Generative model of SARS-CoV-2 variants under immune pressure unveils viral escape potential

Marian Huot[1], Pierre Rosenbaum[2], Cyril Planchais[2], Hugo Mouquet[2,*], Rémi Monasson[1], and 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
[*]Correspondence: hugo.mouquet@pasteur.fr
[**]Correspondence: simona.cocco@phys.ens.fr

SUMMARY

The evolutionary trajectory of SARS-CoV-2 is shaped by competing pressures for ACE2 binding, structural viability, and escape from neutralizing antibodies targeting its receptor-binding domain (RBD). Here, we present EscapeMap, a modular framework that quantifies immune selection and predicts which antibodies are more or less likely to be escaped by future viral evolution. EscapeMap integrates deep mutational scanning data for ACE2 and 31 monoclonal antibodies with a generative sequence model trained on pre-pandemic Coronaviridae. To probe escape potential, we designed RBD variants under pressure from four clinically relevant antibodies (SA55, S2E12, S309, VIR-7229). Among these designs, bearing up to 21 mutations from wildtype, 50% expressed as stable proteins. Binding assays confirm that S309 and VIR-7229 retain recognition across diverse mutation combinations. EscapeMap accurately forecasts which antibodies are vulnerable to escape by our designed sequences. Finally, by identifying negatively correlated escape routes, we prioritize antibody combinations less prone to simultaneous escape, offering a quantitative basis for guiding therapeutic strategies.

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 DMS 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 EscapeMap, 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 EscapeMap'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 experiments, 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 EscapeMap, 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 13) 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. 13).

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. Crucially, EscapeMap is modular: antibody pressure can be excluded (concentration $C = 0$), yielding a generative landscape defined by ACE2 binding and RBM-based viability only (ACE2–RBM), or included to account for immune selection via antibody binding probabilities.

Fit of EscapeMap to sequences from pandemic data

After training an RBM on an MSA containing data from coronaviruses predating SARS-CoV-2 (see SI Figures S1a-b), 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:

$$\Delta \mathrm{LL}_i^{\mathrm{RBM}} = \langle \log P_{\mathrm{RBM}}(v_i, wt_{j\neq i}) \rangle_{v_i} - \log P_{\mathrm{RBM}}(wt_i, wt_{j\neq i}) \tag{2}$$

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 Figure S1d).

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 (see SI Figure S1e). 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. These parameters were then kept constant for all this study. 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 (SI Figure S2b). 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 (SI Figure S2c).

The observed increase in fitted antibody concentrations suggests a progressive shift in the accessible variant space. Early in the pandemic, most residues showed no selective advantage for mutation (Figure 1d), consistent with a small number of mutations in observed variants in variants Alpha and Delta. 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 1e). 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 EscapeMap 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 EscapeMap'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:

$$\Delta LL_i = \langle LL(v_i, \{wt_{j\neq i}\})\rangle_{v_i} - LL(wt_i, \{wt_{j\neq i}\}) \tag{3}$$

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.

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 EscapeMap's 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 EscapeMap to generate new variants (Eq. (1)) via Monte Carlo sampling (see Methods). We first generated 1,000 sequences using the EscapeMap without antibody pressure (ACE2-RBM only), and found that their site-wise entropy (see Methods) 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 EscapeMap 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. (12) 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. Notably, for $\beta = 3$, the generated sequences achieve ESM-IF scores comparable to that of the wild-type RBD (Figure 3c), indicating that they likely adopt the native RBD fold (PDB: 6M0J), despite being $\sim$20–30 mutations away from any sequence in the MSA (see SI Figure S3b). We then explored the influence of antibody selection pressures on the generated sequences (see SI Figure S3c). Mutants generated using EscapeMap with Ab selection pressure of one antibody exhibited ESM-IF scores comparable to those generated without antibody pressure, indicating that the overall structural and functional integrity of sequences is preserved across different selection regimes.

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:

$$F_m = \int_{k_{esc}}^{\infty} f_m(k_m) dk_m, \tag{4}$$

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):

$$f_m(k_m) = \frac{\sum_v \delta(k_m - k_m(v))\; P_{RBM}(v)\; P_{ACE2}(v)}{\sum_v P_{RBM}(v)\; P_{ACE2}(v)}\;. \tag{5}$$

The F_m score represents the probability of sequences sampled by EscapeMap to escape a given antibody, integrating over the full range of log10 dissociation constants k above the escape threshold $k_{esc} = -6$; 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 omicron variant BA.1 escape to antibodies, 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, see Eq. (5), ensuring that only functional variants contribute meaningfully to the score. In contrast, treating all sequences ν as equiprobable in Eq. (5) 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 EscapeMap 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 EscapeMap does.

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

$$\text{Cov}(Ab_1, Ab_2) = \sum_{\{v_i, v_j\}} \Delta k^{(1)}(v_i)\ \Delta k^{(2)}(v_j) \text{Cov}(v_i, v_j), \tag{6}$$

where Cov represents the covariance of amino-acid occurrences when sampling sequences from the distribution $P(v)$, and $\Delta k^{(m)}(v_i)$ 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 $\Delta k^{(1)}(v_i)\Delta k^{(2)}(v_i) < 0$. Furthermore, pairs of mutations (v_i, v_j) that independently contribute to increasing both dissociation constants, characterized by $\Delta k^{(1)}(v_i)\Delta k^{(2)}(v_i) > 0$, should be incompatible and exhibit negative correlations, 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 EscapeMap's prediction of a positive correlation (Cov = 0.2). In contrast, the REGN10933+REGN10987 combination, predicted by EscapeMap 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 EscapeMap 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-Ab-RBM model), which demonstrated strong resilience in EscapeMap's 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).

We achieved a 50% success rate among the tested sequences when evaluating in vitro production of the designed variants (Table 1), a remarkable hit rate given the mutation burden. 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 ACE2 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 ACE2, 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. This discrepancy 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

A key strength of EscapeMap lies in its modular architecture, which integrates distinct biophysical constraints—ACE2 binding, antibody escape, and sequence viability—into a unified energy-based framework. This modularity enables us to isolate and quantify the contribution of immune pressure and receptor binding by fitting antibody- and ACE2-specific concentration parameters from temporal pandemic data. As a result, the model not only predicts which sites are likely to mutate, but also which amino acids are favored at each site under different selective regimes. When immune pressure is added via fitted antibody concentrations, we observe clear shifts in predicted site variability, particularly at epitopes targeted by dominant antibody classes, demonstrating the model's ability to capture adaptive trajectories. Moreover, we find that the entropy of the generated sequence distribution exceeds that observed during real-world SARS-CoV-2 evolution, reflecting the fact that natural selection operates under additional bottlenecks and epidemiological constraints[37]. This capacity to simulate high-fitness, high-diversity escape variants under defined selection pressures offers a valuable tool for forecasting viral adaptation.

In addition, we introduced a rigorous metric for quantifying antibody resistance to viral adaptation, which we call antibody resilience. Traditionally, antibody effectiveness is evaluated based on neutralization assays on past and currently circulating variants, as well as structural epitope mapping[17,38]. 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[39,40]. 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. These insights are particularly relevant for guiding next-generation vaccine strategies. Mosaic vaccines, for instance, co-display multiple diverse RBDs on a single nanoparticle to steer immune responses toward conserved epitopes and promote broad sarbecovirus cross-protection[41]. By mapping viable escape routes, our model offers a principled framework to anticipate immune evasion and support the rational design of such broad-coverage immunogens.

