Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: November 30, 2023

Citation: Lisziewicz J, Lisziewicz R., et al., 2023. HLA-genotype-based Predictive Diagnosis of T-cell Responses to SARS-CoV-2 Infection Powered by Machine Learning. Medical Research Archives, [online] 11(11). https://doi.org/10.18103/mra.v11i11.4533

DOI: https://doi.org/10.18103/mra.v11i11.4533

ISSN: 2375-1924

RESEARCH ARTICLE

HLA-genotype-based Predictive Diagnosis of T-cell Responses to SARS-CoV-2 Infection Powered by Machine Learning

Robert G. Lisziewicz[1], Felipe Oviedo[2], Andras Szasz[1], Juan Lavista Ferres[2], Franco Lori[1,3], Julianna Lisziewicz[1*]

[1]VERDI Solutions; Vienna, Austria
[2]Microsoft AI for Good Research Laboratory; Redmond, USA
[3]Research Institute for Genetic and Human Therapy; Milano, Italy

*julianna.lisziewicz@verdisolution.com

ABSTRACT

Background: The COVID-19 pandemic has necessitated the development of efficient diagnostic tools to predict T-cell responses, which are crucial for viral clearance and protection against reinfection. Current diagnostic tests lack the ability to predict the epitope repertoire of an individual that induces T-cell responses.

Methods: We developed VERDI, a new machine learning-based diagnostic tool that leverages the sequence data of all the six HLA class I alleles of an individual to rank all putative epitopes based on their potential to induce T-cell responses. VERDI was trained on a comprehensive clinical dataset of 920 SARS-CoV-2 epitopes and validated using an independent dataset collected for the FDA-approved T-detect COVID test. We compared VERDI's performance with existing HLA-allele-based models through statistical analyses.

Results: Our findings reveal that VERDI's top-ranked epitopes accurately represent the individual's epitope repertoire that participates in T-cell responses. VERDI outperformed current models, improving T-cell response prediction recall by threefold and precision by eightfold. It exhibited exceptional diagnostic accuracy, precision, and recall in predicting the potency of the top 20 epitopes. Despite experimental limitations that allow testing of only 1% of putative epitopes, VERDI accurately predicted 30% of these, implying a potentially higher accuracy if broader testing were feasible. Notably, the mean potency of the top-ranked epitopes predicted by VERDI, which reflects the strength of an individual's SARS-CoV-2-specific T-cell responses, exhibited a Gaussian distribution.

Conclusions: VERDI is the first diagnostic tool that uses the complete HLA genotype data to predict the breadth and strength of an individual's T-cell responses to SARS-CoV-2 infection. Its ability to accurately identify the potency of epitopes involved in individual T-cell responses and its superior performance compared to the state-of-the-art make it a new resource for personalized vaccine design and disease management.

Introduction

The global outbreak of SARS-CoV-2 has led to widespread hospitalizations and fatalities, especially among individuals with unprepared immune systems. T-cells, especially CD8+ T-cells, have emerged as crucial components in the immune response to the infection, capable of eliminating infected cells even in the absence of protective antibodies. [1,2] While effective vaccines have been developed to induce both antibody and T-cell responses, the continuous mutation of SARS-CoV-2 and the decline of antibody responses over time have underscored the importance of T-cells as the primary defense against severe illness and death. [3,4]

T-cell responses are triggered by specific antigens displayed by Human Leukocyte Antigens (HLA). Following infection, viral proteins are processed into peptides, a subset of which, known as epitopes, bind to HLA class I molecules and are transported to the cell surface to activate T-cell receptors (TCRs). Over recent years, the identification of more than 1,400 epitopes stemming from SARS-CoV-2 has revealed their capacity to induce T-cell responses within at least one out of 1,197 individuals. [5]

However, accurately diagnosing the specific epitopes and their potency in triggering T-cell responses in an individual remains a significant challenge.[6] Experimental methods can test only a few percentage of putative epitopes due to specimen limitations, and the reproducibility of antigen-specific T-cell responses is inconsistent across different individuals. [7] Existing machine learning models designed to predict epitope-HLA and epitope-TCR pairs are most accurate for common HLAs. However, even accurately predicted HLA-allele-binding epitopes rarely induce T-cell responses in HLA-allele-matched individuals. [5,8,9] This discrepancy highlights the unique specificity and potency of T-cell responses in each individual and underscores the need for prediction tools that are not only trained but also validated at the individual level, a feature currently lacking in existing models.

We addressed these challenges by developing a machine learning model named VERDI (Vaccine Epitopes Ranked by Digital Intelligence). VERDI models our hypothesis that T-cells respond to high-density epitopes cooperatively presented by autologous HLA molecules on an individual's cell surface. TCRs are repeatedly stimulated by epitopes on the cell surface that share a common ligand ("core") capable of TCR recognition. The VERDI model addresses the three-phase puzzle by ranking all putative epitopes based on their potential to induce T-cell responses at the individual level: epitope processing and presentation on the cell surface by HLA, TCR recognition of the epitope, and stimulation of the T-cell response.

Substantial clinical data was used to test and validate the VERDI model, leading to a significant discovery - the epitopes that rank highest within the VERDI system directly correspond to the pivotal T-cell antigens. Notably, these top-tier epitopes collectively serve as a defining biomarker of SARS-CoV-2-specific T-cell responses in individuals.

This study provides a comprehensive understanding of the role of T-cell responses in SARS-CoV-2 infection and explores the potential of machine learning in predicting these responses. It contributes to the ongoing debate on the effectiveness of current diagnostic methods and offers a novel predictive diagnosis of T-cell responses, thereby addressing a significant gap in the field.

Methods

Cohorts. Two cohorts were used for developing VERDI - the ABF training cohort and the Adaptive validation cohort. The ABF training and cross-validation cohort is comprised of 79 individuals characterized with a 4-digit genotype of 6 HLA class I alleles representing the sequence of the epitope-binding pocket of the HLA. [10] NetMHCpan4.1 prediction was employed to select 2,204 test epitopes with strong binding to one or more dominant HLA class I alleles of the study subjects. [9] T-cell responses were measured with labeled peptide-MHC-I multimers, which quantify the potency of $CD8^+$ T-cell antigens as "log-fold-change" compared to no antigen or baseline indicating the expansion of antigen-specific T-cells after SARS-CoV-2 infection in the body. We trained VERDI with all the published potency data of an average of 920 epitopes per HLA-allele-matched individuals.

The Adaptive validation cohort comprised 114 individuals characterized by a complete 4-digit HLA-genotype. [11] 545 distinct HLA class I binding epitopes was predicted with the NetMHCpan4.1 to strongly bind to one HLA-allele of the test subject. An average of 169 epitopes per individual were tested with the TCR sequencing method that quantifies the potency of T-cell antigens as "hits" compared to the no antigen control (data was accessed at https://clients.adaptivebiotech.com/pub/covid-2020). A "hit" shows the number of copies of each TCR sequence in the sample that quantifies the expansion of an antigen-specific T-cell clone after SARS-CoV-2 infection in the individual. Since several experiments tested pools of peptides, we postulated that the T-cell responses were dominated by one epitope in the pool having the strongest binding affinity to the dominant HLA-allele of the subject, resulting in an average of 51 tested epitopes per subject.

