Patterns of Volatility Across the Spike Protein Accurately Predict the Emergence of Mutations within SARS-CoV-2 Lineages

Roberth A. Rojas Chávez[1], Mohammad Fili[2], Changze Han[1], Syed A. Rahman[3], Isaiah G. L. Bicar[1], Guiping Hu[4], Jishnu Das[3], Grant D. Brown[5], and Hillel Haim[1#]

[1] Department of Microbiology and Immunology, The University of Iowa, Iowa City, IA.
[2] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA.
[3] Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA.
[4] Department of Sustainability, Rochester Institute of Technology, Rochester, NY.
[5] Department of Biostatistics, College of Public Health, The University of Iowa, Iowa City, IA.

[#]To whom correspondence should be addressed:
Hillel Haim, MD, PhD
Department of Microbiology and Immunology
The University of Iowa
51 Newton Rd, 3-770 BSB
Iowa City, Iowa, 52242
Phone: (319) 335-9989
Email: Hillel-haim@uiowa.edu

Short Title: Forecasting emergence of SARS-CoV-2 variants

Keywords: SARS-CoV-2, COVID-19, Virus evolution, Spike protein, Prediction model.

Abstract word count: 163.

ABSTRACT

New lineages of SARS-CoV-2 are constantly emerging. They contain mutations in the spike glycoprotein that can affect virus infectivity, transmissibility, or sensitivity to vaccine-elicited antibodies. Here we show that the emergence of new spike variants is accurately predicted by patterns of amino acid variability (volatility) in small virus clusters that phylogenetically-precede or chronologically-predate such events. For each spike position, volatility within the virus clusters, volatility at adjacent positions on the three-dimensional structure of the protein, and volatility across the network of co-volatile sites describe its likelihood for mutations. By combining these variables, early-pandemic sequences accurately forecasted mutations in lineages that appeared 6-13 months later. The patterns of mutations in variants Alpha and Delta, as well as the recently-appearing variant Omicron were also predicted remarkably well. Importantly, probabilities assigned to spike positions for within-lineage mutations were lineage-specific, and accurately forecasted the observed changes. Sufficient antecedent warning of the imminent changes in SARS-CoV-2 lineages will allow design of immunogens that address their specific antigenic properties.

SIGNIFICANCE

New variants of SARS-CoV-2 continue to emerge in the population. Due to mutations in the spike protein, some variants exhibit partial resistance to therapeutics and to the immunity provided by COVID-19 vaccines. Thus, there is a need for accurate tools to forecast the appearance of new virus forms in the population. Here we show that patterns of amino acid variability across the spike protein accurately predict the mutational patterns that appeared within SARS-CoV-2 lineages with considerable advance warning time. Interestingly, mutation probabilities varied greatly between lineages, most notably for critical sites in the receptor-binding domain of spike. The high predictive capacity of the model allows design of vaccines that address the properties of variants expected to emerge in the future.

INTRODUCTION

Since emerging in December 2019, SARS-CoV-2 has caused devastating effects worldwide. By December 2021, more than 5 million deaths have been attributed to the infection, and estimated economic losses greater than $10 trillion are expected by the end of 2022 (1, 2). Mutations in the SARS-CoV-2 genome give rise to new forms of its proteins; their emergence is monitored through sequence-based surveillance studies of the population (3). Most mutations that impact SARS-CoV-2 infection are found in the spike protein that adorns the virus surface. Spike mediates fusion with host cells and is the primary target for antibodies elicited by infection or vaccination (4). Mutations in spike can affect disease progression rate, virus transmissibility, and sensitivity to vaccine-elicited antibodies and therapeutics (5). Notably, some mutations have appeared independently in diverse SARS-CoV-2 lineages (6, 7). Such patterns of convergence suggest that similar selective pressures are applied on the virus in different individuals and populations.

COVID-19 vaccines effectively reduce SARS-CoV-2 infection rates and spread. However, the emergence of new SARS-CoV-2 variants with high transmission rates or resistance to vaccine-elicited antibodies has suggested the need to update the currently-applied immunogens (8). While RNA-based vaccines can be rapidly produced relative to protein-based immunogens, several months are required for clinical testing before manufacture and distribution of the vaccine (9). Such timelines limit our ability to rapidly address the appearance of new virus forms in the population. Therefore, there is an urgent need for accurate tools to define the mutational landscape of spike, in order to anticipate the specific changes expected to occur in each lineage. To this end, several approaches have been applied. Most commonly, phylogenetic tools are used to identify codons under positive selection (10). However, since many mutations in spike occur at evolutionarily neutral sites, estimates of positive selective pressures are not sufficient to predict appearance of mutations at all positions of this protein (11, 12). Furthermore, such tools have limited utility to forecast insertion or deletion events, which frequently occur in spike (13). Other approaches have also been used to predict changes in SARS-CoV-2 proteins. A recent study by Maher and colleagues explored multiple predictors, including epidemiological measures of variant spread and effects of the mutations on biological properties of the spike protein (14). Their model based on epidemiological data exhibited good sensitivity and specificity for predicting some mutations up to four months in advance. An interesting study by Rodriguez-Rivas and colleagues applied an epistasis-based model, developed using sequences of non-SARS-CoV-2 coronaviruses (15). Their results

corresponded well with fitness profiles of sites in the receptor-binding domain of spike and with sequence diversity patterns of the protein in the population. Nevertheless, higher-performance tools are needed to predict the precise mutations that appear and to provide greater antecedent warning times (16). Importantly, additional knowledge is required of the lineage specificity of the mutational landscape of the spike protein, to determine if each mutation has a similar likelihood to appear within each of the SARS-CoV-2 variants.

The "noise" in biological systems often contains information that describes future states. For example, we previously described the patterns of in-host variability in antigenic features of the HIV-1 envelope glycoproteins (Envs) (17). We discovered that each feature has a "characteristic" level of variability within the host that is conserved among different individuals. Interestingly, the in-host variability in Env epitopes measured in a small number of patient samples from the 1980s accurately predicted the loss of the epitopes in the population during the next three decades. Thus, the variability in small segments of the population (i.e., within an infected individual) can predict the changes that occur at a system level. Based on this relationship, we hypothesized that the emergence of new lineage-dominant mutations in SARS-CoV-2 spike can be forecasted by patterns of amino acid variability in small groups of viruses that predate or phylogenetically precede the changes. To test this hypothesis, we partitioned spike sequences from early stages of the COVID-19 pandemic into small clusters. Within each cluster, we calculated for each spike position: **(i)** The level of amino acid variability, **(ii)** Amino acid variability at adjacent positions on the three-dimensional structure of the protein, and **(iii)** Amino acid variability at sites that exhibit co-occurrence of variability with the site of interest. These measures of positional and "environmental" variability were applied to a model that assigns a probability to each spike position for emergence as a new lineage-dominant mutation. Using a small number of sequences from the early pandemic, the model exhibited remarkable performance in predicting the mutations that appeared in SARS-CoV-2 lineages 6-13 months later. Our findings suggest that the mutational landscape of spike is diversifying; each position exhibits a distinct likelihood for mutations in each SARS-CoV-2 lineage. This study demonstrates the large amount of information contained in the patterns of variability within small subsets of the virus population. Importantly, we reveal the surprising lineage-specific and predictable nature of the mutations that arise in SARS-CoV-2, which can be applied to address future variants of this virus.

RESULTS

Spike positions with high volatility appear as sites of mutation in SARS-CoV-2 lineages

We considered a model whereby the likelihood for emergence of a new lineage-dominant mutation at any spike position *p* is determined by permissiveness of *p* to accommodate non-ancestral residues. We further hypothesized that this permissiveness is proportional to the level of amino acid variability at *p* in any subgroup of the virus that phylogenetically precedes the emergence event. To calculate sequence variability at each position, we divided all SARS-CoV-2 spike sequences into groups and subgroups (clusters). Nucleotide sequences of 615,374 SARS-CoV-2 spike genes from samples collected worldwide between December 2019 and July 2021 were used. To reduce the impact of sequencing errors, we excluded all sequences with character ambiguities and those that appeared only once, and the remaining dataset was aligned and "compressed" to obtain a single representative for each unique sequence. A unique-sequence approach allowed us to focus on the diversification pattern of the spike protein, independent of its rate of spread in the population. Evolutionary relationships among the 16,808 unique sequences were inferred and a maximum likelihood phylogenetic tree was constructed (see Methods and **Figure 1A**). We then partitioned the tree into discrete groups separated by a minimal distance of 0.004 nucleotide substitutions per site. As expected, many groups corresponded to known SARS-CoV-2 lineages. We define the groups by phylogeny rather than by established designations (e.g., the Pango system) because assignments in the latter are based on mutations in the whole SARS-CoV-2 genome rather than spike alone (e.g., see partition of the Iota variant into three groups in **Figure 1A**). We then distinguished between the baseline groups (collectively colored in grey in **Figure 1A**) and the terminal emergent groups (G_{T1}-G_{T8}) using a threshold of 0.0015 substitutions per site between the centroid of each group and the SARS-CoV-2 spike ancestral sequence. All groups are described in **Table S1**.

