frontiers in Genetics

ORIGINAL RESEARCH
published: 25 March 2021
doi: 10.3389/fgene.2021.602196

AI Aided Design of Epitope-Based Vaccine for the Induction of Cellular Immune Responses Against SARS-CoV-2

Giovanni Mazzocco[1]*, Iga Niemiec[1], Alexander Myronov[1,2], Piotr Skoczylas[1], Jan Kaczmarczyk[1], Anna Sanecka-Duin[1], Katarzyna Gruba[1,2], Paulina Król[1], Michał Drwal[1], Marian Szczepanik[3], Krzysztof Pyrc[4] and Piotr Stępniak[1]

[1] Ardigen, Krakow, Poland, [2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, [3] Department of Medical Biology, Faculty of Health Sciences, Jagiellonian University Medical College, Krakow, Poland, [4] Virogenetics Laboratory of Virology, Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

OPEN ACCESS

Edited by:
Nimisha Ghosh,
Siksha 'O' Anusandhan University,
India

Reviewed by:
Michael Poidinger,
Murdoch Childrens Research
Institute, Royal Children's Hospital,
Australia
Gustavo Fioravanti Vieira,
Universidade La Salle Canoas, Brazil

***Correspondence:**
Giovanni Mazzocco
giovanni.mazzocco@ardigen.com

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 02 September 2020
Accepted: 28 January 2021
Published: 25 March 2021

Citation:
Mazzocco G, Niemiec I, Myronov A, Skoczylas P, Kaczmarczyk J, Sanecka-Duin A, Gruba K, Król P, Drwal M, Szczepanik M, Pyrc K and Stępniak P (2021) AI Aided Design of Epitope-Based Vaccine for the Induction of Cellular Immune Responses Against SARS-CoV-2. Front. Genet. 12:602196. doi: 10.3389/fgene.2021.602196

The heavy burden imposed by the COVID-19 pandemic on our society triggered the race toward the development of therapies or preventive strategies. Among these, antibodies and vaccines are particularly attractive because of their high specificity, low probability of drug-drug interaction, and potentially long-standing protective effects. While the threat at hand justifies the pace of research, the implementation of therapeutic strategies cannot be exempted from safety considerations. There are several potential adverse events reported after the vaccination or antibody therapy, but two are of utmost importance: antibody-dependent enhancement (ADE) and cytokine storm syndrome (CSS). On the other hand, the depletion or exhaustion of T-cells has been reported to be associated with worse prognosis in COVID-19 patients. This observation suggests a potential role of vaccines eliciting cellular immunity, which might simultaneously limit the risk of ADE and CSS. Such risk was proposed to be associated with FcR-induced activation of proinflammatory macrophages (M1) by Fu et al. (2020) and Iwasaki and Yang (2020). All aspects of the newly developed vaccine (including the route of administration, delivery system, and adjuvant selection) may affect its effectiveness and safety. In this work we use a novel *in silico* approach (based on AI and bioinformatics methods) developed to support the design of epitope-based vaccines. We evaluated the capabilities of our method for predicting the immunogenicity of epitopes. Next, the results of our approach were compared with other vaccine-design strategies reported in the literature. The risk of immuno-toxicity was also assessed. The analysis of epitope conservation among other *Coronaviridae* was carried out in order to facilitate the selection of peptides shared across different SARS-CoV-2 strains and which might be conserved in emerging zootic coronavirus strains. Finally, the potential applicability of the selected epitopes for the development of a vaccine eliciting cellular immunity for COVID-19 was discussed, highlighting the benefits and challenges of such an approach.

Keywords: SARS-CoV-2, COVID-19, *Coronaviridae*, vaccines, cellular immunity, epitopes, CD8+, CTL

INTRODUCTION

As of August 6, 2020, more than 19 million cases of COVID-19 were reported worldwide, leading to more than 700 thousands deaths[1]. The disease was first recorded on December 26, 2019, when a 41-year-old patient with no history of hepatitis, tuberculosis, or diabetes was hospitalized at the Central Hospital of Wuhan due to respiratory problems (Wu F. et al., 2020). The metagenomic RNA sequencing of bronchoalveolar lavage (BAL) fluid sample obtained from that patient led to the identification of the seventh coronavirus (CoV) strain known to infect humans.

Coronaviruses are well known human respiratory pathogens associated with the common cold. Until the 21st century they were neglected by the medical world, but the emergence and subsequent spread of the SARS-CoV in the 2002/2003 season raised interest in this virus family and increased awareness of the potential threat. At present, there are four seasonal coronaviruses infecting humans and they cluster within alphacoronaviruses (HCoV-NL63, HCoV-229E) and betacoronaviruses (HCoV-OC43, HCoV-HKU1) genera. Further, three zoonotic strains were reported – severe acute respiratory syndrome coronavirus (SARS-CoV; 2002–2003), the Middle East respiratory syndrome coronavirus (MERS-CoV; 2012-), and SARS-CoV-2 (2019-), all of which belong to the betacoronavirus genus (Wu A. et al., 2020). The highly pathogenic species cluster in two subgenera – sarbecoviruses (SARS-CoVs) and merbecoviruses (MERS-CoVs) (Hu et al., 2018; Wu F. et al., 2020; Zhou et al., 2020).

While generally, viruses infect one host, some have broader specificity or can cross the interspecies borders, causing outbreaks, epidemics, and pandemics. In this context, it is worth mentioning viruses like the Ebola virus, dengue fever virus, Nipah virus, rabies virus, or Hendra virus. However, these are well known and long studied animal viruses that only sometimes enter the human population. Coronaviruses are slightly different, as among the myriads of viral species and subspecies found in animals, it is unlikely to predict the place, the time, and the genotype of the coronavirus that will emerge. The classic transmission route of these viruses encompasses the spillover of the bat species to wild or domesticated animals, rapid evolution in this intermediate host, and subsequent transmission to humans. Coronaviruses emerge at different sites of the globe where the interaction between humans and animals is broad, such as the Asian wet markets and the dromedary camel farms in the Arabian peninsula. While these high-risk regions were identified, the next spillover may occur in Europe or the Americas, as sarbecoviruses are prevalent around the globe (Andersen et al., 2020).

The coronavirus genome is a single-stranded RNA of positive polarity, which ranges in size from 26,000 up to 32,000 bases. Two-thirds of the genome on the 5′ end are occupied by two large open reading frames (ORFs) that may be read along due to the ribosomal slippery site. The resulting polyprotein undergoes subsequent autoproteolysis, and the matured proteins form the complete replicatory machinery and re-shape the microenvironment of the infection. Downstream of the 1ab ORFs, a number of ORFs are found that encode structural and accessory proteins (Cui et al., 2019; Song et al., 2019). Four major structural proteins are: spike surface glycoprotein (S), envelope protein (E), membrane glycoprotein (M), and nucleocapsid phosphoprotein (N). Of them the S protein is the primary determinant of the species and cell tropism, interacting with the receptors and co-receptors on the host cells (Li, 2016; Zhu et al., 2020).

Evolutionary studies indicate that CoV genomes display high plasticity in terms of gene content and recombination (Forni et al., 2016). The long CoV genome expands the sequence space available for adaptive mutations, and the spike glycoprotein used by the virus to engage target cells can adapt with relative ease to exploit homologs of cellular receptors in different species. While coronaviruses are rapidly evolving, their mutation rate is lower than expected for an RNA virus. The large genomes require proofreading machinery to maintain their functions, and proteins required for such activity are among the 1a/1ab proteins.

While sarbecoviruses and merbecoviruses are associated with severe, potentially lethal diseases and are known for their epidemic potential in humans and animals, several years of research did not allow for the development of effective and safe vaccines. In addition to the high variability and ability to elude immune recognition, there are several aspects to be considered. First, the antibody-dependent enhancement (ADE) of the infection was previously reported for some coronaviruses, including sarbecoviruses. ADE is based on the fact that the virus exploits non-neutralizing antibodies to enter the host's cells utilizing the Fc receptor (FcR). The ADE phenomenon was originally observed for antibodies specific to certain dengue virus serotypes developed after a primary infection. During subsequent dengue infections, caused by a different virus serotype, these antibodies were able to recognize the virus but were not capable of neutralizing it. Instead, antibodies bridged the dengue virus and the Fc receptors of the immune cells, such as macrophages and B-cells, mediating viral entry into these cells and transforming the disease from a relatively mild illness to a life-threatening infection. A similar mechanism was later observed for HIV and Ebola infections (Takada et al., 2003, 2001; Guzman et al., 2007; Whitehead et al., 2007; Beck et al., 2008; Dejnirattisai et al., 2010; Willey et al., 2011; Katzelnick et al., 2017). Importantly, ADE has also been reported for some coronaviruses. The best-documented ADE cases are associated with feline infectious peritonitis virus. It was shown that immunization of cats with feline coronavirus spike protein leads to increased severity during future infections due to the induction of infection-enhancing antibodies (Corapi et al., 1992; Hohdatsu et al., 1998). Further, some studies show that antibodies induced by the SARS-CoV spike protein enhance viral entry into FcR-expressing cells (Kam et al., 2007; Jaume et al., 2011; Wang et al., 2014). It was confirmed that this Abs-dependent SARS-CoV entry was independent of the classical ACE2 receptor-mediated entry (Jaume et al., 2011). A recent study investigated the molecular mechanism behind antibody-dependent and receptor-dependent viral entry of MARS-CoV and SARS-CoV pseudoviruses *in vitro* (Wan et al., 2019). The authors demonstrated that MERS-CoV and SARS-CoV neutralizing monoclonal antibodies (mAbs) binding to the receptor-binding domain region of the respective spike protein

[1] https://coronavirus.jhu.edu/map.html

were capable of mediating viral entry into FcR-expressing human cells, confirming the possibility of coronavirus-mediated ADE. Given the critical role of antibodies in host immunity, ADE causes serious concerns in epidemiology, vaccine design, and antibody-based drug therapy.

The consequences of ADE may be dramatic, as it may cause lymphopenia and induce or increase the frequency of the cytokine storm syndrome (CSS). This may result directly from the active infection of immune cells, which in response produce large amounts of the inflammatory markers or indirectly, when virus-antibody complex binds to FcR and activates pro-inflammatory signaling, skewing macrophages responses to the accumulation of pro-inflammatory M1 macrophages in lungs. The macrophages secrete inflammatory cytokines, such as MCP-1 and IL-8, which lead to worsened lung injury (Fu et al., 2020). In both animal models and patients who eventually died from SARS, extensive lung damage was associated with high initial viral loads, increased accumulation of inflammatory monocytes/macrophages in the lungs, and elevated levels of serum pro-inflammatory cytokines and chemokines (IL-1, IL-6, IL-8, CXCL-10, and MCP1) (Channappanavar et al., 2016). Moreover, during the SARS-CoV outbreak in Hong Kong (2003–2004), 80% of the patients developed acute respiratory distress syndrome after 12 days from the diagnosis, coinciding with IgG seroconversion (Peiris et al., 2003). Another study by Huang et al. (2020) highlighted an increased release of IL-1β, IL-4, IL-10, IFNγ, MCP-1, and IP-10 in COVID-19 patients. Interestingly, compared with non-severe cases, severe patients in the intensive care unit showed higher plasma concentrations of TNFα, IL-2, IL-7, IL-10, MIP-1A, MCP-1, and G-CSF, supporting the hypothesis of a possible correlation between CSS and severity of the disease. An extensive study done by Liu et al. (2019) demonstrated that anti-spike IgGs enhanced the induction of pro-inflammatory cytokines (i.e., IL-6, IL-8, and MPC-1) in Chinese rhesus monkeys through the stimulation of alternatively activated monocyte-derived macrophages (MDM) upon SARS-CoV rechallenge. The presence of high MDM infiltrations was shown by histochemical staining of the lung tissue from 3 deceased SARS patients. The blockade of Fc-receptors for IgG (FcγRs) reduced proinflammatory cytokine production, suggesting a potential role of FcγRs for the reprogramming of alternatively activated macrophages. Putting these results in the context of other works in literature (Pahl et al., 2014), one has to consider that anti-S IgG may promote pro-inflammatory cytokine production and, consequently, CSS development.

