Probing the biophysical constraints of SARS-CoV-2 spike N-terminal domain using deep mutational scanning

Wenhao O. Ouyang[1,*], Timothy J.C. Tan[2,*], Ruipeng Lei[1], Ge Song[3,4,5], Collin Kieffer[6], Raiees Andrabi[3,4,5], Kenneth A. Matreyek[7], Nicholas C. Wu[1,2,8,9,§]

[1] Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3] Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA

[4] IAVI Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA 92037, USA

[5] Consortium for HIV/AIDS Vaccine Development (CHAVD), The Scripps Research Institute, La Jolla, CA 92037, USA

[6] Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[7] Department of Pathology, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

[8] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[9] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

* These authors contributed equally to this work

§ Correspondence: nicwu@illinois.edu (N.C.W.)

ABSTRACT

Increasing the expression level of the SARS-CoV-2 spike (S) protein has been critical for COVID-19 vaccine development. While previous efforts largely focused on engineering the receptor-binding domain (RBD) and the S2 subunit, the N-terminal domain (NTD) has been long overlooked due to the limited understanding of its biophysical constraints. In this study, the effects of thousands of NTD single mutations on S protein expression were quantified by deep mutational scanning. Our results revealed that in terms of S protein expression, the mutational tolerability of NTD residues was inversely correlated with their proximity to the RBD and S2. We also identified NTD mutations at the interdomain interface that increased S protein expression without altering its antigenicity. Overall, this study not only advances the understanding of the biophysical constraints of the NTD, but also provides invaluable insights into S-based immunogen design.

INTRODUCTION

The emergence of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has led to the coronavirus disease 2019 (COVID-19) pandemic[1,2]. As the major antigen of SARS-CoV-2, spike (S) glycoprotein plays a critical role in facilitating virus entry[3,4]. Therefore, antibodies to SARS-CoV-2 S are often neutralizing[5,6]. SARS-CoV-2 S protein consists of an N-terminal S1 subunit, which is responsible for engaging the host receptor angiotensin-converting enzyme 2 (ACE2) via the receptor-binding domain (RBD), as well as a C-terminal S2 subunit, which mediates virus-host membrane fusion[4,7,8]. The S1 subunit also contains an N-terminal domain (NTD) in addition to the RBD[4,7]. While the RBD is generally considered to be immunodominant, the NTD is also a target of neutralizing antibodies[9–11]. Structural studies revealed the presence of an antigenic supersite on the NTD that is frequently mutated in SARS-CoV-2 variants of concern (VOCs)[12–17]. In fact, amino acid mutations and indels rapidly accumulate within the NTD during the evolution of SARS-CoV-2 in human, at least partly due to the immune selection pressure[18]. On the other hand, antibodies to NTD epitopes that are conserved across VOCs have also been identified[16,19]. Despite the importance of NTD in immune response against SARS-CoV-2, the biophysical constraints of NTD remain largely elusive.

COVID-19 vaccines, including both recombinant protein-based and mRNA-based, are proven to be highly protective against SARS-CoV-2 infection[20–23]. There is an inverse relationship between the production yield and cost of recombinant protein-based COVID-19 vaccines, such as that from Novavax, which showed promising results in phase 3 clinical trials[22], as well as others that are in earlier phases of clinical trials[24]. High protein expression level is also believed to be critical for the effectiveness of mRNA vaccines[25]. As a result, identifying mutations that increase S protein expression are crucial for optimizing COVID-19 vaccines. While most studies focused on mutating the S2 subunit as well as the RBD to increase S protein expression[7,26–29], little effort has been spent on NTD due to the lack of understanding of its biophysical properties.

Phenotypes of numerous mutations can be measured in a massively parallel manner using deep mutational scanning, which combines saturation mutagenesis and next-generation sequencing[30]. Previous studies have applied deep mutational scanning to evaluate the effects of RBD mutations on protein expression, ACE2-binding affinity, and antibody escape[31–36]. Although deep mutational scanning of the RBD provided important insights into immunogen design and SARS-CoV-2 evolution[29,31,32,35,36], similar studies on other regions of the S protein have not yet been carried out.

Here, we used deep mutational scanning to quantify the effects of thousands of NTD single mutations on S protein expression. One notable observation was that NTD residues, unlike RBD residues, showed a weak correlation between mutational tolerability and relative solvent accessibility (RSA). Instead, the mutational tolerability of NTD residues strongly correlated with their distance to RBD and S2. Residues S50 and G232 were two exceptions, in which they were proximal to S2 and RBD, respectively, and yet had a high mutational tolerability. Subsequently, we functionally characterized two mutations that increased S protein expression, namely S50Q and G232E. These results have important implications towards understanding NTD evolution and S-based immunogen design.

RESULTS

Most NTD mutations have minimal impact on S protein expression

To study how SARS-CoV-2 S protein expression is influenced by NTD mutations, we created a mutant library that contained all possible single amino acid mutations across residues 14-301 of the S protein. Each of these 288 residues was mutated with the choice of all 19 other amino acids and the stop codon, leading to a mutant library with 5,760 single amino acid mutations. The mutant library was expressed using the HEK293T landing pad cell system, such that each transfected cell stably expressed only one mutant[37,38]. Fluorescence-activated cell sorting (FACS) was then

performed using the human anti-S2 antibody CC40.8[39], with PE anti-human IgG Fc as the secondary antibody. Four separated gates were set up based on the PE signals, each covering 25% of the entire population (**Figure S1**). The frequency of each mutant among the entire population was calculated (see Materials and Methods), and a cutoff of 0.0075% was set up to filter out mutants with potentially noisy measurements. Among the 5,760 missense and nonsense mutations, 3,999 (69%) of them satisfied the frequency cutoff for downstream analysis. Of note, the design of our mutant library adopted an internal barcoding strategy that uses synonymous mutations to facilitate sequencing error correction[40]. As described previously[41], the expression score of each mutation was calculated based on their frequency in each of the four gates and normalized such that the average expression score of silent mutations was 1 and that of nonsense mutations was 0.

To evaluate the quality of the deep mutational scanning results, we assessed the expression score distributions of missense, nonsense, and silent mutations (**Figure S2A**). The difference between the expression scores of silent mutations and nonsense mutations was apparent and significant ($P = 6\times10^{-166}$), which validated the selectivity of the deep mutational scanning experiment. Interestingly, silent mutation and missense mutations had similar expression scores, although the difference is statistically significant ($P = 2\times10^{-5}$), indicating that most amino acid mutations in the NTD did not affect S protein expression. In addition, a Pearson correlation of 0.53 was obtained between the expression scores from two independent biological replicates (**Figure S2B**), demonstrating the reproducibility of the deep mutational scanning experiment.

To summarize the expression scores for individual mutations, a heatmap was generated (**Figure 1**). We noticed that high-expressing mutations were enriched within the five NTD loop regions (**Figure S3A**)[12]. High-expressing mutations were also found in residues outside of the loop

regions, such as residues S50 and G232. This observation shows that some NTD mutations can improve the expression of S protein.

Mutational tolerability has minimal correlation with solvent accessibility

While some residues were enriched in high-expression mutations (see above), others were enriched in low-expression mutations (e.g. residues D40, L84, and N234) (**Figure 1**). Consequently, we aimed to identify the biophysical determinants of mutational tolerability in terms of S protein expression. For each residue, we defined the mutational tolerability as the mean expression score of mutations. A higher mutational tolerability would indicate the enrichment of high-expressing mutations at the specified residue. In contrast, a lower mutational tolerability would indicate the enrichment of low-expressing mutations at the specified residue. A total of 243 NTD residues had six or more mutations with expression score available and were included in this analysis.

First, we investigated whether a correlation existed between the mutational tolerability and relative solvent accessibility (RSA). Since buried residues are typically important for protein folding stability, residues with a lower RSA are generally expected to have a lower mutational tolerability. For example, previous deep mutational scanning studies on the RBD have shown a decent correlation between RSA and mutational tolerability (Spearman correlation = 0.73, **Figure 2A**)[34,42]. In contrast, the mutational tolerability of NTD residues had a much weaker correlation with RSA (Spearman correlation = 0.19, **Figure 2B**). These observations indicate that the folding stability of NTD does not have a strong influence on its mutational tolerability, and hence the S expression level.

To investigate whether the mutational tolerability correlated with sequence conservation, we then analyzed the NTD sequences of 27 sarbecovirus strains, including SARS-CoV-2. Less conserved

residues tended to have a higher mutational tolerability, while more conserved residues tended to have a lower mutational tolerability, although the correlation was not strong (Spearman correlation = -0.30) (**Figure 2C**). In comparison, the correlation between sequence conservation and RSA was even weaker (Spearman correlation = -0.16, **Figure 2D**).

