Cite as: M. C. Maher *et al.*, *Sci. Transl. Med.*
10.1126/scitranslmed.abk3445 (2022).

CORONAVIRUS

Predicting the mutational drivers of future SARS-CoV-2 variants of concern

M. Cyrus Maher[1]*, Istvan Bartha[1], Steven Weaver[2], Julia di Iulio[1], Elena Ferri[1], Leah Soriaga[1], Florian A. Lempp[1], Brian L. Hie[3,4], Bryan Bryson[4,5], Bonnie Berger[6,7], David L. Robertson[8], Gyorgy Snell[1], Davide Corti[1], Herbert W. Virgin[1,9,10], Sergei L. Kosakovsky Pond[2], Amalio Telenti[1]*

[1]Vir Biotechnology, San Francisco, California 94158, USA [2]Department of Biology Institute for Genomics and Evolutionary Medicine Temple University, Philadelphia, PA 19122 [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [7]Computer Science and Artificial Intelligence Laboratory. Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow GS1 1QH, UK [9]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA [10]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX 75390, USA

***Corresponding author. Email: cmaher@vir.bio, atelenti@vir.bio**

SARS-CoV-2 evolution threatens vaccine- and natural infection-derived immunity as well as the efficacy of therapeutic antibodies. To improve public health preparedness, we sought to predict which existing amino acid mutations in SARS-CoV-2 might contribute to future variants of concern. We tested the predictive value of features comprising epidemiology, evolution, immunology, and neural network-based protein sequence modeling, and identified primary biological drivers of SARS-CoV-2 intra-pandemic evolution. We found evidence that ACE2-mediated transmissibility and resistance to population-level host immunity has waxed and waned as a primary driver of SARS-CoV-2 evolution over time. We retroactively identified with high accuracy (area under the receiver operator characteristic curve, AUROC=0.92-0.97) mutations that will spread, at up to four months in advance, across different phases of the pandemic. The behavior of the model was consistent with a plausible causal structure wherein epidemiological covariates combine the effects of diverse and shifting drivers of viral fitness. We applied our model to forecast mutations that will spread in the future and characterize how these mutations affect the binding of therapeutic antibodies. These findings demonstrate that it is possible to forecast the driver mutations that could appear in emerging SARS-CoV-2 variants of concern. We validate this result against Omicron, showing elevated predictive scores for its component mutations prior to emergence, and rapid score increase across daily forecasts during emergence. This modeling approach may be applied to any rapidly evolving pathogens with sufficiently dense genomic surveillance data, such as influenza, and unknown future pandemic viruses.

INTRODUCTION

SARS-CoV-2 evolution presents an ongoing challenge to public health. Tens of thousands of mutations have arisen in the SARS-CoV-2 genome as the pandemic has progressed. Understanding the relative importance of mutations in viral proteins, particularly those of relevance for antiviral immunity, is key to allocating preparedness efforts. Mutations in the viral Spike protein have received particular attention because Spike is the target of antibody-mediated immunity and is the primary antigen in current vaccines (*1*). As of December 1st, 2021, there are 10,381 distinct amino acid substitutions, insertions, or deletions in Spike sequences from the GISAID database (*2*). These mutations occur at all but one position in the protein, in different combinations, creating over 160,000 unique Spike protein sequences. A small subset of these mutations are components of "Variants Being Monitored" (VBMs), "Variants of Interest" (VOIs) or "Variants of Concern" (VOCs), as classified by the United States Centers for Disease Control (CDC) (*3*). The distinction between VOIs and the higher alert VOCs is whether a negative clinical impact is suspected or confirmed. VBMs are variants that would be classified as VOCs if not for low prevalence.

Early statistical and algorithmic identification of the key Spike amino acid changes contributing to future putative VBM/VOI/VOCs are of clear benefit to public health strategy. Such predictions could enhance the identification of vulnerabilities for antibody-based therapeutics, vaccines, and diagnostics. Predicting future successful mutations would extend the time available to develop proactive responses at earlier stages of spread. It would also complement existing forecasting efforts which seek to predict overall SARS-CoV-2 incidence, hospitalizations, and death over time (*4*–*6*). Focus on the success of individual mutations rather than genomic variants also facilitates longer-term forecasting. The combinatorics of modeling genomic variants quickly become intractable. As a toy example, for a protein of length 1200, there are over 250 million distinct sequences that differ by only two amino acid changes. By focusing on amino acid success from the outset, we rely on common and largely correct assumptions about independence between mutations, and are able to leverage more information per mutation, thus extending the timeline on which evolution can be meaningfully forecast.

There is a robust and expanding set of analyses characterizing the features of amino acid mutations of SARS-CoV-2. Studies have identified the emergence of new variants with altered biological or antigenic properties (*7*–*9*) and characterized them using low-throughput methods (*10*, *11*). Deep mutational scanning elucidates the in vitro biological effects of all single site amino acid substitutions in a fixed genomic backbone (*12*–*14*). Others have characterized the distribution of immunodominant sites across the viral proteome (*15*, *16*) and estimated the fitness of viral sequences using neural natural language processing (NLP) applied to protein sequences (*17*).

We sought here to build upon these data and approaches to forecast the mutations that will spread from season to season. We hypothesized that this would also allow us to identify the dominant biological drivers of viral evolution over short-term timescales. These two goals are mutually reinforcing: the features that are most useful for forecasting can be inferred as measuring viral fitness. Conversely, a better understanding of evolutionary dynamics can make modeling more accurate and robust. To accomplish these goals we described patterns of rapid mutation spread both globally and within the United States and elucidated the relative predictive importance of amino acid mutational features comprising immunity, transmissibility, evolution, language model, and epidemiology. Next, we utilized data from previous infection waves to train and back-test a forecasting model that anticipates future spreading mutations and illustrated how forecasted mutations could differentially affect clinical antibodies. We extended this analysis to forecast mutations, specifically on the Delta lineage, across the whole SARS-CoV-2 proteome. As the number of Omicron sequences increases, such a targeted analysis could be repeated for that lineage as well.

RESULTS

Biological and epidemiological features of SARS-CoV-2 mutations that spread

For the purpose of developing the models, we defined "spreading" amino acid mutations as a specified fold change in frequency across multiple countries, comparing time windows before and after a chosen date (**Fig. 1**). These mutations could be substitutions, insertions, or deletions. (*2*) Within each country, we tabulated the number of sequences containing the mutation being modeled, versus those that did not, in the three months before and after a date of interest (**Fig. 1A**). For each mutation, we calculated a fold change and an associated comparison-adjusted p-value. Mutations with a significant Benjamini-Hochbert adjusted p-value ($q < 0.05$) from any country were retained. This set was further filtered using the following empirical criteria, all of which had to be met to define a mutation as spreading: a fold change (FC) from baseline of at least 10.0 in at least one country; a FC of at least 2.0 across three or more countries; and a minimum global frequency of 0.1% in the later time window. We highlight that the sequences used to calculate fold change from baseline and minimum frequency were all collected after those used for model training or feature calculation, with no overlap or interleaving between the two datasets. Performance was assessed over time by repeating this analysis in shifting or sliding time windows covering the whole data collection period, which corresponded to the three months prior to the desired forecast start date (**Fig. 1B**). Assessed data windows ranged from January-March 2020 to June-August 2021.

This definition of spreading mutations captured the expansion of VOI/VOCs globally (**fig. S1A**) as well as the growth of a number of lesser-known mutations (**fig. S1B**). Implicit in a mutation-centric approach to forecasting is the assumption that mutations accumulate in a manner that is approximately independent, or at least that their interactions can be averaged out when looking across all genomic backgrounds. To test for significant violations of this implicit assumption, we tested for linkage between all pairs of spreading mutations (**fig. S2**). Enrichment for co-occurrence between pairs of mutations at a rate of greater than 8-fold was observed for fewer than 5% of mutation pairs. Thus, we find that (pairwise) independence between mutations is a useful and approximately correct simplifying assumption.

We next determined which features of amino acid mutations are informative for predicting their spread at baseline (**Table 1, data file S1**). Within the receptor binding domain (RBD) of Spike, we found that ACE2 binding affinity was a useful predictor of mutation spread (area under the receiver operator characteristic curve, AUROC=0.85; **Fig. 1C**). Another useful predictor was the change in in vitro expression of Spike mutants (AUROC=0.82; **fig. S3A**). Among measures of immune escape, the binding contributions of known antibody epitopes (antibody binding score) to anti-SARS-CoV-2 antibodies were predictive of mutation spread (AUROC=0.71; **Fig. 1C**) whereas $CD4^+$ or $CD8^+$ T-cell immunogenicity did not offer substantial explanatory power for mutation spread (AUROC=0.52-0.62; **fig. S3A**). We found that Natural Language Processing (NLP) scores for sequence plausibility (grammaticality) (*17*) were similarly predictive to deep mutational scanning data (AUROC=0.82; **Fig. 1C**). The best evolutionary feature for prediction of spread (AUROC=0.86; **Fig.**

1C) was obtained from Fixed Effects Likelihood (FEL (*18*)) from the Hyphy package [http://www.hyphy.org] (*19*) which tests for pervasive negative or positive selection across the internal branches of a phylogenetic tree.

