Learning from pre-pandemic data to forecast viral antibody escape

by

Sarah Gurev

B.S., Stanford University (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Sarah Gurev
Department of Electrical Engineering and Computer Science
May 18, 2023

Certified by: Debora Marks
Professor of Systems Biology, Harvard University
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Learning from pre-pandemic data to forecast viral antibody escape

by

Sarah Gurev

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Effective pandemic preparedness relies on anticipating viral mutations that are able to evade host immune responses in order to facilitate vaccine and therapeutic design. However, current strategies for viral evolution prediction are not available early in a pandemic – experimental approaches require host polyclonal antibodies to test against and existing computational methods draw heavily from current strain prevalence to make reliable predictions of variants of concern. To address this, we developed EVEscape, a generalizable, modular framework that combines fitness predictions from a deep learning model of historical sequences with biophysical structural information. EVEscape quantifies the viral escape potential of mutations at scale and has the advantage of being applicable before surveillance sequencing, experimental scans, or 3D structures of antibody complexes are available. We demonstrate that EVEscape, trained on sequences available prior to 2020, is as accurate as high-throughput experimental scans at anticipating pandemic variation for SARS-CoV-2 and is generalizable to other viruses including Influenza, HIV, and understudied viruses with pandemic potential such as Lassa and Nipah. We provide continually updated escape scores for all current strains of SARS-CoV-2 and predict likely additional mutations to forecast emerging strains as a tool for ongoing vaccine development (evescape.org).

Thesis Supervisor: Debora Marks
Title: Professor of Systems Biology, Harvard University

Acknowledgments

I would like to thank my advisor Debora Marks for her support. Thanks to my mentor and co-author Nikki Thadani who I am so lucky to have worked with. I would also like to thank Noor Youssef and Pascal Notin whose work and advice I really appreciate. Additionally, thanks to other lab members for their help on this project: Nathan Rollins, Chris Sander, Ralph Estanboulieh, and Daniel Ritter. I also want to thank all members of the Marks Lab: Aaron Kollasch, Tessa Green, Ada Shaw, Alan Amin, Rose Orenbuch, Sam Berry, Han Spinner, Steffan Paul, Courtney Shearer, Hannah Pierce-Hoffman, and Japheth Gado for being supportive labmates and friends. My work has been funded by a MIT-Takeda Fellowship and CEPI. Finally, I am so deeply grateful to my family, especially my Mom, as well as my friends and cat Pika for always being there for me.

Contents

1 Introduction **9**

1.1 Overview of Thesis 10

2 Model **13**

2.1 Overarching framework 13

2.2 Fitness 15

2.3 Accessibility 18

2.4 Dissimilarity 19

3 Results **21**

3.1 Anticipating pandemic variation with pre-pandemic data 21

3.2 Comparative accuracy of EVEscape and high-throughput experiments 26

4 Adaptations **29**

4.1 Insertions and Deletions 29

4.2 Glycosylation 30

4.3 Pandemic Sequencing 31

5 Utility **33**

5.1 Strain forecasting with EVEscape 33

5.2 EVEscape generalizes to other viral families with pandemic potential 36

6 Conclusion **37**

A Supplementary Methods **41**

B Supplementary Figures **51**

C Supplementary Tables **73**

Chapter 1

Introduction

Viral diseases involve a complex interplay between immune detection in the host and viral evasion, often leading to the evolution of viral antigenic proteins. Antibody escape mutations affect viral reinfection rates and the duration of vaccine efficacy. Therefore, anticipating viral variants that avoid immune detection with sufficient lead time is key to developing optimal vaccines and therapeutics.

Ideally, we would be able to anticipate viral immune evasion by using experimental methods such as pseudovirus assays[50, 64] and higher-throughput deep mutational scans[19, 35, 32, 30, 31, 67, 66, 68, 78, 8, 9, 34, 29, 21, 18] (DMSs) that measure the ability of viral variants to bind relevant antibodies. However, these experimental methods require antibodies or sera representative of the aggregate immune selection imposed on the virus, which only become available as large swaths of the population are infected or vaccinated, limiting the impact for early prediction of immune escape. In addition, since pandemic viruses can evolve rapidly (tens of thousands of new SARS-CoV-2 variants are currently sequenced each month), systematically testing all variants as they emerge is intractable, even without considering the effects of potential mutations on currently circulating strains.

It is therefore of interest to build computational methods for predicting viral escape that can be used to identify mutations that may emerge. An ideal model would be able to assess escape likelihood for as-yet-unseen variation throughout the full antigenic protein, would inform the design of targeted experiments, would be updated

with pandemic information, and would make predictions with sufficient lead time for vaccine development (that is, before immune responses to the virus are observed). However, previous computational methods for forecasting viral fitness or immune escape depend critically on real-time sequencing or pandemic antibody structures, limiting their ability to predict unseen variants and making them impractical for vaccine development during the onset of a pandemic[49, 55, 57, 60, 4].

In this work, we introduce EVEscape, a flexible framework that addresses the weaknesses of existing methods by combining a deep generative model trained on historical viral sequences with structural and biophysical constraints. Unlike existing methods, EVEscape does not rely on recent pandemic sequencing or antibodies, making it applicable both in the early stages of a viral outbreak and for ongoing evaluation of emerging SARS-CoV-2 strains. By leveraging functional constraints learned from past evolution, as successfully demonstrated for predicting clinical variant effects[26, 40, 59], EVEscape can capture relevant epistasis[74, 83] and thereby predict mutant fitness within the context of any strain background. Moreover, EVEscape is adaptable to new viruses, as we demonstrate in both our validation on SARS-CoV-2, HIV, and Influenza and in predictions for the understudied Nipah and Lassa viruses. This approach enables advanced warning of concerning mutations, facilitating the development of more effective vaccines and therapeutics. Such an early warning system can guide public health decision-making and preparedness efforts, ultimately minimizing the human and economic impact of a pandemic.

1.1 Overview of Thesis

In Chapter 2 we provide an overview of the EVEscape model. In Chapter 3, we provide results for anticipating pandemic variation. In Chapter 4, we demonstrate how the EVEscape framework can be adapted to pandemic data and new models. In Chapter 5, we show two key applications of EVEscape: strain forecasting and predictions for new viruses with pandemic potential. Lastly, our conclusion is in Chapter 6.

This thesis is adapted from the manuscript:

Learning from pre-pandemic data to forecast viral antibody escape. Nicole Thadani*, **Sarah Gurev***, Pascal Notin*, Noor Youssef, Nathan Rollins, Chris Sander, Yarin Gal, Debora Marks. BioRxiv 2023. [76]

Chapter 2

Model

2.1 Overarching framework

Viral proteins that escape humoral immunity disrupt polyclonal antibody binding while retaining protein expression, protein folding, host receptor binding, and other properties necessary for viral infection and transmission[66]. We built a modeling framework—EVEscape—that incorporates constraints from these different aspects of viral protein function learned from different data sources. We express the probability of a single amino acid substitution to lead to immune escape as the product of three conditional probabilities (Figure 1A):

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

These components are amenable to pre-pandemic data sources, allowing for early warning (Figure 1B). 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

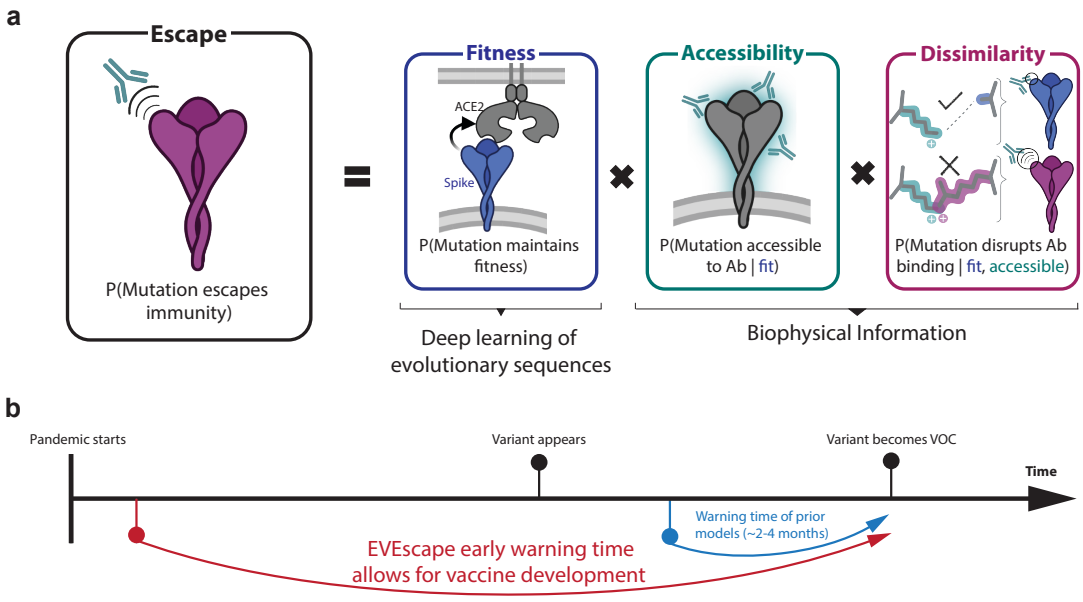

Figure 1: Early prediction of antibody escape from deep generative sequence models, structural and biophysical constraints. EVEscape assesses the likelihood of a mutation to escape the immune response based on the probabilities of a given mutation to maintain viral fitness, to occur in an antibody epitope, and to disrupt antibody binding. It only requires information available early in a pandemic, before surveillance sequencing, antibody-antigen structures or experimental mutational scans are broadly available.

below for details). To support modularity and interpretability of the impact of each component, each term is separately standardized and then fed into a temperature-scaled logistic function:

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

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 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 ($T_{dissimilarity} = 2.0$). We impute missing values of features in EVEscape using the mean value of the feature across the target protein. Moreover, to make strain-level EVEscape predictions, we aggregate across combinations of mutations by summing the EVEscape scores for each mutation.

2.2 Fitness

Firstly, we estimate the fitness effect of substitution mutations (subsequently referred to as mutations) using EVE[26], a deep variational autoencoder trained on evolutionarily-related protein sequences (Table S1-S2) that learns constraints underpinning structure and function for a given protein family. Consequently, EVE considers dependencies across positions (epistasis), capturing the changing effects of mutations as the dominant strain backgrounds diversify from the initial sequence[28, 69, 36].

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 $\mathbf{P}(x|\theta)$, where the parameters θ capture the constraints describing functional variants. A generative model trained on a Multiple Sequence Alignment (MSA) 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:

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

EVE (Evolutionary model of Variant Effects)[26] is a Bayesian variational autoencoder (VAE)[44], 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, which we define as the evolutionary index $E(x)$. 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:

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

The ELBO term itself is estimated via Monte Carlo sampling of the latent space, since the integral over z is intractable, using 20k samples from the approximate posterior distribution $q(z|x, \phi)$. These approximations have been shown to provide strong results in practice[26]. Results are obtained by ensembling scores from 5 independently trained EVE models with different random seeds.

We train the different models following the procedure from the original EVE paper (see Frazer et al.[26], Supplementary Section 3.2), using similarly-sized EVE models and with the same training hyperparameters. The only difference in our training procedure is that we slightly relax the constraint on minimum column coverage for sequences in the training MSAs (50% instead of 70%) as it led to superior fitness prediction performance in our hyperparameter tuning analyses for the different viruses modeled in our work. Due to biases in the our training MSAs from sampling and phylogeny, we reweight each protein sequence s_i by the reciprocal of the number of

sequences in the MSA (with a total of N sequences) that are within a Hamming distance cutoff T. Following prior work, we use a cutoff of 99% sequence identity that has been shown to work well for viral proteins, due to the relatively limited sequence diversity and an expectation that small difference is viral sequences wll have comparatively large impacts on fitness constraints. This weight (π_{s_i}) is therefore calculated as:

$$\pi_{s_i} = \left(\sum_{j=1, j \neq i}^{N} \mathbb{1}[Dist(s_i, s_j) < T] \right)^{-1}$$

We showcase the efficacy of EVE by comparing model predictions and data from mutational scanning experiments that measure multiple facets of fitness for thousands of mutations to viral proteins[20, 84, 36, 62, 22, 70, 11, 25]. Model performance approaches the correlation (ρ) between experimental replicates, including viral replication for influenza[20] ($\rho = 0.53$) and HIV[36] ($\rho = 0.48$) (Figure S1-S2, Table S3). For SARS-CoV-2, we trained EVE across broad pre-pandemic coronavirus sequences, from sarbecoviruses like SARS-CoV-1 to "common cold" seasonal coronaviruses like the Alphacoronavirus NL63 (Table S1), and compared predictions to measures of expression ($\rho = 0.45$) and host receptor ACE2 binding[70] ($\rho = 0.26$) (Figure S1-S3). We note that sites which express in the DMS experiments but are predicted deleterious by EVE are frequently in contact with non-assayed domains of the Spike protein or with the trimer interface- interactions not captured in the RBD yeast-display experiment (Figure S3), suggesting that they are important for full Spike expression but non-essential for RBD folding in the yeast-display system. Moreover, EVE predictions are better correlated with ACE2 binding quantified using a full Spike mammalian cell display assay [11], perhaps because this assay readout combines expression with receptor binding. As a whole, EVE's predictive performance on viral replication experiments and our analysis of model correspondence with RBD biochemical protein assays suggests that EVE captures a combination of the varied constraints on viral protein function. EVE predictions may also complement DMS studies that focus on biochemical protein assays by incorporating information about

non-assayed constraints.

2.3 Accessibility

The second model component, antibody accessibility, is motivated by the need to identify potential antibody binding sites without prior knowledge of B cell epitopes. Accessibility of each residue is computed from its negative weighted residue-contact number across available 3D conformations (without antibodies), which captures both protrusion from the core structure and conformational flexibility[77, 54, 37, 48] (Figure S4, Table S4). Accessibility plays a key role in identifying where antibodies are most likely to contact a protein, and 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 known to bind proteins[37, 77, 54, 48] (Figure S4). When computing antibody-binding likelihood metrics across different structural conformations (i.e., both open and closed structures for SARS-CoV-2 Spike) we used the maximum accessibility (or minimum weighted contact numbers).

We computed weighted contact numbers[48] for each residue from structure as the sum of the square of the reciprocal distance between residue i and all other residues in the full protein (i.e., the full Spike trimer for SARS-CoV-2):

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

where r_{ij} is the distance between the geometric centers of the residue i and residue j side chains. Weighted contact number, beyond capturing surface accessibility, captures protrusion from the core structure and conformational flexibility[77, 54, 37, 48]. By using squared distance, this value focuses on the degree of local interaction, and acts as a measure of exposure to the local environment that would permit antibody binding. It is both a simple and fast metric. 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.

We also explored RSA as a potential accessibility metric. To do so, we first computed accessible surface area based on hypothetical exposure to solvent water molecules using DSSP[42]. 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).

2.4 Dissimilarity

Finally, 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. These are properties known to impact protein-protein interactions[13, 46, 24]. This simple metric correlates with experimentally measured within-site escape more than individual chemical properties, BLOSUM substitution-matrix derived distance[38], or distance in the latent space of the EVE model (Figure S5).

To compute a combined charge-hydrophobicity dissimilarity index, we standard-scaled the charge and hydrophobicity differences and then took the sum of the scaled differences. We use the Eisenberg-weiss hydrophobicity consensus scale[24] and amino acid charge (as 1/0/-1) at physiological pH.

We compared our metric to other chemical properties: differences in size (side-chain mass), hydrophobicity, and charge. We also compared to the BLOSUM62[38] substitution matrix after dropping the null transition diagonal values. We explored latent space differences by examining metric of mutation distance learned by the EVE variational autoencoder. We calculated the L1 distance between the encoded representations of the wild-type sequence and a given single-mutation sequence in the latent space of the model, inspired by a similar approach from Hie et al.[39]

Chapter 3

Results

3.1 Anticipating pandemic variation with pre-pandemic data

Extensive surveillance sequencing and experimentation prompted by the COVID-19 pandemic have presented a unique opportunity to assess EVEscape's ability to predict immune evasion before escape mutations are observed[43, 79]. To test the model's capacity to make early predictions, we carried out a retrospective study using only information available before the pandemic (training on Spike sequences across Coronaviridae available prior to January 2020; Table S1, Data S1). We then evaluated the method by comparing predictions against what was subsequently learned about SARS-CoV-2 Spike immune interactions and immune escape.

The top predicted escape mutations for the whole of Spike are strongly biased towards the receptor-binding domain (RBD) and N-terminal domain (NTD), coincident with the bias for antigenic regions seen in the pandemic[56, 1] (Figures 2A-B, Figure S6). Within these domains, EVEscape scores are biased towards neutralizing regions—the receptor-binding motif of the RBD and the neutralizing supersite[10] in the NTD (Figure 2C). EVEscape's ability to identify the most immunogenic domains of viral proteins without knowledge of specific antibodies or their epitopes could provide crucial information for early development of subunit vaccines in an emerging

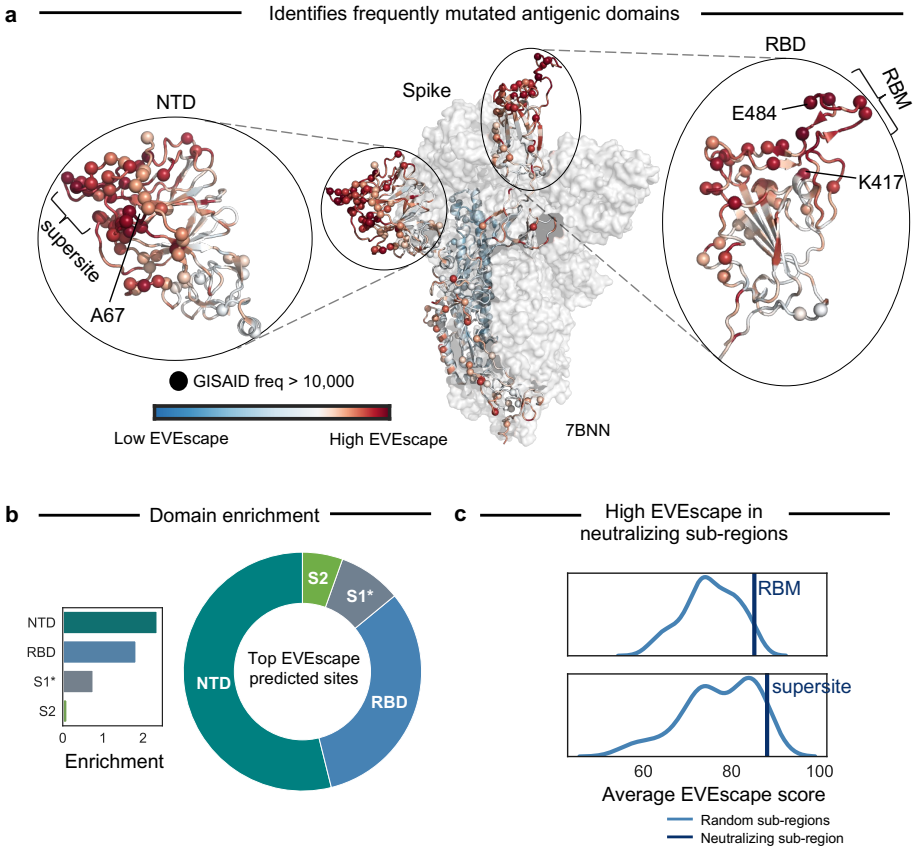

Figure 2: EVEscape identifies antigenic regions without antibody information. a) EVEscape scores mapped onto a representative Spike 3D structure (PDB identifier: 7BNN) highlight high-scoring regions with many observed pandemic variants, both in the RBD (receptor-binding domain) and NTD (N-terminal domain). Spheres indicate sites with mutations observed more than 10,000 times in the GISAID sequence database. b) The top decile of EVEscape predictions span diverse epitope regions across Spike, but the majority of predictions are in the NTD and RBD, which have a disproportionately high number of predicted EVEscape sites relative to their sequence length (enrichment). The regions considered are NTD (sequence positions 14 - 306), RBD (319 - 542), S1* (543 - 685), and S2 (686 - 1273), where S1* refers to the region in S1 between RBD and the S2. c) Neutralizing sub-regions – RBM (receptor-binding motif, 438-506) and NTD supersite50 (14-20,140-158, 245-263) – have significantly higher than average EVEscape scores, relative to a distribution of 150 random contiguous regions of the same length within the RBD and NTD, respectively.

pandemic[80].