What's more, a major strength of our framework is its ability to generate highly mutated RBD variants that remain viable. Among 22 designed sequences, carrying up to 21 mutations from the Wuhan wild-type, we observed a 50% success rate in expression. This is particularly notable given previous study on the same protein reporting that expression drops sharply with increasing mutational load, with hit rates falling below 2% beyond 8 mutations[27]. Strikingly, we do not observe a significant correlation between sequence distance to wild-type and experimental viability, in contrast to trends reported in prior literature where mutational distance predicted loss of function[42]. This suggests that our model does not merely interpolate around known variants but identifies viable configurations even in remote regions of sequence space. We attribute this to the RBM term, which, although trained solely on prepandemic sequences, captures deep co-

evolutionary constraints that implicitly reflect aspects of structural and functional compatibility. While the model lacks explicit tertiary structure modeling, its ability to propose highly divergent yet functional variants highlights its utility as a generative engine for probing the boundaries of immune escape and protein viability. Finally, while approaches based on EVEscape model can suggest plausible RBD variants[12,43], they lack the antibody-level granularity required for experimental control and interpretation. Our model allows direct tuning of immune selection through antibody-specific concentrations, enabling the design and comparison of escape routes under distinct immune pressures.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Simona Cocco (simona.cocco@phys.ens.fr).

Data and code availability

The code used in this study will be made publicly at `https://github.com/m-huot/Trajectories_covid` upon manuscript acceptance. For access requests prior to publication, please contact the corresponding authors.

ACKNOWLEDGMENTS

We acknowledge funding from the Agence Nationale de la Recherche (ANR-19 Decrypted CE30-0021-01 to S.C. and R.M.). H.M. received core grant from the Institut Pasteur. P.R. was supported by a two-years postdoctoral fellowship from the Pasteur-Roux-Cantarini program (Institut Pasteur).

We thank Dianzhuo Wang for fruitful discussions. 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

M.H., H.M., R.M., and S.C. conceived the project, supervised the research, and wrote the manuscript. M.H., R.M., and S.C. developed the computational model and performed data analysis. P.R., C.P. and H.M. designed and carried out the experimental and biological validation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION INDEX

Figures S1–S7. Supplementary figures and legends, included in a separate PDF file.

Main Figures and Tables

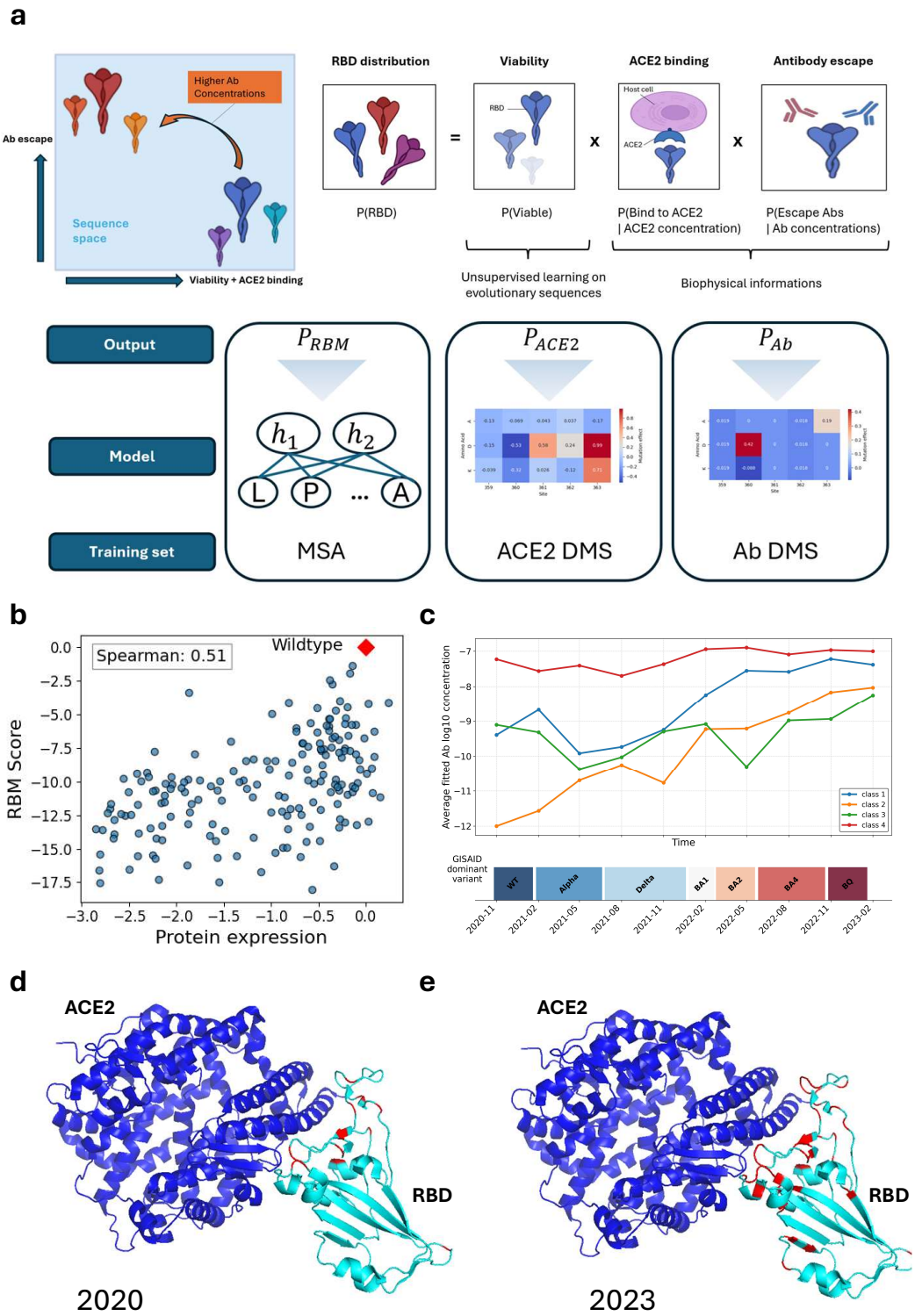

Fig. 1. EscapeMap captures shifting mutational landscape and rising immune pressure during the pandemic. (a) Model architecture. Variants with high likelihood must be viable and show good ACE2 binding and antibody escape. Viability is derived from a RBM fitted on a multiple-sequence alignment of 2610 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) RBD (light blue) complexed with ACE2 (dark blue), colored by model maximal mutational score over each residue compared to wildtype, using EscapeMap with fitted Ab concentration of 2020 (PDB: 6M0J). Red indicates sites where at least one mutation has a higher likelihood than the wildtype amino acid; light blue indicates sites that should be conserved. (e) As in (d), using fitted Ab concentration of 2023.