The highest predictive performance, however, was obtained from epidemiological features, that is, variables which more directly measure sampled mutation counts (**Table 1**). The most predictive variable in this feature category was "Epi Score", the exponentially weighted mean ranking across the other epidemiological variables (mutation frequency, fraction of unique haplotypes in which the mutation occurs, and the number of countries in which it occurs), with AUROC=0.99. This score captures both lineage expansion and recurrent mutation that occurs in multiple variant lineages by convergent evolution. We note that the utility of recurrent mutation signals is consistent with recent findings that convergent evolution plays a substantial role in SARS-CoV-2 adaptation (*20*). As observed for the RBD alone, within Spike we also obtained the best predictive performance with epidemiologic (AUROC=0.96) and evolutionary (AUROC=0.84) measures (**Fig. 1C**). The performance of other feature sets for spike is presented in **fig. S3B**.

We next sought to interrogate the robustness of this approach to changes in the underlying drivers of SARS-CoV-2 evolution. For example, it has been hypothesized that selection due to immune pressure has increased with time as more individuals became immune through infection or vaccination (*20*). For example, the Gamma P.1 lineage is thought to have spread rapidly in Brazil largely due to immune selection in a population with high seroprevalence (*21*). We measured the predictive performance of antibody binding scores, which quantify the predicted percent contribution of each Spike site to antibody affinity. We took this metric as a proxy for B cell immunodominance (**Table 1**) (*22*). Taking the maximum of this value across antibodies at a given site yielded the maximum antibody binding score. The predictiveness of this metric increased from nearly uninformative early in the pandemic (p-value for difference from random=0.53), to an AUROC of 0.75 (p<1e-4; **fig. S2C**) for predicting spreading mutations during the third wave of the pandemic (**Fig. 1D**). Predictiveness subsequently decreased again to 0.64 by summer of 2021 coincidental with the emergence of Delta. However, we found that epidemiological features maintained their performance, achieving an AUROCs of 0.92-0.97 over multiple evaluation periods (**Fig. 1D**).

Last, we trained models to predict spreading mutations using all, or various subsets of, the features identified above. We employed logistic regression with baseline features as inputs. The best predictors were epidemiologic features (AUROC=0.98) and positive selection features (AUROC=0.83; **fig. S4A**). The performance of the full model was comparable to the non-model-based performance of Epi Score (**fig. S4B**). Therefore, to simplify reproducibility and further minimize the risk of overfitting, we used Epi Score to predict mutation spread going forward. We found that taking the top 5% of mutations according to their Epi Score achieved reasonable sensitivity (~50%) and maintained a positive predictive value of between 20 and 60% across time windows (**fig. S5**). Given that an average of ~3% of observed mutations are spreading at any point in time, this represents more than a 300-fold improvement in sensitivity, and a 6- to 20-fold improvement in positive predictive value relative to random selection.

In summary, immunity, transmissibility, evolution, language model, and epidemiologic features all effectively predicted mutation spread. The methodology captured changes to the underlying selective forces over the course of the pandemic. We found that epidemiologic features in particular display superior accuracy and maintain it over time.

Examining global dynamics and the emergence of VOCs

To determine whether local or global dynamics drive mutation spread, we examined whether spreading mutations in the United States were better predicted by global or US-only epidemiological values. We tested the performance of Epi Score across four waves of the pandemic. We found that mutations were predicted with an AUROC above 0.85 up to 11 months in advance, both within the United States and globally. Global epidemiology metrics were best overall and were generally more predictive of country-level mutation spread than the country-level metrics themselves (**fig. S6**).

To illustrate the practical utility of Epi Score using global features, we assessed how early we would have been able to forecast the spread of Spike mutations that define current and former CDC VOCs, VOIs, and VBMs (n=50 defining mutations). To be conservative, we defined the date that a mutation was first forecast as the earliest date at which it was predicted to spread in two subsequent analysis periods. Of the 50 mutations (**Fig. 2A**), the median time between when a mutation was forecast to spread and when it reached 1% frequency was 5 months. The maximum was 20 months, while the minimum was 0 months for D614G, because this mutation had already reached a frequency of 69% by the first forecast period. The distribution of these forecast intervals is presented in **Fig. 2B**.

Of particular note, Y145H was forecast to spread starting in July of 2021. This mutation is now a defining mutation of AY.4.2, a spreading sub-lineage of the Delta VOC. As of October 2021, AY.4.2 accounted for 8.5-11.3% of samples in the UK. Estimated growth rates remain slightly higher for AY.4.2 than for Delta, and the household secondary attack rate was higher for AY.4.2 cases than for other Delta cases (*23*). Based on these observations, we conclude that our approach was able to predict key mutations, across all current and former VOC/VOI/VBMs, several months in advance. Early warning of mutations in current VOCs, VOIs, and VBMs would have been possible before reaching worrisome degrees of global spread.

Understanding performance through a causal lens

Seeking to understand the high predictive performance of epidemiologic features, we constructed a directed acyclic graph to represent the hypothesized causal relationships, and to probe whether relative trends in performance were consistent with the expectations that follow from this model (**Fig. 3A**). We proposed that epidemiologic features mediate the relationship between viral fitness and mutation spread. Our rationale was that if a mutation's contribution to viral fitness was sufficient to drive it to appreciable prevalence at one time point (as measured by global frequency and geographic distribution), and in the context many genetic backgrounds, it would likely drive it to higher prevalence in the future as well (unless it were outcompeted by a more fit adaptation, or the fitness landscape changed). This type of mediated relationship (fitnessÞcurrent prevalenceÞfuture prevalence) implies that epidemiological prevalence features will capture information from both known and unknown drivers of selection.

If the causal model were reasonable, we would expect first that variables whose causal effects are mediated, as defined above, should predict epidemiologic variables at a comparable or even greater accuracy compared to spreading mutations. This is illustrated by comparing the first and second columns of **Fig. 3B**. We observed that, with the exception of the maximal antibody binding score, all top variables predicted Epi Scores better than they predict mutation spread. The lower predictiveness of maximal antibody binding score for Epi Scores would be consistent with a slight time lag effect due to shifting evolutionary pressures.

A second criterion for mediation is that information from these variables should not substantially complement the predictiveness of the epidemiologic variables alone. In other words, there should be little or no additional information that other inputs provide relative to the epidemiologic variables. We assessed this by comparing the AUROCs of two-variable models in column 3 of **Fig. 3B** with the AUROC for Epi Score alone (0.983). The only nominal AUROC increase for a complemented model was observed for the evolutionary measure FEL (0.984). We did not find statistically significant complementarity with Epi Score for this or any other variable, either within the RBD or across full length Spike (see supplemental section "Mediation Analysis", **table S1).**

Our examination of mediated causal relationships begins by assuming a causal graph based on prior knowledge. Such an approach is common to many causal inference methods (*24*) and represents a well-understood limitation of these methods (*24*). Therefore, we considered this as a tool to more systematically analyze the plausibility of our results. Although it is generally difficult to verify the structure of proposed causal graphs, our findings support the concept that epidemiological variables mediate the effects of other classes of explanatory variables, and this may explain their high predictive accuracy.

Emergence and spread of Omicron

While this work was in revision, we were confronted with the emergence in late November 2021 of the Omicron (B.1.1.529/21K) variant. Despite the low frequency of many of the individual mutations that define the major haplotype of Omicron (median allele frequency 0.00046), we observed high Epi Score values across Spike (median Epi Score of 9.51); **Fig. 4A**. A benefit of the computational simplicity of Epi Score is that predictions can easily be updated on a daily basis. We therefore sought to move beyond single time point Epi Scores to examine trends in Epi Score across time for the Omicron mutations. The time-analysis showed that the Omicron Spike mutations had progressively higher Epi Score values long preceding the acceleration that characterized the emergence of Omicron in November 2021 (**Fig. 4B**). We additionally found that the spread of Omicron was rapidly reflected in the raising Epi Scores of its mutations, and that daily forecasts allowed the identification of trending scores.

As an independent approach to assess the singularity of Omicron, we also examined the evolutionary nature of the Omicron mutations using our language model. Omicron had a grammaticality change between that of Alpha and Delta, but the highest semantic change (predicted antigenic shift) of any SARS-CoV-2 lineage (**fig. S7**). Indeed, Omicron's semantic change score was twice that of both Alpha and Delta, consistent with high levels of mutation and immune escape adaptation.

Forecasting spreading mutations in Spike and proteome-wide

Building upon the accurate prediction of spreading mutations across different waves of the pandemic, we next leveraged Epi Score on current data to forecast mutations that may contribute to VOIs and VOCs over the coming months. Because global metrics outperformed metrics restricted to the United States, even for forecasting within the United States, we focused on global forecasting. We considered shortening our feature calculation window to further mitigate the effects of shifting evolutionary dynamics. However, we found that longer feature calculation windows improved performance across all prediction windows (**fig. S8**).

As an application of the forecasting analysis, we examined how forecasted mutations intersected with the binding sites of clinical antibodies as of October 19th, 2021. We found wide variation in the number of forecasted mutations per antibody epitope (**Table 2**), ranging from 10 mutations for Celltrion's CT-P59, to two low-frequency mutations for Vir-7831 (sotrovimab), which was designed to be more robust to viral evolution by targeting a region that is conserved across coronaviruses (*25*). The two mutations in the epitope of sotrovimab, A340S and R346K, do not limit neutralization (*25*, *26*). As an additional proof of concept, we focused our attention on Spike S494P, a mutation reported to have enhanced binding affinity to ACE2 (*27*), and to reduce neutralization by 3-5-fold in some convalescent sera (*27*). We found that the S494P mutation decreases neutralization potential of clinical therapeutic antibodies: Ly-CoV555 (bamlanivimab), CT-P59 and to a lesser extent to REGN10933 (casirivimab) (**Fig. 5**).

