Postscript to "Learning the language of viral evolution and escape"

Authors: Brian Hie[1], Ellen Zhong[2,3], Bonnie Berger[2,4]*, and Bryan Bryson[5,6,]*

Affiliations: [1]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305; [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [3]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [6]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA.

*Corresponding author. Email: bab@mit.edu (B. Berger); bryand@mit.edu (B. Bryson)

Main Text

In our original publication, we previously showed how a language model trained on sequence data alone could predict escape mutations as those that have the highest "grammaticality" (quantified by language model likelihood) and the highest "semantic change" (quantified by distance in language model embedding space) [1]. Since publication, we have performed additional experiments further supporting our hypothesis that grammaticality corresponds to notions of biological fitness and that semantic change corresponds to the ability to evade immunity. We also use our techniques to characterize full-length Spike sequences from the B.1.1.7 lineage (originally observed in the United Kingdom) and the B.1.351 lineage (originally observed in South Africa) compared to a background set of 911 Spike sequences from SARS-CoV-2 surveillance as well as PT188-EM, a laboratory-generated mutant that reduces sensitivity to polyclonal neutralizing sera.

Interpretation of grammaticality and semantic change via cutoff sensitivity experiments

We previously used a bidirectional long short-term memory (BiLSTM) neural network trained on coronavirus sequences to perform what we termed a constrained semantic change search (CSCS), with the goal of predicting escape mutations to the SARS-CoV-2 Spike receptor binding domain (RBD). As validation, we used deep mutational scan (DMS) experiments in which all single-amino acid mutations were made to a yeast-displayed RBD system (ydRBD) to quantify the effect of mutations on protein expression, ACE2 binding [2], and binding to neutralizing antibodies [3]. We use CSCS to rank mutations and quantify enrichment of CSCS-acquired escape mutations using the area-under-the-curve (AUC) of the number of acquired escapes versus the total number of acquired mutants, normalized to be between 0 and 1, and observed a statistically significant and strong enrichment of escape mutants using our CSCS model.

As a follow up experiment, we sought to analyze how our ability to predict escape was dependent on the *relative* strength of antibody binding, ACE2 binding, or protein expression. We therefore varied several cutoffs used to define escape mutations and quantified the effect on the AUC prediction score. First, when we relaxed the antibody escape fraction cutoff from our originally reported cutoff of 0.3 to zero, the lowest value, we observed that more relaxed cutoffs resulted in a reduction of the overall CSCS AUC (**Figure 1A**). Strikingly, this reduction in CSCS AUC was almost completely explained by a reduction in the ability of *semantic change* to predict escape, whereas the ability of grammaticality to predict escape was more robust to cutoff relaxation (**Figure 1A**). This result is consistent with the interpretation of semantic change as modeling a change in the antigenicity of a viral protein: as we relax the antibody binding component in our definition of an “escape mutation” (while preserving ACE2 binding and

expression), the predictive ability of semantic change also diminishes. We also observe that, in absolute terms, grammaticality is not predictive of escape up to a cutoff of 0.35, which we hypothesize to be due to mutants that we count as "escape mutants" but are unviable due to low ACE2 binding or expression.

We therefore conducted a second experiment in which we increased the ACE2 binding score cutoff from a point beyond which ydRBD has no or very weak ACE2 binding (as set by the original study authors [3]) to zero, indicating binding comparable to wildtype protein. The ACE2 binding score approximates the ability for the virus to infect host cells (and therefore approximates replication fitness). In contrast with the results above, increasing the stringency of the ACE2 binding cutoff resulted in an increased ability of *grammaticality* to predict escape, whereas the ability of semantic change to predict escape becomes much less important (**Figure 1B**). These results complement the results described above, and offer additional evidence that grammaticality is related to viral fitness [1]. When we increase the stringency of a separate protein expression cutoff also related to viral fitness, we also observe a greater increase in the AUC of grammaticality compared to semantic change (**Figure 1C**).

