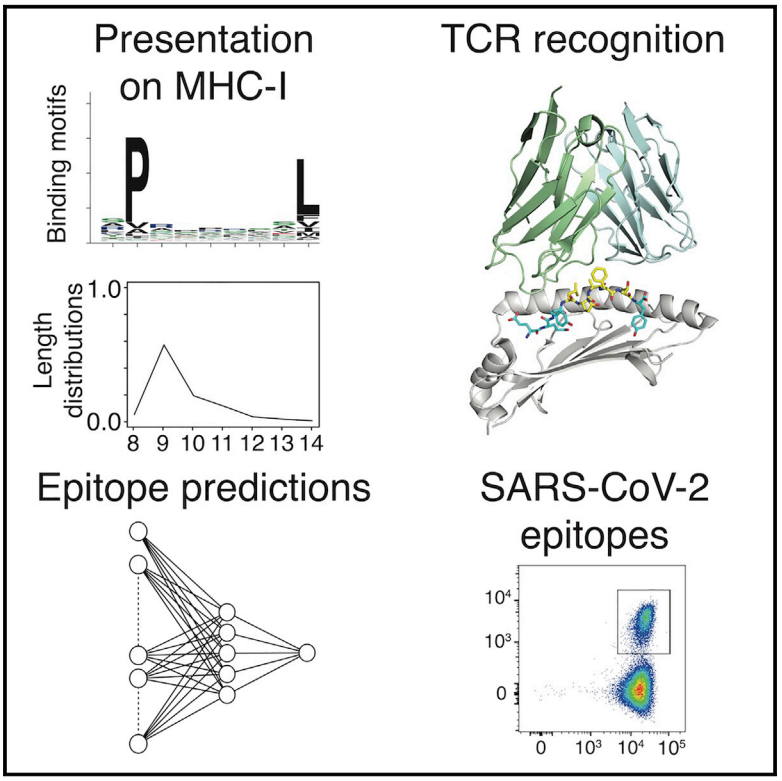
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Cell Systems

Report

Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8$^+$ T-cell epitopes

Graphical abstract



Authors

David Gfeller, Julien Schmidt, Giancarlo Croce, ..., Julien Cesbron, Julien Racle, Alexandre Harari

Correspondence

david.gfeller@unil.ch

In brief

We collected and curated large datasets of HLA-I ligands and neo-epitopes, which were used to train machine learning tools to predict antigen presentation (MixMHCpred2.2) and TCR recognition (PRIME2.0). Applying these tools to SARS-CoV-2 enabled us to identify potent CD8 T cell epitopes with cross-reactivity with other coronaviruses.

Highlights

- Collection and curation of a large dataset of HLA-I ligands and neo-epitopes
- Improved predictions of antigen presentation (MixMHCpred2.2)
- Improved predictions of TCR recognition (PRIME2.0)
- Identification of SARS-Cov-2 CD8$^+$ T cell epitopes

Gfeller et al., 2023, Cell Systems *14*, 72–83
January 18, 2023
https://doi.org/10.1016/j.cels.2022.12.002

CellPress

CellPress

Cell Systems

Report

Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8$^+$ T-cell epitopes

David Gfeller,[1,2,3,5,6,*] Julien Schmidt,[4,5] Giancarlo Croce,[1,2,3,5] Philippe Guillaume,[4,5] Sara Bobisse,[2,4,5] Raphael Genolet,[4,5] Lise Queiroz,[4,5] Julien Cesbron,[4,5] Julien Racle,[1,2,3,5] and Alexandre Harari[2,4,5]

[1]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University of Lausanne, Lausanne, Switzerland
[2]Agora Cancer Research Centre, 1011 Lausanne, Switzerland
[3]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[4]Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University Hospital of Lausanne, Lausanne, Switzerland
[5]Swiss Cancer Center Leman (SCCL), Lausanne, Switzerland
[6]Lead contact
*Correspondence: david.gfeller@unil.ch
https://doi.org/10.1016/j.cels.2022.12.002

SUMMARY

The recognition of pathogen or cancer-specific epitopes by CD8$^+$ T cells is crucial for the clearance of infections and the response to cancer immunotherapy. This process requires epitopes to be presented on class I human leukocyte antigen (HLA-I) molecules and recognized by the T-cell receptor (TCR). Machine learning models capturing these two aspects of immune recognition are key to improve epitope predictions. Here, we assembled a high-quality dataset of naturally presented HLA-I ligands and experimentally verified neo-epitopes. We then integrated these data in a refined computational framework to predict antigen presentation (MixMHCpred2.2) and TCR recognition (PRIME2.0). The depth of our training data and the algorithmic developments resulted in improved predictions of HLA-I ligands and neo-epitopes. Prospectively applying our tools to SARS-CoV-2 proteins revealed several epitopes. TCR sequencing identified a monoclonal response in effector/memory CD8$^+$ T cells against one of these epitopes and cross-reactivity with the homologous peptides from other coronaviruses.

INTRODUCTION

CD8$^+$ T cells have the ability to eliminate infected or malignant cells and play a key role in infectious diseases and cancer immunotherapy. CD8$^+$ T-cell recognition is initiated by the binding of the T-cell receptor (TCR) to peptides displayed on class I human leukocyte antigen (HLA-I) molecules. Detailed knowledge of class I epitopes in cancer and infectious diseases has several translational and clinical applications. Such epitopes can be used to design vaccines that target the most relevant epitopes, including neo-epitopes (i.e., peptides containing non-synonymous genetic alterations) in cancer.[1–3] Class I epitopes can also be used to select TCRs, study them and reinfuse these TCRs into patients as part of adoptive T cell therapy.[4] Unfortunately, identifying epitopes in cancer or infectious diseases is challenging because of the very high number of possible candidates and the diversity of HLA-I alleles. For instance, for each non-synonymous point mutation in a tumor, up to 38 from 8- to 11-mer peptides containing the mutated residue may be immunogenic. Similarly, the number of potential class I epitopes of a given length in a pathogen is roughly equal to the length of the proteome of this pathogen. Major improvements have been done for experimentally screening potential epitope candidates, either with peptide pools[5] or tandem mini-genes.[6,7] Nevertheless, the most common approach to identify new epitopes is to preselect them based on HLA-I ligand predictors.

HLA-I molecules are encoded by three genes (HLA-A, -B and -C). These genes are highly polymorphic in human and different alleles are characterized by specific binding motifs and specific length distributions in their ligands.[8] Binding motifs mainly reflect amino acids favorable for binding to HLA-I molecules at specific positions of the ligands. Peptide length distributions (typically from 8- to 14-mers with a preference for 9-mers for most alleles) capture both the binding preferences of HLA-I molecules as well as the skewed length distribution of peptides available in the endoplasmic reticulum for loading onto HLA-I molecules.[9]

The specificity of HLA-I-binding motifs and peptide length distributions greatly constrains the repertoire of potential epitopes. As such, computational tools that accurately capture these two features of antigen presentation have been developed to narrow down the list of potential epitopes to be experimentally tested. Historically, predictors HLA-I ligands were mainly trained on peptides tested experimentally in binding assays,[10] with the

caveat that many of these peptides had been pre-selected based on previous versions of the predictors. More recently, naturally presented HLA-I ligands identified by mass spectrometry (MS) based HLA-I peptidomics provided a rich source of information about the rules of antigen presentation and the specificity of HLA-I molecules.[11–17] The number of HLA-I ligands identified by this technology, both in mono-allelic and poly-allelic samples, surpasses the one from binding assays and HLA-I peptidomics data are now included in the training of most HLA-I ligand predictors.[17–22] For poly-allelic samples, motif deconvolution has been used to identify HLA-I binding motifs and determine allelic restriction of HLA-I ligands without relying on HLA-I ligand predictors.[12,13,18,23]

Over the years, several attempts have been made to integrate additional features in epitope predictions linked to antigen presentation and TCR recognition. For instance, gene expression and protein abundance were shown to improve HLA-I ligand and class I epitope predictions.[11,17,24] Predictions of cleavage or antigen transport properties were integrated in epitope prediction tools.[25] The concept of antigen presentation hotspot, as determined by the analysis of HLA-I peptidomics data, was also shown to improve predictions.[26] Some studies further attempted to integrate TCR recognition propensities in epitope predictions, for instance by investigating the role of dissimilarity-to-self or foreignness.[27–30] We and others observed that specific amino acids found in epitope residues more likely to interact with the TCR increase the propensity for TCR binding.[31–33]

In this work, we compiled and curated a large dataset of HLA-I ligands and neo-epitopes. We then integrated these data with new algorithmic developments to improve predictions of antigen presentation (MixMHCpred2.2) and TCR recognition propensity (PRIME2.0). Applying these tools to SARS-CoV-2 proteins enabled us to predict and validate several epitopes, which we characterized in terms of TCR functional avidity, clonality, and cross-reactivity.

RESULTS

Integration and curation of HLA-I peptidomics data reveal binding motifs and peptide length distributions for more than hundred alleles

To improve predictions of class I antigen presentation, we manually compiled recent studies of naturally presented HLA-I ligands profiled by MS. Our dataset covers 24 studies for a total of 244 samples (see Table S1). All data were retrieved from the original publications and were not filtered by any HLA-I ligand predictor, ensuring that our dataset is not biased by such filtering. All samples were processed with the motif deconvolution tool MixMHCp and shared motifs across samples containing the same HLA-I allele were annotated to this allele, following our previously established approach (see example in Figure 1A).[13,14] All motifs in each sample were manually verified, and samples or alleles for which motif deconvolution results were ambiguous were not considered (e.g., fourth motif of the first sample in Figure 1A). This enabled us to derive reliable binding motifs and peptide length distributions for 119 HLA-I alleles, supported by a total of 384,070 peptides (Table S2, see examples in Figure 1B). Motifs for HLA-I alleles identified in mono-allelic and poly-allelic samples were highly similar (Figure S1A). Peptide length distributions for alleles in mono-allelic samples displayed a slightly lower fraction of 9-mers and a slightly higher fraction of peptides of other lengths compared with those observed in poly-allelic samples (Figure 1C).

In addition to determining binding motifs and peptide length distributions for the different alleles expressed in a sample, motif deconvolution is useful to identify potential sources of noise in the data.[18,23,34] Noise in HLA-I peptidomics data can consist of peptides from the same sample (e.g., contaminants pulled down together with HLA-I ligands, but not binding to HLA-I molecules), peptides from other samples (e.g., contaminants due to suboptimal cleaning of MS equipment), or wrongly identified peptides occurring during the computational annotation of mass spectra.

To predict contaminants or wrongly identified peptides, we first collected all peptides assigned to the flat motif by MixMHCp[18] in each sample (see example in Figure 1A, right box, and Figure S1B). These are referred to as "unspecific contaminants." We further manually retrieved cases where some motifs predicted by MixMHCp did not match the motifs expected for the alleles given in the HLA-I typing but displayed very high similarity with the motifs of other alleles. These cases are referred to as "allele-specific contaminants."