We next compare model predictions to mutations that were subsequently observed in the pandemic as deposited in GISAID (Global Initiative on Sharing All Influenza Data)[43], which contains over 500,000 unique sequences with over 12,000 missense mutations to Spike. For this analysis we focus on the RBD of Spike as this domain has been the most extensively studied due to its immunodominance[56, 1].

49% of our top RBD predictions were seen in the pandemic by December 2022 (Figure 3A, Figure S7; this proportion is robust to the threshold defining top escape mutations). The more often a mutation occurred in the pandemic, the more likely it is to be predicted by our method — 57% of high frequency observed substitutions are in the top EVEscape predictions (Figures 3B-C). We expect that the highest frequency mutations, seen in historical Variants of Concern (VOCs), will be enriched for escape variants that provide a fitness advantage in an immune population (whilst not expecting that all single substitutions in the VOCs will contribute to escape).

Not surprisingly, the fitness model component alone (here EVE[26]) is better that the full EVEscape model at predicting mutations seen at low frequency in the pandemic – likely because these mutations retain viral function but do not necessarily affect antibody binding or have a strong fitness advantage over other strains (Figures S7-S8). This suggests that EVEscape's immune-specific components reflect important pandemic constraints and allow for mutation interpretability. For instance, VOC mutations R190S and R408S, with high EVEscape but low EVE scores, are in hydrophobic pockets that may facilitate significant immune escape[2] (Figure S8). Meanwhile, the few VOC mutations (i.e., A222V and T547K) with significant EVE—but not EVEscape—scores have functional improvements such as monomer packing and RBD opening but do not impact escape[27, 86] (Figure S8). We also see that the proportion of EVEscape predictions seen during the pandemic increased over time—from 3% in December 2020 to 49% in December 2022 (Figure 3A)—and should continue to increase, an expected trend both as more variants are observed and as adaptive immune pressure increases[45] with the growing vaccinated or previously infected population. Similarly, the fraction of mutations in VOC strains with high EVEscape scores has

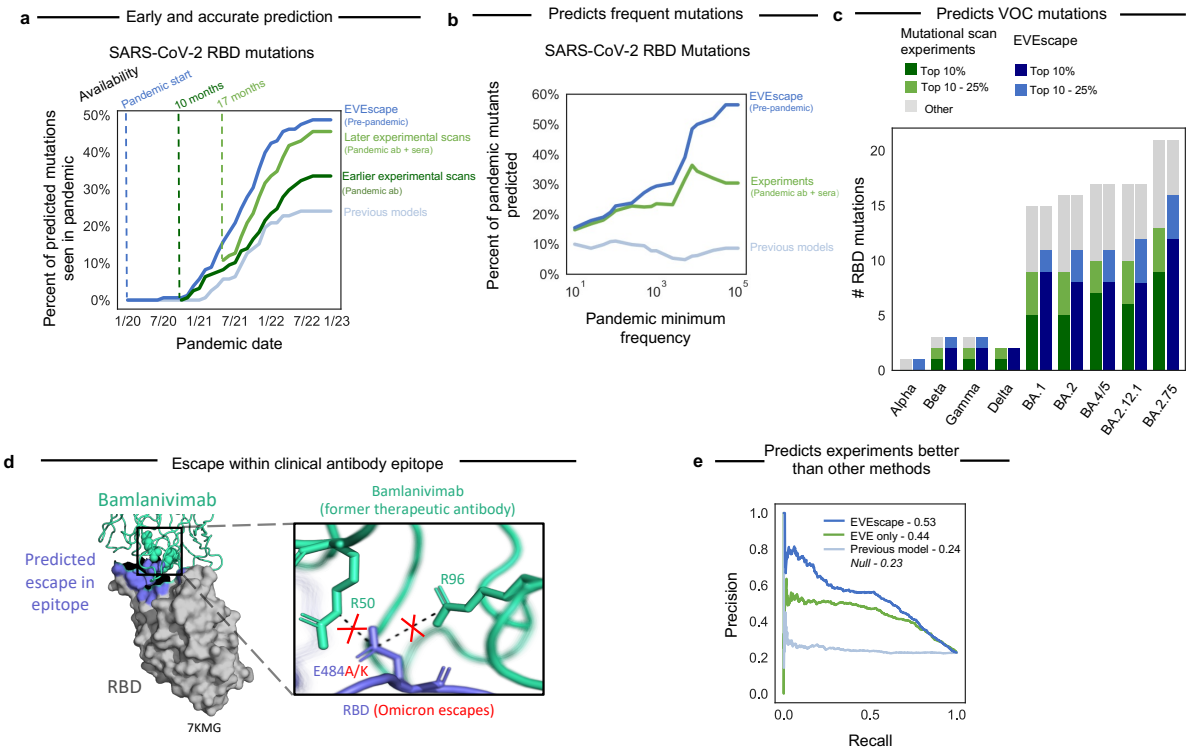

Figure 3: Pre-pandemic EVEscape is as accurate as intra-pandemic experimental scans at anticipating pandemic variation: retrospective analysis. a) Percent of top decile predicted escape mutations by EVEscape, mutational scan experiments (Bloom Set, Table S5), and a previous computational model[39] seen over 100 times in GISAID by each date since the start of the pandemic. EVEscape based on pre-pandemic sequences anticipates pandemic variation at least on par with mutational scan experiments based on antibodies and sera available 10 or 17 months into the pandemic. Analysis focuses only on nonsynonymous point mutations that are a single nucleotide distance away from the Wuhan viral sequence. RBD is the receptor-binding domain of the Spike protein. b) Percent of observed pandemic mutations in top decile of escape predictions by observed frequency during the pandemic. High-frequency mutations in particular are well-captured by EVEscape. c) The majority of RBD mutations observed in VOC strains have high EVEscape scores and somewhat lower scores in the mutational scan experiments against pandemic sera. d) EVEscape can predict escape mutations in the epitope of the former therapeutic antibody bamlanivimab. E484 is involved in a salt bridge with R96 and R50 of bamlanivimab, which lost FDA Emergency Use Authorization due to Omicron's emergence, wherein E484A or E484K mutations (both predicted in the top 1% of EVEscape Spike predictions) escape binding due to the loss of these salt bridges[73]. e) Precision-recall curve of RBD escape predictions of EVEscape, EVEscape fitness component only (EVE model) and previous computational model[39] when compared to DMS escape mutations (AUPRC reported with a comparison to a "null" model where escape mutations are randomly predicted).

also increased (Figure S7).

Our model also predicted escape mutations that were subsequently observed in the pandemic in the epitopes of well-known therapeutic monoclonal antibodies under current or former Emergency Use Authorization[79] (Figure S9), e.g., N440, E484A/K/Q, and Q493R. These predictions demonstrate the interplay of our three model components; for instance, the high accessibility as well as mutability of E484 results in 50% of all possible mutations at this site in the top 2% of EVEscape predictions and includes E484A/K in the top 1%—notable for escape from bamlanivimab[73] (Figure 3D)—because of their high dissimilarity scores. We also identify candidate escape mutations in these therapeutic epitopes that have not yet been observed at frequencies higher than 10,000 – for instance variants to K444 and K417 (Figure S9), a subset of which are beginning to appear. This result suggests that escape sites can be well predicted before a pandemic and may have concrete applications for escape-resistant therapeutic design and early warning of waning effectiveness.

EVEscape represents a significant improvement over past computational methods. EVEscape is more than twice as predictive as prior unsupervised models[39], both at predicting pandemic mutations (49% vs. 24% of top predictions observed in pandemic and 57% vs. 9% of highest frequency mutations predicted) as well as experimental measures of antibody escape (0.53 vs. 0.24 AUPRC) (Figures 3A-B, Figure 3E, Figure S7, Figures S10-S11, Figure S14, Table S5). All EVEscape components play a role in these predictions, with fitness predictions and accessibility metrics identifying sites of escape mutations while dissimilarity identifies amino acids that facilitate escape within sites (Figure S12-13). Moreover, other computational methods[57, 4] focus on near term prediction of strain dominance rather than longer term anticipation of immune evasion as they rely on pandemic sequences, antibody-bound Spike structures, or both, thereby hindering the ability to assess early predictive capacity. It is therefore notable that EVEscape outperforms even supervised approaches at predicting mutations seen in the pandemic (Figure S7).

3.2 Comparative accuracy of EVEscape and high-throughput experiments

We contextualize the performance of EVEscape in comparison to deep mutational scans (DMS), which have been invaluable in identifying and predicting viral variants that may confer immune escape[35, 32, 30, 31, 67, 66, 68, 8, 9, 19, 78]. However, these experiments require polyclonal or monoclonal antibodies from infected or vaccinated people, limiting their early predictive capacity. For example, the DMS experiments conducted by 17 months into the pandemic (using 36 antibodies and 55 sera samples) are a third more predictive (46% vs. 34% observed) than the experiments conducted 7 months prior (using just 10 antibodies) (Figure 3A, Figure S7).

Despite being computed on sequences available more than 17 months earlier, EVEscape is as good as, or better than, the latest DMS scans at anticipating pandemic variation (49% vs. 46% observed, respectively, when considering the top decile of prediction) (Figure 3A). As we consider higher frequency mutations, EVEscape increasingly predicts a greater portion of pandemic variation than experiments (Figure 3B) and predicts a higher fraction of mutations in VOC strains (Figure 3C).

Discrepancies between EVEscape and experiments shed light on the complementary strengths of these approaches. EVEscape and experiments miss 41 and 46 pandemic mutations, respectively, that are predicted by the other method (Figure 4A, Figure 4D). These differences could indicate model inaccuracies or could reflect sparse sampling of host sera response in DMS experiments as well as artifacts from experiments testing only the RBD domain and missing the full set of in vivo constraints. Indeed, as more antibodies are incorporated in experiments, the agreement between EVEscape and experimental predictions increases (Figure S14). The majority of high EVEscape predictions that are not observed in experimental predictions are in known antibody epitopes (Figure 4B, Figure S13). By contrast, those mutations identified by the experiments that are below the threshold in EVEscape predictions are often predicted to have low fitness due to high conservation in the alignment at those positions.

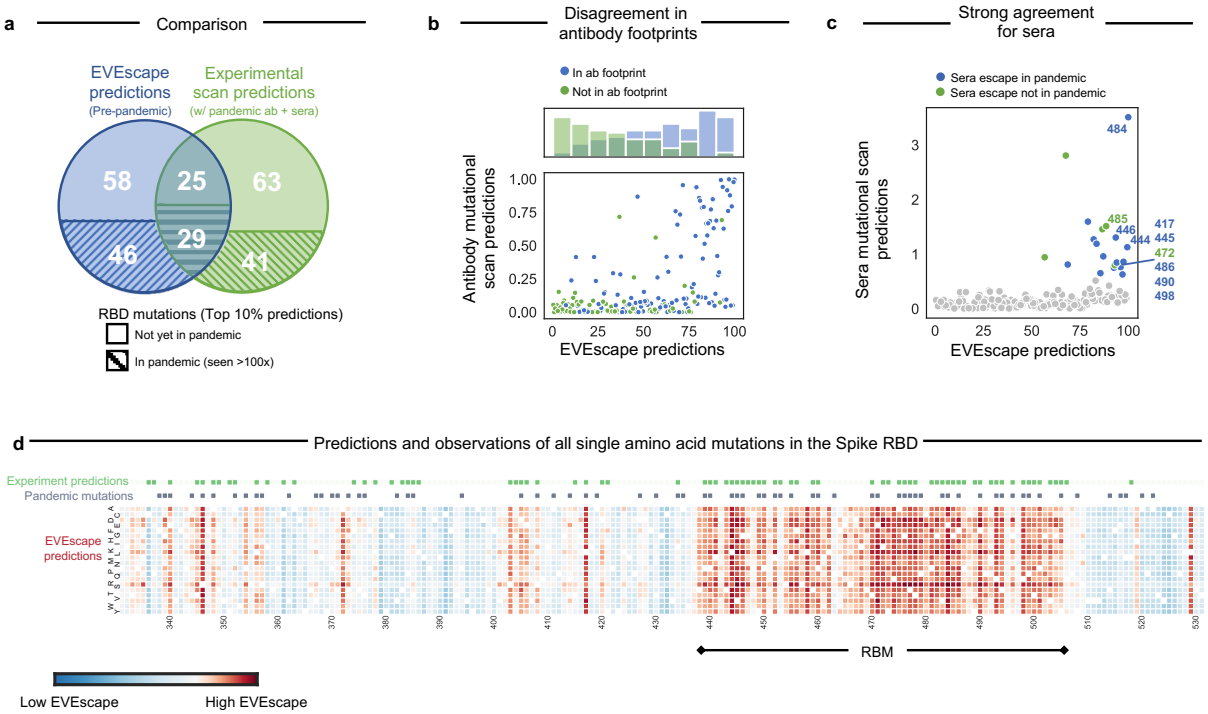

Figure 4: EVEscape and experiments make distinct, complementary escape predictions. a) Share of top decile of predicted escape mutations, predicted using EVEscape or based on mutational scan experiments (Bloom Set, Table S5), seen so far over 100 times in the pandemic. As the virus evolves further, more of the predicted escape mutations are expected to appear. b) RBD site-averaged EVEscape scores agree with site-averaged antibody escape experimental mutational scan measures (Bloom Set, Table S5), with high EVEscape sites that are missing from experimental escape prediction found within known antibody footprints. Hue indicates known antibody footprints from the PDB (information that EVEscape does not use as a pre-pandemic model). c) Predicted escape mutations based on experimental mutational scans (Bloom Set, Table S5) measuring recognition by convalescent sera from patients infected with either Wuhan, Beta, or Delta have high EVEscape scores. Mutations that escape sera are colored by whether they have occurred in the pandemic over 100 times. d) Heatmaps illustrating the EVEscape scores of all single mutations to the Wuhan sequence of SARS-CoV-2 RBD. Top lines are sites with observed pandemic mutation frequency >100 and sites in the top 15% of DMS experimental predictions based on mutational scan experiments. RBM is the receptor-binding motif.

The consensus between EVEscape and experiments is also of interest. We see that agreement is especially strong for polyclonal patient sera (Figure S14); in fact, half of the top 10% of EVEscape RBD sites are sera escape sites from experiments[29, 34, 30, 31, 32] (Figure 4C). These mutants are of particular interest since they escape from the unique composition of antibodies produced by convalescent patients and are thus crucial to considerations of reinfection and vaccine design. For instance, E484, mutated in several VOCs, has the highest experimental sera binding and is the top EVEscape predicted site.

Chapter 4

Adaptations

The modular design of our framework facilitates its adaptability to the specific characteristics of a pandemic and to new data as it becomes available.

4.1 Insertions and Deletions

To consider the effects of insertions and deletions on SARS-CoV-2 Spike immune escape[58], we replace the EVE fitness component with TranceptEVE[53] – a recently developed protein large language model which has previously demonstrated state-of-the-art performance for mutation effects prediction, including indels, which both prior computational models and high-throughput experiments have been unable to capture for SARS-CoV-2. When applied to the pandemic, this model captures the most frequent single insertion and deletion, both at site 144, each in the top decile of pandemic and random indel predictions (Figure S15).

Scores for indels utilize tranceptEVE as the fitness component, negative weighted contact number as the accessibility component, and a maximized dissimilarity component score. TranceptEVE is itself based off of two key components: 1) Tranception[52], a family-agnostic autoregressive transformer trained on a large quantity of unaligned protein sequences from Uniref100[72] from February 2022. 2) A family-specific EVE model that is trained to score sequences for a family of interest, and which acts as a prior distribution over amino acids at each sequence position. The predicted fitness for

a given sequence is then obtained as a weighted average of the log likelihood assigned by these two components – the weights depending on the depth of the alignment used to train the underlying EVE model (deeper alignments implying a larger weight assigned to the EVE log likelihood). We use the same ensemble of 5 EVE models as described above, as well as the large Tranception model checkpoint (∼700M model parameters) made available in Notin et al.[52] which was trained on Uniref100 (see details of the training procedure in the corresponding paper in Appendix B.3).

4.2 Glycosylation

We also show that including glycosylation in the dissimilarity component for HIV Env, where glycans play an important role in immune escape[51, 82, 16, 47], improves model predictions of high-throughput experimental escape[18] (Area under the precision recall curve raises 10% when including glycosylation for HIV; Figure S16). While addition of glycosylation is also important for escape[51, 82, 16, 47], we focus here on loss of glycosylation for simplicity. We adapt the model by maximizing 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. We consider that a mutation is likely to result in loss of glycosylation if the N or S/T is lost. We note that this can be an important factor for real-world escape even when some DMS experiments do not reflect the escape impacts of glycosylation loss, as is the case for SARS-CoV-2 experiments that use yeast display, with glycans different than in mammalian cells[35]. For HIV on the other hand, a significant portion of escape mutations from DMS experiments are a result of escape effects of glycan gains and loss[18].

4.3 Pandemic Sequencing

Additionally, we retrain EVE models with the addition of 11 million new sequences collected during the pandemic, which helps improve agreement with fitness DMS experiments by 20% (Figures S1, S17). This model captures epistatic shifts between Wuhan and BA.2, identifying changes in mutation fitness in the RBD and near BA.2 mutations and predicting positive epistatic shifts for known convergent omicron mutations and likely-epistatic wastewater mutations[65] (Figure S18).

Chapter 5

Utility

5.1 Strain forecasting with EVEscape

A key application of an escape prediction framework is to identify circulating strains with high immune escape potential soon after their emergence, thus enabling the deployment of targeted vaccines and therapeutics before their spread. While the World Health Organization seeks to identify new high-risk variants as they arise, new strains are occurring at an increasing rate with now tens of thousands of novel SARS-CoV-2 strains each month, a scale infeasible for experimental risk assessment[71]. To create strain-level escape predictions, we aggregated EVEscape predictions across all individual Spike mutations in a strain. We evaluated EVEscape strain predictions for their alignment with experimental measures of strain immune evasion as well as their identification of known escape strains from pools of random sequences and from other strains observed at the same pandemic timepoint.

