bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Early Computational Detection of Potential High Risk SARS-CoV-2 Variants

Authors:

Karim Beguir[1]*, Marcin J. Skwark[1], Yunguan Fu[1], Thomas Pierrot[1], Nicolas Lopez Carranza[1], Alexandre Laterre[1], Ibtissem Kadri[1], Bonny Gaby Lui[2], Bianca Sänger[2], Yunpeng Liu[3], Asaf Poran[3], Alexander Muik[2], Ugur Sahin[2]*

Affiliations:

[1]InstaDeep Ltd; 5 Merchant Square, London W2 1AY, UK.

[2]BioNTech SE; An der Goldgrube 12, 55131 Mainz, Germany.

[3]BioNTech US; 40 Erie Street, Cambridge, MA, 02139, USA.

*Corresponding authors. Emails: kb@instadeep.com; ugur.sahin@biontech.de

Abstract: The ongoing COVID-19 pandemic is leading to the discovery of hundreds of novel SARS-CoV-2 variants on a daily basis. While most variants do not impact the course of the pandemic, some variants pose significantly increased risk when the acquired mutations allow better evasion of antibody neutralisation in previously infected or vaccinated subjects, or increased transmissibility. Early detection of such high risk variants (HRVs) is paramount for the proper management of the pandemic. However, experimental assays to determine immune evasion and transmissibility characteristics of new variants are resource-intensive and time-consuming, potentially leading to delayed appropriate responses by decision makers. Here we present a novel *in silico* approach combining Spike protein structure modelling and large protein transformer language models on Spike protein sequences, to accurately rank SARS-CoV-2 variants for immune escape and fitness potential. We validate our immune escape and fitness metrics with *in vitro* pVNT and binding assays. These metrics can be combined into an automated Early Warning System (EWS) capable of evaluating new variants in minutes and risk monitoring variant lineages in near real-time. The EWS flagged 12 out of 13 variants, designated by the World Health Organisation (WHO, Alpha-Omicron) as potentially dangerous, on average two months ahead of them being designated as such, demonstrating its ability to help increase preparedness against future variants. Omicron was flagged by the EWS on the day its sequence was made available, with immune evasion and binding metrics subsequently confirmed through our *in vitro* experiments.

One-Sentence Summary: A COVID-19 Early Warning System combining structural modelling with AI to detect and monitor high risk SARS-CoV-2 variants, identifying >90% of WHO designated variants on average two months in advance.

Main Text:

Main: Despite a relatively slow mutation rate in the human coronaviruses, since the emergence of the human coronavirus SARS-CoV-2 in Wuhan in December 2019, over 250,000 different missense variants (as of November 25, 2021) have been identified in the protein-coding viral sequences deposited in the GISAID (Global Initiative on Sharing All Influenza Data) database and associated with multiple lineages. Of these, over 12,750

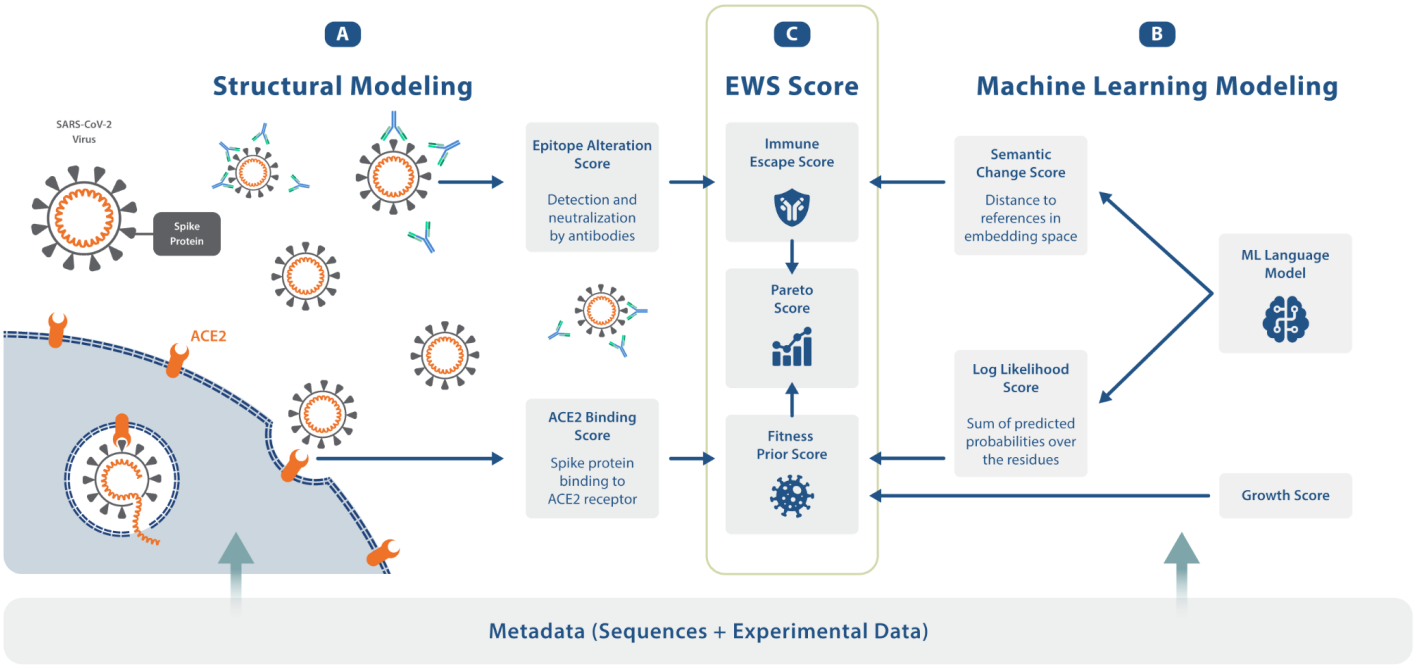
bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

individual missense mutations (including indels) have been observed in the Spike protein, the key target for antibody neutralisation (While most mutations either reduce the overall fitness of the virus, or bear no consequences to its features, some individual or combinations of mutations lead to high risk variants (HRVs), with modified immune evasion capabilities, and/or improved transmissibility. For example, the Alpha (B.1.1.7) variant of concern (VOC) spread widely through higher transmissibility compared with the Wuhan strain, while the Beta (B.1.351) VOC has been shown to be less effectively neutralised by both convalescent sera and antibodies elicited by approved COVID-19 vaccines[1]. The Delta (B.1.617.2) variant characterised by a high transmissibility led to increased mortality and triggered a renewed growth in cases in countries with both high and low vaccination rates (such as the United Kingdom[2] and India[3]). Most recently, the more heavily mutated Omicron (B.1.1.529) variant was amongst the quickest variants to be designated as a VOC by the WHO, due to a combination of widespread dissemination and several concerning mutations in the Spike protein as well as in other proteins[4].

Hundreds of new variants are sequenced daily, some of which are added to the GISAID and other databases[5,6]. As new sequences continue to naturally emerge, the potential for the generation of variants that are both fit and highly immune resistant creates a significant challenge for public health authorities. The transmissibility and immune escape potential of a given variant could be assessed experimentally: evaluating one aspect of the fitness of variants requires experimental measurements of their binding affinity with its human receptor, angiotensin-converting enzyme 2 (ACE2), which is necessary for host cell infection; assessing immune escape potential requires *in vitro* neutralisation tests involving serum from vaccinated subjects or serum from patients previously infected with other variants of SARS-CoV-2. Both methods are resource-intensive and time-consuming, and cannot be scaled to properly address the multitude of emergent variants.

In this work, we present a new method to evaluate SARS-CoV-2 variants based on *in silico* structural modelling and artificial intelligence (AI) language modelling and demonstrate that it captures features of a given variant's fitness as well as its immune escape properties (Fig. 1). This approach is used here to build an Early Warning System (EWS) that trains on the complete (up to a chosen time point) GISAID SARS-CoV-2 sequence database in less than a day and can score novel variants within minutes. It is a non-trivial task, as newly emerging High Risk Variants most often comprise new sets of mutations, and not all combinations of mutations present in previously identified concerning variants actually lead to enhanced immune evasion or transmissibility. In particular, accomplishing this by standard ML methods results in poorer predictive performance (see Supplementary Material, “Detecting HRVs by standard ML methods”). The EWS is fully scalable as new variant data become available, allowing for the continuous risk monitoring of variant lineages. We show that it can flag HRVs months earlier than their designation as such by the WHO, providing an opportunity to shorten the response time of health authorities.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



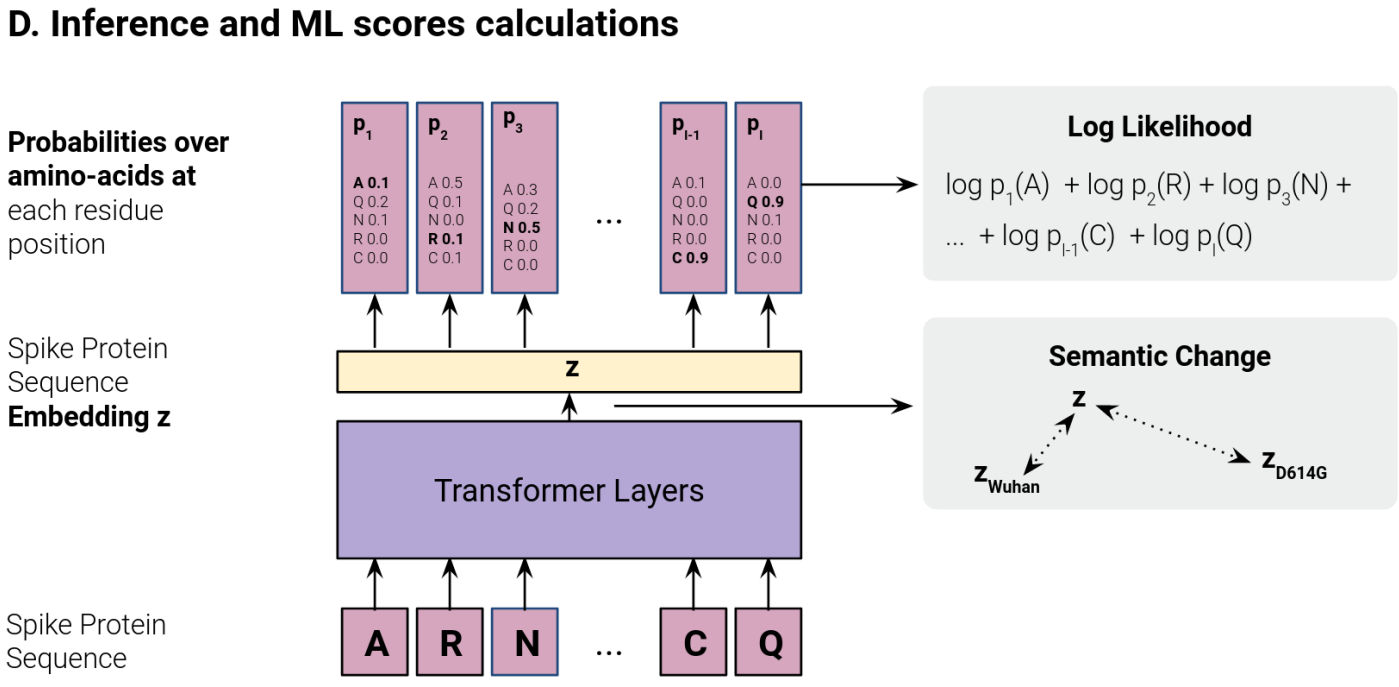


Fig. 1. A schematic of the Early Warning System (EWS): structural modelling methods and natural language processing techniques to enable risk level estimation of SARS-CoV-2 variants in real-time. (A) Structural modelling is used to predict the binding affinity of SARS-CoV-2 Spike protein to host ACE2, and to score the mutated epitope regarding its impact on immune escape. **(B)** Machine Learning modelling is used to

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

extract implicit information from unlabeled data for the hundreds of thousands of registered variants in the GISAID database. (**C**) EWS relies on the information from A and B to compute an *immune escape score* and a *fitness prior score*, which taken together, present a more comprehensive view of the SARS-CoV-2 variant landscape. Both scores can be combined to obtain a single score, based on the notion of Pareto optimality and dubbed Pareto score, that represents a variant's risk. The higher the Pareto score, the fewer variants with higher immune escape and fitness prior scores. **(D)** Schematic of AI model structure for assessing semantic change and log-likelihood. Once trained (Fig. S.1.A), the model receives as input a variant Spike protein sequence, and returns an embedding vector of the Spike protein sequence as well as probabilities over amino-acids for each residue (Fig. S.1.B). The embedding vector is used to calculate semantic change from the Wuhan and D614G variants while the probabilities are used to compute the log-likelihood.

Results :

In silico prediction of immune escape potential.

Mutations in the Spike (S) protein, especially the receptor-binding domain (RBD), are key to the heightened resistance to antibody-mediated neutralisation of new SARS-CoV-2 variants[7]. To evaluate the impact of said mutations on humoral immune evasion, the 336 binding epitopes observed in 310 previously resolved structures of neutralising antibodies (nAbs)[8–11] were mapped onto the S protein based on publicly available resolved 3D structures (Table S.1). An overlay of all nAb:S protein interaction interfaces was used to generate a colour-coded heatmap, indicating which surface-exposed amino acids are located in high epitope density regions (Fig. 2.A). The number of known nAbs whose binding epitope is affected by distinct SARS-CoV-2 variants' mutations was defined as the epitope alteration score (Table S.2).

While this score can be used as a first proxy to evaluate escape from humoral immunity, it is limited by its dependence on the quantity of available antibody structure data. By using deep learning language models, we were able to leverage the information from hundreds of thousands of SARS-CoV-2 S protein sequences deposited to the GISAID database. It was recently demonstrated that these algorithms have the ability to capture the biological properties of proteins through unsupervised learning on large amounts of biological data[12–14]. At inference time, a language model returns the predicted probability distribution of the 20 natural amino acids for each position in the protein, thus leveraging the underlying biology of the large number of sequences seen during training from an evolutionary point of view. Hie et al.[15] showed that language models trained on a dataset of proteins can be used to assess the risk of a viral variant. This risk was measured through two proxies named grammaticality as a measure for fitness and semantic change to assess antigenic variation. In this work, the recurrent neural networks were replaced with attention-based models, namely transformers[16], hence replacing the auto-regressive way of training the model used in[15] with the Bidirectional Encoder Representations from Transformers (BERT) protocol, which has been shown to give better results in many applications[17]. Even though the GISAID dataset contains hundreds of thousands of Spike protein sequences, it is limited in scope to SARS-CoV-2. To learn more general features of protein sequences and address currently unseen viral variants, one would need to use more comprehensive protein sequence resources, such as the UniProtKB database that includes hundreds of millions of protein sequences[18]. To benefit from this large volume of available data, the model was first pre-trained over the large collection of varied proteins included in UniRef100 (non-redundant sequence clusters of UniProtKB and selected UniParc records) and then fine-tuned over S protein sequences (Fig. S.1.A). The transformer model has been re-trained every month on the variants registered in GISAID (up to 250,000 unique S sequences vs. 4,172 S sequences in Hie et al.[15]). Our semantic change calculation differed in that it was computed to estimate the change with respect to the wild type and from the D614G mutation to take into account this mutant that largely replaced the Wuhan strain (Fig. 2.A).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to validate the immune escape *in silico* metrics - semantic change and epitope alteration score - *in vitro* pseudovirus neutralisation test (pVNT) assays were conducted. The cross-neutralizing effect of n≥12 BNT162b2-immune sera was assessed against vesicular stomatitis virus (VSV)-SARS-CoV-2-S pseudoviruses bearing the Spike protein of 19 selected High Risk Variants, including Omicron (B.1.1.529) (Fig. 2.B, Fig. S.2, Table S.3)[19,20]. The SARS-CoV-2 Omicron pseudovirus was by far the most immune escaping with >20-fold reduction of the 50% pseudovirus neutralisation titer ($pVNT_{50}$) compared with the geometric mean titer (GMT) against the Wuhan reference spike–pseudotyped VSV (Fig. S.2.B). The calculated geometric mean ratio with 95% confidence interval (CI) of the Omicron pseudotype and the Wuhan pseudotype GMTs was 0.025 (95% CI; 0.017 to 0.037), indicating another 10-fold drop of the neutralising activity against Omicron compared to the second most immune escaping B.1.1.7+E484K pseudovirus with a geometric mean ratio of 0.253 (95% CI; 0.196 to 0.328) (Fig. S.2.C). This result is in absolute concordance with the *in silico* immune escape score for Omicron which is the highest amongst observed, circulating variants. Across all HRV pseudoviruses tested, both the epitope alteration score and the semantic change score correlate positively with the calculated $pVNT_{50}$ reduction (Fig. 2.B; Pearson r=0.79 and 0.74, respectively). Of note, an average of both *in silico* scores (summarised as the 'immune escape score') exhibits a stronger correlation with the observed reduction in neutralising titers (Pearson r=0.85).