Last, to demonstrate the flexibility and extensibility of our approach, we forecasted the spread of mutations specifically on the Delta genomic background, across the full SARS-CoV-2 proteome. Because the components of Epi Score can be calculated for any mutation where sequencing data are available, extension to the full proteome is trivial and not computationally taxing. It can also be reasonably calculated on any subset of sequences to determine which mutations are most likely to spread based on their characteristics within that subset (or lineage). Therefore, it is also straightforward to adapt this approach to produce lineage-specific forecasts. **Fig. 6A** shows a Manhattan-style plot of Epi Scores across the full SARS-CoV-2 genome. The plot highlights all mutations at positively selected sites (FEL, fixed effects model for detecting site-wise selective pressure, FDR < 0.05) that currently occur at a frequency over 0.1% on a Delta background. We found 151 such mutations, distributed across the proteome. The mutation density was 1.8 per 100 amino acids across the whole proteome, with a rate varying from 0 to 12.3 across SARS-CoV-2 proteins (**Fig. 6B**). By this measure, the highest mutational density was identified in ORF3/NS3, an accessory protein that is reported to modulate autophagosome–lysosome fusion (ORF3a) (*28*) and antagonize interferon (Orf3b) (*29*). Spike was close to average, with a density of 2.3 mutations per 100 amino acids. Based on the Epi Score ranking, the top 5 mutations for potential to spread were Spike:G142D, Spike:T95I, NSP3:A1711V, N:Q9L, and NSP2:K81N. All mutation Epi Scores proteome-wide are presented in **data file S2**.

In summary, we established a method for predicting spreading mutations and applied it to forecast future contributors to putative VOCs/VOIs/VBMs. These predictions yield mutations known to be important from in vitro data. We conclude that this approach can anticipate spreading mutations many months in advance. We find that a subset of forecast mutations could have implications for the continued efficacy of clinical antibodies, but that the level of these effects varies widely. We then extended our analysis to encompass the full SARS-CoV-2 proteome, and to produce Delta and informative Omicron forecasts. This work also suggests that there is considerable potential for spreading mutations located outside of Spike, underlining the importance of forecasting methods that can be applied across the whole viral proteome.

DISCUSSION

We established a working definition for spreading mutations and leveraged this definition to deliver a systematic analysis of amino acid features predictive of mutation spread. This yielded a simple, explainable, and accurate approach for forecasting mutations several months in advance, across multiple pandemic waves. Calculating this scoring was also efficient enough to enable daily forecast updates on millions of sequences using only a laptop. Although this strategy required nothing more than genomic surveillance data, we also highlighted the value of the complete mapping of epitopes, in vitro deep site-directed mutagenesis, and downstream functional experimental validation. Confidence in the prediction of spreading mutations came through retrospectively evaluating multiple waves of the pandemic and verifying consistency with experimental data, and with a plausible causal framework. Furthermore, long observed lags between the earliest warning signals and high population frequency of current mutations in VOCs, VOIs, and VBMs gave further support for using forecasting to anticipate the spread of future concerning mutations. Although this approach will be limited in its ability to anticipate mutations that appear and rise to high frequencies within a short time frame, we found this to be a rare occurrence.

We evaluated epidemiologic features aggregated in the Epi Score such as mutation frequency, and the distribution of mutations across countries and fraction of unique haplotypes across which a mutation occurs. We explored other predictors, including the rate of increase of each of these features, but did not find that they improved performance. We note that the fraction of unique haplotypes shared similarities to phylogenetic measures of recurrent mutation. However, there is considerable lack of phylogenetic resolution in such calculations, so the number of recurrent mutations is a statistically "noisy" measure, depends strongly on the method used to build phylogenies, and is very expensive to compute. The fraction of unique haplotypes, on the other hand, is fast to compute, can be perfectly estimated, and will increase with both recurrent mutation and single-lineage expansion; both of which are indicative of a positive contribution to fitness.

Omicron emerged a*s* the paper was completing the review process. Despite the limited numbers of viral sequences available as of December 2021, we observed a distinctive pattern of Omicron mutations that, despite low frequency of many individual mutations, already had high Epi Score values. It is also notable that for all mutations, high Epi Score values antedated the emergence of Omicron, even though those mutations had not yet converged on the same haplotypes. We interpret these data as indicative that individual mutations were endowed with advantageous properties in the viral genome even before their co-occurrence on the Omicron spike.

There are limits to this study; general prediction of viral evolution is fundamentally an intractable problem. The current work only addresses a simpler question: predicting which mutations will increase in frequency over some threshold in the near future based on the analysis of their recent patterns of spread. Thus, the study predicts spread of existing mutations, but not a true emergence of previously unobserved mutations. In addition, it is difficult to predict which lineages, i.e., a major viral haplotype, will spread because this would require the complex projection of growth of multiple mutations together. These limitations notwithstanding, the data on Omicron suggest that successful lineages may be defined by the convergence of mutations that, individually, exhibited high Epi Score values and other features that signal adaptive evolution.

Although this work forecasts which mutations will spread, the success of a given mutation does not necessarily result in clinical or public health consequences. Therefore, we posit that the value of the predictions is to prioritize mutations for functional screening. Here, we demonstrate how a subset of spreading mutations differentially impact clinical antibodies. We also extended the analysis to encompass the whole viral proteome. By this approach, we identified spreading amino acid replacements in other viral proteins, and highlighted positions under strong positive selection. Given the limited understanding of the role of non-Spike regions of the proteome in driving the pandemic, we believe that those non-Spike mutations should be prioritized for understanding their role in evading innate immunity, increasing the replication of SARS-CoV-2, and more generally for their contribution to viral fitness. We intend for these results to provide a foundation for future improvement. Although we have shown that Epi Score is robust to shifting evolutionary dynamics, performance can be monitored in real-time, and if necessary, re-tuned to capture novel behavior as now shown with the emergence of Omicron. This approach can also be generalized and improved upon to stay ahead of evolutionary cycles for other pathogens (*30*), when sufficiently rich and representative genomic sampling is available.

MATERIALS AND METHODS

Study Design. *Sample size.* The current work to define spreading amino acid mutations was based on viral sequences and metadata obtained from GISAID EpiCoV project (https://www.gisaid.org/). A total of 4,487,305 sequences were analyzed.

Research objectives. We hypothesized that the pattern of spread could be estimated from the large database of GISAID. Next, we hypothesized that one or more variable comprising biological, immunological, epidemiological and genomic (including language) features could be identified as drivers of the spread.

Experimental design. We used predictive models and expressed predictive performance using the area under the receiver operator characteristic curve (AUROC). Prediction was performed using forward feature selection followed by logistic regression. The criterion for forward selection was cross-validated AUROC of the logistic regression model within the training set. Feature selection and model fitting were performed separately within each fold of the outer cross validation loop. Logistic regression was chosen due to its sample efficiency.

Statistical analysis. Spreading mutations were defined based on a Fisher's exact test for frequency fold change per country, adjusted for multiple comparisons, followed by filters for rate of spread (max fold change of at least 10, fold change > 2 in three or more countries), and a minimum prevalence of 0.1%. We estimated epistasis using pointwise mutual information, which corresponds to the log ratio of the observed prevalence of a pair to the expected prevalence assuming independence. The most predictive variable, "Epi Score" was defined as the exponentially weighted mean ranking across the other epidemiological variables (mutation frequency, fraction of unique haplotypes in which the mutation occurs, and the number of countries in which it occurs. For natural language processing (NLP) neural network features, we used the grammaticality and semantic change scores reported by Hie *et al.* (*17*) in which a bidirectional long short-term memory (BiLSTM) model was trained on Spike sequences from GISAID and GenBank. Natural selection features were generated using MEME (*31*) and FEL (*18*) methods implemented in the HyPhy package (*19*) (version 2.5.31). Mediation analysis was based on the Baron and Kenny test. The list of forecast mutations was generated by calculating Epi Scores on the most recent three months of data and taking the top 5% of mutations, a cutoff chosen based empirical analyses.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scitranslmed.abk3445
Materials and Methods
Figs. S1 to S8
Table S1
References (*35*–*41*)
Data Files S1 to S3
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. M. McCallum, A. De Marco, F. A. Lempp, M. A. Tortorici, D. Pinto, A. C. Walls, M. Beltramello, A. Chen, Z. Liu, F. Zatta, S. Zepeda, J. di Iulio, J. E. Bowen, M. Montiel-Ruiz, J. Zhou, L. E. Rosen, S. Bianchi, B. Guarino, C. S. Fregni, R. Abdelnabi, S. C. Foo, P. W. Rothlauf, L.-M. Bloyet, F. Benigni, E. Cameroni, J. Neyts, A. Riva, G. Snell, A. Telenti, S. P. J. Whelan, H. W. Virgin, D. Corti, M. S. Pizzuto, D. Veesler, N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021). doi:10.1016/j.cell.2021.03.028 Medline

2. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017). doi:10.1002/gch2.1018 Medline

3. C. for D. Control, SARS-CoV-2 Variants of Concern (available at https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html).