We quantified amino acid variability at each position of spike within the baseline sequences. To this end, all baseline groups were partitioned into clusters of 50 sequences (**Figure 1B**). For every spike position, we determined in each cluster the absence or presence of variability (assigned values of 0 or 1, respectively). We then calculated the mean variability at each position by averaging these values across all clusters of the baseline. We designate this cluster-averaged measure of amino acid variability "volatility" (*V*). Such a cluster-based approach quantifies the frequency of mutation events rather than frequency of the mutants.

Thus, any cluster of 50 sequences in the baseline group that contains a non-ancestral residue but no variability is assigned a variability value of 0 (see bottom cluster in **Figure 1B**).

Volatility values of spike positions were compared with the emergence of mutations at these sites in the SARS-CoV-2 groups. We define two types of emerging mutations: **(i) A group-dominant mutation (GDM)**, which is found in the group ancestor and in at least 50% of all sequences from that group, and **(ii) A subgroup-emerging mutation (sGEM)**, which is not found in the group ancestor and represents a clonal expansion of less than 50% of all group sequences (see examples in **Figure S1A**). A total of 43 GDMs and 16 sGEMs were detected in the baseline and terminal groups (see **Table S1**). We observed that most positions with high volatility values (as calculated using baseline sequences) emerged as GDMs or sGEMs in the baseline or terminal groups (see positions of spike subunit S1 in **Figure 1C** and of subunit S2 in **Figure S1B**). Of the positions with the highest volatility values, most appeared as GDMs or sGEMs in at least one group, often in both baseline and terminal groups (**Figure 1D**). To verify that GDMs or sGEMs in the baseline do not impact volatility values, we excluded from the baseline all clusters that compose GDMs or sGEMs and then recalculated volatility values. Consistent with our intention to represent the frequency of mutation events in the baseline, depletion of these clusters showed little impact on volatility values (**Figure S1C**).

GDM and sGEM sites were more volatile than sites with no such mutations (**Figure 1E**). Furthermore, non-volatile sites in the baseline did not emerge with GDMs or sGEMs in any baseline or terminal group (**Figure 1F**). In most cases, the minority variant with the highest frequency in the baseline group was also the emergent residue in the terminal groups (**Figure 1G**). Therefore, a high level of positional volatility in the baseline group precedes (as inferred phylogenetically) the emergence of GDMs or sGEMs in the terminal groups. This finding is indeed intuitive – a high frequency of mutations at a given site increases its likelihood to appear in any new emerging lineage.

High volatility at adjacent positions on the spike trimer is associated with appearance of GDMs and sGEMs

We recently developed a machine learning algorithm to characterize the spatial clustering patterns of amino acid variability on the HIV-1 Env protein (unpublished data). We found that the in-host variability at most Env positions can be accurately estimated by the variability at adjacent positions on the three-dimensional structure of the protein. We

hypothesized that SARS-CoV-2 spike positions with high volatility (in the population) may exhibit similar patterns of spatial clustering, and that a high-volatility "environment" may increase the likelihood for emergence of mutations. As expected, mapping of the baseline volatility values onto the structure of the spike trimer (18) demonstrated several clusters of high-volatility positions, most notably in the N-terminal domain (NTD, **Figure 2A**). Many of these positions exhibited significantly higher likelihoods for a volatile state when their adjacent positions were also volatile (see **Figure 2B** for results of the permutation test described in the Methods section). We hypothesized that if such associations are stable over time, then the likelihood for future changes at any position of spike may be associated with volatility of its neighboring positions. To this end, we generated a variable (designated *D*) that describes for each position *p* the total environmental volatility:

$$D_p = \sum_{j=1}^{n} \frac{1}{\Delta_{pj}} \cdot V_j \qquad \textbf{[Eq. 1]}$$

where *n* is the number of positions *j* within 6 Å of position *p*, Δ_{pj} is the distance between the closest two atoms of positions *p* and each position *j*, and V_j is the volatility at each position *j*. Similar to the volatility values (**Fig 1E**), *D* values were higher for positions that emerged with GDMs or sGEMs (**Figure 2C**). Furthermore, none of the positions with a *D* value of zero in the baseline emerged with a GDM or sGEM (**Figure 2D**), suggesting that a high-volatility environment increases the likelihood for their occurrence.

Co-volatility patterns across the spike protein identify positions with high likelihoods for emergence as GDMs or sGEMs

We hypothesized that the co-occurrence of volatility at adjacent positions on the trimer can be generalized to describe associations that are not dependent on physical proximity (i.e., that presence of a volatile state at a given position is associated with presence of a volatile state at a specific set of other positions). To test this hypothesis, we used all 114 baseline clusters to calculate the co-occurrence of volatility at any two spike positions using Fisher's exact test (see schematic in **Figure 2E**). P-values of the test were then used to construct a co-volatility network, whereby the edges that connect the nodes (positions) are defined by the statistical significance of the association between volatility patterns of the positions (see distribution of P-values in **Figure S2A** and example of a network segment in **Figure S2B**).

To determine the significance threshold to apply for network construction, we examined structural properties of the network and its robustness to random deletion of edges. Two network topological metrics were assessed: **(i)** Degree distribution, which describes the average number of connections each node has with other nodes, and **(ii)** Closeness centrality, which describes for each node the sum of the path lengths to all other nodes in the network (more central nodes have lower values) (19). For robust scale-free networks, such random deletions only minimally perturb their topological properties (20). We found that networks defined at a more stringent significance threshold ($P<0.01$) were more robust to edge deletions, with minimal impact on both degree distribution and closeness centrality (**Figure S2C and S2D**). By contrast, when less stringent significance thresholds were used, the number of edges was greater (i.e., they contained more information regarding the co-volatile positions). This suggested that an intermediate significance threshold would provide a sufficiently stable network without losing most information.

We next examined whether, for any position *p* of spike, presence of high volatility at its network-associated co-volatile sites (*q*) increases the likelihood for emergence of mutations. To this end, we generated a simple measure (*R*) designed to capture for each spike position *p* the total volatility of its network "neighbors" *q* (q_1, q_2, … q_n), using a P-value of 0.05 as the threshold:

$$R_p = \sum_{q=1}^{n} w_{pq} \cdot V_q \qquad \textbf{[Eq. 2]}$$

where *n* is the number of network-neighboring positions for position *p*, V_q is the volatility at each position *q* calculated using the baseline sequences, and w_{pq} is the evidence for association between volatility of position *p* and each of its positions *q* (calculated as the $-\log_{10}$(P-value) in Fisher's test). As shown in **Figure 2F**, positions with the highest *R* values in the baseline group emerged with GDMs or sGEMs in the baseline and terminal groups (see values for all spike positions in **Figure S3A**). *R* values were significantly higher for positions with GDMs or sGEMs relative to positions with no such mutations (**Figure 2G**). Furthermore, an *R* value of zero in the baseline was invariably associated with lack of GDM or sGEM appearance in the baseline or terminal groups (**Figure 2H**). Overall, the *V* and *R* values for any position correlated well, and considerably better than their correlation with *D* (**Figure S3B**). Nevertheless, several key positions of spike that emerged with GDMs showed high *R* values but relatively low *V* values, including position 452 in the RBD, positions 141-143 in the N-terminal domain (NTD), position 950 in the S2 subunit and position 679 near the furin cleavage site (data not shown). Therefore,

for any spike position, high volatility at its network-associated sites (calculated using the baseline sequences) describes the likelihood for its emergence as a GDMs or sGEMs.

We compared the volatility-based variables with a measure of the positive selection pressures applied on each site. To quantify positive selection, we used the baseline sequences to calculate for each codon the difference between the nonsynonymous changes (dN) and synonymous changes (dS). All codons with negative dN-dS values were assigned a value of zero. Thus, this variable (designated *S*) quantifies the strength of the positive selective pressures applied on each site (see comparison with a standard dN-dS metric in **Figure S3C**). *S* values were high for many positions with GDMs and sGEMs (**Figure 2I**) and correlated moderately with the *V* and *R* values (**Figure S3D**). Nevertheless, many positions with an *S* value of zero in the baseline still emerged as GDMs or sGEMs (**Figure 2J**). Furthermore, the performance of *S* to predict emergence of GDMs or sGEMs was lower than that of *V* or *R* (**Figure 2K).** A notable limitation of the synonymous and nonsynonymous substitution rates as predictors of changes is their inability to be computed for sites of deletion (e.g., positions 69, 70 and 144 in $G_{T1}(\alpha)$ or positions 156 and 157 in $G_{T3}(\delta)$). By contrast, high *V* and *R* values were assigned to these sites (**Figure 1D** and **Figure 2F**).