The complete HLA-genotype of individuals from the European SDY614 (1,061 subjects) and US SDY28 (1,092 subjects) populations was accessed at https://www.immport.org/shared/home.

VERDI model. VERDI was trained and validated with clinical dataset of SARS-CoV-2-specific T-cell responses from individuals that included information on the HLA genotypes of these individuals as well as the epitopes that were recognized by their T-cells. Using the HLA genotype data VERDI generates a comprehensive list of predicted epitopes ranked by their potency to trigger T cells (See the Supplementary section for details).

Statistical evaluation. To evaluate the binary classification performance of the models, five-fold cross-validation was performed at the antigen-individual level, and the receiver operating characteristic curve (ROC) and precision-recall curve (PR) were computed on the test splits. To evaluate the diagnostic function of VERDI at the individual level, the top-N epitopes with the highest T-cell response per individual were identified. The ranked accuracy (Eq. S.3) was defined as the fraction of top-N antigens predicted correctly for an individual. Precision and recall were also defined at the individual level (Eqs. S.4 and S.5). When only positive T-cell responses were reported in the validation dataset, the "PU-metric" was employed for model comparison (Eq. S.6), which is commonly used in positive-unlabeled learning. The statistical evaluation was summarized in the manuscript.

Results

Comparison of HLA-allele and HLA-genotype-based models of T-cell responses

To model T-cell response prediction, we used the ABF cohort which is the largest published experimental dataset on epitope-specific T-cell responses in HLA-genotyped individuals. [10] As a baseline model following current practices, we used the dominant HLA-allele-based prediction of T-cell responses and selected the highest EL score (EL Max) for every epitope of each subject predicted by NetMHCpan 4.1.[9] The ROC-AUC 0.57 and PR-AUC 0.0135 performance demonstrated that this EL Max model is just marginally better than random guessing to predict the epitopes that induced T-cell responses at the individual level. (**Figures 2 a, b)** Indeed, in the ABF cohort, only 5% of the predicted "strong" HLA-allele-binding epitopes induced T-cell responses in HLA-allele-matched subjects.

To address the diagnostic challenges of T-cell responses, we hypothesized that (i) epitopes may be transported to the cell surface by each autologous HLA molecule albeit to a different extent, and (ii) T-cells may be stimulated by overlapping epitopes having the same "*core*", which is the TCR binding site.[12,13] To model our HLA-genotype-based hypothesis, we developed VERDI, a cloud-based predictive diagnostic test, where the input is the 4-digit HLA genotype of the subject. VERDI reports all putative epitopes ranked by their potency to trigger the T-cells of the HLA-genotyped individual (**Figure 1**). It uses a balanced random forest (B-RF) as a downstream machine learning model that is learned from the HLA-genotype and the epitope-specific T-cell response potency data measured in SARS-CoV-2-infected individuals. Since the model is trained on the ranked probability of epitopes triggering a TCR in the test subject, the output (probability of an epitope being immunogenic) is interpreted as the potency of an epitope to induce T-cell responses in the individual.

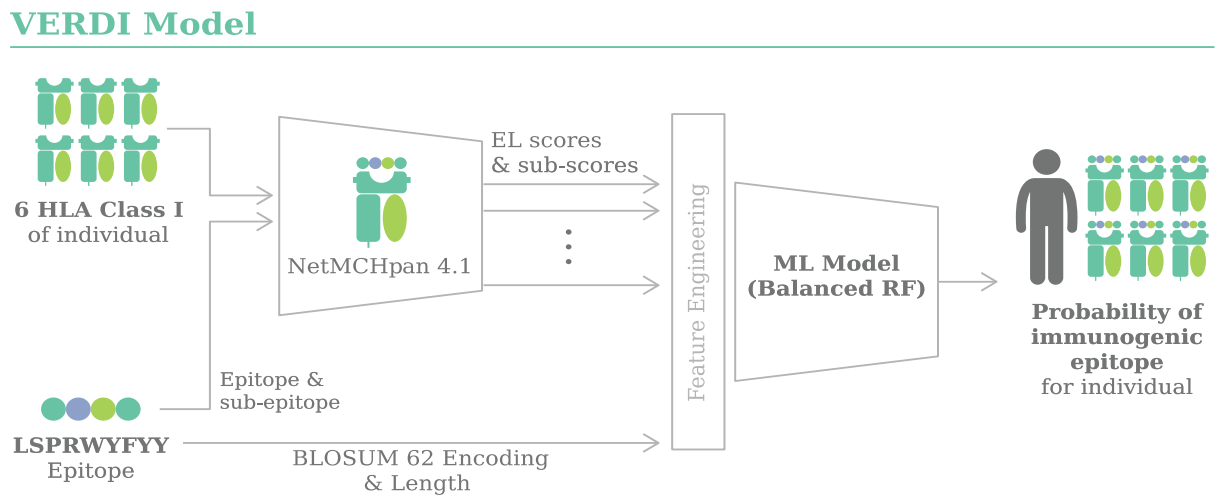


Figure 1: Prediction of T-cell responses of an HLA-genotyped individual with VERDI. 6 HLA class I (input): The test subject's 4-digit HLA class I alleles includes the sequences of the epitope-binding domains of the 6 HLA-alleles. Epitope: VERDI creates all the putative epitopes from input protein sequences and engineers the individualized input features described in the **Supplementary section.**

To compare the performance of the two models, we performed grouped 5-fold cross-validation on the ABF dataset. The epitope-individual ROC and PR curves and the corresponding AUCs showed that the HLA-genotype-based VERDI improved 3-fold the recall and 7.7-fold the precision of the prediction of T-cell responses compared to the HLA-allele-based EL Max model. **(Figures 2 a, b)**

To evaluate the diagnostic performance of VERDI to predict the antigen repertoire of individuals involved in T cell responses to SARS-CoV-2 infection, we computed the top-N ranked accuracy, precision, and recall for the ABF dataset as described in the Supplementary information. The results revealed that the mean ranked accuracy across individuals for VERDI was above 40%, whereas for EL Max it was below 10% (**Figure 2c**). These findings demonstrate the superior diagnostic capability of VERDI in predicting T-cell responses at the individual level. Additionally, VERDI outperformed EL Max in all the evaluation metrics of individualized diagnostics, including accuracy, precision, and recall distribution of the top-20 epitopes, although there was a large spread across individuals due to the limited coverage of certain HLA-alleles in the training dataset (**Figures 2d-2f**). These results suggest that the genetics of both the virus and the individual's HLA-genotype govern SARS-CoV-2-induced T-cell responses, making it impossible to predict these responses accurately at the HLA-allele or population level. Therefore, VERDI represents the first-ever HLA-genotype-based individualized diagnostic test for T-cell responses, offering new opportunities for personalized medicine in infectious disease management.