Mutational tolerability correlates with distance to RBD/S2

We further calculated the distance from each NTD residue to RBD/S2 of the S protein. A positive correlation was observed between the mutational tolerability and the distance to RBD/S2 (Spearman correlation = 0.56) (**Figure 2E**). In other words, the more distant an NTD residue was from the RBD/S2, the higher the mutational tolerability was. Such correlation was apparent when the mutational tolerability of each NTD residue was projected on the S protein structure **(Figure 2G)**. Consistently, the epitopes of two cross-neutralizing antibodies, namely C1717 and C1791, were significantly closer to RBD/S2 ($P \leq 5\times10^{-4}$) and had lower mutational tolerability ($P \leq 0.01$) when compared to the rapidly evolving NTD antigenic supersite (**Figure 2F and Figure S3B**)[14,16,43].

Two buried NTD mutations increase S protein expression

While NTD residues adjacent to RBD/S2 typically had a low mutational tolerability, S50 and G232 were two exceptions **(Figure 2G)**. For example, mutations S50G and G232E had a high expression score in our deep mutational scanning results. To validate this finding, we used the same landing pad system to construct HEK293T cell lines that stably expressed S50Q, G232E, and S50Q/G232E double mutant. As quantified by flow cytometry analysis (**Figure 3A and Figure S4**), the expression level of S50Q and G232E increased from wild type (WT) by 1.7-fold ($P = 0.002$) and 1.5-fold ($P = 9 \times 10^{-4}$), respectively, whereas that of S50Q/G232E increased by 2.5-fold ($P = 2 \times 10^{-6}$).

166 To probe the structural impact of S50Q and G232E, we analyzed their local environments on the
167 structure of S protein and performed structural modelling using Rosetta (**Figure 3C-D**)[44–46]. S50
168 forms a hydrogen bond with K304 and is proximal to the S2 subunit. Structural modelling showed
169 that S50Q not only is able to maintain the hydrogen bond with K304, but also strengthens the van
170 der Waals interaction between the NTD and S2 by pushing K304 towards S2 from the adjacent
171 protomer (**Figure 3C**). G232 is proximal to a positively charged region on the RBD that is featured
172 by R355 and R466 (**Figure 3D**). Structural modeling suggested that G232E could form favorable
173 electrostatic interactions with both R355 and R466. We further recombinantly expressed these
174 mutants and tested their thermostability using a thermal shift assay (**Figure 3B**). Of note, all the
175 recombinantly expressed S proteins contained K986P/V987P mutations in the S2 subunit, which
176 are known to stabilize the prefusion conformation and increase expression[26,47]. The melting
177 temperatures of WT and NTD mutants were almost identical at a T_m of 46 ºC to 46.5 ºC. These
178 observations indicate that despite both S50Q and G232E improve the interaction between NTD
179 and the rest of the S protein, they have minimal impact on the global folding stability of the S
180 protein.
181
182 **S50Q and G232E have minimal effects on the fusion activity and antigenicity**
183 To understand the functional consequences of S50Q and G232E, we further tested whether S50Q,
184 G232E, and S50Q/G232E exhibited a change in fusion activity compared the WT. A fluorescence-
185 based cell-cell fusion assay that relied on the split mNeonGreen2 (mNG2)[48] was performed (see
186 Materials and Methods, **Figure S5**). Briefly, HEK293T landing pad cells that expressed human
187 ACE2 (hACE2) and $mNG2_{1-10}$ were mixed with HEK293T landing pad cells that expressed S
188 proteins and $mNG2_{11}$. Green fluorescence due to mNG2 complementation was generated when
189 fusion between the two cell lines occurred. Fluorescence microscopy analysis showed that all
190 mutants facilitated hACE2-mediated fusion (**Figure 4A-B**). Consistently, flow cytometry analysis
191 at both 3-hour and 24-hour post-mixing indicated that none of the tested mutants diminished the

fusion activity when compared to WT (**Figure 4C-D**). At 3-hour post-mixing, both S50Q (24%, P = 0.03) and G232E (25%, P = 0.01) showed mild, yet significant, increases in fusion activity compared to WT. Similarly, at 24-hour post-mixing, S50Q (19%, P = 0.01), G232E (13%, P = 0.02), and S50Q/G232E double mutant (37%, P = 0.005) all showed an increase in fusion activity compared to WT. Such a mild increase in fusion activity may simply be attributed to the higher expression level of the mutants. Negative control cells expressing the K986P/V987P double mutant, which is known to stabilize the prefusion form of the S protein[26,47], did not show any fusion activity (**Figure 4A-D**).

We then proceeded to investigate whether S50Q, G232E, and S50Q/G232E alter the antigenicity of the S protein. The binding of three antibodies targeting different domains of the S protein were tested, namely CC12.3 (anti-RBD)[49], S2M28 (anti-NTD)[14], and COVA1-07 (anti-S2)[50]. Flow cytometry analysis showed that all three antibodies bound to the tested mutants at a similar level as WT (**Figure 5 and Figure S6**), indicating that S50Q, G232E, and S50Q/G232E did not alter the structural conformation and antigenicity of the S protein.

DISCUSSION

S protein is central to the research of SARS-CoV-2 evolution and COVID-19 vaccines[51–54]. While both the RBD and the NTD on the S protein are targets of neutralizing antibodies and involve in the antigenic drift of SARS-CoV-2[43,55–61], the NTD often receives less attention than the RBD. Using deep mutational scanning, this study shows that many NTD mutations at buried residues do not affect S protein expression. At the same time, the closer an NTD mutation is to RBD/S2, the more likely it is detrimental to S protein expression. These observations imply that for optimum S protein expression, the structural stability at the NTD-RBD and the NTD-S2 interfaces is more critical than the folding stability of the NTD. Our results also at least partly explain why the N1 to N5 loops, which contain the NTD antigenic supersite[62] and are far from the NTD-RBD/S2

218 interfaces, are highly diverse among SARS-CoV-2 variants and sarbecovirus strains. Overall, this
219 study provides crucial biophysical insights into the evolution of the NTD.
220
221 NTD mutations S50Q and G232E, which locate at the interdomain interface and increase S
222 protein expression, represent another important finding of this study. Engineering high expressing
223 S protein can lower the production cost of recombinant COVID-19 vaccine and may improve the
224 effectiveness of mRNA vaccines[25]. Similar to certain previously characterized mutations in the
225 S2[26,27], S50Q and G232E in the NTD increase the expression yield of the S protein without
226 changing its T_m. Consistently, a recent study showed that NTD mutations in BA.1 improve the
227 expression of S protein without increasing its thermostability[63]. Furthermore, S50Q and G232E
228 are not solvent exposed on the S protein surface and do not seem to alter the antigenicity of the
229 S protein. Of note, according to our deep mutational scanning data, S50Q and G232E are just
230 two of many mutations that enhance S protein expression. Therefore, although most studies on
231 S-based immunogen design focus on the mutations in the RBD and S2[7,26–29], our results suggest
232 that mutations in NTD can provide a complementary strategy.
233
234 We acknowledge that S protein expression level does not necessarily correlate with virus
235 replication fitness. For example, NTD mutations that do not affect the S protein expression may
236 be detrimental to the replication fitness of SARS-CoV-2, due to negative impact on NTD
237 functionality. While the functional importance of the NTD in natural infection remains largely
238 unclear, NTD has been proposed to facilitate virus entry by interacting with DC-SIGN, L/SIGN,
239 AXL, ASGR1, and KREMEN1[64–66]. Studies have also shown that the NTD can allosterically evade
240 antibody binding by interacting with a heme metabolite[45], as well as modulate the efficiency of
241 virus-host membrane fusion[67,68]. To fully comprehend the biophysical constraints of NTD, future
242 studies should systematically investigate how different NTD mutations affect virus replication
243 fitness.