Together, these results support our hypothesis that the grammaticality learned by our viral protein language models encodes viral fitness and the semantic change encodes antigenic change. Moreover, at the most stringent cutoffs of fitness (i.e., ACE2 binding or expression) and antigenic change (i.e., antibody escape fraction), we observed that *both* grammaticality and semantic change become useful for predicting escape mutations.

Cutoff sensitivity experiments for other viral proteins

Our cutoff sensitivity results for Spike ydRBD indicate that our models can both predict escape mutations while also capturing information about the relative strength of an escape

mutation, as quantified by a continuous antibody selection score. We then wanted to see if the same trends held across other viruses as well. We therefore computed the AUC while varying the antibody selection cutoff of three other viral protein escape DMS experiments [4]–[6], where we varied the cutoff to consider the full range of positive selection while preserving at least 50 escape mutants (0 to 0.3 for influenza A HA H1, 0 to 7 for influenza A HA H3, 0 to 0.24 for HIV Env, and 0 to 0.5 for SARS-CoV-2 Spike ydRBD). We observed that relaxing the antibody selection cutoff also decreased the ability of our language models to predict escape (**Figure 1D**), consistent with our models having better predictive ability at more stringent antibody escape cutoffs, most likely due to stronger escape potential of these mutants. Our results for ydRBD were also robust when using expression and ACE2 binding cutoffs of -0.4 (meant to preserve escape mutants that emerge in a pseudotyped virus assay [3]) or when using the original study cutoffs (meant to preserve all conceivably viable ydRBD proteins) (**Figure 1D**).

We also performed this experiment for other baseline methods that can be used to model mutational effects from sequence data alone [7]–[12]. While these baselines show the same trends for influenza proteins, their predictions show more limited association with the strength of antibody selection for datasets profiling HIV Env [6] and SARS-CoV-2 Spike ydRBD [3] (**Figure 1D**). Moreover, CSCS consistently outperforms all baseline methods especially at the most stringent antibody selection cutoffs, where the assayed mutations have the strongest experimental evidence of escape potential (**Figure 1D**).

Analysis of SARS-CoV-2 variants-of-concern

The recent emergence of SARS-CoV-2 lineages B.1.1.7 [13] and B.1.351 [14], [15] has generated concern over the effect of these mutations on potential escape from human immunity to previous versions of SARS-CoV-2 [16], [17], particularly since these lineages contain

mutations to the Spike surface protein. We therefore used our language model to quantify the semantic change of B.1.1.7 and B.1.351 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 [16]. 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 B.1.351, 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. B.1.351 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 2**). 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 B.1.351. First, we predict antigenic drift associated with both lineages, a finding consistent with phylogenetic analysis [13], [14]. Interestingly, we predict less semantic change-based escape potential of B.1.1.7 (98.8 percentile rank of semantic change) compared to B.1.351 (99.5 percentile rank). 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* [16].

We can also estimate the individual contributions of mutations to semantic change for each of B.1.1.7, B.1.351, 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 [18], [19] (**Table 1**). The semantic change associated with P681H alone reaches a percentile rank of 48.7 compared to the background sequences.

For B.1.351, 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 [20], [21]. 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 [16]. The E484K prediction provides some opportunities to learn more about that specific RBD mutation.

Discussion

By assessing the ability of our model to predict escape when we increase or relax continuous notions of escape potential (as quantified by mutational scans), we can show that our model is especially predictive of mutants with strong experimental evidence of escape. Moreover, in a ydRBD dataset that separates notions of viral fitness (host receptor binding and protein expression) from antigenicity (antibody binding), we can also show that grammaticality is more predictive of high-fitness mutants and that semantic change is more predictive of antigenically altered mutants. Together, these results suggest that both semantic change and grammaticality are useful for escape prediction and support our hypothesis that our models are indeed learning properties related to escape potential.

