medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

# Stability of SARS-CoV-2 spike antigens against mutations

Ildefonso M. De la Fuente[1,2*], Iker Malaina[2]*, Maria Fedetz[3], Maksymilian Chruszcz[4], Gontzal Grandes[5], Oleg Targoni[6], Antonio A. Lozano-Pérez[7,8], Eyal Shteyer[9], Ami Ben Ya'acov[9], Agustín Gómez de la Cámara[10], Alberto M. Borobia[11], Jose Carrasco-Pujante[12], Jose Ignacio Pijoan[5,13], Carlos Bringas[12], Gorka Pérez-Yarza[5,12], Alberto Ouro[14], Michael J. Crawford[15], Varda Shoshan-Barmatz[16], Vladimir Zhurov[17], José I. López[5], Shira Knafo[18,19], Magdalena Tary-Lehmann[6], Toni Gabaldón[20,21,22,23], Miodrag Grbic[17,24,25].

1. Department of Nutrition, CEBAS-CSIC Institute, Espinardo University Campus, Murcia, 30100, Spain.
2. Department of Mathematics, Faculty of Science and Technology, University of the Basque Country, UPV/EHU, Leioa, 48940, Spain.
3. Department of Cell Biology and Immunology, Institute of Parasitology and Biomedicine "López-Neyra", CSIC, Granada, 18016, Spain.
4. Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC, 29208, USA.
5. Biocruces-Bizkaia Health Research Institute, Barakaldo, 48903, Spain.
6. Cellular Technology Ltd, Cleveland, OH, 44122, USA.
7. Departamento de Biotecnología, Genómica y Mejora Vegetal, Instituto Murciano de Investigación y Desarrollo Agrario y Medioambiental (IMIDA), 30150, La Alberca (Murcia), Spain.
8. Instituto Murciano de Investigación Biosanitaria (IMIB)-Arrixaca, 30120, El Palmar (Murcia), Spain.
9. The Juliet Keidan Institute of Pediatric Gastroenterology Hepatology and Nutrition, Shaare Zedek Medical Center, The Hebrew University School of Medicine, Jerusalem, 9103102, Israel.
10. Department of Clinical Research, Hospital 12 de Octubre, Av. Córdoba, Madrid, Spain
11. Clinical Pharmacology Department, La Paz University Hospital (IdiPAZ). Universidad Autonoma de Madrid, 28029, Spain.
12. Department of Cell Biology and Histology, Faculty of Medicine and Nursing, University of the Basque Country, UPV/EHU, Leioa, 48940, Spain.
13. Unidad de Epidemiología Clínica, Hospital Universitario Cruces, Barakaldo, 48903, Spain.
14. NeuroAging Laboratory Group (NEURAL), Clinical Neurosciences Research Laboratories (LINCs), Health Research Institute of Santiago de Compostela (IDIS), 15706, Santiago de Compostela, Spain
15. Department of Biomedical Sciences, University of Windsor, Windsor Ontario, N9B 3P4, Canada.
16. Department of Life Sciences and the National Institute for Biotechnology in the Negev, Ben-Gurion University, Beer-Sheva, 84105, Israel.
17. Department of Biology University of Western Ontario, Ontario N6A5B7.Canada.
18. Ikerbasque, Basque Foundation for Science, Bilbao, 48013, Spain.

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

19. Department of Physiology and Cell Biology, Faculty of Health Sciences, and The National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, 8410501, Israel.
20. Barcelona Supercomputing Centre (BSC-CNS), Life Sciences Department, Jordi Girona 29, 08034 Barcelona, Spain.
21. Institute for Research in Biomedicine (IRB Barcelona), Mechanisms of Disease Program, The Barcelona Institute of Science and Technology, Baldiri Reixac 10, 08028, Barcelona, Spain.
22. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.
23. Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, 08010, Spain.
24. Department of Agriculture and Food, University of La Rioja, Logroño, La Rioja, 26006, Spain.
25. Department of Biology, University of Belgrade, Belgrade, Serbia.

*** Equal contribution**

**Corresponding Author:**
Ildefonso M. de la Fuente
Instituto CEBAS-CSIC.
Campus Universitario de Espinardo.
Espinardo.
30100 Murcia. ESPAÑA.
TEL: +34 968396200
Email: mtpmadei@ehu.eus ildefonso@cebas.csic.es

## Highlights

• Our novel method predicts SARS-CoV-2 epitopes that are stable against future mutations

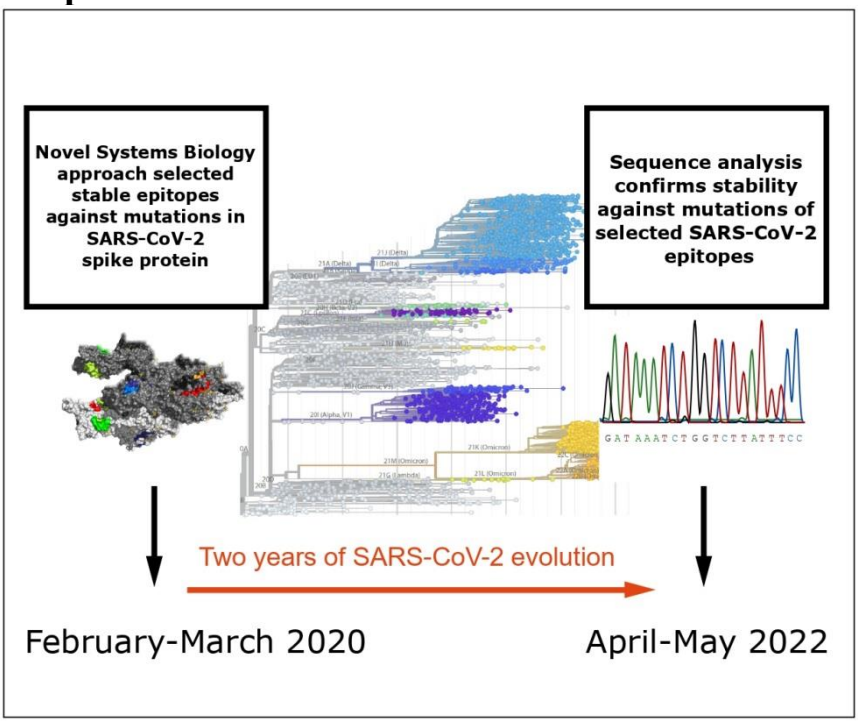
• Genetic analyses (performed 2 years after epitopes were identified) validate the stability of the identified peptides

• These epitopes remained invariable in 97% of all 1,514 known SARS-CoV-2 lineages

• 93.9% of such peptides were conserved in the 43 variants of most interest, including Omicron

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

## Graphical Abstract




## Summary

Modern health care needs preventive vaccines and therapeutic treatments with stability against pathogen mutations to cope with current and future viral infections. At the beginning of the COVID-19 pandemic, our analytic and predictive tool identified a set of eight short SARS-CoV-2 S-spike protein epitopes that had the potential to persistently avoid mutation. Here a combination of genetic, Systems Biology and protein structure analyses confirm the stability of our identified epitopes against viral mutations. Remarkably, this research spans the whole period of the pandemic, during which 93.9% of the eight peptides remained invariable in the globally predominant 43 circulating variants, including Omicron. Likewise, the selected epitopes are conserved in 97% of all 1,514 known SARS-CoV-2 lineages. Finally, experimental analyses performed with these short peptides showed their specific immunoreactivity. This work opens a new perspective on the design of next-generation vaccines and antibody therapies that will remain reliable against future pathogen mutations.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

## Introduction

Mutational processes are an intrinsic characteristic of viruses, and replication errors are the main source of genomic variations. The resulting genetic diversity allows evolutionary adaptation to new environments and immunogenic defenses, and the consequent emergence of different viral variants.

Human coronaviruses have large and complex genomes, and their spike proteins exhibit high plasticity that adapt with relative ease to different cellular receptors and new environmental conditions (Forni et al., 2017). SARS-CoV-2, like all coronaviruses, is less prone to genomic modifications than other RNA viruses with higher mutation rates. The comparatively low $10^{-3}$ to $10^{-5}$ mutations/nucleotide/round of replication (Smith et al., 2013) appear due to the presence of an RNA 3′→5′ exoribonuclease, that maintains greater sequence integrity during replication and transcription processes (Minskaia et al., 2006; Ferron et al., 2018). However, a large number of mutations have accumulated in the spike proteins throughout the recent pandemic. Similar mutational dynamics can be observed in experiments performed with Vero cells using two variants of SARS-CoV-2, in which a spontaneous mutation rate of $1.3 \times 10^{-6} \pm 0.2 \times 10^{-6}$ per-base per-infection cycle was estimated (Amicone et al., 2022). Specifically, the accumulation of S-spike mutations has generated numerous viral variants with different affinities to the human angiotensin-converting enzyme 2, and the ACE2 receptor (Ali et al., 2021, Luan et al., 2021). Many of these variants, especially in the N-terminal Domain (Kubik et al, 2021), have increased the viral infectivity (Davies et al., 2021; Zhang et al., 2021), and/or have reduced the sensitivity to antibody neutralization (Planas et al., 2021).

The majority of mutations observed in the SARS-CoV-2 genome correspond to two types: G →U and C →U, thus leading to extremely frequent homoplasies. Moreover, studies have revealed that the SARS-CoV-2 mutation process is far from equilibrium (De Maio et al., 2021), and that mutations are not symmetric (Roman et al., 2021). On the other hand, large-scale SARS-CoV-2 genome sequencing data analyzed during the pandemic reveal S-spike protein mutational hotspots across phylogenies, which are described as distinct non-synonymous mutations/insertions found at the same positions in several lineages. These mutations appear independently in different geographical zones (Kubik et al 2021; Gerdol et al. 2022). For example, among the Variants of Concern many common mutations are considered to emerge via convergent evolution (van Dorp et al., 2020; Martin et al., 2021; Rochman et al., 2021; Kistler, 2022).

The greatest challenge in the development of successful preventive vaccines lies in identifying antigenic epitopes likely to be resistant to future mutations, and capable of inducing an effective, safe, and long-lasting immune response (Morens et al., 2021). The enormous difficulty of accommodating the inevitable and complex mutational dynamics of viruses has hampered the development of efficient vaccines. Different

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

research groups are trying to develop tools that predict likely mutational hotspots so that they can be avoided. For example, Maher et al. (2022), Obermeyer et al. (2022), and Rodriguez-Rivas et al. (2022) have proposed approaches that rely upon large extant sequence data sets to predict the amino acids most likely to change in SARS-CoV-2. However, so far, no effective strategies exist that can predict regions *likely to remain stable* in a pathogen's genome over a long period of evolution.

Over the last 8 years, we used a System Biology approach to improve the development of mutation-resistant vaccine strategies (Martínez et al., 2015; Martínez et al., 2019). In contrast to the classical methods intending to protect against an individual variant of the virus, our methodological platform, called "Multi-stable string," identifies a small set of peptides with immunogenic capacity and affinity to the HLA Class I and Class II histocompatibility antigens, and that also exhibit a high probability of remaining stable against viral mutation (Martínez et al., 2015). This approach, based upon combinatorial mathematics, advanced computational techniques, artificial intelligence, and immune-informatics tools, provides a new methodology to obtain candidate epitopes most likely to remain stable over a long period of viral evolution.

At the beginning of the COVID-19 pandemic (February-March 2020), when only 22 viral genomes deposited in GISAID project could be analyzed, we identified a set of eight SARS-Cov-2 spike protein epitopes using our approach. Here, we have studied the stability against mutations of these selected peptides throughout the entire two-year period of COVID-19 pandemic.

First, we have carried out an exhaustive genetic analysis of the 3,362 complete genomes available at NextStrain (Hadfield et al., 2018) and GISAID (Khare et al., 2021). This study covers all the S-spike mutations, including very low frequencies (0.0001), considered across 1,514 SARS-CoV-2 variants. The obtained results have shown that our epitopes are located in the S-spike protein cold spots (conserved regions with very low mutational frequency). The genetic analysis has also allowed us to identify additional ten highly stable SARS-CoV-2 spike sequences. Thus, we identified 18 highly conserved SARS-CoV-2 S-protein regions resistant to mutations in this study.

Next, we have analyzed the mutational stability of the eight epitope pool in the main SARS-CoV-2 variants (ECDC), taking into account all the S-spike defining mutations for the considered lineages until April 2022 (GISAID-Outbreak project; Gangavarapu et al., 2022). The percentage of stability of the eight epitope pool after two years of virus evolution was 93.9%. These results have been confirmed in another analysis covering all mutations reported in the CoV-GLUE dataset of the 28 lineages considered of most interest.

Furthermore, we have carried out a fourth stability study of the selected epitopes using all the mutations reported in the CoV-GLUE/GISAID dataset of the 1,514 SARS-CoV-2 lineages that appeared during the two years of Covid-19 pandemics (CoV-GLUE), showing that our epitopes are conserved in at least 97% of all SARS-CoV-2 lineages.

In summary, we designed eight short peptides with the potential to be resistant to mutations, when the pandemic started, and now, more than two years later, this exhaustive analysis confirms the mutational stability of epitopes selected by our quantitative tool.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

We also performed a peptide mapping of the eight short peptides, which indicated that they are all exposed to solvent. Besides, we have studied the structure of the FDA-authorized monoclonal antibodies used in COVID-19 treatment, comparing them with our eight peptides. This study provides additional stable immunogenic targets to increase the range of neutralizing antibodies that can be used for COVID-19 treatment. The overlap between our peptides and the epitopes targeted by some monoclonal antibodies used for COVID-19 treatments advises the selection of these antibodies as a component of future antibody cocktails targeting more stable epitopes.

Finally, we have carried out several experimental analyses using cells of donors recovered from the COVID-19 infection. These analyses have revealed a Th1-specific response after the stimulation with the pool of the eight peptides. Such a response was achieved even at low peptide concentrations. On the other hand, our preliminary experiments have shown that the antibodies generated in patients with the COVID-19 infection recognized all the peptides, indicating their potential as immunogens able to generate humoral response and immunological memory in the human immune system.

To the best of our knowledge, our tool is the first to reliably predict regions in the genome of a pathogen that are likely to remain stable for long periods of time. This approach opens a perspective for new methodological procedures able to serve as a base for the development of universal vaccines, valid for most or all future variants, and more efficient neutralizing antibody therapies through the selection of epitopes stable against mutations.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

# Results

## 1. Selection of eight mutationally stable epitopes at the beginning of the COVID-19 pandemic

To design a combination of epitopes stable against future mutations, we first quantitatively studied the molecular characteristics of the S-spike using a new Systems Biology-based approach (see STAR Methods section). This analysis was performed at the beginning of the COVID-19 pandemic (February-March 2020) and identified more than 3,000 possible epitopes within the S-spike structure. We further constrained epitope selection to those amino acid sequences that were potentially immunogenic: the affinity between the candidate epitopes and the HLA molecules were matched. We took into account the absolute frequency of the different HLA alleles present in the world population to favor the selection of antigens with the highest affinity to the prevalent HLA molecules. Finally, using techniques of combinatorial optimization (Martínez et al., 2015), we winnowed the list of candidate sequences to identify regions of the S-protein structure most likely to remain stable. This analysis enabled selection of the eight short peptides (P1-P8) of the SARS-Cov-2 S-protein with the capacity to preserve maximum stability against variants of SARS-CoV-2. These sequences comprised 10 to 24 amino acids

By virtue of our selection criteria, the epitope pool is characterized by high immunogenicity and affinity to the HLA antigens, covering 85% and 100% of the most frequent Class I and Class II alleles respectively. The location of these peptides within the S-structure, as well as their corresponding amino acid sequences, is shown in Figure 1. Although the algorithm design strategy scanned all S-spike regions without bias, the eight short peptides (P1-P8) identified were mainly localized to domains critical for viral entry into the host cell, including two peptides located near the angiotensin-converting enzyme 2 (ACE2) binding site (Figure 1).