First, we see that pre-pandemic EVEscape-strain scores correlate well with experiments quantifying vaccinated sera neutralization of 21 strains[4] ($\rho = 0.80$; Figure 5A, Data S5), better than an existing computational strain-scoring method ($\rho = 0.77$)[4] even though that method uses 332 pandemic antibody-Spike structures for the prediction. Second, we show that EVEscape-strain scores for VOCs are consistently higher than random sequences at the same mutational depth, and in particular the Beta and later Omicron BA.2, BA.4, BA.2.12.1, BA.2.75, and XBB strains are in the top 1% of

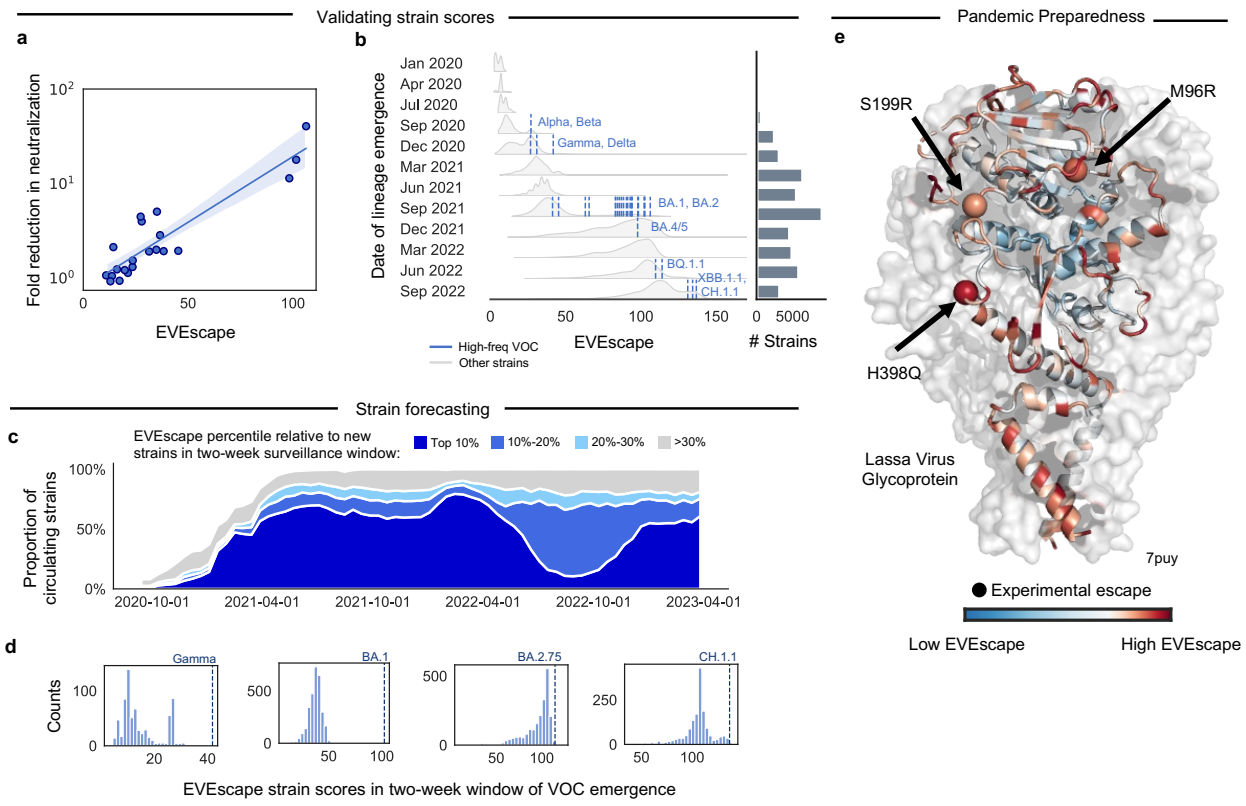

Figure 5: EVEscape applications: Identifying strains with high escape potential and forecasting escape for future pandemics. a) Pre-pandemic EVEscape scores computed for pandemic strains correlate with fold reduction in 50% pseudovirus neutralization titer [4] for each strain relative to Wuhan ($\rho = 0.80, n = 21$). Linear regression line shown with a 95% confidence interval. b) Distributions of newly emerging EVEscape strain scores for non-VOCs (unique combinations of mutations) throughout 12 periods of the pandemic, with counts of the number of unique new strains per period. EVEscape strain scores increase throughout the pandemic. High frequency VOCs (occurring more than 5000 times) are shown in the first period each emerged, depicting that new VOCs are predicted to have higher escape scores than most strains in all previous time periods. c) Pandemic circulating strains are grouped according to their EVEscape decile relative to other strains emerging in the same non-overlapping two-week surveillance window. The relative prevalence of each EVEscape decile over the course of the pandemic is plotted in a stacked lineplot. The majority of circulating strains fall into the top 10% bin. Proportions do not sum to 100% as strains that emerged before the surveillance period of 9/2020 - 3/2023 are not included. d) VOCs (dotted lines) are among the highest scoring of hundreds or thousands of new strains (histograms) within their two-week window of emergence, enabling EVEscape to forecast which strains will dominate as soon as they appear after only a single observation. e) Site-wise maximum EVEscape scores on Lassa Virus Glycoprotein structure (PDB: 7PUY). We show agreement between sites of high EVEscape scores (in red) and escape mutations with experimental evidence (shown with spheres).

these generated sequences (Figure S19). EVEscape strain scores for these VOCs are also in the top 2% 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 (Figure S19).

Lastly, we examine EVEscape's ability to identify immune-evading strains as they emerged in the pandemic. We see that EVEscape scores have increased throughout the pandemic and that they are higher for more recent VOCs, reflecting their increased propensity for immune escape (Figure 5B). Moreover, EVEscape scores for newly emerging VOCs are higher than almost all strains in previous time periods (Figure 5B). Taken together, these results suggest EVEscape's promise as an early-detection tool for picking out the most concerning variants from the large pool of available pandemic sequencing data. We therefore examine EVEscape's utility as a tool to identify strains with high escape potential as they emerge in two-week surveillance windows. We see that the majority of circulating strains were in the top decile of EVEscape scores for their two-week window of emergence (Figure 5C). Moreover, in the two-week windows where the VOC strains Alpha, Beta, Gamma, Omicron BA.1, and Omicron BA.2.75 emerged, each VOC ranked in the top 5 of hundreds or thousands of new strains (Figure 5D, Figure S19). This demonstrates the ability of EVEscape to forecast which strains will dominate as soon as they appear after only a single observation, even as experimental testing of all emerging strains has become intractable. To enable real-time variant escape tracking, we make monthly predictions (Data S5) available on our website (evescape.org), with EVEscape rankings of newly occurring variants from GISAID and interactive visualizations of likely future mutations to our top predicted strains. In sum, the EVEscape model captures relative immune evasion of successful strains and can identify concerning strains from pools of random combinations of mutations as well as from their temporal peers.

5.2 EVEscape generalizes to other viral families with pandemic potential

Most viruses with pandemic potential have far less surveillance and research than SARS-CoV-2[75]. One of the main features of EVEscape is the ability to predict viral antibody escape before a pandemic—without the consequent increase in data during a pandemic—to narrow down vaccine sequences and therapeutics most likely to provide lasting protection, to assess strains as they arise, and to provide a watch list for mutations that might compromise any existing therapies. As one of the first comprehensive analyses of escape in these viruses, we applied the EVEscape methodology to predict escape mutations to the Lassa virus and Nipah virus surface proteins; these viruses cause sporadic outbreaks of Lassa hemorrhagic fever in West Africa and highly lethal Nipah virus infection outbreaks in Bangladesh, Malaysia, and India. Crucially, the three mutants present in the Lassa IV lineage that are known to escape neutralizing antibodies[7] are all in the top 10% of EVEscape predictions, suggesting that EVEscape captures features relevant for Lassa glycoprotein antibody escape (Figure 5E). EVEscape predictions also identify 11 of 12 known escape mutants to Nipah antibodies[6, 81, 85, 14, 15] (Figure S20).

Moreover, we demonstrate generalizability to Influenza Hemagglutinin[21] and HIV Env[18] based on DMS evaluation (Figure S10, Data S3). Based on these findings, we provide all single mutant escape predictions for these proteins (Data S5) to inform active and ongoing vaccine development efforts with the goal of mitigating future epidemic spread and morbidity.

Chapter 6

Conclusion

One of the greatest obstacles for developing vaccines and therapeutics to contain a viral epidemic is the high genetic diversity derived from viral mutation and recombination, especially when under pressure from the host immune system. An early sense of potential escape mutations could inform vaccine and therapeutic design to better curb viral spread. Computational models can learn from the viral evolutionary record available at pandemic-onset and are widely extensible to mutations and their combinations. However, novel pandemic constraints (such as immunity) are unlikely to be captured. To achieve early escape prediction, EVEscape combines a model trained on historical viral evolution with a biologically informed strategy using only protein structure and biophysical constraints to anticipate the effects of immune selection. We demonstrate that EVEscape forecasts pandemic escape mutations and can predict which emerging strains have high escape potential through a retrospective analysis of the SARS-CoV-2 pandemic. This computational approach can preempt predictions from experiments that rely on pandemic antibodies and sera by many months while providing similar levels of accuracy.

EVEscape provides surprisingly accurate early predictions of prevalent escape mutations but cannot anticipate all constraints unique to a new pandemic to determine the precise trajectory of viral evolution. This method will be best leveraged in synergy with experiments developed to measure immune evasion and enhanced with pandemic data as it becomes available. Early in a pandemic, EVEscape can predict

likely escape mutations for prioritized experimental screening with the first available sera samples – validated escape mutations could be strong candidates for multivalent vaccines. EVEscape can also identify structural regions with high escape potential, so therapeutic antibody candidates with few potential escape mutants in their binding footprint may be accelerated. Later in a pandemic, EVEscape can rank emerging strains, as well as mutants on top of prevalent strains, for their escape potential, flagging concerning variants early on for rapid experimental characterization and incorporation into vaccine boosters. The model can also be augmented to leverage current knowledge on virus-specific immune targeting and mutation tolerance from experimental and pandemic surveillance data. In return, our computational framework can inform this collective understanding by proposing escape variant libraries for focused experimental investigations.

Each component of EVEscape may be independently refined over time as additional information is collected during a pandemic. Firstly, the features driving immune escape depend heavily on the regions of the virus targeted by adaptive immunity: as data becomes available about the specific epitopes targeted in vaccinated and convalescent sera, EVEscape can be adapted to emphasize escape mutations at these regions, including for SARS-CoV-2 regions throughout the full viral proteome. Strain-level predictions may also be modified to evaluate the cumulative contribution of variants to likely polyclonal escape. Secondly, structures of antibodies bound to the viral glycoprotein can also provide specific insight on the types of mutants most likely to displace known antibodies and result in successful escape. Lastly, pandemic sequencing data can be incorporated alongside broader viral evolution to provide fine-grained information about the new fitness constraints faced by a pandemic viral species and enhance the ability of our model to extrapolate to new regions of sequence space.

EVEscape is a modular, scalable, and interpretable probabilistic framework designed to predict escape mutations early in a pandemic and to identify observed strains and their mutants that are most likely to thrive in a populace with widespread pre-existing immunity as the pandemic progresses. To this end, we provide EVEscape

scores for all single mutation variants of SARS-CoV-2 Spike to Wuhan as well as scores for all observed strains and predictions of single mutation effects on the most concerning emerging strain backgrounds, with plans to continuously 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. To further accelerate broad and effective vaccine development, we provide EVEscape mutation predictions for all single mutations to Influenza, HIV, Lassa virus and Nipah virus surface proteins.

Appendix A

Supplementary Methods

Data Acquisition

Training Data

Multiple sequence alignments for fitness models

For each viral protein, we construct multiple sequence alignments performing 5 iterations of the profile-HMM based homology search tool jackhmmer[41] against the UniRef100 database[72]. 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[40, 59, 26]. 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 optimize 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 (theta) 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 (Table S2).

Alignments with pandemic sequences

We construct an "evolutionary alignment" with non-SARS-CoV-2 sequences as described above using jackhmmer (with at least 50% sequence coverage, at least 30% column coverage, and theta of 0.01). We extract the full sequences pulled into the jackhammer alignment and re-align the sequences using super5[23], then remove gapped positions relative to the Wuhan sequence. We also construct a "pandemic alignment" with all unique Spike sequences (with count >100) seen up until 11/27/21 (when BA.2 first appeared in GISAID). We then concatenate that "pandemic alignment" with the "evolutionary alignment" to create the final alignment.

Protein structures for accessibility calculation

For each viral surface protein, we selected crystal structures representing known structural states available to B-cell and antibody interactions (extracellular conformations) (Table S4). All heteroatoms and protein chains not part of the multimeric viral surface protein were removed.

Evaluation data

Antibody footprints

To identify known antibody footprints of viral surface proteins in the RCSB PDB[5], 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 3.5Å 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. We use this same method to identify antibody footprints for specific clinical antibodies. For experimental evidence of clinical antibody escape susceptibility, we used the Stanford Coronavirus Antiviral & Resistance Database

(CoV-RDB) susceptibility summary for monoclonal antibodies under emergency use authorization[79].

Deep mutational scans

We benchmark our models on a series of viral protein deep mutational scans[19, 35, 30, 32, 34, 31, 29, 68, 67, 78, 66, 21, 18, 8, 36, 62, 22, 70, 11, 20, 84, 9, 25] (Table S3, Table S5). 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[17]. 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[66]. We use the RBD dimeric ACE2 binding and expression DMS data for analysis[70].

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)[43] on 10/24/22. We further processed this data to get counts of combinations of mutations, the date of emergence, and PANGO lineage, as well as to get the month of emergence for each single mutation in Spike. We also downloaded consensus mutations for each PANGO lineage on 10/31/22 and mutation frequencies on 10/26/22 from Covid-19 CG[12].

Lassa virus and Nipah virus antibody escape data

We aggregated data on single mutations resulting in escape from known Lassa and Nipah virus antibodies from literature studies with experimentally determined reduction in antibody binding, reduction in antibody neutralization, or emergence in growth selection experiments[14, 15, 6, 81, 85, 7].

Epistasis mutation sets

Our convergent omicron mutation set is defining mutations in Omicron lineages at sites 346, 444, 452, 460, and 486. This set is: L452R, N460K, F486V, K444N, L452M, F486I, R346T, F490S, K444M, K444T. Our wastewater mutation set is the set of mutations from Smyth et al.[65], which are mutations that were frequent in wastewater, but had rarely been seen clinically (pre-Omicron, mid 2021), so may be likely epistatic. This set is: Q493K, Q498Y, Q498H, T572N, H519N, H519Q.

Strain Neutralization data

We download neutralization data from Beguir et al.[4], which contains the observed 50% pseudovirus neutralization titer (pVNT50) for 21 SARS-CoV-2 S protein variants. The pVNT50 reduction is relative to Wuhan. Neutralization is measured for $n \geq 12$ sera collected after primary 2-dose vaccination by the Pfizer BioNTech vaccine (BNT162b2) and assessed against vesicular stomatitis virus (VSV)-based pseudoviruses with each S protein variant.

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[40, 59].

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[18]. 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 S11). 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 S3). We use thresholds of 0.57 for Bloom RBD, 0.9 for Xie RBD, 0.054 for Flu, and 0.138 for HIV to make model comparisons; mutations designated as escape by these experimental thresholds are almost all within 5Å of the antibody they escape (Figure S11). Note that the downloaded RBD escape datasets were already filtered using thresholds on expression and ACE2 binding of -1 and -2.35, respectively[33]. To define a site-wise escape value, we averaged across the maximum escape values for each mutant at the site. 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 S3). As the scales are different for the Bloom and Xie datasets, we focus on the original Bloom RBD DMS data when we need to consider the top fraction of escape mutations. 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.

Metrics

To compare computational 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). 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. 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. Note that this is not information that is available to the model during training.

Comparison to pandemic 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. The date of lineage emergence is the 1st percentile of dates for that variant (to avoid issues with outliers from GISAID data entry). Variants are marked as high frequency VOCs if their count is greater than 5,000 and it occurs in the first time period (pandemic divided into 12 periods) that any strain of that PANGO lineage appears. 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 90% of strain sequences, ignoring insertions and deletions. Number of occurrences in the pandemic is defined by raw counts of GISAID records with a given substitution or set of substitutions.

Metrics

We calculate the fraction of predicted mutations (top 10%) seen in the pandemic over 100 times. We expect to see an increase in this fraction over the course of the pandemic, as more variants are observed and adaptive immune pressure increases with a growing vaccinated or previously infected population. We also calculate for each observed pandemic frequency minimum threshold, the percentage of pandemic mutants seen above that observed threshold that are predicted in the top 10%. We do not expect all pandemic mutants to be captured in the top 10% of predictions, because not all pandemic mutants are related to escape. Even amongst very frequent pandemic mutations mostly present in Variants of Concern, which we expect to be more enriched for high escape potential, we do not expect all of these mutations to be related to escape as some instead influence ACE2 binding or structural changes. To evaluate strain scores, we calculate the number of strains (and the corresponding percentile) that would need to be tested to have detected selected VOCs from all new

strains in the two-week window they emerged. Unique new strains are defined by unique sets of Spike substitution mutations.

Escape within clinical antibody epitopes

We look at EVEscape predictions in the footprints (within 3.5Å) of six different clinical antibody epitopes. We then notate which of these mutations have already occurred in the pandemic (observed more than 10,000 times) and which have experimental evidence of escape for those clinical antibodies as seen in CoV-RDB[79]. We list all possible mutations, not just those a single nucleotide distance from Wuhan.

Comparison to strain neutralization

We show spearman correlation with experimental strain neutralization data as well as the linear regression line shown with a 95% confidence interval. EVEscape scores for these strains are calculated based on the mutations used in the experiment for each strain, ignoring indels. We convert percent reduction in neutralization (x) to fold reduction (1/1-x).

Regional Enrichment

We examine the distribution of EVEscape predictions throughout the Spike protein and, within the RBD, between the known footprints of different antibody classes[3]. We analyze enrichment of regions by comparing the average EVEscape score for the region 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. We similarly compare scores of known neutralizing subregions to random regions in their respective full regions. We also compare enrichment of number of sites in the top 15% of EVEscape scores in each region relative to the length of the region. We consider the regions: NTD (sequence positions 14 - 306), RBD (319 - 542), S1* (543 - 685),

and S2 (686 - 1273), where S1* refers to the region in S1 between RBD and S2. NTD and RBD are enriched in antibody sites. We also calculate the mutational tolerance of each region, the average EVE fitness score.

Epistasis

We analyze epistasis by comparing EVE scores on a Wuhan full Spike model (using a pre-pandemic alignment) and on an omicron (BA.2) full Spike model (using an alignment with data up to BA.2). The BA.2 epistatic shift is the Wuhan linear regression residual for a model fit to the two sets of EVE scores for all single mutations to full Spike. We compare the epistatic shift of two subsets of mutations, convergent omicron mutations and wastewater mutations[65], to the full set of single mutations to full Spike. We also analyze the locations of the maximum epistatic shift, in relation to the Spike structure and to the set of sites mutated within BA.2.

Comparison to other computational models

We compare published SARS-CoV-2 RBD and Spike models predictions[61, 57, 39, 4] using metrics from above relevant to the intended purpose of each model (fitness or escape of either single mutations, sites, or strains).

