bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Design of T cell epitope-based vaccine candidate for SARS-CoV-2 targeting nucleocapsid and spike protein escape variants

Gabriel Jabbour[1], Samantha Rego[2], Vincent Nguyenkhoa[3], Sivanesan Dakshanamurthy[4, *].

[1]School of Medicine, Georgetown University Medical Center, Washington DC 20057.

[2]Georgetown University, Washington DC 20057.

[3]University of Virginia, Charlottesville, VA.

[4]Molecular and Experimental Therapeutic Research in Oncology (METRO) Program, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington DC 20057.

*Corresponding Author

Dr. Sivanesan Dakshanamurthy, Ph.D, MBA.

e-mail: sd233@georgetown.edu.

Phone: (202) 687 2347

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The current COVID-19 pandemic continues to spread and devastate in the absence of effective treatments, warranting global concern and action. Despite progress in vaccine development, the rise of novel, increasingly infectious SARS-CoV-2 variants makes it clear that our response to the virus must continue to evolve along with it. The use of immunoinformatics provides an opportunity to rapidly and efficiently expand the tools at our disposal to combat the current pandemic and prepare for future outbreaks through epitope-based vaccine design. In this study, we validated and compared the currently available epitope prediction tools, and then used the best tools to predict T cell epitopes from SARS-CoV-2 spike and nucleocapsid proteins for use in an epitope-based vaccine. We combined the mouse MHC affinity predictor and clinical predictors such as HLA affinity, immunogenicity, antigenicity, allergenicity, toxicity and stability to select the highest quality CD8 and CD4 T cell epitopes for the common SARS-CoV-2 variants of concern suitable for further preclinical studies. We also identified variant-specific epitopes to more precisely target the Alpha, Beta, Gamma, Delta, Cluster 5 and US variants. We then modeled the 3D structures of our top 4 N and S epitopes to investigate the molecular interaction between peptide-MHC and peptide-MHC-TCR complexes. Following in vitro and in vivo validation, the epitopes identified by this study may be used in an epitope-based vaccine to protect across all current variants, as well as in variant-specific booster shots to target variants of concern. Immunoinformatics tools allowed us to efficiently predict epitopes in silico most likely to prove effective in vivo, providing a more streamlined process for vaccine development in the context of a rapidly evolving pandemic.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

In late 2019, a novel coronavirus was detected in patients in Wuhan, China, and was later classified as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The disease caused by SARS-CoV-2 infection, Coronavirus Disease 2019 (COVID-19), was declared a pandemic by the WHO in March 2020 and has since rapidly spread across the globe. The effects of this pandemic have been devastating, causing over 218 million infections and 4.5 million deaths worldwide as of September 2, 2021 [1], and these numbers continue to grow. While treatment options remain limited, the development of effective vaccines is crucial to curbing the spread of SARS-CoV-2 and combating the COVID-19 pandemic. The FDA has provided Emergency Use Authorizations for two vaccines as well as full approval for one vaccine, all of which are helping to lessen the toll of COVID-19 [2,3,4]. However, the rise of new, increasingly infectious variants along with a lack of global vaccine access means that the devastating effects of COVID-19 continue to be felt worldwide. It is therefore necessary to continue development of new vaccines that can combat current variants, as well as establishing a streamlined process for the rapid development of effective vaccines against novel variants. We believe that a viral epitope-based vaccine could fulfill this need as an alternative to the current vaccines. An epitope-based vaccine may be able to more precisely target certain mutations of variants of concern, as well as allowing for swift development of new vaccines through immunoinformatics as novel variants emerge in the future.

Immunoinformatics provides a more rapid, efficient, and reliable approach to vaccine design, critical to increase the speed of vaccine development in the context of a swiftly spreading pandemic. Specifically, immunoinformatics offers a cost- and time-effective strategy for designing epitope-based vaccines, which have several advantages over conventional types of vaccines. They are more specific, able to generate long-lasting immunity, and can minimize adverse reactions [5]. Given these advantages and the urgency of the current COVID-19 pandemic, we sought to predict immunogenic T cell epitopes of SARS-CoV-2 for an epitope-based vaccine candidate against the common SARS-CoV-2 variants, including the variants of concern (VOC).

T cell epitopes are recognized by Major Histocompatibility Complex (MHC) molecules on the surface of particular cells, which display the antigen to T cells, necessary for inducing an immune response. Several tools have been developed for the prediction of T cell epitopes, many of which are available on the freely available Immune Epitope Database [6]. However, the accuracy of these tools in predicting immunogenic epitopes is widely variable [7]. Therefore, to ensure that the prediction tool used in our analysis was the most accurate available on IEDB.org, we systematically analyzed the available T cell epitope prediction tools by comparing their match rate with experimentally determined

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

immunogenic SARS-CoV-2 epitopes. This allowed us to determine the best performing epitope prediction tool for our own analysis. In addition, our analysis also validates the epitope prediction tools available on IEDB for use in predicting SARS-CoV-2 epitopes and provides guidance on how to use this resource to predict high quality SARS-CoV-2 epitopes. There are two classes of MHC molecules, MHC class I, which are expressed on all nucleated cells and activate cytotoxic T cell lymphocytes (CD8+ T cells), and MHC class II, which are expressed on antigen presenting cells and activate helper T cell lymphocytes (CD4+ T cells). To ensure our vaccine would induce both a CD8+ and CD4+ T cell response, we included epitopes that can be recognized by both MHC-I and MHC-II molecules. This is necessary to confer long-term protective immunity by inducing cytotoxic T cells and neutralizing antibodies [8]. Therefore, we used the best performing epitope prediction tools to predict epitopes of the SARS-CoV-2 Nucleocapsid (N) and Spike (S) proteins predicted to bind both MHC class I and class II.

We predicted epitopes from the SARS-CoV-2 structural proteins, namely N and S protein, for our HLA peptide-based vaccine design due to their role as major antigens of SARS-CoV-2, likely to be immunogenic. In convalescent COVID-19 patients, a robust T-cell immunity against the N protein has been detected and N protein epitopes have been found to induce responses in T cells derived from recovered patients [9]. The spike (S) glycoprotein has been reported to be a crucial surface protein of SARS-CoV-2 that facilitates viral entry into the host cell. This makes it a key target for antiviral efforts, and all current COVID-19 vaccines contain the S protein. DNA vaccines containing the S gene derived from SARS-CoV-1 have also been shown to induce T-cell responses in mice [10,11]. These proteins therefore provide ideal candidates for sources of immunogenic epitopes needed for an effective vaccine.

While the currently approved vaccines have been shown to offer effective protection against the original SARS-CoV-2 strain in clinical trials [12,13,14], multiple variants that differ from those present during clinical trials are now spreading worldwide and continue to emerge. Of current concern is the Delta (B.1.617.2) variant of SARS-CoV-2. Even countries with high vaccination rates against COVID-19, such as Israel, have seen increases in new COVID-19 cases as the variant spreads [15]. Data from Israel's health ministry suggests that the Pfizer-BioNTech vaccine is much less effective at preventing infection by the Delta SARS-CoV-2 variant than previous strains, reporting a 39% efficacy rate compared with 95% against the original strain [15]. This raise concerns that the current vaccines, designed for previous strains, may be less effective against current and future variants. We believe it is necessary to develop vaccines that can target the mutations of variants of concern, potentially for use as variant-specific booster shots. Therefore, we sought to aid in designing these epitope-based booster vaccines that target certain variants by identifying epitopes specific to the mutations of variants of concern.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In this study, we combined careful analysis of the available epitope prediction tools, epitope prediction across multiple SARS-CoV-2 antigens, MHC binding properties, and variants, as well as HLA population coverage analysis to find immunogenic CD8 and CD4 T cell epitopes for all SARS-CoV-2 variants of concern. This epitope combination presents a candidate for use in an effective epitope-based vaccine that, following experimental validation, could aid in the fight against the COVID-19 pandemic. The variant-specific epitopes identified could also be useful in designing booster shots to target particular variants. This streamlined process may prove an efficient and reliable approach to vaccine development in fighting new outbreaks as new variants arise.

Methods

Retrieval of SARS-CoV-2 sequence

The SARS-CoV-2 reference N and S protein sequences (accession numbers YP_009724397 and YP_009724390 respectively) [16] were downloaded from the NCBI GenBank. The lineage-defining N and S protein mutations for the common variants: alpha variant (VOC-202012/01, B.1.1.7, UK variant), Beta variant (VOC-202012/02, B.1.351, South-African variant), Gamma variant (VOC 202101/02, P.1, Brazil-Japan variant) [17], US variants (N [18] and S [19] protein mutations), Delta variant (VUI B.1.617, Indian variant) [20] and Cluster 5 mink variant [21] were used to create the variant specific N and S lineage-defining protein sequences.

Retrieval of N protein SARS-CoV-2 experimentally determined epitopes

<u>CD8 experimental epitopes</u>

An experimental epitope is selected if its immunogenicity has been experimentally determined and its length is between 8 and 14 amino acids [22]. We selected 67 epitopes from various papers [23-31], 6 from RCSB.org (7KGT, 7LG2, 7LG3, 7LFZ, 7KGR and 7KGS) and 15 epitopes from IEDB.org and ViPR.org with positive T cell assay and 100% match with the SARS-CoV-2 N reference sequence, for a total of 88 high quality epitopes with experimentally determined immunogenicity (Supplementary Material 1).