MATERIALS AND METHODS

Construction of the NTD mutant library

SARS-CoV-2 S NTD mutant library was constructed based on the HEK293T landing pad system[37,38]. The template for constructing the NTD mutant library was a plasmid that encoded (from 5' to 3') an attB site, a codon-optimized SARS-CoV-2 S (GenBank ID: NC_045512.2) with the PRRA motif in the furin cleavage site deleted, an internal ribosome entry site (IRES), and a puromycin-resistance marker. This plasmid was used as a PCR template to generate a linearized vector and a library of mutant NTD inserts. The linearized vector was generated using 5'-TGC TCG TCT CTA CAA CTC CGC CAG CTT CAG CAC C-3' and 5'-TGC TCG TCT CTT CAC TGG CCG TCG TTT TAC AAC G-3' as primers. Inserts were generated by two separate batches of PCRs to cover the entire NTD. The first batch of PCRs consisted of 36 reactions, each containing one cassette of forward primers as well as the universal reverse primer 5'-TGC TCG TCT CGT TGT ACA GCA CGG AGT AGT CGG C-3'. Each cassette contained an equal molar ratio of eight forward primers that had the same 21 nt at the 5' end and 15 nt at the 3' end. Each primer within a cassette were also encoded with an NNK (N: A, C, G, T; K: G, T) sequence at a specified codon positions for saturation mutagenesis. In addition, each primer also carried unique silent mutations (also known as synonymous mutations) to help distinguish between sequencing errors and true mutations in downstream sequencing data analysis as described previously[40]. The forward primers, named as CassetteX_N (X: cassette number, N: primer number), are listed in **Table S1**. The second batch of PCR consisted of another 36 PCRs, each with a universal forward primer 5'-TGC TCG TCT CAG TGA ATT GTA ATA CGA CTC ACT A-3' and a unique reverse primer as listed in **Table S2**. Subsequently, 36 overlapping PCRs were performed using the universal forward and reverse primers, as well as a mixture of 10 ng each of the corresponding products from the first and second batches of PCR. The 36 overlap PCR products were then mixed at equal molar ratio to generate the final insert of the NTD mutant library. All PCRs were performed using

PrimeSTAR Max polymerase (Takara Bio) per manufacturer's instruction, followed by purification using Monarch Gel Extraction Kit (New England Biolabs). The final insert and the linearized vector were digested by BsmBI-v2 (New England Biolabs) and ligated using T4 DNA Ligase (New England Biolabs). Ligation product was purified by PureLink PCR Purification Kit (Thermo Fisher Scientific) and then transformed into MegaX Dh10B T1R cells (Thermo Fisher Scientific). At least half a million colonies were collected. Plasmid mutant library were purified from the bacteria colonies using PureLink HiPure Plasmid Midiprep Kit (Invitrogen). All primers in this study were ordered from Integrated DNA Technologies.

Construction of stable cell lines using HEK293T landing pad cells

Human embryonic kidney 293T (HEK293T) landing pad cells[37,38] were used to display the NTD mutant library for deep mutational scanning. Landing pad cells were maintained using complete growth medium consisting of Dulbecco's Modified Eagle's Medium (DMEM) (Corning), 10% v/v FBS (VWR), Pen-Strep (Gibco), non-essential amino acid (Gibco), and 2 μg/mL doxycycline. 1.2 μg of plasmid was transfected into 6×10^5 landing pad cells. For the deep mutational scanning experiment, eight transfection reactions were carried out in parallel to minimize loss of mutant diversity at the transfection step. Transfected cells were then incubated at 37 °C with 5% CO_2. After 48 hours, 10 nM AP1903 was supplemented to carry out negative selection. At 72 hours after the negative selection, positive selection antibiotic (1 μg/mL puromycin for NTD cell lines or 100 μg/mL hygromycin for hACE2 cell lines) was supplemented to the medium to carry out positive enrichment of cells with successful recombination. Constructed cell lines would remain in the complete growth medium supplemented with doxycycline and the positive selection antibiotics.

Sorting the NTD mutant library based on S protein expression level

Four T-75 flasks (Corning) that were 90% confluent with cells that carried the NTD mutant library were washed with 1× PBS, harvested with warm versene and pelleted via centrifugation at 300 ×

g for 5 mins at room temperature. Cells were then resuspended in FACS buffer (2% v/v fetal bovine serum, 5 mM EDTA in DMEM supplemented with glucose, L-glutamine and HEPES but without phenol red (Gibco)). Subsequently, cells were incubated with 5 µg/mL of CC40.8 at 4 °C with gentle shaking for 1 hour. Cells were washed once with ice-cold FACS buffer and incubated with 1 µg/mL PE anti-human IgG Fc (Biolegend) at 4 °C with gentle shaking in the dark for 1 hour. Cells were washed once and resuspended in ice-cold FACS buffer. Cells were then filtered using a 40 µm cell strainer (VWR) before cell sorting. FACS were performed using a BD FACSAria II cell sorter (BD) with a 561 nm laser and a 582/15 bandpass filter. Cells were collected into ice-cold D10 medium (Dulbecco's modified Eagle medium with 4.5 g/L glucose, 4 mM L-glutamine and 110 mg/L sodium pyruvate (Corning), supplemented with 10% v/v fetal bovine serum (VWR), 1× penicillin-streptomycin (Gibco), 1× non-essential amino acids (Gibco)) and binned into no (bin 0), low (bin 1), medium (bin 2) and high (bin 3) expression according to PE signal, where each bin contains 25% of the singlet population (**Figure S1**). A biological replicate of the deep mutational scanning experiment was performed, starting from the transfection step.

Next-generation sequencing of the NTD mutant library

Sorted cells from each bin were pelleted at 300 × g, 4 ºC for 15 mins and then resuspended in 200 µL PBS (Corning). Genomic DNA extraction was then performed using DNA Blood and Tissue Kit (QIAGEN) according to the manufacturer's instructions with a modification: cells were incubated at 56 °C for 30 min instead of 10 min. The NTD mutant library was amplified from the genomic DNA in two non-overlapping fragments using KOD DNA polymerase (MilliporeSigma) per manufacturer's instruction with the following two primer sets, respectively (also see **Table S3**). Set 1: 5'- CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CTC TGC TGC CTC TGG TGT CCA GC-3' (NTD-DMS-recover-1F) and 5'-GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CTG TTG GCG CTG CTG TAC ACC CG-3' (NTD-DMS-recover-1R)

Set 2: 5'-CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CTA GCT GGA TGG AAA GCG AGT TC-3' (NTD-DMS-recover-2F) and 5'-GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CTC ACG GTG AAG GAC TTC AGG GT-3' (NTD-DMS-recover-2R)

A second round of PCR was carried out to add the adapter sequence and index to the amplicons as described previously[69]. The final PCR products were submitted for next-generation sequencing using Illumina MiSeq PE300.

Analysis of next-generation sequencing data

Next-generation sequencing data were obtained in FASTQ format. Forward and reverse reads of each paired-end read were merged by PEAR[70]. The merged reads were parsed by SeqIO module in BioPython[71]. Primer sequences were trimmed from the merged reads. Trimmed reads with lengths inconsistent with the expected length were discarded. The trimmed reads were then translated to amino acid sequences, with sequencing error correction performed at the same time as previously described[40]. Amino acid mutations were called by comparing the translated reads to the WT amino acid sequence. Frequency (*F*) of a mutant *i* at position *s* within bin *n* of replicate *k* was computed for each replicate as follows:

$$F_{i,s,n,k} = \frac{readcount_{i,s,n,k}+1}{\sum_s \sum_i (readcount_{i,s,n,k}+1)} \quad (1)$$

A pseudocount of 1 was added to the read counts of each mutant to avoid division by zero in subsequent steps. We then calculated the total frequency (F_{total}) of mutant *i* at position *s* as follows:

$$F_{total,i,s} = \frac{\sum_{k=1}^{2} \sum_{n=0}^{3} F_{i,s,n,k}}{8} \quad (2)$$

Mutants with a F_{total} of equal or greater than 0.0075% were selected for downstream analysis. Subsequently, the weighted average (*W*) of each mutant among 4 bins (bin 0 to bin 3) in each replicate was computed as described previously[41]:

344 $$W_{i,s,k} = \frac{F_{i,s,0,k}\times 0.25 + F_{i,s,1,k}\times 0.5 + F_{i,s,2,k}\times 0.75 + F_{i,s,3,k}\times 1}{\sum_{n=0}^{3} F_{i,s,n,k}} \quad (3)$$

345 Selected mutants were then categorized based on the mutation types (missense, nonsense, and
346 silent). The mean value of weighted average for nonsense as well as silent mutations were
347 calculated. Expression score (*ES*) of a mutant *i* at position s of replicate *k* was calculated as
348 described previously[41]:

349 $$ES_{i,s,k} = \frac{W_{i,s,k} - \overline{W_{nonsense,k}}}{\overline{W_{silent,k}} - \overline{W_{nonsense,k}}} \quad (4)$$

350 Final expression score of a mutant *i* at position *s* was calculated by taking the average of the
351 expression scores between replicates. Mutational tolerability of position *s* was then calculated by
352 taking the average of the expression scores of all mutants at that position:

353 $$mutational\ tolerability_s = \frac{\sum_{i\in s} ES_{i,s}}{\sum i\in s} \quad (5)$$

354

355 **Structural analysis of deep mutational scanning results**

356 DSSP[72,73] was used to calculate the solvent exposure surface area (SASA) of each residue in
357 NTD and RBD on the S trimer (PDB 6ZGE)[44]. Deep mutational scanning result of RBD was
358 extracted from a previous study[42]. Relative solvent accessibility (RSA) was computed by dividing
359 the SASA by the theoretical maximum allowed solvent accessibility of the corresponding amino
360 acid[74].