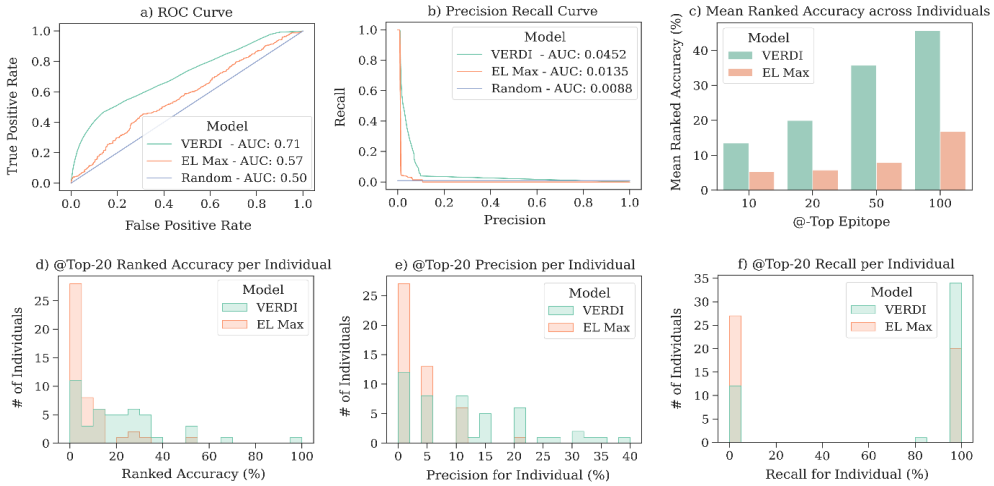


Figure 2. Performance of VERDI to predict the specificity and potency of T-cell antigens of an individual compared to the EL Max (state-of-the-art) and Random baseline models. a) Individual-epitope responses for Receiving Operating Characteristic (ROC) curves for each model across all test folds, along with mean AUCs. b) Precision-Recall (PR) curves for each model across all test folds, along with mean AUCs. Both ROC and PR curves are computed by predicting T-cell responses at the epitope-individual level on the test folds. c) Mean Ranked Accuracy across individuals for the EL Max and VERDI models. d)-f) Distribution of at @Top-20 ranked accuracy, precision, and recall across individuals. (For precision and accuracy calculations we use the same decision threshold, but we explored other thresholds for the model, all yielding similar results.) Subfigures a and b are computed across all epitopes and individuals in the dataset. Subfigures c-f are computed using aggregated metrics across individuals as defined in the Methods section.

Validation of VERDI with clinical data used for FDA approval of the T-detect COVID test

To evaluate the clinical performance of VERDI in individuals outside the ABF cohort, we used an independent dataset collected for the development of the first T-cell response test approved by the FDA.[14] Compared to the ABF dataset, the Adaptive dataset encompassed not only different epitopes but also different cohorts of individuals living in different parts of the world (US vs. Denmark), different methods to quantify T-cells (TCR sequencing vs. multimer staining), and tested a different number of SARS-CoV-2 epitopes per subject (mean of 51 vs. 920).

VERDI ranked all 70,000 SARS-CoV-2-derived putative epitopes for each individual in the Adaptive cohort and we evaluated a smaller top-ranked dataset per individual since fewer epitopes were tested by Adaptive. VERDI predicted the top N=1,2,3 with an average accuracy of around 30% across individuals and outperformed the EL Max model for the most restrictive case predicting the most potent T-cell antigens. (**Figure 3a)** However, it appeared that EL Max outperformed VERDI as we relaxed the ranking requirements for larger N rankings. Adaptive reported only the positive T-cell responses on epitopes predicted to be immunogenic, and consequently, a potential excess of false positives predicted by EL Max was not adequately penalized. To correctly control for false positives, we used the "PU-Metric" from positive-unlabeled ML and observed that VERDI is superior to the EL Max model for all top-N values of interest. (**Figure 3b**) Importantly, VERDI predicted at least 1 of the 3 most potent T-cell antigens identified by the FDA-approved T-cell response diagnostic for most subjects. (**Figure 3c**) The PU-Metric distribution indicates that VERDI has a good predictive performance in most individuals. (**Figure 3d**) The validation of this challenging out-of-distribution dataset confirmed that VERDI provides a better predictive diagnosis for T-cell responses than the EL Max.

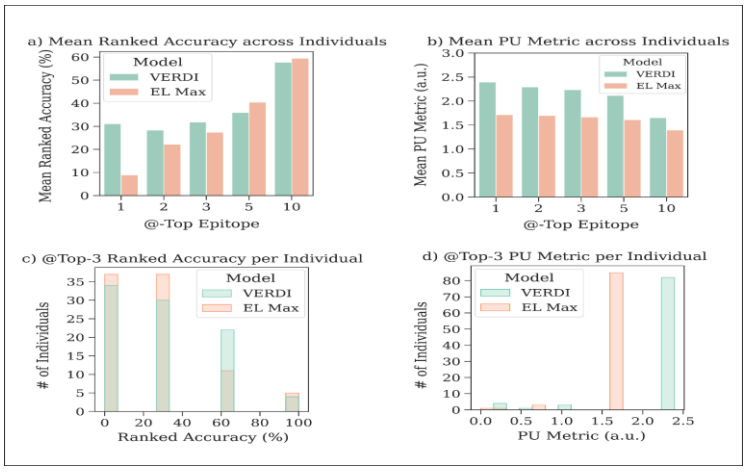


Figure 3. Clinical validation of VERDI in diagnosing T-cell responses to SARS-CoV-2 in the individuals of the Adaptive cohort. EL Max: state-of-the-art prediction; VERDI: our ML model trained with clinical data of individuals in the ABF cohort. a) Mean Ranked Accuracy at various top-N epitopes. Lower N values compared to Figure 2 are used, given the limited number of reported epitopes per patient in the Adaptive Cohort. b) PU metric corrects the limitations of the Mean Ranked Accuracy, as explained in the methods sections. c) Distribution of @Top-3 Ranked Accuracy per Individual for VERDI and EL Max model. d) Distribution of @Top-3 PU Metric per Individuals for VERDI and EL Max model.

Diagnosing the breadth and strength of SARS-CoV-2-specific T-cell responses of individuals

The top-50 epitopes ranked by VERDI include the T-cell antigens responding to SARS-CoV-2 infection since individuals respond to an average of 30 epitopes.[15] We evaluated the T-cell epitope repertoire in 2,346 individuals and found that immunogenic epitopes (antigens) spread throughout the SARS-CoV-2 proteome with extreme variability. Only a small subset of antigens was shared by a subset of individuals, and no single antigen was immunogenic in all individuals.