<u>CD4 experimental epitopes</u>

To find experimentally determined epitopes, we searched for SARS-CoV-2 N protein epitopes whose immunogenicity had been validated experimentally. We selected 28 epitopes from previous papers that had been found to be immunogenic in convalescent COVID-19 patients [27, 31-33] and 20 epitopes

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from ViPR with positive assay and 100% match with the SARS-CoV-2 N protein reference sequence. We also searched the RCSB Protein Data Bank for known SARS-CoV-2 N protein epitopes binding to MHC class II, which returned no results. This provided a total of 48 high quality epitopes with experimentally determined immunogenicity (Supplementary Material 1).

Choosing the best tool

The 5 common CD8 T cell epitope prediction tools available on IEDB: ANN 4.0 [34-39], IEDB consensus [40], IEDB recommended for proteasome-TAP-MHC1 binding combined predictor (IEDB Prot-TAP-MHC) [41,42], NetMHCpan EL 4.1 (IEDB recommended 2020.09) and NetMHCpan BA 4.1 [43-47] and the 4 common CD4 T cell epitope prediction tools available on IEDB: IEDB recommended 2.22, IEDB consensus 2.22 [48,49], NetMHCIIpan EL 4.0 and NetMHCIIpan BA 4.0 [50,51] were used to predict SARS-CoV-2 N protein epitopes. Only MHC alleles with a frequency > 1% were selected. For MHCII epitope prediction, the default size of 15 amino acids per epitope was selected.

The duplicate-free peptides were first ordered according to each of the available attributes in each tool, then matched to the experimental epitopes according to 3 stringency levels: lenient, intermediate and stringent (Supplementary Material 1). Receiver operating characteristic (ROC) analysis was used to assess the quality of select peptide prediction tools. ROC evaluates binary data as positive and negative values, which was achieved by having defined thresholds based on the predicted peptides' binding affinity. It is important to note that the measure of binding affinity (rank, IC50, score, or percentile rank) varied amongst prediction tools, which was an additional consideration when determining the best tool. The positive and negative experiment test set was obtained by assessing each tool based on the matches of predicted epitopes to experimentally determined epitopes. More specifically, a match between a predicted and experimental epitope based on the Lenient, Intermediate, or Stringent method was a positive (1), while a non-match was a negative (0).

To determine the quality of each prediction tool, specificity and sensitivity calculations were needed for the construction of the ROC curves. The tool that maximizes both specificity and sensitivity would be regarded as the best tool. This is more easily quantified by measuring the area under the curve (AUC), where an area of 0.5 indicates an entirely random prediction and an area of 1.0 indicates a perfect prediction. All analysis and data visualizations were constructed from R scripts. The pROC package was used for ROC and AUC calculations.

Epitope prediction

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Using the best tool-attribute combination, we predicted CD8 and CD4 T cell epitopes for both N and S proteins for each variant.

MHCI epitope immunogenicity prediction

Immunogenicity scores for the top 300 epitopes from each variant were calculated using the MHC I immunogenicity tool in IEDB.org [52]. Only CD8 common epitopes with an immunogenicity score > 0 and CD8 variant-specific epitopes with an immunogenicity score > - 0.5 were selected for further analysis.

MHCII epitope antigenicity prediction

Vaxijen 2.0 [53-55] was used to select the best CD4 epitopes due to the lack of a reliable immunogenicity prediction tool on IEDB.org.

Toxicity prediction

Using ToxinPred [56], CD8 and CD4 peptides were found to be either toxic or non-toxic. Only non-toxic epitopes were selected for further studies.

Allergenicity prediction

AllerTop 2.0 [57,58] was used to predict allergenicity of both CD8 and CD4 epitopes. Only “probable non-allergenic” epitopes were selected.

Instability index prediction

ProtParam [59] was used to predict the instability index [60] to determine whether an epitope was stable or unstable in vivo.

Mouse MHC affinity prediction

NetH2pan 4.0 [61,62] was used to predict mouse MHC affinity of CD8 epitopes using MHCI alleles of the most commonly used mouse strains to study COVID-19 vaccines: C57BL/6 and BALB/c rodents. Only Weak Binding and Strong Binding epitopes were selected for potential further preclinical studies.

Population coverage analysis

The population coverage analysis tool available on IEDB.org [63] was used to predict the percentage of the world population predicted to present the epitopes with known MHC restrictions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

p-MHC 3D structure prediction

The tertiary structure of the selected epitopes was predicted using PEPstrMOD [64,65]. The 3D structures of MHC molecules were downloaded from Protein Database Bank on RCSB.org and processed on PyMOL to remove the epitope. The predicted epitope along with the empty MHC molecule were energy minimized using Swiss-PdbViewer [66]. The two molecules were then docked using FlexPepDock online server [67,68].

Results

Choosing the best tool

For the prediction of common SARS-CoV-2 CD8 epitopes for both N and S proteins, the results suggested the use of IEDB recommended (NetMHCpan EL 4.1) ordered according to “rank”, as opposed to “score” as recommended by IEDB. However, this difference was not statistically significant. IEDB recommended ranked first in AUC at 0.7423 and 0.9192 for the Lenient and Stringent method, respectively (Figure 2). For the prediction of common SARS-CoV-2 CD4 epitopes, there was no statistically significant difference between the tools, so we decided to use the IEDB recommended ordered according to “rank,” since it was a top performer with AUC of 0.6969, 0.7078 and 0.767 for the Lenient, Intermediate and Stringent methods respectively (Figure 3).

CD8 epitope prediction

Using NetMHCpan EL 4.1, MHCI immunogenicity predictor, ToxinPred, AllerTop 2.0, we only selected CD8 epitopes that were immunogenic, non-toxic and non-allergenic. We obtained 31 common N and 31 common S protein epitopes across all variants (Supplementary Table S1), 1 N and 2 S epitopes specific for the beta variant, 2 N and 4 S epitopes specific for the gamma variant, 2 N and 4 S epitopes specific for the alpha variant, 3 N epitopes specific for the US-Australian variants, 2 S epitopes specific for the US variant, 2 S epitopes specific for the Cluster 5 variant and 5 S epitopes specific for the delta variant and 1 S epitope specific for both the delta and US variants (Supplementary Table S3 and Supplementary Material 2).

Adding the stability filter, we obtained 15 N and 21 S CD8 epitopes in total (Supplementary Table S2), 1 N epitope specific for the alpha variant, 1 S epitope specific for the beta variant, 3 S epitopes specific for the gamma variant, 4 S epitopes specific for the alpha variant, 2 S epitopes specific for the US variant, 2 S epitopes specific for the Cluster 5 variant, 1 S epitope specific for the delta variant and 1 S epitope specific for both the US and delta variants (Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CD4 epitope prediction

Using IEDB recommended 2.22, Vaxijen 2.0, ToxinPred and AllerTop 2.0, we only selected the antigenic non-toxic and non-allergenic CD4 epitopes. We obtained 5 common N and 11 common S protein epitopes across all variants (Supplementary Table S4), 1 S epitope specific for the beta variant, 3 N and 8 S epitopes specific for the gamma variant, 4 S epitopes specific for the alpha variant, 1 N epitopes specific for the US-Australia variant, 5 S epitopes specific for the US variant, 3 N and 7 S epitopes specific for the delta variant, 1 S epitope common to the delta and US variants, and 3 S epitopes specific for the Cluster 5 variant (Supplementary Table S5 and Supplementary Material 3).

With stability prediction, we obtained 1 common N and 7 common S epitopes (Table 3), 1 S epitope specific for the beta variant, 7 S epitopes specific for the gamma variant, 1 S epitope specific for the alpha variant, 4 S epitopes specific for the US variant, 3 N and 3 S epitopes specific for the delta variant, 1 S epitope common to the delta and US variants, and 3 S epitopes specific for the Cluster 5 variant (Table 4).

Population coverage analysis

The 31 N and 31 S CD8 epitopes cover 98.46% and 92.02% of the world population respectively. However, since epitope-based vaccine constructs usually contain fewer epitopes, we also found the regional and world coverage of our top 11 N and top 15 S CD8 epitopes to be 97.87% and 89.92% respectively (Table 1, Figure 4 and Supplementary Material 4). These 23 highly immunogenic, non-toxic and non-allergenic peptides are predicted to elicit a protective immune response in 98.42% of the world population. After adding the instability index prediction and filtering out the unstable epitopes, the population coverage decreased to 78.28% and 87.13% for N and S protein epitopes respectively (Supplementary Table S2).

The 5 N and 11 S CD4 epitopes have world population coverages of 45.40% and 97.11% respectively (Supplementary Table S4). After adding the stability prediction, the top 1 N and top 7 S CD4 epitopes cover 34.55% and 95.13% of the world population respectively (Table 3 and Figure 4). Together, the top 8 N and S CD4 epitopes cover 96.81% of the world.

p-MHC 3D structure prediction

Next, 3 N (NTASWFTAL, TPSGTWLTY and LPNNTASWF) and 1 S (VVFLHVTYV) immunogenic, non-toxic, non-allergenic common CD8 epitopes with significant affinity to at least one mouse MHC haplotype were selected for 3D modeling. These epitopes were docked with their respective predicted HLA alleles: NTASWFTAL-HLA-A*68, TPSGTWLTY-HLA-A*01:01, LPNNTASWF-HLA-B*07:02 and

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VVFLHVTYV-HLA-A*02:01 (Figure 5). Furthermore, the VVFLHVTYV-HLA-A*02:01 complex was superimposed with the 7N1E RCSB structure to get an insight into the interaction between the p-MHC complex and the pRLQ3 TCR molecule (Figure 5 and Supplementary Figure S1). To avoid clash the p-MHC-pRLQ3 TCR complex was visually analyzed, and energy-minimized with the SANDER module of AMBER [69].