361

362 Each NTD residue's distance to RBD/S2 was calculated based on the S trimer structure (PDB
363 6ZGE)[44] with the NTD replaced by the high resolution crystal structure (PDB 7B62)[45]. For each
364 NTD residue, the distances to all RBD and S2 residues were measured. The shortest distance
365 was then recorded as the "distance to RBD/S2". Residue-residue distance was defined as the
366 distance between the centroid coordinates of two residues.

To visualize the mutational tolerability of each NTD residue, the crystal structure of SARS-CoV-2 S protein NTD (PDB 7B62) was used[45]. The NTD crystal structure was then aligned with the S trimer to generate the figures (PDB 6ZGE)[44].

NTD sequence conservation analysis

The sequence conservation analysis of NTD was based on 27 sarbecovirus strains (**Table S6**)[1,75–79]. S sequences of these stains were retrieved from GenBank and Global Initiative for Sharing Avian Influenza Data (GISAID)[80]. Their NTD sequences were then identified using tBlastn search using the amino acid sequence of SARS-CoV-2 Hu-1 NTD (Gene ID: 43740568) as the query sequence. The BlastXML output of the tBlastn was then parsed and used as the input for multiple sequence alignment using MAFFT[81,82]. For each residue position, sequence conservation was defined as the proportion of strains that contains the same amino acid variant as SARS-CoV-2 Hu-1.

Rosetta-based mutagenesis

The structure of the spike protein was obtained from Protein Data Bank (PDB 6ZGE)[44]. Water molecules and N-acetyl-D-glucosamine were removed using PyMOL (Schrödinger). Then, the amino acids were renumbered using pdb-tools[83]. Fixed backbone point-mutagenesis for S50Q and G232E was performed using the 'fixbb' application in Rosetta (RosettaCommons). One-hundred poses were generated for each mutagenesis. Using the lowest-scoring structure from fixed backbone mutagenesis as input, a constraint file was obtained using the minimize_with_cst application in Rosetta. Fast relax was then performed via the 'relax' application in Rosetta[46] with the corresponding constraint file. The lowest-scoring structure out of eight was then used for structural analysis. Code and source files for structural modelling are available in https://github.com/nicwulab/SARS-CoV-2_NTD_DMS/tree/main/rosetta.

393

394 **Split mNeonGreen2-based cell-cell fusion assay**

395 Human ACE2 (hACE2) construct was constructed in a previous study[38]. A split mNeonGreen2
396 (mNG2) reporter system was integrated into the S plasmid (see above) and the hACE2 plasmid[48].
397 Specifically, a gene fragment that encoded (from 5' to 3') a GCN4 leucine zipper, a GS linker,
398 $mNG2_{1-10}$, and a 2A self-cleaving peptide was inserted into the hACE2 plasmid between the IRES
399 and the hygromycin resistance marker. Similarly, a gene fragment that encodes (from 5' to 3') a
400 GCN4 leucine zipper, a GS linker, $mNG2_{11}$, and a 2A self-cleaving peptide was inserted into the
401 S plasmid between the IRES and the puromycin resistance marker. Each plasmid construct was
402 transfected and recombined into HEK293T landing pad cells per steps described above.

403

404 Once the stable cell lines were created, 5×10^5 landing pad cells expressing hACE2 with $mNG2_{1-10}$
405 were seeded in 6-well plates (Fisher Scientific). The cells were then then incubated at 37 °C
406 with 5% CO_2 for 15 mins to allow seeding. Subsequently, 5×10^5 landing pad cells expressing the
407 S with $mNG2_{11}$ were then added dropwise to the seeded hACE2 cells. Both cells were filtered
408 through 40 μm cell strainer (VWR) prior to seeding. At 3-hour and 24-hour post-mixing, fusion
409 events in each well were qualitatively assessed with an ECHO Revolve epifluorescence
410 microscope (ECHO) in inverted format. Overlayed images were captured on white light and FITC
411 filter channels using an UPlanFL N 10X/0.30NA objective (Olympus) with identical light intensity
412 and exposure settings for all conditions. Cells in each well were then collected using 0.5 mM
413 EDTA, pelleted via centrifugation at 300 × *g* for 5 mins at room temperature, and resuspended in
414 the FACS buffer. LSRII flow cytometry (BD) was used to quantify the fusion events of each sample.
415 Negative controls were measured first to set up proper gating strategies (**Figure S5**). Then, the
416 flow cytometry analysis was performed on 10^5 live cells for each sample. Data were analyzed

using FCS Express 6 software (De Novo Software). The percentage of mNG2 positive population of each sample were used for normalization (**Table S4**).

Flow cytometry analysis for the protein expression assay and antibody binding assay

Approximately 1×10^6 cells that carried the selected SARS-CoV-2 S NTD mutant were washed with 1× PBS, harvested with warm versene and pelleted via centrifugation at 300 × g for 5 mins at room temperature. The cells were resuspended in the FACS buffer. Subsequently, cells were incubated with 5 μg/mL of the selected antibodies at 4 °C with gentle shaking for 1 hour. Cells were then washed once with ice-cold FACS buffer and incubated with 2 μg/mL PE anti-human IgG Fc (BioLegend) at 4 °C with gentle shaking in the dark for 1 hour. Cells were washed once, pelleted via centrifugation at 300 × g for 5 mins at room temperature, and resuspended in ice-cold FACS buffer. LSRII flow cytometry (BD) was used to measure the PE signal of each sample. Negative controls were measured first to set up proper gating strategies (**Figure S4 and S6**). Then, the flow cytometry analysis was performed on 10^5 singlets for each sample. Data were analyzed using FCS Express 6 software (De Novo Software).

Normalization of the expression assay results

The mean fluorescence intensity (MFI) of the entire population was recorded for each sample, followed by the normalization as previously described[42]. For a given sample *i*, the following equation was used to compute the normalized expression (*NE*):

$$NE = \frac{MFI_i - MFI_{negative\ control}}{MFI_{wildtype} - MFI_{negative\ control}} \quad (6)$$

Normalizations were performed for each sample within a given biological replicate (**Table S4**).

Recombinant expression and purification of soluble S protein

SARS-CoV-2 S ectodomain with the PRRA motif in the furin cleavage site deleted and mutations K986P/V987P, which are known to stabilize the prefusion conformation and increase expression[26,47], was cloned into a phCMV3 vector. The S ectodomain construct contained a trimerization domain and a 6×His-tag at the C-terminal. Expi293F cells (Gibco), which were maintained using Expi293 expression medium (Gibco), were used to express soluble S protein. Briefly, 20 μg of the plasmid was transfected into 20 mL of Expi293F cells at 3×10^6 cells mL^{-1} using ExpiFectamine 293 Transfection Kit (Thermo Fisher Scientific) following the manufacturer's instructions. Transfected cells were then incubated at 37 °C, 8% CO_2 and shaking at 125 rpm for 6 days. Cell cultures were then harvested and centrifuged at 4000 × *g* at 4 ºC for 15 mins. The supernatant was clarified using a 0.22 μm polyethersulfone filter (Millipore). S protein in the clarified supernatant was then purified using Nickel Sepharose Excel resin (Cytiva), with 20 mM imidazole in PBS as wash buffer, and 300 mM imidazole in PBS as elution buffer. Three rounds of 2 mL elutions were performed. The eluted protein was then concentrated and analyzed by SDS-PAGE reducing gels (Bio-Rad) (**Figure S7A**). Concentrated protein solution was further purified using Superdex 200 XK 16/100 size exclusion column (Cytiva) in 20 mM Tris-HCl pH 8.0 and 150 mM NaCl (**Figure S7B**). Selected elution fractions were combined and concentrated. Final protein concentration was measured using NanoDrop One (Thermo Fisher Scientific).

Protein thermostability assay

5 μg of purified protein was mixed with 5× SYPRO orange (Thermo Fisher Scientific) in 20 mM Tris-HCl pH 8.0, 150 mM NaCl at a final volume of 25 μL. The sample mixture was then transferred into an optically clear PCR tube (VWR). SYPRO orange fluorescence data in relative fluorescence unit (RFU) was collected from 10 °C to 95 °C using CFX Connect Real-Time PCR Detection System (Bio-Rad). The temperature corresponding to the lowest point of the first derivative,

465 −d(RFU)/dT, was defined as the melting temperature (T_m). Data were analyzed using OriginPro
466 2020b (Origin Lab). Raw data are shown in **Table S5**.
467
468 **Code availability**
469 Custom python and R scripts for data analysis and plotting in this study have been deposited to
470 https://github.com/nicwulab/SARS-CoV-2_NTD_DMS.
471
472 **Data availability**
473 Raw sequencing data have been submitted to the NIH Short Read Archive under accession
474 number: BioProject PRJNA792013. Biological materials including plasmids, constructed NTD
475 mutant library as well as individual HEK293T landing pad cell lines are available by contacting
476 the corresponding author (N.C.W.). Source data are provided with this paper.
477
478 **ACKNOWLEDGEMENT**
479 We thank Meng Yuan, Huibin Lv, and Qi Wen Teo for helpful discussion and the Roy J. Carver
480 Biotechnology Center at the University of Illinois at Urbana-Champaign for assistance with
481 fluorescence-activated cell sorting and next-generation sequencing. This work was supported by
482 National Institutes of Health (NIH) R00 AI139445 (N.C.W.), DP2 AT011966 (N.C.W.), R01
483 AI167910 (N.C.W.), the Michelson Prizes for Human Immunology and Vaccine Research
484 (N.C.W.).
485
486 **AUTHOR CONTRIBUTIONS**
487 W.O.O, T.J.C.T., and N.C.W. conceived and designed the study. G.S. and R.A. provided the
488 CC40.8 antibody. K.A.M. provided the HEK293T landing pad cell line and assisted in experimental
489 design. W.O.O and T.J.C.T. performed the deep mutational scanning and flow cytometry
490 experiments. T.J.C.T. performed structural modelling using Rosetta. W.O.O, T.J.C.T., and R.L.