As previously reported,[18] the fraction of predicted unspecific contaminants is especially large in 8-mers and 11- to 14-mers both in mono-allelic and poly-allelic samples, representing for instance more than 50% of 14-mers (Figure 1D, see specific examples in Figure S1B). Predicted unspecific contaminants show a relatively flat motif with small preference for arginine and lysine at the last residue (Figure 1E), which is consistent with the hypothesis that some of these peptides correspond to peptides that underwent trypsin-based digestion used in standard proteomics. The predicted unspecific contaminants had a peptide length distribution markedly different from the one of HLA-I ligands (Figure 1F), further indicating that most of these peptides are not bona fide HLA-I ligands. Allele-specific contaminants were observed in some mono-allelic samples (Figure S1C) as well as one poly-allelic sample with erroneous HLA-I typing (Figure S1D).

These observations demonstrate the importance of performing careful quality-control before using mono- or poly-allelic HLA-I peptidomics data to train HLA-I ligand predictors.[35]

Models of HLA-I binding specificities and peptide length distributions improve predictions of naturally presented HLA-I ligands

To improve predictions of class I antigen presentation, we integrated these data into the training of our HLA-I ligand predictor MixMHCpred and further refined the modeling of peptide length distributions (see STAR Methods). As with most HLA-I ligand predictors, the final score of a peptide is expressed as a %rank, which represents how the predicted binding of a peptide compares with the one of random peptides from the human proteome (see STAR Methods). To benchmark the new version of MixMHCpred (v2.2) we used two external datasets. The first one consists of ten HLA-I peptidomics datasets from meningioma samples.[18] The second one consists of eleven recently published HLA-I peptidomics datasets.[20] These datasets were not included in the training of any predictor considered in this work.

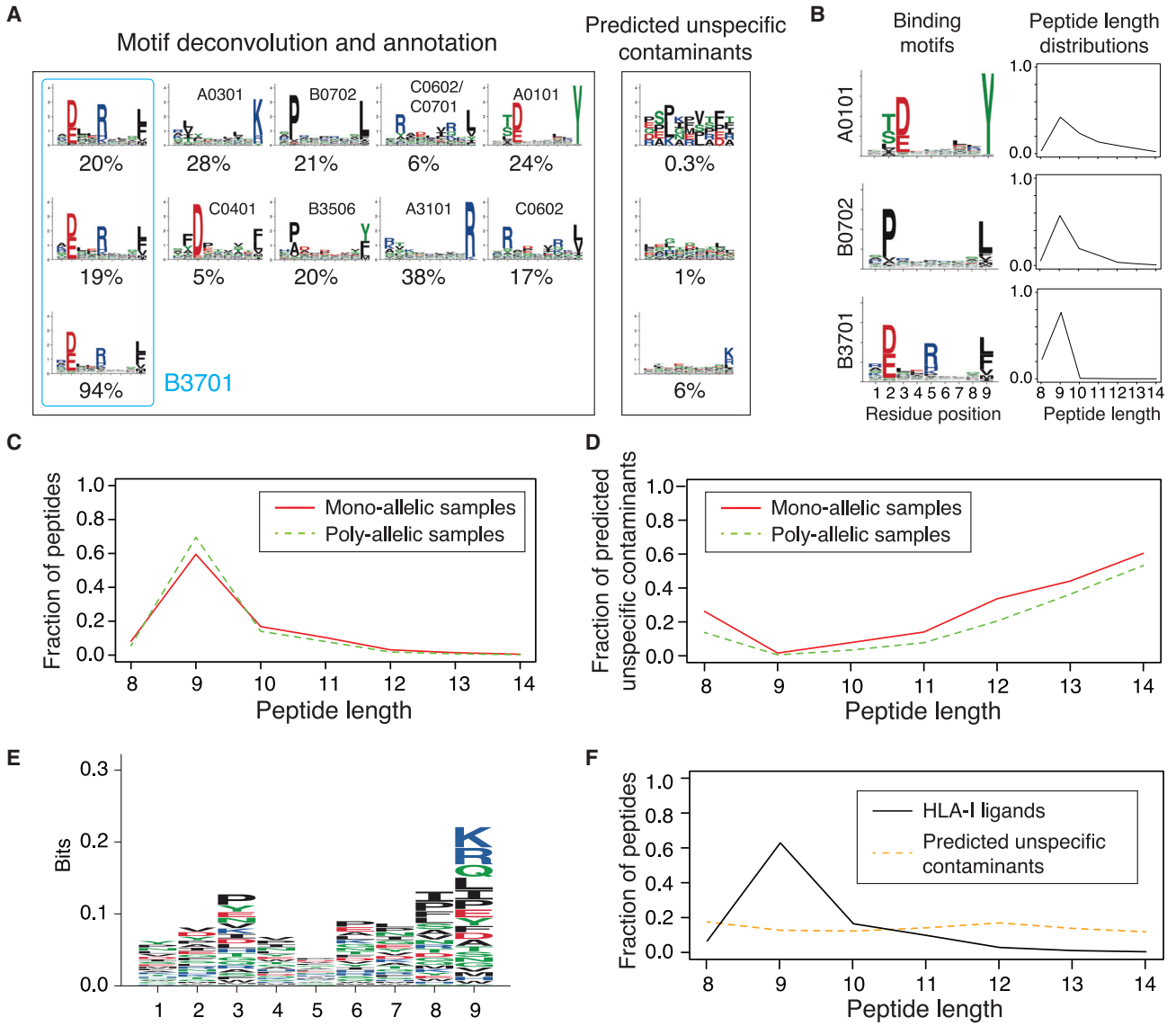


Figure 1. Motif deconvolution across HLA-I peptidomics datasets reveal binding motifs and peptide length distributions of HLA-I molecules, as well as predicted contaminants in mono- and poly-allelic datasets.

(A) Example of motif deconvolution with MixMHCp in three HLA-I peptidomics samples. This includes determination of HLA-I motifs and predicted unspecific contaminants (i.e., peptides assigned to the flat motif of MixMHCp), as well as motif annotation by identifying shared motifs across samples sharing the same allele. The example shows the deconvolved motifs in two poly-allelic samples that share the HLA-B*37:01 allele ("donor1" and "HCC1143" in Table S1), as well as the mono-allelic HLA-B*37:01 sample.

(B) Examples of binding motifs and peptide length distributions obtained by motif deconvolution.

(C) Peptide length distributions in mono-allelic and poly-allelic HLA-I peptidomics data. Each curve represents the average peptide length distribution across alleles with both mono- and poly-allelic HLA-I peptidomics data.

(D) Fraction of predicted unspecific contaminants across different lengths (average over all samples).

(E) Motif of the predicted unspecific contaminants (9-mers) identified by MixMHCp across all samples.

(F) Comparison of the length distribution of peptides assigned to HLA-I alleles and predicted unspecific contaminants.

In total they contain 78,011 HLA-I ligands (Table S3). 4-fold excess of randomly selected peptides from the human proteome were used as negatives to compute receiver operating curves (ROCs) and positive predictive values (PPVs) (see STAR Methods and Table S3). Both the area under the ROC curve (AUC) and the PPV were higher for MixMHCpred2.2, compared with NetMHCpan4.1,[21] MHCflurry2.0,[19] HLAthena[17] and MixMHCpre2.0.2[18] (Figures 2A, 2B, and S2A–S2D).

Different performance in predicting naturally presented HLA-I ligands could originate from differences in modeling either the amino acid specificity or the peptide length specificity of naturally presented HLA-I ligands. To explore these two aspects of

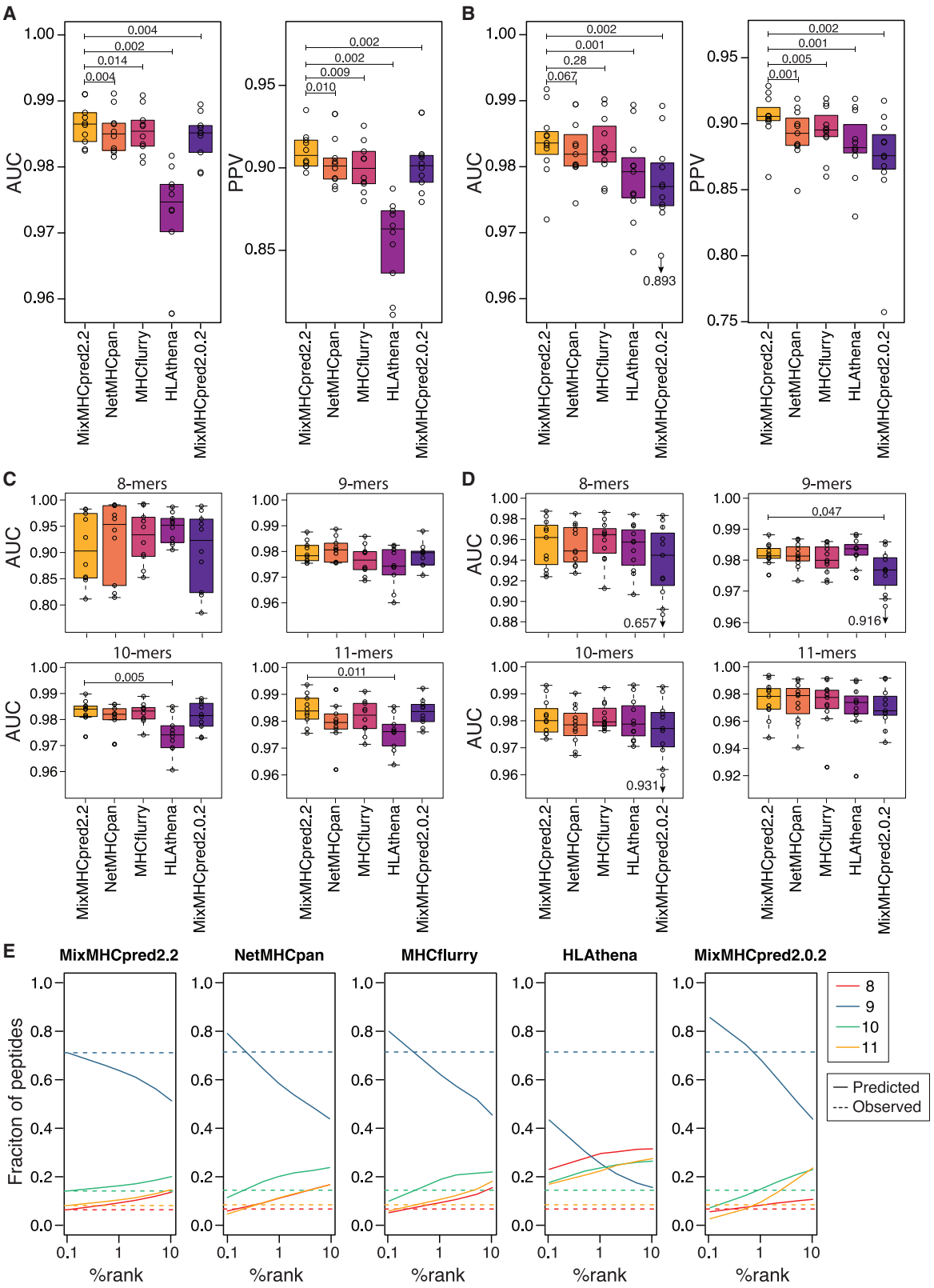


Figure 2. Models of HLA-I binding specificities and peptide length distributions improve predictions of naturally presented HLA-I ligands

(A) Boxplot of AUC and PPV values for the different predictors considered in this study applied on the 10 HLA-I peptidomics samples from Gfeller et al.[18]

(B) AUC and PPV values obtained for HLA-I peptidomics samples from Pyke et al.[20]

(legend continued on next page)

HLA-I ligand predictors, we first benchmarked the different predictors on each peptide length separately. We observed only few cases with statistically significant differences in AUC values (Figures 2C and 2D). This suggests that all methods considered in this work capture relatively well the binding motifs of HLA-I alleles and that the differences observed in Figures 2A and 2B may come from the modeling of peptide length distributions.

