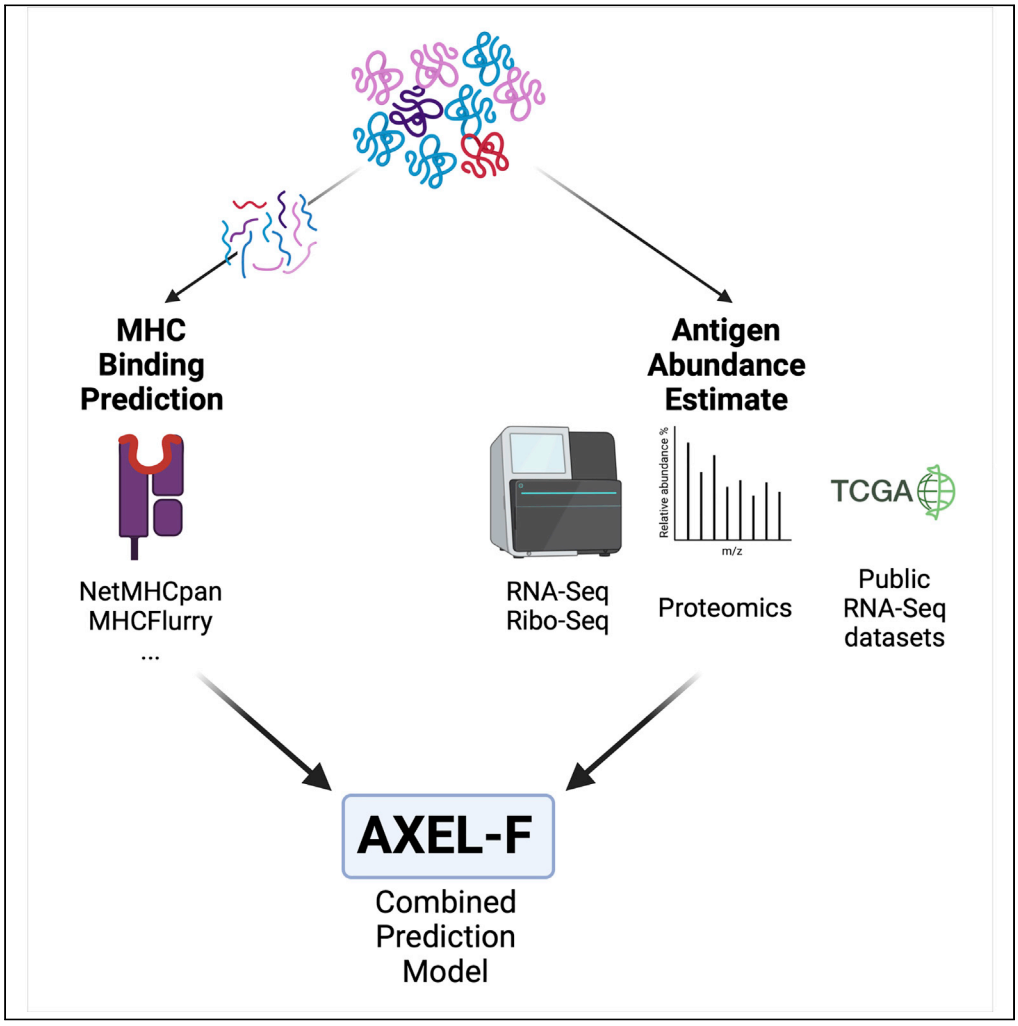
iScience

CellPress
OPEN ACCESS

Article

Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions



Zeynep Koşaloğlu-Yalçın, Jenny Lee, Jason Greenbaum, ..., Alessandro Sette, Morten Nielsen, Bjoern Peters

bpeters@lji.org

Highlights

HLA ligands originate from highly expressed transcripts

Antigen abundance and HLA binding are independent predictors of ligands and epitopes

Utilizing RNA-Seq, Ribo-Seq, or proteomic data improves epitope predictions

Cancer-type-matched TCGA RNA-Seq data can be used to estimate gene expression in patient

Koşaloğlu-Yalçın et al., iScience 25, 103850 February 18, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.103850

iScience

CellPress
OPEN ACCESS

Article

Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions

Zeynep Koşaloğlu-Yalçın,[1] Jenny Lee,[1] Jason Greenbaum,[1] Stephen P. Schoenberger,[2,3] Aaron Miller,[2,3] Young J. Kim,[4] Alessandro Sette,[1,5] Morten Nielsen,[6,7] and Bjoern Peters[1,5,8,*]

SUMMARY

Many steps of the MHC class I antigen processing pathway can be predicted using computational methods. Here we show that epitope predictions can be further improved by considering abundance levels of peptides' source proteins. We utilized biophysical principles and existing MHC binding prediction tools in concert with abundance estimates of source proteins to derive a function that estimates the likelihood of a peptide to be an MHC class I ligand. We found that this combination improved predictions for both naturally eluted ligands and cancer neoantigen epitopes. We compared the use of different measures of antigen abundance, including mRNA expression by RNA-Seq, gene translation by Ribo-Seq, and protein abundance by proteomics on a dataset of SARS-CoV-2 epitopes. Epitope predictions were improved above binding predictions alone in all cases and gave the highest performance when using proteomic data. Our results highlight the value of incorporating antigen abundance levels to improve epitope predictions.

INTRODUCTION

Presentation of peptides on the cell surface by major histocompatibility complex (MHC) class I molecules is crucial for $CD8^+$ T-cell-mediated immune responses, including those against viral infections and tumors. The MHC class I antigen processing and presentation pathway consists of multiple steps during which proteins are degraded into peptides, loaded on MHC class I molecules, and presented on the cell surface (Leone et al., 2013). Recognition of these peptide-MHC complexes on the cell surface as foreign by $CD8^+$ T cells prompts an immune response, which can lead to the eradication of affected cells. Accurate identification of which specific peptides are presented on MHC class I has applications in developing diagnostics and therapeutic interventions for infectious diseases and cancer (Soria-Guerra et al., 2015; Patronov and Doytchinova, 2013; Schumacher et al., 2019).

