Generative model of SARS-CoV-2 variants under functional and immune pressure unveils viral escape potential and antibody resilience

Marian Huot[1], Pierre Rosenbaum[2], Cyril Planchais[2], Hugo Mouquet[2*], Rémi Monasson[1], Simona Cocco[1*]

[1]*Laboratory of Physics of the École Normale Supérieure, CNRS UMR 8023 and PSL Research, Sorbonne Université, 24 rue Lhomond, Paris, France*
[2]*Institut Pasteur, Université Paris Cité, Humoral Immunology Unit, F-75015 Paris, France*
[*]*hugo.mouquet@pasteur.fr, simona.cocco@phys.ens.fr*

Abstract:

The evolutionary trajectory of SARS-CoV-2 variants is shaped by the selective pressures exerted by host immunity, in particular neutralizing antibodies targeting the receptor-binding domain (RBD). Here, we introduce a data-driven model that quantifies the impact of antibody pressure on RBD evolution and assesses antibody resilience beyond single mutations and known variants. We integrate deep mutational scanning of ACE2 and 31 antibodies with a generative model trained on pre-pandemic Coronaviridae sequences. We then use our generative model to design viable RBD variants under immune pressure from four monoclonal antibodies—SA55, S2E12, S309, and VIR-7229. Experimental validation of 22 variants, with up to 21 mutations from Wuhan wild-type, confirms 50% expression rate. Binding assays further reveal that S309 and VIR-7229 maintain binding across diverse mutational combinations, while SA55 is escaped by one variant and S2E12 exhibits lower resilience. In addition, our model captures correlated escape across antibody pairs (R = 0.65), guiding the selection of negatively correlated cocktails to reduce shared escape routes. By quantitatively linking viral adaptation to antibody resistance profiles, this framework provides a predictive foundation for optimizing therapeutic strategies and enhancing long-term pandemic preparedness.

Keywords: Restricted Boltzmann Machines, SARS-CoV-2, antibody escape, deep mutational scanning, antibody resilience, viral adaptation.

Introduction

Anticipating the evolution of SARS-CoV-2 has become a cornerstone of pandemic preparedness, given the virus's ability to acquire mutations that enhance transmissibility, evade immunity, or both [1]. The receptor-binding domain (RBD) of the spike protein, which mediates viral entry into host cells, is a key target for neutralizing antibodies [2]. Variants with mutations in the RBD that reduce antibody neutralization can evade immune responses, posing significant public health challenges and leading to new pandemic waves. Consequently, identifying such mutations early is imperative for guiding vaccine design and therapeutic interventions to avoid rapid decrease of vaccine and antibody therapy effectiveness [3].

Evolutionary data-driven approaches have proven effective in identifying mutations that preserve viral protein stability and maintain functionality [4,5]. However, variant selection is predominantly driven by binding affinity to cellular receptor ACE2 [6,7] and antibody escape [8,9]. Indeed, population immunity, built through vaccinations and prior infections, has been shown to shape SARS-CoV-2 evolutionary trajectory [10,11]. Pinpointing viable mutations that enable antibody escape is substantially challenging. Work from Thadani et al. [12] demonstrates a significant advancement in this regard, as it leverages structural information to assess the accessibility of epitopes and the likelihood of mutations disrupting antibody binding. In parallel, Wang et al. [13] introduced UniBind, a deep-learning framework that integrates binding affinity predictions to assess the evolutionary trajectory of SARS-CoV-2. By leveraging multi-task learning and heterogeneous biological data, their model accurately predicts the effects of spike protein mutations on ACE2 binding and neutralizing antibody escape, providing a computational approach to anticipate immune evasion.

Although machine learning models are effective at predicting the effects of mutations, experimental data remains the gold standard for assessing viral fitness and immune escape. Experimental techniques such as deep mutational scanning (DMS) and pseudovirus assays provide high-resolution, functional insights into how mutations impact ACE2 binding and antibody neutralization. DMS systematically introduces mutations across a protein and measures their effects on binding, expression, or function using high-throughput selection assays coupled with deep sequencing. This approach has been applied to SARS-CoV-2 to map escape mutations for multiple antibodies [14–22], paving the way for a deeper understanding of RBD evolution. As a matter of fact, several studies [23–26] successfully leveraged experimental data to enhance the prediction of high-fitness SARS-CoV-2 variants, by incorporating deep mutational scanning (DMS) or combinatorial scanning results into their models.

Despite numerous research on SARS-CoV-2 antigenic changes, to which extent antibody pressure has quantitatively shaped viral adaptation—and the fundamental limits of this process—remain poorly understood. While existing studies have mapped single RBD mutations that confer escape and tested antibody binding against multiple known variants, they have largely overlooked antibody resilience—the ability of an antibody or antibody repertoire to remain effective against viral evolution that deviates significantly from the wild-type. A major limitation in the literature is the lack of a quantitative framework to assess whether intrinsic biochemical or structural constraints restrict immune escape. Addressing this is essential for optimizing therapeutic strategies and designing antibody treatments that not only counteract existing variants but also anticipate and withstand future viral adaptations.

Our study addresses this gap by integrating evolutionary data-driven modeling with DMS data to quantify immune pressure on SARS-CoV-2 evolution. We develop a probabilistic framework rooted in biophysical principles and statistical physics to predict combinations of mutations that enhance ACE2 binding, maintain structural stability, and drive antibody escape. Beyond prediction, we experimentally validate our model by testing recombinant modeled RBD variants generated under controlled antibody pressure, demonstrating how immune selection influences mutational trajectories. This validation confirms the model's ability to identify escape mutations and highlights the constraints

imposed by ACE2 binding and protein stability. Finally, we leverage this framework to evaluate antibody resilience, identifying vulnerabilities in SARS-CoV-2 neutralizing monoclonal antibodies, including those developed in clinics, and optimizing cocktail designs that mitigate viral escape. By systematically linking computational predictions with experimental insights, our approach offers a quantitative perspective that can inform the design of more resilient antibody therapies and support efforts to anticipate future viral evolution.

Results

RBD mutational landscape from Deep Mutational Scans and Pre-pandemic sequence data

We developed a computational model to investigate the landscape governing mutations in the receptor-binding domain (RBD) of SARS-CoV-2, balancing protein viability—including expression, stability and overall functionality— with the capacity to bind to the human host cells, and antibody escape, which is a critical factor driving viral adaptation. The model evaluates the likelihood of a RBD sequence by integrating three main components (Figure 1a): ACE2 affinity, antibody escape, and sequence viability.

We denote by $P_{ACE2}(v)$ the probability of the free RBD protein with sequence v to bind ACE2. The ACE2 binding probability is parametrized by the ACE2 concentration, which is treated as a parameter to be fitted from pandemic sequence data, and by the RBD-ACE2 dissociation constant. The RBD-ACE2 dissociation constant is estimated for each RBD sequence v with an additive model (Eq 12) inferred from deep-mutational scans (DMS) data providing the dissociation constants of single-site mutants of the Wuhan wild-type RBD [27] (see Methods).

The probability of antibody escape captures the selective pressure imposed by the immune system. High antibody concentrations correspond to large pressures, and promote the emergence of escape mutations in the RBD. For each antibody Ab_m considered in the model, the escape probability $P_{Ab_m}(v)$ of a RBD variant v is parametrized by the antibody concentration and the dissociation constant between RBD and Ab_m. Similarly to ACE2 binding affinity, the dissociation constant between RBD and antibody depends on the underlying RBD sequence and is estimated using DMS data [14–20].

We first focus our model on a panel of SARS-CoV-2 neutralizing antibodies targeting the RBD and belonging to classes 1 to 4 [20]. Specifically, our initial dataset includes 7 class 1, 9 class 2, 5 class 3, and 8 class 4 antibodies. Because these antibodies were discovered at the beginning of the pandemic, they most likely have influenced the evolutionary trajectory of SARS-CoV-2 during its initial spread. In addition, they were extensively characterized in DMS experiments, providing rich quantitative data necessary to estimate dissociation constants for every variant (Eq. 12).

To account for the evolutionary constraints inherent in naturally occurring protein sequences, the model incorporates the likelihood $P_{RBM}(v)$ that a sequence belongs to the Coronaviridae RBD protein family. This term is derived using a Restricted Boltzmann Machine (RBM), a generative unsupervised neural network with two layers, trained on multiple sequence alignments (MSA) of natural prepandemic RBD sequences in the Coronaviridae family, which are not restricted to human Coronavirus only. The RBM model captures latent factors of co-evolution and mutational tolerance by forming a joint probability landscape over observable sequence states and their hidden representations [28].

The total log-likelihood of a RBD sequence is expressed, up to an additive constant to ensure normalization, as:

$$LL(v) = \beta_{RBM} \log P_{RBM}(v) + \log P_{ACE2}(v) + \sum_m \log P_{Ab_m}(v) \ , \tag{1}$$

where β_{RBM} (similar to an inverse temperature in statistical physics) is a parameter that allows us to adjust the relative strength of far homologous evolutionary constraints with respect to ACE2 binding and Abs escape from selective pressures in human SARS-CoV-2.

This framework integrates biophysical properties and mutational constraints sampled through natural evolution, providing a comprehensive approach to understanding how mutations influence the emergence of SARS-CoV-2 variants.

Fit of the model to sequences from pandemic data

After training a Restricted Boltzmann Machine (RBM) on an MSA containing data from coronaviruses predating SARS-CoV-2 (see SI), we first assessed whether the RBM log-likelihood score serves as a reliable proxy for a protein viability including in particular its expression, and stability. Importantly, because the RBM is trained solely on pre-pandemic sequences, it can be applied immediately at the onset of a pandemic, providing insight into mutational constraints on the RBD sequence without requiring any prior knowledge of the emerging virus's evolution. For each site i, we defined the RBM log-likelihood score of mutating site i in the wild-type background:

$$\Delta LL_i^{RBM} = \langle \log P_{RBM}(v_i, \{wt_{j\neq i}\})\rangle_{v_i} - \log P_{RBM}(wt_i, \{wt_{j\neq i}\})$$

where $\langle\rangle_{v_i}$ denotes the average over the set of all single mutants with amino-acid v_i at site i . This score reflects the average impact of a site mutation on the protein viability, with a positive score indicating mutation is beneficial compared to wildtype. Consistent with the findings of Rodriguez-Rivas et al. [4], we observed a strong correlation between protein expression (measured by Starr et al. [19] on single mutants) and the RBM score obtained with our epistatic model (Figure 1b, Spearman's $\rho = 0.51$), which significantly outperformed the independent model ($\rho = 0.24$, see SI).

We next investigated how ACE2 binding affinity and antibody escape contributed to viral evolution during the pandemic, in conjunction with viability constraints captured by the RBM log-likelihood score. To assess temporal adaptation, we divided the pandemic into three-month intervals spanning the October 2020 to January 2023 period. For each interval, we split the data into training (70%) and test (30%) sets and optimized the model parameters to maximize the likelihood of the training sequences. The ACE2 concentration and the RBM inverse temperature parameter were globally fitted across the entire dataset, reflecting, respectively, stable selection pressure on ACE2 binding and viability constraints. In contrast, the concentrations of the 29 antibodies were re-estimated independently for each time interval to capture evolving immune pressures over the course of the pandemic.

Our findings reveal that the fitted average log10 concentration of antibodies from classes 1, 2 and 3 increased over time, by, respectively, about two, four, and one order of magnitude, indicating that viral evolution was progressively driven by the need to evade these antibody classes (Figure 1c). In contrast, class 4 antibodies maintained consistently high concentrations, suggesting that variants emerging during the pandemic exhibited greater escape from class 4 antibodies relative to pre-pandemic viruses, even in the early stages of the pandemic. Furthermore, prior to the widespread administration of the booster dose (third vaccination dose) of COVID-19 vaccine, we observed limited antibody diversity, with a small number of different antibodies—primarily from classes 1 and 2—exhibiting higher concentrations compared to the early pandemic (Figure 1d). However, post-booster data showed increased antibody diversity, consistent with experimental findings from Andreano et al. [29], which documented a broader range of antibodies in donors after the third (SN3) vaccination doses (Figure 1e).

The observed increase in fitted antibody concentrations suggests a progressive shift in the accessible variant space. Early in the pandemic, the antibody escape term provided little discrimination between variants of concern (VOCs), as immune selection pressure had not yet strongly shaped viral evolution. However, as the likelihood of observed variants was maximized period after period, the selective advantage conferred by antibody escape became more pronounced. By January 2022, the antibody escape log-probability ($\sum_m \log P_{Ab_m}(v)$) had become significantly more discriminative, with wild-type SARS-CoV-2 showing markedly smaller scores compared to subsequent VOCs (Figure 1f).

Early in the pandemic, most residues showed no selective advantage for mutation. However, under late-pandemic antibody concentration parameters, residues at the ACE2 binding interface, which overlaps with class 1 and class 2 antibody epitopes, exhibited a significant increase in mutational benefit, reflecting strong immune-driven selection (Figure 1g). In contrast, core residues remained highly conserved in both early and late-pandemic conditions, consistent with their essential role in

maintaining protein stability. This trend confirms that viral evolution, increasingly driven by the need to escape neutralization, led to a greater-than-expected accumulation of mutations in key epitopes relative to predictions made without antibody selection pressure.

Assessing model predictive power for viral evolution

We assessed the predictive performance of our model by evaluating its ability to identify the most variable sites and the most probable mutations across different pandemic periods. Precisely, we used fitted concentration parameters of a time period to predict mutations of the subsequent time period. Figure 2a highlights the high AUC values achieved by the full model in predicting single amino-acid variants on a site (label is True if the amino acid was observed at least once), while Figure 2b further showcases the model's ability to predict site mutability (label is True if the site is in the top 30% in terms of number of mutated amino acids observed). In practice, to quantify the mutational tolerance at each residue, we compute the average cost of a mutation over all possible amino acids on the site:

$$\Delta LL_i = \langle LL(v_i, \{wt_{j\neq i}\})\rangle_{v_i} - LL(wt_i, \{wt_{j\neq i}\}) \tag{2}$$

where the average is computed on the ensemble of all single mutants at site i in a wild-type background. This score reflects the log-likelihood difference between the single mutants and the wild-type, with larger values indicating greater mutability. While some overlap between mutations in the training and test sets is expected, the high AUC values remain significant given that our model requires to fit 31 parameters only (29 Ab concentrations, plus ACE2 concentration and β_{RBM}), whereas the full mutation distribution encompasses 178×20 amino-acid frequency parameters.

We stress that incorporating the RBM term in the full model improves predictive accuracy with respect to a simpler model that considers ACE2 binding and antibody terms only—such as those developed in prior studies [24, 26] (Figure 2c). This enhancement is observed for both site mutability and single amino-acid variants prediction. Next, we sought to identify highly mutable sites under antibody pressure by comparing model predictions to mutations observed in the work of Baum et al. [30] (Figure 2d). The full model, when calibrated with high antibody concentrations ($\log_{10} C = -6$), accurately predicts mutable sites under antibody pressure, whereas the antibody-only Ab model captures these sites only partially. Notably, sites 446 and 489 are mistakenly identified as favorable mutation sites, despite their known detrimental effects on, respectively, protein expression and ACE2 binding [27]. These findings highlight the necessity of integrating evolutionary constraints with antibody interactions to improve the accuracy of SARS-CoV-2 evolutionary predictions.