To further investigate this hypothesis, we computed the predicted peptide length distributions at different %rank thresholds (see STAR Methods). We then compared these predicted peptide length distributions with those observed in HLA-I peptidomics samples (Figure 2E). Overall, we observed that MixMHCpred2.2 predictions had the best agreement with the experimental peptide length distributions across different thresholds. Both NetMHCpan4.1 and MHCflurry2.0 displayed good agreement, although with some more pronounced under-representation of 9-mers at %ranks larger than 1. MixMHCpred2.0.2 displayed less stable distributions across % rank thresholds, including an over-representation of longer peptides for high %rank (i.e., %rank between 2% and 10%) and an under-representation of such peptides for small %rank (i.e., % rank < 1). HLAthena displayed a very clear under-representation of 9-mers, and over-representation of 8-, 10-, and 11-mers across all %rank thresholds. The discrepancy was also observed when considering peptide length distributions from mono-allelic HLA-I peptidomics data (Figure S2E). These observations suggest that integrating peptide lengths, either by stable renormalization of the raw scores (MixMHCpred2.2), as separate input nodes in neural networks (NetMHCpan), or by using padding (MHCflurry), is important to accurately capture the length distribution of naturally presented HLA-I ligands across different alleles. The length distribution of naturally presented HLA-I ligands is a result of both preferences of HLA-I alleles and a skewed length distribution among peptides available for loading on HLA-I in the endoplasmic reticulum.[9] In particular, peptide length distributions computed based on naturally presented HLA-I ligands show higher values for 9-mers and lower values for other peptide lengths compared with peptide length distributions computed based on the results of binding assays.[9] This may explain the lower predicted distribution of 9-mers for NetMHCpan and MHCflurry at %rank larger than 1%, since these two methods include results of binding assays in their training set.

Models of TCR recognition propensity improve predictions of neo-epitopes

To expand upon previous attempts to capture biochemical properties of epitopes that increase TCR recognition propensities,[31–33] we collected data from 70 recent neo-antigen studies. This resulted in 596 verified immunogenic neo-epitopes, as well as 6,084 non-immunogenic peptides tested experimentally (see STAR Methods and Table S4). Most of the immunogenic and non-immunogenic peptides were previously selected based on HLA-I ligand predictors and, as a result, show much higher predicted binding to HLA-I compared with random peptides (Figure 3A). To correct for this bias in our data, we further included for each neo-epitope 99 peptides randomly selected from the same source protein as additional negatives (see STAR Methods and Table S4). We then used these data to train a PRedictor of Immunogenic Epitope (PRIME2.0). PRIME2.0 is based on a neural network and uses as input features (1) the predicted HLA-I presentation score (−log(%rank) of MixMHCpred2.2), (2) the amino acid frequency at positions with minimal impact on binding to HLA-I and more likely to face the TCR,[33] and (3) the length of the peptide (Figure 3B; see STAR Methods). Compared with our previous work (PRIME1.0[33]), the training set of PRIME2.0 is more realistic in terms of predicted HLA-I binding of the negatives (i.e., broad coverage of the range of %rank values without enrichment in predicted ligands). Moreover, the use of neural networks can capture potential correlations between different input features (see below). We performed multiple cross-validations based on randomly splitting the data (standard 10-fold cross-validation), iteratively excluding specific alleles (leave-one-allele-out cross-validation), or iteratively excluding data from specific studies (leave-one-study-out cross-validation) (see STAR Methods). Overall, we observed improved predictions with PRIME2.0 (Figures 3C and S3A), even if most of the neo-epitopes considered in this work had been predicted by NetMHCpan, and several of them are part of the training of NetMHCpan, MHCflurry, and PRIME1.0. We also restricted our benchmark to peptides experimentally tested (i.e., excluding random negatives from the test sets) (Figures 3D and S3B). These peptides typically show relatively good predicted binding to HLA-I (Figure 3A) since most of them were pre-selected based on HLA-I ligand predictors. On this test set, PRIME2.0 displayed in general better performance than HLA-I ligand predictors. In this case, PRIME1.0 had roughly similar performance as PRIME2.0, consistent with the fact that PRIME1.0 was mainly trained on peptides (both immunogenic and non-immunogenic) with high predicted affinity to HLA-I molecules.

To investigate the robustness of our results with respect to the use of other predictors of HLA-I ligands in PRIME, we retrained PRIME2.0 using %ranks from NetMHCpan, MHCflurry, or HLAthena. Overall, we observed similar results (Figures S3C and S3D), which demonstrates that the framework of PRIME2.0 is robust and compatible with different HLA-I ligand predictors.

To explore the impact of the use of neural networks in PRIME2.0, we retrained PRIME2.0 using a logistic regression (Figures 3C, 3D, S3A, and S3B, cyan bars). When considering peptides spanning the whole range of possible binding affinities to HLA-I, PRIME2.0 based on a neural network outperformed

(C) AUC values computed separately for each peptide length for the ten samples from Gfeller et al.[18] Only p values between MixMHCpred2.2 and other tools and smaller than 0.05 are indicated.

(D) AUC values computed separately for each peptide length for eleven samples from Pyke et al.[20] Only p values between MixMHCpred2.2 and other tools and smaller than 0.05 are indicated.

(E) Predicted peptide length distributions at different %rank thresholds for each HLA-I ligand predictor (average over alleles available in all predictors). Dashed lines show the peptide length distributions observed in naturally presented HLA-I ligands (average over all alleles). Boxplots in (A)–(D) represent the median and lower/upper quartiles. p values were computed with paired Wilcoxon test.

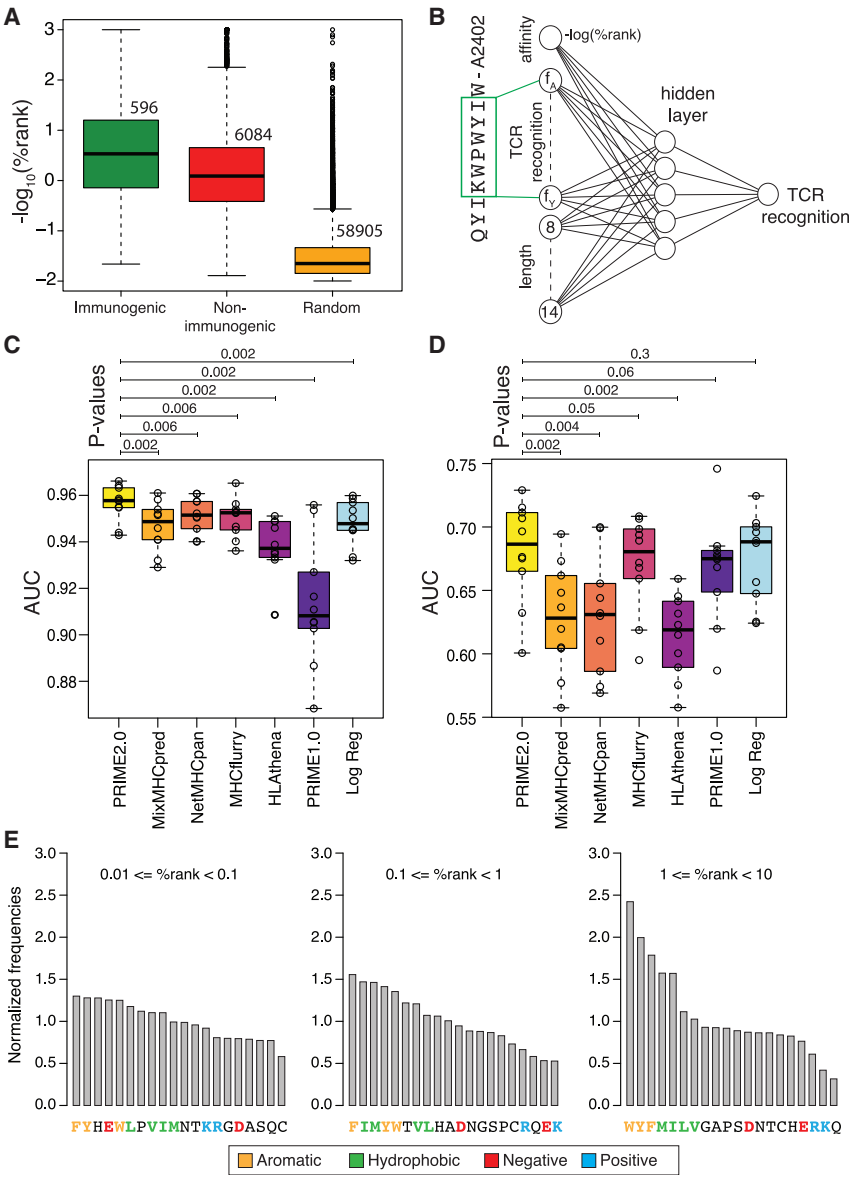


Figure 3. Models of TCR recognition propensity improve predictions of neo-epitopes

(A) Predicted binding affinity to HLA-I (based on % rank of MixMHCpred2.2.) of experimentally validated immunogenic (green) and non-immunogenic (red) peptides, as well as random peptides (orange) used to train PRIME.

(B) Architecture of neural network of PRIME2.0. The first input node corresponds to the predicted binding to the HLA-I allele (−log(%rank) from MixMHCpred2.2). The next 20 nodes correspond to amino acid frequencies on residues with minimal impact on predicted affinity to the HLA-I allele (green box). These positions were determined as previously described.[33] The last seven nodes correspond to the length of the peptide (i.e., 8–14, one-hot encoding).

(C) Benchmarking of PRIME2.0 based on 10-fold cross-validation. "Log Reg" indicates the model trained on the same data as PRIME2.0 but with a logistic regression instead of a neural network.

(D) Same cross-validation as in (C) after excluding randomly generated negatives in the test set.

