Supplementary Materials for Learning from pre-pandemic data to forecast viral antibody escape

Authors

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

Affiliations

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

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

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

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

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

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

*These authors contributed equally to this work

**Corresponding author: debbie@hms.harvard.edu

Present addresses:

†Seismic Therapeutic, Watertown, MA, USA

Contents:

Supplemental Note S1: EVE model captures viral protein mutant fitness effects

In this work, we use the deep unsupervised sequence model EVE as the fitness component of EVEscape, based on its prior state-of-art performance at predicting the effects of viral protein mutation[18]. We validate EVE's performance on viral proteins and specifically on the SARS-CoV-2 Spike by (1) comparing to a broader set of viral protein DMS experiments, (2) examining EVE performance compared to recently published sequence models at predicting SARS-CoV-2 RBD mutation effects, and (3) examining sites of disagreement between EVE predictions and SARS-CoV-2 RBD DMS experiments.

We evaluated EVE model performance by comparing to high-throughput fitness experiments—deep mutational scans (DMSs)—of the effects of mutations on SARS-CoV-2, HIV and influenza viral function (Table S2)[51–59] EVE generally outperforms linear site-independent and EVmutation models[19] in predicting viral protein mutant effects (Figure S6, Figure S7). Despite the limited natural sequence diversity available for the coronavirus family, especially prior to the pandemic, EVE predictions for SARS-CoV-2 RBD are correlated with observed experimental phenotypes such as expression and binding to the ACE2 human cellular receptor in yeast-display and mammalian cell-surface expression systems[54,55] – pre-pandemic EVE predictions are superior to or on par with other published unsupervised sequence models[19,21,22] (Figure 3B, Figures S6-S8). In the yeast-display assay, a subset of RBD sites in the expression assay tolerate mutations that are predicted as deleterious by EVE (red rectangle in Figure S8A). Several of these positions are in contact with non-assayed domains of the Spike protein, or with other Spike subunits in the trimer assembly, suggesting that they are important for full Spike expression but non-essential for RBD folding in the yeast-display system. EVE predictions are better correlated with ACE2 binding quantified using a full Spike mammalian cell display assay[55], perhaps because this assay readout combines expression with receptor binding (Figure S6-S7). 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.

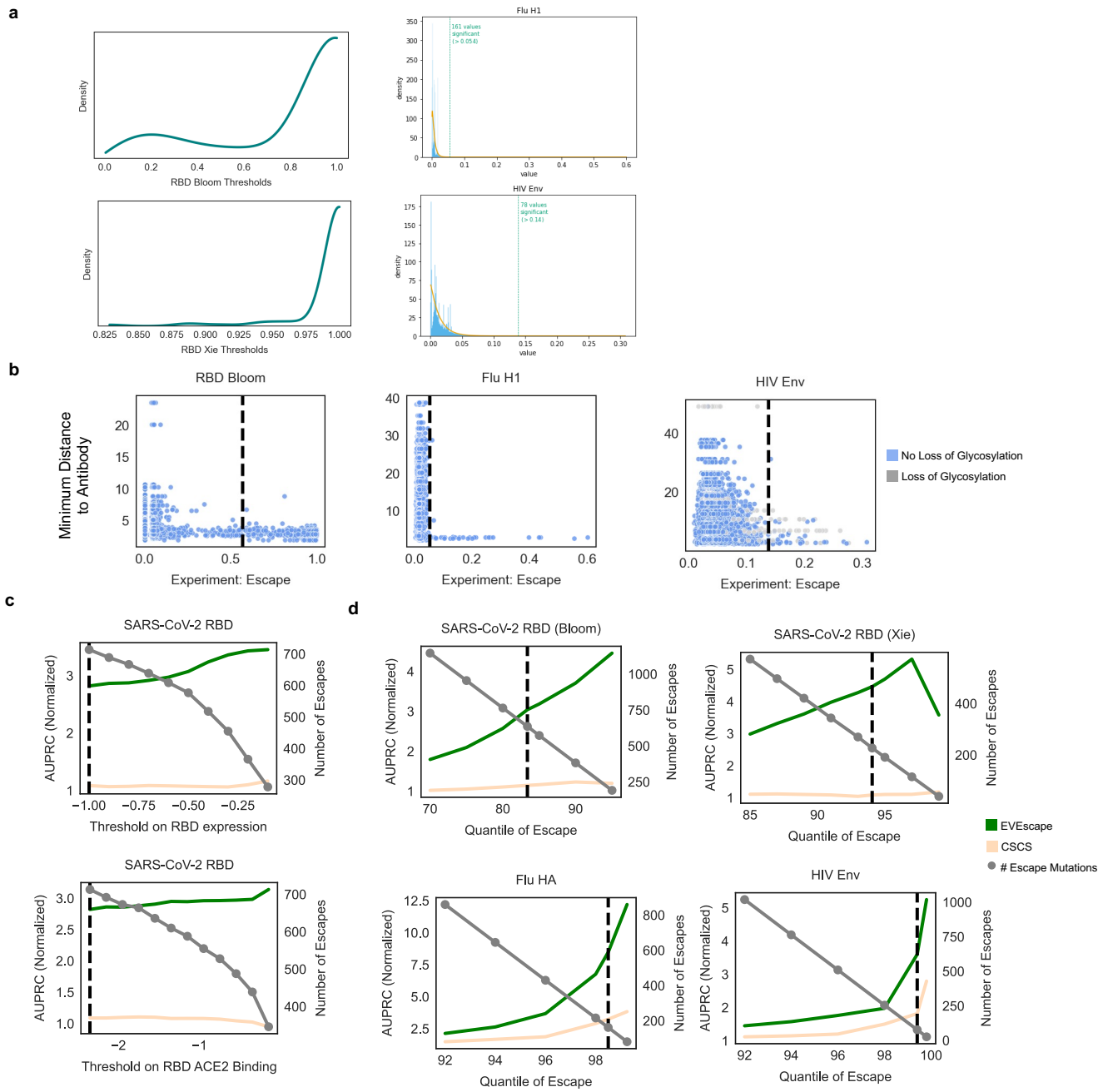


Figure S1: EVEscape performance is robust across data thresholds. a) Distribution of escape thresholds from bootstrapping 8 antibodies 1000 times and fitting a gamma distribution to each sample for Bloom and Xie RBD escape data (left) and gamma distributions to select Flu and HIV escape thresholds (right). **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 within 5Å of an 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 # 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 # of escape mutations (dashed line uses escape threshold chosen by our paper).

