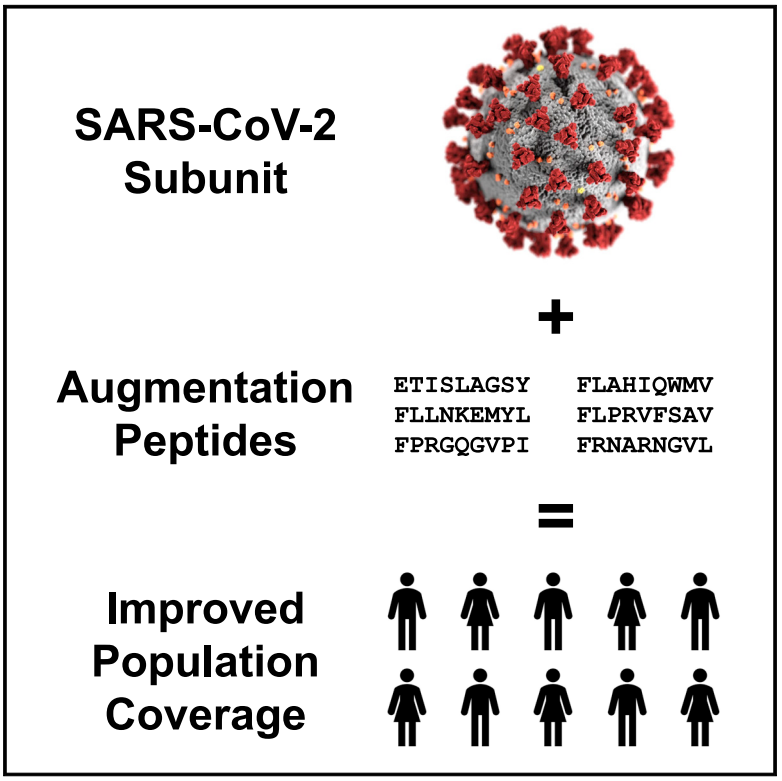
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Cell Systems

Brief Report

Predicted Cellular Immunity Population Coverage Gaps for SARS-CoV-2 Subunit Vaccines and Their Augmentation by Compact Peptide Sets

Graphical Abstract



Authors

Ge Liu, Brandon Carter, David K. Gifford

Correspondence

gifford@mit.edu

In Brief

An integrated model of peptide-HLA immunogenicity based on data from COVID-19 patients and machine learning is used to evaluate and design COVID-19 vaccines for cellular immunity. Human population coverage gaps are predicted for COVID-19 subunit vaccines, and augmentation strategies are proposed to close these gaps. Compact peptide vaccine designs that do not include a subunit are presented that have good population coverage.

Highlights

- Clinical data and machine learning predict SARS-CoV-2 peptide-HLA immunogenicity
- Human population coverage gaps of COVID-19 subunit vaccines are predicted
- Subunit augmentation improves vaccine population coverage for cellular immunity
- Subunit-free peptide vaccines are predicted to have high population coverage

Liu et al., 2021, Cell Systems *12*, 102–107
January 20, 2021
https://doi.org/10.1016/j.cels.2020.11.010

CellPress

CellPress

Cell Systems

Brief Report

Predicted Cellular Immunity Population Coverage Gaps for SARS-CoV-2 Subunit Vaccines and Their Augmentation by Compact Peptide Sets

Ge Liu,[1,2] Brandon Carter,[1,2] and David K. Gifford[1,2,3,4,*]
[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[2]MIT Electrical Engineering and Computer Science, Cambridge, MA, USA
[3]MIT Biological Engineering, Cambridge, MA, USA
[4]Lead Contact
*Correspondence: gifford@mit.edu
https://doi.org/10.1016/j.cels.2020.11.010

SUMMARY

Subunit vaccines induce immunity to a pathogen by presenting a component of the pathogen and thus inherently limit the representation of pathogen peptides for cellular immunity-based memory. We find that severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) subunit peptides may not be robustly displayed by the major histocompatibility complex (MHC) molecules in certain individuals. We introduce an augmentation strategy for subunit vaccines that adds a small number of SARS-CoV-2 peptides to a vaccine to improve the population coverage of pathogen peptide display. Our population coverage estimates integrate clinical data on peptide immunogenicity in convalescent COVID-19 patients and machine learning predictions. We evaluate the population coverage of 9 different subunits of SARS-CoV-2, including 5 functional domains and 4 full proteins, and augment each of them to fill a predicted coverage gap.

INTRODUCTION

All reported current efforts for the COVID-19 vaccine design that are part of the United States Government's Operation Warp Speed use variants of the spike subunit of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to induce immune memory (Table S7). Subunit vaccines seek to reduce the safety risks of attenuated or inactivated pathogen vaccines by optimizing the portion of a pathogen that is necessary to produce durable immune memory (Moyle and Toth, 2013). Suggested coronavirus subunit vaccine components include the spike (S) protein, the receptor-binding domain (RBD) of S, the S1 domain of S, the S2 domain of S, the nucleocapsid (N), the membrane (M), the envelope (E), the N-terminal domain (NTD) of S, and the fusion peptide (FP) of S (Wang et al., 2020; Yu et al., 2020; Dai et al., 2020). Subunit vaccines have been enabled by our ability to engineer and express pathogen surface components that retain their three-dimensional structure to induce neutralizing antibodies and a corresponding B cell memory. However, the production of durable immune memory rests in part upon help from T cells, which get their cues from peptides displayed by human leukocyte antigen (HLA) molecules encoded by the major histocompatibility complex (MHC) of genes. Since a subunit vaccine does not fully represent a pathogen, vaccine excluded pathogen peptides will not be observed during vaccination by an individual's T cells.

We find that proposed SARS-CoV-2 subunit vaccines exhibit population coverage gaps in their ability to generate a robust number of predicted peptide-HLA hits in every individual. A *peptide-HLA hit* is the potential immunogenic display of a peptide by a single HLA allele. Subunit vaccine-based simulation of a T cell response is limited because of their limited representation of pathogen peptides, and the preferences of each individual's HLA molecules for the peptides they will bind and display. Since HLA loci exhibit linkage disequilibrium, we use the frequencies of population haplotypes in our coverage computations. Each haplotype describes the joint appearance of HLA alleles. Cytotoxic $CD8^+$ T cells observe peptides displayed by molecules encoded by an individual's classical class I loci (HLA-A, HLA-B, and HLA-C), and helper $CD4^+$ T cells observe peptides displayed by molecules encoded by an individual's classical class II loci (HLA-DR, HLA-DQ, and HLA-DP).