Our observations revealed a surprising consistency in the potency of the T-cell antigen repertoire within the same individual **(Figure 4 a, b).** We quantified the mean potency of the top-50 T-cell antigens as the measure of an individual's SARS-CoV-2 specific T-cell responses. For instance, the T-cell response strength in individual ADAP-142 was 68% [64-80] (HLA-A24:02, HLA-A24:02, HLA-B15:35, HLA-B40:02, HLA-C08:01, HLA-C15:02). In contrast, for ADAP-6359, the strength was 82% [74-96] (HLA-A02:01, HLA-A24:02, HLA-B07:02, HLA-B07:02, HLA-C07:02, HLA-C07:02). Interestingly, the strength of T-cell responses across individuals appeared to follow a Gaussian distribution in different populations (**Figure 4 c, d**). These findings suggest that the breadth and strength of SARS-CoV-2 specific T-cell responses vary among individuals. However, the strength remains fairly consistent within the same individual, following a Gaussian distribution across different populations. This predictive diagnostic tool could potentially guide physicians in determining which patients may require preventative measures or treatment.

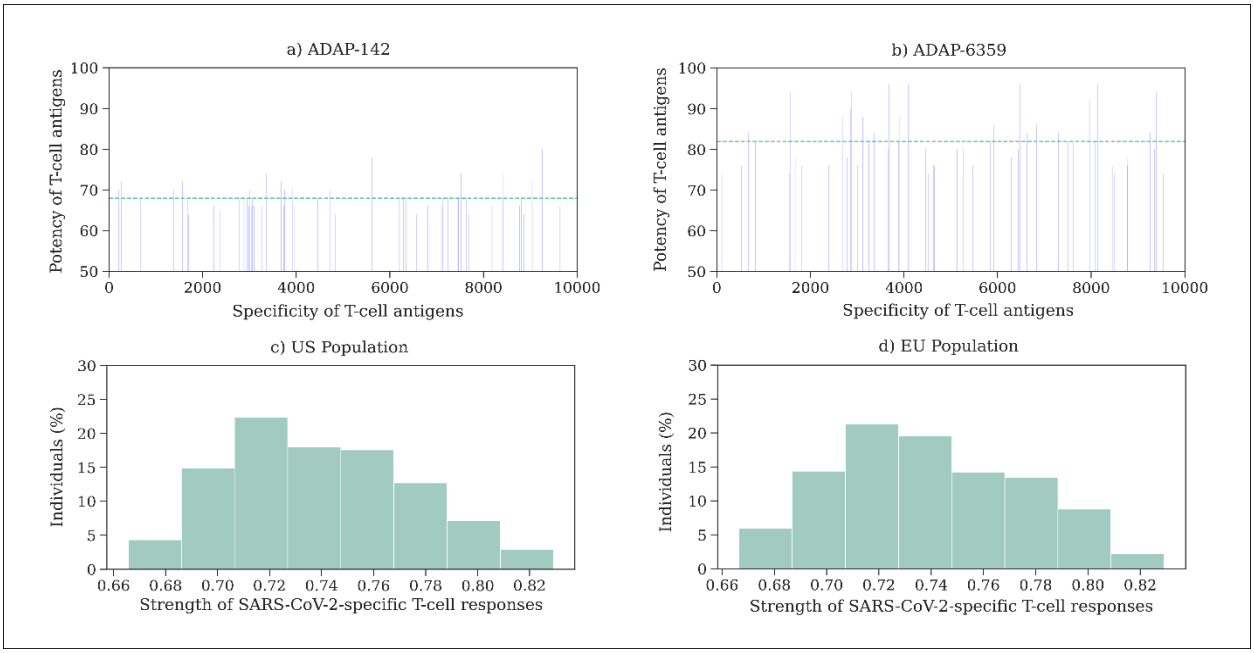


Figure 4. Characterization of SARS-CoV-2-specific cytotoxic T-cell responses at the individual level. a, b) Illustration of the specificity (indicated as the location in the proteome) and potency (likelihood of T-cell response) of the top-50 SARS-CoV-2-specific T-cell antigens in two individuals from the Adaptive Cohort. Dotted lines represent the strength of SARS-CoV-2-specific T-cell responses of the individual defined as the mean potency of the top-50 high-ranked T-cell antigens illustrated as 9-mer "cores". c and d) Frequency distribution of the strength of T-cell responses in the EU populations.

Discussion

VERDI represents the first predictive diagnostic tool for SARS-CoV-2 specific T-cell responses in individuals, based on their 4-digit HLA-genotype data. Our model posits that an epitope can be transported to the cell surface by any autologous HLA molecules, and T-cells respond to a high-density ligand on the epitopes, referred to as the "core," which stimulates the same TCR (**Figure 5**). The VERDI model aligns with experimental data showing that clusters of HLA-epitope complexes on the cell surface stimulate a cluster of TCRs. [16,17] Furthermore, our HLA-genotype-dependent framework elegantly complements prior findings that underscore the significance of epitope presentation by HLA-alleles, acknowledging that while necessary, it alone is insufficient to elicit T-cell responses.[5,18] The VERDI model also explains the promiscuity of epitopes binding to various HLA-alleles and the pivotal role they play in eliciting T-cell responses that remained previously went unnoticed at an individual level. [19] VERDI has not only elucidated this mechanism but has also paved a strategic avenue for the predictive diagnosis of the breadth and strength of T-cell responses at the individual level, which had previously remained elusive and unattainable.

