Learning from pre-pandemic data to forecast viral antibody escape

Authors

Nicole N. Thadani[1,*], Sarah Gurev[1,2,*], Pascal Notin[3,*], Noor Youssef[1], Nathan J. Rollins[1,†], Chris Sander[4,5,6], Yarin Gal[3], Debora S. Marks[1,6,**]

Affiliations

[1]Marks Group, Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA

[3]OATML Group, Department of Computer Science, University of Oxford, Oxford, UK

[4]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

[5]Department of Cell Biology, Harvard Medical School, Boston, MA, USA

[6]Broad Institute of Harvard and MIT, Cambridge, MA, USA

*These authors contributed equally to this work
**Corresponding author: debbie@hms.harvard.edu

Present addresses:
†Seismic Therapeutic, Watertown, MA, USA

Abstract

From early detection of variants of concern to vaccine and therapeutic design, pandemic preparedness depends on identifying viral mutations that escape the response of the host immune system. While experimental scans are useful for quantifying escape potential, they remain laborious and impractical for exploring the combinatorial space of mutations. Here we introduce a biologically grounded model to quantify the viral escape potential of mutations at scale. Our method - EVEscape - brings together fitness predictions from evolutionary models, structure-based features that assess antibody binding potential, and distances between mutated and wild-type residues. Unlike other models that predict variants of concern based on newly observed variants, EVEscape has no reliance on recent community prevalence, and is applicable before surveillance sequencing or experimental scans are broadly available. We validate EVEscape predictions against experimental data on H1N1, HIV and SARS-CoV-2, including data on immune escape. For SARS-CoV-2, we show that EVEscape anticipates mutation frequency, strain prevalence, and escape mutations. Drawing from GISAID, we provide continually updated escape predictions for all current strains of SARS-CoV-2.

Introduction

Viral diseases are characterized by the interplay between immune detection and evasion, leading to rapid evolution and changes in virulence. Viral escape mutations influence reinfection rates and the duration of vaccine-induced immunity, shaping population prevalence over time. To develop optimal vaccines and therapeutics we need to anticipate variants that avoid immune detection with sufficient lead time. New experimental technologies that can test thousands of individual mutations simultaneously (deep mutational scanning (DMS) experiments) can identify mutations that escape existing patient sera or antibodies[1–13]. However, current experimental technologies are limited in a number of ways: (i) They are restricted to testing a miniscule proportion of the 20^L possible sequence variants and combinations of mutations are often non-additive and epistatic[14]. Emerging "Variants of Concern" (VOCs) often contain multiple newly seen mutations whose combined effect is difficult to extrapolate from DMS experiments. (ii) Mutational scanning experiments for a novel pathogen are delayed by challenges in developing high-throughput protein assays and in obtaining appropriate sera or antibody samples – to date, such data is only available for the SARS-CoV-2 receptor binding domain (RBD). (iii) High throughput experimental assays are not necessarily faithful to factors influencing *in vivo* infectivity and immune escape.

The limitations of experimental methods motivate the development of computational approaches. An ideal computational model would be able to assess escape likelihood for as-yet-unseen variation, would be interpretable, and would make predictions with sufficient time for vaccine development. Unfortunately, recent computational methods for forecasting growth of lineages depend critically on real-time pandemic sequencing[15–17]. Therefore, these models can predict reasonably only up to 4 months ahead and cannot assess unseen variants, making them impractical for vaccine development. In contrast, other computational methods training on sequences from evolutionarily distant species have shown surprising success for predicting variant fitness[18–20], including the impact of mutations on human health and on viral replication. These latter methods and related generative probabilistic models of sequences have shown some concordance with mutations in current VOCs[21] and with variants that affect antibody binding[22]. However, broad viral evolution may contain minimal information about host-specific immunity, and these models do not leverage other information crucial for predicting antibody escape.

In this work, we introduce EVEscape, a scalable, modular approach that does not rely on recent pandemic sequencing for predicting viral antibody escape, so it can be used at the start of the pandemic and for continuous evaluation of the risk posed by emerging variants. We combine models of evolutionary data, structural features, and residue dissimilarity properties in an interpretable, biologically grounded framework that captures epistatic effects and is therefore extensible to combinations of variants. EVEscape outperforms current methods and is generalizable across different viruses—in this work we focus on SARS-CoV-2, HIV and H1N1. Our model's significant warning time for concerning mutations could allow for the development of more effective vaccines, antibody therapeutics and diagnostics, as well as help guide public

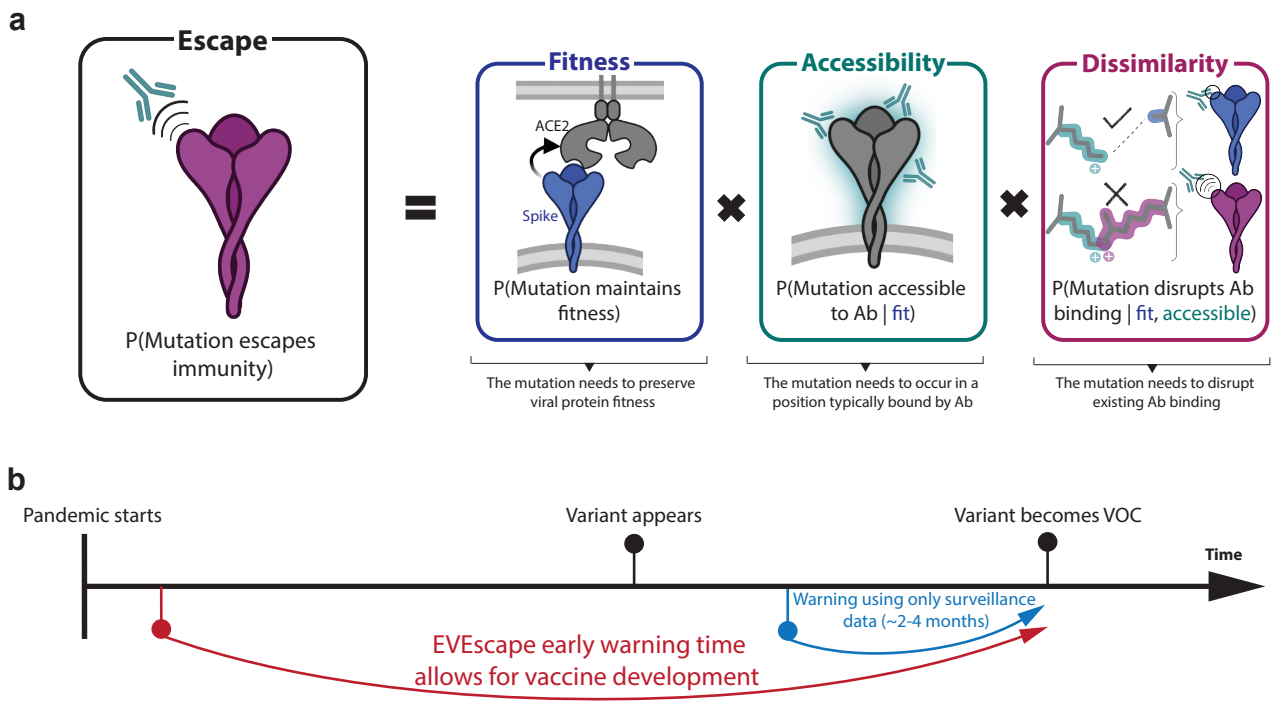

Figure 1: EVEscape assesses the likelihood of a mutation to escape immune response based on the probabilities of a given mutation to maintain viral fitness, to occur in an antibody epitope, and to disrupt antibody binding, using information available early in a pandemic.

health decisions and preparedness efforts with a potentially large impact on the human and economic burden of a pandemic (Figure 1B).