491 expressed and purified the recombinant proteins. W.O.O, T.J.C.T., and C.K. performed the
492 microscopy analysis. W.O.O, T.J.C.T., and N.C.W. performed data analysis. W.O.O., T.J.C.T.
493 and N.C.W. wrote the paper and all authors reviewed and/or edited the paper.
494

495 **COMPETING INTERESTS**

496 The authors declare no competing interests.

FIGURES

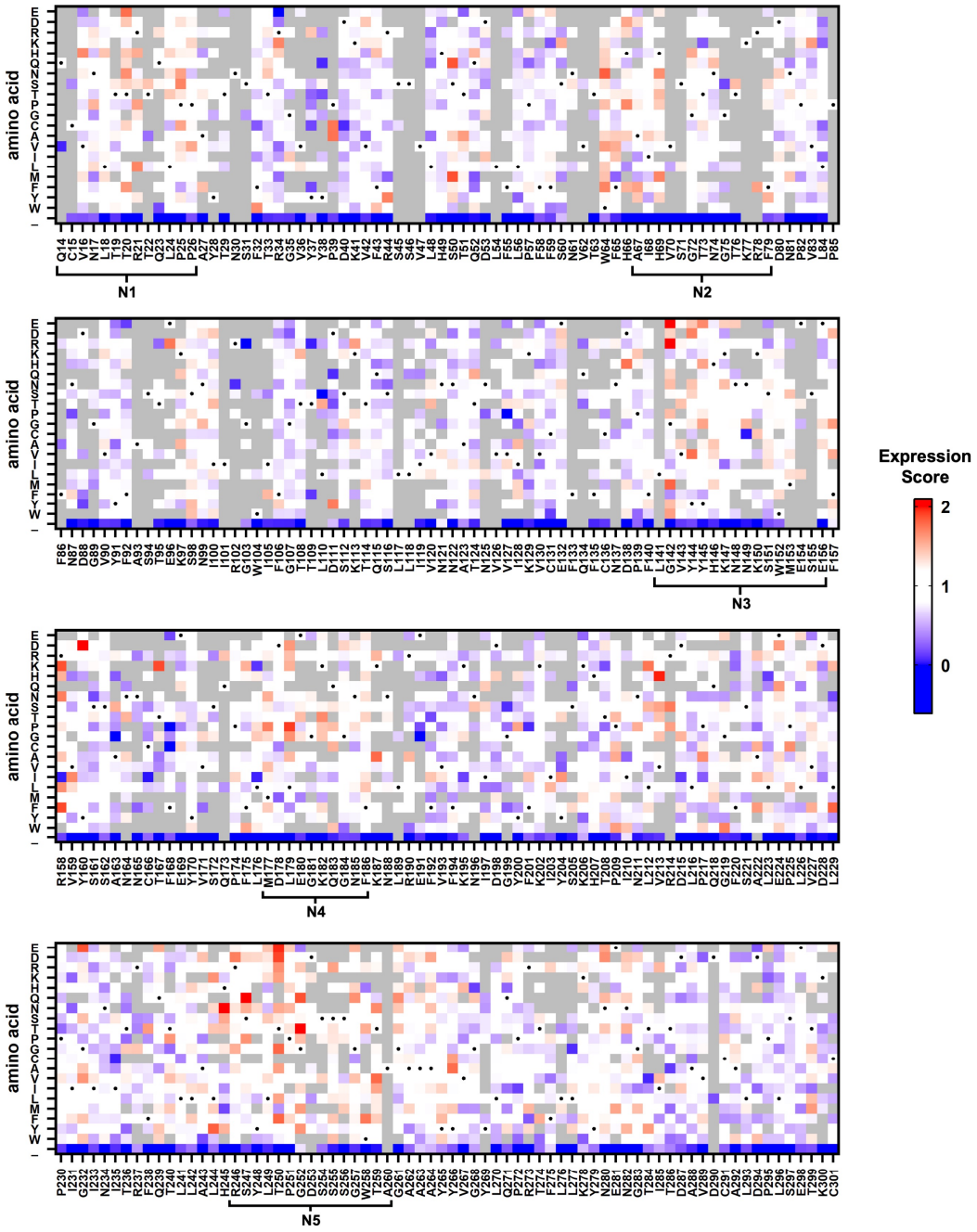
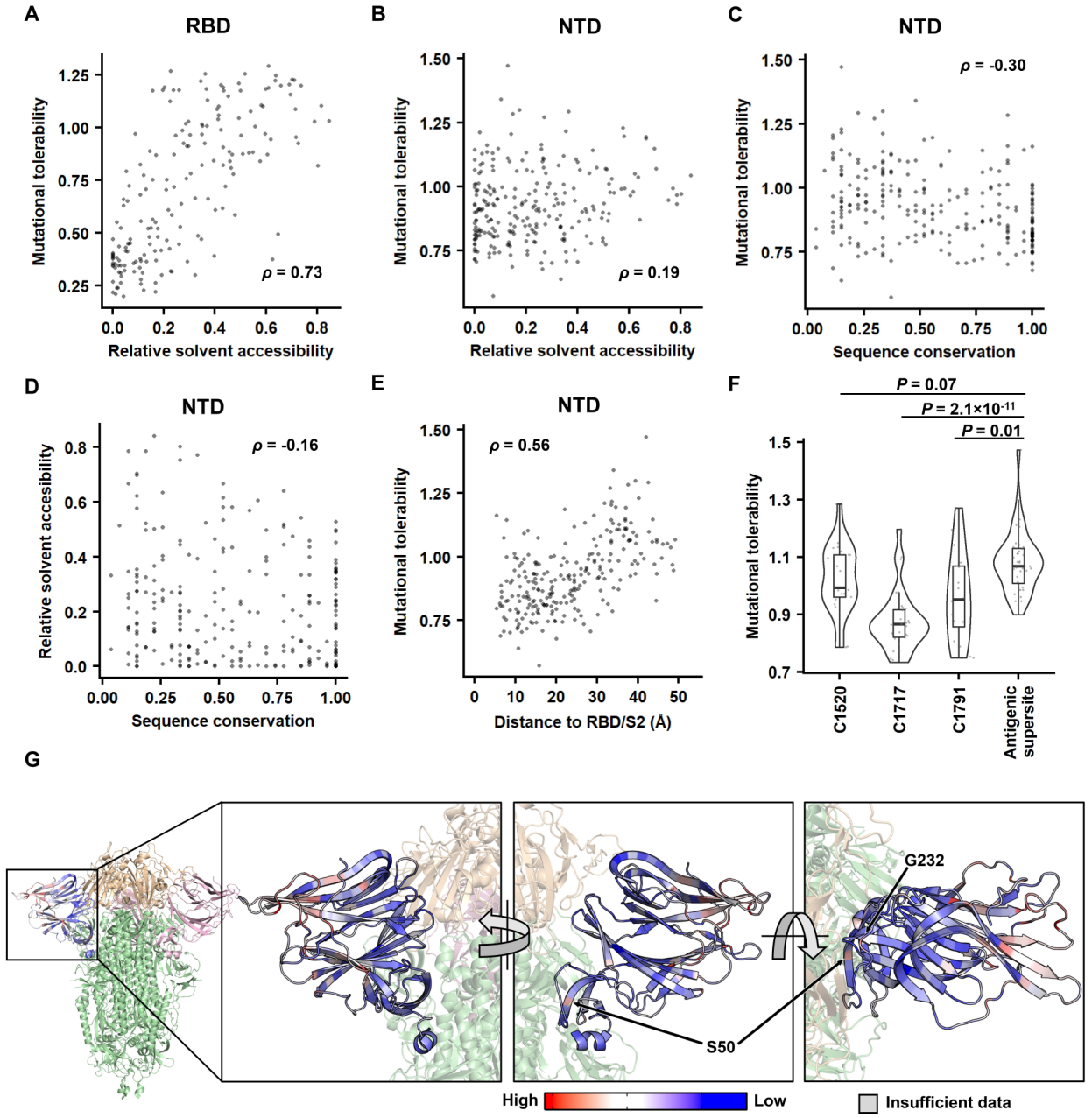


Figure 1. Effects of NTD single mutations on S protein expression. The expression scores of individual NTD mutations are shown as a heatmap. X-axis represents the residue position. Y-axis represents different amino acids as well as the stop codon (_). Amino acids corresponding

to the WT sequence are indicated by the black dots. Mutations with a total frequency of <0.0075% were excluded from the analysis and shown in grey. Regions corresponding to the N1-N5 loops were defined as previously described[12].