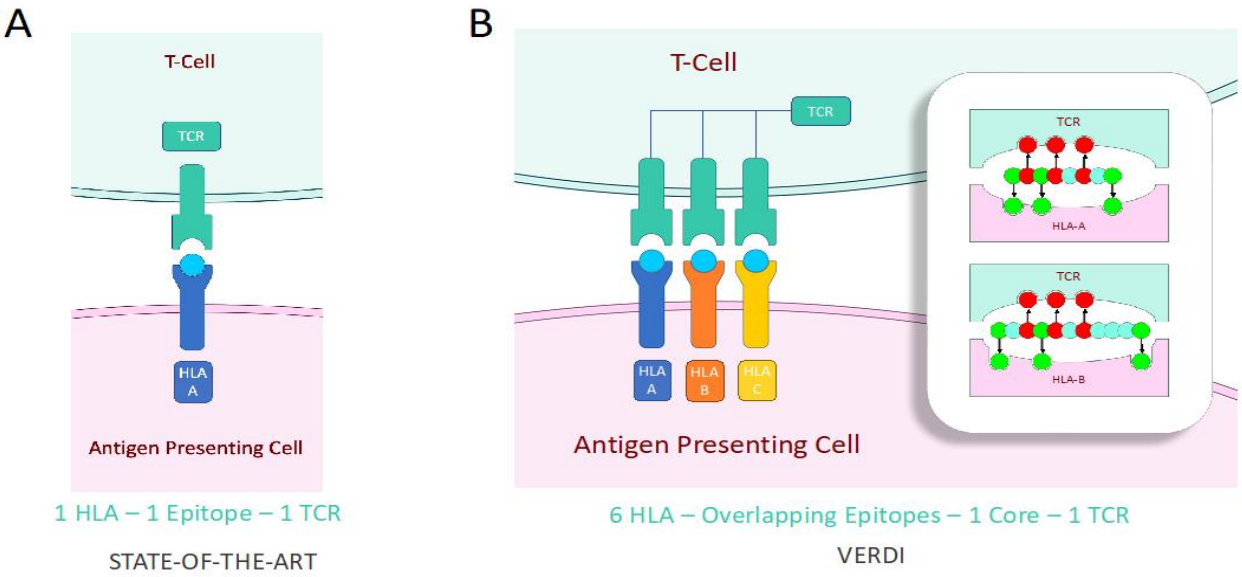


Figure 5. Models for the mechanism of induction of cytotoxic T-cell responses in an individual. a) The state of the art, modeled by EL Max assumes that T-cells are activated by an epitope that strongly binds to an individual's HLA-allele. b) The VERDI model assumes that T-cells are activated by the top-ranked high-density epitopes transported to the cell surface by any HLA-allele expressed in the cell. Overlapping epitopes can also stimulate the same T cell, which are represented as red boxes in the illustration.

Clinical studies have shown that, on average, SARS-CoV-2 infection induces T-cell responses through approximately 30 epitopes per individual.[15] Consequently, the top 50 epitopes ranked by VERDI can accurately diagnose the extent of T-cell responses to SARS-CoV-2, which may be relevant to the outcome of infection and vaccine protection at the individual level.[3] We propose a concept of sequential T-cell activation orchestrated by these top-ranked epitopes. Epitopes possessing greater potency exert a more robust stimulus on T-cells, thereby

intensifying their proliferation and accelerating the eradication of infected cells. Following the successful clearance of the infection, a subset of activated T-cells forms a memory pool, while others succumb to programmed cell death. Conversely, epitopes with comparatively diminished potency elicit a milder response in T-cell proliferation, leading to a protracted process of eliminating infected cells. As long as the infection persists, additional epitopes continue to trigger fresh T-cell clones, culminating in a diverse T-cell response. This extended immune engagement simultaneously prompts the expression of inhibitory molecules that foster viral persistence.[20] Importantly, it is notable that individuals harboring weaker epitope repertoire are more prone to experiencing elevated viral loads and the onset of severe COVID-19 symptom.[21] Experimental evidence confirms that individuals, who cleared the infection quickly and effectively, had more potent T-cell responses and fewer SARS-CoV-2-specific T-cell epitopes compared to individuals with severe COVID-19.[22] The extreme heterogeneity of T-cell responses explains the diverse clinical findings among SARS-CoV-2-infected individuals.

VERDI faces its primary limitation in the scarcity of high-quality data. Presently, experimental techniques can assess only a limited number of epitopes per individual, reaching a maximum of a few hundred. Moreover, T-cell response tests lack standardization and reproducibility among individuals. Additionally, the complete HLA-genotype of participants is often not published in clinical trials reporting the potency of T-cell epitopes. Despite these limitations, VERDI holds promise for continuous improvement through the incorporation of additional data.

VERDI holds immediate medical utility in diagnosing epitope repertoire that elicit T-cell responses in individuals, as 4-digit HLA-genotype data can be sourced from clinically validated tests employed in transplantation.[23] Moreover, VERDI's predictive diagnostic potential delivers invaluable insights to physicians, enabling a deeper comprehension of cross-reactive T-cell responses that occur between circulating viruses and vaccines. The prospect of predictive diagnostics extends towards personalized treatment strategies, enhancing the ability to manage expectations concerning the clinical trajectory of SARS-CoV-2 infection for individuals. This stems from the correlation observed between T-cell responses and clinical outcomes, as supported by previous research.[1,3,5] Furthermore, the identification of epitopes inducing T-cell responses through VERDI diagnosis could exert significant influence over personalized vaccine design. By pinpointing prevalent epitopes that are conserved across diverse SARS-CoV-2 variants, a fertile ground emerges for the creation of precision vaccines and T-cell therapies, promising advancements in the field of personalized healthcare.

Conclusions

VERDI is an innovative machine learning-based diagnostic tool, utilizing an individual's complete HLA genotype to predict T-cell responses to SARS-CoV-2. Its improved capability to assess the potency of viral antigens in individual T-cell responses sets it apart and positions it as an advancement in immunology. When compared to current state-of-the-art tools, VERDI emerges as an improved option, demonstrating an increased capacity for accuracy, sensitivity, and specificity.

This predictive diagnostic tool offers potential in reshaping vaccine development approaches, moving towards a more individualized paradigm where vaccines can be tailored based on one's unique genetic makeup. Such personalized vaccines have the capacity to stimulate more focused and robust immune responses, providing a key strategy in achieving comprehensive and lasting protection against not only SARS-CoV-2 but potentially other infectious agents.

Beyond vaccine design, VERDI holds promise in transforming the landscape of disease management and prognosis. By providing personalized insights into an individual's immune response, it might facilitate risk assessment, treatment planning, and ongoing disease monitoring, thus optimizing patient care and potentially leading to improved clinical outcomes.

Moreover, VERDI's emergence marks the initiation of a host of future research avenues and applications. The intersection of genetics and immune responses, as probed by VERDI, is poised to broaden our understanding of various diseases, extending far beyond SARS-CoV-2, and offering insights into the fundamental mechanisms underlying immune responses in health and disease.