Discussion

Immunoinformatics provides an efficient and reliable strategy for designing epitope-based vaccines to combat the evolving COVID-19 pandemic, especially as new SARS-CoV-2 variants emerge and spread. However, with the ever-increasing number of machine learning tools in the field of immunoinformatics, the selection of epitope prediction tools is becoming increasingly arbitrary. It is necessary to validate the performance of a given tool before its predictions can be utilized with confidence. Our analysis of the performance of multiple epitope prediction tools against experimentally validated SARS-CoV-2 epitopes allowed us to identify the best IEDB tool for prediction of SARS-CoV-2 epitopes. We showed that among the commonly used IEDB prediction tools, NetMHCpan EL 4.1 and CD4 IEDB recommended 2.22 are ordered according to “rank”, are the most reliable combinations to predict SARS-CoV-2 CD8 and CD4 epitopes, respectively. More studies are needed to calculate a rank threshold for the epitope prediction tools to prevent the arbitrary selection of thresholds for high affinity epitope selection.

Next, we used immunoinformatics tools to predict immunogenic T cell epitopes of SARS-CoV-2 for use in a multi-epitope vaccine against the common SARS-CoV-2 variants, including the variants of concern (VOC). Our study identified 31 common N and 31 common S CD8 epitopes (Supplementary Table S1), and 5 common N and 11 common S CD4 epitopes (Supplementary Table S4), all of which are predicted to be immunogenic (or antigenic for CD4 epitopes), non-toxic and non-allergenic. Epitopes showed a world population coverage of 97.87% for the selected top 11 N, 89.92% for the selected top 15 S CD8 epitopes (Table 1), 34.55% for the selected top 1 N CD4 epitope and 95.13%% for the selected top 7 S CD4 epitopes (Table 3). These epitopes present ideal candidates for use in a candidate vaccine that could provide protection across current SARS-CoV-2 variants. B cell epitopes will also need to be identified to complete the vaccine formulation. Additionally, we identified variant-specific immunogenic, non-toxic, non-allergenic, and stable CD8 (Table 2) and CD4 (Table 4) epitopes that can more pointedly target the mutations particular to each variant of concern. Given the rise of increasingly infectious variants, as well as concerns of decreasing vaccine efficacy against new variants, these epitopes may be useful in developing variant-specific booster shots to provide stronger protection against variants.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Despite the success of current vaccines, it remains necessary to expand the tools at our disposal to combat the current COVID-19 pandemic and prepare for future variants and outbreaks. Specifically, epitope-based vaccines present a useful tool with advantages over the current mRNA vaccines. Firstly, epitope-based vaccines may be able to confer immunity that lasts longer than natural immunity or immunity induced by current vaccines. Natural immunity from previous infection has been shown to be unreliable in older individuals [70] and fades relatively quickly with a half-life of 200 days for T cells [71], making it ineffective in providing long term immunity against reinfection, with T cell immunity fading faster than B cell immunity [72]. Epitope-based vaccines can be designed to target every aspect of the immune system by including epitopes that activate both CD8 and CD4 T cells as well as B cells, all of which are necessary to confer long-term immunity. While all current COVID-19 vaccines only include the spike protein of SARS-CoV-2, it has been shown that only N protein T cell epitopes for SARS-CoV lasted at least 17 years, making them the most long-lasting epitopes [32]. This suggests an opportunity to generate longer lasting immunity by making use of the longevity of N protein epitopes in vaccines. Multi-epitope vaccines can include epitopes of both the S and N proteins of SARS-CoV-2, such as those suggested by our study, to provide the most effective and long-lasting protection against SARS-CoV-2. More studies are also needed to investigate how to prolong T cell immunity in a potentially universal coronavirus vaccine.

Another advantage of epitope-based vaccines is their specificity, which provides them with a unique ability to precisely target certain areas of a protein of interest. This ability is especially important in the context of an evolving virus such as SARS-CoV-2 as novel, increasingly infectious variants are spreading and continue to emerge. Epitope-based vaccines can be designed to target certain variants by including epitopes containing variant-specific mutations, such as those identified by our study. This feature may be of particular use in booster shots designed to target specific variants against which the current vaccines may be less effective [73]. T cells, IgG and IgM antibodies have been shown to decrease substantially 6 months after administering the 2nd dose of the “low-dose” Moderna vaccine [74], suggesting that booster shots may be necessary to confer long-lasting immunity by reactivating cellular and humoral immunity. Of particular concern currently is the Delta variant that is spreading globally, as recent data suggests that both the AstraZeneca and Pfizer vaccines are slightly less effective against this variant in preventing COVID-19 related symptoms compared to the Alpha variant [75]. A recent large-scale study of Delta infections in the UK found that the protection offered against the variant by the AstraZeneca and Pfizer vaccines wanes with time, further supporting the need for variant-specific vaccines to boost immunity against variants of concern [76]. Despite global vaccination efforts, the rise in COVID-19 cases around the world as the Delta variant of SARS-CoV-2 spreads implies that variant-specific booster shots may be needed to combat emerging variants. Our study

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identified high-quality epitopes specific to current variants of concern that can be included in epitope-based vaccines to target most variants of concern.

The specificity of epitope-based vaccines also provides a potential method for preventing Original Antigenic Sin (OAS). According to this theory, the development of immunity against a pathogen is shaped by first exposure to that pathogen, limiting the flexibility of the immune system [77]. Upon subsequent infection with a novel strain of the pathogen, the immune system produces antibodies to the original strain it was first exposed to rather than mounting a response to novel antigens of the new strain. This phenomenon has been seen in response to influenza vaccines, demonstrating how OAS can allow novel influenza strains to evade the immune system despite vaccination [78]. OAS has also been seen in T cells, impairing the ability of cytotoxic T lymphocytes to respond effectively to variant viruses [79]. OAS is an important consideration in developing vaccines against SARS-CoV-2, as it suggests that immunity developed in response to infection or vaccination with one strain may not protect against all novel variants. Additionally, a booster vaccine that includes epitopes from the original reference sequence, such as using a variant of the SARS-CoV-2 spike protein, will likely reactivate B and T cells against the original epitopes without mounting a variant-specific immune response. For booster vaccines to induce a variant-specific immune response that can confer additional immunity against novel variants, they must be carefully formulated to avoid inducing an immune response against the original strain, a technique which requires great specificity. While some methods have been suggested to alleviate OAS [80], we hereby hypothesize a new method of preventing OAS through the use of epitope-based booster vaccines. We suggest that the specificity of epitope-based booster vaccines may allow them to prevent OAS and better activate variant-specific immune cells, thus providing stronger immunity against novel SARS-CoV-2 variants.

Conclusion

In this study, we identified the best epitope prediction tools available on IEDB for predicting CD8 and CD4 binding epitopes of SARS-CoV-2. We then used these tools to predict SARS-CoV-2 N and S protein T cell epitopes. Prioritizing epitopes that were found to be common across the current SARS-CoV-2 variants, predicted to be immunogenic, non-toxic, non-allergenic, with high cell penetrability and with a high population coverage based on HLA alleles allowed us to identify high-quality T cell epitopes. We recommend these epitopes for a multi-epitope vaccine against the common SARS-CoV-2 variants, including variants of concern. We also present variant-specific immunogenic CD8 and CD4 T cell epitopes that can more precisely target variants of concern. Given the rise of increasingly infectious variants with the potential to evade natural immunity and current vaccines, these epitopes may prove useful in developing variant-specific booster shots to combat the most concerning variants. Together,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these predicted epitopes may provide both broad protection against common SARS-CoV-2 variants, as well as more targeted protection against specific variants. While other studies have attempted to use immunoinformatics to design epitope-based vaccines against SARS-CoV-2, our comparison of available epitope prediction tools to experimentally validated SARS-CoV-2 epitopes validates that our predicted epitopes come from the most accurate available tool on IEDB.org. Additionally, our comparison of predicted epitopes across all current variants increases confidence that the effectiveness of our designed vaccine will not be undermined by the mutations of certain variants. Our epitope prediction together with our mouse MHC affinity prediction warrant further preclinical studies to validate their efficacy for use in designing epitope-based vaccines against SARS-CoV-2. While this experimental validation remains necessary, immunoinformatics has allowed us to predict epitopes in silico that are most likely to prove effective in vitro and in vivo. This will greatly reduce the cost and time of traditional vaccine development as it is urgent to develop new therapeutics and vaccines to combat this rapidly evolving pandemic as efficiently as possible.