Results

Modeling approach

Viral protein variants that escape humoral immunity must disrupt antibody binding (often by mutating residues in epitopes), while retaining protein expression and folding, host receptor binding, and other properties necessary for viral infection and transmission[9]. Our approach—EVEscape—predicts escape from data sources available pre-pandemic: sequence likelihood predictions from broader viral evolution, antibody accessibility information from protein structures, and changes in binding interaction propensity from residue chemical properties. More specifically, we express the probability of a mutation to induce immune escape as the product of three probabilities: likelihood to maintain fitness ('fitness' term), likelihood of the mutation to target an antibody epitope ('accessibility' term) and likelihood of the mutation to disrupt antibody binding ('dissimilarity' term) (Figure 1A).

The fitness term is computed from a deep unsupervised generative model of mutation effects. The accessibility term of a protein site is modeled via the negative of the weighted contact number computed from viral protein structures (or if unavailable, predicted structures from homologs) (Table S3), while the dissimilarity term depends on the difference in charge and hydrophobicity between the mutant and wildtype residues. Each term is standardized and then

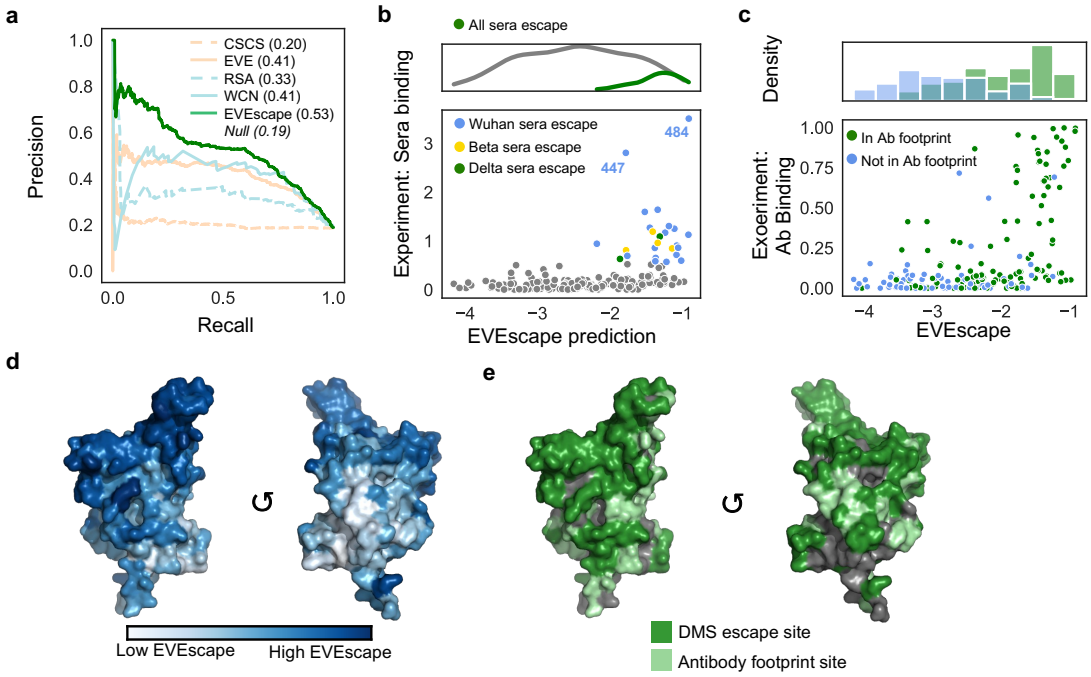

Figure 2: EVEscape captures antibody footprints and escape potential. a) Precision-Recall of RBD DMS escape mutations (AUPRC reported compared to "null" model). b) EVEscape captures mutations that effect recognition by convalescent sera from patients infected with different VOCs. c) RBD site-averaged EVEscape predictions plotted against site-averaged Bloom DMS escape, with hue indicating known antibody footprints. d) RBD site-averaged EVEscape predictions (PDB: 7BNN). e) RBD sites of DMS escape mutants and of known antibody footprints (PDB: 7BNN).

fed into a temperature-scaled logistic function (Methods, Data S3). The EVEscape score is then obtained as the log transform of the product of the three terms.

EVEscape predicts experimental antibody evasion

To evaluate EVEscape's performance and generalizability across viruses we compare to experimental deep mutational scans measuring the ability of mutations to permit viral cell entry in the presence of antibodies (for influenza H1 and HIV envelope proteins)[10,11] or to inhibit sera and antibody binding (for SARS-CoV-2 RBD)[1–9,12,13] (Table S4, Data S4). We focus on the SARS-CoV-2 RBD, as RBD is the primary target for viral neutralization[23] and a large number of antibodies and sera have been screened. We examine performance on Flu and HIV as a secondary analysis to confirm generalizability, as fewer antibodies have been tested and the distribution of these antibodies does not reflect known immunodominant domains. We binarize the experimental measurement by taking the maximum value across antibodies and sera tested, then applying a threshold to define mutations as 'escape' or 'not escape' (Figure S1, Methods). We find that our main conclusions are robust to the threshold choice and that mutations designated as escape by our threshold are almost all within 5Å of the antibody they escape (Figure S1). A key feature of an escape mutant predictor is the quality of its positive 'escape' predictions, as in practice, the positive predictive value will influence costly experimental screening efforts and selection of a limited number of variants for vaccine incorporation. To reflect this, we focus on the area under the precision-recall curve (AUPRC) as a performance

metric (reported relative to the AUPRC of a "null" model), although other measures of overall statistical performance (e.g., AUROC) are provided in supplementary information.

EVEscape outperforms deep unsupervised sequence models and metrics of surface accessibility in the quality of its top escape predictions (Figure 2A, Figure S2) across datasets with diverse experimental methods (Figure S3, Table S4). For instance, when focusing on the top decile of predictions for all single substitution mutants for RBD, EVEscape captures 2.5x more actual escape mutants (31% of total measured), and the overall AUPRC of EVEscape (0.53) is 2.7 times higher than the prior state of art (0.2)[22]. EVEscape is especially strong at identifying escape mutants from polyclonal patient sera (Figure S3); in fact, nearly 50 percent of sera escape sites from patients infected with the original Wuhan strain or the Beta or Delta VOCs are in the top 10% of EVEscape predictions (Figure 2B). These mutants are of particular interest since they show a significant degree of escape from the unique composition of antibodies produced by convalescent patients, and so are crucial to considerations of reinfection and vaccine design. Mutations at sites E484 and G447 are particularly notable for escaping sera binding (E484, mutated in several VOCs, is the top EVEscape predicted site). While EVEscape frequently predicts escape mutations that are not observed in the escape DMS, this is likely due to sparse sampling of the antibodies that bind these viral proteins in DMS experiments. Indeed, EVEscape performance has improved as more DMS data has become available (Figure S3) and the majority of EVEscape positive predictions that are not observed escape mutants in DMS are in known antibody epitopes (Figure 2C-E, Figure S4).

Model components provide complementary information for escape

We next examine the roles each EVEscape component plays in identifying potential escape mutations. We find that sequence model predictions and surface accessibility metrics play the strongest role in performance and are both informative of antibody binding footprints as well as which sites have the highest number of escape mutations (Figure 3A/C, Figure S5), while dissimilarity is useful for distinguishing escape mutations within sites (Figure 3D).