Numerous computational tools have been developed to predict the various steps in the MHC class I antigen processing and presentation pathway (reviewed in Peters et al. (2020)), including prediction of proteasomal cleavage (Nielsen et al., 2005; Eggers et al., 1995), transport into the ER by the transporter associated with antigen processing (TAP) (Peters et al., 2003; Bhasin and Raghava, 2004), peptide-MHC binding (reviewed in Nielsen et al. (2020)), and predicting the stability of the peptide-MHC complex (Rasmussen et al., 2016; Jorgensen et al., 2014). Among these, tools predicting peptide-MHC binding has been proven to be the most discriminative in predicting immunogenic epitopes, i.e. presented peptides that are recognized by T cells (Kosaloglu-Yalcin et al., 2018; Bjerregaard et al., 2017; Nielsen et al., 2020; Peters et al., 2020; Paul et al., 2020). These tools generally consist of machine learning methods that have been trained with experimentally generated peptide-MHC binding data. Such experimental data are, for example, available in the Immune Epitope Database (IEDB) (Vita et al., 2019).

One drawback of using predictions solely based on peptide-MHC binding data is that it ignores the antigen processing and presentation pathway. This drawback can be overcome by using ligand elution data for training. These ligands are naturally found presented by MHC molecules on the cell surface, which means they passed through the natural antigen processing and presentation pathway. Ligand elution data inherently contain information on sequence motifs associated with processing that is not available when only

[1]Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA 92037, USA

[2]Division of Hematology and Oncology, Center for Personalized Cancer Therapy, San Diego Moore's Cancer Center, University of California, San Diego, San Diego, CA, USA

[3]Laboratory of Cellular Immunology, La Jolla Institute for Immunology, La Jolla, CA 92037, USA

[4]Department of Otolaryngology-Head & Neck Surgery, Vanderbilt Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[5]Department of Medicine, University of California, San Diego, La Jolla Institute for Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

[6]Department of Health Technology, Technical University of Denmark, DK Lyngby, 2800, Denmark

[7]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP San Martín, B1650, Argentina

[8]Lead contact

*Correspondence: bpeters@lji.org

https://doi.org/10.1016/j.isci.2022.103850

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

peptide-MHC binding is considered (Peters et al., 2020). Recent advances in high-throughput ligand elution assays allow identifying thousands of natural ligands with a single experiment (Shao et al., 2018; Vaughan et al., 2017). Importantly, the eluted ligands are not biased by any pre-selection using prediction methods. These large and unbiased datasets of eluted ligands provide more power when training machine learning methods. In fact, machine learning methods that have been trained on a combination of peptide-MHC binding and ligand elution data outperform methods that have been trained on peptide-MHC binding data alone in predicting epitopes (Jurtz et al., 2017; O'Donnell et al., 2020; Kosaloglu-Yalcin et al., 2018; Nielsen et al., 2020; Peters et al., 2020).

One aspect in the antigen processing and presentation pathway that is still often ignored by epitope prediction methods is the abundance of epitope source proteins. Proteomic studies have previously reported correlations between protein abundance and MHC-peptide presentation (Juncker et al., 2009; Hickman et al., 2004; Milner et al., 2006), and more recently, it was reported that MHC-peptide presentation is strongly correlated with mRNA expression of the ligand's source protein (Abelin et al., 2017; Fortier et al., 2008; Bassani-Sternberg et al., 2015; Juncker et al., 2009), underlining the potential value of including information about source protein abundance into epitope predictions.

Some might argue that MHC ligand elution data already contain information about antigen abundance. However, it is the hallmark of multicellular organisms to have tissue- and cell-type-dependent expression patterns of genes. Ligand elution data are typically generated from cell lines or specifically isolated tissues and therefore only reflects the expression patterns of the specific cell type from which ligands were eluted. Although a significant subset of genes, such as actin or ubiquitin, are expressed across all cell and tissue types at comparable levels, other genes such as insulin or keratin are only expressed by specific cell types in specific tissues (Uhlen et al., 2015).

Cell type and tissue-specific expression patterns can, for example, be better captured by RNA sequencing (RNA-Seq). Abelin et al. and Sarkizova et al. reported increased performance in predicting eluted ligands, Sarkizova et al. additionally reported increased performance in predicting peptides that were observed experimentally in patient-derived tumor cell lines, and Bulik-Sullivan et al. reported increased performance in predicting immunogenic neoantigens when they included RNA expression of epitope source antigens in their respective machine learning models (Bulik-Sullivan et al., 2018; Abelin et al., 2017; Sarkizova et al., 2020).

In this study, we wanted to formally describe the interplay of peptide-MHC binding and the abundance of the peptide's source protein. We took advantage of the publicly available, highly accurate peptide-MHC binding prediction tool NetMHCpan 4.0 and developed a model that combines these predictions with the RNA expression of the peptide's source protein in a biophysically meaningful fashion to estimate the likelihood of the peptide being presented on a given MHC class I molecule. Our model named **A**ntigen e**X**pression based **E**pitope **L**ikelihood-**F**unction (AXEL-F) outperformed NetMHCpan 4.0 in discriminating eluted ligands from random background peptides as well as in predicting neoantigens that are recognized by T cells. We also showed that in cases where cancer-patient-specific RNA-Seq data is not available, cancer-type-matched expression data from TCGA can be used to accurately estimate patient-specific gene expression. Using AXEL-F together with TGCA expression data, we were able to improve the prediction of neoantigens that are recognized by T cells. We furthermore showed that SARS-CoV-2 epitopes can be more effectively predicted when abundance levels of virus proteins are taken into account. Epitope predictions were improved when we used AXEL-F together with RNA-Seq of SARS-CoV-2 infected cells, and predictions were even further improved when ribosome profiling or proteomic data was utilized to measure antigen abundance.