Our mathematical epitope selection was developed by February-March 2020 when only the 22 viral genomes had been deposited in the GISAID project database (Khare et al., 2021). The sequences of the eight peptides (P1-P8) were submitted to a formal registration procedure (May 2020, patent register #P202030467).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

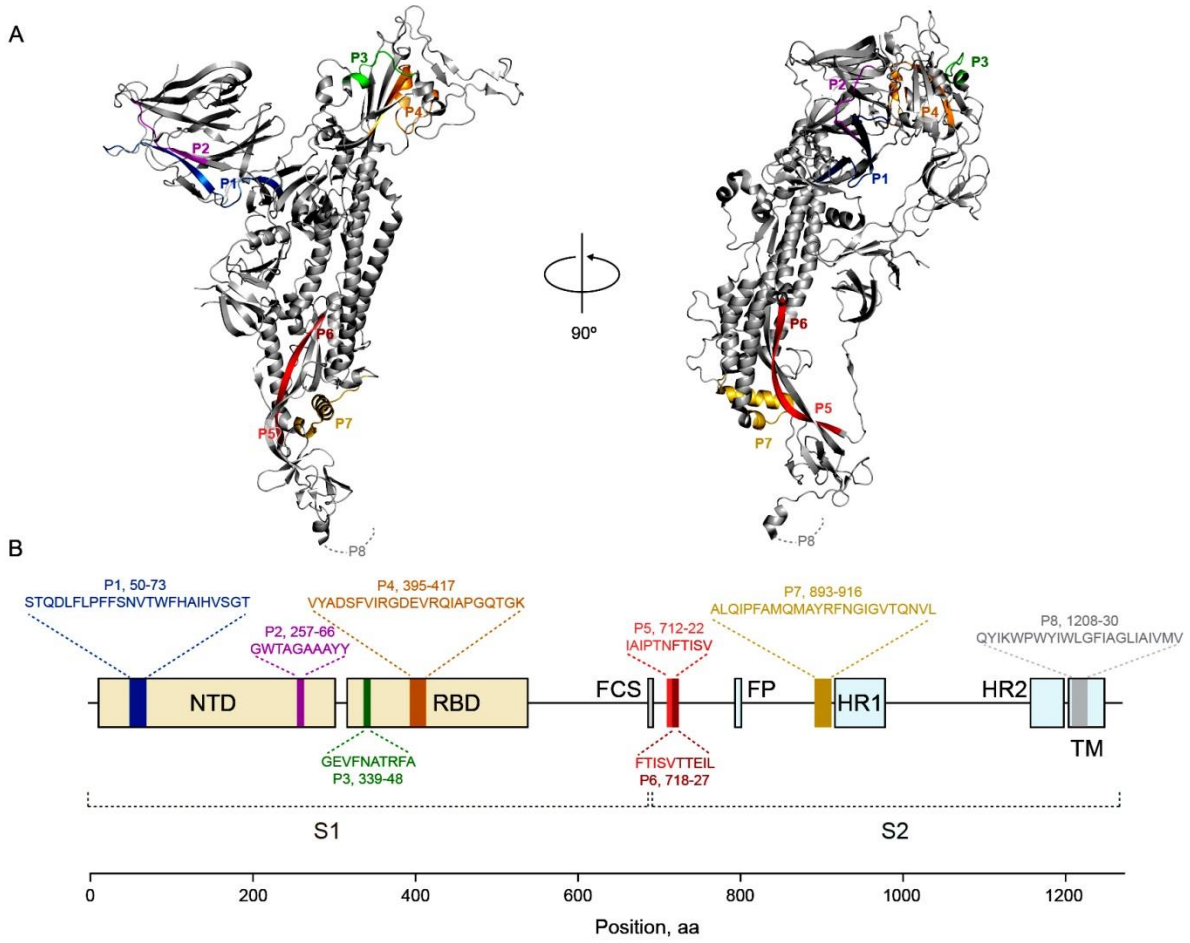



**Figure 1. Peptide sequences and location of the eight regions stable against mutations identified at the beginning of the COVID-19 pandemic**
(**A**) The structure of the Spike protein and locations of the eight identified peptides. The complete sequences of the eight epitopes and their respective locations are detailed in (B), where peptides P3 and P4 are located in the RBD (Receptor Binding Domain). Noteworthy, our methodology studied the S-structure as a whole, without preselecting any determined region. S1 – Receptor Binding Domain; S2 – Membrane Fusion Domain; NTD – N-Terminal Domain; FCS – Furin Cleavage Site; FP – Fusion Peptide; HR1 and 2 - Heptapeptide Repeat Sequence; TM – Transmembrane domain.

## 2. Stability against mutations of the eight epitopes: a two year post hoc analysis

To test the predictive capacity of our tool to determine epitope stability we interrogated sequence databases of SARS-Cov-2 evolution over the subsequent two years. We analyzed the public data of GISAID (Khare et al., 2021), which included 3,362 complete genomes across 1,514 SARS-CoV-2 lineages sampled between December 2019 and January 2022 (available at NextStrain; Hadfield et al., 2018). We also computed entropy per site, the number of mutational events across the strain tree, and protein variation using GISAID hCoV-19 sequences available at the CoV-GLUE database (Singer et al., 2020), which includes very rare mutations (0.0001).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Figure 2A-D shows an analysis of the Spike sequence mutations covering the distribution of non-synonymous mutation frequencies, site entropies, and a number of inferred mutational events per site across the SARS-CoV-2 phylogeny. To study whether each residue had the same mutation probability, we performed a goodness of fit test of the distribution of probabilities using $\chi^2$ test scores based on values obtained from the number of non-synonymous mutation events (see Figure 2C). We obtained a $p - value = 0$ and a $\chi^2 - statistic = 2 \cdot 10^4$, rejecting the null hypothesis. This indicates the probability of mutation is not the same in every region of the protein. The non-synonymous mutations are not uniformly distributed over the SARS-Cov-2 S-protein, and as a consequence, we can observe regions with high or low mutation frequency.

Likewise, our analysis indicated that the selected epitopes (P1-P8) are represented by specific residues that, on average, have comparable site entropy and mutational event values across the SARS-CoV-2 phylogeny, and low mutation rates compared with other randomly selected epitopes of similar size. For example, we chose 10,000 random combinations of short peptides of similar size and compared them with the set of eight selected epitopes (P1-P8) (STAR Methods). The selected epitope pool presented a frequency of 0.0023 mutations per site, and relative entropy of 0.022. This pool P1-P8 epitopes is expected to be invariable in 71.4% of the sequences, as compared with 45.0% in 10,000 randomly sampled combinations of similar fragments from the Spike protein. The results indicate that P1-P8 epitopes are not located in mutational hotspots of the S-protein.

In this post-hoc sequence analysis, we found another ten short peptides that have a low mutation rate (Figure 2 A-D, marked in blue R1-R10). The complete sequence of these peptides (R1-R10) is indicated in the STAR Methods. In total, all 18 mutation-stable regions have an average amino-acid length of 16.22±5.54 (mean $\pm$ SD), with a confidence interval $I_{\mu}^{0.95} = (13.47, 18.98)$. There were no significant differences between the amino-acid lengths in P1-P8 and R1-R10 groups (STAR Methods section).

Our Systems Biology-based approach selected eight short SARS-CoV-2 S-spike peptides (P1-P8), among millions of combinations of possible amino acid sequences including the previously mentioned ten highly conserved peptide sequences (R1-R10) (Figure 2). This latter set of stable S-protein peptides was not predicted by our tool because mathematical and computational calculations excluded them on the basis of lower potential to induce an immune response (see STAR Methods section).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

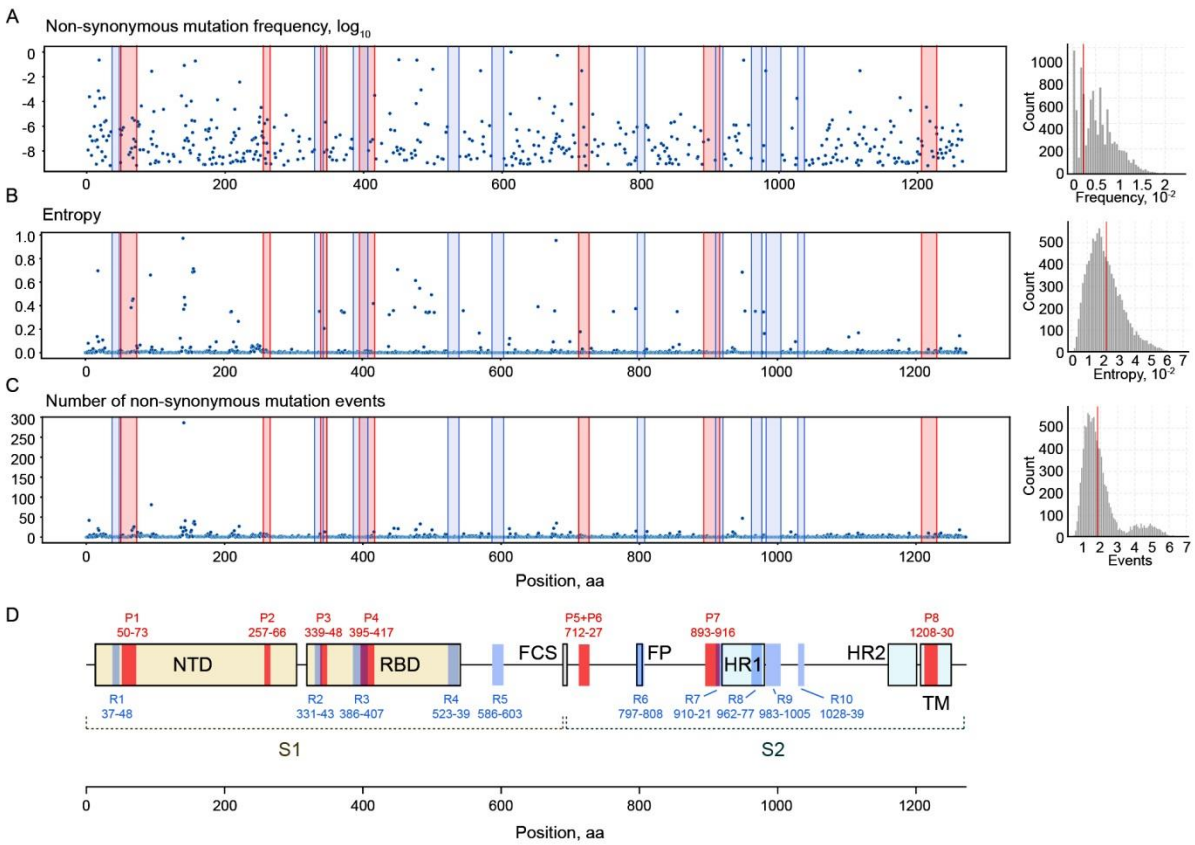



**Figure 2. Stable SARS-CoV-2 S-spike sequences against mutations**
Distribution of non-synonymous mutation frequencies (A), site entropies, normalized Shannon entropy where 1 is maximal entropy (B), and a number of inferred mutational events per site across the SARS-CoV-2 phylogeny (C) overlayed by schematics of S-protein selected peptides (D) are shown as dot plots for all residues of the S-spike protein. Bar plots on the right show the distribution of average values for 10,000 randomly chosen combinations of epitopes of similar size. Vertical red bars indicate the value for our combination of eight epitopes (P1-P8): frequency of 0.0023 mutations per site, 1.82 mutational events per site, and relative entropy of 0.022. Vertical blue bars correspond to the set of ten highly conserved peptide sequences (R1-R10).

## 3. Stability of the eight selected epitopes in the SARS-CoV-2 variants

Evolution of the SARS-CoV-2 genome has generated a complex cascade of lineages derived from a common ancestor (Figure 3A, left). To evaluate the mutational stability of the eight selected peptides (P1-P8) we have first analyzed the stability level of our epitope pool in the 43 SARS-CoV-2 variants of most interest, taking into account the defining S-spike mutations described until April 30, 2022 (GISAID-Outbreak project).

Figure 3A shows the phylogeny of the SARS-CoV-2 virus (nextstrain.org), where 3,544 representative genomes are considered. The origin of the mutational cascade was established in the first viral samples obtained in China at the beginning of the pandemic (Wu et al., 2020). On the right in Figure 3A, the 41 SARS-CoV-2 variants of most interest as determined by the European Centre for Disease Prevention and Control (ECDC) are shown. The list includes variants of concern (VOC), Variants Under Monitoring, and De-escalated variants (ecdc.europa.eu). The ECDC traces a wider number of variants than the 32 considered by the World Health Organization (who.int). The variants defined by the ECDC exhibit "significant potential for transmissibility,

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

severity, and/or immunity likely to have an impact on the epidemiological situation based on properties analyzed by combined genomic, epidemiological, and in vitro pieces of evidence" (ecdc.europa.eu). In addition, we have added to the analysis two other Delta variants (21I and 21J) given their importance in some areas during the second half of 2021. The total number of variants considered in our study (n=43), the corresponding Pango lineage, WHO-ECDC nomenclature, and the percentages of stability for the eight epitopes in each variant, are also shown in Figure 3A. All the characteristic mutations for the analyzed lineages (non-synonymous substitutions or deletions that occur in > 75% of sequences within each variant) were obtained from the GISAID-Outbreak project.

Figure 3 A and B show that the stability percentage of the eight-epitope pool (P1-P8) in the 43 SARS-CoV-2 variants is 93.9% with a confidence interval of $I_{\mu}^{0,95} = (90.8\%,\ 96.2\%)$. 55.8% of all of variants display 100% stability of the eight selected peptides, with 41.9% displaying only a single epitope appears mutated. The Omicron variant has five out of eight epitopes conserved (62.5%). Specifically, peptide P1 in the Omicron variant has two mutations, epitope P3 only displays a substitution of the first residue, and epitope P4 presents a unique change in the last residue, which is illustrated in Figure 6.

The dynamics of stability against mutations of each peptide in every variant analyzed are displayed in Figure 3B. Epitopes P2, P5, P6, P7, and P8 never exhibit mutational modification as catalogued by ECDC (GISAID-Outbreak project). Epitope P4 appears mutated in eight variants out of 43 (Beta B.1.351, B.1.214.2, B.1.351+P384L, Gamma P.1, B.1.351+E516Q, P.1+P681H, B.1.617.2 + K417N and Omicron B.1.1.529), displaying a mutational probability of 0.186, with a confidence interval of $I_{p}^{0,95} = (0.084,0.334)$. Epitope P3 is mutated in four variants (A.23+E484K, Mu B.1.621, B.1.640 and Omicron B.1.1.529) with a mutational probability of 0.095 and a confidence interval of $I_{p}^{0,95} = (0.027, 0.226)$. Finally, epitope P1 is mutated in nine variants (alpha B.1.1.7, A.28, B.1.1.7+E484K, eta B.1.525, B.1.1.7+L452R, B.1.1.7+S494P, B.1.616(c), B.1.620 and Omicron B.1.1.529) and exhibits a mutational probability of 0.214 with a confidence interval of $I_{p}^{0,95} = (0.103, 0.368)$.

Aside from analyzing the stability of our S-spike-derived peptides with the defining mutations (Figure 3), we assessed the stability of the selected epitopes against all non-synonymous mutations in the 28 lineages of most interest reported in the CoV-GLUE dataset (Figure 4A). The results indicated that the peptides (P1-P8) exhibit a very low average probability of mutation. Specifically, epitope P1 is conserved in 95% of the lineages; epitope P2 in 99.8%; epitope P3 in 89.1%; epitope P4 in 89.2%; the overlapping epitope P5-6, in 96.1%; epitope P7 in 98.2%; and finally, epitope P8 in 91.6% of the lineages. This translates to an overall preservation ratio of 94.1%, with a confidence interval $I_{\mu}^{0.95} = (90.19\%, 98.08\%)$. Therefore, in the subset of 28 lineages (Figure 4A) we observe a strong agreement with the previous analysis focused on defining mutations of SARS-CoV-2 variants catalogued by ECDC (GISAID-Outbreak project) (Figure 3).