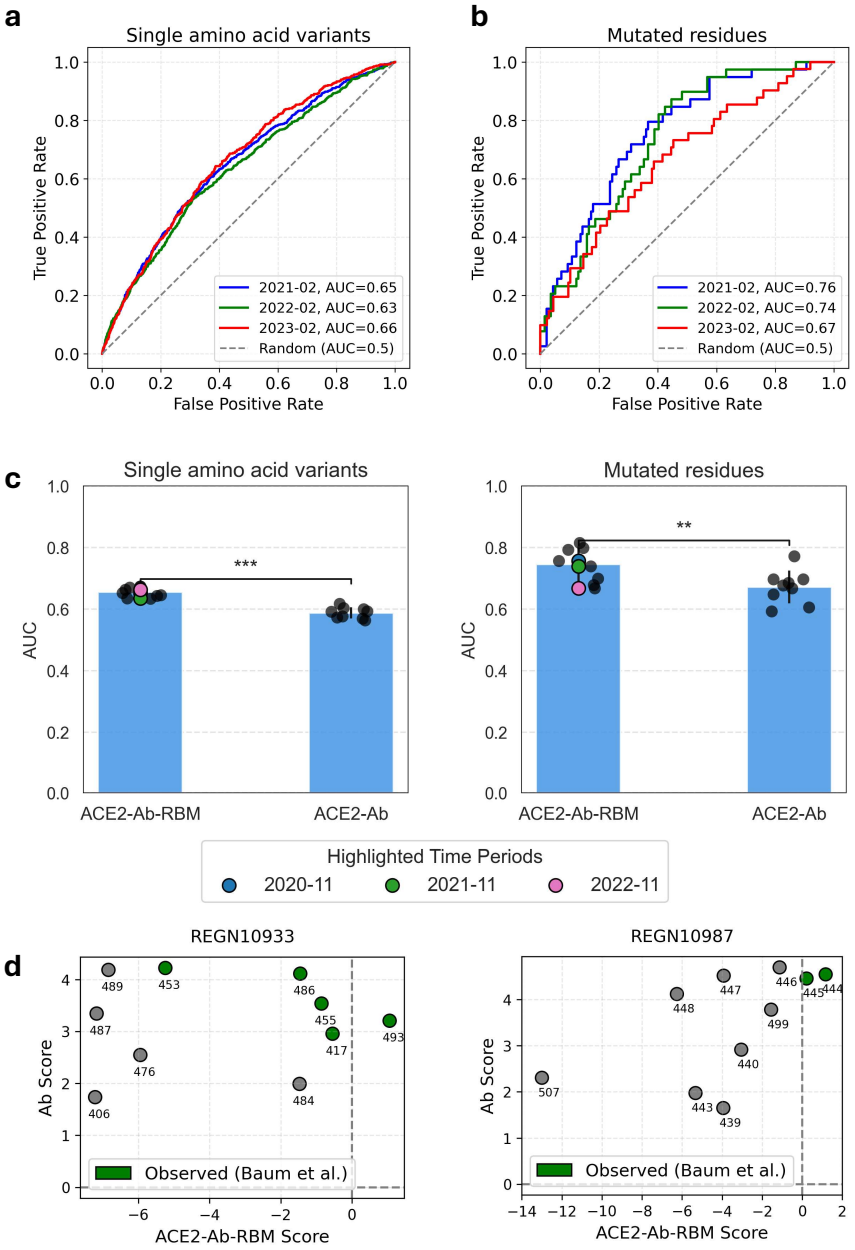

Fig. 2. EscapeMap predicts future mutations and site-level tolerance under immune pressure. (a) AUC of EscapeMap prediction for the appearance of mutations at different times. (b) AUC of EscapeMap prediction for the site-specific tolerance to a high number of unique mutations at different times. (c) Predictive performance (AUC) of EscapeMap, which incorporates ACE2–Ab–RBM probabilities, compared to a baseline model that includes only ACE2–Ab terms and omits RBM-based viability, 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 site scores from EscapeMap (ACE2–Ab–RBM) versus antibody-only scores (Ab), for the top 10 escape sites ranked by DMS-derived antibody escape. The Ab model reflects escape potential without viability or ACE2 constraints, where a positive score indicates that mutations at the site strongly promote antibody escape. In contrast, a positive score in EscapeMap (ACE2–Ab–RBM) indicates that mutations at the site enable escape while maintaining viability and ACE2 binding. Observed escape sites from serial passaging (Baum et al.) are highlighted in green.

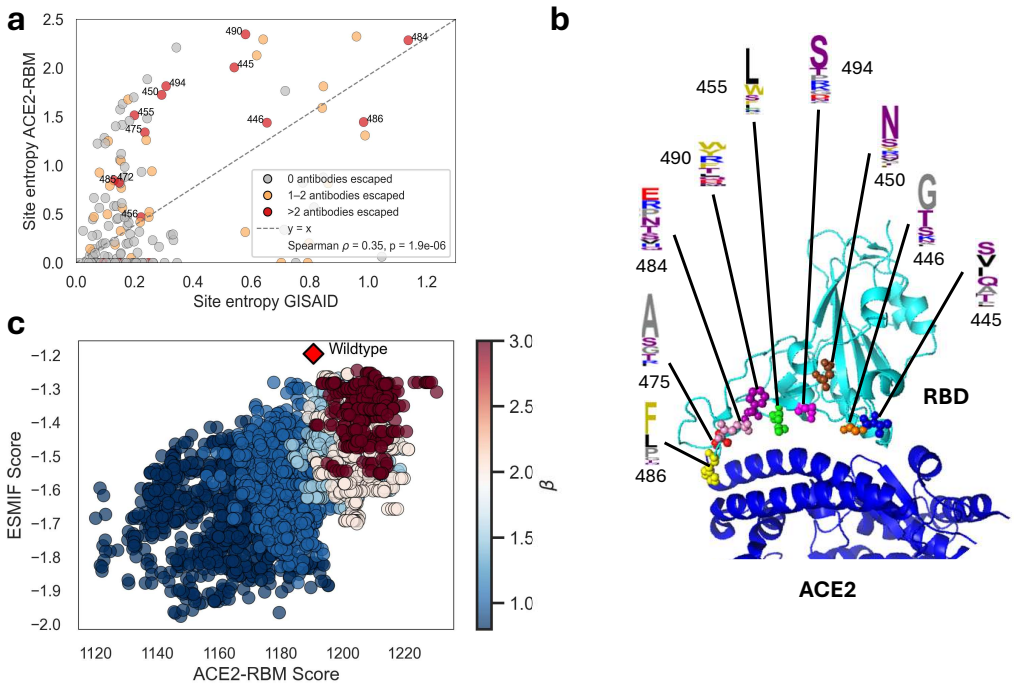

Fig. 3. Sequences generated by EscapeMap capture mutational constraints. (a) Comparison of site entropies of unique sequences generated using EscapeMap without antibody pressure (ACE2–RBM) and sequences found in GISAID. Colors indicate how many antibodies can be escaped (site max escape¿0.9 in DMS) when mutating that site. (b) Structural representation of the RBD (light blue) in complex with ACE2 (dark blue) (PDB: 6M0J). Colored spheres indicate sites involved in escape to at least 3 antibodies, and with entropy above 1. Amino acids show representative mutations at each site. (c) Scatter plot of ESM Inverse Folding (ESMIF) score versus EscapeMap score (ACE2–RBM) for generated variants at inverse temperatures $\beta \in \{1, 1.5, 2, 3\}$ without antibody pressure.