Therefore, the likelihood for emergence of a GDM or sGEM at any spike position is associated with its volatility, as well as the volatility at adjacent positions on the protein and at associated sites on the co-volatility network.

Volatility profiles in sequence clusters from the early pandemic predict appearance of mutations in the lineage-emerging phase

We examined the ability of the four variables (*V*, *R*, *D* and *S*) to forecast changes in spike. Specifically, we tested whether viruses that temporally preceded emergence of SARS-CoV-2 lineages can predict appearance of lineage-dominant mutations at future time points. To this end, sequences were classified by their Pango lineage designations rather than our spike-based group definitions. We first determined the formation time of each lineage, defined here as the date by which 26 unique nucleotide sequences from the lineage were detected (see **Figure 3A** and **Table S2**). Based on lineage formation timelines, we decided to apply sequences from samples collected between December 30th 2019 and September 19th 2020 as the "early-phase" group that is used to predict emergence of mutations in lineages that formed between October 10th 2020 and June 12th 2021 (**Table S2**). We designate these latter **lineage-defining**

mutations (LDMs). The early-phase group was composed of 1,760 unique sequences, which included only one sequence from SARS-CoV-2 lineage B.1.1.7 (WHO variant designation Alpha) and none from the major variants Epsilon, Iota, Gamma or Delta. Six minor lineages emerged relatively early in the pandemic, which contained mutations at positions 614, 222 and 477 (see **Table S2**). To avoid a potential bias, the three positions were excluded from these analyses. A total of 67 LDM sites were identified in the lineage-emerging phase.

We then divided the early-phase sequences into 36 clusters of 50 unique sequences, which were used to calculate *V*, *R* and *D* values for all spike positions. We also calculated the *S* value using all early-phase sequences. These values were compared between the LDM sites of different SARS-CoV-2 lineages and sites with no such mutations (**Figure 3B**-**3E**). For LDM sites in some variants, the *V* and *R* values were modestly higher than the values in the no-mutation sites. No differences were observed between *D* or *S* values at LDM sites in any of the variants and the no-mutation sites. We hypothesized that a combination of the volatility-based variables (*V*, *R* and *D*) would exhibit higher performance as a predictor of emerging mutations than each of them separately. To this end, we used a logistic regression model that applies *V*, *R* and *D* values of the early-phase sequences to calculate the probability of each site to emerge with an LDM in the lineage-emerging phase (see Methods). Remarkably, for all SARS-CoV-2 variants, the probabilities calculated for LDM sites were significantly higher than probabilities assigned to the no-mutation sites (**Figure 3F**).

To examine the evolution of the volatility-based variables in the early stages of the COVID-19 pandemic, we calculated *V*, *R* and *D* values at different time points of the early phase. In addition, we examined the changes in the probabilities assigned by the combined model. We observed that the pattern of emerging LDMs was predicted with high sensitivity and specificity by the time 5 clusters were formed (249 unique sequences), corresponding to samples collected before April 1st 2020 (**Figure 4, A-C**). Of the individual predictors, *R* exhibited the highest performance, modestly lower than the combined model. We further analyzed the changes in *R* values assigned to the specific sites-of-emergence in the highly-prevalent SARS-CoV-2 variants Alpha and Delta (B.1.617.2). For variant Alpha, five of the nine sites exhibited *R* values in the 95th percentile by April 1st 2020 (see **Figure 4D** and all variables in **Figure S4A**). For variant Delta, four of the nine sites-of-emergence also showed high *R* values at the above early time point (**Figure 4E** and **Figure S4B**).

We further examined the predictive performance of the first 249 unique sequences. Higher probabilities were assigned by these sequences to LDM sites of lineages that emerged

at earlier stages of the pandemic (**Figure 4F** and **Figure S4C**). Higher probabilities were also assigned to convergent sites (i.e., those that emerged with LDMs in multiple lineages) (**Figure 4G** and **Figure S4D**). We examined the classification metrics for the probability values assigned by the first 249 or all 1,760 early-phase sequences. Using a probability of 0.5 as the cutoff value (i.e., the decision threshold of the algorithm), high levels of sensitivity, specificity, accuracy and recall were observed, indicating a low false-negative rate (**Figure 4H**). By comparison, the level of precision calculated for this threshold was low, reflecting an apparently large number of false-positive predictions. We note that the indicated precision over-estimates the false positive rate due to our definition of LDMs, whereby only mutations that are contained in more than 50% of all lineage strains are considered LDM sites. Thus, many sites that are emerging *within* lineages (i.e., equivalent to the sGEMs in the phylogeny-indexed analyses) were classified as "non-emergent". We also note that the false positive rate decreased with increasing probability values, resulting in a gradual increase in precision (**Figure 4I**). For positions assigned probability values within the 98th percentile, a precision level of approximately 0.5 was observed.

Taken together, these findings show that a high level of volatility at any site and at its spatial- and network-associated sites precedes (temporally) emergence of LDMs in the population. Volatility profiles calculated using a small number of unique sequences (e.g., 249 collected until April 1st 2020) can predict with high sensitivity and specificity the LDMs that would appear 6 to 13 months later. Thus, clear indications of the sites-of-emergence can be identified at very early stages of the pandemic.

Mutations in the SARS-CoV-2 Omicron variant are accurately predicted by the combined model

The SARS-CoV-2 variant Omicron (lineage B.1.1.529) emerged in November 2021. The first known case of infection occurred in South Africa; since then, it has rapidly spread worldwide (21). This variant contains a staggering 37 mutations in the spike protein, approximately two-thirds of which were not observed as LDMs in other SARS-CoV-2 lineages (22, 23). We examined the ability of the volatility-based model to predict emergence of these LDMs using sequences from samples collected in South Africa. Since the NCBI database, which served as the source for all sequences used in this study, contained only five SARS-CoV-2 sequences from South Africa, we applied data from the GISAID database (24). Sequences collected between March 6th 2020 and November 21st 2021 were used. All Omicron and Omicron-probable sequences were removed from this dataset. The final dataset was composed

of 269 unique nucleotide sequences, which were used to calculate *V*, *R* and *D* values that were applied as input for the logistic regression model. **Figure 5A** shows the probability percentiles assigned to the 36 LDM sites of Omicron. The insertion at position 214 was not included since our analyses focused on the 1,273 spike positions of the SARS-CoV-2 ancestral sequence. Of the 36 mutation sites in Omicron, 25 were assigned probabilities higher than the 0.5 decision threshold of the algorithm; of these, 15 sites were assigned probabilities in the 95th percentile and 12 in the 99th percentile. Fourteen of the mutation sites also appeared as LDMs in other SARS-CoV-2 lineages (see symbols above bars in **Figure 5A**). Of the remaining 22 Omicron-unique LDM sites, 15 were assigned probability values higher than the 0.5 decision threshold.

We examined the predictive capacity of the combined model using different sequence datasets as input. For predicting the 36 LDMs in Omicron, the 269 sequences from South Africa performed modestly better than the 5,700 baseline sequences (see black and grey bars in **Figure 5A** and classification metrics in **Figure 5B**). We compared this performance with predictions of the LDMs that appeared in variants Alpha and Delta, using the 249 early-phase sequences as input. Most classification metrics were higher for prediction of changes in lineages Alpha and Delta relative to Omicron (**Figure 5B**). Nevertheless, the distribution of probability percentiles assigned to the LDM sites in the variants differed considerably. For example, 33 and 44 percent of LDMs in the Omicron and Alpha variants, respectively, were assigned probabilities in the 99th percentile relative to 11 percent in the Delta variant (**Figure 5C**). Nevertheless, the overall performance of the volatility-based model to predict all lineage changes was still lower for the Omicron variant, reflecting a higher proportion of LDM sites with low *V*, *R* and *D* values.

Therefore, volatility patterns in 269 sequences from samples collected in South Africa until November 2021 predicted well most mutations in the Omicron variant. One-third of the Omicron LDM sites were assigned to the 99th probability percentile. However, relative to other variants, a higher proportion of the Omicron mutations exhibited low probability values.