Fitness: Sequence likelihood predictions from generative sequence models reflect mutation fitness effects as well as site mutability, which is often an indicator of a protein region targeted by neutralizing antibodies[24] like the SARS-CoV-2 RBD. We selected EVE[18] to predict the impact of viral protein mutation effects on function due to its generally superior performance on a range of viral function DMS experiments as compared to other sequence-based models[19,21,22] (Figure S6-S7, Note S1, Data S2, Table S2). Despite limited known natural sequence diversity for the coronavirus family, EVE predictions trained on pre-pandemic coronavirus sequences for SARS-CoV-2 RBD (Table S1, Data S1) are correlated with both expression and binding to the ACE2 host receptor (Figure 3B, Figure S8). Predictions improve with the incorporation of pandemic sequence data (Figure S6).

Antibody accessibility: Surface accessibility plays a key role in identifying where antibodies are most likely to contact a protein. While relative solvent accessibility (RSA) and weighted contact number (WCN) both reflect features of accessibility, we selected WCN as this metric also captures protrusion from the core structure that corresponds with where antibodies are

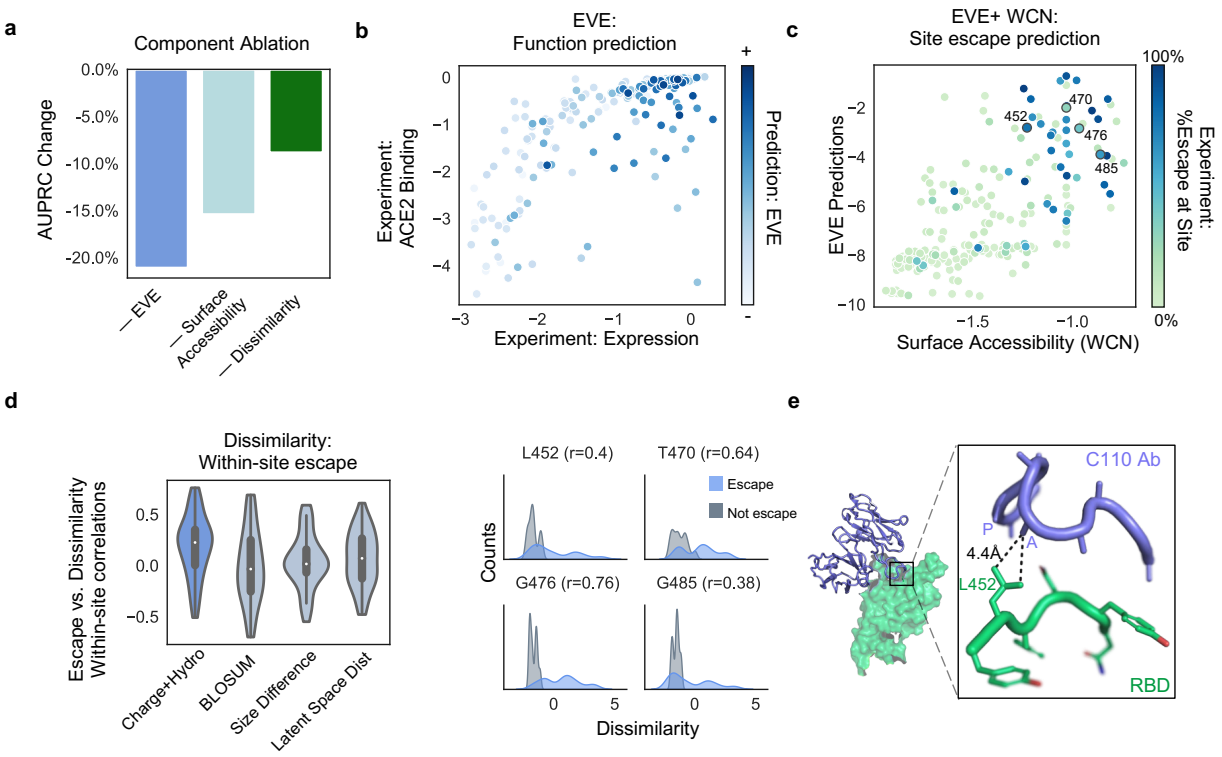

Figure 3: Surface accessibility metrics and mutation effect models provide complementary information for predicting antibody epitopes, while residue dissimilarity reflects within-site escape likelihood. a) All features of EVEscape contribute to performance in predicting RBD escape mutants. b) EVE prediction captures a combination of SARS2 RBD yeast expression and ACE2 binding - features both necessary for successful immune escape (EVE spearman with expression = 0.45, EVE spearman with ACE2 binding = 0.27). c) Sites with either high accessibility or high EVE fitness predictions have a greater percent of escape mutants. d) Chemical property dissimilarity is more indicative of mutant escape likelihood within a site than substitution matrices or other distance metrics. Summary of within-site escape point biserial correlations with charge-hydrophobicity dissimilarity for sites with some escape (3-17 escape mutations) (left), plots illustrating charge-hydrophobicity performance in key sites (right). e) The L452 RBD site is an example of decrease in hydrophobicity displacing the proximal alanine and proline (side chain not resolved) in the RBD C110 antibody interaction. (PDB: 7K8V)

known to bind proteins[25–27] (Figure S9). To capture all protein conformations potentially accessible to antibodies, we consider the maximum accessibility across structures (Table S3).

Residue dissimilarity: Within sites targeted by antibodies, mutations must disrupt antibody binding while retaining a certain minimum fitness to facilitate escape. EVE scores select for the most likely mutation under learned constraints from viral evolution, but may not capture antibody evasion. Thus, we incorporated a term summarizing residue dissimilarity in properties known to impact protein-protein interactions (hydrophobicity and charge[28,29]) to identify the most likely escape mutations. We find that within sites that have escape mutants, a high charge-hydrophobicity dissimilarity is predictive of escape, and that this simple metric correlates with within-site escape more than individual chemical properties, substitution-matrix derived distance, or distance in the latent space of the EVE model (Figure 3D, Figure S10A). For instance, L452K/R/E/D mutations that decrease hydrophobicity have high escape potential (Figure 3C), with the residue dissimilarity metric boosting the EVEscape scores from 80th to 90th percentile compared to L452I/A/G/V. These predictions are in concordance with observed C110 antibody binding in the DMS experiment (Figure 3E, Data S4). In general, a low charge-

hydrophobicity dissimilarity does not preclude escape (Figure 3D), plausibly because any mutation to a residue with a large number of antibody contacts is likely to disrupt antibody binding. Correlations between within-site maximum escape and charge-hydrophobicity distance are higher for sites targeted by multiple screened antibodies, suggesting that data capturing a larger diversity of antibody paratopes improves dissimilarity performance (Figure S10).

Another potential escape strategy is to alter the glycans shielding the protein surface. We experimented with maximizing the dissimilarity factor if a mutation can remove a known surface glycan. While addition of glycosylation is also important for antibody escape[30], we focused on loss of glycosylation because mutations removing glycan sites are readily identified by their alteration of surface N-linked glycosylation motifs. For the HIV envelope protein, this improves prediction of escape mutants—unsurprisingly, as surface glycosylation changes are a common escape strategy for HIV variants[31] (Figure S11). While DMS experiments do not reflect escape impacts of glycosylation loss for the other viral proteins, this factor is an important consideration for real-world escape[32–34].