In silico estimation of fitness.

The immune escape score predicts if a given viral variant may evade neutralisation by the immune system, but it does not capture protein changes that either enhance the efficacy of viral cell entry, or negatively impact its structure or function. Capturing the full transmissibility potential of the virus (*fitness*) is beyond the scope of this work as it comprises many complex dynamics, however, we can propose three informative priors contributing toward it: *ACE2 binding score*, *conditional log-likelihood score* and *growth*.

A key determinant of viral spread is the effectiveness with which virus particles can attach to and invade target host cells. This characteristic is especially important when considering individuals without pre-existing immunity or viral variants which are able to better evade immune responses. In order to infect the human host cell, the RBD of the viral S protein associates with ACE2, the cellular receptor for SARS-CoV-2. Therefore, we assessed the fitness prior based on the predicted impact of sets of mutations on the binding affinity of the variant S protein to the human ACE2 receptor, here referred to as the *ACE2 binding score*. The interaction between a variant S protein and the ACE2 protein was computed through repeated, fully flexible, *in silico* docking experiments, allowing for an unbiased sampling of the binding landscape. In order to reduce the required computational resources, the Spike protein modelling was restricted to its RBD domain, i.e. the domain known to be directly binding to the ACE2 receptor. To calculate the ACE2 binding score, we used the median binding energy (that is the difference estimated Gibbs free energy δG between bound and unbound states), which acts as a proxy for global complex affinity (Fig. S.3).

In order to assess the validity of the ACE2 binding score, the simulation results were compared with *in vitro* results. Surface plasmon resonance (SPR) kinetic analysis was performed to determine the affinity (K_D, dissociation constant) of 19 RBD variants to the ACE-2 receptor. Notably, the assay measures observable association rates, which are a result of a dynamic process, while simulations measure aggregated, static binding affinity, thus marginalising the contribution of mutations toward the flexibility and kinetics of the spike protein. Despite this, the ACE2 binding score used herein shows meaningful correlation with the K_D values with a Pearson correlation coefficient of 0.45 (Fig. 2.C).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Another aspect that partially models the fitness of a variant, is how similar a given variant is to the other variants which have been known to grow rapidly. This is not achievable by simple sequence comparison, due to epistatic interactions between sites of polymorphism, in which certain mutation combinations enhance fitness while being deleterious when they occur separately. The same trained transformer model described previously was leveraged to calculate the log-likelihood of the input sequence. From a language model perspective, the higher the log-likelihood of a variant, the more probable is the variant to occur. In particular, the log-likelihood metric supports substitutions, insertions, and deletions without requiring a reference sequence to measure against, unlike the grammaticality of Hie et al.[15] that requires a reference sequence. Our language model was not provided with explicit sequence count data in the training phase, yet on average assigned higher log-likelihood values to sequences with higher actual observed count. High log-likelihood may indicate features common in the general variant population, which are likely to be fitness-related, thus allowing strains harbouring these to sustain additional such mutations.

However, the values of log-likelihood tend to diminish with the increasing number of mutations, which is expected given the definition of this metric; this over-emphasizes variants with low mutation counts. Considering that all the samples used for training have been detected in patients, and as such have likely satisfied a minimal fitness criteria, we introduced the *conditional log-likelihood score*, measuring how the log-likelihood of the variant in question compares to other variants with similar mutational loads, as opposed to the entire population. This metric sheds more light on highly mutated, potentially concerning variants like B.1.1.529 (Omicron). Due to its high mutational load, this variant might be perceived by raw log-likelihood as highly unlikely. However, relative to other variant sequences with a similar number of mutations, it becomes clear that it stands out, leading to a high conditional log-likelihood score (Fig. S.4).

None of the metrics discussed above capture the entirety of factors affecting the frequency of viral variants. Additionally, conditional log-likelihood is a metric measuring similarity to already known, rapidly increasing samples. By its nature, it cannot fully assess the variants which exhibit completely new sequence features, until these features are observed more often. Therefore, the fitness prior metric includes *growth,* an empirical term of the quantified change in the fraction of the sequences in the database that a variant in question comprises. This addresses the intuitive notion that variants which are increasing in prevalence are more imminently interesting than those which are not. The growth metric also reflects the overall fitness of a given variant, implicitly taking into account the impact of mutations beyond the RBD domain and Spike protein. It complements the ACE2 binding score which models the RBD domain only.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

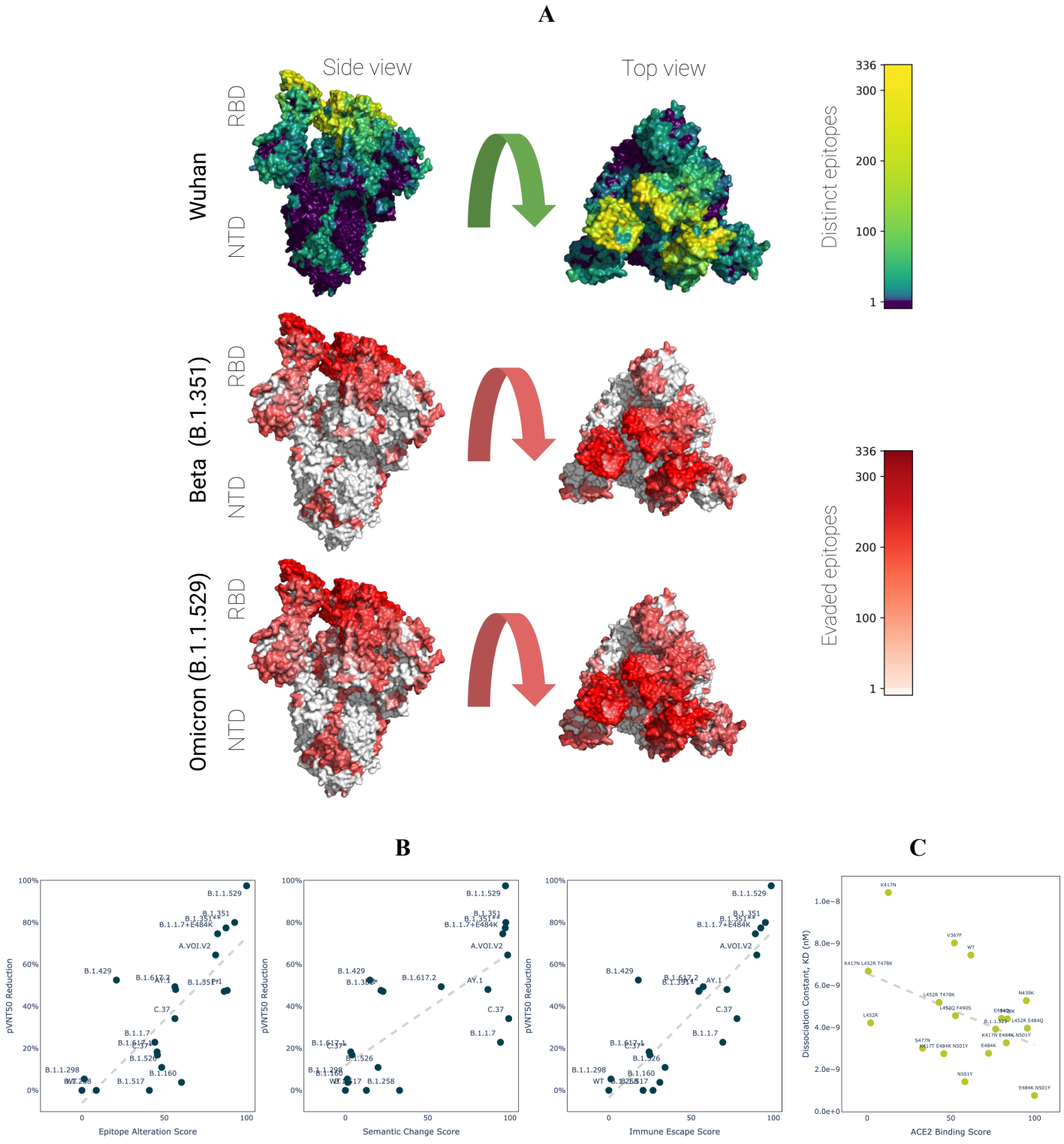


Fig. 2. *In silico* predicted scores for immune escape and fitness prior correlate with *in vitro* data. (A) The surface of a SARS-CoV-2 Spike protein in 'one RBD up' conformation (PDB id: 7kdl), in top colored by the frequency of contact of surface residues with neutralising antibodies (brighter. warmer colour corresponds to

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more antibody binding). Middle and bottom row depict the number of evaded epitopes in a Beta (B.1.351) and Omicron (B.1.1.529). Left column: side view. Right column: top view. **(B)** Validation of the immune escape metric with pseudovirus neutralisation test (pVNT) results. Relationships of the epitope alteration score, semantic change score, and combined immune escape score with the observed 50% pseudovirus neutralisation titer ($pVNT_{50}$) reduction are shown across n=19 selected SARS-CoV-2 variants of interest, including Omicron (B.1.1.529). Cross-neutralization of n=12-40 BNT162b2-immune sera was assessed against vesicular stomatitis virus (VSV)-SARS-CoV-2-S pseudoviruses. $pVNT_{50}$ reduction compared to wild-type SARS-CoV-2 (Wuhan strain) Spike pseudotyped VSV is given in percent. Variants for which experimentally measured geometric mean $pVNT_{50}$ increased compared to the Wuhan strain have been assigned a $pVNT_{50}$ reduction of 0 (equal to wild type). Epitope alteration score (based on structural modelling) indicates the number of known neutralising antibodies (max. n=310) whose binding epitope is affected by the SARS-CoV-2 variants' mutations. Semantic change score (based on machine learning) indicates the predicted variation in the biological function between a variant and wild-type SARS-CoV-2. For the semantic change score, the distance in embedding space between the sequence in question and a reference (WT+D614G Spike protein) is compared. Sequences have then been ranked with respect to this distance and the resultant rank has been scaled in the range of [0, 100]. The immune escape score is calculated as the average of the scaled epitope alteration score and the scaled semantic change score. Dashed lines represent the linear regression. **(C)** Validation of a component of the fitness prior metric, capturing the ACE2 binding propensity. Relationships of the ACE2 binding score with the experimentally determined ACE-2 binding affinity (K_D, dissociation constant) are shown across n=19 RBD variants, along with a linear regression dash line. The ACE2 binding score is ranked and scaled analogously to fitness prior components, such that variants with the lowest energy are assigned a score of 100, highest - 0.

Combining fitness prior and immune escape scores to continuously monitor high risk variants.

Different selective pressures on the virus evolution lead to variants with high immune escape and fitness, since a virus must remain evolutionarily competent to successfully spread. We hypothesise that a system that keeps track of immune escape and fitness factors (as depicted in Fig. 3.A-B), could potentially continuously monitor HRVs on a near real-time basis, since new sequences get evaluated and added to the data pool in minutes. The ranking of any sequence, and consequently - lineage, depends on other circulating sequences.

As seen in Fig. 3C, variants of concern start off relatively far into the upper-right corner (i.e. are comparatively highly immune escaping and have satisfactory fitness prior score for their immune escape value). Newly emerging variants diversify over time, as there are more circulating sequences observed. Their aggregated immune escape score decreases, while fitness prior score (based partially on prior observations) - increases for truly fit variants (Alpha: B.1.1.7 and Q lineages, and Delta: B.1.617.2 and AY lineages). Lineages such as Beta progressively decrease in aggregated fitness, closely recapitulating their lack of global growth, despite continued prevalence. The effect of perceptible global growth of B.1.351 (Beta) in April 2021, as well as its drop in prevalence and acquisition of further diversifying mutations, are all visible in the plot. The case of B.1.351 (Beta) illustrates a variant that is - according to the data and statistical models - unlikely to regain global significance. Simultaneously, we notice the effects of fitness-enhancing events (either mutations or emergence of evolutionary niche), such as the increase in metrics for Alpha lineages in Summer 2021, which was possibly due to the competitive effect of B.1.617.2 emergence. This evolutionary pressure could have been one of the factors behind the near eradication of B.1.1.7. The remaining sequences are under evolutionary pressure to adapt to the changing circumstances, and while currently not significant globally, still pose a tangible threat.

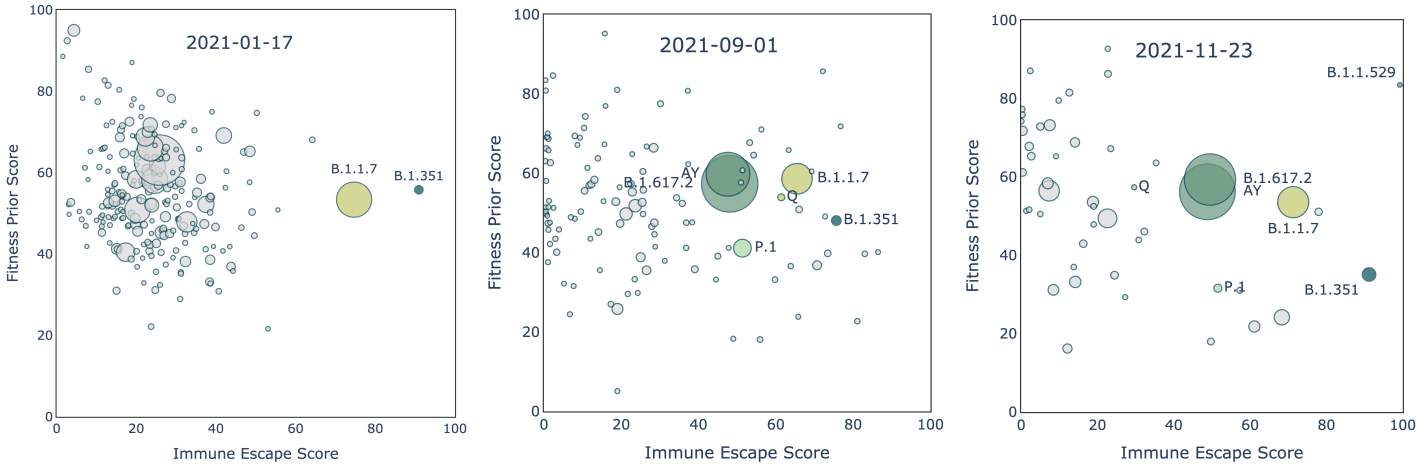
bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To jointly score the relative risks of variants using immune escape potential and fitness prior, an optimality score, termed Pareto score, was used to assess variants. The Pareto score is a mathematically robust way to identify lineages that are both immune escaping and infectious, and captures the relative evolutionary advantage of a given strain (see Methods section for calculation details). For each lineage, as defined by the Pango nomenclature system [21], scores were calculated by averaging the scores of the individual sequences belonging to a given lineage. A high Pareto score at a given time for a specific lineage indicates that only a few other lineages have higher scores for fitness and immune escape at that time.