We carried out a fourth study of genetic stability of the P1-P8 peptides by analyzing the non-synonymous mutations reported in the GISAID dataset in all the 1,514 known SARS-CoV-2 lineages generated during the Covid-19 pandemic (CoV-GLUE lineages).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

In Figure 4B, the k-means clustering analysis of mutation probabilities of all lineages indicates that the vast majority (1384 out of 1514, or 91.4%; cluster 8) exhibit an extremely low average probability of mutation within the selected epitopes (P1-P8). Among the rest of the lineages (130), 85 exhibit low mutational probability, while only 45 present a high probability of change, limited to a specific region of a single epitope (Figure 4B). Our results indicate that epitope P1 was stable in 97.4% of the lineages; epitope P2, 97.9%; epitope P3, 99.5%; epitope P4, 98%; the overlapping epitope P5-6, 98.8%; epitope P7, 99.9%; and, finally, epitope P8 was stable in 98% of the lineages. The percentage of conservation of all short peptides (P1-P8) indicates an overall preservation ratio of 98.5%, with a confidence interval $I_{\mu}^{0.95} = (97.65\%,\ 99.34\%)$. Therefore, we demonstrate, with 95% of confidence, that the selected epitopes (P1-P8) have been conserved at least in 97% of all SARS-CoV-2 lineages that arose during two years of the Covid-19 pandemics.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

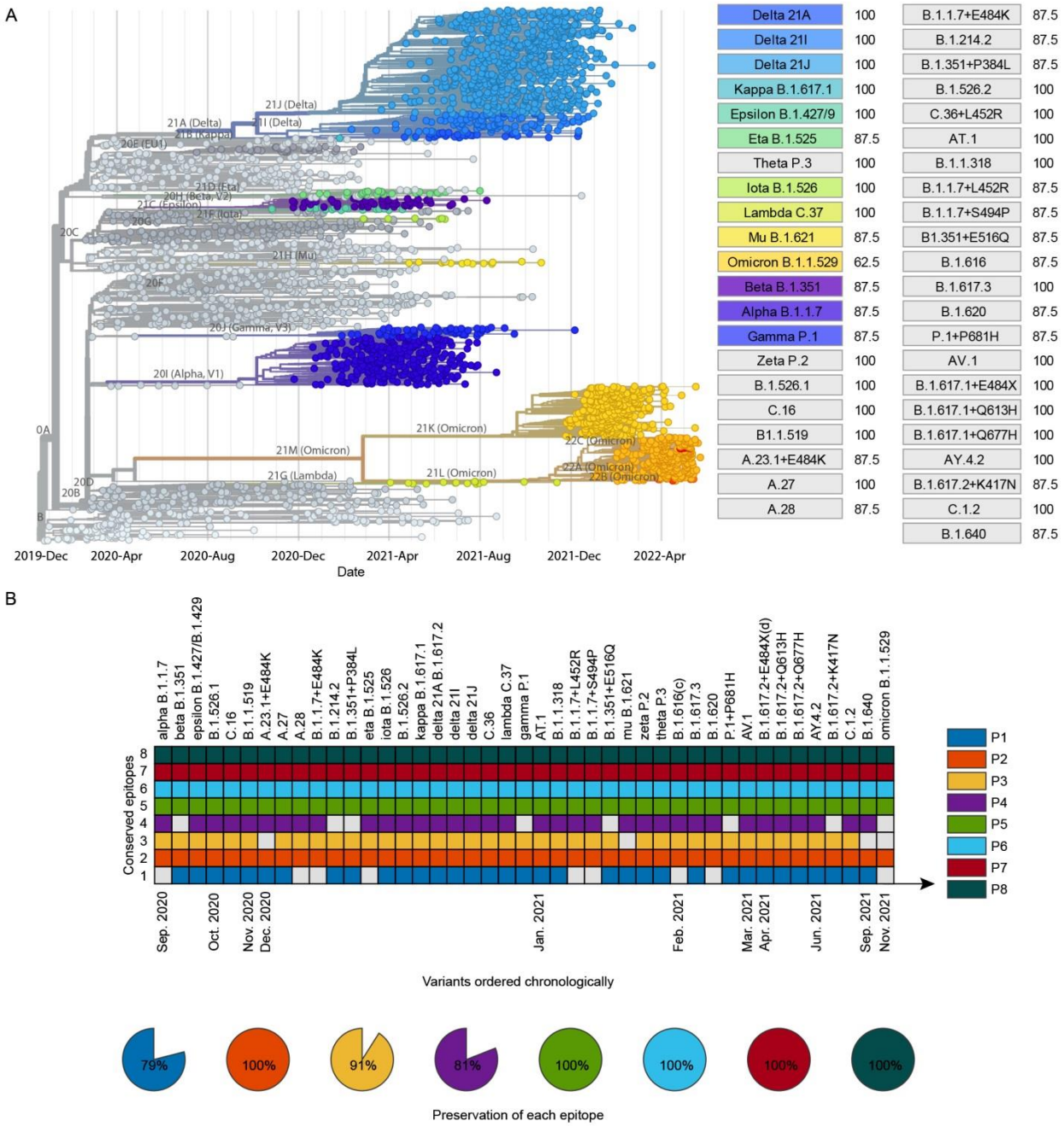



**Figure 3. Identified epitope stability against all defining mutations in the main 43 SARS-CoV-2 variants**
(A) The time elapsed from the original design of the epitopes (February to March 2020) until the mutational stability analysis performed in April 2022 (two years after design), as well as a summary of the SARS-CoV-2 phylogenetic evolution is depicted. On the right side, the percentage of stability of the set of eight peptides (P1-P8) in each of the 43 variants considered (93.9% in total) is shown. The confidence interval evaluating the stability of the P1-P8 epitopes is $I_{\mu}^{0,95} = (90.8\%,\ 96.2\%)$. (B) illustrates the mutational stability of the selected epitopes in each variant of most interest and the percentage of preservation of each peptide sequence.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

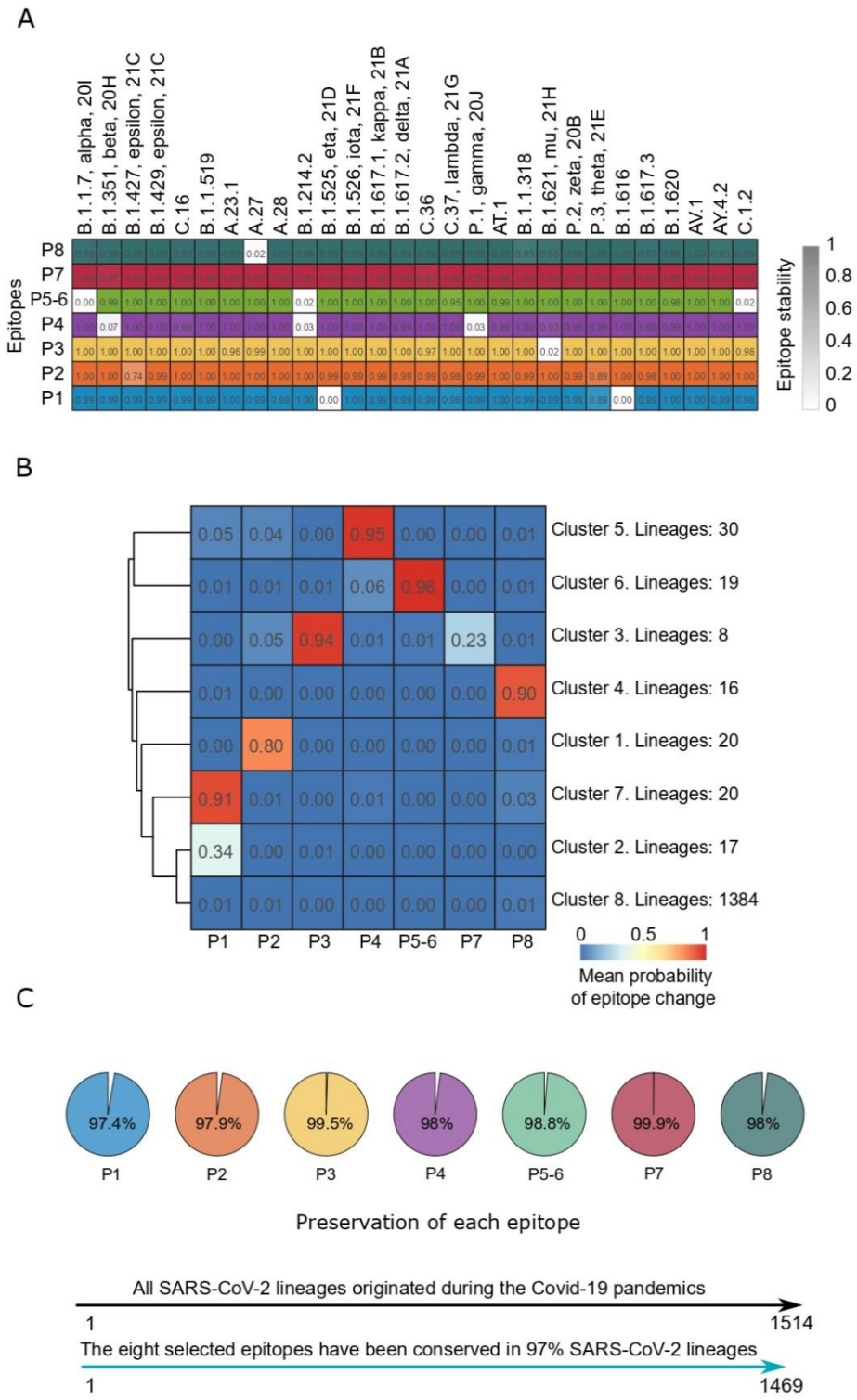



**Figure 4. Stability of the selected epitopes against all non-synonymous mutations in the 28 lineages considered of most interest, and in 1514 lineages originated during the Covid-19 pandemics**

(A) Stability of the selected epitope regions against all reported mutations in the GISAID dataset in the 28 lineages of most interest. The percentage of stability of the set of eight peptides in each of the 28 lineages considered is 93.9%, with a confidence interval $I_{\mu}^{0.95} = (90.19\%, 98.08\%)$. This result agrees with the analysis focused on defining mutations (Figure 3). (B) k-means clustering of probabilities of observation of non-synonymous mutation within regions of selected epitopes across 1514 lineages in the GISAID dataset. The distance is Euclidean, the linkage is Ward's, k = 8. Our selected epitopes are conserved in 97% of all SARS-CoV-2 lineages (1384 total). (C) The percentage of conservation of the epitopes (P1-P8).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

## 4. Peptide mapping and solvent exposure of the eight epitopes identified stable against mutations

To identify, understand, and differentiate characteristics of the selected epitopes (P1-P8) and the peptides (R1-R10) (which also exhibited low mutation rates but were rejected by our Systems Biology methodology), we mapped both groups of peptides on the three-dimensional structure of SARS-CoV-2 S-protein. Available experimental models of the S-spike allowed us to map seven of the selected peptides (P5-P6 are partially overlapped, see Figure 1) on the protein's structure. Peptide mapping (Figure 5, A and B) indicates that at least some of the peptides' residues are solvent-exposed, with residues of peptides P2 and P7 being the most buried. Peptide P8 is located in the C-terminal region of the protein that is disordered and therefore is most likely exposed to solvent. Notably, these residues are solvent exposed in both "closed" and "up" conformations (Mehra and Kepp 2022). Moreover, the change from "closed" to "up" conformation significantly increases solvent accessibility of peptide P4 (Figure 5, B and C).

Our peptide selection methodology does not assume anything about protein oligomerization. Method robustness is illustrated by the fact that none of the peptides are buried completely in the S-protein trimer, sites instead remain accessible to antibodies. Another interesting feature of the selected peptides is that despite the relatively large sequence distances separating them, some of them are quite close to each other in the mature protein. For example, peptides P1 and P2 are next to each other (Figure 5A and B), and peptide P3 and a fragment of peptide P4 are in close vicinity in the closed conformation of the S-spike. In addition, peptides P5 and P6 (which partially overlap) together with peptide P7 are also relatively close in the three-dimensional protein structure.

The analysis of the S-protein ("up" conformation) in a complex with ACE2 (Zhou et al., 2020) illustrates how the residues forming peptides P3 and P4, which correspond to the receptor binding domain (RBD) region, remain exposed to solvent (Figure 6A). This suggests that they might remain epitopes recognized by human antibodies, and they are not critical for the formation of the S-protein-ACE2 interface (for modifications and properties of the peptides, see STAR Methods). However, we cannot exclude the possibility that the ACE2 carbohydrate moieties block access to potential epitopes formed by peptide P3 or P4 residues.

Only three peptides (P1, P3 and P4) out of the eight originally identified are affected by mutations present in the Omicron variant (Figure 6; for details, see STAR Methods). Similarly, only three peptides (R2, R3 and R8) from the second group of stable peptides are mutated in Omicron strains. However, these mutations have a relatively small impact on the spatial conformation of our peptides: only peptide P1 seems to be markedly affected. In this instance, there is not only a single mutation (A67V) in peptide P1, but also the residues H69-V79 are missing in the Omicron/BA.1 strain. The impact of conformational changes, ACE2 binding, and mutations in the Omicron variant on the selected peptides are shown in Figure 6.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

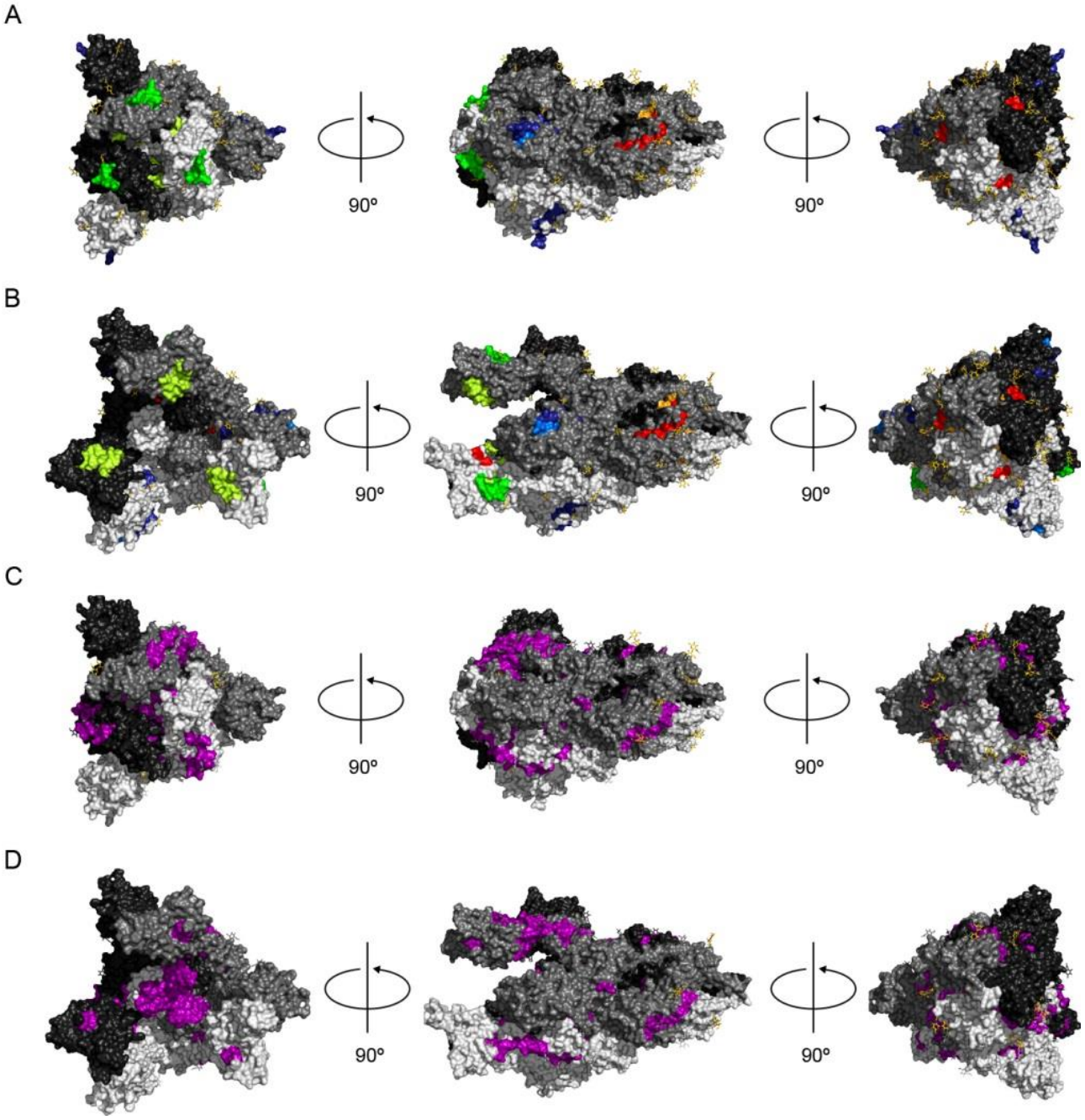



**Figure 5. Peptide mapping and solvent exposure.**
Peptides (P1-P8) are marked using a color scheme indicated in Figure 1. Partially overlapped peptides P5 and P6 are marked in red. Panel (A) shows an experimental model of the Spike protein in a "closed" conformation (PDB code: 6VXX). The protein chains forming the trimmer are displayed in different shades of grey. Carbohydrate moieties are shown in stick representation with carbon atoms in yellow. Panel (B) presents an experimental model of the S-spike protein in "up" conformation (PDB code: 7KMS). This conformation of the S-spike protein allows for ACE2 binding. Panel (C) shows an experimental model of the spike protein in "closed" conformation with positions of our non-selected peptides (R1-R10) shown in magenta. Panel (D) represents an experimental model of the S-spike protein in "up" conformation (PDB code: 7KMS) with R1-R10 peptides shown in magenta.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