Appendix B

Supplementary Figures

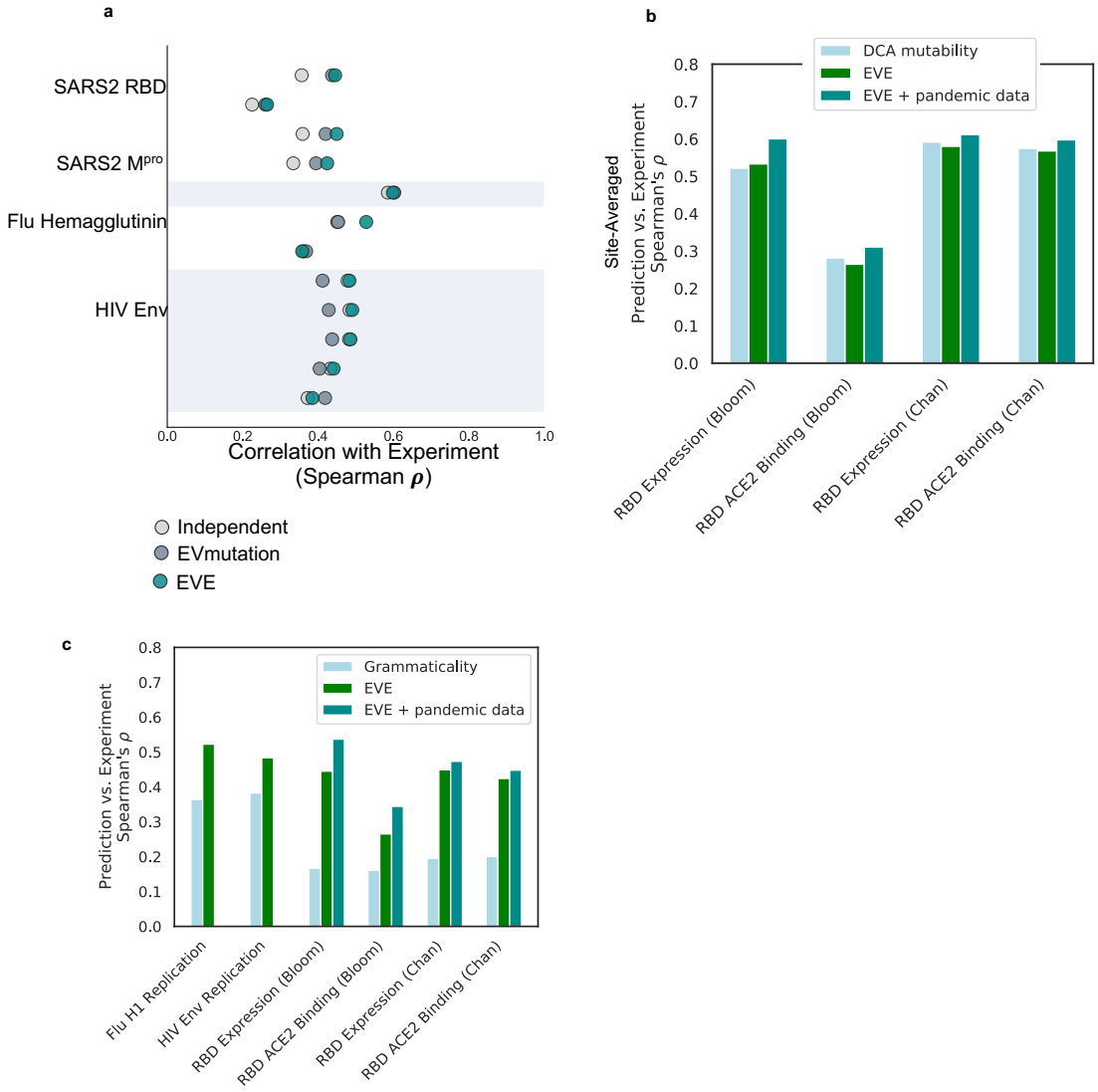

Figure S1: Fitness effects of viral proteins predicted from evolutionary sequence models. a) EVE predictions are well correlated with a broad range of viral surface protein deep mutation scanning experiments surveying protein replication and function, SARS-CoV-2 RBD [70, 11] and Mpro [25], H1N1 hemagglutinin [20, 84] and HIV env [36, 62, 22]. b) Site-averaged EVE predictions have similar correlations with site-averaged SARS-CoV-2 RBD DMS experiments as Potts model DCA [61] or EVmutation [40]. c) EVE predictions have higher correlations with Flu H1, HIV Env, and SARS-CoV-2 RBD DMS experiments than grammaticality in CSCS [39].

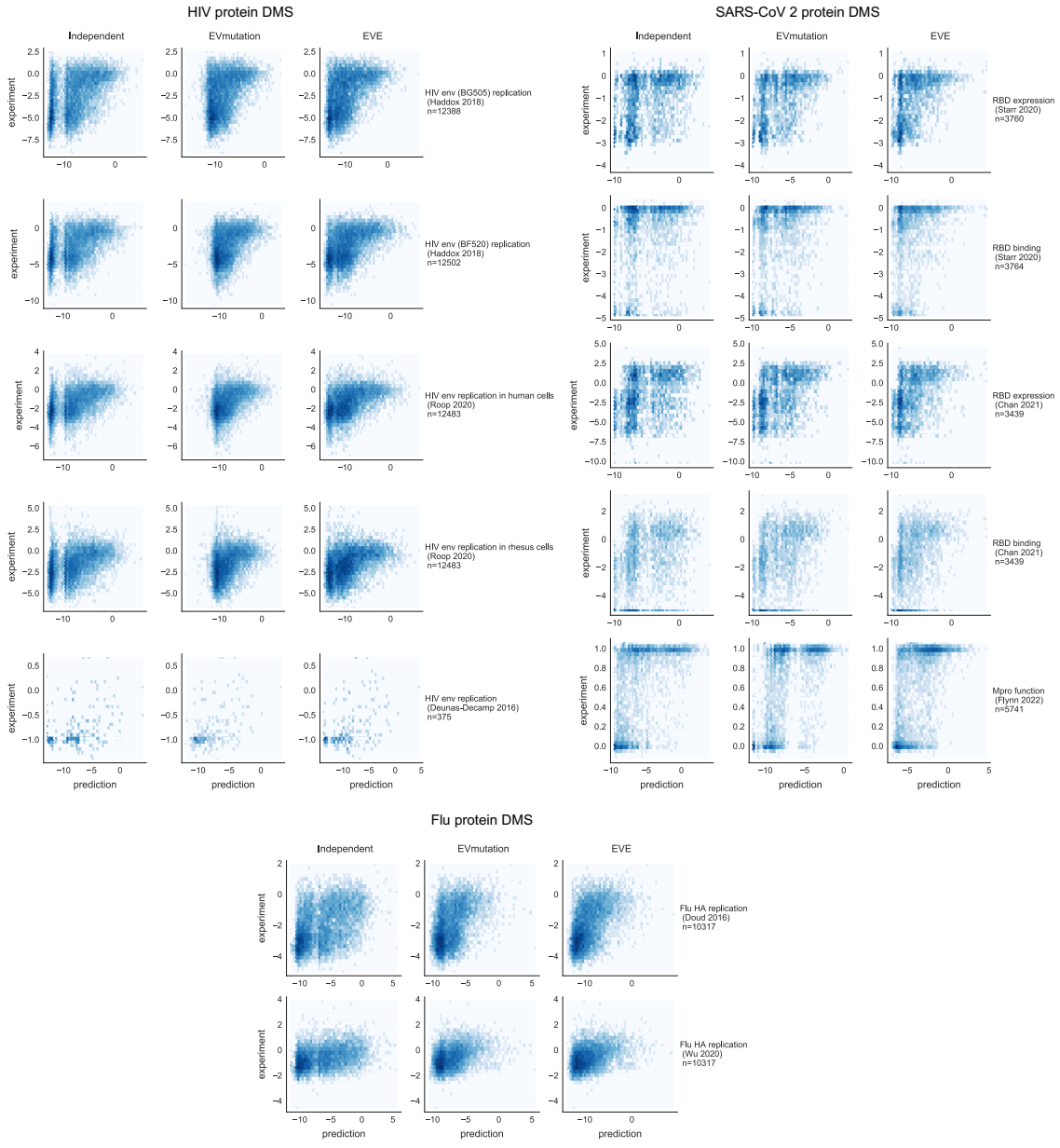

Figure S2: Comparison of effects of mutations from experiment and computation. Measurements of viral protein functions such as expression, replication and receptor binding in deep mutational scans versus predictions from a site-independent model, EVmutation, and EVE model.

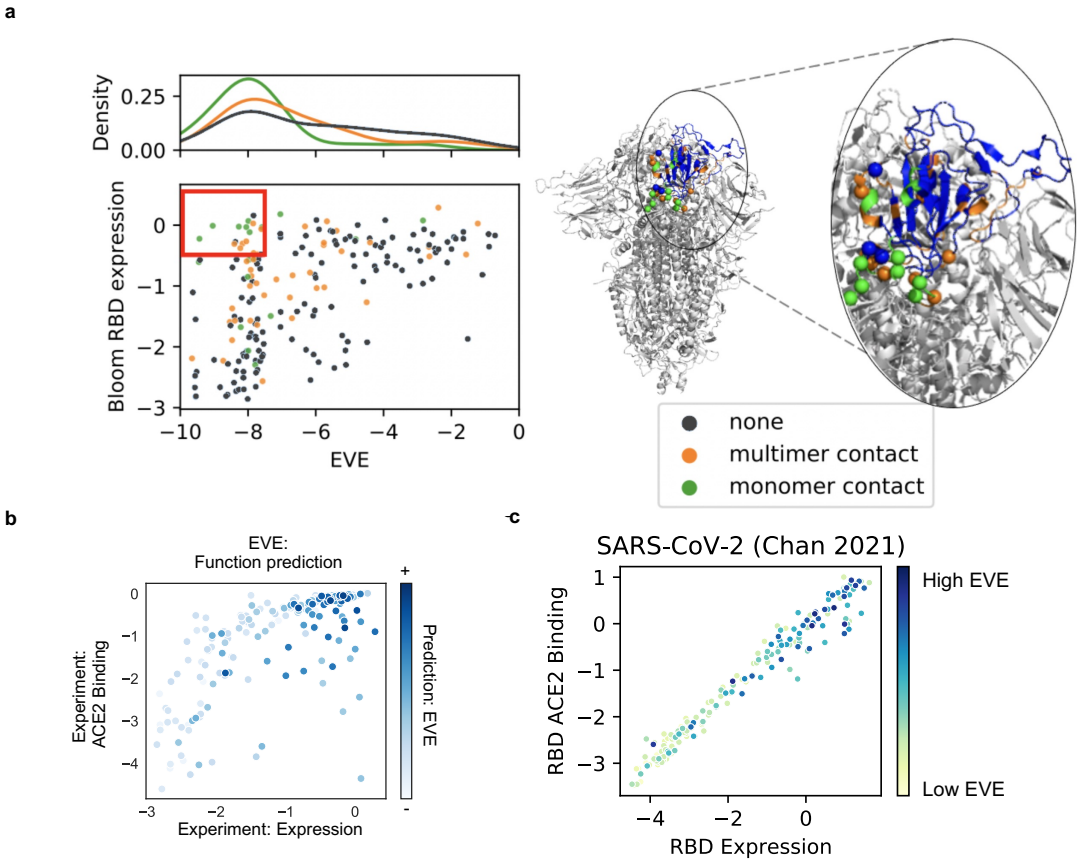

Figure S3: EVE captures constraints beyond RBD expression assay. a) Site-averaged EVE scores predict several sites that tolerate mutants in the yeast-display RBD expression assay [70] to be deleterious (red box)–many of these mutants are located at the interface between RBD and the rest of Spike protein. Sites in the red box in scatterplot are shown as spheres on the Spike structure (PDB: 7CAB). b) EVE prediction captures a combination of SARS-CoV-2 RBD yeast expression and ACE2 binding - features both necessary for successful immune escape (EVE spearman with expression = 0.45, EVE spearman with ACE2 binding = 0.38 when low expressed are removed)35 c) The mammalian-cell RBD expression and ACE2 binding experiments are highly correlated, likely due to the alternate FACS-binning strategy and metric used for this ACE2 binding experiment [11]. EVE predictions are correlated with both measures.

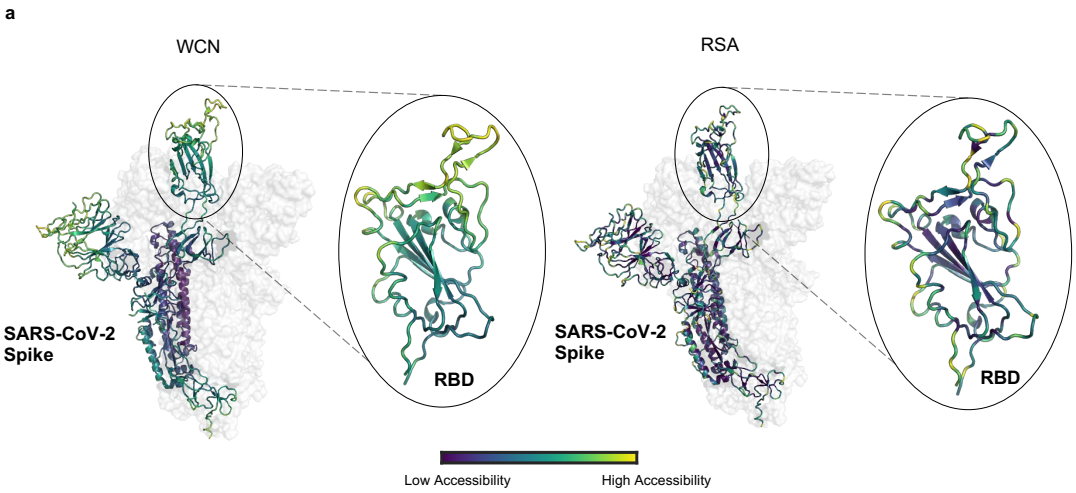

Figure S4: Weighted contact number captures flexible regions. a) WCN and RSA values visualized on the SARS-CoV-2 Spike structures show different distributions, particularly in the RBD (PDB: 7BNN), as WCN captures protrusion from the core structure. [77, 54, 37, 48]

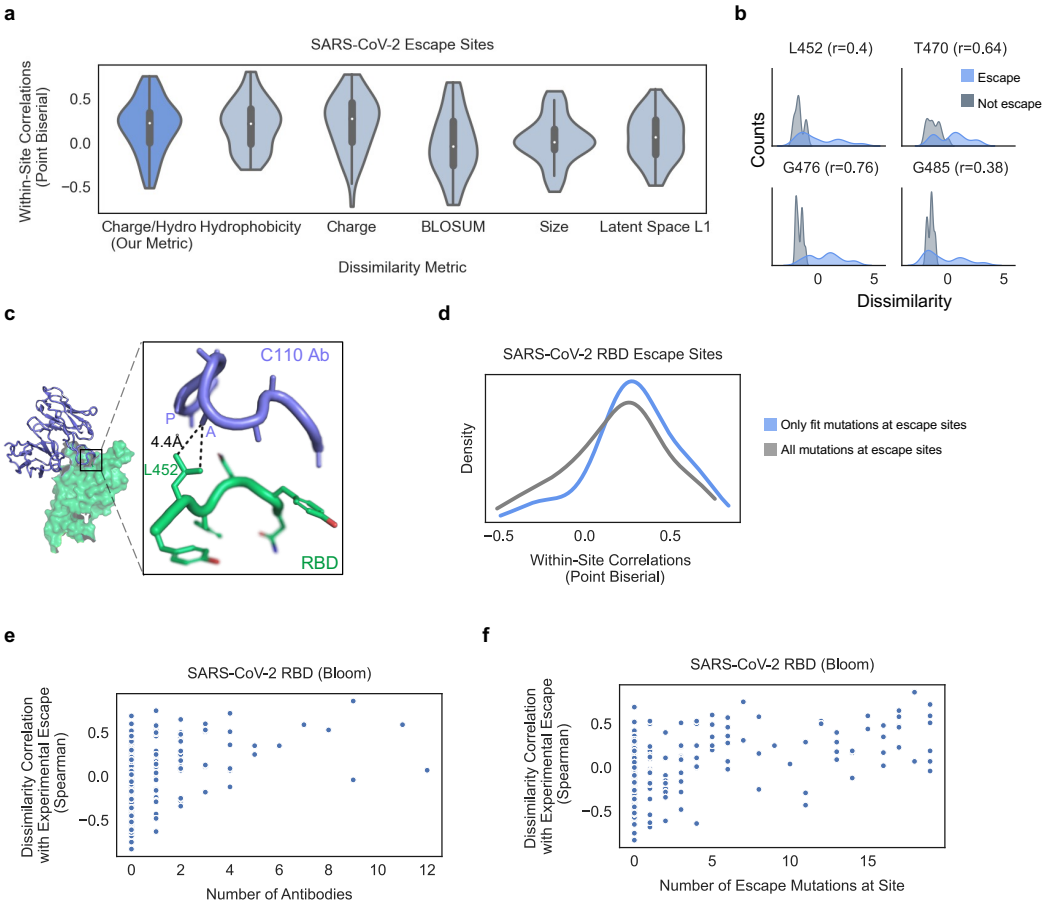

Figure S5: Charge-hydrophobicity metric captures residue dissimilarity relevant for loss of antibody binding. a) Within-site point biserial correlations between residue dissimilarity metrics and SARS-CoV-2 DMS escape data at escape sites (sites with 3-17 escape mutations). More sites have a higher correlation for our charge-hydrophobicity metric than charge or hydrophobicity alone, BLOSUM62, residue size, or EVE latent space (L1) distance. b) Charge-hydrophobicity dissimilarity performance in key sites c) The L452 RBD site is an example of decrease in hydrophobicity displacing the proximal alanine in the RBD C110 antibody interaction. (PDB: 7K8V) d) Within-site correlations at RBD escape sites increase when considering only mutations where fitness is maintained (passes Bloom lab's RBD expression and ACE2 binding cutoffs) e) Within-site correlations between residue dissimilarity and escape increase when more antibodies have escape mutations at that site. f) Within-site correlations between residue dissimilarity and escape increase when more mutations escape at site (and there can be no correlation with binarized escape when every mutation escapes).