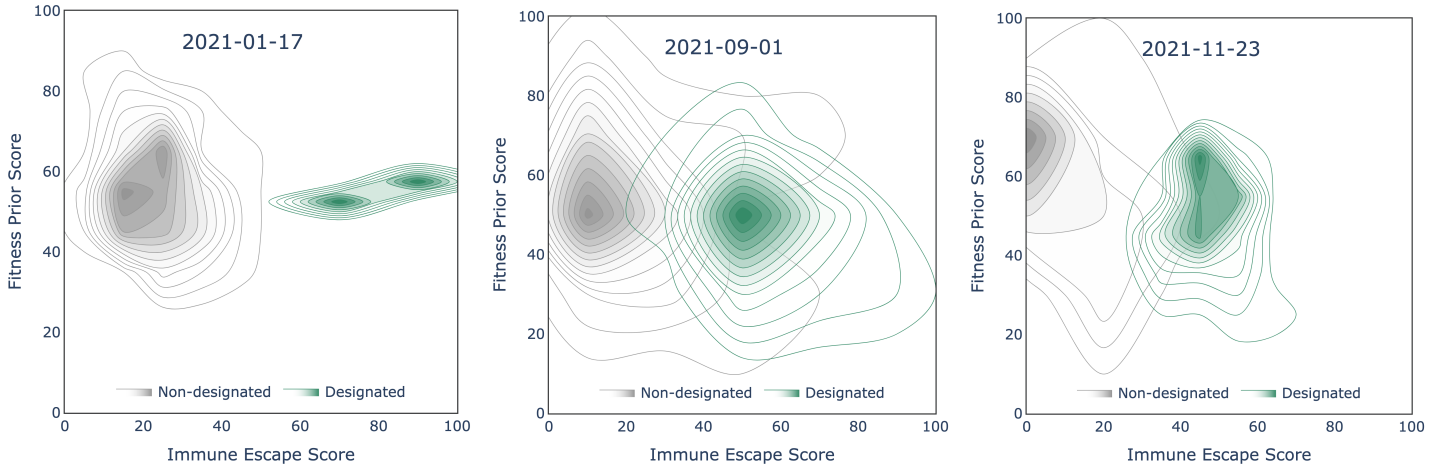
In order to validate that the Pareto score separates WHO designated variants from non-designated variants, Welch's t-tests were conducted over the registered variants population every week from January 2021 to November 2021 (Table S.5). The null hypothesis can be rejected with a p-value < 0.05, for all 50 weeks, thus demonstrating that the Pareto score can separate designated from non-designated variants continuously through time. For visualisation purposes, a focus was made on three dates of interest: the 17th of January 2021, the 1st of September 2021 and the 23rd of November 2021. At these dates, p-values = 2E-143, 6E-4 and 4E-4 were reported. Density contour estimates conducted on January 17th, 2021, September 1st, 2021 and November 23rd, 2021 also demonstrate clear separability between WHO designated variants and non-designated ones (per lineage: Fig. 3.B and per sequence: Fig. S.5). Importantly, they suggest that immune escape significantly contributes to this separability, more so than the fitness prior score.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

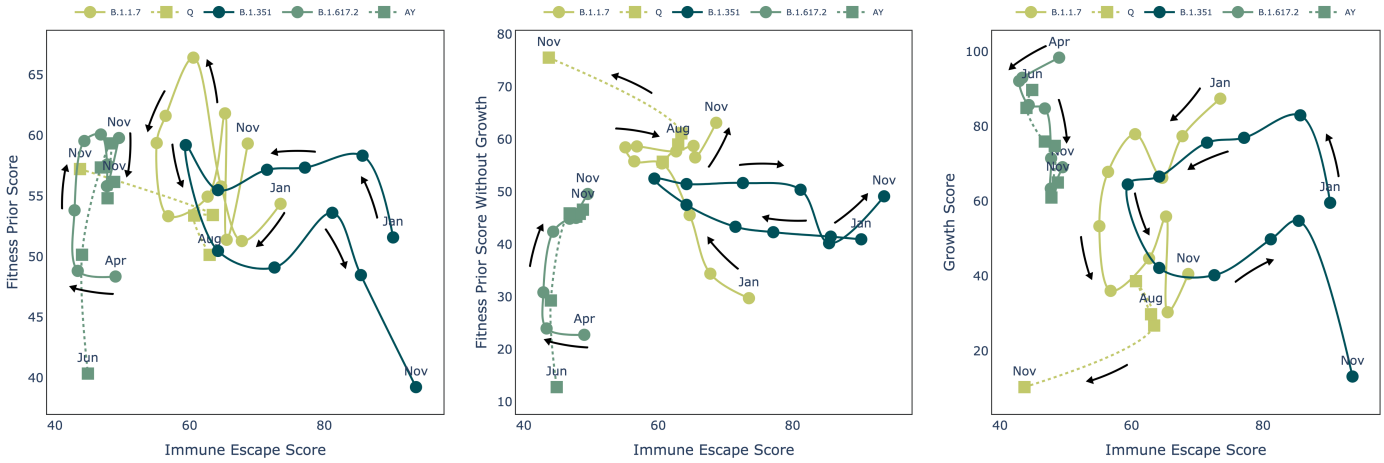
A



B



C



bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 3. Combining immune escape and fitness prior for continuous monitoring. (A) Snapshot of lineages in terms of fitness prior and immune escape score on respectively from left to right January 17th, 2021, September 1st 2021 and November 23rd 2021. Marker size indicates the number of submissions of each lineage. **(B)** Given a large number of lineages, we used densities instead of points clouds for visualisation. Densities of non-designated and designated variants on January 17th, 2021, September 1st 2021 and November 23rd 2021 are represented. The density contour plot is computed by grouping points specified by their coordinates into bins and calculating contours using counts. **(C)** Progression of the fitness prior and immune escape scores of main lineages designated by WHO through time from the early snapshot (January 2021) to the later snapshot (September 2021). Each dot represents the position of the centre of mass of a given lineage on each month. The left and centre plot demonstrates the progression using fitness prior score with and without growth respectively. The right plot shows the progression using only growth.

Detection of potentially high risk variants prior to substantial spread in the population.

Experimental assays aiming to determine a given variant's immune evasion and fitness are time and resource-intensive. Available data show that approximately thousands of new variants are emerging every week at an increasing rate (on average ~250 per week in September 2020, 7,000 in August 2021 and 10,000 in October 2021). Moreover, this number is likely an underestimate given limited viral sequencing and data deposition in many countries. It is therefore not feasible for health authorities to perform preventative experimental assessments whenever a new variant is identified, despite the benefits of a proactive stance detecting HRVs before their spread.

As seen in the previous section, our EWS immune escape score does help separate WHO designated variants from non-designated variants and has demonstrated a significant correlation to *in vitro* neutralisation test results.

In addition, our immune escape score is computed from sequence alone and unlike our fitness prior score does not require growth metrics, which are not available when a novel variant is sequenced. This means that an early detection version of our system, operating based on immune escape score alone, could potentially spot HRVs.

Moreover, it was recently proposed that viral evolution in immunocompromised patients generates intrapatient viral variants with increased immune evasion, rather than increased fitness and constitutes a significant factor contributing to variant spread[22,23]. Some of the new variants reside on long branches of phylogenetic tree, which suggests they could have undergone an extensive intrapatient evolution enabled by the immunocompromised status of the host. These results, together with increased vaccination rates worldwide, put an added emphasis on immune evasion as a key risk factor in newly emerging variants, which further motivated our approach to use only immune escape for early HRV detection.

A systematic analysis was conducted where for every week between September 16th, 2020 and November 23rd 2021 the EWS ranked all new sequences on immune escape to compile a weekly flagged HRV watch-list. The models were only trained on variants up to the previous month's start date and any other data used were prior to the analysis date. To assess the system's sensitivity, we focused on the detection of the 13 variants designated by WHO (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Eta, Theta, Iota, Kappa, Lambda, Mu, and Omicron).

When using a weekly watch-list with a size of 20 variants (less than 0.5% of the weekly average of new variant sequences), EWS flagged 12 WHO designated variants out of 13 (Fig. 4.A), with an average of 58 days of lead time (i.e two months) before these were designated as such by the WHO (Table S.4). For variants Alpha to Mu for which we have sufficient data, on average only 0.5% of cases of that variant were already recorded at the

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

time of their detection by the EWS (25 sequences on average), to be contrasted with the WHO announcements that happened on average when 18% of cases for that variant were already recorded (1,593 sequences on average). Different watch-list sizes ranging from 1 up to 200 variants were assessed to evaluate detection sensitivity (Fig. 4.B). The number of named variants detected remained stable when varying the weekly watch list size between 10 and 200. While a longer list compromises specificity, it leads to an increase in the detection lead time (the number of days ahead of WHO designation) (Fig. S.6).

Our system however does not accurately pinpoint the emergence of the B.1.617.2 Delta family of variants. Delta is known to be neutralised by vaccines[24] and its global prevalence can be attributed to other fitness-enhancing factors. These factors, such as P681R mutation, which abrogates O-glycosylation, thus further enabling furin cleavage, are outside of the scope of our approach. Delta variants also first emerged in India, a vast country with a diverse population and relatively limited sequencing capabilities, hence available samples may have been insufficient to fully describe the epidemiological landscape in time. Government regulations prohibiting the export of biological data out of the country may have also further restricted sequence data from reaching global databases like GISAID.

Strikingly, WHO-designated variants Alpha, Beta, Gamma, Theta, and Omicron are detected by the EWS on the same week they are first reported, even in the extreme case where the weekly watch-list allows only one variant, meaning they were the top-scoring sequences among all emerging variants that week (Fig. 4.B). Using a larger weekly watch-list of 20 variants, Epsilon, Zeta, Eta and Lambda are also detected in the same week they are first reported, on top of the previously mentioned WHO-designated variants (9 in total detected the first week).

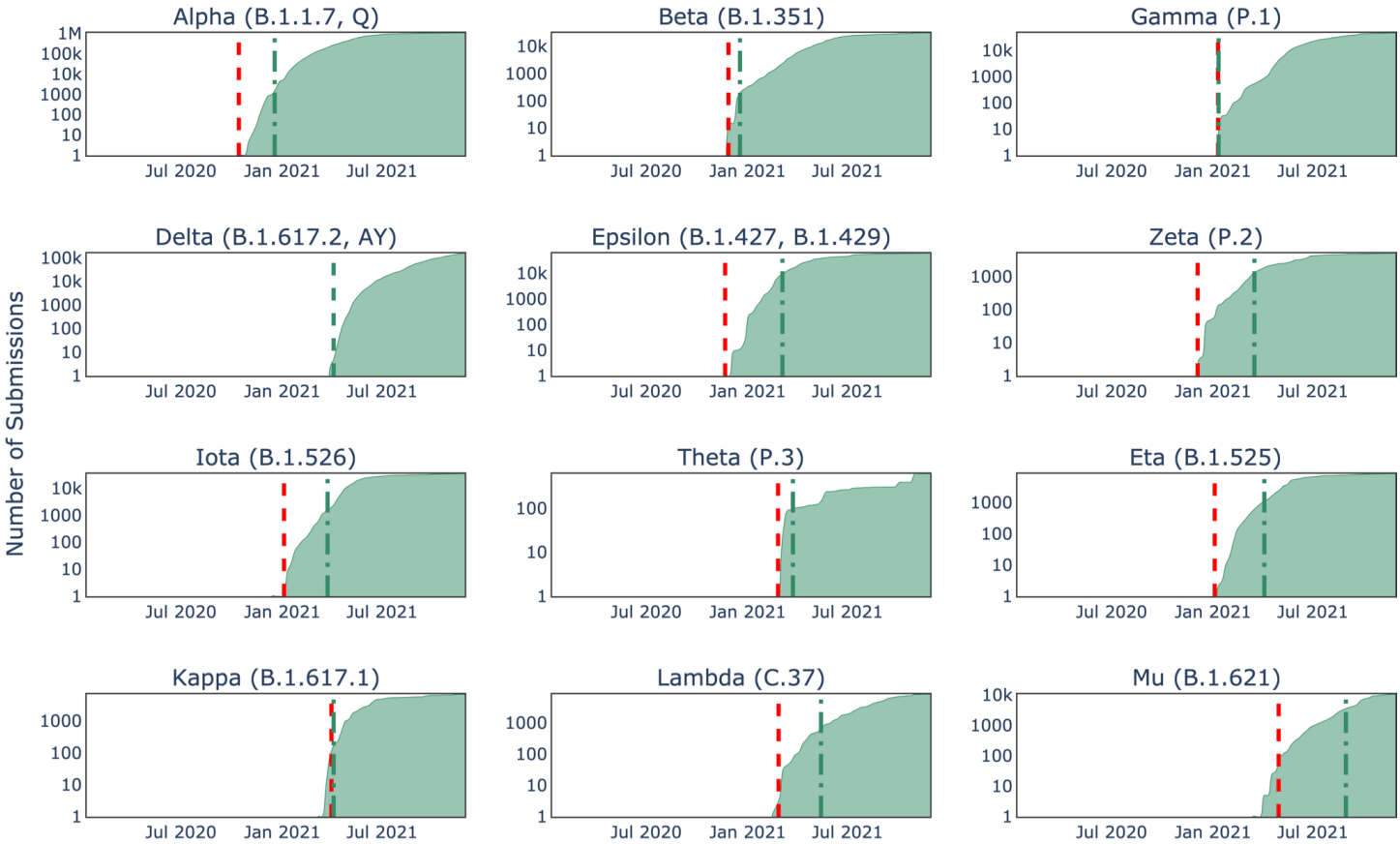
Specifically, the EWS identified Omicron as the highest immune escaping variant over more than 70,000 variants discovered between early October and late November 2021. This variant combines frequent RBD mutations (K417N, S477N, N501Y), with less frequent ones (G339D, S371L, S373P, S375F, Q498R) to potentially evade RBD-targeting antibodies. The NTD indels in positions 69-70, 143-145, 211-214 alter known antibody recognition sites as well. These mutations, together with over 20 others, led to an exceptional epitope alteration score, the highest recorded since the beginning of the pandemic and a high semantic change score, which combined rank Omicron in the top 0.005% of variants on immune escape since the pandemic started.

One can consider the growth score alone as a plausible metric that requires neither AI nor simulation to early detect HRVs with yet better than random results. However, the immune escape score implemented in the EWS outperforms the growth score across the variants where a comparison is available, in terms of lead time ahead of WHO designation. The growth score also fails to detect Delta ahead of time despite the established fitness of this variant, which may be another consequence of incomplete or delayed sequencing data (Fig. 4.C).