As a real-world example of the interplay of the three model components, more than half of mutations to the highly accessible and mutable E484 site are in the top 2% of EVEscape predictions, nicely confirmed by the appearance of E484A/K (with high charge dissimilarity scores) in numerous VOCs including Omicron. E484 is involved in a salt bridge with R96 and R50 of LY-CoV555 (therapeutic antibody bamlanivimab), which lost FDA authorization because Omicron (E484A) escapes binding[35].

EVEscape predicts broad, non-neutralizing antibodies as least escapable

An ideal escape predictor will reflect the likelihood of a mutant to escape polyclonal serum composed of antibodies that are most prevalent in the convalescent and vaccinated population. To evaluate EVEscape's coverage of diverse antibody epitopes, we examined predictions of escape mutants across the four structurally defined classes of RBD-targeting antibodies identified by Barnes et. al[36]. EVEscape top predictions incorporate escape mutants across diverse epitope regions, with heavy coverage of antibody classes 1-3 and sparse coverage of class 4 (Figure 4A-B). These top predictions include known Class 1 and 2 immunodominant sites E484, K417, and L452. Class 2 antibodies have been identified as immunodominant in the sera of convalescent patients, followed by class 3 and 1[4]. Class 4 antibodies bind to a cryptic epitope only revealed when 2 of the 3 RBDs on the Spike protein trimer are in the 'up' conformation, and are capable of binding SARS-CoV-1 as well as SARS-CoV-2 but are less potent neutralizers[37].

Generally, EVEscape scores are lower for escape mutations from antibodies that bind to a diverse pool of sarbecoviruses ("broad" antibodies) (Figure 4C) – these mutants are also less likely to be seen in the pandemic[9]. This is likely due to the typical high conservation of broad antibody epitopes in viral evolution, as well as the propensity of broad antibodies to bind cryptic epitopes[38] (Figure S12). Broad antibodies are of great interest due to their potential as pan-coronavirus therapeutics but tend to be less neutralizing than the most potent antibodies that bind near ACE2-contacting RBD surfaces. Within broad antibodies, EVEscape better captures

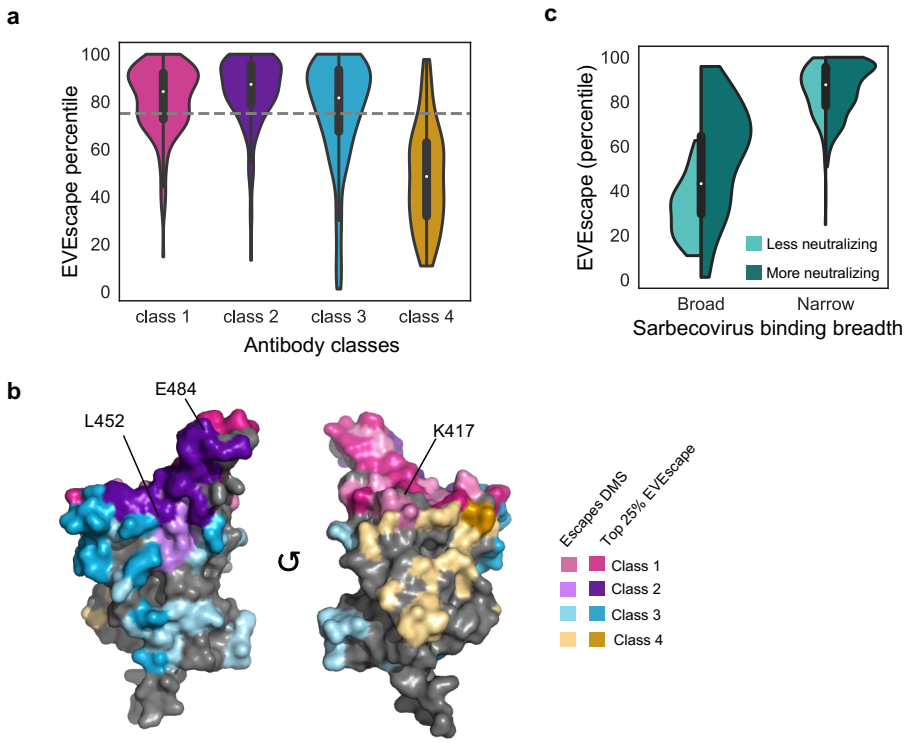

Figure 4: EVEscape predictions identify escape-resistant classes of antibodies. a-b) EVEscape scores of observed escape mutations cover diverse epitope regions across antibody classes including known immunodominant sites (E484, K417, L452) (PDB: 7BNN). c) EVEscape scores observed escape mutants from narrow antibodies and broad neutralizing antibodies higher than those from broad, non-neutralizing antibodies. Miniature boxplots within violin plots indicate median and interquartile range.

escape from neutralizing antibodies, due to its accessibility component that captures protrusion from the core structure (Figure 4C, Figure S12). Broad antibodies are relatively rare in convalescent sera, so evasion of these antibodies is likely not essential for variant spread (though this may change as broad antibody therapeutics are more widely adopted)[39,40]. These results further underline the importance of broad antibodies as therapeutics that may retain utility throughout the pandemic, as opposed to neutralizing antibodies that may be encumbered by significant immune escape.

EVEscape anticipates mutations in pandemic and future strains

We then investigated whether EVEscape's predictions (trained entirely on data available pre-pandemic) of likely escape mutations in the Spike protein correspond with trends observed in SARS-CoV-2 evolution through the course of the pandemic. Over 11 million SARS-CoV-2 sequences have been deposited in the GISAID (Global Initiative on Sharing All Influenza Data)[41] database, including more than 6500 mutations to Spike that cover more than 99% of sites. We focus on mutations with a one nucleotide substitution distance from the Wuhan sequence, as these mutations are more likely to occur and make up 93% of the residue mutations observed in GISAID. The frequency of mutations observed in GISAID corresponds with EVEscape scores—mutations in the top quartile of EVEscape are almost four times as likely to be observed in the pandemic than the bottom quartile, and mutations with high EVEscape scores are increasing in