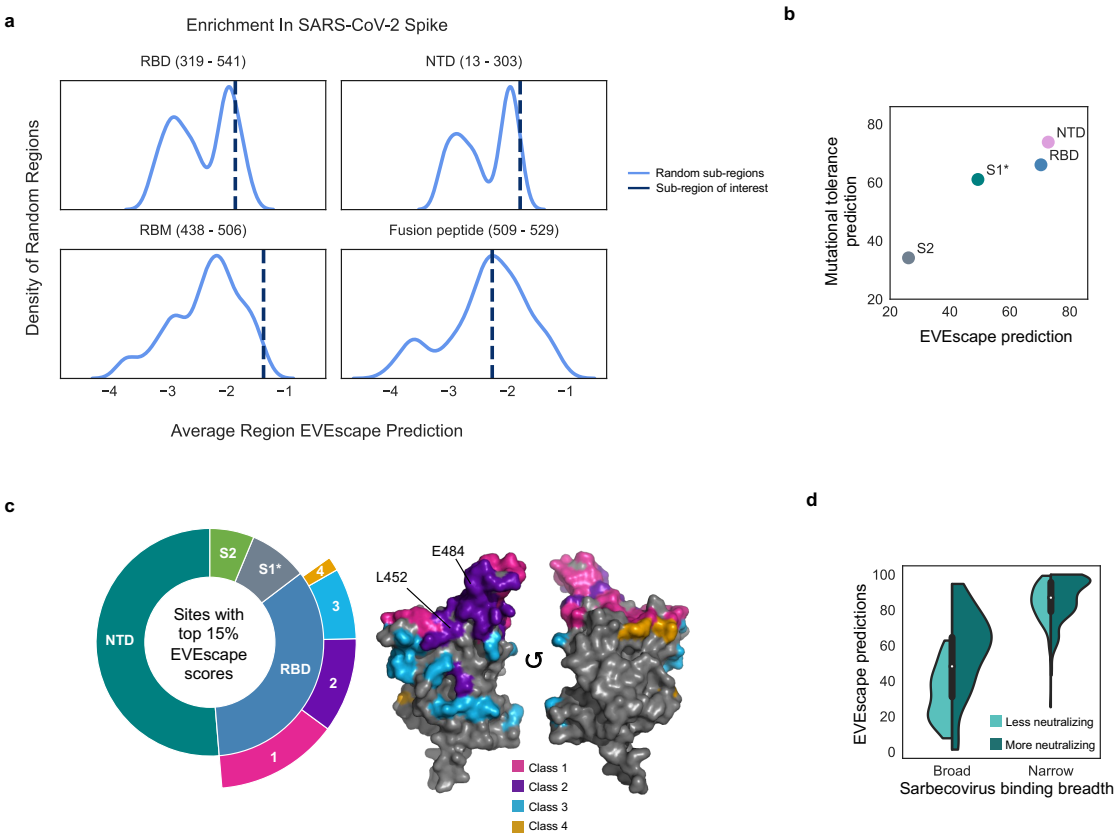

Figure S6: EVEscape enrichment in regions of SARS-CoV-2 Spike. a) RBD (particularly receptor binding motif (RBM)) and N-terminal domain (NTD) have significantly enriched average EVEscape scores, relative to a distribution of 500 random contiguous regions of the same length from full Spike. b) Average region EVEscape predictions are highest in RBD and NTD, though NTD is more mutationally tolerant with a higher average region EVE fitness score. c) EVEscape predictions cover diverse epitope regions across Spike and diverse RBD antibody classes [3] (3D structure of RBD on the right), including known immunodominant sites (E484, K417, L452) (PDB ID: 7BNN). The regions considered are NTD (sequence positions 14 - 306), RBD (319 - 542), S1* (543 - 685), and S2 (686 - 1273), where S1* refers to the region in S1 between RBD and S2. d) EVEscape scores experimental escape mutants from narrow antibodies and broad neutralizing antibodies higher than those from broad, non-neutralizing antibodies. Sarbecovirus binding breath and neutralization from Starr et al. [66]

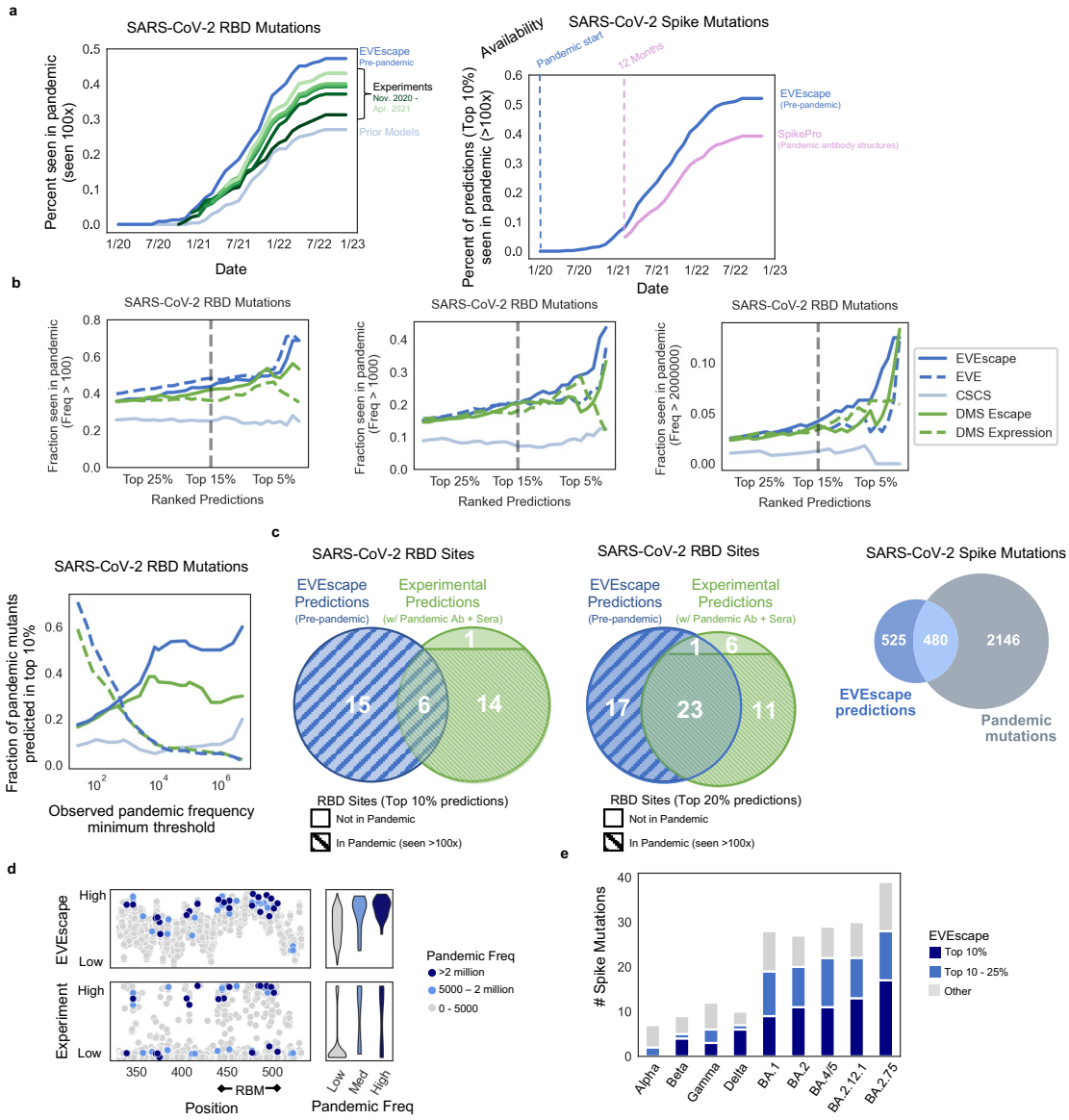

Figure S7: EVEscape as accurate as experimental scans at anticipating pandemic variation: retrospective analysis. a) Fraction of RBD predictions in top 15% of EVEscape, DMS experiments (Bloom Set, Table S4), and prior models52 seen by each date over 100 times in GISAID (left). DMS experiments are separated into which studies were available by each starting date. EVEscape predictions for for full Spike and prior SpikePro model [60] (right). b) Fraction of mutations seen 100, 1000, or 2 million times over different thresholds of top ranked predictions (Top) and share of predicted escape mutations in top decile of prediction based on their observed frequency (Bottom). c) Venn diagram of RBD sites seen in the top 10% (left) or top 20% (middle) of EVEscape and all DMS experimental predictions (Bloom Set Table S4), with markings for whether the site was seen over 100 times in GISAID over the full pandemic. Venn diagram of full Spike sites seen in top 10% of EVEscape and seen >100 times over the full pandemic (right). d) Comparison of EVEscape computational model predictions (top panel, y axis EVEscape score) and DMS experimental predictions (bottom panel, y axis experimental score) to frequency of mutations. e) The majority of Spike mutations in VOC strains have high EVEscape scores.

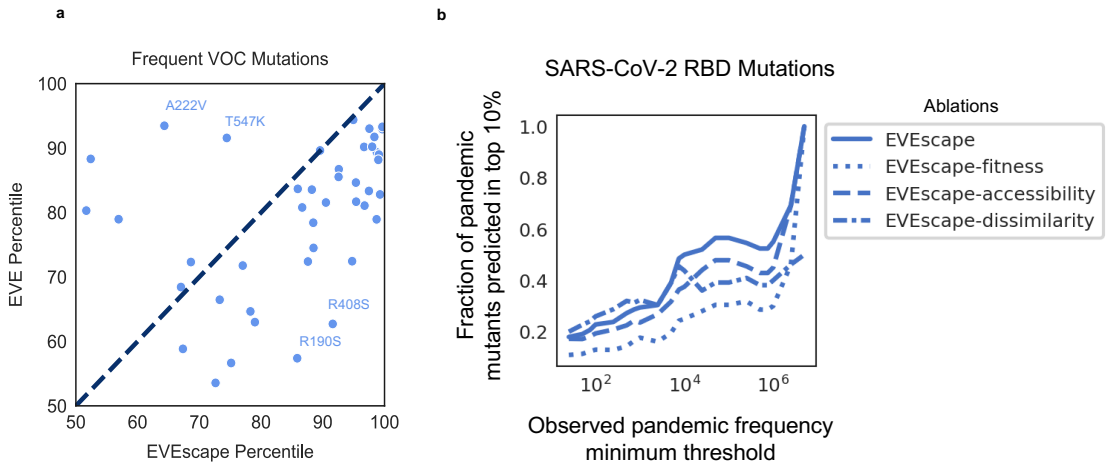

Figure S8: The role of EVEscape components in capturing pandemic variant mutations. a) EVEscape is more predictive than EVE alone at capturing frequent VOC mutations in full Spike. VOC mutations with high EVE scores and lower EVEscape scores (i.e., A222V and T547K) are known to impact structure and to not escape sera neutralization. Mutations with the highest EVEscape but low EVE scores (i.e., R190S and R408S) are in hydrophobic pockets that may promote antibody binding [2]. b) EVEscape is more predictive of high-frequency pandemic mutations than ablations of any of its 3 components. Notably, the ablation of the dissimilarity term leads to similar performance at identifying low-frequency mutations, but inferior performance at identifying high-frequency mutations.

Clinical therapeutic antibody	Pre-pandemic forecasting of mutations in epitopes
Bamlanivimab	T470R, T470K, **S494R**, Q493R, **Q493L**, **Q493K**, **Q493H**, **G485R**, G485E, G482R, G482D, **E484V**, E484Q, E484K, E484G, E484A
Sotrovimab	K444T, K444N, K444M, K444I, K444E, G485R, G485E, G446R, G446E, G446D, E484V, E484Q, E484K, E484G, E484A
Etesevimab	S443R, N440K, L441R, L441H
Imdevimab	R346W, R346T, R346S, R346P, R346M, R346L, R346I, R346G, R346C, Q498R, Q498K, N440K, L441R, L441H, **K444T**, **K444N**, **K444M**, **K444I**, **K444E**, **G446R**, G446E, **G446D**
Casirivimab	Q493R, **Q493L**, **Q493K**, **Q493H**, L455R, K417M, K417I, **K417E**, G485R, G485E, **E484V**, E484Q, E484K, E484G, E484A
Regdanvimab	S494R, Q493R, Q493L, Q493K, Q493H, L455R, L452R, K417M, K417I, K417E, E484V, E484Q, E484K, E484G, E484A

Pandemic mutations (Freq >10,000) colored & Experimental evidence **bolded**

Figure S9: Forecasting of clinical antibody epitope escape mutations. Forecasted mutations from the pre-pandemic model intersected with six clinical therapeutic monoclonal antibody epitopes. Epitopes are defined by sites within 3.5Å. Experimental evidence is from CoV-RDB [79]. All possible mutations are considered (not just those a nucleotide distance of one from Wuhan).

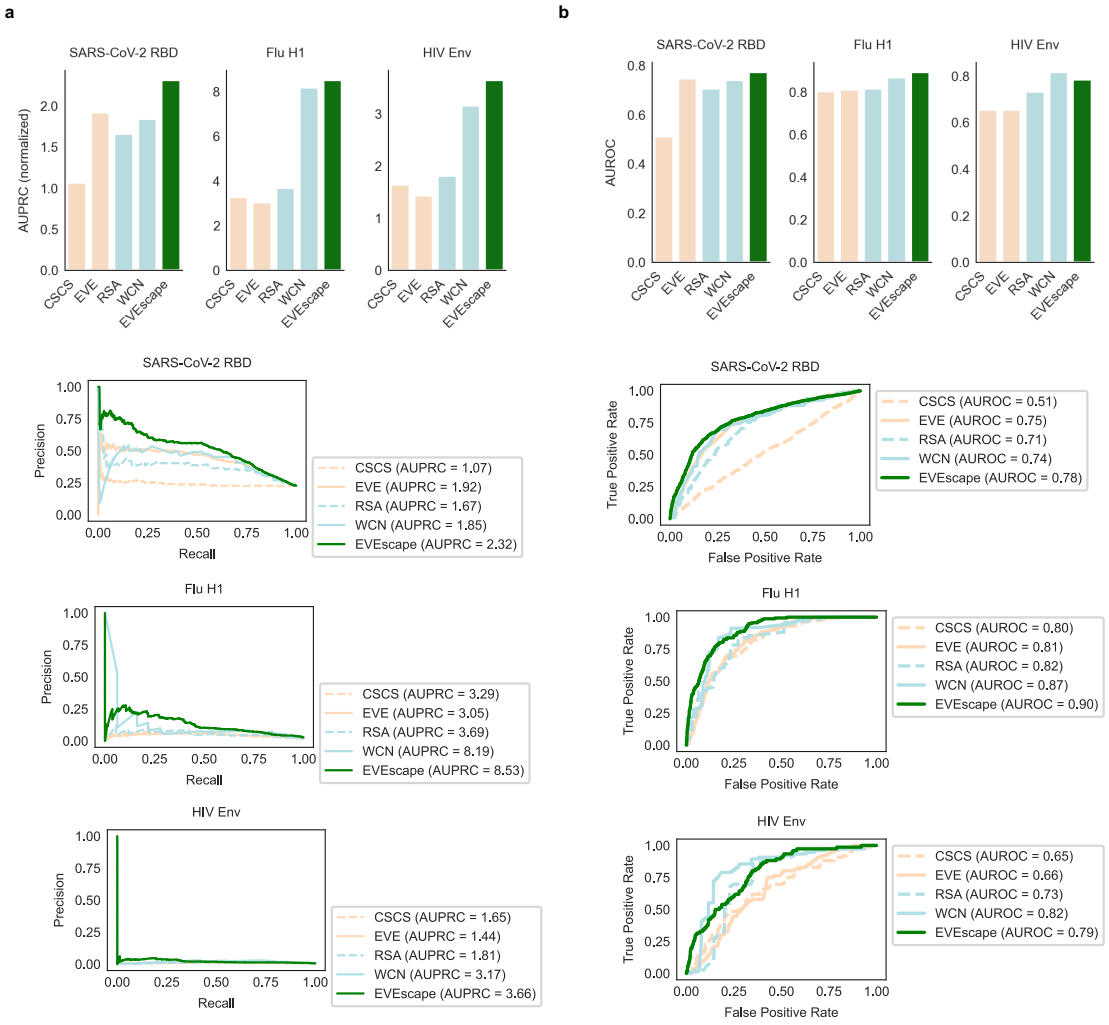

Figure S10: EVEscape performance on escape DMS data is generalizable across viruses. Precision-Recall (with AUPRC normalized by "null" model) (a) and AUROC (b) of predicting DMS escape mutations, for SARS-CoV-2 RBD, Flu H1, and HIV Env. Note: The "null" model AUPRC is equivalent to the fraction of observed escapes, and therefore AUPRC values are not comparable between viral proteins with different fractions of escape mutations (i.e. RBD and HIV Env). The fraction of observed escapes in the DMS experiments are 0.19 for RBD, for 0.015 for Flu, and 0.006 for HIV – Flu and HIV data examined far fewer antibody and sera samples (Table S5).

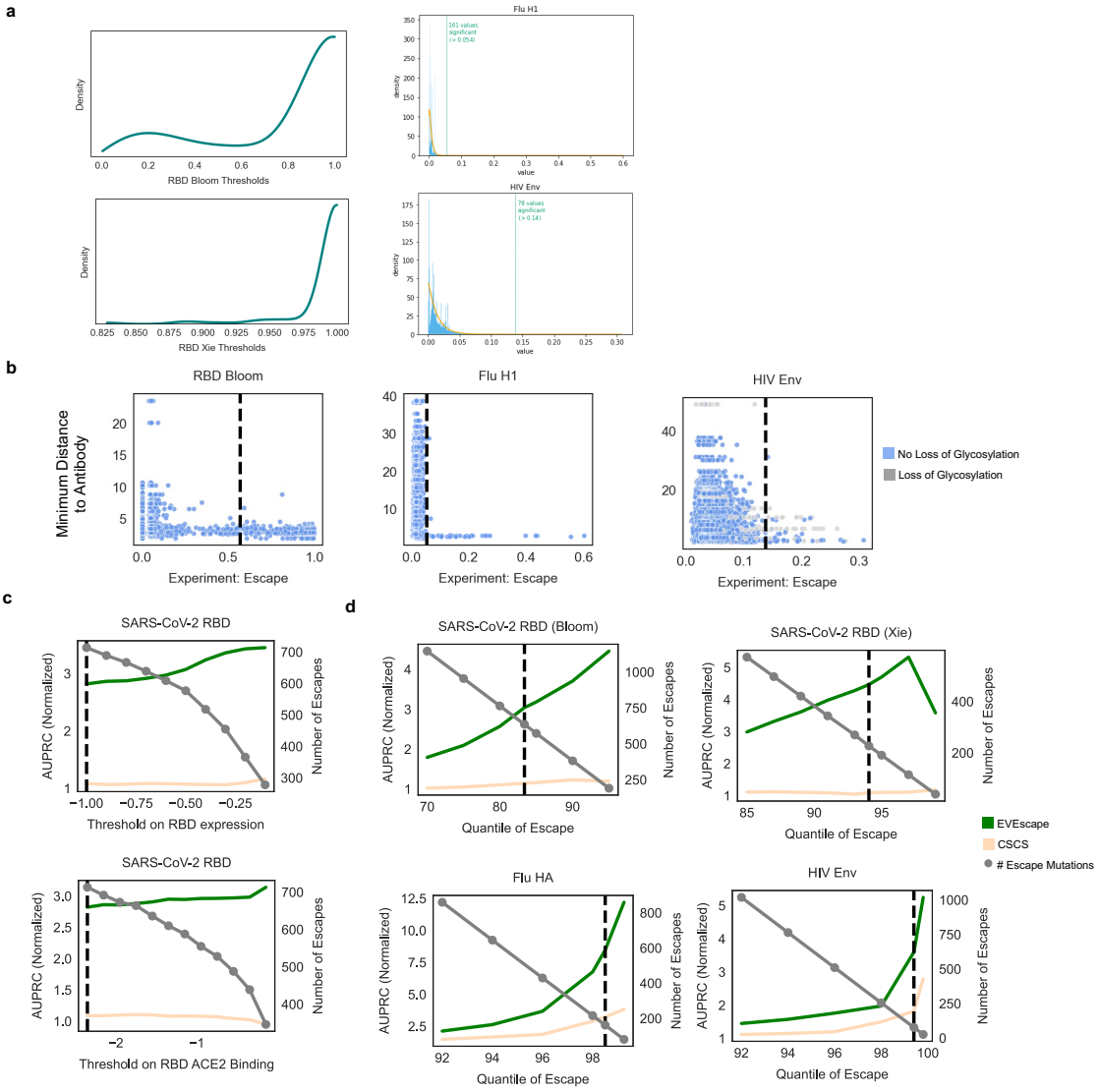

Figure S11: EVEscape performance is robust across data thresholds. a) Distribution of escape thresholds from bootstrapping 8 antibodies 1000 times and fitting a gamma distribution to each sample for Bloom and Xie RBD escape data (left) and gamma distributions to select Flu and HIV escape thresholds (right). b) Maximum escape values (over set of antibodies with PDB structures) for each mutation vs. the minimum distance to an antibody—most escape mutations (to the right of dashed line) are to residues with atoms to within 5Å of any residue on the antibody. For HIV, this is true for the mutations that do not involve loss of glycosylation. c) Impact of choice of RBD expression and ACE2 binding thresholds (dashed line uses thresholds chosen by Bloom escape papers and our paper) on AUPRC (normalized by "null" model – fraction of observed escapes) and number of mutations considered as escape. d) Impact of choice of escape threshold on RBD (Bloom and Xie data separated), Flu, and HIV AUPRC (normalized) and number of escape mutations (dashed line uses escape threshold chosen by our paper).