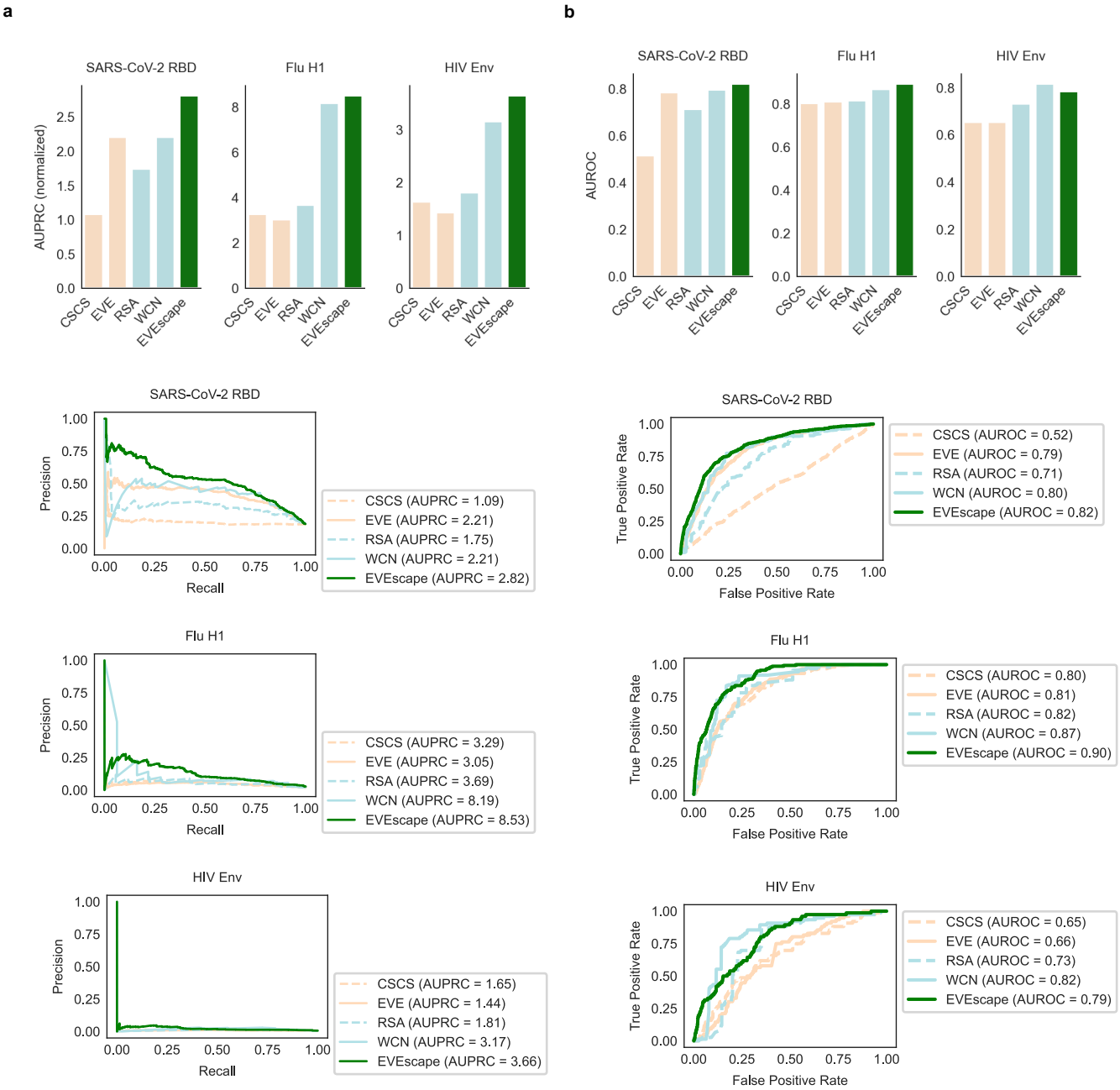


Figure S2: EVEscape performance on escape DMS data is generalizable across viruses. Precision-Recall (with AUPRC normalized by "null" model) (**a**) and AUROC (**b**) of predicting DMS escape mutations, for SARS-CoV-2 RBD, Flu H1, and HIV Env.

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 S4).

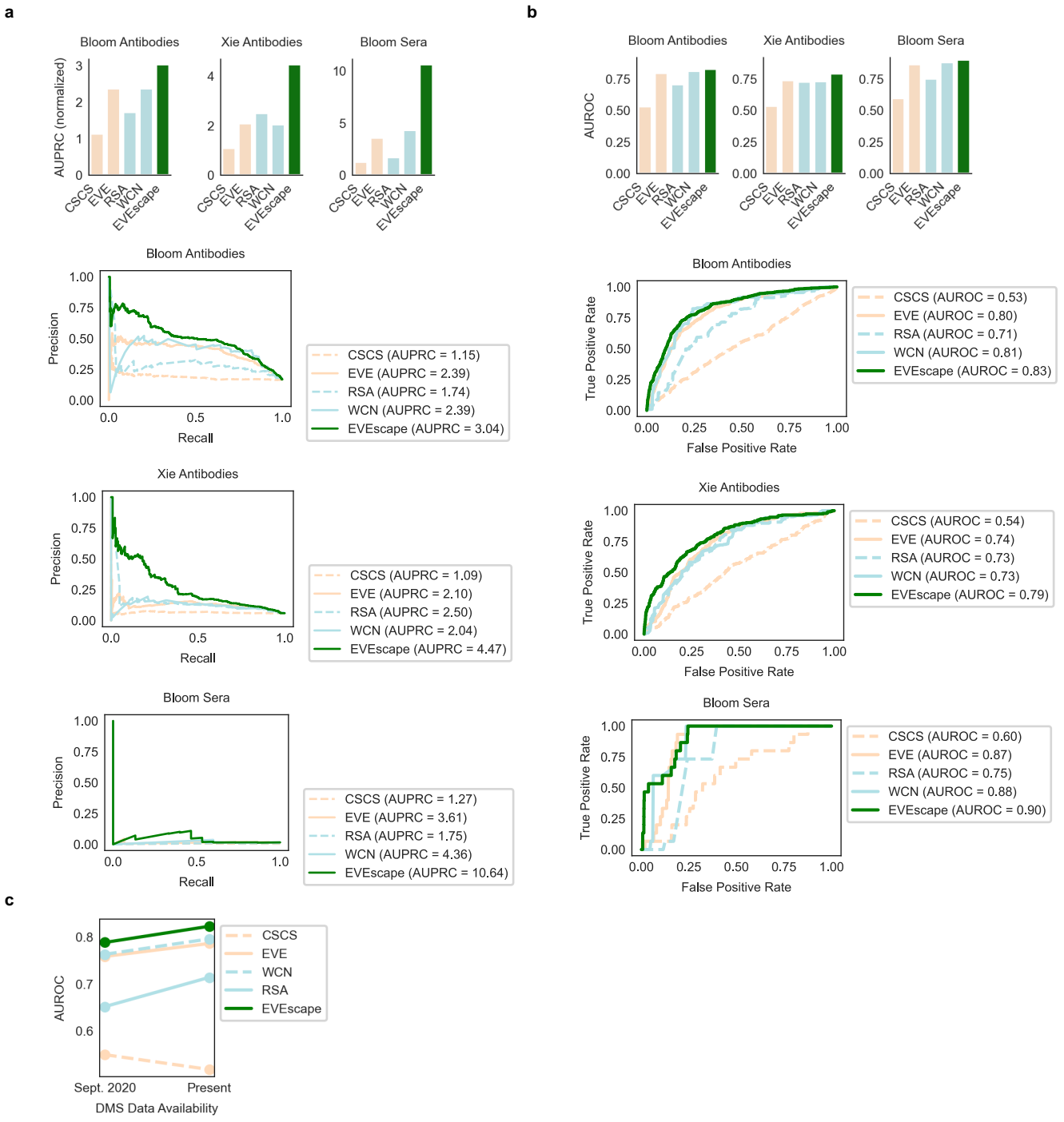


Figure S3: EVEscape RBD performance is robust to antibody and sera samples and improves with more available data for validation. Precision-Recall (with AUPRC normalized by "null" model) (**a**) and AUROC (**b**) of predicting RBD DMS escape mutations, for Bloom and Xie antibodies and Bloom sera. **c**) Comparison of model performance (AUROC) between data available in September 2020 with the first escape DMS study (10 antibodies)[2] 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 S4). The fraction of observed escapes in the DMS experiments are 0.17 for Bloom Ab, for 0.06 for Xie Ab, and 0.003 for Bloom sera.