4. A. Adiga, L. Wang, B. Hurt, A. Peddireddy, P. Porebski, S. Venkatramanan, B. Lewis,

M. Marathe, All Models Are Useful: Bayesian Ensembling for Robust High Resolution COVID-19 Forecasting. *Medrxiv*, 2021.03.12.21253495 (2021).

5. H. Zhao, N. N. Merchant, A. McNulty, T. A. Radcliff, M. J. Cote, R. S. B. Fischer, H. Sang, M. G. Ory, COVID-19: Short term prediction model using daily incidence data. *PLOS ONE* **16**, e0250110 (2021). doi:10.1371/journal.pone.0250110 Medline
6. E. L. Ray, N. Wattanachit, J. Niemi, A. H. Kanji, K. House, E. Y. Cramer, J. Bracher, A. Zheng, T. K. Yamana, X. Xiong, S. Woody, Y. Wang, L. Wang, R. L. Walraven, V. Tomar, K. Sherratt, D. Sheldon, R. C. Reiner, B. A. Prakash, D. Osthus, M. L. Li, E. C. Lee, U. Koyluoglu, P. Keskinocak, Y. Gu, Q. Gu, G. E. George, G. España, S. Corsetti, J. Chhatwal, S. Cavany, H. Biegel, M. Ben-Nun, J. Walker, R. Slayton, V. Lopez, M. Biggerstaff, M. A. Johansson, N. G. Reich, Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. *Medrxiv*, 2020.08.19.20177493 (2020).
7. A. Padane, A. Kanteh, N. Leye, A. Mboup, J. Manneh, M. Mbow, P. A. Diaw, B. P. Ndiaye, G. Lo, C. I. Lo, A. Ahoudi, A. Gueye-Gaye, J. J. N. Malomar, A. Dia, Y. A. Dia, N. D. Diagne, D. Wade, A. K. Sesay, N. C. Toure-Kane, U. Dalessandro, S. Mboup, First detection of the British variant of SARS-CoV-2 in Senegal. *New Microbes New Infect.*, 100877 (2021). doi:10.1016/j.nmni.2021.100877 Medline
8. A. L. Valesano, K. E. Rumfelt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, A. S. Lauring, Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLOS Pathog.* **17**, e1009499 (2021). doi:10.1371/journal.ppat.1009499 Medline
9. R. Charkiewicz, J. Nikliński, P. Biecek, J. Kiśluk, S. Pancewicz, A. M. Moniuszko-Malinowska, R. Flisiak, A. J. Krętowski, J. Dzięcioł, M. Moniuszko, R. Gierczyński, G. Juszczyk, J. Reszeć, The first SARS-CoV-2 genetic variants of concern (VOC) in Poland: The concept of a comprehensive approach to monitoring and surveillance of emerging variants. *Adv. Med. Sci.* **66**, 237–245 (2021). doi:10.1016/j.advms.2021.03.005 Medline
10. W. Dejnirattisai, D. Zhou, P. Supasa, C. Liu, A. J. Mentzer, H. M. Ginn, Y. Zhao, H. M. E. Duyvesteyn, A. Tuekprakhon, R. Nutalai, B. Wang, C. López-Camacho, J. Slon-Campos, T. S. Walter, D. Skelly, S. A. Costa Clemens, F. G. Naveca, V. Nascimento, F. Nascimento, C. Fernandes da Costa, P. C. Resende, A. Pauvolid-Correa, M. M. Siqueira, C. Dold, R. Levin, T. Dong, A. J. Pollard, J. C. Knight, D. Crook, T. Lambe, E. Clutterbuck, S. Bibi, A. Flaxman, M. Bittaye, S. Belij-Rammerstorfer, S. C. Gilbert, M. W. Carroll, P. Klenerman, E. Barnes, S. J. Dunachie, N. G. Paterson, M. A. Williams, D. R. Hall, R. J. G. Hulswit, T. A. Bowden, E. E. Fry, J. Mongkolsapaya, J. Ren, D. I. Stuart, G. R. Screaton, Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939–2954.e9 (2021). doi:10.1016/j.cell.2021.03.055 Medline
11. D. A. Collier, A. De Marco, I. A. T. M. Ferreira, B. Meng, R. P. Datir, A. C. Walls, S. A. Kemp, J. Bassi, D. Pinto, C. Silacci-Fregni, S. Bianchi, M. A. Tortorici, J. Bowen, K. Culap, S. Jaconi, E. Cameroni, G. Snell, M. S. Pizzuto, A. F. Pellanda, C. Garzoni, A. Riva, A. Elmer, N. Kingston, B. Graves, L. E. McCoy, K. G. C. Smith, J. R. Bradley, N. Temperton, L. Ceron-Gutierrez, G. Barcenas-Morales, W. Harvey, H. W. Virgin, A. Lanzavecchia, L. Piccoli, R. Doffinger, M. Wills, D. Veesler, D. Corti, R. K. Gupta; CITIID-NIHR BioResource COVID-19 Collaboration; COVID-19 Genomics UK (COG-UK) Consortium, Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021). doi:10.1038/s41586-021-03412-7 Medline
12. T. N. Starr, A. J. Greaney, A. S. Dingens, J. D. Bloom, Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Reports Medicine* **2**, 100255 (2021). doi:10.1016/j.xcrm.2021.100255 Medline
13. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020). doi:10.1016/j.cell.2020.08.012 Medline
14. T. N. Starr, A. J. Greaney, A. Addetia, W. W. Hannon, M. C. Choudhary, A. S. Dingens, J. Z. Li, J. D. Bloom, Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021). doi:10.1126/science.abf9302 Medline
15. B. Agerer, M. Koblischke, V. Gudipati, L. F. Montaño-Gutierrez, M. Smyth, A. Popa, J.-W. Genger, L. Endler, D. M. Florian, V. Mühlgrabner, M. Graninger, S. W. Aberle, A.-M. Husa, L. E. Shaw, A. Lercher, P. Gattinger, R. Torralba-Gombau, D. Trapin, T. Penz, D. Barreca, I. Fae, S. Wenda, M. Traugott, G. Walder, W. F. Pickl, V. Thiel, F. Allerberger, H. Stockinger, E. Puchhammer-Stöckl, W. Weninger, G. Fischer, W. Hoepler, E. Pawelka, A. Zoufaly, R. Valenta, C. Bock, W. Paster, R. Geyeregger, M. Farlik, F. Halbritter, J. B. Huppa, J. H. Aberle, A. Bergthaler, SARS-CoV-2 mutations in MHC-I-restricted epitopes evade $CD8^+$ T cell responses. *Sci. Immunol.* **6**, eabg6461 (2021). doi:10.1126/sciimmunol.abg6461 Medline
16. A. Tarke, J. Sidney, N. Methot, E. D. Yu, Y. Zhang, J. M. Dan, B. Goodwin, P. Rubiro, A. Sutherland, E. Wang, A. Frazier, S. I. Ramirez, S. A. Rawlings, D. M. Smith, R. da Silva Antunes, B. Peters, R. H. Scheuermann, D. Weiskopf, S. Crotty, A. Grifoni, A. Sette, Impact of SARS-CoV-2 variants on the total $CD4^+$ and $CD8^+$ T cell reactivity in infected or vaccinated individuals. *Cell Reports Medicine* **2**, 100355 (2021). doi:10.1016/j.xcrm.2021.100355 Medline
17. B. Hie, E. D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021). doi:10.1126/science.abd7331 Medline
18. S. L. Kosakovsky Pond, S. D. W. Frost, Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005). doi:10.1093/molbev/msi105 Medline
19. S. L. Kosakovsky Pond, A. F. Y. Poon, R. Velazquez, S. Weaver, N. L. Hepler, B. Murrell, S. D. Shank, B. R. Magalis, D. Bouvier, A. Nekrutenko, S. Wisotsky, S. J. Spielman, S. D. W. Frost, S. V. Muse, HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020). doi:10.1093/molbev/msz197 Medline
20. D. P. Martin, S. Weaver, H. Tegally, J. E. San, S. D. Shank, E. Wilkinson, A. G. Lucaci, J. Giandhari, S. Naidoo, Y. Pillay, L. Singh, R. J. Lessells, R. K. Gupta, J. O. Wertheim, A. Nekturenko, B. Murrell, G. W. Harkins, P. Lemey, O. A. MacLean, D. L. Robertson, T. de Oliveira, S. L. Kosakovsky Pond; NGS-SA; COVID-19 Genomics UK (COG-UK), The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7 (2021). doi:10.1016/j.cell.2021.09.003 Medline
21. N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. D. S. Candido, S. Mishra, M. A. E. Crispim, F. C. S. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. G. Kraemer, N. Gaburo Jr., C. D. C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T. Araujo, F. S. V. Malta, A. B. de Lima, J. D. P. Silva, D. A. G. Zauli, A. C. S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. Dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. Dos Santos, R. S. Aguiar, J. L. Proença-Modena, B. Nelson, J. A. Hay, M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, C. A. Prete Jr., V. H. Nascimento, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. D. P. S. S. Carvalho, O. G. Pybus, S. Flaxman, S. Bhatt, E. C. Sabino, Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021). doi:10.1126/science.abh2644 Medline
22. S. Vilar, G. Cozza, S. Moro, Medicinal chemistry and the molecular operating environment (MOE): Application of QSAR and molecular docking to drug discovery. *Curr. Top. Med. Chem.* **8**, 1555–1572 (2008). doi:10.2174/156802608786786624 Medline
23. U. H. Security, *SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 27* (2021; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1029715/technical-briefing-27.pdf).
24. N. Pearce, D. A. Lawlor, Causal inference-so much more than statistics. *Int. J. Epidemiol.* **45**, 1895–1903 (2016). doi:10.1093/ije/dyw328 Medline
25. A. L. Cathcart, C. Havenar-Daughton, F. A. Lempp, D. Ma, M. A. Schmid, M. L. Agostini, B. Guarino, J. D. iulio, L. E. Rosen, H. Tucker, J. Dillen, S. Subramanian, B. Sloan, S. Bianchi, D. Pinto, C. Saliba, J. A. Wojcechowskyj, J. Noack, J. Zhou, H. Kaiser, A. Chase, M. Montiel-Ruiz, E. Dellota, A. Park, R. Spreafico, A. Sahakyan, E. J. Lauron, N. Czudnochowski, E. Cameroni, S. Ledoux, A. Werts, C. Colas, L. Soriaga, A. Telenti, L. A. Purcell, S. Hwang, G. Snell, H. W. Virgin, D. Corti, C. M. Hebner, The dual function monoclonal antibodies VIR-7831 and VIR-7832 demonstrate potent in vitro and in vivo activity against SARS-CoV-2. *Biorxiv*, 2021.03.09.434607 (2021).
26. T. N. Starr, N. Czudnochowski, Z. Liu, F. Zatta, Y.-J. Park, A. Addetia, D. Pinto, M. Beltramello, P. Hernandez, A. J. Greaney, R. Marzi, W. G. Glass, I. Zhang, A. S. Dingens, J. E. Bowen, M. A. Tortorici, A. C. Walls, J. A. Wojcechowskyj, A. De Marco,