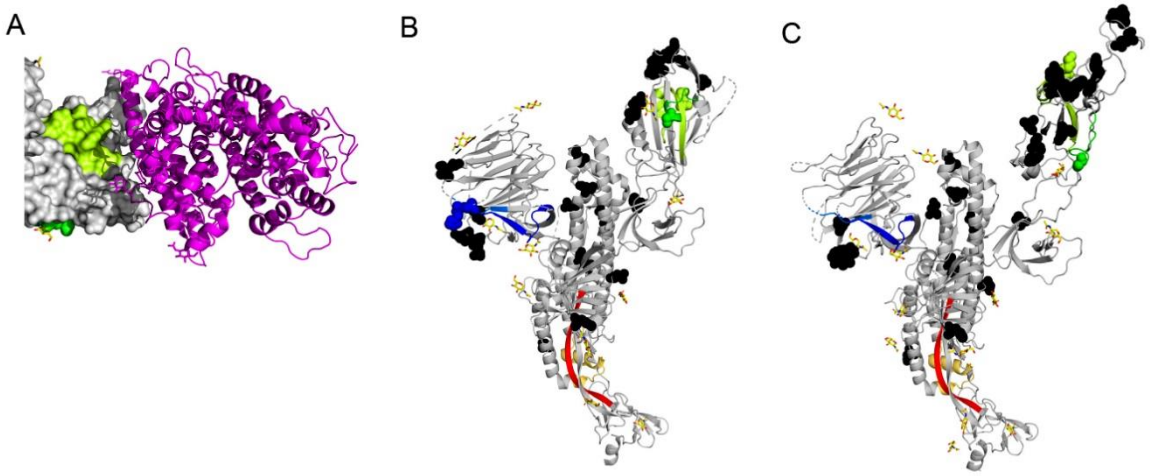



**Figure 6. Impact of conformational changes, ACE2 binding and mutations on the selected peptides and in the Omicron variant**
Identified peptides are marked using a color scheme indicated in Fig. 1. Partially overlapped peptides P5 and P6 are marked in red. The interaction between the RBD (shown in surface representation) and human ACE2 (in magenta; shown in ribbon representation) is demonstrated in panel (A). Residues forming peptides P3 and P4 are displayed in green and lemon green, respectively. The experimental model of the protein in "closed" conformation (PDB code: 6VXX) is shown in panel (B). The protein chains forming the trimmer are displayed in different shades of grey. Carbohydrate moieties are shown in stick representation with carbon atoms in yellow. An experimental model of the S-spike protein in "up" conformation (PDB code: 7KMS) is presented in panel (C). This conformation of the S-spike proteins allows the ACE2 binding. Carbohydrate residues are indicated in stick representation. The PDB deposit (7KMS) corresponding to the S-spike-protein-ACE2 complex was used to generate this figure. Residues that are mutated, deleted, or inserted in the Omicron variant are marked as black spheres. Only very small fragments of peptide P1 (blue spheres), peptide P3 (green spheres), and peptide P4 (lemon green spheres) are affected by the mutations observed in the Omicron/BA.1 variant.

## 5. Analysis of the amino-acid composition in the eight identified versus 10 rejected epitopes

To investigate what differentiates the identified (P1-P8) from rejected (R1-R10) peptides, we studied the structure and amino-acid composition, comparing them with random peptides derived from the SARS-CoV-2 S-spike protein (Figure 5A and B; see also our complete analysis in Figure S1). The composition of the sequences covered by P1-P8 was enriched in tryptophan (4.58%) and depleted in lysine (1.53%) as compared to 1.01% and 4.94% respectively in 10.000 randomly-sampled fragment combinations controls. No other amino acid had significantly statistically deviated frequencies. However, while none of the peptides P1-P8 contained cysteine residues, half of the R1-R10 peptides included this amino acid. The cysteine residues observed in R2, R3, R4, R5 and R10 peptides form disulfide bridges in the S-spike protein. Consequently, it is not surprising that they fall into highly conserved regions of the protein. Most cysteine residues are associated with non-epitope regions.

Tryptophan was enriched in peptides P1, P2 and P8, which is somewhat surprising because this residue is usually not present in epitopes, but rather in paratopes (Soga et

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

al., 2010). By contrast, there is no tryptophan residue present in peptides R1-R10. Moreover, the paucity of lysine residues in peptides P1-P8 is surprising as well because this residue is quite often observed in epitopes (Zheng et al., 2015).

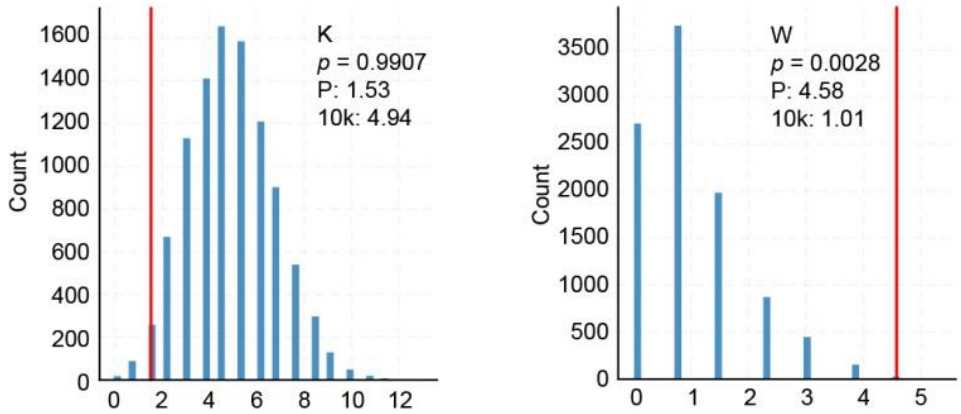



**Figure 7. Distribution of K and W amino acid frequencies in random sets of peptides from the S-spike protein**
X-axis represents the frequency (in percentage) of the indicated amino acid (on letter code, in the top right of each distribution), each bar represents the count of sets of random peptides showing a given frequency across 10,000 random samples (see STAR Methods). The vertical red line indicates the frequency in the set of peptides identified by our algorithm. p, indicates the probability of a random peptide set to have a frequency higher than one of the selected peptides; P indicates the frequency of the indicated amino acid in the selected peptide set, 10k indicates the average frequency of the indicated amino acid in the 10,000 random peptide sets.

## 6. Experimental testing of cellular response to identified peptides P1-P8

T-cell memory, the recall response, was interrogated using the eight peptides in an ELISPOT assay. The samples probed were from peripheral blood mononuclear cells (PBMC) samples derived from convalescent COVID-19 patients and unexposed donors (blood drawn in 2016 and 2017; Figure S2). All the samples from convalescent donors developed an IFNγ response following stimulation with the eight-peptide pool (P1-P8) even at low concentrations. Two samples of unexposed donors responded with T cell activation to the peptide pool. An explanation for the specific response of unexposed donors' samples could be related to a preexistent cross-reactive T-cell memory. Different studies have shown the existence of this immune memory prior to SARS-CoV-2, caused by exposure to previously circulating coronaviruses responsible for common colds. This reaction has been observed in roughly 28-60% of healthy people sampled in years before the start of SARS-CoV-2 pandemic (Grifoni et al., 2020; Mateus et al., 2020; Weiskopf et al., 2020). Neither IL-17 nor IL-5 cytokines were detected in the analyses of the response generated by the PBMCs of the convalescent and healthy donors.

## 7. Experimental test of peptides – patients antibody interactions

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

The functionality of the eight peptides was then tested by ELISA assay. The interaction of each peptide with serum IgG of patients is shown in Figure S3. All the data were normalized against readings of the serum control pool reaction with S-spike-RBD.

The patient samples' reactions with the S-spike-RBD of SARS-CoV-2 were approximately twice more intense than in controls. The sera of patients interact with all the peptides. Thus, the designed peptides were recognized by antibodies produced by the immune system in response to naturally occurring S-spike protein, which indicated that they could generate the immune response in humans.

## 8. Structural analysis of the anti-spike monoclonal antibodies' targets and comparison with the eight stable epitopes

Neutralizing antibodies are being used as a therapeutical tool against SARS-CoV-2 infection. We compared the epitopes-targets of FDA-authorized monoclonal antibodies to the P1-P8 epitopes (this study is detailed in the STAR Methods section). All these antibodies bind to the S-spike protein receptor binding domain (Corti et al., 2021; Hansen et al., 2020; Hastie et al., 2021; Jones et al., 2021; Shi et al., 2020). Bamlanivimab and Imdevimab (REGN10987) targets do not overlap with any residues of the peptides P1-P8. However, Etesevimab and Casirimab (REGN10933) targets partially overlap with residues from peptides P4 and R3. In addition, there is a significant overlap between the Sotrovimab targeted epitope and peptides P3 and R2. This overlap between our peptides and some targets of several monoclonal antibodies used for treating patients with COVID-19 infection confirms that our tool parameters are accurate, robust, and identify the immunogenic potential of epitopes (see STAR Methods section, for more details).

# Discussion

An essential challenge in the development of preventive vaccines is to design an efficient and safe antigenic composition that ensures stability and resistance against future mutations. Unfortunately, no effective strategies that can predict stable regions in a pathogen's genome over a long period of evolution exist so far.

At the beginning of the COVID-19 pandemic, we used our novel tool, "Multi-stable string," to identify a set of eight SARS-Cov-2 S-spike protein epitopes (P1-P8), which were predicted to be stable against long-term future mutations. Here we show that, more than two years after the identification of these short peptides, using a combination of different analyses, including genetic, Systems Biology and protein structure studies, our selected epitopes (P1-P8) are stable against viral mutations. This research spans the whole period of the pandemic.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Genetic analysis allows us to draw the following three main conclusions:

i. Mutations occurring in the SARS-CoV-2 S-spike do not follow a uniform distribution: some peptide sequences display high mutation rates while other small regions are remarkably conserved.

ii. Post hoc analysis shows that after two years of evolution, at least 18 SARS-CoV-2 S-spike peptide sequences appear stable against mutations (Figure 2). These mutational "cold spots" exhibit a short amino-acid length 16.875$\pm$7.1 (mean $\pm$ SD).

iii. Of the 18 polypeptides identified by post hoc lineage sequence analysis, the eight epitopes (P1-P8) originally identified in February-March 2020 (Figure 1) are all included and remained stable against mutations. The remaining ten highly conserved sequences (R1-R10) had been discarded since a priori mathematical and computational calculations predicted a sub-optimal capacity to induce an immune response (see STAR Methods section).

The eight selected epitopes (P1-P8) were interrogated against the SARS-CoV-2 variants of most interest (ECDC) taking into account all the S-spike defining mutations for the considered lineages until April 2022. Analysis shows that after more than two years of evolution of the COVID-19 pandemics, the percentage of stability of our designed epitope pool was 93.9%.

This result has been confirmed by analyzing all mutations reported in the Cov-GLUE dataset which lists 28 lineages considered of most interest. Specifically, our results show that the percentage of stability of the total epitope pool was 94.1%.

Finally, stability of the eight epitopes (P1-P8), was measured against the Nextstrain/GISAID dataset which encompass 1,514 SARS-CoV-2 variants generated during the Covid-19 pandemics (CoV-GLUE): epitopes P1-P8 are conserved at least in 97% of all these sequenced SARS-CoV-2 lineages.

Thus, we conclude that our tool functions with fidelity to identify stable sequences. The least stable epitopes were P4, which displayed a mutational probability of 0.186, and P1 which showed a mutational probability of 0.214. Peptide 3 was also mutated in three variants with a mutational probability of 0.095. In short, the probability that all epitopes may mutate simultaneously is extremely low.

Peptide mapping confirmed that our epitopes (P1-P8) are exposed to solvent in the S-spike protein structure (Figure 5). The comparison of our peptides with the targets of the monoclonal antibodies authorized by the FDA has shown that our peptides could increase the range of neutralizing antibodies used for SARS-CoV-2 treatment. Indeed, the overlap among our peptides and the targets of monoclonal antibodies used for SARS-CoV-2 treatments both confirms tool accuracy and suggests future utility in the development of antibody cocktails to target a greater number of stable epitopes.

Preliminary experiments were performed to assess specific immune recognition of the identified peptides. Peripheral blood mononuclear cells of donors recovered from COVID-19 infection demonstrated an IFN gamma response after stimulation with the pool of eight peptides in ELISPOT assays, both at high and low concentrations. Also,

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

IgG antibodies from convalescent COVID-19 patients interacted with all selected peptides in a manner similar to the intact Spike antigen. The predicted peptides have the potential to induce an immune response in humans.

Our "Multi-stable string" approach, based fundamentally on mathematical combinatorial methods, advanced computational techniques, artificial intelligence, and immune-informatics tools, allows the design of a number of specific epitopes that can direct strategies to accommodate universal variants, thereby enabling the design of long-lasting prophylactic and therapeutic products. For example, one therapeutic monoclonal antibody, directed against an S-protein of SARS-CoV-2, has a remarkable neutralizing capacity including against current Omicron variants. It has been approved for use by the European Medicines Agency, EMA (www.ema.europa.eu) since May 2021. The antibody targets an S-protein sequence (Gupta, et al., 2021; Wu, et al. 2022; Gupta, et al., 2022; Iketani et al., 2022) which is located in a small region that is very stable against mutations in all the variants known to date. This peptide sequence fully includes our epitope P3 (aa339-348).

Our approach can be extended in several ways. We are improving our procedures to consider more variables in the combinatorial optimization process. The "Multi-stable string" concept will be extended to compare "sets" of solutions, evaluating the number of epitopes of each solution, or their total length. While indicating strategic avenues to pursue, this work will need to expand immunogenicity analysis to assess the "in vivo" immunogenicity and safety profile of our peptides in animal models using as a carrier a biologically safe nanoparticle on which we have been working also for several years.

In brief, this work is a Systems Biology pioneer study that opens a new perspective for the development of advanced vaccines and therapies with stable epitopes against future pathogens that exhibit a high capacity to mutate and evolve. We consider that designing universal vaccines and neutralizing antibodies with antigens resilient to pathogen mutations is crucial for the development of next-generation preventive vaccines, valid for most or all future variants, and more efficient post-infection therapies.