RESULTS

We model peptide-HLA immunogenicity by combining data from convalescent COVID-19 patients as measured by the multiplexed identification of T cell receptor antigen specificity (MIRA) assay with machine learning predictions (Snyder et al., 2020; Klinger et al., 2015). The display of a peptide by an HLA molecule is necessary, but not sufficient, for the peptide to be immunogenic and causes T cell activation and expansion. The combined model predicts which HLA molecule displayed a peptide that was observed to be immunogenic in a MIRA experiment, and uses machine learning predictions of peptide display for HLA alleles not observed or peptides not tested in MIRA data (STAR Methods). We use this combined model of peptide immunogenicity to compute our estimates of vaccine population coverage and to propose augmentation peptides to close population coverage gaps.

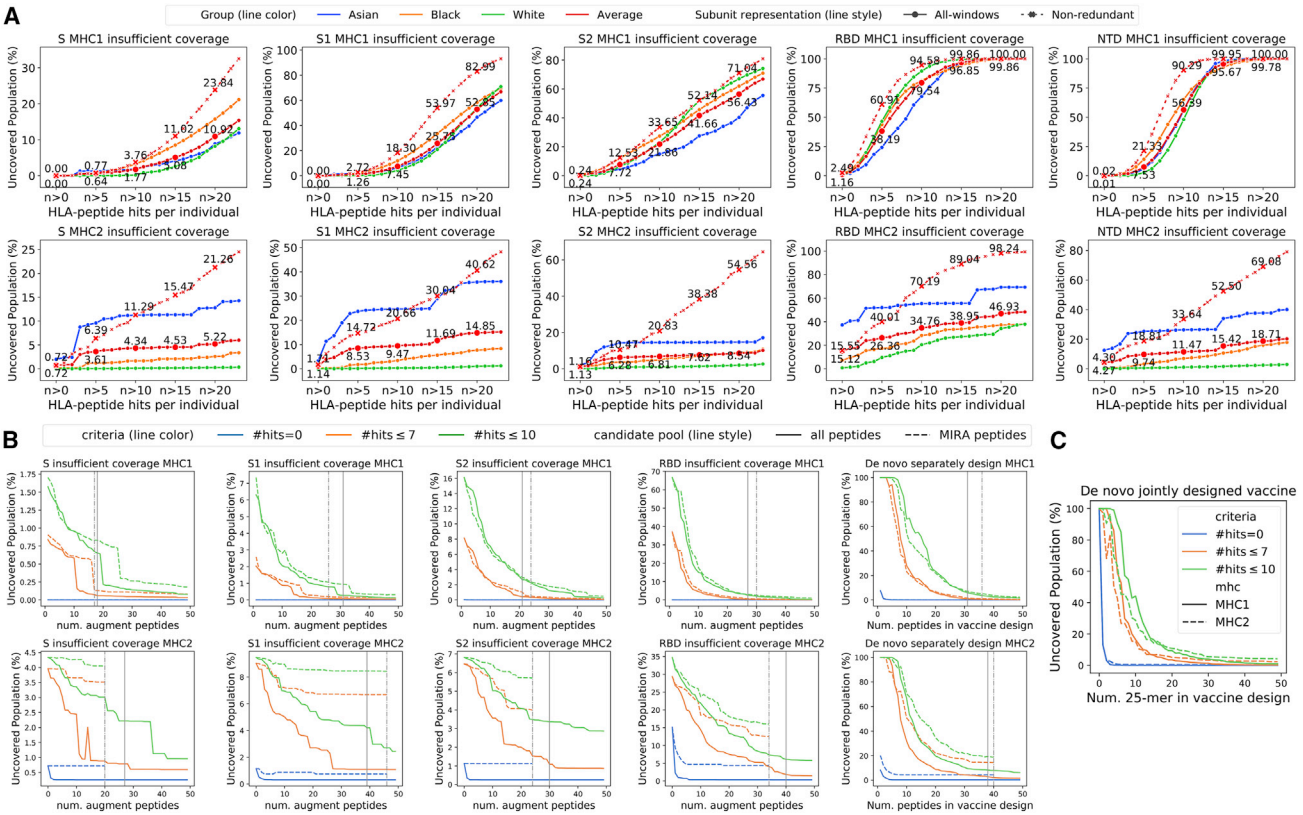


Figure 1. Predicted Human Population Coverage Gaps and Improvement with Proposed Vaccines

(A) Predicted uncovered percentage of populations as a function of the minimum number of peptide-HLA hits in an individual. Annotated percentages are the average across populations self-reporting as Asian, Black, and white. A redundant sampling of peptides is depicted by solid lines. A nonredundant sampling of peptides is depicted by dotted lines.

(B) Predicted uncovered percentage of the population for a subunit plus augmentation peptides or a subunit free design as a function of the number of augmentation peptides, MHC class I (top row) and class II (bottom row).

(C) Uncovered population for a joint class I and class II *de novo* vaccine design that does not include a subunit. Dotted graph lines in (B) utilize only MIRA validated peptides. In (B) vertical lines show the peptide count used to evaluate Table S1, dotted lines are MIRA peptides only.

We use two different candidate sets of peptides for subunit augmentation and *de novo* vaccine design—known immunogenic peptides and all possible peptides. First, we exclusively use the set of peptides that were observed to be immunogenic in the MIRA assay (Snyder et al., 2020; Klinger et al., 2015). Second, we utilize peptides from the SARS-CoV-2 genome that have a mutation rate of < 0.001 and a zero glycosylation probability predicted by NetNGlyc (Gupta et al., 2004) (STAR Methods).

For our vaccine coverage predictions, we use previously reported estimates of HLA haplotype frequencies from Liu et al., (2020) for HLA-A, HLA-B, and HLA-C (classical class I) and HLA-DR, HLA-DQ, and HLA-DP (classical class II) to score the number of peptide-HLA hits observed for various subunits of SARS-CoV-2 with and without additional augmentation peptides. We also evaluate subunit population coverage for MHC class I using HLA haplotype frequencies from Gragert et al., (2013) (STAR Methods).

SARS-CoV-2 Subunit Population Coverage Analysis

We first used our model of peptide immunogenicity to compute a baseline of the predicted number of peptide-HLA hits that would result from an infection by the SARS-CoV-2 virus using the HLA haplotype frequencies from Liu et al., (2020). For this task, we extracted all peptides of length 8–10 (MHC class I) and 13–25 (MHC class II) inclusive from the SARS-CoV-2 proteome (STAR Methods).