Acknowledgements

We wish to acknowledge for the support by Lombardi Comprehensive Cancer Center, Georgetown University Medical Center.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. WHO Coronavirus (COVID-19) Dashboard. Accessed Sept 2, 2021. https://covid19.who.int

2. Commissioner O of the. FDA Takes Additional Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for Second COVID-19 Vaccine. FDA. Published December 21, 2020. Accessed June 22, 2021. https://www.fda.gov/news-events/press-announcements/fda-takes-additional-action-fight-against-covid-19-issuing-emergency-use-authorization-second-covid

3. Commissioner O of the. FDA Issues Emergency Use Authorization for Third COVID-19 Vaccine. FDA. Published March 2, 2021. Accessed June 22, 2021. https://www.fda.gov/news-events/press-announcements/fda-issues-emergency-use-authorization-third-covid-19-vaccine

4. Commissioner O of the. FDA Approves FIrst COVID-19 Vaccine. FDA.Published August 23, 2021. Accessed August 31, 2021. https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine

5. Ahmad TA, Eweida AE, Sheweita SA. B-cell epitope mapping for the design of vaccines and effective diagnostics. *Trials in Vaccinology*. 2016;5:71-83. Accessed July 10, 2021. doi:https://doi.org/10.1016/j.trivac.2016.04.003

6. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019;47(D1):D339-D343. Accessed July 10, 2021. doi:10.1093/nar/gky1006

7. Parvizpour S, Pourseif MM, Razmara J, Rafi MA, Omidi Y. Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. *Drug Discovery Today*. 2020;25(6):1034-1042. Accessed June 22, 2021. doi:https://doi.org/10.1016/j.drudis.2020.03.006

8. Rosa DS, Ribeiro SP, Cunha-Neto E. CD4+ T Cell Epitope Discovery and Rational Vaccine Design. *Arch Immunol Ther Exp*. 2010;58(2):121-130. Accessed June 22, 2021. doi:10.1007/s00005-010-0067-0

9. Le Bert N, Tan AT, Kunasegaran K, et al. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*. 2020;584(7821):457-462. Accessed July 20, 2021. doi:10.1038/s41586-020-2550-z

10. Zhao K, Yang B, Xu Y, Wu C. CD8+ T cell response in HLA-A*0201 transgenic mice is elicited by epitopes from SARS-CoV S protein. *Vaccine*. 2010;28(41):6666-6674. Accessed July 20, 2021. doi:10.1016/j.vaccine.2010.08.013

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Huang J, Ma R, Wu C. Immunization with SARS-CoV S DNA vaccine generates memory CD4+ and CD8+ T cell immune responses. *Vaccine*. 2006;24(23):4905-4913. Accessed July 20, 2021. doi:10.1016/j.vaccine.2006.03.058

12. Polack FP, Thomas SJ, Kitchin N, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*. 2020;383(27):2603-2615. Accessed July 18, 2021. doi:10.1056/NEJMoa2034577

13. Baden LR, El Sahly HM, Essink B, et al. Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *N Engl J Med*. 2021;384(5):403-416. Accessed July 18, 2021. doi:10.1056/NEJMoa2035389

14. Sadoff J, Gray G, Vandebosch A, et al. Safety and Efficacy of Single-Dose Ad26.COV2.S Vaccine against Covid-19. *N Engl J Med*. 2021; 384:2187-2201. Accessed July 18, 2021. doi:10.1056/NEJMoa2101544

15. Israel Ministry of Health. Coronavirus in Israel - General Situation. Accessed July 8, 2021. https://datadashboard.health.gov.il/COVID-19/general

16. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in china. *Nature*. 2020;579(7798):265-269. Accessed Apr 24, 2021. doi: 10.1038/s41586-020-2008-3.

17. Investigation of novel SARS-CoV-2 variant, Variant of Concern 202012/01, Technical Briefing 5. Public Public Health England. Announced 13th February, 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf

18. Vahed M, Calcagno TM, Quinonez E, Mirsaeidi M. Impacts of 203/204: RG>KR mutation in the N protein of SARS-CoV-2. *bioRxiv*. 2021:2021.01.14.426726. http://biorxiv.org/content/early/2021/01/14/2021.01.14.426726.abstract. doi: 10.1101/2021.01.14.426726.

19. Tchesnokova V, Kulakesara H, Larson L, et al. Acquisition of the L452R mutation in the ACE2-binding interface of spike protein triggers recent massive expansion of SARS-cov-2 variants. *bioRxiv*. 2021:2021.02.22.432189. http://biorxiv.org/content/early/2021/03/11/2021.02.22.432189.abstract. doi: 10.1101/2021.02.22.432189.

20. Yadav PD, Sapkal GN, Abraham P, et al. Neutralization of variant under investigation B.1.617 with sera of BBV152 vaccinees. *bioRxiv*. 2021:2021.04.23.441101.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

http://biorxiv.org/content/early/2021/04/23/2021.04.23.441101.abstract. doi: 10.1101/2021.04.23.441101.

21. Lassaunière R, Fonager J, Rasmussen M, et al. Working paper on SARS-CoV-2 spike mutations arising in Danish mink, their 2 spread to humans and neutralization data. Announced 5th November, 2020. https://files.ssi.dk/Mink-cluster-5-short-report_AFO2, accessed 23rd February 2021

22. Schumacher TN, De Bruijn ML, Vernie LN, et al. Peptide selection by MHC class I molecules. *Nature*. 1991;350(6320):703-706. Accessed Apr 21, 2021. doi: 10.1038/350703a0.

23. Quadeer A. A., Ahmed S. F., Mckay M. R. Epitopes targeted by T cells in convalescent COVID-19 patients. *bioRxiv*. 2020. https://www.mckayspcb.com/SARS2TcellEpitopes/.

24 Prachar M, Justesen S, Steen-Jensen DB, et al. Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools. *Scientific Reports*. 2020;10(1):1-8. https://www.nature.com/articles/s41598-020-77466-4. Accessed Apr 21, 2021. doi: 10.1038/s41598-020-77466-4.

25. Sohail MS, Ahmed SF, Quadeer AA, McKay MR. In silico T cell epitope identification for SARS-CoV-2: Progress and perspectives. *Adv Drug Deliv Rev*. 2021;171:29-47. Accessed Apr 21, 2021. doi: 10.1016/j.addr.2021.01.007.

26. Schulien I, Kemming J, Oberhardt V, et al. Characterization of pre-existing and induced SARS-CoV-2-specific CD8 + T cells. *Nature Medicine*. 2021;27(1):78-85. https://www.nature.com/articles/s41591-020-01143-2. Accessed Apr 21, 2021. doi: 10.1038/s41591-020-01143-2.

27. Nelde A, Bilich T, Heitmann JS, et al. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nature immunology*. 2021;22(1):74-85. https://www.ncbi.nlm.nih.gov/pubmed/32999467. doi: 10.1038/s41590-020-00808-x.

28. Kared H, Redd AD, Bloch EM, et al. CD8+ T cell responses in convalescent COVID-19 individuals target epitopes from the entire SARS-CoV-2 proteome and show kinetics of early differentiation. *bioRxiv : the preprint server for biology*. 2020. http://europepmc.org/abstract/MED/33052343 https://doi.org/10.1101/2020.10.08.330688 https://europepmc.org/articles/PMC7553170 https://europepmc.org/articles/PMC7553170?pdf=render. doi: 10.1101/2020.10.08.330688.

29. Saini SK, Hersby DS, Tamhane T, et al. SARS-CoV-2 genome-wide mapping of CD8 T cell recognition reveals strong immunodominance and substantial CD8 T cell activation in COVID-19 patients. *bioRxiv*. 2020:2020.10.19.344911. http://biorxiv.org/content/early/2020/10/19/2020.10.19.344911.abstract. doi: 10.1101/2020.10.19.344911.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

30. Tarke A, Sidney J, Kidd CK, et al. Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *bioRxiv*. 2020. Accessed Apr 25, 2021. doi: 10.1101/2020.12.08.416750.

31. Tarke A, Sidney J, Kidd CK, et al. Comprehensive analysis of t cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Rep Med*. 2021;2(2):100204. Accessed Apr 21, 2021. doi: 10.1016/j.xcrm.2021.100204.

32. Le Bert N, Tan AT, Kunasegaran K, et al. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*. 2020;584(7821):457-462. Accessed Apr 25, 2021. doi: 10.1038/s41586-020-2550-z.

33. Peng Y, Mentzer AJ, Liu G, et al. Broad and strong memory CD4.sup.+ and CD8.sup.+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. *Nature immunology*. 2020;21(11):1336. https://www.nature.com/articles/s41590-020-0782-6. doi: 10.1038/s41590-020-0782-6.

34. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: Application to the MHC class I system. *Bioinformatics*. 2016;32(4):511-517. Accessed Jun 13, 2021. doi: 10.1093/bioinformatics/btv639.

35. Buus S, Lauemøller SL, Worning P, et al. Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*. 2003;62(5):378-384. Accessed Jun 13, 2021. doi: 10.1034/j.1399-0039.2003.00112.x.

36. Lundegaard C, Nielsen M, Lund O. The validity of predicted T-cell epitopes. *Trends Biotechnol*. 2006;24(12):537-538. Accessed Jun 13, 2021. doi: 10.1016/j.tibtech.2006.10.001.

37. Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*. 2008;24(11):1397-1398. Accessed Jun 13, 2021. doi: 10.1093/bioinformatics/btn128.

38. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: Accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*. 2008;36(Web Server issue):509. Accessed Jun 13, 2021. doi: 10.1093/nar/gkn202.

39. Nielsen M, Lundegaard C, Worning P, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. 2003;12(5):1007-1017. Accessed Jun 13, 2021. doi: 10.1110/ps.0239403.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

40. Moutaftsi M, Peters B, Pasquetto V, et al. A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat Biotechnol*. 2006;24(7):817-819. Accessed Jun 13, 2021. doi: 10.1038/nbt1215.

41. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhütter H. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J Immunol*. 2003;171(4):1741-1749. Accessed Jun 13, 2021. doi: 10.4049/jimmunol.171.4.1741.

42. Tenzer S, Peters B, Bulik S, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci*. 2005;62(9):1025-1037. Accessed Jun 13, 2021. doi: 10.1007/s00018-005-4528-2.

43. Hoof I, Peters B, Sidney J, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. 2009;61(1):1-13. Accessed Jun 13, 2021. doi: 10.1007/s00251-008-0341-z.

44. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199(9):3360-3368. Accessed Jun 13, 2021. doi: 10.4049/jimmunol.1700893.

45. Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One*. 2007;2(8):e796. Accessed Jun 13, 2021. doi: 10.1371/journal.pone.0000796.

46. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med*. 2016;8(1):33. Accessed Jun 13, 2021. doi: 10.1186/s13073-016-0288-x.

47. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020;48(W1):W449-W454. Accessed Jun 13, 2021. doi: 10.1093/nar/gkaa379.

48. Wang P, Sidney J, Dow C, Mothé B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol*. 2008;4(4):e1000048. Accessed Jun 16, 2021. doi: 10.1371/journal.pcbi.1000048.

49. Wang P, Sidney J, Kim Y, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*. 2010;11:568. Accessed Jun 16, 2021. doi: 10.1186/1471-2105-11-568.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

50. Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics*. 2015;67(11-12):641-650. Accessed Jun 16, 2021. doi: 10.1007/s00251-015-0873-y.

51. Jensen KK, Andreatta M, Marcatili P, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. 2018;154(3):394-406. Accessed Jun 16, 2021. doi: 10.1111/imm.12889.

52. Calis JJA, Maybeno M, Greenbaum JA, et al. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol*. 2013;9(10):e1003266. Accessed Jun 13, 2021. doi: 10.1371/journal.pcbi.1003266.

53. Doytchinova Irini A, Flower Darren R. Bioinformatic approach for identifying parasite and fungal candidate subunit vaccines. *Open Vaccines Journal*. 2008;1:22-26.

54. Doytchinova IA, Flower DR. Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine*. 2007;25(5):856-866. Accessed Jul 30, 2021. doi: 10.1016/j.vaccine.2006.09.032.

55. Doytchinova IA, Flower DR. VaxiJen: A server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*. 2007;8:4. Accessed Jul 30, 2021. doi: 10.1186/1471-2105-8-4.

56. Gupta S, Kapoor P, Chaudhary K, et al. In silico approach for predicting toxicity of peptides and proteins. *PLOS ONE*. 2013;8(9):e73957. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073957. Accessed Jul 10, 2021. doi: 10.1371/journal.pone.0073957.

57. Dimitrov I, Flower DR, Doytchinova I. AllerTOP - a server for in silico prediction of allergens. *BMC Bioinformatics*. 2013;14(Suppl 6):S4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3633022/. Accessed Jul 11, 2021. doi: 10.1186/1471-2105-14-S6-S4.

58. Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v.2--a server for in silico prediction of allergens. *J Mol Model*. 2014;20(6):2278. Accessed Jul 11, 2021. doi: 10.1007/s00894-014-2278-5.

59. Gasteiger E, Hoogland C, Gattiker A, et al. Protein identification and analysis tools on the ExPASy server. In: John M. Walker, ed. *The proteomics protocols handbook.* Totowa, N.J: Humana Press; 2005:571-607.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

60. Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng*. 1990;4(2):155-161. Accessed Jul 29, 2021. doi: 10.1093/protein/4.2.155.

61. DeVette CI, Andreatta M, Bardet W, et al. NetH2pan: A computational tool to guide MHC peptide prediction on murine tumors. *Cancer Immunol Res*. 2018;6(6):636-644. Accessed Jul 15, 2021. doi: 10.1158/2326-6066.CIR-17-0298.

62. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199(9):3360-3368. Accessed Jul 15, 2021. doi: 10.4049/jimmunol.1700893.

63. Bui H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*. 2006;7:153. Accessed Jun 13, 2021. doi: 10.1186/1471-2105-7-153.

64. Kaur H, Garg A, Raghava GPS. PEPstr: A de novo method for tertiary structure prediction of small bioactive peptides. *Protein Pept Lett*. 2007;14(7):626-631. Accessed Aug 9, 2021. doi: 10.2174/092986607781483859.

65. Singh S, Singh H, Tuknait A, et al. PEPstrMOD: Structure prediction of peptides containing natural, non-natural and modified residues. *Biol Direct*. 2015;10:73. Accessed Aug 9, 2021. doi: 10.1186/s13062-015-0103-4.

66. Guex N, Peitsch MC. SWISS-MODEL and the swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*. 1997;18(15):2714-2723. Accessed Jun 14, 2021. doi: 10.1002/elps.1150181505.

67. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O. Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Res*. 2011;39(Web Server issue):249. Accessed Jul 4, 2021. doi: 10.1093/nar/gkr431.

68. Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins*. 2010;78(9):2029-2040. Accessed Jul 4, 2021. doi: 10.1002/prot.22716.

69. D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2021), Amber 2021, University of California, San Francisco.

70. Hansen CH, Michlmayr D, Gubbels SM, Mølbak K, Ethelberg S. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in denmark in 2020: A population-level observational study. *Lancet*. 2021;397(10280):1204-1212. Accessed Jun 22, 2021. doi: 10.1016/S0140-6736(21)00575-4.

71. Cohen KW, Linderman SL, Moodie Z, et al. Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells. *medRxiv*. 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8095229/. Accessed Jul 22, 2021. doi: 10.1101/2021.04.19.21255739.

72. Dan JM, Mateus J, Kato Y, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science*. 2021;371(6529):eabf4063. Accessed Jun 22, 2021. doi: 10.1126/science.abf4063.

73. Lopez Bernal J, Andrews N, Gower C, et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N Engl J Med*. 2021; 385:585-594. Accessed Aug 20, 2021. doi:10.1056/NEJMoa2108891

74. Mateus J, Dan JM, Zhang Z, et al. Low dose mRNA-1273 COVID-19 vaccine generates durable T cell memory and antibodies enhanced by pre-existing crossreactive T cell memory. *medRxiv*. 2021:2021.06.30.21259787. http://medrxiv.org/content/early/2021/07/05/2021.06.30.21259787.abstract. doi: 10.1101/2021.06.30.21259787.

75. Stowe J, Andrews N, Gower, C, et al. Effectiveness of COVID-19 Vaccines Against Hospital Admission with the Delta (B.1.617.2) Variant. *Public Health England*. Preprint published online June 14, 2021. Accessed Aug 27, 2021. https://go.nature.com/3gnqwxr

76. Pouwels KB, Pritchard E, Matthews PC, et al. Impact of Delta on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *medRxiv*. Preprint published online August 24, 2021. Accessed Aug 27, 2021. doi:10.1101/2021.08.18.21262237

77. Francis T. On the Doctrine of Original Antigenic Sin. *Proceedings of the American Philosophical Society*. 1960;104(6):572-578. Accessed Jul 22, 2021 from http://www.jstor.org/stable/985534

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

78. Kim JH, Skountzou I, Compans R, Jacob J. Original antigenic sin responses to influenza viruses. *J Immunol*. 2009;183(5):3294-3301. Accessed Aug 20, 2021. doi:10.4049/jimmunol.0900398

79. Klenerman P, Zinkernagel RM. Original antigenic sin impairs cytotoxic T lymphocyte responses to viruses bearing variant epitopes. *Nature*. 1998;394(6692):482-485. Accessed Aug 20, 2021. doi:10.1038/28860

80. Kim JH, Davis WG, Sambhara S, Jacob J. Strategies to alleviate original antigenic sin responses to influenza viruses. *Proc Natl Acad Sci U S A*. 2012;109(34):13751-13756. Accessed Aug 20, 2021. doi:10.1073/pnas.0912458109

