Complete Combinatorial Mutational Enumeration of a protein functional site enables sequence-landscape mapping and identifies highly-mutated variants that retain activity

Mireia Solà Colom
Institute for Protein Innovation, Boston, Massachusetts, 02115, USA; Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; AI Proteins, Boston, Massachusetts, 02215, USA

Jelena Vucinic
Université Fédérale de Toulouse, ANITI, IRIT-CNRS UMR 5505, Université Toulouse I Capitole, 31000 Toulouse, France

Jared Adolf-Bryfogle
Institute for Protein Innovation, Boston, Massachusetts, 02115, USA; Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA

James W. Bowman
Institute for Protein Innovation, Boston, Massachusetts, 02115, USA; Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; AI Proteins, Boston, Massachusetts, 02215, USA

Sébastien Verel
Université Littoral Côte d'Opale, UR 4491, LISIC, F-62100 Calais, France

Isabelle Moczygemba
Institute for Protein Innovation, Boston, Massachusetts, 02115, USA; Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; AI Proteins, Boston, Massachusetts, 02215, USA

Thomas Schiex
Université Fédérale de Toulouse, ANITI, INRAE-UR 875, 31000 Toulouse, France

David Simoncini (David.Simoncini@gmail.com)
Université Fédérale de Toulouse, ANITI, IRIT-CNRS UMR 5505, Université Toulouse I Capitole, 31000 Toulouse, France

Christopher D. Bahl (cdbahl@gmail.com)
Institute for Protein Innovation, Boston, Massachusetts, 02115, USA; Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, 02115, USA; AI Proteins, Boston, Massachusetts, 02215, USA https://orcid.org/0000-0002-3652-3693

Biological Sciences - Article

Keywords:

Posted Date: September 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2248327/v2

Additional Declarations: Yes there is potential Competing Interest. MSC, JTB, IM and CDB own stock in AI Proteins, Inc. CDB owns stock in Oncopep, Inc., and is a scientific advisor for Oncopep, Inc. and Applied Photophysics Ltd.

Complete Combinatorial Mutational Enumeration of a protein functional site enables sequence-landscape mapping and identifies highly-mutated variants that retain activity

Mireia Solà Colom[1,2,3], Jelena Vucinic[4], Jared Adolf-Bryfogle[1,2], James W. Bowman[1,2,3], Sébastien Verel[5], Isabelle Moczygemba[1,2,3], Thomas Schiex[6], David Simoncini[4]*, Christopher D. Bahl[1,2,3]*

Affiliations:
[1] Institute for Protein Innovation; Boston, Massachusetts, 02115, USA
[2] Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School; Boston, Massachusetts, 02115, USA
[3] current address: AI Proteins; Boston, Massachusetts, 02215, USA
[4] Université Fédérale de Toulouse; ANITI, IRIT-CNRS UMR 5505, Université Toulouse Capitole, 31000 Toulouse, France
[5] Université Littoral Côte d'Opale; UR 4491, LISIC, F-62100 Calais, France
[6] Université Fédérale de Toulouse; ANITI, INRAE-UR 875, 31000 Toulouse, France

* Corresponding authors:

Christopher D. Bahl
Email: chris@aiproteins.bio

David Simoncini
Email: david.simoncini@ut-capitole.fr

Abstract:

Understanding how proteins evolve under selective pressure is a longstanding challenge. The immensity of the search space has limited efforts to systematically evaluate the impact of multiple simultaneous mutations, so mutations have typically been assessed individually. However, epistasis, or the way in which mutations interact, prevents accurate prediction of combinatorial mutations based on measurements of individual mutations. Here, we use artificial intelligence to define the entire functional sequence landscape of a protein binding site *in silico*, and we call this approach Complete Combinatorial Mutational Enumeration (CCME). By leveraging CCME, we are able to construct a comprehensive map of the evolutionary connectivity within this functional sequence landscape. As a proof of concept, we applied CCME to the ACE2 binding site of the SARS-CoV-2 spike protein receptor binding domain. We selected representative variants from across the functional sequence landscape for testing in the laboratory. We identified variants that retained functionality to bind ACE2 despite changing over 40% of evaluated residue positions, and the variants now escape binding and neutralization by monoclonal antibodies. This work represents a crucial initial stride towards achieving precise predictions of pathogen evolution, opening avenues for proactive mitigation.

Main Text:

Protein evolution is a complex process that has shaped the diversity of life, and it is essential to understand because it impacts how our environment is likely to respond to climate change, how infectious diseases evolve, and how we can engineer proteins for industrial and therapeutic applications. One of the major challenges in studying protein evolution is the vastness of sequence space. The total number of possible amino acid sequences for a typical protein is astronomical [1], and it is currently impossible to explore all of this space experimentally or computationally. The functional sequence landscape of a protein can be defined as the set of all amino acid sequences that are able to carry out that protein's biological activity, and it is a substantially reduced search space when compared to the total possible sequence landscape. Therefore, the functional sequence landscape is more tractable to fully explore and is the focus of the work we report here.

To date, methods to study the functional sequence landscape of a protein have largely relied on assessing the impact of individual mutations on activity [2]. However, individual amino acid substitutions often interact non-linearly when combined. Thus, the effect of a mutation at one site in a protein depends on the sequence at other sites. Evaluating amino acid substitutions at a functional site must therefore be performed combinatorially to account for these effects, but this search space is still too vast for traditional computational or experimental approaches to fully enumerate. Here, we describe a novel computational approach utilizing automated reasoning artificial intelligence that we call Complete Combinatorial Mutational Enumeration (CCME) that is capable of enumerating a functional sequence landscape, and we use it to map the functional sequence landscape the protein which mediates human cell entry by the virus SARS-CoV-2.

To infect human cells, the SARS-CoV-2 spike (S) protein homotrimer must bind to its receptor, the angiotensin-converting enzyme 2 (ACE2) homodimer, on the host cell surface [3]. The receptor binding domain (RBD) on the S protein contains all amino acids which directly contact ACE2, and blocking the ACE2:RBD interaction prevents the virus from infecting cells [4–6]. Thus, the RBD is the target of most known neutralizing antibodies. We chose to focus our studies on the ACE2 binding site of the RBD from the L strain, which is the first strain of SARS-CoV-2, originally identified in December of 2019.

Enumeration of functional sequence space

CCME is performed using a 3D protein structure, a pairwise decomposable energy function, the cost function network prover toulbar2 [7,8] and a dedicated sequence enumeration algorithm (Fig. S1). To begin, we used Rosetta to identify the interacting residues at the ACE2:RBD interface from the first high-resolution structure of this complex [4] (Fig. 1). We then evaluated all combinatorial mutations at the 27 interface residue positions on RBD, and simultaneously, we allowed the 25 interface residues on ACE2 to explore alternative rotamers. This defines a search space of 1.3×10^{35} sequences and more than 5×10^{87} side-chain conformations (because we allow for all common rotamers at each residue position). This is greater than the number of atoms in the observable universe, so full enumeration of this search space is not possible using naive brute-force computation. Toulbar2 is able to find the best solution, prove its optimality and exhaustively enumerate all sequences with energy within a threshold of the optimum [9–11]. Performing sequence enumerations with toulbar2 has several advantages. Whereas sampling methods would run multiple individual trajectories and gather as many different sequences as possible without any knowledge on the size of the functional sequence landscape, toulbar2 systematically discards all unfit sequences in order to retain the exact ensemble of all functional sequences.

First, we assessed the binding energy between RBD and ACE2, which we approximated by the difference in kcal/mol between the bound and unbound RBD conformations, or ΔG. We used Pompd, a computational protein design program which uses toulbar2 for sequence enumeration [12], to compute an exhaustive list of all variant sequences capable of adopting an ACE2-bound conformation within 8 kcal/mol of the global energy minimum; this yielded over 91 million sequences (Fig. S2). Next, we evaluated the impact of these sequence changes on RBD stability. To preserve the competent-for-binding structure of the RBD, we tolerate only minimally destabilizing mutations, defined as a < 1 kcal/mole increase in energy. For binding, we required the $\Delta\Delta$G, the difference of ΔG between the L strain model and the variant model, to be positive. The intersection of these two ensembles resulted in about 4.5 million sequence variants. To reduce the size of this sequence space, we analyzed the fitness landscape defined by our 4.5 million variants, using $\Delta\Delta$G as the fitness function and connecting two variants when they differ only by one mutation [13]. In this landscape, we identified 3,272 locally optimal sequences, meaning that mutating to any of their neighbors wouldn't improve the interface $\Delta\Delta$G. Because we wanted to assess how distant from each other functional variants could be, we looked for the most diverse subset of local optimal sequences. We clustered these by sequence similarity with MMseqs2 (using maximum E-value of gap-corrected Karlin-Altschul statistics, minimum

coverage and sequence identity as criterions). We visually represented the spatial distribution of these clusters on a t-distributed stochastic neighbor embedding map (Fig. 2). We selected the medoid of each of the 59 clusters that we obtained for characterization in the wet laboratory, and we will refer to these sequences henceforth as the Potential Variants (PVs).

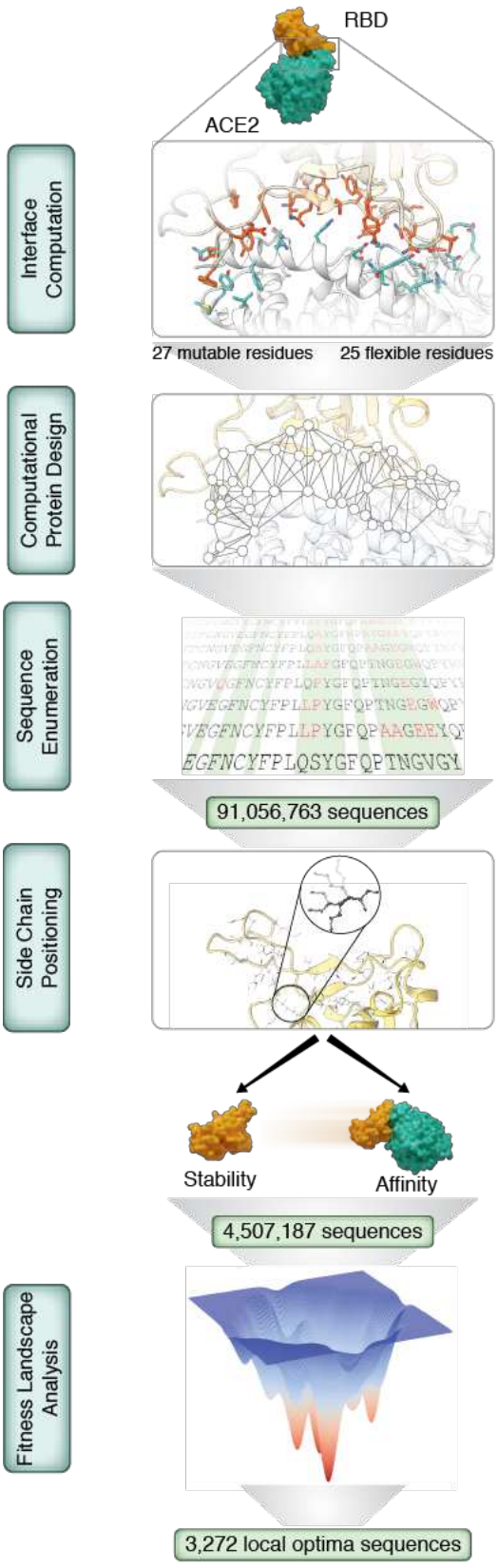

Figure 1. Computational workflow for Complete Combinatorial Mutational Enumeration Interface computation: Interface residues are computed on the ACE2/L strain RBD complex. **Computational Protein Design:** the global minimum energy conformation is computed. **Sequence enumeration:** all sequences within a 8 kcal/mol threshold are enumerated. **Side chain positioning:** the energy of all sequences enumerated on the complex form is computed on the RBD apo form. Next, filters are applied in order to only keep sequences with stable apo conformation and good affinity towards ACE2. **Fitness Landscape Analysis:** the $\Delta\Delta G$ mutational fitness landscape is analyzed and 3,272 local optima sequences are retained.

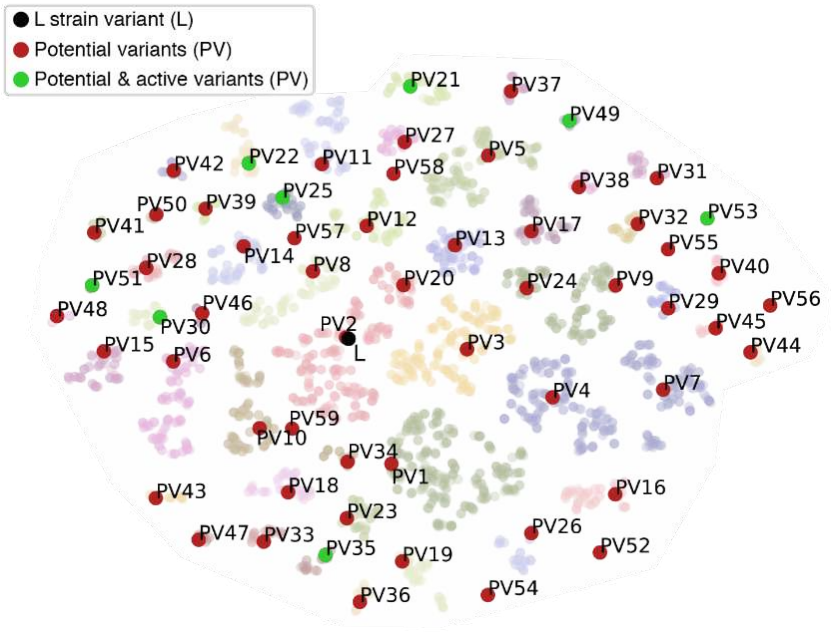

Figure 2. Clustering identification of the 59 SARS-CoV-2 potential variants (PVs). Spatial distribution of the 59 identified clusters on a t-distributed stochastic neighbor embedding (t-SNE) map. The L strain sequence is highlighted in black, the 59 PVs (i.e., medoids of each cluster) are highlighted in red, and the active potential variants (*i.e.*, validated for Fc-Ace2 binding and/or infectivity) are highlighted in green.

