Science
AAAS

Supplementary Materials for

Unsupervised evolution of protein and antibody complexes with a structure-informed language model

Varun R. Shanker, Theodora U.J. Bruun, Brian L. Hie, Peter S. Kim

Corresponding authors: bhie@stanford.edu, kimpeter@stanford.edu

The PDF file includes:

Materials and Methods
Figs. S1 to S11
Tables S1 to S3
References

Other Supplementary Materials for this manuscript include the following:

Data S1 to S5
Supplementary Information

Materials and Methods

Structure-informed language model description and scoring of sequences

As input to the model, we provide a protein structure $\mathbf{Y} \in \mathbb{R}^{N\times 3\times 3}$, where N is the number of amino acids, and each amino acid is featurized by the three-dimensional physical coordinates of all three atoms in the protein backbone: the α-carbon, β-carbon, and nitrogen atoms in the protein backbone (hence the dimensionality $N \times 3 \times 3$). The structure-informed language model learns the probability distribution p of a protein sequence $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ (where $\mathcal{X}$ is the alphabet of amino acids) given a structure $\mathbf{Y}$ via the chain rule of probability

$$p(\mathbf{x}|\mathbf{Y}) = p(x_1|\mathbf{Y})p(x_2|x_1, \mathbf{Y}) \ldots p(x_N|x_1, \ldots, x_{N-1}, \mathbf{Y}).$$

The probability distribution at each position is defined over $\mathcal{X}$, such that it is a 20-dimensional vector with all constituent entries summing to 1.

Thus, for a given sequence $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_N)$ and its corresponding given structure $\widehat{\mathbf{Y}}$, we can score the probability of $\hat{\mathbf{x}}$ folding into $\mathbf{Y}$ under the inverse folding model by computing the value of $p(\mathbf{x} = \hat{\mathbf{x}}|\mathbf{Y})$, which we can do autoregressively as

$$p\left(\mathbf{x} = \hat{\mathbf{x}}\middle|\widehat{\mathbf{Y}}\right) = p\left(x_1 = \hat{x}_1\middle|\widehat{\mathbf{Y}}\right) \ldots p\left(x_N = \hat{x}_N\middle|\hat{x}_1, \ldots, \hat{x}_{N-1}, \widehat{\mathbf{Y}}\right).$$

This is evaluated output is a likelihood between 0 and 1, inclusive. The computed score $p\left(\mathbf{x} = \hat{\mathbf{x}}\middle|\widehat{\mathbf{Y}}\right)$ is used as prediction for "fitness" (e.g., binding affinity or enzymatic activity). Importantly, the model does not have any explicit access to "fitness" during either training or evaluation, which we refer to as "zero shot" fitness prediction.

In this work, we also extend the structure-informed language model beyond single chain structures by considering the joint probability all sequences will fold into the backbone coordinates of their corresponding chains together, such that the computed likelihood is evaluated on the entire complex.

We use the inverse folding model checkpoint of ESM-IF1 GVP-Transformer as of April 10, 2022 (*11*).

Diverse proteins benchmarking experiment with scanning mutagenesis data

We analyzed the effectiveness of using the structure-informed language model, ESM-IF1 model to identify high fitness variants from protein mutational scans as a proxy for the ability to guide evolution without explicitly modeling a protein's function. We also compared its performance to ESM-1v, a sequence-only general protein language model. To do so, we used all deep mutational scanning (DMS) datasets from the benchmarking study by Livesey and Marsh (*21*) profiling over 100 variants and reported to have 90% or higher coverage of DMS results across the corresponding curated PDB structure (**Supplementary Table 1**). From this set of 12 proteins, Cas9 was excluded because its sequence length was larger than the maximum allowable length of 1024 amino acids by ESM-1v and ccdB was excluded because the experimental values were discretized within a small range. For each of the 10 mutagenesis datasets, all the sequence likelihood of all variants with coverage in the structure were determined using the structure-informed language model. For ESM-1v, the average masked marginals likelihood score across all five models in the ESM-1v group was used. The experimental data distribution was binarized for high-fitness classification using a percentile-based threshold. The enrichment of high fitness variants was then determined by using the metric of fraction high fitness as defined by the fraction of the top 10 model-predicted variants with experimental values above the high fitness threshold. The analysis was performed at three different percentile thresholds, top 5th percentile (95th percentile), top 10th percentile (90th percentile), and top 20th percentile (80th percentile), to determine sensitivity of the result based on the stringency of the selected cutoff parameter.

Benchmarking of antibody mutagenesis

We use five antibody mutagenesis datasets (*42*, *43*) to benchmark the performance of modeling variant effects on antibody binding using the structure-informed language model against three sequence-only methods, ESM-1v (*41*), AbLang (*47*), and abYsis (*48*). Variant sequences were scored using the model with three different forms of structure input: i) variable region of mutated antibody chain only ii) variable regions of both antibody chains iii) variable regions of both antibody chains in complex with antigen. The autoregressive scoring of sequences enables evaluation of sequences with multiple mutations. The Spearman correlation was determined between the log likelihood scores across all sequences and corresponding reported experimental binding measurements: -log(K_D) for CR9114 and CR6261; log(binding enrichment) g6.31. The following structures were used for input backbone coordinates of the VH, VL, and antigen: PDB 4FQI (*44*), CR9114-H5; PDB 3GBN (*45*), CR6261-H1; PDB 2FJG (*46*), g6.31-VEGF.

ESM-1v, AbLang, and abYsis were scored using the variant sequence of the antibody variable region. For variants with multiple mutations, the average effect of all mutant amino acids in the sequence was considered, namely

$$p(\mathbf{x}) = \frac{1}{|\mathcal{M}|} \sum_{i \in M} [\log\ p(\mathbf{x}_i = \mathbf{x}_i^{\mathrm{mt}})\ -\ \log p(\mathbf{x}_i = \mathbf{x}_i^{\mathrm{wt}})]$$

where $\mathcal{M}$ is defined as the set of all mutations in the input sequence $\mathbf{x}$. For abYsis, individual mutation likelihoods were determined using the frequency of amino acids at each position based on multiple sequence alignment provided by the webtool version 3.4.1 (http://www.abysis.org/abysis/index.html). We aligned VH and VL protein sequences using the default settings provided in the 'Annotate' tool, with the database of 'Homo sapiens' sequences as of April 1, 2023. The contribution of additional sequence context of antigen information and paired antibody chain were also evaluated using ESM-1v to determine if sequence-only general protein language models can learn binding.

Computational benchmarking on the antibody binding datasets was also conducted with ProteinMPNN (*49*), an alternate model for scoring sequences against a target backbone structure. Spearman correlations were computed using the global score across all chains in the input protein complex, identical to the whole-complex scoring strategy used for ESM-IF1.

Acquisition of antibody amino acid substitutions using structure-informed language model and sequence-only protein language models

We select amino acid substitutions recommended by the structure-informed language model to test in our directed evolution campaigns for LY-CoV1404 and SA58. For a given wild-type antibody variable region sequence, $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^N$, where $\mathcal{X}$ is the set of amino acids and N is the sequence length, we score all possible single amino acid substitutions against a corresponding structure of the variable regions of both antibody chains in complex with the RBD of SARS-CoV-2 Spike protein, $\widehat{\mathbf{Y}}$ by computing $p\left(\mathbf{x} = \hat{\mathbf{x}} \middle| \widehat{\mathbf{Y}}\right)$. Protein structures used are reported in Supplementary Table 1. We then select the set of top ten predicted single amino acid substitutions at unique residues in each antibody variable region prior to the final framework segment to test in the first round of evolution.

After testing individual amino acid mutations in a pseudovirus neutralization screen, in Round 2, beneficial mutations (defined as IC_{50} fold-change > 1.1) were combined to assess the combinatorial effects and potential for further neutralization improvement. We tested up to four

combinations of single amino acid mutations on each chain (two total mutations to the antibody). We also used the model to score a library of all possible combinations of the beneficial mutations to an antibody chain (For example, VH LY-CoV1404 has 8 beneficial mutations resulting in 255 total candidate sequences), and selected the top five scoring designs (or less if there were a fewer number of total possible combinations). Lastly, we tested a maximum of two variants consisting of the best single-chain designs together. In total, 31 variants were tested for LY-CoV1404 and 25 variants were tested for SA58.

Antibody variants tested in the first round of evolution in **Supplementary Figure 8** were recommended by the ensemble of protein language models as described in Hie et al (*58*). The top ten predictions were selected, and, if needed, the prediction stringency was decreased to allow for ten predictions. To support a fair comparison to structure-guided evolution, we performed an identical directed evolution campaign using protein language models. We use the same experimental algorithm outlined above to advance mutations to the second round and select a maximum number of round 2 mutations to screen. Combining multiple mutations on a single chain was performed by ranking and acquiring the set of top additive single variants.