Generating new variants

The ability to predict likely mutations should translate into the capacity to generate viable viral sequences, enabling the evaluation of treatment efficacy and vaccine resilience across diverse viral evolutionary scenarios.

We used our model (Eq. (1)) to generate new variants via Monte Carlo sampling (see Methods). We first generated 1,000 sequences using the full model without antibody pressure (ACE2-RBM only), and found that their site-wise entropy shows some correlation with that observed during the pandemic (Spearman $\rho = 0.35$, Figure 3a), though generally exhibiting higher variability. Notably, several sites associated with escape from multiple antibodies—among the 29 antibodies studied—show substantial mutational tolerance, with some sites accommodating over four viable amino acids (Figure 3b). This diversity stems from the RBM component being trained on homologous RBD sequences, including those evolutionarily distant from the wild type. These results suggest that the model captures a broader mutational landscape than that explored during the pandemic, allowing it to serve as a generative tool to benchmark antibody performance against a wide panel of plausible future variants.

A key objective in the generation of sequence is to balance sequence diversity with viability. Figure 3c evaluates the structural consistency of the mutants generated at varying overall model inverse temperatures, $\beta \in \{1, 1.5, 2, 3\}$, using ESM Inverse Folding (ESM-IF) [31] scores. ESM-IF is a deep learning model trained on millions of protein structures that estimates the likelihood of a given sequence folding into a specific structural template. Raising the overall inverse temperature parameter β in Eq. (11) strengthens the selective pressures across all components of the model, biasing the sampling towards sequences that better satisfy structural constraints and maintain the binding of ACE2, thus favoring functional and evolutionarily plausible variants. For $\beta = 3$, the generated sequences achieve ESM-IF scores comparable to that of the wild-type RBD, indicating that they likely adopt the native RBD fold (PDB: 6M0J), despite being ∼20–30 mutations away from any sequence in the MSA (see SI Figure S3).

We then explored the influence of antibody selection pressures on the generated sequences (Figure 3d). Mutants generated using the full model with strong Ab selection pressure exhibited ESMIF scores comparable to those generated with the ACE2-RBM model only, indicating that the overall structural and functional integrity of sequences is preserved across different selection regimes. These results demonstrate that at high overall inverse temperatures, the model generates plausible viral sequences that respect biophysical constraints while incorporating selective pressures from both ACE2 binding and antibody escape. Experimental tests of the generated RBD are reported below.

Predicting Antibody Resilience for new variants

Assuming that generated RBD sequences with high RBM and ACE2-binding likelihoods are functional, our model provides a quantitative framework for assessing how readily the RBD can adapt to immune pressure and facilitate antibody escape. The more functional escape variants exist, the greater the probability that an antibody will lose efficacy. To quantify this, we compute the average fraction F_m of escape variants to a specific antibody, say, Ab_m:

$$F_m = \int_{k^{th}}^{\infty} f_m(k_m) dk_m, \tag{3}$$

where $f_m(k_m)\, dk_m$ is the fraction of variants whose dissociation constant for antibody m is comprised in the range $[10^k; 10^{k+dk}]$ (in g/mL):

$$f_m(k_m) = \frac{\sum_v \delta(k_m - k_m(v))\; P_{RBM}(v)\; P_{ACE2}(v)}{\sum_v P_{RBM}(v)\; P_{ACE2}(v)}\,. \tag{4}$$

The F_m score represents the probability of sequences sampled by the model to escape a given antibody, integrating over the full range of log10 dissociation constants k above the escape threshold $k^{th} = -6$; this value is consistent with the detection limit reported by Cao et al. [32], who were unable to measure IC_{50} values above the μg/mL range. As illustrated in Figure 4a, we find that LY-CoV555 exhibits high F_m compared to S309, indicating that LY-CoV555 is more susceptible to immune escape.

To accurately estimate the tail of the escape probability distribution, we employ thermodynamic integration (see Methods). This approach allowed us to compute the escaping fraction F_m for the 29 antibodies considered up to now. We also add to the Abs panel two recently characterized, broad, and potent neutralizers, SA55 and VIR-7229 [33, 34]. Remarkably, the fractions of escaping variants serve as strong predictors of antibody escape against the omicron variant BA.1, as validated by experimental data from Cao et al. [32] (Figure 4b, AUC=0.75). Crucially, our method assigns higher weight to sequences that are more likely to be viable and binding ACE2, based on the ACE2-RBM model, see Eq. (4), ensuring that only functional variants contribute meaningfully to the score. In contrast, treating all sequences v as equiprobable in Eq. (4) fails (AUC=0.55) to identify which antibodies are vulnerable to escape by BA.1.

Interestingly, evolutionary constraints imposed by RBD viability and ACE2 binding substantially limit the overall potential for antibody escape. As shown in Figure 4c, the proportion of functional escape variants is significantly lower than that predicted under a uniform sequence distribution. Among the antibodies with the highest resilience (i.e., lowest F_m), we identify VIR-7229, S2E12 and S309 and SA55 from the panel of 31 antibodies, as well as SA55. These antibodies are selected for experimental benchmarking in the later stages of this study.

Next, we show how the inferred model can be used to evaluate the feasibility of designing effective antibody cocktails. A crucial factor in determining cocktail efficacy is the overlap of epitopes, as antibodies targeting similar epitopes can be escaped by the same viral mutation [30, 35, 36]. However, beyond epitope overlap, it is important to incorporate, in the prediction of the impact of antibody cocktails, a finer description of the effect of escaping mutations on the variant fitness, as our model does.

To quantify antibody synergy, we compute the covariance between dissociation constants associated with antibodies 1 and 2:

$$\mathrm{Cov}(k_1, k_2) = \sum_{\{v_i, v_j\}} \Delta G^1_{v_i}\, \Delta G^2_{v_j} \mathrm{Cov}(v_i, v_j), \tag{5}$$

where Cov represents the covariance of amino-acid occurrences when sampling sequences from the distribution $P(v)$, and $\Delta G^m_{v_i}$ denotes the effect on the dissociation constant for antibody m of the amino acid v on site i, .

The goal in designing an antibody cocktail is to ensure that no single mutation (v_i) (for amino acid v in site i) has the same effect on both antibodies, which requires $\Delta G^1_{v_i} \Delta G^2_{v_i} < 0$. Furthermore, pairs of mutations (v_i, v_j) that independently contribute to increasing both dissociation constants, characterized by $\Delta G^1_{v_i} \Delta G^2_{v_j} > 0$, should be incompatible and exhibit negative correlations, $\mathrm{Cov}(v_i, v_j) < 0$.

Antibody cocktails combining class 1 and class 2 antibodies are suboptimal due to their positive correlations, which make them susceptible to shared escape mutations. In contrast, class 3 antibodies display anticorrelated escape profiles when paired with class 1 or class 2 antibodies, reducing the likelihood of shared escape mutations and thereby improving cocktail robustness (Figure 4e). This result aligns with experimental findings from Starr et al. [18], which demonstrated the lack of resilience of the LY-CoV555+LY-CoV016 antibody cocktail—consistent with our model's prediction of a positive correlation (Cov = 0.2). In contrast, the REGN10933+REGN10987 combination, predicted by the model to have a negative correlation (Cov = -0.2), appears more promising as was experimentally validated by Baum et al. [30].

In vitro experimental validation

To experimentally validate our model, we generated 6 sequences from the distribution without antibody pressure (ACE2-RBM model) at inverse temperature $\beta = 3$, as well as 16 sequences under selective pressure from VIR-7229, SA55, S2E12 and S309 (ACE2-Abs-RBM model), which demonstrated strong resilience in the model predictions. To generate sequences under high selective pressure we fixed the antibody concentration to the high value $C = 10^{-6}$ g/mL in the model probability P_{Ab} (Methods).

Condition	ACE2-RBM	S2E12	S309	SA55	VIR-7229	Neg. Ctr. ACE2	Neg. Ctr. RBM	WT
Nb sequences tested	6	4	4	4	4	2	2	1
Nb sequences expressed	3	3	3	1	1	0	0	1

Table 1. Number of sequences tested and expressed under various experimental conditions.

We achieved a 50% success rate among the tested sequences when evaluating in vitro production of the designed variants (Table 1). As a control, we also generated 4 sequences with either reduced importance of RBM log-likelihood or ACE2 affinity (see Methods). None of these control sequences could be expressed as recombinant proteins, which emphasizes the importance of maintaining structural integrity and binding potential.

Interestingly, most functional sequences show good ACE-2 binding (Figures 5a-b), although they spanned a broad range of mutations across the RBD, particularly at the ACE2 binding interface (Figure 5c for structural representation of RBD S2E12c, see SI Figure S6 for representation the other sequences that could be expressed). We observe nearly no correlation between ACE2 binding affinity and distance from the wild-type sequence, indicating that our model is capable of generating sequences that are both functionally viable and evolutionarily distant from the wild-type (see SI Figures S7c-d). However, RBM log-likelihood score does not reliably predict why certain sequences failed to be produced (see SI Figure S7d). This limitation is likely due to the absence of structural information in the model, which is trained solely on an MSA. The ESMIF score, expected to be a finer predictor of the stability of RBD in complex with ACE-2, provides a slightly improved and significant discrimination between expressed and unexpressed proteins (p-value = 0.016).

We have experimentally tested the Abs binding for the 12 expressed RBD proteins for a panel of 9 antibodies encompassing the 4 antibodies (VIR-7229, SA55, S2E12 and S309) used in the design of the new variants under antibody selection pressure (Figures 5a-b). Figure 5d demonstrates that the fraction of escaping sequences in the model distribution is an effective proxy for antibody resilience. Experimental results confirm in particular the existence of multiple sequences capable of escaping class 1 REGN10933 and class 2 and 3 antibodies; no RBD was able to escape class 3 S309 and class 1 VIR-7229.

Surprisingly, S2E12—despite ranking as highly resilient in our model—was easily escaped in vitro (Figure 5d), likely due to epistatic interactions which are not taken into account in the Abs dissociation constant. Notably, all sequences that escaped S2E12 carried a mutation at site 486. Deep mutational scanning data indicates that such mutation alone increases escape by a 25-fold factor, with escape fractions approaching 99%. However, because S2E12 binds strongly to the wild-type RBD, our additive model for k estimated that this single mutational effect would not be sufficient to confer full escape. In practice, the presence of additional mutations in the generated sequences amplified the impact of the site 486 mutation, ultimately abolishing binding—highlighting a limitation of additive escape models in capturing context-dependent effects.

A key advantage of our approach is that it provides a framework for testing antibody resilience against highly mutated, yet probable sequences, which allows for the evaluation of escape potential under complex mutational landscapes. VIR-7229, in particular, proved to be highly resilient, as it retained binding even when residue 456—previously identified as important for binding and already mutated in EG.5 and KP variants—was mutated (Figure 5e).

The escape mechanism observed for SA55 involves a mutation at residue G504E (Figure 5f). Notably, this mutation remains viable even in highly mutated RBD backgrounds, indicating substantial tolerance. To assess whether this mutation alone can drive escape, we introduced G504E into both the wild-type (WT) and KP.3 backgrounds. In both cases, the resulting variants maintained strong ACE2 binding while consistently escaping SA55 neutralization (Figures 5g-h). This demonstrates that G504E is a potent single-point escape mutation for SA55, and that its effect is not dependent on a specific genetic context. The fact that this escape occurs across divergent backgrounds, including the clinically relevant KP.3 variant, reveals a critical vulnerability in SA55's resilience. To further dissect the specificity of this escape route, we evaluated whether G504E conferred escape from other antibodies. We observed no loss of binding to S309 or VIR-7229—neither in the WT nor KP.3 backgrounds—indicating that G504E selectively compromises SA55 binding without affecting other antibodies. Together, these results

suggest that G504E constitutes a viable and highly specific escape route for SA55, underscoring the importance of considering such context-independent mutations when evaluating antibody robustness.

Importantly, the $\log_{10}$ dissociation constants predicted by our linear models (Methods) exhibit strong correlations with experimentally measured binding scores (Figures 6a–b) and accurately identify antibody escape events (AUC = 0.9; see SI Figure S7a for per-antibody metrics). By setting the predicted escape threshold to $k^{\text{th}} = -6$, we obtain a true positive rate of 61% and a false positive rate of 0%. Under this threshold, the model correctly predicts that none of the generated variants escape S309 (Figure 6b), while COV2-2130 is escaped by most of the expressed sequences.

Last of all, we show in Figures 6c-d that the predicted correlations between antibodies escaping were matching well those in experiments (Pearson correlation: 0.65). Notably, class 1 antibodies systematically showed the same binding behavior towards functional generated sequences. We find, both in the model and in the experimental data, that VIR-7229+REGN10987 and S309+REGN10987 are good cocktails as they dissociation constants are anti-correlated, despite REGN10987 not being a powerful antibody alone. Moreover SA55+VIR-7229 results in a good cocktail according to the measured negative correlation on the tested variants, in disagreement with the model predicting a positive correlation. On the contrary S309+VIR-7229 is negatively correlated in the model but not in data. These discrepancies may be explained by differences in sampling and statistical resolution. In our analysis, Pearson correlation coefficients were computed to assess the co-variation of antibody dissociation constants across sequence variants. For antibody pairs with limited or no observed escape among the tested variants, the variance in dissociation constants is extremely small, inflating the Pearson correlation regardless of biological signal. This effect likely contributes to discrepancies between experimental and model-based correlations.

Overall, these results provide overall a robust experimental validation of our model, demonstrating its potential for predicting viral escape and guiding the design of more resilient antibody therapeutics.

Discussion

Understanding the evolutionary landscape of SARS-CoV-2 requires a combination of experimental and computational methods. Deep mutational scanning (DMS) has provided invaluable insights into the binding properties of viral proteins by experimentally measuring the impact of thousands of mutations [14–16, 18, 19, 27]. However, DMS alone does not predict how the viral protein can best balance functionality and antibody escape. To bridge this gap, our approach integrates DMS data with a probabilistic evolutionary model trained on prepandemic sequences. By fitting model hyperparameters using sequences observed throughout the pandemic, we then capture how the mutational landscape has shifted toward antibody escape. This finding is consistent with previous studies showing that SARS-CoV-2 variants have evolved under strong immune pressure, leading to increased neutralization resistance over time [21, 32].