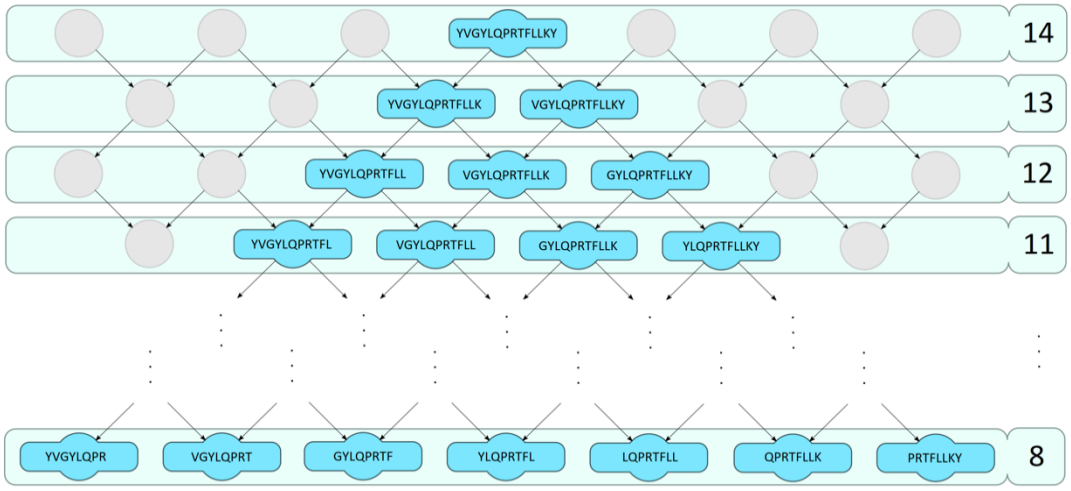
Supplementary Material

Individualized inputs of VERDI and feature engineering

VERDI uses the following features for each epitope-individual pair: Epitope Weight Scores (EWS), Subepitope Weight Scores (SWS), Blosum-62 epitope encoding, and epitope length. EL scores (ELS) were predicted with the NetMCHPan4.1 for each epitope-HLA pair. We interpret these scores at the individual level as the likelihood of the HLA-allele presenting an epitope on the cell surface. We define EWS are the aggregated ELS of an epitope, approximating the density of the epitope on the cells of the test subject. Thus, the model features EWS_A, EWS_B, and EWS_C are the separately aggregated ELS on the HLA*A, HLA*B, and HLA*C locus, respectively.

To account for the contribution of overlapping epitope relations, a multi-leveled directed graph network was introduced **(Figure S1).** Every node in the graph identifies a peptide which is distributed into different levels based on their amino acid length. Every node that is not on the lowest level is considered to be a parent node, and every parent node has an edge pointing toward its left-child and right child. The left child is one amino acid shorter Prefix of the parent peptide. The right child is one amino acid shorter Suffix of the parent peptide. We call every epitope which is below the selected epitope in the graph, a sub-epitope. The overlapping epitopes are considered by calculating the Sub-epitope Weight Scores (SWS) which are based on the EWS for the sub-epitopes in the left and right subgraph (S-Eq1, S-Eq2). When the selected epitope has a length greater than the minimum length, the SWS1 is defined by the sum of the EWS of each sub-epitope in the left subgraph of the selected epitope. The SWS2 is defined by the sum of EWS of each sub-epitope in the right subgraph, except the common nodes with the left subgraph (left subgraph of the right child) of the selected epitope.

Figure S1: Multi-leveled directed graph network used for the identification of T-cell antigens from overlapping epitopes



The left sub-epitope score (*SWS1*) and the right sub-epitope score (*SWS2*) may be calculated from:

$$SWS1_{(x_i)} = EWS_{(lx_{i-1})} + SWS1_{(lx_{i-1})} + SWS2_{(lx_{i-1})} \quad \text{[S-Eq1]}$$
$$SWS2_{(x_i)} = EWS_{(rx_{i-1})} + SWS2_{(rx_{i-1})} \quad \text{[S-Eq2]}$$

where x_i is an epitope with the length of *i*, where *i* is between 9 and 14 amino acids, lx_{i-1} is the left sub-epitope of x_i and rx_{i-1} is the right sub-epitope of x_i . The initial values for HLA class 1 are $S1(x_8) := 0$, and $SWS2(x_8) := 0$.

Finally, the epitope length is passed to the model as an integer, and BLOSUM 62 is used to encode the epitope sequence.

EL Max and VERDI Model validation at the individual level

To demonstrate the binary classification performance of the EL Max and VERDI models we performed two evaluations at the antigen-individual and the aggregated-individual levels.

First, we considered the predictions made independently for each antigen and individual. The subject-grouped five-fold cross-validation procedure uses the ABF clinical data separated into 5 random folds of 80% and 20% training and test splits, respectively. These splits were stratified to maintain the ratios of immunogenic and non-immunogenic antigens in each subject and the grouped data were not shared between train and test splits. In each split, we train the model and compute the receiver operating characteristic curve (ROC) and the precision-recall curve (PR) on the test split. The PR curve is an alternative metric to the ROC curve in this class imbalanced scenario.

Second, as we are interested in diagnosing the T-cell antigens of individuals, we compute aggregated performance metrics for each individual. Therefore, for a given individual, we rank T-cell epitopes according to the predicted likelihood of being immunogenic (potency). Then, we threshold the top-N epitopes as

antigens for each individual according to the ranking. For validating the ABF cohort, we considered the top-20 epitopes as T-cell antigens of the individual. Then, we compute two metrics: Top-20 precision and Top-20 recall, which correspond to the precision and recall metrics computed for the top-20 antigens of each individual, compared to the (non-ranked) ground-truth immunogenic epitopes in the ABF cohort (S-Eq4, S-Eq5). Ideally, we would also like to validate the congruence of our ranked likelihood (potency) with the ranked epitopes for each individual in the ABF cohort. This ranking corresponds to the "log-fold-change" for each epitope according to the ABF cohort. In this setting, we define the *Top-N ranked accuracy* for each individual, which corresponds to the percentage of the top-N experimentally ranked T-cell antigens that are correctly predicted as eliciting T-cell responses by the epitopes ranked by potency (S-Eq. 3). We explore this aggregated ranked accuracy metric for different N values and choose to explore the distribution of top-20 individual ranked accuracy.

To demonstrate the performance of VERDI with other individuals, we trained VERDI with the complete ABF cohort and performed T-cell antigen prediction on the independent Adaptive Cohort. Adaptive only reported immunogenic epitopes (T-cell antigens) and we could not obtain the results of the experiments with no T-cell responses. Therefore, we focused on the ranked accuracy at the individual level, as defined in the ABF cohort, except that the "hits" variable is used to experimentally rank epitopes for each individual. Because of the smaller data size per subject, we choose the top-3 ranked accuracy as our metric of interest. In the absence of true negative examples, the top-N precision and recall metrics are not correct in this setting. For this purpose, we computed a metric commonly used in positive unlabeled learning (PU-learning), the "PU metric", which corresponds to the recall in the dataset scaled by the fraction of positive predictions of a given model (S-Eq6).[24] This can be seen as a penalized f-score approximation that avoids benefiting models that produce an excess of false positives. A model with a higher PU metric is preferred.

Evaluation metrics of the individualized diagnostic test

We use two classes of evaluation metrics: population and individual metrics. The population metrics are computed across the dataset considering each data point in the dataset regardless of the individual and correspond to the common dataset-wise definition of ROC-AUC and PR-AUC classification metrics. These metrics provide useful comparative information on model performance.

However, individual-level metrics are more consistent with the diagnostic function of VERDI. At the individual level, we are interested in identifying the Top-N epitopes with the highest T-cell response, where N is defined according to the number of epitopes of interest per individual in each dataset. In our case, we choose N=20 in the ABF dataset and N=3 in the Adaptive dataset, as the latter contains less characterized epitopes per individual.

In each Top-N case, we define the ranked accuracy A_R as:

$$A_R = \frac{TP_R}{N} \quad \text{[S-Eq3]}$$

TP_R is computed by ranking the top-N epitopes by the ground-truth potency measure in each dataset for each individual, and by ranking the top-N antigens predicted by VERDI. TP_R corresponds to the number of correct predictions in the VERDI ranking that are present in the ground-truth ranking. Thus, A_R can be interpreted as the fraction of top-N antigens that were predicted correctly for an individual.