Taking into account the risk associated with the improper humoral response and high variability of sites targeted by the neutralizing antibodies, together with the low effectiveness of IgG-mediated immunity during mucosal infection, it is of importance to consider the anticoronaviral vaccine in a broader perspective. This may include alternative delivery systems/routes based on, e.g., virus-like particles and intranasal delivery for the IgA mediated response, but it is also important to consider combining the humoral response with the cell-mediated response. Ideally, such an approach might allow for the design of a vaccine carrying carefully selected epitopes to induce only the neutralizing antibodies and epitopes targeted for induction of the cellular response. While neutralizing antibodies impair the virus entry, activated CD8+ cytotoxic T-cells can identify and eliminate infected cells. Moreover, CD4+ helper T-cells are required to stimulate the production of antibodies. Antibody response was found to be short-lived in convalescent SARS-CoV patients (Tang et al., 2011) in contrast to T-cell responses, which have been shown to provide long-term protection (Peng et al., 2006; Fan et al., 2009; Tang et al., 2011), up to 11 years post-infection (Ng et al., 2016). The activation of CD8+ cytotoxic T-cells capable of recognizing and destroying infected cells represents a crucial second line of defense against the virus that should be considered. The importance of both CD8+ and CD4+ T-cell activation has been reported in several SARS-CoV studies for both animal models and humans (Channappanavar et al., 2014). Moreover, several recent studies indicate a strong correlation between the reduction of lymphocyte counts (CD4+ and CD8+) and the severity of COVID-19 cases (Chen et al., 2020; Liao et al., 2020; Wan et al., 2020).

The selection of epitopes capable of eliciting either B-cell or T-cell responses is a critical step for the development of subunit vaccines. Most of the efforts in this area are directed toward the stimulation of neutralizing antibodies, whereas the cellular immune response is less explored. Considering the importance of T-cell activation for vaccine efficacy, the focus of the work here presented is on the latter. Despite the apparent similarity between SARS-CoV and SARS-CoV-2, there is still a considerable genetic variation between these two. Thus, it is not trivial to assess if epitopes eliciting an immune response against previous coronaviruses are likely to be effective against SARS-CoV-2, with the exception of identical peptides shared among subgenera. A restricted list of SARS-CoV epitopes identical to those present in SARS-CoV-2 and resulting positive in immunoassays, has been recently reported (Ahmed et al., 2020). Nonetheless, the 29 T-cell epitopes described therein are mostly limited to S, N, and M antigens and encompass an exiguous number of Class I Human Leukocyte Antigen (HLA) alleles. In order to extend the search area to other epitopes, computational predictive models might be applied. Methods for the selection of vaccine peptides are typically based on the predicted binding affinity (or probability of presentation on the cell surface) of peptide-HLA (pHLA) complexes or defined by the physicochemical properties of the peptides (Baruah and Bose, 2020; Grifoni et al., 2020; Lee and Koohy, 2020). These methods take into account only restricted parts of processes contributing to the final immunogenicity of an epitope, and thus their prediction capabilities are limited. In addition to pHLA binding, proteasome cleavage, pHLA loading, and presentation, as well as direct activation of CD8+ T-cell to the pHLA complex should be taken into account.

Here, we use a machine learning model for the prediction of epitope immunogenicity. The model is trained on data including the experimental T-cell immunogenicity data of viral epitopes. We validate our model on publicly available immunogenicity data of epitopes from the *Coronaviridae* virus family (held out from training). Assessment of the risk of immuno-toxicity and the analysis of epitope conservation among different strains are also performed.

MATERIALS AND METHODS

Presentation Data

A curated dataset containing peptides presented by class I HLAs on the surface of host cells was extracted from publicly available databases (Abelin et al., 2017; Di Marco et al., 2017; Sarkizova et al., 2020). The presentation of each peptide within the dataset was experimentally confirmed by mass-spectroscopy experiments. All peptides were of human origin and were presented on the surfaces of monoallelic human cell lines (see **Figure 1** and **Table 1**). Synthetic negative data (non-presented peptides) were also prepared based on human proteome (GRChg38, release 98).

Immunogenicity Data

All peptides collected from the IEDB database (Vita et al., 2019) were of viral origin and were confirmed in experimental immunoassays. Similar data were extracted from selected publications (Wang et al., 2004; Chen et al., 2005; Tsao et al., 2006; Liu et al., 2010; Zhang, 2013; Ogishi and Yotsuyanagi, 2019). The number of pHLAs (per immunoassay category) used for training is given in **Table 2**. Most of the peptides were obtained from human hosts, with a minority obtained from transgenic mice. Only peptides containing 8–11 amino acids were included in the analysis. In some cases, multiple experimental settings and protocols were used to validate immunogenicity for a given pHLA, occasionally leading to non-consensual results. Each pHLA was considered immunogenic if at least one experiment conducted on human cells positively confirmed that immunological event. If no experiments conducted on human cells were available, the pHLA was considered immunogenic, if at least one such confirming experiment was conducted in transgenic mice. The remaining pHLAs were used as negative examples. From this dataset we held out the *Coronaviridae* family as a separate test set.

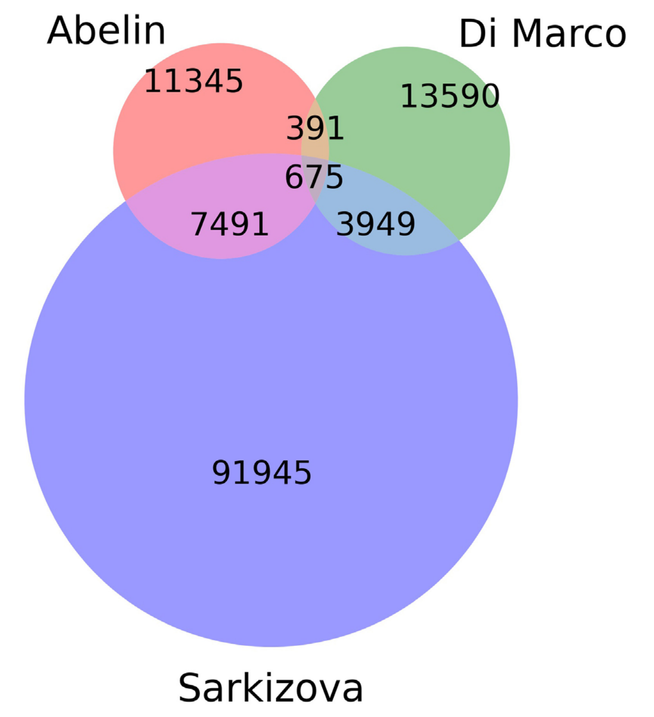

FIGURE 1 | Venn diagram showing the number of unique and common peptides among datasets.

TABLE 1 | The total number of pHLAs included in our model from each dataset.

Source publication	No. pHLAs
Abelin et al. (2017)	22,999
Di Marco et al. (2017)	22,889
Sarkizova et al. (2020)	146,739

Predictive Model Design

Our computational methods are based on machine learning and predict (1) the probability of pHLAs to be presented on the host's cell surface and (2) the immunogenicity of such complexes. The model for pHLA presentation is based on artificial neural networks and has been trained on a curated collection of peptide presentation data (Abelin et al., 2017; Di Marco et al., 2017; Sarkizova et al., 2020). Both peptide sequence and HLA type were taken into consideration as separate inputs. We use bootstrapping and select 80% of positive examples during training with the remaining ones used for early stopping. We then ensemble the results of a collection of 27 such neural networks. Our model is pan-specific and can be used to generate predictions for any peptide and any canonical class I HLA (i.e., A, B, and C). Note, that the accuracy of our method depends on the considered HLA type, as in the case of other machine learning methods for predicting pHLA properties.

The model mentioned above was also used as a starting point for training the immunogenicity model. The latter was fine-tuned using the viral peptide immunogenicity data collected from IEDB (Vita et al., 2019) and Ogishi and Yotsuyanagi (2019). The immunogenicity model was validated using a Leave One Group Out (LOGO) cross-validation scheme with groups defined by viral families. The final model is an ensemble of 11 models – one per each LOGO split. An additional group "others" was defined by aggregating data from viruses that belong to several families, having a small number of observations. Such an approach provides data splits according to the virus families and leads to a better measure of performance on virus families not seen in training (e.g., *Coronaviridae*). Moreover, it reveals the differences in model performance on various virus families.

TABLE 2 | The number of pHLA complexes used for training per immunogenic assay group.

Source publication	Negative	Positive
IFN(γ)	23,249	2,598
Cytotoxicity	218	524
Proliferation	7	34
cytokines/chemokines	0	13
TNF(α)	1	8

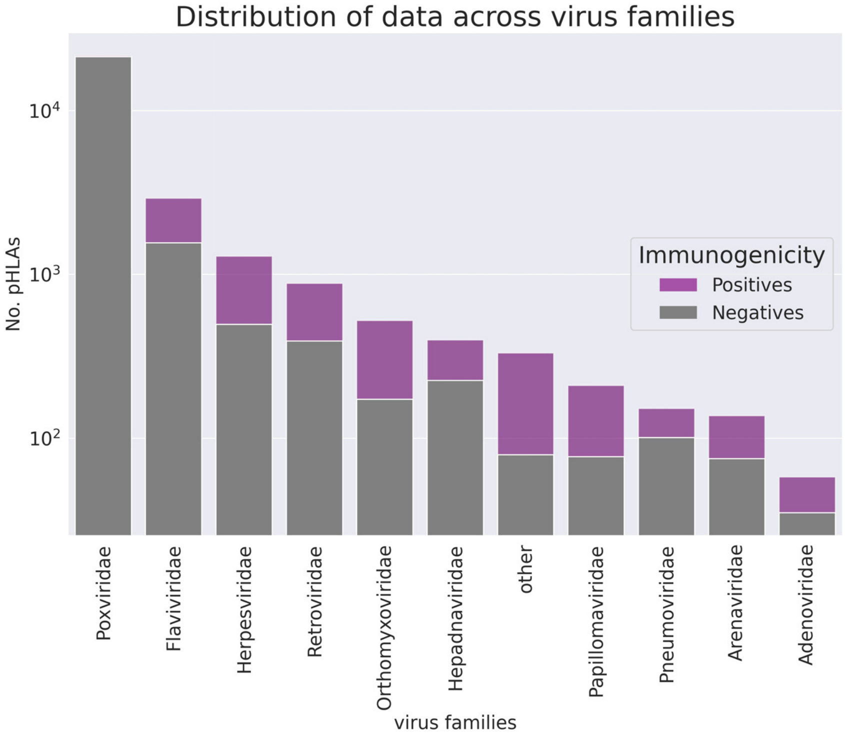

FIGURE 2 | The number of pHLA complexes with confirmed immunogenicity in the curated database per virus family (logarithmic scale). Families counting less than 55 observations are aggregated in the "other" group.

The final predictions of our model (called ArdImmune Rank) are obtained by combining the predictions of both models (i.e., the pHLA presentation and the immunogenicity model).

Both models were implemented in Python 3.7 using the keras 2.4.3 package, which is a high-level API of TensorFlow. For our usage TensorFlow with GPU support was deployed, i.e., tensorflow-gpu 2.2.0. For GPU-based computations we used *cudnn* 7.6.5 and *cudatoolkit* 10.2.89 and a machine equipped with NVIDIA Tesla V100 GPU card with CUDA® 7.0 architecture, 640 Tensor Cores, 5,120 CUDA® Cores and 32 GB HBM2 GPU Memory. Additionally, *scikit-learn*, *pandas,* and *numpy* were used to perform standard machine learning tasks while images were produced using *matplotlib* and *seaborn*.