505

506 **Figure 2. identifying the biophysical determinants of mutational tolerability. (A-B)** The
507 relationship between relative solvent accessibility (RSA) and the mutational tolerability is shown
508 for **(A)** RBD and **(B)** NTD. The deep mutational scanning data on RBD expression was from a
509 previous study[42]. **(C-D)** The relationship between sequence conservation among 27 sarbecovirus
510 strains (**Table S6**) and **(C)** the mutational tolerability, or **(D)** RSA of each NTD residue is shown.

511 **(E)** The relationship between the distance to RBD/S2 and the mutational tolerability of each NTD
512 residue is shown. **(A-E)** Each datapoint represents one residue. The Spearman's rank correlation
513 coefficient (ρ) is indicated. **(F)** The mutational tolerability of residues within the cross-neutralizing
514 NTD antibody epitopes (C1520, C1717, C1791)[16] is compared to that within the antigenic
515 supersite[14] using a violin plot. Each datapoint represents one residue. P-values were computed
516 by two-tailed t-test. **(G)** The mutational tolerability of each NTD residue is projected on one NTD
517 of the S trimer structure (PDB 6ZGE[44] and PDB 7B62[45]). Red indicates residues with higher
518 mutational tolerability, while blue indicates residues with lower mutational tolerability. Residues
519 with insufficient data to calculate mutational tolerability are colored in grey. Two residues of
520 interests, namely S50 and G232, are indicated. RBDs are colored in wheat, the two other NTDs
521 are in pink, and the rest of the S1 and S2 subunits are in green.

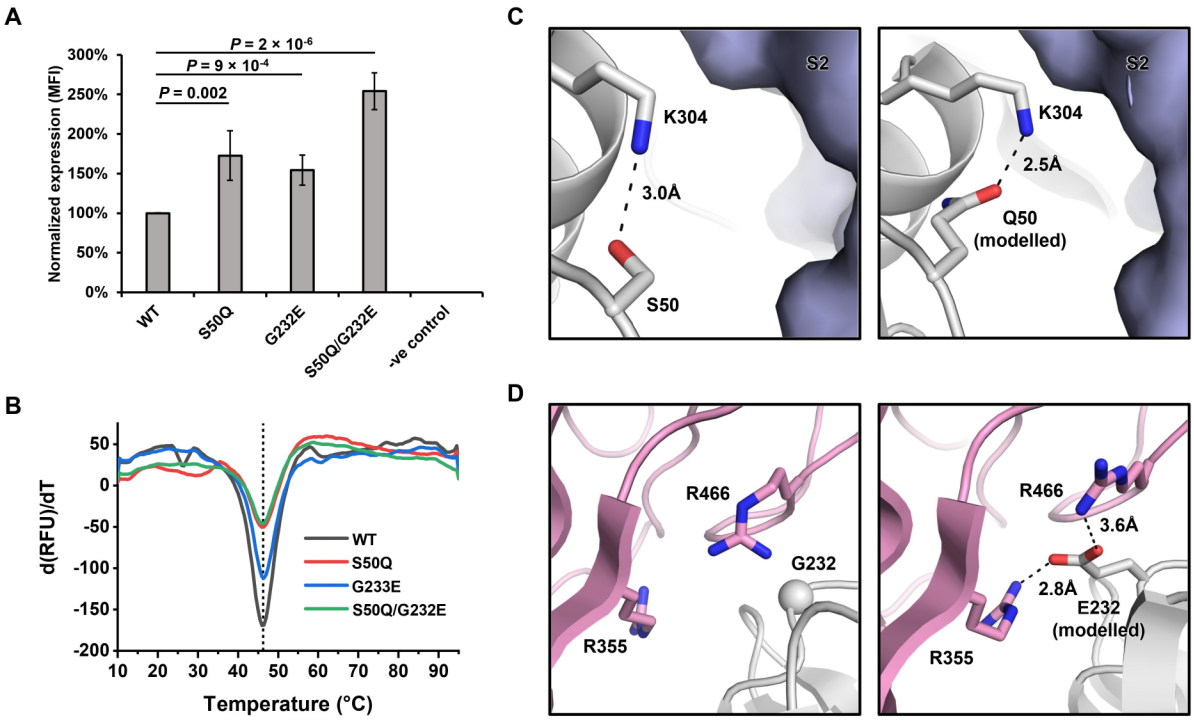
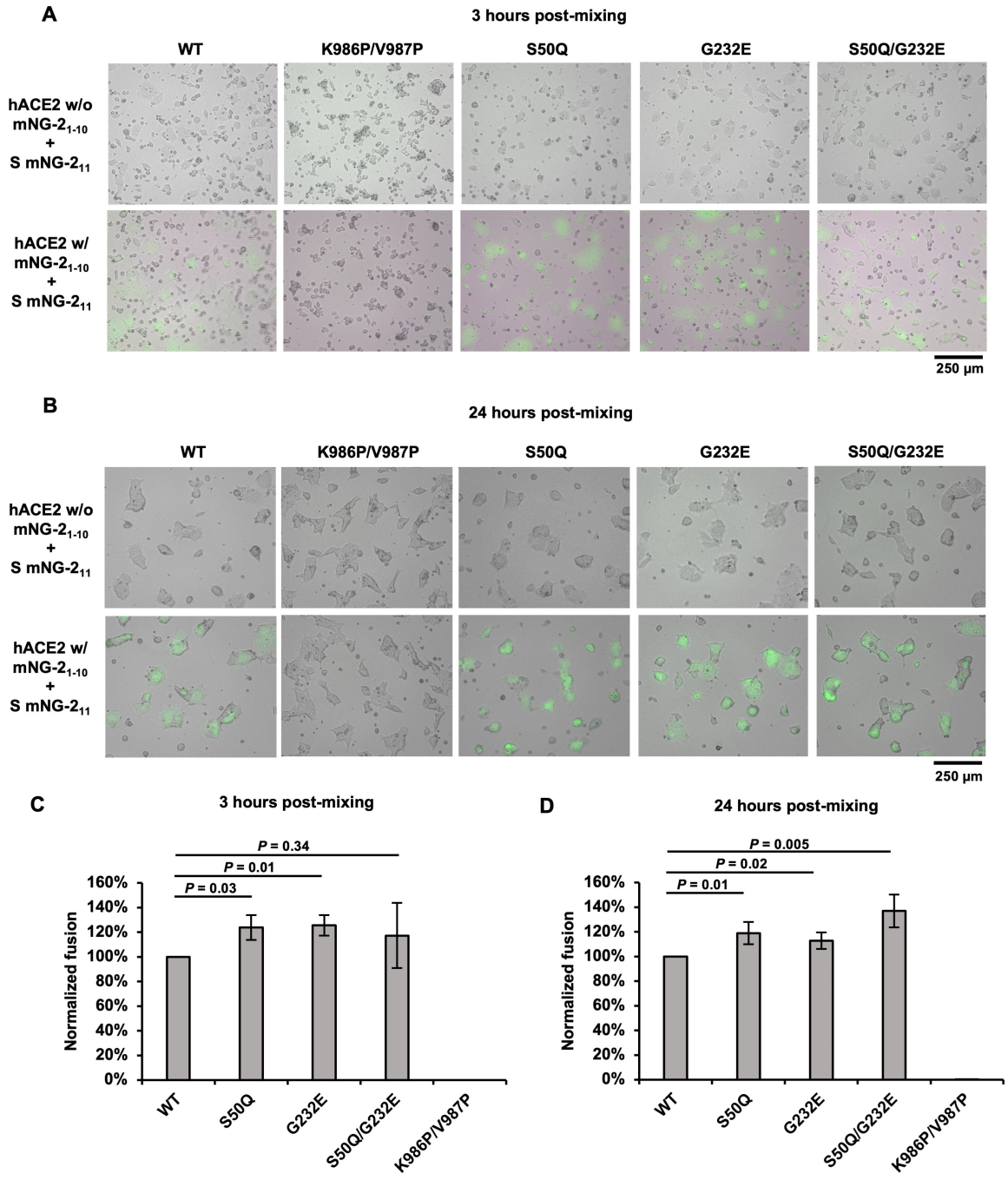