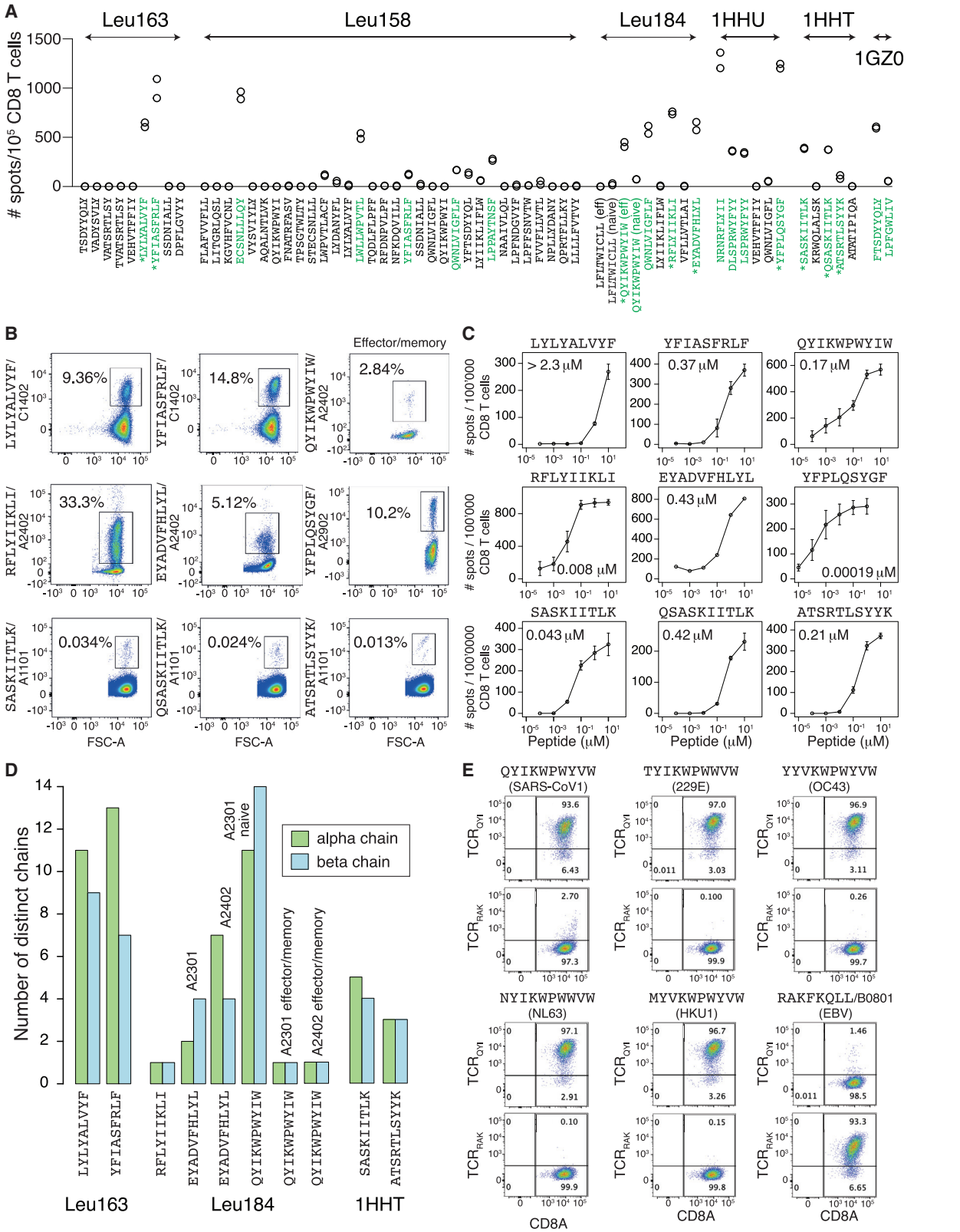
(E) Normalized amino acid frequencies at positions with minimal impact on predicted affinity to HLA-I for immunogenic versus non-immunogenic peptides used to train PRIME2.0 within different ranges of predicted HLA-I binding (%rank of MixMHCpred). Boxplots in (A), (C), and (D) represent the median and lower/upper quartiles. p values were computed with paired Wilcoxon test.

PRIME2.0 based on a logistic regression, demonstrating that neural networks are useful to handle peptides with different predicted binding affinities to HLA-I (Figure 3C). Reversely, when considering only experimentally validated peptides in the test set (i.e., peptides with similar predicted binding to HLA-I), we did not observe significant differences (Figure 3D).

To further interpret the impact of amino acids found in epitope positions more likely to impact TCR recognition (green box in Figure 3B), we compared the frequency of these amino acids between the positives (i.e., epitopes) and the negatives (i.e., non-immunogenic peptides) used in the training of PRIME2.0 for different ranges of MixMHCpred %ranks (Figure 3E). Our results reveal an enrichment in aromatic and hydrophobic residues among epitopes and a depletion of charged or polar residues, which is consistent with previous studies.[31,33] The enrichment is especially pronounced for epitopes with predicted low binding to HLA-I (%rank between 1% and 10%). These results are robust to the use of other HLA-I ligand predictors (Figure S3E). These observations support the following model of TCR recognition propensity. For high affinity HLA-I ligands the presence of specific amino acids in the epitope residues which are more likely to directly interact with the TCR is less important because the high stability of the peptide-HLA-I complex increases the probability of stable TCR binding. Conversely, for low affinity HLA-I ligands, the presence of specific amino acids favoring TCR recognition becomes more important and helps counterbalancing the lower stability/affinity of the peptide-HLA-I complexes. This correlation pattern between different features of PRIME2.0 (i.e., predicted affinity and amino acid frequency at positions more likely to interact with the TCR) provides a plausible explanation for the improvement obtained when training PRIME2.0 with a neural network instead of a logistic regression (Figure 3C).

Immunogenicity predictions reveal SARS-CoV-2 CD8$^+$ T-cell epitopes

To explore whether PRIME2.0 could be used in a prospective way to identify immunologically relevant epitopes, we applied it to the proteome of SARS-CoV-2 and selected a list of 213 peptides with PRIME2.0 %rank lower or equal to 0.5 with at least one of the 15 most common HLA-I alleles (see STAR Methods and Tables S5A and S5B). We then *in vitro* primed



(legend on next page)

CD8$^+$ T cells from 6 donors (Table S5C) with pools of the predicted peptides and deconvolved the IFNγ ELISpot responses to the level of single epitopes (see STAR Methods and Figure 4A). Three donors had been tested positive for SARS-CoV-2 (i.e., 1GZ0, 1HHU, and 1HHT), and no donor had been vaccinated (samples collected early 2020). In total, we could identify 18 immunogenic peptides, with 2 of them (YFIASFRLF and QWNLVIGFLF) identified in two different donors (Table 1). Eight of these epitopes had not been observed in previous studies including the two identified in multiple donors (Table 1). Three additional epitopes had been reported with other allelic restrictions (LYLYALVYF, FTSDYYQLY, and YFPLQSYGF). To validate these observations, we used peptide-HLA multimers to stain CD8$^+$ T cells recognizing nine of these epitopes in four donors for which enough cells were available (Leu163, Leu184, 1HHT, and 1HHU). All epitopes could be confirmed (Figure 4B). We then measured the functional avidity (EC_{50}) of the CD8$^+$ T cells recognizing these epitopes. The functional avidity displayed some variability, ranging from low micro-molar to sub-nano-molar values (Figure 4C). The highest avidity was observed for the HLA-A*29:02 restricted YFPLQSYGF epitope in a SARS-CoV-2 positive donor (1HHU). This epitope had been previously observed in patients, with a restriction to HLA-A*24:02.[36] HLA-A*29:02 and HLA-A*24:02 are part of the same HLA-I supertype (A24) and display some overlap in their binding motifs, including preference for F at both P2 and PΩ anchor residues. This suggests that the YFPLQSYGF epitope may be immunogenic in several patients with HLA-I alleles of the A24 supertype.

We next investigated how well these peptides would have been predicted by different tools. To this end, we first computed the best score for the alleles of each donor for the different predictors. The immunogenic peptides had highest scores with PRIME2.0, as expected since they had been selected with this tool (Figure S4A). We then computed AUC values in each donor (see STAR Methods). All predictors had AUC values larger than 0.5 in all donors (Figure S4B). There was substantial variability across samples, and no method had statistically significantly better AUC values. However, the limited number of samples and of immunogenic peptides, as well as the biased selection of the initial list of 213 peptides with only one algorithm make it difficult to draw robust conclusions about the performance of the different algorithms.

To gain insights in the clonality of the CD8$^+$ T cell populations recognizing these epitopes, we sorted CD8$^+$ T cells recognizing seven of these epitopes and sequenced separately the alpha and beta chains of their TCRs (Table S6). Different epitopes were recognized by different numbers of TCRs (Figure 4D). For epitopes recognized by several TCRs, one or two alpha and beta chains had significantly higher frequency, suggesting that the recognition may be primarily driven by the pairing of such chains (Figure S4C). For the QYIKWPWYIW epitope from the spike protein (donor Leu184), naive and effector/memory CD8$^+$ T cells recognizing this epitope were sorted separately (Figure S4D). We observed a high diversity of TCR chains among naive CD8$^+$ T cells. Reversely, a unique clone (TCR_{QYI}: TRAV20*01-CAALNYGGATNKLIF-TRAJ32*01 and TRBV4-3*01-CASSPSGGAYEQYF-TRBJ2-7*01) was found in the effector/memory CD8$^+$ T cells. This unique TCR was identified in effector/memory CD8$^+$ T cells recognizing QYIKWPWYIW displayed both on HLA-A*23:01 and HLA-A*24:02 (Figure S4E; Table S6), as expected since HLA-A*23:01 and HLA-A*24:02 have very high sequence similarity and almost identical binding motifs (Figure S1A). We next asked if the same TCR could be found in other individuals. The same beta chain was found in 19 SARS-CoV-2$^+$ donors in the ImmuneCODE database, which is a large repertoire of TCRβ chains from SARS-CoV-2$^+$ donors.[37] The same alpha and beta chains were also found in the TCR repertoire of one of the two SARS-CoV-2$^+$ donors analyzed in a recent study.[38] Moreover, the same alpha chain and a highly similar beta chain (same CDR3β sequence) were found in the other donor of the Minervina et al. study (Figure S4F). Both donors were HLA-A*24:02$^+$. These observations suggest that the recognition of the QYIKWPWYIW epitope may be mediated by the same TCR in multiple donors.

The donor where recognition of the QYIKWPWYIW epitope was observed (Leu184) had not been tested positive for SARS-CoV-2 and had not received any SARS-CoV-2 vaccine. Therefore, we hypothesized that the monoclonal population of effector/memory CD8$^+$ T cells recognizing this epitope could originate from previous exposure to other coronaviruses. This hypothesis is supported by the fact that the QYIKWPWYIW epitope is quite well conserved in the spike protein of other coronaviruses (Table 2). In particular, non-conserved residues are either found at P1, which has little impact on HLA-A*24:02 binding and TCR recognition, or involve amino acids with similar biophysical properties (I → V, Y → W). To further verify our hypothesis, we stained TCR_{QYI} transfected cells with multimers consisting of these homologous peptides in complex with HLA-A*24:02. Our results demonstrate that TCR_{QYI} is able to recognize all of these peptides (Figure 4E). These results are consistent with the observation that previous exposure to other coronaviruses can confer some immunity to SARS-CoV-2.[39,40]

Figure 4. PRIME2.0 identifies SARS-CoV-2 CD8$^+$ T-cell epitopes

(A) IFNγ ELISpot results for the peptides tested individually (i.e., after deconvolution of the pools). Immunogenic peptides are shown in green. Stars indicate peptides for which enough CD8$^+$ T cells were available for peptide-HLA multimer validation and functional avidity assays. Donors are indicated above each peptide group. For donor Leu184, two epitopes (LFLTWICLL and QYIKWPWYIW) were tested with both effector/memory and naive CD8$^+$ T cells.