Validation Scheme

In order to validate the ArdImmune Rank model over different virus families not seen during the training procedure, a LOGO strategy was applied. The peptides associated with coronaviruses were held out from the dataset and left for testing purposes only. At each LOGO iteration, the dataset was split into training and validation sets, and the model was tested accordingly. Peptides within the training set highly similar to the ones in the validation set were removed from the training set. The similarity of peptides was assessed using a clustering algorithm classifying their sequences into groups of peptides sharing a common root (differing only by short prefixes or suffixes of lengths of at most three amino acids). The number of pre-processed peptides in each group is given in **Figure 2**. Finally, the immunogenicity model (an ensemble of 11 models from the LOGO scheme) was validated on the held-out *Coronaviridae* dataset.

SARS-CoV-2 Data Analysis

Selection of HLA Alleles

Class I HLA types were chosen based on their frequency of occurrence in the United States and Europe. HLA-allele frequency data were downloaded from[2], accounting for all the populations within the regions of choice and all ethnicities. The overall frequency for each allele was computed as the weighted average with weights corresponding to the size of each population, separately for the United States and Europe, encompassing all ethnic populations. All HLA-alleles with frequency ≥ 0.01 were chosen for the study.

Toxicity/Tolerance Evaluation

In order to evaluate the risk for a given pHLA to be cross-reactive or tolerogenic with respect to self-epitopes within the human proteome, a procedure for the evaluation of potential toxicity/tolerance was implemented. Initially, each SARS-CoV-2 peptide was queried against the reference human proteome (GRChg38, release 100) using the BLASTp algorithm and a BLOSUM45 substitution matrix. All matches with *E*-values less than or equal to four were included in the analysis. The selected peptides are available in **Supplementary Data 1**.

[2] http://www.allelefrequencies.net/

Selection of Peptides

The dataset consisting of SARS-CoV-2 peptides was generated according to the following procedure: (1) all the reference sequences of the virus proteins were collected from the NCBI database[3], (2) from each protein, all possible peptides of length 8–11 amino acids were selected. In addition, for proteins encoded by the ORF1a and ORF1ab genes (i.e., pp1a, pp1ab, respectively), the peptides within the cleavage sites were excluded. Finally, all the peptide duplicates were removed from the dataset. A total of 47,612 peptide sequences were collected.

Estimation of SARS-CoV-2 Genome Diversity

The analysis of conservation of SARS-CoV-2 genomic sequences was performed using 8,639 complete genomic sequences obtained from the GISAID database[4] and GenBank[5]. All sequences were aligned to the SARS-CoV-2 reference genome (NCBI Reference Sequence: NC_045512.2). The R DECIPHER package (Wright, 2015) v2.14.0 was used to perform the multiple sequence alignment (MSA) of long SARS-CoV-2 whole genome sequences. The following parameters were applied: AlignSeqs(sequences, iterations = 2, refinements = 1, gapOpening = c(−18, −16), gapExtension = c(−2, −1), FUN = AdjustAlignment, processors = 18). In order to align short sequences with partial fragments of the SARS-CoV-2 genome, the R Biostrings v2.54.0 package was used, adopting the following parameters: Biostrings:pairwiseAlignment(pattern = sequences, subject = reference_genome, type = "local," and scoreOnly = F). Next, all the nucleotides within the coding cDNA sequence (CDS) regions of the reference genome were translated into amino acids using the *translate* function available in the R Biostring package v2.45.0 (Pagès and DebRoy, 2020) with the following parameters: Biostrings:translate [DNAStringSet(sequences), if.fuzzy.codon = "solve"]. The Standard Genetic Code provided by default was used for the encoding. All the fuzzy codons were marked as unknown amino acids by setting the *if.fuzzy.codon* = "*solve*" parameter. For each protein, all sequences containing indels or being inconveniently aligned were removed. Inconvenient sequences include those having short reading frameshifts, marked as transcription artifacts. Mutation frequencies for both long and short genomics fragments were computed for each amino acid in the SARS-CoV-2 proteome. The mutation frequency of each amino acid was defined as the ratio between the number of translated protein sequences containing the mutation and the number of sequences containing a valid nucleotide (sequences containing unknown nucleotides in this position were excluded). The maximum mutation frequency score for each peptide was computed as the maximum value of the mutation frequency scores among all amino acid positions of the peptide. Mutation frequency values for all positions within SARS-CoV-2 proteome are available in **Supplementary Data 2**.

Datasets for External Comparison

In order to highlight similarities and differences of our approach with respect to other methods, we compare the scores of our model with scores relative to the same pHLAs reported in a list of selected studies. A peptide missing from the reference proteome ("QSADAQSFLNR") was removed. Only peptides between 8 and 11 amino acids were considered. The peptides arising from the cleavage sites of the ORF1a/ab polyprotein were also removed from the datasets. These sites are defined as nucleic acids within the NCBI reference sequence: NC_045512.2 but outside the range of the ORF1a/ab coding sequences.

The ORF1a and ORF1ab cleavage sites were corrected for reading frameshift which occurs for ORF1ab (as opposed to ORF1a), when independently translating RNA polymerase and nsp11, respectively.

1. Baruah and Bose (2020): Five epitopes from the surface glycoprotein of SARS-CoV-2 and their corresponding HLA class I supertype representative were reported by the authors (Table 1 in the reference publication). Bioinformatics protocols, machine learning methods, and structural analysis were applied in the original paper for the selection of these pHLAs.
2. Lee and Koohy (2020): 19 A*02:01 restricted epitopes were selected applying TCR-specific Position Weight Matrices (PWM) previously published by the authors. The geometric mean of the three scores was used as an estimator for immunogenicity (Tables 4, 5 in the reference publication).
3. Grifoni et al. (2020):
    a. 1st dataset: 386 SARS-CoV-2 CD8+ predicted epitopes were collected (Supplementary Table 6 in the reference publication) and 41 peptides were excluded as a result of our filtering procedure.
    b. 2nd dataset: 28 SARS-CoV-2 CD8+ epitopes mapped to immunodominant SARS-CoV epitopes were selected (Table 5 in the reference publication). One peptide was excluded as a result of our filtering procedure.
4. Gupta et al. (2020): 10 HLA-A*11:01 restricted peptides from the surface glycoprotein of SARS-CoV-2 were selected by the authors (Table 4 in the reference publication). Bioinformatics protocols, machine learning methods, and structural analysis were used for the selection of those pHLAs. A candidate with an optimal docking score is reported.
5. Prachar et al. (2020): 138 peptides with pHLA complex stability measurements performed using Immunotrack's NeoScreens® assay were made available by the authors. A peptide absent in our dataset was excluded from the comparison.
6. Rammensee et al. (2020): 5 HLA class I peptides were used by the authors for the experimental vaccination of self-experimenting healthy volunteers. IFNγ ELISPOT assays for the measurement of CD8+ activation were negative for all these peptides.

[3] https://www.ncbi.nlm.nih.gov/search/all/?term=SARS-CoV-2

[4] https://bigd.big.ac.cn/ncov/release_genome?lang=en

[5] https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/

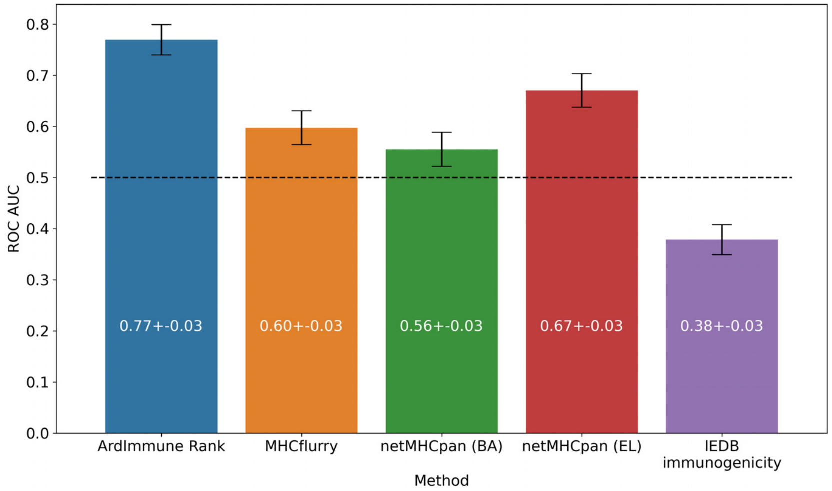

FIGURE 3 | Predictive performance of the selected models on the *Coronaviridae* dataset. ArdImmune Rank, blue bars; MHCflurry, orange bars. netMHCpan, green and red bars for the predicted binding affinity (BA) and ligand likelihood (EL); IEDB immunogenicity, purple bars.

7. Smith et al. (2020): Predictions for ∼615 k peptides were extracted from the Supplementary Table 1 of the reference publication. Approximately 7,600 peptides were excluded as a result of our filtering procedure.

The ArdImmune Rank percentile rank for the pHLAs described in the above datasets was computed for groups of peptides according to their HLA allele. Only pHLAs with a binding affinity percentile rank score < 0.02 (computed using NetMHCpan 4.0) were considered. The predictions were calculated separately for peptides of structural and non-structural origin.

RESULTS

Model Performance

The performance of our method on the test set encompassing *Coronaviridae* epitopes (excl. SARS-CoV-2 epitopes) is shown in **Figure 3**. In addition, the results of our approach are compared to those obtained by other commonly used pHLA binding affinity and pHLA presentation probability predictors, namely netMHCpan 4.0 (Jurtz et al., 2017) and MHCflurry (O'Donnell et al., 2018), as well as the IEDB immunogenicity predictor, version 3.0 (Calis et al., 2013). For both binding affinity tools [MHCflurry and netMHCpan (BA)], the binding affinity predictions in nanomoles (nM) are converted into (0, 1) range with a widely used logarithmic transformation [i.e., first the predictions are bounded from above by 50,000 nM and from below by 1 nM and then transformed with $\left(1 - \frac{\log_{10}x}{log_{10}50{,}000}\right)$. The difference in the predictive performance (measured with ROC AUC) of our model with respect to the other methods is statistically significant (and ranges from 0.10 to 0.39). Moreover (as verified on our training dataset across virus families), the high Pearson correlation between the results produced by the binding predictors (corr. coeff. $\rho = 0.88$) and the low correlation of such results with the predictions of our model ($\rho = 0.45$ and $\rho = 0.53$) demarcate substantial differences between our approach and the approaches based on those methods for predicting immunogenic epitopes (see **Figure 4**).

We apply the LOGO cross-validation scheme according to the procedure described in the Materials and methods section. While we observe a significant variation in ROC AUC scores depending on the tested groups (i.e., virus families), the performance of each method is not correlated with the number of observations within each group. The *Pneumoviridae* family might be an outlier in our dataset as the predictive performance of all the considered models are substantially different for this family than those observed for the other families. Although some groups display a noticeable correlation between pHLA immunogenicity and pHLA binding affinity predictions (e.g., *Pneumoviridae* and *Orthomyxoviridae*), this trend is not confirmed across all groups. The performance (median ROC AUC across virus families) of our method is comparable to those obtained for binding affinity and ligand likelihood predictors, usually with a smaller variance of prediction performance (see **Figures 5**, **6**).