Antibody cloning

We cloned the antibody sequences into the CMV/R plasmid backbone for expression under a CMV promoter. The heavy chain or light chain sequence was cloned between the CMV promoter and the bGH poly(A) signal sequence of the CMV/R plasmid to facilitate improved protein expression. Variable regions were cloned into the human IgG1 backbone; LY-CoV1404 variants were cloned with a lambda light chain, whereas variants of SA58 were cloned with a kappa light chain. The vector for both heavy and light chain sequences also contained the HVM06_Mouse (UniProt: P01750) Ig heavy chain V region 102 signal peptide (MGWSCIILFLVATATGVHS) to allow for protein secretion and purification from the supernatant. VH and VL segments were ordered as gene blocks from Integrated DNA Technologies and were cloned into linearized CMV/R backbones with 5× In-Fusion HD Enzyme Premix (Takara Bio).

Antigen cloning

RBD sequences were cloned into a pADD2 vector between the rBeta-globin intron and β-globin poly(A). All RBD constructs contain an AviTag and 6×His tag. RBD sequences were based off wild-type Wuhan-Hu-1 (GenBank: BCN86353.1), Omicron BA.1 (GenBank: UFO69279.1), BQ.1.1 (GenBank: OP412163.1), XBB.1.5 (GenBank: OP790748.1).

DNA preparation

Plasmids were transformed into Stellar competent cells (Takara Bio), and transformed cells were plated and grown at 37 °C overnight. Colonies were mini-prepped per the manufacturer's recommendations (GeneJET, K0502, Thermo Fisher Scientific) and sequence confirmed (Sequetech) and then maxi-prepped per the manufacturer's protocols (ZymoPure II Plasmid Maxiprep Kit, Zymo Research). Plasmids were sterile filtered using a 0.22-µm syringe filter and stored at 4 °C.

Protein expression

All proteins were expressed in Expi293F cells (Thermo Fisher Scientific, A14527). Proteins containing a biotinylation tag (AviTag) were also expressed in the presence of a BirA

enzyme, resulting in spontaneous biotinylation during protein expression. Expi293F cells were cultured in media containing 66% FreeStyle/33% Expi media (Thermo Fisher Scientific) and grown in TriForest polycarbonate shaking flasks at 37 °C in 8% carbon dioxide. The day before transfection, cells were pelleted by centrifugation and resuspended to a density of 3×10^6 cells per milliliter in fresh media. The next day, cells were diluted and transfected at a density of approximately 3–4 × 10^6 cells per milliliter. Transfection mixtures were made by adding the following components: maxi-prepped DNA, culture media and FectoPRO (Polyplus) would be added to cells to a ratio of 0.5 µg: 100 µl: 1.3 µl: 900 µl. For example, for a 100-ml transfection, 50 µg of DNA would be added to 10 ml of culture media, followed by the addition of 130 µl of FectoPRO. For antibodies, we divided the transfection DNA equally among heavy and light chains; in the previous example, 25 µg of heavy chain DNA and 25 µg of light chain DNA would be added to 10 ml of culture media. After mixing and a 10-min incubation, the example transfection cocktail would be added to 90 ml of cells. The cells were harvested 3–5 days after transfection by spinning the cultures at 10,000*g* for 10 min. Supernatants were filtered using a 0.45-µm filter.

Antibody purification

We purified antibodies using a 5-ml MabSelect Sure PRISM column on the ÄKTA pure fast protein liquid chromatography (FPLC) instrument (Cytiva). The ÄKTA system was equilibrated with line A1 in 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.4, 150 mM sodium chloride (NaCl), line A2 in 100 mM glycine pH 2.8, line B1 in 0.5 M sodium hydroxide, Buffer line in 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.4, 150 mM sodium chloride (NaCl) and Sample lines in water. The protocol washes the column with A1, followed by loading of the sample in the Sample line until air is detected in the air sensor of the sample pumps, followed by five column volume washes with A1, elution of the sample by flowing of 20 ml of A2 directly into a 50-ml conical containing 2 ml of 1 M tris(hydroxymethyl)aminomethane (Tris) pH 8.0, followed by five column volumes of A1, B1 and A1 and then a wash step of the fraction collector with A1. We concentrated the eluted samples using 50-kDa cutoff centrifugal concentrators, followed by buffer exchange using a PD-10 column (Sephadex) that had been pre-equilibrated into 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.4, 150 mM sodium chloride (NaCl). Purified antibodies were used directly in experiments or flash-frozen and stored at −20 °C.

Antigen purification

All RBD antigens were His-tagged and purified using HisPur Ni-NTA resin (Thermo Fisher Scientific, 88222). Cell supernatants were diluted with 1/3 volume of wash buffer (20 mM imidazole, 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.4, 150 mM sodium chloride (NaCl), and the Ni-NTA resin was added to diluted cell supernatants. For all antigens, the samples were then incubated at 4 °C while stirring overnight. Resin/supernatant mixtures were added to chromatography columns for gravity flow purification. The resin in the column was washed with wash buffer (20 mM imidazole, 20 mM HEPES pH 7.4, 150 mM NaCl), and the proteins were eluted with 250 mM imidazole, 20 mM HEPES pH 7.4, 150 mM NaCl. Column elutions were concentrated using centrifugal concentrators at 10-kDa cutoff, followed by size-exclusion chromatography on an ÄKTA pure system (Cytiva). ÄKTA pure FPLC with a Superdex 200 Increase (S200) gel filtration column was used for purification. Then, 1 ml of sample was injected using a 2-ml loop and run over the S200, which

had been pre-equilibrated in degassed 20 mM HEPES, 150 mM NaCl before use and flash-frozen before storage at −20 °C.

BLI binding experiments

All reactions were run on an Octet RED96 at 30 °C, and samples were run in 1× PBS with 0.1% BSA and 0.05% Tween 20 (Octet buffer). IgGs were assessed for binding to biotinylated antigens using streptavidin biosensors (Sartorius/ForteBio). Antigen was loaded at a concentration of 200nM. Tips were then washed and baselined in wells containing only Octet buffer. Samples were then associated in wells containing IgG at 100 nM concentration. A control well with loaded antigen but that was associated in a well containing only 200 µl of Octet buffer was used as a baseline subtraction for data analysis. Association and dissociation binding curves were fit in Octet System Data Analysis Software version 9.0.0.15 using a 1:2 bivalent model for IgGs to determine apparent K_d. Fold-change in apparent K_d were determined by computing the ratio of wildtype K_d to variant K_d. Averages of K_d fold-change values from at least two independent experiments are reported to two significant figures in **Supplementary Data 2**. To estimate measurement error, we computed the standard deviation for each antibody−antigen K_d pair.

Polyspecificity Particle assay

Polyspecificity reagent (PSR) was obtained as described by Xu et al(*74*). Soluble membrane proteins were isolated from homogenized and sonicated Expi 293F cells followed by biotinylation with Sulfo-NHC-SS-Biotin (Thermo Fisher Scientific, 21331) and stored in PBS at −80 °C. The PolySpecificity Particle (PSP) assay was performed as described in Makowski et al.(*75*). Protein A magnetic beads (Invitrogen, 10001D) were washed three times in PBSB (PBS with 1 mg ml^{-1} BSA) and diluted to 54 µg ml^{-1} in PBSB. Then, 30 µl of the solution containing the beads was incubated with 85 µl of antibodies at 15 µg ml^{-1} overnight at 4 °C with rocking. The coated beads were then washed twice with PBSB using a magnetic plate stand (Invitrogen, 12027) and resuspended in PBSB. We then incubated 50 µl of 0.1 mg ml^{-1} PSR with the washed beads at 4 °C with rocking for 20 min. Beads were then washed with PBSB and incubated with 0.001× streptavidin-APC (BioLegend, 405207) and 0.001× goat anti-human Fab fragment FITC (Jackson ImmunoResearch, 109-097-003) at 4 °C with rocking for 15 min. Beads were then washed and resuspended with PBSB. Beads were profiled via flow cytometry using a Sony SH800 cell sorter. Data analysis was performed with FlowJo software version 10.9.0 to obtain median fluorescence intensity (MFI) values, which are reported for each antibody across three or more replicate wells. Elotuzumab (Fisher Scientific) and ixekizumab (Fisher Scientific) are also included in each assay as controls.