Mutations that occurred within SARS-CoV-2 lineages are accurately predicted by the combined model

We tested the ability of the model to predict occurrence of within-lineage mutations. For this purpose, we indexed sequences by phylogeny rather than time (i.e., we applied our group-based assignments rather than the Pango lineage-based designations of LDMs). We focused

these studies on groups $G_{T1}(\alpha)$ and $G_{T3}(\delta)$. Both groups contain mutations that affect virus infectivity, neutralization sensitivity or transmission efficacy (25, 26). According to data collected until the end of July 2021, $G_{T1}(\alpha)$ contains six sGEMs (**Figure 6A**, right). In $G_{T3}(\delta)$, four sGEMs emerged until July 2021 (**Table S1**). To address the rapid expansion of $G_{T3}(\delta)$ between July and September (from 674 to 4,283 unique sequences), we used an extended $G_{T3}(\delta)$ dataset that includes sequences from samples collected until September 5th 2021. All emergent sublineages within $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ (i.e., clusters that contain the sGEMs as the dominant-cluster residues) were excluded from our datasets, and the remaining sequences were used to calculate the predictors *V*, *R* and *D*. These values were applied to the logistic regression model to assign a probability to each position for emergence as an sGEM within $G_{T1}(\alpha)$ or $G_{T3}(\delta)$. **Figures 6A** and **6B** show the 35 positions with the highest probabilities for mutations in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$, respectively. Remarkably, five of the six sGEM sites that appeared in $G_{T1}(\alpha)$ were among the top 16 mutations predicted to occur (see blue bars in **Figure 6A**). For $G_{T3}(\delta)$, 6 of the 12 sGEMs were among the sites assigned the highest probability scores (**Figure 6B**). We note that all sGEM sites in $G_{T1}(\alpha)$ were assigned higher probabilities by the $G_{T1}(\alpha)$ sequences than the probabilities assigned to them by the $G_{T3}(\delta)$ or baseline sequences (**Figure S5A**). Most sGEM sites in $G_{T3}(\delta)$ exhibited a similar pattern, suggesting that the likelihood for emergence of sGEMs is group specific. Lineage specificity of the predictions is described in the next section.

We also compared the predicted and observed residues at the sites of emergence. Consistent with the results shown in **Figure 1G**, for all sGEMs in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$, the minority variant with the highest frequency in each group also appeared as the new emergent residue (see characters above bars in **Figure 6**). Interestingly, high probabilities were assigned for reversion of several GDM sites in $G_{T1}(\alpha)$ to the SARS-CoV-2 ancestral residue (indicated by filled star symbols). For example, the sites of deletion in $G_{T1}(\alpha)$, at positions 69, 70 and 144, showed high probabilities for insertions (see sequence alignment of selected variants in **Figure S6B**). This finding is consistent with the high mutation rates at these sites (13). Several GDM sites in $G_{T3}(\delta)$ also showed high probabilities for reversion to the SARS-CoV-2 ancestral residue, including predicted changes D142G, N950D, del156E and G158R.

Many of the positions assigned high probabilities for emergence have known effects on SARS-CoV-2 infectivity, neutralization or transmission. For $G_{T1}(\alpha)$, such sites include: **(i)** L18F in the NTD, which increases resistance to antibodies (27), **(ii)** P479S, F490P and S494P in the RBD, which are also associated with resistance to antibodies (28, 29), and **(iii)** D427N and V367L in the RBD, which increase virus infectivity (30, 31). For $G_{T3}(\delta)$, many of the high-

probability mutations are also associated with resistance to neutralizing antibodies, including D80Y, Y28H, Y144del and H146Y in the NTD (27) or S494P in the RBD.

An example of the high performance of the combined model to predict within-lineage changes is the new lineage of the Delta variant designated AY4.2. This lineage appeared in October 2021 and contains two mutations in the NTD, namely A222V and Y145H. Notably, both sites exhibit high probabilities for emergence of mutations, and the highest-frequency minority variants in $G_{T3}(\delta)$ were the same as the emergent residues of AY4.2 (**Figure 6B**). Position 222 shows a high *S* value in $G_{T3}(\delta)$, whereas position 145 shows no indication of positive selection (see purple inverted bars in **Figure 6B**). Indeed, several sGEM sites in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ were assigned high probabilities but low non-significant *S* values. These sites, as well as the high-probability insertion events that cannot be assigned *S* values, highlight the contribution of volatility patterns to predicting the emerging mutations in SARS-CoV-2.

The mutational landscape of spike is lineage specific

To better understand the lineage specificity of the predictions, we examined the distribution of sites with high mutation probabilities on the cryo-EM structure of spike. Specifically, we compared the location of sites within the 95th probability percentile, as calculated using the baseline and $G_{T3}(\delta)$ sequence datasets (**Figures 7A**). As expected, many high-probability sites were located in the NTD. This domain contains an epitope that is targeted by multiple potent antibodies and is thus designated the "NTD supersite" (27, 32-35). The epitope is composed of loops N1, N3 and N5 of the S1 subunit (see **Figure 7B**). Interestingly, the sites with high probabilities for mutations in the baseline group and $G_{T1}(\delta)$ formed three clusters on the NTD supersite (**Figure 7C**): **(i)** Positions within the 95th percentile only in the baseline group, **(ii)** Positions within the 95th percentile only in $G_{T3}(\delta)$, and **(iii)** Positions within the 95th percentile in both $G_{T3}(\delta)$ and in the baseline. In most cases, considerable differences were observed between the mutation probabilities assigned by the baseline and $G_{T3}(\delta)$ sequences (see boxed regions comparing percentiles in **Figure 7C**).

We also compared the location of high-probability sites in the RBD, as calculated using sequences from the $G_{T1}(\alpha)$, $G_{T3}(\delta)$ and baseline groups. Again, considerable differences were observed in the probabilities assigned to each site by the three datasets (**Figure 7D**). Interestingly, all major RBD sites that impact antibody sensitivity showed lower probabilities for mutations to occur within $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ relative to their probabilities to occur from the

baseline (see also **Table S3**). For example, position 484 in the RBD, which impacts virus sensitivity to vaccine-induced immune sera (36), exhibits a high probability for mutations in the baseline but a low probability for mutations within the two lineages (**Table S3**). Similarly, position 501 that is converging to Tyr in diverse SARS-CoV-2 lineages (6), shows a lower probability in $G_{T3}(\delta)$ (the N501Y mutation is already found in the ancestor of $G_{T1}(\alpha)$). Such differences reflect the divergent volatility profiles of spike in these groups, which is also manifested by the distinct topologies of their co-volatility networks (**Figure S5C**). These patterns suggest a shift to a new state in the emergent lineages. This notion was further supported by the considerable differences in the inferred positive selective pressures applied on spike positions in the above groups. Indeed, many positions in the RBD that affect infectivity or antibody sensitivity exhibit lower *S* values in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ relative to the baseline group (**Table S3**). Analysis of the GDM sites in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ also revealed considerable changes in *S*. Interestingly, while several sites showed a decrease in S values upon transition from the baseline to the emergent groups, other sites showed dramatic increases in these values (**Table S4**). Therefore, similar to the distinct profiles of volatility, these results conform to a lineage specific state of spike.

Taken together, these findings show that patterns of volatility among strains that phylogenetically precede emergence of new sublineages can accurately predict the sites and identity of the mutations. The vast differences in the volatility profiles and selective pressures applied on spike positions suggest that the mutational landscape of this protein is evolving. Each position has a unique likelihood for emergence of mutations that is distinct for the viruses of each SARS-CoV-2 lineage.

DISCUSSION

New variants of SARS-CoV-2 are constantly appearing in the population. The mutations they contain in the spike glycoprotein impact virus infectivity, transmissibility or sensitivity to immune sera. To address the antigenic pattern of these new forms, including the recently-appearing hyper-mutated Omicron variant (37, 38), there are increasing calls for the design of new variant-specific vaccines (39, 40). Assuming persistence of SARS-CoV-2 in the population, and continuing emergence of new spike forms, the arms race between virus and vaccine is expected to be lengthy and costly. Thus, there is a clear need for accurate tools to forecast the antigenicity of variants expected to emerge in the future within each lineage. Standard phylogenetic tools can identify sites subjected to positive selective pressures; however, these

only constitute a minority of the mutations observed. At most other sites, mutations appear to be random and are thus regarded as unpredictable. Here we show that, in contrast to the above perception, the large majority of mutations that define SARS-CoV-2 lineages and those that are emerging as sublineages within them can be accurately forecasted using a small number of sequences that precede the emergence events. To this end, we apply a novel approach to calculate the likelihood of each position to appear as a lineage-dominant mutation. We show that the volatility profile of each position and volatility of its environment (i.e., network- and spatial-neighbors) contain sufficient information to predict such events with high sensitivity and specificity. Importantly, the predicted changes differ among the SARS-CoV-2 lineages. The surprising predictability of the mutations suggests that immunogens and therapeutics can be tailored to future population-dominant forms of spike expected to appear.