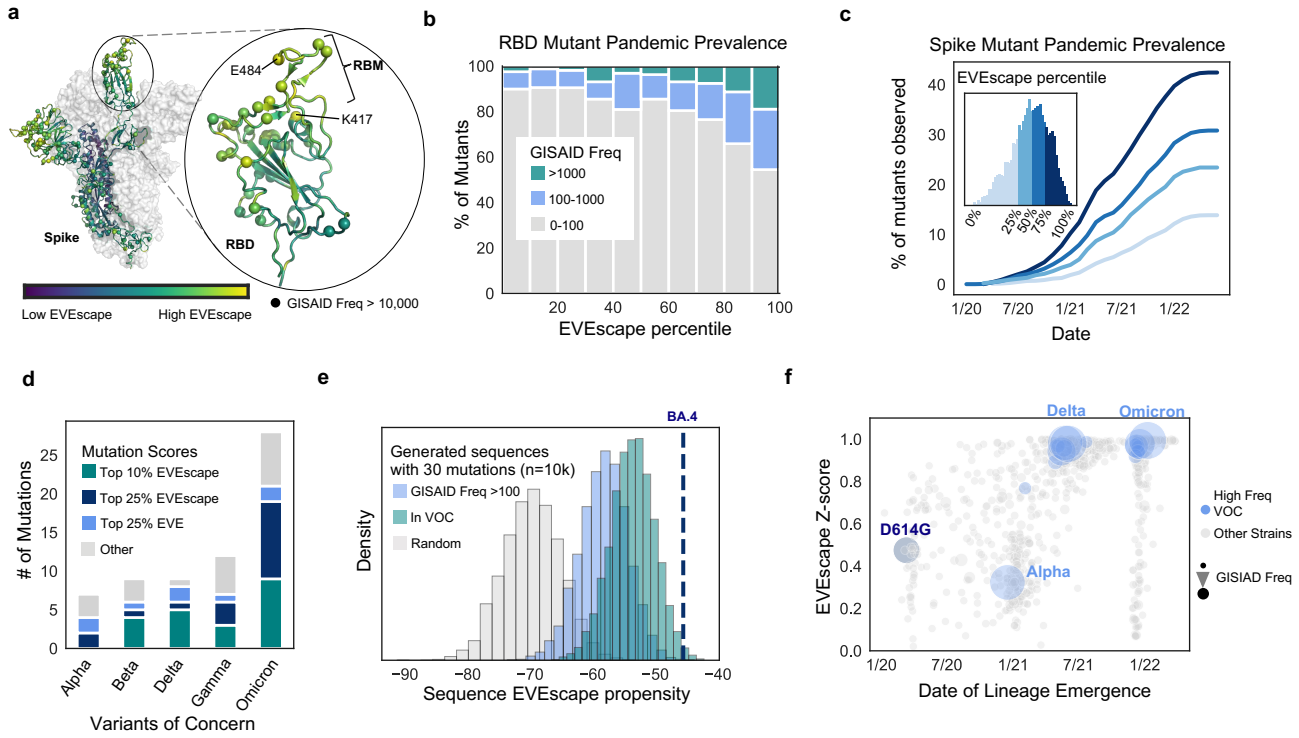

Figure 5: EVEscape anticipates single mutations and strains observed in the SARS-CoV-2 pandemic. a) Site-averaged EVEscape scores on Spike structure (PDB: 7BNN) depict regions of high EVEscape scores (RBD, particularly the ACE2 binding region (RBM), and NTD). Spheres indicate sites with GISAID mutations observed more than 10K times. b) EVEscape mutation scores correspond to observed GISAID mutations in the RBD. c) Percentage of mutations in each EVEscape quartile seen >100 times in GISAID over time d) The majority of mutations in VOC strains are in the top quartile of Spike EVEscape predictions. e) BA.4 has a high EVEscape score compared to random samples of mutations at the same mutation depth. f) EVEscape z-scores increase throughout the pandemic, particularly for VOCs. Z-scores are relative to random combinations at the same mutation depth of single mutations seen over 1000 times in GISAID. Note: EVEscape percentiles have been adjusted to consider only mutations that are a nucleotide distance of one away from Wuhan.

prevalence faster than those with lower scores (Figure 5B-C). Moreover, EVEscape's predictions of the relative escapability of regions throughout Spike are consistent with the immunodominance of RBD and NTD antibodies in convalescent and vaccinated sera[42,43], as escape mutation predictions are significantly enriched in the RBD (particularly the ACE2 contacting loop, 40% of which are in the predicted top 10% of Spike mutations) and the NTD (Figure 5A, Figure S13).

We examined performance specifically on VOC mutations, which have been extensively characterized for their fitness and immune evasion. Mutations found in the VOCs are predominantly in the top 25% of Spike EVEscape predictions, likely reflecting their neutral or positive impact on general viral fitness as well as immune evasion (Figure 5D). Of the remaining VOC mutations, i.e., S375F and T376A at the bottom of EVEscape's predictions (Figure S14A), many are actually known to decrease fitness, including by impairing infectivity, and are often reverted[44]. EVEscape's success at predicting VOC mutations is indicative of its capacity to learn a broad range of viral fitness and immune escape constraints. For instance, many VOCs exhibit higher ACE2 affinities, which itself can be a mechanism of immune escape through binding

competition[45]. EVEscape predicts many VOC mutations with enhanced ACE2 binding (i.e., Q493R and L452R that have reduced antibody binding, and G339D with neutral impact on antibody binding) in the top 10% (Figure S14A). EVEscape is in general more predictive of frequent VOCs than EVE alone (Figure S14B). The few VOCs mutations (i.e., A222V and T547K) with significant EVE—but not EVEscape—scores are known both for functional improvements such as monomer packing and RBD opening and for not impacting escape[46,47]. On the other hand, mutations with the highest EVEscape but low EVE scores include R190S and R408S, which are in hydrophobic pockets that likely facilitate significant immune escape[48].

For circulating viral strains, escape requires all mutations in the sequence to preserve viral function while facilitating immune evasion. To create strain-level escape propensity predictions, we aggregated EVEscape predictions across all individual Spike mutations in a strain, then compared the aggregate EVEscape score to that of other sequences at the same mutational distance from the original Wuhan strain. We find that VOCs compare favorably to random sequences at the same mutational depth, and in particular the Delta and Omicron strains known for their fitness and immune evasion are at the top end of score distributions (Figure 5E-F, Figure S14C, Data S5). VOCs like Omicron BA.4 compare favorably against not only sequences of random mutations at the same depth, but also against sequences composed only of mutations already known to be favorable — those seen more than 100 times in GISAID, and even more strikingly, against combinations of mutations sampled from other VOCs. These results illustrate EVEscape's promise as an early-detection tool for identifying the most concerning variants in the large pool of available pandemic sequencing data. Future results will be available on our website (evescape.org) that enables real-time variant escape tracking through EVEscape rankings of newly occurring variants from GISAID and interactive visualization of likely future mutations to our top predicted escape variants.

Discussion

Predicting viral evolution under shifting immune constraints is crucial to pandemic preparedness. Experimental methods craft assays to capture viral functional properties while computational sequence models learn constraints from the evolutionary record – approaches we view as orthogonal but synergistic towards this end. While experimental methods can be tailored to key protein functions, they are time-consuming and thus limited in scope, and often miss aspects of natural virus function (particularly high-throughput methods). Computational sequence models learn a full picture of constraints from the course of natural evolution, but are subject to limitations of their training data. Pandemic surveillance data is restricted to making predictions for observed mutations, leaving blind spots regarding future mutational avenues as well as bias due to sampling and epidemiological effects[49]. Historical viral evolution data is immediately available at pandemic onset, and due to greater sequence diversity, is widely extensible to potential mutations and their combinations. However, novel pandemic constraints (such as immunity) are unlikely to be captured. To achieve early prediction, EVEscape combines historical viral evolution data with a biologically-informed strategy using only protein structure to anticipate immune selection and escape. Our model identifies key escape mutations and strains in high-throughput experiments of antibody binding and throughout the SARS-CoV-2 pandemic.

Later in a pandemic, EVEscape is flexible to incorporating antibody binding footprints, experimental screens and pandemic surveillance data to match current knowledge on SARS-CoV-2 specific immune targeting and mutation tolerance. EVEscape can also enhance this understanding by proposing escape variant libraries for experimental investigation, as well as suggesting viral proteins and regions with significantly high or low potential for escape to inform future therapeutics.

EVEscape is a modular, scalable, and interpretable probabilistic framework that may be used to identify observed pandemic strains most likely to thrive in conditions of widespread pre-existing immunity, as well as to propose the most concerning new mutations on any circulating strain. To this end, we provide predictions for all single mutation variants of SARS-CoV-2 Spike as well as aggregate strain-level predictions for all strains that have been observed 1000 times or more in GISAID and will continue to update with new strains. As the framework is generalizable across viruses, EVEscape can be used from the start for future pandemics as well as to better understand and prepare for emerging pathogens.