## Acknowledgments

We would like to thank Andrea O'Malley for reading the manuscript and valuable comments. Dr. Lozano-Pérez acknowledges the European Commission ERDF/FEDER Operational Program 'Murcia' CCI Nº 2007ES161PO001 (Project No. 14-20/20). M.G acknowledges support from the NSERC Discovery grant (Canada) and the epidemiology input of M. Gotovac. This work also has received funding from the Department of Education of the Basque Government via the Consolidated Research Group MATH MODE (IT1456-22). Besides, I.M.D.F and I.M were supported by the UPV/EHU and Basque Center of Applied Mathematics, grant US21/27.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

**Author contributions:** IMDF: conceived, designed and directed the investigation; IMDF and IM peptide design; IM optimization and programming; MF data interpretation; MT-L, OT, ES, ABY, AO, VSB, and SK experimental testing; MG and AAL-P performed complementary experiments; MC, TG, VZ and MG mapping of peptides, genetic and structural analysis; JIL, GG, IM and MF data analysis; MG, AB, AGC, MJC, JP, MT-L, OT, AAL-P, SK and GPY research mapping; CB and JC-P collected data; all authors wrote the manuscript and agreed with its submission.

**Competing interests:** The authors declare no competing financial interests.

**Data and materials availability:** All data is available in the manuscript or the STAR Methods.

## REFERENCES

Ali, F., Kasry, A., and Amin, M. (2021). The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. Med. Drug. Discov. *10*, 100086. 10.1016/j.medidd.2021.100086.

Amicone, M., Borges, V., Alves, M.J., Isidro, J., Zé-Zé, L., Duarte, S., Vieira, L., Guiomar, R., Gomes, J.P., and Gordo, I. (2022). Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. Evol. Med. Public Health *10*, 142-155. 10.1093/emph/eoac010.

Chruszcz, M., Pomés, A., Glesner, J., Vailes, L.D., Osinski, T., Porebski, P.J., Majorek, K.A., Heymann, P.W., Platts-Mills, T.A.E., Minor, W., et al. (2012). Molecular determinants for antibody binding on group 1 house dust mite allergens. J. Biol. Chem. *287*, 7388-7398. 10.1074/jbc.M111.311159.

Corti, D., Purcell, L.A., Snell, G., and Veesler, D. (2021). Tackling COVID-19 with neutralizing monoclonal antibodies. Cell *184*, 3086-3108. 10.1016/j.cell.2021.05.005.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., et al. (2021). Estimated

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science *372*, eabg3055. 10.1126/science.abg3055.

De Maio, N., Walker, C.R., Turakhia, Y., Lanfear, R., Corbett-Detig, R., and Goldman, N. (2021). Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol. Evol. *13*, evab087. 10.1093/gbe/evab087.

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges *1*, 33-46. 10.1002/gch2.1018.

European Centre for Disease Prevention and Control (ECDC). SARS-CoV-2 variants of concern as of 9 June 2022. https://www.ecdc.europa.eu/en/covid-19/variants-concern.

European Medicines Agency (EMA): EMA/304600/2021 - GlaxoSmithKline use of sotrovimab (VIR-7831/GSK4182136) for the treatment of COVID-19. https://www.ema.europa.eu/en/documents/referral/sotrovimab-also-known-vir-7831-gsk4182136-covid19-article-53-procedure-assessment-report_en.pdf, (June 2022).

Ferron, F., Subissi, L., Silveira De Morais Ana, T., Le Nhung Thi, T., Sevajol, M., Gluais, L., Decroly, E., Vonrhein, C., Bricogne, G., Canard, B., and Imbert, I. (2018). Structural and molecular basis of mismatch correction and ribavirin excision from coronavirus RNA. Proc. Natl. Acad. Sci. *115*, E162-E171. 10.1073/pnas.1718806115.

Forni, D., Cagliani, R., Clerici, M., and Sironi, M. (2017). Molecular evolution of human coronavirus genomes. Trends Microbiol. *25*, 35-48. 10.1016/j.tim.2016.09.001.

Gangavarapu, K., Latif, A.A., Mullen, J.L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Zeller, M., Aceves, C.M., Zaiets, K., et al. (2022). Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Preprint at medRxiv, 10.1101/2022.01.27.2226996

Gerdol, M., Dishnica, K., and Giorgetti, A. (2022). Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. Virus Res. *310*, 198674. 10.1016/j.virusres.2022.198674.

Global Initiative on Sharing Avian Influenza Data (GISAID). https://platform.gisaid.org/epi3/frontend#

Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A., Sutherland, A., Premkumar, L., Jadi, R.S., et al. (2020). Targets of T

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell *181*, 1489-1501.e1415. 10.1016/j.cell.2020.05.015.

Gupta, A., Gonzalez-Rojas, Y., Juarez, E., Crespo Casal, M., Moya, J., Falci, D.R., Sarkis, E., Solis, J., Zheng, H., Scott, N., et al. (2021). Early treatment for Covid-19 with SARS-CoV-2 neutralizing antibody sotrovimab. N. Engl. J. Med. *385*, 1941-1950. 10.1056/NEJMoa2107934.

Gupta, A., Gonzalez-Rojas, Y., Juarez, E., Crespo Casal, M., Moya, J., Rodrigues Falci, D., Sarkis, E., Solis, J., Zheng, H., Scott, N., et al. (2022). Effect of sotrovimab on hospitalization or death among high-risk patients with mild to moderate COVID-19: a randomized clinical trial. JAMA *327*, 1236-1246. 10.1001/jama.2022.2832.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121-4123. 10.1093/bioinformatics/bty407.

Hansen, J., Baum, A., Pascal, K.E., Russo, V., Giordano, S., Wloga, E., Fulton, B.O., Yan, Y., Koon, K., Patel, K., et al. (2020). Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. Science *369*, 1010-1014. 10.1126/science.abd0827.

Hastie, K.M., Li, H., Bedinger, D., Schendel, S.L., Dennison, S.M., Li, K., Rayaprolu, V., Yu, X., Mann, C., Zandonatti, M., et al. (2021). Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: a global consortium study. Science *374*, 472-478. 10.1126/science.abh2315.

Iketani, S., Liu, L., Guo, Y., Liu, L., Chan, J.F.W., Huang, Y., Wang, M., Luo, Y., Yu, J., Chu, H., et al. (2022). Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature *604*, 553-556. 10.1038/s41586-022-04594-4.

Jones, B.E., Brown-Augsburger, P.L., Corbett, K.S., Westendorf, K., Davies, J., Cujec, T.P., Wiethoff, C.M., Blackbourne, J.L., Heinz, B.A., Foster, D., et al. (2021). The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. Sci. Transl. Med. *13*, eabf1906. 10.1126/scitranslmed.abf1906.

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T.C., Yeo, W., et al. (2021). GISAID's role in pandemic response. China CDC Weekly *3*, 1049-1051. 10.46234/ccdcw2021.255.

Kistler, K.E., Huddleston, J., and Bedford, T. (2022). Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host Microbe *30*, 545-555.e544. 10.1016/j.chom.2022.03.018.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Kubik, S., Arrigo, N., Bonet, J., and Xu, Z. (2021). Mutational hotspot in the SARS-CoV-2 spike protein N-terminal domain conferring immune escape potential. Viruses *13*, 2114. 10.3390/v13112114.

Luan, B., Wang, H., and Huynh, T. (2021). Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations. FEBS Lett. *595*, 1454-1461. 10.1002/1873-3468.14076.

Maher, M.C., Bartha, I., Weaver. S., di Iulio J, Ferri E, Soriaga L, Lempp FA, Hie BL, Bryson B, Berger B, et al., (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med. 2022 Feb 23;14(633):eabk3445. 10.1126/scitranslmed.abk3445.

Martin, D.P., Weaver, S., Tegally, H., San, J.E., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., et al. (2021). The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell *184*, 5189-5200.e5187. 10.1016/j.cell.2021.09.003.

Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., and de la Fuente, I.M. (2015). A combinatorial approach to the design of vaccines. J. Math. Biol. *70*, 1327-1358. 10.1007/s00285-014-0797-4.

Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Pérez, M.-B., and de la Fuente, I.M. (2019). Weighted lambda superstrings applied to vaccine design. PLoS One *14*, e0211714. 10.1371/journal.pone.0211714.

Mateus, J., Grifoni, A., Tarke, A., Sidney, J., Ramirez Sydney, I., Dan Jennifer, M., Burger Zoe, C., Rawlings Stephen, A., Smith Davey, M., Phillips, E., et al. (2020). Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science *370*, 89-94. 10.1126/science.abd3871.

Matyášek, R., Řehůřková, K., Berta Marošiová, K., and Kovařík, A. (2021). Mutational asymmetries in the SARS-CoV-2 genome may lead to increased hydrophobicity of virus proteins. Genes *12*, 826. 10.3390/genes12060826.

Mehra, R., and Kepp, K.P. (2022). Structure and mutations of SARS-CoV-2 spike protein: a focused overview. ACS Infect. Dis. *8*, 29-58. 10.1021/acsinfecdis.1c00433.

Minskaia, E., Hertzig, T., Gorbalenya Alexander, E., Campanacci, V., Cambillau, C., Canard, B., and Ziebuhr, J. (2006). Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc. Natl. Acad. Sci. *103*, 5108-5113. 10.1073/pnas.0508200103.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Morens, D.M., Taubenberger, J.K., and Fauci, A.S. (2021). A centenary tale of two pandemics: the 1918 influenza pandemic and COVID-19, part I. Am. J. Public Health *111*, 1086-1094. 10.2105/ajph.2021.306310.

Nextstrain. Genomic epidemiology of SARS-CoV-2 with subsampling focused globally since pandemic start. https://nextstrain.org/ncov/gisaid/global/all-time.

Obermeyer. F., Jankowiak. M., Barkas. N., Schaffner SF, Pyle JD, Yurkovetskiy L, Bosso M, Park DJ, Babadi M, MacInnis BL, et al., (2022). Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 17;376(6599):1327-1332.10.1126/science.abm1208.

Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. SARS-CoV-2 (hCoV-19) mutation reports. Lineage | Mutation tracker. https://outbreak.info/situation-reports?pango=BA.1.

Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M.M., Planchais, C., Porrot, F., Robillard, N., Puech, J., et al. (2021). Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature *596*, 276-280. 10.1038/s41586-021-03777-9.

Pomés, A., Mueller, G.A., and Chruszcz, M. (2020). Structural aspects of the allergen-antibody interaction. Front. Immunol. *11*. 10.3389/fimmu.2020.02067.

Rochman, N.D., Wolf, Y.I., Faure, G., Mutz, P., Zhang, F., and Koonin, E.V. (2021). Ongoing global and regional adaptive evolution of SARS-CoV-2. Proc. Natl. Acad. Sci. *118*, e2104241118. 10.1073/pnas.2104241118.

Rodriguez-Rivas. J., Croce G, Muscat M, Weigt M. (2022). Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. Proc Natl Acad Sci U S A. ;119(4):e2113118119. 10.1073/pnas.2113118119.

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., et al. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. Nature *584*, 120-124. 10.1038/s41586-020-2381-y.

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveillance *22*, 30494. 10.2807/1560-7917.ES.2017.22.13.30494.

Singer, J., Gifford, R., Cotten, M., and Robertson, D. (2020). CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. Preprint at https:/www.preprints.org, 10.20944/preprints202006.0225.v1.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Smith, E.C., Blanc, H., Vignuzzi, M., and Denison, M.R. (2013). Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog. *9*, e1003565. 10.1371/journal.ppat.1003565.

Soga, S., Kuroda, D., Shirai, H., Kobori, M., and Hirayama, N. (2010). Use of amino acid composition to predict epitope residues of individual antibodies. Protein Eng. Des. Sel. *23*, 441-448. 10.1093/protein/gzq014.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351. 10.1016/j.meegid.2020.104351.

Weiskopf, D., Schmitz, K.S., Raadsen, M.P., Grifoni, A., Okba, N.M.A., Endeman, H., Akker, J.P.C.v.d., Molenkamp, R., Koopmans, M.P.G., Gorp, E.C.M.v., et al. (2020). Phenotype and kinetics of SARS-CoV-2–specific T cells in COVID-19 patients with acute respiratory distress syndrome. Sci. Immunol. *5*, eabd2071. 10.1126/sciimmunol.abd2071.

World Health Organization (WHO). Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265-269. 10.1038/s41586-020-2008-3.

Wu, M., Wall, E.C., Carr, E.J., Harvey, R., Townsley, H., Mears, H.V., Adams, L., Kjaer, S., Kelly, G., Warchal, S., et al. (2022). Three-dose vaccination elicits neutralising antibodies against Omicron. The Lancet *399*, 715-717. 10.1016/S0140-6736(22)00092-7.

Zhang, L., Jackson, C.B., Mou, H., Ojha, A., Peng, H., Quinlan, B.D., Rangarajan, E.S., Pan, A., Vanderheiden, A., Suthar, M.S., et al. (2020). SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat. Commun. *11*, 6013. 10.1038/s41467-020-19808-4.

Zheng, W., Ruan, J., Hu, G., Wang, K., Hanlon, M., and Gao, J. (2015). Analysis of conformational B-Cell epitopes in the antibody-antigen complex using the depth function and the convex Hull. PLoS One *10*, e0134835. 10.1371/journal.pone.0134835.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Zhou, T., Tsybovsky, Y., Gorman, J., Rapp, M., Cerutti, G., Chuang, G.-Y., Katsamba, P.S., Sampson, J.M., Schön, A., Bimela, J., et al. (2020). Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains. Cell Host Microbe *28*, 867-879.e865. 10.1016/j.chom.2020.11.004.

# STAR+METHODS

-Methods Details
-Figures S1 to S3
-References

## Methods Details

### Systems Biology-based design of spike protein epitopes stable against future mutations

In order to obtain vaccine candidates with elevated stability against long-term mutations, we require the solutions to fulfill the $\lambda$ superstring condition (Martinez et al., 2015). In particular, with our first approach, we seek to solve the *shortest $\lambda$ superstring* problem (which generalizes the *shortest common superstring problem*, and the *set cover problem*, whose computational complexity is NP-hard). In short, we summarize the $\lambda$ superstring condition and the *shortest $\lambda$ superstring* problem as follows:

Let $A$ be a finite alphabet (in our case, formed by 20 amino acids) and $A^* = \bigcup_{n=1}^{\infty} A^n \cup \{\theta\}$ the set of all the possible strings formed by elements of $A$, where $\theta$ is the empty string. The set $A^*$ is a semigroup for the concatenation operation (denoted by $+$) where $t + t' = (t_1, \ldots, t_n) + (t_1', \ldots, t_m') = (t_1, \ldots, t_n, t_1', \ldots, t_m')$. We consider that $t = (t_1, \ldots, t_m)$ is a substring of another string $h = (h_1, \ldots, h_n)$ when $\exists k \in \{1, \ldots, n - m + 1\} |\, t_{k+i-1} = h_i, \forall i \in \{1, \ldots, m\}$. Then, the overlapping between two strings $t$ and $t'$ is defined as $overlapping(t, t') = \max\{i \in \{0,1, \ldots, min\{m, n\}\} \mid t_{n-i+j} = t'_j, for\ j = 1, \ldots, i\}$. Additionally, if we consider $T \subseteq A^*$ the set of target strings (in our case, corresponding to potential epitopes), and $\lambda \in \mathbb{N}$, we can define a $\lambda$ superstring as: let $H_1, \ldots, H_k \subseteq A^*$ and $T \subseteq A^*$, if we denote as $C(h, v)$ the set of all common substrings

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

of $h$ and $v$, a $\lambda$ superstring for the set $(H_1, \dots, H_k, T)$ is a string $v \in A^* | \, |C(H_i, v) \cap T| \geq \lambda, \forall\, i = 1, \dots, k$.

Now we can enunciate the *shortest $\lambda$ superstring* problem as: let $H_1, \dots, H_n \subseteq A^*$ be a finite set of strings of the alphabet $A$, let $T \subseteq A^*$ the set of target strings, and let $\lambda \in \mathbb{N}$, find a $\lambda$ superstring $v \in A^*$ for $(H_1, \dots, H_k, T)$ with minimum length.

