Viral Evolution and Antibody Escape Mutations using Deep Generative Models

Nicole Thadani [*1] **Nathan Rollins** [*1] **Sarah Gurev** [2] **Pascal Notin** [3] **Yarin Gal** [3] **Debora Marks** [1 4]

Abstract

Mutations in viruses can result in zoonosis, immune escape, and changes in pathology. To control evolving pandemics, we wish to predict likely trajectories of virus evolution. Here we predict the probability of SARS-CoV-2 protein variants by using deep generative models to capture constraints on broader evolution of coronavirus sequences. We validate against lab measurements of mutant effects on replication and molecular function (e.g. receptor binding). We then apply our predictor to evaluate the potential of mutational escape from known antibodies, a strategy which can facilitate the development of antibody therapeutics and vaccines to mitigate immune evasion.

1. Introduction

Viral diseases are characterized by the interplay between immune detection and evasion, leading to rapid evolution and changes in virulence. Viral escape mutations influence reinfection rates and the duration of vaccine-induced immunity, shaping population prevalence over time. To control viral epidemics, we wish to predict plausible variants with enhanced transmissibility or virulence. Methods for predicting viral mutation effects could be used to anticipate the antigenic ramifications of viral evolution and develop targeted vaccines and therapeutics that mitigate resulting viral spread.

A variety of deep mutational scanning (DMS) methods have emerged for quantifying the effects of viral protein mutations in parallel. (Doud & Bloom, 2016; Lee et al., 2018; Haddox et al., 2018; Starr et al., 2020; 2021; Mattenberger et al., 2021). Each assay quantifies one of several phenotypes including viral replication in cell culture, neutralization of viral replication by antibodies or drugs, and cellular receptor or antibody binding. While these assays have greatly informed our understanding of the potential impacts of viral evolution, they are restricted to experimentally tractable phenotypes. Viral replication assays are a relatively comprehensive assay of virus function, but are limited to culture-friendly cell lines and do not capture effects on transmissibility or behavior in tissues. Replication assays are also challenging for viruses that lack efficient cell-growth protocols. In lieu of replication assays for SARS-CoV-2, yeast-display assays of the Spike receptor binding domain (RBD) have been used to measure surface expression, binding to the ACE2 receptor, and binding to antibodies (Starr et al., 2020; 2021). However, the relationship between these biochemical phenotypes and overall viral function are unclear.

There is also a limit to the number of sequences assayed in a DMS. Although DMS can often measure all possible single mutants to a target protein, those mutations are in context of a single background sequence and cannot always extrapolate to the fast-changing context of circulating variants (Lee et al., 2018). DMS can measure only a vanishingly small fraction of the possible multi-mutation combinations.

Alternative insights into virus evolution can be inferred from sequence databases of natural proteins. Generative models of natural sequences can capture the evolutionary constraints on a protein and predict mutant effects without supervision. Natural sequence models are independent of wet-lab assays and can aide interpretation of DMS measurements. Rather than single phenotypes of single mutants, these models can capture the ensemble of constraints in evolution and predict effects of multiple mutants and sequence contexts. The state-of-art models consider interactions between mutations and the full sequence context, outperforming models that rely solely on sitewise conservation (Shin et al., 2021; Riesselman et al., 2018; Hopf et al., 2017; Frazer et al., 2020).

Here, we apply these models to predict the effects of mutations to viral proteins and investigate both agreement and contrast to experimentally-determined phenotypes. We then predict the mutability of antibody epitopes to SARS-CoV-2 Spike protein and the likelihood of escape mutants.

[*]Equal contribution [1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA [3]OATML Group, Department of Computer Science, University of Oxford, Oxford, United Kingdom [4]Broad Institute of Harvard and MIT, Cambridge, MA, USA. Correspondence to: Debora Marks <debbie@hms.harvard.edu>.

The 2021 ICML Workshop on Computational Biology.

2. Methods

2.1. Multiple sequence alignments

For each viral protein, multiple sequences alignments were constructed using jackhmmer (Eddy), an iterative profile-HMM based search tool, against the uniref100 database augmented with coronavirus and influenza sequences from GISAID and HIV sequences from the Los Alamos National Labs HIV database. We optimized search depth to maximize sequence coverage and the effective number of sequences included after clustering similar sequences as previously reported (Hopf et al., 2017; Riesselman et al., 2018), achieving greater numbers of effective sequences per protein length for influenza and HIV proteins than for coxsackievirus capsid and Spike RBD (Table 1).