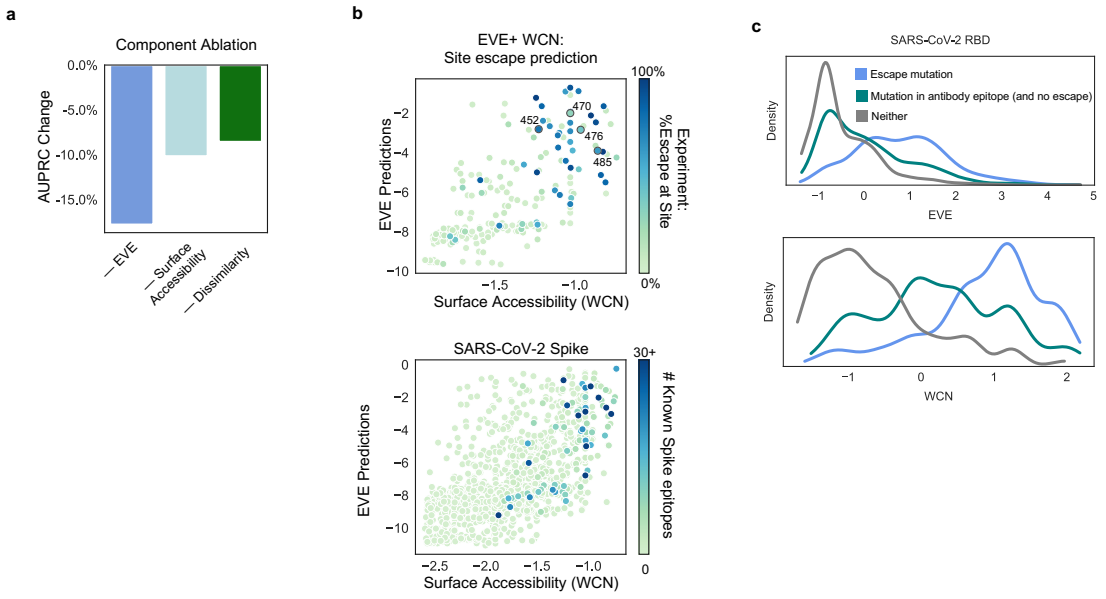

Figure S12: Surface accessibility metrics, mutation effect models, and dissimilarity provide complementary information for predicting antibody epitopes and escape mutations. a) All features of EVEscape contribute to performance in predicting RBD escape mutants. b) Sites with either high WCN accessibility or high EVE fitness predictions have a greater percent of escape mutants (upper). WCN and EVE predictions provide similar information about the location of Spike epitopes as represented in antibody-Spike crystal structures in RCSB PDB (lower). c) Density of standard-scaled EVEscape components differ for SARS-CoV-2 RBD escape (and antibody epitopes) and non-escape mutations.

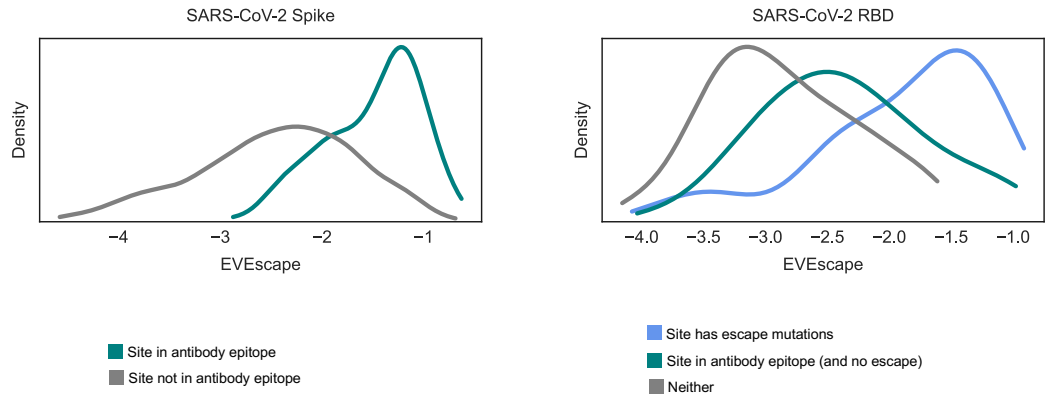

Figure S13: Top EVEscape predicted sites are known escape and in antibody footprints. Density of site-averaged EVEscape for SARS-CoV-2 full Spike (left) and RBD (right) shows success of EVEscape at distinguishing sites with observed escape mutations, as well as sites in known antibody epitopes, from sites with no evidence of antibody binding or escape. All but 2 sites in the top 20% of EVEscape scores are in known antibody footprints or have escape mutations in experiments.

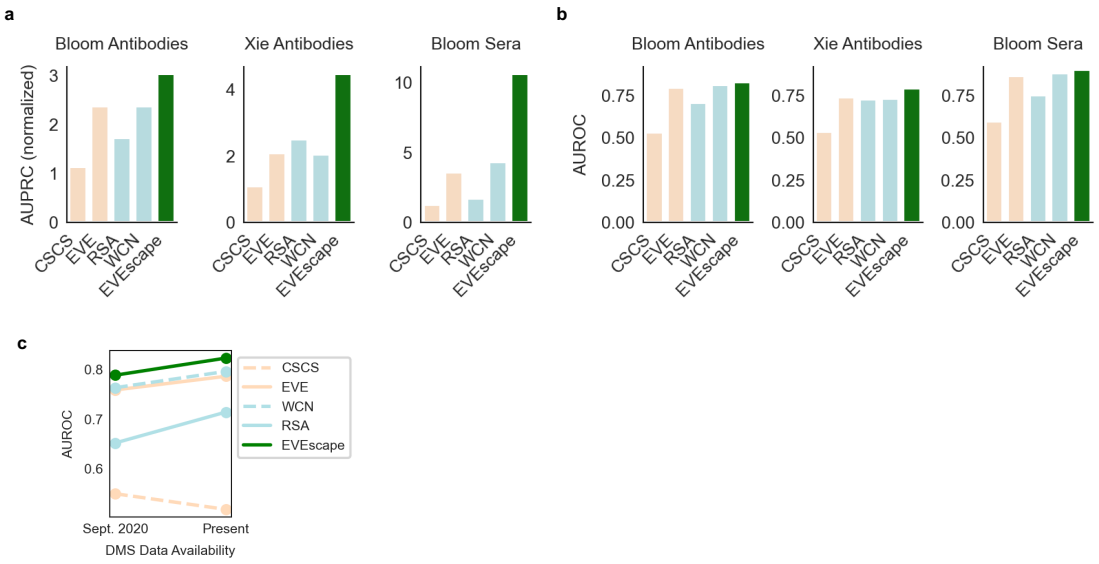

Figure S14: EVEscape RBD performance is robust to antibody and sera samples and improves with more available data for validation. Precision-Recall (with AUPRC normalized by "null" model) (a) and AUROC (b) of predicting RBD DMS escape mutations, for Bloom and Xie antibodies and Bloom sera. c) Comparison of model performance (AUROC) between data from first escape DMS study (10 antibodies – Sept. 2020)[35] and data available at present (338 antibodies, 55 sera samples).

Note: The "null" model AUPRC is equivalent to the fraction of observed escapes, and therefore AUPRC values are not comparable between data samples with different fractions of escape mutations (i.e., Bloom sera vs. Bloom antibodies, Table S5). The fraction of observed escapes in the DMS experiments are 0.17 for Bloom Ab, 0.06 for Xie Ab, and 0.003 for Bloom sera.

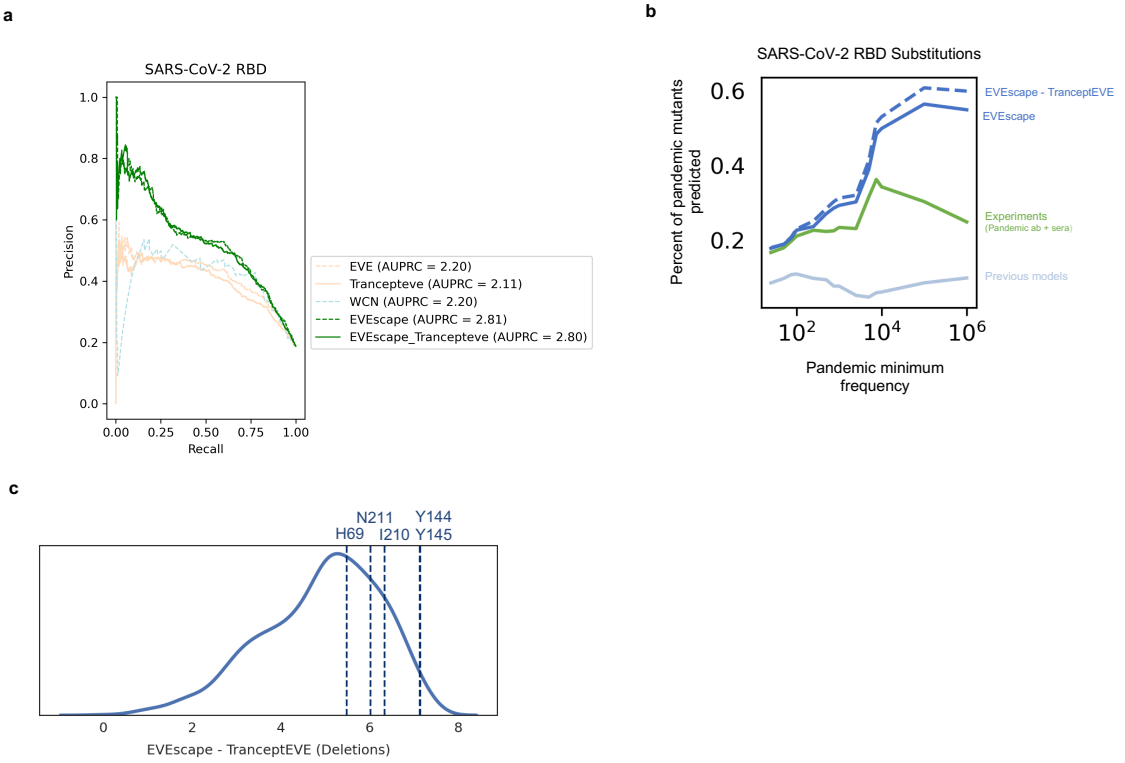

Figure S15: EVEscape adapts to new transformer model of mutation fitness capable of scoring indels. The EVEscape fitness component can be substituted with a new generative model, Trancept-EVE59 that is capable of scoring substitutions as well as insertions and deletions. a) EVEscape using TranceptEVE as the fitness model performs equivalently to EVEscape using EVE at predicting substitutions from from deep mutational scans that escape antibody binding. b) Percent of predicted substitutions in top decile of prediction based on their observed frequency during the pandemic shows EVEscape with TranceptEVE is just as good as, or better than, EVEscape using EVE at predicting pandemic substitutions. c) Histogram of EVEscape scores with TranceptEVE as a fitness model for all single deletions to Spike. Single deletions seen in the pandemic more than 1000 times are predicted higher than most other single deletions, especially the very frequent pandemic deletion Y144- (seen more than a million times).

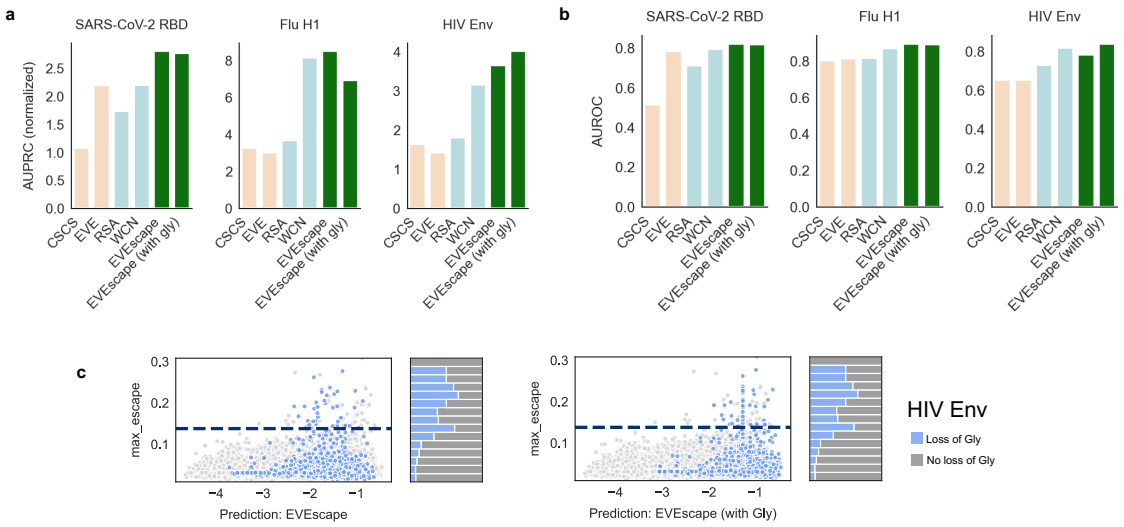

Figure S16: Incorporating glycosylation in EVEscape improves performance on HIV Env. Precision-Recall (with AUPRC normalized by "null" model – fraction of observed escapes) (a) and AUROC (b) of EVEscape and EVEscape+Gly predicting DMS escape mutations for SARS-CoV-2 RBD, Flu H1, and HIV Env. c) Scatterplot of HIV Env maximum escape at each mutation vs. EVEscape predictions with and without glycosylation. Hue indicates mutations that cause cause loss of glycosylation. The majority of HIV Env escape mutations involve glycosylation loss, and EVEscape+Gly performs better on these mutations.
Note: In the limited HIV Env dataset examining 8 antibodies, 50% of all escape mutations are likely due to removal of a glycan17. The effects of glycosylation changes may not be reflected in the SARS-CoV-2 Spike experiments as these experiments were conducted in a yeast system with different surface glycan types8. While SARS-CoV-2 Spike (22 glycosylation sites) and Flu H1 (up to 11 glycosylation sites) are much less extensively glycosylated than HIV Env (up to 30 glycosylation sites), some glycosylation changes in these proteins facilitate escape [51, 82, 16, 47].

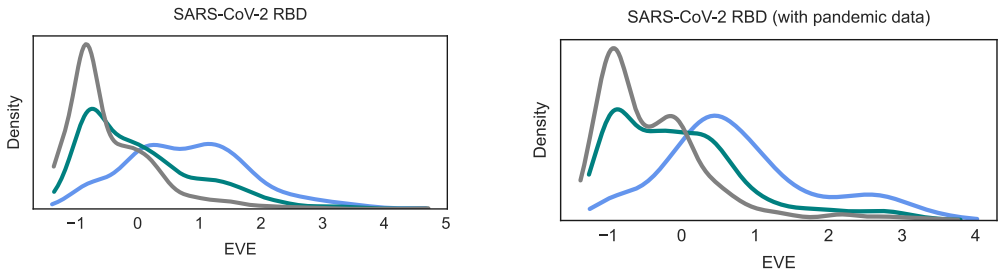

Figure S17: Incorporating pandemic data into EVE improves prediction of escape DMS. Incorporating pandemic sequences in EVE training data results in a greater distinction between escape and non-escape mutations with high EVE scores.

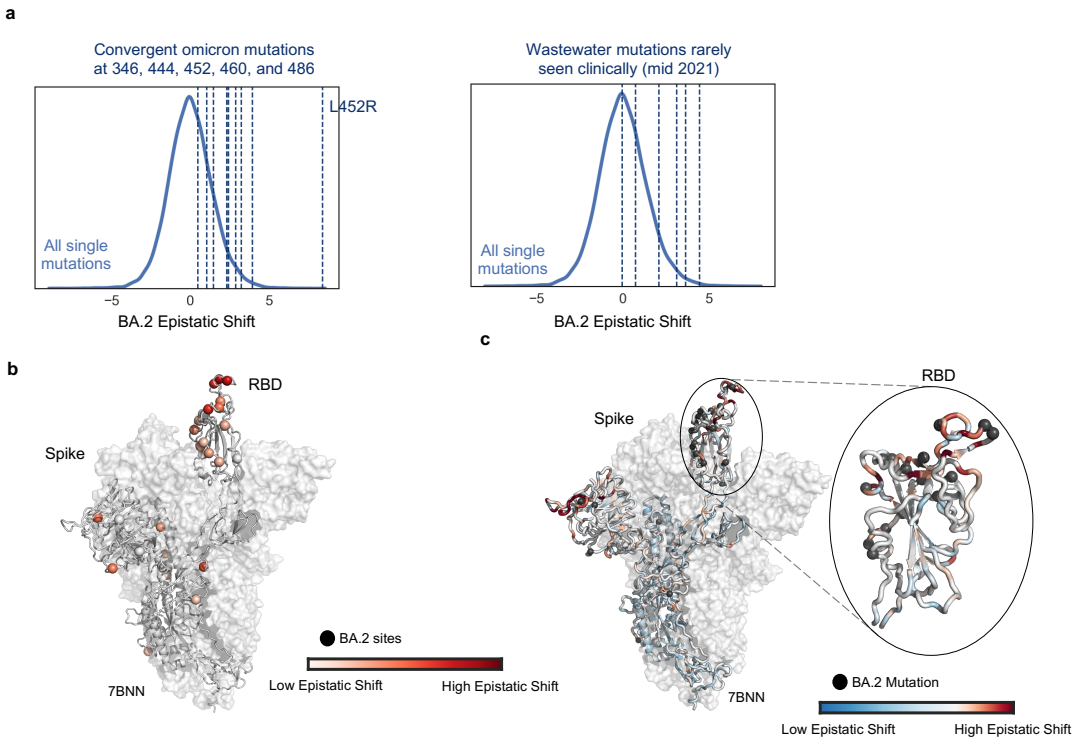

Figure S18: EVEscape captures the epistatic shift between Wuhan and BA.2. a) Histogram of epistatic shift values between Wuhan and BA.2 EVE models for all single mutations, calculated as linear regression residuals. Convergent mutations that arise multiple times in Omicron lineages (mutations at sites 346, 444, 452, 460, and 486) highlighted on the left. Wastewater mutations seen mid-2021 [65] that were rarely seen clinically in patients, and so likely epistatic (right). b) Max epistatic shift magnitudes of mutations at sites in BA.2 shows high epistatic shifts concentrated in RBD. c) Large epistatic shifts for mutations on Wuhan and BA.2 strains concentrated at sites proximal to BA.2 mutations.

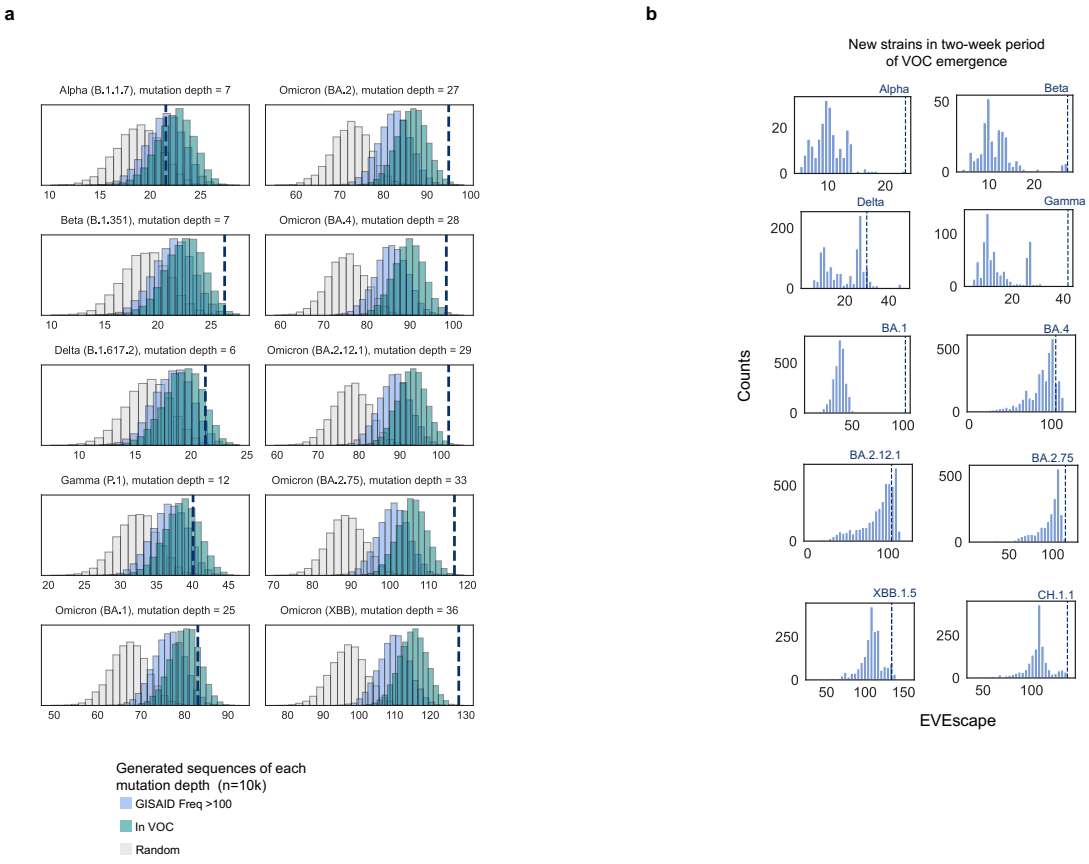

Figure S19: EVEscape strains. a) VOCs have high EVEscape scores compared to random mutations at the same mutation depth, particularly Beta and later Omicron strains. b) VOCs are among the highest scoring new strains for their two-week period of emergence using a pre-pandemic EVEscape model.