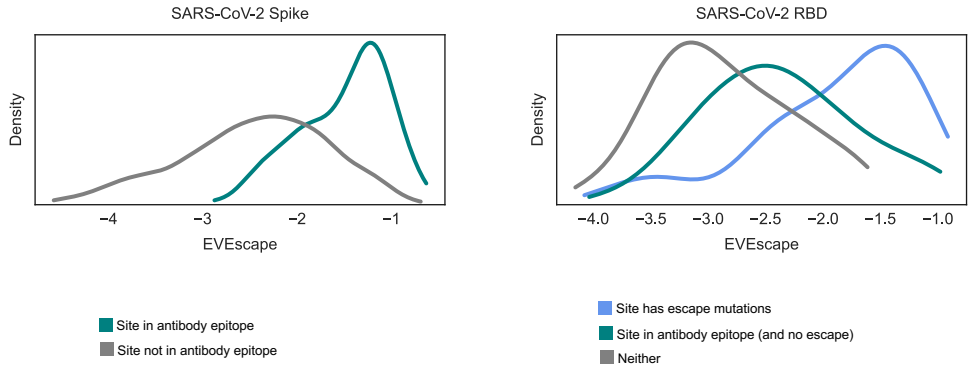


Figure S4: Almost all EVEscape predicted escape sites have escape mutations or are in antibody footprints. Density of site-averaged EVEscape for SARS-CoV-2 full Spike (left) and RBD (right) shows success of EVEscape at distinguishing sites with observed escape mutations, as well as sites in known antibody epitopes, from sites with no evidence of antibody binding or escape.

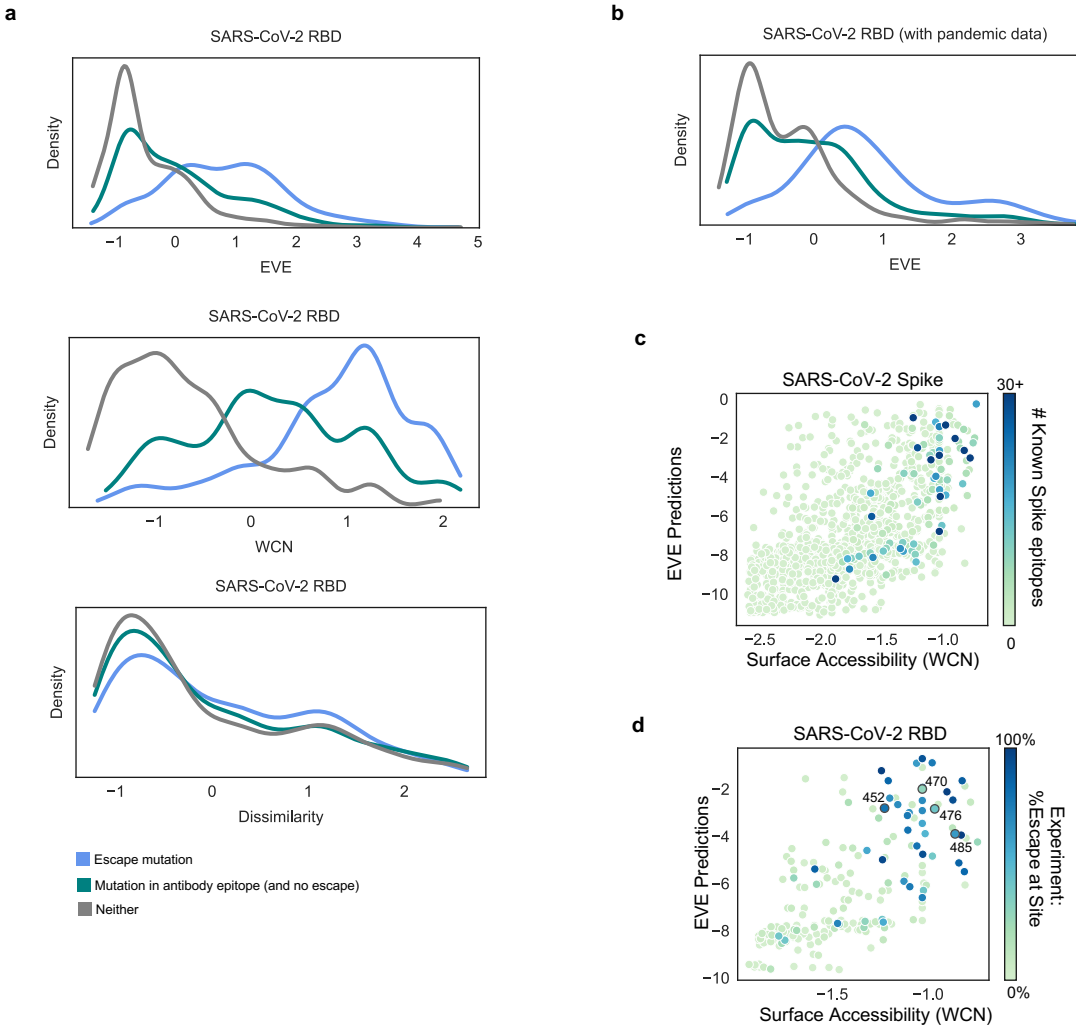


Figure S5: Fitness and accessibility model components separate escape mutants and antibody epitopes from other mutations. a) Density of standard-scaled EVEscape components differ for SARS-CoV-2 RBD escape (and antibody epitopes) and non-escape mutations. **b**) Incorporating pandemic sequences in EVE training data results in a greater distinction between escape and non-escape mutations with high EVE scores. **c**) WCN and EVE predictions provide similar information about the location of Spike epitopes as represented in antibody-Spike crystal structures in RCSB PDB. **d**) Sites with either high accessibility or high EVE fitness predictions have a greater percentage of escape mutants.