Lentivirus production

We produced SARS-CoV-2 Spike (Wuhan-Hu-1, BA.1, and BQ.1.1 variants) pseudotyped lentiviral particles. Viral transfections were done in HEK293T cells (American Type Culture Collection, CRL-3216) using BioT (BioLand) transfection reagent. Six million cells were seeded in D10 media (DMEM + additives: 10% FBS, L-glutamate, penicillin, streptomycin and 10 mM HEPES) in 10-cm plates one day before transfection. A five-plasmid system was used for viral production, as described in Crawford et al(*76*). The Spike vector contained the 21-amino-acid truncated form of the SARS-CoV-2 Spike sequence from the Wuhan-Hu-1 strain of SARS-CoV-2 (GenBank: BCN86353.1), BA.1 variant of concern

(GenBank: OL672836.1), or BQ.1.1 variant of concern (GenBank: OP412163.1. The other viral plasmids, used as previously described(*76*), are pHAGE-Luc2-IRS-ZsGreen (NR-52516), HDM-Hgpm2 (NR-52517), pRC-CMV-Rev1b (NR-52519) and HDM-tat1b (NR-52518). These plasmids were added to D10 medium in the following ratios: 10 µg pHAGE-Luc2-IRS-ZsGreen, 3.4 µg FL Spike, 2.2 µg HDM-Hgpm2, 2.2 µg HDM-Tat1b and 2.2 µg pRC-CMV-Rev1b in a final volume of 1,000 µl.

After adding plasmids to medium, we added 30 µl of BioT to form transfection complexes. Transfection reactions were incubated for 10 min at room temperature, and then 9 ml of medium was added slowly. The resultant 10 ml was added to plated HEK cells from which the medium had been removed. Culture medium was removed 24 h after transfection and replaced with fresh D10 medium. Viral supernatants were harvested 72 h after transfection by spinning at 300*g* for 5 min, followed by filtering through a 0.45-µm filter. Viral stocks were aliquoted and stored at −80 °C.

Pseudovirus neutralization

The target cells used for infection in SARS-CoV-2 pseudovirus neutralization assays are from a HeLa cell line stably overexpressing human angiotensin-converting enzyme 2 (ACE2) as well as the protease known to process SARS-CoV-2: transmembrane serine protease 2 (TMPRSS2). Production of this cell line is described in detail by Rogers et al (*77*). with the addition of stable TMPRSS2 incorporation. ACE2/TMPRSS2/HeLa cells were plated 1 day before infection at 8,000 cells per well. Ninety-six-well, white-walled, white-bottom plates were used for neutralization assays (Thermo Fisher Scientific).

On the day of the assay, purified IgGs in 1× PBS were made into D10 medium (DMEM + additives: 10% FBS, L-glutamate, penicillin, streptomycin and 10 mM HEPES). A virus mixture was made containing the virus of interest (for example, SARS-CoV-2) and D10 media. Virus dilutions into media were selected such that a suitable signal would be obtained in the virus-only wells. A suitable signal was selected such that the virus-only wells would achieve a luminescence of at least >1,000,000 relative light units (RLU). Then, 60 µl of this virus mixture was added to each of the antibody dilutions to make a final volume of 120 µl in each well. Virus-only wells were made, which contained 60 µl of D10 and 60 µl of virus mixture. Cells-only wells were made, which contained 120 µl of D10 media.

The antibody/virus mixture was left to incubate for 1 h at 37 °C. After incubation, the medium was removed from the cells on the plates made one day prior. This was replaced with 100 µl of antibody/virus dilutions and incubated at 37 °C for approximately 48 h. Infectivity readout was performed by measuring luciferase levels. Medium was removed from all wells, and cells were lysed by the addition of 100 µl of BriteLite assay readout solution (PerkinElmer) into each well. Luminescence values were measured using an Infinite 200 PRO Microplate Reader (Tecan) using i-control version 2.0 software (Tecan) after shaking for 30 sec. Each plate was normalized by averaging the cells-only (0% infection) and virus-only (100% infection) wells. Neutralization titer was defined as the sample dilution at which the RLU was decreased by 50% as compared with the RLU of virus-only control wells after subtraction of background RLUs in wells containing cells only. Normalized values were fitted with a three-parameter nonlinear regression inhibitor curve in GraphPad Prism 9.1.0 to determine the half-maximal inhibitory concentration (IC_{50}). Antibody variant half-maximal inhibitory concentrations are compared to per-assay wildtype controls to compute fold-changes. Both average half-maximal inhibitory concentrations and average half-maximal inhibitory concentration fold-changes are

reported in **Supplementary Data 1**. Neutralization assays were performed in biological duplicates with technical duplicates.

Computing frequency of changes to antibody protein sequences

We computed the frequency of residues involved in affinity-enhancing substitutions using the abYsis webtool, which also computes the frequency of amino acids at each position based on a multiple sequence alignment. We aligned VH and VL protein sequences using the default settings provided in the 'Annotate' tool, using the database of 'All' sequences as of April 1, 2023. We also used the Kabat region definition provided by abYsis webtool version 3.4.1 to annotate the framework regions and CDRs within the VH and VL sequences which are reported in **Supplementary Table 3**.

Comparing efficiency of machine learning-guided directed evolution methods

To compare the performance of our experimental campaigns with the structure-informed language model against other machine learning methods for protein evolution, we compared the fraction of variants tested in the protein engineering campaign to the number of assay-labeled training data points used to inform the predictions. Data was sourced from Biswas et al. (*63*) and made contemporaneous by the addition of recently published studies as indicated in **Supplementary Data 5**. The fraction improved, or hit rate, refers to experimentally tested predictions which have improved functional activity relative to either a wildtype protein that is used as a starting point for directed evolution or the protein used as a reference template for design.

Supplementary Figures and Tables

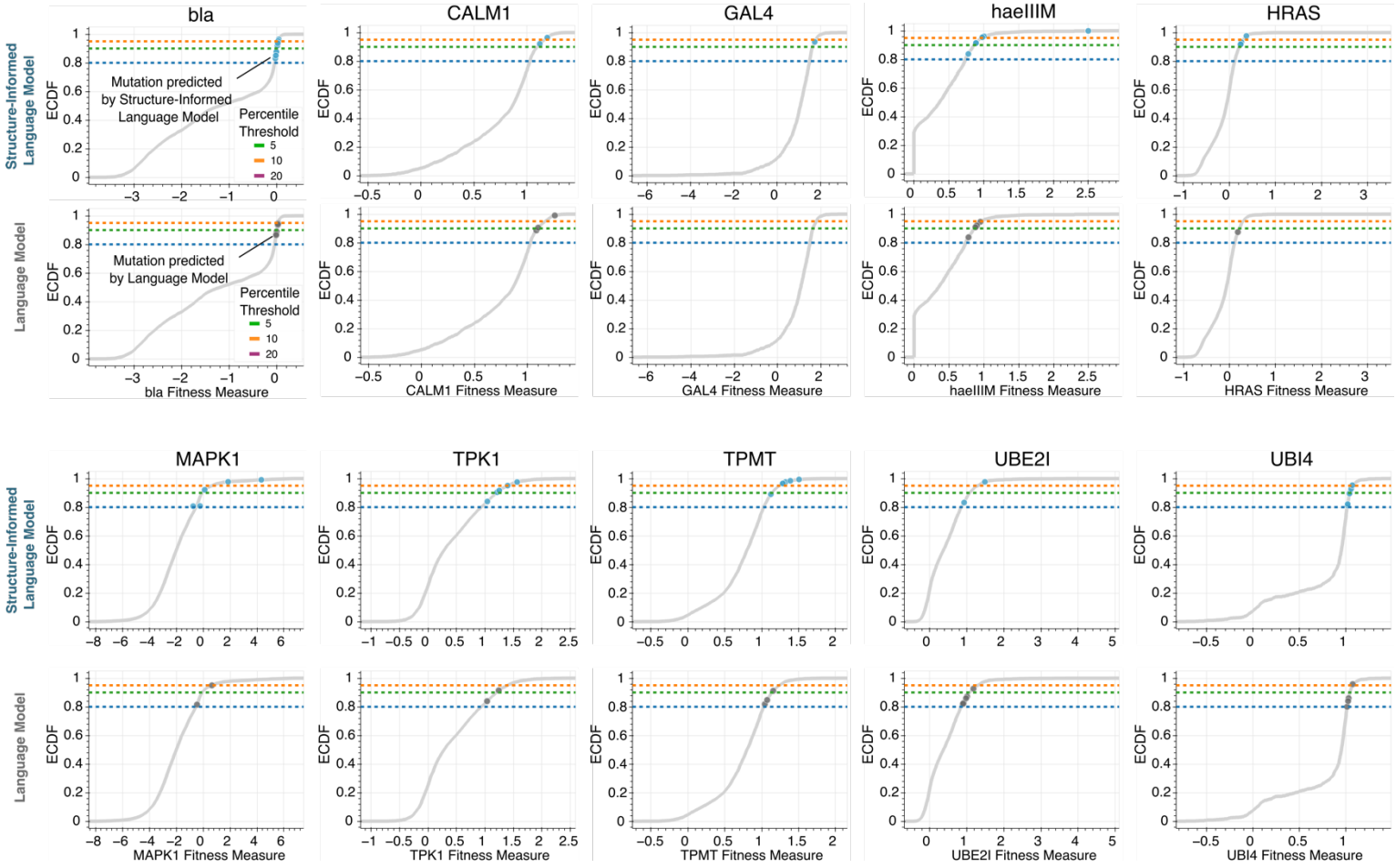


Fig. S1. Evolutionary prediction with sequence-informed language model identifies high fitness variants across proteins with diverse functions