The immune escape score used by the EWS to early detect HRVs combines the epitope alteration score and semantic change score (Fig.1. A, B, C). We evaluated the early detection performance of each one of these two components separately and combined: while the Epitope Alteration Score detects 11 out of 13 WHO designated variants ahead of time, the Semantic Change score detects only 8 out of 13. Their combination, however, flags 12 out of 13 WHO designated variants (Fig. 4.D). We have also applied standard machine learning techniques, both supervised and unsupervised, denoted respectively “GLM with mutations” and “UMAP with mutations”, corresponding conceptually to Epitope Alteration Score and Semantic Change score (see Supplementary Material section “Detecting HRVs by standard ML methods”) . These standard machine learning techniques do not reach the same predictive performance as the methods proposed in this work. This validates our approach associating protein structure modelling and transformer language models on protein sequence to accurately rank SARS-CoV-2 variants.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A



bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

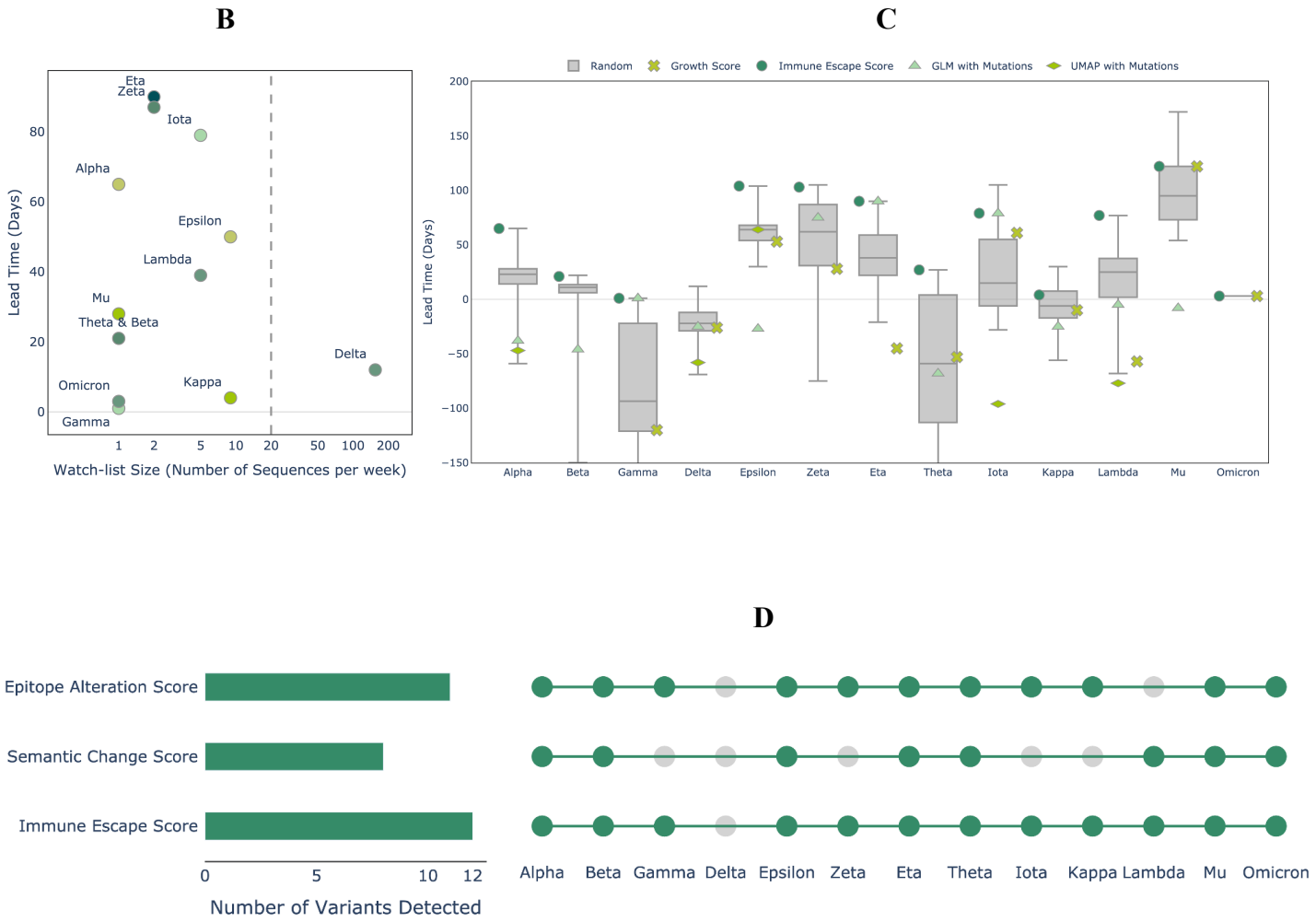


Fig. 4. EWS flags High Risk Variants ahead of their WHO designation. (A) Cumulative sum of all cases of a given variant lineage (in log scale) over time. Vertical lines indicate the date of WHO designation of a given variant (green dot-dashed) vs. date of flagging by the EWS (red dashed, using a weekly watch-list size of 20 variants). (**B**). Lead time of EWS detection ahead of WHO designation vs. minimum weekly watch-list size required (in log scale). (**C**). Detection results (measured in days of lead time vs. WHO designation) from selecting 20 variants per week at random (repeated 100 times) compared with selecting top 20 variants by growth score (light-green cross) and immune escape score (green circle). Boxplots borders indicate 25th and 75th percentiles, horizontal lines indicate median, and whiskers indicate minimal and maximal values. If a variant cannot be detected with growth or immune escape score, the marker is not displayed. (**D**) Variants detected when using Epitope Alteration Score, Semantic Score and Immune Escape Score components of the EWS. The left bar chart displays the number of variants detected by EWS using different scores. The right part visualises whether a WHO designated variant is detected in advance using different scores, where green dots indicate early detections and grey dots mean the variants are not detected in advance.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Through validation of our immune escape and fitness prior scores using published and newly generated data, we show that the combination of structural simulations, AI, and genomic sequencing of SARS-CoV-2 variants allows for continuous risk monitoring and sensitive early detection of HRVs.

Importantly, EWS flags Omicron on the day it is uploaded to GISAID (November 23rd, 2021) based on its sequence alone, as one of the highest immune escaping variants ever documented for SARS-CoV-2 sequence variants. Moreover, EWS assigns Omicron a high fitness prior score based on the combination of predicted ACE2 binding and a substantial conditional log-likelihood score. However, we would like to point out that a limitation for the calculation of the latter score for Omicron is the relatively small number of variants that have such a high number of mutations.

The Early Warning System can sensitively detect HRVs months ahead of the official WHO designation, sometimes within the same week a sequenced variant enters the database. The only variant with delayed flagging by the EWS compared with WHO designation, the Delta variant, suffers from significant underrepresentation of the lineage in GISAID. This partial and delayed representation of the lineage in the database prevents even growth-based detection, emphasising the importance of extensive, robust, and timely sequencing of SARS-CoV-2 genomic samples globally.

Combining comprehensive sequencing with structural modelling and AI can provide unprecedented insights into the COVID-19 pandemic which could be harnessed by public health authorities and governments worldwide to increase their preparedness to HRVs and potentially alleviate the associated human and economic costs. Future development of our approach could be extended to further functionalities such as assessment of known and predicted T cell epitopes, as well as the projection of prospective variant evolution scenarios.

References

1. Liu, Y. *et al.* Neutralizing Activity of BNT162b2-Elicited Serum. *N. Engl. J. Med.* **384**, 1466–1468 (2021).

2. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect. Dis.* (2021) doi:10.1016/S1473-3099(21)00475-8.

3. Singh, J., Rahman, S. A., Ehtesham, N. Z., Hira, S. & Hasnain, S. E. SARS-CoV-2 variants of concern are emerging in India. *Nat. Med.* **27**, 1131–1133 (2021).

4. The Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE). Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-con

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cern (2021).

5. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data--from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
6. Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482–D490 (2017).
7. Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, (2020).
8. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
9. Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115–119 (2020).
10. Dejnirattisai, W. *et al.* The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell* **184**, 2183-2200.e22 (2021).
11. Yan, R. *et al.* Structural basis for bivalent binding and inhibition of SARS-CoV-2 infection by human potent neutralizing antibodies. *Cell Res.* **31**, 517–525 (2021).
12. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
13. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* **118**, (2021).
14. Elnaggar, A. *et al.* ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* (2020).
15. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
16. Vaswani, A. *et al.* Attention is all you need. in *Advances in neural information processing systems*

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5998–6008 (2017).

17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
18. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
19. Muik, A. *et al.* Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine-elicited human sera. *Science* **371**, 1152–1153 (2021).
20. Sahin, U. *et al.* BNT162b2 vaccine induces neutralizing antibodies and poly-specific T cells in humans. *Nature* **595**, 572–577 (2021).
21. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
22. Corey, L. *et al.* SARS-CoV-2 Variants in Patients with Immunosuppression. *N. Engl. J. Med.* **385**, 562–566 (2021).
23. Weigang, S. *et al.* Within-host evolution of SARS-CoV-2 in an immunosuppressed COVID-19 patient as a source of immune escape variants. *Nat. Commun.* **12**, 6405 (2021).
24. Liu, J. *et al.* BNT162b2-elicited neutralization of B.1.617 and other SARS-CoV-2 variants. *Nature* **596**, 273–275 (2021).

Acknowledgments

Supported by BioNTech and InstaDeep. We thank the BioNTech German clinical trial (NCT04380701, EudraCT: 2020-001038-36) participants, from whom the post-immunization human sera for the cross-neutralization analysis were obtained.

Author contributions

U.S., K.B., M.J.S., A.M., Y.F, T.P. and A.P. conceived and conceptualized the work. K.B. conceived the AI

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

scoring models. K.B., Y.F., T.P., and A.L. conceived the AI training procedure. K.B., M.J.S., Y.F., N.L.C., and I.K. conceived and developed the data pipeline, software, and visuals. Y.F., T.P., and I.K. performed the AI experiments. M.J.S. conceived and developed the *in silico* epitope alteration score and structural bioinformatics methodology. A.M. and B.G.L. planned and supervised the *in vitro* experiments. A.M., B.G.L. and B.S. performed *in vitro* experiments. A.M., B.G.L. and B.S. analyzed *in vitro* experimental data. U.S., K.B., M.J.S., Y.F., T.P., N.L.C., A.M., A.P., and Y.L. interpreted data and wrote the manuscript. All authors supported the review of the manuscript.

Competing interests

U.S. is a management board member and employee at BioNTech SE. A.M., B.G.L. and B.S. are employees at BioNTech SE. A.P. and Y. L. are employees at BioNTech US.

U.S. and A.M. are inventors on patents and patent applications related to RNA technology and the COVID-19 vaccine. U.S., A.M., B.G.L., and B.S. have securities from BioNTech SE.

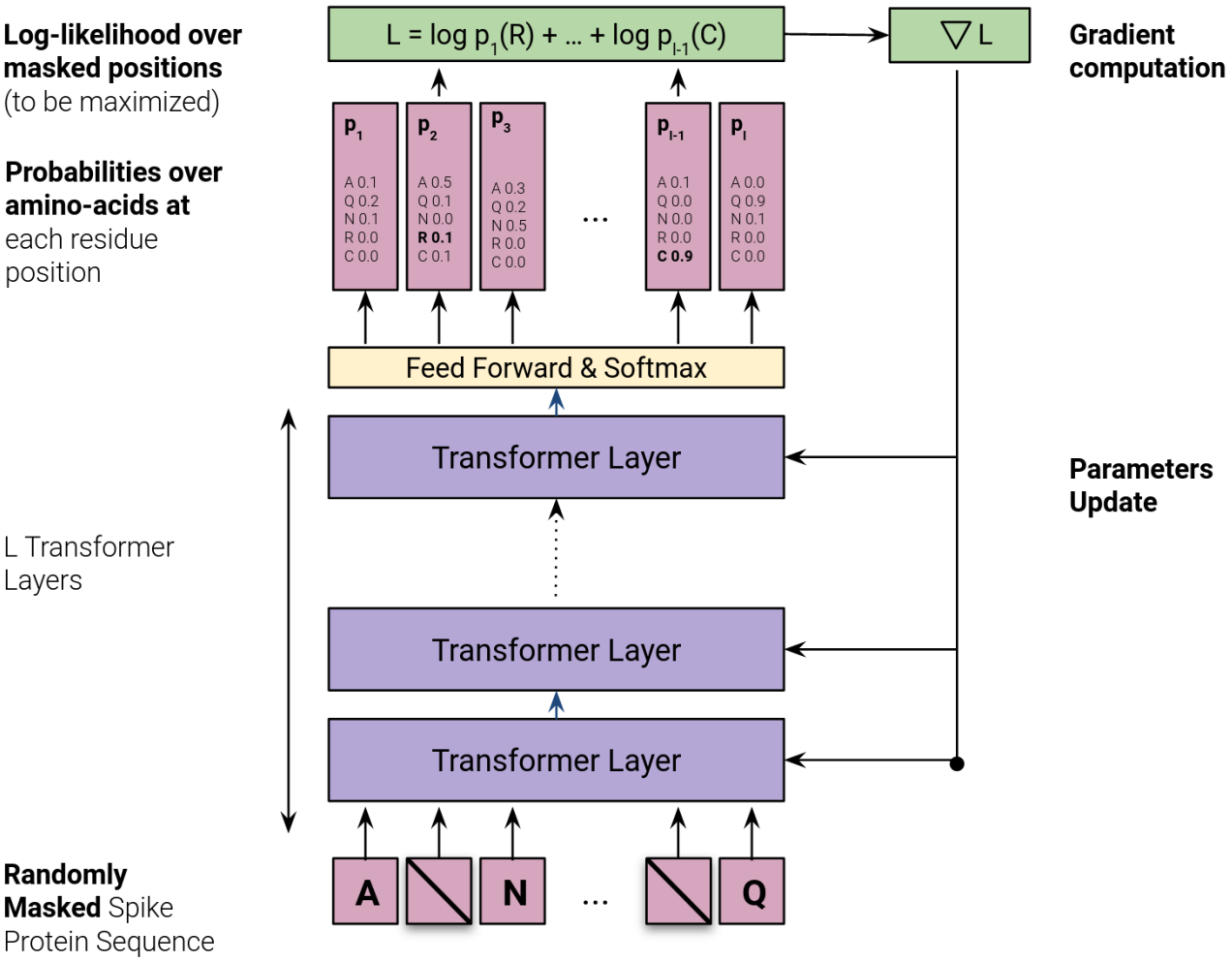
K.B. is a management board member and employee at InstaDeep Ltd. M.J.S., Y.F., T.P., N.L.C., A.L. and I.K. are employees of InstaDeep Ltd or its subsidiaries.