Note that the Coronaviridae dataset (**Figure 3**) is the most relevant dataset to the problem at hand, but it also is a very small dataset containing 67 epitopes. Hence, the variation of performance of the selected methods is expected to be high and their performance on the training set (**Figures 5**, **6**) might be different (note also that in the LOGO validation – in **Figures 5**, **6** – we use a single immunogenicity model instead of 11 models, as in **Figure 3**). On the other hand, evaluation on the Coronaviridae dataset might still reflect performance of the selected methods on the epitopes from the SARS-CoV-2 genome. The dataset encompassing all other virus families used in our LOGO cross-validation procedure (training dataset) is much larger, but is also very heterogeneous. For example the Poxviridae family contains predominantly Vaccinia virus, which is a model organism with mostly non-immunogenic epitopes reported in IEDB. Namely, there are 1.6% immunogenic observations in the Poxviridae family, whereas for Herpesviridae 62% of the

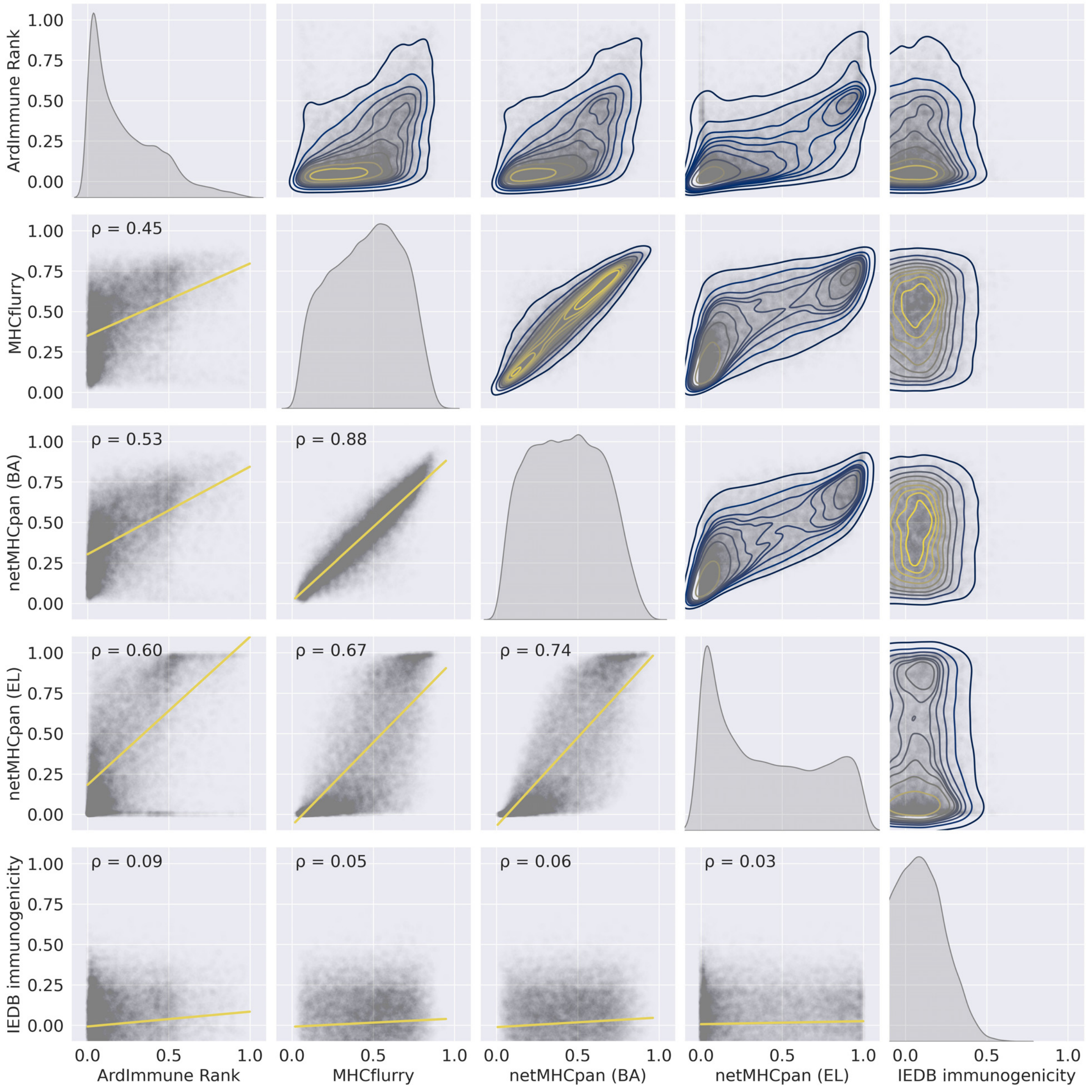

FIGURE 4 | The pairwise relationships between the predictions of the selected models on the training set – (1) ArdImmune Rank, (2) MHCflurry, (3) netMHCpan (BA), (4) netMHCpan (EL), and (5) IEDB immunogenicity. Lower triangle – scatterplots with linear regression models fitted (yellow lines) and Pearson's correlation coefficients (PCC) that measure linear correlations between two variables. Diagonal and upper triangle – the prediction distributions obtained by kernel density estimations (1D-KDE and 2D-KDE, respectively).

observations are immunogenic. Moreover, IEDB observations are very small in size within some families (e.g., Adenoviridae with $N = 58$) and much larger in others (e.g., Poxviridae with $N = 21709$). In such a situation, the large variance of performance of predictive methods when evaluated on different viral families is expected and originates from both the underlying biological and experimental factors, as well as from the small number of observations for some virus families.

The model was then used to predict the immunogenicity of peptides from the SARS-CoV-2 proteome. Target peptides and HLA types considered for the analysis were selected according to the procedure described in the "Selection of peptides" and "Selection of HLA alleles" sections, respectively. A considerable number of peptides with high scores are observed in both structural and non-structural proteins, encompassing different HLA alleles. Structural epitopes are dominated by the

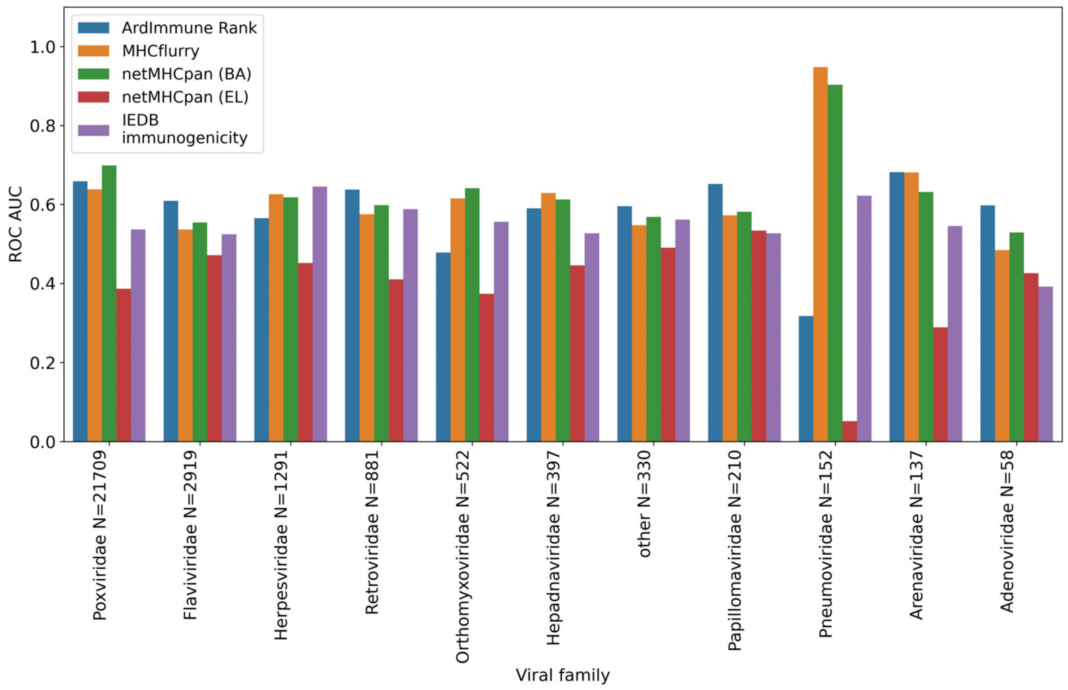

FIGURE 5 | Predictive performance of the selected models obtained in a LOGO cross validation and measured with ROC AUC. ArdImmune Rank, blue bars; MHCFlurry, orange bars; NetMHCpan (BA), green bars; NetMHCpan (EL), red bars; IEDB immunogenicity, purple bars.

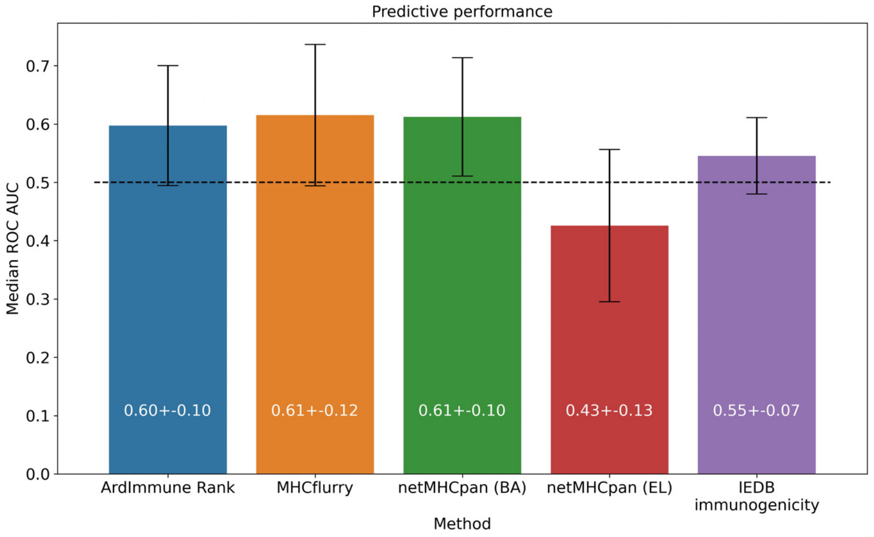

FIGURE 6 | Predictive performance of the selected models, averaged across virus groups in the training dataset.

Spike protein, whereas the non-structural ones mostly originate from the ORF1a/ORF1ab-encoded polyproteins. Peptides with percentile rank ≤ 2 presented across the selected HLAs, were considered for both structural (**Table 3**) and non-structural (**Table 4**) viral proteins. We noticed that some HLA alleles exhibit a large number of highly-ranked peptides, in particular A*02:01, A*11:01, A*24:41 and C*12:03. Interestingly, the presence of some of these alleles was earlier reported to be statistically correlated with the immune protection in SARS cases. Namely, A*02:01 was found to present immunogenic peptides (Ahmed et al., 2020; Lee and Koohy, 2020) whereas A*11:01-restricted epitopes were proposed to be included in a SARS-CoV vaccine by Sylvester-Hvid et al. (2004). Groups of peptides predicted to be associated with multiple HLAs are shown in **Figure 7**. These epitopes originate from both structural and non-structural antigens.