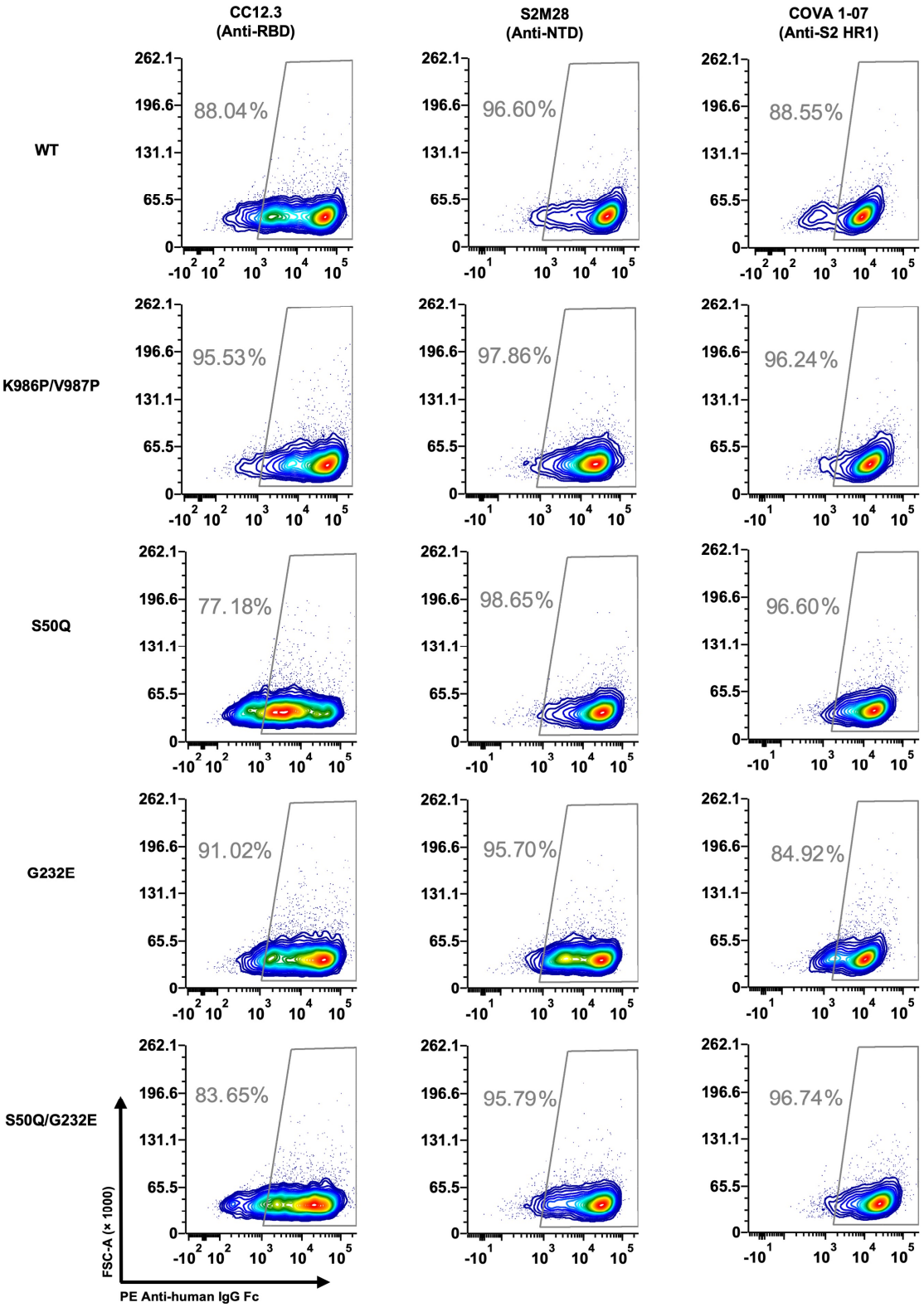
Figure 3. S50Q and G232E at the interdomain interface increase S protein expression. (A) Cell surface S protein expression of WT and NTD mutants was quantified using flow cytometry analysis with CC40.8 as the primary antibody. Untransfected HEK293T landing pad cells were used as a negative control (-ve control). S protein expression level was defined as the mean fluorescence intensity (MFI) of the positive gated population. S protein expression level was normalized to WT. The error bar indicates the standard deviation of six independent experiments. P-values were computed by two-tailed t-test. **(B)** Thermostability of WT S protein and selected NTD mutants was measured using differential scanning fluorimetry. The black vertical dotted line indicates the melting temperature of WT (T_m = 46.2 ºC). **(C-D)** Rosetta-based structural modelling of **(C)** S50Q and **(D)** G232E was performed using the structure of S protein (PDB 6ZGE)[45]. The three protomers of the S protein are colored in white, light blue, and pink. Potential interactions are represented by black dashed lines with distance labeled.



535

536 **Figure 4. S50Q, G232E, and S50Q/G232E do not diminish fusion activity. (A-B)** Fluorescence
537 microscopy analysis of the fusion events at **(A)** 3-hour and **(B)** 24-hour post-mixing of cells
538 expressing S and mNG2$_{11}$ (S cells) and cells expressing hACE2 and mNG2$_{1-10}$ (hACE2 cells).
539 Cells with green fluorescence signals are the fused cells. Scale bars are shown at the bottom

right corner. **(C-D)** Fusion activity of WT and selected NTD mutants at **(C)** 3-hour and **(D)** 24-hour post-mixing was quantified using flow cytometry analysis. Fusion activity was normalized to WT. The error bar indicates the standard deviation of at least four independent experiments. P-values were computed by two-tailed t-test.



544

545 **Figure 5. S50Q, G232E, and S50Q/G232E do not alter antibody binding.** Three antibodies

546 targeting different domains on the S were tested for binding to cells expressing WT, K986P/V987P,
547 S50Q, G232E, or S50Q/G232E S protein. Binding was measured by flow cytometry analysis.
548 Gating was set up using untransfected HEK293T landing pad cells, which served as a negative
549 control (**Figure S6**).

REFERENCES

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

2. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).

3. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**, 562–569 (2020).

4. Walls, A. C. *et al.* Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281-292.e6 (2020).

5. Strohl, W. R. *et al.* Passive immunotherapy against SARS-CoV-2: from plasma-based therapy to single potent antibodies in the race to stay ahead of the variants. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.* (2022) doi:10.1007/s40259-022-00529-7.

6. Yuan, M., Liu, H., Wu, N. C. & Wilson, I. A. Recognition of the SARS-CoV-2 receptor binding domain by neutralizing antibodies. *Biochem. Biophys. Res. Commun.* **538**, 192–203 (2021).

7. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).

8. Huang, Y., Yang, C., Xu, X., Xu, W. & Liu, S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).

9. Premkumar, L. *et al.* The receptor-binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci. Immunol.* **5**, eabc8413 (2020).

10. Piccoli, L. *et al.* Mapping neutralizing and immunodominant sites on the sars-cov-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* **183**, 1024-1042.e21 (2020).

11. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).

12. 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).

13. Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450–456 (2020).

14. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e16 (2021).

15. Suryadevara, N. *et al.* Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* **184**, 2316-2331.e15 (2021).

16. Wang, Z. *et al.* Analysis of memory B cells identifies conserved neutralizing epitopes on the N-terminal domain of variant SARS-Cov-2 spike proteins. *Immunity* S1074-7613(22)00174–1 (2022) doi:10.1016/j.immuni.2022.04.003.

17. Cerutti, G. *et al.* Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe* **29**, 819-833.e7 (2021).

18. Wolf, K. A., Kwan, J. C. & Kamil, J. P. Structural dynamics and molecular evolution of the SARS-CoV-2 spike protein. *mBio* **13**, e02030-21 (2022).

19. Lin, W.-S., Chen, I.-C., Chen, H.-C., Lee, Y.-C. & Wu, S.-C. Glycan masking of epitopes in the ntd and rbd of the spike protein elicits broadly neutralizing antibodies against SARS-CoV-2 variants. *Front. Immunol.* **12**, (2021).

20. Tenforde, M. W. *et al.* Effectiveness of mRNA Vaccination in Preventing COVID-19-Associated Invasive Mechanical Ventilation and Death - United States, March 2021-January 2022. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 459–465 (2022).

21. Tartof, S. Y. *et al.* Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *The Lancet* **398**, 1407–1416 (2021).

22. Heath, P. T. *et al.* Safety and efficacy of NVX-CoV2373 COVID-19 vaccine. *N. Engl. J. Med.* **385**, 1172–1183 (2021).