Data and Code Availability

All data is provided in supplementary materials. Code is available at https://github.com/OATML-Markslab/EVEscape and future updates will be available at evescape.org.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Debora Marks.

Supplementary Information: This file contains supplementary note S1, supplementary figures S1-S14 and supplementary tables S1-S4.

Supplementary Data 1: Alignments used for EVE models for SARS-CoV-2, HIV, and Flu.

Supplementary Data 2: EVE and EVmutation scores for DMS fitness experiments.

Supplementary Data 3: EVEscape performance for selection of factor-specific temperature scaling.

Supplementary Data 4: EVEscape scores and escape and fitness data for all Spike, HIV, and Flu mutations.

Supplementary Data 5: EVEscape sequence propensity for all SARS-CoV-2 PANGO lineages.

Supplementary Data 6: Acknowledgements for all GISAID sequences.

Methods

Data acquisition:

Multiple sequence alignments:

For each viral protein, we construct multiple sequence alignments performing 5 iterations of the profile-HMM based homology search tool jackhmmer[50] against the UniRef100 database. As previously reported for EVE, DeepSequence, and EVcouplings, we generally keep sequences that align to at least 50% of the target sequence and columns with at least 70% coverage, except in the case of SARS-CoV-2 Spike where we use lower column coverage as needed (30-70%) to maximally cover experimental positions and significant pandemic sites[18–20]. For our pre-pandemic (pre-2020) alignment used as the primary model throughout this paper, we remove pandemic sequences using the "date of creation" variable from UniRef. We optimized search depth to maximize sequence coverage and the effective number of sequences (Neff) included after re-weighting similar protein sequences in the alignment within a Hamming distance cutoff of 0.01. To select sequence depth, we prioritized alignments with coverage >0.7L and Neff/L>1, or if this was not attainable, relaxed the requirements for Neff/L.

Structure and structure-based calculations

Selecting structures for surface accessibility calculations:

For each viral surface protein, we selected crystal structures representing known structural states available to B-cell and antibody interactions (extracellular conformations) (Table S3). All heteroatoms and protein chains not part of the trimeric viral surface protein were removed.

Antibody footprints:

To identify known antibody footprints of viral surface proteins in the RCSB PDB, we queried the database with the protein name and the word "antibody" and required that the source organism contain both "Homo sapiens" and the given virus name. Then for each structure we identified antibody and viral protein polymer entities and computed the antibody footprint as any residue with any atom within 3.5 angstroms of the antibody. Finally, we mapped footprints to the target viral protein sequence by using SIFTS to renumber all hits according to a UniProt ID, then used a MUSCLE multiple sequence alignment of the different UniProt sequences to map those hits to the target viral protein sequence.

Deep mutational scans

We benchmark our models on a series of viral protein deep mutational scans[1–13,51–58] (Table S2, Table S4). For each viral mutational scan, we select the variable or variables of protein fitness or antibody escape treated as primary in the publications. For mutants where the result is provided as residue frequencies observed at a given site (such as results expressed as preferences and processed by dms_tools2), we normalize the data at each site by dividing by the value of the wild-type residue. For the HIV analysis, we exclude antibody VRC34.01 due to its large spread of escape mutation distal to the epitope[59]. For SARS-CoV-2 RBD, we use only antibodies/sera escape data from the Wuhan sequence for our primary results. We also utilize data provided about the antibodies tested for the SARS-CoV-2 escape DMS studies, including the class of each antibody as well as the SARS-CoV-2 neutralization potency and sarbecovirus binding breadth[9]. We use the RBD dimeric ACE2 binding and expression DMS data for analysis[54].

Pandemic sequencing data

We downloaded data on Spike variants and their deposit dates in the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV project database (www.gisaid.org)[41] on 5/23/22. We further processed this data to get counts of combinations of mutations, the date of emergence, and PANGO lineage, as well as get the month of emergence for each single mutation in Spike. We also downloaded consensus mutations for each PANGO lineage from Covid-19 CG[60] on 7/1/22.

Modeling approach:

Overarching framework

We express the probability of a single amino acid substitution to lead to immune escape as the product of three conditional probabilities (Figure 1A):

$$\begin{aligned}\boldsymbol{P}(Mutation\ escapes\ immunity) \\ &= \boldsymbol{P}(Mutation\ maintains\ fitness) * \boldsymbol{P}(Mutation\ accessible\ to\ antibody \mid fit) \\ &* \boldsymbol{P}(Mutation\ disrupts\ antibody\ binding \mid fit, accessible)\end{aligned}$$

The EVEscape index estimates the log likelihood of escape as per the above equation. The fitness factor is obtained via a deep generative model for fitness prediction, while the accessibility and dissimilarity factors are features derived respectively from the known 3D structures for the viral protein and chemical characteristics of the amino acids involved in the mutation compared to the wild-type (see below for details).

Once selected, each factor is standardized and fed into a temperature scaled logistic function:

$$\begin{aligned}\boldsymbol{P}(Mutation\ escapes\ immunity) \\ &= logistic\left(\frac{1}{T_{fitness}} * standardize\left(F_{fitness}\right)\right) \\ &* logistic\left(\frac{1}{T_{accessibility}} * standardize\left(F_{accessibility}\right)\right) \\ &* logistic\left(\frac{1}{T_{dissimilarity}} * standardize\left(F_{dissimilarity}\right)\right)\end{aligned}$$

where the standardize(.) operator corresponds to standard scaling. We then take the log transform of the product of the 3 terms to obtain the final EVEscape scores.

Factor-specific temperature scaling helps recalibrate probability estimates for each term. We provide our hyperparameter grid search of these temperature hyperparameters across viruses in Data S3, examining versions of the model where we either include or do not include glycosylation in the dissimilarity term. We find that the fitness and accessibility components are already properly calibrated ($T_{fitness}$ = $T_{accessibility}$ = 1.0), while the dissimilarity component benefits from being slightly rescaled ($T_{dissimilarity}$ = 2.0).

Fitness metric

Observed viral protein sequences reflect evolution under selection constraints for functional and infectious viruses. Generative sequence models express the probability that a sequence x would be generated by this process as $p(x|\theta)$, where the parameters θ capture the constraints describing functional variants. A generative model trained on observed viral protein variants can then be used to estimate the relative plausibility of a given mutant sequence as compared to wild-type by using the log ratio of sequence likelihoods as a heuristic:

$$log\frac{p(x^{mutant}|\theta)}{p(x^{wildtype}|\theta)}$$

EVE (Evolutionary model of Variant Effects) is a Bayesian variational autoencoder (VAE)[61], capable of capturing complex higher-order interactions across sequence positions. The fitness of a given protein sequence is measured via the log likelihood ratio of the mutated sequence *x* over that of the reference wild-type sequence *w*. Since an exact computation of the log likelihood of a sequence is intractable, we approximate it with the Evidence Lower Bound (ELBO) loss used to optimize the VAE:

$$E_{EVE}(x) = -log\frac{p(x|\theta)}{p(w|\theta)} \sim ELBO(w) - ELBO(x)$$

The ELBO term itself is estimated via Monte Carlo sampling, using 20k samples from the approximate posterior distribution. These approximations have been shown to provide strong results in practice[18]. Final results are obtained by ensembling scores from 5 independently trained EVE models with different random seeds.

Accessibility metric

To predict likely antibody binding sites, we used weighted contact number (WCN) and compared performance to relative surface accessibility (RSA).