A key strength of our model lies in its ability to function both as a predictive and a generative tool. While many previous models have focused solely on predicting viral escape mutations and mutable sites [4, 12], our framework extends this by generating a distribution of potential escape variants, similar to recent approach by Youssef et al. [37]. Predictive models like Beth-1 and VaxSeer have been successfully applied in influenza research to forecast dominant strains and optimize vaccine selection by leveraging mutation fitness dynamics and antigenicity-dominance predictions [38, 39]. However, these models primarily focus on predicting dominant variants rather than exploring the full mutational landscape. In contrast, our approach generates a broader distribution of potential variants, revealing an interesting pattern: the entropy of our model-generated sequences is higher than that observed in real pandemic evolution. This discrepancy arises because real-world viral evolution is constrained by factors such as transmission bottlenecks, epidemiological dynamics, and stochastic events, which limit the diversity of viable mutations that successfully spread in the population [40]. By leveraging this generative capacity, our model can assist in designing preemptive countermeasures against future viral evolution.

In addition, we introduced a rigorous metric for quantifying antibody resilience. Traditionally, antibody effectiveness is evaluated based on neutralization assays on past and currently circulating variants, as well as structural epitope mapping [17,41]. However, these approaches do not systematically quantify how robust an antibody is to viral escape across the full mutational landscape. Our resilience metric provides a probabilistic estimate of how many viable viral sequences retain ACE2 binding while escaping a given antibody, thereby offering a quantitative measure of antibody robustness. Moreover, as SARS-CoV-2 evolves under immune pressure, it accumulates combinations of mutations whose effects are often shaped by epistatic interactions [42, 43]. In this context, our model enables the systematic evaluation of antibody cocktails by accounting for these epistatic constraints, ensuring that the selected antibodies are only jointly escaped through mutually incompatible mutations. This goes beyond conventional cocktail design strategies, which typically focus on minimizing epitope overlap without accounting for co-evolutionary constraints [30, 35, 36]. Our experimental results with SA55 demonstrate the utility of this approach: while G504E is a known escape mutation in a wild-type background, our findings show that it remains viable even when introduced into highly mutated RBD sequences, as well as recent variant KP.3. This highlights the model's capacity to sample escape mutations in diverse, evolutionarily distant contexts, enabling the identification of escape routes that retain function despite extensive background variation—an important consideration for anticipating antibody vulnerabilities under future immune pressure. This approach is especially relevant for the development of polyvalent [44] and mosaic vaccines [45], which seek to induce broad, cross-protective immunity by covering a wide range of potential escape routes.

Despite its strengths, our model has limitations. Notably, only half of the experimentally tested sequences were expressed, indicating that additional constraints—beyond those captured by our

framework—play a role in determining protein viability. The RBM-loglikelihood, trained solely on few thousands of prepandemic coronavirus sequences, captures co-evolutionary patterns and long-term sequence constraints across the RBD protein family. This allows the model to infer which mutations are likely to be tolerated based on evolutionary precedent, providing insight into the viability of novel variants even without knowledge of pandemic-era evolution. However, while this statistical approach implicitly reflects aspects of protein stability, it cannot account for detailed tertiary structural effects.

Incorporating structure-based deep learning models, such as Protein MPNN [46] and ESM Inverse Folding [31], or Rosetta-based energy calculations [47], could improve hit rates in experimental validation. However, this would come at the cost of increased computational complexity and increasing need of training data and make exploration of whole mutational landscape harder.

Methods

Sequence data

In our study, we utilized multiple sequence alignments (MSAs) of 2610 sequences acquired in a previous study from Rodrigues et al [4]. The sequence data in FASTA format were sourced from several databases: GISAID (release 16 May 2021), Uniref90 (December 2020 release), ViPR (downloaded in September 2020), NCBI viral genomes (downloaded in September 2020), and the MERS coronavirus database (downloaded in September 2020). As part of quality control, sequences containing nonstandard amino acids and repeated sequences were removed. Additionally, to distinguish training data from test data, any sequences with greater than 90% sequence identity to the Wuhan-Hu-1 reference were filtered out, excluding all SARS-CoV-2 sequences, including closely related sequences from nonhuman hosts.

All 15,371,428 spike sequences on GISAID [48] as of 14-April-2023 were downloaded and aligned, following approach in the work of Starr et al. [19]. Sequences from non-human origins and with lengths outside $[1260, 1276]$ were removed. They were then aligned via mafft [49] and sequences containing unicode errors, gap or ambiguous characters were removed. To avoid regional bias, we filtered to only keep sequences observed in the USA. Overall, we retained 3,712,556 submissions represented by 10,453 unique RBD sequences. RBD amino-acid mutations were enumerated compared to the reference Wuhan-Hu-1 SARS-CoV-2 RBD sequence (Genbank MN908947, residues N331-T531).

To align with the MSA data, the model was finally applied to sequences of length 178, spanning residues S349 to G526, leading to 8,167 unique RBD sequences.

Background distribution of RBD sequences from pre-pandemic data

We use Restricted Boltzmann Machines [28] to model the probability distribution over RBD sequences, which we can infer from homologous sequence data. This distribution encodes essential evolutionary, functional and structural constraints acting on RBD. A Restricted Boltzmann Machine (RBM) is a joint probabilistic model for sequences and representations. It is formally defined on a bipartite, two-layer graph. Protein sequences $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ are displayed on the Visible layer, and representations $\mathbf{h} = (h_1, h_2, \ldots, h_M)$ on the Hidden layer. Each visible unit takes one out of 21 values (20 amino acids +1 alignment gap). Hidden-layer unit values h_μ are real. The joint probability distribution of $\mathbf{v}, \mathbf{h}$ is:

$$P_{\mathrm{RBM}}(\mathbf{v}, \mathbf{h}) \propto \exp\left(\sum_{i=1}^{N} g_i\left(s_i\right) - \sum_{\mu=1}^{M} \mathcal{U}_\mu\left(h_\mu\right) + \sum_{i,\mu} h_\mu w_{i\mu}\left(s_i\right)\right) \tag{6}$$

where g_i and U_μ denote, respectively, local biases and potentials acting on the visible and hidden units, and the $w_{i,\mu}$'s represent the interaction weights between these layers. We use the dReLU potential $\mathcal{U}_\mu(h) = \frac{1}{2}\gamma_{\mu,+}h_+^2 + \frac{1}{2}\gamma_{\mu,-}h_-^2 + \theta_{\mu,+}h_+ + \theta_{\mu,-}h_-$, where $h_+ = \max(h, 0)$, $h_- = \min(h, 0)$. The parameters $(\gamma_{\mu,+}, \gamma_{\mu,-}, \theta_{\mu,+}, \theta_{\mu,-})$ are optimized by maximizing the product of the marginal probabilities

$$P_{\mathrm{RBM}}(v) \propto \int_h dh\, P_{\mathrm{RBM}}(v, h) = \exp(\sum_{i=1}^{N} g_i\left(v_i\right) + \sum_{\mu=1}^{M} \Gamma_\mu\left(I_\mu(\mathbf{v})\right)), \tag{7}$$

over the sequences v within a MSA of the protein family. The cumulant generating function $\Gamma_\mu(I) = \ln \int_{h_\mu} \mathrm{d}h_\mu\, e^{I_\mu h_\mu - \mathcal{U}_\mu(h_\mu)}$ is computed as described in work from Tubiana et al. [28].

We trained the RBMs using the (PGM package) [28] and optimized them through the Persistent Contrastive Divergence method. The number of hidden units was fixed to 100, with activation functions set to dReLU. Training was carried out employing 10 Markov chains and performing 10 Monte Carlo updates per gradient computation, for 50 iterations. To introduce regularization, we applied an L1/2

penalty on the weight parameters with a coefficent 0.12. The MSA used for training was obtained from the work of Rodrigues et al. [4] concatenated with the wildtype sequence.

Binding to ACE2 and Antibodies

We define the probabilities that the RBD encoded by sequence v is bound to ACE2 and escapes antibody i (Ab_i) through, respectively,

$$P_{\mathrm{ACE2}}(v) = \frac{C_{\mathrm{ACE2}}\, e^{-G_{\mathrm{ACE2}}(v)}}{C_{\mathrm{ACE2}}\, e^{-G_{\mathrm{ACE2}}(v)} + e^{-G_{\mathrm{free}}}} = \frac{1}{1 + 10^{k_{\mathrm{ACE2}} - c_{\mathrm{ACE2}}}} \,, \tag{8}$$

$$P_{\mathrm{Ab}_m}(v) = \frac{e^{-G_{\mathrm{free}}}}{C_m\, e^{-G_m(v)} + e^{-G_{\mathrm{free}}}} = \frac{1}{1 + 10^{c_m - k_m(v)}} \,, \tag{9}$$

where G_{free}, G_{ACE2} and G_m are the (dimensionless) free energies associated with the free, ACE2-bound and Ab_m-bound states. $k_m(v) = (G_m(v) - G_{\mathrm{free}})/\log 10$ and $k_A(v) = (G_{\mathrm{ACE2}}(v) - G_{\mathrm{free}})/\log 10$ are log10 dissociation constants for antibody m and ACE2, which depend on the sequence v, while c_m and c_{ACE2} represent log10 concentrations of antibody m and ACE2 driving selection pressures over RBD.

Full model

Multiplying the background RBD distribution in Eq. (7) with the binding probabilities in Eqs. (8) and (9) defines the full probability distribution over the sequences v, with the result

$$P(v) = \frac{\exp(-E(v|\theta))}{Z(\theta)} \,, \tag{10}$$

where θ denotes the set of all model parameters, and the effective energy E reads

$$\begin{aligned} E(v) &= -\beta_{RBM}\, \log P_{\mathrm{RBM}}(v) - \log P_{ACE2}(v) - \sum_m \log P_{Ab_m}(v) \\ &= -\beta_{RBM}\Big(\sum_i g_i(s_i) + \sum_\mu \Gamma_\mu(s)\Big) + \log\left(1 + 10^{k_{\mathrm{ACE2}}(v) - c_{\mathrm{ACE2}}}\right) \\ &\quad + \sum_m \log\left(1 + 10^{c_m - k_m(s)}\right) \end{aligned}$$

The presence of the inverse temperature parameter, β_{RBM}, allows us to adjust the strength of evolutionary constraints, modulating the balance between selective pressures and sequence flexibility.

Finally, we can increase the strength of selection by raising the overall model inverse temperature β, which amplifies the contribution of the effective energy $E(v)$ to the sampling distribution. Accordingly, we sample sequences from the Boltzmann-like distribution

$$P(v) = \frac{\exp(-\beta\, E(v|\theta))}{Z(\theta)} \,, \tag{11}$$

where the partition function $Z(\theta)$ ensures normalization and is defined as

$$Z(\theta) = \sum_v \exp(-\beta\, E(v|\theta)) \,.$$

Experimental dissociation constants

IC50 of all 29 antibodies to wildtype were obtained in work from Starr et al, Nature 2021 [20] for S304, S2H14, S2H97, S2H13, S2X35, S309, S2X58, S2X227, S2X16, S2H58, S2D106, S2E12; Zost et al, Nature 2020 [50] for COV2-2677, COV2-2165, COV2-2082, COV2-2094, COV2-2130, COV2-2050,

COV2-2832, COV2-2096, COV2-2479, COV2-2499, COV2-2196; Starr et al, Science 2021 [19] for REGN10987, REGN10933 and LY-CoV016; Wang et al, Nature 2021 [51] for LyCov555; Tortorici et al, Nature 2021 [17] for S2X259; Greaney et al, Cell Host & Microbe 2021 [14] for CR3022.

IC50 of antibodies S2H14, COV2-2196, S2E12, LY-CoV016, S2H13, LY-CoV555, COV2-2479, COV2-2499, COV2-2130, S309, S304, COV2-2677, S2H97, S2X259, S2X35 to BA1 where obtained in work from Cao et al [32].

Finally, wildtype IC50 as well as DMS for SA55 and VIR7229 were extracted from work of Rosen et al. [33].

To model effect of single mutations on dissociation constants, we use deep mutational scans : study from Starr et al. [27] for ACE2 DMS; data provided in bloomlab reposatory, that integrates data several previous studies [14–20]. DMS data provided escape ratios for nearly all single mutants on a Wuhan background, to all 29 antibodies. Noting that the dissociation constant of RBD writes:

$$K = \frac{[RBD_{free}][Ab]}{[RBD_{bound}]} \propto \frac{[RBD_{free}]}{[RBD_{bound}]} ,$$

we assume $k = log_{10}K$ variations behave like log_{10} escape ratios variations.

We approximate the wild-type dissociation constant to each antibody using its measured IC_{50} value:

$$k(\mathrm{wt}) := \log_{10}\left(\mathrm{IC}_{50}(\mathrm{wt})\right)$$

By combining these baseline values $k(\mathrm{wt})$ with experimentally measured changes in k for single mutations obtained from deep mutational scanning, we construct a dataset of log dissociation constants for (almost all, except for missing entries) single amino acid variants against each antibody.

Prediction of log K for multiple variants

Although dissociation constants were measured for wildtype RBD and single variants, we need to extrapolate to variants with multiple mutations. Since the free energy of binding ΔG is related to the dissociation constant by $\Delta G \sim \log_{10} K$ (up to a constant), we define the effect of each site in terms of ΔG_i. To model the effects of all sites of sequence v on log dissociation constants, we use an additive model:

$$\Delta G(v_1 ... v_N) := \Delta G_0 + \sum_i^N \Delta G_i(v_i) \ . \tag{12}$$

Choosing such an independent model allows us to leverage experimental deep muational scans to predict mutational effect on dissociation constants not only to single mutations but also to sequences far from wildtype. Previous study from Moulana et al. [52] on combinatorial datasets has shown such a non-epistatic model could well represent variant dissociation constants (Pearson correlation $\in [0.6 - 1.0]$). In practice, we compute the log dissociation constant using a ridge regression (L2 penalty $\lambda = 0.1$) on the one hot encoding of sequence amino acids. A unique property of ridge regression applied to one-hot encoding is that the predicted mutational effect for unknown amino acids is the average known mutational effect of mutations on the same site [31]. This mechanism allows the model to effectively generalize to novel amino acids at specific positions from knowledge of the average impact of mutations on the dissociation constant. This is necessary as deep mutational scans may not cover the entirety of possible single variants.

Fitting ACE2-AB-RBM model on pandemic data

We want to find the parameter set θ that maximizes the model average log-likelihood $\langle \log P(s) \rangle_{\mathrm{dataset}}$ of the sequences in the training set, which includes all pandemic data. For simplicity, we collectively

denote as θ the set of parameters encompassing antibody and ACE2 log-concentrations $\{c_i\}$, dissociation constants $\{k_i\}$, and the inverse temperature β of the RBM component.

The gradient of the log-likelihood is given by

$$\frac{\partial \mathrm{LL}}{\partial \theta} = -\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{dataset}} + \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{model}} \tag{13}$$

In practice, we choose a minibatch with a single sequence, randomly drawn from the dataset at each step of the learning algorithm. We then sample another sequence from the model distribution P, and compute the energy difference between these two sequences. Using PyTorch's autodifferentiation tools we evaluate the gradient. Last of all, we perform a gradient update step of θ based on this computation.