AXEL-F is publicly available and free to use for the academic community at http://axelf-beta.iedb.org/axelf.

RESULTS

HLA class I eluted ligands originate from highly expressed genes and are predicted good HLA binders

We wanted to assess the performance of expression level, predicted binding affinity, and their combination in distinguishing eluted ligands from a set of random background peptides. We utilized a previously published dataset of 15,090 HLA class I ligands eluted from five different HLA class I alleles (hereafter referred

to as Trolle set) (Trolle et al., 2016). Importantly, this dataset was not preselected or filtered in any way. We compared the eluted ligands against a set of background peptides. The background dataset was generated as follows: for each peptide in the Trolle set, ten peptides were randomly picked from the human proteome. As RNA expression data were not included in the Trolle study, we retrieved RNA-Seq data of HeLa cells from another previously published study (Cantarella et al., 2019) and used the provided UniProt identifiers of the ligand source proteins to annotate the corresponding expression level measured as transcripts per million (TPM). Next, we performed HLA class I binding predictions for each ligand and random background peptide using the NetMHCpan 4.1 algorithm (Jurtz et al., 2017). For each peptide, we retrieved the predicted binding affinity provided in IC50 together with the corresponding percentile rank (BA_Rank), as well as the eluted ligand score (EL_Score) and the corresponding percentile rank (EL_Rank). The complete Trolle dataset is provided in Table S1.

We compared RNA expression levels of the genes from which eluted ligands originated with expression levels of the genes from which background peptides were retrieved. We analyzed each of the five alleles separately (Figure 1A). As expected, this analysis showed that RNA expression levels of ligands are significantly higher than those of random background peptides ($p < 2.2 \times 10^{-16}$, Wilcoxon Test). These results confirm that MHC I eluted ligands are preferentially derived from abundant proteins, as previously reported (Abelin et al., 2017). We next compared predicted IC50 values of eluted ligands with background peptides separately for each of the five alleles and found that eluted ligands were predicted to bind at significantly higher levels ($p < 2.2 \times 10^{-16}$, Wilcoxon Test, Figure 1A). These results are in concordance with previously reported studies (Bulik-Sullivan et al., 2018; Abelin et al., 2017; Sarkizova et al., 2020). Similar results were obtained when this analysis was performed based on BA_Rank and EL_Rank (Figure S1). Results were also similar when we only used the subset of ligands and background peptides that were predicted binders (IC50 < 500 nM, Figure S2).

As a next step, we wanted to further investigate the relationship and interplay between HLA binding and RNA expression levels. Directly comparing HLA binding and RNA expression showed that there is no correlation (R = 0.17, Pearson's correlation, Figure S3). To further investigate this, we separated the binding affinity and TPM values in our dataset into ranges to create a 2-dimensional matrix with the TPM on the x axis and the predicted binding on the y axis, analogous to what was reported by Abelin et al. (Abelin et al., 2017) (Figures 1B and S4). Visual inspection of the resulting matrix revealed that certain IC50 and TPM ranges were enriched for eluted ligands, namely the part of the matrix representing high binding affinity and substantial RNA expression, as already discussed earlier. The matrix, however, showed additional interplay between IC50 and TPM: ligands originating from lower expressed RNA transcripts seemed to bind HLA strongly, whereas ligands that were not able to strongly bind HLA seemed to be derived from highly expressed RNA transcripts.

These observations, which are in concordance with others (Abelin et al., 2017), indicated that HLA binding of a ligand and the expression of its source protein might compensate for each other. Abundant expression of a source protein will generate more peptides, which in turn might enhance the chances of these abundant peptides to bind HLA even if their HLA binding capacity is weak, simply by being available in high numbers. Conversely, a peptide with high binding affinity might still be presented on HLA even if it is not abundantly expressed by outcompeting other more abundant peptides available for HLA binding.

HLA binding and expression level are independent predictors of HLA class I eluted ligands

Having established that eluted ligands are highly expressed and are predicted good binders, we analyzed the predictive performance of these metrics in distinguishing ligands (positives) from background peptides (negatives). We considered all four metrics provided by NetMHCpan as well as the TPM of the source protein as a measure of expression and performed a receiver operating characteristic (ROC) analysis to assess prediction performance in terms of the area under the ROC curve (AUC) as well as partial AUC at 10% false positive (pAUC). All NetMHCpan metrics were excellent predictors of eluted ligands: all AUC and pAUC values were above 0.99. With an AUC value of 0.812 and a pAUC of 0.629, TPM alone was also a good predictor for eluted ligands (Figure 2).

Integrating HLA binding of the ligand and expression of its source protein using a Boltzmann distribution improves prediction of eluted ligands

To integrate the HLA binding capacity of the ligand and the abundance of its source protein into a function to more accurately predict ligand elution, we first applied a naive approach to combine HLA binding and