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and Tables

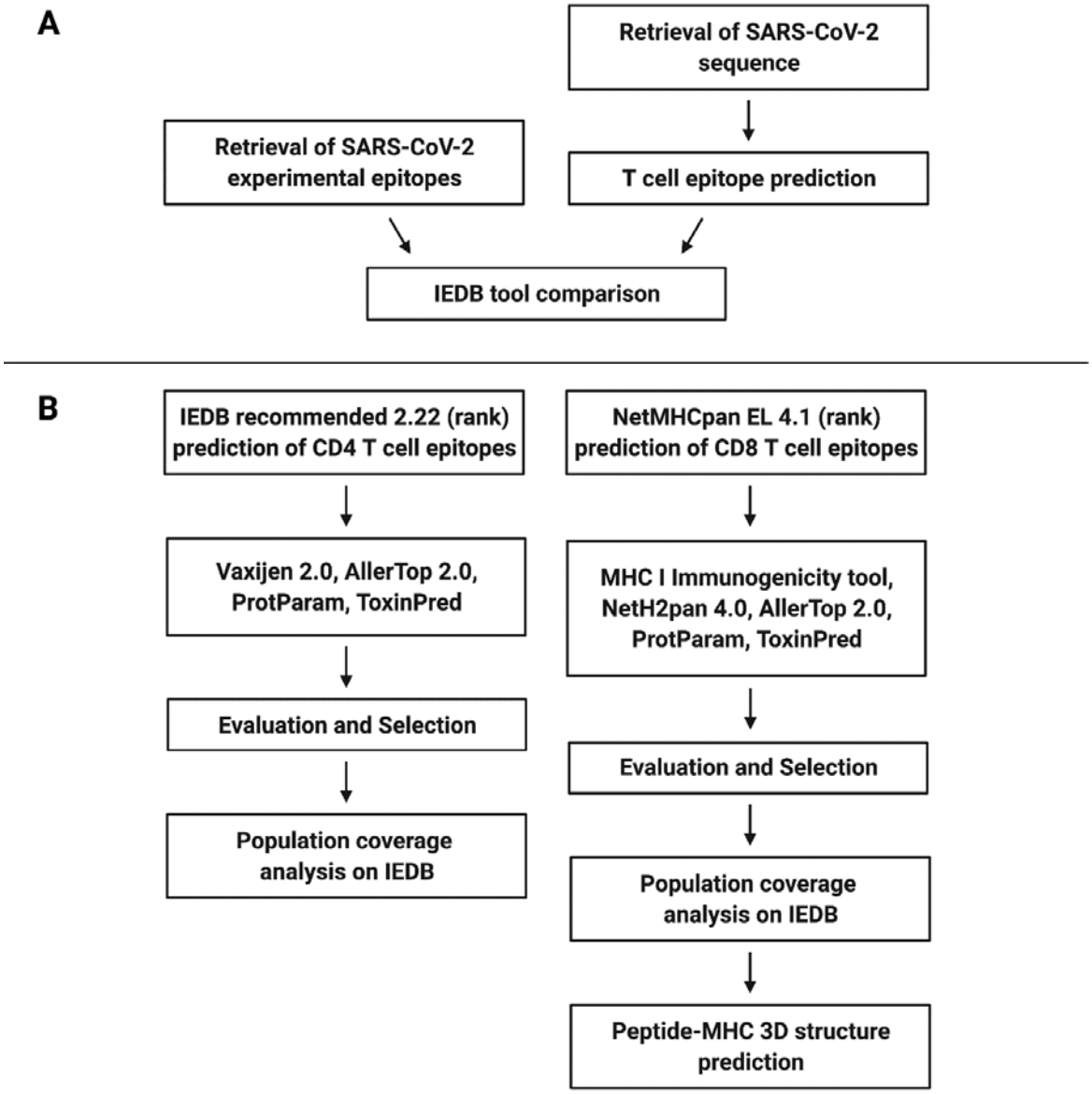

Figure 1. Schematic flow chart diagram of the in silico peptide prediction process. (A) Selection of the best MHC-I and -II binding epitope prediction tools based on its classification performance. (B) In silico identification and evaluation of novel CD4 and CD8 peptides for vaccine design.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

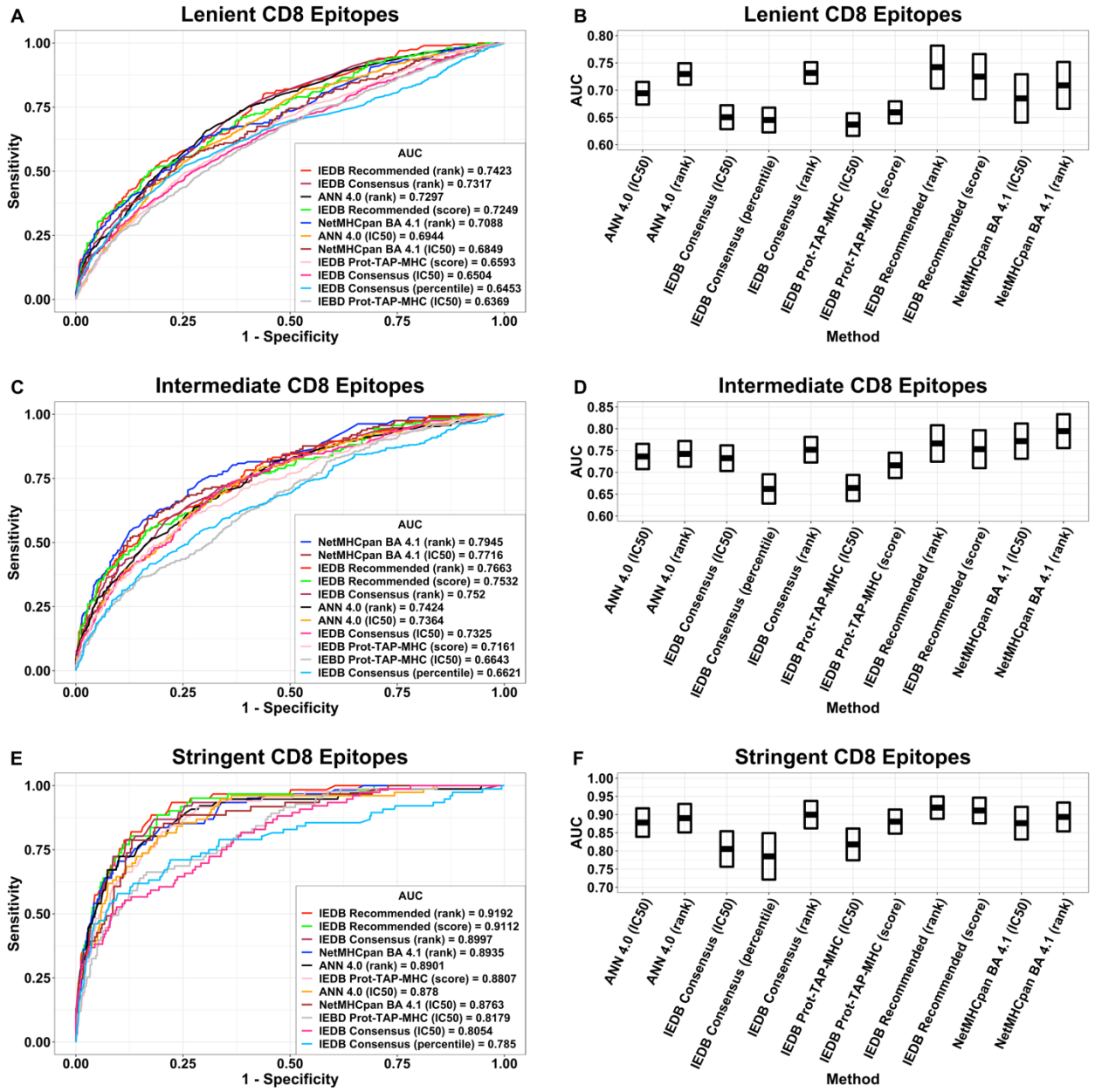

Figure 2. Comparison of MHC class I binding epitope prediction tools' binary classification performance. (A, C, E) ROC curves with AUC values were shown for each tool-attribute combination that was evaluated for its experimental epitope prediction power. Cutoff thresholds for classification were established through the attribute, ordered from highest to lowest binding affinity, for each tool. (B, D, F) Boxplots show the upper and lower 95% confidence interval boundaries of each tool-attribute's AUC values.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