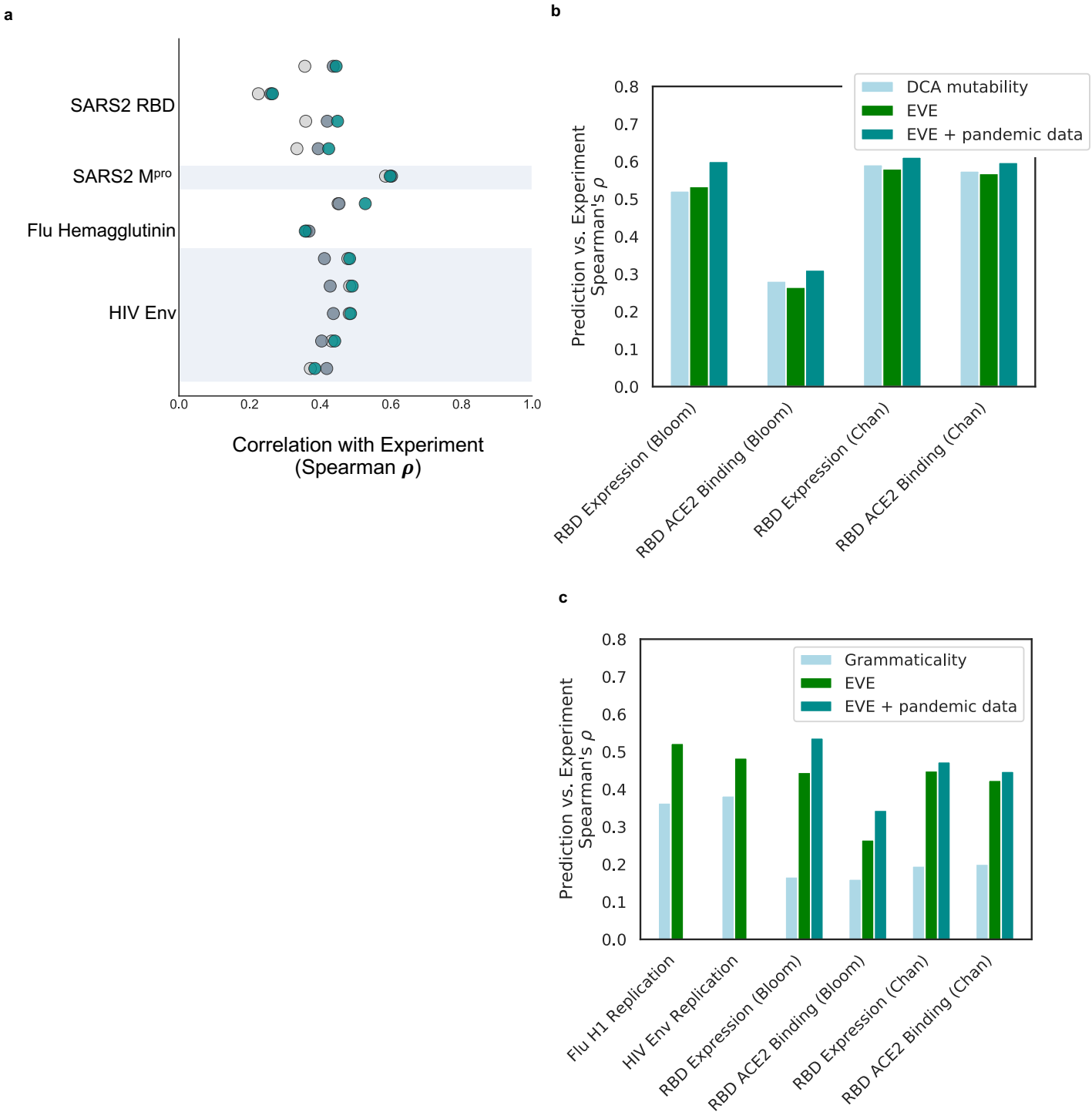


Figure S6: Comparisons of EVE DMS predictions to other sequence-based model predictions. **a**) EVE predictions are correlated with a broad range of viral surface protein DMS experiments surveying protein replication and function, including SARS-CoV-2 RBD and M^{pro}. **b**) Site-averaged EVE predictions have similar correlations with site-averaged SARS-CoV-2 RBD DMS experiments as DCA mutability[21]. **c**) EVE predictions have higher correlations with Flu H1, HIV Env, and SARS-CoV-2 RBD DMS experiments than grammaticality[22].

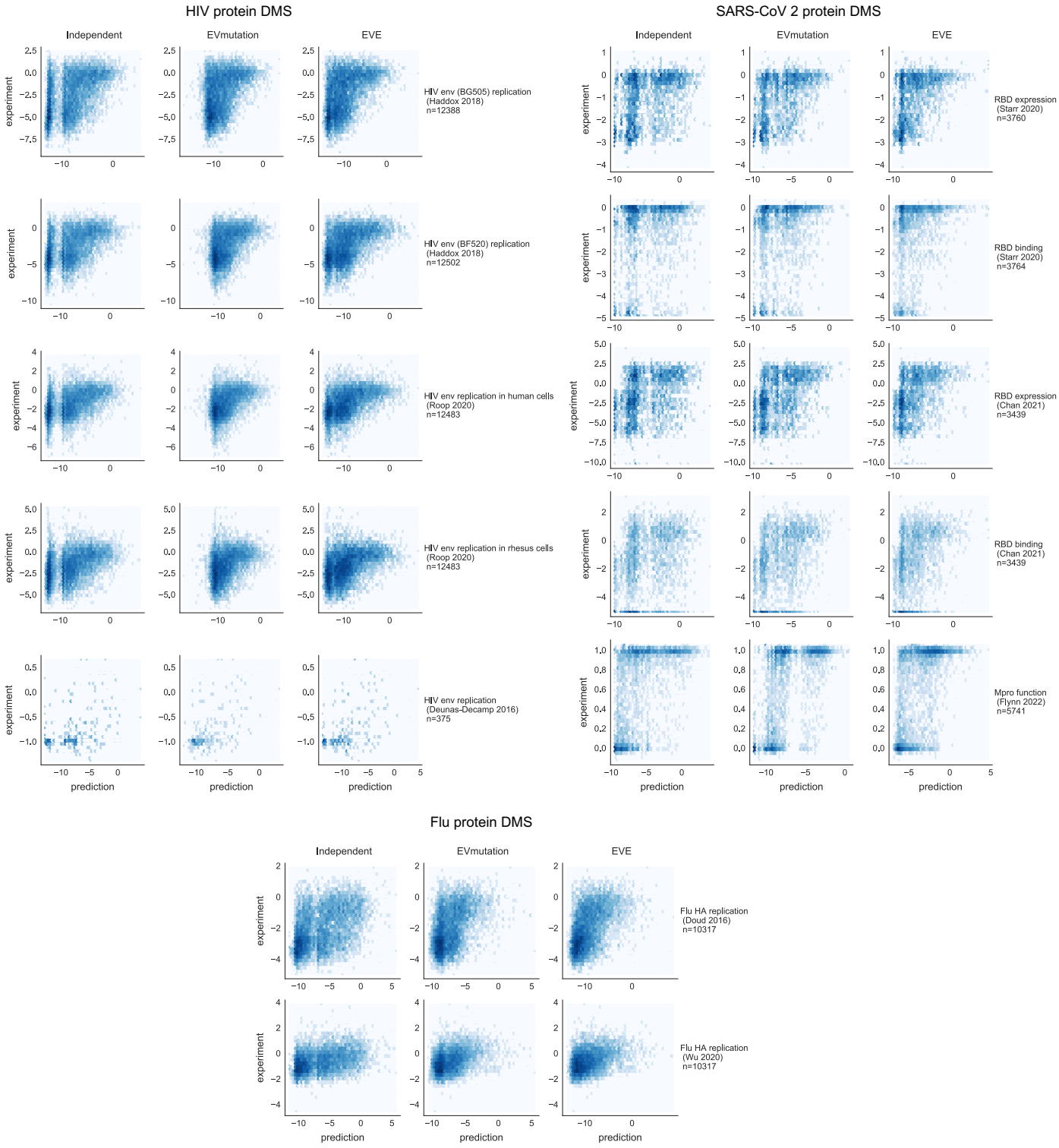
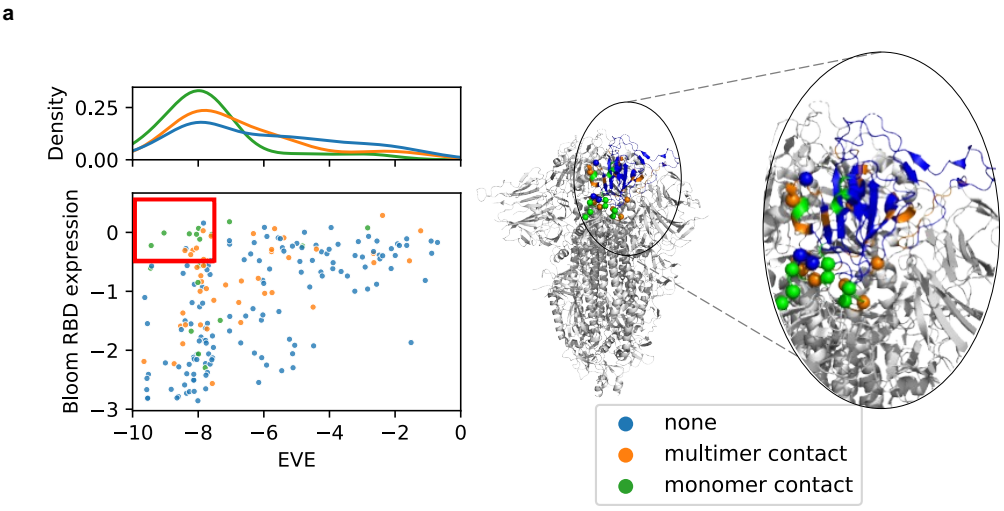