In addition to higher hit rates of high fitness variants, the structure-informed language model generally identifies variants with greater magnitude of improvements in fitness. The top ten predicted variants with experimental fitness values ranking in the 20th percentile of all variants profiled in the deep mutational screen are shown. The grey curve shows the empirical cumulative distribution function (ECDF) of all experimental fitness values determined in the screen. The dotted lines correspond to the three percentile-based thresholds used in the sensitivity analysis (**Figure 1d)** to classify high fitness variants. bla, Beta-lactamase TEM; CALM1, Calmodulin-1; haeIIIM, Type II methyltransferase M.HaeIII; HRAS, GTPase HRas; MAPK1, Mitogen-activated protein kinase; TMPT, Thiopurine S-methyltransferase; TPK1, Thiamin pyrophosphokinase 1; UBI4, Polyubiquitin; UBE2I, SUMO-conjugating enzyme UBC9

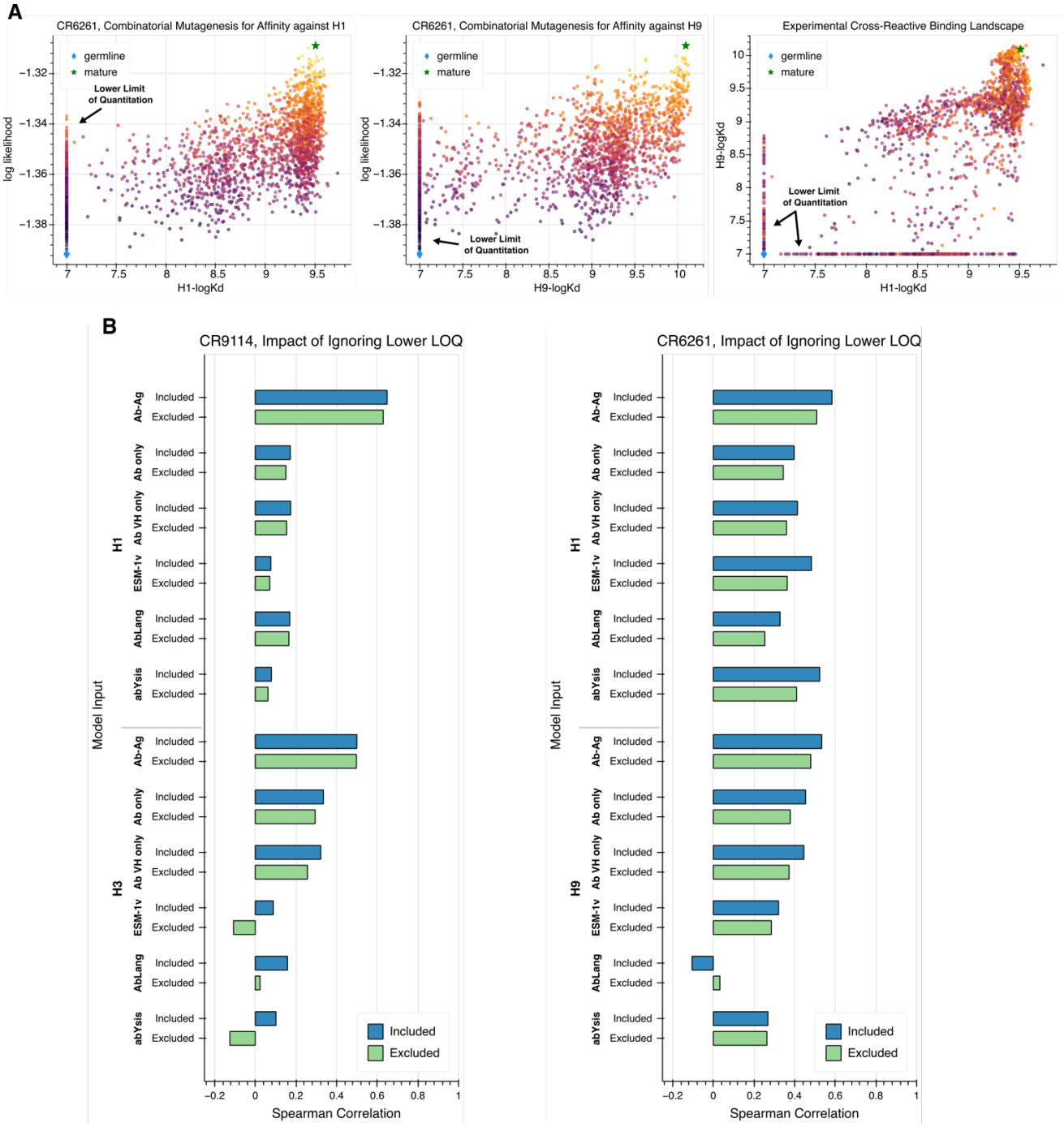


Fig. S2. Impact of lower limit of quantitation of binding assay on predictive performance
(A) Scatter plots showing CR6261 variant sequences scored with the structure-informed language model compared to experimental binding data and inclusive of the assay's lower limit of quantitation, which is omitted for visualization in **Figure 3b**. **(B)** Comparative bar plots showing the impact of removing sequences with experimental measurements bounded artificially by the assay to dataset-wide correlation. While Spearman correlations shown in Figure 3a are computed without any modification to the data, trends in prediction and comparison among modeling methods are robust to filtering sequences affected by this assay artifact.