To sample from the model distribution, the conventional method involves performing a complete Markov Chain Monte Carlo (MCMC) simulation for each gradient step, continuing until equilibrium is reached. A more computationally efficient alternative is Persistent Contrastive Divergence (PCD), a technique introduced by Tieleman for training Restricted Boltzmann Machines [53]. Instead of initializing each MCMC run from scratch, PCD reuses samples from the previous iteration, updating them with a limited number of MCMC steps before the next gradient evaluation. When the model parameters change gradually, this approach effectively maintains equilibrium, providing accuracy comparable to standard MCMC sampling while significantly reducing computational cost.

Fraction of escaping variants

The density of variants that are viable, bind to ACE2, and exhibit a $\log_{10}$ dissociation constant equal to k_m to antibody m is given by $f_m(k_m)$ in Eq. (4). To estimate the log-density $S(k_m) := \log P(k_m)$, one could in principle sample variants from the joint ACE2-RBM model and compute a histogram of their dissociation constants. However, this naive approach may suffer from undersampling and high variance, particularly in the tails of the distribution.

To circumvent this sampling issue, we used thermodynamic integration and generated sequences v according to the biased distribution

$$P_{\mathrm{RBM}}(v)\, P_{\mathrm{ACE2}}(v)\, e^{\beta_m\, k_m(v)}$$

which induces a sampling distribution over k_m of the form $\propto e^{S(k_m)+\beta_m k_m}$, for values of $\beta_m \in [-3, 3]$. For each β_m, we compute the expected dissociation constant $\langle k_m \rangle = \bar{k}_m$.

Assuming that the observed mean value $\bar{k}_m$ corresponds to the mode of the biased distribution (an assumption justified in the large-L limit), we get

$$\bar{k}_m \approx \arg\max\,(S(k_m) + \beta_m\, k_m) \quad \Rightarrow \quad S'(\bar{k}_m) = -\beta_m.$$

We then reconstruct $S(k_m)$ by integrating the estimated derivative $S'(k_m)$, up to a normalization constant.

Sequence Generation under Different Selective Pressures

For each experimental condition, we generated a batch of 1,000 RBD sequences by sampling from the gap-less distribution $P(v)$ (Eq. 10). Samples were obtained using Markov Chain Monte Carlo (MCMC) with 500 burn-in steps, 20 steps between each sample, and 20 independent chains to ensure adequate exploration of the sequence space.

To simulate immune pressure, sequences were drawn from the full model (Eq. 10), which includes the ACE2-RBM energy landscape and the antibody term favoring escape from a given antibody with antibody concentration parameter were set to $\log_{10} c = -6$.

As controls, we also sampled from perturbed distributions. To test the effect of reducing structural constraints, we increased the RBM temperature ($\beta_{\mathrm{RBM}} \in \{0.1, 0.2\}$) to relax evolutionary penalties.

Conversely, to test reduced ACE2 affinity, we used a lower ACE2 concentration ($\log_{10} C_{\mathrm{ACE2}} \in \{-6, -4\}$), diminishing the selection for tight receptor binding.

Sequence scoring with ESM Inverse Folding

To evaluate sequence likelihoods by a state of the art machine-learning method learned on structures, we use *ESM Inverse Folding* available in ESMIF repository, which assigns a log-likelihood score to each sequence given a protein structure [31]. In this approach, the structural context is provided by the PDB file, while the FASTA file contains the sequences to be assessed. The log-likelihood scores quantify how well each sequence aligns with the given structure (6M0J).

Recombinant RBD proteins and monoclonal antibodies

Codon-optimized nucleotide fragments encoding the RBD proteins followed by C-terminal tags (Hisx8-tag and AviTag) were synthesized and cloned into pcDNA3.1/Zeo(+) expression vector (Thermo Fisher Scientific). Human angiotensin-converting enzyme 2 (ACE2) ectodomain and RBD proteins were produced by transient transfection of exponentially growing Freestyle 293-F suspension cells (Thermo Fisher Scientific) using polyethylenimine (PEI) precipitation method, purified from culture supernatants by high-performance chromatography using the Ni Sepharose® Excel Resin according to manufacturer's instructions (Citivia), dialyzed against PBS using Slide-A-Lyzer® dialysis cassettes (Thermo Fisher Scientific), quantified using NanoDrop 2000 instrument (Thermo Fisher Scientific), and controlled for purity by SDS-PAGE using NuPAGE 3-8% Tris-acetate gels (Life Technologies) as previously described (PMID: 35704748). Purified AviTagged-ACE2 protein was biotinylated using the Enzymatic Protein Biotinylation Kit following the manufaturer's procedures (Sigma-Aldrich). Human anti-SARS-CoV2 RBD and negative control (mGO53) IgG1 antibodies were produced by transient co-transfection of Freestyle™ 293-F suspension cells (Thermo Fisher Scientific), purified by affinity chromatography using Protein G Sepharose® 4 Fast Flow (GE Healthcare) according to manufacturer's instructions (Citivia), and quantified using NanoDrop 2000 instrument (Thermo Fisher Scientific).

ELISA

ELISAs were performed as previously described (PMID: 35704748). Briefly, high-binding 96-well ELISA plates (Costar, Corning) were coated overnight with 250 ng/well of purified recombinant RBD proteins. After washings with 0.05% Tween 20-PBS (washing buffer), plates were blocked 2 h with 2% BSA, 1 mM EDTA, 0.05% Tween 20-PBS (Blocking buffer), washed, and incubated with serially diluted purified IgG mAbs in PBS. Recombinant IgG1 antibodies were tested at 10 µg/ml, and 7 consecutive 1:4 dilutions in PBS. After washings, plates were revealed by incubation for 1 h with goat HRP-conjugated anti-human IgG (Jackson ImmunoReseach, 0.8 µg/ml final) and by adding 100 µl of HRP chromogenic substrate (ABTS solution, Euromedex) after washing steps. Optical densities were measured at 405 nm (OD405nm) after 1 h incubation at room temperature, and background values given by incubation of PBS alone in coated wells were subtracted. For ACE2 binding to immobilized RBD proteins, plates were incubated with recombinant biotinylated ACE2 ectodomain at 50 µg/ml and 7 consecutive 1:2 dilutions. After washings, plates were revealed by incubation for 30 min with streptavidin HRP-conjugated (BD Biosciences) as described above. Experiments were performed using HydroSpeed™ microplate washer and Sunrise™ microplate absorbance reader (Tecan Männedorf, Switzerland). Aera under the curve (AUC) values were calculated from ELISA titration curves obtained in duplicate.

References

1. A. M. Carabelli, T. P. Peacock, L. G. Thorne, W. T. Harvey, J. Hughes, COVID-19 Genomics UK Consortium, T. I. De Silva, S. J. Peacock, W. S. Barclay, T. I. De Silva, G. J. Towers, and D. L. Robertson, "SARS-CoV-2 variant biology: immune escape, transmission and fitness," Nat. Rev. Microbiol. (2023).
2. C. B. Jackson, M. Farzan, B. Chen, and H. Choe, "Mechanisms of SARS-CoV-2 entry into cells," Nat. Rev. Mol. Cell Biol. **23**, 3–20 (2022).
3. A. Tuekprakhon, R. Nutalai, A. Dijokaite-Guraliuc, D. Zhou, H. M. Ginn, M. Selvaraj, C. Liu, A. J. Mentzer, P. Supasa, H. M. Duyvesteyn, R. Das, D. Skelly, T. G. Ritter, A. Amini, S. Bibi, S. Adele, S. A. Johnson, B. Constantinides, H. Webster, N. Temperton, P. Klenerman, E. Barnes, S. J. Dunachie, D. Crook, A. J. Pollard, T. Lambe, P. Goulder, N. G. Paterson, M. A. Williams, D. R. Hall, E. E. Fry, J. Huo, J. Mongkolsapaya, J. Ren, D. I. Stuart, G. R. Screaton, C. Conlon, A. Deeks, J. Frater, L. Frending, S. Gardiner, A. Jämsén, K. Jeffery, T. Malone, E. Phillips, L. Rothwell, and L. Stafford, "Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum," Cell **185**, 2422–2433.e13 (2022).
4. J. Rodriguez-Rivas, G. Croce, M. Muscat, and M. Weigt, "Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes," Proc. National Acad. Sci. **119**, e2113118119 (2022).
5. B. Hie, E. D. Zhong, B. Berger, and B. Bryson, "Learning the language of viral evolution and escape," Science **371**, 284–288 (2021).
6. S. Ozono, Y. Zhang, H. Ode, K. Sano, T. S. Tan, K. Imai, K. Miyoshi, S. Kishigami, T. Ueno, Y. Iwatani, T. Suzuki, and K. Tokunaga, "SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity," Nat. Commun. **12**, 848 (2021).
7. M. I. Barton, S. A. MacGowan, M. A. Kutuzov, O. Dushek, G. J. Barton, and P. A. Van Der Merwe, "Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics," eLife **10**, e70658 (2021).
8. K. G. Nabel, S. A. Clark, S. Shankar, J. Pan, L. E. Clark, P. Yang, A. Coscia, L. G. A. McKay, H. H. Varnum, V. Brusic, N. V. Tolan, G. Zhou, M. Desjardins, S. E. Turbett, S. Kanjilal, A. C. Sherman, A. Dighe, R. C. LaRocque, E. T. Ryan, C. Tylek, J. F. Cohen-Solal, A. T. Darcy, D. Tavella, A. Clabbers, Y. Fan, A. Griffiths, I. R. Correia, J. Seagal, L. R. Baden, R. C. Charles, and J. Abraham, "Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain," Science **375**, eabl6251 (2022).
9. Q. Wang, S. Iketani, Z. Li, L. Liu, Y. Guo, Y. Huang, A. D. Bowen, M. Liu, M. Wang, J. Yu, R. Valdez, A. S. Lauring, Z. Sheng, H. H. Wang, A. Gordon, L. Liu, and D. D. Ho, "Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants," Cell **186**, 279–286.e8 (2023).
10. N. A. Raharinirina, N. Gubela, D. Börnigen, M. R. Smith, D.-Y. Oh, M. Budt, C. Mache, C. Schillings, S. Fuchs, R. Dürrwald, T. Wolff, M. Hölzer, S. Paraskevopoulou, and M. Von Kleist, "SARS-CoV-2 evolution on a dynamic immune landscape," Nature **639**, 196–204 (2025).
11. M. Meijers, D. Ruchnewitz, J. Eberhardt, M. Łuksza, and M. Lässig, "Population immunity predicts evolutionary trajectories of SARS-CoV-2," Cell **186**, 5151–5164.e13 (2023).
12. N. N. Thadani, S. Gurev, P. Notin, N. Youssef, N. J. Rollins, D. Ritter, C. Sander, Y. Gal, and D. S. Marks, "Learning from prepandemic data to forecast viral escape," Nature **622**, 818–825 (2023).
13. G. Wang, X. Liu, K. Wang, Y. Gao, G. Li, D. T. Baptista-Hon, X. H. Yang, K. Xue, W. H. Tai, Z. Jiang, L. Cheng, M. Fok, J. Y.-N. Lau, S. Yang, L. Lu, P. Zhang, and K. Zhang, "Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution," Nat. Med. **29**, 2007–2018 (2023).
14. A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. Crawford, A. S. Dingens, R. S. Nargi, R. E. Sutton, N. Suryadevara, P. W. Rothlauf, Z. Liu, S. P. Whelan, R. H. Carnahan, J. E. Crowe, and J. D. Bloom, "Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition," Cell Host & Microbe **29**, 44–57.e9 (2021).
15. A. J. Greaney, T. N. Starr, C. O. Barnes, Y. Weisblum, F. Schmidt, M. Caskey, C. Gaebler, A. Cho, M. Agudelo, S. Finkin, Z. Wang, D. Poston, F. Muecksch, T. Hatziioannou, P. D. Bieniasz, D. F. Robbiani, M. C. Nussenzweig, P. J. Bjorkman, and J. D. Bloom, "Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies," Nat. Commun. **12**, 4196 (2021).
16. A. J. Greaney, A. N. Loes, K. H. Crawford, T. N. Starr, K. D. Malone, H. Y. Chu, and J. 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**, 463–476.e6 (2021).
17. M. A. Tortorici, N. Czudnochowski, T. N. Starr, R. Marzi, A. C. Walls, F. Zatta, J. E. Bowen, S. Jaconi, J. Di Iulio, Z. Wang, A. De Marco, S. K. Zepeda, D. Pinto, Z. Liu, M. Beltramello, I. Bartha, M. P. Housley, F. A. Lempp, L. E. Rosen, E. Dellota, H. Kaiser, M. Montiel-Ruiz, J. Zhou, A. Addetia, B. Guarino, K. Culap, N. Sprugasci, C. Saliba, E. Vetti, I. Giacchetto-Sasselli, C. S. Fregni, R. Abdelnabi, S.-Y. C. Foo, C. Havenar-Daughton, M. A. Schmid, F. Benigni, E. Cameroni, J. Neyts, A. Telenti, H. W. Virgin, S. P. J. Whelan, G. Snell, J. D. Bloom, D. Corti, D. Veesler, and M. S. Pizzuto, "Broad sarbecovirus neutralization by a human monoclonal antibody," Nature **597**, 103–108 (2021).
18. T. N. Starr, A. J. Greaney, A. S. Dingens, and J. D. Bloom, "Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016," Cell Reports Med. **2**, 100255 (2021).
19. T. N. Starr, A. J. Greaney, A. Addetia, W. W. Hannon, M. C. Choudhary, A. S. Dingens, J. Z. Li, and J. D. Bloom, "Prospective mapping of viral mutations that escape antibodies used to treat COVID-19," Science **371**, 850–854 (2021).
20. T. N. Starr, N. Czudnochowski, Z. Liu, F. Zatta, Y.-J. Park, A. Addetia, D. Pinto, M. Beltramello, P. Hernandez, A. J. Greaney, R. Marzi, W. G. Glass, I. Zhang, A. S. Dingens, J. E. Bowen, M. A. Tortorici, A. C. Walls, J. A. Wojcechowskyj, A. De Marco, L. E. Rosen, J. Zhou, M. Montiel-Ruiz, H. Kaiser, J. R. Dillen, H. Tucker, J. Bassi, C. Silacci-Fregni, M. P. Housley, J. Di Iulio, G. Lombardo, M. Agostini, N. Sprugasci, K. Culap, S. Jaconi, M. Meury, E. Dellota Jr, R. Abdelnabi, S.-Y. C. Foo, E. Cameroni, S. Stumpf, T. I. Croll, J. C. Nix, C. Havenar-Daughton, L. Piccoli, F. Benigni, J. Neyts, A. Telenti, F. A. Lempp, M. S. Pizzuto, J. D. Chodera, C. M. Hebner, H. W. Virgin, S. P. J. Whelan, D. Veesler, D. Corti, J. D. Bloom, and G. Snell, "SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape," Nature **597**, 97–102 (2021).
21. Y. Cao, F. Jian, J. Wang, Y. Yu, W. Song, A. Yisimayi, J. Wang, R. An, X. Chen, N. Zhang, Y. Wang, P. Wang, L. Zhao, H. Sun, L. Yu, S. Yang, X. Niu, T. Xiao, Q. Gu, F. Shao, X. Hao, Y. Xu, R. Jin, Z. Shen, Y. Wang, and X. S. Xie, "Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution," Nature (2022).
22. A. Yisimayi, W. Song, J. Wang, F. Jian, Y. Yu, X. Chen, Y. Xu, S. Yang, X. Niu, T. Xiao, J. Wang, L. Zhao, H. Sun, R. An, N. Zhang, Y. Wang, P. Wang, L. Yu, Z. Lv, Q. Gu, F. Shao, R. Jin, Z. Shen, X. S. Xie, Y. Wang, and Y. Cao, "Repeated Omicron exposures override ancestral SARS-CoV-2 immune imprinting," Nature **625**, 148–156 (2024).
23. J. Ito, A. Strange, W. Liu, G. Joas, S. Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium, and K. Sato, "A Protein Language Model for Exploring Viral Fitness Landscapes," (2024).
24. D. Wang, M. Huot, V. Mohanty, and E. I. Shakhnovich, "Biophysical principles predict fitness of sars-cov-2 variants," Proc. National Acad. Sci. **121**, e2314518121 (2024).
25. M. C. Maher, I. Bartha, S. Weaver, J. Di Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, D. L. Robertson, G. Snell, D. Corti, H. W. Virgin, S. L. Kosakovsky Pond, and A. Telenti, "Predicting the mutational drivers of future SARS-CoV-2 variants of concern," Sci. Transl. Med. **14**, eabk3445 (2022).
26. M. Huot, D. Wang, J. Liu, and E. Shakhnovich, "Few-shot viral variant detection via bayesian active learning and biophysics," bioRxiv (2025).
27. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, and J. D. Bloom, "Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2