We predict that SARS-CoV-2 will have 318 (white), 307 (Black), and 391 (Asian) peptide-HLA hits for MHC class I on average in the respective self-reporting human populations. For an MHC class II redundant sampling, we predict that SARS-CoV-2 will have 5,180 (white), 3,871 (Black), and 2,070 (Asian) peptide-MHC hits. Thus, the average number of predicted SARS-CoV-2 peptide-HLA hits for MHC class I is 338 and for MHC class II 3,707.

We found that all subunits of SARS-CoV-2 have gaps in their predicted human population coverage for robust peptide MHC display using EvalVax (STAR Methods). We computed the predicted uncovered population percentage of the SARS-CoV-2 subunits S, S1, S2, RBD, and NTD as a function of the minimum required predicted number of peptide-HLA hits displayed by an individual (Figure 1). An individual is *uncovered* if they are not predicted to have a specified number of peptide-HLA hits. Results for the FP, M, N, and E subunits are shown in Figure S3. We observed a negative correlation between subunit size and

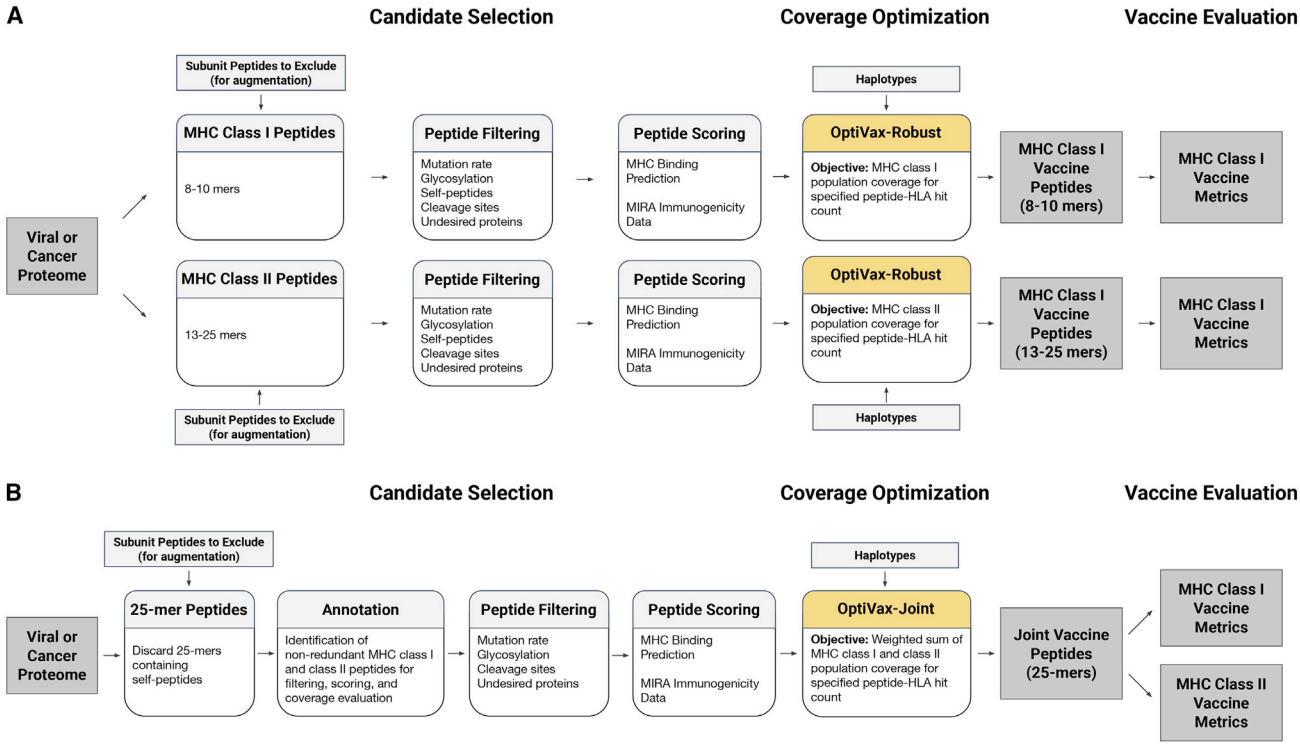


Figure 2. The Separate and Joint Design Methods for Peptide Vaccines

(A) In the separate method, windowed pathogen proteomes are filtered for acceptable peptides and MHC class I and class II vaccine designs are chosen to optimize population coverage at specified levels of peptide-HLA hits.

(B) In the joint method, 25-mer pathogen peptides are annotated with their MHC class I and class II peptides, which are filtered, scored, evaluated for population coverage, and used to optimize the selection of their parent 25-mers into a joint vaccine.

predicted population gap (Pearson $r = -0.65$ for MHC class I and Pearson $r = -0.63$ for MHC class II, Figure S4). The predicted fraction of the uncovered population was greatest for the smallest subunits, which is a direct consequence of their elimination of the largest number of pathogen peptides. For example, of the S protein subunits, RBD is the smallest and has the largest predicted population coverage gap. While the significance of the reduced immune footprint of subunit vaccines remains to be fully elucidated when no or very few peptide-HLA hits are predicted, a corresponding reduction in T cell activation, expansion, and memory function would be expected. Based on our prediction, the RBD subunit had no MHC class II peptides displayed in 15.12% of the population (averaged across Asian, Black, and white self-reporting individuals). We noticed that the uncovered population of RBD with no predicted display of MHC class II peptides ranges from 0.811% for the population self-reporting as white, to a high of 37.287% for the population self-reporting as Asian. The high uncovered population in the Asian population is caused by the HLA haplotype frequencies in the Asian population. Thus, clinical trials need to carefully consider ancestry in their study designs to ensure that efficacy is measured across an appropriate population. For the RBD subunit, 26.357% of the population had fewer than six MHC class II peptide-HLA hits. For RBD MHC class I, the coverage gap is 1.163% for no hits and 38.186% for fewer than six hits. EvalVax predicted that on average for an S subunit vaccine the uncovered population would be 0.001% (class I) and 0.721% (class II) for no display, and 0.642% (class I) and 3.610% (class II) for fewer than 6 peptide-HLA hits. Table S9 provides a list of diplotypes with zero predicted peptide-HLA hits for each SARS-CoV-2 subunit.