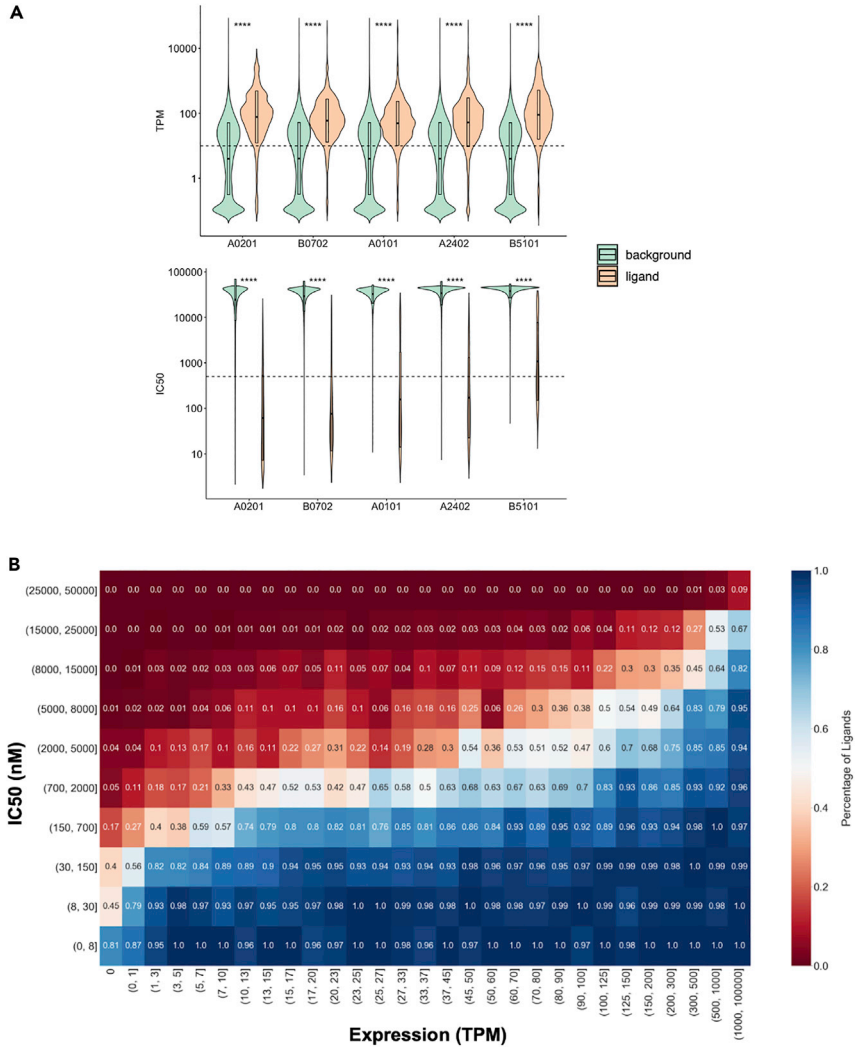


Figure 1. HLA binding and abundance of source proteins of HLA class I eluted ligands

(A) HLA class I eluted ligands originate from highly expressed genes and are predicted good HLA binders. The quartile ranges and density of TPM (top) and predicted IC50 (bottom) values are displayed for the five alleles included in the dataset. Ligands (displayed in tan) are expressed at significantly higher levels than random background peptides (displayed in green) and are predicted to bind at significantly higher levels ($p < 2.2 \times 10^{-16}$, Wilcoxon Test). Dashed lines indicate TPM 10 and IC50 500 nM, respectively.

(B) Interplay between HLA binding of eluted ligands and expression of their source proteins. The binding affinities and TPM values were separated into ranges to create a 2-dimensional matrix with the TPM on the x axis and the IC50 on the y axis. Each peptide was assigned to a cell in this matrix according to its IC50 and TPM values. For each cell, the percentage of ligands among all peptides that fall into the corresponding IC50 and TPM ranges was determined, and the cell was colored accordingly.

RNA expression by simply assigning a poor predictive value to each peptide that was derived from a non-expressed source protein (i.e. TPM = 0). This approach was based on the biological assumption that a peptide cannot be an eluted ligand if its source protein is not expressed. We considered all peptides from the Trolle set and the background peptides and assigned each peptide the worst possible prediction score if the corresponding source protein was not expressed. An ROC analysis was performed for each metric provided by NetMHCpan, and corresponding AUC values summarized in Table S2 clearly indicated that this naive method does not improve predictive performance across NetMHCpan predictions.

Next, we tested a more complex model to capture the effect of quantitative expression differences and combine them with peptide-HLA binding using the Boltzmann distribution, which is often used to describe

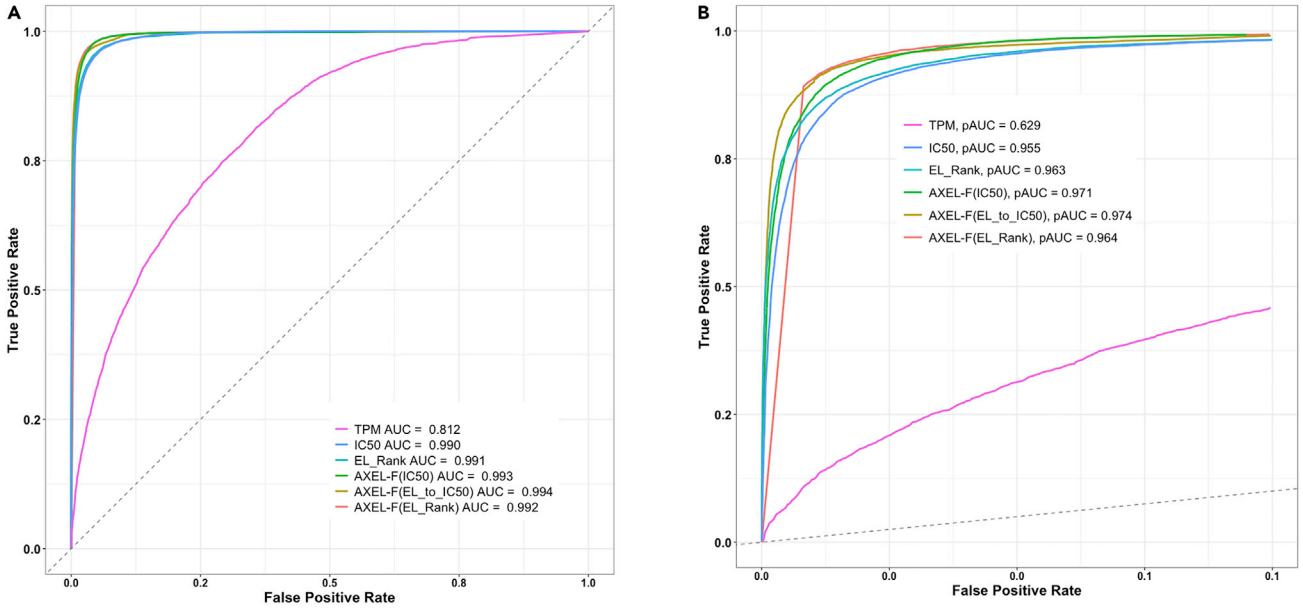


Figure 2. Performance of different predictors in identifying HLA class I eluted ligands in the Trolle dataset