2.2. Models

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

$$\log \frac{p(x^{\text{mutant}}|\theta)}{p(x^{\text{wildtype}}|\theta)} \tag{1}$$

We compared three alignment-based probabilistic models of sequences. The sitewise model and EVmutation are undirected graphical models with probability

$$p(x|\theta) = \frac{1}{Z}\exp(E(x)) \tag{2}$$

where $E(x)$ is the log-potential of a given sequence and Z normalizes over possible sequences (Hopf et al., 2017). For the sitewise model, $E(x)$ is additive over sites.

$$E_{site}(x) = \sum_i h_i(x_i) \tag{3}$$

The EVmutation model also includes terms for pairwise combinations of sites.

$$E_{pair}(x) = \sum_i h_i(x_i) + \sum_{i<j} J_{ij}(x_i, x_j) \tag{4}$$

The EVE (Evolutionary model of Variant Effects) model is a Bayesian variational autoencoder (VAE), capable of capturing complex higher-order interactions across sequence positions (Riesselman et al., 2018; Frazer et al., 2020).

Table 1. Number of effective sequences over length for selected viral protein alignments.

FLU H1	HIV ENV	COXSACKIEVIRUS	SARS2 RBD
16.9	50.3	1.76	0.573

2.3. Experimental data collection

We assembled a dataset of viral mutational scans combining those analyzed in DeepSequence with additional recent studies (Riesselman et al., 2018). Studies of influenza, HIV, zika, and coxsackievirus assayed viral replication, while studies of SARS-CoV-2 Spike RBD assayed RBD expression in a yeast-display platform and RBD binding to the ACE2 receptor.

To identify plausible viral escape mutations from antibodies, we collected cryo-EM structures of spike RBD in complex with antibodies and considered all residues with any atom within 4 Å of the antibody as its epitope (Piccoli et al., 2020; Chi et al., 2020; Wang et al., 2021). We also used available antibody-RBD DMS binding studies to compare measured antibody escape with predicted mutant functionality (Starr et al., 2021).

3. Results

3.1. Natural sequence models predict the effects of mutations in viral proteins

We compare predictions from generative models of viral sequences to mutational scans of viral surface proteins, with assays measuring viral replication as well as specific protein phenotypes such as cell receptor binding.

The EVE Bayesian VAE is on par with or outperforms linear sitewise and EVmutation models (Figure 1). For influenza hemagglutinin and HIV envelope proteins, correlation between model predictions and viral replication measurements is similar to the correlation between independent experimental replicates (Table 2). Despite the limited natural sequence diversity available for Spike RBD (Table 1), EVE predictions are moderately correlated with observed experimental phenotypes. EVE mutation effect performance on RBD phenotypes may improve as more diverse coronavirus sequences and SARS-CoV-2 variants are identified and incorporated in training data.

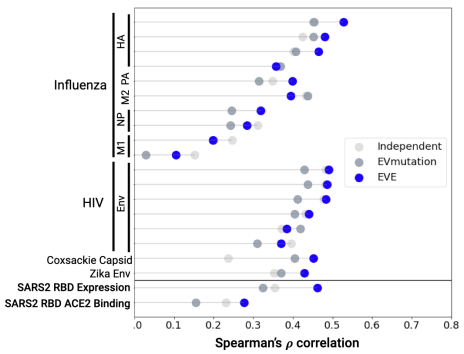

Figure 1. EVE is on par with or outperforms other models in viral missense mutation effect prediction.

Table 2. Pearson correlations between viral replication experiment replicates (as reported in (Doud & Bloom, 2016; Haddox et al., 2018)) and model predictions.

EXPERIMENT	REPLICATES	EVE (MUTATIONS)	EVE (SITES)
FLU H1	0.59-0.66	0.51	0.62
HIV ENV	0.59-0.64	0.52	0.64

3.2. Natural sequence models capture protein sequence constraints not measured in SARS-CoV-2 RBD DMS

To date, DMSs of SARS-CoV-2 have focused on subunits of the Spike protein, primarily the receptor-binding domain (RBD). We examined disagreements between natural sequence model predictions and DMS experiments measuring RBD expression and binding to the ACE2 human cellular receptor in a yeast-display system.