We found that predicted subunit coverage gaps for MHC class I were largely consistent when we utilized HLA haplotype frequencies from Gragert et al., (2013) (Figure S7) (STAR Methods).

SARS-CoV-2 Subunit Augmentation with Peptide Sets for MHC Class I and II

We used Optivax-Robust to compute separate MHC class I and II augmentation sets of SARS-CoV-2 peptides to be combined with each subunit to maximize the predicted population coverage for a target minimum number of MHC class I and class II peptide hits in every individual (Figure 2). We used two sets of candidate peptides for these augmentation sets: (1) peptides that were found to be immunogenic in MIRA assay data, and (2) all filtered peptides from the entire SARS-CoV-2 proteome (STAR Methods). The use of peptides immunogenic in MIRA data is intended to ensure that vaccine peptides are immunogenic while limiting population coverage by not considering other peptides that may cover rare MHC alleles. We predicted the uncovered fraction of the population as a function of MHC class I or II peptides set size for both candidate sets (Figure 1; Table S1). The computed sets of augmentation peptides were predicted to substantially reduce the populations predicted to be insufficiently covered by each subunit. Post augmentation, the predicted uncovered population for RBD with no peptide-MHC hits is reduced to 0.003% (MHC class I) and 4.351% (MHC class II) with MIRA positive peptides only, and

0.0% (MHC class I) and 0.309% (MHC class II) with all filtered peptides from SARS-CoV-2 (Table S1). We were conservative and did not consider potential MHC class II trans-isoforms that could improve coverage during vaccine design.

A Peptide-Only SARS-CoV-2 Vaccine for MHC Class I and II

We designed *de novo* peptide-only vaccines that did not assume an associated subunit component. We proposed either peptide vaccines that separately optimize for MHC class I and class II population coverage or a single joint peptide vaccine that optimizes MHC class I and class II coverage simultaneously (STAR Methods). As candidates for vaccine inclusion, we considered (1) only peptides that were found to be immunogenic in the MIRA data, and (2) all peptides from the SARS-CoV-2 proteome for separate vaccine designs. For joint vaccine design, we included all peptides from the SARS-CoV-2 proteome as candidates as the MIRA positive peptides do not overlap sufficiently to do joint optimization. We explored the predicted decrease of the uncovered population as a function of peptide count and found that *de novo* vaccine designs are predicted to simultaneously produce a large number of predicted hits for both MHC class I and class II display (Figure 1; Tables S3 and S4). The predicted population coverage of our peptide only designs exhibited a diverse display of peptides across populations self-reporting as Black, white, and Asian (Figure S5). Peptide-only vaccine designs have been found to be effective (Herst et al., 2020).

SARS-CoV-2 Joint *De Novo* Designs Are More Compact than Separate Designs

We found that a joint *de novo* design that uses a single set of peptides for MHC class I and class II coverage requires fewer peptides than separate vaccines for MHC class I and class II coverage (Figure 1). With 9 jointly selected peptides, more than 93% of the population was predicted to have more than 4 peptide-HLA hits. A 24-peptide joint design was predicted to produce more than 4 peptide HLA hits in more than 99% of the population (Table S3). We set a series of population coverage goals for MHC class I and MHC class II coverage with more than 7 peptide-HLA hits per individual. We considered 25 evenly spaced coverage levels between 0% and 100% coverage. We computed the total number of peptides needed to reach each set of MHC class I and MHC class II coverage goals simultaneously, where the number of MHC class I and MHC class II peptides are summed for separately designed peptide sets. We computed the total number of amino acids needed for a construct with a typical mRNA delivery platform with 10-amino acid linkers (Sahin et al., 2017) (Supplemental Information). We found that joint designs reduce the total number of required amino acids and peptides required to achieve each level of population coverage (Figure S6). A single mRNA delivery construct has been demonstrated to work for both MHC class I and class II peptides (Kreiter et al., 2008; Sahin et al., 2017).

DISCUSSION

We augment subunit vaccines with a compact set of peptides to improve the display and immunogenicity of a vaccine on HLA class I and II molecules across a population of people. Subunit vaccines offer safety advantages over inactivated or attenuated pathogen vaccines, but their ability to fully mimic a pathogenic infection to train cellular immunity is limited. Immunity to a pathogen may rest in part upon T-cell-based adaptive immunity and corresponding T memory cells. We expect that a vaccine that provides a diverse display of a pathogen's peptides will create reservoirs of $CD4^+$ and $CD8^+$ memory cells that will assist in establishing immunity to the pathogen. SARS-CoV-2 infection elicits a robust memory T cell response even in antibody-seronegative individuals, suggesting a T cell response is an important component of immunity to COVID-19 (Sekine et al., 2020).

We found that for SARS-CoV-2 the joint optimization of predicted MHC class I and class II pathogen peptide display achieves population coverage criteria with a more compact vaccine design than designing separate peptide sets for MHC class I and class II. Using a simpler design with shorter constructs may contribute to the effectiveness of a vaccine by providing an equivalent diversity of peptide display in a population with a less complex mixture of vaccine peptides.

Augmentation peptides can be delivered using the same vehicle as their associated subunit vaccine or they can be delivered separately. Nucleic acid-based vaccines can incorporate RNA or DNA sequences that encode class I and class II augmentation peptides with desired signal sequences, linkers, and protease cleavage sites (Kreiter et al., 2008; Sahin et al., 2017) (examples in Supplemental Information, Tables S5 and S6). The peptides can be expressed as part of the subunit or separately and can be encoded on the same or different molecules as the primary subunit. When augmentation peptides are added as a new subunit domain a vaccine designer can improve population coverage as described in Figure 1B. Nucleic acid constructs carrying augmentation peptides can be delivered by injection in lipid nanoparticle particle carriers or directly (Dowdy, 2017; Wolff et al., 1990). Protein-based vaccines can include independent augmentation peptides into the vaccine formulation. The delivery of independent augmentation peptides can be accomplished using nanoparticles (Herst et al., 2020).

Our computational objective function encodes the two key goals of our augmentation strategy—population coverage and the display of a highly diverse set of peptides in each individual. Our population coverage goal is ensured by optimizing predicted display coverage over population haplotype frequencies. The display of a diverse set of peptides is established by setting augmentation design goals for the number of peptides that need to be displayed by each individual.