Our model highlights B.1.1.7 and B.1.351 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 B.1.351. 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 [16]. 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. 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, associated data, and pretrained models are available at https://github.com/brianhie/viral-mutation/. Representative mutations for the B.1.1.7 and B.1.351 lineages were obtained from reference [17]. We used pretrained models as described by Hie et al. [1].

Benchmarking sweeps analysis

Code for conducting benchmarking sweeps across multiple escape cutoff parameters is available at https://github.com/brianhie/viral-mutation/blob/63e1538d67335fc8b84903a834ce1a5c42052a1c/bin/benchmark_escape.sh. Code for sweeping ydRBD expression cutoff can be found at https://github.com/brianhie/viral-mutation/blob/4fd6ae14199de54506f7dcadd34987f51eeaaf57/bin/benchmark_sweep_expression.sh and code for sweeping ydRBD-ACE2 binding cutoff can be found at https://github.com/brianhie/viral-mutation/blob/4fd6ae14199de54506f7dcadd34987f51eeaaf57/bin/benchmark_sweep_binding.sh.

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, B.1.351, 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 2**, 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, B.1.351, 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.

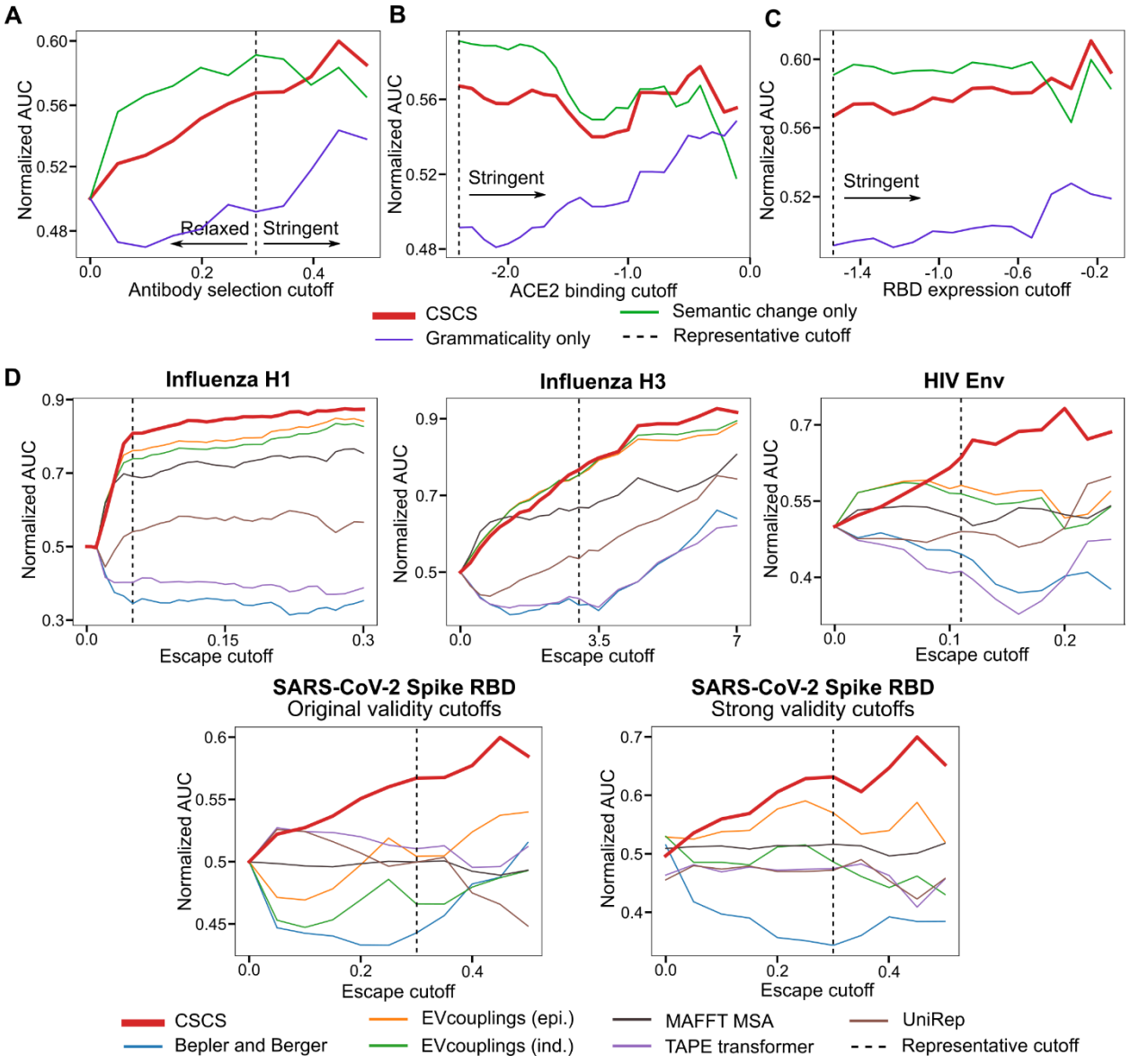

Figure 1: AUC cutoff sensitivity experiments