Figure S7: Predictions of viral protein DMS fitness experiments. Scatterplots of viral protein fitness DMS against independent, EVmutation, and EVE model predictions.



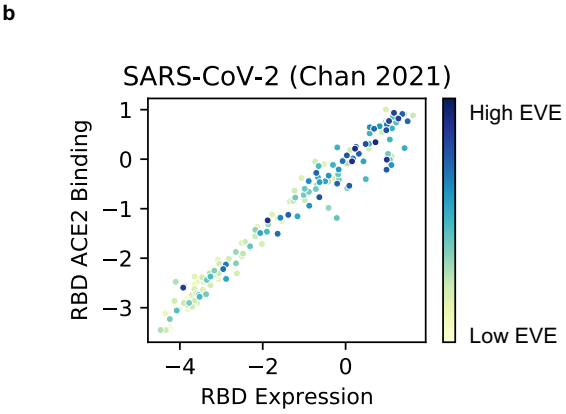


Figure S8: EVE captures structural constraints beyond RBD expression assay. a) Site-averaged EVE scores predict several sites that tolerate mutants in the yeast-display RBD expression assay to be deleterious (red box)–many of these mutants are located at the interface between RBD and the rest of Spike protein. Sites in the red box in scatterplot are shown as spheres on the Spike structure (PDB: 7CAB). **b**) 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[55]. EVE predictions are correlated with both measures.