(A and B) Receiver operating characteristic (ROC) curves (A) and ROC curves at 10% false-positive rate (B) for different NetMHCpan predictors, TPM and AXEL-F scores are displayed

biophysical systems (Moore, 1972). We adapted the Boltzmann formalism by empirically adding parameters to describe peptide presentation (described in detail in the methods section). Our final model estimates, for a given peptide with IC50 and TPM value, its likelihood of being presented on HLA and being an epitope. We named this model AXEL-F, standing for **A**ntigen e**X**pression based **E**pitope **L**ikelihood-**F**unction.

The model had three free parameters, namely α, which is a scaling factor for TPM, kT, which is a scale that mimics the product of the Boltzmann's constant k and the thermodynamic temperature T, and minTPM, which is a parameter that accounts for the detection limit of RNA-Seq. We used the Trolle dataset to identify the optimal value for these free parameters based on the predictive performance measured by AUC. The parameters obtained in this way fell into a consistent range between the five subsets corresponding to the five alleles in the Trolle set (Table S3).

AXEL-F outperformed IC50 consistently for all five alleles as well as for the complete dataset (p value = 0.002, De-Long's test), as shown by increased AUC values (Table S3). AUC values for AXEL-F scores were also slightly higher than those for EL_Rank for most alleles and were on par when applied to the complete dataset. As the gain in performance was significantly higher when parameters were trained on the complete dataset (p value $< 2.2 \times 10^{-16}$, De-Long's test), we decided to present subsequent analyses for the complete dataset only.

It is widely known (Jurtz et al., 2017) that the NetMHCpan EL_Rank and EL_Score predictions generally perform better in predicting eluted ligands than predicted IC50. The neural network that performs these EL predictions has been trained on eluted ligand data and is thus more capable of capturing eluted ligands. We thus used EL_Rank instead of IC50 in our model and fit the free parameters as described earlier. This model had lower AUC and pAUC values than the model using IC50 in predicting the ligands in the Trolle dataset (Table 1). This might be due to the fact that the EL_Score is an output score of the neural network architecture and is an abstract value that cannot be directly translated into a biological context. IC50 values in contrast are defined as the concentration that inhibits 50% binding of a labeled reference peptide, and if the assay is performed under appropriate conditions, the log(IC50) values are proportional to binding free energies and can be directly used in our biological model. We still wanted to take advantage of the known superior predictive performance of EL_Rank and incorporate it into our model. To achieve this, we mapped each EL_Rank to a corresponding IC50 value by comparing the percentile ranks of the two metrics. This

Table 1. Prediction performance of different predictors for ligand elution datasets

Predictor	Trolle AUC	Trolle pAUC	Abelin AUC	Abelin pAUC	Pyke cell line AUC	Pyke cell line pAUC	Pyke tissue AUC	Pyke tissue pAUC
TPM	0.812	0.629	0.763	0.607	0.704	0.583	0.694	0.580
IC50	0.990	0.955	0.969	0.940	0.951	0.932	0.961	0.870
EL_Rank	0.991	0.963	0.977	0.961	0.965	0.940	0.979	0.932
AXEL-F (IC50)	0.993	0.971	0.976	0.949	0.954	0.924	0.965	0.883
AXEL-F (EL_Rank)	0.992	0.964	0.980	0.960	0.959	0.916	0.967	0.865
AXEL-F (EL_to_IC50)	0.994	0.974	0.980	0.961	0.964	0.941	0.978	0.925
MHCflurry	0.986	0.957	0.988	0.969	0.969	0.936	0.970	0.909
AXEL-F (MHCflurry)	0.991	0.970	0.989	0.975	0.941	0.824	0.974	0.918
MixMHCpred	0.990	0.959	0.975	0.963	0.909	0.900	0.969	0.923
HLAthena	0.869	0.899	0.911	0.926	0.960	0.934	0.955	0.920

allowed us to use the improved performance of EL predictions while still being able to biologically explain the scores and integrate them into our model. We named this new metric EL_to_IC50 and used it as an input for AXEL-F instead of the IC50 values we used earlier and again used the Trolle dataset to train the parameters α, kT, and minTPM. The values we obtained were very similar to those obtained when using IC50: α was fit to 1.228, kT to 0.222, and minTPM to 0.544. We were able to further boost the performance of AXEL-F and achieved an AUC of 0.994 and a pAUC of 0.974 (Table 1, Figure 2). AXEL-F (EL_to_IC50) significantly outperformed all NetMHCpan predictions (p value $< 2.2 \times 10^{-16}$, De-Long's test).

To compare AXEL-F performance, we also performed predictions using three additional state-of-the-art prediction tools, MHCflurry 2.0 (O'Donnell et al., 2020), MixMHCpred 2.0 (Bassani-Sternberg et al., 2017), and HLAthena (Sarkizova et al., 2020). All tools were trained with eluted ligand data, and HLAthena also included antigen RNA expression levels in its machine learning model. Importantly, the Trolle dataset was part of the training data of all tools, including NetMHCpan. As expected, all tools perform well in predicting eluted ligands from the Trolle dataset; however, HLAthena only supports predictions for peptides of length 8–11 which left ~2,500 ligands of length 12–14 (16%) without predictions, resulting in lower performance (Table 1). AXEL-F outperformed MHCflurry, MixMHCpred, and HLAthena significantly (p value $< 2 \times 10^{-13}$, De-Long's test).