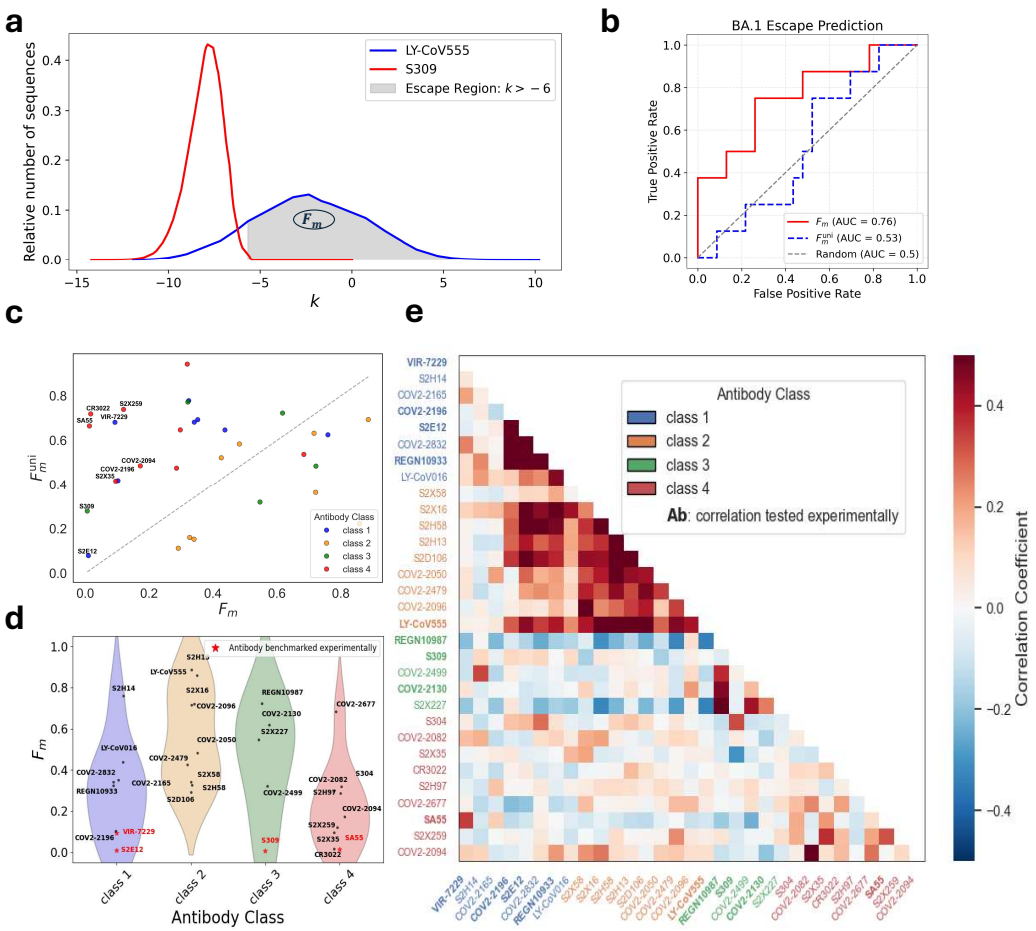

Fig. 4. Fraction of escaping sequences unveils antibody vulnerability. (a) Fraction F_m of sequences escaping antibody S309 and LY-CoV555. F_m is the integral of the escape region $k > -6$. (b) AUC for predicting antibody escape of BA.1 variants (Cao et al., 15 antibodies), using either F_m computed by EscapeMap, which weights sequences by ACE2 binding and viability, or F_m^{uni} baseline, computed under uniform sampling over sequence space. (c) Scatter plots comparing F_m and F_m^{uni} baseline across antibodies. Lower values correspond to a lower estimated number of escaping sequences. (d) Computed F_m values for each antibody, grouped by antibody class. Antibodies with the lowest values—SA55, VIR-7229, S2E12, and S309—are underlined and selected for in vitro testing under high antibody pressure. (e) Model correlations between antibody dissociation constants. Red indicates positive correlations; blue indicates negative correlations. Experimentally validated correlations are highlighted.

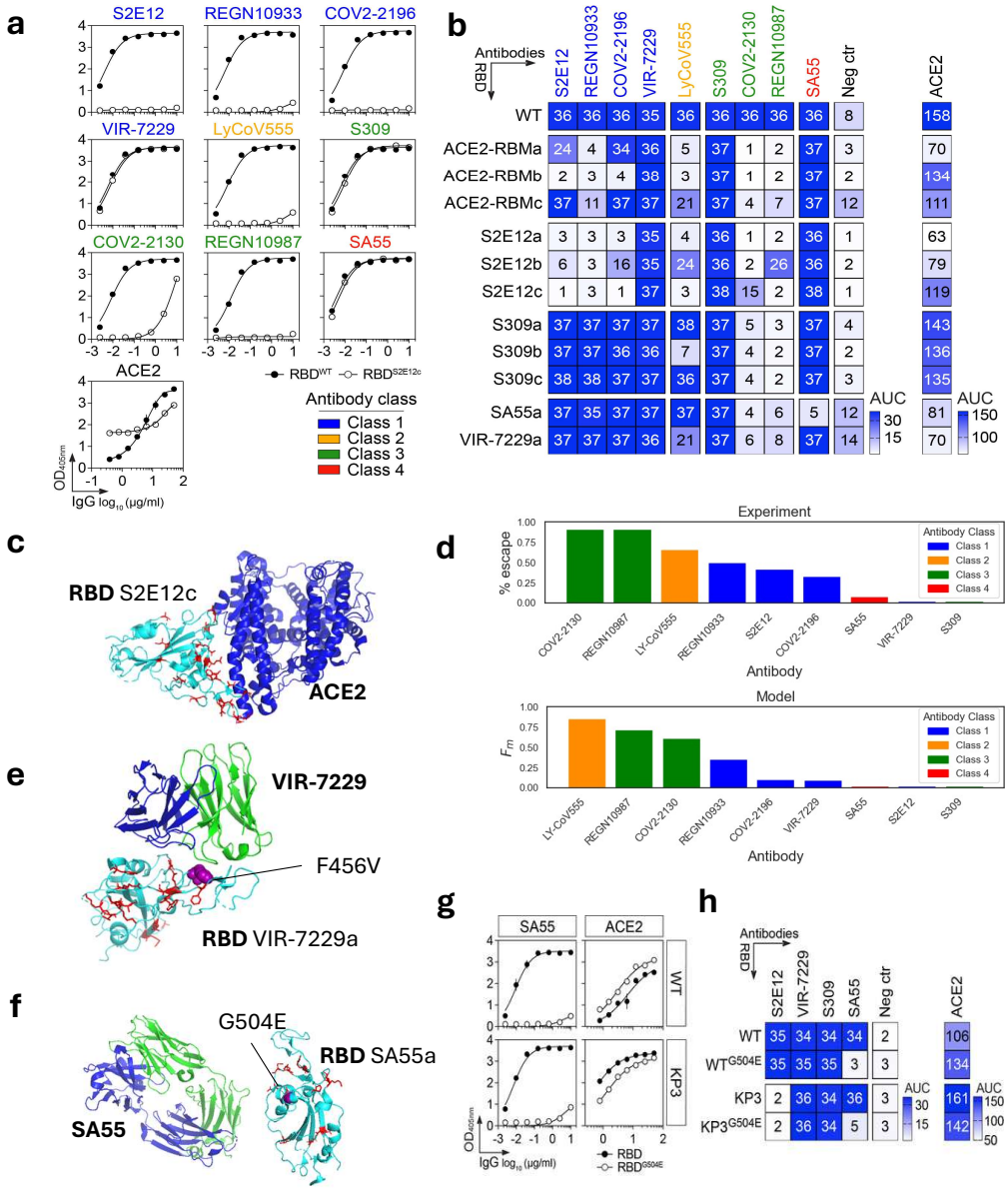

Fig. 5. Sequence design under immune pressure reveals antibody escape routes in vitro. (a) Representative ELISA graphs showing the binding of anti-RBD antibodies and ACE2 to purified wildtype and mutated RBD S2E12c, generated with S2E12 immune pressure. (b) Heatmap of ELISA binding of anti-RBD antibodies 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.