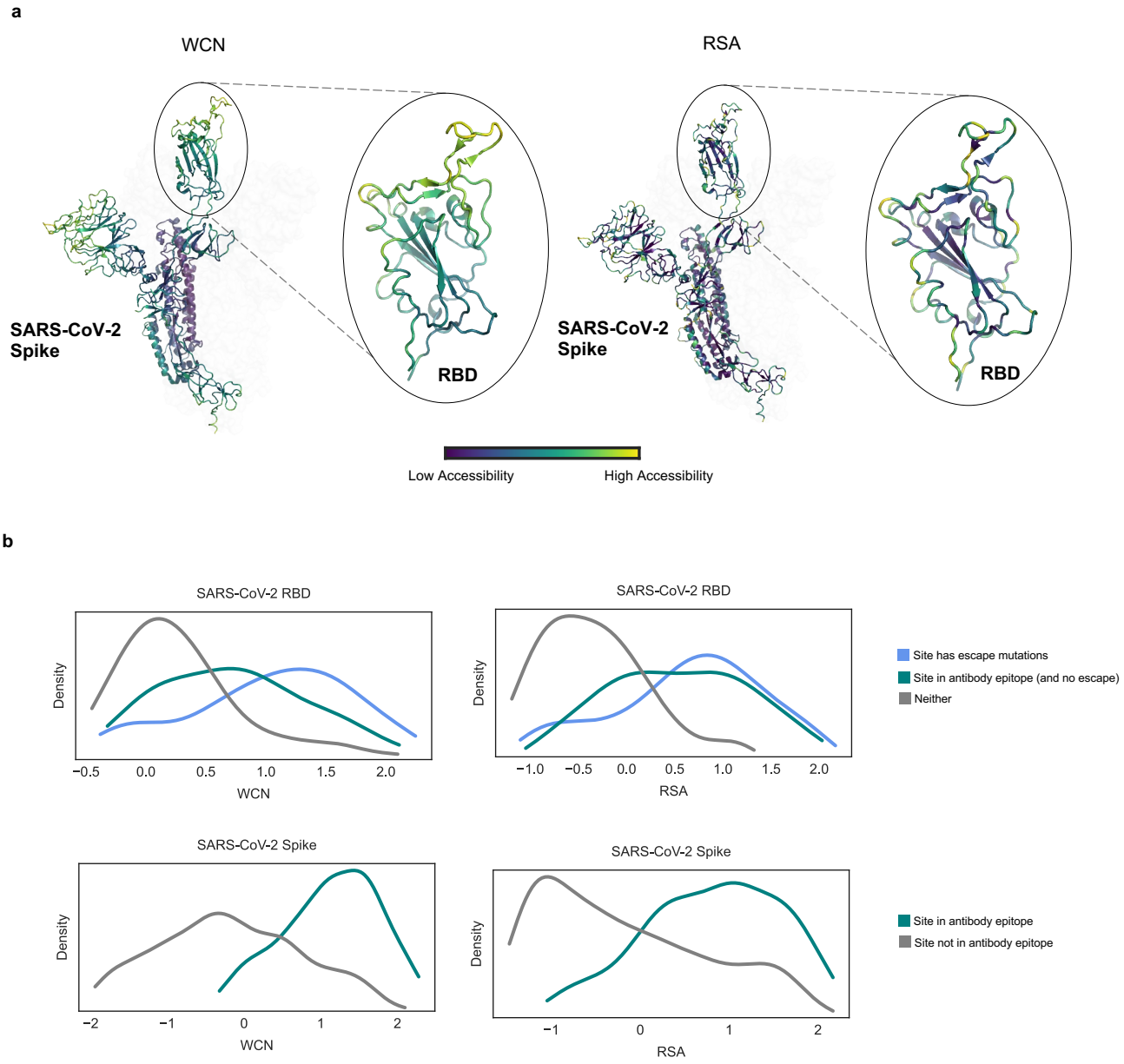


Figure S9: Antibody accessibility metrics distinguish between sites with escape mutants and known antibody epitopes and other sites. a) WCN and RSA values visualized on the SARS-CoV-2 Spike structures show different distributions, particularly in the RBD (PDB: 7BNN), as WCN captures protrusion from the core structure.
b) Distributions of RSA and WCN (standard-scaled) illustrates the success of accessibility at distinguishing between sites with escape mutations or within antibody epitopes and sites with neither - SARS-CoV-2 RBD (top) and full Spike (bottom).

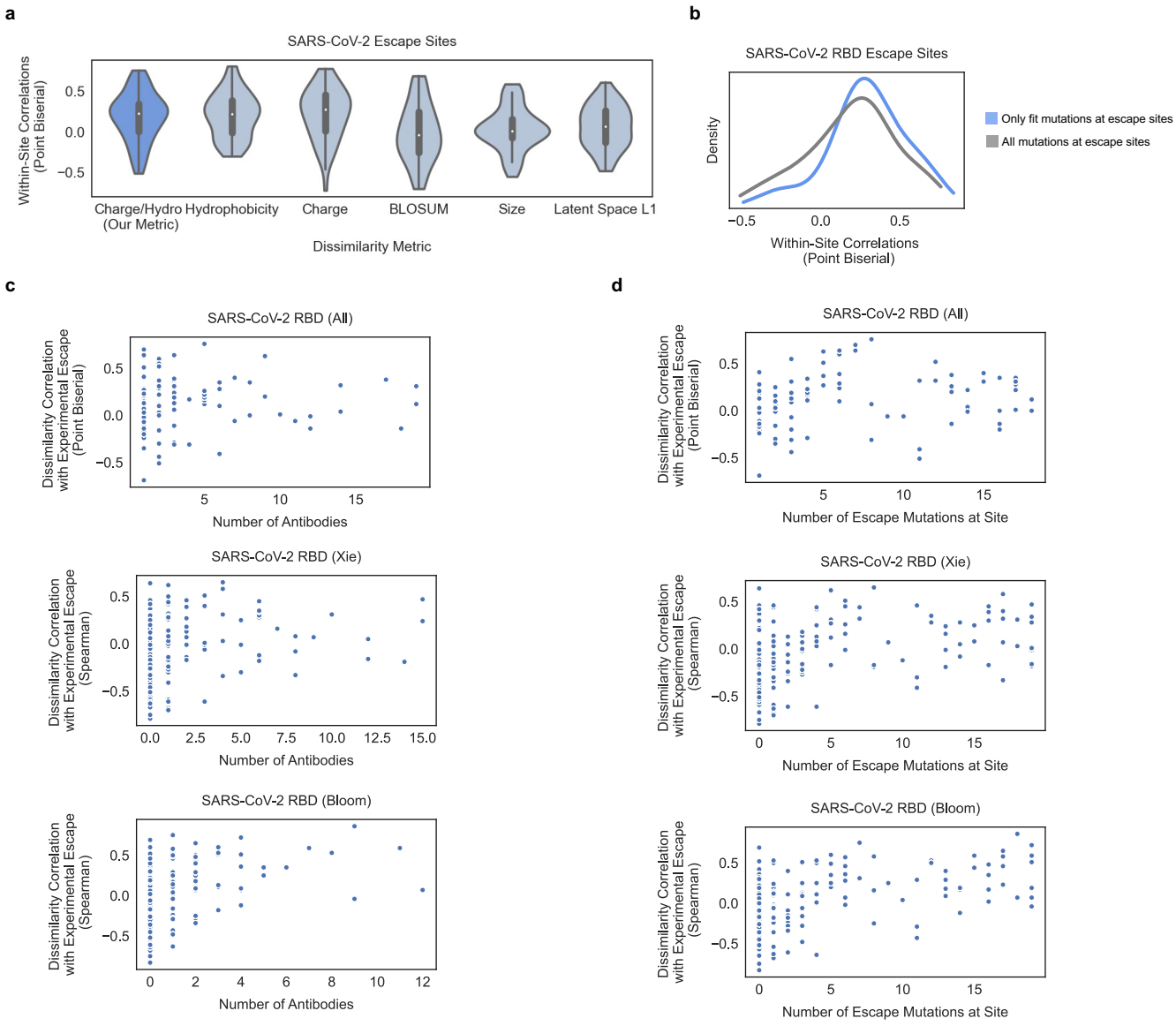


Figure S10: Charge-hydrophobicity metric captures residue dissimilarity relevant for loss of antibody binding. a) Within-site point biserial correlations between residue dissimilarity metrics and SARS-CoV-2 DMS escape data at escape sites (sites with 3-17 escape mutations). More sites have a higher correlation for our charge-hydrophobicity metric than charge or hydrophobicity alone, BLOSUM62, residue size, or EVE latent space (L1) distance. **b**) 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) **c**) Within-site correlations between residue dissimilarity and escape increase when more antibodies have escape mutations at that site. **d**) 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).