(B) Staining of CD8$^+$ T cells with peptide-HLA multimers for nine epitopes from donors for which enough CD8$^+$ T cells could be obtained (i.e., donors Leu163, Leu184, 1HHT, and 1HHU, see Table 1). For QYIKWPWYIW, effector/memory CD8$^+$ T cells were used.

(C) Functional avidity (effective concentration 50%, EC_{50}). Error bars represent the standard deviation of two replicate, except for EYADVFHLYL where only one replicate could be performed due to limited amount of CD8$^+$ T cells.

(D) Number of distinct alpha and beta chains identified in TCRs recognizing the seven epitopes for which TCR sequencing could be performed.

(E) Multimers consisting of different homologs of QYIKWPWYIW found in the spike protein of other coronaviruses (see Table 2) in complex with HLA-A*24:02 were used for staining of TCR$^-$ Jurkat cells transfected with TCR_{QYI}. The control TCR_{RAK} represents TCR$^-$ Jurkat cells transfected with a TCR recognizing the Epstein-Barr virus epitope RAKFKQLL in complex with HLA-B*08:01.

Table 1. List of immunogenic SARS-CoV-2 epitopes

Donor	HLA-I typing	Epitope sequence	Source protein	Known epitopes with their reported allelic restriction	Cells available for multimer validation	Alleles used in the multimer validation	TCR-seq
Leu163	A0102,A0201, B4901,B5101, C0701,C1402	LYLYALVYF	AP3A	A2402	yes	C1402	yes
		YFIASFRLF	VME1	–	yes	C1402	yes
Leu158	A1101,A2402, B1801,B3501, C0401,C1203	ECSNLLLQY	SPIKE	–	no	–	no
		LWLLWPVTL	VME1	A2402	no	–	no
		YFIASFRLF	VME1	–	no	–	no
		QWNLVIGFLF	VME1	–	no	–	no
		LPPAYTNSF	SPIKE	B0702, B3501, B5301	no	–	no
Leu184	A2301,A2402, B3502,B4901, C0401,C0701	QYIKWPWYIW	SPIKE	A2301	yes	A2402/A2301	yes
		QWNLVIGFLF	VME1	–	no	–	no
		RFLYIIKLI	VME1	–	yes	A2402	yes
		EYADVFHLYL	R1AB	–	yes	A2402/A2301	yes
1GZ0	A0102,A0201, B0801,B5101, C0701,C1502	FTSDYYQLY	AP3A	A0101, A2402, A2902	no	–	no
		LPFGWLIV	AP3A	B5101	no	–	no
1HHU	A0103,A2902, B4403,B7301, C1505,C1601	NRNRFLYII	VME1	–	no	–	no
		DLSPRWYFYY	NCAP	A0201, A2902	no	–	no
		LSPRWYFYY	NCAP	–	no	–	no
		YFPLQSYGF	SPIKE	A2402	yes	A2902	no
1HHT	A1101,A3201, B4002,B4402, C0202,C0501	SASKIITLK	AP3A	A0301, A1101, B5701	yes	A1101	yes
		QSASKIITLK	AP3A	–	yes	A1101	no
		ATSRTLSYYK	VME1	A1101, A3001	yes	A1101	yes

In a recent study, the 9-mer peptide (QYIKWPWYI) fully overlapping with the 10-mer epitope recognized by TCR_{QYI} was shown to elicit an immuno-dominant $CD8^+$ T-cell response, and the QYIKWPWYI – HLA-A*24:02 complex was crystallized.[41] This structure shows that the three non-anchor aromatic sidechains shared with the 10-mer investigated in our work (i.e., W5, W7, and Y8) are all facing outside of the HLA-I-binding site and therefore are likely to interact with the TCR (Figure S4G). The presence and orientation of the aromatic sidechains in this immuno-dominant epitope of the spike protein are consistent with the model of improved TCR recognition propensity of aromatic residues, which underlies the PRIME algorithm.

DISCUSSION

$CD8^+$ T-cell epitopes play central roles in immune responses against infectious diseases as well as cancer and represent promising targets for personalized cancer immunotherapy treatments. In this work, we trained both a predictor of antigen presentation (MixMHCpred2.2) and a predictor of immunogenicity (PRIME2.0). By expanding the training set and optimizing the algorithms, we could demonstrate improved predictions for both HLA-I ligands and class I neo-epitopes.

A key aspect of any machine learning predictor is the quality and depth of the training data. Consistent with previous studies,[12,18,35] our results reveal that different types of putative contaminants can be found in both poly- and mono-allelic HLA-I peptidomics data. Contaminants include peptides with trypsin-like motifs or peptides coming from other HLA-I alleles, including in samples that were assumed to be mono-allelic. These results emphasize the importance of carefully applying quality controls before using such data for training HLA-I ligand predictors.[35] Another important aspect is the choice of the algorithm. Our benchmark of HLA-I ligand predictors suggest that accurate modeling of peptide length distribution is important for such predictions. In particular, we observed important difference in the benchmarking of HLA-I ligand predictors when combining peptides of all lengths (Figures 2A and 2B) and when treating separately peptides of different lengths (Figures 2C and 2D). By construction, the benchmarking of each peptide length separately cannot inform us on whether a predictor accurately models peptide length distributions. Considering that this is an important aspect of naturally presented HLA-I ligands and class I epitopes, we advocate for systematically combining HLA-I ligands of different lengths and using as negatives random peptides with uniform length distribution when training and benchmarking HLA-I ligand predictors.

The analysis of the data used to train our predictor of immunogenicity (PRIME2.0) confirmed the importance of aromatic residues, especially tryptophan. In line with previous studies and crystal structures of TCR-peptide-MHC complexes,[31,33,41,42] we suggest that this preference reflects the ability of tryptophan (or other aromatic residues) to engage into stable molecular interactions with the TCR. However, we cannot exclude that other factors play a role in the importance given to tryptophan in PRIME. First, tryptophan tends to be slightly depleted in MS-based HLA-I peptidomics studies.[11,14] This may bias HLA-I ligand predictors trained on such data, and PRIME may be correcting for this bias. Second, recent studies have demonstrated

Table 2. Sequences of the Spike peptides homologous to QYIKWPWYIW in other coronaviruses

Organism	Sequence	Differences with SARS-CoV-2
SARS-CoV-2	QYIKWPWYIW	–
SARS-CoV-1	QYIKWPWYVW	I9V
229E	TYIKWPWWVW	Q1T, Y8W, I9V
OC43	YYVKWPWYVW	Q1Y, I3V, I9V
NL63	NYIKWPWWVW	Q1N,Y8W, I9V
HKU1	MYVKWPWYVW	Q1M, I3V, I9V

Differences with the SARS-CoV-2 epitope are underlined.

that peptides genomically encoded with a W can undergo a W>F substitution during protein synthesis.[43] Considering that several studies used tandem mini-genes to identify neo-epitopes, we cannot exclude that some of the W-containing epitopes used in the training of PRIME were actually presented on HLA-I molecules with a W>F substitution, which contributed to overcome central tolerance and increase their immunogenicity. This supports a model where tryptophan containing protein segments are especially promising neo-epitope candidates, both in terms of improving TCR recognition and overcoming central tolerance.

In our benchmark of PRIME2.0, we observed improved predictions when training PRIME2.0 with neural networks compared with logistic regressions and when considering negatives spanning the whole range of predicted affinity to HLA-I (Figure 3C). The analysis of amino acid frequencies in Figure 3E suggests that the importance of specific residues at position interacting with the TCR changes depending on the affinity of the epitopes to the HLA-I molecules. This may explain why models that can capture such correlations (e.g., neural networks) outperform linear models (e.g., logistic regressions).

Applying our tool to the SARS-CoV-2 proteome, we could uncover several epitopes, including one (QYIKWPWYIW) recognized by a monoclonal population of antigen-experienced $CD8^+$ T cells with an effector/memory phenotype. This epitope has very high homology with other coronaviruses and is 100% conserved in all common variants of SARS-CoV-2. This suggests that $CD8^+$ T-cell responses elicited against this epitope by previous infection, vaccination, or cross-reactivity with other coronaviruses may be effective across all SARS-CoV-2 variants. Due to limitations in the available samples and the number of SARS-CoV-2 peptides that could be tested, we restricted our experimental validation to predictions generated with PRIME2.0. As such, this part of the work shows that PRIME2.0 can be prospectively applied for epitope discovery but should not be used to draw conclusions about the performance of other predictors.

Overall, our work provides improved predictions for both antigen presentation (MixMHCpred2.2) and TCR recognition (PRIME2.0) of class I epitopes. In terms of HLA-I ligand predictions, a decent accuracy had already been reached by many existing tools.[18,19,21] Much harder is the task of predicting immunogenicity, both because of the smaller size of the training data and because of the multiple other factors that influence T-cell recognition (e.g., co-receptors, cytokines, etc.). Efforts focusing on generating high-quality immunogenicity training data and developing machine learning frameworks to harness these data will be key to further improve class I epitope predictions.

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
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data collection and curation
  - Training of MixMHCpred
  - Comparing predicted and experimental peptide length distributions
  - Training of PRIME
  - Predictions of SARS-CoV-2 epitopes
  - Identification of SARS-CoV-2 epitopes
  - Predictability of SARS-CoV-2 epitopes
  - Peptide-HLA multimer validation of SARS-CoV-2 epitopes and sorting of $CD8^+$ T cells
  - Functional avidity assay
  - Bulk TCR sequencing
  - TCR sequence analyses
  - TCR_{QYI} transfection in Jurkat cells and recognition of the homologous peptides from other coronaviruses
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.12.002.

ACKNOWLEDGMENTS

D.G. acknowledges support from the Swiss Cancer Research Foundation (KFS-4961-02-2020). G.C. is supported by the Marie-Curie fellowship (H2020-MSCA-IF-2020, no 101027973). We thank Rachel M. Pyke for confirming the issue with the HLA-I typing of one sample in Pyke et al., 2021.