The volatility-based variables quantify the likelihood for occurrence of a mutation at each site. A high frequency of independent substitution events at a given site (quantified by volatility) is expected to increase the likelihood for its appearance in any emerging clonal lineage. In addition, we show that the emergence of mutations at spike positions is associated with volatility of their spatially-adjacent and network-associated sites (quantified by D and R, respectively). The spatial clustering of volatile sites is intuitive. Indeed, clustering on the linear sequence of the protein can be explained by mutational hotspots due to properties of the viral RNA (41, 42) or protein segments with high permissiveness for changes due to their limited impact on fitness (31). Clustering on the three-dimensional structure can also be explained by spike regions that are subjected to fitness or immune selective pressures. By contrast, the association between volatility of sites separated by larger distances on the protein is less intuitive. We propose that such associations describe the epistasis network of spike (i.e., the relationships between fitness profiles of different spike positions). Indeed, the volatility of each position likely captures its fitness profile; low volatility describes a state with a single high-fitness residue, whereas high volatility describes the presence of multiple residues with high fitness. Accordingly, we hypothesize that co-volatility patterns may capture the associations between the fitness profiles of the different sites. For example, a high R value for any position p describes its propensity for sequence variability due to permissiveness of its associated epistatic sites q. Therefore, such relationships may capture the adaptive sites q required to facilitate changes at site p. Comparisons of co-volatility network structure with structure of the epistasis network of spike, as determined by deep mutational scanning (31), will reveal the accuracy of the above hypothesis.

The mutational landscape of the spike protein was surprisingly lineage-specific; different patterns of changes were predicted for the baseline group, $G_{T1}(\alpha)$ and $G_{T3}(\delta)$. For example, different segments of the NTD neutralization supersite were assigned distinct probabilities for mutations (**Figure 7C**). Similarly, all major sites in the RBD that affect sensitivity to antibodies show high probabilities to occur from the baseline group but low probabilities to occur from $G_{T1}(\alpha)$ or $G_{T3}(\delta)$ (**Figure 7D** and **Table S3**). Furthermore, most sGEM sites in $G_{T1}(\alpha)$ or $G_{T3}(\delta)$ were assigned the highest mutation probabilities by sequences of the same group (**Figure S5A**). Based on the lineage-specific probabilities, the changes that occurred *within* them were predicted well: 5 of the 6 sGEMs in Alpha and 6 of the 12 sGEMS in Delta were assigned probability values in the 99th percentile. Similarly, the changes in the AY4.2 lineage of Delta were also assigned high probabilities for occurrence within this variant. These findings suggest that the fitness landscape of the spike protein is diversifying. Supportive of this notion are the considerable differences between the inferred positive selective pressures applied on spike positions in the different lineages (**Table S3** and **Table S4**) and the distinct structures of their co-volatility networks (**Figure S5C**). Such differences may reflect properties of the virus, but also the immune pressures applied by the host (e.g., by different proportions of vaccinated individuals in the groups).

Some lineage-dominant mutations allow the virus to adapt to fitness and immune selective pressures, whereas others are "hitchhikers" on the driver mutations (43). The drivers are subject to positive selection whereas the hitchhikers are mostly evolutionarily neutral or can exhibit reduced fitness (44, 45). In variants Alpha, Delta and Omicron, most mutations show no evidence for positive selection. Using our model, both drivers and hitchhikers are readily predicted by small numbers of sequences that phylogenetically precede or chronologically predate their appearance as lineage-dominant changes. Many of the LDMs in variants Alpha and Omicron were assigned probability values in the 99th percentile (44 and 33 percent of their LDMs, respectively; **Figure 5C**). However, performance of the model to predict the entire mutational profile (i.e., all LDMs) was lower for the Omicron variant. Indeed, 8 of the 36 LDM sites in Omicron had both *V* and *R* values of zero, whereas none of the 32 sites in variants Alpha, Gamma, Delta, Epsilon or Iota exhibited such a pattern (data not shown). The basis for appearance of mutations at such low-volatility sites raises questions regarding the origin of the Omicron variant: Is it derived from a host with unique selective pressures, or from a sublineage of the virus that has expanded in a poorly characterized population? Increased sequence surveillance as well as data accessibility of SARS-CoV-2 isolated from human and non-human

hosts may provide the information required to understand the rare pattern that appeared in variant Omicron.

Several of sites with high probabilities for mutations have been characterized for their effects on infectivity and antigenicity whereas the effects of others, and specifically in the context of existing mutations in each lineage, are still unknown. Advance notice of the imminent changes in each lineage allows testing of their impact on virus fitness and sensitivity to vaccine-elicited antibodies, for tailoring vaccines to the mutations expected to emerge within each lineage. Knowledge of the sites that are not expected to change is as important as the prediction of positions that are likely to mutate. For example, most mutations in the RBD that affect virus infectivity or sensitivity to antibodies, including E484K, L452R, S477N and N501Y are assigned high likelihoods to occur from the baseline group but low likelihoods to occur in $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ (**Table S3**). These findings clearly suggest that immunogens should be designed according to the mutational landscape that is specific to each lineage.

We note that, despite the high predictive capacity of the model described, these studies constitute a relatively simple framework to demonstrate predictability of the changes in SARS-CoV-2. Our forecasts can likely be improved by the use of more sophisticated learners to combine *V*, *R* and *D* values, alternative methods to define architecture of the co-volatility networks, and incorporation of additional statistics that describe the positive (and negative) selective pressures applied on each site. Furthermore, the use of more homogenous donor populations (e.g., vaccinated versus non-vaccinated individuals) will likely improve the ability of the models to predict emergence of lineage-dominating changes in SARS-CoV-2.

METHODS

Sequence alignment

Nucleotide sequences of SARS-CoV-2 isolated from humans were downloaded from the National Center for Biotechnology Information (NCBI) database and the Virus Pathogen Database and Analysis Resource (ViPR). For analysis of variant Omicron, sequences were downloaded from the GISAID repository (24). The following processing steps and analyses were performed within the Galaxy web platform (46). To facilitate alignment of sequences that contain more nucleotides than those corresponding to the spike gene, we trimmed excess bases with Cutadapt, using 5'-ATGTTTGTT-3' and 3'-TACACATAA-5 "adapters" that flank the spike gene. Adapter sequences were allowed to match once with a minimum overlap of 5

bases, an error rate of 0.2 with a sequence length between 3,700 and 3,900 bases. To ensure accuracy of the data, all sequences with any nucleotide ambiguities were removed by replacing the non-standard bases to 'N' with snippy-clean_full_aln, followed by filtration of N-containing sequences using Filter FASTA. Sequences that cause frameshift mutations were excluded using Transeq. Nucleotide sequences were aligned by MAFFT, using the FFT-NS-2 method (47). The aligned sequences were then "compressed" using Unique.seqs to obtain a single representative for each unique nucleotide sequence (48). Nucleotide sequences were then translated with Transeq and amino acid sequences were aligned with MAFFT, FFT-NS-2 (47). The first position of each PNGS motif triplet (Asn-X-Ser/Thr, where X is any amino acid except Pro) was assigned a distinct identifier from Asn. Our phylogenetic analyses were performed using the full-length spike protein, which contained several sequences with amino acid insertions. To maintain consistent numbering of spike positions, all calculations described in this work were performed for the 1,273 positions of the spike protein in the SARS-CoV-2 reference strain (accession number NC_045512).

Phylogenetic tree construction and analyses

A maximum-likelihood tree was constructed for the aligned compressed nucleotide sequences using the generalized time-reversible model with CAT approximation (GTR-CAT) nucleotide evolution model with FASTTREE (49). The tree was rooted to the sequence of the SARS-CoV-2 reference strain (NC_045512) with MegaX (50). To divide the tree into "Groups" of sequences, we used an in-house code in Python (see link to GitHub repository in the Data Availability section). This tool uses the Newick file to divide the dataset into sequence groups with a user-defined genetic distance between their centroids. All analyses described in this work were performed using a distance of 0.004 nucleotide substitutions per site for group partitioning. Groups that did not contain at least 50 unique sequences were excluded from our analyses. To discern between baseline groups and terminal groups, we used a distance of 0.0015 nucleotide substitutions per site between each group centroid and the SARS-CoV-2 reference strain. A total of 20 groups were obtained, composed of 12 baseline and 8 terminal groups.