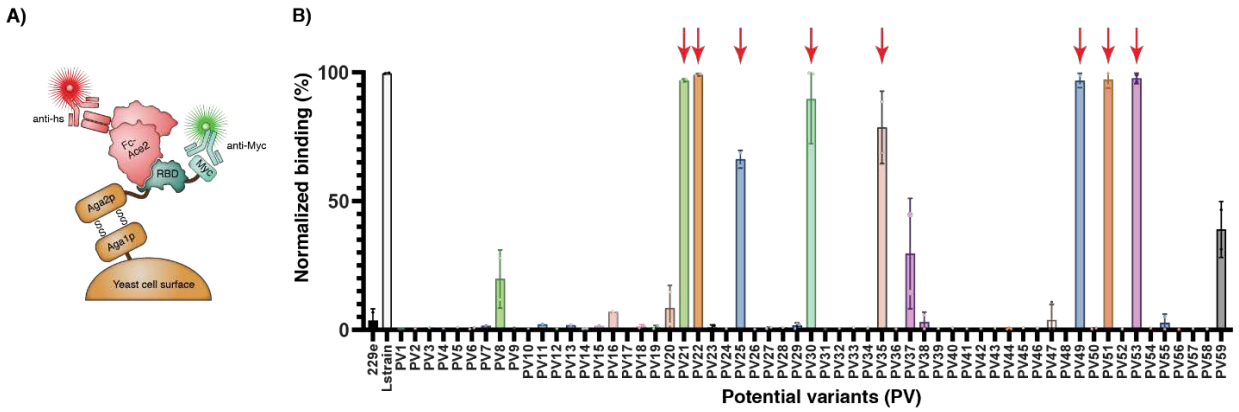

Figure 3. Screening of the SARS-CoV-2 potential variants (PVs) for binding to Fc-ACE2 using yeast display. (A) The RBDs of the PVs were expressed on the surface of *Saccharomyces cerevisiae* cells as genetic fusions to the AGA2 surface protein. An anti-Myc antibody (Ab) and an anti-human secondary Ab labeled with different fluorophores were subsequently used to label the RBD-expressing cells and the cells bound to Fc-ACE2, respectively. **(B)** *Saccharomyces cerevisiae* cells displaying the indicated RBD PV were incubated with 40 nM Fc-ACE2. Binding to Fc-ACE2 was detected with an Alexa647-conjugated anti-human secondary Ab. The RBD variants highlighted with red arrows were chosen for further experimental characterization.

The 59 PVs span a wide region of the sequence space. They each have 10 to 15 amino acid changes compared to the L strain, which in many cases is over half of all residues that interact with ACE2 (Fig. S3, Table S1). When compared to one another, the closest PVs have as few as 4 amino acid differences, and the furthest have up to 15 differences (Table S1). Interestingly, some residue positions are highly conserved across PVs, while others are highly variable. Notably, aromatic residues in the L strain are more likely to be conserved between our PVs, and this is consistent with the variability observed in the GISAID database of clinical isolates (Fig. S4). Overall, our PVs exhibit higher sequence entropy than currently identified clinical isolates (Fig. S5).

Potential variant RBDs bind ACE2

We next sought to assess the fidelity of our *in silico* predictions using *in vitro* experiments. We obtained synthetic genes encoding the 59 PVs and expressed them from *Saccharomyces cerevisiae* as fusions to AGA2 on the cell surface. This system enabled us to rapidly evaluate all of the PV RBDs for the ability to bind their receptor ACE2 using yeast display and fluorescence activated cell sorting (Fig. 3A). We used the RBD from the L strain as the positive control, and the RBD from human coronavirus 229E, which binds to a different receptor, as a negative

control[14]. ACE2 was expressed, purified and used as an Fc fusion, as this construct recapitulates the dimer that this protein forms endogenously.

From the 59 PVs, 11 showed binding at 40 nM Fc-ACE2 (Fig. 3B), which is the K_D for the soluble RBD of the L strain, and 8 of these PV RBDs bound at levels comparable to the L strain. Some PVs exhibited binding at Fc-ACE2 concentrations as low as 1 nM (fig. S6). This result was encouraging, especially considering the high amount of protein sequence changes.

We obtained mammalian expression constructs for the 8 PV RBDs that showed best binding, as well as the L strain and 229E. RBDs from seven PVs could be purified with yields similar to those of the L strain. PV35 was not able to be expressed in this system. We used biolayer interferometry (BLI) to measure the binding affinity of each RBD to Fc-ACE2. Binding to the L strain RBD was concordant with previously reported affinities [15,16], and the PV RBDs bound to Fc-ACE2 with varying affinities (Fig. 4). PV30 bound with the tightest affinity, which was similar to that of the L strain. No binding was detectable for PV25.

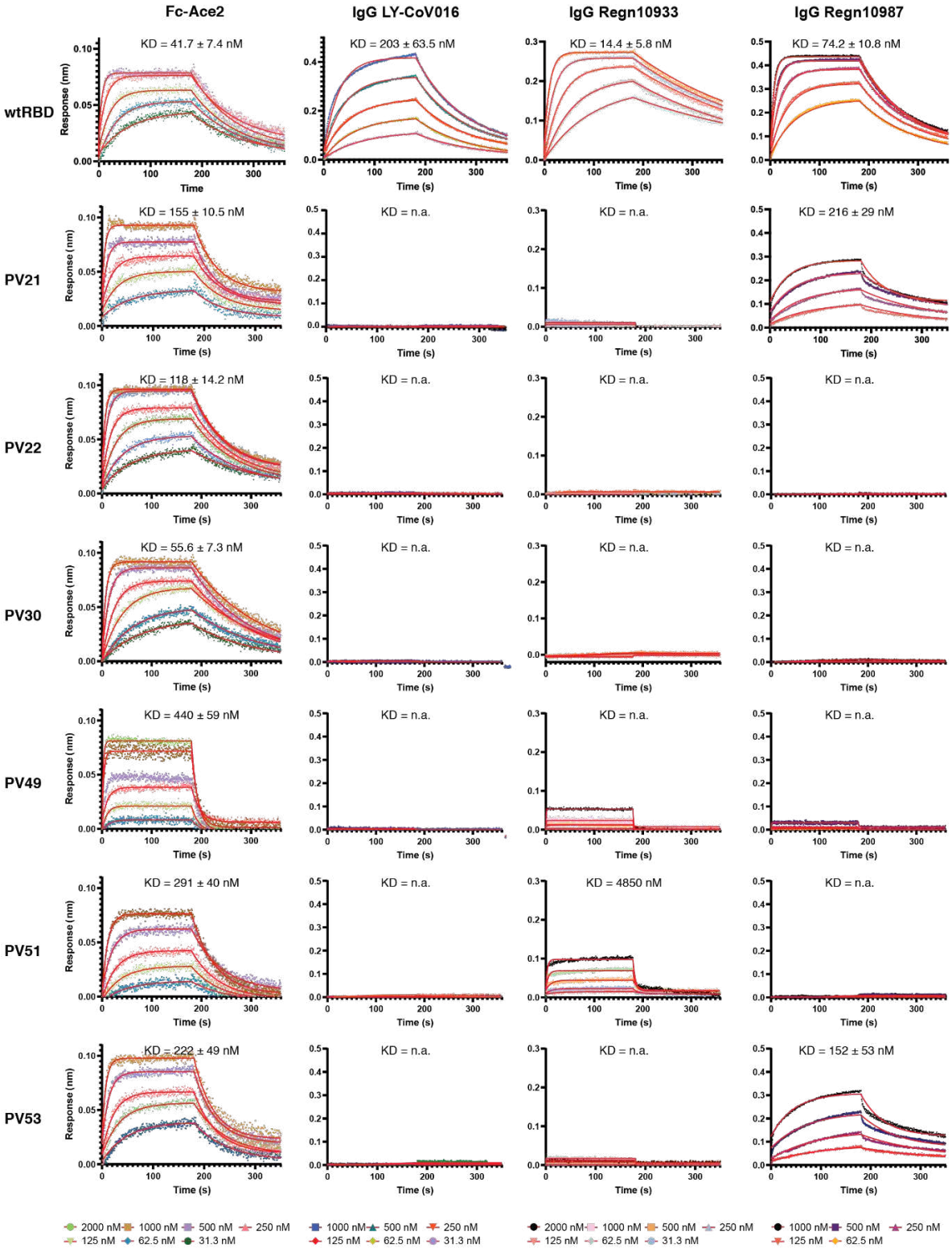
Fc-Ace2
IgG LY-CoV016
IgG Regn10933
IgG Regn10987
wtRBD
PV21
PV22
PV30
PV49
PV51
PV53
KD = 41.7 ± 7.4 nM
KD = 203 ± 63.5 nM
KD = 14.4 ± 5.8 nM
KD = 74.2 ± 10.8 nM
KD = 155 ± 10.5 nM
KD = n.a.
KD = n.a.
KD = 216 ± 29 nM
KD = 118 ± 14.2 nM
KD = n.a.
KD = n.a.
KD = n.a.
KD = 55.6 ± 7.3 nM
KD = n.a.
KD = n.a.
KD = n.a.
KD = 440 ± 59 nM
KD = n.a.
KD = n.a.
KD = n.a.
KD = 291 ± 40 nM
KD = n.a.
KD = 4850 nM
KD = n.a.
KD = 222 ± 49 nM
KD = n.a.
KD = n.a.
KD = 152 ± 53 nM
Response (nm)
Time
Time (s)
2000 nM 1000 nM 500 nM 250 nM 125 nM 62.5 nM 31.3 nM
1000 nM 500 nM 250 nM 125 nM 62.5 nM 31.3 nM
2000 nM 1000 nM 500 nM 250 nM 125 nM 62.5 nM 31.3 nM
1000 nM 500 nM 250 nM 125 nM 62.5 nM

Figure 4. Binding affinities of the SARS-CoV-2 potential variants (PVs) for the Fc-ACE2 receptor three neutralizing antibodies as measured by biolayer interferometry (BLI). Binding affinities of the indicated RBD to Fc-ACE2 and the neutralizing antibodies as analyzed by biolayer interferometry (BLI).

Potential variant pseudoviruses

We sought to confirm whether functional sequences identified using CCME could form infectious viruses. We used non-infectious pseudovirus particles to model viral cell entry, as this approach is safe and has been demonstrated to reliably recapitulate this stage of the viral life cycle[17]. We produced SARS-CoV-2 S-pseudotyped lentiviral particles in which the different PV RBDs replace the L strain RBD (Fig. 5A). Following purification, we quantified the pseudovirus particles using real-time PCR (fig. S7), and equal amounts were used to transduce both ACE2-negative (ACE2-) and ACE2-expressing (ACE2+) HEK293 cells (fig. S8). Not surprisingly, the pseudotyped viral particles expressing the six PVs that recognized Fc-ACE2 in BLI experiments were also able to transduce ACE2+ cells. The PV with the lowest Fc-ACE2 affinity, PV49, showed the least efficient transduction (fig. 5B). Interestingly, PV35, which could not be expressed as a soluble RBD (see above), was able to efficiently transduce ACE2+ cells, suggesting that it can be stably expressed in the context of the S protein.