Early results from clinical studies of subunit vaccines for SARS-CoV-2 show that some vaccine recipients did not develop positive $CD8^+$ T cell responses (Jackson et al., 2020). It is difficult to fully evaluate these results because the HLA types of study participants are not provided by these early studies. Thus, these study populations may not be reflective of HLA types in the general world population. The BNT162b1 RBD subunit vaccine produced a less robust $CD8^+$ response than $CD4^+$ response (Sahin et al., 2020), and this was also noted in the mRNA-1273 S subunit vaccine results (Jackson et al., 2020). Further clinical data are required to fully assess the T cell immunogenicity of various subunits and delivery methods. Clinical trials should select their participants to have representative HLA type distributions to test for population coverage. Future studies will need to examine the

durability of immunity in individuals with a minimal T cell response.

By simultaneously achieving the twin goals of coverage and diversity with peptides derived from a pathogen, we effectively compress the cellular immunologic fingerprint of a pathogen into a vaccine. To produce an antibody response, the subunit component of a vaccine can encode a three-dimensional epitope to stimulate neutralizing antibody production by B cells. Taken together, these two designed components, a pathogen subunit and its augmentation, will provide both B cell and T cell epitopes of a pathogen while permitting epitope selection to mitigate deleterious effects and improve population coverage.

All of our software and data are freely available as open source to allow others to use and extend our methods.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - SARS-CoV-2 Proteome and Candidate Peptides
  - EvalVax Subunit Vaccine Evaluation
  - Design of a Single Set of Peptides to Maximize MHC Class I and II Population Coverage
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.11.010.

ACKNOWLEDGMENTS

This work was supported in part by Schmidt Futures, a Google Cloud Platform grant, NIH grant R01CA218094, and a C3.AI DTI Award to D.K.G. We benefited from thoughtful comments from Michael Birnbaum, Mary Carrington, and Brooke Huisman. Ge Liu's contribution was made prior to joining Amazon.

AUTHOR CONTRIBUTIONS

G.L., B.C., and D.G. contributed to problem definition and solution. G.L. designed and implemented the optimization procedure, with advice from B.C. and D.G.; G.L., B.C., and D.G. wrote the paper.

DECLARATION OF INTERESTS

Brandon Carter is an employee of Think Therapeutics, and Ge Liu is an employee of Amazon. David Gifford and Brandon Carter have a financial interest in Think Therapeutics, Inc.

Received: August 4, 2020
Revised: October 18, 2020
Accepted: November 19, 2020
Published: November 27, 2020

REFERENCES

Dai, L., Zheng, T., Xu, K., Han, Y., Xu, L., Huang, E., An, Y., Cheng, Y., Li, S., Liu, M., et al. (2020). A universal design of Betacoronavirus vaccines against COVID-19, MERS, and SARS. Cell *182*, 722–733.e11.

Dowdy, S.F. (2017). Overcoming cellular barriers for RNA therapeutics. Nat. Biotechnol. *35*, 222–229.

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

Gragert, L., Madbouly, A., Freeman, J., and Maiers, M. (2013). Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire us donor registry. Hum. Immunol. *74*, 1313–1320.

Gupta, R., Jung, E., and Brunak, S. (2004). Prediction of N-glycosylation sites in human proteins. http://www.cbs.dtu.dk/services/NetNGlyc/.

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.

Herst, C.V., Burkholz, S., Sidney, J., Sette, A., Harris, P.E., Massey, S., Brasel, T., Cunha-Neto, E., Rosa, D.S., Chao, W.C.H., et al. (2020). An effective CTL peptide vaccine for Ebola zaire based on survivors' CD8+ targeting of a particular nucleocapsid protein epitope with potential implications for COVID-19 vaccine design. Vaccine *38*, 4464–4475.

Jackson, L.A., Anderson, E.J., Rouphael, N.G., Roberts, P.C., Makhene, M., Coler, R.N., McCullough, M.P., Chappell, J.D., Denison, M.R., Stevens, L.J., et al. (2020). An mRNA vaccine against SARS-CoV-2 — preliminary report. N. Engl. J. Med. *383*, 1920–1931, https://doi.org/10.1056/NEJMoa2022483.

Jensen, K.K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J.A., Yan, Z., Sette, A., Peters, B., and Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. Immunology *154*, 394–406.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol. *199*, 3360–3368.

Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., Moorhead, M., and Faham, M. (2015). Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. PLoS One *10*, e0141561.

Kreiter, S., Selmi, A., Diken, M., Sebastian, M., Osterloh, P., Schild, H., Huber, C., Türeci, O., and Sahin, U. (2008). Increased antigen presentation efficiency by coupling antigens to MHC class I trafficking signals. J. Immunol. *180*, 309–318.

Liu, G., Carter, B., Bricken, T., Jain, S., Viard, M., Carrington, M., and Gifford, D.K. (2020). Computationally optimized SARS-CoV-2 MHC class I and II vaccine formulations predicted to target human haplotype distributions. Cell Syst. *11*, 131–144.e6.

Moyle, P.M., and Toth, I. (2013). Modern subunit vaccines: development, components, and research opportunities. ChemMedChem *8*, 360–376.

O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. Cell Syst. *7*, 129–132.e4.

O'Donnell, T.J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell Syst. *11*, 42–48.e7.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Ramsuran, V., Naranbhai, V., Horowitz, A., Qi, Y., Martin, M.P., Yuki, Y., Gao, X., Walker-Sperling, V., Del Prete, G.Q., Schneider, D.K., et al. (2018). Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells. Science *359*, 86–90.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020a). 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. *48*, W449–W454.

Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., and Nielsen, M. (2020b). Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J. Proteome Res. *19*, 2304–2315.

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Löwer, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrörs, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature *547*, 222–226.

Sahin, U., Muik, A., Derhovanessian, E., Vogler, I., Kranz, L.M., Vormehr, M., Baum, A., Pascal, K., Quandt, J., Maurus, D., et al. (2020). COVID-19 vaccine BNT162b1 elicits human antibody and TH1 T cell responses. Nature *586*, 594–599, https://doi.org/10.1038/s41586-020-2814-7.

Sekine, T., Perez-Potti, A., Rivera-Ballesteros, O., Strålin, K., Gorin, J.B., Olsson, A., Llewellyn-Lacey, S., Kamal, H., Bogdanovic, G., Muschiol, S., et al. (2020). Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19. Cell *183*, 158–168.e14.