Binding," Cell **182**, 1295–1310.e20 (2020).
28. J. Tubiana, S. Cocco, and R. Monasson, "Learning protein constitutive motifs from sequence data," eLife **8**, e39397 (2019).
29. E. Andreano, I. Paciello, G. Pierleoni, G. Piccini, V. Abbiento, G. Antonelli, P. Pileri, N. Manganaro, E. Pantano, G. Maccari, S. Marchese, L. Donnici, L. Benincasa, G. Giglioli, M. Leonardi, C. De Santi, M. Fabbiani, I. Rancan, M. Tumbarello, F. Montagnani, C. Sala, D. Medini, R. De Francesco, E. Montomoli, and R. Rappuoli, "B cell analyses after SARS-CoV-2 mRNA third vaccination reveals a hybrid immunity like antibody response," Nat. Commun. **14**, 53 (2023).
30. A. Baum, B. O. Fulton, E. Wloga, R. Copin, K. E. Pascal, V. Russo, S. Giordano, K. Lanza, N. Negron, M. Ni, Y. Wei, G. S. Atwal, A. J. Murphy, N. Stahl, G. D. Yancopoulos, and C. A. Kyratsous, "Antibody cocktail to sars-cov-2 spike protein prevents rapid mutational escape seen with individual antibodies," Science **369**, 1014–1018 (2020).
31. C. Hsu, H. Nisonoff, C. Fannjiang, and J. Listgarten, "Learning protein fitness models from evolutionary and assay-labeled data," Nat. Biotechnol. **40**, 1114–1122 (2022).
32. Y. Cao, A. Yisimayi, F. Jian, W. Song, T. Xiao, L. Wang, S. Du, J. Wang, Q. Li, X. Chen, Y. Yu, P. Wang, Z. Zhang, P. Liu, R. An, X. Hao, Y. Wang, J. Wang, R. Feng, H. Sun, L. Zhao, W. Zhang, D. Zhao, J. Zheng, L. Yu, C. Li, N. Zhang, R. Wang, X. Niu, S. Yang, X. Song, Y. Chai, Y. Hu, Y. Shi, L. Zheng, Z. Li, Q. Gu, F. Shao, W. Huang, R. Jin, Z. Shen, Y. Wang, X. Wang, J. Xiao, and X. S. Xie, "BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection," Nature **608**, 593–602 (2022).
33. L. E. Rosen, M. A. Tortorici, A. De Marco, D. Pinto, W. B. Foreman, A. L. Taylor, Y.-J. Park, D. Bohan, T. Rietz, J. M. Errico, K. Hauser, H. V. Dang, J. W. Chartron, M. Giurdanella, G. Cusumano, C. Saliba, F. Zatta, K. R. Sprouse, A. Addetia, S. K. Zepeda, J. Brown, J. Lee, E. Dellota, A. Rajesh, J. Noack, Q. Tao, Y. DaCosta, B. Tsu, R. Acosta, S. Subramanian, G. D. De Melo, L. Kergoat, I. Zhang, Z. Liu, B. Guarino, M. A. Schmid, G. Schnell, J. L. Miller, F. A. Lempp, N. Czudnochowski, E. Cameroni, S. P. Whelan, H. Bourhy, L. A. Purcell, F. Benigni, J. Di Iulio, M. S. Pizzuto, A. Lanzavecchia, A. Telenti, G. Snell, D. Corti, D. Veesler, and T. N. Starr, "A potent pan-sarbecovirus neutralizing antibody resilient to epitope diversification," Cell **187**, 7196–7213.e26 (2024).
34. Y. Cao, F. Jian, Z. Zhang, A. Yisimayi, X. Hao, L. Bao, F. Yuan, Y. Yu, S. Du, J. Wang, T. Xiao, W. Song, Y. Zhang, P. Liu, R. An, P. Wang, Y. Wang, S. Yang, X. Niu, Y. Zhang, Q. Gu, F. Shao, Y. Hu, W. Yin, A. Zheng, Y. Wang, C. Qin, R. Jin, J. Xiao, and X. S. Xie, "Rational identification of potent and broad sarbecovirus-neutralizing antibody cocktails from SARS convalescents," Cell Reports **41**, 111845 (2022).
35. Z. Ku, X. Xie, E. Davidson, X. Ye, H. Su, V. D. Menachery, Y. Li, Z. Yuan, X. Zhang, A. E. Muruato, A. G. I Escuer, B. Tyrell, K. Doolan, B. J. Doranz, D. Wrapp, P. F. Bates, J. S. McLellan, S. R. Weiss, N. Zhang, P.-Y. Shi, and Z. An, "Molecular determinants and mechanism for antibody cocktail preventing SARS-CoV-2 escape," Nat. Commun. **12**, 469 (2021).
36. T. C. Yu, Z. T. Thornton, W. W. Hannon, W. S. DeWitt, C. E. Radford, F. A. Matsen, and J. D. Bloom, "A biophysical model of viral escape from polyclonal antibodies," Virus Evol. **8**, veac110 (2022).
37. N. Youssef, S. Gurev, F. Ghantous, K. P. Brock, J. A. Jaimes, N. N. Thadani, A. Dauphin, A. C. Sherman, L. Yurkovetskiy, D. Soto, R. Estanboulieh, B. Kotzen, P. Notin, A. W. Kollasch, A. A. Cohen, S. E. Dross, J. Erasmus, D. H. Fuller, P. J. Bjorkman, J. E. Lemieux, J. Luban, M. S. Seaman, and D. Marks, "Protein design for evaluating vaccines against future viral variation," (2023).
38. W. Shi, J. Wohlwend, M. Wu, and R. Barzilay, "VaxSeer: Selecting influenza vaccines through evolutionary and antigenicity models," (2023).
39. J. Lou, W. Liang, L. Cao, I. Hu, S. Zhao, Z. Chen, R. W. Y. Chan, P. P. H. Cheung, H. Zheng, C. Liu, Q. Li, M. K. C. Chong, Y. Zhang, E.-k. Yeoh, P. K.-S. Chan, B. C. Y. Zee, C. K. P. Mok, and M. H. Wang, "Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations," Nat. Commun. **15**, 2546 (2024).
40. P. V. Markov, M. Ghafari, M. Beer, K. Lythgoe, P. Simmonds, N. I. Stilianakis, and A. Katzourakis, "The evolution of SARS-CoV-2," Nat. Rev. Microbiol. **21**, 361–379 (2023).
41. C. Planchais, I. Fernández, B. Chalopin, T. Bruel, P. Rosenbaum, M. Beretta, J. D. Dimitrov, L. Conquet, F. Donati, M. Prot, F. Porrot, D. Planas, I. Staropoli, F. Guivel-Benhassine, E. Baquero, S. Van Der Werf, A. Haouz, E. Simon-Lorière, X. Montagutelli, B. Maillère, F. A. Rey, P. Guardado-Calvo, H. Nozach, O. Schwartz, and H. Mouquet, "Broad sarbecovirus neutralization by combined memory B cell antibodies to ancestral SARS-CoV-2," iScience **27**, 110354 (2024).
42. N. D. Rochman, Y. I. Wolf, G. Faure, P. Mutz, F. Zhang, and E. V. Koonin, "Ongoing global and regional adaptive evolution of sars-cov-2," Proc. National Acad. Sci. **118**, e2104241118 (2021).
43. N. D. Rochman, G. Faure, Y. I. Wolf, L. Freddolino, F. Zhang, and E. V. Koonin, "Epistasis at the sars-cov-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape," MBio **13**, e00135–22 (2022).
44. S. Chang, K.-S. Shin, B. Park, S. Park, J. Shin, H. Park, I. K. Jung, J. H. Kim, S. E. Bae, J.-O. Kim *et al.*, "Strategy to develop broadly effective multivalent covid-19 vaccines against emerging variants based on ad5/35 platform," Proc. National Acad. Sci. **121**, e2313681121 (2024).
45. A. A. Cohen, N. van Doremalen, A. J. Greaney, H. Andersen, A. Sharma, T. N. Starr, J. R. Keeffe, C. Fan, J. E. Schulz, P. N. Gnanapragasam *et al.*, "Mosaic rbd nanoparticles protect against challenge by diverse sarbecoviruses in animal models," Science **377**, eabq0839 (2022).
46. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. De Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker, "Robust deep learning–based protein sequence design using ProteinMPNN," Science **378**, 49–56 (2022).
47. M. Philipp, C. Moth, N. Ristic, J. Tiemann, F. Seufert, A. Panfilova, J. Meiler, P. Hildebrand, A. Stein, D. Wiegreffe, and R. Staritzbichler, "MutationExplorer : a webserver for mutation of proteins and 3D visualization of energetic impacts," Nucleic Acids Res. **52**, W132–W139 (2024).
48. S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy," Glob. Challenges **1**, 33–46 (2017).
49. K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," Mol. Biol. Evol. **30**, 772–780 (2013).
50. S. J. Zost, P. Gilchuk, J. B. Case, E. Binshtein, R. E. Chen, J. P. Nkolola, A. Schäfer, J. X. Reidy, A. Trivette, R. S. Nargi, R. E. Sutton, N. Suryadevara, D. R. Martinez, L. E. Williamson, E. C. Chen, T. Jones, S. Day, L. Myers, A. O. Hassan, N. M. Kafai, E. S. Winkler, J. M. Fox, S. Shrihari, B. K. Mueller, J. Meiler, A. Chandrashekar, N. B. Mercado, J. J. Steinhardt, K. Ren, Y.-M. Loo, N. L. Kallewaard, B. T. McCune, S. P. Keeler, M. J. Holtzman, D. H. Barouch, L. E. Gralinski, R. S. Baric, L. B. Thackray, M. S. Diamond, R. H. Carnahan, and J. E. Crowe, "Potently neutralizing and protective human antibodies against SARS-CoV-2," Nature **584**, 443–449 (2020).
51. P. Wang, M. S. Nair, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, B. S. Graham, J. R. Mascola, J. Y. Chang, M. T. Yin, M. Sobieszczyk, C. A. Kyratsous, L. Shapiro, Z. Sheng, Y. Huang, and D. D. Ho, "Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7," Nature **593**, 130–135 (2021).
52. A. Moulana, T. Dupic, A. M. Phillips, J. Chang, A. A. Roffler, A. J. Greaney, T. N. Starr, J. D. Bloom, and M. M. Desai, "The landscape of antibody binding affinity in SARS-CoV-2 omicron BA.1 evolution," Elife **12** (2023).
53. T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the 25th International Conference on Machine Learning,* (Association for Computing Machinery, New York, NY, USA, 2008), ICML '08, p. 1064–1071.

Code Availability

The code used in this study will be made publicly available upon manuscript acceptance. For access requests prior to publication, please contact the corresponding authors.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Author contributions

Competing interests

The authors declare no competing interests.