Calculating weighted contact number

We computed weighted contact numbers[64] for each residue from structure as follows:

$$WCN_i = \sum_{j \neq i} \frac{1}{{r_{ij}}^2}$$

where r_{ij} is the distance between the geometric centers of the residue i and residue j side chains. We impute missing values in WCN due to gaps in the protein structure using the mean of WCN values of the residues preceding and following the gap.

RSA:

To compute RSA, we first computed accessible surface area based on hypothetical exposure to solvent water molecules using DSSP[62]. To calculate relative accessible surface area (RSA), we divided accessible surface area by the residue maximum accessibilities determined in Sander et al[63]. We impute missing values in RSA due to gaps in the protein structure by using the mean of RSA values of the residues preceding and following the gap (counting residues adjacent to the gap with RSA values>1 as part of the gap).

Aggregating across structures:

When computing antibody-binding likelihood metrics across different structural conformations we used the maximum accessibility (or minimum weighted contact numbers).

Dissimilarity metric

To predict the likelihood of a given mutation displacing an antibody interaction, we used a charge-hydrophobicity based measure of functional dissimilarity between the wild-type residue and the mutation residue. We compare our metric to individual chemical properties, substitution matrices, and the distance in the latent space of a VAE. We also experiment with incorporating glycosylation in our dissimilarity metric.

Charge-hydrophobicity

To compute a combined charge-hydrophobicity dissimilarity index, we standard-scaled the charge and hydrophobicity[67] differences and then took the sum of the scaled differences.

Chemical properties

We compared our metric to differences in residue size (side-chain mass), hydrophobicity, and charge.

Substitution Matrices

We compared our metric to the BLOSUM62[66] matrix after dropping the null transition diagonal values.

Latent space distances

We also compared our metric to a metric of mutation distance learned by the EVE variational autoencoder. We calculated the L1 distance between the encoded representations of the wild-type viral protein sequence and a given single-mutation sequence in the latent space of the model, inspired by a similar approach first introduced by Hie et al.[22]

Glycosylation

We developed a version of our model considering glycosylation loss as a contributor to dissimilarity. In this version, we maximize the charge-hydrophobicity dissimilarity term if a mutation is likely to result in loss of a surface N-glycan site. We identified surface N-glycan sites as NxS/T sequons (where x is any amino acid except proline) with the N residue having an RSA>0.2. A mutation is likely to result in loss of glycosylation if the N or S/T is lost.

Imputing missing data

We impute missing values of features in EVEscape using the mean value of the feature across the target protein.

Strain-level EVEscape propensity predictions

We aggregate across combinations of mutations by summing the EVEscape scores for each mutation.

Evaluation:

Comparison to functional assays

We compared model predictions to continuous experimental metrics of viral function using spearman's rank correlation coefficient as our main evaluation metric, as previously described[19,20].

Comparison to escape DMS

Data processing

As escape data is noisy at levels of low escape and a relatively low fraction of mutants exhibit escape, we chose to treat the escape outcome variable as binary. We selected a threshold for escape by fitting a gamma distribution to the data (combined across all screened antibodies and sera) and selecting the threshold corresponding to a 5% false discovery rate[11]. As the number of antibodies tested for RBD is much higher than for Flu and HIV, we bootstrapped the RBD data selecting 8 antibodies 1000 times and fitting a gamma distribution to these samples, then selected the average 5% false discovery rate threshold. As these thresholds are subject to our choice of a false discovery rate, we also plot performance for a range of thresholds (Figure S4). We identified a mutant as "escape" if its maximum escape value across any antibody tested exceeded the threshold (so a mutation for RBD is "escape" if it exceeds the threshold for any antibodies/sera in the Bloom or the Xie datasets (Data S4)). We use thresholds of 0.57 for Bloom RBD, 0.9 for Xie RBD, 0.054 for Flu, and 0.138 for HIV. Note that the downloaded RBD escape datasets were already filtered using thresholds on expression and ACE2 binding of -1 and -2.35, respectively[68].

To define a site-wise escape value, we averaged across the maximum escape values for each mutant at the site. For the sera RBD DMS data, we define a mutation as escaping Wuhan sera if it passes the Bloom RBD escape threshold for any Wuhan sera. Otherwise, if a mutation does not escape any Wuhan sera, but does escape Beta or Delta sera, it is labeled with the corresponding variant sera. For the antibody RBD DMS data, we define the antibody class of each mutation/site by determining the maximum number of antibodies for a given class that escape that mutation/site (Data S4).

Metrics

To quantify model performance in classifying escape mutants, we computed two metrics. We consider area under the receiver operating curve (AUROC) and area under the precision-recall curve (AUPRC). AUROC summarizes the tradeoff between true positives and false positives over a range of thresholds on the continuous model prediction score but is overly permissive in cases of imbalanced datasets–-although still suitable for assessing relative performance. The AUPRC metric summarizes the tradeoff between capturing all escape mutants (recall) and not incorrectly predicting escape mutants (precision). This approach is suitable for evaluating classification of imbalanced datasets but penalizes false positive predictions. In the case of escape predictors, false positive predictions may be due to insufficient sampling of the human antibody repertoire against the virus of interest, so this penalization is potentially too stringent. We normalize AUPRC by the "null" precision model AUPRC, which is equivalent to the fraction of escapes observed in the mutations experimentally screened. Therefore, AUPRC values are not comparable between viral proteins or subsets of DMS datasets with different fractions of escape mutations.

Comparison to known antibody footprints

We also evaluated the model's ability to predict sites of antibody binding, as quantified by looking at antibody footprints in the RCSB PDB within a minimum all-atom distance of 3.5Å.

Comparison to pandemic sequencing data

Data Processing

We evaluate the model against occurrence of single mutations and strains in GISAID. In determining the set of Spike mutations to compare EVEscape scores to GISAID data, we consider only those mutations that are a single RNA nucleotide mutation distance from Wuhan. Our set of pandemic strains are the combinations of mutations that have occurred together greater than 500 times in GISAID. The date of lineage emergence is the 5th percentile of dates for that variant (to avoid issues with outliers from GISAID data entry) and the variants are marked as high frequency VOCs if their mode lineage is a VOC and their count is greater than 50,000, using the processed GISAID data table. We define PANGO lineages for the VOCs by the nonsynonymous Spike consensus mutations for that strain from COVID-19 CG that occur in greater than 10% of strain sequences, ignoring insertions and deletions.

Metrics

To evaluate sequence EVEscape propensities of the strains observed in GISAID, we use a z-score to compare the EVEscape propensity to the scores of 10,000 sequences at the same mutation depth, randomly generated by sampling single mutations seen over 1000 times in GISAID.

Regional Enrichment

We analyze enrichment of regions by the location of the average EVEscape score for the region as compared to a distribution of the average EVEscape score of random regions. For comparison to full Spike, we compare to the scores of 500 random contiguous regions (of the same length as the region of interest) within Spike. For comparison to RBD, we compare to scores of 100 contiguous regions, using the full Spike model.