Snyder, T.M., Gittelman, R.M., Klinger, M., May, D.H., Osborne, E.J., Taniguchi, R., Zahid, H.J., Kaplan, I.M., Dines, J.N., Noakes, M.N., et al. (2020). Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. medRxiv. https://doi.org/10.1101/2020.07.31.20165647.

Srinivasan, S., Cui, H., Gao, Z., Liu, M., Lu, S., Mkandawire, W., Narykov, O., Sun, M., and Korkin, D. (2020). Structural genomics of SARS-CoV-2 indicates evolutionary conserved functional regions of viral proteins. Viruses *12*, 360.

Tang, M., Lautenberger, J.A., Gao, X., Sezgin, E., Hendrickson, S.L., Troyer, J.L., David, V.A., Guan, L., Mcintosh, C.E., Guo, X., et al. (2012). The principal genetic determinants for nasopharyngeal carcinoma in China involve the HLA class I antigen recognition groove. PLoS Genet. *8*, e1003103.

UniProt, Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res *47*, D506–D515.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods *17*, 261–272.

Wang, N., Shang, J., Jiang, S., and Du, L. (2020). Subunit vaccines against emerging pathogenic human coronaviruses. Front. Microbiol. *11*, 298.

Wolff, J.A., Malone, R.W., Williams, P., Chong, W., Acsadi, G., Jani, A., and Felgner, P.L. (1990). Direct gene transfer into mouse muscle in vivo. Science *247*, 1465–1468.

Yu, J., Tostanoski, L.H., Peter, L., Mercado, N.B., McMahan, K., Mahrokhian, S.H., Nkolola, J.P., Liu, J., Li, Z., Chandrashekar, A., et al. (2020). DNA vaccine protection against SARS-CoV-2 in rhesus macaques. Science *369*, 806–811.

Zeng, H., and Gifford, D.K. (2019). Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst. *9*, 159–166.e3.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Computational model predictions	This paper	Mendeley Data: https://doi.org/10.17632/gs8c2jpvdn.1
HLA haplotype population frequencies	Liu et al., 2020	Mendeley Data: https://doi.org/10.17632/cfxkfy9zp4.1
Tables S8 and S9	This paper	Mendeley Data: https://doi.org/10.17632/gs8c2jpvdn.1
SARS-CoV-2 proteome	GISAID (Elbe and Buckland-Merrett, 2017)	Sequence entry Wuhan/IPBCAMS-WH-01/2019, used data as processed and provided by (Liu et al., 2020)
Human proteome	(UniProt, 2019)	UniProt: UP000005640 (Proteome ID)
HLA class I haplotype frequencies	Gragert et al., 2013	NMDP full 2011 dataset (HLA-A~C~B) from http://frequency.nmdp.org
MIRA COVID-19 Immunogenicity data (MHC class I and II)	Snyder et al., 2020	ImmuneCODE-MIRA-Release002.1; https://adaptivepublic.blob.core.windows.net/publishedproject-supplements/covid-2020/ImmuneCODE-MIRA-Release002.1.zip
Software and Algorithms		
OptiVax and EvalVax	This paper	https://github.com/gifford-lab/optivax/tree/master/augmentation_paper
NetMHCpan-4.0	Jurtz et al., 2017	http://www.cbs.dtu.dk/services/NetMHCpan-4.0/
NetMHCpan-4.1	Reynisson et al., 2020a	http://www.cbs.dtu.dk/services/NetMHCpan-4.1/
NetMHCIIpan-3.2	Jensen et al., 2018	http://www.cbs.dtu.dk/services/NetMHCIIpan-3.2/
NetMHCIIpan-4.0	Reynisson et al., 2020b	http://www.cbs.dtu.dk/services/NetMHCIIpan-4.0/
PUFFIN	Zeng and Gifford, 2019; https://github.com/gifford-lab/PUFFIN	GitHub commit a63f6c563b7e2f7b04eac28a6cf09d8078ac3a2a with pre-trained model
MHCflurry 2.0	O'Donnell et al., 2020	Version 2.0, https://github.com/openvax/mhcflurry

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, David K. Gifford (gifford@mit.edu).

Materials Availability

This study did not generate new materials.

Data and Code Availability

- All source data for generating population coverage curves have been deposited and are publicly available at https://github.com/gifford-lab/optivax. The peptide scoring predictions and processed haplotype frequencies have been deposited to Mendeley Data: https://doi.org/10.17632/gs8c2jpvdn.1. This paper analyzes existing, publicly available data. These datasets' accession numbers are provided in the Key Resources Table.
- All original code is publicly available at https://github.com/gifford-lab/optivax.
- The scripts used to generate the figures reported in this paper are available at https://github.com/gifford-lab/optivax
- Any additional information required to reproduce this work is available from the Lead Contact.

METHOD DETAILS

SARS-CoV-2 Proteome and Candidate Peptides

The SARS-CoV-2 proteome is comprised of four structural proteins (E, M, N, and S) and open reading frames (ORFs) encoding nonstructural proteins (Srinivasan et al., 2020). We obtained the SARS-CoV-2 viral proteome from GISAID (Elbe and Buckland-Merrett, 2017) sequence entry Wuhan/IPBCAMS-WH-01/2019, the first documented case, as processed and provided by Liu et al.

(2020). Nextstrain (Hadfield et al., 2018) was used to identify ORFs and translate the sequence. We use sliding windows to extract all peptides of length 8–10 (MHC class I) and 13–25 (MHC class II) inclusive from the SARS-CoV-2 proteome, resulting in 29,403 peptides for MHC class I and 125,593 peptides for MHC class II.

For vaccine augmentation we use two different candidate sets: known immunogenic peptides, and all possible peptides. First, we exclusively use the set of peptides that were observed to be immunogenic in the MIRA assay (Snyder et al., 2020; Klinger et al., 2015). In this case we use the MIRA sets identified for MHC class I and II separately. Second, we use the same filtered candidate peptide set as Liu et al. (2020), in which peptides with mutation rate > 0.001 or non-zero glycosylation probability predicted by NetNGlyc (Gupta et al., 2004) are filtered.