L. E. Rosen, J. Zhou, M. Montiel-Ruiz, H. Kaiser, J. R. Dillen, H. Tucker, J. Bassi, C. Silacci-Fregni, M. P. Housley, J. di Iulio, G. Lombardo, M. Agostini, N. Sprugasci, K. Culap, S. Jaconi, M. Meury, E. Dellota Jr., R. Abdelnabi, S. C. Foo, E. Cameroni, S. Stumpf, T. I. Croll, J. C. Nix, C. Havenar-Daughton, L. Piccoli, F. Benigni, J. Neyts, A. Telenti, F. A. Lempp, M. S. Pizzuto, J. D. Chodera, C. M. Hebner, H. W. Virgin, S. P. J. Whelan, D. Veesler, D. Corti, J. D. Bloom, G. Snell, SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature* **597**, 97–102 (2021). doi:10.1038/s41586-021-03807-6 Medline

27. S. Chakraborty, Evolutionary and structural analysis elucidates mutations on SARS-CoV2 spike protein with altered human ACE2 binding affinity. *Biochem. Biophys. Res. Commun.* **534**, 374–380 (2021). doi:10.1016/j.bbrc.2020.11.075 Medline

28. Y. Zhang, H. Sun, R. Pei, B. Mao, Z. Zhao, H. Li, Y. Lin, K. Lu, The SARS-CoV-2 protein ORF3a inhibits fusion of autophagosomes with lysosomes. *Cell Discov.* **7**, 31 (2021). doi:10.1038/s41421-021-00268-z Medline

29. Y. Konno, I. Kimura, K. Uriu, M. Fukushi, T. Irie, Y. Koyanagi, D. Sauter, R. J. Gifford, S. Nakagawa, K. Sato; USFQ-COVID19 Consortium, SARS-CoV-2 ORF3b is a potent interferon antagonist whose activity is increased by a naturally occurring elongation variant. *Cell Rep.* **32**, 108185 (2020). doi:10.1016/j.celrep.2020.108185 Medline

30. J. W. Tang, T. T. Lam, H. Zaraket, W. I. Lipkin, S. J. Drews, T. F. Hatchette, J.-M. Heraud, M. P. Koopmans, A. M. Abraham, A. Baraket, S. Bialasiewicz, M. A. Caniza, P. K. S. Chan, C. Cohen, A. Corriveau, B. J. Cowling, S. J. Drews, M. Echavarria, R. Fouchier, P. L. A. Fraaij, T. F. Hachette, J.-M. Heraud, H. Jalal, L. Jennings, A. Kabanda, H. A. Kadjo, M. R. Khanani, E. S. C. Koay, M. P. Koopmans, M. Krajden, T. T. Lam, H. K. Lee, W. I. Lipkin, J. Lutwama, D. Marchant, H. Nishimura, P. Nymadawa, B. A. Pinsky, S. Rughooputh, J. Rukelibuga, T. Saiyed, A. Shet, T. Sloots, J. J. M. Tamfum, J. W. Tang, S. Tempia, S. Tozer, F. Treurnicht, M. Waris, A. Watanabe, E. O. Wemakoy; INSPIRE investigators, Global epidemiology of non-influenza RNA respiratory viruses: Data gaps and a growing need for surveillance. *Lancet Infect. Dis.* **17**, e320–e326 (2017). doi:10.1016/S1473-3099(17)30238-4 Medline

31. B. Murrell, J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S. L. Kosakovsky Pond, Detecting individual sites subject to episodic diversifying selection. *PLOS Genet.* **8**, e1002764 (2012). doi:10.1371/journal.pgen.1002764 Medline

32. I. Manfredonia, C. Nithin, A. Ponce-Salvatierra, P. Ghosh, T. K. Wirecki, T. Marinus, N. S. Ogando, E. J. Snijder, M. J. van Hemert, J. M. Bujnicki, D. Incarnato, Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.* **48**, 12436–12452 (2020). doi:10.1093/nar/gkaa1053 Medline

33. A. Tarke, J. Sidney, C. K. Kidd, J. M. Dan, S. I. Ramirez, E. D. Yu, J. Mateus, R. da Silva Antunes, E. Moore, P. Rubiro, N. Methot, E. Phillips, S. Mallal, A. Frazier, S. A. Rawlings, J. A. Greenbaum, B. Peters, D. M. Smith, S. Crotty, D. Weiskopf, A. Grifoni, A. Sette, Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Reports Medicine* **2**, 100204 (2021). doi:10.1016/j.xcrm.2021.100204 Medline

34. A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. D. Crawford, A. S. Dingens, R. S. Nargi, R. E. Sutton, N. Suryadevara, P. W. Rothlauf, Z. Liu, S. P. J. Whelan, R. H. Carnahan, J. E. Crowe Jr., J. D. Bloom, Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44–57.e9 (2021). doi:10.1016/j.chom.2020.11.007 Medline

35. T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, Sylvain, P. Ivanov, D. Avila, Abdalla, S. Abdalla, Willing, J. development team, in *20th International Conference on Electronic Publishing*, (2016).

36. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature* **585**, 357–362 (2020). doi:10.1038/s41586-020-2649-2 Medline

37. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

38. S. Seabold, J. Perktold, in *9th Python in Science Conference*, (2010).

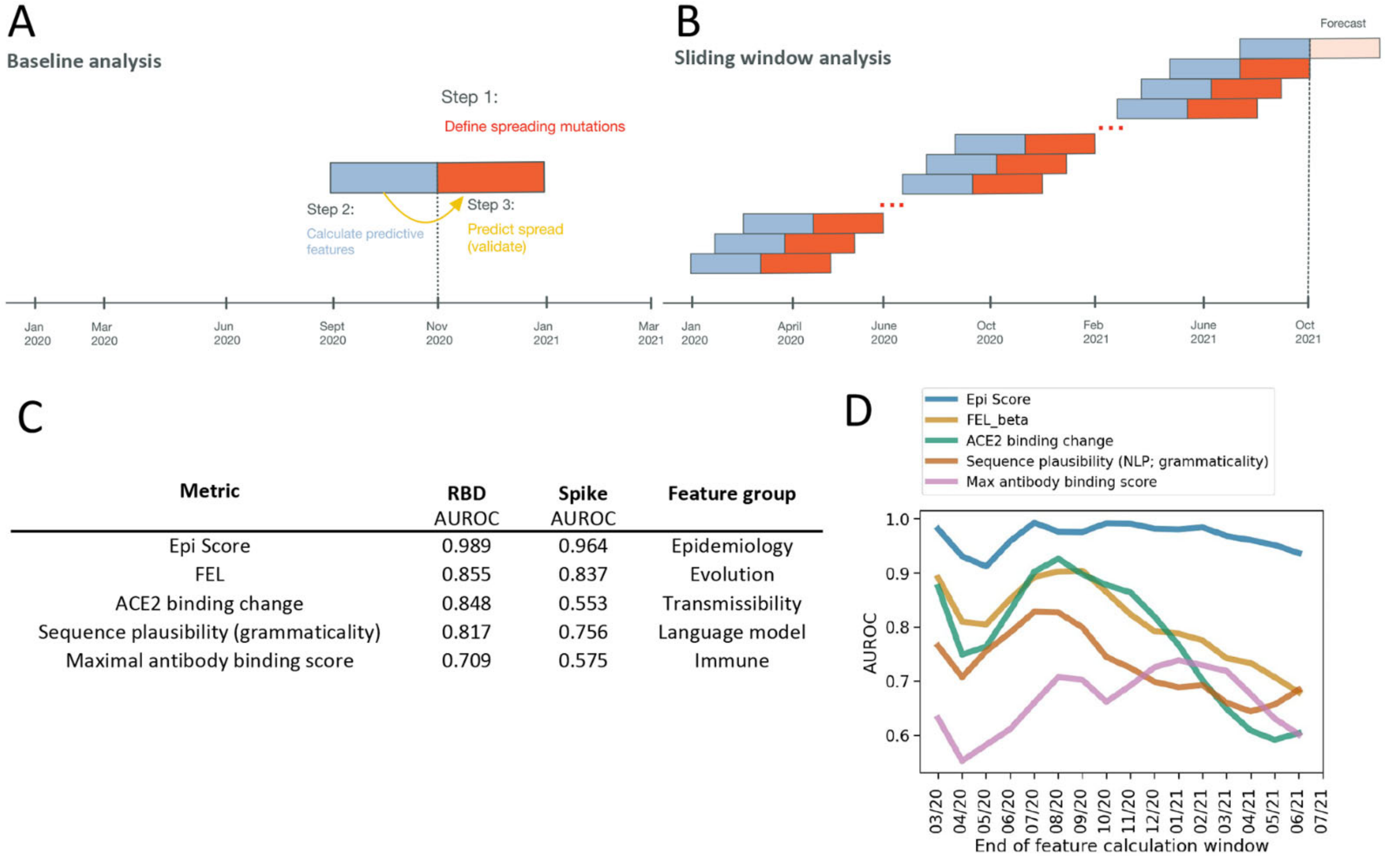
39. W. McKinney, in *Proceedings of the 9th Python Science Conference*, S. van der Walt, J. Millman, Eds. (2010), pp. 56–61.