(**A**) With lower (more relaxed) values of the antibody binding escape fraction cutoff, CSCS AUC decreases, which is almost entirely explained by a decrease in the AUC of semantic change, whereas grammaticality remains robust to relaxation in the parameter. Increasing the stringency of this parameter removes unviable proteins enabling grammaticality to be more predictive of escape as well. (**B**) With higher (more stringent) values of the ACE2 binding cutoff, grammaticality becomes much more predictive of escape, whereas the predictive value of

semantic change falls. (**C**) With higher (more stringent) values of the protein expression score cutoff, the grammaticality AUC increases by 0.03, while semantic change AUC decreases by 0.009. (**D**) As antibody selection cutoffs increase, i.e., become more stringent, the ability for our BiLSTM-implemented CSCS to predict the escape mutations also increases consistently across all four viral protein DMS datasets. In contrast, the pattern is not observed consistently for any of the baseline methods tested outside of influenza proteins, indicating that these do not capture antibody sensitivity or continuous escape potential in their predictions. “Original validity cutoffs” indicates cutoffs to ACE2 binding and protein expression used by original study authors (-2.35 and -1.5, respectively) whereas “strong validity cutoffs” indicates more stringent cutoffs (-0.4 for both) that still preserve escape mutants observed in a pseudotyped virus assay [3] .

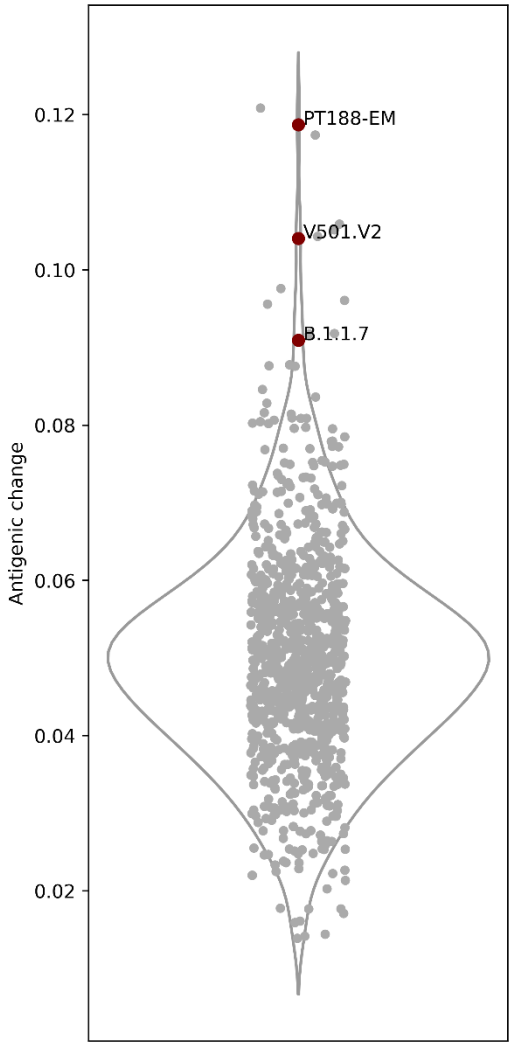

Figure 2: Predicted semantic change of mutant SARS-CoV-2 Spike sequences.