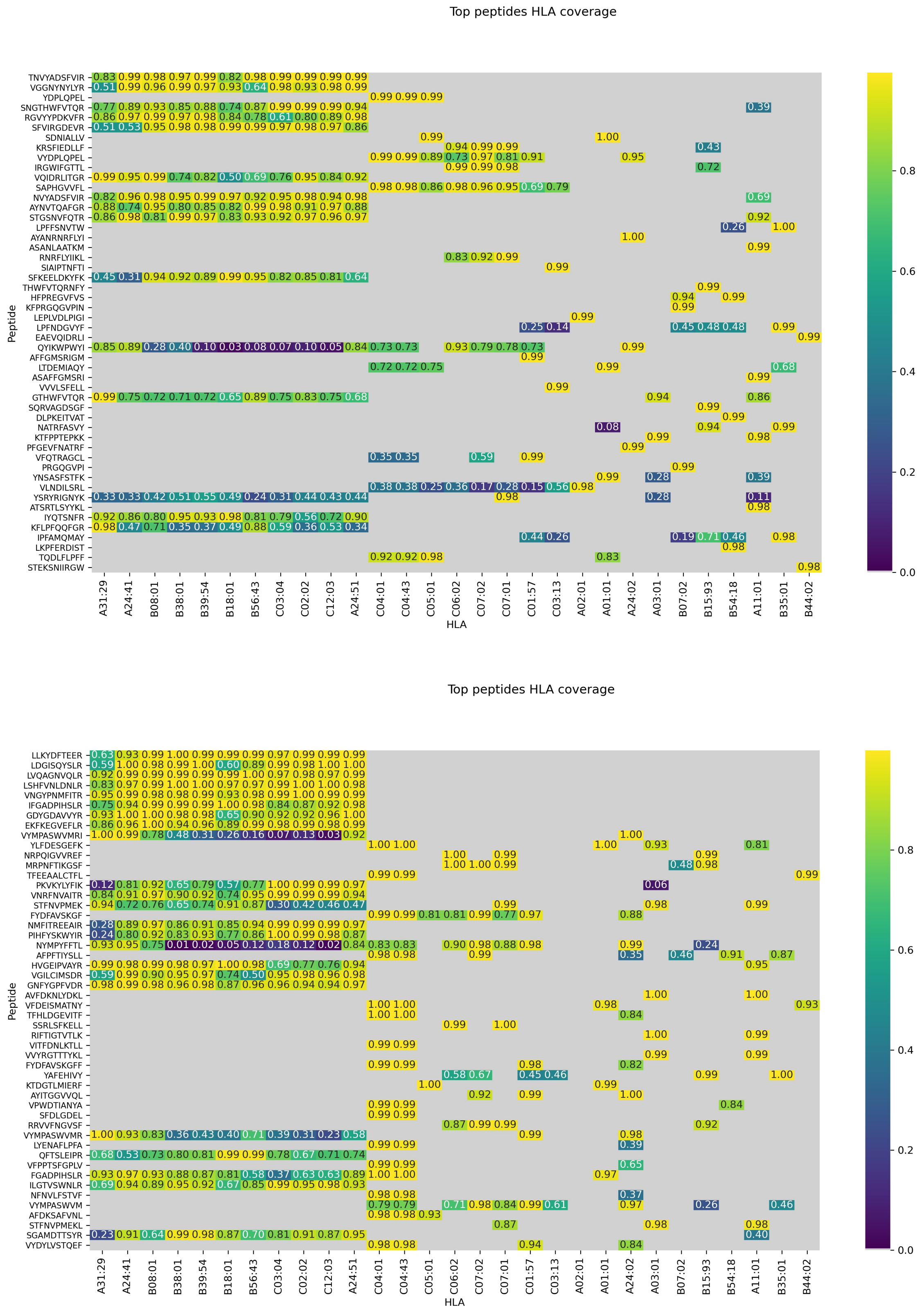

FIGURE 7 | Peptides presented across multiple HLAs. Immunogenicity scores are reported for epitopes from both structural **(top)** and non-structural **(bottom)** proteins. Peptide-HLA combinations marked in gray are predicted non-binders (netMHCpan 4.0 percentile rank > 2). For the remaining pHLAs, the color relates to the percentile rank of our predictions for a given HLA type (0.95 means that the prediction is among top 5% of the predictions for that particular HLA allele).

TABLE 3 | Peptides with ArdImmune Rank percentile rank ≤ 2 obtained from SARS-CoV-2 structural proteins, sorted by (1) the number of HLA types capable of binding and presenting given peptide and (2) the median rank across different HLA types.

No.	Peptide	Prot. start	Prot. end	Protein	HLA% rank ≤ 2	Median HLA%_rank	Max mut. freq
1	TNVYADSFVIR	393	403	S	0.994	A24:41| A24:51| B39:54| C02:02| C03:04| C12:03	0.00012
2	VGGNYNYLYR	445	454	S	0.989	A24:41| A24:51| B38:01| C12:03	0.00013
3	YDPLQPEL	1,138	1,145	S	0.995	C04:01| C04:43| C05:01	0.00049
4	SNGTHWFVTQR	1,097	1,107	S	0.989	C02:02| C03:04| C12:03	0.00012
5	RGVYYPDKVFR	34	44	S	0.983	A24:51| B08:01| B39:54	0.00012
6	SFVIRGDEVR	399	408	S	0.989	B18:01| B56:43| C02:02	0.00012
7	SDNIALLV	214	221	M	0.995	A01:01| C05:01	0.00047
8	KRSFIEDLLF	814	823	S	0.99	C07:01| C07:02	0.00024
9	VYDPLQPEL	1,137	1,145	S	0.989	C04:01| C04:43	0.00049
10	IRGWIFGTTL	101	110	S	0.992	C06:02| C07:02	0.00012
11	VQIDRLITGR	991	1,000	S	0.992	A31:29| B08:01	0.00000
12	SAPHGVVFL	1,055	1,063	S	0.984	C04:01| C04:43	0.00024
13	NVYADSFVIR	394	403	S	0.986	B08:01| B39:54	0.00012
14	AYNVTQAFGR	267	276	N	0.989	B56:43| C03:04	0.00035
15	STGSNVFQTR	637	646	S	0.986	A24:41| B38:01	0.00000
16	LPFFSNVTW	56	64	S	0.996	B35:01	0.00024
17	AYANRNRFLYI	38	48	M	0.995	A24:02	0.00058
18	ASANLAATKM	1,020	1,029	S	0.995	A11:01	0.00024
19	RNRFLYIIKL	42	51	M	0.995	C07:01	0.00023
20	SIAIPTNFTI	711	720	S	0.995	C03:13	0.00024
21	SFKEELDKYFK	1,147	1,157	S	0.994	B18:01	0.00049
22	THWFVTQRNFY	1,100	1,110	S	0.994	B15:93	0.00012
23	HFPREGVFVS	1,088	1,097	S	0.994	B54:18	0.00012
24	KFPRGQGVPIN	65	75	N	0.993	B07:02	0.00035
25	LEPLVDLPIGI	223	233	S	0.992	A02:01	0.00000
26	LPFNDGVYF	84	92	S	0.991	B35:01	0.00049
27	EAEVQIDRLI	988	997	S	0.991	B44:02	0.00000
28	QYIKWPWYI	1,208	1,216	S	0.991	A24:02	0.00024
29	AFFGMSRIGM	313	322	N	0.991	C01:57	0.00071
30	LTDEMIAQY	865	873	S	0.99	A01:01	0.00024
31	ASAFFGMSRI	311	320	N	0.99	A11:01	0.00012
32	VVVLSFELL	510	518	S	0.989	C03:13	0.00013
33	GTHWFVTQR	1,099	1,107	S	0.989	A31:29	0.00012
34	SQRVAGDSGF	184	193	M	0.989	B15:93	0.00000
35	DLPKEITVAT	163	172	M	0.988	B54:18	0.00012
36	NATRFASVY	343	351	S	0.987	B35:01	0.00024
37	KTFPPTEPKK	361	370	N	0.993	A03:01	0.00036
38	PFGEVFNATRF	337	347	S	0.986	A24:02	0.00024
39	VFQTRAGCL	642	650	S	0.986	C01:57	0.00012
40	PRGQGVPI	67	74	N	0.986	B07:02	0.00035
41	YNSASFSTFK	369	378	S	0.986	A01:01	0.00025
42	VLNDILSRL	976	984	S	0.984	A02:01	0.00012
43	YSRYRIGNYK	196	205	M	0.984	C07:01	0.00012
44	ATSRTLSYYKL	171	181	M	0.984	A11:01	0.02876
45	IYQTSNFR	312	319	S	0.983	B18:01	0.00014
46	KFLPFQQFGR	558	567	S	0.983	A31:29	0.00036
47	IPFAMQMAY	896	904	S	0.982	B35:01	0.00000
48	LKPFERDIST	461	470	S	0.982	B54:18	0.00025
49	TQDLFLPFF	51	59	S	0.982	C05:01	0.00292
50	STEKSNIIRGW	94	104	S	0.982	B44:02	0.00073

Peptides marked in red are considered as highly variable (HV) due to maximum mutation frequency score ≥ 0.05.

TABLE 4 | Peptides with model percentile rank ≤ 2 obtained from SARS-CoV-2 non-structural proteins, sorted by (1) the number of HLA types capable of binding and presenting given peptide and (2) the median rank across different HLA types.

No.	Peptide	Prot. start	Prot. end	Protein	HLA% rank ≤ 2	Median HLA%_rank	Max mut. freq
1	LLKYDFTEER	4,662	4,671	ORF1ab	0.991	A24:51| B08:01| B18:01| B38:01| B39:54| B56:43| C02:02| C12:03	0.00012
2	LDGISQYSLR	570	579	ORF1a	0.997	A24:41| A24:51| B08:01| B38:01| B39:54| C03:04| C12:03	0.00372
3	LVQAGNVQLR	3,330	3,339	ORF1a	0.993	A24:41| A24:51| B08:01| B18:01| B38:01| B39:54| B56:43	0.00565
4	LSHFVNLDNLR	2,518	2,528	ORF1a	0.997	A24:51| B08:01| B38:01| B39:54| C02:02| C03:04| C12:03	0.00414
5	VNGYPNMFITR	5,991	6,001	ORF1ab	0.995	A24:41| A24:51| B39:54| C02:02| C03:04| C12:03	0.00036
6	IFGADPIHSLR	1,153	1,163	ORF1a	0.993	B08:01| B18:01| B38:01| B39:54| B56:43	0.00332
7	GDYGDAVVYR	5,527	5,536	ORF1ab	0.997	A24:41| A24:51| B08:01| B38:01| B39:54	0.00084
8	EKFKEGVEFLR	633	643	ORF1a	0.986	A24:51| B08:01| B56:43| C02:02| C03:04	0.00371
9	VYMPASWVMRI	3,653	3,663	ORF1a	0.998	A24:02| A24:41| A31:29	0.00412
10	YLFDESGEFK	906	915	ORF1a	0.995	A01:01| C04:01| C04:43	0.00413
11	NRPQIGVVREF	5,813	5,823	ORF1ab	0.993	B15:93| C06:02| C07:01	0.00024
12	MRPNFTIKGSF	3,393	3,403	ORF1a	0.997	C06:02| C07:01| C07:02	0.00425
13	TFEEAALCTFL	3,174	3,184	ORF1a	0.992	B44:02| C04:01| C04:43	0.00399
14	PKVKYLYFIK	4,223	4,232	ORF1a	0.993	C02:02| C03:04| C12:03	0.00398
15	VNRFNVAITR	5,882	5,891	ORF1ab	0.991	C02:02| C03:04| C12:03	0.00000
16	STFNVPMEK	2,600	2,608	ORF1a	0.989	A03:01| A11:01| C07:01	0.00550
17	FYDFAVSKGF	4,811	4,820	ORF1ab	0.988	C04:01| C04:43| C07:02	0.00048
18	NMFITREEAIR	5,996	6,006	ORF1ab	0.99	C02:02| C03:04| C12:03	0.00060
19	PIHFYSKWYIR	38	48	ORF8	0.988	C02:02| C03:04| C12:03	0.00023
20	NYMPYFFTL	2,167	2,175	ORF1a	0.981	A24:02| C01:57| C07:02	0.00415
21	AFPFTIYSLL	8	17	ORF10	0.98	C04:01| C04:43| C07:02	0.00168
22	HVGEIPVAYR	110	119	ORF1a	0.991	A31:29| B08:01| B18:01	0.00206
23	VGILCIMSDR	5,894	5,903	ORF1ab	0.983	A24:41| A24:51| C02:02	0.00132
24	GNFYGPFVDR	3,442	3,451	ORF1a	0.983	A24:41| A31:29| B08:01	0.00467
25	AVFDKNLYDKL	1,176	1,186	ORF1a	0.998	A03:01| A11:01	0.00386
26	VFDEISMATNY	5,696	5,706	ORF1ab	0.998	C04:01| C04:43	0.00024
27	TFHLDGEVITF	1,543	1,553	ORF1a	0.997	C04:01| C04:43	0.00440
28	SSRLSFKELL	4,755	4,764	ORF1ab	0.996	C06:02| C07:01	0.00012
29	RIFTIGTVTLK	6	16	ORF3a	0.995	A03:01| A11:01	0.01995
30	VITFDNLKTLL	1,550	1,560	ORF1a	0.994	C04:01| C04:43	0.00385
31	VVYRGTTTYKL	5,533	5,543	ORF1ab	0.993	A03:01| A11:01	0.00024
32	FYDFAVSKGFF	4,811	4,821	ORF1ab	0.993	C04:01| C04:43	0.00048
33	YAFEHIVY	6,682	6,689	ORF1ab	0.993	B15:93| B35:01	0.00024
34	KTDGTLMIERF	5,241	5,251	ORF1ab	0.992	A01:01| C05:01	0.00000
35	AYITGGVVQL	599	608	ORF1a	0.991	A24:02| C01:57	0.00427
36	VPWDTIANYA	2,133	2,142	ORF1a	0.991	C04:01| C04:43	0.00401
37	SFDLGDEL	142	149	ORF1a	0.99	C04:01| C04:43	0.00014
38	RRVVFNGVSF	3,163	3,172	ORF1a	0.989	C07:01| C07:02	0.00399
39	VYMPASWVMR	3,653	3,662	ORF1a	0.992	A31:29| C01:57	0.00412
40	LYENAFLPFA	3,606	3,615	ORF1a	0.987	C04:01| C04:43	0.17819
41	QFTSLEIPR	5,910	5,918	ORF1ab	0.987	B18:01| B56:43	0.00060
42	VFPPTSFGPLV	4,712	4,722	ORF1ab	0.986	C04:01| C04:43	0.55016
43	FGADPIHSLR	1,154	1,163	ORF1a	0.999	C04:01| C04:43	0.00332
44	ILGTVSWNLR	1,367	1,376	ORF1a	0.985	C03:04| C12:03	0.00398
45	NFNVLFSTVF	4,704	4,713	ORF1ab	0.985	C04:01| C04:43	0.00012
46	VYMPASWVM	3,653	3,661	ORF1a	0.985	C01:57| C07:02	0.00412
47	AFDKSAFVNL	6,355	6,364	ORF1ab	0.984	C04:01| C04:43	0.00029
48	STFNVPMEKL	2,600	2,609	ORF1a	0.983	A03:01| A11:01	0.00550
49	SGAMDTTSYR	3,218	3,227	ORF1a	0.984	B38:01| B39:54	0.00508
50	VYDYLVSTQEF	3,810	3,820	ORF1a	0.983	C04:01| C04:43	0.00412

Peptides marked in red are considered as Highly Variable (HV) due to maximum mutation frequency score ≥ 0.05.