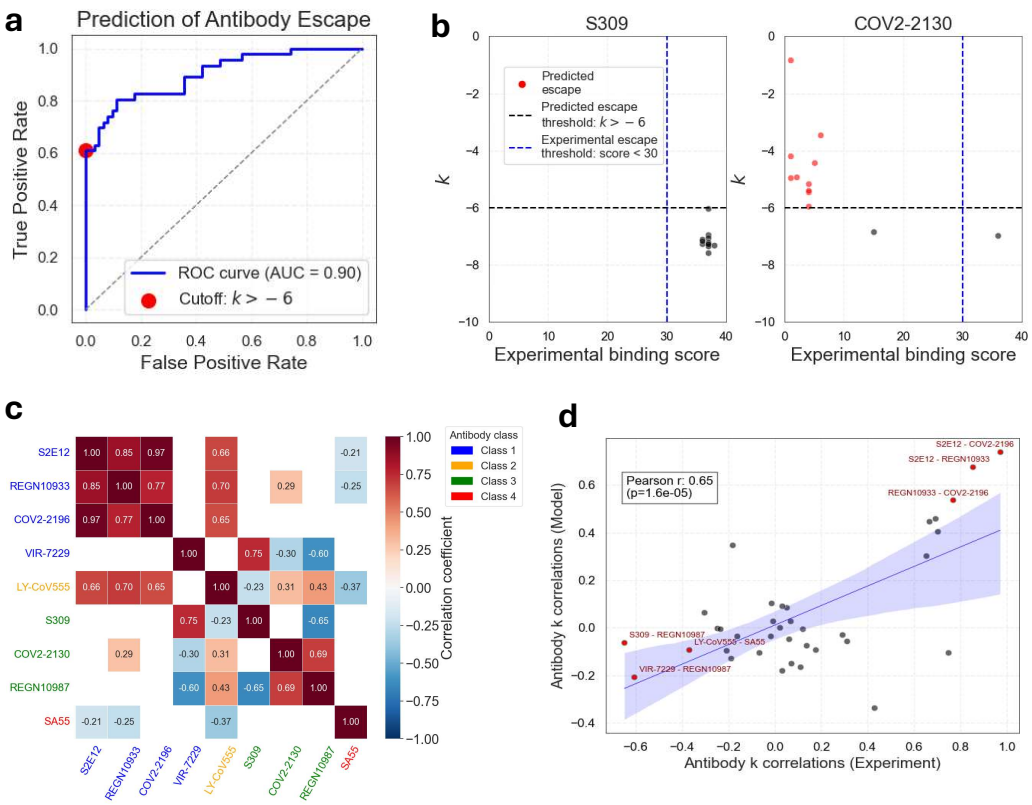

Fig. 6. EscapeMap predicts observed antibody escape profiles and binding correlations. (a) ROC curve for model prediction of antibody escape (binding score < 30). (b) Comparison of measured and predicted binding scores for antibodies S309 and COV2-2130. Each point corresponds to a functional sequence for which antibody dissociation constants were measured. (c) Correlation of antibody binding scores in experiment. Small correlations between -0.2 and 0.2 are masked. (d) Correlations between antibody dissociation constants for cocktail design, by experiment vs. model. Pairs with highest/lowest experimental correlations are highlighted.

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

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

References

1. Carabelli, A.M., Peacock, T.P., Thorne, L.G., Harvey, W.T., Hughes, J., COVID-19 Genomics UK Consortium, De Silva, T.I., Peacock, S.J., Barclay, W.S., De Silva, T.I., Towers, G.J., and Robertson, D.L. (2023). SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nature Reviews Microbiology. URL: https://www.nature.com/articles/s41579-022-00841-7. doi: 10.1038/s41579-022-00841-7.

2. Jackson, C.B., Farzan, M., Chen, B., and Choe, H. (2022). Mechanisms of SARS-CoV-2 entry into cells. Nature Reviews Molecular Cell Biology *23*, 3–20. URL: https://www.nature.com/articles/s41580-021-00418-x. doi: 10.1038/s41580-021-00418-x.

3. Tuekprakhon, A., Nutalai, R., Dijokaite-Guraliuc, A., Zhou, D., Ginn, H.M., Selvaraj, M., Liu, C., Mentzer, A.J., Supasa, P., Duyvesteyn, H.M., Das, R., Skelly, D., Ritter, T.G., Amini, A., Bibi, S., Adele, S., Johnson, S.A., Constantinides, B., Webster, H., Temperton, N., Klenerman, P., Barnes, E., Dunachie, S.J., Crook, D., Pollard, A.J., Lambe, T., Goulder, P., Paterson, N.G., Williams, M.A., Hall, D.R., Fry, E.E., Huo, J., Mongkolsapaya, J., Ren, J., Stuart, D.I., Screaton, G.R., Conlon, C., Deeks, A., Frater, J., Frending, L., Gardiner, S., Jämsén, A., Jeffery, K., Malone, T., Phillips, E., Rothwell, L., and Stafford, L. (2022). Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. Cell *185*, 2422–2433.e13. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867422007103. doi: 10.1016/j.cell.2022.06.005.

4. Rodriguez-Rivas, J., Croce, G., Muscat, M., and Weigt, M. (2022). Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proceedings of the National Academy of Sciences *119*, e2113118119. URL: https://pnas.org/doi/full/10.1073/pnas.2113118119. doi: 10.1073/pnas.2113118119.

5. Hie, B., Zhong, E.D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. Science *371*, 284–288.

6. Ozono, S., Zhang, Y., Ode, H., Sano, K., Tan, T.S., Imai, K., Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y., Suzuki, T., and Tokunaga, K. (2021). SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. Nature Communications *12*, 848. URL: https://www.nature.com/articles/s41467-021-21118-2. doi: 10.1038/s41467-021-21118-2.

7. Barton, M.I., MacGowan, S.A., Kutuzov, M.A., Dushek, O., Barton, G.J., and Van Der Merwe, P.A. (2021). 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. URL: https://elifesciences.org/articles/70658. doi: 10.7554/eLife.70658.

8. Nabel, K.G., Clark, S.A., Shankar, S., Pan, J., Clark, L.E., Yang, P., Coscia, A., McKay, L.G.A., Varnum, H.H., Brusic, V., Tolan, N.V., Zhou, G., Desjardins, M., Turbett, S.E., Kanjilal, S., Sherman, A.C., Dighe, A., LaRocque, R.C., Ryan, E.T., Tylek, C., Cohen-Solal, J.F., Darcy, A.T., Tavella, D., Clabbers, A., Fan, Y., Griffiths, A., Correia, I.R., Seagal, J., Baden, L.R., Charles, R.C., and Abraham, J. (2022). Structural basis for continued antibody evasion by the SARS-CoV-2 receptor binding domain. Science *375*, eabl6251. URL: https://www.science.org/doi/10.1126/science.abl6251. doi: 10.1126/science.abl6251.

9. Wang, Q., Iketani, S., Li, Z., Liu, L., Guo, Y., Huang, Y., Bowen, A.D., Liu, M., Wang, M., Yu, J., Valdez, R., Lauring, A.S., Sheng, Z., Wang, H.H., Gordon, A., Liu, L., and

Ho, D.D. (2023). Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants. Cell *186*, 279–286.e8. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867422015318. doi: 10.1016/j.cell.2022.12.018.

10. Raharinirina, N.A., Gubela, N., Börnigen, D., Smith, M.R., Oh, D.Y., Budt, M., Mache, C., Schillings, C., Fuchs, S., Dürrwald, R., Wolff, T., Hölzer, M., Paraskevopoulou, S., and Von Kleist, M. (2025). SARS-CoV-2 evolution on a dynamic immune landscape. Nature *639*, 196–204. URL: https://www.nature.com/articles/s41586-024-08477-8. doi: 10.1038/s41586-024-08477-8.

11. Meijers, M., Ruchnewitz, D., Eberhardt, J., Łuksza, M., and Lässig, M. (2023). Population immunity predicts evolutionary trajectories of SARS-CoV-2. Cell *186*, 5151–5164.e13. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867423010760. doi: 10.1016/j.cell.2023.09.022.

12. Thadani, N.N., Gurev, S., Notin, P., Youssef, N., Rollins, N.J., Ritter, D., Sander, C., Gal, Y., and Marks, D.S. (2023). Learning from prepandemic data to forecast viral escape. Nature *622*, 818–825. URL: https://www.nature.com/articles/s41586-023-06617-0. doi: 10.1038/s41586-023-06617-0.

13. Wang, G., Liu, X., Wang, K., Gao, Y., Li, G., Baptista-Hon, D.T., Yang, X.H., Xue, K., Tai, W.H., Jiang, Z., Cheng, L., Fok, M., Lau, J.Y.N., Yang, S., Lu, L., Zhang, P., and Zhang, K. (2023). Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. Nature Medicine *29*, 2007–2018. URL: https://www.nature.com/articles/s41591-023-02483-5. doi: 10.1038/s41591-023-02483-5.

14. Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H., Dingens, A.S., Nargi, R.S., Sutton, R.E., Suryadevara, N., Rothlauf, P.W., Liu, Z., Whelan, S.P., Carnahan, R.H., Crowe, J.E., and Bloom, J.D. (2021). Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. Cell Host & Microbe *29*, 44–57.e9. URL: https://linkinghub.elsevier.com/retrieve/pii/S1931312820306247. doi: 10.1016/j.chom.2020.11.007.

15. Greaney, A.J., Starr, T.N., Barnes, C.O., Weisblum, Y., Schmidt, F., Caskey, M., Gaebler, C., Cho, A., Agudelo, M., Finkin, S., Wang, Z., Poston, D., Muecksch, F., Hatziioannou, T., Bieniasz, P.D., Robbiani, D.F., Nussenzweig, M.C., Bjorkman, P.J., and Bloom, J.D. (2021). Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. Nature Communications *12*, 4196. URL: https://www.nature.com/articles/s41467-021-24435-8. doi: 10.1038/s41467-021-24435-8.

16. Greaney, A.J., Loes, A.N., Crawford, K.H., Starr, T.N., Malone, K.D., Chu, H.Y., and Bloom, J.D. (2021). 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. URL: https://linkinghub.elsevier.com/retrieve/pii/S1931312821000822. doi: 10.1016/j.chom.2021.02.003.

17. Tortorici, M.A., Czudnochowski, N., Starr, T.N., Marzi, R., Walls, A.C., Zatta, F., Bowen, J.E., Jaconi, S., Di Iulio, J., Wang, Z., De Marco, A., Zepeda, S.K., Pinto, D., Liu, Z., Beltramello, M., Bartha, I., Housley, M.P., Lempp, F.A., Rosen, L.E., Dellota, E., Kaiser, H., Montiel-Ruiz, M., Zhou, J., Addetia, A., Guarino, B., Culap, K., Sprugasci, N., Saliba, C., Vetti, E., Giacchetto-Sasselli, I., Fregni, C.S., Abdelnabi, R., Foo, S.Y.C., Havenar-Daughton, C.,

Schmid, M.A., Benigni, F., Cameroni, E., Neyts, J., Telenti, A., Virgin, H.W., Whelan, S.P.J., Snell, G., Bloom, J.D., Corti, D., Veesler, D., and Pizzuto, M.S. (2021). Broad sarbecovirus neutralization by a human monoclonal antibody. Nature *597*, 103–108. URL: https://www.nature.com/articles/s41586-021-03817-4. doi: 10.1038/s41586-021-03817-4.

18. Starr, T.N., Greaney, A.J., Dingens, A.S., and Bloom, J.D. (2021). Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell Reports Medicine *2*, 100255. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666379121000719. doi: 10.1016/j.xcrm.2021.100255.

19. Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z., and Bloom, J.D. (2021). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. Science *371*, 850–854. URL: https://www.science.org/doi/10.1126/science.abf9302. doi: 10.1126/science.abf9302.

20. Starr, T.N., Czudnochowski, N., Liu, Z., Zatta, F., Park, Y.J., Addetia, A., Pinto, D., Beltramello, M., Hernandez, P., Greaney, A.J., Marzi, R., Glass, W.G., Zhang, I., Dingens, A.S., Bowen, J.E., Tortorici, M.A., Walls, A.C., Wojcechowskyj, J.A., De Marco, A., Rosen, L.E., Zhou, J., Montiel-Ruiz, M., Kaiser, H., Dillen, J.R., Tucker, H., Bassi, J., Silacci-Fregni, C., Housley, M.P., Di Iulio, J., Lombardo, G., Agostini, M., Sprugasci, N., Culap, K., Jaconi, S., Meury, M., Dellota Jr, E., Abdelnabi, R., Foo, S.Y.C., Cameroni, E., Stumpf, S., Croll, T.I., Nix, J.C., Havenar-Daughton, C., Piccoli, L., Benigni, F., Neyts, J., Telenti, A., Lempp, F.A., Pizzuto, M.S., Chodera, J.D., Hebner, C.M., Virgin, H.W., Whelan, S.P.J., Veesler, D., Corti, D., Bloom, J.D., and Snell, G. (2021). SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. Nature *597*, 97–102. URL: https://www.nature.com/articles/s41586-021-03807-6. doi: 10.1038/s41586-021-03807-6.

21. Cao, Y., Jian, F., Wang, J., Yu, Y., Song, W., Yisimayi, A., Wang, J., An, R., Chen, X., Zhang, N., Wang, Y., Wang, P., Zhao, L., Sun, H., Yu, L., Yang, S., Niu, X., Xiao, T., Gu, Q., Shao, F., Hao, X., Xu, Y., Jin, R., Shen, Z., Wang, Y., and Xie, X.S. (2022). Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. Nature. URL: https://www.nature.com/articles/s41586-022-05644-7. doi: 10.1038/s41586-022-05644-7.

22. Yisimayi, A., Song, W., Wang, J., Jian, F., Yu, Y., Chen, X., Xu, Y., Yang, S., Niu, X., Xiao, T., Wang, J., Zhao, L., Sun, H., An, R., Zhang, N., Wang, Y., Wang, P., Yu, L., Lv, Z., Gu, Q., Shao, F., Jin, R., Shen, Z., Xie, X.S., Wang, Y., and Cao, Y. (2024). Repeated Omicron exposures override ancestral SARS-CoV-2 immune imprinting. Nature *625*, 148–156. URL: https://www.nature.com/articles/s41586-023-06753-7. doi: 10.1038/s41586-023-06753-7.

23. Ito, J., Strange, A., Liu, W., Joas, G., Lytras, S., The Genotype to Phenotype Japan (G2P-Japan) Consortium, and Sato, K. (2024). A Protein Language Model for Exploring Viral Fitness Landscapes. . URL: http://biorxiv.org/lookup/doi/10.1101/2024.03.15.584819. doi: 10.1101/2024.03.15.584819.

24. Wang, D., Huot, M., Mohanty, V., and Shakhnovich, E.I. (2024). Biophysical principles predict fitness of sars-cov-2 variants. Proceedings of the National Academy of Sciences *121*, e2314518121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2314518121. doi: 10.1073/pnas.2314518121. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2314518121.

25. Maher, M.C., Bartha, I., Weaver, S., Di Iulio, J., Ferri, E., Soriaga, L., Lempp, F.A., Hie, B.L., Bryson, B., Berger, B., Robertson, D.L., Snell, G., Corti, D., Virgin, H.W., Kosakovsky Pond,

S.L., and Telenti, A. (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Science Translational Medicine *14*, eabk3445. URL: https://www.science.org/doi/10.1126/scitranslmed.abk3445. doi: 10.1126/scitranslmed.abk3445.

26. Huot, M., Wang, D., Liu, J., and Shakhnovich, E. (2025). Few-shot viral variant detection via bayesian active learning and biophysics. bioRxiv. URL: https://www.biorxiv.org/content/early/2025/03/13/2025.03.12.642881. doi: 10.1101/2025.03.12.642881. arXiv:https://www.biorxiv.org/content/early/2025/03/13/2025.03.12.642881.full.pdf.

27. Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., King, N.P., Veesler, D., and Bloom, J.D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell *182*, 1295–1310.e20. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867420310035. doi: 10.1016/j.cell.2020.08.012.

28. Tubiana, J., Cocco, S., and Monasson, R. (2019). Learning protein constitutive motifs from sequence data. eLife *8*, e39397. URL: https://elifesciences.org/articles/39397. doi: 10.7554/eLife.39397.