A subset of RBD sites in the expression assay tolerate mutations that are predicted as deleterious by EVE. (Figure 2 A-B). Several of these positions are in contact with non-assayed domains of the spike protein, or with other spike subunits in the trimer assembly. Possibly these sites are critical for spike protein folding and trimer assembly, but non-essential for RBD folding and display in the yeast-display system. When these sites are removed, EVE spearman correlation with RBD expression improves (Table 3).

The RBD ACE2 binding assay and EVE predictions are not well-correlated. This may be because EVE predicts several mutations as deleterious that are not proximal to ACE2 (receptor binding motif (RBM)) in the bound structure (Figure 2 C-D). These mutations may be in sites that have other roles in spike structural stability or function. When we only consider the RBM, EVE spearman correlation with RBD ACE2 binding improves (Table 3).

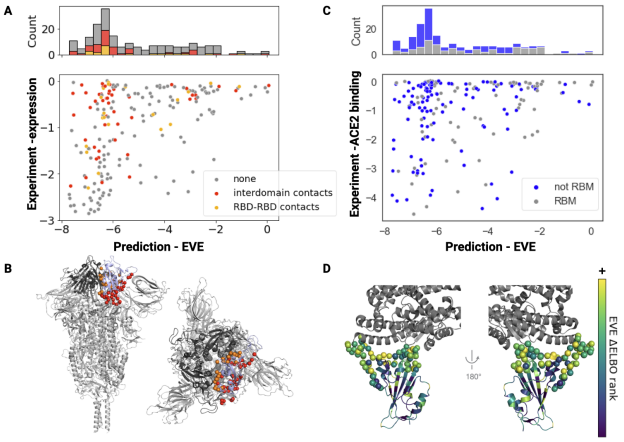

Figure 2. EVE prediction differences from RBD yeast-display experiments are biologically interpretable. A) EVE score is deleterious in contrast to expression in sites proximal to non-assayed Spike domains B) Sites where RBD (light blue) is proximal to other spike domains (red spheres) or other RBDs in the spike trimer (orange spheres). PDB: 6XR8 C) EVE score is deleterious in contrast to ACE2 binding in sites not located in the receptor binding motif (RBM). D) Sites in the RBM (spheres) are predicted to be more tolerant to mutation by EVE. PDB: 6M0J

Table 3. Spearman correlations between Spike RBD experiments and EVE predictions on subsets of residues

EXPERIMENT	REGION	EVE
RBD EXPRESSION	RBD	0.46
RBD EXPRESSION	RBD W/O SPIKE CONTACTS	0.55
RBD ACE2 BINDING	RBD	0.27
RBD ACE2 BINDING	RBM	0.45

As a whole, EVE's predictive performance on viral replication experiments and our analysis of model disagreements with RBD biochemical protein assays suggests that EVE learns a combination of the varied constraints on viral protein function. EVE predictions may complement DMS studies that focus on biochemical protein assays by incorporating information about non-assayed constraints.

3.3. Natural sequence models can be used to predict the mutability of antibody epitopes

Viral escape mutations that avoid immune recognition influence the likelihood of reinfection and the duration of vaccine-induced immunity. Mutation effect predictions can inform the design of vaccines and therapeutics to target protein regions intolerant to mutation, reducing the chances of