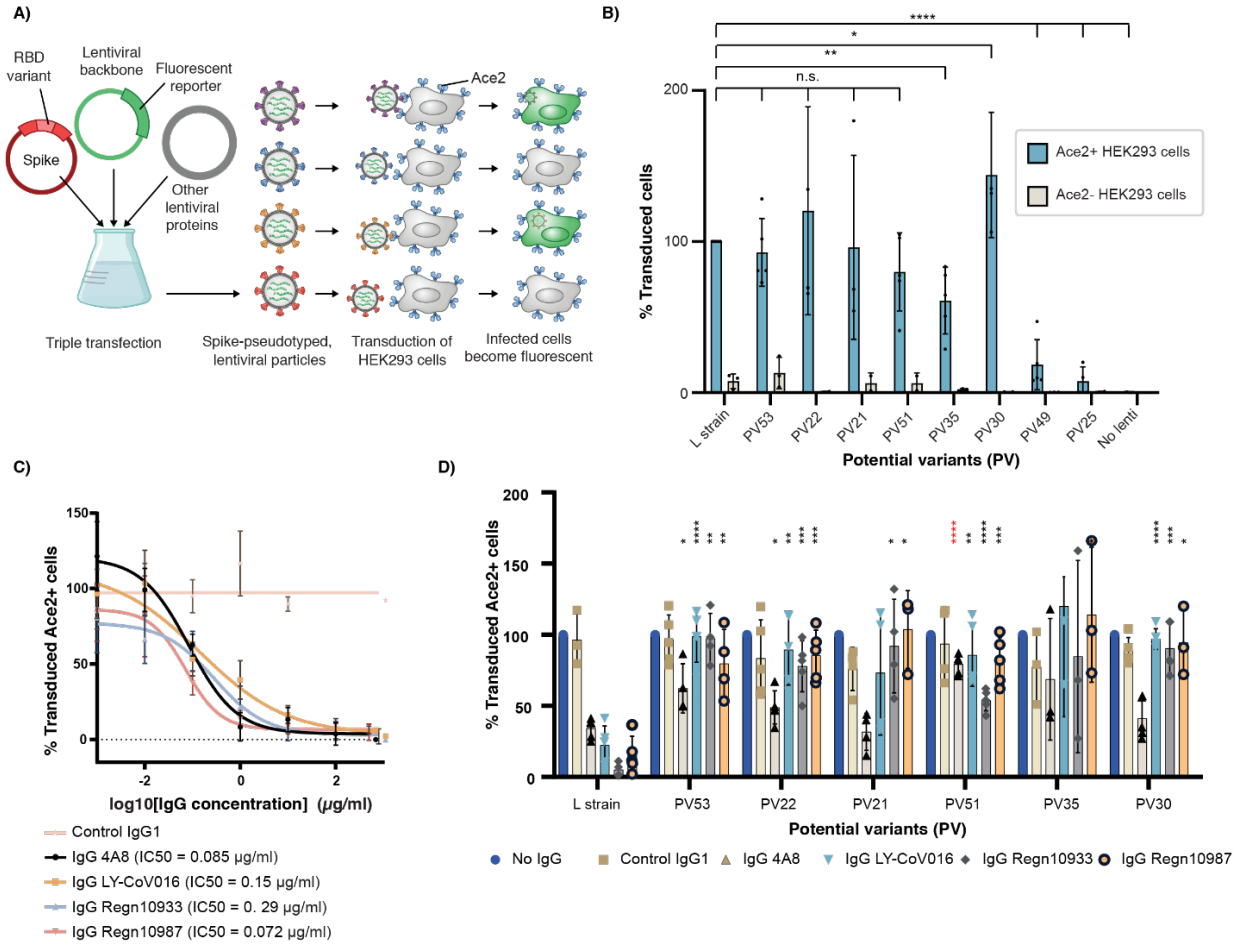

Figure 5. Infectivity of Ace2-expressing cells by the SARS-CoV-2 PVs and neutralization by neutralizing antibodies. (A) Fluorescent lentivirus pseudotyped with the SARS-CoV-2 S protein containing the different RBD potential variants (PV) were used to transduce ACE2-expressing HEK293 cells. In this setup, cell entry is dependent on ACE2 expression and cell fluorescence can be measured as a read-out for lentivirus transduction. **(B)** Equal amounts of lentivirus expressing the different PVs were used to transduce ACE2-expressing (ACE2+) and wild-type (ACE2-) HEK293 cells. Cell fluorescence was measured by flow cytometry and normalized by the % of cells transduced by lentivirus pseudotyped with the L strain. Data from 3 biological replicates is shown. **(C)** To assess the neutralization capacity of the indicated therapeutic antibodies, IC50s were first determined for neutralization of the L strain. **(D)** Ten-fold excess of the estimated IC50 concentrations of each antibody were pre-incubated with the different pseudotyped lentivirus variants before adding them to ACE2+ cells. Fluorescence values are normalized by the no IgG control. P-values < 0.05 as compared to the L strain RBD lentiviral particles are shown: n.s.: no significant. *, **, ***, ****: p values < 0.05, 0.005, 0.0005, 0.00005, respectively.

Neutralizing antibody escapeçlol

Next, we assessed how efficacious the current FDA-approved neutralizing monoclonal antibody therapeutics would be for treating infection by a forecasted PV. We expressed and purified the two antibodies from the Regeneron cocktail (Regn-10933 and Regn-10987[18,19]), as well as Eli Lilly Ly-CoV016[19], which recognizes an overlapping but not identical epitope on the RBD [20]. We used BLI to measure the binding of these antibodies to purified RBD. While all three antibodies bound the L strain RBD, the PVs exhibited substantially diminished binding. All tested PVs escape Ly-CoV016, while only PV51 is recognized by the Regn10933 antibody at a very low affinity. In addition, only two PVs (PV21 and PV53) are recognized by the Regn10987 antibody, but also at a decreased affinity compared to the L strain RBD (fig. 4). Finally, we evaluated the neutralization capacity of the therapeutic antibodies on the pseudovirus particles expressing the different PVs. In addition to Regn10933, Regn10987 and Ly-CoV016, we included the neutralizing antibody 4A8, which neutralizes SARS-CoV-2 infection by binding to the NTD of the S protein [21]. The neutralization capacity of the Regn10933, Regn10987 and Ly-COV016 antibodies decreased for pseudovirus particles expressing all PVs in comparison to those expressing the RBD of the L strain (Fig 5C, D). In contrast, the 4A8 antibody was still able to neutralize most of the PVs, as expected. Together, these data indicate that forecasted receptor binding sites on viral cell entry proteins are capable of forming infectious virions that evade extant therapeutics.

Mapping the functional sequence landscape

A major advantage of CCME is that it enables mapping of protein sequence space, and this can be used to identify sequences to target for improved therapeutics and vaccines. To do this, we need to evaluate the probability of all mutational paths through the accessible sequence landscape, *i.e.*, the series of amino acid substitutions that a protein may accrue over time. When a protein's function is absolutely required for an organism's fitness, as is the case with viral spike proteins, it is reasonable to assume that non-functional sequences are highly unlikely to propagate. Furthermore, an important consequence of DNA encoding is that some amino acid substitutions are more likely to occur than others [22], and this also depends on the genome GC contents (the SARS-CoV2- genome is 62% AU/AT rich). We therefore derived amino acid-level mutation probabilities from RNA mutation probabilities. This was preferred over a BLOSUM based estimation or an estimation that could be derived from a large protein language model because these estimations are not specialized for the considered organism (with its high AU/AT contents [23]). Also, these estimations all capture fitness for function. In our case, fitness for function is represented in $\Delta\Delta G$ and we need a purely physical probability of the mutation event.

Based on these assumptions, we constructed a graph of the sequence fitness landscape where edges between nodes are weighted by mutational probability (Fig. 6). We looked for the most probable mutational paths from the L strain to several PVs in the $\Delta\Delta G$ fitness landscape. For all of them, we found paths with lengths varying from 7 to 16 single mutations, in which all intermediates remain in the functional sequence landscape (Table S3). To have a higher-order view of the viral functional variant graph topology, we computed communities, which are

subsets of tightly connected variants. Such communities are separated from the others by a few edges that must be traversed in order to reach them. For the SARS-CoV-2 RBD, the resulting community graph shows three isolated “sequence islands.” Thus, it is highly unlikely that the virus will be able to mutate across the gap between islands. Only four communities are strongly connected to the L strain community (Fig 4). One of our active PVs (PV30) lies in the L strain community and another (PV51) belongs to a neighboring community. Therefore, one way to confine viral evolution could be to design vaccines that protect against the most probable and infectious variants linking the L strain community to these neighboring communities.

Limitations and future outlook

The novel betacoronavirus SARS-CoV-2 emerged in late 2019 and caused a global pandemic that has resulted in more than 6 million deaths thus far [24]. While effective vaccines [25] and therapeutics [26] were developed with unprecedented speed, this wasn’t sufficient to keep up with the pace of viral evolution. New strains quickly emerged that were able to infect vaccinated persons and escape neutralizing monoclonal antibody treatments [27–29]. Inevitably, SARS-CoV-2 will continue to evolve [30].

All current vaccines and therapeutics for infectious diseases work by targeting pathogen virulence factors that already exist. For example, our response to endemic pathogens like influenza is to continuously develop and administer seasonal vaccines to protect against newly emerged variants [31]. Similarly, it is all too common for monoclonal antibody treatments for infectious disease to decline in efficacy as pathogens inevitably evolve resistance [32]. Much like we do with weather, humanity needs the ability to forecast pathogen evolution and predict amino acid changes to virulence factor proteins long before they occur. A first step towards this goal is to be able to map the functional mutational landscape of a pathogen’s virulence factor proteins. This ability could enable the design of vaccines and therapeutics which are able to retain their efficacy against novel variants and prevent them from spreading.

The work presented here represents an important first step towards anticipation of pathogen evolution. Still, improvements remain to be made and additional factors have to be included. First, many drivers of selective pressure during evolution such as antibody escape are not taken into account here. Second, the false-positive prediction rate must improve, as many PVs did not bind to ACE2 in our experiments. Third, the false-negative rate must also improve. We performed these studies before the emergence of the Omicron variant, and retrospective analysis revealed that 4 out of the 7 sampled Omicron mutations did not appear in our enumerations, even if epistasis is taken into account, including Q498R and N501Y. The Q498R and N501Y mutations have a significantly higher energy than their wild-type counterparts in our calculations, due to steric clashes (Fig. S9). Proteins are not static molecules, and solving these challenges will likely require accounting for protein dynamics during enumeration, as well as enumerating a larger number of residues at one time to account for long-range epistatic effects that are not directly part of the functional site.

In contrast to other approaches [33,34], CCME enumerates the entire sequence space of a functional protein site. Longitudinal genetic sequencing is not required, so CCME can be applied immediately after a novel pathogen is discovered. A unique and important advantage of this approach is its ability to predict epistatic effects (Table S5). This enabled us to identify several highly-mutated PVs that support viral fitness, yet are evolutionarily isolated by non-functional sequences on an inaccessible sequence island. This indicates that myriad functional protein sequences exist that are largely inaccessible to life via Darwinian evolution.

CCME is a promising path towards computationally designed vaccines that would need to be updated less frequently, and may enable near complete eradication of rapidly evolving viruses like coronavirus and influenza to the same degree that humanity was able to achieve for slowly evolving viruses like pox or polio.

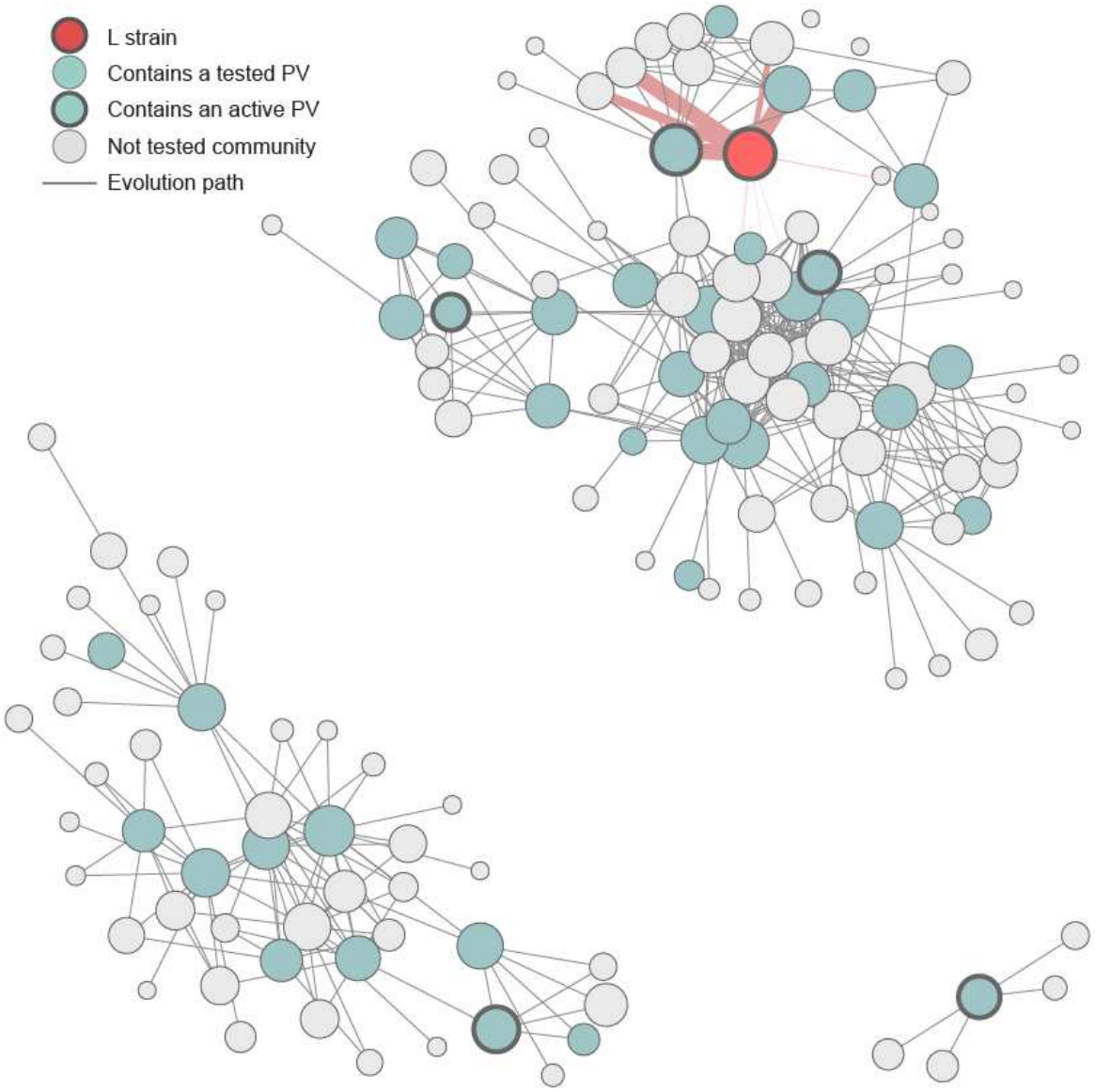

Figure 6. **Sequence community graph.** Each node represents a community of sequences. The red node represents the community that contains the L strain variant and blue nodes represent communities that contain tested PVs. Nodes with thick circles represent communities that contain active PVs. Thickness of red edges shows how connected the L strain community is.