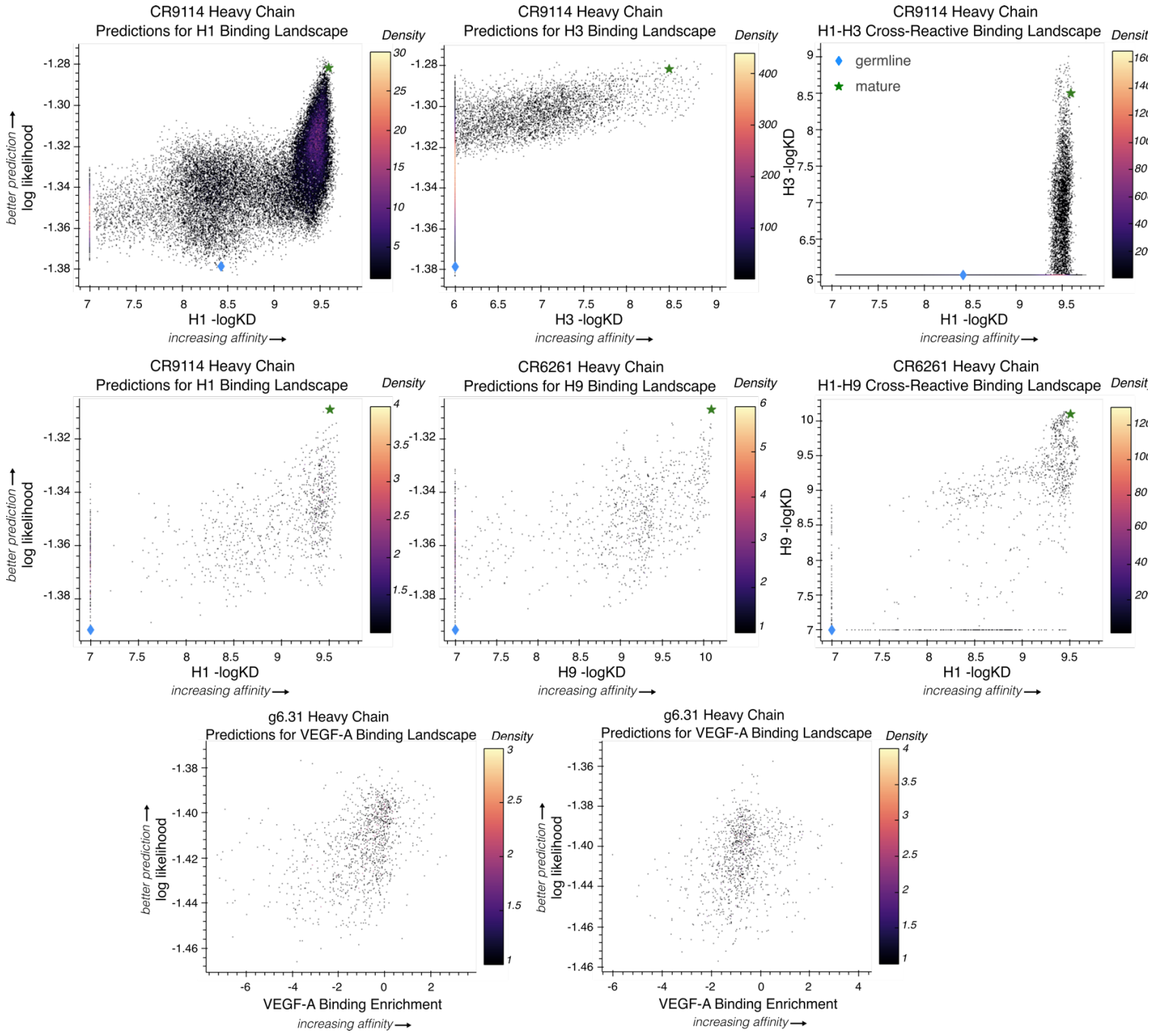


Fig. S3. Primary antibody mutagenesis data with structure-informed language model log likelihood predictions

Scatter plots showing the sequence log likelihood prediction and corresponding experimental binding measures for each sequence tested used to compute Spearman correlations in **Figure 2a** 'Ab-Ag' condition. For CR9114 and CR6261 (top and middle), the final column (right) shows the cross-reactive binding landscape, that is experimental values are plotted on both axes. Points are colored based on the density of the rasterized plot, which was used to accommodate the large number of sequences interrogated within these datasets.

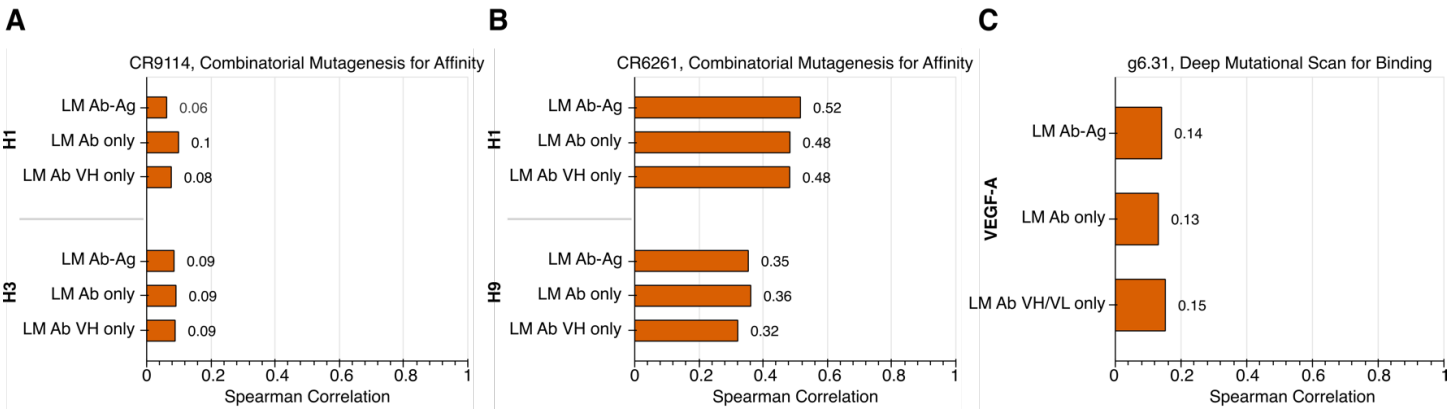


Fig. S4. Evaluating the ability for language models to learn binding with additional sequence context

Bar plots showing the Spearman correlation between ESM-1v language-model (LM) predictions for mutational landscapes of **(A)** CR9114 **(B)** CR6261 and **(C)** g6.31 and experimentally determined measurements of binding to the indicated antigens using. Language model variant prediction was evaluated in three different settings: i) providing the entire antibody variable region and antigen complex (Ab-Ag) ii) providing only the antibody variable region (Ab only), and iii) providing only the single antibody variable region of the chain responsible for binding or being mutated (Ab VH only or Ab VH/VL only). In contrast to performance improvements in binding predictions when antigen information is provided for the structure-informed language model, no benefits are observed with sequence-only general protein language models.

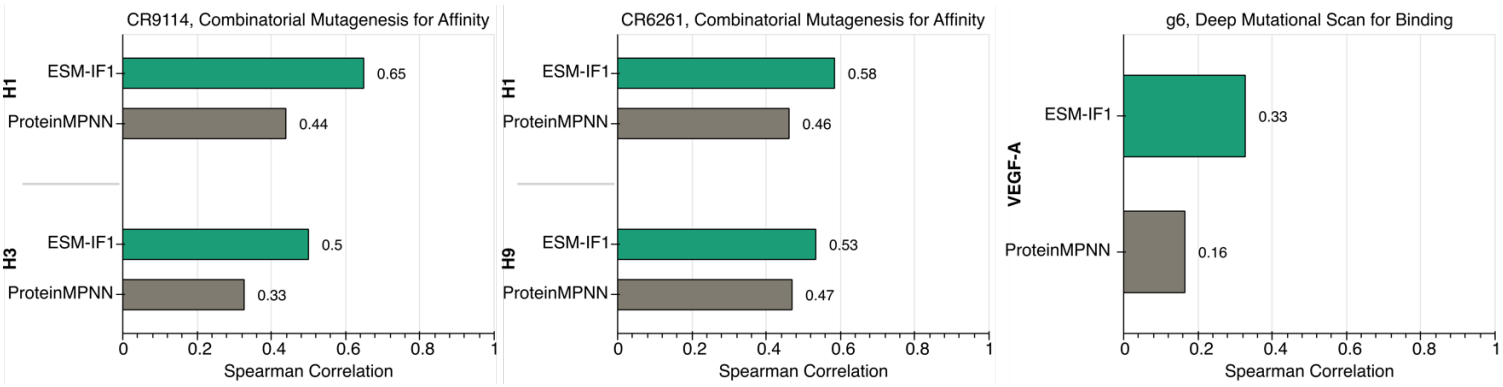


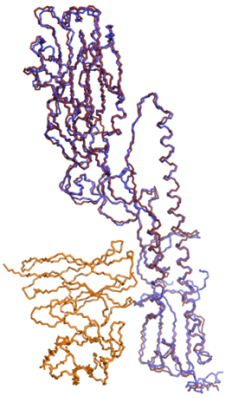
Fig. S5. Comparison of antibody binding prediction to an alternate model for structure-based sequence scoring

The structure-informed language model, ESM-IF1, performs better for antibody binding prediction than the message-passing neural network-based model, ProteinMPNN (*49*), which can also be used to score sequences for a given target protein structure. Notably, ESM-IF1 is only trained on single chain protein structures while the training dataset for ProteinMPNN includes multichain protein complexes. Spearman correlations between model prediction scores and experimental binding data are shown for each of the datasets presented in **Figure 2**. Both models were evaluated using the same input target structures and scoring method (**Methods**).