K.B., M.J.S., Y.F., T.P., N.L.C., and A.L. are inventors on patents and patent applications related to AI technology. K.B., M.J.S., Y.F., T.P., N.L.C., and A.L. have securities from InstaDeep Ltd.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Materials

A. Training



bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B. Inference and ML scores calculations

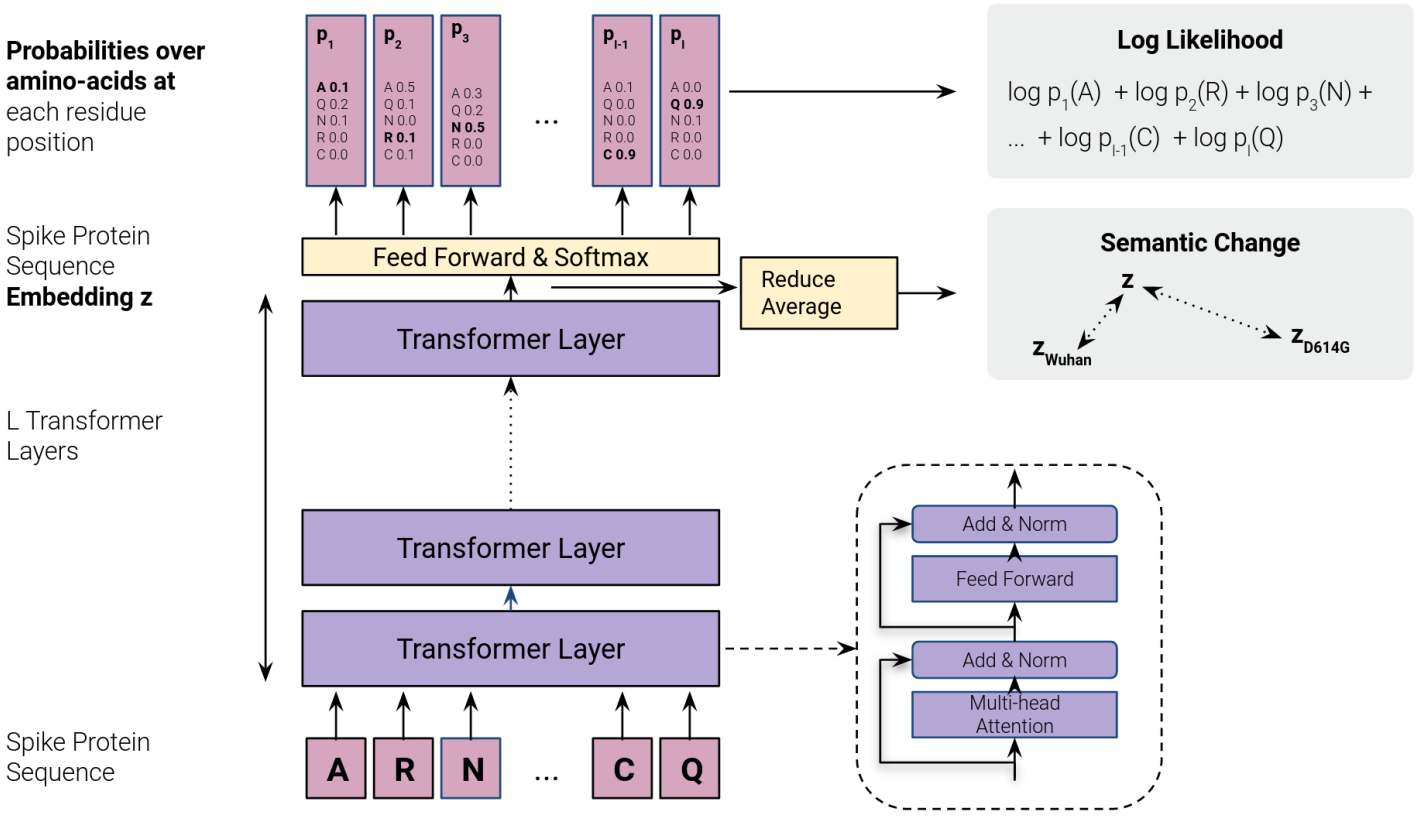


Figure S.1. Machine learning modelling. (A) A transformer language model is pre-trained on all the protein sequences registered in the UniRef100 dataset. Every week, the model is fine-tuned over all the Spike protein sequences registered at least twice, so far by the GISAID initiative. Both the pre-training and fine-tuning use the same protocol. Amino acids of a protein sequence are randomly masked. The model predicts probabilities over amino-acids at each residue position, both for residues that were masked and not masked. A loss function evaluates the sum over the masked residues of the log-probability of the correct predictions. A gradient of this loss is computed and used to update the model's parameters so as to increase the loss function. **(B)** Once fine-tuned, the model is used to compute the semantic change and the log-likelihood to characterise a Spike protein sequence. The output of the last transformer layer is averaged over the residues to obtain an embedding z of the protein sequence. The embedding of the Wuhan strain z_{wuhan} and the embedding of the D614G variant z_{D614G} are computed once for all. The semantic change is computed as the sum of the L1 distance between the z and and z_{wuhan} the L1 distance between z and z_{D614G}. The log-likelihood is computed from the probabilities over the residues returned by the model. It is calculated as the sum of the log-probabilities over all the positions of the Spike protein amino acids.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

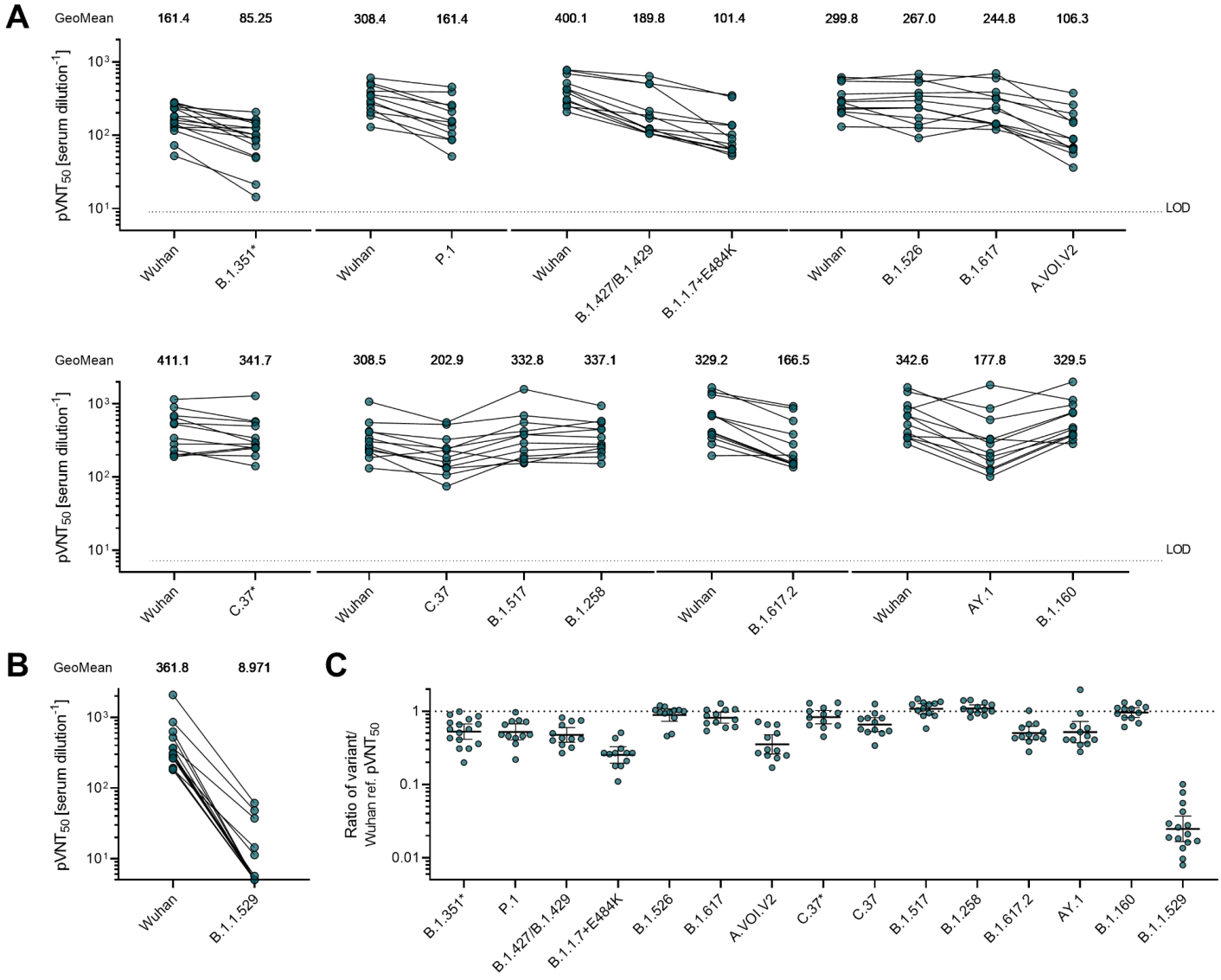


Figure S.2. Cross-neutralization of BNT162b2-immune sera against VSV-SARS-CoV-2-S pseudoviruses bearing the Spike protein of selected SARS-CoV-2 variants. Serum samples were obtained from participants in the BNT162b2 vaccine phase-I/II trial on day 28 or day 43 (7 or 21 days after Dose 2). A recombinant vesicular stomatitis virus (VSV)-based SARS-CoV-2 pseudovirus neutralisation assay was used to measure neutralisation. The pseudoviruses tested incorporated the ancestral SARS-CoV-2 Wuhan Hu-1 Spike or Spikes with substitutions present in B.1.1.7+E484K (Alpha), B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta), AY.1 (Delta), B.1.427/B.1.429 (Epsilon), B.1.526 (Iota), B.1.617 (Kappa), C.37 (Lambda), C.37* (Lambda), A.VOI.V2, B.1.517, B.1.258, B.1.160, and B.1.1.529 (Omicron) (Table S.3). **(A)** Pseudovirus 50% neutralisation titers ($pVNT_{50}$) are shown. Dots represent results from individual serum samples. Lines connect paired neutralisation analyses performed within one experiment. In total 8 experiments were performed covering the listed SARS-CoV-2 variants always referencing variant-specific neutralisation to the Wuhan reference. **(B)** $pVNT_{50}$ against B.1.1.529 (Omicron) are shown. Dots represent results from individual serum samples. Lines connect paired neutralisation analyses performed within one experiment. **(C)** Ratio of $pVNT_{50}$ between SARS-CoV-2 variant and Wuhan reference strain Spike-pseudotyped VSV. Dots represent results from

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

individual serum samples. Horizontal bars represent geometric mean ratios, error bars represent 95% confidence intervals.

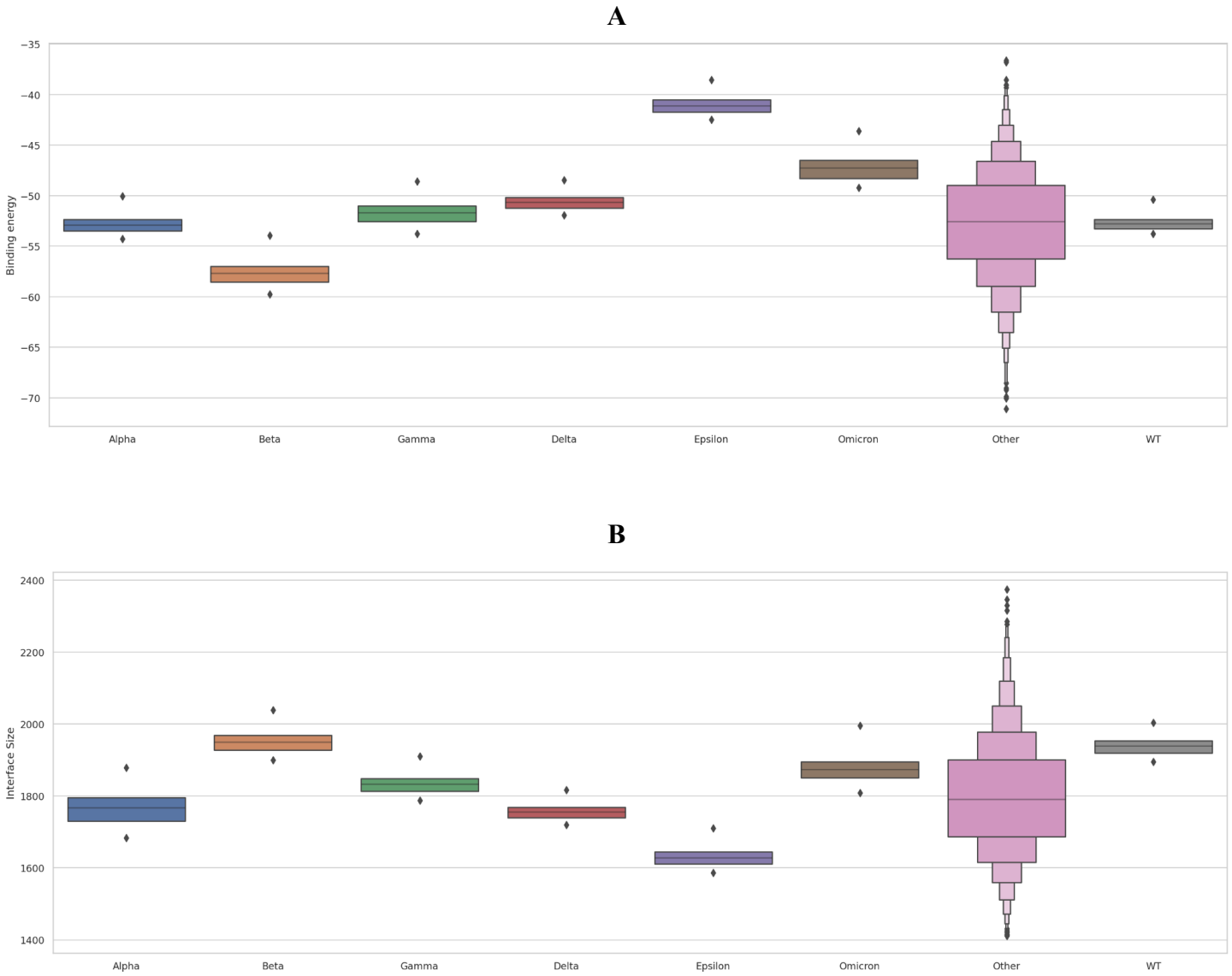


Figure S.3. Results of molecular simulations of RBD binding. The efficiency of Spike protein RBD binding to the ACE2 receptor is dictated by the combination of binding energy (A; the lower the better) and size of the interface (B). Both boxen plots depict distribution of these values across performed RBD binding simulations for circulating spike protein variants. Note, that while larger interfaces may be difficult to form, they are also more difficult to break. Strikingly, Omicron, despite its heavily mutated RBD has a relatively large interface and a binding affinity around the 25th percentile of the background distribution ('Other').