29. Andreano, E., Paciello, I., Pierleoni, G., Piccini, G., Abbiento, V., Antonelli, G., Pileri, P., Manganaro, N., Pantano, E., Maccari, G., Marchese, S., Donnici, L., Benincasa, L., Giglioli, G., Leonardi, M., De Santi, C., Fabbiani, M., Rancan, I., Tumbarello, M., Montagnani, F., Sala, C., Medini, D., De Francesco, R., Montomoli, E., and Rappuoli, R. (2023). B cell analyses after SARS-CoV-2 mRNA third vaccination reveals a hybrid immunity like antibody response. Nature Communications *14*, 53. URL: https://www.nature.com/articles/s41467-022-35781-6. doi: 10.1038/s41467-022-35781-6.

30. Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., Wei, Y., Atwal, G.S., Murphy, A.J., Stahl, N., Yancopoulos, G.D., and Kyratsous, C.A. (2020). Antibody cocktail to sars-cov-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science *369*, 1014–1018. URL: https://www.science.org/doi/abs/10.1126/science.abd0831. doi: 10.1126/science.abd0831. arXiv:https://www.science.org/doi/pdf/10.1126/science.abd0831.

31. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. (2022). Learning inverse folding from millions of predicted structures. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds. Proceedings of the 39th International Conference on Machine Learning vol. 162 of *Proceedings of Machine Learning Research*. PMLR pp. 8946–8970. URL: https://proceedings.mlr.press/v162/hsu22a.html.

32. Cao, Y., Yisimayi, A., Jian, F., Song, W., Xiao, T., Wang, L., Du, S., Wang, J., Li, Q., Chen, X., Yu, Y., Wang, P., Zhang, Z., Liu, P., An, R., Hao, X., Wang, Y., Wang, J., Feng, R., Sun, H., Zhao, L., Zhang, W., Zhao, D., Zheng, J., Yu, L., Li, C., Zhang, N., Wang, R., Niu, X., Yang, S., Song, X., Chai, Y., Hu, Y., Shi, Y., Zheng, L., Li, Z., Gu, Q., Shao, F., Huang, W., Jin, R., Shen, Z., Wang, Y., Wang, X., Xiao, J., and Xie, X.S. (2022). BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. Nature *608*, 593–602. URL: https://www.nature.com/articles/s41586-022-04980-y. doi: 10.1038/s41586-022-04980-y.

33. Rosen, L.E., Tortorici, M.A., De Marco, A., Pinto, D., Foreman, W.B., Taylor, A.L., Park, Y.J., Bohan, D., Rietz, T., Errico, J.M., Hauser, K., Dang, H.V., Chartron, J.W., Giurdanella, M., Cusumano, G., Saliba, C., Zatta, F., Sprouse, K.R., Addetia, A., Zepeda,

S.K., Brown, J., Lee, J., Dellota, E., Rajesh, A., Noack, J., Tao, Q., DaCosta, Y., Tsu, B., Acosta, R., Subramanian, S., De Melo, G.D., Kergoat, L., Zhang, I., Liu, Z., Guarino, B., Schmid, M.A., Schnell, G., Miller, J.L., Lempp, F.A., Czudnochowski, N., Cameroni, E., Whelan, S.P., Bourhy, H., Purcell, L.A., Benigni, F., Di Iulio, J., Pizzuto, M.S., Lanzavecchia, A., Telenti, A., Snell, G., Corti, D., Veesler, D., and Starr, T.N. (2024). A potent pan-sarbecovirus neutralizing antibody resilient to epitope diversification. Cell *187*, 7196–7213.e26. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867424010845. doi: 10.1016/j.cell.2024.09.026.

34. Cao, Y., Jian, F., Zhang, Z., Yisimayi, A., Hao, X., Bao, L., Yuan, F., Yu, Y., Du, S., Wang, J., Xiao, T., Song, W., Zhang, Y., Liu, P., An, R., Wang, P., Wang, Y., Yang, S., Niu, X., Zhang, Y., Gu, Q., Shao, F., Hu, Y., Yin, W., Zheng, A., Wang, Y., Qin, C., Jin, R., Xiao, J., and Xie, X.S. (2022). Rational identification of potent and broad sarbecovirus-neutralizing antibody cocktails from SARS convalescents. Cell Reports *41*, 111845. URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124722017375. doi: 10.1016/j.celrep.2022.111845.

35. Ku, Z., Xie, X., Davidson, E., Ye, X., Su, H., Menachery, V.D., Li, Y., Yuan, Z., Zhang, X., Muruato, A.E., I Escuer, A.G., Tyrell, B., Doolan, K., Doranz, B.J., Wrapp, D., Bates, P.F., McLellan, J.S., Weiss, S.R., Zhang, N., Shi, P.Y., and An, Z. (2021). Molecular determinants and mechanism for antibody cocktail preventing SARS-CoV-2 escape. Nature Communications *12*, 469. URL: https://www.nature.com/articles/s41467-020-20789-7. doi: 10.1038/s41467-020-20789-7.

36. Yu, T.C., Thornton, Z.T., Hannon, W.W., DeWitt, W.S., Radford, C.E., Matsen, F.A., and Bloom, J.D. (2022). A biophysical model of viral escape from polyclonal antibodies. Virus Evolution *8*, veac110. URL: https://academic.oup.com/ve/article/doi/10.1093/ve/veac110/6889254. doi: 10.1093/ve/veac110.

37. Markov, P.V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N.I., and Katzourakis, A. (2023). The evolution of SARS-CoV-2. Nature Reviews Microbiology *21*, 361–379. URL: https://www.nature.com/articles/s41579-023-00878-2. doi: 10.1038/s41579-023-00878-2.

38. Planchais, C., Fernández, I., Chalopin, B., Bruel, T., Rosenbaum, P., Beretta, M., Dimitrov, J.D., Conquet, L., Donati, F., Prot, M., Porrot, F., Planas, D., Staropoli, I., Guivel-Benhassine, F., Baquero, E., Van Der Werf, S., Haouz, A., Simon-Lorière, E., Montagutelli, X., Maillère, B., Rey, F.A., Guardado-Calvo, P., Nozach, H., Schwartz, O., and Mouquet, H. (2024). Broad sarbecovirus neutralization by combined memory B cell antibodies to ancestral SARS-CoV-2. iScience *27*, 110354. URL: https://linkinghub.elsevier.com/retrieve/pii/S2589004224015797. doi: 10.1016/j.isci.2024.110354.

39. Rochman, N.D., Wolf, Y.I., Faure, G., Mutz, P., Zhang, F., and Koonin, E.V. (2021). Ongoing global and regional adaptive evolution of sars-cov-2. Proceedings of the National Academy of Sciences *118*, e2104241118.

40. Rochman, N.D., Faure, G., Wolf, Y.I., Freddolino, L., Zhang, F., and Koonin, E.V. (2022). Epistasis at the sars-cov-2 receptor-binding domain interface and the propitiously boring implications for vaccine escape. MBio *13*, e00135–22.

41. Cohen, A.A., van Doremalen, N., Greaney, A.J., Andersen, H., Sharma, A., Starr, T.N., Keeffe, J.R., Fan, C., Schulz, J.E., Gnanapragasam, P.N. et al. (2022). Mosaic rbd nanoparticles protect against challenge by diverse sarbecoviruses in animal models. Science *377*, eabq0839.

42. Gonzalez Somermeyer, L., Fleiss, A., Mishin, A.S., Bozhanova, N.G., Igolkina, A.A., Meiler, J., Alaball Pujol, M.E., Putintseva, E.V., Sarkisyan, K.S., and Kondrashov, F.A. (2022). Heterogeneity of the gfp fitness landscape and data-driven protein design. eLife *11*, e75842. URL: https://doi.org/10.7554/eLife.75842. doi: 10.7554/eLife.75842.