A

CR9114, PDB 4FQI
HA of H5N1 influenza, PDB 4FQI
HA of H1N1 influenza, PDB 7MFG
Backbone RMSD =1.07Å (1290 atoms)

B



CR6261, PDB 3GBN
HA of H1N1 influenza, PDB 3GBN
HA of H1N1 influenza, PDB 7MFG
Backbone RMSD =0.76Å (1326 atoms)

C

CR9114

Strain	HA Subtype	Use in Fig.2	Sequence Identity (%)	Epitope Identity
Influenza A/New Caledonia/20/1999 (H1N1)	H1	Experimental Screening	63.1	14/21
Influenza A/Wisconsin/67/2005 (H3N2)	H3	Experimental Screening	40.2	10/21
Sampled with Inverse Folding; T=1e-6	H5	-	53.3	12/21
Influenza A/Vietnam/1203/2004 (H5N1)	H5	Target Structure	-	-

CR6261

Strain	HA Subtype	Use in Fig.2	Sequence Identity (%)	Epitope Identity
Influenza A/New Caledonia/20/99 (H1N1)	H1	Experimental Screening	86.9	21/21
Influenza A/Hong Kong/1073/1999 (H9N2)	H9	Experimental Screening	53.9	13/21
Sampled with Inverse Folding; T=1e-6	H1	-	57.5	17/21
Influenza A/Brevig Mission/1/1918 (H1N1)	H1	Target Structure	-	-

Fig. S6. Structural and sequence similarity of antigens used in antibody protein complexes **(A)** For cross-reactive antibodies, inclusion of the antigen structure is informative even for predicting binding to a different antigen. In **Figure 2a**, we report a correlation of 0.65 between structure-informed language model log likelihoods of CR9114 variants and experimental affinity measurements to H1 despite using a structure solved with CR9114 in complex with H5. We use both the protein sequence and backbone structure coordinates of the entire complex as input. Across both HA subunits, H5 and H1 have considerable sequence differences, yet only 1.07 Å root mean square deviation (RMSD) across the entire protein backbone. **(B)** Comparison of antigenic structural similarity between the target protein complex structure used as input (PDB 3GBN (*45*)) to compute predictions and structure of H1 from the strain experimentally tested (PDB 7MFG (*78*)), **(C)** Table summarizing the sequence identity of residues composing the antibody epitope on HA and entire protein sequence for relevant strains used in this study. Additionally, sequence sampling with a low temperature was performed to assess recovery of native-like HA sequences. Sequence recovery of sequences sampled for HA, given an input protein complex with the corresponding bnAb, is within the range of natural influenza HA sequences.

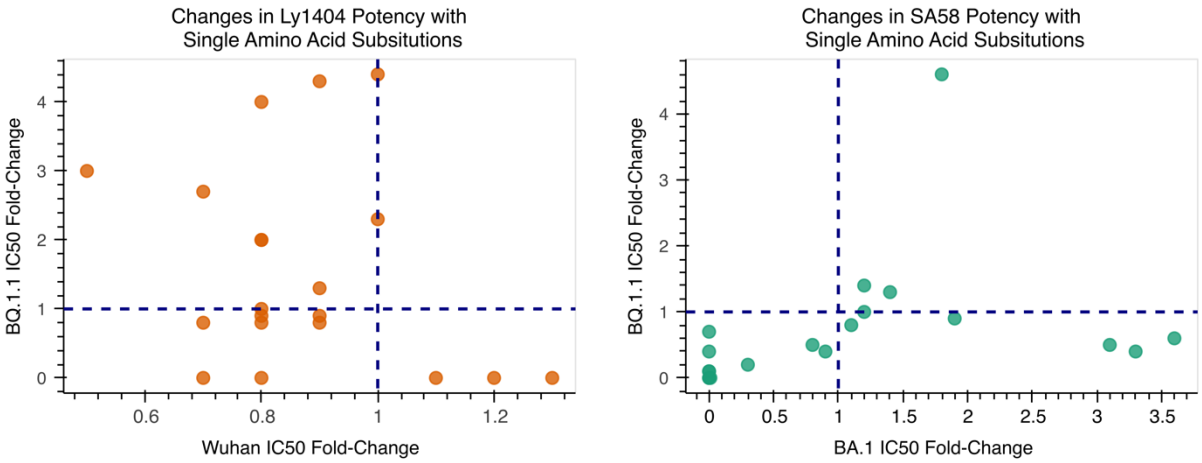


Fig. S7. Functional diversity of structure-informed language model-recommended mutations

Among the 20 single amino acid substitutions tested for LY-CoV1404, 12 of 20 = 60% improve neutralization against at least one of the two strains tested. Similarly, 9 of 20 = 45% of the single amino acid substitutions tested for SA58 improve neutralization. While some variants improve function against both pseudovirus strains, others overwhelmingly only improve against one. This suggests that focusing sequence exploration to structurally compatible mutations does not compromise functional diversity.

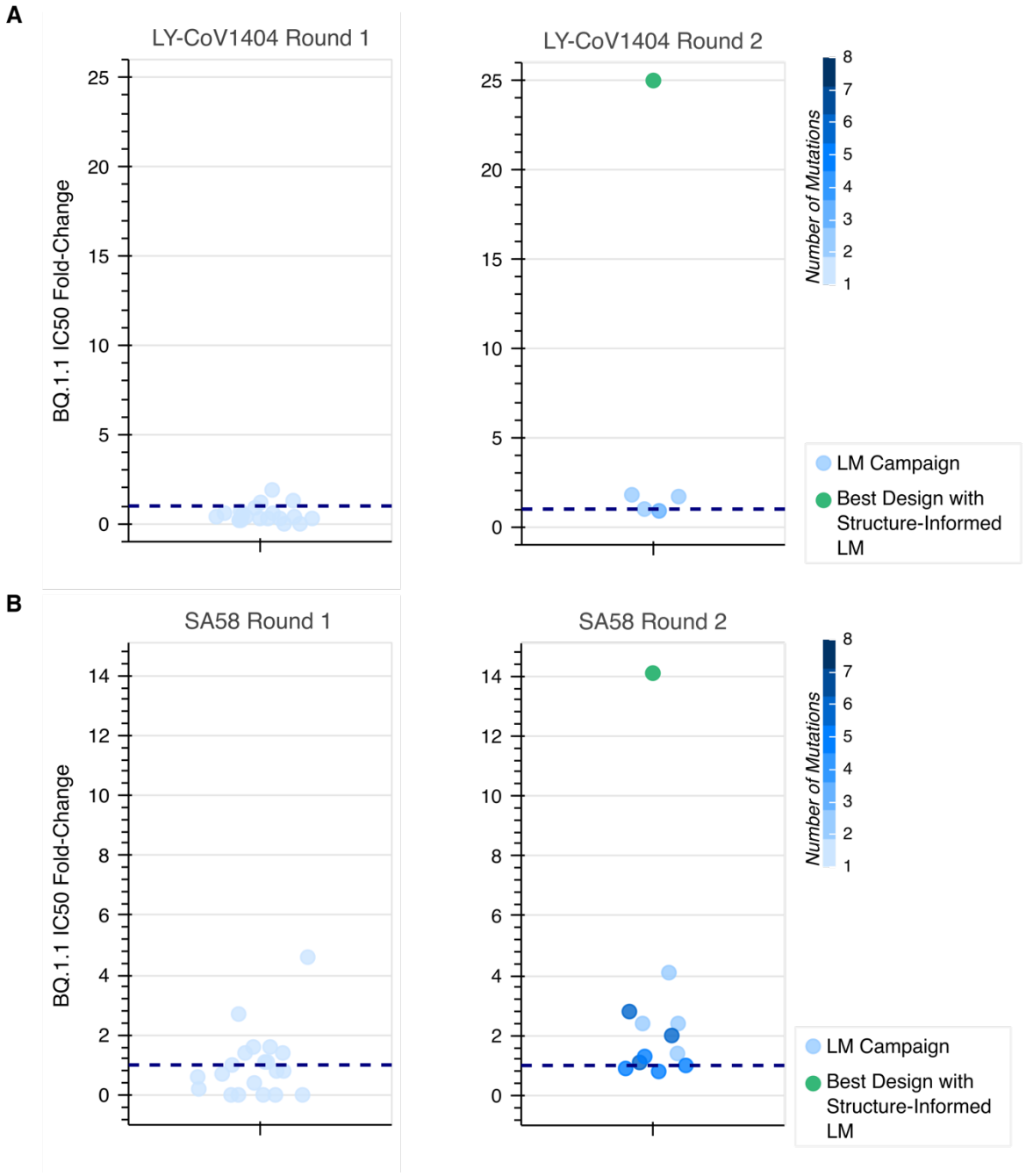


Fig. S8. Sequence-only language model guided evolution of LY-CoV1404 and SA58
Strip plots showing two rounds of directed evolution for antibodies **(A)** LY-CoV1404 and **(B)** SA58 using the same experimental algorithm with variants recommended by an ensemble of sequence-only protein language models (*58*), a method which has previously been demonstrated to improve antibody affinity. The top final design using the structure-informed language model (shown in green), as presented in **Figure 3c**, has substantially greater order of magnitude improvements to the final design achieved with the language models. Given the improved experimental outcomes in comparison to the competitive baseline, these results strongly support the use of incorporating structural information.