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

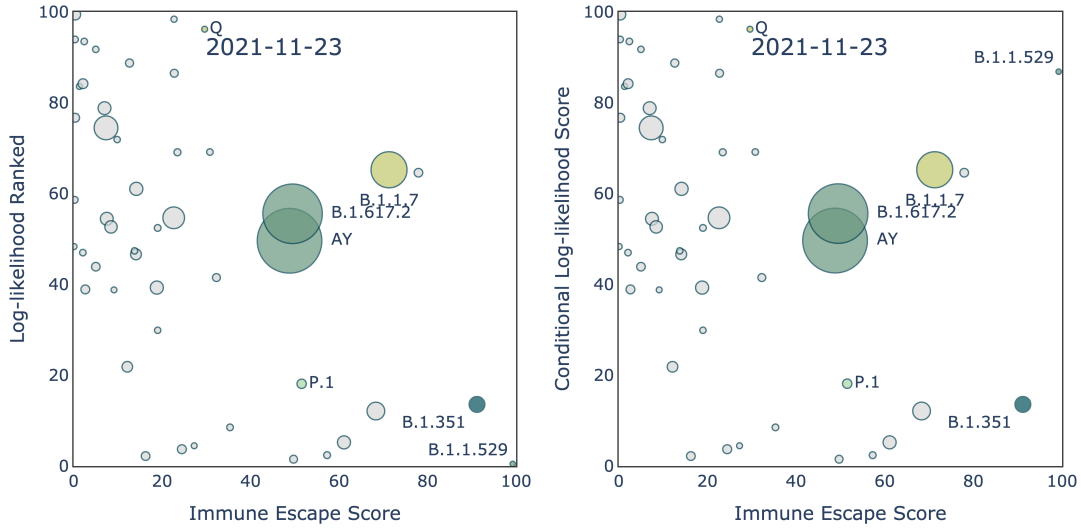


Figure S.4. Log-likelihood score corrects for large mutation count. (Left) Snapshot of lineages in terms of log-likelihood ranked without correction for large mutation count and immune escape score on November 23rd 2021. Marker size indicates the number of submissions of each lineage. **(Right)** Same plot where the log-likelihood ranked without correction has been replaced by its corrected version. Note, that both plots are nearly identical, as highly mutated sequences comprise less than 1% of the entire data set. We observe nearly no change between the plots, with concerning lineages residing on the second Pareto front, except for the emergence of B.1.1.529 (Omicron) as a clear outlier, practically alone in the first Pareto front, due to its high immune escape and extraordinary log-likelihood, given its high number of mutations. Additionally, the conditional log-likelihood score is nearly collinear with expected prevalence of sequence in population (see Fig. S.8)

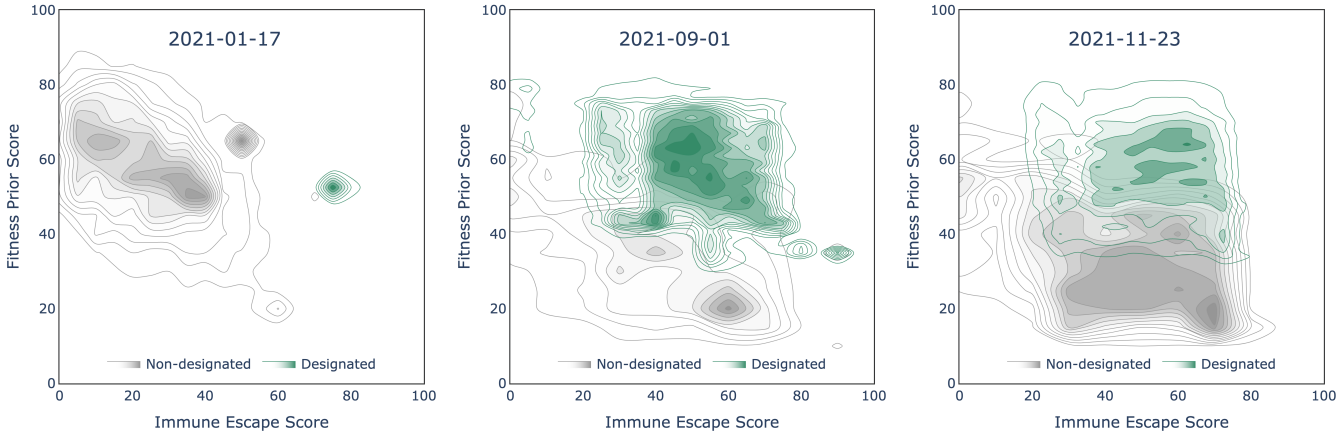


Figure S.5. Combining immune escape and fitness prior for continuous monitoring. (Left) Density contour plot of *sequences* on January 17th, 2021. Sequences are split into two groups: WHO designated ones and other non-designated ones. **(Centre)** Density contour plot on September 1st, 2021. **(Right)** Density contour plot on November 23rd, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

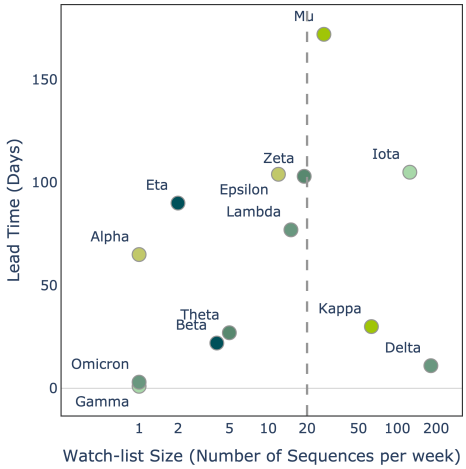


Figure S.6. The maximum lead time of EWS detection ahead of WHO designation vs. required weekly watch-list size. With a weekly watch list of 200 sequences, all WHO designated variants are detected, including Delta.

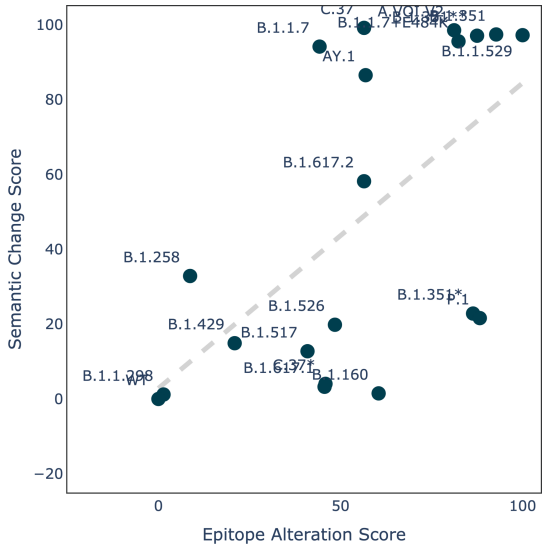


Figure S.7. Metrics of anticipated reduction of the immune response. Semantic change and Epitope alteration score accurately segment the variant landscape, allowing to discriminate between variants that do not have immune escape propensity (B.1.429, WT), highly mutated, but neutralizable variants (P.1, B.1.160), and those with high potential for evading immune response (B.1.1.7, AY.1, B.1.351).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

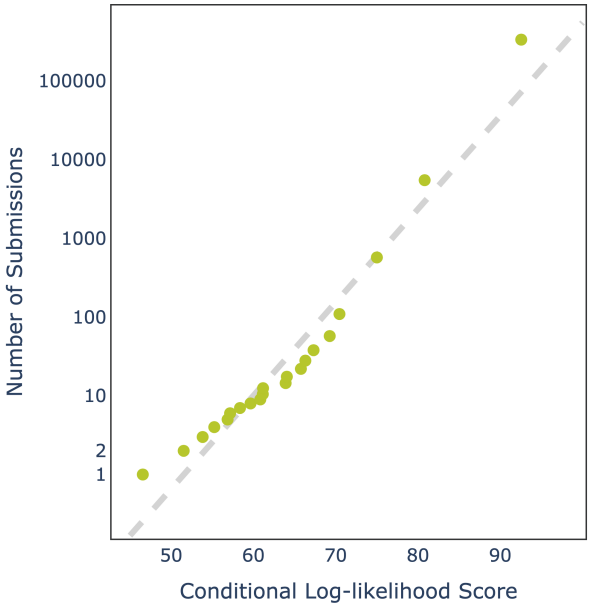


Figure S.8. Validation of the conditional log-likelihood score. Sequences are grouped into bins based on their submission count and the conditional log-likelihood scores and number of submissions were averaged per bin. The first ten bins correspond to count 1 to 10. The next 10 bins are equally split between counts 11 and 1000 such that each bin has a similar number of sequences. The last two bin contains all sequences having a submission count from 1000 to 10,000 and sequences having more than 10,000 submissions. This shows that the mean conditional log-likelihood of sequences that are observed frequently in circulation is much higher than that of outlier, infrequent sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S.1. N=310 nAbs Resolved 3D Structures PDB Identifiers.

Code	Accession Date	Code	Accession Date	Code	Accession Date	Code	Accession Date	Code	Accession Date
6W41	2020-03-09	7MMO	2021-04-30	7KXJ	2020-12-04	7E5Y	2021-02-21	7CH4	2020-07-05
6WPS	2020-04-27	7MY2	2021-05-20	7KXK	2020-12-04	7E7X	2021-02-28	7CH5	2020-07-05
6WPT	2020-04-27	7MY3	2021-05-20	7KZB	2020-12-10	7E7Y	2021-02-28	7CHB	2020-07-05
6XC2	2020-06-08	7N0G	2021-05-25	7L02	2020-12-10	7E86	2021-03-01	7CHC	2020-07-05
6XC3	2020-06-08	7N0H	2021-05-25	7L06	2020-12-11	7E88	2021-03-01	7CHE	2020-07-05
6XC4	2020-06-08	7N0I	2021-05-25	7L09	2020-12-11	7E8C	2021-03-01	7CHF	2020-07-05
6XC7	2020-06-08	7N0R	2021-05-25	7L0N	2020-12-11	7E8F	2021-03-01	7CHH	2020-07-05
6XCM	2020-06-08	7N3C	2021-05-31	7L2C	2020-12-16	7E8M	2021-03-02	7CHO	2020-07-06
6XCN	2020-06-08	7N3D	2021-05-31	7L2D	2020-12-16	7EAM	2021-03-07	7CHP	2020-07-06
6XDG	2020-06-10	7N3I	2021-06-01	7L2E	2020-12-16	7EAN	2021-03-07	7CHS	2020-07-06
6XE1	2020-06-11	7N62	2021-06-07	7L2F	2020-12-16	7EJ4	2021-04-01	7CJF	2020-07-10
6XEY	2020-06-14	7N64	2021-06-07	7L3N	2020-12-18	7EJ5	2021-04-01	7CM4	2020-07-24
6XKP	2020-06-26	7N8H	2021-06-14	7L56	2020-12-21	7F62	2021-06-24	7CR5	2020-08-12
6XKQ	2020-06-26	7N8I	2021-06-14	7L57	2020-12-21	7F63	2021-06-24	7CWL	2020-08-29
6YLA	2020-04-06	7N9A	2021-06-17	7L58	2020-12-21	7JMO	2020-08-02	7CWM	2020-08-29
6YM0	2020-04-07	7N9B	2021-06-17	7L5B	2020-12-21	7JMP	2020-08-02	7CWN	2020-08-29
6YZ5	2020-05-06	7N9C	2021-06-17	7LAA	2021-01-06	7JMW	2020-08-03	7CWO	2020-08-29
6YZ7	2020-05-06	7N9E	2021-06-17	7LAB	2021-01-06	7JV2	2020-08-20	7CWS	2020-08-31
6Z2M	2020-05-17	7N9T	2021-06-18	7LCN	2021-01-11	7JV4	2020-08-20	7CWT	2020-08-31
6Z43	2020-05-22	7ND4	2021-01-30	7LD1	2021-01-12	7JV6	2020-08-20	7CWU	2020-08-31
6ZBP	2020-06-08	7ND5	2021-01-30	7LDJ	2021-01-13	7JVA	2020-08-20	7CYH	2020-09-03
6ZCZ	2020-06-12	7ND6	2021-01-30	7LJR	2021-01-30	7JVB	2020-08-20	7CYP	2020-09-04
6ZDG	2020-06-14	7ND7	2021-01-30	7LM8	2021-02-05	7JVC	2020-08-20	7CZP	2020-09-09
6ZDH	2020-06-14	7ND8	2021-01-30	7LOP	2021-02-10	7JW0	2020-08-24	7CZQ	2020-09-09
6ZER	2020-06-16	7ND9	2021-01-30	7LQV	2021-02-15	7JX3	2020-08-26	7CZR	2020-09-09

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

6ZFO	2020-06-17	7NDA	2021-01-30	7LQW	2021-02-15	7K43	2020-09-14	7CZS	2020-09-09
6ZH9	2020-06-21	7NDB	2021-01-30	7LRS	2021-02-17	7K45	2020-09-14	7CZT	2020-09-09
6ZHD	2020-06-22	7NDC	2021-01-30	7LRT	2021-02-17	7K4N	2020-09-15	7CZU	2020-09-09
6ZLR	2020-07-01	7NDD	2021-01-30	7LS9	2021-02-17	7K8M	2020-09-27	7CZV	2020-09-09
6ZXN	2020-07-30	7NEG	2021-02-04	7LSS	2021-02-18	7K8S	2020-09-27	7CZW	2020-09-09
7A25	2020-08-16	7NEH	2021-02-04	7LX5	2021-03-03	7K8T	2020-09-27	7CZX	2020-09-09
7A29	2020-08-16	7NKT	2021-02-18	7LXW	2021-03-05	7K8U	2020-09-27	7CZY	2020-09-09
7A5R	2020-08-21	7NTC	2021-03-09	7LXX	2021-03-05	7K8V	2020-09-27	7CZZ	2020-09-09
7A5S	2020-08-21	7NX6	2021-03-17	7LXY	2021-03-05	7K8W	2020-09-27	7D00	2020-09-09
7AKD	2020-09-30	7NX7	2021-03-17	7LXZ	2021-03-05	7K8X	2020-09-27	7D03	2020-09-09
7B14	2020-11-23	7NX8	2021-03-17	7LY0	2021-03-05	7K8Y	2020-09-27	7D0B	2020-09-09
7B17	2020-11-23	7NX9	2021-03-17	7LY2	2021-03-05	7K8Z	2020-09-27	7D0C	2020-09-09
7B18	2020-11-24	7NXA	2021-03-17	7LY3	2021-03-05	7K90	2020-09-27	7D0D	2020-09-09
7B27	2020-11-26	7NXB	2021-03-17	7M3I	2021-03-18	7K9Z	2020-09-29	7D2Z	2020-09-17
7B3O	2020-12-01	7OAN	2021-04-19	7M42	2021-03-19	7KFV	2020-10-15	7D30	2020-09-17
7BEH	2020-12-23	7OAO	2021-04-19	7M6D	2021-03-25	7KFW	2020-10-15	7D4G	2020-09-23
7BEI	2020-12-23	7OAP	2021-04-19	7M6E	2021-03-25	7KFX	2020-10-15	7DCC	2020-10-24
7BEJ	2020-12-23	7OAQ	2021-04-20	7M6F	2021-03-25	7KFY	2020-10-15	7DCX	2020-10-27
7BEK	2020-12-23	7OAU	2021-04-20	7M6G	2021-03-25	7KGJ	2020-10-16	7DD2	2020-10-27
7BEL	2020-12-23	7OAY	2021-04-20	7M6H	2021-03-25	7KGK	2020-10-16	7DD8	2020-10-28
7BEM	2020-12-24	7OLZ	2021-05-20	7M6I	2021-03-25	7KKK	2020-10-27	7DEO	2020-11-04
7BEN	2020-12-24	7OR9	2021-06-04	7M71	2021-03-26	7KKL	2020-10-27	7DET	2020-11-05
7BEO	2020-12-24	7ORA	2021-06-04	7M7B	2021-03-27	7KLG	2020-10-30	7DEU	2020-11-05
7BEP	2020-12-24	7ORB	2021-06-04	7M7W	2021-03-29	7KLH	2020-10-30	7DJZ	2020-11-22
7BWJ	2020-04-14	7P77	2021-07-19	7M8J	2021-03-29	7KLW	2020-11-01	7DK0	2020-11-22
7BYR	2020-04-24	7P78	2021-07-19	7MDW	2021-04-06	7KM5	2020-11-02	7DK4	2020-11-23
7BZ5	2020-04-26	7P79	2021-07-19	7ME7	2021-04-06	7KMG	2020-11-02	7DK5	2020-11-23
7C01	2020-04-29	7P7A	2021-07-19	7MEJ	2021-04-06	7KMH	2020-11-02	7DK6	2020-11-23

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7C2L	2020-05-08	7R6W	2021-06-23	7MF1	2021-04-08	7KMI	2020-11-02	7DK7	2020-11-23
7C8V	2020-06-03	7R6X	2021-06-23	7MFU	2021-04-11	7KMK	2020-11-03	7DPM	2020-12-20
7C8W	2020-06-03	7R7N	2021-06-25	7MJI	2021-04-20	7KML	2020-11-03	7DZX	2021-01-26
7CAC	2020-06-08	7R8L	2021-06-26	7MJJ	2021-04-20	7KN5	2020-11-04	7DZY	2021-01-26
7CAH	2020-06-08	7R8M	2021-06-26	7MJK	2021-04-20	7KN6	2020-11-04	7E23	2021-02-04
7CAI	2020-06-08	7R8N	2021-06-26	7MJL	2021-04-20	7KN7	2020-11-04	7KSG	2020-11-22
7CAK	2020-06-08	7R8O	2021-06-26	7MKL	2021-04-24	7KQB	2020-11-14	7MM0	2021-04-29
7CAN	2020-06-09	7R98	2021-06-28	7MKM	2021-04-24	7KQE	2020-11-15	7RAL	2021-07-01
7CDI	2020-06-19	7RA8	2021-06-30	7MLZ	2021-04-29	7KS9	2020-11-21	7CDJ	2020-06-19

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S.2. EWS Scores. EWS has in total five sub-scores grouped into immune escape and fitness prior scores. Each sub-score is normalised ranks that range between 0 and 100%. The average of the sub-scores in each score category is computed to define immune escape and fitness prior scores.