AUTHOR CONTRIBUTIONS

D.G. designed the study. D.G. performed the bioinformatics analyses. D.G. and J.R. developed the bioinformatics tools. D.G. and G.C performed the TCR sequence analyses. J.S., S.B., R.G., P.G., L.Q., and J.C. performed the experiments. A.H. supervised the experiments. D.G wrote the manuscript. G.C., J.S., S.B., R.G., J.R., and A.H. provided materials and feedback on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 23, 2022
Revised: October 12, 2022
Accepted: December 8, 2022
Published: January 4, 2023

REFERENCES

1. Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. Nature *547*, 217–221.
2. 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.
3. Sahin, U., Oehm, P., Derhovanessian, E., Jabulowsky, R.A., Vormehr, M., Gold, M., Maurus, D., Schwarck-Kokarakis, D., Kuhn, A.N., Omokoko, T., et al. (2020). An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. Nature *585*, 107–112. https://doi.org/10.1038/s41586-020-2537-9.
4. Rosenberg, S.A., and Restifo, N.P. (2015). Adoptive cell transfer as personalized immunotherapy for human cancer. Science *348*, 62–68.
5. Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., da Silva Antunes, R., Moore, E., Rubiro, P., et al. (2021). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Rep. Med. *2*, 100204. https://doi.org/10.1016/j.xcrm.2021.100204.
6. Parkhurst, M., Gros, A., Pasetto, A., Prickett, T., Crystal, J.S., Robbins, P., and Rosenberg, S.A. (2017). Isolation of T-cell receptors specifically reactive with mutated tumor-associated antigens from tumor-infiltrating lymphocytes based on CD137 expression. Clin. Cancer Res. *23*, 2491–2505. https://doi.org/10.1158/1078-0432.CCR-16-2680.
7. Tran, E., Ahmadzadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P.F., Gartner, J.J., Zheng, Z., Li, Y.F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. Science *350*, 1387–1390. https://doi.org/10.1126/science.aad1253.
8. Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting antigen presentation-what could we learn from a million peptides? Front. Immunol. *9*, 1716. https://doi.org/10.3389/fimmu.2018.01716.
9. Trolle, T., McMurtrey, C.P., Sidney, J., Bardet, W., Osborn, S.C., Kaever, T., Sette, A., Hildebrand, W.H., Nielsen, M., and Peters, B. (2016). The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. J. Immunol. *196*, 1480–1487. https://doi.org/10.4049/jimmunol.1501721.
10. Peters, B., Nielsen, M., and Sette, A. (2020). T cell epitope predictions. Annu. Rev. Immunol. *38*, 123–145. https://doi.org/10.1146/annurev-immunol-082119-124838.
11. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity *46*, 315–326. https://doi.org/10.1016/j.immuni.2017.02.007.
12. Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M., and Nielsen, M. (2019). NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. Mol. Cell. Proteomics *18*, 2459–2477. https://doi.org/10.1074/mcp.TIR119.001658.
13. Bassani-Sternberg, M., and Gfeller, D. (2016). Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide-HLA interactions. J. Immunol. *197*, 2492–2499. https://doi.org/10.4049/jimmunol.1600808.
14. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P.O., Kandalaft, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allostery regulating HLA specificity. PLoS Comput. Biol. *13*, e1005725. https://doi.org/10.1371/journal.pcbi.1005725.
15. Bravi, B., Tubiana, J., Cocco, S., Monasson, R., Mora, T., and Walczak, A.M. (2021). RBM-MHC: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by HLA-I alleles. Cell Syst. *12*. 195–202.e9. https://doi.org/10.1016/j.cels.2020.11.005.
16. Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H.-G., and Stevanović, S. (2017). Unveiling the peptide motifs of HLA-C and HLA-G from naturally presented peptides and generation of binding prediction matrices. J. Immunol. *199*, 2639–2651. https://doi.org/10.4049/jimmunol.1700938.
17. Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209. https://doi.org/10.1038/s41587-019-0322-9.
18. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018). The length distribution and multiple specificity of naturally presented HLA-I ligands. J. Immunol. *201*, 3705–3716. https://doi.org/10.4049/jimmunol.1800914.
19. 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. https://doi.org/10.1016/j.cels.2020.06.010.
20. Pyke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., McClory, R., et al. (2021). Precision neoantigen discovery using large-scale immunopeptidomes and composite modeling of MHC peptide presentation. Mol. Cell. Proteomics *20*, 100111. https://doi.org/10.1016/j.mcpro.2021.100111.
21. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). 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. https://doi.org/10.1093/nar/gkaa379.
22. Shao, X.M., Bhattacharya, R., Huang, J., Sivakumar, I.K.A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., et al. (2020). High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. Cancer Immunol. Res. *8*, 396–408. https://doi.org/10.1158/2326-6066.CIR-19-0464.
23. Andreatta, M., Alvarez, B., and Nielsen, M. (2017). GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. *45*, W458–W463. https://doi.org/10.1093/nar/gkx248.
24. Koşaloğlu-Yalçın, Z., Lee, J., Greenbaum, J., Schoenberger, S.P., Miller, A., Kim, Y.J., Sette, A., Nielsen, M., and Peters, B. (2022). Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions. iScience *25*, 103850. https://doi.org/10.1016/j.isci.2022.103850.
25. Stranzl, T., Larsen, M.V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics *62*, 357–368. https://doi.org/10.1007/s00251-010-0441-4.
26. Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017). "Hotspots" of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. Front. Immunol. *8*, 1367.
27. Balachandran, V.P., Łuksza, M., Zhao, J.N., Makarov, V., Moral, J.A., Remark, R., Herbst, B., Askan, G., Bhanot, U., Senbabaoglu, Y., et al. (2017). Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature *551*, 512–516. https://doi.org/10.1038/nature24462.
28. Duan, F., Duitama, J., Al Seesi, S., Ayres, C.M., Corcelli, S.A., Pawashe, A.P., Blanchard, T., McMahon, D., Sidney, J., Sette, A., et al. (2014). Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. J. Exp. Med. *211*, 2231–2248.
29. Łuksza, M., Riaz, N., Makarov, V., Balachandran, V.P., Hellmann, M.D., Solovyov, A., Rizvi, N.A., Merghoub, T., Levine, A.J., Chan, T.A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature *551*, 517–520. https://doi.org/10.1038/nature24473.

30. Wells, D.K., van Buuren, M.M., Dang, K.K., Hubbard-Lucey, V.M., Sheehan, K.C.F., Campbell, K.M., Lamb, A., Ward, J.P., Sidney, J., Blazquez, A.B., et al. (2020). Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. Cell *183*. 818–834.e13. https://doi.org/10.1016/j.cell.2020.09.015.

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

32. Chowell, D., Krishna, S., Becker, P.D., Cocita, C., Shu, J., Tan, X., Greenberg, P.D., Klavinskis, L.S., Blattman, J.N., and Anderson, K.S. (2015). TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. Proc. Natl. Acad. Sci. USA *112*, E1754–E1762. https://doi.org/10.1073/pnas.1500973112.

33. Schmidt, J., Smith, A.R., Magnin, M., Racle, J., Devlin, J.R., Bobisse, S., Cesbron, J., Bonnet, V., Carmona, S.J., Huber, F., et al. (2021). Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. Cell Rep. Med. *2*, 100194. https://doi.org/10.1016/j.xcrm.2021.100194.

34. Sricharoensuk, C., Boonchalermvichien, T., Muanwien, P., Somparn, P., Pisitkun, T., and Sriswasdi, S. (2022). Unsupervised mining of HLA-I peptidomes reveals new binding motifs and potential false positives in the community database. Front. Immunol. *13*, 847756. https://doi.org/10.3389/fimmu.2022.847756.

35. Fritsche, J., Kowalewski, D.J., Backert, L., Gwinner, F., Dorner, S., Priemer, M., Tsou, C.-C., Hoffgaard, F., Römer, M., Schuster, H., et al. (2021). Pitfalls in HLA ligandomics—how to catch a li(e)gand. Mol. Cell. Proteomics *20*, 100110. https://doi.org/10.1016/j.mcpro.2021.100110.

36. Saini, S.K., Hersby, D.S., Tamhane, T., Povlsen, H.R., Amaya Hernandez, S.P., Nielsen, M., Gang, A.O., and Hadrup, S.R. (2021). SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8+ T cell activation in COVID-19 patients. Sci. Immunol. *6*, eabf7550. https://doi.org/10.1126/sciimmunol.abf7550.

37. Nolan, S., Vignali, M., Klinger, M., Dines, J.N., Kaplan, I.M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., Gittelman, R.M., et al. (2020). A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. https://doi.org/10.21203/rs.3.rs-51964/v1.

38. Minervina, A.A., Komech, E.A., Titov, A., Bensouda Koraichi, M., Rosati, E., Mamedov, I.Z., Franke, A., Efimov, G.A., Chudakov, D.M., Mora, T., et al. (2021). Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. eLife *10*, e63502. https://doi.org/10.7554/eLife.63502.

39. Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., Hippenstiel, S., Dingeldey, M., Kruse, B., Fauchere, F., et al. (2020). SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. Nature *587*, 270–274. https://doi.org/10.1038/s41586-020-2598-9.

40. Loyal, L., Braun, J., Henze, L., Kruse, B., Dingeldey, M., Reimer, U., Kern, F., Schwarz, T., Mangold, M., Unger, C., et al. (2021). Cross-reactive CD4+ T cells enhance SARS-CoV-2 immune responses upon infection and vaccination. Science *374*, eabh1823. https://doi.org/10.1126/science.abh1823.

41. Shimizu, K., Iyoda, T., Sanpei, A., Nakazato, H., Okada, M., Ueda, S., Kato-Murayama, M., Murayama, K., Shirouzu, M., Harada, N., et al. (2021). Identification of TCR repertoires in functionally competent cytotoxic T cells cross-reactive to SARS-CoV-2. Commun. Biol. *4*, 1365. https://doi.org/10.1038/s42003-021-02885-6.

42. Devlin, J.R., Alonso, J.A., Ayres, C.M., Keller, G.L.J., Bobisse, S., Vander Kooi, C.W., Coukos, G., Gfeller, D., Harari, A., and Baker, B.M. (2020). Structural dissimilarity from self drives neoepitope escape from immune tolerance. Nat. Chem. Biol. *16*, 1269–1276. https://doi.org/10.1038/s41589-020-0610-1.

43. Pataskar, A., Champagne, J., Nagel, R., Kenski, J., Laos, M., Michaux, J., Pak, H.S., Bleijerveld, O.B., Mordente, K., Navarro, J.M., et al. (2022). Tryptophan depletion results in tryptophan-to-phenylalanine substitutants. Nature *603*, 721–727. https://doi.org/10.1038/s41586-022-04499-2.

44. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645–3647. https://doi.org/10.1093/bioinformatics/btx469.

45. Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The immune epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343. https://doi.org/10.1093/nar/gky1006.

46. Bergmeir, C., and Benítez, J.M. (2012). Neural networks in R using the Stuttgart neural network simulator: RSNNS. J. Stat. Software *46*, 1–17.

47. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

48. Viganò, S., Utzschneider, D.T., Perreau, M., Pantaleo, G., Zehn, D., and Harari, A. (2012). Functional avidity: A measure to predict the efficacy of effector T cells? Clin. Dev. Immunol. *2012*, 153863. https://doi.org/10.1155/2012/153863.

49. Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., Bolotin, D.A., Staroverov, D.B., Putintseva, E.V., Plevova, K., et al. (2014). Towards error-free profiling of immune repertoires. Nat. Methods *11*, 653–655. https://doi.org/10.1038/nmeth.2960.