TABLE 5 | The most frequently mutated positions within the SARS-CoV-2 proteome.

No.	Protein	Protein position	Mutation frequency
1	ORF1ab	4,715	0.5502
2	S	614	0.5478
3	ORF3a	57	0.1789
4	ORF1a	3,606	0.1781
5	N	203	0.1770
6	N	204	0.1765
7	ORF1a	265	0.1646
8	ORF3a	251	0.1439
9	ORF8	84	0.1384
10	ORF1ab	5,865	0.0926
11	ORF1ab	5,828	0.0924
12	ORF1a	765	0.0668
13	ORF1a	739	0.0590

SARS-CoV-2 Genome Diversity Analysis

In order to enable the exclusion of peptides originating from genetically highly variable areas, the mutation frequency of each amino acid within the SARS-CoV-2 genome was computed (see section "Materials and Methods" for details). The genes that those peptides originate from are likely to mutate, hence the inclusion of such peptides might lower the vaccine efficacy over time. From the analysis of 8,639 complete genome sequences, obtained from different SARS-CoV-2 isolates, which then were translated into protein sequences, the mutation frequency at each amino acid position was computed.

For each peptide in the SARS-CoV-2 proteome, the maximum mutation frequency was calculated (see section "Materials and Methods"), and peptides with the resulting score ≥ 0.05 (marked in color in **Tables 3**, **4**) are considered to be highly variable (HV) and should be disregarded as vaccine components. 13 amino acid positions were observed to contain mutations in at least 5% of the selected sequences. Among these, as many as nine amino acid positions were mutated in more than 10% of the selected sequences, while two positions showed mutations in fully half of the samples (more than 50%). In **Table 5** we present the most frequently mutated positions within the SARS-CoV-2 proteome. Mutation frequency values for all positions are available in the **Supplementary Data 2**. Figures presenting distribution of mutation frequency are available in the **Supplementary Data 3**.

Within the top-50 immunogenic peptides originating from the SARS-CoV-2 structural and non-structural proteins (NSPs), 1 and 3 HV peptides were found, respectively.

Toxicity/Tolerance Results

Each peptide derived from the SARS-CoV-2 proteome was studied to ascertain the lack of similarity with peptides present in the reference human proteome. When administered in a vaccine, epitopes highly similar to peptides presented by the host's healthy tissues could either trigger an unwanted immune self-reaction or be tolerated by the immune system. In both cases, these peptides should be eliminated from the vaccine composition. A total of 11 SARS-CoV-2-derived peptides with moderate similarity to human proteins were found (*E*-value $\leq$ 4). Of these, four were significantly similar (*E*-value $\leq$ 1) and thus should be avoided (see **Supplementary Data 1**). None of these peptides were found within the top-100 ranked peptides.

Comparison With Other Methods

Results from a list of selected publications were compared with percentile ranks computed by our method for the same pHLAs. We did not find any significant correlation with the *in silico* predictions from Grifoni et al. (2020), Lee and Koohy (2020), and Gupta et al. (2020) highlighting a clear distinction between our methodology and the procedures used in these studies. Although the best candidate selected by Gupta et al. is not among our best candidates for HLA-A*11:01, it is scored by the model as the top candidate among those proposed by the authors. A moderate negative correlation ($\rho \cong -0.45$) was observed between the percentile rank scores of our method and the scores presented by Smith et al. (2020). Although our top peptide candidates associated with the HLAs proposed by Baruah and Bose (2020) do not include any of the five peptides proposed by the authors, we noticed a consensus between the HLA percentile rank of the pHLAs selected by the authors, and our percentile rank scores (**Figure 8**).

The immunogenicity scores predicted by our model were then compared with the experimental measurement of pHLA binding stability done by Prachar et al. (2020). Peptide candidates with low immunogenicity ranks are enriched in regions with a low stability percentage. The results are shown in **Figure 9**, on the left. The immunogenicity score is expressed as the complement to 100 of the immunogenicity percentile rank. The stability percentage is defined relative to reference peptides (see Prachar et al., 2020 for details). The concordance between high immunogenicity (or low immunogenicity rank) and high stability percentage is more noticeable after the exclusion of peptides with low predicted binding affinity (**Figure 9**, right). The Spearman correlation between pHLA stability percentage and the predicted immunogenicity ($\rho = 0.392$) is higher than the correlation between the stability percentage and the predicted binding affinity ($\rho = 0.313$). The binding affinity was computed using NetMHCpan 4.0 (Jurtz et al., 2017).

A noticeable difference in the distributions of experimentally measured pHLA stability percentage was obtained by ranking using binding affinity predictors and our immunogenicity predictions. A clear distinction between stable and unstable pHLAs was obtained through the selection of the top-10% and the bottom-10% scores predicted by the immunogenicity model, whereas the use of filters relying on standard binding affinity thresholds (e.g., 100 nM) leads to a less defined separation (**Figure 10**).

Finally, we report low scores for all the five class I pHLAs which were experimentally confirmed to be non-immunogenic by Rammensee et al. (2020). None of these peptides were recommended by ArdImmune Rank as a candidate to be included in a vaccine formulation against SARS-CoV-2.

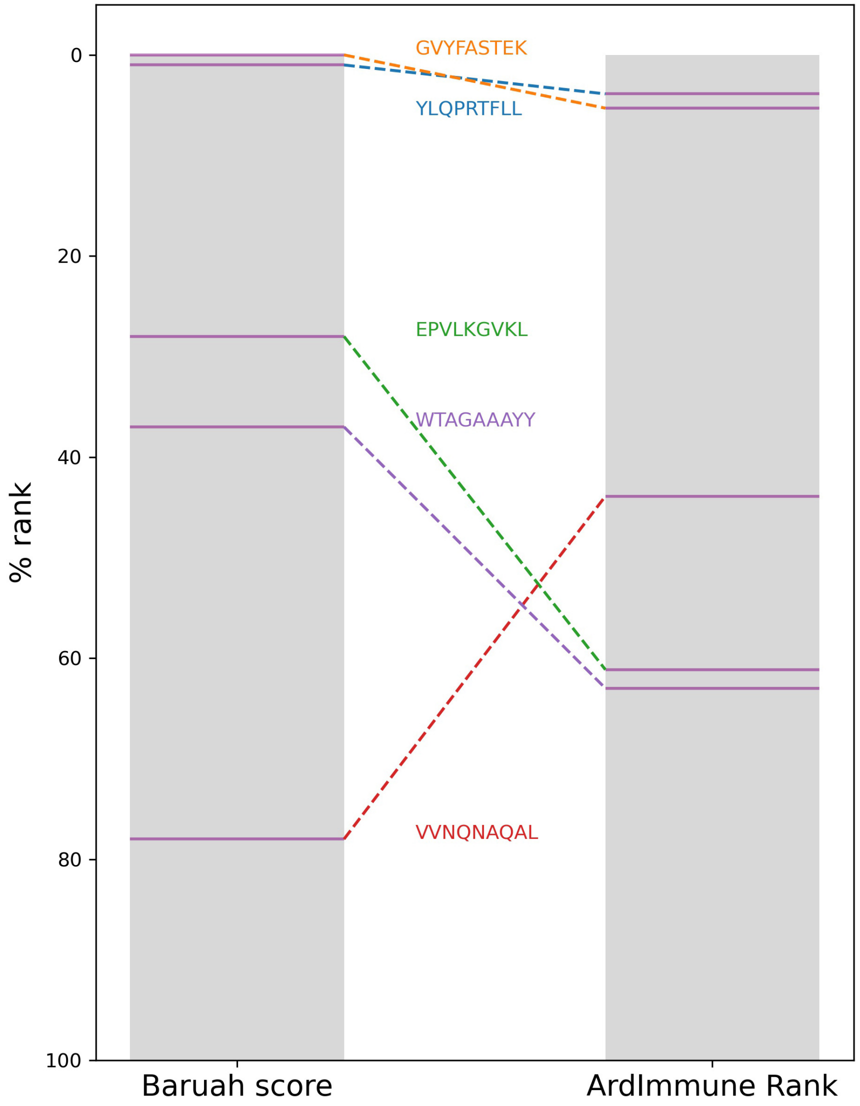

FIGURE 8 | The HLA percentile ranks of the five peptides selected by Baruah et al. as computed from Baruah score and ArdImmune Rank.

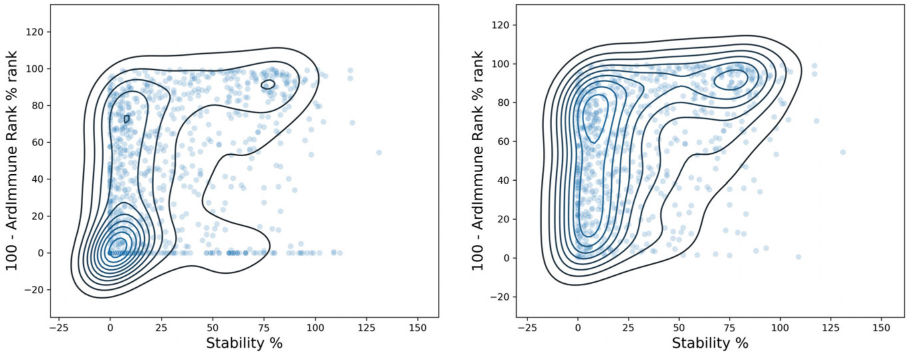

FIGURE 9 | Comparison between ArdImmune Rank percentile ranks for pHLA immunogenicity and pHLA stability data measured by Prachar et al. (2020) Scatter plots and kernel density estimations are shown with **(right)** and without **(left)** the exclusion of pHLA predicted non-binders (K_d percentile rank ≥ 2). The complement of the ArdImmune Rank percentile rank is shown on the y-axis (higher value = lower rank), while the stability percentage as reported by Prachar et al. (2020) is shown on the x-axis.