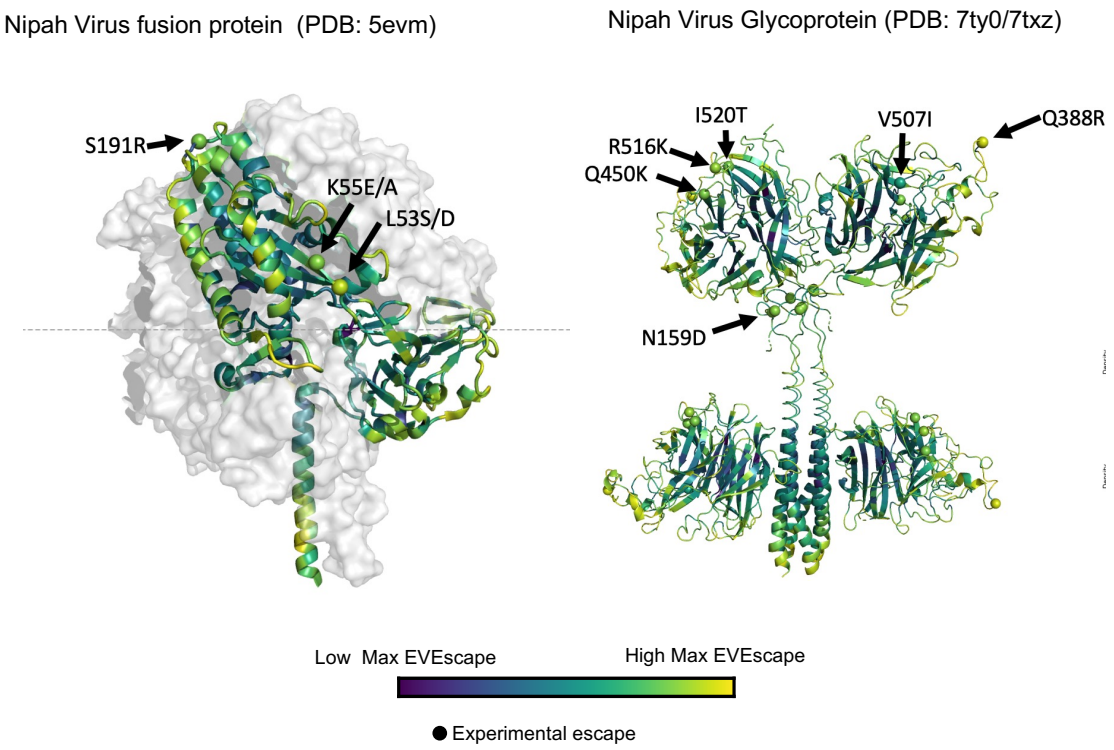

Figure S20: EVEscape predictions for potential pandemics. Site-maximum EVEscape scores on Nipah Virus fusion protein (left) and Glycoprotein (right) structures depict regions of high EVEscape scores and known escape mutations with experimental evidence [6, 81, 85, 14, 15] (little is known for this understudied virus with pandemic potential) are highlighted with spheres.

Appendix C

Supplementary Tables

	RBD (+pandemic)	**RBD**	Spike (+pandemic)	**Spike**
SARS-CoV-2	1394	1	1292	1
SARS-CoV-1	37	24	27	23
Other SARS-like	110	101	108	99
MERS	305	265	293	259
Betacoronavirus 1 (OC43)	577	416	504	394
Alphacoronavirus 1	0	0	240	175
229E	0	0	131	95
NL63	0	0	52	47
HKU1	57	27	54	27
HKU15	0	0	212	141
Avian coronavirus	0	0	668	581
Porcine epidemic diarrhea virus	0	0	1800	1440
Other coronavirus	252	175	486	347
Other/unknown	0	0	1	0
Total	**2732**	**1009**	**5868**	**3629**

Table S1: Taxa of sequences in Spike and RBD training alignments. RBD and Spike without pandemic data are the primary alignments used.

Protein	# Seqs	Seq Length	% Coverage	N_eff
Flu H1	71463	565	97.0%	12238.5
HIV Env	109050	690	97.2%	48082.1
SARS-CoV-2 RBD (+pandemic)	2732	221	98.6%	277.7
SARS-CoV-2 RBD	1009	221	98.6%	195.4
SARS-CoV-2 Spike (+pandemic)	5868	1273	98.9%	1639.1
SARS-CoV-2 Spike	3629	1273	94.7%	1345.6
Lassa Glycoprotein	1093	491	99.8%	536.2
Nipah Glycoprotein	5036	602	88.7%	1328.1
Nipah Fusion Protein	6155	546	94.1%	1105.6

Table S2: EVE training alignment summary statistics.

Virus	Protein	Study	Strain	Assay variable	N	Ind	EVH	EVE
Flu	H1	Doud 2016 [20]	A/WSN/1933	replication	10317	0.45	0.45	**0.53**
		Wu 2020 [84]	H1 (strain)	replication	10317	0.36	**0.37**	0.36
HIV	Env	Haddox 2018 [36]	BG505	replication	12388	**0.48**	0.41	**0.48**
			BF520	replication	12502	0.48	0.43	**0.49**
		Roop 2020 [62]	BG505	replication (human)	12483	0.48	0.44	**0.49**
				replication (rhesus)	12483	0.43	0.40	**0.44**
		Duenas-Decamp 2016 [22]	BG505	replication	375	0.37	**0.42**	0.38
SARS2	Spike RBD	Starr 2020[70]	Wuhan-Hu-1	yeast expression (RBD)	3798	0.36	0.33	**0.45**
				ACE2 binding	3802	0.23	0.16	**0.26**
		Chan 2021[11]	Wuhan-Hu-1	human cell expression	3458	0.33	0.32	**0.45**
				ACE2 binding	3458	0.31	0.30	**0.42**
	Mpro	Flynn 2022[25]	Wuhan-Hu-1	yeast growth	5741	0.58	**0.60**	**0.60**

Table S3: Experimental details and EVE, EVmutation, and independent model performance (spearman correlations) for DMS fitness experiments.

	PDB ID	Description
SARS-CoV-2 Spike	6VXX	Spike (closed state)
	6VYB	Spike (open state)
	7CAB	Spike (closed state with higher sequence coverage)
	7BNN	Spike (open state with higher sequence coverage)
Flu H1	1RVX	1934 H1 Hemagglutinin (similar to Bloom DMS sequence)
HIV Env	5FYL	BG505 SOSIP.664 Env (prefusion) [18]
	7TFO	BG505 SOSIP.664 Env (CD4-bound open state)
Lassa Glycoprotein	7PUY	Lassa Virus Josiah Strain Glycoprotein
Nipah Fusion Protein	5EVM	Nipah Fusion Protein (prefusion)
Nipah Glycoprotein	7TY0	Nipah Glycoprotein Malaysian Strain
	7TXZ	Nipah Glycoprotein Malaysian Strain

Table S4: PDB structures capturing diverse protein conformations used for accessibility calculations.

	Papers	Assay Details	#Muts	#Escape Muts	# Ab/Sera
SARS2 RBD (Wuhan-Hu-1)	Bloom Lab: Dong 2021 [19] Greaney 2021 [35] Greaney 2021 [32] Greaney 2021 [31] Starr 2021 [67] Starr 2021 [66] Tortorici 2021 [78] Starr 2021 [68]	FACS-based yeast display, antibody binding	3819	635	36
	Bloom Lab: Greaney 2021 [32] Greaney 2021 [30] Greaney 2021 [31]	FACS-based yeast display, sera binding	3819	15	55
	Xie Lab: Cao 2022 [8]	MACS-based yeast display, antibody binding	3819	227	247
Flu H1 (A/WSN/1933)	Doud 2018 [21]	Viral cell entry with antibodies	10735	161	6
HIV Env (BG505)	Dingens 2019 [18]	Viral cell entry with antibodies	12730	76	8

Table S5: Escape DMS data used for EVEscape validation.

Bibliography

[1] Fatima Amanat, Mahima Thapa, Tinting Lei, Shaza M Sayed Ahmed, Daniel C Adelsberg, Juan Manuel Carreño, Shirin Strohmeier, Aaron J Schmitz, Sarah Zafar, Julian Q Zhou, Willemijn Rijnink, Hala Alshammary, Nicholas Borcherding, Ana Gonzalez Reiche, Komal Srivastava, Emilia Mia Sordillo, Harm van Bakel, Personalized Virology Initiative, Jackson S Turner, Goran Bajic, Viviana Simon, Ali H Ellebedy, and Florian Krammer. SARS-CoV-2 mRNA vaccination induces functionally diverse antibodies to NTD, RBD, and S2. *Cell*, 184(15):3936–3948.e10, July 2021.

[2] Sandhya Bangaru, Gabriel Ozorowski, Hannah L Turner, Aleksandar Antanasijevic, Deli Huang, Xiaoning Wang, Jonathan L Torres, Jolene K Diedrich, Jing-Hui Tian, Alyse D Portnoff, Nita Patel, Michael J Massare, John R Yates, 3rd, David Nemazee, James C Paulson, Greg Glenn, Gale Smith, and Andrew B Ward. Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. *Science*, 370(6520):1089–1094, November 2020.

[3] Christopher O Barnes, Claudia A Jette, Morgan E Abernathy, Kim-Marie A Dam, Shannon R Esswein, Harry B Gristick, Andrey G Malyutin, Naima G Sharaf, Kathryn E Huey-Tubman, Yu E Lee, Davide F Robbiani, Michel C Nussenzweig, Anthony P West, Jr, and Pamela J Bjorkman. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, 588(7839):682–687, December 2020.

[4] Karim Beguir, Marcin J Skwark, Yunguan Fu, Thomas Pierrot, Nicolas Lopez Carranza, Alexandre Laterre, Ibtissem Kadri, Abir Korched, Anna U Lowegard, Bonny Gaby Lui, Bianca Sänger, Yunpeng Liu, Asaf Poran, Alexander Muik, and Uğur Şahin. Early computational detection of potential high-risk SARS-CoV-2 variants. *Comput. Biol. Med.*, 155(106618):106618, March 2023.

[5] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000.

[6] Viktoriya Borisevich, Benhur Lee, Andrew Hickey, Blair DeBuysscher, Christopher C Broder, Heinz Feldmann, and Barry Rockx. Escape from monoclonal antibody neutralization affects henipavirus fitness in vitro and in vivo. *J. Infect. Dis.*, 213(3):448–455, February 2016.

[7] Tierra K Buck, Adrian S Enriquez, Sharon L Schendel, Michelle A Zandonatti, Stephanie S Harkins, Haoyang Li, Alex Moon-Walker, James E Robinson, Luis M Branco, Robert F Garry, Erica Ollmann Saphire, and Kathryn M Hastie. Neutralizing antibodies against lassa virus lineage I. *MBio*, 13(4):e0127822, August 2022.

[8] Yunlong Cao, Jing Wang, Fanchong Jian, Tianhe Xiao, Weiliang Song, Ayijiang Yisimayi, Weijin Huang, Qianqian Li, Peng Wang, Ran An, Jing Wang, Yao Wang, Xiao Niu, Sijie Yang, Hui Liang, Haiyan Sun, Tao Li, Yuanling Yu, Qianqian Cui, Shuo Liu, Xiaodong Yang, Shuo Du, Zhiying Zhang, Xiaohua Hao, Fei Shao, Ronghua Jin, Xiangxi Wang, Junyu Xiao, Youchun Wang, and Xiaoliang Sunney Xie. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature*, 602(7898):657–663, February 2022.

[9] Yunlong Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, Shuo Du, Jing Wang, Qianqian Li, Xiaosu Chen, Yuanling Yu, Peng Wang, Zhiying Zhang, Pulan Liu, Ran An, Xiaohua Hao, Yao Wang, Jing Wang, Rui Feng, Haiyan Sun, Lijuan Zhao, Wen Zhang, Dong Zhao, Jiang Zheng, Lingling Yu, Can Li, Na Zhang, Rui Wang, Xiao Niu, Sijie Yang, Xuetao Song, Yangyang Chai, Ye Hu, Yansong Shi, Linlin Zheng, Zhiqiang Li, Qingqing Gu, Fei Shao, Weijin Huang, Ronghua Jin, Zhongyang Shen, Youchun Wang, Xiangxi Wang, Junyu Xiao, and Xiaoliang Sunney Xie. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by omicron infection. *Nature*, 608(7923):593–602, August 2022.

[10] Gabriele Cerutti, Yicheng Guo, Tongqing Zhou, Jason Gorman, Myungjin Lee, Micah Rapp, Eswar R Reddem, Jian Yu, Fabiana Bahna, Jude Bimela, Yaoxing Huang, Phinikoula S Katsamba, Lihong Liu, Manoj S Nair, Reda Rawi, Adam S Olia, Pengfei Wang, Baoshan Zhang, Gwo-Yu Chuang, David D Ho, Zizhang Sheng, Peter D Kwong, and Lawrence Shapiro. Potent SARS-CoV-2 neutralizing antibodies directed against spike n-terminal domain target a single supersite. *Cell Host Microbe*, 29(5):819–833.e7, May 2021.

[11] Kui K Chan, Timothy J C Tan, Krishna K Narayanan, and Erik Procko. An engineered decoy receptor for SARS-CoV-2 broadly binds protein S sequence variants. *Sci Adv*, 7(8), February 2021.

[12] Albert Tian Chen, Kevin Altschuler, Shing Hei Zhan, Yujia Alina Chan, and Benjamin E Deverman. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *Elife*, 10, February 2021.

[13] C Chothia and J Janin. Principles of protein-protein recognition. *Nature*, 256(5520):705–708, August 1975.

[14] Ha V Dang, Yee-Peng Chan, Young-Jun Park, Joost Snijder, Sofia Cheliout Da Silva, Bang Vu, Lianying Yan, Yan-Ru Feng, Barry Rockx, Thomas W Geisbert, Chad E Mire, Christopher C Broder, and David Veesler. An antibody

against the F glycoprotein inhibits nipah and hendra virus infections. *Nat. Struct. Mol. Biol.*, 26(10):980–987, October 2019.

[15] Ha V Dang, Robert W Cross, Viktoriya Borisevich, Zachary A Bornholdt, Brandyn R West, Yee-Peng Chan, Chad E Mire, Sofia Cheliout Da Silva, Antony S Dimitrov, Lianying Yan, Moushimi Amaya, Chanakha K Navaratnarajah, Larry Zeitlin, Thomas W Geisbert, Christopher C Broder, and David Veesler. Broadly neutralizing antibody cocktails targeting nipah virus and hendra virus fusion glycoproteins. *Nat. Struct. Mol. Biol.*, 28(5):426–434, May 2021.

[16] Suman R Das, Scott E Hensley, Alexandre David, Loren Schmidt, James S Gibbs, Pere Puigbò, William L Ince, Jack R Bennink, and Jonathan W Yewdell. Fitness costs limit influenza a virus hemagglutinin glycosylation as an immune evasion strategy. *Proc. Natl. Acad. Sci. U. S. A.*, 108(51):E1417–22, December 2011.

[17] Adam S Dingens, Priyamvada Acharya, Hugh K Haddox, Reda Rawi, Kai Xu, Gwo-Yu Chuang, Hui Wei, Baoshan Zhang, John R Mascola, Bridget Carragher, Clinton S Potter, Julie Overbaugh, Peter D Kwong, and Jesse D Bloom. Complete functional mapping of infection- and vaccine-elicited antibodies against the fusion peptide of HIV. *PLoS Pathog.*, 14(7):e1007159, July 2018.

[18] Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. An antigenic atlas of HIV-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity*, 50(2):520–532.e3, February 2019.

[19] Jinhui Dong, Seth J Zost, Allison J Greaney, Tyler N Starr, Adam S Dingens, Elaine C Chen, Rita E Chen, James Brett Case, Rachel E Sutton, Pavlo Gilchuk, Jessica Rodriguez, Erica Armstrong, Christopher Gainza, Rachel S Nargi, Elad Binshtein, Xuping Xie, Xianwen Zhang, Pei-Yong Shi, James Logue, Stuart Weston, Marisa E McGrath, Matthew B Frieman, Tyler Brady, Kevin M Tuffy, Helen Bright, Yueh-Ming Loo, Patrick M McTamney, Mark T Esser, Robert H Carnahan, Michael S Diamond, Jesse D Bloom, and James E Crowe, Jr. Genetic and structural basis for SARS-CoV-2 variant neutralization by a two-antibody cocktail. *Nat Microbiol*, 6(10):1233–1244, October 2021.

[20] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all Amino-Acid mutations on influenza hemagglutinin. *Viruses*, 8(6), June 2016.

[21] Michael B Doud, Juhye M Lee, and Jesse D Bloom. How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Nat. Commun.*, 9(1):1386, 2018.

[22] Maria Duenas-Decamp, Li Jiang, Daniel Bolon, and Paul R Clapham. Saturation mutagenesis of the HIV-1 envelope CD4 binding loop reveals residues controlling distinct trimer conformations. *PLoS Pathog.*, 12(11):e1005988, November 2016.

[23] Robert C Edgar. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.*, 13(1):6968, November 2022.

[24] D Eisenberg, R M Weiss, and T C Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U. S. A.*, 81(1):140–144, January 1984.

[25] Julia M Flynn, Neha Samant, Gily Schneider-Nachum, David T Barkan, Nese Kurt Yilmaz, Celia A Schiffer, Stephanie A Moquin, Dustin Dovala, and Daniel N A Bolon. Comprehensive fitness landscape of SARS-CoV-2 mpro reveals insights into viral resistance mechanisms. *Elife*, 11, June 2022.

[26] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021.

[27] Tiziana Ginex, Clara Marco-Marín, Miłosz Wieczór, Carlos P Mata, James Krieger, Paula Ruiz-Rodriguez, Maria Luisa López-Redondo, Clara Francés-Gómez, Roberto Melero, Carlos Óscar Sánchez-Sorzano, Marta Martínez, Nadine Gougeard, Alicia Forcada-Nadal, Sara Zamora-Caballero, Roberto Gozalbo-Rovira, Carla Sanz-Frasquet, Rocío Arranz, Jeronimo Bravo, Vicente Rubio, Alberto Marina, IBV-Covid19-Pipeline, Ron Geller, Iñaki Comas, Carmen Gil, Mireia Coscolla, Modesto Orozco, José Luis Llácer, and Jose-Maria Carazo. The structural role of SARS-CoV-2 genetic background in the emergence and success of spike mutations: The case of the spike A222V mutation. *PLoS Pathog.*, 18(7):e1010631, July 2022.

[28] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, 2:e00631, May 2013.

[29] Allison J Greaney, Rachel T Eguia, Tyler N Starr, Khadija Khan, Nicholas Franko, Jennifer K Logue, Sandra M Lord, Cate Speake, Helen Y Chu, Alex Sigal, and Jesse D Bloom. The SARS-CoV-2 delta variant induces an antibody response largely focused on class 1 and 2 antibody epitopes. *PLoS Pathog.*, 18(6):e1010592, June 2022.