References

1. Dong, J. *et al.* Genetic and structural basis for SARS-CoV-2 variant neutralization by a two-antibody cocktail. *Nat Microbiol* **6**, 1233–1244 (2021).
2. Greaney, A. J. *et al.* Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44-57.e9 (2021).
3. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).
4. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nature Communications* vol. 12 (2021).
5. Greaney, A. J. *et al.* Antibodies elicited by mRNA-1273 vaccination bind more broadly to the receptor binding domain than do those from SARS-CoV-2 infection. *Sci. Transl. Med.* **13**, (2021).
6. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. *Cell Rep Med* **2**, 100255 (2021).
7. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021).
8. Tortorici, M. A. *et al.* Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature* **597**, 103–108 (2021).
9. Starr, T. N. *et al.* SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature* **597**, 97–102 (2021).
10. Doud, M. B., Lee, J. M. & Bloom, J. D. How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Nat. Commun.* **9**, 1386 (12/2018).
11. Dingens, A. S., Arenz, D., Weight, H., Overbaugh, J. & Bloom, J. D. An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes. *Immunity* **50**, 520-532.e3 (2019).
12. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
13. Greaney, A. J. *et al.* A SARS-CoV-2 variant elicits an antibody response with a shifted immunodominance hierarchy. *PLoS Pathog.* **18**, e1010248 (2022).
14. Frei, L., Metcalfe, S. W., Yermanos, A., Kelton, W. & Reddy, S. T. Predictive profiling of SARS-CoV-2 variants by deep mutational learning. *bioRxiv* (2021).
15. Maher, M. C. *et al.* Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *bioRxiv* (2021) doi:10.1101/2021.06.21.21259286.
16. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

17. Beguir, K. *et al.* Early Computational Detection of Potential High Risk SARS-CoV-2 Variants. *bioRxiv* 2021.12.24.474095 (2021) doi:10.1101/2021.12.24.474095.
18. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
19. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
20. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
21. Rodriguez-Rivas, J., Croce, G., Muscat, M. & Weigt, M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
22. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
23. Wang, Z. *et al.* Analysis of memory B cells identifies conserved neutralizing epitopes on the N-terminal domain of variant SARS-Cov-2 spike proteins. *Immunity* **55**, 998-1012.e8 (2022).
24. Andrews, S. F. *et al.* Immune history profoundly affects broadly protective B cell responses to influenza. *Sci. Transl. Med.* **7**, 316ra192 (2015).
25. Thornton, J. M., Edwards, M. S., Taylor, W. R. & Barlow, D. J. Location of “continuous” antigenic determinants in the protruding regions of proteins. *EMBO J.* **5**, 409–413 (1986).
26. Novotný, J. *et al.* Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. U. S. A.* **83**, 226–230 (1986).
27. Haste Andersen, P., Nielsen, M. & Lund, O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* **15**, 2558–2567 (2006).
28. Chothia, C. & Janin, J. Principles of protein-protein recognition. *Nature* **256**, 705–708 (1975).
29. Kringelum, J. V., Nielsen, M., Padkjær, S. B. & Lund, O. Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.* **53**, 24–34 (2013).
30. Moore, P. L. *et al.* Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. *Nat. Med.* **18**, 1688–1692 (2012).
31. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422**, 307–312 (2003).
32. Liu, H. *et al.* A combination of cross-neutralizing antibodies synergizes to prevent SARS-CoV-2 and SARS-CoV pseudovirus infection. *Cell Host Microbe* **29**, 806-818.e6 (2021).
33. Das, S. R. *et al.* Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1417-22 (2011).
34. Li, Y. *et al.* The importance of glycans of viral and host proteins in enveloped virus infection. *Front. Immunol.* **12**, (2021).
35. Tada, T. *et al.* Increased resistance of SARS-CoV-2 Omicron variant to neutralization by vaccine-elicited and therapeutic antibodies. *EBioMedicine* **78**, 103944 (2022).
36. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
37. Yuan, M. *et al.* A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630–633 (2020).
38. Crowe, J. E. Principles of Broad and Potent Antiviral Human Antibodies: Insights for

Vaccine Design. *Cell Host Microbe* **22**, 193–206 (08/2017).
39. Vanshylla, K. *et al.* Kinetics and correlates of the neutralizing antibody response to SARS-CoV-2 infection in humans. *Cell Host Microbe* **29**, 917-929.e4 (2021).
40. Vanshylla, K. *et al.* Discovery of ultrapotent broadly neutralizing antibodies from SARS-CoV-2 elite neutralizers. *Cell Host Microbe* **30**, 69-82.e10 (2022).
41. Khare, S. *et al.* GISAID's role in pandemic response. *China CDC Wkly* **3**, 1049–1051 (2021).
42. Piccoli, L. *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* **183**, 1024-1042.e21 (2020).
43. Amanat, F. *et al.* SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD, and S2. *Cell* **184**, 3936-3948.e10 (2021).
44. Pastorio, C. *et al.* Determinants of Spike infectivity, processing and neutralization in SARS-CoV-2 Omicron subvariants BA.1 and BA.2. *bioRxiv* 2022.04.13.488221 (2022) doi:10.1101/2022.04.13.488221.
45. Bachmann, M. F., Mohsen, M. O. & Speiser, D. E. Increased receptor affinity of SARS-CoV-2: a new immune escape mechanism. *NPJ Vaccines* **7**, 56 (2022).
46. Ginex, T. *et al.* The structural role of SARS-CoV-2 genetic background in the emergence and success of spike mutations: the case of the spike A222V mutation. *bioRxiv* (2021) doi:10.1101/2021.12.05.471263.
47. Zhao, L. P. *et al.* Rapidly identifying new Coronavirus mutations of potential concern in the Omicron variant using an unsupervised learning strategy. *Res Sq* (2022) doi:10.21203/rs.3.rs-1280819/v1.
48. Bangaru, S. *et al.* Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. *Science* **370**, 1089–1094 (2020).
49. DeGrace, M. M. *et al.* Defining the risk of SARS-CoV-2 variants on immune protection. *Nature* **605**, 640–652 (2022).
50. Eddy, S. HMMER: biosequence analysis using profile hidden Markov models. *HMMER* www.hmmer.org.
51. Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J. & Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife* **7**, e34420 (2018).
52. Roop, J. I., Cassidy, N. A., Dingens, A. S., Bloom, J. D. & Overbaugh, J. Identification of HIV-1 Envelope Mutations that Enhance Entry Using Macaque CD4 and CCR5. *Viruses* **12**, (2020).
53. Duenas-Decamp, M., Jiang, L., Bolon, D. & Clapham, P. R. Saturation Mutagenesis of the HIV-1 Envelope CD4 Binding Loop Reveals Residues Controlling Distinct Trimer Conformations. *PLoS Pathog.* **12**, e1005988 (2016).
54. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
55. Chan, K. K., Tan, T. J. C., Narayanan, K. K. & Procko, E. An engineered decoy receptor for SARS-CoV-2 broadly binds protein S sequence variants. *Sci Adv* **7**, (2021).
56. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. *bioRxiv* (2022) doi:10.1101/2022.01.26.477860.

57. Doud, M. B. & Bloom, J. D. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses* **8**, (2016).
58. Wu, N. C. *et al.* Different genetic barriers for resistance to HA stem antibodies in influenza H3 and H1 viruses. *Science* **368**, 1335–1340 (2020).
59. Dingens, A. S. *et al.* Complete functional mapping of infection- and vaccine-elicited antibodies against the fusion peptide of HIV. *PLoS Pathog.* **14**, e1007159 (2018).
60. Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A. & Deverman, B. E. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife* **10**, (2021).
61. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
62. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
63. Rost, B. & Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216–226 (1994).
64. Lin, C.-P. *et al.* Deriving protein dynamical properties from weighted protein contact number. *Proteins* **72**, 929–935 (2008).
65. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biol.* **14**, e1002452 (2016).
66. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
67. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 140–144 (1984).
68. Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evol.* **8**, veac021 (2022).