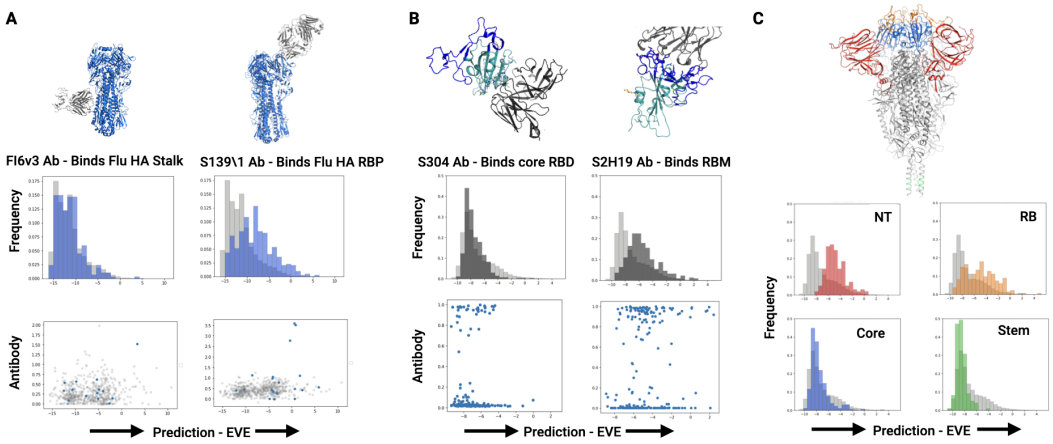

Figure 3. EVE predicts the plausibility of mutations in influenza HA and SARS2 Spike. A) Comparison of two HA antibodies (top), EVE predictions (middle), and measured antibody binding (bottom). B) Comparison of two Spike RBD antibodies (top), EVE predictions (middle), and measured antibody binding (bottom). C) Spike structure highlighting the NTD (red), RBM (orange), core RBD (blue) and stem helix (green) (top). EVE score distributions of mutations to residues of epitopes in different Spike domains (bottom).

successful escape. We applied EVE to evaluate the chance of escape from known antibody epitopes in the influenza hemagglutinin protein (HA) and the SARS-CoV-2 Spike RBD (Figure 3A-C).

The antibody FI6v3 binds to the stalk of the HA protein, while S139\1 binds to the receptor binding pocket. EVE predicts that FI6v3 mutations are more deleterious than S139\1 mutations, corresponding to results from DMS experiments quantifying replication fitness in the presence of antibodies - these studies found that stalk-targeted antibodies are more resistant to escape (Doud et al., 2018).

The antibody S304 binds in the core of the RBD, while S2H19 binds to the RBM (Piccoli et al., 2020). EVE predicts that S304 epitope mutations (average EVE score: -7.55) are more deleterious than S2H19 epitope mutations (average EVE score: -5.42). DMS quantifying antibody binding in a yeast-display platform identified potential escape mutations to both antibodies (Starr et al., 2021). EVE predicts that all 41 potential escape mutants (escape fraction $>$ 0.5) to S304 are deleterious (EVE score $<$ -4), while 29 of the 92 potential escape mutants to S2H19 may be tolerated (Figure 3B, bottom). This strongly contrasts the DMS data interpreted as suggested in (Starr et al., 2021), which permits all 41 mutants in S304 and all 92 mutants in S2H19. This application of EVE contextualizes RBD antibody binding experiments with information about other constraints on protein function that may limit emergence of escape variants.

EVE can also be applied to regions of SARS-CoV-2 Spike that currently lack DMS characterization. We used EVE to predict mutation effects for antibody epitopes identified in the N-terminal domain (NTD) and stem helix (Chi et al., 2020; Wang et al., 2021) (Figure 3C). EVE predicts that the NTD epitope is most tolerant to mutation, followed by the RBM. Conversely, the stem helix and core RBD epitopes are predicted to be less tolerant to mutation. The 2D89 antibody to the stem helix has demonstrated cross-reactivity to five human coronaviruses, supporting the hypothesis that this region is a useful target for universal coronavirus vaccines and therapeutic antibodies (Wang et al., 2021). While here we highlight one antibody per structural domain, we found consistent results for 135 antibody epitopes throughout Spike.

4. Conclusion

Models of natural sequence variation learn constraints on viral proteins from the evolutionary record, not limited to experimentally tractable phenotypes or the number of sequences that can be assayed. The deep Bayesian VAE, EVE, accurately predicts protein-level mutation effects on viral replication, and is moderately correlated with DMSs of SARS-CoV-2 Spike RBD expression and binding. Sites of model disagreement with the spike RBD experiments suggest that EVE sheds light on constraints unobserved in DMS studies and learns a combination of the varied constraints on viral protein function. Mutation effect predictions by natural sequence models can identify antibody epitopes intolerant to mutation, adding information about experimentally unobserved constraints on protein function and extending beyond regions explored by DMSs. Such mutation-intolerant epitopes are ideal regions to target for strain-universal vaccination and therapeutics.