23. Sridhar, S. *et al.* Safety and immunogenicity of an AS03-adjuvanted SARS-CoV-2 recombinant protein vaccine (CoV2 preS dTM) in healthy adults: interim findings from a phase 2, randomised, dose-finding, multicentre study. *Lancet Infect. Dis.* **22**, 636–648 (2022).

24. Pollet, J., Chen, W.-H. & Strych, U. Recombinant protein vaccines, a proven approach against coronavirus pandemics. *Adv. Drug Deliv. Rev.* **170**, 71–82 (2021).

25. Schlake, T., Thess, A., Fotin-Mleczek, M. & Kallen, K.-J. Developing mRNA-vaccine technologies. *RNA Biol.* **9**, 1319–1330 (2012).

26. Juraszek, J. *et al.* Stabilizing the closed SARS-CoV-2 spike trimer. *Nat. Commun.* **12**, 244 (2021).

27. Hsieh, C.-L. *et al.* Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501–1505 (2020).

28. Olmedillas, E. *et al.* Structure-based design of a highly stable, covalently-linked SARS-CoV-2 spike trimer with improved structural properties and immunogenicity. *bioRxiv* 2021.05.06.441046 (2021) doi:10.1101/2021.05.06.441046.

29. Ellis, D. *et al.* Stabilization of the SARS-CoV-2 spike receptor-binding domain using deep mutational scanning and structure-based design. *Front. Immunol.* **12**, (2021).

623 30. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat.*
624 *Methods* **11**, 801–807 (2014).

625 31. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by
626 different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021).

627 32. Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to
628 the SARS-CoV-2 receptor-binding domain. *Virus Evol.* **8**, veac021 (2022).

629 33. Starr, T. N. *et al.* SARS-CoV-2 RBD antibodies that maximize breadth and resistance to
630 escape. *Nature* **597**, 97–102 (2021).

631 34. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain
632 reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295-1310.e20 (2020).

633 35. Starr, T. N., Greaney, A. J., Dingens, A. S. & Bloom, J. D. Complete map of SARS-CoV-
634 2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-
635 CoV016. *Cell Rep. Med.* **2**, 100255 (2021).

636 36. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to
637 treat COVID-19. *Science* **371**, 850–854 (2021).

638 37. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of
639 large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).

640 38. Shukla, N., Roelle, S. M., Suzart, V. G., Bruchez, A. M. & Matreyek, K. A. Mutants of
641 human ACE2 differentially promote SARS-CoV and SARS-CoV-2 spike mediated infection.
642 *PLoS Pathog.* **17**, e1009715 (2021).

643 39. Zhou, P. *et al.* A human antibody reveals a conserved site on beta-coronavirus spike
644 proteins and confers protection against SARS-CoV-2 infection. *Sci. Transl. Med.* **14**, eabi9215.

645 40. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise
646 epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).

647 41. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively
648 parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).

649 42. Chan, K. K., Tan, T. J. C., Narayanan, K. K. & Procko, E. An engineered decoy receptor
650 for SARS-CoV-2 broadly binds protein S sequence variants. *Sci. Adv.* **7**, eabf1738 (2021).

651 43. McCallum, M. *et al.* Molecular basis of immune evasion by the Delta and Kappa SARS-
652 CoV-2 variants. *Science* **374**, 1621–1626 (2021).

653 44. Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform
654 on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020).

655 45. Rosa, A. *et al.* SARS-CoV-2 can recruit a heme metabolite to evade antibody immunity.
656 *Sci. Adv.* **7**, eabg7607 (2021).

657 46. Kellogg, E. H., Leaver-Fay, A. & Baker, D. Role of conformational sampling in computing
658 mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011).

659 47. Sanders, R. W. & Moore, J. P. Virus vaccines: proteins prefer prolines. *Cell Host*
660 *Microbe* **29**, 327–333 (2021).

661 48. Feng, S. *et al.* Improved split fluorescent proteins for endogenous protein labeling. *Nat.*
662 *Commun.* **8**, 370 (2017).

663 49. Yuan, M. *et al.* Structural basis of a shared antibody response to SARS-CoV-2. *Science*
664 **369**, 1119–1123 (2020).

665 50. Claireaux, M. *et al.* A public antibody class recognizes a novel S2 epitope exposed on
666 open conformations of SARS-CoV-2 spike. *bioRxiv* 2021.12.01.470767 (2021)
667 doi:10.1101/2021.12.01.470767.

668 51. Yewdell, J. W. Antigenic drift: Understanding COVID-19. *Immunity* **54**, 2681–2687
669 (2021).

670 52. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat.*
671 *Rev. Microbiol.* **19**, 409–424 (2021).

672 53. Yuan, M. *et al.* Structural and functional ramifications of antigenic drift in recent SARS-
673 CoV-2 variants. *Science* **373**, 818–823 (2021).

674 54. Telenti, A. *et al.* After the pandemic: perspectives on the future trajectory of COVID-19.
675 *Nature* **596**, 495–504 (2021).

676 55. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7.
677 *Nature* **593**, 130–135 (2021).

678 56. Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal
679 and serum-derived polyclonal antibodies. *Nat. Med.* **27**, 717–726 (2021).

680 57. McCallum, M. *et al.* SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of
681 concern. *Science* **373**, 648–654 (2021).

682 58. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune
683 evasion. *Nature* **599**, 114–119 (2021).

684 59. Cameroni, E. *et al.* Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron
685 antigenic shift. *Nature* **602**, 664–670 (2022).

686 60. Singh, Y. *et al.* N-terminal domain of SARS CoV-2 spike protein mutation associated
687 reduction in effectivity of neutralizing antibody with vaccinated individuals. *J. Med. Virol.* **93**,
688 5726–5728 (2021).

689 61. Gobeil, S. M.-C. *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure,
690 conformation, and antigenicity. *Science* **373**, eabi6226 (2021).

691 62. Lok, S.-M. An NTD supersite of attack. *Cell Host Microbe* **29**, 744–746 (2021).

692 63. Javanmardi, K. *et al.* Antibody escape and cryptic cross-domain stabilization in the
693 SARS-CoV-2 Omicron spike protein. *bioRxiv* 2022.04.18.488614 (2022)
694 doi:10.1101/2022.04.18.488614.

64. Wang, S. *et al.* AXL is a candidate receptor for SARS-CoV-2 that promotes infection of pulmonary and bronchial epithelial cells. *Cell Res.* **31**, 126–140 (2021).

65. Gu, Y. *et al.* Receptome profiling identifies KREMEN1 and ASGR1 as alternative functional receptors of SARS-CoV-2. *Cell Res.* **32**, 24–37 (2022).

66. Soh, W. T. *et al.* The N-terminal domain of spike glycoprotein mediates SARS-CoV-2 infection by associating with L-SIGN and DC-SIGN. *bioRxiv* 2020.11.05.369264 (2020) doi:10.1101/2020.11.05.369264.

67. Qing, E. *et al.* Inter-domain communication in SARS-CoV-2 spike proteins controls protease-triggered cell entry. *Cell Rep.* **39**, 110786 (2022).

68. Qing, E. *et al.* Dynamics of SARS-CoV-2 spike proteins in cell entry: control elements in the amino-terminal domains. *mBio* **12**, e01590-21 (2021).

69. Wang, Y., Lei, R., Nourmohammad, A. & Wu, N. C. Antigenic evolution of human influenza H3N2 neuraminidase is constrained by charge balancing. *eLife* **10**, e72516 (2021).

70. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

71. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

72. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

73. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39**, D411-419 (2011).

74. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PloS One* **8**, e80635 (2013).

75. Wu, N. C. *et al.* A natural mutation between sars-cov-2 and sars-cov determines neutralization by a cross-reactive antibody. *PLOS Pathog.* **16**, e1009089 (2020).

76. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336 (2022).

77. Zhou, H. *et al.* A novel bat coronavirus closely related to sars-cov-2 contains natural insertions at the s1/s2 cleavage site of the spike protein. *Curr. Biol.* **30**, 2196-2203.e3 (2020).

78. Lam, T. T.-Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).

79. Tao, Y. & Tong, S. Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from kenyan bats. *Microbiol. Resour. Announc.* **8**, e00548-19 (2019).

80. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **22**, 30494 (2017).

731 81. Ream, D. & Kiss, A. J. NCBI / GenBank BLAST output XML parser tool.
732 https://www.semanticscholar.org/paper/NCBI-%2F-GenBank-BLAST-Output-XML-Parser-Tool-
733 Ream-Kiss/3ead0ae31b91d3096369de11f3488024f752bdc5 (2013).

734 82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
735 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

736 83. Rodrigues, J. P. G. L. M., Teixeira, J. M. C., Trellet, M. & Bonvin, A. M. J. J. pdb-tools: a
737 swiss army knife for molecular structures. *F1000Research* **7**, 1961 (2018).

738