Figures

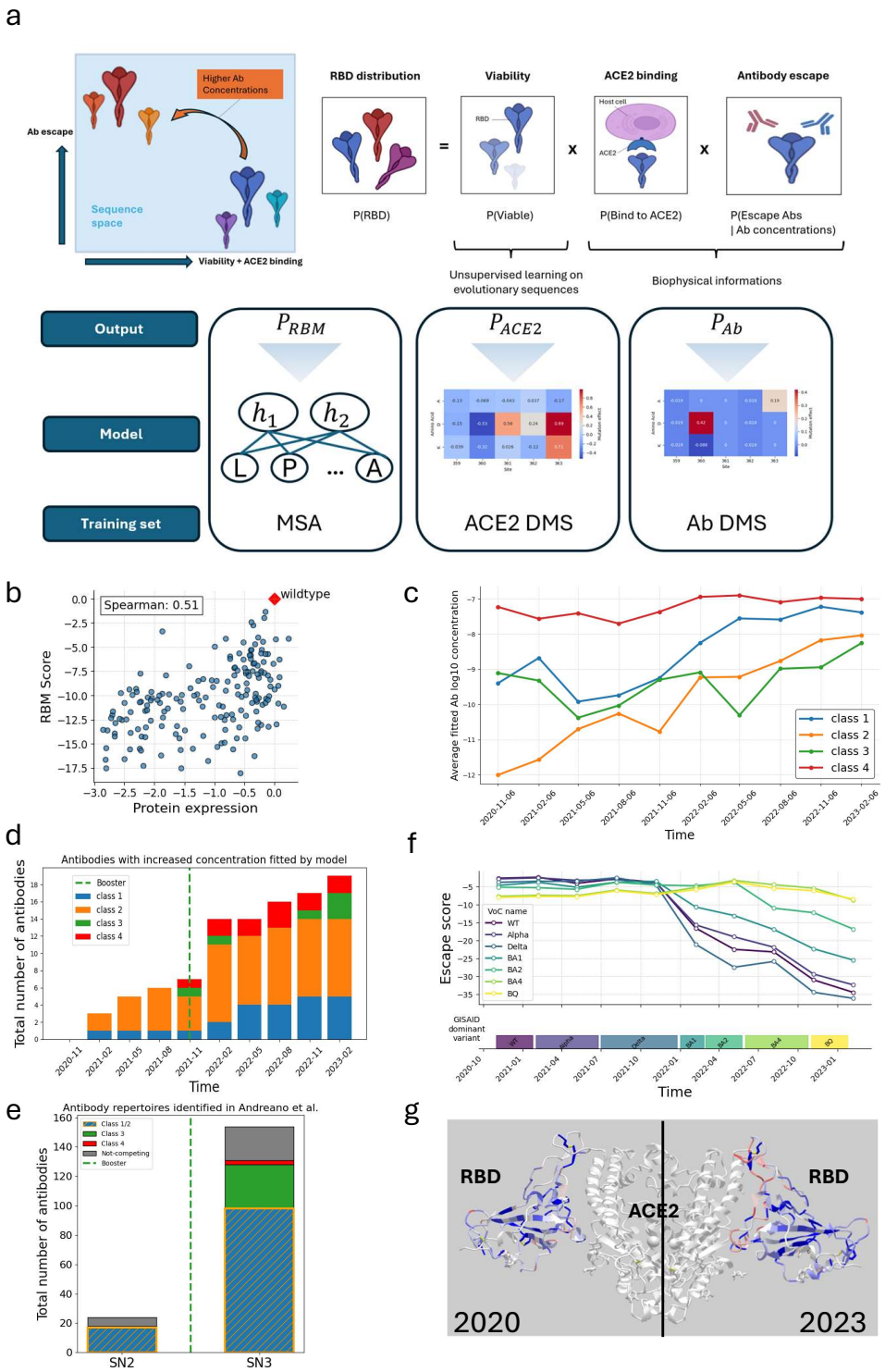

Fig. 1. Modeling of the SARS-CoV-2 RBD sequence landscape. (a) Model architecture. Variants with high likelihood must be viable and show good ACE2 binding and antibody escape. Viability is derived from a RBM fitted on a multiple-sequence alignment of 2610 prepandemic Coronaviridae sequences, mainly SARS-CoV-2 homologs. Dissociation constants with ACE2 and antibodies are obtained with a linear model fitted on deep mutational scans. (b) Correlation between RBM mutational scores and experimentally measured changes in expression (Starr et al.) of single-site mutants, averaged per site, in the Wuhan background. (c) Average $\log_{10}$ Ab concentrations (g/ml) fitted by the model as a function of time by antibody class. (d) Number of antibodies with fitted concentration that increased by a factor of 10 compared to the initial time period, per class and across times. (e) Diversity of antibodies found in serum after 2 vaccination (SN2) and 3 vaccination (SN3) doses (Andreano et al.). (f) Antibody escape score computed on all antibodies of VOCs across times. (g) RBD complexed with ACE2 (white), colored by model maximal mutational score over each residue with respect to wildtype, using the full ACE2–AB–RBM model, and fitted Ab concentration of 2020 vs. 2023. Red indicates sites where mutation could benefit the protein (score>0); blue indicates sites that should be conserved (score<0).

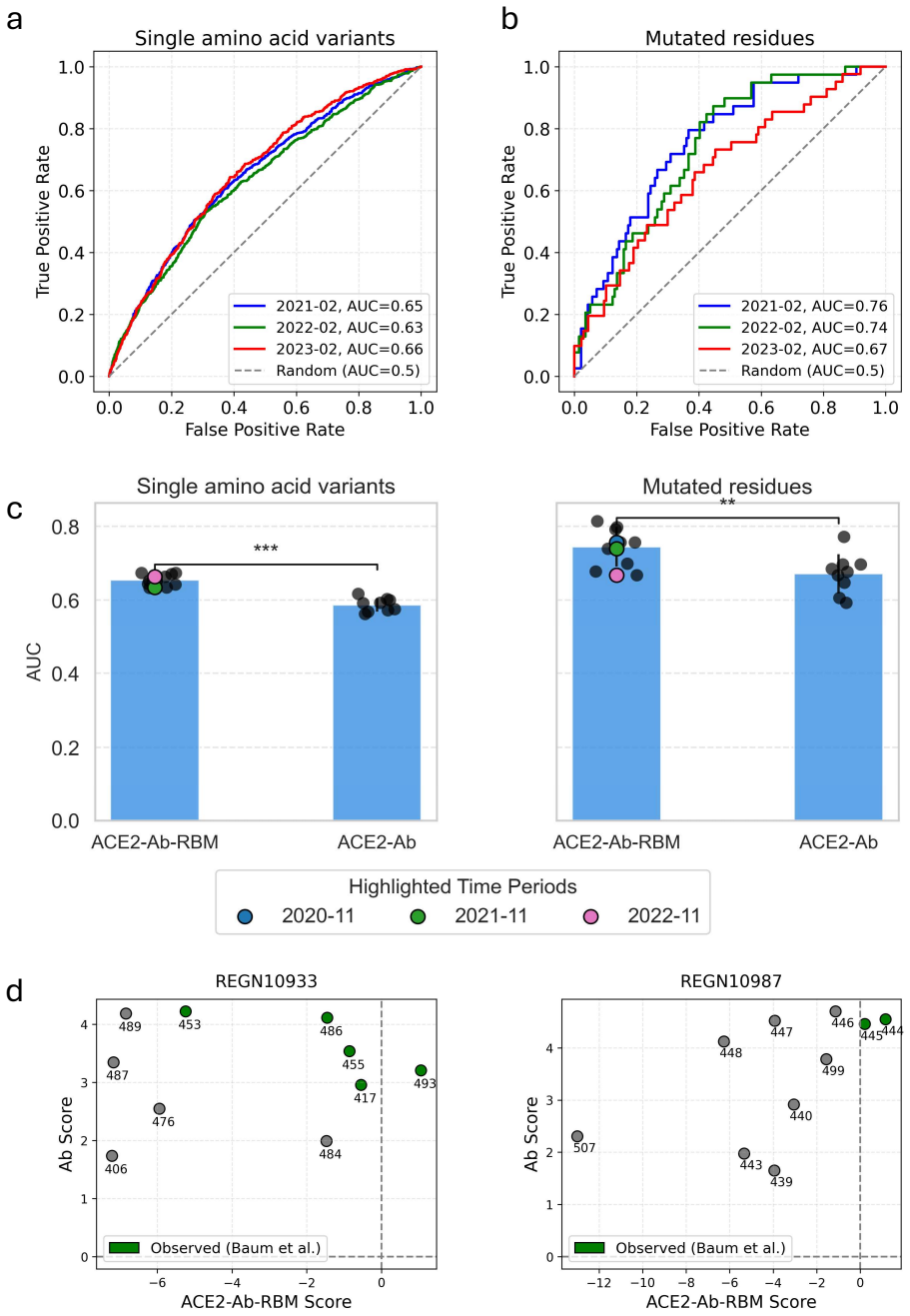

Fig. 2. Model predictive power. (a) AUC of model prediction for the appearance of mutations at different times. (b) AUC of model prediction for the site-specific tolerance to a high number of unique mutations at different times. (c) Predictive performance (AUC) of the full model including ACE2–Ab–RBM probabilities versus the ACE2–Ab baseline across different timepoints. Performance is assessed on two tasks: prediction of single-amino-acid variant effects (left) and site-level mutational tolerance (right). Asterisks indicate statistically significant differences based on paired t-tests ($p<0.01$ for both tasks). (d) Scatter plot of scores (complete model—ACE2–AB–RBM— vs. antibody escape—AB score—) of the top 10 escape sites (ranked by antibody escape score). Sites with positive scores have a high likelihood to mutate. Observed escape sites from serial passaging (Baum et al.) are highlighted in green.

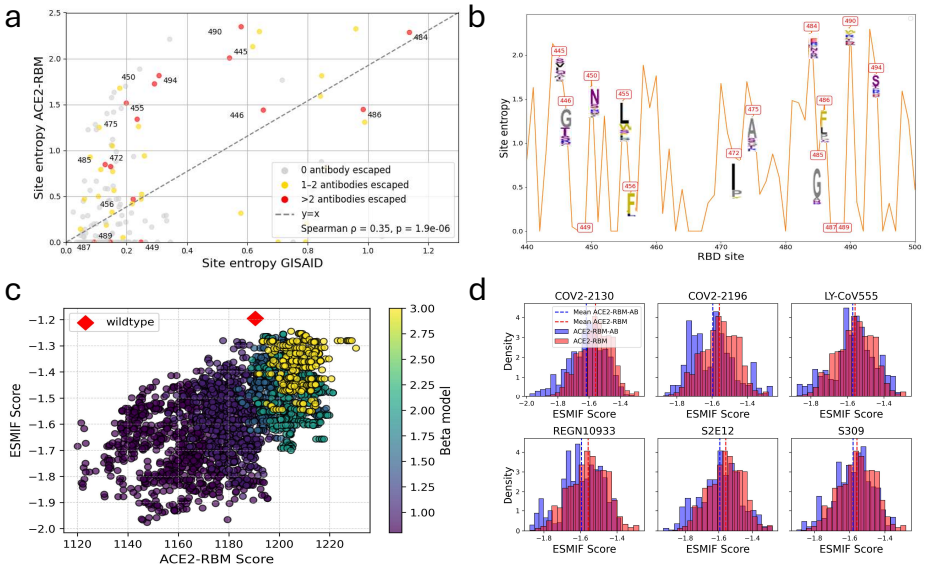

Fig. 3. Model generative power. (a) Comparison of site entropies of unique sequences generated using the ACE2–RBM model and sequences found in GISAID. Colors indicate how many antibodies can be escaped (site max escape>0.9 in DMS) when mutating that site. (b) Site entropy on generated sequences for residues 440–500. Sites involved in escape for at least three antibodies are highlighted in red, with the corresponding amino acids shown when different from wildtype. (c) Scatter plot of ESM Inverse Folding (ESMIF) score versus ACE2–RBM score for generated variants at inverse temperatures $\beta \in \{1, 1.5, 2, 3\}$. (d) Histograms of ESMIF scores of variants generated using the complete model (ACE2–RBM–AB) with one antibody (indicated) at high concentration $\log_{10} C = -6$, compared to histograms for sequences generated with the ACE2–RBM model (no antibody pressure).

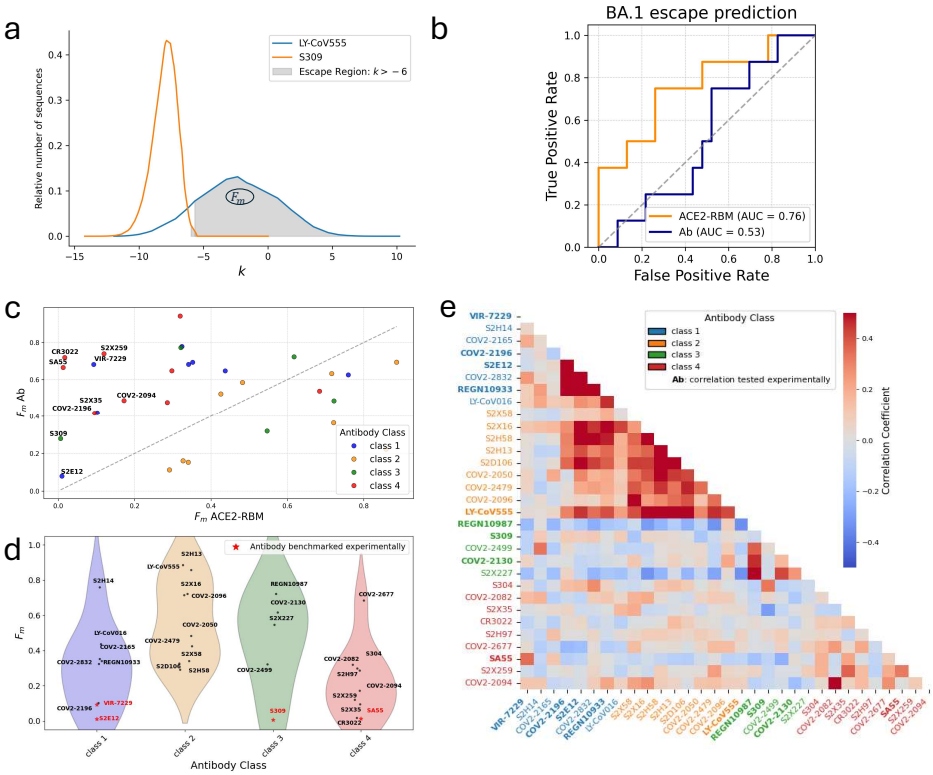

Fig. 4. Antibody resilience. (a) Fraction F_m of sequences escaping antibody S309 and LY-CoV555. F_m is the integral of the escape region $k > -6$. (b) AUC for predicting model BA.1 escape (Cao et al., 15 antibodies) using either the ACE2–RBM model or the AB model based on DMS data. (c) Scatter plots of fraction F_m of sequences escaping antibody, when sampling from the ACE2–RBM model vs. the AB model. Low F_m values correspond to antibodies resilient to viral adaptation. (d) Fraction F_m of sequences escaping antibody, when sampling from the ACE2–RBM model, grouped by antibody class. Resilient antibodies are underlined (SA55, VIR-7229, S2E12, S309), selected for in vitro tests under high antibody pressure. (e) Model correlations between antibody dissociation constants. Red indicates positive correlations; blue indicates negative correlations. Experimentally validated correlations are highlighted.