Calculations of volatility

To calculate volatility of spike positions, we divided all sequences in each group into clusters of 50 sequences. Sequence variability in each cluster was quantified using two approaches. To calculate volatility (*V*) values, we used a binary approach, whereby every position in a 50-sequence cluster was assigned a value of 1 it if contains any diversity in amino

603 acid sequence, or a value of 0 if all sequences in the cluster contain the same amino acid. Thus,
604 each cluster is assigned a 1,273-feature vector that describes the absence or presence of
605 volatility at each position of spike. Volatility was then calculated by averaging values by position
606 across all clusters tested. For calculations of *D* or *R* values for each position *p*, we used a
607 quantitative approach to define volatility at positions associated with *p* (i.e., at positions *j* and *q*
608 in **Equation 1** and **Equation 2**, respectively). Briefly, sequence variability within each cluster
609 was measured by assigning amino acids hydropathy scores according to a modified Black and
610 Mould scale (17). Each amino acid is assigned a distinct value. The Asn residue in PNGS motifs
611 and deletions are also assigned unique values. The values assigned were: PNGS, 0; Arg,
612 0.167; Asp, 0.19; Glu, 0.203; His, 0.304; Asn, 0.363; Gln, 0.376; Lys, 0.403; Ser, 0.466; Thr,
613 0.542; Gly, 0.584; Ala, 0.68; Cys, 0.733; Pro, 0.759; Met, 0.782; Val, 0.854; Trp, 0.898; Tyr, 0.9;
614 Leu, 0.953; Ile, 0.958; Phe, 1; deletion site, 1.5. Variability in each cluster was calculated as the
615 standard deviation in hydropathy values among the 50 sequences, and variability values of all
616 clusters were averaged to obtain the volatility value for each position *j* or *q* (i.e., V_j or V_q,
617 respectively).
618

619 **Co-volatility calculations and network analyses**

620 To determine the propensity for co-volatility of any two spike positions, we generated a
621 matrix that contains binary volatility values in all clusters of the tested group (rows) for all 1,273
622 spike positions (columns). The co-occurrence of a volatile state between any two spike positions
623 was calculated using Fisher's exact test and the associated P-value determined using a custom
624 Java script. To construct the network of co-volatility, we used as input the matrix that describes
625 the $-\log_{10}$(P-value) between the volatility profiles of any two spike positions, whereby nodes are
626 the positions of spike and the edges that connect them reflect the P-values of their association.
627 Network structure was visualized using the open-source software Gephi (51). Networks were
628 generated using different P-value thresholds (i.e., an edge was assigned only if the P-value was
629 lower than 0.1, 0.05 or 0.01). To determine robustness of network structure, we randomly
630 deleted 10, 20 or 30 percent of all edges for each of the networks, and network topological
631 properties were computed using the Cytoscape Network Analyzer tool (52). Two metrics were
632 calculated for the complete and depleted networks: **(i)** Degree distribution, and **(ii)** Closeness
633 centrality (19).
634

635 **Calculation of total weighted volatility at network-associated sites (*R*)**

636 The variable *R* describes for each spike position the total weighted volatility at all
637 positions that are associated with it on the co-volatility network. To calculate *R* for each position
638 *p*, we first identified all positions *q* (q_1, q_2, … q_n) that are associated with *p* on the co-volatility
639 network, as defined by a P-value of less than 0.05 in the Fisher's exact test. We then calculated
640 for each position *p* the *R* value:

641 $$R_p = \sum_{q=1}^{n} w_{pq} \cdot V_q$$

642 where *n* is the number of *q* positions for each position *p*, w_{pq} is the association index between
643 volatility of position *p* and each position *q* (calculated as the $-\log_{10}$(P-value) in Fisher's test), and
644 V_q is the volatility at each position *q*.

645

646 **Calculations of the positive selection measure *S***

647 We estimated for each codon of spike the number of inferred synonymous (S) and
648 nonsynonymous (N) substitutions using the Mega7 platform (53). Estimates were generated
649 using the joint Maximum Likelihood reconstructions of ancestral states under a Muse-Gaut
650 model of codon substitution (54) and a Felsenstein 1981 model of nucleotide substitution (55).
651 The input phylogenetic tree was constructed using FASTTREE. The dN-dS metric was used to
652 detect codons that have undergone positive selection, where dS is the number of synonymous
653 substitutions per site and dN is the number of nonsynonymous substitutions per site. dN-dS
654 values were normalized using the expected number of substitutions per site. Maximum
655 Likelihood computations of dN and dS were conducted using the HyPhy software package (56).
656 Sites of deletion within groups $G_{T1}(\alpha)$ and $G_{T3}(\delta)$ were excluded from the analyses. For all
657 calculations, negative dN-dS values were assigned an *S* value of 0.

658

659 **Permutation test to determine spatial clustering of volatility**

660 We performed a permutation test to determine the spatial clustering of volatile sites
661 around each spike position. To this end, for each position *p*, we identified the 10 closest
662 positions on the trimer, using coordinates of the cryo-EM structure of the cleavage-positive
663 spike (PDB ID 6ZGI) (18). We then calculated for each position *p* the statistic T_p^0:

$$T_p^0 = \sum_{j \in \varphi_p} V_p^0 * V_j^0$$

where V_p^0 describes the volatility at position *p*, V_j^0 is the volatility at the j^{th} neighboring position to *p*, and φ_p denotes the positions numbers of the 10 closest neighbors to position *p*. We then permuted all positions identifiers other than *p* and calculated the statistic T_p^k:

$$T_p^k = \sum_{j \in \varphi_p} V_p^0 * V_j^k$$

where V_j^k is the volatility at the j^{th} adjacent position in the k^{th} permutation (*k=1,2, … 5,000*).

Under the null hypothesis of no spatial clustering, we would expect the neighbor labels to be arbitrary. We therefore test this null hypothesis by estimating the probability of observing a positive departure from the null distribution via:

$$P = \frac{\sum_{k=1}^{N} I_{\{T_p^k \geq T_p^0\}}}{N}$$

where N is the total number of permutations (5,000) and I is the indicator function. Therefore, the P-value quantifies the fraction of times the volatility of the surrounding residues is larger for the permuted values relative to the non-permuted values.

Calculations of total weighted volatility at adjacent positions on the spike trimer (*D*)

We calculated for each position *p* of spike the total volatility at all sites that are within a distance of 6 Å on the spike trimer structure. The coordinates of the cryo-electron microscopy structure of the cleaved spike protein in the closed conformation (PDB ID 6ZGI) were used (18). Coordinates of all atoms were included; N-acetyl-glucosamine atoms were assigned the same position number as their associated Asn residues. We then determined for each spike position the minimal distance between its atoms and the closest atoms of all other spike positions using coordinates of the three spike protomers. This information was used to calculate for each position *p* the weighted sum of volatility values at all spike positions that are within 6 Å distance on the spike trimer: $D_p = \sum_{j=1}^{n} \frac{1}{\Delta_{pj}} \cdot V_j$

where Δ_{pj} is the distance (in Å) between position *p* and each of the neighboring positions *j* on the trimer, V_j is the volatility value at each position *j*, and *n* is the number of *j* positions for

position *p*. We note that the 6ZGI structure is missing the following spike residues (numbered according to the SARS-CoV-2 reference strain): 1-13, 71-75, 618-632, 677-688, 941-943 and 1146-1273. To calculate *D* values for these positions, we applied the volatility values of the positions immediately adjacent on the linear sequence of spike (i.e., positions -1 and +1).

Combined model to predict emergence of dominant-group and subgroup-emerging mutations

To assign a probability for each position to emerge with a mutation, we used a logistic regression model that applies *V*, *R* and *D* values. The model was trained using *V*, *R* and *D* values calculated using the 5,700 sequences of the baseline group, with the positive outcome being the 43 GDM and 16 sGEM sites described in **Figure 1**. To this end, we first created interaction terms between the initial predictors (i.e., *V*, *R* and *D*). To address the class imbalance in our datasets (59 of the 1,273 spike positions contained a GDM or sGEM) we used the adaptive synthetic sampling approach (ADASYN) (57). Nested cross-validation was used to tune the model while estimating the metrics of interest. This procedure was also used to generate the prediction probabilities for each position. Five folds were used for both the inner and outer parts of the nested cross-validation. Grid search was utilized to optimize hyperparameters with the area under the receiver operating characteristic curve (ROC) as the objective for optimization. The model-specific parameters that we incorporated into the hyperparameter tuning procedure are the inverse of the regularization strength C and the penalty type. For this purpose, we used a set of values from 0.001 to 100 for parameter C, and for penalization we used L1 norm, L2 norm, elastic net, or no penalty in the parameter space. Since we used ADASYN to handle the class imbalance, we also added the number of positions with similar feature values as another hyperparameter to the search grid. The number of positions with similar feature values was set between 5 and 45. As classification metrics, we used sensitivity, specificity, precision, recall, AUC and balanced accuracy. The balanced accuracy metric, which is the average of sensitivity and specificity, was used due to the relative imbalance in the datasets.

DATA AVAILABILITY

The following data used in our analyses are available on the Mendeley Data repository at doi: 10.17632/wn7jwk9n22.1.

1. Sequence GenBank IDs of all 615,374 nucleotide spike sequences isolated from samples collected between December 2019 and July 2021.
2. Nucleotide alignment of the 16,808 unique spike sequences derived from the above.
3. Nucleotide alignment of 4,283 unique spike sequences of variant Delta isolated from samples collected between December 2019 and September 5th, 2021.
4. Sequence GISAID IDs of all 24,054 spike sequences isolated from samples collected in South Africa between March 6th 2020 and November 21st 2021.