40. S. L. K. Pond, S. D. W. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005). doi:10.1093/bioinformatics/bti079 Medline

41. A. J. Greaney, A. N. Loes, K. H. D. Crawford, T. N. Starr, K. D. Malone, H. Y. Chu, J. D. Bloom, Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476.e6 (2021). doi:10.1016/j.chom.2021.02.003 Medline

Acknowledgments: We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID used in the current study. The full GISAID acknowledgement list can be found at http://data.hyphy.org/web/SARS-CoV-2/gisaid.csv.gz. We thank Darren Martin and Emma Hodcroft for useful discussion. **Funding:** This research is funded by Vir Biotechnology. D.L.R. is funded by the MRC (MC_UU_12014/12). S.L.K.P was supported in part by the NIH/NIAID (AI134384). **Author contributions:** M.C.M., A.T. conceived the study. M.C.M, I.B., S.W., J.dI., F.A.L., B.L.H., performed experiments. M.C.M, I.B., E.F., L.S., F.A.L., B.L.H., S.L.K.P., A.T. analyzed and interpreted data. Br.B., Bo.B., D.L.R., G.S., D.C., H.W.V., S.L.K., A.T. supervised research. M.C.M., D.C., H.W.V., S.L.K.P., A.T. wrote the manuscript. **Competing interests:** M.C.M., I.B., J.dI., F.A.L., E.F., L.S., G.S., D.C., H.W.V., and A.T are employees of Vir Biotechnology and may hold shares of the company. H.W.V. is a founder of PierianDx and Casma Therapeutics and holds stock or stock options in these companies. Neither company funded the present work. H.W.V. holds a number of patents from work at Washington University School of Medicine. The present work does not involve these patents. Vir has filed a patent application on this work: 63/212,945; "Predicting mutational drivers of future pathogen spread". **Data and materials availability:** All data associated with this study are present in the paper or supplementary materials. **Data file S1** describes the sources for all features used for prediction. **Data file S2** provides the Epi Scores proteome-wide. **Data file S3** provides all the figure data for figures S3-S7. Code used in this study is available at DOI: https://zenodo.org/badge/latestdoi/440943417.

Submitted 7 July 2021
Accepted 6 January 2022
Published First Release 11 January 2022
10.1126/scitranslmed.abk3445



Metric	RBD AUROC	Spike AUROC	Feature group
Epi Score	0.989	0.964	Epidemiology
FEL	0.855	0.837	Evolution
ACE2 binding change	0.848	0.553	Transmissibility
Sequence plausibility (grammaticality)	0.817	0.756	Language model
Maximal antibody binding score	0.709	0.575	Immune

Fig. 1. Predicting mutation spread. (A) Analyzing performance at baseline and over time. The core analysis consists of three steps. First, creating a working definition for spreading mutations. Second, calculating features that can predict future spread using a window of prior data. Third, having constructed models on training data, run prediction of future spread (Forecast), and interpret the results. (B) Performance was assessed over time by repeating this analysis in sliding time windows covering the whole data collection period. (C) The most predictive metrics within each feature group at baseline (see Table 1 and table S1) were ranked by performance within the receptor binding domain (RBD), where the most data are available and for the Spike. (D) RBD classification accuracy over time for the top GISAID-based feature (Epi Score), and the top transmission and immune variables (Table 1). AUROCs in panel D are smoothed with a rolling window of two analysis periods. AUROC, area under the receiver operating characteristic curve. FEL, fixed effects model for detecting site-wise selective pressure.

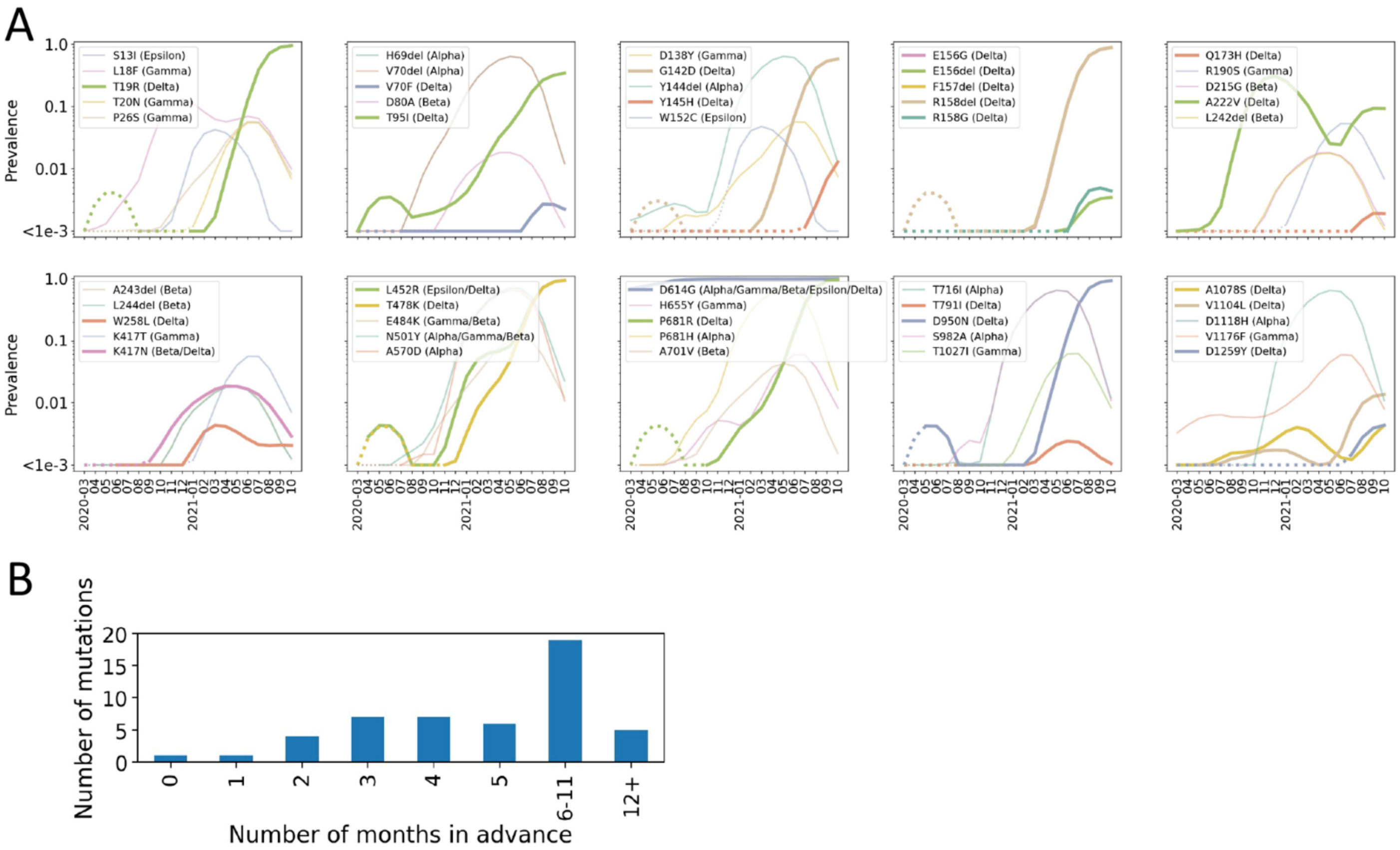
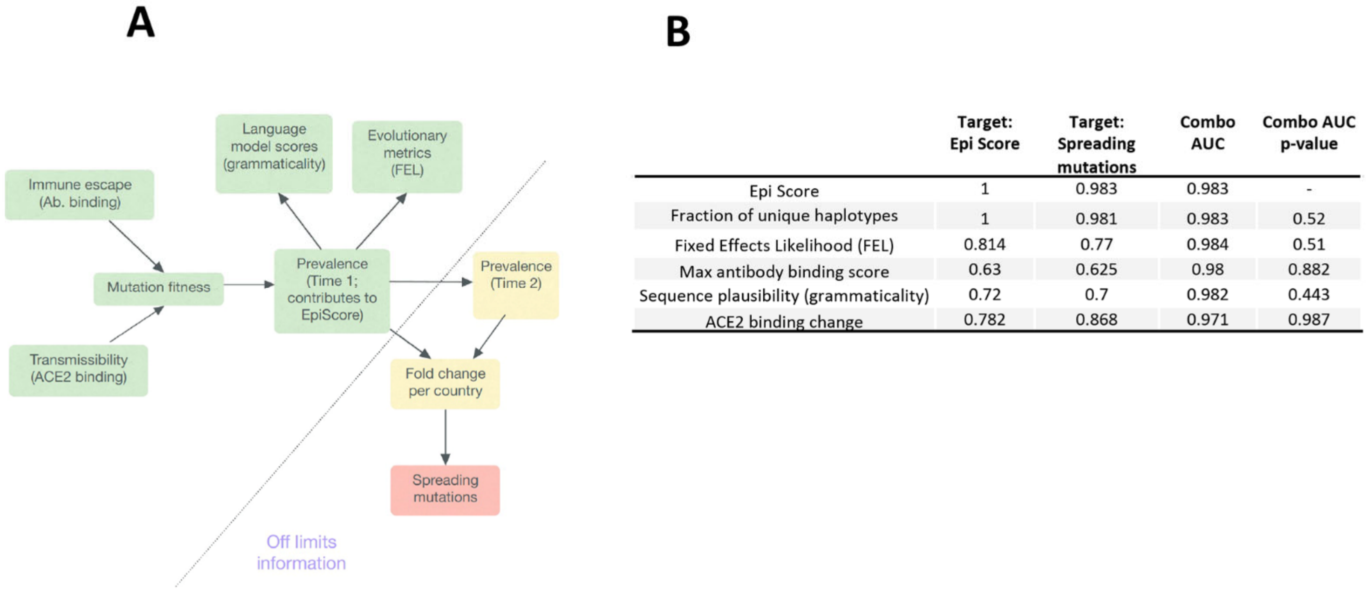