Materials and Methods

Protein models preparation

The crystal structure of the SARS-CoV-2 spike receptor-binding domain (RBD) bound to the ACE2 receptor was retrieved from the Protein Databank (pdb code 6M0J) and used as a starting point for protein models preparation. Two protein models were derived from this crystal structure : the RBD/ACE2 complex form and the RBD unbound form. Both structures were relaxed 100 times using rosetta modeling suite version 3.12 [35,36] and the lowest scoring models were kept. The relaxations were made with coordinate constraints ensuring that the models do not deviate by more than 0.15 Angstroms from the initial crystal structure [37]. Additional flags were set in order to account for glycosylated amino acid residues.
After relaxation, the RBD/ACE2 interface residues were computed on the complex form using Rosetta scripts and the InterfaceByVector residues selector with default settings. 27 and 25 residues were selected respectively on the RBD and ACE2 side.

Computational protein design and exhaustive sequence enumeration

Computational protein design and exhaustive suboptimal sequence enumeration tasks were performed on the RBD/ACE2 complex protein model using POMPd [12]. POMPd relies on PyRosetta to compute energy matrices. PyRosetta version r245 was used in this project. Mutable, flexible and rigid residues were defined as follows: the 27 interface residues on the RBD side were mutable, the 25 interface residues on the ACE2 side were flexible and all other residues were rigid. Mutable residues are allowed to mutate to any of the 20 natural amino acid types, flexible residues can reorient their side chain without changing their amino acid type and rigid residues are completely frozen. The Dunbrack 2010 rotamer library [38] was used to define the conformational search space. The basic level of rotamer discretization was used, no extra rotamers were added on chi angles. The Rosetta genpot energy function was used, and additional flags were set in order to account for glycosylated amino acid residues. For design, POMPd calls toulbar2 with flags “-dee: -hbfs: -m -A -s --cpd“. For enumeration the additional flags “--scpbranch -a -ub <E_{max}>” are added. All calculations were performed on the CALMIP high performance computing cluster, using Intel Skylake 6140 2.3 Ghz CPUs.

Calculations on the RBD/ACE2 complex form

A side chain positioning task was performed on the L strain structure in order to compute its optimal energy, which was determined to be -1694.57 kcal/mol. The Global Minimum Energy Conformation (GMEC), the optimal sequence using our settings, was computed. The GMEC was found to have an energy of -1704.42 kcal/mol. An exhaustive enumeration of suboptimal sequences was then performed using an energy threshold of 8 kcal/mol above the GMEC. 91,056,763 different sequences satisfying the threshold were identified. The energy threshold of 8 kcal/mol was determined to be the maximum value for which results could be obtained in one day time with the computational resources used approximately 400 gigabytes of RAM, and given that the number of sequences grows exponentially with the size of the enumeration

threshold (Fig S2). The L strain sequence is not present in the enumerated sequences, it is located at 9.85 kcal/mol from the global optimum.

Calculations on the RBD unbound form

A side chain positioning task was performed on the L strain RBD unbound form using toulbar2. Its optimal energy, using our settings, was determined to be -386.74 kcal/mol.
For each one of the 91 million sequences found in the enumeration, the RBD unbound form energy was also computed by solving 91 million NP-complete side chain positioning problems to optimality. The 27 interface residues were defined as flexible and all other residues were kept rigid. The computation was performed using a parallel implementation MPI-based variant of toulbar2. It was completed in less than 2 days on 200 CPU cores.

$\Delta\Delta$G fitness landscape

$\Delta\Delta G$ values were computed as

$$\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt} = (E_{mut}^{complex} - E_{mut}^{apo}) - (E_{wt}^{complex} - E_{wt}^{apo})$$

where $E_{mut}^{complex}$ and E_{mut}^{apo} are the energies of each mutant in the enumeration, respectively in complex and apo forms, $E_{wt}^{complex}$ and E_{wt}^{apo} are the energies of the L strain respectively in complex and apo forms. Prior to computing the fitness landscape, the 91 million sequences were filtered in order to retain only sequences having a sufficiently stable RBD unbound form, with energy less than 1 kcal/mol worse than the L strain and a negative $\Delta\Delta G$ energy (with increased predicted affinity towards ACE2). The remaining 6,390,176 sequences were further filtered in order to remove all mutants exhibiting unpaired cysteine mutations. The final set includes 4,507,187 different sequences. The fitness landscape was computed on the final set of sequences, using a Hamming distance of 1 as neighborhood and $\Delta\Delta G$ energy as the fitness function. We could include the L strain variant in the fitness landscape since it is a neighbor of two sequences.

Local optima cluster representatives calculation

The fitness landscape contained 3,272 local optima, which were clustered with mmseqs using a sequence identity threshold of 80%:

```
mmseqs easy-cluster in_fasta  out_clusters tmp --min-seq-id 0.8
```

The clustering produced 59 clusters. Each cluster medoid was then identified, and the 59 corresponding sequences were selected for experimental analysis. A sequence logo representing all local optima was computed using Weblogo[39].

Most probable paths from the L strain to active potential variants

We calculated shortest paths between the L strain and active potential variants in the $\Delta\Delta G$ fitness landscape graph in which edges were weighted by mutational probabilities. We used nucleic acid level mutation rates estimated by maximum likelihood using MEGA [40] with the

General Time Reversible model (best fit under AIC and BIC regularization) extracted from [41] on coronavirus genomic sequences. From this, we computed a transition probability matrix at the nucleic acid level using matrix exponentiation, with a time parameter adjusted to get an expected number of nucleic acid mutations of around one mutation over the designed RBD region (with 27 residues). Matrix exponentiation was computed using the Pade approximation available in Python scikit as the scipy.linalg .expm function. The resulting transition matrix gives access to transition probabilities $P(M|W)$ that a given nucleic acid base W (in the L strain) will mutate to a base M in the next time-slice. To compute the amino acid level mutation rates induced by this nucleic acid transition matrix, we first computed a codon to codon transition probability matrix, assuming independent identically distributed mutation rates given by the previous matrix. For a given amino acid A, let $lc(A)$ be the set of synonymous codons representing amino acid A, the a priori probability that a given codon c in $lc(A)$ is used to represent A is simply $f(c) = r(c)/|lc(A)|$ where $r(c)$ is the Relative Synonymous Codon Usage (RSCU) of the synonymous codon c, as computed for SARS-CoV-2 coronavirus[42]. The probability for an amino acid W, represented by a latent codon variable, to mutate in an amino acid M (represented by any of its synonymous codon) is then

$$P(M|W) = \sum_{c_W \in lc(W)} f(c_W) \sum_{c_M \in lc(M)} P(c_M|c_W)$$

The negated logarithm of the above transition probability matrix was used to weight the edges connecting two sequences in our variant landscape. The weight of a minimum cost path between two variants then defines a most likely path from the source variant to the target variant. Dijkstra's algorithm was used to compute the shortest paths from the L strain to active potential variants.

Sequence community graph

The community graph was calculated from 4,507,188 sequences (including L strain variant). It was partitioned using the Leiden algorithm and modularity as a quality measure. Each node in the graph represents a community of sequences. Edges between the nodes were weighted with mutational probabilities described previously. The size of the nodes is proportional to the log of the size of communities. The thickness of edges connecting the L strain community is proportional to the log sum of all L strain community outgoing edges weights. All nodes with a degree smaller than 3 were removed. Self edges were removed and the graph was made undirected. The Leiden algorithm was run for 50 iterations and appears to be stable with a modularity of 0,93. Calculations were done using python leidenalg library. The partition was computed with the following command:

```
la.find_partition(g, weights = 'weights', partition_type=la.ModularityVertexPartition,
    n_iterations=50)
```

Theoretical affinity of antibodies towards L strain RBD and potential variants.

Three different antibodies in complex with L strain RBD were used for ΔΔG calculations (pdb codes : 6XDG and 7C01). These complexes were relaxed 100 times using rosetta modeling

suite version 3.12 (genpot scoring function) and the lowest scoring models were kept. Coordinate constraints were set in order to ensure that the models do not deviate by more than 0.15 Angstroms from the initial crystal structure. Additional flags were set in order to account for glycosylated amino acid residues. The energy of each of these complexes was calculated. The energies of the potential variants in complex with antibodies were obtained by side chain positioning calculations.

Identification of matching natural RBD sequences

We downloaded natural spike protein sequences from GISAID (https://www.gisaid.org/), using the spikeprot0125.tar archive containing 7,352,708 and only kept the 7,241,769 sequences with length above 1620, representing putative full-length sequences (possibly containing wildcard characters 'X'). Identifying the subset of our 4,507,187 RBD motifs that appears in the database would require more than 20,000 billions pairwise alignments. Suspecting that only few of these would appear in the natural diversity, we exploited the gap structure of the motifs to look for matches of partial dense sub-motifs in the GISAID set. The submotif defined by the 18 last residues of our RBD motif contains only short gaps (the longest being 7 residues long). Our 4,507,187 RBD motifs contain only 152,487 different combinations of these 18 residues. We sorted this set alphabetically and divided it into 50 subsets. Exploiting the fact that any finite language is regular, we built a regular expression containing the disjunction of the motifs appearing in it and compiled it to a Deterministic Finite State Automata (DFA). By bringing similar motifs closer together, sorting before splitting increases the likelihood that the automata size will be small. Search was performed only in the subregion of the full sequences starting at position 451 and ending at position 519. DFA compilation and search was performed using Google's re2 library (https://github.com/google/re2), as available in the Python API pyre2 (https://pypi.org/project/pyre2/). The sets of RBD motifs that yielded no hit were discarded and the same process of division in subsets, disjunction, automata compilation and search repeated recursively until singleton sequences with hits in GISAID were identified. We then selected full RBDs containing one of these 18-residue motifs with GISAID-hits and repeated the same search process. We found no occurrence of these in the GISAID sequences. We therefore repeated the same overall process, allowing this time for precisely one mismatch. With this added matching flexibility, our set of designed RBDs had 4,905,597 hits in GISAID (67.7%), covering the L strain, Delta and Lambda VOCs (Variants Of Concern), from a total of 51 designed RBDs (see Table S5-GISAID matches). With one extra mismatch allowed, the Alpha, Kappa, Eta and Iota VOCs are also covered.

Extraction of 27-residues RBD motifs from GISAID

From all sequences of the spike protein in the spikeprot0125.tar archive, we kept the 7,241,769 sequences with length above 1,620, representing putative full-length sequences (possibly containing wildcard characters 'X'). From each sequence, we extracted the region from position 454 to 555 that was expected to contain the RBD design region (positions 404 to 505 in the L strain). We removed all sequences containing 'X' and removed duplicate sequences. A multiple sequence alignment was computed with mafft, using the L strain protein S sequence as a reference, preserving length (using mafft flags --6merpair --thread -1 --keeplength --addfragments). The 27 residues of interest were extracted from each sequence, resulting in a

set of 826 different 27-mers. All sequences with remaining gaps were removed, resulting in a set of 774 unique gapless 27-mers. A sequence logo was computed from this set using Weblogo.

2D map of ΔΔG local minima landscape

We projected the $\Delta\Delta G$ local minima landscape on a 2D map using t-distributed stochastic neighbor embedding (t-SNE) as implemented in the python sk-learn package. We defined a customized distance metric in order to ensure that local minima clusters computed with mmseqs2 are correctly identified by t-SNE:

$$d(s_1, s_2) = Hamming(s_1, s_2) + \lambda\, SameCluster(s_1, s_2)$$

Where $Hamming$ is the Hamming distance between two sequences (*i.e.,* number of mutations), $SameCluster$ is a function which returns 1 if two sequences belong to the same cluster and 0 otherwise, and λ is a control parameter ($\lambda = 20$ in our calculations). The t-SNE algorithm was run for 1000 iterations with a learning rate of 50, a perplexity value of 6 and an early exaggeration value of 12.

Sequence entropy

The sequence entropy of the 27 residues of the RBD interface were computed on the 3272 local optima of the $\Delta\Delta G$ fitness landscape, as well as on the 774 unique sequences extracted from GISAID. The Shannon entropy was calculated after normalizing the frequency of occurrence of each amino acid type at each position by the natural frequency of occurrence of amino acids as estimated in the literature [43].