CODE AVAILABILITY

The custom code used in our studies is publicly available within the following hub repository:

https://github.com/RoberthAnthonyRojasChavez/SARS2-Volatility

Instructions to the use of the code can be found in the following folders:

1. For calculation of *V*, *R* and *D* values, see the accordingly named folders.

2. For grouping sequences based on genetic distance cutoffs, see the 'Tree' folder.

3. For performing Fisher's exact test to determine the relationship between the volatility profile of any two spike positions, see the 'R' folder.

4. For calculating the minimal distance between any two residues on the spike protein based on coordinates of the trimer structure, see the 'D' folder.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest with the contents of this article.

ACKNOWLEDGEMENTS

We are grateful to Dr. Wendy Maury and Dr. Stanley Perlman for critical reading of this manuscript. We are also grateful to Dr. Benjamin Darbro for helpful discussions. This work was supported by intramural funds to HH, by grant 110028-67-RGRL to HH from the American Foundation for AIDS Research (amfAR), and by National Institutes of Health grant 1DP2AI164325 to JD.

REFERENCES

1. Cutler DM & Summers LH (2020) The COVID-19 Pandemic and the $16 Trillion Virus. *JAMA* 324(15):1495-1496.
2. Dong E, Du H, & Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 20(5):533-534.
3. Meredith LW, *et al.* (2020) Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 20(11):1263-1271.
4. Dai L & Gao GF (2021) Viral targets for vaccines against COVID-19. *Nat Rev Immunol* 21(2):73-82.
5. Harvey WT, *et al.* (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19(7):409-424.
6. Martin DP, *et al.* (2021) The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* 184(20):5189-5200 e5187.
7. Peacock TP, Penrice-Randal R, Hiscox JA, & Barclay WS (2021) SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *J Gen Virol* 102(4).
8. Callaway E (2021) Omicron likely to weaken COVID vaccine protection. *Nature* 600(7889):367-368.
9. Ball P (2021) The lightning-fast quest for COVID vaccines - and what it means for other diseases. *Nature* 589(7840):16-18.
10. Korber B, *et al.* (2020) Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182(4):812-827 e819.
11. Dearlove B, *et al.* (2020) A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc Natl Acad Sci U S A* 117(38):23652-23662.
12. MacLean OA, *et al.* (2021) Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol* 19(3):e3001115.
13. McCarthy KR, *et al.* (2021) Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* 371(6534):1139-1142.
14. Maher MC, *et al.* (2021) Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *medRxiv*.
15. Rodriguez-Rivas J, Croce G, Muscat M, & Weigt M (2022) Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc Natl Acad Sci U S A* 119(4).

16. Janse M, Brouwers T, Claassen E, Hermans P, & van de Burgwal L (2021) Barriers Influencing Vaccine Development Timelines, Identification, Causal Analysis, and Prioritization of Key Barriers by KOLs in General and Covid-19 Vaccine R&D. *Front Public Health* 9:612541.
17. DeLeon O, *et al.* (2017) Accurate predictions of population-level changes in sequence and structural properties of HIV-1 Env using a volatility-controlled diffusion model. *PLoS Biol* 15(4):e2001549.
18. Wrobel AG, *et al.* (2020) SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat Struct Mol Biol* 27(8):763-767.
19. Barabasi AL & Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509-512.
20. Albert R, Jeong H, & Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* 406(6794):378-382.
21. Karim SSA & Karim QA (2021) Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* 398(10317):2126-2128.
22. Wang L & Cheng G (2021) Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa. *J Med Virol*.
23. Wang Y, *et al.* (2022) The significant immune escape of pseudotyped SARS-CoV-2 variant Omicron. *Emerg Microbes Infect* 11(1):1-5.
24. Shu Y & McCauley J (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 22(13).
25. Yang TJ, *et al.* (2021) Effect of SARS-CoV-2 B.1.1.7 mutations on spike protein structure and function. *Nat Struct Mol Biol* 28(9):731-739.
26. Planas D, *et al.* (2021) Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* 596(7871):276-280.
27. McCallum M, *et al.* (2021) N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* 184(9):2332-2347 e2316.
28. Li Q, *et al.* (2020) The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* 182(5):1284-1294 e1289.
29. Liu Z, *et al.* (2021) Landscape analysis of escape variants identifies SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *bioRxiv*.
30. Patino-Galindo JA, *et al.* (2021) Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. *Genome Med* 13(1):124.

31. Starr TN, *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* 182(5):1295-1310 e1220.
32. Cerutti G, *et al.* (2021) Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell host & microbe* 29(5):819-833 e817.
33. Chi X, *et al.* (2020) A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* 369(6504):650-655.
34. Suryadevara N, *et al.* (2021) Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* 184(9):2316-2331 e2315.
35. Liu L, *et al.* (2020) Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* 584(7821):450-456.
36. Chen RE, *et al.* (2021) Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat Med* 27(4):717-726.
37. Ai J, *et al.* (2021) Omicron variant showed lower neutralizing sensitivity than other SARS-CoV-2 variants to immune sera elicited by vaccines after boost. *Emerg Microbes Infect*:1-24.
38. Dejnirattisai W, *et al.* (2021) Reduced neutralisation of SARS-CoV-2 omicron B.1.1.529 variant by post-immunisation serum. *Lancet*.
39. Cohen J (2021) Omicron sparks a vaccine strategy debate. *Science* 374(6575):1544-1545.
40. Fan S, *et al.* (2021) Preclinical immunological evaluation of an intradermal heterologous vaccine against SARS-CoV-2 variants. *Emerg Microbes Infect*:1-45.
41. Geller R, *et al.* (2015) The external domains of the HIV-1 envelope are a mutational cold spot. *Nat Commun* 6:8571.
42. Vandelli A, *et al.* (2020) Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Res* 48(20):11270-11283.
43. Buskirk SW, Peace RE, & Lang GI (2017) Hitchhiking and epistasis give rise to cohort dynamics in adapting populations. *Proc Natl Acad Sci U S A* 114(31):8330-8335.
44. Jungreis I, Sealfon R, & Kellis M (2021) SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nat Commun* 12(1):2642.
45. Meng B, *et al.* (2021) Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep* 35(13):109292.

46. Afgan E, *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46(W1):W537-W544.
47. Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.
48. Schloss PD, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537-7541.
49. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
50. Stecher G, Tamura K, & Kumar S (2020) Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* 37(4):1237-1239.
51. Jacomy M, Venturini T, Heymann S, & Bastian M (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* 9(6):e98679.
52. Lotia S, Montojo J, Dong Y, Bader GD, & Pico AR (2013) Cytoscape app store. *Bioinformatics* 29(10):1350-1351.
53. Kumar S, Stecher G, & Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* 33(7):1870-1874.
54. Muse SV & Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11(5):715-724.
55. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17(6):368-376.
56. Pond SL, Frost SD, & Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5):676-679.
57. He HB, Bai Y, Garcia EA, & Li ST (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *Ieee Ijcnn*:1322-1328.