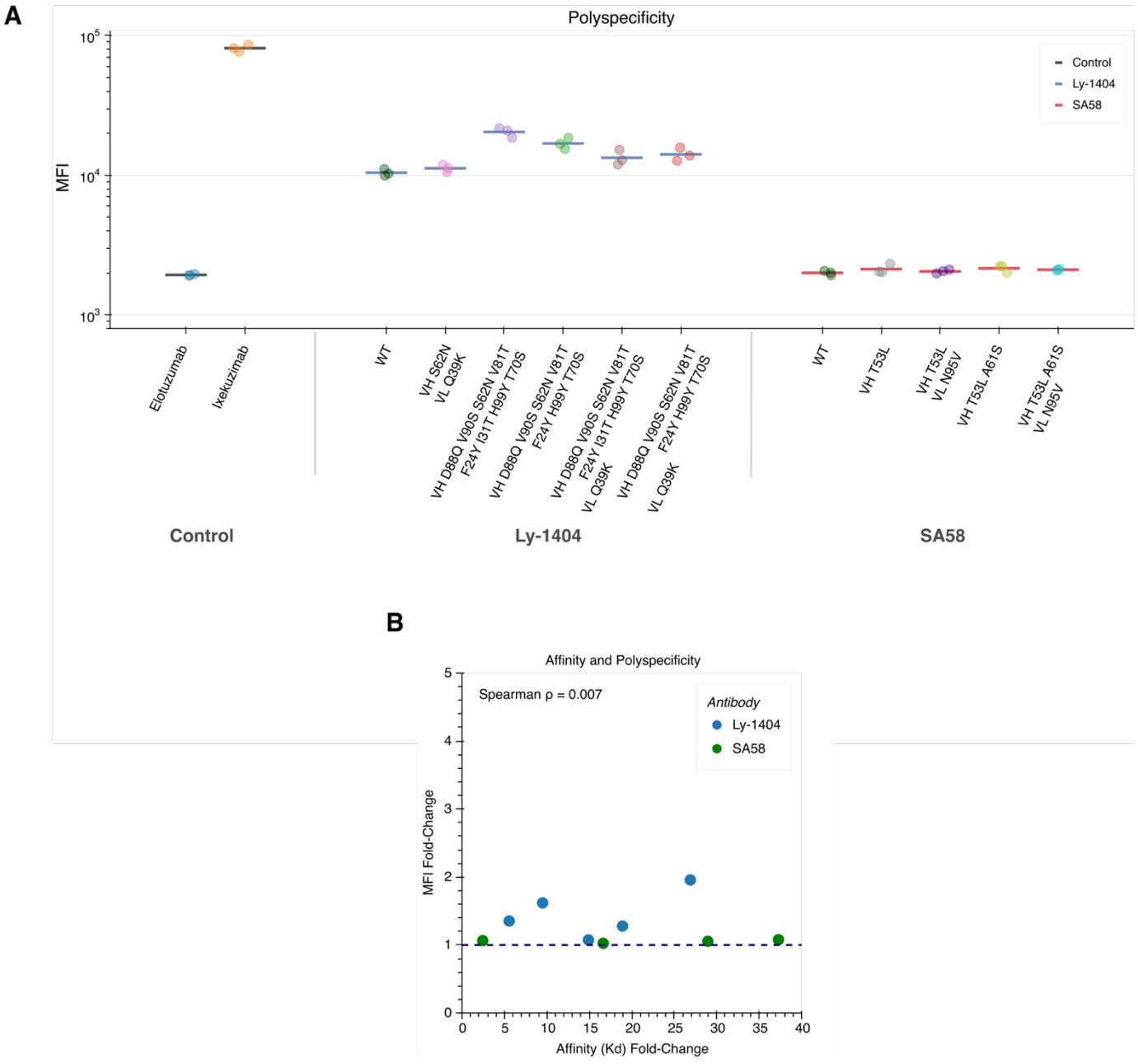


Fig. S9. Polyspecificity of evolved antibodies
(A) The median fluorescence intensity (MFI) signal obtained from flow cytometry is shown for several evolved antibodies with improved affinity and compared to two clinical monoclonal antibodies with high and low polyspecificity used to define a clinically viable range. **(B)** Fold-change in polyspecificity signal is plotted against fold-change in affinity as IgG against BQ.1.1 for LY-CoV1404 and XBB.1.5 for SA58. There is no correlation between the improvements in on-target improvements in affinity and off-target nonspecific changes in polyspecificity (Spearman $\rho = 0.007$).

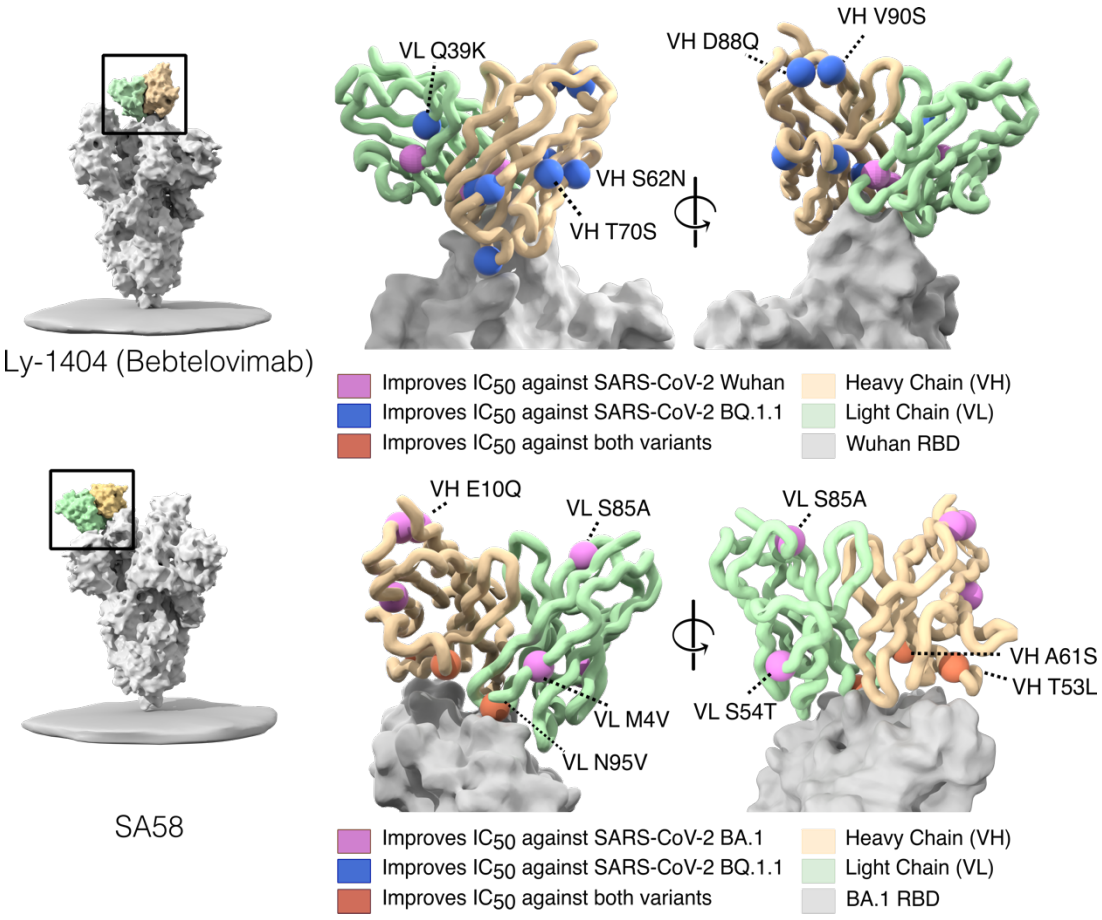


Fig. S10. Mapping neutralization-enhancing substitutions
Neutralization-enhancing mutations are labeled on the structure of the wild-type antibody in complex with the RBD of SARS-CoV-2 spike protein (LY-CoV1404: PDB 7MMO (*50*); SA58: PDB 7Y0W (*54*)). Notably, several mutations are identified to have significant beneficial impacts on binding neutralization and affinity (**Supplementary Data 1 & 2)** despite located away from the binding interface.