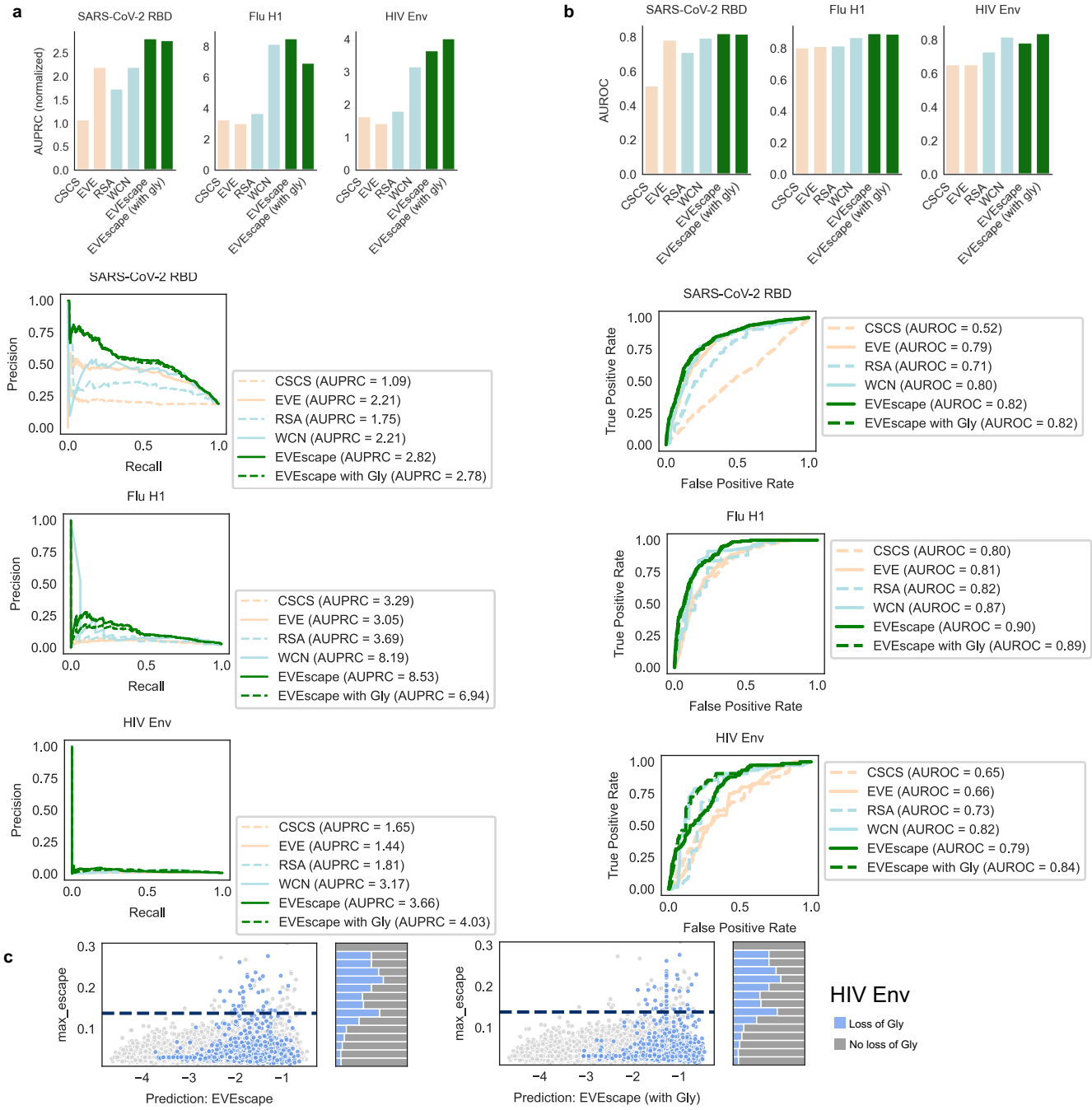


Figure S11: Incorporating glycosylation in EVEscape improves performance on HIV Env. Precision-Recall (with AUPRC normalized by "null" model – fraction of observed escapes) (**a**) and AUROC (**b**) of EVEscape and EVEscape+Gly predicting DMS escape mutations for SARS-CoV-2 RBD, Flu H1, and HIV Env. **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 glycan[11]. 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 types[4]. 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[31–34].

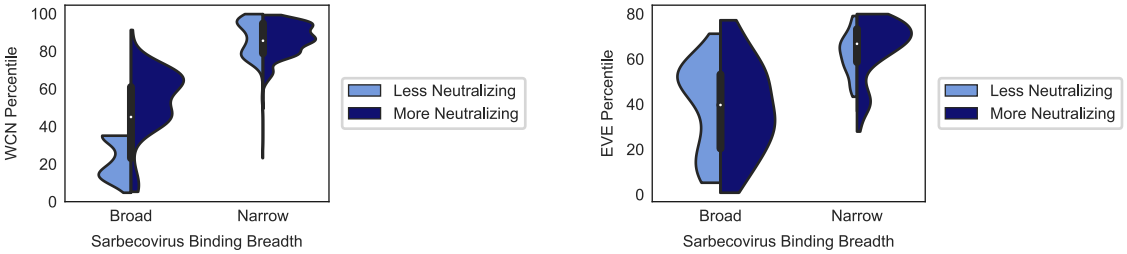


Figure S12: EVEscape prediction of escape mutants to antibodies with varying sarbecovirus breadth and neutralization potency. WCN predicts less escape for broad antibodies with lower neutralization (left). While EVE also predicts less escape for antibodies with broad sarbecovirus binding breadth, EVE does not distinguish between neutralizing and non-neutralizing broad antibodies (right).

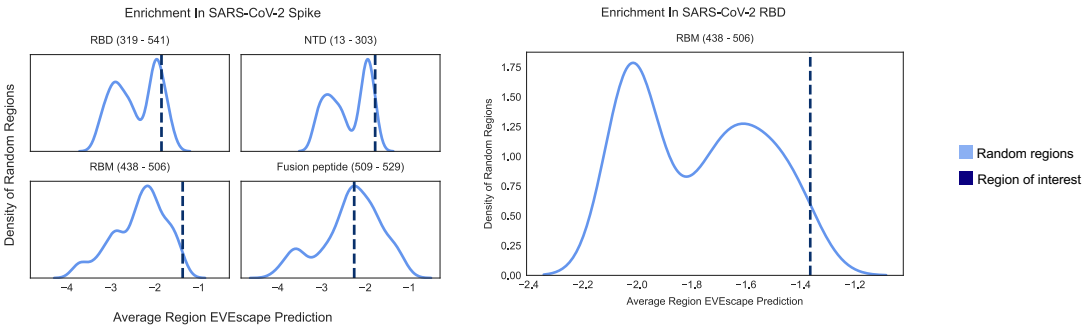


Figure S13: **EVEscape enrichment in regions of SARS-CoV-2 Spike.** RBD (particularly receptor binding motif (RBM)) and N-terminal domain (NTD) have significantly enriched average EVEscape scores, relative to a distribution of 500 random contiguous regions of the same length from full Spike (left). The RBM is significantly enriched within the RBD (from full Spike model), relative to 100 contiguous regions of the same length in the RBD (right).