In addition to ranked accuracy, we were interested in measuring the precision and recall of each individual. For this, we ranked the Top-N antigens by VERDI. Then, the Precision *P* and recall *R* for an individual is defined as:

$$P = \frac{TP}{TP+FP} \quad \text{[S-Eq4]}$$

$$R = \frac{TP}{TP+FN} \quad \text{[S-Eq5]}$$

Where *TP* is the number of true positives in the epitopes of interest, *FP* is the number of false positives and *FN* is the number of false negatives at the individual level.

In the Adaptive dataset, only positive values are reported. In this case, we use a common metric in positive-unlabeled learning, computed for each individual at the top-N epitopes ranked by predicted likelihood:

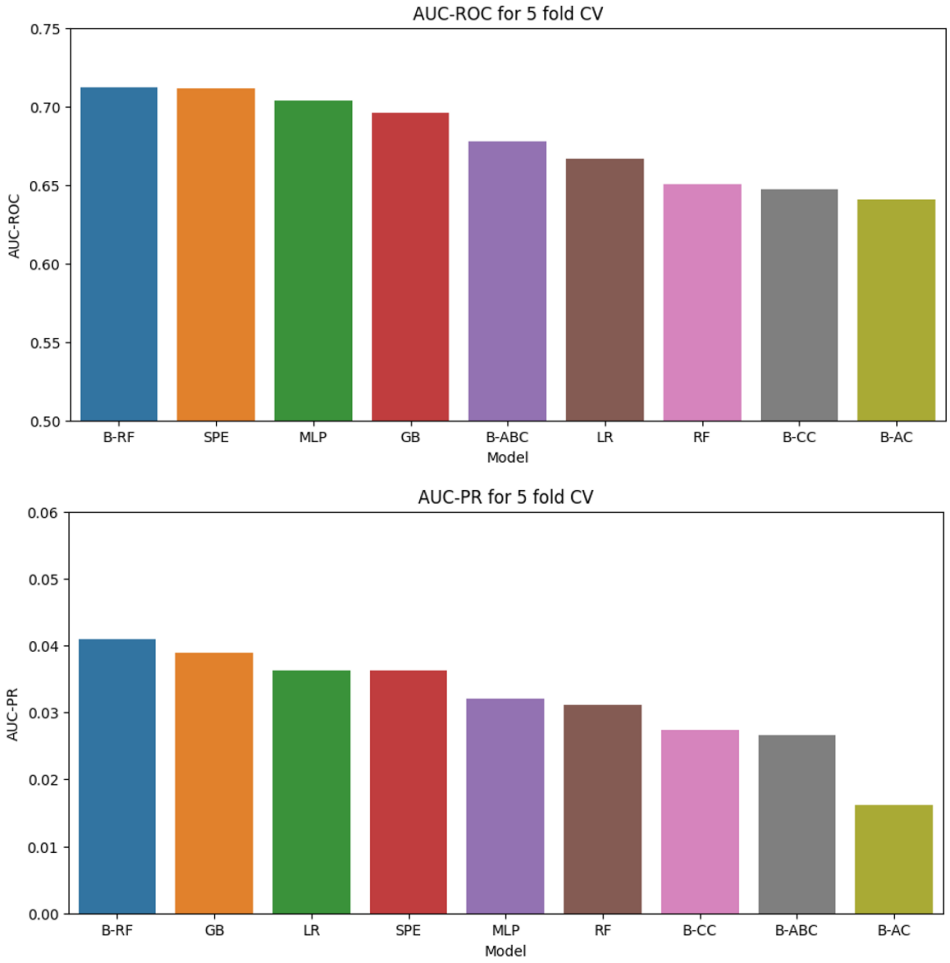
$$PU = \frac{R \cdot R}{\Pr(y'=1)} \quad \text{[S-Eq6]}$$

Where R is the recall for the Top-N ranked epitopes and, $\Pr(y'=1)$ corresponds to the fraction of the positive predictions made by the model. This metric approximates the f-score when true negatives are not available. [5]

Model selection

We used VERDI's engineered features to train and select the best-performing model based on cross-validation. The following models were evaluated by five-fold patient-grouped cross-validation. We consider several ML classification models, [25–28] including Logistic regression (LR), Gradient Boosting Regression Trees (GB), Random Forest (RF), and Multilayer Perceptron (MLP), as implemented in the sci-kit-learn package. We also considered the deep neural network architecture TabNet [29] with class weighting corresponding to the inverse of the frequency of each class. Finally, given that the dataset suffers from substantial class imbalance, we considered several machine learning models for imbalanced classification, as implemented in the imbalanced-ensemble Python package, including Balanced Random Forest (B-RF), Balanced Cascade Classifier (B-CC), AdaUBoost (B-ABC), AdaCost (B-AC), and Self-Paced Ensemble (SPE) Classifier. **Figure S2** presents the results for mean ROC-AUC and PR-AUC for each model at the individual-epitope level across 5 folds. The model with the best performance is a Balanced Random Forest classifier, which we use for validation and comparison to the EL Max model.

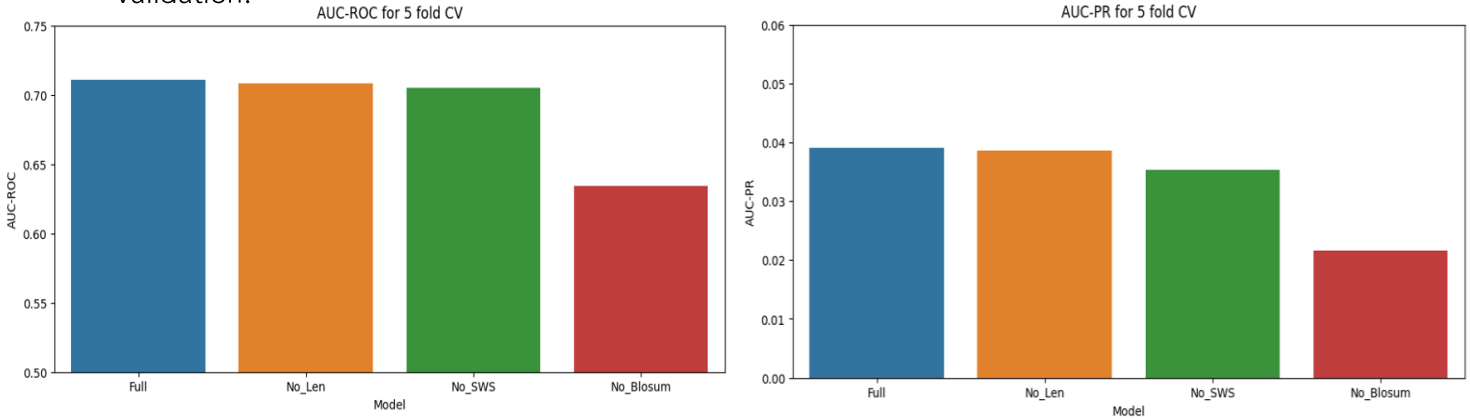
Figure S2: Mean AUC-ROC and AUC-PR for various ML algorithms using five-fold cross-validation



Individualized inputs and feature selection

Figure S3 summarizes the impact on individual-epitope performance metrics for a subset of model features. "No_Len" is a model without the epitope length feature, "No_SWS" does not include all SWS features, and "No_Blosum" is a model without epitope encoding. The largest impact on model performance is caused by removing the epitope encoding.

