each of the individual mutations to obtain the values in **Tables 1-3**. These scores are computed at https://github.com/brianhie/viral-mutation/blob/a54c297cb4aecc95d03c32c7a2c2c5b556307490/bin/cov.py#L394.

To generate **Figure** 1, the distribution of the background set semantic changes was visualized as a strip plot with random jitter along the *x*-axis and semantic change on the *y*-axis, as well as a violin plot of the same distribution; these were implemented using the seaborn Python package (version 0.11.1). On the same plot, the semantic change values for B.1.1.7, V501.V2, and PT188-EM were also plotted for comparison. Code for producing this plot is at https://github.com/brianhie/viral-mutation/blob/a54c297cb4aecc95d03c32c7a2c2c5b556307490/bin/cov.py#L397.

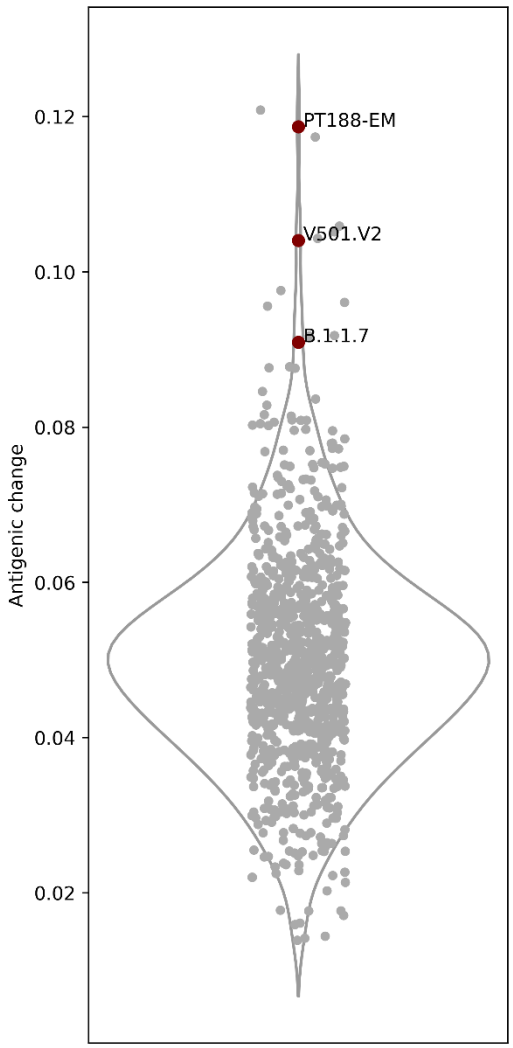

Figure 1: Predicted semantic change of mutant SARS-CoV-2 Spike sequences.

A background set of Spike sequences is plotted as gray circles with random jitter on the *x*-axis and language model-quantified semantic change on the *y*-axis. Mutants-of-interest (B.1.1.7, V501.V2, and PT188-EM) are plotted on the same axes in red. B.1.1.7, V501.V2, and PT188-EM all have outlier escape potential (98.8, 99.5, and 99.9 percentile ranks, respectively) and suggest the highest escape potential for PT188-EM, which has been experimentally confirmed to induce escape against polyclonal sera from multiple patients (Andreano et al., 2021).

Mutation	Semantic change, percentile rank
H69del	6.81%
V70del	20.86%
Y145del	13.61%
N501Y	11.20%
A570D	14.38%
P681H	48.74%
T716I	33.15%
S982A	14.38%
D1118H	4.06%

Table 1: Semantic change of constituent mutations of B.1.1.7.

The highest semantic change is associated with P681H, a mutation in the furin cleavage site and part of a 4-residue sequence that is unique to SARS-CoV-2.

Mutation	Semantic change, percentile rank
D80A	30.68%
D215G	31.83%
K417N	9.33%
E484K	25.14%
N501Y	11.20%
A701V	50.16%

Table 2: Semantic change of constituent mutations of V501.V2.

The highest semantic change is associated with A701V, which is of increasing concern in viral surveillance in Malaysia and Singapore.

Mutation	Semantic change, percentile rank
F140del	78.70%
E484K	25.14%
$_{248a}$KTRNKSTSRRE$_{248k}$	99.23%

Table 3: Semantic change of constituent mutations of PT188-EM.

Two out of three mutations have high semantic change: F140del, a deletion in the N3 loop of Spike, and insertion $_{248a}$KTRNKSTSRRE$_{248k}$ in the N5 loop of Spike. Interestingly, E484K, which has been implicated in mutational escape *in vitro* (Andreano et al., 2021; Greaney et al., 2021), has the lowest predicted semantic change.

References

Abdullah, N.H. (2020). The current situation and Information on the Spike protein mutation of Covid-19 in Malaysia. Facebook post. https://www.facebook.com/DGHisham/posts/the-current-situation-and-information-on-the-spike-protein-mutation-of-covid-19-/3948439941846531/.

Andreano, E., Piccini, G., Licastro, D., Casalino, L., Johnson, N. V., Paciello, I., Monego, S.D., Pantano, E., Manganaro, N., Manenti, A., et al. (2021). SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma. bioRxiv doi:10.1101/2020.12.28.424451.

Garry, R.F. (2021). Mutations arising in SARS-CoV-2 spike on sustained human-to-human transmission and human-to-animal passage. Virological https://virological.org/t/mutations-arising-in-sars-cov-2-spike-on-sustained-human-to-human-transmission-and-human-to-animal-passage/578.

Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H.D., et al. (2020). Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. Cell Host Microbe. In press.

Greaney, A.J., Loes, A.N., Crawford, K.H., Starr, T.N., Malone, K.D., Chu, H.Y., and Bloom, J.D. (2021). Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. bioRxiv doi:10.1101/2020.12.31.425021.

Harun, H.N. (2020). Dr Noor Hisham: UK Covid-19 mutation not detected in Malaysia. New Straits Times.

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

Hoffmann, M., Kleine-Weber, H., and Pöhlmann, S. (2020). A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol. Cell *78*, 779–784.e5.

Peacock, T.P., Goldhill, D.H., Zhou, J., Baillon, L., Frise, R., Swann, O.C., Kugathasan, R., Penn, R., Brown, J.C., Sanchez-David, R.Y., et al. (2020). The furin cleavage site of SARS-CoV-2 spike protein is a key determinant for transmission due to enhanced replication in airway cells. bioRxiv doi:10.1101/2020.09.30.318311.

Pond, S.L.K., Wilkison, E., Weaver, S., James, S.E., Tegally, H., Oliveira, T. de, and Martin, D. (2021). A preliminary selection analysis of the South African V501.V2 SARS-CoV-2 clade. Virological https://virological.org/t/a-preliminary-selection-analysis-of-the-south-african-v501-v2-sars-cov-2-clade/573.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D.L., and Volz, E. (2021). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

Starr, T. N., Greaney, A.J., Hilton, S. K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M. A., Walls, A.C., et al. (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell, *182*, 1295-1310.e20.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh,

D., Pillay, S., San, E.J., Msomi, N., et al. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv doi:10.1101/2020.12.21.20248640.