Yeast display experiments

DNA sequences encoding for the receptor binding domains (RBDs) of L strain SARS-CoV-2, the human coronavirus 229E and the 59 potential variants (PVs) were synthesized by Twist Bioscience. Next, they were amplified by PCR to introduce 50 nucleotides long flanking sequences complementary to the yeast display plasmid (RRID:Addgene_41522). The amplified DNA sequences and the linearized yeast display plasmid were transformed into *Saccharomyces cerevisiae* cells (Strain EBY100; ATCC) so that the yeast homologous recombination machinery ligated the DNA sequences encoding for the RBDs at the N-terminal of the Myc tag. Transformed cells were selectively grown in tryptophan-free minimal (SD-Trp-Ura) media (6.7g/L Yeast Nitrogen Base, 5.0g/L Casamino acids, 1.065 g/L MES acid, and 2% w/v dextrose) for 24 h at 30 C, with shaking. Next day, cell media was changed to SG-CAA media (2% Galactose, 0.67% Yeast Nitrogen Base, 0.5% Casamino Acids, 0.54% Sodium Phosphate Dibasic, 0.856% Sodium Phosphate Monobasic Monohydrate) to induce RBD expression for 24 h at 30 C, with shaking. Next day, induced cells were spun down for 2 mins at 2,000 x g, resuspended in HBS blocking buffer (20 mM Hepes 7.4, 150 mM NaCl, 1% (w/v) BSA) and incubated with recombinant Fc-ACE2 for 45 mins at RT, with shaking. Next, plates were washed twice with HBS blocking buffer and incubated with 1:250 diluted FITC-conjugated anti c-Myc (Immunology Consultants Lab, CMYC-45F) and Alexa647-conjugated anti-human- antibodies for 30 mins at RT, with shaking. Cells were washed twice with HBS blocking buffer and cell

fluorescence was measured using an IntelliCyt high throughput flow cytometer. Cells were gated to exclude non-single cells, FITC labeling was used to select RBD-expressing cells, and Alexa647 labeling was used to quantify Fc-ACE2-binding cells. Fc-ACE2-binding is reported as percentage within the FITC+ population and was gated according to the Alexa647 signal of the positive (L strain RBD) and negative (229e RBD) controls.

Expression and purification of recombinant soluble proteins

The DNA constructs for the RBDs of L strain SARS-CoV-2, the human coronavirus 229E and the eight PVs that showed Fc-ACE2 binding in yeast experiments were codon-optimized for mammalian cell expression, synthesized by Twist Bioscience and cloned into a mammalian expression vector as C-terminal genetic fusions to a 10xHis, a siderocalin module and a 3C protease cleavage site. The DNA construct encoding Fc-ACE2 was acquired from Addgene (#164222) and the DNA constructs encoding the four neutralizing antibodies were synthesized by Genscript. 24 µg of the respective DNA constructs were used to transfect 30 ml of suspension Expi293F (Thermo Scientific) cells at a density of 2.5E6 cells/ml in Expi293 media (Thermo Scientific) and cells were grown at 37 C in a humidified 8% CO2 incubator, with 130 rpm shaking. After 24 h, cells were feeded with 3 mM valproic acid and 0.45% glucose. After 5 days, cells were harvested for 10 mins at 1,000 x g. All RBD variants were purified using a sepharose Ni-IMAC resin (Pierce, Thermo Scientific) and eluted by 3C protease cleavage. The expressed IgG and Fc-ACE2 were purified using a protein A resin and eluted with 150 mM NaCl, 100 mM glycine (pH 2.8).

Kinetic analyses by biolayer interferometry (BLI)

BLI experiments were performed using an Octet 8-channel system (Sartorius) using HBS blocking buffer supplemented with 0.05% (w/v) Tween-20. 30 nM Fc-ACE2 or 20 nM of the three tested therapeutic antibodies were immobilized on Octet protein A biosensors. The biosensors were dipped into wells containing purified L strain RBD, 229E or the respective PV at 2000, 1000, 500, 250, 125, 62.5 and 31.3 nM concentrations for 200 seconds, and subsequently dipped into wells containing HBS blocking buffer supplemented with 0.05% (w/v) Tween-20 for 200 seconds. Data were reference-subtracted, and curves were fitted using the GraphPad Prism association dissociation model (https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_equaton_association_then_disso.htm).

ACE2-expressing HEK293 cell lines

All transduction and neutralization experiments were performed using an Ace2-negative (Ace2-) HEK293 cell line, and two different Ace2-positive (Ace2+) HEK 293 cell lines. One ACE2+ cell line was transiently transfected with an ACE2 plasmid (Addgene #141185) followed by three rounds of hygromycin selection. Next, ACE2 expression was validated by binding of an anti-Myc antibody (Fig S.10). The other ACE2+ cell line was created by lentiviral transduction and previously published (Wu et al. 2021). Significant differences were not observed in the results obtained with both cell lines.

Production of spike-pseudotyped lentivirus

To produce lentiviral particles pseudotyped with the spike (S) protein, the RBDs of the chosen PVs were cloned in the context of the S protein into the Addgene # 145032 plasmid. Subsequently, 3 µg of the corresponding spike plasmid, 12 µg of a lentiviral backbone expressing neonGreen (Addgene #162034) and 9 µg of a 2nd generation lentiviral packaging plasmid (Addgene #122600) were used for triple-transfections of 30 ml Exp293F cells. After 24 hours, cells were feeded with 3 mM valproic acid and 0.45% glucose, and lentivirus production proceeded for 4 additional days. After that, cells were pelleted for 5 mins at 1,000 x g and supernatants were filtered using 0.45 um filters (Sartorius) and stored at 4C.
The generated lentiviral particles were quantified by RT–PCR using a commercial kit (Biovision cat. No. K1471) that includes lysis buffer, reverse transcriptase, DNA polymerase and oligos annealing to the lentiviral scaffold. A standard curve was built with the provided standards and used to quantify the lentivirus amounts. Serial 10-fold dilutions of all pseudotyped lentivirus were run.

Pseudotyped lentiviral particles transduction assays
Ace2+ and Ace2- HEK293 cells were seeded in 96 well plates at 5x10E4 cells per well in DMEM media supplemented with 10% FBS and incubated at 37 C. After 24 h, media was removed and replaced by equal amounts of all pseudovirus diluted in fresh DMEM media. After 18 hours, cell media was removed and cells were washed three times with HBS blocking buffer. Next, viral transduction was measured as neonGreen fluorescence using an IntelliCyt high throughput flow cytometer. Uninfected controls and the lentiviral particles expressing the RBD of the L strain were used to set the gates. All experiments were done in technical replicates and repeated in 3 different days, and statistical significances were calculated by unpaired t tests using the GraphPad Prism software version 9.0.

Neutralization assays
First, we estimated the IC50 of the four purified neutralizing antibodies on pseudotyped lentiviruses carrying the L strain RBD (fig. 3C), and then used a concentration corresponding to 10 x the respective IC50s for the neutralization experiments. To estimate the IC50s of the neutralizing antibodies with pseudotyped lentivirus expressing the RBD of the L strain, 10-fold dilutions of the antibodies were incubated with L strain pseudotyped lentivirus for 1 hour at 37 C and subsequently added to ACE2+ HEK293 cells. NeonGreen fluorescence was analyzed using an IntelliCyt high throughput flow cytometer and the data from 4 different experiments were used to estimate the IC50 values using the Graphpad Prism version 9.0. For the neutralization experiments, equal amounts of the lentivirus pseudotyped with the L strain or the corresponding PV variants were pre-incubated with 10 x IC50 concentrations of the 4 antibodies for 1 hour at 37 C and subsequently added to ACE2+ cells. Technical duplicates and at least 3 biological replicates of each sample were performed. Statistical significance was calculated by unpaired t tests using the GraphPad Prism software version 9.0.

References

1. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
2. Leander, M., Liu, Z., Cui, Q. & Raman, S. Deep mutational scanning and machine learning reveal structural and molecular rules governing allosteric hotspots in homologous proteins. *Elife* **11**, (2022).
3. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23**, 3–20 (2022).
4. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
5. Wang, J., Lan, J., Wang, X. Q. & Wang, H. W. Cryo-EM structure of SARS-CoV2 RBD-ACE2 complex. Preprint at https://doi.org/10.2210/pdb7dqa/pdb (2021).
6. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020).
7. Cooper, M. C. *et al.* Soft arc consistency revisited. *Artif. Intell.* **174**, 449–478 (2010).
8. Hurley, B. *et al.* Multi-language evaluation of exact solvers in graphical model discrete optimization. *Constraints* **21**, 413–434 (2016).
9. Traoré, S. *et al.* Fast search algorithms for computational protein design. *J. Comput. Chem.* **37**, 1048–1058 (2016).
10. Simoncini, D. *et al.* Guaranteed Discrete Energy Optimization on Large Protein Design Problems. *J. Chem. Theory Comput.* **11**, 5980–5989 (2015).
11. Hallen, M. A. & Donald, B. R. Protein Design by Provable Algorithms. *Commun. ACM* **62**, 76–84 (2019).
12. Vucinic, J., Simoncini, D., Ruffini, M., Barbe, S. & Schiex, T. Positive multistate protein design. *Bioinformatics* **36**, 122–130 (2020).

13. Simoncini, D., Barbe, S., Schiex, T. & Verel, S. Fitness landscape analysis around the optimum in computational protein design. *Proceedings of the Genetic and Evolutionary Computation Conference* Preprint at https://doi.org/10.1145/3205455.3205626 (2018).
14. Li, Z. *et al.* The human coronavirus HCoV-229E S-protein structure and receptor binding. *Elife* **8**, (2019).
15. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
16. Chan, K. K. *et al.* Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
17. Crawford, K. H. D. *et al.* Protocol and Reagents for Pseudotyping Lentiviral Particles with SARS-CoV-2 Spike Protein for Neutralization Assays. *Viruses* **12**, (2020).
18. Baum, A. *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
19. Shi, R. *et al.* A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120–124 (2020).
20. Barnes, C. O. *et al.* SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
21. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650–655 (2020).
22. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
23. Wang, Y. *et al.* Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. *Sci. Rep.* **10**, 12331 (2020).
24. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/.
25. Krammer, F. SARS-CoV-2 vaccines in development. *Nature* vol. 586 516–527 Preprint at https://doi.org/10.1038/s41586-020-2798-3 (2020).

26. Pomplun, S. Targeting the SARS-CoV-2-spike protein: from antibodies to miniproteins and peptides. *RSC Med Chem* **12**, 197–202 (2020).
27. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).
28. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
29. Wang, P. *et al.* Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe* **29**, 747–751.e4 (2021).
30. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
31. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* **16**, 60 (2018).
32. Bates, J. T. *et al.* Escape from neutralization by the respiratory syncytial virus-specific neutralizing monoclonal antibody palivizumab is driven by changes in on-rate of binding to the fusion protein. *Virology* **454-455**, 139–144 (2014).
33. Taft, J. M. *et al.* Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor binding domain. *Cell* (2022) doi:10.1016/j.cell.2022.08.024.
34. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.e20 (2020).
35. Pavlovicz, R. E., Park, H. & DiMaio, F. Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. *PLoS Comput. Biol.* **16**, e1008103 (2020).
36. Park, H., Zhou, G., Baek, M., Baker, D. & DiMaio, F. Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein–Ligand Docking. *Journal of Chemical Theory and Computation* vol. 17 2000–2010 Preprint at

https://doi.org/10.1021/acs.jctc.0c01184 (2021).

37. Nivón, L. G., Moretti, R. & Baker, D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS One* **8**, e59004 (2013).
38. Shapovalov, M. V. & Dunbrack, R. L., Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).
39. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
40. Kumar, S., Nei, M., Dudley, J. & Tamura, K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinform.* **9**, 299–306 (2008).
41. Sohpal, V. K. Computational analysis of SARS-CoV-2, SARS-CoV, and MERS-CoV genome using MEGA. *Genomics Inform.* **18**, e30 (2020).
42. Hou, W. Characterization of codon usage pattern in SARS-CoV-2. *Virol. J.* **17**, 138 (2020).
43. Carugo, O. Amino acid composition and protein dimension. *Protein Sci.* **17**, 2187–2191 (2008).
44. Cooper, M. C., de Givry, S. & Schiex, T. Graphical Models: Queries, complexity, algorithms. in vol. 154 4:1–4:22 (Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020).

Acknowledgements
We would like to thank Jesse Bloom for sharing the antibody DNA constructs, and Shang-Chuen Wu for sharing the genomically-integrated ACE2+ cell line. We thank CALcul en MIdi. Pyrénées (CALMIP, Toulouse, France) for providing computational resources.

ANR grant number ANR-20-CE45-0016 (JV, DS).
ANR grant number ANR-19-P3IA-0004 - ANITI - Artificial and Natural Intelligence Toulouse Institute - 3IA (2019) (TS,DS).
National Institutes of Health grant R21 AI156570 (MSC, JTB, IM, JAB, CDB)
National Institutes of Health grant R21 EB028342 (MSC, JTB, IM, JAB, CDB)
National Science Foundation grant 2031785 (MSC, JTB, IM, JAB, CDB)

Author Contributions
DS and JAB performed landscape enumeration. JV, TS, SV and DS performed analyses of the enumerated landscape. MSC, JTB and IM performed the laboratory experiments on PV RBDs. MSC, JV, TS, DS and CDB wrote the manuscript and prepared figures with input from all authors. DS and CDB supervised the study, and CDB conceptualized the study.

Competing interests
MSC, JTB, IM and CDB own stock in AI Proteins, Inc.

Materials & Correspondence
Supplementary Information is available for this paper.