MIRA provides immunogenicity data with peptide-detail data that summarizes for each individual (MIRA experiment) the peptide sets that were found to cause T cell activation. A MIRA peptide set can be a single peptide, or a group of highly related peptides that are samples from slightly offset positions in the proteome. The MIRA subject-metadata contains the HLA types for individuals. While the HLA type of an individual provides us with the candidate HLA alleles that could display a given peptide, it does not tell us which allele displayed the peptide. The MIRA data used in this study includes 119 (MHC class I) or 8 (MHC class II) convalescent HLA-typed COVID-19 patients that were queried for $CD8^+$ T cell activation for 269 peptide pools (generated from 545 peptides) or $CD4^+$ activation for 56 peptide pools (generated from 251 peptides). Each peptide pool contains at most 13 MHC class I peptides or up to 6 MHC class II peptides. The patient population had 110 (MHC class I) and 22 (MHC class II) HLA alleles (Snyder et al., 2020). We included all MIRA immunogenic peptides for vaccine analysis and design, as to date no peptide has been observed to cause immunopathology that exacerbates disease severity. The MIRA data are incomplete, as they are a sample of the peptide complement of SARS-CoV-2 and convalescent patient HLA alleles. We use these data to select and calibrate machine learning models to make immunogenicity predictions for all peptide and HLA combinations.

Since the MIRA assay does not identify the patient HLA allele that presents a peptide and does not distinguish between individual peptides in a given pool, we built a combined model of MIRA observations and machine learning predictions to model peptide immunogenicity when presented by a specific HLA. We did not use the predicted HLA restrictions from Snyder et al. (2020)'s Table S2 as it identified pools of peptides, and not individual peptides and is only for MHC class I. For an HLA allele that appeared in the MIRA data and peptides that were tested, a peptide was predicted to be immunogenic when displayed by that HLA allele if (1) it was immunogenic in the MIRA data in 38% (MHC class I) or 40% (MHC class II) of individuals that had the HLA allele, and (2) it was predicted to bind to the HLA allele with an affinity of at least 500 nM. We used the prevalence of immunogenic peptides across individuals as criteria (1) as it performed better than using the prevalence of TCR sequences of immunogenic peptides. Other criteria that we explored that did not perform as well are included in Table S2. The selected criteria maximized the AUROC for prediction of the MIRA data that contained both positive and negative examples of peptide pool immunogenicity for individuals with a given HLA type (Table S2). Criteria (2) allowed us to predict the specific HLA allele(s) that displayed a peptide since MIRA data provides all of the HLA alleles for a given individual and does not provide information on which allele(s) displayed a peptide. We evaluate a peptide-HLA immunogenicity model using the MIRA data, and score a MIRA pool-individual pair positive if at least one peptide in the pool is predicted by the model to be immunogenic when displayed by one of the HLAs of the individual. When computing ROC or PRC curves where a variable decision boundary is employed, the maximum score across all pool peptides and HLAs is utilized for evaluation. Our combined model of HLA specific peptide immunogenicity predictions has a precision of 0.581 and AUROC of 0.833 (MHC class I) and precision of 0.849 and AUROC of 0.923 (MHC class II) (Table S2; Figure S1).

For HLA alleles not present or peptides not tested in MIRA data we use machine learning predictions of peptide immunogenicity. We evaluated machine learning methods by their ability to predict MIRA peptides that are immunogenic in an individual based upon the HLA type of the individual (Figure S1). For a given individual we used both positive and negative sets to characterize their performance, and we prioritized precision for conservative vaccine design (Table S2). We found for MHC class I the best method utilized a 50 nM threshold from an ensemble that outputs the mean predicted binding affinity of NetMHCpan-4.0 (Jurtz et al., 2017), PUFFIN (Zeng and Gifford, 2019), and MHCflurry 2.0 (O'Donnell et al., 2020, 2018). We selected this ensemble as it is more robust to errors by a single method. For MHC class II the method we selected used a 50 nM threshold and NetMHCIIpan-4.0 (Reynisson et al., 2020b). Our machine learning predictions of HLA specific peptide display have a precision of 0.447 and AUROC of 0.715 (MHC class I) and a precision of 0.869 and AUROC of 0.701 (MHC class II) for immunogenicity (Figure S1). Other methods we explored included NetMHCpan-4.1 (Reynisson et al., 2020a) (MHC class I), PUFFIN (Zeng and Gifford, 2019) (MHC class II) and NetMHCIIpan-3.2 (Jensen et al., 2018) (MHC class II) (Table S2).

We use HLA class I and class II haplotype frequencies provided by Liu et al. (2020). HLA haplotype frequencies were generated from previously published next-generation sequencing data generated in the Carrington lab and their collaborators (Tang et al., 2012; Ramsuran et al., 2018). All of these HLA data are based upon genome sequencing that provides the highest resolving power for HLA typing. The self-reporting ancestries were: Asian descent: Southern Chinese; White descent: 1.8% German, 98.2% European American; and Black descent: 29.6% African American, 36.6% South African, 27% Ugandan, and 6.8% West African. For the HLA class I locus, this dataset contains 2,138 distinct haplotypes spanning 230 HLA-A, HLA-B, and HLA-C alleles. For HLA class II, this dataset contains 1,711 distinct haplotypes spanning 280 HLA-DP, HLA-DQ, and HLA-DR alleles. Population frequencies are provided for three populations self-reporting as having White, Black, or Asian ancestry. We used these data for vaccine evaluation and design as they included the HLA-DQA and HLA-DPA/DPB alleles for MHC class II that are not present in Gragert et al. (2013).

We also predicted MHC class I population coverage using Gragert et al. (2013). For this analysis we used a combined immunogenicity model of peptide-HLA immunogenicity with an ensemble of NetMHCpan-4.0 and MHCflurry 2.0 (Jurtz et al., 2017; O'Donnell et al., 2020, 2018) for machine learning predictions.

We consider nine subunit vaccines for SARS-CoV-2: the full envelope (E), membrane (M), nucleocapsid (N), and spike (S) proteins as well as the S1, S2, receptor binding domain (RBD), N-terminal domain (NTD), and fusion peptide (FP) domains from S. The amino acid positions for each of the S protein subunits are shown in Figure S2. When evaluating these subunit vaccines we include all peptides of length 8–10 (MHC class I) and 13–25 (MHC class II) spanning the corresponding regions of the proteome.

EvalVax Subunit Vaccine Evaluation

We evaluate population coverage of SARS-CoV-2 subunit vaccines using EvalVax-Robust (Liu et al., 2020). EvalVax-Robust computes population coverage of a given peptide set using the HLA haplotype frequencies in each population of individuals self-reporting as having Black, Asian, or White ancestry. Population coverage $P(n)$ is defined as the fraction of individuals predicted to have $\geq n$ peptide-HLA hits using our model of peptide-HLA immunogencity. EvalVax-Robust computes the frequency of diploid HLA genotypes, and accounts for both homozygous and heterozygous HLA loci. We compute the average population coverage as an unweighted average of population coverage over the three populations. Insufficient coverage of $\leq n$ hits is defined as $100\% - P(n+1)$.