50. Corrie, B.D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F.M.W., Christley, S., Scott, J.K., et al. (2018). iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol. Rev. *284*, 24–41. https://doi.org/10.1111/imr.12666.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-hCD8	Biolegend	RRID: AB_2044010
Anti-CCR7	Biolegend	RRID: AB_11219587
Anti-CD45RA	Biolegend	RRID: AB_314412
Anti-Flag	Merck	A9469
pNPP SIGMAFast	Merck	N2770
Pacific blue mouse anti human CD8 (clone: RPA-T8)	BD bioscience	Cat# 558207; RRID:AB_397058
APC/Fire™ 750 anti-human CD3 Antibody (clone: SK7)	BioLegend	cat# 344840; RRID:AB_2572114
Aqua Fluorescente reactive dye	Invitrogen	L34966-A
Biological samples		
Healthy donors blood mononuclear cells	Biobank from the Center of Experimental Therapeutics, Department of Oncology, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.	Protocol 235/14 and 2016-02094, 2016-02166 and 2017-00490
Chemicals, peptides, and recombinant proteins		
SARS-CoV-2 peptides	ThermoFischer	N/A
pMHC multimers	Peptide & Tetramer Core Facility, University of Lausanne	N/A
RPMI	GIBCO	61870-010
MEM NEAA	GIBCO	11140-035
2-Mercaptoethanol	GIBCO	31350-010
Sodium Pyruvat	GIBCO	11360-033
HEPES	BioConcept	5-31F00H
Pen/Strep	BioConcept	4-01F00H
Human Serum	Biowest	S419H-100
FBS	Biowest	S-1810-500
IL2	Novartis	PZN02238131
DAPI	Sigma-Aldrich	10236276001
Critical commercial assays		
IFNγ Enzyme-Linked ImmunoSpot Assay	Mabtech	3420-2APT-10
Human CD8 isolation kit	Miltenyi	130-045-201
Human CD4 isolation kit	Miltenyi	130-096-533
Deposited data		
TCR sequencing data	this paper	GSE201212
Supplementary Tables	this paper	https://doi.org/10.17632/2kmmjp4tmm.1
Experimental models: Cell lines		
Jurkat cell line	Promega	T Cell Activation Bioassay (NFAT) # J1601
Schneider's Drosophila Line 2	ATCC	CRL-1963
CD4 blasts	In house	N/A
Software and algorithms		
MixMHCpred2.0.2	Gfeller et al.[18]	https://github.com/GfellerLab/MixMHCpred/releases/tag/v2.0.2
PRIME1.0	Schmidt et al.[33]	https://github.com/GfellerLab/PRIME/releases/tag/v1.0

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NetMHCpan4.1	Reynisson et al.[21]	http://www.cbs.dtu.dk/services/NetMHCpan/
MHCflurry2.0.1	O'Donnell et al.[19]	https://github.com/openvax/mhcflurry
HLAthena	Sarkizova et al.[17]	http://hlathena.tools/ (executable shared by the authors, private communications)
MixMHCpred2.2	This paper	https://doi.org/10.5281/zenodo.7375748
PRIME2.0	This paper	https://doi.org/10.5281/zenodo.7375740
Prism version 8.0.0	GraphPad Software, Inc	N/A
FlowJo X	FlowJo, LLC	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, David Gfeller (david.gfeller@unil.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- TCR sequencing data were deposited at Gene Expression Omnibus (GEO: GSE201212). The HLA-I ligand and neo-epitope datasets used to train MixMHCpred and PRIME are available in Tables S2 and S4. Additional Supplemental Items are available from Mendeley Data: https://doi.org/10.17632/2kmmjp4tmm.1
- MixMHCpred (v2.2, Zenodo: https://doi.org/10.5281/zenodo.7375748) and PRIME (v2.0, Zenodo: https://doi.org/10.5281/zenodo.7375740) are available at https://github.com/GfellerLab/ and through the web interface http://prime.gfellerlab.org/.
- Any additional information required to reproduce this work is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Six donors were recruited (Leu163, Leu158, Leu184, 1GZ0, 1HHT, 1HHU). The HLA-I typing was known for all six donors and the last three donors (1GZ0, 1HHT, 1HHU) had been tested positive for SARS-CoV-2 (PCR tests) (Table S5C). The recruitment and blood withdrawal were approved by regulatory authorities and all donors signed informed consents (Protocol 235/14 and 2016-02094, 2016-02166 and 2017-00490).

Jurkat cells (T cell activation bioassay NFAT, Promega) were used to transfect the TCR isolated in this work.

Primary $CD8^+$ T cells were cultured in RPMI 1640 Glutamax media (GIBCO) supplemented with 8 % human serum (Biowest), non-essential amino acids (GIBCO), 2-mercaptoethanol (GIBCO), sodium pyruvate (GIBCO), HEPES (GIBCO), penicillin/streptomycin (BioConcept) and 150 $IU.mL^{-1}$ of rhIL2 (Novartis). CD4 blasts were cultured in RPMI supplemented with 10 % FBS (Biowest) and penicillin/streptomycin (BioConcept). All cells were maintained at 37°C under 5 % CO_2.

METHOD DETAILS

Data collection and curation

Naturally presented HLA-I ligands of length 8 to 14 were collected from 244 samples, coming from 24 different HLA-I peptidomics studies (see Table S1). These comprise both mono- and poly-allelic samples. All data were retrieved from the original studies to prevent any filtering based on HLA-I ligand predictors. When only filtered data had been published in the original studies, access to unfiltered data was kindly provided to us by the authors of these studies. All samples were processed with the motif deconvolution algorithm MixMHCp in order to identify shared HLA-I motifs across samples sharing the same alleles, following our previously established procedure.[13,14,18] All samples were manually reviewed and peptides assigned to motifs that could not be unambiguously assigned to one allele were not considered. Peptides assigned to the flat motif in MixMHCp or to motifs corresponding to alleles not supposed to be in the sample were considered as predicted contaminants. The final dataset of naturally presented HLA-I ligands consists of 258,814 unique peptides, representing 384,070 peptide-HLA-I interactions with 119 different HLA-I alleles. 59 alleles were observed in both mono- and poly-allelic samples, 22 only in poly-allelic samples, and 38 only in mono-allelic samples. This curated set of naturally presented HLA-I ligands is available in Table S2. Binding motifs of HLA-I alleles were plotted with ggseqlogo.[44]

Immunogenic neo-epitopes were retrieved from several neo-antigen studies and were completed by neo-epitope data from IEDB[45] (tcell_full_v3.csv file, downloaded on March 27, 2021), excluding potential overlap. Both immunogenic and non-immunogenic mutated peptides were considered. This resulted in 596 experimentally verified neo-epitopes (10 8-mers, 391 9-mers, 148 10-mers and 47 11-mers) and 6084 experimentally verified non-immunogenic peptides (Table S4).

Computing peptide length distributions.

Peptide length distributions were established by computing the fraction of naturally presented HLA-I ligands of each $l \in [8, 14]$ length for each allele. For each of the 59 alleles found in both mono-allelic and poly-allelic samples, peptide length distributions were also computed separately for ligands coming from mono-allelic and poly-allelic samples (Figure 1C).

Training of MixMHCpred

MixMHCpred2.2 was trained based on our curated set of naturally presented HLA-I ligands, following the procedure described in Gfeller et al.[18] The main difference consists of a more stable modelling of peptide length distributions. In mathematical terms, the score of a peptides X of length L with allele h is given by:

$$S^h(X) = \frac{M^{(h,L)}(X) - C^{(h,L)}}{D^{(h,L)}}$$

$M^{(h,L)}(X)$ represents the raw score of peptide X given by the Position Weight Matrices representing the motif of allele h for L-mers, including normalization by background frequencies and BLOSUM62 based pseudocounts, as described in Gfeller et al.[18] The correction factors $D^{(h,L)}$ were computed so $S^h(X)$ that has a standard deviation of 1 over a set of 100'000 peptides of length L randomly selected from the human proteome (i.e., $D^{(h,L)}$ represents the standard deviation of the scores of these peptides). The correction factors $C^{(h,L)}$ were computed so that the length distribution of the top 0.1% of 700'000 random peptides (taken from the human proteome with uniform length distribution between 8- and 14-mers) follows exactly the peptide length distribution of allele h observed in HLA-I peptidomics data. In mathematical terms, defining $P^h(L)$ as the experimental peptide length distribution for a given allele h, $C^{(h,L)}$ corresponds to the raw score $M^{(h,L)}(\widehat{X})$, where $\widehat{X}$ represents the L-mer peptide ranked $100'000 \times 0.001 \times (L^{\max} - L^{\min} + 1) \times P^h(L)$ among the set of 100'000 random L-mer peptides, with and $L^{\min} = 8$ and $L^{\max} = 14$. Given the observed discrepancies between peptide length distributions from mono and poly-allelic samples (Figure 1C), peptide length distributions from poly-allelic samples were always used, when available. %ranks given as output of MixMHCpred2.2 were estimated based on the distribution of scores $S^h(X)$ of a set of 700'000 random peptides (100'000 of each length from 8 to 14), as done in other HLA-I ligand predictors. Consistent with recommendations for other tools,[21] these %rank should be used for ranking candidates to be experimentally validated. The new version of MixMHCpred (2.2) was benchmarked against NetMHCpan4.1,[21] MHCflurry2.0,[19] HLAthena[17] and MixMHCpred2.0.2[18] using naturally presented HLA-I ligands identified in unmodified tissues. To ensure that the HLA-I peptidomics samples used for this benchmark were not part of the training of any of these tools, we used (i) HLA-I peptidomics datasets coming from 10 menigioma samples measured in Gfeller et al.[18] that were not integrated in the training of any version of MixMHCpred, NetMHCpan, MHCflurry or HLAthena and were not uploaded in IEDB and (ii) HLA-I peptidomics samples from Pyke et al.[20] which were published after the latest release of these tools, excluding sample '1180157F' due to ambiguity in HLA-I typing. 4-fold excess of random negatives were added by randomly selecting peptides from the human proteome. For this comparison, we restricted to peptides of length 8 to 11, since HLAthena cannot be run for longer peptides. These peptides consist of 78,011 HLA-I ligands (counting duplicates across different samples) and 312,004 random peptides (Table S3). PPV in the top 20% (which is equivalent to recall with 4-fold excess of random negatives) and AUC values were computed for each sample and each predictor.

Comparing predicted and experimental peptide length distributions

As with predicted motifs, peptide length distributions predicted by each predictor at different %rank thresholds were computed based on 100'000 randomly selected peptides from the human proteome for each length $l \in [8, 11]$. The average predicted peptide length distributions over all alleles available in all predictors is shown in Figure 2E for each predictor and different %rank thresholds.

Training of PRIME

The new version of PRIME (v2.0) was trained using a fully connected neural network with 5 hidden nodes (*mlp* package in R[46]). The input layer consists of 28 nodes (Figure 3B). The first input node encodes the predicted binding to the HLA-I molecule (-log(%rank), predicted by MixMHCpred). The twenty next input nodes encode amino acid frequencies at positions with minimal impact on predicted affinity to HLA-I and more likely to interact with the TCR.[33] The last seven input nodes encode the length of the peptide (one-hot encoding of lengths 8 to 14).

The set of experimentally verified immunogenic and non-immunogenic peptides was used to train PRIME2.0. As this set of peptides is heavily skewed towards peptides with high predicted affinity (Figure 3A), 99-fold excess of negatives were further added by randomly selecting for each immunogenic neo-epitope 99 peptides from the same source protein (non-mutated), for a total of 58,905 random peptides (for one neo-epitope, the source protein could not be found, and no random peptide was included for this neo-epitope). The length of these negatives was randomly chosen between 8 and 14. The use of only human (mutated) peptides in both positives and negatives prevents potential biases in amino acid frequencies due to different GC content across different organisms.