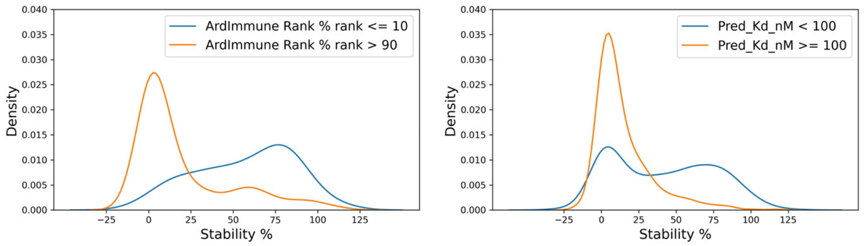

FIGURE 10 | Distribution of stability percentage for different filtering procedures. The respective pHLA stability score densities of the 10% top ranked and the 10% lowest ranked peptides in terms of predicted immunogenicity is shown on the **left**. The pHLA stability score densities computed according to the binding affinity ranges reported by Prachar et al. (2020). ($K_d \geq 100$ nM, $K_d < 100$ nM, based on predicted binding affinity) is shown on the **right**.

DISCUSSION

The high selective pressure exerted upon coronaviruses, caused by the need of a viable host for survival, together with their high genetic variability, facilitates their rapid evolution and the prompt generation of escape mutants. Despite the vigorous effort of the industry, vaccine design, clinical trials, and production require at least several months and most likely several years. Many investigations aimed at developing vaccines protecting humans and animals from coronaviruses were initiated in the last few decades, setting the basis for the recent scientific advancement in COVID-19 treatment. Nonetheless, a limiting aspect associated with the approval and commercialization of a vaccine is that the demand for a vaccine is limited to the outbreak period, and its market value is proportional to the number of people affected. This represented a major issue for the development of vaccines for SARS and MERS (Du et al., 2009; Dhama et al., 2020). In addition, the majority of coronavirus biotherapeutics (i.e., antibodies and vaccines) are designed to leverage neutralizing antibodies directed against the S protein. Safety issues such as those associated with the ADE and CSS events, make the development of vaccine and antibody-based therapies even more problematic.

In combination with the stimulation of humoral immune response, which is aimed at the direct neutralization of the virus, the targeted elimination of infected cells is a crucial element of the immune response against viruses. This might be induced either by the administration of a vaccine eliciting protective CD8+ Cytotoxic T Lymphocyte (CTL) or by transferring CD8+ cells engineered to recognize viral antigens specifically. Previous studies have confirmed a strong correlation between the depletion and exhaustion of T-cells and worse prognosis in critical coronavirus patients (Diao et al., 2020) highlighting the potential of vaccines inducing T-cell responses for COVID-19 prevention. This strategy has beneficial features such as a lower risk of stimulating ADE and CSS with respect to antibody-based strategies (Jaume et al., 2011; Channappanavar et al., 2016) and the stimulation of the immune response against intracellular epitopes not reachable by the antibodies but potentially highly immunogenic. In both cases, the selection of effective immunogenic epitopes is of paramount importance.

The aim of this study was to identify SARS-CoV-2 epitopes for the development of a vaccine composition focused on T-cell activation. We investigated several aspects predetermining whether viral epitopes may induce an effective T-cell response, including the MHC-I peptide presentation and immunogenicity potential, SARS-CoV-2 genome variability, and possible toxicity/immune tolerance of the peptides considered.

In contrast to the majority of works on this topic either relying of pHLA binding and presentation events or modeling single pHLA structural interactions, the model applied herein was designed to leverage simultaneously information about the propensity of a peptide to be presented by its cognated HLA and the probability that such pHLA is immunogenic, inferred from similar experimental data. As we show in **Figure 3** when evaluated on the experimentally-validated *Coronaviridae* immunogenicity data, our approach has higher performance than the widely-used predictors assessing pHLA binding affinity, presentation or immunogenicity (i.e., IEDB).

By applying our method, a considerable amount of highly scored T-cell epitopes was found across the SARS-CoV-2 proteome, encompassing the structural proteins and NSPs, as shown in **Tables 3**, **4**. The majority of selected epitopes were conserved across different SARS-CoV-2 isolates. Only 16 epitopes were excluded because of their significant mutability (see **Table 5**). The availability of epitopes from NSPs allows for the design of vaccine components dedicated to T-cell responses, and might be further integrated with other components focused on B-cell responses. The adoption of such a compartmentalized strategy might help to lower the risk of non-neutralizing antibody production, which constituted a reason of concern during the development of a vaccine formulation for SARS. Moreover, during the early stages of viral infection, the expression of non-structural proteins is significantly higher than the expression of structural ones. The targeted stimulation of the immune response toward epitopes originating from non-structural proteins might be useful to induce an immune response at the early phase of the disease. Some highly ranked peptides were found to be presented across multiple HLAs and could be used to increase population coverage while decreasing the number of epitopes needed to be included in the vaccine formulation. This aspect could be particularly relevant for solutions relying on delivery systems of limited capacity.

The risk of eliciting potentially harmful and sometimes deadly (Linette et al., 2013) cross-reactivities is an issue to be carefully addressed in vaccine design. On the other hand, epitopes shared with proteins from the host could also be tolerated by the host's immune system, being not useful for vaccine purposes. Considering the importance of such an aspect, the analysis of potential toxicity and tolerance was addressed in this study, leading to the identification of four highly ranked epitopes having a certain degree of similarity with proteins within the human proteome. Such peptides were removed for safety and efficacy reasons.

The substantial difference between the selection of pHLA candidates performed by our methodology with respect to those presented by Grifoni et al. (2020), Lee and Koohy (2020), and Gupta et al. (2020) highlights a clear distinction between these approaches. Nonetheless, our method supported the selection of top candidates in small datasets obtained by applying hand-crafted filtering stages (Baruah and Bose, 2020; Gupta et al., 2020). The mild correlation with the results from Smith et al. (2020) might indicate the usage of equivalent components during some steps of the selection process. A relative concordance between the pHLA stability scores from Prachar et al. (2020) and the associated immunogenic scores computed by our method was observed (**Figure 9**). Moreover, we show that the peptide ranks produced by our immunogenicity model have a higher correlation with the experimentally measured pHLA stability than the ranks obtained by methods relying solely on binding affinity or ligand likelihood predictions. This observation is consistent with works reported in the literature (Harndahl et al., 2012). We also obtained low immunogenicity scores for all five peptides which have been experimentally confirmed by Rammensee to be unable to activate CD8+ lymphocytes.

CONCLUSION

In this paper we suggested a SARS-CoV-2 vaccine composition in the form of the list of epitopes optimized for their (predicted) immunogenicity and HLA population coverage. Our motivation is that cellular immune response is fundamental for an effective SARS-CoV-2 vaccine and it also mitigates the risks of ADE and CSS which are typically associated with modalities relying on the activation of humoral immune response. We showed that the predictive model, on which our methodology is based outperforms, on *Coronaviridae* data, other methods used to date for the design of epitope-based vaccines against SARS-CoV-2. Our approach differs from other existing methods and shows a higher correlation with the measured pHLA stability in comparison with methods based solely on binding affinity predictions. The limitations of our method have the same roots as those found in other *in silico* approaches based on predicting various pHLA properties, i.e., the accuracy of these predictive methods. We expect that with the increasing amount of experimentally validated data and with further algorithmic enhancements in the field of artificial intelligence, the accuracy of such models and the effectiveness of vaccine design will continue to improve. Computational methods have proven to be of considerable support in optimizing the vaccine design process on several occasions. Moreover, a notable improvement in the predicting skills of such methods has been recorded in recent years, admittedly due to the increasing advancements in machine learning coupled with a surge in the availability of powerful computational resources. However, it is important to mention that such tools do not represent a substitute for the laboratory experiments necessary to verify and optimize the safety and efficacy of vaccines. Their role is to support the design of such experiments in order to reduce their number, the time needed and cost.

DATA AVAILABILITY STATEMENT

The lists containing the predicted immunogenic peptides with percentile rank ≤ 2 are included in this study (**Tables 3** and **4**). The lists of all the predicted immunogenic peptides generated during this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

GM wrote the article with contributions from IN, PSk, and JK. AM, PSk, IN, and KG performed the analyses and generated figures and tables included in the article. GM, IN, AM, PSk, JK, AS-D, KG, PK, and MD developed the applied methodology. PSt conceived the idea for the project and coordinated the work. AS-D, MS, and KP gave essential contributions to the interpretation of immunological and virological aspects of the study. All the authors reviewed, edited, contributed to the article and approved the submitted version.

FUNDING

The study was sponsored by Ardigen. The applied methodology was in part developed prior to this study with support from the regional Polish grant RPMP.01.02.01-12-0301/17 (European Funds, Regional Program) approved by the Małopolska Centre for Entrepreneurship.

ACKNOWLEDGMENTS

Ardigen and COVID-19 Vaccine Corporation (CVC) announced that they entered a research collaboration aimed at the development of SARS-CoV-2 vaccine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.602196/full#supplementary-material

REFERENCES

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46, 315–326. doi: 10.1016/j.immuni.2017.02.007

Ahmed, S. F., Quadeer, A. A., and McKay, M. R. (2020). Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* 12:254. doi: 10.3390/v12030254

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9

Baruah, V., and Bose, S. (2020). Immunoinformatics-aided identification of T Cell and B Cell Epitopes in the surface glycoprotein of 2019-nCoV. *J. Med. Virol.* 92, 495–500. doi: 10.1002/jmv.25698

Beck, Z., Prohászka, Z., and Füst, G. (2008). Traitors of the immune system—enhancing antibodies in HIV infection: their possible implication in HIV vaccine development. *Vaccine* 26, 3078–3085. doi: 10.1016/j.vaccine.2007.12.028

Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., et al. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 9:e1003266. doi: 10.1371/journal.pcbi.1003266

Channappanavar, R., Fehr, A. R., Vijay, R., Mack, M., Zhao, J., Meyerholz, D. K., et al. (2016). Dysregulated Type I interferon and inflammatory monocyte-macrophage responses cause lethal pneumonia in SARS-CoV-infected mice. *Cell Host Microb.* 19, 181–193. doi: 10.1016/j.chom.2016.01.007

Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K., and Perlman, S. (2014). Virus-specific memory CD8 T cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *J. Virol.* 88, 11034–11044. doi: 10.1128/JVI.01505-14

Chen, H., Hou, J., Jiang, X., Ma, S., Meng, M., Wang, B., et al. (2005). Response of Memory CD8 $^+$ T cells to severe acute respiratory syndrome (SARS) Coronavirus in recovered SARS patients and healthy individuals. *J. Immunol.* 175, 591–598. doi: 10.4049/jimmunol.175.1.591

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel Coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395, 507–513. doi: 10.1016/S0140-6736(20)30211-7

Corapi, W. V., Olsen, C. W., and Scott, F. W. (1992). Monoclonal antibody analysis of neutralization and antibody-dependent enhancement of feline infectious peritonitis virus. *J. Virol.* 66, 6695–6705. doi: 10.1128/JVI.66.11.6695-6705.1992

Cui, J., Li, F., and Shi, Z. (2019). Origin and evolution of pathogenic Coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9

Dejnirattisai, W., Jumnainsong, A., Onsirisakul, N., Fitton, P., Vasanawathana, S., Limpitikul, W., et al. (2010). Cross-reacting antibodies enhance dengue virus infection in humans. *Science* 328, 745–748. doi: 10.1126/science.1185181

Dhama, K., Sharun, K., Tiwari, R., Dadar, M., Malik, Y. S., Singh, K. P., et al. (2020). COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. *Hum. Vacc. Immunotherap.* 16, 1232–1238. doi: 10.1080/21645515.2020.1735227

Di Marco, M., Schuster, H., Backert, L., Ghosh, M., Rammensee, H. G., and Stevanoviæ, 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. doi: 10.4049/jimmunol.1700938

Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., et al. (2020). Reduction and functional exhaustion of T cells in patients with Coronavirus disease 2019 (COVID-19). *Front. Immunol.* 11:827. doi: 10.3389/fimmu.2020.00827

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B. J., and Jiang, S. (2009). The spike protein of SARS-CoV — a target for vaccine and therapeutic development. *Nat. Rev. Microbiol.* 7, 226–236. doi: 10.1038/nrmicro2090

Fan, Y. Y., Huang, Z. T., Li, L., Wu, M. H., Yu, T., Koup, R. A., et al. (2009). Characterization of SARS-CoV-specific memory T cells from recovered individuals 4 years after infection. *Archiv. Virol.* 154, 1093–1099. doi: 10.1007/s00705-009-0409-6