Figure S3: Impact of various features on mean AUC-ROC and AUC-PR using five-fold cross-validation.



Conflict of Interest Statement:

The authors affiliated with Microsoft declare no conflict of interest. Authors JL and RL have assigned their patent ownership to VERDI Solutions, a company in which they hold ownership interests.

Acknowledgement Statement:

None

Funding Statement:

None

References:

1. Sette A, Crotty S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell*. 2021;184(4): 861-880. doi:10.1016/j.cell.2021.01.007

2. Swadling L, Diniz MO, Schmidt NM, et al. Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2. *Nature*. 2022;601(7891):110-117. doi:10.1038/s41586-021-04186-8

3. Moss P. The T cell immune response against SARS-CoV-2. *Nat Immunol*. 2022;23(2):186-193. doi:10.1038/s41590-021-01122-w

4. Almendro-Vázquez P, Laguna-Goya R, Paz-Artal E. Defending against SARS-CoV-2: The T cell perspective. *Front Immunol*. 2023;14:1107803. doi:10.3389/fimmu.2023.1107803

5. Grifoni A, Sidney J, Vita R, et al. SARS-CoV-2 human T cell epitopes: Adaptive immune response against COVID-19. *Cell Host & Microbe*. 2021;29(7):1076-1092. doi:10.1016/j.chom.2021.05.010

6. Hudson D, Fernandes RA, Basham M, Ogg G, Koohy H. Can we predict T cell specificity with digital biology and machine learning? *Nat Rev Immunol*. 2023;23(8):511-521. doi:10.1038/s41577-023-00835-3

7. Lee E, Sandgren K, Duette G, et al. Identification of SARS-CoV-2 Nucleocapsid and Spike T-Cell Epitopes for Assessing T-Cell Immunity. Subbarao K, ed. *J Virol*. 2021;95(6):e02002-20. doi:10.1128/JVI.02002-20

8. Bukhari SNH, Jain A, Haq E, Mehbodniya A, Webber J. Machine Learning Techniques for the Prediction of B-Cell and T-Cell Epitopes as Potential Vaccine Targets with a Specific Focus on SARS-CoV-2 Pathogen: A Review. *Pathogens*. 2022;11(2):146. doi:10.3390/pathogens11020146

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

10. Saini SK, Hersby DS, Tamhane T, et al. SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8 $^{+}$ T cell activation in COVID-19 patients. *Sci Immunol*. 2021;6(58):eabf7550. doi:10.1126/sciimmunol.abf7550

11. Snyder TM, Gittelman RM, Klinger M, et al. *Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels*. Infectious Diseases (except HIV/AIDS); 2020. doi:10.1101/2020.07.31.20165647

12. Stone JD, Chervin AS, Kranz DM. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology*. 2009;126(2):165-176. doi:10.1111/j.1365-2567.2008.03015.x

13. Hennecke J, Wiley DC. T Cell Receptor–MHC Interactions up Close. *Cell*. 2001;104(1):1-4. doi:10.1016/S0092-8674(01)00185-4

14. FDA. Fact sheet for healthcare providers. Published September 2, 2021. https://www.fda.gov/media/146479/download

15. Tarke A, Sidney J, Kidd CK, et al. Comprehensive analysis of T cell immunodominance and immunoprevalence

of SARS-CoV-2 epitopes in COVID-19 cases. *Cell Reports Medicine*. 2021;2(2):100204. doi:10.1016/j.xcrm.2021.100204

16. Blumenthal D, Edidin M, Gheber LA. Trafficking of MHC molecules to the cell surface creates dynamic protein patches. *Journal of Cell Science*. Published online January 1, 2016:jcs.187112. doi:10.1242/jcs.187112

17. Anikeeva N, Fischer NO, Blanchette CD, Sykulev Y. Extent of MHC Clustering Regulates Selectivity and Effectiveness of T Cell Responses. *The Journal of Immunology*. 2019;202(2):591-597. doi:10.4049/jimmunol.1801196

18. Assarsson E, Sidney J, Oseroff C, et al. A Quantitative Analysis of the Variables Affecting the Repertoire of T Cell Specificities Recognized after Vaccinia Virus Infection. *The Journal of Immunology*. 2007;178(12):7890-7901. doi:10.4049/jimmunol.178.12.7890

19. Jiang S, Wu S, Zhao G, et al. Identification of a promiscuous conserved CTL epitope within the SARS-CoV-2 spike protein. *Emerging Microbes & Infections*. 2022;11(1):730-740. doi:10.1080/22221751.2022.2043727

20. Aghbash PS, Eslami N, Shamekh A, Entezari-Maleki T, Baghi HB. SARS-CoV-2 infection: The role of PD-1/PD-L1 and CTLA-4 axis. *Life Sciences*. 2021;270:119124. doi:10.1016/j.lfs.2021.119124

21. Cevik M, Kuppalli K, Kindrachuk J, Peiris M. Virology, transmission, and pathogenesis of SARS-CoV-2. *BMJ*. Published online October 23, 2020:m3862. doi:10.1136/bmj.m3862

22. Nelde A, Bilich T, Heitmann JS, et al. SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat Immunol*. 2021;22(1):74-85. doi:10.1038/s41590-020-00808-x

23. Ameratunga R, Woon ST, Jordan A, et al. Response to letter to the editor: the clinical utility of diagnostic T cell assays for COVID-19. *Expert Review of Clinical Immunology*. 2021;17(11):1159-1161. doi:10.1080/1744666X.2021.1982386

24. Jaskie K, Spanias A. *Positive Unlabeled Learning*. Springer International Publishing; 2022. doi:10.1007/978-3-031-79178-9

25. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis*. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

26. Vladimir Vovk. Kernel Ridge Regression - Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. *Springer*. 2013:105-116.

27. Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*. 2016;316(5):533. doi:10.1001/jama.2016.7653

28. Chen C, Breiman L. Using Random Forest to Learn Imbalanced Data. *University of California, Berkeley*. Published online January 2004.

29. Arik SÖ, Pfister T. TabNet: Attentive Interpretable Tabular Learning. *AAAI*. 2021;35(8):6679-6687. doi:10.1609/aaai.v35i8.16826