To benchmark the new version (2.0) of PRIME, we first performed a standard 10-fold cross-validation, by randomly splitting the data in ten groups, iteratively training the model on nine groups and testing on the remaining one. Given that our dataset of immunogenic neo-epitopes is skewed towards frequent HLA-I alleles and towards studies where many neo-epitopes had been reported, we also performed a leave-one-allele-out, respectively a leave-one-study-out, cross-validation, using iteratively as test set each allele, respectively each study, with more than two experimentally validated immunogenic and two experimentally validated non-immunogenic peptides. PRIME2.0 was benchmarked against MixMHCpred2.2 developed in this work, NetMHCpan4.1,[21] MHCflurry,[19] HLAthena[17] and PRIME1.0[33] (https://github.com/GfellerLab/PRIME/releases/tag/v1.0). PRIME2.0 was also benchmarked against a model trained exactly on the same data but using a logistic regression (*glmnet* package in R, with *family*="binomial" and *lambda* = 1[47]).

Predictions of SARS-CoV-2 epitopes

The SARS-CoV-2 reference proteome was downloaded from UniProt on March 22, 2020 and peptides of length 8 to 11 were retrieved. The list of HLA-I alleles was established by taking the top 15 most frequent alleles in the TCGA cohort (Table S5B). Only peptides with a %rank lower or equal to 0.5 for PRIME2.0 for at least one allele and coming from the five proteins SPIKE, VME1, VEMP, NCAP, AP3A were considered. A few peptides from R1AB were further manually included as they came from regions with several predicted epitopes for multiple alleles, and some peptides were manually removed from the list. The final list consists of 213 peptides (Table S5A).

Identification of SARS-CoV-2 epitopes

The 213 peptides were purchased at ThermoFisher (>80 % purity), solubilized in DMSO at 10 mM, aliquoted and kept at -80°C. $CD8^+$ T cells were isolated (ref 130-045-201, Miltenyi) from cryopreserved PBMC (for SARS-CoV-2 positive donors) or fresh leukapheresis (for SARS-CoV-2 negative donors). $CD4^+$ T cells were isolated (ref 130-096-533) and used to generate CD4 blasts. For SARS-CoV-2 positive donors (1HHU, 1HHT, 1GZ0), due to the limited number of PBMCs, total $CD8^+$ T cells were used for further *in vitro* stimulation. For the other three donors (Leu163, Leu158. Leu184), naïve and effector/memory $CD8^+$ T cells were isolated by Fluorescence-activated Cell Sorting (FACS) upon staining with anti-CD8 antibody (344710 BioLegend), anti-CCR7 antibody (353227 BioLegend) and anti-CD45RA antibody (304108 BioLegend) for 30 min at 4°C. After three washes with FACS buffer (PBS 0.5 % FBS 2 mM EDTA) cells were incubated 10 min with DAPI (Sigma 10236276001) at 250 nM and washed again three times. Total $CD8^+$ T cells (donors 1GZ0, 1HHT, 1HHU), naïve ($CCR7^+$ and $CD45RA^+$) $CD8^+$ T cells (donors Leu163, Leu158, Leu184) and effector/memory ($CD45RA^-$) $CD8^+$ T cells (donor Leu184 – not enough effector/memory cells available for the other donors) were collected separately and then co-incubated (10^6 mL^{-1}) with autologous irradiated CD8-depleted PBMCs and pools of 11 to 24 peptides (1μM) in RPMI supplemented with 8 % human serum and IL-2 (50 IU mL^{-1} for 48h and then switch 1mL of media with 150 IU mL^{-1} every 48h, split as necessary to get minimum 10^6 Cell.mL^{-1}). IFNγ Enzyme-Linked ImmunoSpot (ELISpot) was performed at day 12 post-stimulation. One day before ELIspot, cells were incubated in RPMI supplemented with 8 % human serum without IL2. ELISpot assays were performed using pre-coated 96-well ELISpot plates (Mabtech 3420-2APT-10) and counted with Bioreader-6000-E (BioSys). Briefly, 100,000 $CD8^+$ T cells were incubated for 16h with 30,000 $CD4^+$ T cell blasts pulsed for 1h with 1μM peptide pools. All peptide pools giving a specific response (considered if at least 10 spots for 100 000 incubated cells and 2 times the background signal, obtained by incubation of cells without peptide) were deconvoluted by repeating ELISpot assays with individual peptides.

Predictability of SARS-CoV-2 epitopes

The %rank of the 213 SARS-CoV-2 peptides tested for immunogenicity was computed with the different predictors in each donor. The final score for each peptide in a given donor was taken as the best %rank across all alleles of this donor. The distributions of the scores for the 18 immunogenic peptides in their respective donors are shown in Figure S4A for each predictor. For each patient, AUC values were also computed based on these scores to illustrate how different tools would have performed (Figure S4B). It should still be emphasized that this analysis has some biases since, for practical reasons, the initial list of 213 peptides was based on PRIME2.0 predictions and without considering the HLA-I alleles of the actual donors.

Peptide-HLA multimer validation of SARS-CoV-2 epitopes and sorting of $CD8^+$ T cells

Peptides found as immunogenic in the ELISpot assays were resynthesized with a purity >95 % and used for production of peptide-HLA multimers (Peptide and Tetramer Core Facility of the University Hospital of Lausanne). $CD8^+$ T cells were incubated with multimers (1/50 dilution) 45 min at 4°C in FACS buffer (PBS supplemented with 0.5 % FBS and 2mM EDTA), isolated by FACS and either directly used for TCR sequencing or expanded with autologous irradiated CD8-depleted feeders in RPMI supplemented with 8% human serum, phytohemagglutinin (1 μg mL^{-1}) and IL2 (150 IU mL^{-1}).

Functional avidity assay

Functional avidity of antigen-specific $CD8^+$ T-cell responses was assessed by performing *in vitro* IFNγ Enzyme-Linked ImmunoSpot (Mabtech) assay with limiting peptide dilutions (ranging from 10μM to 100pM) as described earlier.[48] For all peptide concentrations, ELISpot signals were measured in two replicates and the average of the two replicates was used to compute EC_{50} values. EC_{50} values reported in Figure 4C were computed by fitting sigmoid curves with the "ec50estimator" package in R (https://github.com/AlvesKS/ec50estimator). For EYADVFHLYL, enough cells were available for only one replicate. For the HLA-A*29:02 restricted

YFPLQSYGF epitope, single clones were isolated and the EC_{50} values represent the average over all clones coming from two different pools (error bars represent the standard deviation between the average values in the two pools). For this epitope, peptide concentrations ranging from 10^{-11} to 10^{-6} M were used, as the first response was stronger than for other epitopes (Figure 4C).

Bulk TCR sequencing

mRNA was extracted using the Dynabeads mRNA DIRECT purification kit according to the manufacturer instructions (ThermoFisher) and was then amplified using the MessageAmp II aRNA Amplification Kit (Ambion) with the following modifications: *in vitro* transcription was performed at 37°C for 16 h. First strand cDNA was synthesized using the Superscript III (Thermofisher) and a collection of TRAV/TRBV specific primers. Unique Molecular identifiers (UMI) of length 9 were added to each read. TCRs were then amplified by PCR (20 cycles with the Phusion from NEB) with a single primer pair binding to the constant region and the adapter linked to the TRAV/TRBV primers added during the reverse transcription. A second round of PCR (25 cycles with the Phusion from NEB) was performed to add the Illumina adapters containing the different indexes. The TCR products were purified with AMPure XP beads (Beckman Coulter), quantified and loaded on the MiSeq instrument (Illumina) for deep sequencing of the TCRα/TCRβ chain.

TCR sequence analyses

The fastq files were processed with MIGEC,[49] using default parameters to demultiplex them and identify the TCRα and TCRβ clonotypes. For each sample, the frequency of each TCR chain was computed based on UMI corrected counts. Only TCRs with more than one UMI count and representing more than 1% of the total UMI counts were considered (Table S6). TCRs with the same amino acid sequences were merged in Figures 4D and S4C.

The beta chain of the TCR_{QYI} (TRBV4-3*01-CASSPSGGAYEQYF-TRBJ2-7*01) recognizing the QYIKWPWYIW epitope and found in effector/memory $CD8^+$ T cells of Leu184 was used to search TCRβ repertoires in the ImmunoCode database[37] through the iReceptor web platform.[50] Both the alpha and beta chains were used to query separately the TCRα and TCRβ repertoires of the two SARS-CoV-2+ patients in Minervina et al.[38] The closest hits (Figure S4F) were defined as those having the same CDR3 sequence and the most similar CDR1 and CDR2, based on sequence identity (100% identity for the alpha chain in both donors, 100% identity for the beta chain in donor M).

TCR_{QYI} transfection in Jurkat cells and recognition of the homologous peptides from other coronaviruses

TCR_{QYI} full-length α and β chains were in silico designed and obtained by Thermo Fisher Scientific as strings. Strings have been amplified and purified by silica membrane columns (NucleoSpin PCR Clean-up, Macherey-Nagel) and used as individual templates for mRNA in vitro transcription using the HiScribe T7 In vitro transcription kit (NEB), followed by lithium chloride precipitation, as instructed by the manufacturer. RNA polyadenylation and molecular size were assessed by gel electrophoresis in denaturing conditions. Purified RNA was quantified using a Qubit BR Assay kit (Thermo Fisher Scientific) and resuspended in H_2O at 1-2μg/mL followed by storage at -80°C, until used.

TCRα and TCRβ pairs were transfected into a recipient Jurkat cell line (T cell activation bioassay NFAT, Promega) that was further engineered by knocking out the endogenous TCRα and TCRβ chains using CRISPR/Cas9 and by stable transduction with CD8A and CD8B. Cells were propagated following the manufacturer's instructions. For TCR transfection, $1x10^6$ Jurkat cells were co-electroporated with 1.5μg of each TCR chain using a Neon Transfection System 100μl kit (Thermo Fisher Scientific) with the following parameters: 1325V, 10ms, 3 pulses. After electroporation, cells were immediately resuspended in complete medium and incubated at 37°C for 18-20 hours before staining. TCR_{QYI} electroporated Jurkat cells were stained with a PE conjugated QYIKWPWYVW-, TYIKWPWWVW-, YYVKWPWYVW-, NYIKWPWWVW-, MYVKWPWYVW-HLA-A*24:02 multimers (Peptide and Tetramer Core Facility of the University Hospital of Lausanne), washed once and further stained with anti-CD3 APC-Fire (Biolegend) and -CD8 FITC (BD Biosciences) fluorophore-conjugated anti-human antibodies. Aqua live dye (Thermo Fisher Scientific) was used to access viability. As control, Jurkat cells electroporated with a TCR recognizing the EBV epitope RAKFKQLL displayed on HLA-B*08:01 were used. The samples were acquired by LSR Fortessa (BD Biosciences) and analysed by FlowJoX.

QUANTIFICATION AND STATISTICAL ANALYSIS

P-values for the comparison between AUC values obtained for different predictors in the different cross-validation schemes were computed with paired Wilcoxon test.

IFNγ ELISpot results were considered as positives if the number of spots was larger than 10 for 100 000 incubated cells and larger than 2 times the background signal (obtained by incubation of cells without peptide).