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, B.1.351 [i.e., "V501.V2"], and PT188-EM) are plotted on the same axes in red. B.1.1.7, B.1.351, 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 [16].

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 B.1.351.

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* [3], [16], has the lowest predicted semantic change.

References

[1] B. Hie, E. Zhong, B. Berger, and B. Bryson, "Learning the language of viral evolution and escape," *Science.*, vol. 371, no. 6526, pp. 284–288, 2021.

[2] T. N. Starr *et al.*, "Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding," *Cell*, vol. 182, no. 5, pp. 1295-1310.e20, 2020.

[3] A. J. Greaney *et al.*, "Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition," *Cell Host Microbe*, vol. 29, no. 1, pp. 44-57.e9, 2021.

[4] M. B. Doud, J. M. Lee, and J. D. Bloom, "How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin," *Nat. Commun.*, vol. 9, no. 1, p. 1386, 2018.

[5] J. M. Lee *et al.*, "Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin," *eLife*, vol. 27, no. 8, p. e49324, 2019.

[6] A. S. Dingens, D. Arenz, H. Weight, J. Overbaugh, and J. D. Bloom, "An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes," *Immunity*, vol. 50, no. 2, pp. 520-532.e3, 2019.

[7] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, 2013.

[8] T. A. Hopf *et al.*, "The EVcouplings Python framework for coevolutionary sequence analysis," *Bioinformatics*, vol. 35, no. 9, pp. 1582–1584, 2019.

[9] T. A. Hopf *et al.*, “Mutation effects predicted from sequence co-variation,” *Nat. Biotechnol.*, vol. 35, no. 2, pp. 128–135, 2017.

[10] R. Rao *et al.*, “Evaluating Protein Transfer Learning with TAPE,” *Adv. Neural Inf. Process. Syst.*, pp. 9686–9698, 2019.

[11] T. Bepler and B. Berger, “Learning protein sequence embeddings using information from structure,” in *7th International Conference on Learning Representations*, 2019, vol. arXiv, no. cs.LG, p. 1902.08661.

[12] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, “Unified rational protein engineering with sequence-based deep representation learning,” *Nat. Methods*, vol. 16, no. 12, pp. 1315–1322, 2019.

[13] A. Rambaut *et al.*, “Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations,” *Virological*, 2021.

[14] S. L. K. Pond *et al.*, “A preliminary selection analysis of the South African V501.V2 SARS-CoV-2 clade,” *Virological*, 2021.

[15] H. Tegally *et al.*, “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*, p. 10.1101/2020.12.21.20248640, 2020.

[16] E. Andreano *et al.*, “SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma,” *bioRxiv*, p. 10.1101/2020.12.28.424451, 2021.

[17] R. F. Garry, “Mutations arising in SARS-CoV-2 spike on sustained human-to-human transmission and human-to-animal passage,” *Virological*, 2021.

[18] M. Hoffmann, H. Kleine-Weber, and S. Pöhlmann, “A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells,” *Mol. Cell*,

vol. 78, no. 4, pp. 779–784.e5, 2020.

[19] T. P. Peacock *et al.*, “The furin cleavage site of SARS-CoV-2 spike protein is a key determinant for transmission due to enhanced replication in airway cells,” *bioRxiv*, p. 10.1101/2020.09.30.318311, 2020.

[20] N. H. Abdullah, “The current situation and Information on the Spike protein mutation of Covid-19 in Malaysia,” *Facebook post*, 2020.

[21] H. N. Harun, “Dr Noor Hisham: UK Covid-19 mutation not detected in Malaysia,” *New Straits Times*, 2020.