Fig. 2. Early detection of variant mutations. (A) Depiction of where in their growth trajectories current and former VOC/VOI mutations were first forecast to spread. Dotted lines denote the part of the curve where the variant had not yet been forecast to spread. Solid lines denote the period after first forecast. Delta-defining variants are shown by thick lines. Mutations are presented in genomic order. (B) The number of months between when the mutations presented in (A) were forecast and when they reached a prevalence of 1% globally.


A
Language model scores (grammaticality)
Evolutionary metrics (FEL)
Immune escape (Ab. binding)
Mutation fitness
Prevalence (Time 1; contributes to EpiScore)
Prevalence (Time 2)
Transmissibility (ACE2 binding)
Fold change per country
Spreading mutations
Off limits information
B

	Target: Epi Score	Target: Spreading mutations	Combo AUC	Combo AUC p-value
Epi Score	1	0.983	0.983	-
Fraction of unique haplotypes	1	0.981	0.983	0.52
Fixed Effects Likelihood (FEL)	0.814	0.77	0.984	0.51
Max antibody binding score	0.63	0.625	0.98	0.882
Sequence plausibility (grammaticality)	0.72	0.7	0.982	0.443
ACE2 binding change	0.782	0.868	0.971	0.987

Fig. 3. Epi Score mediates effects captured by other data sources. (A) Causal model: mutation fitness drives viral prevalence at time 1 (as measured by global frequency, and geographic and haplotype distribution, Epi Score). Language model score or evolutionary metrics are summaries of GISAID data and therefore are shaped by mutation prevalence. Prevalence at time 1 predicts prevalence at time 2, which ultimately leads to mutation being defined as spreading. Therefore, prevalence at time 1 (as captured by Epi Score) mediate the effects of the biological variables that enhance viral fitness through transmissibility or escape adaptation. **(B)** To quantitatively test for mediation, we assessed whether variables were better at predicting mutations in the top 5% of Epi Scores, compared to spreading mutations for time 2 versus time 1. "Combo AUC" refers to the combined AUC of that variable with Epi Score. Significant improvements of the combined model over that of Epi Score alone would indicate complementarity, and therefore predictive information not captured by Epi Score alone.

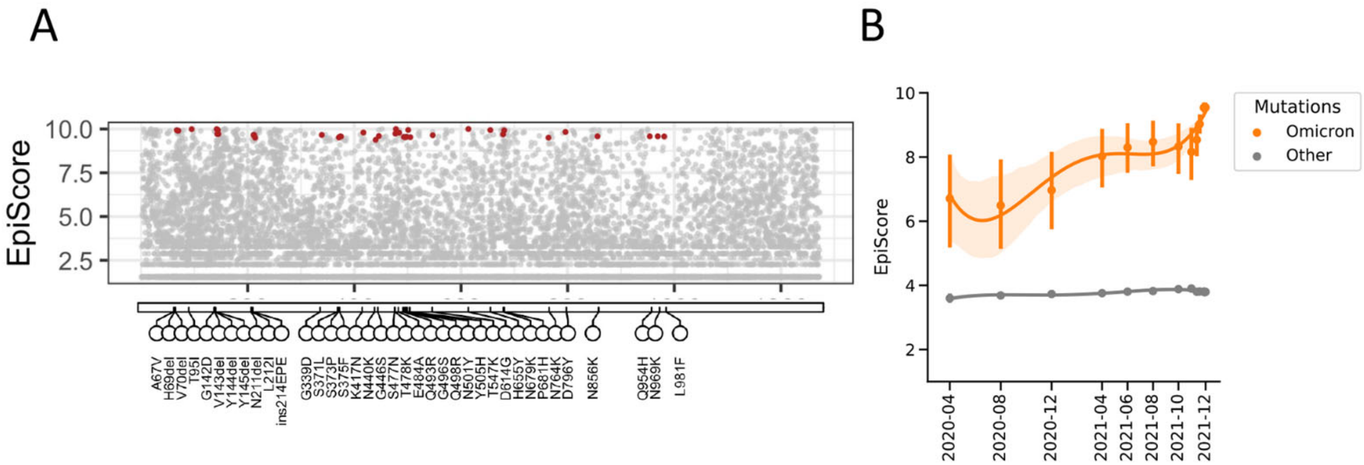
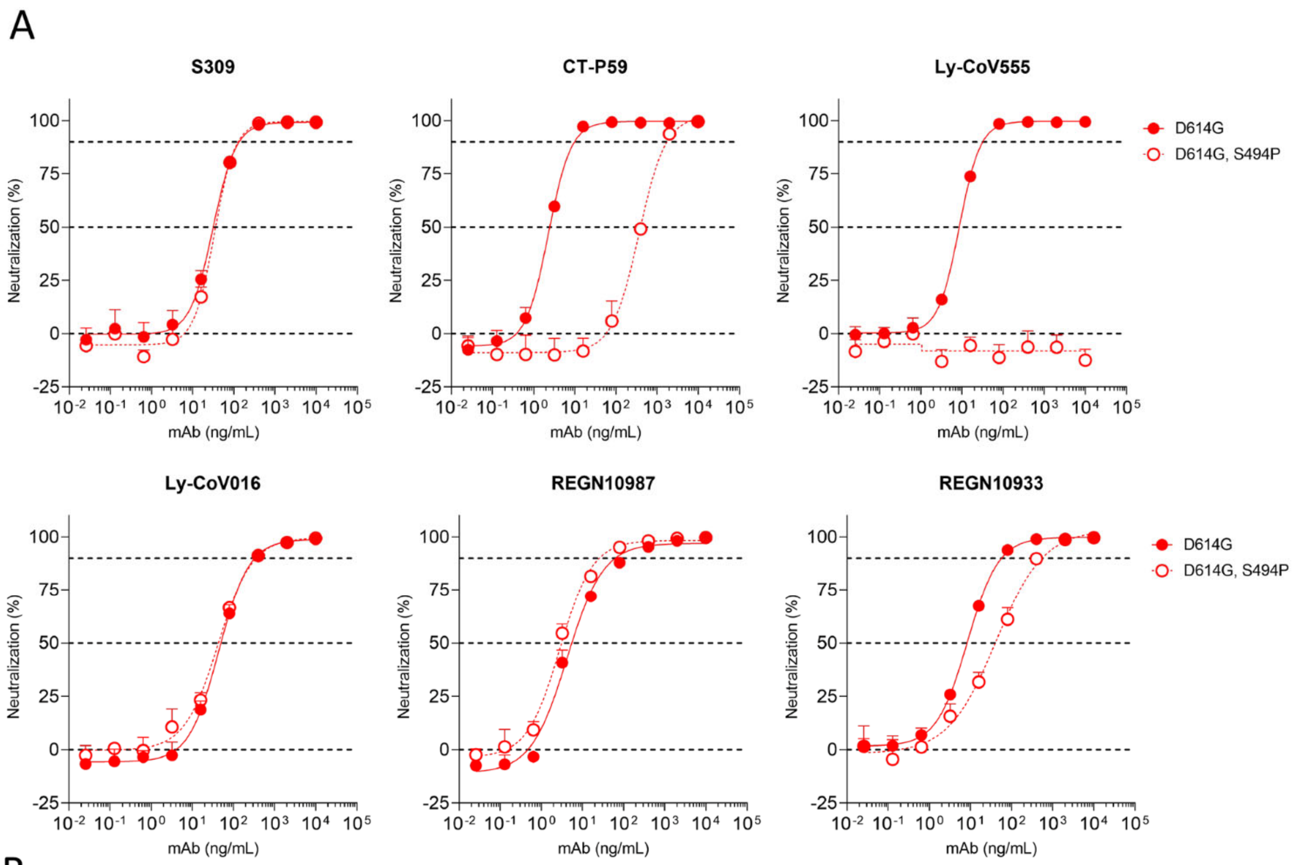