Figure 1

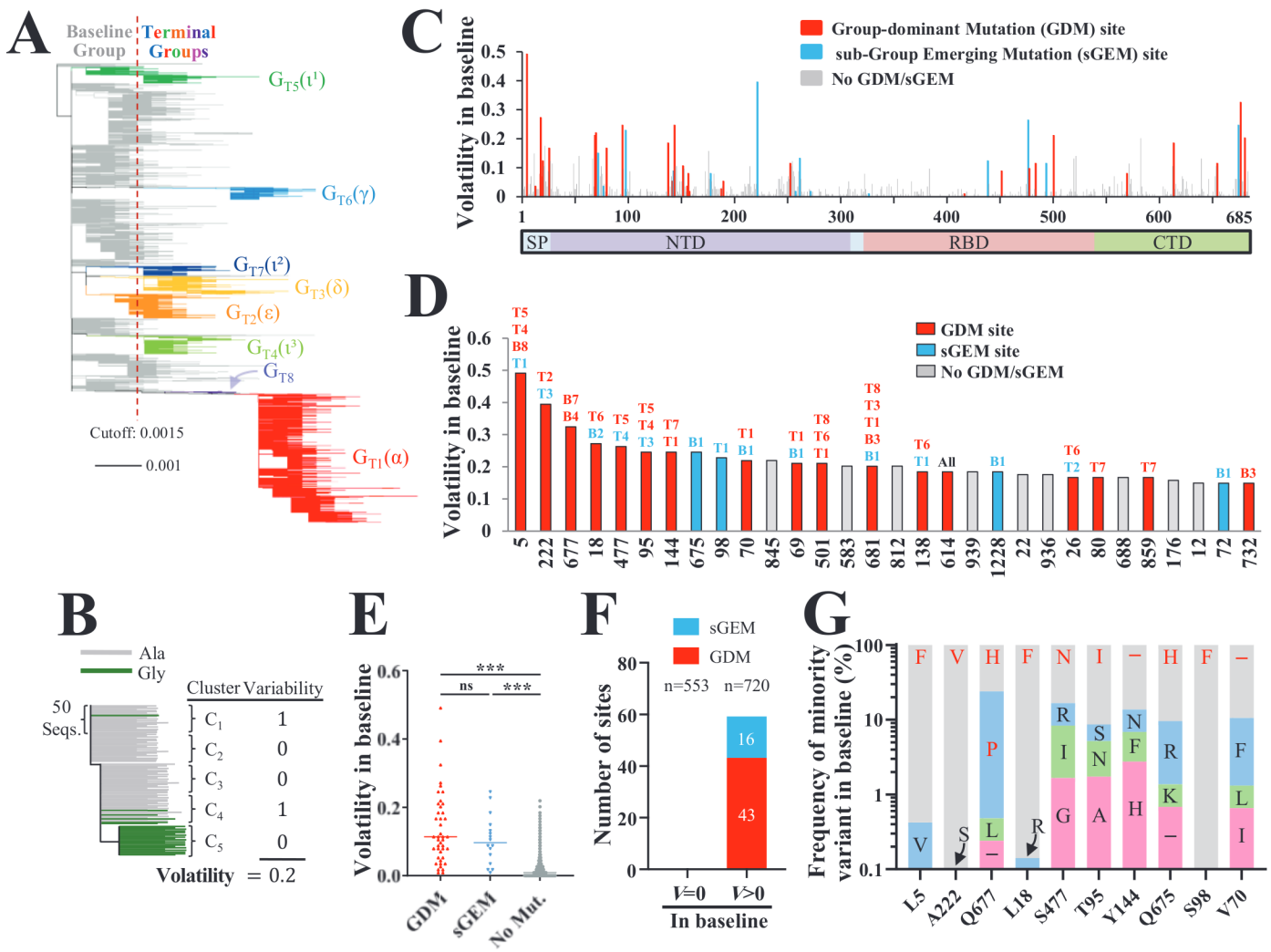

Figure 1. Spike positions with high volatility appear as sites of group-dominant or subgroup-emerging mutations. (A) Phylogenetic tree based on 16,808 unique spike sequences. Terminal groups are colored and labeled, with their WHO variant designations in parentheses. **(B)** Schematic of our approach to calculate volatility for each position of spike. **(C)** Volatility values for all positions of spike subunit S1, calculated using the 114 baseline clusters (see values for S2 subunit in **Figure S1C**). **(D)** Thirty spike positions with the highest volatility values. The baseline ("B") or terminal ("T") groups that contain mutations at these positions are indicated. **(E)** Comparison of volatility values for spike positions that emerged with a GDM, sGEM or no such mutations. P-values in an unpaired T test: ***, P<0.0005; ****, P<0.00005; ns, not significant. **(F)** Number of sites that appeared with GDMs and sGEMs when volatility (*V*) in the baseline group was zero or larger than zero. The number of site in each subset (n) is indicated. **(G)** Frequencies of minority variants (non-ancestral residues) at the ten positions of spike with the highest volatility values (see panel D). Frequencies are expressed as a percent of all sequences with a non-ancestral residue at the indicated position. The residues that emerged as GDMs or sGEMs are indicated in red font.

Figure 2

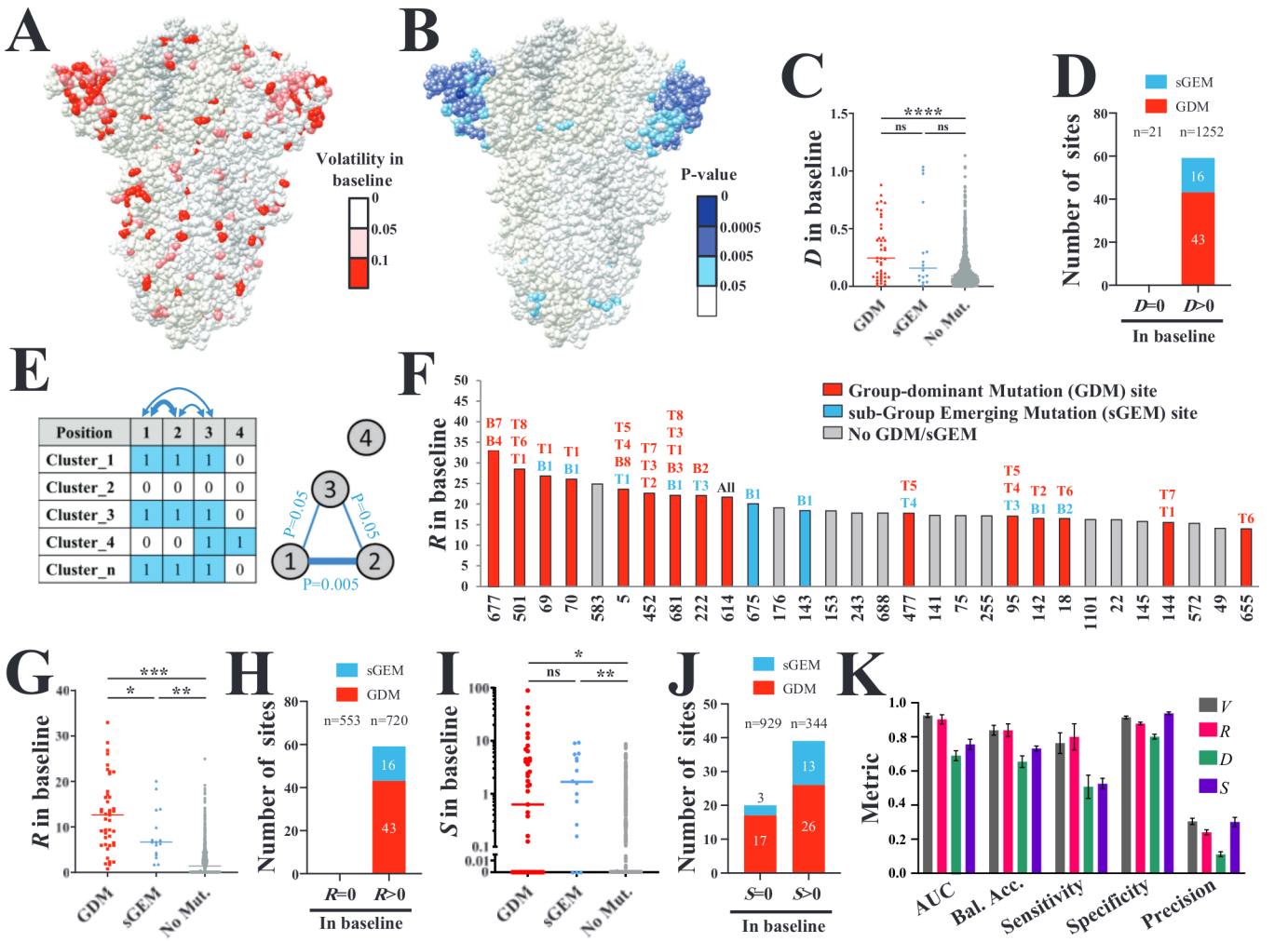

Figure 2. High volatility at spatially-adjacent and network-neighboring sites is associated with emergence of GDMs and sGEMs. (A) Cryo-EM structure of the spike trimer (PDB 6ZGI). Residues are colored by positional volatility values in the baseline group. **(B)** Results of a permutation test to identify sites that are more likely to be volatile when their 10 closest positions are volatile. **(C)** A measure of the total volatility at adjacent positions on the spike trimer. The variable *D* describes for each position *p* the sum of the volatilities at all positions within a distance of 6Å, weighted by their proximity to *p* (see **Equation 1**). *D* values are compared between positions with GDMs, sGEMs or no such mutations. **(D)** The number of sites that emerged with GDMs or sGEMs when the *D* value was zero or larger than zero. **(E)** Schematic of our approach to calculate co-volatility of spike positions. The absence (0) or presence (1) of amino acid variability was determined in each cluster of 50 sequences for all positions of spike. The co-occurrence of a volatile state at all position pairs was determined using Fisher's test, and the P-values were used to construct the network of co-volatility between all positions. **(F)** Thirty spike positions with the highest *R* values (see all in **Figure S3A**). Sites of GDMs or sGEMS are indicated by bar color and the groups of emergence are indicated above the bars. **(G)** *R* values for spike positions that emerged with a GDM, sGEM or with no such mutations. **(H)** Number of GDMs and sGEMs that emerged at spike positions when *R* in the baseline group was equal to or greater than zero. **(I)** Comparison of the positive selection metric *S* between positions that emerged with a GDM, sGEM or with no such mutations, as calculated using the baseline group. **(J)** Number of sites that emerged with GDMs or sGEMs when *S* in the baseline group was zero or larger than zero. **(K)** Classification metrics for evaluating performance of the indicated variables to predict presence of a mutation (either GDM or sGEM) in any group (baseline or terminal). Probabilities were calculated using a logistic regression model that applies the baseline group of sequences. Error bars, standard errors of the means for five-fold cross validation.