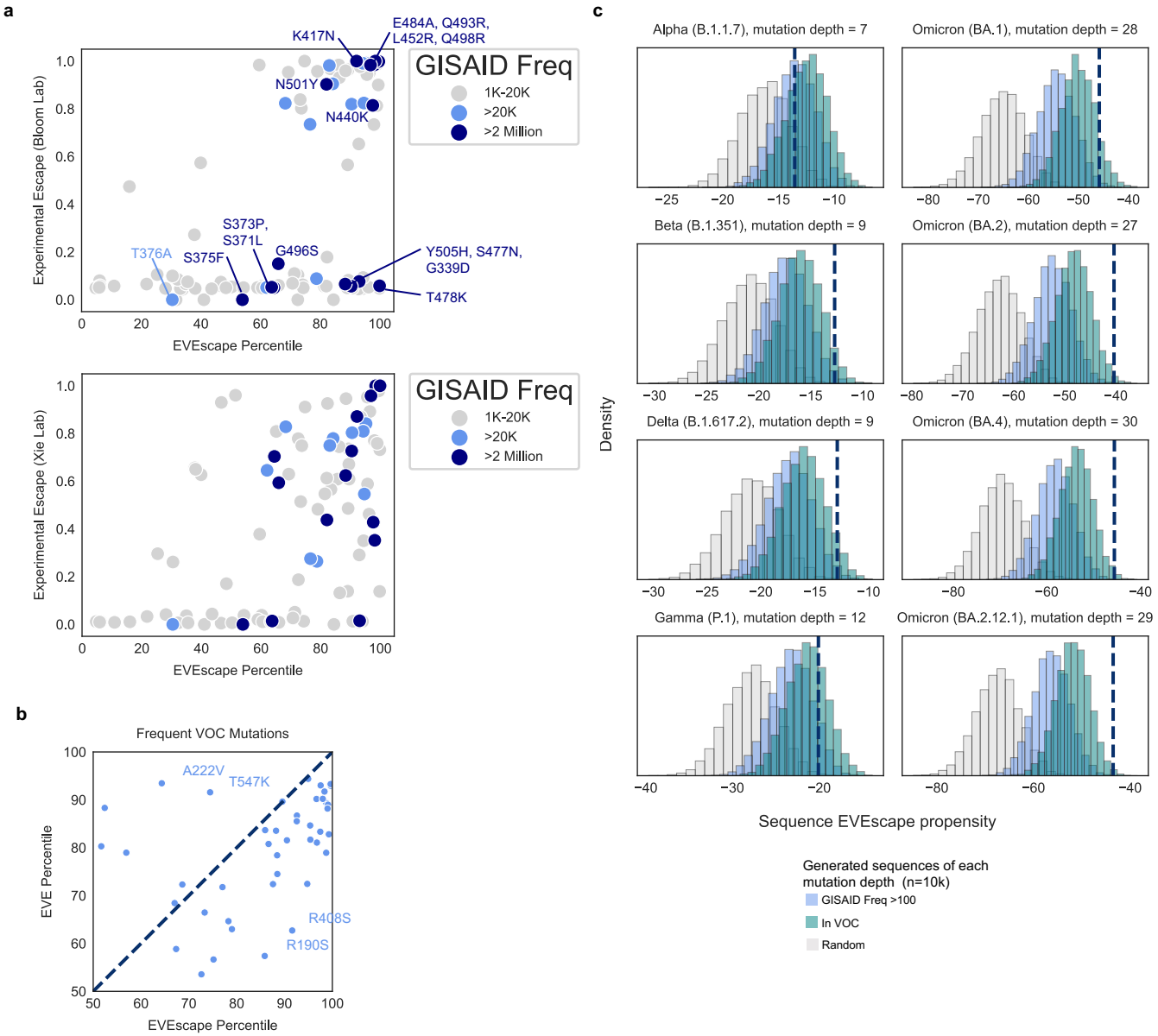


Figure S14: EVEscape anticipates SARS-CoV-2 frequent mutations/VOCs and is more predictive than EVE alone. a) All mutations with observed antibody escape from DMS experiments that are observed in GISAID (>1000 times) have high EVEscape scores. Of observed mutations without experimental escape, EVEscape better captures mutations of very high frequency. **b**) 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[46-48]. **c**) VOCs have high EVEscape scores compared to random mutations at the same mutation depth, particularly Delta and Omicron.

	RBD (+ pandemic data)	**RBD**	Spike (+ pandemic data)	**Spike**
SARS-CoV-2	1398	1	1751	1
SARS-CoV-1	38	24	34	23
Other SARS-like	113	101	115	99
MERS	308	265	316	259
Betacoronavirus 1 (OC43)	610	416	675	394
Alphacoronavirus 1	0	0	529	175
229E	0	0	142	95
NL63	0	0	70	47
HKU1	64	27	65	27
HKU15	0	0	216	141
Avian coronavirus	0	0	4142	581
Porcine epidemic diarrhea virus	0	0	2388	1440
Other coronavirus	255	175	561	347
Other/unknown	0	0	2	0
Total	**2786**	**1009**	**11006**	**3629**

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

Virus	Protein	Study	Strain	Alignment	Assay variable	N	$\rho_{Independent}$	$\rho_{EVmutation}$	ρ_{EVE}
Influenza	H1	Doud 2016[57]	A/WSN/1933	A0A2Z5U3Z0_9INFA_b0.1	replication	10317	0.45	0.45	**0.53**
		Wu 2020[58]	H1 (strain)	A0A6H1V8E8_9PLVG_Y373S_b0.1	replication	10317	0.36	**0.37**	0.36
HIV	Env	Haddox 2018[51]	BG505	A0A192B1T2_9HIV1_b0.1	replication	12388	**0.48**	0.41	**0.48**
			BF520	ENV_HV1B9_S364P-M373R_b0.1	replication	12502	0.48	0.43	**0.49**
		Roop 2020[52]	BG505	A0A192B1T2_9HIV1_b0.1	replication (human cells)	12483	0.48	0.44	**0.49**
					replication (rhesus cells)	12483	0.43	0.40	**0.44**
		Duenas-Decamp 2016[53]	BG505	A0A192B1T2_9HIV1_b0.1	replication	375	0.37	**0.42**	0.38
SARS-CoV-2	Spike RBD	Starr 2020[54]	Wuhan-Hu-1	P0DTC2_321-541_b0.3_pre2020.a2m	yeast expression (RBD)	3798	0.36	0.33	**0.45**
					ACE2 binding	3802	0.23	0.16	**0.26**
		Chan 2021[55]			human cell expression (full Spike)	3458	0.33	0.32	**0.45**
					ACE2 binding	3458	0.31	0.30	**0.42**
	M^{pro}	Flynn 2022[56]	Wuhan-Hu-1	nsp5-YP_009725301_b0.1.a2m	yeast growth	5741	0.58	**0.60**	**0.60**

Table S2: 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) Trimer structure created using structural symmetry in Pymol (adapted from Dingens et al.)[11]
	7TFO	BG505 SOSIP.664 Env (CD4-bound open state)

Table S3: PDB structures capturing diverse protein conformations used for surface accessibility calculations.

	Papers	Assay Details	# of Mutations	# of Escape Mutations (using thresholds from our paper)	# of Antibodies/ Sera	Alignment
SARS-CoV-2 RBD (Wuhan-Hu-1)	**Bloom Lab (antibodies):** Dong 2021[1] Greaney 2021[2] Greaney 2021[4] Greaney 2021[5] Starr 2021[6] Starr 2021[7] Tortorici 2021[8] Starr 2021[9]	FACS-based yeast display screening of antibody binding	3819	635	91	P0DTC2_321-541_b0.3_pre2020.a2m
	Bloom Lab (sera): Greaney 2021[3] Greaney 2021[4] Greaney 2021[5]	FACS-based yeast display screening of sera binding	3819	15	55	P0DTC2_321-541_b0.3_pre2020.a2m
	Xie Lab: Cao 2022[12]	MACS-based yeast display screening of antibody binding	3819	227	247	P0DTC2_321-541_b0.3_pre2020.a2m
Flu H1 (A/WSN/1933)	Doud 2018[10]	Screening viral cell entry in the presence of antibodies	10735	161	6	I4EPC4_t0.99_b0.1.a2m
HIV Env (BG505)	Dingens 2019[11]	Screening viral cell entry in the presence of antibodies	12730	76	8	Q2N0S5_20-709_b0.1_t0.99.a2m

Table S4: Escape DMS data used for EVEscape validation.