43. Youssef, N., Gurev, S., Ghantous, F., Brock, K.P., Jaimes, J.A., Thadani, N.N., Dauphin, A., Sherman, A.C., Yurkovetskiy, L., Soto, D., Estanboulieh, R., Kotzen, B., Notin, P., Kollasch, A.W., Cohen, A.A., Dross, S.E., Erasmus, J., Fuller, D.H., Bjorkman, P.J., Lemieux, J.E., Luban, J., Seaman, M.S., and Marks, D. (2023). Protein design for evaluating vaccines against future viral variation. . URL: http://biorxiv.org/lookup/doi/10.1101/2023.10.08.561389. doi: 10.1101/2023.10.08.561389.

44. Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. Global Challenges *1*, 33–46. URL: https://onlinelibrary.wiley.com/doi/10.1002/gch2.1018. doi: 10.1002/gch2.1018.

45. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mst010. doi: 10.1093/molbev/mst010.

46. Zost, S.J., Gilchuk, P., Case, J.B., Binshtein, E., Chen, R.E., Nkolola, J.P., Schäfer, A., Reidy, J.X., Trivette, A., Nargi, R.S., Sutton, R.E., Suryadevara, N., Martinez, D.R., Williamson, L.E., Chen, E.C., Jones, T., Day, S., Myers, L., Hassan, A.O., Kafai, N.M., Winkler, E.S., Fox, J.M., Shrihari, S., Mueller, B.K., Meiler, J., Chandrashekar, A., Mercado, N.B., Steinhardt, J.J., Ren, K., Loo, Y.M., Kallewaard, N.L., McCune, B.T., Keeler, S.P., Holtzman, M.J., Barouch, D.H., Gralinski, L.E., Baric, R.S., Thackray, L.B., Diamond, M.S., Carnahan, R.H., and Crowe, J.E. (2020). Potently neutralizing and protective human antibodies against SARS-CoV-2. Nature *584*, 443–449. URL: https://www.nature.com/articles/s41586-020-2548-6. doi: 10.1038/s41586-020-2548-6.

47. Wang, P., Nair, M.S., Liu, L., Iketani, S., Luo, Y., Guo, Y., Wang, M., Yu, J., Zhang, B., Kwong, P.D., Graham, B.S., Mascola, J.R., Chang, J.Y., Yin, M.T., Sobieszczyk, M., Kyratsous, C.A., Shapiro, L., Sheng, Z., Huang, Y., and Ho, D.D. (2021). Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature *593*, 130–135. URL: https://www.nature.com/articles/s41586-021-03398-2. doi: 10.1038/s41586-021-03398-2.

48. Moulana, A., Dupic, T., Phillips, A.M., Chang, J., Roffler, A.A., Greaney, A.J., Starr, T.N., Bloom, J.D., and Desai, M.M. (2023). The landscape of antibody binding affinity in SARS-CoV-2 omicron BA.1 evolution. Elife *12*.

49. Tieleman, T. (2008). Training restricted boltzmann machines using approximations to the likelihood gradient. In Proceedings of the 25th International Conference on Machine Learning. ICML '08. New York, NY, USA: Association for Computing Machinery. ISBN 9781605582054 pp. 1064–1071. URL: https://doi.org/10.1145/1390156.1390290. doi: 10.1145/1390156.1390290.

STAR 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[44] 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[45] 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 $v = (v_1, v_2, \ldots, v_N)$ are displayed on the Visible layer, and representations $h = (h_1, h_2, \ldots, h_M)$ 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 v, h is:

$$P_{\text{RBM}}(v, h) \propto \exp\left(\sum_{i=1}^{N} g_i\left(v_i\right) - \sum_{\mu=1}^{M} \mathcal{U}_\mu\left(h_\mu\right) + \sum_{i,\mu} h_\mu w_{i\mu}\left(v_i\right)\right) \tag{7}$$

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

$$P_{\text{RBM}}(v) \propto \int_h dh\, P_{\text{RBM}}(v, h) = \exp\left(\sum_{i=1}^{N} g_i\left(v_i\right) + \sum_{\mu=1}^{M} \Gamma_\mu\left(I_\mu(v)\right)\right), \tag{8}$$

over the sequences v within a MSA of the protein family. The cumulant generating function $\Gamma_\mu(I) = \ln \int \mathrm{d}h_\mu \, e^{I_\mu h_\mu - \mathcal{U}_\mu(h_\mu)}$, where $I_\mu = \sum_i w_{i\mu}(v_i)$, 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,

$$P_{\mathrm{ACE2}}(v) = \frac{C_{\mathrm{ACE2}}\, e^{-G_{\mathrm{ACE2}}(v)}}{C_{\mathrm{ACE2}}\, e^{-G_{\mathrm{ACE2}}(v)} + e^{-G_{\mathrm{free}}}} = \frac{1}{1 + 10^{k_{\mathrm{ACE2}} - c_{\mathrm{ACE2}}}} \ , \tag{9}$$

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

where $G_{\mathrm{free}}, G_{\mathrm{ACE2}}$ 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. (8) with the binding probabilities in Eqs. (9) and (10) defines the full probability distribution over the sequences v, with the result

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

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

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

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

$$P_{Gen}(v) = \frac{\exp(-\beta\, E(v|\theta))}{Z(\theta)} \ , \tag{12}$$

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

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

Experimental dissociation constants

IC_{50} values for the 29 antibodies against the wild-type RBD were taken from[20] for S304, S2H14, S2H97, S2H13, S2X35, S309, S2X58, S2X227, S2X16, S2H58, S2D106 and S2E12;[46] for COV2-2677, COV2-2165, COV2-2082, COV2-2094, COV2-2130, COV2-2050, COV2-2832, COV2-2096, COV2-2479, COV2-2499 and COV2-2196;[19] for REGN10987, REGN10933 and LY-CoV016;[47] for LY-CoV555;[17] for S2X259; and[14] for CR3022.

IC_{50} values against BA.1 for S2H14, COV2-2196, S2E12, LY-CoV016, S2H13, LY-CoV555, COV2-2479, COV2-2499, COV2-2130, S309, S304, COV2-2677, S2H97, S2X259 and S2X35 were obtained from[32].

Wild-type IC_{50} values and deep-mutational-scanning data for SA55 and VIR-7229 were taken from[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:

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

By combining these baseline values $k(\text{wt})$ 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 logarithm of the dissociation constant $k = \log_{10} K$ is proportional to the free energy of binding, we model $\log_{10} K$ using an additive model in terms of site-specific contributions Δk_i. Specifically, for a sequence $v = (v_1, \ldots, v_N)$, we write:

$$k(v_1, \ldots, v_N) := \Delta k_0 + \sum_{i=1}^{N} \Delta k_i(v_i) \ . \tag{13}$$

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.[48] 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 $\langle \log P(s) \rangle_{\text{dataset}}$ 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

$$\frac{\partial \text{LL}}{\partial \theta} = -\left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{dataset}} + \left\langle \frac{\partial E}{\partial \theta} \right\rangle_{\text{model}} \tag{14}$$

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[49]. 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

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

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

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

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. 11). 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. 11), 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 ($\beta_{\text{RBM}} \in \{0.1, 0.2\}$) to relax evolutionary penalties. Conversely, to test reduced ACE2 affinity, we used a lower ACE2 concentration ($\log_{10} C_{\text{ACE2}} \in \{-6, -4\}$), diminishing the selection for tight receptor binding.

Site entropy

To quantify the mutational variability at each site i of the RBD, we compute the Shannon entropy:

$$S_i = -\sum_{v_i} P_i(v_i) \log P_i(v_i),$$

where $P_i(v_i)$ is the marginal probability of observing amino acid v_i at position i. High entropy indicates that the site tolerates diverse mutations, while low entropy reflects evolutionary constraints or functional importance.

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). A single point mutation, G504E, was introduced into the RBD-encoding constructs using the QuickChange Site-Directed Mutagenesis kit (Agilent Technologies) following the manufacturer's instructions. 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.