Fig. 4. Emergence and spread of Omicron. (A) The Epi Scores of 37 Omicron-defining mutations are shown as of December 8, 2021 (red dots). **(B)** Although some of the mutations in Omicron already had very high Epi Scores and were widely spread, emergent mutations were distinguished by the progressively increasing Epi Score between April 2020 and August 2021 preceding the rapid acceleration at the end of 2021. Shown are mean and confidence interval Epi Score values. Other: Epi Score of all other mutations in the SARS-CoV-2 spike.



B

Antibody	EC50 (ng/mL)		Fold-change
	D614G	D614G, S494P	
S309	35.6	40.7	1.1
CT-P59	2.3	393.0	172.1
Ly-CoV555	8.2	>10,000	>1213
Ly-CoV016	47.1	42.6	0.9
REGN10987	6.5	3.8	0.6
REGN10933	6.7	38.6	5.9

Fig. 5. S494P mutation decreases neutralization potential of three clinically approved therapeutic antibodies. (A) VSV-SARS-CoV-2 pseudovirus was generated based on the "Wuhan-Hu-1" sequence with either the D614G mutation or D614G and S494P mutations. Virus neutralization was measured in a microneutralization assay on Vero E6 cells. Example results from one repeat are shown. **(B)** EC50 values and fold-changes were calculated from two independent experiments. S309 is the parent molecule of VIR-7831, which had been previously evaluated on the S494P variant and showed no change in neutralization (*25*).

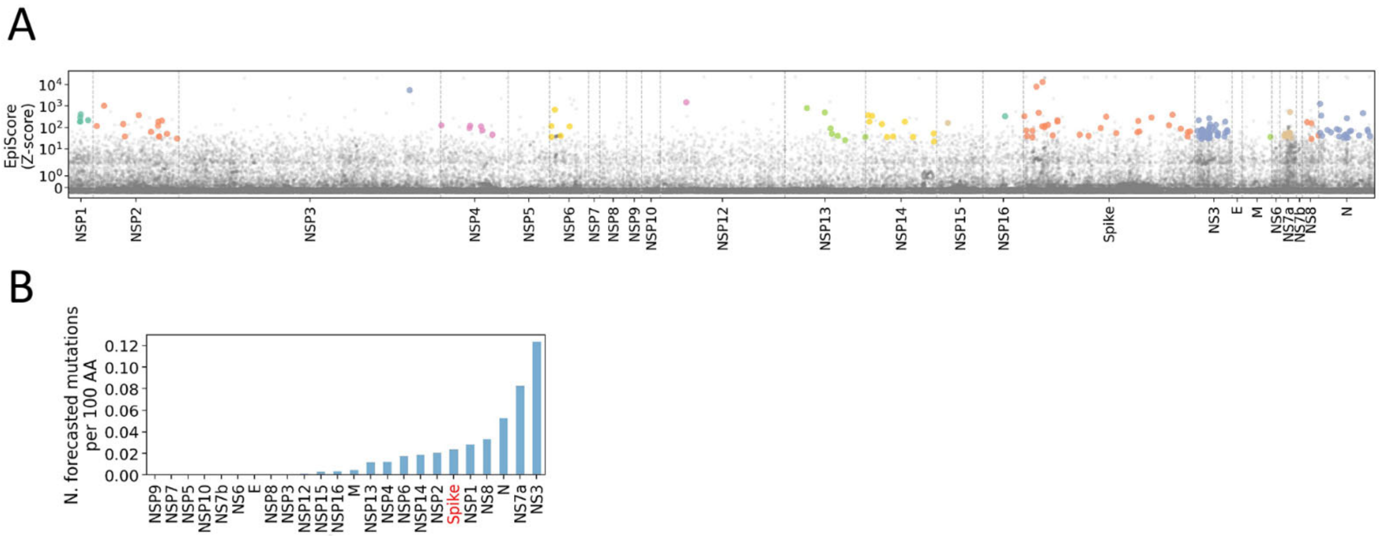


Fig. 6. Manhattan-style plot of Epi Scores across the SARS-CoV-2 Delta proteome. (A) For visualization purposes, Epi Scores have been calculated as Z-scores, which correlate to the default, rank-based calculation as a spearman R > 0.99. Points highlighted in color occur at a frequency over 0.1% on a Delta background (B.1.617.2 + AY lineages) and occur at significantly positively selected sites (FEL FDR-adjusted q-value < 0.05). All mutations occurring at over 80% frequency, in the lineages accounting for >80% of all Delta cases, were excluded from the visualization. Thus, the plot serves to highlight variants predicted to spread and under positive selection in the current Delta background. For a complete listing, Suppl. File S2. **(B)** The rate per 100 amino acids of highlighted forecasted mutations from panel A, per gene in the SARS-CoV-2 proteome.

Table 1. Summary of analytical features. A total of 48 parameters for 14 variables were created for 5 feature groups. These features capture evolutionary, immune, epidemiologic, transmissibility, and language model predictors of mutation spread. A detailed description of all parameters is included in **data file S1**.

Feature group	Variable	Meaning	Source or reference	Number of parameters
Evolution	Positive selection (FEL, MEME)	Parameters from Fixed Effects Likelihood (FEL) and Mixed Effects Model of Evolution (MEME)	HyPhy (*19*)	11
	Codon-SHAPE	RNA SHAPE constraint	Manfredonia *et al.* 2020 (*32*)	3
	Viral entropy	Shannon entropy at each codon position for an amino acid site	This work	3
Immune	CD8 epitope escape	The frequency of SARS CoV-2 mutations in cytotoxic lymphocyte (CTL) epitopes	Agerer *et al.* 2021 (*15*)	1
	CD8 response	The percent and average CD8+ T cell response to an epitope in patients	Tarke *et al.* 2021 (*33*)	2
	CD4 response	The percent and average CD4+ T cell response to an epitope in patients	Tarke *et al.* 2021 (*33*)	2
	Antibody binding score	The estimated percent contribution of a site to binding of the indicated antibody, as estimated by Molecular Operating Environment (MOE)	This work	17
	Maximum escape fraction in vitro	The maximum escape fraction across all conditions for that mutation	Greaney *et al.* 2021 (*34*)	1
Epidemiology	Variant frequency	The percent of sequences with the mutation	Calculated from GISAID (*2*)	1
	Fraction of unique haplotypes	The fraction of unique Spike haplotypes in which a mutation is observed	Calculated from GISAID (*2*)	1
	Number of countries	The number of countries where it has been observed.	Calculated from GISAID (*2*)	1
	Epi Score	The exponentially weighted mean rank across the other epidemiology variables	Calculated from GISAID (*2*)	1
Transmissibility	RBD expression change	Change in RBD expression due to the mutation	Starr *et al.* 2020 (*13*)	1
	ACE2 binding change	The change in binding affinity for ACE2	Starr *et al.* 2020 (*13*)	1
Language model	Language model	Grammaticality and semantic change of a mutation	Hie *et al.* 2021 (*17*)	2

Table 2. Forecasted mutations for therapeutic antibodies. Forecasted mutations, as of October 19th (including VOC mutations) were intersected with the binding epitopes of therapeutic monoclonal antibodies. Mutations were included if they were in sites contributing at least 1% of the total binding energy for a given antibody, as estimated by Molecular Operating Environment (MOE) program. Mutations known to decrease antibody EC50 more than five-fold are marked with asterisks. Mutations with daggers indicate neutralization is decreased less than five-fold (https://covdb.stanford.edu/page/susceptibility-data/), whereas values with double daggers indicate untested antibody, mutation combinations.

Clinical therapeutic antibody	**Forecasted mutations in epitopes**
VIR-7831 (sotrovimab)	A344S†, R346K†
LY-CoV016 (etesevimab)	K417T‡, K417N*, L455F‡
REGN10987 (imdevimab)	R346K†, K444N*, G446V*
LY-CoV555 (bamlanivimab)	L452R*, L452Q‡, V483F†, E484K*, E484Q*, F490S*, S494L‡, S494P*
REGN10933(casirivimab)	K417T*, K417N*, L455F*, G476S*, S477I‡, T478K‡, E484K*, E484Q*, F490S*
CT-P59	K417T‡, K417N†, L452R*, L452Q‡, L455F‡, E484K*, E484Q‡, F490S‡, S494L‡, S494P‡