Code and scripts used in this work are available at: https://github.com/deep-evo4cast/deep-evo4cast

Correspondence and requests for materials should be addressed to Chris Bahl (chris@aiproteins.bio) or David Simoncini (david.simoncini@ut-capitole.fr)

SUPPLEMENTARY INFORMATION

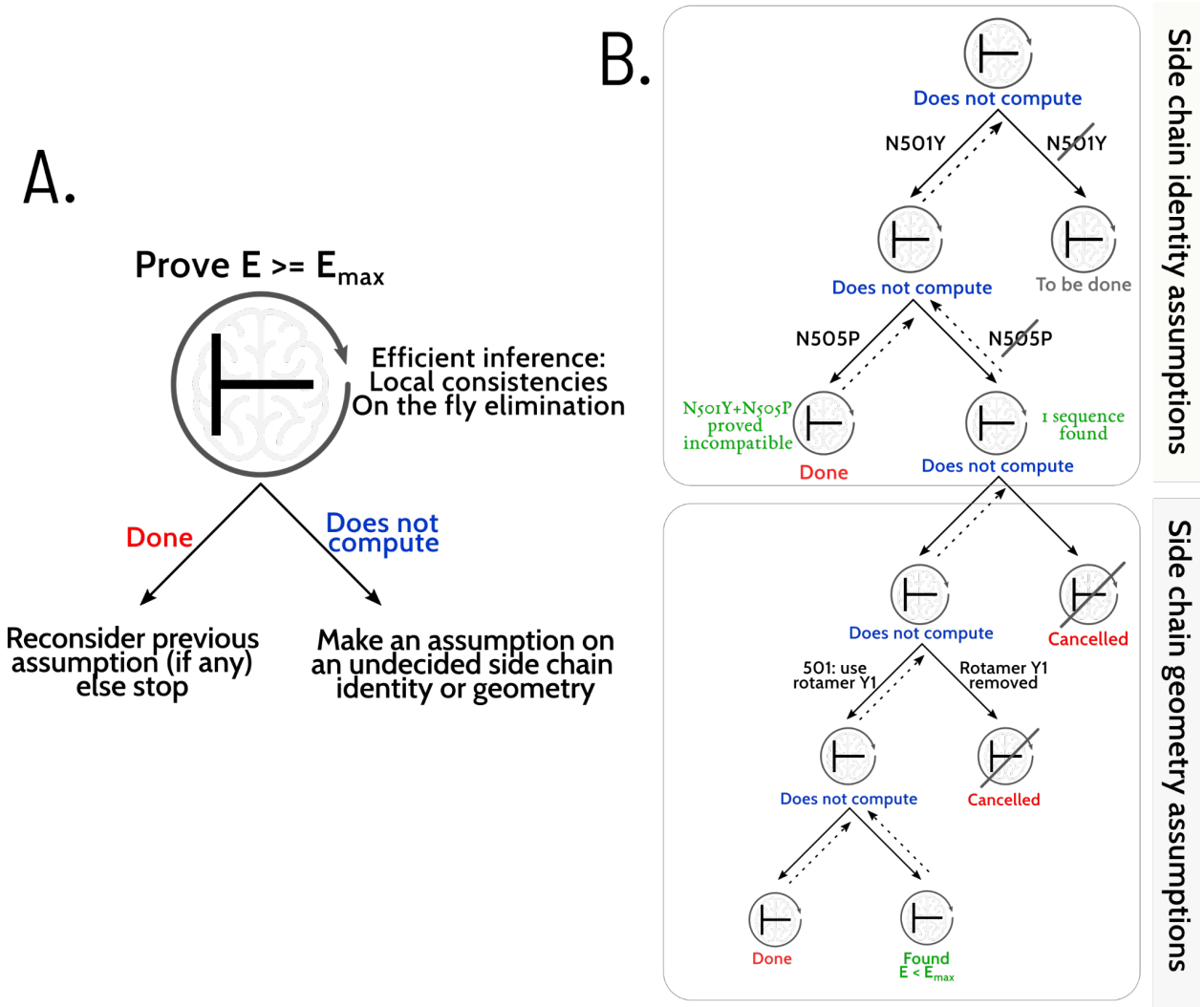

Figure S1. Toulbar2 sequence enumeration algorithm. High level description of the proof mechanisms used by toulbar2 for fast enumeration of sequences with at least one energy conformation below a threshold E_{max}.

A. Initially asked to find all sequences which can have an energy less than E_{max}, toulbar2 proceeds instead *ad absurdo*, trying to prove that all sequences must have energy above E_{max} in all their possible geometries and collects sequence counter-examples as the proof proceeds. The proof relies on efficient massive inference (symbol ⊢) by local consistencies and variable elimination [44]. Alone, these proof systems are only able to solve relatively simple problems. When the proof is out of reach, an assumption on the identity or geometry of a yet undecided side-chain is made. This makes the problem simpler and eventually solvable. If, instead, the local proof can be directly achieved, this branch of the proof is done and previous assumptions are reconsidered.

B. For sequence enumeration [9], the proof is built in two layers. In the first layer, only side-chain identities are decided. Once this is done, side-chain geometries (rotamers) are explored. As soon as a geometry of energy below E_{max} is found, a counter example (and a suitable sequence) is found and pending geometry assumptions explorations are canceled. toulbar2 therefore limits the combinatorial explosion of the protein sequence fitness landscape explored thanks to two proof pruning mechanisms: massive local inference (“Done” nodes) and counter-example based geometry pruning (“Canceled” nodes). Local inferences are also used to guide the search for counter-examples: when an assumption needs to be made, toulbar2 selects the assumption for which the last local inference was the farther away from the E_{max} target.

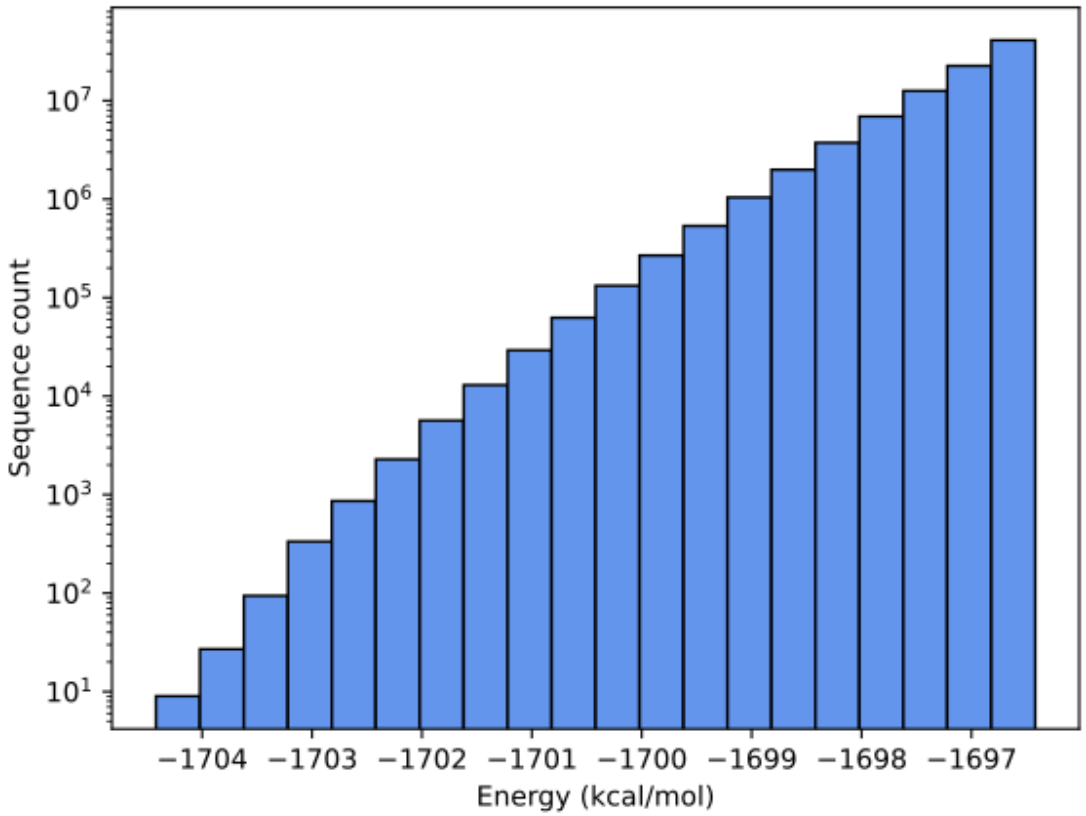

Figure S2. Energy distribution of all 91 million enumerated sequences. Non-cumulative energy distribution of sequences enumerated within 8 kcal/mol of the global minimum on the ACE2/RBD complex. The scale on the y axis is logarithmic.

```
L strain      CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
Alpha         CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
Delta         CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
Omicron       CPFDEVFNATRFASVYAWNRKRISNCVADYSVLYNLAPFFTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
PV21          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGG 70
PV22          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGS 70
PV25          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
PV30          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
PV35          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
PV49          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70
PV51          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGG 70
PV53          CPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGD 70

L strain      EVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQA 140
Alpha         EVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYRYRLFRKSNLKPFERDISTEIYQA 140
Delta         EVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYRYRLFRKSNLKPFERDISTEIYQA 140
Omicron       EVRQIAPGQTGNIADYNYKLPDDFTGCVIAWNSNKLDSKVSGNYNYLYRLFRKSNLKPFERDISTEIYQA 140
PV21          EVRQIAPGQTGLIADYNYKLPDDFTGCVIAWNSNNLDSKWGGNYNYLFRMFRKSNLKPFERDISTEIFQA 140
PV22          EVRQIAPGQTGVIADYNYKLPDDFTGCVIAWNSNNLDSKEGGNYNYLFRKFRKSNLKPFERDISTEIFQA 140
PV25          EVRQIAPGQTGAIADYNYKLPDDFTGCVIAWNSNNLDSKEGGNYNYLYRKFRKSNLKPFERDISTEIYQA 140
PV30          EVRQIAPGQTGLIADYNYKLPDDFTGCVIAWNSNNLDSKEGGNYNYLFRLFRKSNLKPFERDISTEIFQA 140
PV35          EVRQIAPGQTGAIADYNYKLPDDFTGCVIAWNSNNLDSKEGGNYNYLYRKFRKSNLKPFERDISTEIFQA 140
PV49          EVRQIAPGQTGWIADYNYKLPDDFTGCVIAWNSNNLDSKFGGNYNYLYRLFRKSNLKPFERDISTEIYQA 140
PV51          EVRQIAPGQTGEIADYNYKLPDDFTGCVIAWNSNNLDSKDGGNYNYLYRLFRKSNLKPFERDISTEIYQA 140
PV53          EVRQIAPGQTGLIADYNYKLPDDFTGCVIAWNSNNLDSKWGGNYNYLFRLFRKSNLKPFERDISTEIYQA 140

L strain      GSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPA 187
Alpha         GSKPCNGVEGFNCYFPLQSYGFQPTYGVGYQPYRVVVLSFELLHAPA 187
Delta         GSKPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPA 187
Omicron       GNKPCNGVAGFNCYFPLRSYSFRPTYGVGHQPYRVVVLSFELLHAPA 187
PV21          GSTPCNGVEGFNCYFPLLPYGFQPAAGEEYQPYRVVVLSFELLHAPA 187
PV22          GSTPCNGVEGFNCYFPLLPYGFQPTNGEGWQPYRVVVLSFELLHAPA 187
PV25          GSTPCNGVEGFNCYFPLLPFGFTPTAGEGWQPYRVVVLSFELLHAPA 187
PV30          GSTPCNGVQGFNCYFPLQPYGFQPTNGEGYQPYRVVVLSFELLHAPA 187
PV35          GSTPCNGVEGFNCYFPLLAFGFQPTNGEGWQPYRVVVLSFELLHAPA 187
PV49          GSTPCNGVEGFNCYFPLQAYGFQPAAGEGWQPYRVVVLSFELLHAPA 187
PV51          GSTPCNGVEGFNCYFPLQAYGFHPATGEEYQPYRVVVLSFELLHAPA 187
PV53          GSTPCNGVEGFNCYFPLVPYGFQPAAGEGYQPYRVVVLSFELLHAPA 187
```

Figure S3. Sequence alignment of the main natural variants of concern and potential variants validated in this study. Sequence alignment of the L strain, alpha, delta and omicron variants of concern together with the PVs showing the best binding to Fc-ACE2. RBD interface positions are highlighted in green, residues that differ from the L strain are shown in red. Most omicron mutations are not found in our predicted PVs. However, this is likely due to the fact that we only sampled the 27 RBD residues that contact Ace2, and other residues away from the interface might also contribute to RBD stability and indirectly to Ace2 binding. Also, the other spike domains play important roles in the viral life cycle, and viral fitness consists of a complex combination of factors that go beyond the spike. In the future, we will consider additional factors to improve our algorithm.

Figure S4. Sequence logos of GISAID variants and ΔΔG fitness landscape local minima. RBD interface residues sequence logo representations of 774 unique GISAID sequences (A) and ΔΔG fitness landscape local minima sequences (B).