As our model is based on the biophysical principles of antigen presentation, it should be applicable to any method predicting HLA binding. To demonstrate this, we used MHCflurry predictions and fitted the free parameters of AXEL-F α, kT, and minTPM in the same way we did earlier for NetMHCpan predictions. The values we obtained were very similar to those obtained when using NetMHCpan: α was fit to 1.296, kT to 0.119, and minTPM to 0.533. Importantly, with an AUC of 0.991, AXEL-F (MHCflurry) significantly outperformed MHCflurry alone (p value $< 2 \times 10^{-13}$, De-Long's test, Table 1).

Model evaluation on independent datasets of eluted ligands

We next validated our results with an independent dataset (Abelin et al., 2017) that contained 26,089 eluted ligands from 16 HLAs, with only four overlapping with those present in the Trolle dataset. Abelin et al. used mono-allelic cell lines to elute ligands from and also performed RNA-Seq and provided the data in the form of TPM values. We again generated a background dataset and performed NetMHCpan predictions for all 260,890 peptides. The complete Abelin dataset is provided in Table S4.

We calculated AXEL-F likelihood scores by using both IC50 and EL_to_IC50 as inputs and compared the performance with NetMHCpan predictions in discriminating eluted ligands from the random background set. The performance metrics summarized in Table 1 and the ROC curves shown in Figure S5 indicate that AXEL-F likelihood scores significantly outperformed IC50 and BA_Rank when likelihood scores are calculated using IC50 (AXEL-F (IC50), p value $< 2.2 \times 10^{-16}$, De-Long's test). When likelihood scores were

calculated using EL_to_IC50 (AXEL-F (EL_to_IC50)) it also significantly outperformed EL_Score and EL_Rank (p value < 2.2×10^{-16}, De-Long's test). We again considered the three tools MHCflurry, MixMHCpred, and HLAthena, and AXEL-F (EL_to_IC50) outperformed MixMHCpred and HLAthena significantly, whereas MHCflurry performed significantly best on this dataset (p value < 2.2×10^{-16}, De-Long's test, Table 1). Again, AXEL-F using MHCflurry also slightly improved performance of MHCflurry alone. Importantly, the Abelin dataset was also included in the training of all tools.

To further validate our results, we obtained two additional datasets from a recently published study (Pyke et al., 2021). Pyke et al. used mono-allelic cell lines to elute ligands from 25 different HLA. As no expression data was provided for these cell lines, we obtained precalculated TPM values for the corresponding cell line (K562) from the Cancer Cell Line Encyclopedia (Ghandi et al., 2019). Pyke et al. also eluted ligands from 12 tissue samples of colorectal and lung cancer patients. As it was not indicated which tissue sample corresponded to which cancer type, we used the TCGA pan-cancer expression dataset to estimate antigen abundance in this set.

Interestingly, although AXEL-F using IC50 significantly outperformed IC50 alone in predicting eluted ligands from the Pyke datasets (p value < 2.2×10^{-16}, De-Long's test), AXEL-F (EL_to_IC50) performed slightly worse than EL_Rank alone (Table 1). Similarly, MHCflurry alone performed significantly better than AXEL-F (MHCflurry) on the Pyke cell line dataset. (p value < 2.2×10^{-16}, De-Long's test, Table 1). On the Pyke cancer patients dataset, however, AXEL-F (MHCflurry) significantly outperformed MHCflurry predictions alone (p value < 2.2×10^{-16}, De-Long's test, Table 1).

Overall, these data showed that AXEL-F outperformed all NetMHCpan predictions alone on both the original Trolle dataset and the independent Abelin dataset. On the Abelin dataset, AXEL-F even performed well for alleles that were not included in the training dataset (Table S5). We also achieved similar results when using AXEL-F with MHCflurry instead of NetMHCpan predictions, highlighting that our model is potentially applicable to any HLA binding prediction method. When tested on two additional independent elution datasets, AXEL-F improved NetMHCpan IC50 predictions but EL predictions were not improved, whereas MHCflurry predictions were only improved in one of the datasets.

Cancer neoantigens can be more accurately predicted by integrating cancer expression data

We next wanted to analyze how our model AXEL-F performed in predicting epitopes, specifically cancer neoepitopes that arise from somatic mutations. We utilized a previously published study by Parkhurst et al. (Parkhurst et al., 2019) that reported immunogenicity screening results of neoantigens from 75 patients with various gastrointestinal cancers. The group performed whole-exome sequencing to detect somatic mutations and determined which neoantigens were recognized by tumor-infiltrating lymphocytes (Parkhurst et al., 2019). We chose this dataset, as unlike many other studies, Parkhurst et al. did not preselect the peptides for immunogenicity screening based on binding predictions or expression thresholds, and the group also provided some RNA-Seq information that we wanted to explore with our model.

As our study is based on HLA class I predictions, we only considered the 54 neoantigens that were recognized by $CD8^+$ T cells and the 7,529 peptides that were not recognized at all. We further filtered the dataset by only retaining peptides for which RNA-Seq information was provided, which resulted in a final set of 46 patients with 28 recognized neoantigens and 1,298 peptides that were not recognized. We named this dataset the NCI dataset (Table S6) and wanted to compare the performance of NetMHCpan predictions alone and in combination with neoantigen source protein abundance in distinguishing immunogenic neoantigens (positives) from peptides that were not recognized by tumor-infiltrating lymphocytes (negatives).