**Weighed $\lambda$ superstring problem**

This concept was generalized in (Martinez et al., 2019), where we included weights for each target string (epitopes), therefore giving more importance to the most interesting ones. A weighted $\lambda$ superstring for $(H_1, \dots, H_k\,, T, w)$ is defined as a string $v \in A^*$ such as

$$\sum_{t \in C(h,v) \cap T} w(t) \geq \lambda, \forall h \in H_i, i = 1, \dots, k,$$

and solving the shortest weighted $\lambda$ superstring problem is to find, for a given $\lambda$, a weighted $\lambda$ superstring for $(H_1, \dots, H_k\,, T, w)$ of minimum length.

**Multi-stable string problem**

In order to obtain a combination of short peptides to build a multi-peptide vaccine, we have adapted the problem of weighted $\lambda$ superstring to the *Multi-stable string* problem, which is to find a group of short weighted $\lambda$ superstrings obtained sequentially, each from a smaller subset. To do this, first, we obtained a weighted $\lambda$ superstring $v_1$ for $(H_1, \dots, H_k\,, T_1, w)$; next, we removed from $T_1$ the target strings $t \in v_1 \cap T_1$, obtaining the set $T_2$, and we obtained a weighted $\lambda$ superstring $v_2$ for $(H_1, \dots, H_k\,, T_2, w)$; this procedure was repeated until we obtained the group of strings creating the final solution.

**Combinatorial optimization**

To solve the *Multi-stable string* problem, first, we approached the *shortest common superstring problem* by developing an algorithm based on Estimation of Distribution Algorithms (EDA). This family of algorithms searches for the probability distribution of the best solution to a given problem with respect to an objective function (in our case, the $\lambda$ parameter), starting with an initial distribution and evolving during the learning process, where the probability distribution is improved.
In short, the algorithm for solving the *shortest common superstring problem* goes as follows:
Given a set of strings $s_1, \dots, s_n \subseteq A^*$ over an alphabet $A$, we define the weight matrix as $W^0_{i,j} = (w^0_{i,j})$, where $w^0_{i,j} := \left(1 + \left|o(s_i, s_j)\right|\right)^{expo}$. This matrix $W^0$ will determine the initial estimation of the probability distribution of the *common superstring*, and the *expo* parameter is a control parameter of the algorithm.
Now, given a number of iterations *nit*, iterating, we build the weight matrix $W^k$ for $k = 1, \dots, nit$, which will let us estimate the probability distribution of the string of the $k + 1$ iteration. Given a population of size *spop*, we build the matrix $W^k$ by sampling *spop*

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

permutations and using the probability distribution given by $W^{k-1}$, i.e., we generate $\pi_1, \dots, \pi_{spop}$ permutations of the set $\{1, .., n\}$.

In more detail, the procedure is the following:

For each element of the population (*spop* times), we obtain a permutation $\pi_i$ as follows: first, we obtain a permutation $\pi(1)$ from $\{1, .., n\}$ using the continuous uniform distribution, and we proceed iteratively:

1. If $\pi(1), \dots, \pi(r)$ permutations (with $r \leq n-1$) have been chosen, we chose randomly a value for $b \in \{0,1\}$.
    a. If $b = 0$ (element on the left), we chose an element $u \in G = \{1, .., n\} - \{\pi(1), \dots, \pi(r)\}$ following the probability distribution $p$ defined $\forall x \in G$ as $p(x) := \frac{w_{y,\pi(1)}^{k-1}}{z}$, where $z = \sum_{y \in G} w_{y,\pi(1)}^{k-1}$. Using the value obtained for $u$, we redefine the set of permutations as $(\pi(1), \dots, \pi(r+1)) := \big(u, \pi(1), \dots, \pi(r)\big)$, being $\pi(1), \dots, \pi(r)$ the previous values.
    b. If $b = 1$ (element on the right), we chose an element $u \in G = \{1, .., n\} - \{\pi(1), \dots, \pi(r)\}$ following the probability distribution $p$ defined $\forall x \in G$ as $p(x) := \frac{w_{\pi(r),y}^{k-1}}{z}$, where $z = \sum_{y \in G} w_{\pi(r),y}^{k-1}$. Using the value obtained for $u$, we redefine the set of permutations as $(\pi(1), \dots, \pi(r+1)) := (\pi(1), \dots, \pi(r), u)$, being $\pi(1), \dots, \pi(r)$ the previous values.
2. If $\pi(1), \dots, \pi(n-1)$ permutations have been chosen and $\{u\} = \{1, .., n\} - \{\pi(1), \dots, \pi(n-1)\}$, then
    a. If $w_{u,\pi(1)}^{k-1} > w_{\pi(n-1),u}^{k-1}$, we redefine $(\pi(1), \dots, \pi(n)) := \big(u, \pi(1), \dots, \pi(n-1)\big)$.
    b. Otherwise, $(\pi(1), \dots, \pi(n)) := (\pi(1), \dots, \pi(n-1), u)$.

Once $\pi_1, \dots, \pi_{spop}$ are obtained, we build the string $t_i := StrS_{\pi_i}$ for each permutation $\pi_i$, merging (with overlap) the strings $s_{\pi(1)}, \dots, s_{\pi(n)}$, and we evaluate the length of $t_i$. Next, after fixing an acceptance *ratio*, we chose a total of $m = \lfloor ratio \cdot spop \rfloor$ permutations $\pi_{i_1}, \dots, \pi_{i_m}$ for which the lengths of $t_{i_1}, \dots, t_{i_m}$ are the smallest ones.

Finally, to obtain $W^k$, we begin assigning $W^k := W^{k-1}$, and carry out the next readjustment for $j = 1, \dots, m$ and for $l = 1, \dots, n-1$:

$$w_{\pi_{i_j}(l),\pi_{i_j}(l+1)}^{k} := w_{\pi_{i_j}(l),\pi_{i_j}(l+1)}^{k} + w_{\pi_{i_j}(l),\pi_{i_j}(l+1)}^{0}.$$

Thus, the factor $\big(1 + |o(s_i, s_j)|\big)^{expo}$ is implicitly considered. The reason for considering this factor is twofold, first, it guarantees that the described process of sampling randomly the superstrings makes sense and no division by 0 occurs. Additionally, it preserves the diversity in the population, allowing the appearance of

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

two consecutive disjoint strings. Last, the output of the algorithm is the shortest $t_i$ obtained after *nit* iterations.

The pseudocode of this algorithm is the following:

**Algorithm for solving the *shortest common superstring* problem**

**Input:** $S = \{s_1, \dots, s_n\}$, *expo*, *nit*, *spop*, *ratio*.
**Output:** A common superstring $s$ of $S$ with minimum length.
$i := 1;$
**while** $i \leq n$ **do**
    $j := 1;$
    **while** $j \leq n$ **do**
        $w_{i,j}^0 := \left(1 + \left|o(s_i, s_j)\right|\right)^{expo};$
        $j := j + 1;$
    $i := i + 1;$
$k := 1;$
**while** $k \leq nit$ **do**
$i := 1;$
**while** $i \leq spop$ **do**
    $\pi(1) := rnd_unif(1, n);$
    $r := 1;$
    **while** $r \leq n - 2$ **do**
        $b := rnd_unif(0,1);$
        **if** $b = 0$ **then**
            $G := \{1, .., n\} - \{\pi(1), \dots, \pi(r)\};$
            $u := rnd_distr(x \in P \mid p(x) := \frac{w_{y,\pi(1)}^{k-1}}{z}, z := \sum_{y \in G} w_{y,\pi(1)}^{k-1});$
            $(\pi(1), \dots, \pi(r+1)) := \left(u, \pi(1), \dots, \pi(r)\right);$
        **else**
            $G := \{1, .., n\} - \{\pi(1), \dots, \pi(r)\};$
            $u := rnd_distr(x \in P \mid p(x) := \frac{w_{\pi(r),y}^{k-1}}{z}, z := \sum_{y \in G} w_{\pi(r),y}^{k-1});$
            $(\pi(1), \dots, \pi(r+1)) := (\pi(1), \dots, \pi(r), u);$
        $r := r + 1;$
    $\{u\} := \{1, .., n\} - \{\pi(1), \dots, \pi(n-1)\};$
    **if** $w_{u,\pi(1)}^{k-1} > w_{\pi(n-1),u}^{k-1}$ **then**
        $(\pi(1), \dots, \pi(n)) := \left(u, \pi(1), \dots, \pi(n-1)\right);$
    **else**
        $(\pi(1), \dots, \pi(n)) := (\pi(1), \dots, \pi(n-1), u);$
    $\pi_i := \pi;$
    $t_i := StrS_{\pi_i};$
    $i := i + 1;$
$m := \lfloor ratio \cdot spop \rfloor;$
Take the permutations $\pi_{i_1}, \dots, \pi_{i_m}$ for which the lengths of $t_{i_1}, \dots, t_{i_m}$ are the smallest.
$W^k := W^{k-1};$
$j := 1;$
**while** $j \leq m$ **do**

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

$l := 1;$
**while** $l \leq n - 1$ **do**
$w^{k}_{\pi_{i_j}(l),\pi_{i_j}(l+1)} := w^{k}_{\pi_{i_j}(l),\pi_{i_j}(l+1)} + w^{0}_{\pi_{i_j}(l),\pi_{i_j}(l+1)};$
$l := l + 1;$
$j := j + 1;$
$k := k + 1;$
Take the shortest string $s$ obtained along the process.

Finally, before applying the algorithm over the host string, target string, and weight sets, we adapted it, first, to solve the weighted $\lambda$ superstring problem (requiring the combinatorial condition to our solutions; for more information, see Martinez et al., 2015; Martinez et al., 2019), and then, to solve the *Multi-stable string* problem (by applying the algorithm sequentially, and altering the target string set).

**Weighting the epitopes**

To select the most promising epitopes to include in the vaccine candidates, we have studied the amino acid sequences of the aforementioned 22 variants of the S-spike protein. The objective was to consider the minimum number of amino acids that make up epitopes with a possible immunogenic response. In addition, these sequences of amino acids should be recognized by the molecules of the major histocompatibility complex classes I and II. Specifically, we used a sliding window technique to extract all the possible sequences of length 9 and 15, storing the unmatched peptides. The final result of this analysis gave us a total of 3112 possible epitopes.

Next, we analyzed the class I immunogenicity levels of each of the epitopes of length 9. To do this, we used a bioinformatic tool developed by the IEDB team (http://www.iedb.org/), the main database of epitopes. Specifically, this tool allows us to classify epitopes according to their immunogenicity (Class I), which is estimated through their amino acid composition, and order of amino acids (Calis et al., 2013). We applied this tool to all the 9-mers, and ranked them from highest values to lowest. Since negative estimation meant low probabilities of being immunogenic, we only maintained the values for positive estimations and assigned 0 to all the rest.

To define the final set of potential epitopes of length 9, we evaluated the degree of affinity to the histocompatibility molecules of class I. In particular, we evaluated the affinity of class I molecules to the reference allele of IEDB, composed by 27 HLA molecules (A*01:01, A*02:01, A*02:03, A*02:06, A*03:01, A*11:01, A*23:01, A*24:02, A*26:01, A*30:01, A*30:02, A*31:01, A*32:01, A*33:01, A*68:01, A*68:02, B*07:02, B*08:01, B*15:01, B*35:01, B*40:01, B*44:02, B*44:03, B*51:01, B*53:01, B*57:01, B*58:01) (Weiskopf et al., 2013). Only those that obtained the best computational results were considered i.e., those below the 1% cutoff on the percentile rank). Finally, we removed all the 9-mers which did not pass either the immunogenicity threshold, or the HLA-I binding threshold.

To select the potential epitopes of length 15, we used the tool for class II molecules (Wang et al., 2010). Likewise, we studied the affinity of each epitope to the most representative set of HLA alleles in the population (DRB1*03:01, DRB1*07:01,

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

DRB1*15:01, DRB3*01:01, DRB3*02:02, DRB4*01:01, DRB5*01:01) (Paul, et al., 2015), and considered as potential epitopes those that obtained the best computational results (i.e., those below the 10% cutoff on the percentile rank).

In order to develop a vaccine candidate as universal as possible, epitopes obtained in the previous step, were weighted by the absolute frequency of the different HLA alleles that occur in the population. This approach would favour the selection of epitopes with higher affinity to the most prevalent histocompatibility molecules in the general population. This analysis was carried out considering experimental data (http://www.allelefrequencies.net/). The epitopes selected in this phase of the process were those that, on the one hand, could be recognized by as many different alleles as possible, and, on the other hand, had an affinity to the most frequent alleles in society, thus increasing the possibilities of vaccine success.

## Systems Biology-based analysis of the ten highly conserved antigenic sequences not selected (R1-R10).

Even if our methodological system selected eight short SARS-CoV-2 spike peptides (P1-P8), it took into account millions of possible amino acid combinations, and in particular, it evaluated and discarded the aforementioned ten highly conserved antigenic sequences. This second set of 10 S-spike peptides (R1-R10) was not selected due to the following criteria:

First, based on the 22 S-spike sequence considered, besides fulfilling the lambda-superstring criterion, another main characteristic was maximized to obtain the best epitopes. This point corresponds to the estimation of the class-I immunogenicity by the tool "T cell class-I pMHC immunogenicity predictor" (http://tools.iedb.org/immunogenicity), which classifies epitopes according to their immunogenic response, amino acid composition, and order of their amino acids sequence (Calis et al., 2013). After estimating the values of the "second set of ten stable epitopes", we observed that seven of them scored less than the smallest score obtained by our main set of eight selected stable epitopes (which scored higher than 0.153). Consequently, since a higher value is related to a higher probability of generating an immune response, those seven were outperformed by the chosen ones because of their lower estimated immunogenicity.

Therefore, only three of the "second set of ten stable epitopes (R1-R10)" were possibly immunogenically good candidates. Two of them (the sequences NITNLCPFGEVFN and KLNDLCFTNVYADSFVIRGDEV) overlapped with two of our main set of eight stable epitopes (P3 and P4). NITNLCPFGEVFN of length 13, overlapped with our selected sequence P3 (GEVFNATRFA) of length ten, was not considered because, despite of being shorter, the estimated immunogenic response of sequence P3 scored 0.312, while the sequence belonging to the "second set of ten stable epitopes" only scored 0.276.

Besides considering the score of the tool described above, we also took into account the affinity of short peptides for class I molecules (http://tools.iedb.org/mhci; Moutaftsi et al., 2006) and the affinity of long peptides to class II molecules (http://tools.iedb.org/mhcii; Wang et al., 2010). Additionally, in order to develop an antigenic set as universal as possible, each MHC molecule was weighted by the

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

absolute frequency of the different HLA alleles that occur in the population (http://www.allelefrequencies.net). Consequently, when the values obtained with these tools were taken into account, we maximized a combined score.

The 22 residue epitope KLNDLCFTNVYADSFVIRGDEV that overlapped with our identified 24 residue epitope P4 (VYADSFVIRGDEVRQIAPGQTGK) was close in estimated immunogenic response (0.481 against 0.471 of the peptide P4). However, when the combined score was taken into account, the former scored 7.136, while peptide P4 exhibited 7.778. This indicates that, if both estimated immunogenicity and binding affinities were taken into account, peptide P4 outperformed the epitope of length 22.

Next, the last sequence belonging to the set of ten stable epitopes (R1-R10), namely the 23 residue peptide RLDKVEAEVQIDRLITGRLQSLQ was compared with identified peptide P7 (ALQIPFAMQMAYRFNGIGVTQNVL). They exhibited similar length and estimated immunogenicity. When the estimated immunogenicity was calculated, the first scored 0.1666 and the second 1.586. However, the differences when considering the binding affinity in the combined score were much higher, yielding values of 6.289 for the 23-length epitope, and 8.038 for the peptide P7.