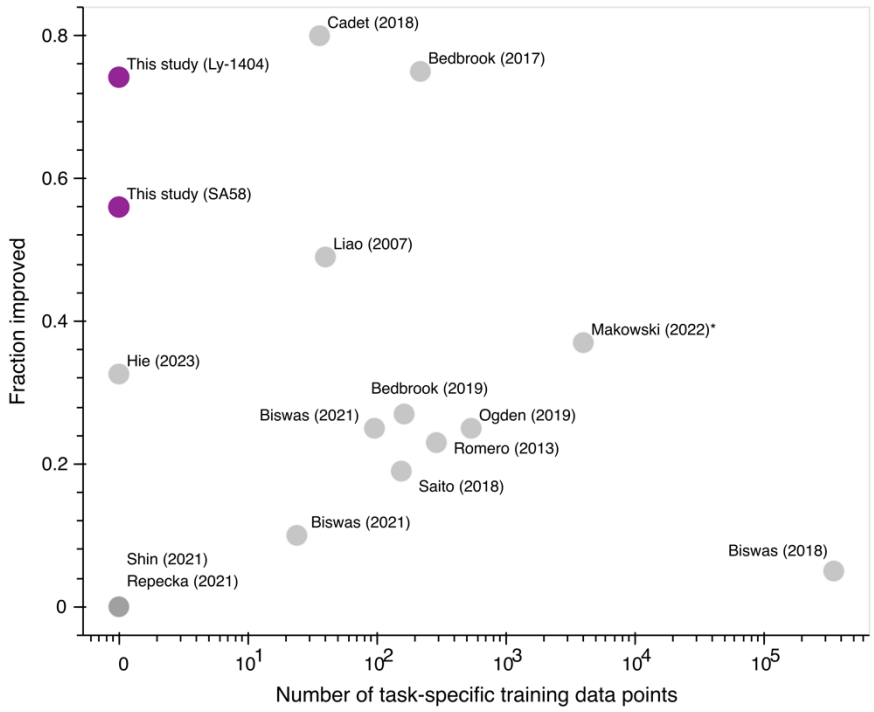


Fig. S11. Comparison to other machine learning-guided directed evolution campaigns
'Fraction improved' refers to the hit rate of variants tested that are improved relative to a wildtype protein used as a starting point for directed evolution or a reference protein used as a design template. Higher hit rates indicate more efficient experimental exploration. Our experimental campaigns achieve among the highest hit rates with the lowest number of assay-labeled training data points to-date (*8*, *56*, *58*, *63–73*).

Protein(s) (Uniprot ID)	Organism	Functional Assay	Mutagenesis Method	Utilized assay	PDB Structure	Total coverage of DMS (%)	Access date*	Reference
UBE2I (P63279)	Human	POPCode, a variant of multiple-site directed mutagenesis.	Competitive growth assay in yeast.	score	5F6E chain A	100	12/10/2018	(Weile *et al*, 2017)
TPK1 (Q9H3S3)				score	3S4Y chain A	92.46		
CALM1 (P0DP23)				score	5V03 chain R	100		
HRas (P01112)	Human	Systematic site-directed mutagenesis.	Two-hybrid assay.	unregulated	2CE2 chain X	100	12/10/2018	(Bandaru *et al*, 2017)
MAPK1 (P28482)	Human	Systematic site-directed mutagenesis.	Competitive growth assay.	VRT	4ZZN chain A	99.44	12/10/2018	(Brenan *et al*, 2016)
TPMT (P51580)	Human	Systematic site-directed mutagenesis.	Fluorescence of a GFP fusion protein.	score	2BZG chain A	92.9	12/10/2018	(Matreyek *et al*, 2018)
UBI4(b) (P0CG63)	Yeast	Site directed mutagenesis by cassette ligation.	Fluorescence activated cell sorting (FACS).	Relative_E1-activity_limiting	4Q5E chain B	100	12/10/2018	(Roscoe & Bolon, 2014)
GAL4 (P04386)	Yeast	Systematic site-directed mutagenesis.	Two-hybrid assay.	Nonselection_24	3COQ chain B	90.64	12/10/2018	(Kitzman *et al*, 2015)
bla(b) (P62593)	E. coli	Systematic site-directed mutagenesis.	Antibiotic resistance.	Ampicillin_2500	1M40 chain A	100	12/10/2018	(Stiffler *et al*, 2015)
haeIIIM (P20589)	H. aegyptius	Random mutagenesis.	Competitive growth assay.	DMS_G3	3UBT chain B	99.37	12/10/2018	(Rockah-Shmuel *et al*, 2015)

Table S1. List of proteins, protein structures, and assay information for deep mutational scanning experiments. Summary of the DMS datasets used in this analysis, including functional assay, method of mutagenesis, and structure used for inverse folding scoring. We also note the specific DMS assay from each study we use for calculating correlation with inverse folding log likelihoods.

*Access date is as reported in *Livesey & Marsh*, 2020 study from which these data were sourced and this table was adapted

CR9114					HA1							HA2													
Strain	HA Subtype	Use in Fig.2	Sequence Identity (%)	Epitope Identity	38	40	41	42	291	292	293	18	19	20	21	36	38	41	42	45	46	48	49	52	56
Influenza A/New Caledonia/20/1999 (H1N1)	H1	Experimental Screening	63.1	14/21	H	V	N	L	S	L	P	V	D	G	W	A	Q	T	Q	I	N	I	T	V	I
Influenza A/Wisconsin/67/2005 (H3N2)	H3	Experimental Screening	40.2	10/21	N	T	E	L	D	K	P	V	D	G	W	A	L	T	Q	I	N	I	N	L	I
Sampled Sequence; T=1e-6	H5	-	53.3	12/21	N	L	N	I	S	M	P	T	D	G	L	P	K	T	Q	I	D	I	D	V	V
Influenza A/Vietnam/1203/2004 (H5N1)	H5	Target Structure	-	-	H	Q	D	I	S	M	P	V	D	G	W	A	K	T	Q	I	D	V	T	V	I

CR6261					HA1									HA2											
Strain	HA Subtype	Use in Fig.2	Sequence Identity (%)	Epitope Identity	18	38	40	41	42	291	292	293	318	19	20	21	38	41	42	45	46	49	52	52	56
Influenza A/New Caledonia/20/99 (H1N1)	H1	Experimental Screening	86.9	21/21	H	H	V	N	L	S	L	P	T	D	G	W	Q	T	Q	I	D	T	V	N	I
Influenza A/Hong Kong/1073/1999 (H9N2)	H9	Experimental Screening	53.9	13/21	Q	H	K	E	L	T	L	P	V	A	G	W	K	T	Q	I	D	T	V	N	V
Sampled Sequence; T=1e-6	H1	-	57.5	17/21	H	H	V	N	L	S	L	P	T	D	G	F	R	T	Q	I	N	T	V	N	K
Influenza A/Brevig Mission/1/1918 (H1N1)	H1	Target Structure	-	-	H	H	V	N	L	S	L	P	T	D	G	W	Q	T	Q	I	D	T	V	N	I

Table S2. Conservation analysis of cross-reactive antibodies used in computational benchmarking. Conservation analysis of cross-reactive antibodies used in computational benchmarking. Epitopes were sourced from Dreyfus et al., *Science* (2012) for CR9114 and the Immune Epitope Database & Tools for CR6261.

LY-CoV1404

Chain Mutated	Design	Region	WT Amino Acid Frequency	Mutant Amino Acid Frequency
HC	D88Q	HFR3	0.03333	0.00382
HC	V90S	HFR3	0.03316	0.05155
HC	S62N	CDR-H2	0.13159	0.16299
HC	V81T	HFR3	0.03432	0.00205
HC	F24Y	HFR1	0.01738	0.00002
HC	I31T	CDR-H1	0.00933	0.09048
HC	H99Y	HFR3	0.01593	0.00138
HC	T70S	HFR3	0.88405	0.06153
HC	I105L	CDR-H3	0.02764	0.05760
LC	A98I	CDR-L3	0.02297	0.03198
LC	Q39K	LFR2	0.92316	0.00238
LC	T5Q	LFR1	0.89340	0.00933
LC	K47E	LFR2	0.52285	0.01490
LC	M49L	LFR2	0.05585	0.77076

SA58

Chain Mutated	Design	Region	WT Amino Acid Frequency	Mutant Amino Acid Frequency
HC	T53L	CDR-H2	0.03814	0.00963
HC	A61S	CDR-H2	0.59797	0.13159
HC	E10Q	HFR1	0.24182	0.01366
LC	N95V	CDR-L3	0.13399	0.00685
LC	S85A	LFR3	0.01109	0.00698
LC	S54T	CDR-L2	0.65138	0.05372
LC	M4V	LFR1	0.29424	0.03348

Table S3. Analysis of neutralization-enhancing mutations. Single amino acid substitutions with beneficial effects on neutralization are reported alongside the region of the variable domain they are located within, as well as the wild-type and mutant amino acid frequencies in observed human antibody sequences.

Supplementary Data 1: Neutralization data with IC_{50} values of evolved antibodies across both evolutionary campaigns

Supplementary Data 2: Binding data with IgG K_D values of evolved antibodies

Supplementary Data 3: Antibody variant prediction benchmarking results

Supplementary Data 4: MFI values for polyspecificity experiments

Supplementary Data 5: Efficiency comparison of machine learning-guided directed evolution methods

Supplementary Information: Antibody sequences