To do so, we first performed NetMHCpan predictions on the complete dataset and found that both IC50 and EL_Rank could discriminate immunogenic neoantigens with AUC values of 0.698 and 0.688, respectively (Table 2). The RNA-Seq information that was provided with the NCI dataset included the number of reads overlapping the mutation site (tumor_rna_depth), the number of reads overlapping the mutation site and confirming the mutation (tumor_rna_alt_reads), and the relative frequency of reads confirming the mutation among all reads overlapping the mutation site (tumor_rna_alt_freq). Unfortunately, TPM values were not provided as part of the dataset. As a first analysis, we assessed the predictive performance of these RNA metrics in predicting immunogenic neoantigens and found that all three metrics had some predictive value (Table 2 and Figure S6). With an AUC of 0.642, the number of reads supporting the mutation

Table 2. Prediction performance (AUC) of different predictors in predicting immunogenic neoantigens

Predictor	NCI set	Literature set
tumor_rna_alt_freq	0.593	–
tumor_rna_depth	0.586	–
tumor_rna_alt_reads	0.642	–
TCGA_TPM_subtype_matched	0.641	0.641
TCGA_TPM_pancancer	0.613	0.613
TCGA_TPM_subtype_mismatched	0.520	0.541
IC50	0.723	0.628
EL_Rank	0.729	0.614
AXEL-F (EL_to_IC50, tumor_rna_alt_reads)	0.753	—
AXEL-F (EL_to_IC50, TCGA_TPM_pancancer)	0.735	0.632
AXEL-F (EL_to_IC50, TCGA_TPM_subtype_mismatched)	0.669	0.606
AXEL-F (EL_to_IC50, TCGA_TPM_subtype_matched)	0.754	0.646
MHCflurry	0.779	0.639
AXEL-F (MHCflurry, TCGA_TPM_subtype_matched)	0.799	0.659
MixMHCpred	0.659	0.645
HLAthena (TCGA_TPM_subtype_matched)	0.756	0.657

(tumor_rna_alt_reads) had the best performance among the three RNA metrics. We presumed that the tumor_rna_alt_reads could be used as a proxy for describing the expression of mutated transcripts and used this metric instead of a TPM together with the EL_to_IC50 to calculate AXEL-F likelihood scores. The AUC values for AXEL-F scores that were obtained this way (AXEL-F (tumor_rna_alt_reads)) were higher than those of both NetMHCpan predictions as well as the tumor_rna_alt_reads alone, with an AUC of 0.753 (Table 2). The difference in AUC, however, was not significant (De-Long's test).

TPM values from TCGA can be used to accurately estimate gene expression in a given patient sample

As TPM values were not provided as part of the NCI dataset, we wanted to analyze whether it is possible to estimate RNA expression levels in a given patient sample by using expression data retrieved from The Cancer Genome Atlas (TCGA). We downloaded precalculated TCGA TPM values, and for each cancer type included, we calculated the median RNA expression for each gene across all samples in the corresponding cancer-type-specific subset. We utilized in-house RNA-Seq data of 25 patients with nine different cancer types and analyzed how well gene expression of these patients can be estimated by using the gene expression data retrieved from TCGA. For each patient from our in-house cohort, we matched the TPM from in-house RNA-Seq with TPM values from TCGA. This was done for each cancer type separately so that our in-house data was matched with all available cancer types from TCGA. We then analyzed how well TCGA median TPM values correlate with the in-house RNA-Seq TPM values. This analysis showed that cancer-type-matched TPM values correlate very well (Pearson's $r^2 > 0.6$, Figure S7).

Having established that TPM values from TCGA can be used to estimate RNA expression in a given patient sample, we proceeded to utilize this approach to estimate TPM values for the NCI dataset. We matched the cancer type and gene name for each peptide in the NCI dataset and assigned the corresponding cancer-type-specific median TPM from TCGA. With an AUC of 0.641, this TCGA_TPM alone was almost as predictive for immunogenic neoantigens as tumor_rna_alt_reads alone. When we used the TCGA_TPM together with the EL_to_IC50 as inputs for AXEL-F, we achieved the best performance reaching AUC values of 0.754 (Table 2 and Figure S6). Of note, when we did not match the cancer types and used TPM values calculated from the entire TCGA dataset (TCGA PANCAN), the AUC was 0.735 and thereby lower compared with when cancer types were matched. When we furthermore mismatched cancer subtypes, the AUC dropped

significantly to 0.669 (p value < 0.01, De-Long's test), highlighting the potentially cancer and tissue-specific expression patterns of neoantigen source proteins.

We also applied HLAthena to the NCI dataset together with the cancer-type-matched TCGA data, and the tool achieved an AUC of 0.757, outperforming AXEL-F, however not significantly (De-Long's test, Table 2). Interestingly, MHCflurry alone already outperformed all other prediction tools, and performance was further improved when AXEL-F was applied with MHCflurry predictions and the cancer-type-matched TCGA data (Table 2).

To validate these results we assembled an additional dataset of neoantigens from the literature, not considering if they were preselected based on binding or not. This dataset consisted of 222 validated neoantigens and 1,918 negatives (Table S7). As expression data were not provided, we again used the TCGA expression data to assign TPM values to each peptide. As expected, prediction performance on this set was slightly lower for all tools. Overall, the results mimicked what we observed on the NCI neoantigen dataset.

These results underline the general applicability of our model and furthermore, its potential to predict cancer neoantigens accurately. Even when patient-specific expression data are not available, which often occurs due to the many technical challenges of RNA-Seq, it is possible to estimate the expression of neoantigen source proteins from TCGA and perform more accurate predictions.

Prediction of SARS-CoV-2 epitopes can be improved by considering the abundance of viral proteins

The public health importance of the SARS-CoV-2 pandemic has led to the rapid generation of high-quality datasets on the T cell epitopes targeted by infected individuals, as well as the antigen expression and abundance associated with this virus, which makes it an ideal test case for our prediction approach. We used the most comprehensive map available for $CD4^+$ and $CD8^+$ T cell epitopes that Tarke et al. (Tarke et al., 2021) identified across the entire SARS-CoV-2 viral proteome. Tarke et al. used NetMHCpan EL predictions to select 5,600 peptides corresponding to the top 200 predicted binders for the most common 28 HLA class I (Tarke et al., 2021) alleles and ultimately discovered 523 peptide epitopes that elicited $CD8^+$ responses. As expected (Kim et al., 2014), given that the peptides chosen for testing were based on the EL predictions, a post-hoc evaluation of the predictive performance of NetMHCpan EL_Rank on this dataset in discriminating epitopes from the other 5,047 peptides was low with an AUC of 0.521 (Table 3). Similarly, the performance of MHCflurry and MixMHCpred was also low (Table 3).