Finally, when we compared the two epitope pools (P1-P8 and R1-R10), both the estimated immunogenicity and the combined score achieved by our main set of eight stable epitopes P1-P8 (0.418±0.344, mean±std; and 4.581±3.8, respectively) were higher than the scores calculated for R1-R10 (-0.097±0.415; and 2.847±2.381, respectively).

## Analysis of the peptides

The Spike protein sequence that was used in our analysis was obtained from Uniprot (code: P0DTC2; SPIKE_SARS2), and it is shown below. The sequences corresponding to our identified peptides are highlighted.

```
         10         20         30         40         50
MFVFLVLLPL VSSQCVNLTT RTQLPPAYTN SFTRGVYYPD KVFRSSVLHS
        60         70         80         90        100
TQDLFLPFFS NVTWFHAIHV SGTNGTKRFD NPVLPFNDGV YFASTEKSNI
110        120        130        140        150
IRGWIFGTTL DSKTQSLLIV NNATNVVIKV CEFQFCNDPF LGVYYHKNNK
        160        170        180        190        200
SWMESEFRVY SSANNCTFEY VSQPFLMDLE GKQGNFKNLR EFVFKNIDGY
        210        220        230        240        250
FKIYSKHTPI NLVRDLPQGF SALEPLVDLP IGINITRFQT LLALHRSYLT
        260        270        280        290        300
PGDSSSGWTA GAAAYYVGYL QPRTFLLKYN ENGTITDAVD CALDPLSETK
        310        320        330        340        350
```

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

```
CTLKSFTVEK GIYQTSNFRV QPTESIVRFP NITNLCPFGE VFNATRFASV
       360        370        380        390        400
YAWNRKRISN CVADYSVLYN SASFSTFKCY GVSPTKLNDL CFTNVYADSF
       410        420        430        440        450
VIRGDEVRQI APGQTGKIAD YNYKLPDDFT GCVIAWNSNN LDSKVGGNYN
       460        470        480        490        500
YLYRLFRKSN LKPFERDIST EIYQAGSTPC NGVEGFNCYF PLQSYGFQPT
       510        520        530        540        550
NGVGYQPYRV VVLSFELLHA PATVCGPKKS TNLVKNKCVN FNFNGLTGTG
       560        570        580        590        600
VLTESNKKFL PFQQFGRDIA DTTDAVRDPQ TLEILDITPC SFGGVSVITP
       610        620        630        640        650
GTNTSNQVAV LYQDVNCTEV PVAIHADQLT PTWRVYSTGS NVFQTRAGCL
       660        670        680        690        700
IGAEHVNNSY ECDIPIGAGI CASYQTQTNS PRRARSVASQ SIIAYTMSLG
       710        720        730        740        750
AENSVAYSNN SIAIPTNFTI SVTTEILPVS MTKTSVDCTM YICGDSTECS
       760        770        780        790        800
NLLLQYGSFC TQLNRALTGI AVEQDKNTQE VFAQVKQIYK TPPIKDFGGF
       810        820        830        840        850
NFSQILPDPS KPSKRSFIED LLFNKVTLAD AGFIKQYGDC LGDIAARDLI
       860        870        880        890        900
CAQKFNGLTV LPPLLTDEMI AQYTSALLAG TITSGWTFGA GAALQIPFAM
       910        920        930        940        950
QMAYRFNGIG VTQNVLYENQ KLIANQFNSA IGKIQDSLSS TASALGKLQD
       960        970        980        990       1000
VVNQNAQALN TLVKQLSSNF GAISSVLNDI LSRLDKVEAE VQIDRLITGR
      1010       1020       1030       1040       1050
LQSLQTYVTQ QLIRAAEIRA SANLAATKMS ECVLGQSKRV DFCGKGYHLM
      1060       1070       1080       1090       1100
SFPQSAPHGV VFLHVTYVPA QEKNFTTAPA ICHDGKAHFP REGVFVSNGT
      1110       1120       1130       1140       1150
HWFVTQRNFY EPQIITTDNT FVSGNCDVVI GIVNNTVYDP LQPELDSFKE
      1160       1170       1180       1190       1200
ELDKYFKNHT SPDVDLGDIS GINASVVNIQ KEIDRLNEVA KNLNESLIDL
      1210       1220       1230       1240       1250
QELGKYEQYI KWPWYIWLGF IAGLIAIVMV TIMLCCMTSC CSCLKGCCSC
      1260       1270
GSCCKFDEDD SEPVLKGVKL HYT
```

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

The Uniprot entry was also used to summarize information on modification of residues that are present in the identified peptides. The summary of these modifications is listed below.

**Peptide P1**

Residues 50-73 (STQDLFLPFFSNVTWFHAIHVSGT)
N61 – glycosylation; T73 is before a glycosylation site at N74 (not in the peptide); H69-V79 are deleted for example in Omicron, this is causing increased incorporation of cleaved spike into virions.

**Peptide P2**

Residues 257-266 (GWTAGAAAYY)
No modifications for this peptide are reported in the Uniprot entry.

**Peptide P3**

Residues 339-348 (GEVFNATRFA)
G339D mutation in Omicron; N343 is glycosylated; N343Q mutation – reduced viral infectivity.

**Peptide P4**

Residues 395-417 (VYADSFVIRGDEVRQIAPGQTGK)
K417N mutation in strain: Beta/B.1.351, Gamma/P.1; May enhance affinity to human ACE2 receptor. K417N mutation in Omicron; K417T mutation in strain: Gamma/P.1.

**Peptides P5 & P6**

Peptide 5 – residues 712-722 (IAIPTNFTISV)
Peptide 6 – residues 718-727 (FTISVTTEIL)
T716I mutation is present in strain: Alpha/B.1.1.7.; N717 is glycosylated.

**Peptide P7**

Residues 893-916 (ALQIPFAMQMAYRFNGIGVTQNVL)
No modifications for this peptide are reported in the Uniprot entry.

**Peptide P8**

Residues 1208-1230 (QYIKWPWYIWLGFIAGLIAIVMV)
G1219V mutation in strain 19B/501Y.

**Peptide R1**

Residues 37-48 (YYPDKVFRSSVL).

**Peptide R2**

Residues 331-343 (NITNLCPFGEVFN)
G339D mutation is present in strains Omicron/BA.1 and Omicron/BA.2.
331N is glycosylated, and N331Q mutation reduced viral infectivity.
343N is glycosylated
Cys336 ↔ Cys361 – disulfide bridge.

**Peptide R3**

Residues 386-407 (KLNDLCFTNVYADSFVIRGDEV)
Cys391 ↔ Cys525 - disulfide bridge
D405N mutation is present in strain Omicron/BA.2.

**Peptide R4**

Residues 523-539 (VCGPKKS TNLVKNKCV)
Cys538 ↔ Cys590 – disulfide bridge.

**Peptide R5**

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Residues 586-603 (DITPCSFGGVSVITPGTN)
Cys538 ↔ Cys590 – disulfide bridge
N603 is glycosylated.

**Peptide R6**

Residues 797-808 (FGGFNFSQILPD)
N801 is glycosylated.

**Peptide R7**

Residues 910-921 (GVTQNVLYENQK).

**Peptide R8**

Residues 962-977 (LVKQLSSNFGAISSVL)
N969K mutation is present in strains Omicron/BA.1 and Omicron/BA.2.

**Peptide R9**

Residues 983-1005 (RLDKVEAEVQIDRLITGRLQSLQ)
Residues K986 and V987 are mutated to prolines (in vaccine BNT162b2/Pfizer-Biontech, vaccine mRNA-1273/Moderna, vaccine Ad26.COV2.S/Janssen Pharmaceutical), which causes stabilization in prefusion state.

**Peptide R10**

Residues 1028-1039 (KMSECVLGQSKR)
Cys1032 ↔ Cys1043 – disulfide bridge.

## Aminoacid lengths of SARS-CoV-2 mutation-stable peptide regions.

The amino acid length of the eight selected epitopes (P1-P8) was 16.875±7.1 (mean ± SD), with a confidence interval $I_{\mu}^{0.95} = (10.94, 22.81)$ while the length for the ten highly conserved peptide sequences (R#1-R#10) was 15.7±4.24, with a confidence interval $I_{\mu}^{0.95} = (12.66, 18.74)$. Next, we analyzed if there were significant differences between the length of both groups, obtaining a confidence interval of $I_{\mu_1-\mu_2}^{0.95} = (-5.1, 7.45)$, a $t - statistic = 0.413$ and $a\,p - value = 0.688$: there are no significant differences between the lengths of the group of eight selected epitopes (P1-P8), and the group of ten highly conserved ones (R1-R10). The 18 mutation-stable regions in total have an average amino-acid length of 16.22±5.54 (mean ± SD), with a confidence interval $I_{\mu}^{0.95} = (13.47, 18.98)$.

# Structural analysis

Analysis of the S-spike protein structures as well as structures of complexes with therapeutic antibodies was performed using COOT (1) and Pymol (2). Pymol was used to prepare Fig. 5. PDBePISA and structures of antibodies complexed with the Spike protein were used to identify residues forming epitopes. The summary of this analysis with information on overlaps between the peptides and the identified epitopes by us is presented below.

**Bamlanivimab** blocks ACE2 and binds to the spike protein RBD in “up” and “down” conformations. Analysis of the structure (PDB code: 7KMG) indicates that there is no overlap between Bamlanivimab binding epitope and the peptides P1-P8.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

**Etesevimab -** Analysis of the structure (PDB code: 7C01) using PDBePISA shows that the epitope includes residues 415-417, 420, 421, 453, 455-460, 486, 487, 489, 490, 493 that are interacting with heavy chain, as well as residues **403**, **405**, **406**, 408, 409, 449, 453, 493-496, 498, 500-502, 504 and 505, which interact with the light chain. Residues highlighted in red are present in peptide P4. Residues marked in bold are present in peptide R3.

**Casirimab (REGN10933)-** Analysis of the structure (PDB code: 6XDG) indicates that the epitope includes residues: **403**, **406**, 417, 421, 473, 475-478, 484-490, 492-496 and 501. There is a partial overlap between the residues forming the epitope and residues from peptide P#4 (**403**, **406** & 417). Residues highlighted in red are present in peptide P4. Residues marked in bold are present in peptide R3.

**Imdevimab (REGN10987) - Analysis of the structure (**PDB code: 6XDG) shows that the epitope includes residues: 346, 439-441, 443-450, 498, 499, 500 and 501. There is no overlap between Imdevimab binding epitope and the peptides that we have identified.

**Sotrovimab –** Analysis was done using the structure of S309, (Sotrovimab precursor, PDB code: 6WPT). The epitope includes residues **333-337**, **339-341**, **343**-346, 440, 441 and 509. Peptide P3 (residues 339-348) has a significant overlap with Sotrovimab (S309) epitope and residues highlighted in red are present in peptide P3. Residues marked in bold are present in peptide R2.

## Sequence variation analysis

To test the stability of the epitopes against mutations, we analyzed publicly available variation data from GISAID (Khare et al., 2021). These include 3,362 complete genomes, across 1,514 SARS-CoV-2 lineages, sampled between December 2019 and January 2022 available at NextStrain (Hadfield et al., 2018), from which the entropy per site, and the number of mutational events across the strain tree are computed. Likewise, the protein variation analysis of GISAID hCoV-19 sequences, available at CoV-GLUE database (Singer et al., 2020), have been taken into account in our study. Data from CovGlue was downloaded for all SARS-CoV-2 variants separately and disregarding mutations with a frequency lower than 0.0001. The probability for a peptide of being invariable (e.g. identical to the reference protein sequence) in a given SARS-CoV-2 variant was approximated by assuming independence of the sites and multiplying the expected non-mutated frequencies (1 - mutation frequency). We compared these results with 10,000 random samples of seven non-overlapping peptides of the same size as the ones selected by our algorithm (note that peptides 5 and 6 overlap and are considered as a single peptide spanning 712-727 in this analysis). The amino acid composition for these combinations of peptides was computed in terms of amino acid frequencies (%) and compared to the selected epitopes. Significant deviations were considered when observed values in the selected epitope set was higher or lower than the 95% and 5% percentile values in the distribution obtained for the 10,000 random peptide samples.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

## ELISPOT assay

The experiments were performed at Cellular Technology Ltd (CTL, Cleveland, USA). Designed peptides were synthesized at Mimotopes Pty Ltd (Australia). The peptides were diluted at CTL and studied in the IFNγ detecting single color- and IFNγ/IL5/IL-17 detecting triple color ELISPOT assays. Peptides P1 through P8 were tested individually, at different concentrations (100 µg/ml, 25 µg/ml and 2.5 µg/ml) and in a pool of eight peptides. The ELISPOT assay was performed using CTL ELISPOT protocols and CTL kits for the single color enzymatic (Cat. # hIFNg-1M-red) and triple color ELISPOT fluorescent (Cat. # hT3015F/275/hT58/hT32) assays.
**Procedure.** Briefly, ELISPOT plates were coated with the appropriate coating antibody and stimulated for 24h with peptides in the single color ELISPOT assay and for 72h in the triple color ELISPOT assay. Medium alone served as the negative control and PHA (tested at 5 µg/ml) served as the positive control. At the end of the stimulation the cells were discarded from the plates and the corresponding detecting antibodies were added for the overnight incubation at 4ºC. Afterwards, various tertiary reagents were added for the overnight incubation at 4ºC. The ImmunoSpot® UV Analyzer/Cell Counter S6 Ultimate (CTL) was used to quantify the spot forming units. The frequency of peptide-reactive cells is expressed as spot forming units (SFU) for 400,000 cells.
An antigen specific positive response was defined as = X - (SD of X) – [Y + (2SD of Y)] and is greater than the numerical value 3.
X: Average Spot # induced by a given Ag (cells and antigen)
Y: Average Spot # in the medium control (cells and no antigen)
O: Average Spot # in the negative control (no cells and no antigen).
Positive response was determined as greater than (Y + O + 2SD of Y) number of spots after exposure to Ag (X) and more than 10 spots.
In Cellular Technology Ltd experiments, all donors provided written informed consent, Institutional Review Board (Pro00043178).

## ELISA assay

Sera from healthy controls and severely affected patients 3 months after the diagnosis of SARS-CoV-2 infection were obtained at the Shaare Zedek Medical Center (SCMZ) in Jerusalem, Israel, and tested for specific IgG antibodies in the ELISA assay. The experiments were performed at Department of Life Sciences and the National Institute for Biotechnology in the Negev, Ben-Gurion University, Beer-Sheva, Israel. Designed peptides were synthesized at Mimotopes Pty Ltd (Australia). Reagents: SARS-CoV-2 Spike RBD (40592-V08B, Sino Biological Inc.); Donkey Anti-Human IgG, Fc HRP (709-035-098, Jackson ImmunoResearch Europe Ltd); BSA (Thermo Fisher). Samples were diluted at 1:25 in PBS. The protocols indicated by the company (Thermo Fisher) were followed. The virus present in the samples was inactivated by incubating the samples at 60 degrees Celsius for 30 minutes. 96-well ELISA plates were coated overnight at 4 degrees with 100 µL of Cov S-spike RBD at 0.5 µg/mL or the different peptides at 10 µg/ml in coating buffer (0.1 M $Na_2HPO_4$ pH=9). The next day, the plates were washed three times with washing buffer PBST (PBS with tween at 0.1% w/v). The wells were blocked with 1% BSA in PBS for 1 hour at 37 degrees. Then, the plates were washed with PBST 3 times. 100 µL of the samples diluted 1:400 in 0.1% BSA in PBS were added and incubated for 1 hour at 37 degrees. The wells were washed six

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

times. Donkey Anti-Human IgG, Fc HRP was added at 1:5000 in 0.1% BSA -PBS, and the plates were incubated for 1 hour at 37 degrees. The wells were washed six times. 3,3',5,5'-tetramethylbenzidine solution was added as a chromogenic substrate. The samples' absorbance was measured at 650 nm 10 minutes after in a Spark plate reader (Tecan). The study was approved by the ADVARRA. In SZMC the study was approved by the institutional review board of SZMC (permit 0181-20-SZMC).