References

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y., Chen, Z., Guo, Y., Zhang, J., Li, Y., Song, X., Chen, Y., Xia, L., Fu, L., Hou, L., Xu, J., Yu, C., Li, J., Zhou, Q., and Chen, W. A neutralizing human antibody binds to the n-terminal domain of the spike protein of SARS-CoV-2. *Science*, 369(6504):650–655, August 2020.

Doud, M. B. and Bloom, J. D. Accurate measurement of the effects of all Amino-Acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, 2016.

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

Eddy, S. HMMER: biosequence analysis using profile hidden markov models. www.hmmer.org.

Frazer, J., Notin, P., Dias, M., Gomez, A., Brock, K., Gal, Y., and Marks, D. S. Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. *bioRxiv*, pp. 2020.12.21.423785, December 2020.

Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., and Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife*, 7:e34420, March 2018.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, February 2017.

Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., and Bloom, J. D. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proc. Natl. Acad. Sci. U. S. A.*, 115(35):E8276–E8285, August 2018.

Mattenberger, F., Latorre, V., Tirosh, O., Stern, A., and Geller, R. Globally defining the effects of mutations in a picornavirus capsid. *Elife*, 10:e64256, January 2021.

Piccoli, L., Park, Y.-J., Tortorici, M. A., Czudnochowski, N., Walls, A. C., Beltramello, M., Silacci-Fregni, C., Pinto, D., Rosen, L. E., Bowen, J. E., Acton, O. J., Jaconi, S., Guarino, B., Minola, A., Zatta, F., Sprugasci, N., Bassi, J., Peter, A., De Marco, A., Nix, J. C., Mele, F., Jovic, S., Rodriguez, B. F., Gupta, S. V., Jin, F., Piumatti, G., Lo Presti, G., Pellanda, A. F., Biggiogero, M., Tarkowski, M., Pizzuto, M. S., Cameroni, E., Havenar-Daughton, C., Smithey, M., Hong, D., Lepori, V., Albanese, E., Ceschi, A., Bernasconi, E., Elzi, L., Ferrari, P., Garzoni, C., Riva, A., Snell, G., Sallusto, F., Fink, K., Virgin, H. W., Lanzavecchia, A., Corti, D., and Veesler, D. 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.

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

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1): 1–11, April 2021.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., and Bloom, J. D. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.e20, 2020.

Starr, T. N., Czudnochowski, N., Zatta, F., Park, Y.-J., Liu, Z., Addetia, A., Pinto, D., Beltramello, M., Hernandez, P., Greaney, A. J., Marzi, R., Glass, W. G., Zhang, I., Dingens, A. S., Bowen, J. E., Wojcechowskyj, J. A., Marco, A. D., Rosen, L. E., Zhou, J., Montiel-Ruiz, M., Kaiser, H., Tucker, H., Housley, M. P., Iulio, J. D., Lombardo, G., Agostini, M., Sprugasci, N., Culap, K., Jaconi, S., Meury, M., Dellota, E., Cameroni, E., Croll, T. I., Nix, J. C., Havenar-Daughton, C., Telenti, A., Lempp, F. A., Pizzuto, M. S., Chodera, J. D., Hebner, C. M., Whelan, S. P. J., Virgin, H. W., Veesler, D., Corti, D., Bloom, J. D., and Snell, G. Antibodies to the SARS-CoV-2 receptor-binding domain that maximize breadth and resistance to viral escape. *bioRxiv*, pp. 2021.04.06.438709, April 2021.

Wang, C., van Haperen, R., Gutiérrez-Álvarez, J., Li, W., Okba, N. M. A., Albulescu, I., Widjaja, I., van Dieren, B., Fernandez-Delgado, R., Sola, I., Hurdiss, D. L., Daramola, O., Grosveld, F., van Kuppeveld, F. J. M., Haagmans, B. L., Enjuanes, L., Drabek, D., and Bosch, B.-J. A conserved immunogenic and vulnerable site on the coronavirus spike protein delineated by cross-reactive monoclonal antibodies. *Nat. Commun.*, 12(1): 1715, March 2021.