[30] Allison J Greaney, Andrea N Loes, Katharine H D Crawford, Tyler N Starr, Keara D Malone, Helen Y Chu, and Jesse D Bloom. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe*, 29(3):463–476.e6, March 2021.

[31] Allison J Greaney, Andrea N Loes, Lauren E Gentles, Katharine H D Crawford, Tyler N Starr, Keara D Malone, Helen Y Chu, and Jesse D Bloom. 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(600), June 2021.

[32] Allison J Greaney, Tyler N Starr, Christopher O Barnes, Yiska Weisblum, Fabian Schmidt, Marina Caskey, Christian Gaebler, Alice Cho, Marianna Agudelo, Shlomo Finkin, Zijun Wang, Daniel Poston, Frauke Muecksch, Theodora Hatziioannou, Paul D Bieniasz, Davide F Robbiani, Michel C Nussenzweig, Pamela J Bjorkman, and Jesse D Bloom. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies, 2021.

[33] Allison J Greaney, Tyler N Starr, and Jesse D Bloom. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evol.*, 8(1):veac021, May 2022.

[34] Allison J Greaney, Tyler N Starr, Rachel T Eguia, Andrea N Loes, Khadija Khan, Farina Karim, Sandile Cele, John E Bowen, Jennifer K Logue, Davide Corti, David Veesler, Helen Y Chu, Alex Sigal, and Jesse D Bloom. A SARS-CoV-2 variant elicits an antibody response with a shifted immunodominance hierarchy. *PLoS Pathog.*, 18(2):e1010248, February 2022.

[35] Allison J Greaney, Tyler N Starr, Pavlo Gilchuk, Seth J Zost, Elad Binshtein, Andrea N Loes, Sarah K Hilton, John Huddleston, Rachel Eguia, Katharine H D Crawford, Adam S Dingens, Rachel S Nargi, Rachel E Sutton, Naveenchandra Suryadevara, Paul W Rothlauf, Zhuoming Liu, Sean P J Whelan, Robert H Carnahan, James E Crowe, Jr, and Jesse D Bloom. Complete mapping of mutations to the SARS-CoV-2 spike Receptor-Binding domain that escape antibody recognition. *Cell Host Microbe*, 29(1):44–57.e9, January 2021.

[36] Hugh K Haddox, Adam S Dingens, Sarah K Hilton, Julie Overbaugh, and Jesse D Bloom. Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife*, 7:e34420, March 2018.

[37] Pernille Haste Andersen, Morten Nielsen, and Ole Lund. Prediction of residues in discontinuous b-cell epitopes using protein 3D structures. *Protein Sci.*, 15(11):2558–2567, November 2006.

[38] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89(22):10915–10919, November 1992.

[39] Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021.

[40] Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, February 2017.

[41] L Steven Johnson, Sean R Eddy, and Elon Portugaly. Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431, August 2010.

[42] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

[43] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael Tc Lee, Winston Yeo, Gisaid Core Curation Team, and Sebastian Maurer-Stroh. GISAID's role in pandemic response. *China CDC Wkly*, 3(49):1049–1051, December 2021.

[44] Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. December 2013.

[45] Kathryn E Kistler, John Huddleston, and Trevor Bedford. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe*, 30(4):545–555.e4, April 2022.

[46] Jens Vindahl Kringelum, Morten Nielsen, Søren Berg Padkjær, and Ole Lund. Structural analysis of b-cell epitopes in antibody:protein complexes. *Mol. Immunol.*, 53(1-2):24–34, January 2013.

[47] Yuqing Li, Dongqi Liu, Yating Wang, Wenquan Su, Gang Liu, and Weijie Dong. The importance of glycans of viral and host proteins in enveloped virus infection. *Front. Immunol.*, 12, April 2021.

[48] Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *Proteins*, 72(3):929–935, August 2008.

[49] M Cyrus Maher, Istvan Bartha, Steven Weaver, Julia di Iulio, Elena Ferri, Leah Soriaga, Florian A Lempp, Brian L Hie, Bryan Bryson, Bonnie Berger, David L Robertson, Gyorgy Snell, Davide Corti, Herbert W Virgin, Sergei L Kosakovsky Pond, and Amalio Telenti. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. June 2021.

[50] David C Montefiori. Measuring HIV neutralization in a luciferase reporter gene assay. *Methods Mol. Biol.*, 485:395–405, 2009.

[51] Penny L Moore, Elin S Gray, C Kurt Wibmer, Jinal N Bhiman, Molati Nonyane, Daniel J Sheward, Tandile Hermanus, Shringkhala Bajimaya, Nancy L Tumba, Melissa-Rose Abrahams, Bronwen E Lambson, Nthabeleng Ranchobe, Lihua Ping, Nobubelo Ngandu, Quarraisha Abdool Karim, Salim S Abdool Karim, Ronald I Swanstrom, Michael S Seaman, Carolyn Williamson, and Lynn Morris. Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. *Nat. Med.*, 18(11):1688–1692, November 2012.

[52] Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. May 2022.

[53] Pascal Notin, Lood Van Niekerk, Aaron W Kollasch, Daniel Ritter, Yarin Gal, and Debora S Marks. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. December 2022.

[54] J Novotný, M Handschumacher, E Haber, R E Bruccoleri, W B Carlson, D W Fanning, J A Smith, and G D Rose. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. U. S. A.*, 83(2):226–230, January 1986.

[55] Fritz Obermeyer, Martin Jankowiak, Nikolaos Barkas, Stephen F Schaffner, Jesse D Pyle, Leonid Yurkovetskiy, Matteo Bosso, Daniel J Park, Mehrtash Babadi, Bronwyn L MacInnis, Jeremy Luban, Pardis C Sabeti, and Jacob E Lemieux. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, 376(6599):1327–1332, June 2022.

[56] Luca Piccoli, Young-Jun Park, M Alejandra Tortorici, Nadine Czudnochowski, Alexandra C Walls, Martina Beltramello, Chiara Silacci-Fregni, Dora Pinto, Laura E Rosen, John E Bowen, Oliver J Acton, Stefano Jaconi, Barbara Guarino, Andrea Minola, Fabrizia Zatta, Nicole Sprugasci, Jessica Bassi, Alessia Peter, Anna De Marco, Jay C Nix, Federico Mele, Sandra Jovic, Blanca Fernandez Rodriguez, Sneha V Gupta, Feng Jin, Giovanni Piumatti, Giorgia Lo Presti, Alessandra Franzetti Pellanda, Maira Biggiogero, Maciej Tarkowski, Matteo S Pizzuto, Elisabetta Cameroni, Colin Havenar-Daughton, Megan Smithey, David Hong, Valentino Lepori, Emiliano Albanese, Alessandro Ceschi, Enos Bernasconi, Luigia Elzi, Paolo Ferrari, Christian Garzoni, Agostino Riva, Gyorgy Snell, Federica Sallusto, Katja Fink, Herbert W Virgin, Antonio Lanzavecchia, Davide Corti, and David Veesler. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell*, 183(4):1024–1042.e21, November 2020.

[57] Fabrizio Pucci and Marianne Rooman. Prediction and evolution of the molecular fitness of SARS-CoV-2 variants: Introducing SpikePro. *Viruses*, 13(5):935, May 2021.

[58] R Shyama Prasad Rao, Nagib Ahsan, Chunhui Xu, Lingtao Su, Jacob Verburgt, Luca Fornelli, Daisuke Kihara, and Dong Xu. Evolutionary dynamics of indels in SARS-CoV-2 spike glycoprotein. *Evol. Bioinform. Online*, 17:11769343211064616, December 2021.

[59] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018.

[60] Nash D Rochman, Guilhem Faure, Yuri I Wolf, Peter L Freddolino, Feng Zhang, and Eugene V Koonin. Epistasis at the SARS-CoV-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape. *MBio*, 13(2):e0013522, April 2022.

[61] Juan Rodriguez-Rivas, Giancarlo Croce, Maureen Muscat, and Martin Weigt. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. U. S. A.*, 119(4):e2113118119, January 2022.

[62] Jeremy I Roop, Noah A Cassidy, Adam S Dingens, Jesse D Bloom, and Julie Overbaugh. Identification of HIV-1 envelope mutations that enhance entry using macaque CD4 and CCR5. *Viruses*, 12(2), February 2020.

[63] B Rost and C Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20(3):216–226, November 1994.

[64] Fabian Schmidt, Yiska Weisblum, Frauke Muecksch, Hans-Heinrich Hoffmann, Eleftherios Michailidis, Julio C C Lorenzi, Pilar Mendoza, Magdalena Rutkowska, Eva Bednarski, Christian Gaebler, Marianna Agudelo, Alice Cho, Zijun Wang, Anna Gazumyan, Melissa Cipolla, Marina Caskey, Davide F Robbiani, Michel C Nussenzweig, Charles M Rice, Theodora Hatziioannou, and Paul D Bieniasz. Measuring SARS-CoV-2 neutralizing antibody activity using pseudotyped and chimeric viruses. *J. Exp. Med.*, 217(11), November 2020.

[65] Davida S Smyth, Monica Trujillo, Devon A Gregory, Kristen Cheung, Anna Gao, Maddie Graham, Yue Guan, Caitlyn Guldenpfennig, Irene Hoxie, Sherin Kannoly, Nanami Kubota, Terri D Lyddon, Michelle Markman, Clayton Rushford, Kaung Myat San, Geena Sompanya, Fabrizio Spagnolo, Reinier Suarez, Emma Teixeiro, Mark Daniels, Marc C Johnson, and John J Dennehy. Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.*, 13(1):635, February 2022.

[66] Tyler N Starr, Nadine Czudnochowski, Zhuoming Liu, Fabrizia Zatta, Young-Jun Park, Amin Addetia, Dora Pinto, Martina Beltramello, Patrick Hernandez, Allison J Greaney, Roberta Marzi, William G Glass, Ivy Zhang, Adam S Dingens, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, Jason A Wojcechowskyj, Anna De Marco, Laura E Rosen, Jiayi Zhou, Martin Montiel-Ruiz, Hannah Kaiser, Josh R Dillen, Heather Tucker, Jessica Bassi, Chiara Silacci-Fregni, Michael P Housley, Julia di Iulio, Gloria Lombardo, Maria Agostini, Nicole Sprugasci, Katja Culap, Stefano Jaconi, Marcel Meury, Exequiel Dellota, Jr, Rana Abdelnabi, Shi-Yan Caroline Foo, Elisabetta Cameroni, Spencer Stumpf, Tristan I Croll, Jay C Nix, Colin Havenar-Daughton, Luca Piccoli, Fabio Benigni, Johan Neyts, Amalio Telenti, Florian A Lempp, Matteo S Pizzuto, John D Chodera, Christy M Hebner, Herbert W Virgin, Sean P J Whelan, David Veesler, Davide Corti, Jesse D Bloom, and Gyorgy Snell. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature*, 597(7874):97–102, September 2021.

[67] Tyler N Starr, Allison J Greaney, Amin Addetia, William W Hannon, Manish C Choudhary, Adam S Dingens, Jonathan Z Li, and Jesse D Bloom. Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science*, 371(6531):850–854, February 2021.

[68] Tyler N Starr, Allison J Greaney, Adam S Dingens, and Jesse D Bloom. 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(4):100255, April 2021.

[69] Tyler N Starr, Allison J Greaney, William W Hannon, Andrea N Loes, Kevin Hauser, Josh R Dillen, Elena Ferri, Ariana Ghez Farrell, Bernadeta Dadonaite, Matthew McCallum, Kenneth A Matreyek, Davide Corti, David Veesler, Gyorgy Snell, and Jesse D Bloom. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science*, 377(6604):420–424, July 2022.

[70] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine H D Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, Neil P King, David Veesler, and Jesse D Bloom. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.e20, September 2020.

[71] Lorenzo Subissi, Anne von Gottberg, Lipi Thukral, Nathalie Worp, Bas B Oude Munnink, Surabhi Rathore, Laith J Abu-Raddad, Ximena Aguilera, Erik Alm, Brett N Archer, Homa Attar Cohen, Amal Barakat, Wendy S Barclay, Jinal N Bhiman, Leon Caly, Meera Chand, Mark Chen, Ann Cullinane, Tulio de Oliveira, Christian Drosten, Julian Druce, Paul Effler, Ihab El Masry, Adama Faye, Simani Gaseitsiwe, Elodie Ghedin, Rebecca Grant, Bart L Haagmans, Belinda L Herring, Shilpa S Iyer, Zyleen Kassamali, Manish Kakkar, Rebecca J Kondor, Juliana A Leite, Yee-Sin Leo, Gabriel M Leung, Marco Marklewitz, Sikhulile Moyo, Jairo Mendez-Rico, Nada M Melhem, Vincent Munster, Karen Nahapetyan, Djin-Ye Oh, Boris I Pavlin, Thomas P Peacock, Malik Peiris, Zhibin Peng, Leo L M Poon, Andrew Rambaut, Jilian Sacks, Yinzhong Shen, Marilda M Siqueira, Sofonias K Tessema, Erik M Volz, Volker Thiel, Sylvie van der Werf, Sylvie Briand, Mark D Perkins, Maria D Van Kerkhove, Marion P G Koopmans, and Anurag Agrawal. An early warning system for emerging SARS-CoV-2 variants. *Nat. Med.*, 28(6):1110–1115, June 2022.

[72] B E Suzek, H Huang, P McGarvey, R Mazumder, and C H Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[73] Takuya Tada, Hao Zhou, Belinda M Dcosta, Marie I Samanovic, Vidya Chivukula, Ramin S Herati, Stevan R Hubbard, Mark J Mulligan, and Nathaniel R Landau. Increased resistance of SARS-CoV-2 omicron variant to

neutralization by vaccine-elicited and therapeutic antibodies. *EBioMedicine*, 78(103944):103944, April 2022.

[74] Joseph M Taft, Cédric R Weber, Beichen Gao, Roy A Ehling, Jiami Han, Lester Frei, Sean W Metcalfe, Max D Overath, Alexander Yermanos, William Kelton, and Sai T Reddy. Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain. *Cell*, 185(21):4008–4022.e14, October 2022.

[75] Julian W Tang, Tommy T Lam, Hassan Zaraket, W Ian Lipkin, Steven J Drews, Todd F Hatchette, Jean-Michel Heraud, Marion P Koopmans, Ashta Mary Abraham, Amal Baraket, Seweryn Bialasiewicz, Miguela A Caniza, Paul K S Chan, Cheryl Cohen, André Corriveau, Benjamin J Cowling, Steven J Drews, Marcela Echavarria, Ron Fouchier, Pieter L A Fraaij, Todd F Hachette, Jean-Michel Heraud, Hamid Jalal, Lance Jennings, Alice Kabanda, Herve A Kadjo, Mohammed Rafiq Khanani, Evelyn S C Koay, Marion P Koopmans, Mel Krajden, Tommy T Lam, Hong Kai Lee, W Ian Lipkin, Julius Lutwama, David Marchant, Hidekazu Nishimura, Pagbajabyn Nymadawa, Benjamin A Pinsky, Sanjiv Rughooputh, Joseph Rukelibuga, Taslimarif Saiyed, Anita Shet, Theo Sloots, J J Muyembe Tamfum, Julian W Tang, Stefano Tempia, Sarah Tozer, Florette Treurnicht, Matti Waris, Aripuana Watanabe, and Emile Okitolonda Wemakoy. Global epidemiology of non-influenza RNA respiratory viruses: data gaps and a growing need for surveillance. *Lancet Infect. Dis.*, 17(10):e320–e326, October 2017.

[76] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Chris Sander, Yarin Gal, and Debora S Marks. Learning from pre-pandemic data to forecast viral escape. April 2023.

[77] J M Thornton, M S Edwards, W R Taylor, and D J Barlow. Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J.*, 5(2):409–413, February 1986.

[78] M Alejandra Tortorici, Nadine Czudnochowski, Tyler N Starr, Roberta Marzi, Alexandra C Walls, Fabrizia Zatta, John E Bowen, Stefano Jaconi, Julia Di Iulio, Zhaoqian Wang, Anna De Marco, Samantha K Zepeda, Dora Pinto, Zhuoming Liu, Martina Beltramello, Istvan Bartha, Michael P Housley, Florian A Lempp, Laura E Rosen, Exequiel Dellota, Jr, Hannah Kaiser, Martin Montiel-Ruiz, Jiayi Zhou, Amin Addetia, Barbara Guarino, Katja Culap, Nicole Sprugasci, Christian Saliba, Eneida Vetti, Isabella Giacchetto-Sasselli, Chiara Silacci Fregni, Rana Abdelnabi, Shi-Yan Caroline Foo, Colin Havenar-Daughton, Michael A Schmid, Fabio Benigni, Elisabetta Cameroni, Johan Neyts, Amalio Telenti, Herbert W Virgin, Sean P J Whelan, Gyorgy Snell, Jesse D Bloom, Davide Corti, David Veesler, and Matteo Samuele Pizzuto. Broad sarbecovirus neutralization by a human monoclonal antibody. *Nature*, 597(7874):103–108, September 2021.

[79] Philip L Tzou, Kaiming Tao, Sergei L Kosakovsky Pond, and Robert W Shafer. Coronavirus resistance database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. *PLoS One*, 17(3):e0261045, March 2022.

[80] Ning Wang, Jian Shang, Shibo Jiang, and Lanying Du. Subunit vaccines against emerging pathogenic human coronaviruses. *Front. Microbiol.*, 11:298, February 2020.

[81] Zhaoqian Wang, Moushimi Amaya, Amin Addetia, Ha V Dang, Gabriella Reggiano, Lianying Yan, Andrew C Hickey, Frank DiMaio, Christopher C Broder, and David Veesler. Architecture and antigenicity of the nipah virus attachment glycoprotein. *Science*, 375(6587):1373–1378, March 2022.

[82] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, Natalia L Komarova, Martin A Nowak, Beatrice H Hahn, Peter D Kwong, and George M Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312, March 2003.

[83] Leander Witte, Viren A Baharani, Fabian Schmidt, Zijun Wang, Alice Cho, Raphael Raspe, Camila Guzman-Cardozo, Frauke Muecksch, Marie Canis, Debby J Park, Christian Gaebler, Marina Caskey, Michel C Nussenzweig, Theodora Hatziioannou, and Paul D Bieniasz. Epistasis lowers the genetic barrier to SARS-CoV-2 neutralizing antibody escape. *Nat. Commun.*, 14(1):302, January 2023.

[84] Nicholas C Wu, Andrew J Thompson, Juhye M Lee, Wen Su, Britni M Arlian, Jia Xie, Richard A Lerner, Hui-Ling Yen, Jesse D Bloom, and Ian A Wilson. Different genetic barriers for resistance to HA stem antibodies in influenza H3 and H1 viruses. *Science*, 368(6497):1335–1340, June 2020.

[85] Kai Xu, Barry Rockx, Yihu Xie, Blair L DeBuysscher, Deborah L Fusco, Zhongyu Zhu, Yee-Peng Chan, Yan Xu, Truong Luu, Regina Z Cer, Heinz Feldmann, Vishwesh Mokashi, Dimiter S Dimitrov, Kimberly A Bishop-Lilly, Christopher C Broder, and Dimitar B Nikolov. Crystal structure of the hendra virus attachment G glycoprotein bound to a potent cross-reactive neutralizing human monoclonal antibody. *PLoS Pathog.*, 9(10):e1003684, October 2013.

[86] Lue Ping Zhao, Terry Lybrand, Peter Gilbert, Thomas H Payne, Chul-Woo Pyo, Daniel Geraghty, and Keith Jerome. Rapidly identifying new coronavirus mutations of potential concern in the omicron variant using an unsupervised learning strategy. *Res. Sq.*, February 2022.