## References

Calis, J.J.A., Maybeno, M., Greenbaum, J.A., Weiskopf, D., De Silva, A.D., Sette, A., Keşmir, C., and Peters, B. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. PLoS Comp. Biol. *9*, e1003266. 10.1371/journal.pcbi.1003266.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121-4123. 10.1093/bioinformatics/bty407.

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T.C., Yeo, W., et al. (2021). GISAID's role in pandemic response. China CDC Weekly *3*, 1049-1051. 10.46234/ccdcw2021.255.

Martínez, L., Milanič, M., Legarreta, L., Medvedev, P., Malaina, I., and de la Fuente, I.M. (2015). A combinatorial approach to the design of vaccines. J. Math. Biol. *70*, 1327-1358. 10.1007/s00285-014-0797-4.

Martínez, L., Milanič, M., Malaina, I., Álvarez, C., Pérez, M.-B., and M. de la Fuente, I. (2019). Weighted lambda superstrings applied to vaccine design. PLoS One *14*, e0211714. 10.1371/journal.pone.0211714.

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D.C., Sidney, J., Bui, H.-H., Grey, H., and Sette, A. (2006). A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. Nat. Biotechnol. *24*, 817-819. 10.1038/nbt1215.

Paul, S., Lindestam Arlehamn, C.S., Scriba, T.J., Dillon, M.B.C., Oseroff, C., Hinz, D., McKinney, D.M., Carrasco Pro, S., Sidney, J., Peters, B., and Sette, A. (2015). Development and validation of a broad scheme for prediction of HLA class II restricted T cell epitopes. J. Immunol. Methods *422*, 28-34. 10.1016/j.jim.2015.03.022.

Singer, J., Gifford, R., Cotten, M., and Robertson, D. (2020). CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. Preprint at https:/www.preprints.org, 10.20944/preprints202006.0225.v1

The Allele Frequency Net Database (AFND). http://www.allelefrequencies.net/

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

The Immune Epitope Database (IEDB) Analysis resource: peptide binding to MHC class I molecules prediction tool. http://tools.iedb.org/mhci/

The Immune Epitope Database (IEDB) Analysis resource: peptide binding to MHC class II molecules prediction tool. http://tools.iedb.org/mhcii/

The Immune Epitope Database (IEDB) Analysis resource: T cell class I pMHC immunogenicity predictor. http://tools.iedb.org/immunogenicity/

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., and Peters, B. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics *11*, 568. 10.1186/1471-2105-11-568.

Weiskopf, D., Angelo Michael, A., de Azeredo Elzinandes, L., Sidney, J., Greenbaum Jason, A., Fernando Anira, N., Broadwater, A., Kolla Ravi, V., De Silva Aruna, D., de Silva Aravinda, M., et al. (2013). Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. Proc. Natl. Acad. Sci. *110*, E2046-E2053. 10.1073/pnas.1305227110.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

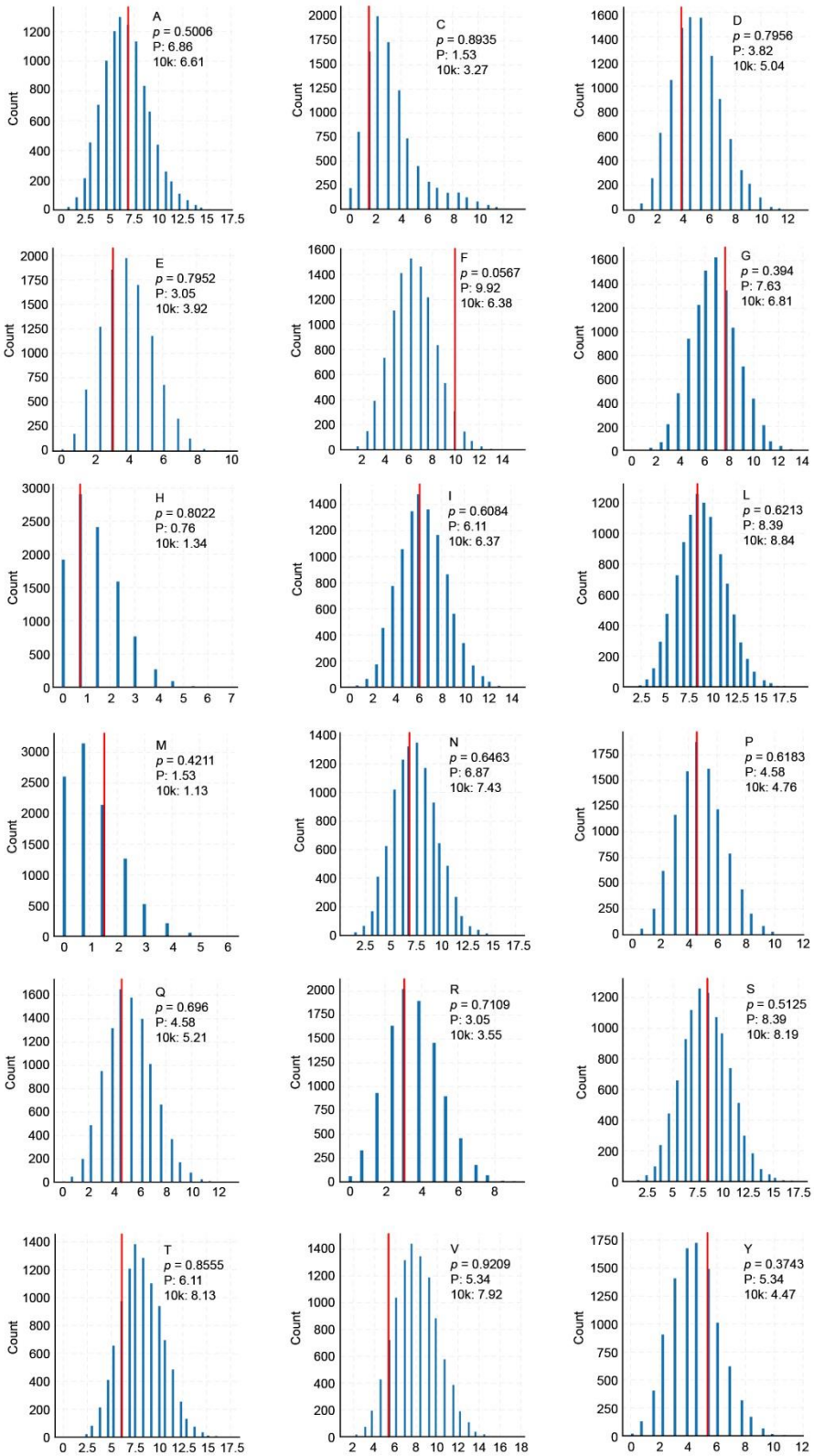



**Figure S1. Distribution of amino acid frequencies in random sets of peptides from the S-protein**
X axis represents the frequency (in percentage) of the indicated amino acid (on letter code, in the top right of each distribution), each bar represents the count of sets of random peptides showing a given frequency across 10,000 random samples (see methods). The vertical red line indicates the frequency in the set of peptides selected by our algorithm. p, indicates the probability of a random peptide set to have a frequency higher than the one of the selected peptide set; P indicates the frequency of the indicated amino acid in the selected peptide set, 10k indicates the average frequency of the indicated amino acid in the 10,000 random peptide sets.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

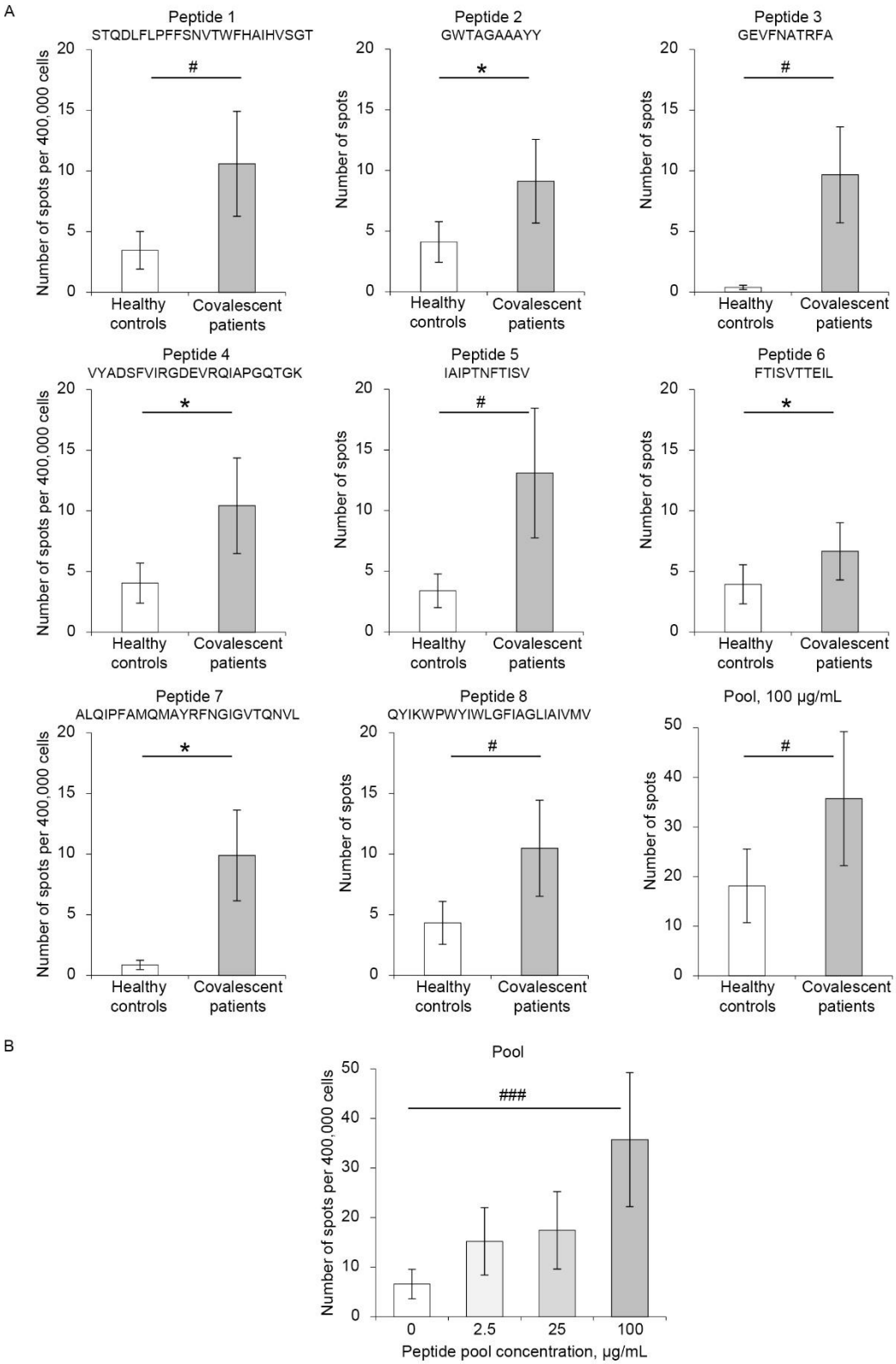



**Figure S2. Recall response to the selected peptides (P1-P8)**

(A) IFNγ ELISPOT data analysis for individual and pooled peptides at a concentration of 100 µg/ml. (B) IFNγ ELISPOT data analysis for pooled peptides (P1-P8) at different concentrations. Immune response of PBMCs samples from convalescent COVID-19 patients and from unexposed donors (with the blood drawn in 2016 and 2017), stimulated in vitro with designed peptides for 24 hours was studied in the ELISPOT assay. The frequency of peptide-reactive cells, expressed as spot-forming units for 400,000 cells, is shown by circles for each donor and the pointed lines represent the median frequency. p values derived from Mann–Whitney's *U*-test and Kruskal-Wallis's test: * >0.05; # <0.05; ## <0.01; ### <0.005.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.13.22280980; this version posted October 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

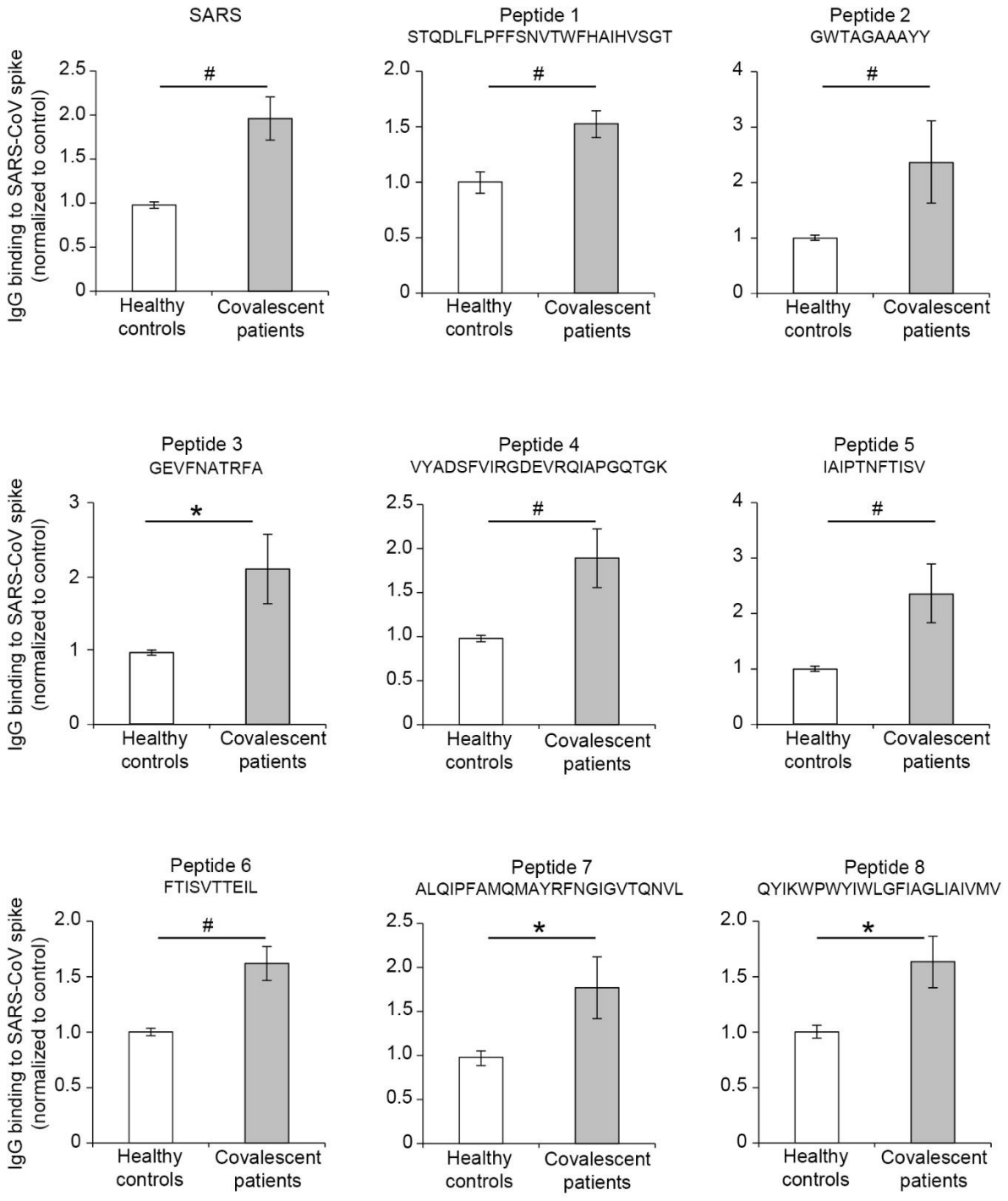



**Figure S3. Binding of IgG from controls and positives for Covid-19 to SARS-CoV-2 S-spike RBD and selected peptides (P1-P8)**

First panel shows IgG binding to SARS-CoV-2 S-spike RBD. The rest of the panels illustrate IgG binding to each of eight peptides (P1-P8). The data are expressed relative to the control in each case. p values derived from Mann–Whitney's *U*-test: * >0.05; # <0.05.