Our subunit population coverage estimates are not lowered by discarding subunit peptides as unsuitable. We consider all peptides that result from a windowing of the subunit proteome, and include the redundant peptides caused by using varying window sizes at the same proteome start position. In addition, we do not filter peptides for mutation rate or glycosylation during evaluation.

Design of Separate MHC Class I and II Peptide Sets to Augment Subunit Vaccine Population Coverage

In the separate design method we use OptiVax-Robust (Liu et al., 2020) to augment subunit vaccines with additional peptides to produce separate sets of peptides for class I and class II augmentation (Figure 2A). The candidate peptides for vaccine inclusion are chosen from either: (1) all peptides observed to be immunogenic in a MIRA assay, or (2) all filtered peptides from the SARS-CoV-2 proteome. All filtered peptides are selected from the remaining SARS-CoV-2 proteome (all peptides except those spanning the subunit), excluding peptides that are likely to mutate (have mutation rate > 0.001) or have non-zero predicted probability of glycosylation. All candidate peptides considered during augmentation must be predicted to be immunogenic using our model of peptide-HLA immunogenicity.

The augmentation algorithm uses a starting peptide set which is extracted from the subunit vaccine to maximize the coverage of the subunit while removing redundant peptides resulting from overlapping sliding windows using the redundancy elimination algorithm found in Liu et al. (2020). Using a non-redundant starting peptide set ensures that augmentation does not depend upon redundant peptides for population coverage support. OptiVax-Robust performs vaccine augmentation by adding peptides to this starting set to improve the population coverage at each peptide-HLA hits cutoff n. At each iteration redundant peptides are removed from consideration, and redundancy is defined with an edit distance metric (Liu et al., 2020). OptiVax-Robust uses a beam search algorithm that iteratively expands the solution by one peptide and gradually optimizes population coverage from $n = 1$ to the targeting level of per-individual peptide-HLA hits (Liu et al., 2020). We use a beam size of 5 for the augmentation of subunit vaccines.

For each desired budget of augmentation peptides, OptiVax produces an augmentation set. Larger augmentation sets are not necessarily supersets of smaller augmentation sets, as the underlying combinatorial optimization problem is complex. A vaccine designer can evaluate how many peptides they wish to use to realize a predicted population coverage. For the augmentation sets in Table S1 for $n = 7$ we targeted 99.3% coverage for MHC class I augmentation and 98% coverage for MHC class II. The exceptions were S and S1, where we targeted for MHC class I 99.9% coverage (all peptides) or 99.7% (MIRA peptide only), and for class II 98.5% (all peptides) or 98% (MIRA peptides). Class II is more difficult to cover with MIRA peptides alone, and thus we accept the best coverage possible. Augmentation sets are computed starting with non-redundant subunits to avoid peptide-hit credit for windowing induced redundancies. For the evaluation of original and augmented subunit vaccines in Table S1, we provide results for all window derived subunit peptides and the non-redundant set of subunit peptides. All window peptides can include the same HLA binding epitope multiple times from its sampling by multiple windows, and thus serves as the predicted lower bound on population insufficient coverage. The non-redundant results are the predicted upper bound of population insufficient coverage.

Design of a Single Set of Peptides to Maximize MHC Class I and II Population Coverage

We developed the OptiVax-Joint method to produce a minimal set of 25-mer peptides to reach a target population coverage probability at a threshold of n predicted hits for each individual for both MHC class I and class II (Figure 2B). The 25-mer candidate peptides are produced by windowing the pathogen proteome that is not part of a selected subunit, using a window step size of 8 amino acids between candidate peptides. Each of the candidate 25-mer peptides is annotated with its non-redundant peptides of length 8–10 (MHC class I) and 13–25 (MHC class II) and the HLA alleles where they are predicted to be immunogenic. Peptide redundancy is defined with an edit distance metric for the elimination of overlapping peptides (Liu et al., 2020).

OptiVax-Joint begins with the empty set, and performs vaccine augmentation by adding candidate 25-mer peptides to this starting set to improve both MHC class I and class II population coverage at a target number of peptide-HLA hits n. When OptiVax-Joint is started with an empty set of peptides it produces a *de novo* peptide vaccine design without an associated subunit component. Each

25-mer is scored based on its contained annotated class I and class II peptides for its improvement in the number of per-individual peptide-HLA hits (Liu et al., 2020) over the haplotypes of the target population. Contained peptides are not counted towards population coverage if they have an observed mutation rate > 0.001 or have a non-zero predicted probability of glycosylation. OptiVax-Joint uses a beam search algorithm that iteratively expands the solution by one 25-mer peptide and gradually optimizes population coverage from $n = 1$ peptide hit to the targeted level of per-individual peptide-HLA hits for both MHC class I and class II (Liu et al., 2020). We use a beam size of 5 for the augmentation of subunit vaccines.

For each desired budget of peptides, OptiVax-Joint produces a vaccine peptide set. Larger sets are not necessarily supersets of smaller augmentation sets, as the underlying combinatorial optimization problem is complex. A vaccine designer can evaluate how many peptides they wish to use to realize a predicted population coverage. For the joint sets in Table S1, we targeted 99% coverage at $n = 7$ for MHC class I augmentation and 97% coverage at $n = 7$ for MHC class II augmentation.

As a point of comparison, we also computed separate MHC class I and class II vaccine designs using OptiVax-Robust, using candidate sets drawn either from MIRA immunogenic peptides or all filtered peptides.

QUANTIFICATION AND STATISTICAL ANALYSIS

Classification performance of peptide-MHC scoring models was calculated using scikit-learn (Pedregosa et al., 2011) in Python using the *sklearn.metrics.roc_auc_score* (AUROC), *sklearn.metrics.average_precision_score* (Average Precision), *sklearn.metrics.accuracy_score* (Accuracy), *sklearn.metrics.precision_recall_fscore_support* (Precision, Recall and F1 score), and *sklearn.metrics.classification_report* (Sensitivity and Specificity) functions. AUROC and average precision are computed using raw predictions, and the remaining metrics are computed using binarized predictions based on the respective binding criteria. Pearson r correlation was computed using scipy (Virtanen et al., 2020) in Python using the *scipy.stats.pearsonr* function.