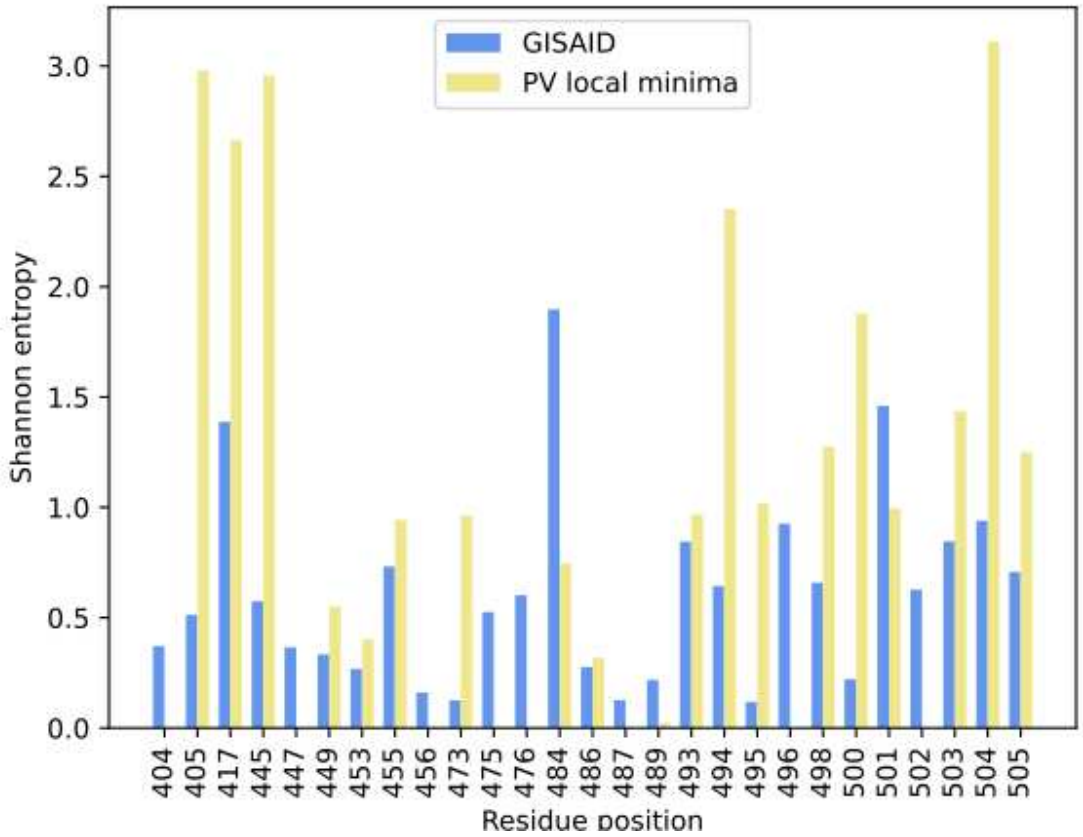

Figure S5. Shannon entropy of GISAID variants and ΔΔG fitness landscape local minima. Amino acid composition entropy of RBD interface residues for 774 unique GISAID sequences (blue) and 3272 local optima sequences (yellow).

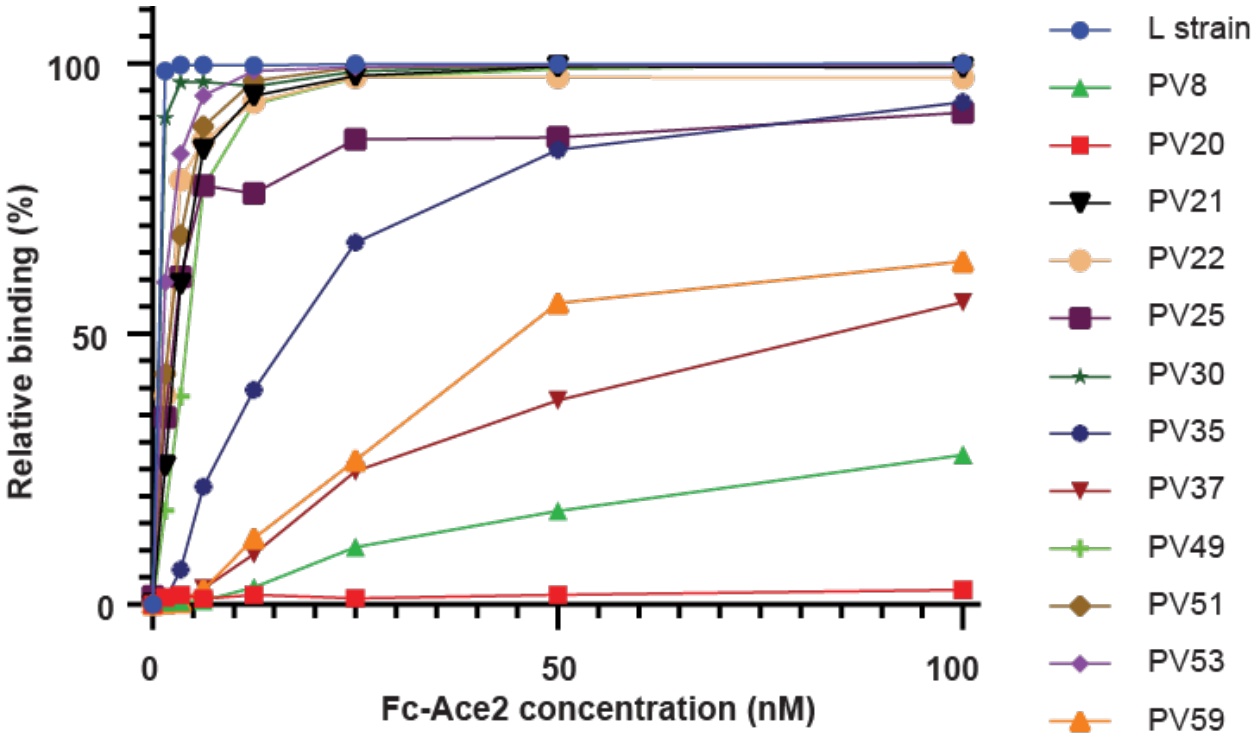

Figure S6. Dose response curves of yeast cells displaying the indicated RBD potential variants (PV) and Fc-ACE2 at decreasing concentrations. Relative binding is shown as the % of RBD-expressing cells.

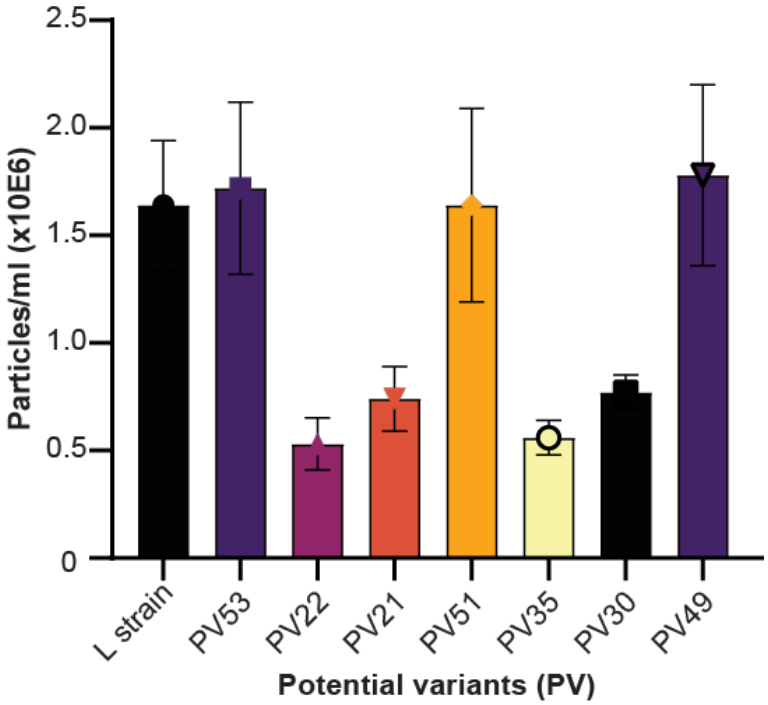

Figure S7. **Quantification of pseudotyped lentiviral particles by real time-PCR**. The viral titers were quantified by real time PCR and normalized to use equal amounts of all RBD variants in the transduction experiments. The obtained viral titers are consistent with these from Cronshaw et al. 2020.

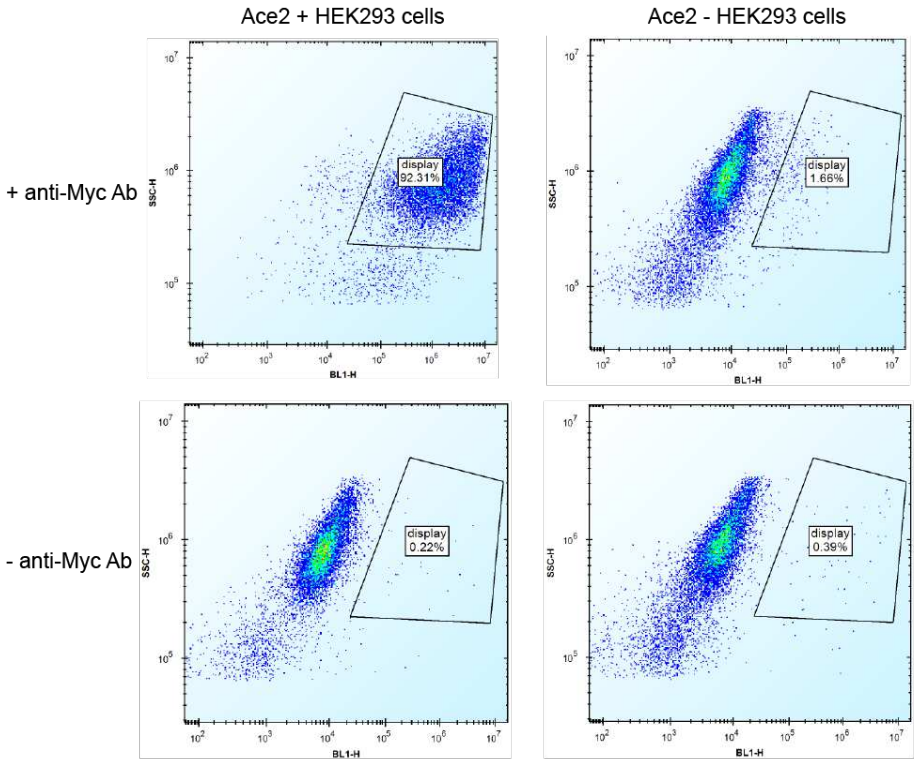

Figure S8. Validation of the Ace2+ transient cell line. To obtain the ACE2+ transient cell line, HEK293 cells were transiently transfected with a mammalian expression plasmid encoding for human ACE2 with an N-terminal Myc tag (Addgene #141185). Thus, ACE2 expression could be validated by staining HEK293 with a FITC-conjugated anti-Myc antibody.

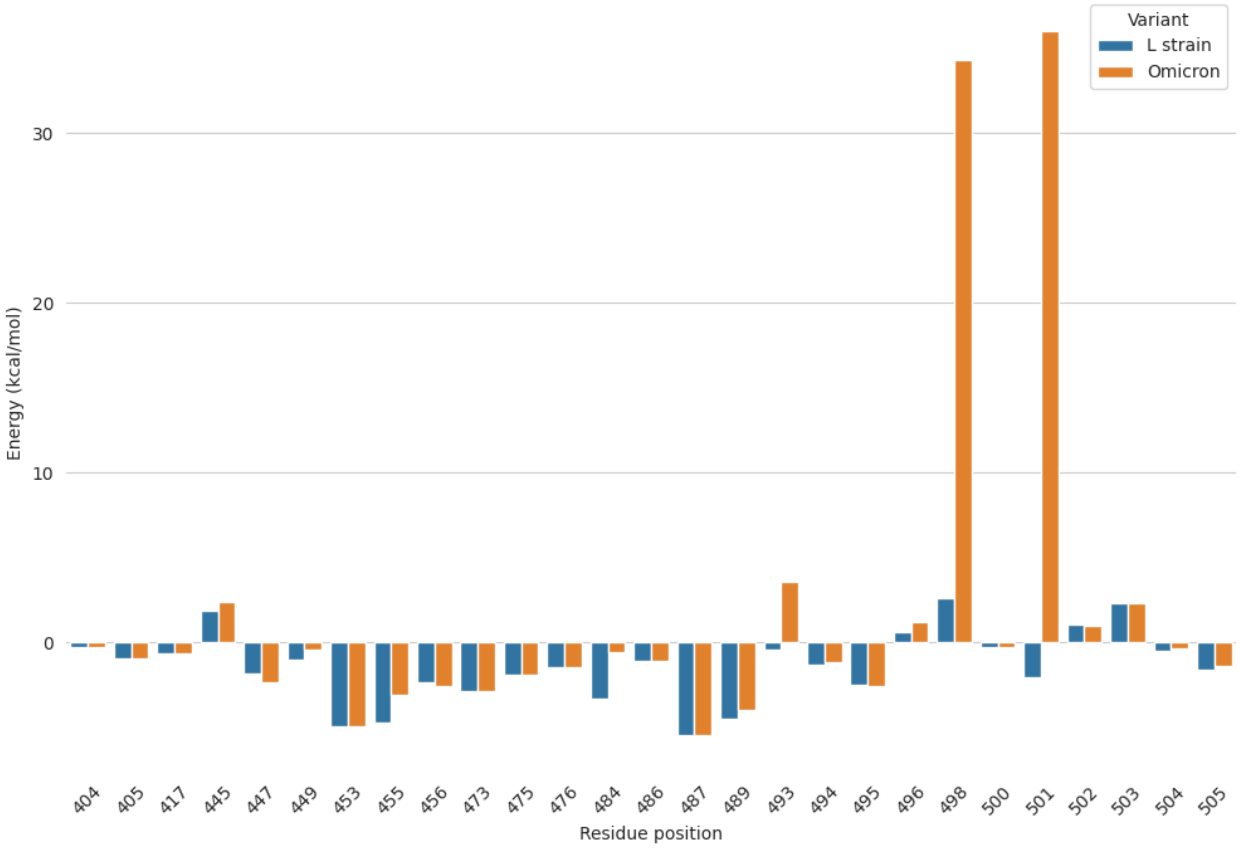

Figure S9. Per residue score breakdown of the L strain and omicron RBD/ACE2 complex form. Scoring was done on the initial backbone of the L strain RBD/ACE2 complex. Mutable interface residues are shown.