Forni, D., Cagliani, R., Mozzi, A., Pozzoli, U., Al-Daghri, N., Clerici, M., et al. (2016). Extensive positive selection drives the evolution of nonstructural proteins in Lineage C Betacoronaviruses. *J. Virol.* 90, 3627–3639. doi: 10.1128/JVI.02988-15

Fu, Y., Cheng, Y., and Wu, Y. (2020). Understanding SARS-CoV-2-mediated inflammatory responses: from mechanisms to potential therapeutic tools. *Virol. Sin.* 35, 266–271. doi: 10.1007/s12250-020-00207-4

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., and Sette, A. (2020). A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microb.* 27, 671–680.e2. doi: 10.1016/j.chom.2020.03.002

Gupta, E., Mishra, R. K., and Niraj, R. R. K. (2020). Identification of potential vaccine candidates against *SARS-CoV-2*, a step forward to fight novel Coronavirus 2019-NCoV: a reverse vaccinology approach. *bioRxiv* [Preprint], doi: 10.1101/2020.04.13.039198

Guzman, M. G., Alvarez, M., Rodriguez-Roche, R., Bernardo, L., Montes, T., Vazquez, S., et al. (2007). Neutralizing antibodies after infection with dengue 1 virus. *Emerg. Infect. Dis.* 13, 282–286. doi: 10.3201/eid1302.060539

Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., Sørensen, M., Nielsen, M., et al. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity: antigen processing. *Eur. J. Immunol.* 42, 1405–1416. doi: 10.1002/eji.201141774

Hohdatsu, T., Yamada, M., Tominaga, R., Makino, K., Kida, K., and Koyama, H. (1998). Antibody-dependent enhancement of feline infectious peritonitis virus infection in feline alveolar macrophages and human monocyte cell line U937 by Serum of cats experimentally or naturally infected with feline Coronavirus. *J. Veter. Med. Sci.* 60, 49–55. doi: 10.1292/jvms.60.49

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., et al. (2018). Genomic characterization and infectivity of a novel SARS-like Coronavirus in Chinese Bats. *Emerg. Micro. Infect.* 7, 1–10. doi: 10.1038/s41426-018-0155-5

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel Coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Iwasaki, A., and Yang, Y. (2020). The potential danger of suboptimal antibody responses in COVID-19. *Nat. Rev. Immunol.* 20, 339–341. doi: 10.1038/s41577-020-0321-6

Jaume, M., Yip, M. S., Cheung, C. Y., Leung, H. L., Li, P. H., Kien, F., et al. (2011). Anti-Severe acute respiratory syndrome Coronavirus spike antibodies trigger infection of human immune cells via a PH- and cysteine protease-independent Fc R pathway. *J. Virol.* 85, 10582–10597. doi: 10.1128/JVI.00671-11

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. doi: 10.4049/jimmunol.1700893

Kam, Y. W., Kien, F., Roberts, A., Cheung, Y. C., Lamirande, E. W., Vogel, L., et al. (2007). Antibodies against Trimeric S glycoprotein protect hamsters against SARS-CoV challenge despite their capacity to mediate FcγRII-dependent entry into B Cells in vitro. *Vaccine* 25, 729–740. doi: 10.1016/j.vaccine.2006.08.011

Katzelnick, L. C., Gresh, L., Halloran, M. E., Mercado, J. C., Kuan, G., Gordon, A., et al. (2017). Antibody-dependent enhancement of severe dengue disease in humans. *Science* 358, 929–932. doi: 10.1126/science.aan6836

Lee, C. H., and Koohy, H. (2020). In silico identification of vaccine targets for 2019-NCoV. *F1000Research* 9:145. doi: 10.12688/f1000research.22507.2

Li, F. (2016). Structure, function, and evolution of Coronavirus spike proteins. *Ann. Rev. Virol.* 3, 237–261. doi: 10.1146/annurev-virology-110615-042301

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., et al. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* 26, 842–844. doi: 10.1038/s41591-020-0901-9

Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in Myeloma and melanoma. *Blood* 122, 863–871. doi: 10.1182/blood-2013-03-490565

Liu, J., Sun, Y., Qi, J., Chu, F., Wu, H., Gao, F., et al. (2010). The membrane protein of severe acute respiratory syndrome Coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. *J. Infect. Dis.* 202, 1171–1180. doi: 10.1086/656315

Liu, L., Wei, Q., Lin, Q., Fang, J., Wang, H., Kwok, H., et al. (2019). Anti-Spike IgG causes severe acute lung injury by skewing macrophage responses during acute SARS-CoV infection. *JCI Insight* 4:e123158. doi: 10.1172/jci.insight.123158

Ng, O. W., Chia, A., Tan, A. T., Jadi, R. S., Leong, H. N., Bertoletti, A., et al. (2016). Memory T cell responses targeting the SARS Coronavirus persist up to 11 years post-infection. *Vaccine* 34, 2008–2014. doi: 10.1016/j.vaccine.2016.02.063

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. doi: 10.1016/j.cels.2018.05.014

Ogishi, M., and Yotsuyanagi, H. (2019). Quantitative prediction of the landscape of T Cell epitope immunogenicity in sequence space. *Front. Immunol.* 10:827. doi: 10.3389/fimmu.2019.00827

Pagès, A., and DebRoy, G. (2020). *Biostrings: Efficient Manipulation of Biological Strings. R Package Version 2.56.0.*

Pahl, J. H. W., Kwappenberg, K. M. C., Varypataki, E. M., Santos, S. J., Kuijjer, M. L., Mohamed, S., et al. (2014). Macrophages inhibit human osteosarcoma cell growth after activation with the bacterial cell wall derivative Liposomal Muramyl tripeptide in combination with Interferon-γ. *J. Exper. Clin. Cancer Res.* 33:27. doi: 10.1186/1756-9966-33-27

Peiris, J. S. M., Chu, C. M., Cheng, V. C. C., Chan, K. S., Hung, I. F. N., Poon, L. L. M., et al. (2003). Clinical progression and viral load in a community outbreak of Coronavirus-associated SARS pneumonia: a prospective study. *Lancet* 361, 1767–1772. doi: 10.1016/S0140-6736(03)13412-5

Peng, H., Yang, L. T., Wang, L. Y., Li, J., Huang, J., Lu, Z. Q., et al. (2006). Long-lived memory T lymphocyte responses against SARS Coronavirus nucleocapsid Protein in SARS-recovered patients. *Virology* 351, 466–475. doi: 10.1016/j.virol.2006.03.036

Prachar, M., Justesen, S., Steen-Jensen, D. B., Thorgrimsen, S., Jurgons, E., Winther, O., et al. (2020). COVID-19 vaccine candidates: prediction and validation of 174 SARS-CoV-2 epitopes. *bioRxiv* [Preprint], doi: 10.1101/2020.03.20.000794

Rammensee, H. S., Stevanoviæ, S., Gouttefangeas, C., Heidu, S., Klein, R., Preuß, B., et al. (2020). Designing a therapeutic SARS-CoV-2 T-cell-inducing vaccine for high-risk patient groups. *bioRxiv* [Preprint], doi: 10.21203/rs.3.rs-27316/v1

Sarkizova, S., Klaeger, S., Le, P. M., Li, L. W., Oliveira, G., Keshishian, H., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* 38, 199–209. doi: 10.1038/s41587-019-0322-9

Smith, C. C., Entwistle, S., Willis, C., Vensko, S., Beck, W., Garness, J., et al. (2020). Landscape and selection of vaccine epitopes in SARS-CoV-2. *bioRxiv* [Preprint], doi: 10.1101/2020.06.04.135004

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., et al. (2019). From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses* 11:59. doi: 10.3390/v11010059

Sylvester-Hvid, C., Nielsen, M., Lamberth, K., Roder, G., Justesen, S., Lundegaard, C., et al. (2004). SARS CTL vaccine candidates; HLA Supertype-, genome-wide scanning and biochemical validation. *Tissue Antig.* 63, 395–400. doi: 10.1111/j.0001-2815.2004.00221.x

Takada, A., Feldmann, H., Ksiazek, T. G., and Kawaoka, Y. (2003). Antibody-dependent enhancement of ebola virus infection. *J. Virol.* 77, 7539–7544. doi: 10.1128/JVI.77.13.7539-7544.2003

Takada, A., Watanabe, S., Okazaki, K., Kida, H., and Kawaoka, Y. (2001). Infectivity-enhancing antibodies to ebola virus glycoprotein. *J. Virol.* 75, 2324–2330. doi: 10.1128/JVI.75.5.2324-2330.2001

Tang, F., Quan, Y., Xin, Z. T., Wrammert, J., Ma, M. J., Lv, H., et al. (2011). Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. *J. Immunol.* 186, 7264–7268. doi: 10.4049/jimmunol.0903490

Tsao, Y. P., Lin, J. Y., Jan, J. T., Leng, C. H., Chu, C. C., Yang, Y. C., et al. (2006). HLA-A*0201 T-cell epitopes in severe acute respiratory syndrome (SARS) Coronavirus nucleocapsid and spike proteins. *Biochem. Biophys. Res. Commun.* 344, 63–71. doi: 10.1016/j.bbrc.2006.03.152

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The Immune Epitope Database (IEDB), 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006

Wan, S., Xiang, Y., Fang, W., Zheng, Y., Li, B., Hu, Y., et al. (2020). Clinical features and treatment of COVID-19 patients in northeast Chongqing. *J. Med. Virol.* 92, 797–806. doi: 10.1002/jmv.25783

Wan, Y., Shang, J., Sun, S., Tai, W., Chen, J., Geng, Q., et al. (2019). Molecular mechanism for antibody-dependent enhancement of coronavirus entry. edited by tom Gallagher. *J. Virol.* 94:e02015-19.

Wang, S. F., Tseng, S. P., Yen, C. H., Yang, J. Y., Tsao, C. H., Shen, C. W., et al. (2014). Antibody-dependent SARS Coronavirus infection is mediated by antibodies against spike proteins. *Biochem. Biophys. Res. Commun.* 451, 208–214. doi: 10.1016/j.bbrc.2014.07.090

Wang, Y. D., Sin, W. Y. F., Xu, G. B., Yang, H. H., Wong, T. Y., Pang, X. W., et al. (2004). T-Cell Epitopes in severe acute respiratory syndrome (SARS) Coronavirus spike protein elicit a specific T-Cell immune response in patients who recover from SARS. *J. Virol.* 78, 5612–5618. doi: 10.1128/JVI.78.11.5612-5618.2004

Whitehead, S. S., Blaney, J. E., Durbin, A. P., and Murphy, B. R. (2007). Prospects for a dengue virus vaccine. *Nat. Rev. Microbiol.* 5, 518–528. doi: 10.1038/nrmicro1690

Willey, S., Aasa-Chapman, M. M. I., O'Farrell, S., Pellegrino, P., Williams, I., Weiss, R. A., et al. (2011). Extensive complement-dependent enhancement of HIV-1 by autologous non-neutralising antibodies at early stages of infection. *Retrovirology* 8:16. doi: 10.1186/1742-4690-8-16

Wright, E. S. (2015). DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinform.* 16:322. doi: 10.1186/s12859-015-0749-z

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome composition and divergence of the novel Coronavirus (2019-NCoV) originating in China. *Cell Host Microb.* 27, 325–328. doi: 10.1016/j.chom.2020.02.001

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new Coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi: 10.1038/s41586-020-2008-3

Zhang, X. W. (2013). A Combination of Epitope prediction and molecular docking allows for good identification of MHC class I restricted T-Cell epitopes. *Comput. Biol. Chem.* 45, 30–35. doi: 10.1016/j.compbiolchem.2013.03.003

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new Coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel Coronavirus from patients with pneumonia in China, 2019. *New Engl. J. Med.* 382, 727–733. doi: 10.1056/NEJMoa2001017

Conflict of Interest: GM, IN, AM, PSk, JK, AS-D, KG, PK, MD, and PSt are employees at Ardigen or were in the past.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.