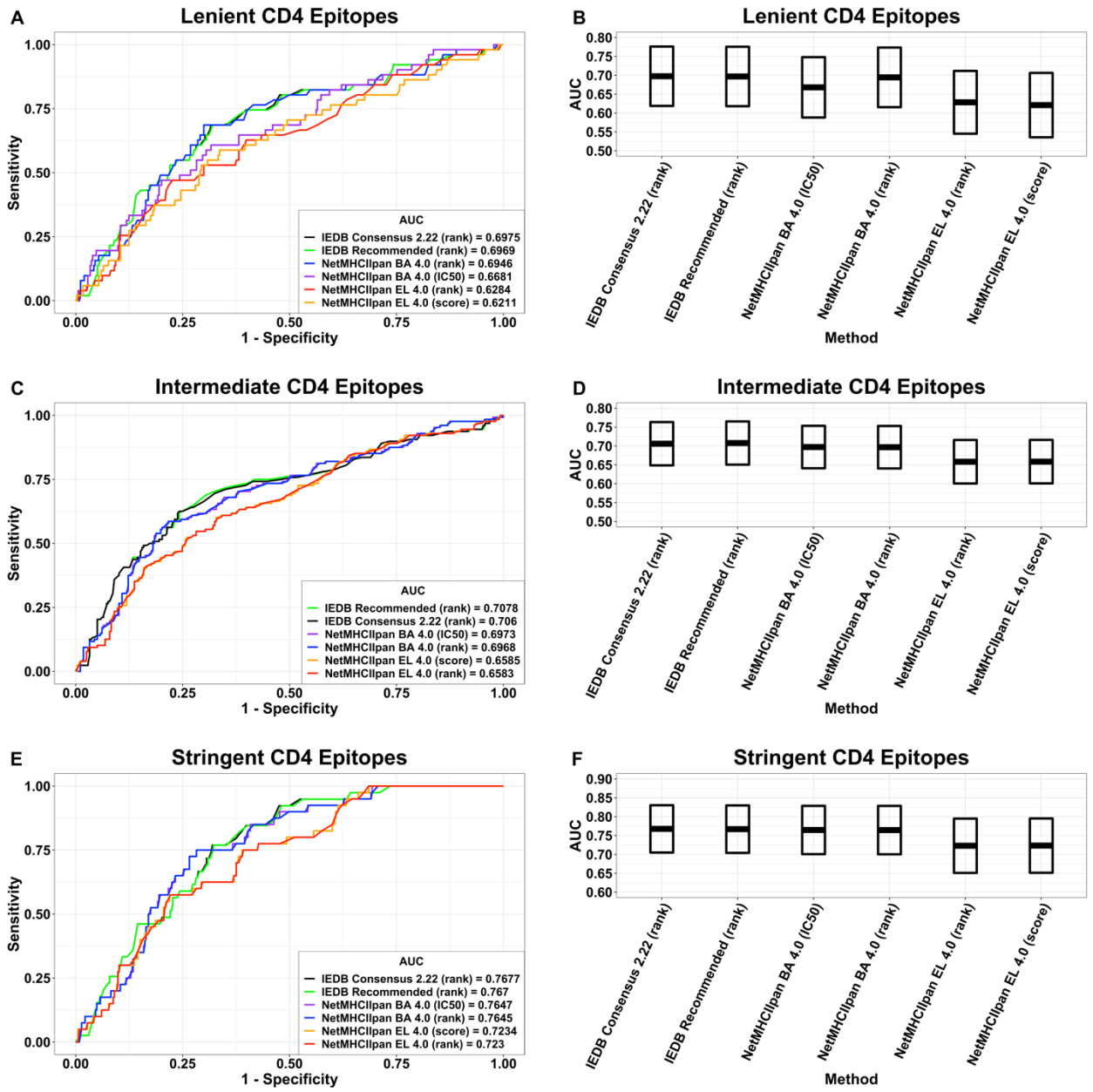

Figure 3. Comparison of MHC class II binding epitope prediction tools' binary classification performance. (A, C, E) ROC curves with AUC values were shown for each tool-attribute combination that was evaluated for its experimental epitope prediction power. Cutoff thresholds for classification were established through the attribute, ordered from highest to lowest binding affinity, for each tool. (B,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

D, F) Boxplots show the upper and lower 95% confidence interval boundaries of each tool-attribute's AUC values.

Protein	Population Coverage	Epitopes	Common HLA restriction	Mouse MHC restriction
N	97.87%	LSPRWYFYY	HLA-A*01:01, HLA-A*30:02, HLA-B*57:01	H-2-Dd
		SPRWYFYYL	HLA-B*07:02, HLA-B*08:01	H-2-Ld
		TPSGTWLTY	HLA-B*35:01, HLA-B*53:01, HLA-A*26:01, HLA-A*01:01	H-2-Ld
		NTASWFTAL	HLA-A*68:02, HLA-A*26:01, HLA-A*02:06	H-2-Kb, H-2-Dd
		YYRRATRRI	HLA-A*24:02, HLA-A*23:01	H-2-Kd
		TGPEAGLPY	HLA-B*35:01	H-2-Dd
		LPNNTASWF	HLA-B*53:01, HLA-B*35:01, HLA-B*07:02	H-2-Ld
		KDLSPRWYF	HLA-B*44:02, HLA-A*32:01, HLA-B*44:03	H-2-Ld
		GMSRIGMEV	HLA-A*02:03, HLA-A*02:01	Not available
		KTFPPTEPK	HLA-A*11:01, HLA-A*03:01, HLA-A*30:01, HLA-A*31:01, HLA-A*68:01, HLA-A*32:01	Not available
		ELIRQGTDY	HLA-A*26:01, HLA-B*15:01	Not available
S	89.92%	AEIRASANL	HLA-B*40:01, HLA-B*44:03, HLA-B*44:02	H-2-Dd
		DAVRDPQTL	HLA-B*51:01	H-2-Db, H-2-Dd, H-2-Kd
		FTISVTTEI	HLA-A*68:02, HLA-A*02:06	H-2-Db
		GEVFNATRF	HLA-B*44:03, HLA-B*44:02	H-2-Db
		INITRFQTL	HLA-B*08:01	H-2-Db, H-2-Kb, H-2-Dd, H-2-Kd, H-2-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

				Ld
		LAGTITSGW	HLA-B*58:01, HLA-B*57:01, HLA-B*53:01	H-2-Ld
		LPFNDGVYF	HLA-B*35:01, HLA-B*53:01, HLA-B*51:01	H-2-Ld
		VVFLHVTYV	HLA-A*02:03, HLA-A*02:06, HLA-A*68:02, HLA-A*02:01	H-2-Db, H-2-Kb, H-2-Dd
		YQPYRVVVL	HLA-B*08:01	H-2-Dd
		LPIGINITRF	HLA-B*53:01, HLA-B*35:01	H2-Ld
		NFTISVTTEI	HLA-A*68:02	H2-Kd
		VLPFNDGVYF	HLA-B*53:01	H2-Db
		HWFVTQRNF	HLA-A*23:01, HLA-A*24:02	Not available
		WTAGAAAYY	HLA-A*26:01, HLA-A*01:01	Not available
		QLTPTWRVY	HLA-A*30:02, HLA-B*15:01	Not available

Table 1. Top 11 N and top 15 S protein CD8 epitopes that are most immunogenic, non-toxic, non-allergenic and common across all studied SARS-CoV-2 variants with their world population coverage and their mouse MHC haplotype when available.

Protein	Variant	Epitope	Human HLA restriction	Mouse MHC restriction
N	**Alpha**	SLNGPQNQR	HLA-A*31:01, HLA-A*03:01, HLA-A*33:01, HLA-A*68:01, HLA-A*11:01	Not available
S	**Beta**	TPINLVRGL	HLA-B*07:02	H-2-Ld
	Gamma	GQTGTIADY	HLA-B*15:01	Not available
		ASANLAAIK	HLA-A*11:01	Not available
		GTIADYNYK	HLA-A*11:01	Not available
	Alpha	VLNDILARL	HLA-A*02:03, HLA-A*02:01, HLA-A*02:06	H-2-Kb, H-2-Dd
		SVLNDILARL	HLA-A*02:03	H-2-Kb
		QIITTHNTF	HLA-B*15:01	H-2-Db, H-2-Dd, H-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

				2-Kd, H-2-Ld
		SVLNDILAR	HLA-A*11:01	Not available
	US	YNYRYRLFR	HLA-A*33:01	Not available
		LPLVSIQCV	HLA-B*51:01	H-2-Ld
	Cluster 5	YNYLFRLFR	HLA-A*33:01	Not available
		NYNYLFRLFR	HLA-A*33:01	Not available
	Delta	TYVPAHEKNF	HLA-A*24:02, HLA-A*23:01	H-2-Kd
	US and delta	KVGGNYNYR	HLA-A*31:01	Not available

Table 2. All immunogenic non-toxic non-allergenic stable CD8 variant-specific epitopes for each SARS-CoV-2 variant N and S protein, with the variant-specific mutations written in red (see Supplementary Table S3 for full list of immunogenic non-toxic non-allergenic variant-specific CD8 epitopes).

Protein	Population coverage	Epitopes	Common HLA restriction
N	34.55%	GTRNPANNAAIVLQL	HLA-DQA1*01:02, HLA-DQB1*06:02
S	95.13%	YRVVVLSFELLHAPA	HLA-DPA1*03:01, HLA-DPB1*04:02, HLA-DPA1*02:01, HLA-DPB1*01:01, HLA-DPA1*01:03, HLA-DPB1*02:01
		VVVLSFELLHAPATV	HLA-DRB1*01:01
		VVLSFELLHAPATVC	HLA-DRB1*01:01
		VLSFELLHAPATVCG	HLA-DRB1*01:01
		RVVVLSFELLHAPAT	HLA-DRB1*01:01
		QSLLIVNNATNVVIK	HLA-DRB1*13:02, HLA-DRB3*02:02
		GWTFGAGAALQIPFA	HLA-DRB1*09:01

Table 3. All common N and S CD4 antigenic non-toxic non-allergenic stable epitopes across all studied SARS-CoV-2 variants and their world population coverage.

Protein	Variant	Epitope	HLA restriction
N	**Delta**	KKKKAYETQALPQRQ	HLA-DPA1*02:01, HLA-DPB1*14:01