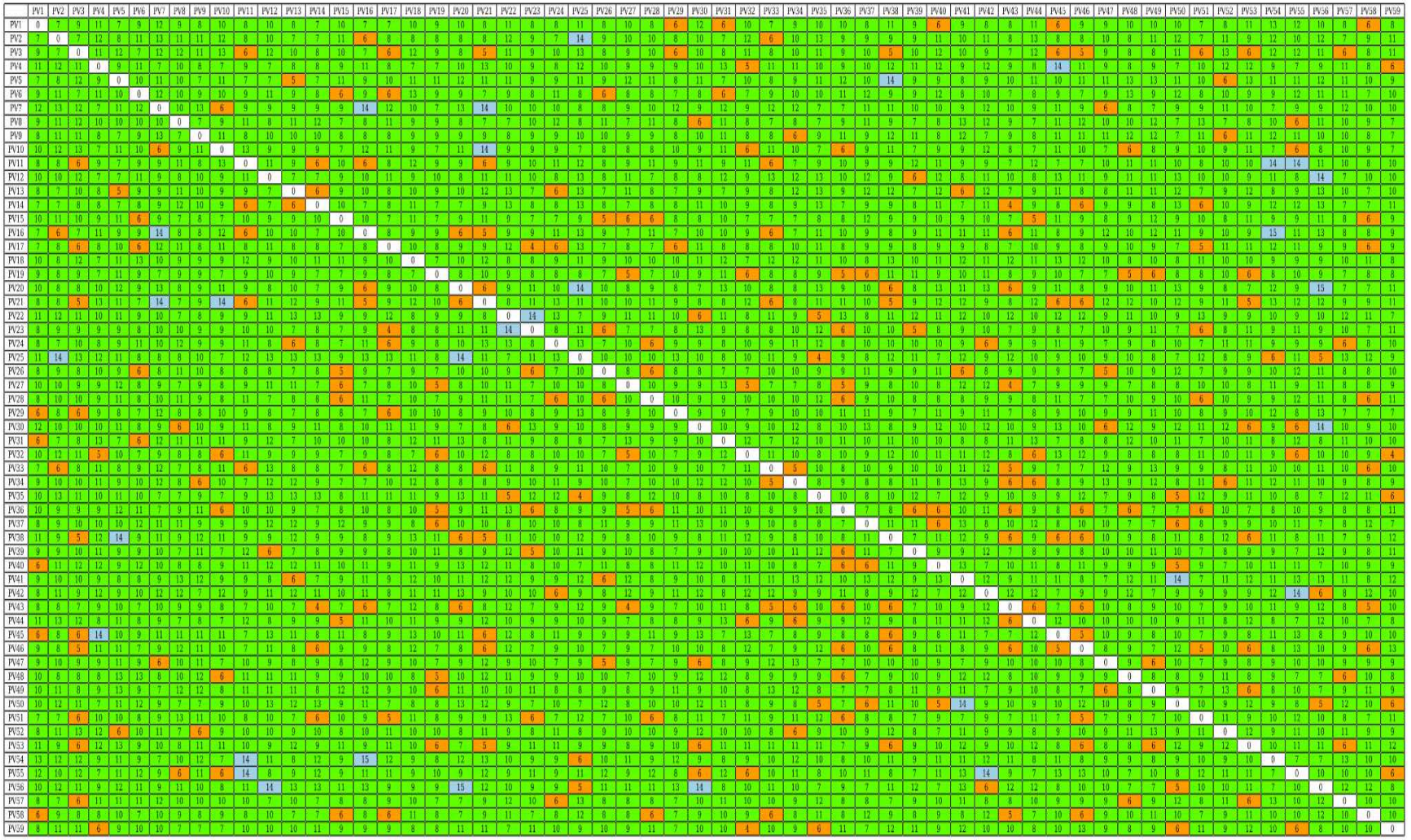

Table S1. Distance matrix of the 59 PVs. Pairwise Hamming distances between all potential variants.

		With local minimas (3,272)		With all predicted RBDs (4,507,187)	
Variant	**RBD (27 AAs)**	**Min. # of mutations**	**# of matches**	**Min. # of mutations**	**# of matches**
Beta, V2 (B.1.351)	GDNVGYYLFYAGKFNYQSYGQT YGVGY	6	2	3	1
Alpha, V1 (B.1.1.7)	GDKVGYYLFYAGEFNYQSYGQT YGVGY	5	2	2	4
Gamma, V3 (P.1)	GDTVGYYLFYAGKFNYQSYGQTY GVGY	6	2	4	40
Delta (B.1.617.2)	**GDKVGYYLFYAGEFNYQSYGQT NGVGY**	4	2	1	2
Kappa (B.1.617.1)	GDKVGYYLFYAGQFNYQSYGQT NGVGY	5	5	2	8
Eta (B.1.525)	GDKVGYYLFYAGKFNYQSYGQT NGVGY	5	2	2	2
Iota (B.1.526)	GDKVGYYLFYAGKFNYQSYGQT NGVGY	5	2	2	2
Lambda (C.37)	GDKVGYYLFYAGEFNYQSYGQT NGVGY	4	2	1	2
Mu (B.1.621)	GDKVGYYLFYAGKFNYQSYGQT YGVGY	6	2	3	4
Omicron (BA.1)	GDNVGYYLFYAGAFNYRSYSRTY GVGH	>8	None (<=8)	7	21
Omicron (BA.2)	GNNVGYYLFYAGAFNYRSYGRTY GVGH	>8	None (<=8)	6	2
L strain (WT)	GDKVGYYLFYAGEFNYQSYGQT NGVGY	4	2	1	2

Table S2. Distances between variants of concern, local minima and potential variants in the filtered fitness landscape.

Potential Variant (PV)	Mutations on the path	Distance from wild type	Path length
PV49	V503E,V445F,S494A,Y505W,T500A,N501A,K417W	7	7
PV30	V503E,S494P,V445E,E484Q,K417L,Y453F,Y473F	7	7
PV53	V503E,T500A,V445L,L445W,Y453F,S494P,K417L,N501A,Q493V	8	9
PV21	V503E,D405G,G504E,K417R,V445L,T500A,L445W,S494P,Q498V,L455M,Q493L,Y473F,Y453F,R417L,N501A,V498Q	12	16
PV22	V503E,D405G,S494P,K417E,G405R,V445E,Y505W,L455K,Q493L,R405S,Y453F,E417V,Y473F	10	13
PV51	V503E,D405G,G504E,T500A,S494A,V445D,Q498H,K417E,N501T	9	9
PV25	V503E,S494P,V445E,Y495F,K417E,Y505W,L455K,Q493L,Q498T,N501A,E417A	10	11

Table S3. Most probable mutational paths from L strain to antibody escaping PVs. For each PV, we show all mutations on the most probable path in the order in which they appear as well as the Hamming distance from the L strain and the path length.

Mismatch pos.	Potential PV in the filtered fitness landscape	# of hits in GISAID	Comment
494	GDKVGYYLFYAGEFNYQAYGQTNGDGY	7	
498	GDKVGYYLFYAGEFNYQSYGVTNGLGY	8	
498	GDKVGYYLFYAGEFNYQSYGVTNGDGY	7	
498	GDKVGYYLFYAGEFNYQSYGVTNGAGY	48	
498	GDKDGYYLFYAGEFNYQSYGVTNGVGY	7	
405	GGKVGYYLFYAGEFNYQSYGQTNGVEY	1	
445	GDKEGYYLFYAGEFNYQSYGQTNGAGY	48	
445	GDKHGYYLFYAGEFNYQSYGQTNGAGY	48	
445	GDKEGYYLFYAGEFNYQSYGQTNGDGY	7	
445	GDKHGYYLFYAGEFNYQSYGQTNGDGY	7	
445	GDKMGYYLFYAGEFNYQSYGQTNGDGY	7	
445	GDKEGYYLFYAGEFNYQSYGQTNGLGY	8	
445	GDEEGYYLFYAGEFNYQSYGQTNGVGY	29	
503	GAKVGYYLFYAGEFNYQSYGQTNGEGY	14	
503	GDAVGYYLFYAGEFNYQSYGQTNGEGY	1	
503	GDEVGYYLFYAGEFNYQSYGQTNGEGY	29	
503	GDKAGYYLFYAGEFNYQSYGQTNGEGY	187	
503	GDKDGYYLFYAGEFNYQSYGQTNGEGY	7	
503	GDKDGYYLFYAGEFNYQSYGQTNGH	7	

	GY		
503	GDKFGYYLFYAGEFNYQSYGQTNGEGY	218	
503	GDKFGYYLFYAGEFNYQSYGQTNGSGY	218	
503	GDKGGYYLFYAGEFNYQSYGQTNGEGY	6	
503	GDKIGYYLFYAGEFNYQSYGQTNGEGY	211	
503	GDKIGYYLFYAGEFNYQSYGQTTGEGY	1	
503	GDKLGYYLFYAGEFNYQSYGQTNGEGY	2	
503	GDKSGYYLFYAGEFNYQSYGQTNGEGY	1	
503	GDKVGFYLFYAGEFNYQSYGQTNGEGY	5	
503	GDKVGYYLFFAGEFNYQSYGQTNGEGY	44	
503	GDKVGYYLFYAGEFNYQAYGQTNGDGY	148	
503	GDKVGYYLFYAGEFNYQAYGQTNGEGY	148	
503	GDKVGYYLFYAGEFNYQAYGQTNGHGY	148	
503	GDKVGYYLFYAGEFNYQAYGQTNGSGY	148	
503	GDKVGYYLFYAGEFNYQPFGQTNGEGY	1	
503	GDKVGYYLFYAGEFNYQPYGQTNGEGY	7 268	
503	GDKVGYYLFYAGEFNYQPYGQTNGSGY	7 268	
503	GDKVGYYLFYAGEFNYQPYGQTTGEGY	6	
503	GDKVGYYLFYAGEFNYQSFGQTNGEGY	1	
503	GDKVGYYLFYAGEFNYQSYGQANGEGY	14	
503	GDKVGYYLFYAGEFNYQSYGQTNGEGW	18	

503	GDKVGYYLFYAGEFNYQSYGQTNGEGY	4 884 812	L/Delta/Lambda
503	GDKVGYYLFYAGEFNYQSYGQTNGSGY	4 884 812	L/Delta/Lambda
503	GDKVGYYLFYAGEFNYQSYGQTTGEGY	4 593	
503	GDKYGYYLFYAGEFNYQSYGQTNGEGY	1	
503	GDMVGYYLFYAGEFNYQSYGQTNGEGY	16	
503	GDNVGYYLFYAGEFNYQPYGQTNGEGY	2	
503	GDNVGYYLFYAGEFNYQSYGQTNGEGY	7 817	
503	GDQVGYYLFYAGEFNYQSYGQTNGEGY	4	
503	GDRVGYFLFYAGEFNYQSYGQTNGEGY	2	
503	GDRVGYYLFYAGEFNYQSYGQTNGEGY	70	
503	GDSVGYYLFYAGEFNYQSYGQTNGEGY	1	
503	GGKVGYYLFYAGEFNYQSYGQTNGEGY	96	
	Total unique hits at distance 1	4 905 597	

Table S4. Number of GISAID RBD sequences that would fit in the filtered fitness landscape. List of all predicted RBDs from the filtered sequence landscape having a match within at most one mutation in the spike protein GISAID database, filtered for complete sequences. A total of 4,905,597 GISAID sequences would fit in our filtered sequence landscape, representing 67.7% of all GISAID sequences.

Variant	Sequence	Changes in ACE2 affinity according to DMS
PV21	GGLWGYFMFFAGEFNYLPYGQAAGEEY	-87
PV22	GSVEGYFKFFAGEFNYLPYGQTNGEGW	nd
PV30	GDLEGYFLFFAGQFNYQPYGQTNGEGY	-3.7
PV35	GDAEGYYKFFAGEFNYLAFGQTNGEGW	nd
PV49	GDWFGYYLFYAGEFNYQAYGQAAGEGW	-11.7
PV51	GGEDGYYLFYAGEFNYQAYGHATGEEY	-162.18
PV53	GDLWGYFLFYAGEFNYVPYGQAAGEGY	-7.24

Table S5. Deep Mutational Scanning (DMS) misses some of the mutations contained in the identified PVs, since it does not consider epistatic effects. DMS identified the single-point mutations contained in some of our infective PVs (e.g. PV30, PV53). However, the mutations found in e.g. PV21 and PV51 are highly detrimental to ACE2 binding according to DMS, while we showed that these PVs are as infectious as the L strain. Single mutation effects were obtained from Bloom lab's github repository: https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS, and each PV was scored by summing up the values of each individual mutation. A negative value means weaker predicted binding for human ACE2. PV22 and PV35 could not be scored because at least one of their mutations was not observed in the DMS libraries.