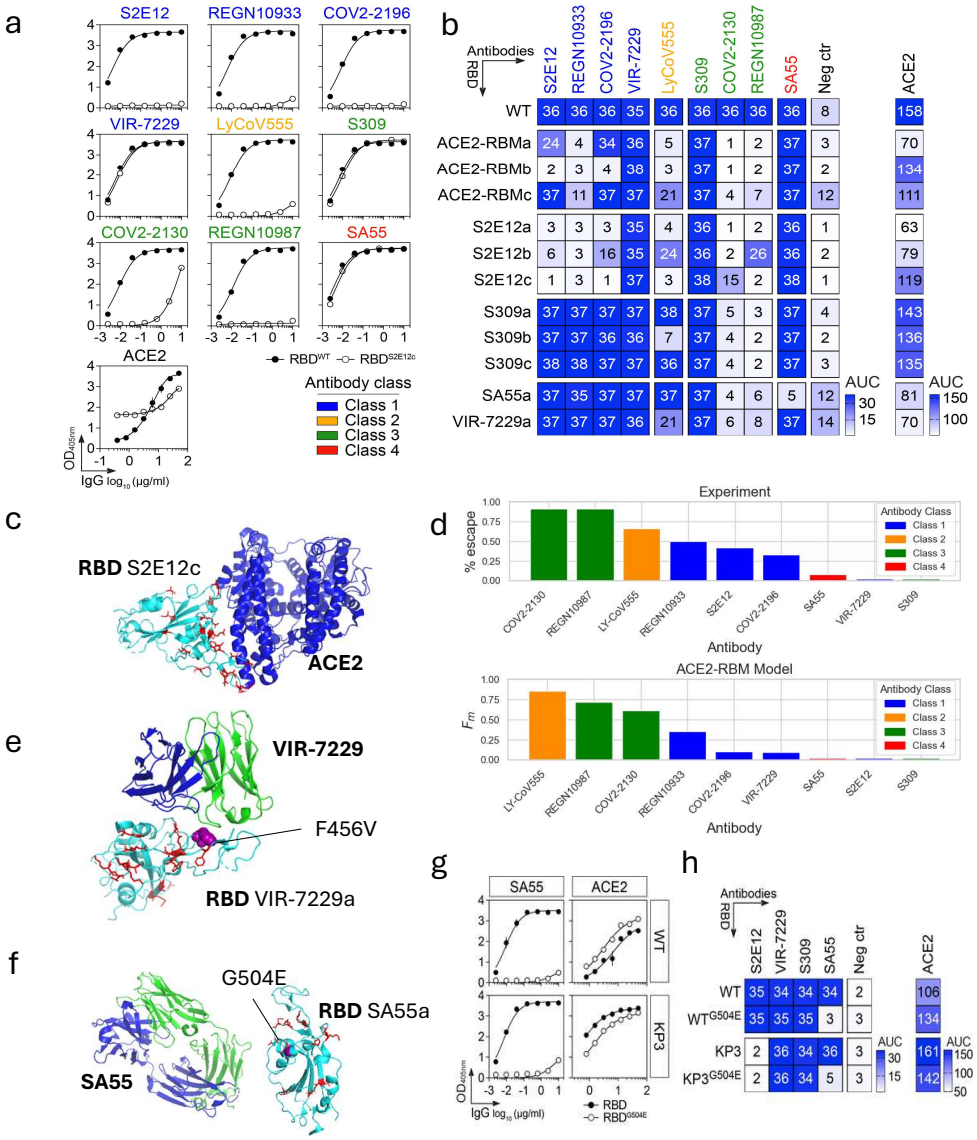

Fig. 5. In-vitro model validation. (a) Representative ELISA graphs showing the binding of anti-RBD antibodies and ACE2 to purified wildtype and mutated RBD S2E12c, generated with S2E12 immune pressure. (b) Heatmap of ELISA binding of anti-RBD antibodies and ACE2 to purified RBDs (higher scores = higher binding). Conditions: wildtype, ACE2–RBM (no antibody), ACE2–Ab-RBM ($c = -6$). (c) Structural representation of the RBD (light blue) in complex with ACE2 (dark blue) (PDB: 6M0J). Red highlights indicate sites with a mutation in the tested and expressed variant S2E12c. (d) Percentage of functional sequences escaping each antibody in experiment, compared to model-derived fraction F_m of sequences escaping antibody. (e) Structural representation of RBD in complex with ACE2 and the VIR-7229 antibody (PDB: 9ASD). Red = mutations in functional sequences; purple spheres = residue 456 (critical for binding). (f) As in (e), for the SA55 antibody (PDB: 7Y0W); purple spheres = residue 504 (critical for binding). (g) Representative ELISA graphs showing the binding of antibody SA55 and ACE2 to purified wildtype, new variant KP.3, and their G504E mutants. (h) Heatmap of ELISA binding of anti-RBD antibodies and ACE2 to purified wildtype, new variant KP.3, and their G504E mutants.

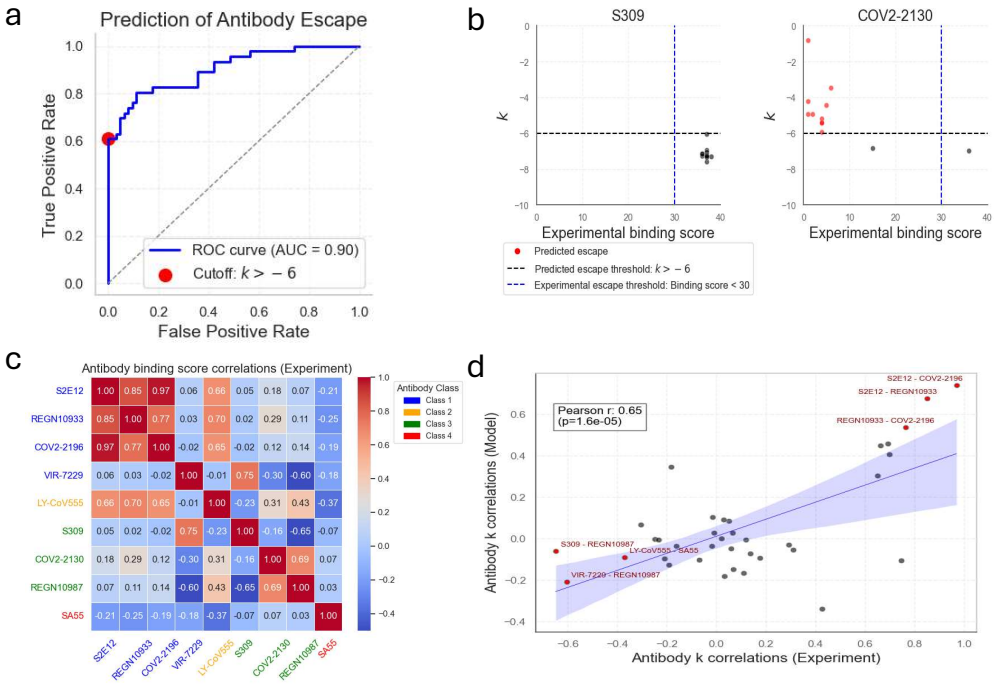

Fig. 6. Antibody escape prediction. (a) ROC curve for model prediction of antibody escape (binding score < 30). (b) Comparison of measured and predicted binding scores for antibodies S309 and COV2-2130. Each point corresponds to a functional sequence for which antibody dissociation constants were measured. (c) Correlation of antibody binding scores in experiment. (d) Correlations between antibody dissociation constants for cocktail design, by experiment vs. model. Pairs with highest/lowest experimental correlations are highlighted.

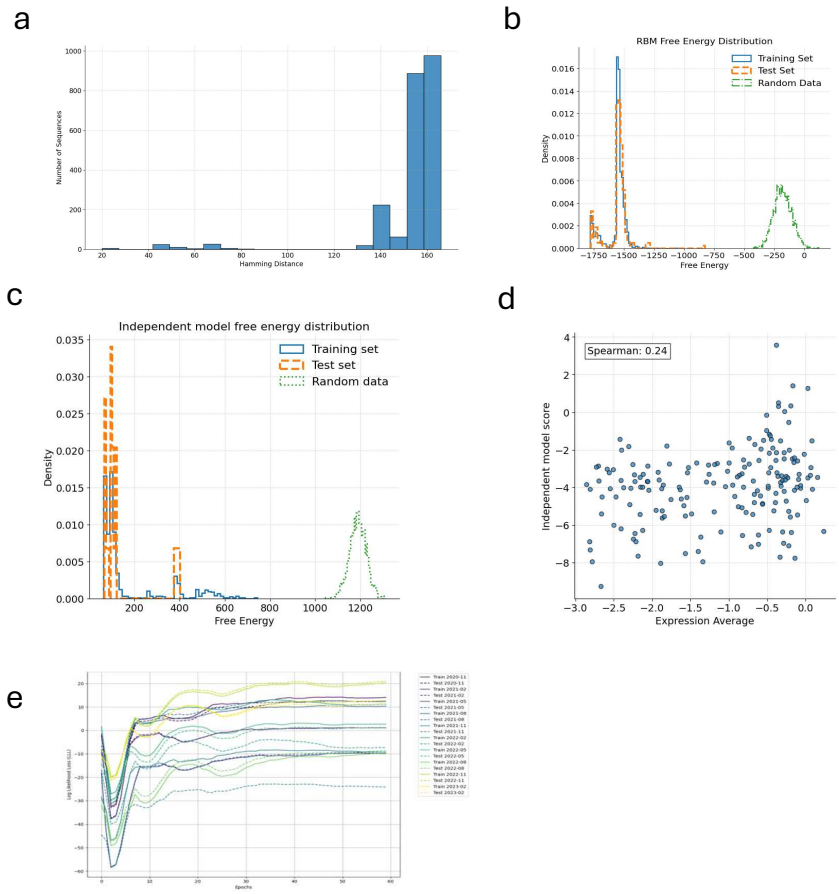

Fig. S1. Model fit. (a) Distances between RBD-related sequences in the MSA used to fit the RBM and wildtype RBD (Wuhan). (b) Free energy of train (90%), test (10%), and random sequences after fitting the RBM on the MSA (lower energy = higher likelihood). (c) As in (b) for an independent model on the MSA. (d) Correlation between independent-model mutability score and experimentally measured expression changes, averaged per site on Wuhan background. (e) Likelihoods of train/test sets per period during full-model training.

a

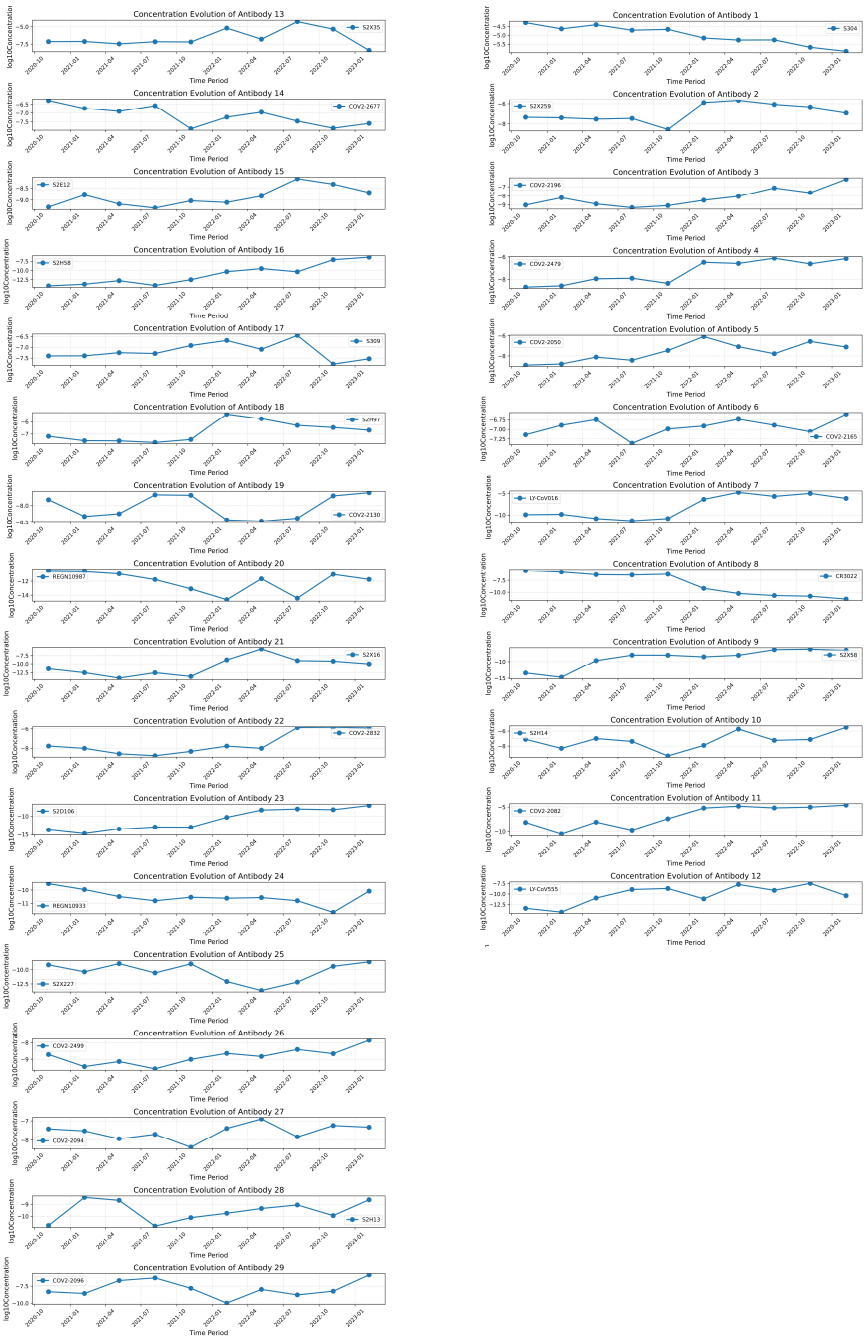

Fig. S2. Fitted antibody concentrations. (a) Antibody concentrations fitted by the full model during each pandemic period.

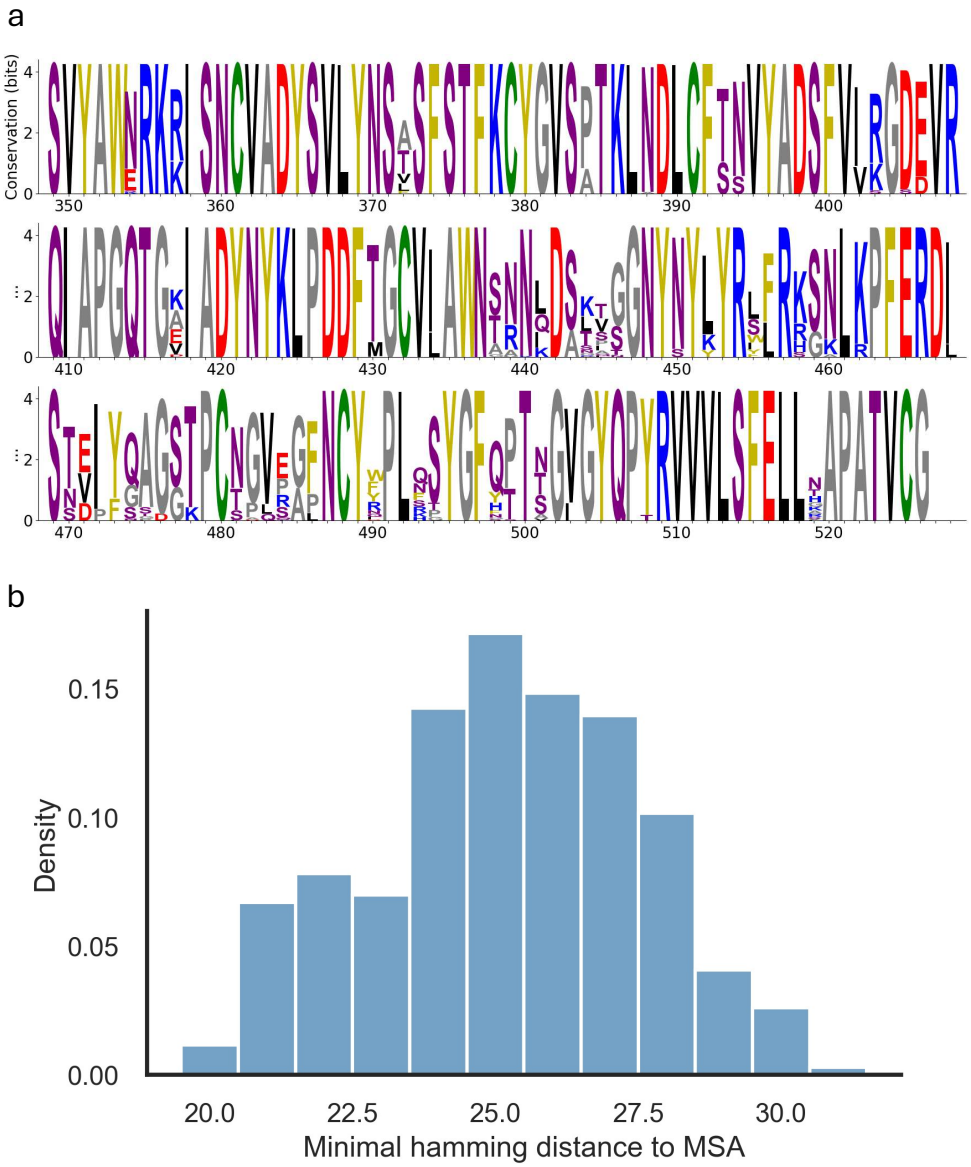

Fig. S3. Generated sequences at $\beta = 3$. (a) Sequence logos of sequences generated with the ACE2–RBM model. (b) Distribution of minimal Hamming distances between generated sequences with the ACE2–RBM model and the natural MSA, quantifying how far generated variants deviate from evolutionarily observed sequences.