Score Name	Score Category	Description
Epitope Alteration	Immune Escape	Measures the alteration of the Spike protein at epitope positions by counting the number of antibodies potentially escaped.
Semantic Change	Immune Escape	Measures the functional change of the Spike protein using Transformer-derived embedding differences with respect to the wild type and D614G.
ACE2 Binding	Fitness Prior	Approximates the binding affinity using binding energies estimated from in-silico simulations.
Conditional Log-Likelihood	Fitness Prior	Measures the relative rank of the measured existence probability of the Spike protein using Transformer-derived log-likelihoods with reference to other sequences with similar mutation count.
Growth	Fitness Prior	Calculated lineage-level growth using deposited sequence metadata (retrospective data).

Table S.3. Spike mutations in SARS-CoV-2 Spike pseudoviruses and observed reduction of neutralising antibody response in pseudovirus neutralisation assay.

Lineage	WHO Nomenclature	Mutation	Reduction
B.1.1.7	Alpha	H69- V70- Y144- N501Y A570D D614G P681H T716I S982A D1118H	22.92%
B.1.1.7+E484K	Alpha	H69- V70- Y144- E484K N501Y A570D D614G P681H T716I S982A D1118H	74.65%
B.1.351	Beta	L18F D80A D215G L242- A243- L244- R246I K417N E484K N501Y D614G A701V	80.04%
B.1.351*	Beta	D80A D215G L242H K417N E484K N501Y D614G A701V	47.19%
B.1.351**	Beta	D80A D215G L242- A243- L244- K417N E484K N501Y D614G A701V	77.45%
P.1	Gamma	L18F T20N P26S D138Y R190S K417T E484K N501Y H655Y T1027I V1176F	47.66%
B.1.617.2	Delta	T19R G142D E156G F157- R158- K417N L452R T478K D614G P681R D950N	49.43%
AY.1	Delta	T19R T95I G142D E156G F157- R158- W258L K417N L452R T478K K558N D614G P681R D950N	48.09%
B.1.427/B.1.429	Epsilon	S13I W152C L452R D614G	52.57%

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B.1.526	Iota	L5F T95I D253G E484K D614G A701V	10.94%
B.1.617.1	Kappa	L452R E484Q D614G P681R	18.34%
C.37	Lambda	G75V T76I R246- S247- Y248- L249- T250- P251- G252- D253N L452Q F490S D614G T859N	34.22%
C.37*	Lambda	G75V T76I L452Q F490S D614G T859N	16.88%
B.1.1.529	Omicron	A67V H69- V70- T95I G142D V143- Y144- Y145- N211I L212V ins214EPE G339D S371L S373P S375F K417N N440K G446S S477N T478K E484A Q493R G496S Q498R N501Y Y505H T547K D614G H655Y N679K P681H N764K D796Y N856K Q954H N969K L981F	97.52%
A.VOI.V2		D80Y Y144- I210- D215G R246- S247- Y248- L249M W258L R346K T478R E484K H655Y P681H Q957H	64.54%
B.1.1.298		Y453F D614G I692V M1229I	5.42%
B.1.160		S477N S494P D614G K1191N	3.81%
B.1.258		H69- V70- L189F N439K D614G V772I	-9.28%
B.1.517		G181V G252V N501T D614G P812L	-7.87%

Table S.4. Early detection of variants of concerns. The summary table shows that the EWS can detect WHO designated variants months before the WHO official designation date. The average lead time for early detection across is 58 days.

WHO Label	Lineage	EWS Detection Date	WHO Designation Date	First Submission Date	Lead Time (Days)	Number of registered sequences upon WHO Detection	Number of registered sequences upon EWS Designation
Alpha	B.1.1.7	2020-10-14	2020-12-18	2020-10-14	65	1881	1
Beta	B.1.351	2020-11-27	2020-12-18	2020-11-26	21	218	15
Gamma	P.1	2021-01-10	2021-01-11	2021-01-10	1	29	4
Delta	B.1.617.2		2021-04-04	2021-03-23		5	
Epsilon	B.1.429	2020-11-21	2021-03-05	2020-11-21	104	9957	1
Zeta	P.2	2020-12-04	2021-03-17	2020-12-02	103	1450	3
Eta	B.1.525	2021-01-04	2021-04-04	2021-01-04	90	1377	6
Theta	P.3	2021-02-25	2021-03-24	2021-02-25	27	102	1
Iota	B.1.526	2021-01-04	2021-03-24	2020-12-09	79	1120	2
Kappa	B.1.617.1	2021-03-31	2021-04-04	2021-03-05	4	167	167
Lambda	C.37	2021-02-26	2021-05-14	2021-02-26	77	719	4
Mu	B.1.621	2021-04-30	2021-08-30	2021-03-11	122	3671	88
Omicron	B.1.1.529	2021-11-23	2021-11-26	2021-11-23	3	11	10

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S.5. Welch's t-test p-values. Every week, all registered variants are scored with the Pareto score. Welch's t-tests are conducted to assess if respectively WHO designated variants and VOCs can be separated from others. p-values are reported every week between January 2020 and November 2021 .

	Welch t-test p-values			Welch t-test p-values			Welch t-test p-values	
Week	WHO Designated Variants	VOC	Week	WHO Designated Variants	VOC	Week	WHO Designated Variants	VOC
2020-12-13	8E-118	8E-118	2021-04-11	3E-05	4E-18	2021-08-08	7E-06	5E-09
2020-12-20	7E-07	7E-07	2021-04-18	5E-06	2E-20	2021-08-15	2E-05	9E-06
2020-12-27	5E-128	5E-128	2021-04-25	6E-06	4E-08	2021-08-22	4E-07	3E-21
2021-01-03	7E-06	7E-06	2021-05-02	2E-12	5E-07	2021-08-29	3E-05	4E-05
2021-01-10	5E-140	5E-140	2021-05-09	3E-08	5E-06	2021-09-05	1E-04	5E-04
2021-01-17	2E-07	2E-07	2021-05-16	4E-09	8E-06	2021-09-12	5E-05	1E-04
2021-01-24	4E-06	4E-06	2021-05-23	1E-06	4E-04	2021-09-19	1E-04	2E-04
2021-01-31	5E-144	5E-144	2021-05-30	6E-07	4E-04	2021-09-26	2E-06	9E-06
2021-02-07	2E-06	2E-06	2021-06-06	2E-04	2E-04	2021-10-03	1E-04	1E-05
2021-02-14	6E-28	4E-19	2021-06-13	4E-05	2E-05	2021-10-10	2E-06	7E-06
2021-02-21	7E-21	9E-08	2021-06-20	8E-06	2E-04	2021-10-17	1E-07	4E-07
2021-02-28	1E-16	2E-08	2021-06-27	3E-05	3E-05	2021-10-24	7E-07	3E-08
2021-03-07	8E-07	2E-19	2021-07-04	1E-04	2E-03	2021-10-31	2E-05	2E-05
2021-03-14	2E-05	2E-07	2021-07-11	1E-05	1E-05	2021-11-07	2E-04	3E-04
2021-03-21	1E-05	1E-20	2021-07-18	1E-04	3E-04	2021-11-14	4E-03	3E-04
2021-03-28	1E-05	1E-22	2021-07-25	1E-03	1E-07	2021-11-21	5E-05	7E-06
2021-04-04	4E-06	2E-19	2021-08-01	2E-04	6E-09			

Table S.6. Comparison between EWS detection capabilities and three baselines. Two baselines are based on unsupervised learning (UMAP) and one baseline is supervised (GLM).

HRV	Number Days Ahead UMAP	Number Days Ahead GLM	Number Days Ahead EWS Semantic Change only	Number Days Ahead EWS
Alpha	-47	-38	65	65
Beta	-	-46	21	21
Gamma	-	1	-121	1
Delta	-58	-25	-78	-
Epsilon	64	-27	66	104

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zeta	-	75	-	103
Eta	-	90	90	90
Theta	-	-68	13	27
Iota	-96	79	-49	79
Kappa	-	-25	-29	4
Lambda	-77	-5	77	77
Mu	-	-8	12	122
Omicron	-	-	3	3

Materials and Methods.

We describe the proposed methodologies in detail in the following sections.

Variant Notations

We refer to a variant as a protein sequence of a coronavirus' Spike protein that differs from the original Wuhan Spike protein that we refer to as wild type. We represent a variant in terms of its mutations with respect to the Wuhan strain. For instance, the notation N501Y represents a substitution in position 501, replacing N with Y; the notation ins214AR represents inserting AR after position 213; and notation H69- V70- represents a deletion of H and V at positions 69 and 70.

VSV-SARS-CoV-2 S pseudovirus neutralisation assay

A recombinant replication-deficient VSV vector that encodes green fluorescent protein (GFP) and luciferase (Luc) instead of the VSV-glycoprotein (VSV-G) was pseudotyped with SARS-CoV-2 Spike (S) protein derived from either the Wuhan reference strain (NCBI Ref: 43740568) or variants of interest according to published pseudotyping protocols[1]. The mutations found in S of the VOCs are listed in Table S.3. In brief, HEK293T/17 monolayers transfected to express SARS-CoV-2 S with the C-terminal cytoplasmic 19 amino acids truncated (SARS-CoV-2-S[CΔ19]) were inoculated with the VSVΔG-GFP/Luc vector. After incubation for 1 hour at 37 °C, the inoculum was removed, and cells were washed with PBS before medium supplemented with anti-VSV-G antibody (clone 8G5F11, Kerafast) was added to neutralise residual input virus. VSV-SARS-CoV-2 pseudovirus-containing medium was collected 20 hours after inoculation, 0.2 µm filtered and stored at −80 °C.

For pseudovirus neutralisation assays, 40,000 Vero 76 cells were seeded per 96-well. Sera were serially diluted 1:2 in culture medium starting with a 1:15 dilution (dilution range of 1:15 to 1:7,680). VSV-SARS-CoV-2-S pseudoparticles were diluted in culture medium to obtain either ~1,000 or ~200 transducing units (TU) in the assay. Same input virus amounts for all pseudoviruses were used within an individual experiment. In total 8 experiments were performed covering the SARS-CoV-2 variants listed in Table S.3 always taking Wuhan S pseudovirus as reference. Serum dilutions were mixed 1:1 with pseudovirus for 30 minutes at room temperature prior to addition to Vero 76 cell monolayers and incubation at 37 °C for 24 hours. Supernatants were removed,

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the cells were lysed with luciferase reagent (Promega). Luminescence was recorded, and neutralisation titres were calculated by generating a four-parameter logistic fit of the percent neutralisation at each serial serum dilution. The $pVNT_{50}$ is reported as the interpolated reciprocal of the dilution yielding a 50% reduction in luminescence. If no neutralisation yielding a 50% reduction in luminescence was observed, an arbitrary titer value of 7.5, half of the limit of detection (LO), was reported.

Binding kinetics of RBD variants to ACE2 using surface plasmon resonance spectroscopy

Binding kinetics of RBD variants was determined using a Biacore T200 device (Cytiva) with HBS-EP+ running buffer (BR100669, Cytiva) at 25 °C. Carboxyl groups on the CM5 sensor chip matrix were activated with a mixture of 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) and N-hydroxysuccinimide (NHS) to form active esters for the reaction with amine groups. anti-mouse-Fc-antibody (BR100838, Cytiva) was diluted in 10 mM sodium acetate buffer pH 5 (30 µg/mL) for covalent coupling to immobilisation level of ~6,000 response units (RU). Free N-hydroxysuccinimide esters on the sensor surface were deactivated with ethanolamine-HCl.

Human ACE2-mFc (10108-H05H, Sino Biological Inc.) was diluted to 5 µg/mL with HBS-EP+ buffer and applied at 10 µL/min for 15 seconds to the active flow cell for capture by immobilised antibody, while the reference flow cell was treated with buffer. Binding analysis of captured hACE2-mFc to RBD variants (40592-V08B, 40592-V08H113, 40592-V08H115, 40592-V08H82, 40592-V08H59, 40592-V08H84, 40592-V08H85, 40592-V08H112, 40592-V08H28, 40592-V08H81, 40592-V08H90, 40592-V08H91, 40592-V08H88, 40592-V08H86, 40592-V08H111, 40592-V08H80, 40592-V08H1, 40592-V08H14, 40592-V08H46, 40592-V08H121 **40592-V08H121,** Sinobiological Inc.) was performed using a multi-cycle kinetic method with concentrations ranging from 3.125 to 50 nM. An association period of 120 seconds was followed by a dissociation period of 300 seconds with a constant flow rate of 30 µL/min and a final regeneration step. Binding kinetics were calculated using a global kinetic fit model (1:1 Langmuir, Biacore T200 Evaluation Software Version 3.1, Cytiva).

Data

The genomic sequences and Spike protein sequences were collected from GISAID. For each Spike protein sequence, the missing amino acids were filled using the next known amino acid and the lineage assignment was performed using PANGOLIN[2]. Mutations with respect to the wild type were calculated using Clustal Omega[3] and HH-suite[4].

The GISAID dataset is imbalanced towards some lineages that have been more prevalent and because certain regions have performed more sequencing than others. To mitigate this bias in the dataset during training, we weighed the importance of each sequence differently in the loss calculation. The importance of a sequence is defined as

$$w_s \propto \sum_{(s,l)\sim D} \log_{10}(c_s) + \log_{10}(c_{(s,l)}) + \log_{10}(c_{l|s})$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where the values c_s and $c_{s,l}$ are the numbers of occurrences in the dataset of the sequence s and the sequence-laboratory pair (s, l), respectively. The value $c_{l|s}$ corresponds to the number of laboratories having reported sequence s, which measure the prevalence across regions of the variant.

The model excludes from training all the sequences which have been observed only once in the data set. In this way it eliminates spurious changes, due to sequencing errors, as well as samples of viruses of subpar evolutionary fitness, which do not spread between patients.