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		DKKKKAYETQALPQR	HLA-DPA1*02:01, HLA-DPB1*01:01
		KKKAYETQALPQRQK	HLA-DPA1*02:01, HLA-DPB1*01:01
S	**Beta**	HTPINLVRGLPQGFS	HLA-DRB4*01:01, HLA-DRB1*11:01, HLA-DRB1*08:02, HLA-DRB1*04:05
	Gamma	LVLLPLVSSQCVNFT	HLA-DRB4*01:01, HLA-DRB1*08:02, HLA-DRB1*04:01, HLA-DRB1*15:01, HLA-DRB1*12:01, HLA-DRB1*04:05
		FLVLLPLVSSQCVNF	HLA-DRB1*01:01, HLA-DRB4*01:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*04:05, HLA-DRB1*15:01, HLA-DRB5*01:01, HLA-DRB1*04:01
		VLLPLVSSQCVNFTN	HLA-DRB4*01:01, HLA-DRB1*04:05
		LLPLVSSQCVNFTNR	HLA-DRB4*01:01
		LPLVSSQCVNFTNRT	HLA-DRB4*01:01
		EYVNNSYECDIPIGA	HLA-DQA1*01:01, HLA-DQB1*05:01, HLA-DRB3*01:01
		GAEYVNNSYECDIPI	HLA-DQA1*01:01, HLA-DQB1*05:01
	Alpha	AIPINFTISVTTEIL	HLA-DRB1*07:01, HLA-DRB1*09:01, HLA-DRB3*02:02, HLA-DQA1*03:01, HLA-DQB1*03:02, HLA-DRB1*04:01, HLA-DRB1*13:02, HLA-DQA1*04:01, HLA-DQB1*04:02, HLA-DQA1*05:01, HLA-DQB1*02:01, HLA-DRB1*04:05, HLA-DRB3*01:01
	US	FVFLVLLPLVSIQCV	HLA-DRB1*01:01, HLA-DPA1*03:01, HLA-DPB1*04:02, HLA-DRB1*11:01, HLA-DPA1*02:01, HLA-DPB1*01:01, HLA-DRB5*01:01, HLA-DRB1*04:01, HLA-DRB1*15:01
		MFVFLVLLPLVSIQC	HLA-DRB1*01:01, HLA-DPA1*03:01, HLA-DPB1*04:02, HLA-DRB1*11:01, HLA-DPA1*02:01, HLA-DPB1*01:01, HLA-DRB5*01:01, HLA-DRB1*04:01, HLA-DRB1*15:01, HLA-DPA1*01:03, HLA-DPB1*02:01
		VLLPLVSIQCVNLTT	HLA-DRB4*01:01
		LVLLPLVSIQCVNLT	HLA-DRB4*01:01
	Cluster 5	IWLGFIAGLIAIVIV	HLA-DQA1*05:01, HLA-DQB1*03:01, HLA-DRB1*01:01, HLA-DRB1*12:01, HLA-DQA1*01:01, HLA-DQB1*05:01, HLA-DRB1*09:01, HLA-DPA1*03:01, HLA-DPB1*04:02
		YIWLGFIAGLIAIVI	HLA-DQA1*05:01, HLA-DQB1*03:01, HLA-DQA1*01:01,

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

			HLA-DQB1*05:01, HLA-DRB1*01:01, HLA-DRB1*12:01, HLA-DRB1*09:01, HLA-DPA1*03:01, HLA-DPB1*04:02
		GFIAGLIAIVIVTIM	HLA-DQA1*05:01, HLA-DQB1*03:01, HLA-DPA1*03:01, HLA-DPB1*04:02, HLA-DQA1*01:02, HLA-DQB1*06:02, HLA-DRB1*01:01, HLA-DRB1*12:01
	Delta	NTSNQVAVLYQGVNC	HLA-DQA1*01:02, HLA-DQB1*06:02
		GVVFLHVTYVPAHEK	HLA-DRB1*04:05
		VVFLHVTYVPAHEKN	HLA-DRB1*04:05
	US and delta	YNYRYRLFRKSNLKP	HLA-DRB1*11:01, HLA-DPA1*02:01, HLA-DPB1*05:01

Table 4. All antigenic non-toxic non-allergenic and stable CD4 epitopes per mutation for each SARS-CoV-2 variant N and S protein, with the variant-specific mutations written in red.

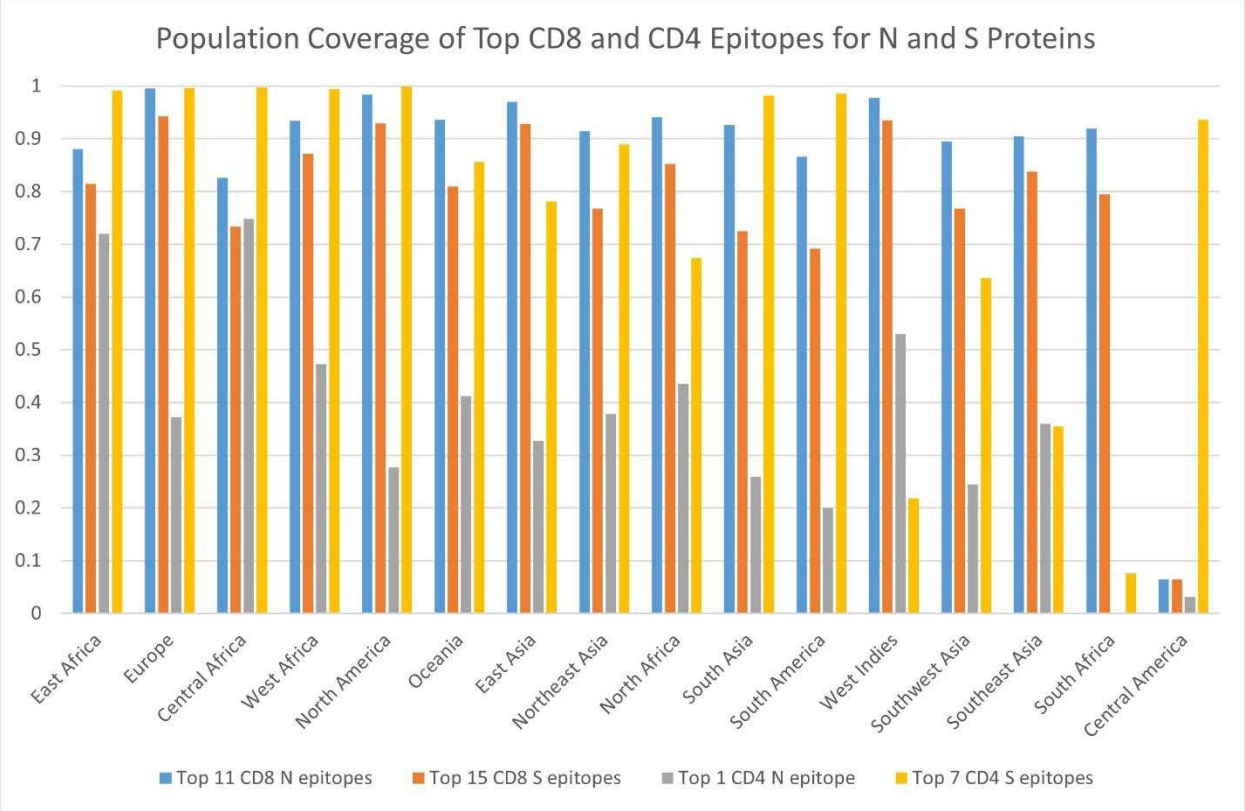

Figure 4. Regional population coverage of the top 11 N CD8 epitopes, top 15 S CD8 epitopes, top 1 N CD4 epitope and top 7 S CD4 epitopes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.11.459907; this version posted September 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

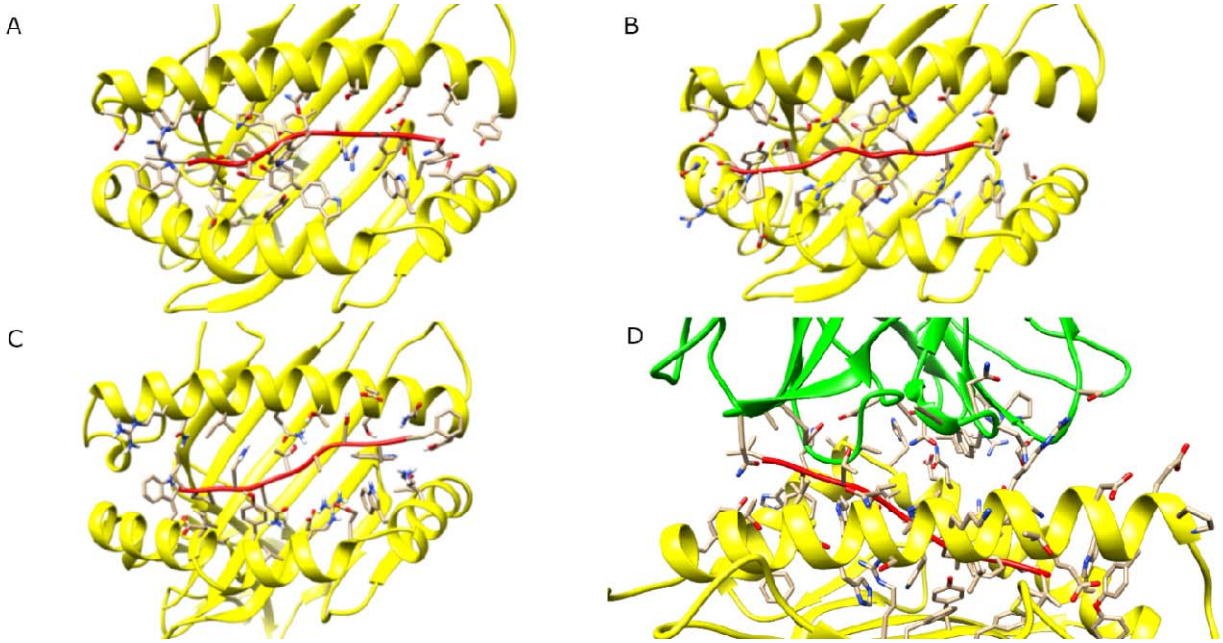

Figure 5. Close-up shots of the predicted 3D structure for the interaction of 3 SARS-CoV-2 N epitopes (red) with their selected MHC I molecules (yellow): NTASWFTAL-HLA-A*68 **(A),** TPSGTWLTY-HLA-A*01:01 **(B)**, LPNNTASWF-HLA-B*07:02 **(C)** and 1 SARS-CoV-2 S epitope (red) with its selected MHC (yellow) and TCR molecule (green): VVFLHVTYV-HLA-A*02:01-pRLQ3 TCR **(D)**.