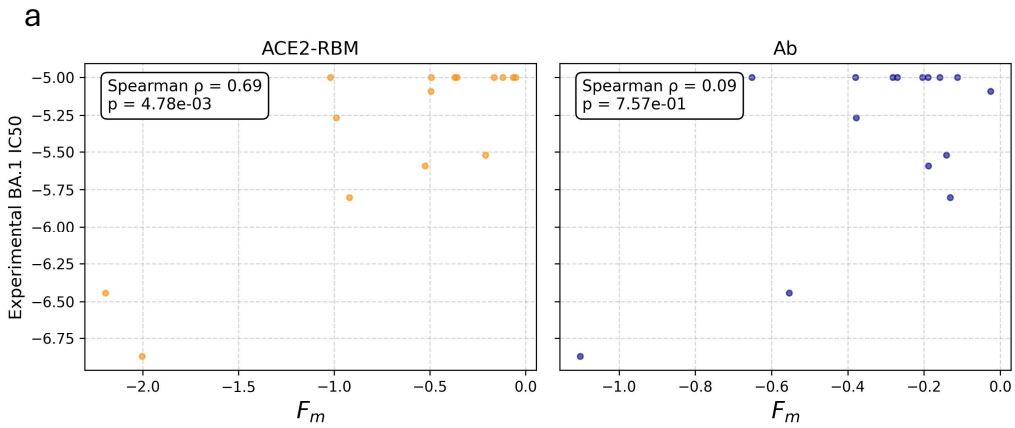

Fig. S4. Prediction of antibody escape. (a) Comparison of experimentally measured IC_{50} on BA.1 (Cao et al., 15 antibodies) and fraction F_m of sequences escaping antibody when sampling from ACE2–RBM or AB models.

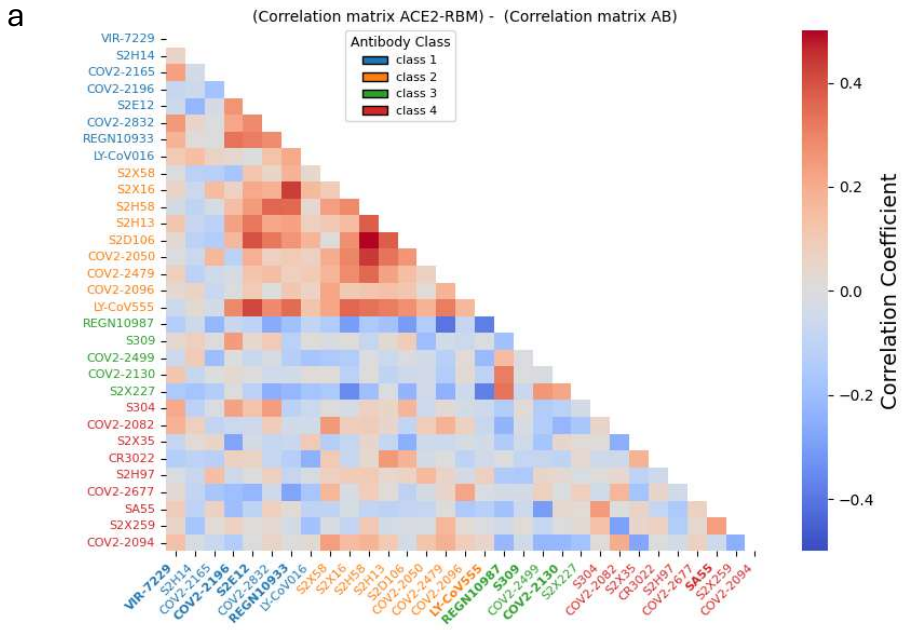

Fig. S5. Correlation between antibody dissociation constants. (a) Difference between correlation matrices of antibody dissociation constants, for sequences sampled using ACE2–RBM vs. uniform distribution.

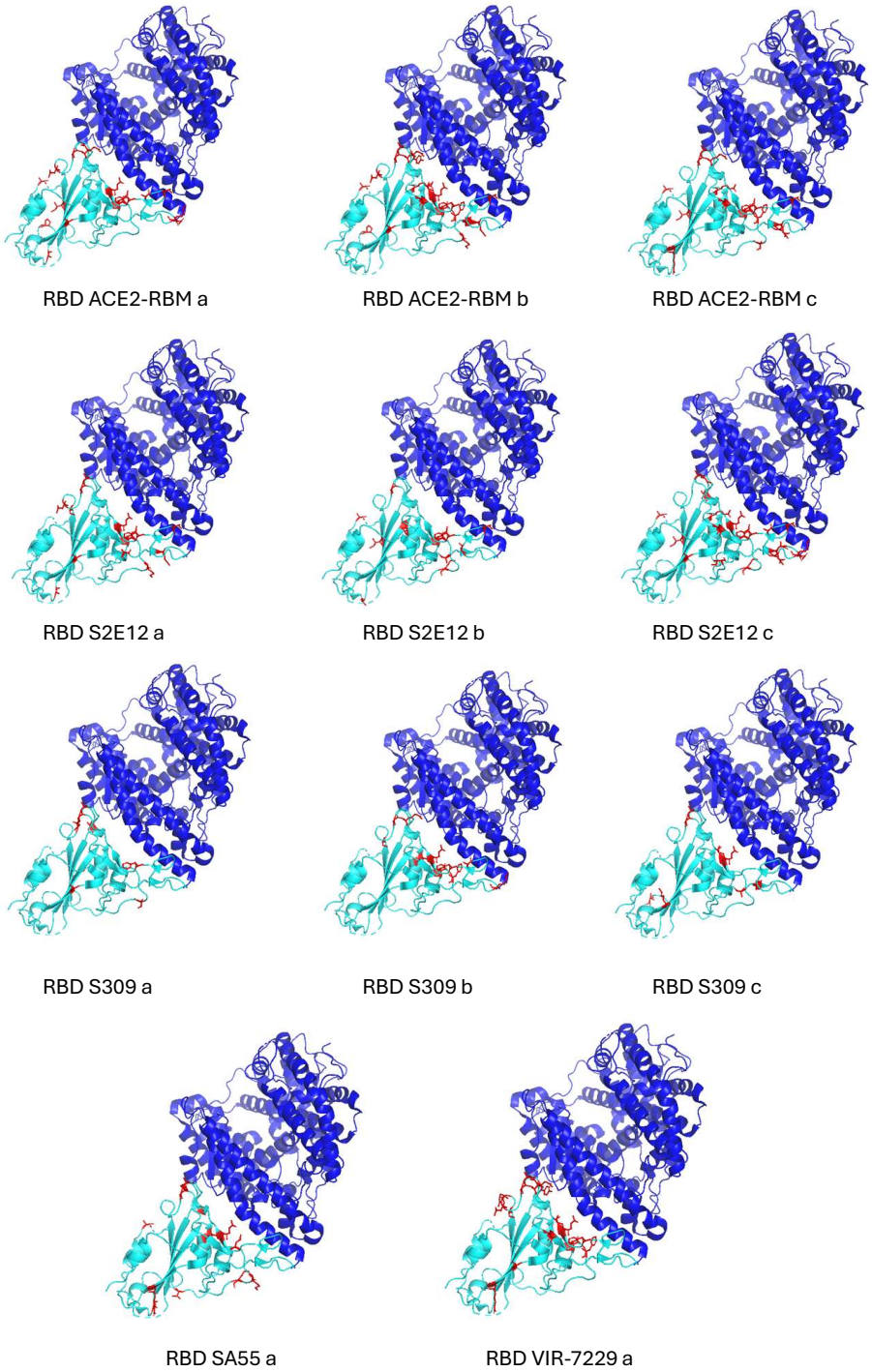

Fig. S6. Mutated sites in expressed sequences. Structural representation of RBD, in complex with ACE2 (PDB: 6M0J), for every tested sequence that could be expressed. Red = mutations in sequence.

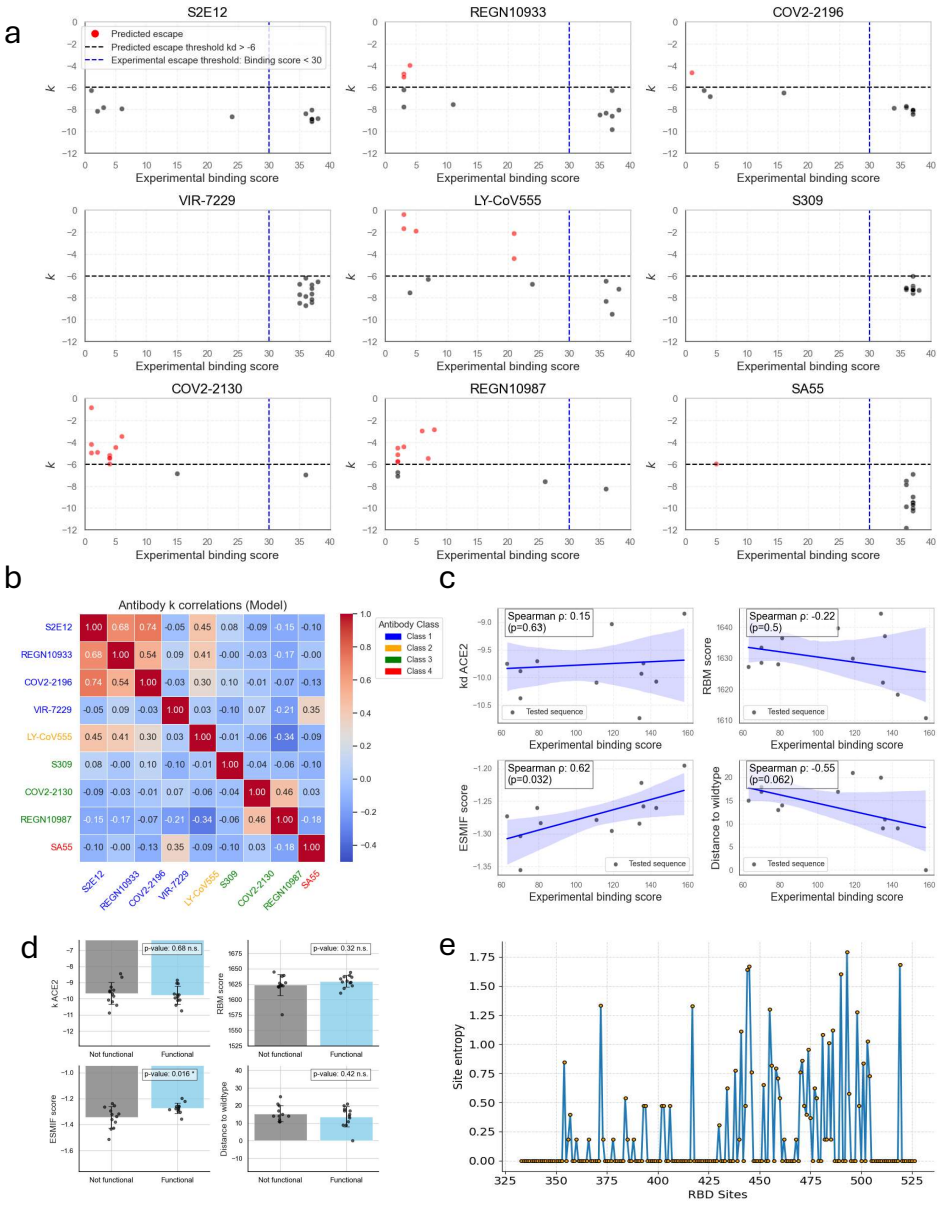

Fig. S7. Measured binding for generated sequences. (a) Comparison of measured and predicted binding scores for all antibodies. Each point corresponds to a functional sequence for which antibody dissociation constants were measured. (b) Correlation between antibody binding scores when sampling sequences with the ACE2–RBM model. (c) Correlation between ACE2 binding measured in experiment for functional sequences and (i) distance to wildtype, (ii) RBM energy, (iii) predicted ACE2 log10 dissociaiton constant, and (iv) ESMIF score. (d) Comparison of RBM score, predicted ACE2 log10 dissociaiton constant , ESMIF score, and distance to wildtype between functional sequences that could be expressed and non-functional sequences. Each bar shows the mean ± standard deviation, with individual data points overlaid. Statistical significance is indicated using two-sided t-tests ($p < 0.05$: *, $p > 0.05$: ns). (e) Site entropy in the designed sequences, plotted along the RBD sequence.

a

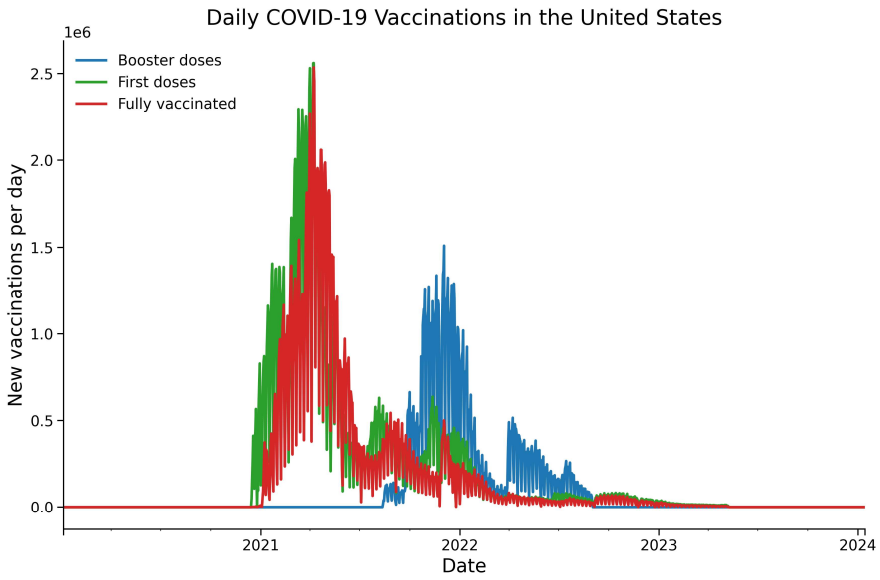

Fig. S8. Vaccinations. (a) Daily new vaccinations in the United States. Fully vaccinated individuals received two doses, while booster corresponds to the third dose.