Language modelling

The domain of Natural Language Processing (NLP) has experienced several breakthroughs in the past years. The emergence of recurrent and attention-based deep neural networks led to impressive results for text generation and translation. Recently, this technology has been leveraged to learn the language of biology[5,6]. It works with a simple analogy where protein sequences are considered as sentences and the amino acids as words. The models are trained on large datasets of known protein sequences in an unsupervised manner. In other words, there is no need to label the data and any newly registered protein sequence can be exploited.

Information about protein properties is stored at two positions inside the model once it is trained. On one side, the probabilities returned by the model indicate how likely this sequence is to be natural/viable/feasible. On the other hand, the outputs of the model's layers and notably the last layer provide a high dimensional representation for each sequence, which we call embedding of the protein. The embedding of the protein contains information about the protein properties and can be used either directly or to train a classification or regression model. Recently, (Meier et al.) demonstrated that these models also capture the effects of mutations on protein function [7].

Model architecture

In this work, the input of the model consists of the sequence characters corresponding to the amino acids forming the protein. Each amino acid is first tokenized, i.e. mapped to their index in the vocabulary containing the 20 natural amino acids and then projected to an embedding space. The sequence of embeddings is then fed to the transformer model[8] consisting of a series of blocks, each composed of a self-attention operation followed by a position-wise multi-layer network (Fig. S.1).

Self-attention modules explicitly construct pairwise interactions between all positions in the sequence which enable them to build complex representations that incorporate context from across the sequence. Because the self-attention operation is permutation-equivariant, a positional encoding must be added to the embedding of each token to distinguish its position in the sequence.

Self-supervised Training

Given a large database of protein sequences, the model can be trained using the masked language modelling objective presented in [31]. Each input sequence is corrupted by replacing a fraction of the amino acids with a special mask token. The network is then trained to predict the missing tokens from the corrupted sequence. In practice, for each sequence x, we randomly sample a set of indices $i \in M$, for which the amino acid tokens are

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

replaced by a mask token, resulting in a corrupted sequence $\hat{\mathbf{x}}$. During pre-training, the set M is defined such that 15% of the amino acids in the sequence get corrupted. When corrupted an amino acid has 10% to be replaced by another randomly selected amino acid and 80% being masked. During fine-tuning these probabilities do not change, however, only 2.5% of the amino acids in the sequence get corrupted. This probability was selected in order to enable the model to become more accurate for Spike protein sequences while keeping its performance on diverse sequences from UniRef100. The training objective corresponds to the negative log-likelihood of the true sequence at the corrupted positions.

$$L_\theta(\hat{\mathbf{x}}|\mathbf{x}) = -\sum_{i \in M} \log p(\mathbf{x}_i|\hat{\mathbf{x}})$$

To minimise this loss, the model must learn to identify dependencies between the corrupted and uncorrupted elements of the sequence. Consequently, the learned representations of the proteins, taken as the average of the embeddings of each amino acid (Fig. S.1), must successfully extract generic features of the biological language of proteins. These features can then be used to fine-tune the model on downstream tasks.

In this work, we used the transformer model from (Rives et al, 2021) (esm1_t34_670M_UR100) which was trained using the aforementioned procedure on the UniRef100 dataset[9], containing >277M representative sequences. The pre-trained model was then fine-tuned every month on all the Spike protein sequences registered in the GISAID data bank at the training date.

Gradient descent is used to minimise the loss function. We relied on the Adam optimizer[10] and used a learning rate schedule. Batch size is 1. The fine-tuning started with a warm-up period of 100 mini-batches where the learning rate increased linearly from 10^{-7} to 10^{-5}. After the warm-up period, the learning rate decreased following $10^{-6}\sqrt{k}$ where k represents the number of mini-batches.

Inference and ML Scores Calculations

Once fine-tuned, the model is used to compute the semantic change and the log-likelihood to characterise a Spike protein sequence.

Formally, an input sequence is represented by a sequence of tokens defined as $\mathbf{x} = (x_1, \ldots, x_n)$ where n is the number of tokens and $\forall i \in [1, n], \quad x_i \in \mathcal{X}$ where $\mathcal{X}$ is a finite alphabet that contains the amino-acids and other tokens such as class and mask tokens. In this work, a class token is appended to all sequences before feeding them to the network, as such x_1 represents the class token, while $x_2, \ldots, x_n$ represents the amino-acids, or masked amino-acids, in the spike protein sequence. The sequence $\mathbf{x}$ is passed through attention layers. We note $\mathbf{z} = (z_1, \ldots, z_n)$ the output of the last attention layer where z_i is the sequence embedding vector at position i.

Please note that in our architecture, embedding vector z_i is a function of all input tokens $(x_j)_{j \in [1,n]}$. In opposition, in Bi-LSTM architectures[11], z_i would be a function of all inputs tokens except the one at the position i, $(x_j)_{j \in [1,n], j \neq i}$.

In order to represent a protein sequence through a single embedding vector, which size does not depend on the protein sequence length, we consider

$$\bar{\mathbf{z}} = \frac{1}{n-1} \sum_{i=2}^{n} z_i,$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that we call embedding vector of the variant represented by sequence $\mathbf{x}$. Note that the summation starts at the second position to ensure that the class token's embedding, which is at the first position, does not contribute to the sequence embedding.

The embedding of the Wuhan strain $\bar{\mathbf{z}}_{\text{wuhuan}}$ and the embedding of the D614G variant $\bar{\mathbf{z}}_{\text{D614G}}$ are computed once for all. In this work, the semantic change of a variant $\mathbf{x}$ is computed as:

$$\Delta\bar{\mathbf{z}} = ||\bar{\mathbf{z}} - \bar{\mathbf{z}}_{\text{wuhuan}}||_1 + ||\bar{\mathbf{z}} - \bar{\mathbf{z}}_{\text{D614G}}||_1,$$

where $||.||_1$ is the L1 norm.

The last attention layer output $\mathbf{z}$ is transformed by a feed-forward layer and a softmax activation into a vector of probabilities over tokens at each positions $\mathbf{p} = (p_1, \ldots, p_n)$ where p_i is a vector of probabilities at position i, $p_i = (p(x_i = x_1 \mid \mathbf{x}), \ldots, p(x_i = x_n \mid \mathbf{x}))$.

The log-likelihood of a variant $l(\mathbf{x})$ is computed from these probabilities. It is calculated as the sum of the log probabilities over all the positions of the Spike protein amino acids. Formally,

$$l(\mathbf{x}) = \sum_{i=2}^{n} \log p(x_i = x_i \mid \mathbf{x}).$$

This quantity measures the likelihood of observing a variant sequence $\mathbf{x}$ according to the model. Therefore, the more sequences in the training data that are similar to a considered variant, the higher the log-likelihood of this variant will be. The proposed log-likelihood metric supports substitution, insertion, and deletion without the requirement of a reference.

Implementation Details

The method is implemented using the Pytorch[12] deep learning framework. Model training and inferences are performed on InstaDeep's high performance computing infrastructure. The average training and inference time is <4 GPU days and <12 GPU hours, respectively, using Nvidia A100-SXM4-40GB GPUs.

Epitope Alteration Score

Epitope alteration score attempts to capture the impact of mutations in the variant in question on recognition by experimentally assessed antibodies. To compute this score, we enumerate the number of *unique* epitopes involving altered positions, as measured across all the antibody-Spike complex structures deposited in Protein Data Bank.

This score, by design, highly emphasises the effect of mutations on highly antigenic sites, such as the receptor-binding domain (RBD). This allows to approximate the expected weight of mutations, and to ascribe importance to non-RBD mutations, if and only if sufficient escape potential with regard to RBD-targeting antibodies is achieved.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ACE2 Binding Score

We selected 279 receptor-binding domain (RBD) differentiated variants, including the wide type, for in-silico simulation. For each variant, we generated a putative structure from which we generated at least 500 structures through a conformational sampling algorithm. These structures were further optimised with a probabilistic optimization algorithm, a variant of simulated annealing, aiming to overcome local energy barriers and follow a kinetically accessible path toward an attainable deep energy minimum with respect to a knowledge-based, protein-oriented potential. This results in 214,142 structures in total. For each structure, we calculated the change of binding energy when the interface forming chains are separated, versus when they are complexed. These measurements were aggregated per RBD variant using medians. Each metric is normalised by the metric on wild type, corresponding to no mutation on RBDs, such that the metrics for the wild type are all ones. Sequences having other RBD mutation combinations, representing very rare RBDS, corresponding to <9% of all known sequences, were excluded from this analysis, due to reasons of computational efficiency.

Growth Score

The growth score is computed from the GISAID metadata. At a given date, we considered only sequences that have been submitted within the last eight weeks. For each lineage, its proportion among all submissions was calculated for the eight-week window and for the last week, denoted by r_{win} and r_{last} correspondingly. The growth of the lineage is defined by their ratio, r_{last} / r_{win}, measuring the change of the proportion. Having values larger than one means that the lineage is rising and smaller than one indicates declining.

Scores scaling and merging

The semantic change, log-likelihood, epitope alteration score, ACE2 binding score, and growth have all different scales and units. Thus, they can not be compared directly. To make comparisons possible, a scaling strategy is introduced. For a given metric m, all the variants considered are ranked according to this metric. In the ranking system used, the higher rank the better. Variants with the same value will get the same rank. The ranks are then transformed into values between 0 and 100 through a linear projection to obtain the values for the scaled metric. All reported scores, except for log-likelihood, have been scaled according to this strategy.

Log-likelihood was observed to strongly penalise variants with a large number of mutations. An increased number of mutations may strongly affect fitness, thus explaining decreased log-likelihood. However, as the variants that are scored by EWS have been registered, which implies that they managed to infect hosts and replicate sufficiently to be detected, we hypothesise that they have at least minimal fitness. A variant with two mutations, whose log-likelihood is in the bottom 20th percentile globally, is less likely to survive the evolutionary competition. A variant, with analogous log-likelihood, but with twenty mutations is more likely among others, similarly mutated ones. Thus, we introduced a group-based ranking strategy where each variant is ranked only among variants with a similar number of mutations. For each variant, having N mutations, its conditional log-likelihood score is ranked among all variants having at least M mutations, with $M = \min(\max(0, N-10), 50)$. Deletions at N-terminal or C-terminal are considered as one single mutation for grouping. In each group, as for the other ranking technique, the ranks are then transformed into values between 0 and 100 through a linear projection to obtain the values for the scaled metric. Although this work uses all the samples that have no less than ten mutations fewer than the query, results are largely robust to the choice of a threshold.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The immune escape score is computed as the average of the scaled semantic change and of the scaled epitope alteration score. The fitness prior score is computed as the average of the scaled conditional log-likelihood, the scaled ACE2 binding score, and the scaled growth.

Pareto Score

Pareto optimality is defined over a set of lineages. Lineages are Pareto optimal within that set if there are no lineages in the set with both higher immune escape and higher fitness prior scores. The Pareto score is a measure of the degree of Pareto optimality. Lineages with the highest Pareto score are Pareto optimal. Lineages with the second-best Pareto score would be Pareto optimal, if the Pareto optimal lineages were removed from the set, and so on.

To compute the Pareto score, we first compute all the Pareto fronts that exist in the considered set of lineages. The first Pareto front corresponds to the set of lineages for which there does not exist any other lineage with both higher immune escape and fitness prior score. The second Pareto front is computed as the Pareto front over the set of lineages remaining when removing the ones from the first Pareto front. Successive Pareto fronts are computed until all the lineages are assigned to a front. Then, a linear projection is used so that the lineages from the first front obtain a Pareto score of 100 and the ones from the last front get a Pareto score of 0.

Semantic Change vs Epitope Alteration Score

Semantic change is a measure of *how different* is the variant in question with regard to the underlying statistical model (large ML model fine-tuned on Spike protein sequences observed until a given time point). This value depends on the observations. Epitope alteration score is a measure of *how many distinct epitopes are evaded* by the variant in question in comparison to wild type. This score, on the other hand, is computed purely based on known binding sites of antibodies, as reported in Protein Data Bank. It too changes with time with new discoveries of anti-Spike antibodies, but to a lesser extent and is expected to converge to a stable value.

We observe that these scores, while intuitively correlated, are not collinear. In particular, we see that most of HRVs regarded as immune escaping (and denoted as VoCs, VoIs etc.) indeed exhibit high semantic change, but are rather diverse in terms of Epitope Alteration Score (Fig. S.7).

Detection

During retrospective early detection analysis, each week EWS considers only the new sequences reported during that week. Therefore, each sequence is considered only once at the time of its first report. Furthermore, to prevent consistently detecting sequences of prevalent lineages such as Alpha, EWS does not consider sequences of the Variants of Concern that were designated as such at the time of evaluation.

Detecting HRVs by standard ML methods

In this section, we compare the capabilities detection of EWS to standard ML methods to highlight both the difficulty of the task and the need for deep learning representations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For all the baselines we consider protein sequences are represented by a vector of **N** binary components. To compute the representation for a protein sequence **S** deposited at time **t**, we calculate the **N** most prevalent mutations in all deposited sequences up to time **t**, inclusive. Each binary component of the representation equals **1** if the mutation is present in **S** and **0** otherwise. We consider **N**=1280 for all the baselines, to permit for a direct comparison with the methods proposed in the paper..

As our method learns from unlabeled data, we first consider the unsupervised learning baseline: Uniform Manifold Approximation and Projection (UMAP). This is an intuitive approach, as it has been successfully applied to analogous problems in biology, as well as it is known to render meaningful insights when applied in life science settings [13]. It is performed each week over the representations of all sequences available up to this week. A metric equivalent to the semantic change is computed as amean L1 distance between the sequence projection and the projections of the Wuhuan and D614G strains in UMAP spaces. The same detection technique as performed by EWS is then used to flag every week a set of 20 variants suspected to be dangerous. We report that only 5 out of 13 variants detected with an average lead time of -43 days, see Table S.6. In comparison, analogous techniques applied in EWS detect 12 out of 13 variants (8 ahead of time), with a mean lead time of 6 days. Both results highlight the need for more involved representations, such as the ones learned by Transformers models, especially in tasks like this one, where significance of the novel findings is difficult to approximate *a priori*.

Second, a supervised learning baseline was considered. Each week, we label all protein sequences that have been registered by **1** if it has been named as HRV anytime before or during the week considered and **0** otherwise. A Generalised Linear Model (GLM) is built over the same 1280-dimensional representations of the sequences. The probability of belonging to the HRV class returned by the GLM is then used to rank sequences. Subsequently, the same detection technique performed by EWS is used to flag every week a set of 20 variants suspected to be dangerous. We report that 12 over 13 variants are eventually detected, with only 4 over 13 being detected before WHO naming, with an average lead time of 0.3 days, see Table S.6. We advise the reader that not only is this approach less performing, it is also less generic than the one proposed herein. This makes it implicitly not fully applicable for pandemics that would attract less worldwide attention and thus less or no labelled data. In addition, this approach cannot be used early in a pandemics as there are no labels available and hallmark mutations are unlikely to be among the most common ones in population early on.

Methods References

1. Berger Rentsch, M. & Zimmer, G. A vesicular stomatitis virus replicon-based bioassay for the rapid and sensitive determination of multi-species type I interferon. *PLoS ONE* **6**, e25858 (2011).

2. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.24.474095; this version posted December 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

4. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

5. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* **118**, (2021).

6. Elnaggar, A. *et al.* ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* (2020).

7. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *BioRxiv* (2021) doi:10.1101/2021.07.09.450648.

8. Vaswani, A. *et al.* Attention is all you need. in *Advances in neural information processing systems* 5998–6008 (2017).

9. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

10. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

12. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 8024–8035 (Curran Associates, Inc., 2019).

13. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).