We retrieved SARS-CoV-2 RNA expression data from another previously published study. Finkel et al. (Finkel et al., 2021) transfected Vero E6 cells with SARS-CoV-2 and performed RNA-Seq 5 and 24 h posttransfection. We chose to present results using data from 5 h posttransfection. Our results showed that SARS-CoV-2 epitopes are originating from significantly higher expressed RNA transcripts than peptides that were not recognized by $CD8^+$ T cells ($p < 2.2 \times 10^{-16}$, Wilcoxon Test, Figure S8). Furthermore, TPM values alone were predictive for epitopes with an AUC of 0.682. However, AXEL-F combining these expression values with binding predictions did not further improve AUC (Table 3, Figure 3).

Finkel et al. (Finkel et al., 2021) also performed ribosome profiling (Ribo-Seq) of the same SARS-CoV-2 transfected Vero E6 cells. As Ribo-Seq was designed to capture ORFs that are being actively translated (Ingolia, 2014), we hypothesized that it might measure antigen abundance more accurately than RNA-Seq. In fact, with an AUC of 0.683, TPM values derived from Ribo-Seq performed slightly better than those derived from RNA-Seq in predicting epitopes. AXEL-F combining Ribo-Seq and binding predictions improved performance to an AUC of 0.695.

Having observed the improvement from RNA-Seq to Ribo-Seq, we wanted to investigate the next step in the translation process and obtained SARS-CoV-2 proteomic datasets to measure the abundance of viral source proteins. We utilized data from a study by Poran et al. (Poran et al., 2020) that provided quantification of SARS-CoV-2 proteins from three publicly available proteomic datasets. With an AUC of 0.710, this proteomic quantification performed significantly better than TPM values derived from RNA-Seq and Ribo-Seq (p value $< 2.2 \times 10^{-16}$, De-Long's Test). Again, Axel-F combining proteomic quantification and binding predictions slightly improved performance to an AUC of 0.715 (Table 3, Figure 3).

Table 3. Prediction performance (AUC) of different predictors in predicting SARS-CoV-2 epitopes

Predictor	Tarke	Tarke with random	Peng
TPM_RNASeq	0.682	0.703	0.766
TPM_RiboSeq	0.683	0.649	0.776
Proteomic	0.710	0.681	0.773
EL_Rank	0.521	0.606	0.808
AXEL-F (RNA-Seq)	0.663	0.722	0.866
AXEL-F (Ribo-Seq)	0.695	0.749	0.867
AXEL-F (Proteomic)	0.715	0.766	0.892
MHCflurry	0.514	0.599	0.733
AXEL-F (MHCflurry, RNA-Seq)	0.561	0.524	0.791
AXEL-F (MHCflurry, Ribo-Seq)	0.614	0.520	0.794
AXEL-F (MHCflurry, Proteomic)	0.602	0.509	0.808
HLAthena (RNA-Seq)	0.575	0.565	0.727
HLAthena (Ribo-Seq)	0.633	0.613	0.727
HLAathena (Proteomic)	0.629	0.670	0.916
MixMHCpred	0.528	0.610	0.798

When we used HLAthena and AXEL-F (MHCflurry) with RNA-Seq, Ribo-Seq, and Proteomic data, prediction performance consistently dropped below the performance of AXEL-F and the abundance measures alone (Table 3). On this dataset, AXEL-F significantly outperformed all other tools (p value < 1.2e-10, De-Long's Test).

As mentioned earlier, Tarke et al. selected all peptides based on binding predictions. To simulate a more realistic scenario of the SARS-CoV-2 proteome, we randomly selected 1,000 peptides from the proteome and added them to the negatives. As expected, prediction performance of EL_Rank and AXEL-F improved (Table 3).

To validate these results, we used a second dataset of validated SARS-CoV-2 epitopes. Peng et al. tested 18-mer peptides spanning the SARS-CoV-2 proteome and identified 11 unique peptide-HLA with confirmed $CD8^+$ T cell responses (Peng et al., 2020). We used these 11 peptide-HLA as positives and the remaining 18-mer peptides spanning the SARS-CoV-2 proteome together with the 36 HLA from the cohort used in the study as negatives and assessed prediction performance using the same abundance measurements as discussed earlier. As expected, all tools performed better on this set, as peptides were not preselected using any prediction tools (Table 3). AXEL-F using RNA-Seq again outperformed EL_Rank alone, and AXEL-F using Ribo-Seq and proteomics data performed slightly better. We observed similar results for AXEL-F (MHCflurry). Surprisingly, HLAthena using proteomics data significantly outperformed all other methods (p value = 0.02, De-Long's Test) in this dataset (Table 3).

DISCUSSION

In this study, we used a biophysically inspired model to describe how antigen abundance and peptide-MHC binding affinity interact to drive MHC peptide presentation. We developed our model AXEL-F based on the hypothesis that the likelihood of a peptide being presented on HLA class I and subsequently being recognized by $CD8^+$ T cells is dependent on both the abundance of its source protein and its HLA binding capacity. AXEL-F clearly improved NetMHCpan predictions for predicting neoantigens that are recognized by T cells. We also showed that SARS-CoV-2 epitopes can be more effectively predicted when abundance levels of virus proteins are taken into account. Epitope predictions were improved when we combined binding predictions with RNA-Seq of SARS-CoV-2-infected cells, and predictions were even further improved when ribosome sequencing or proteomic data were utilized to measure antigen abundance.

We showed that the expression level of source proteins alone is already a good predictor of ligand elution. The predictive value is even more pronounced in the case of neoantigens: even though patient-specific expression data were not available and we used publicly available cancer type matched expression data

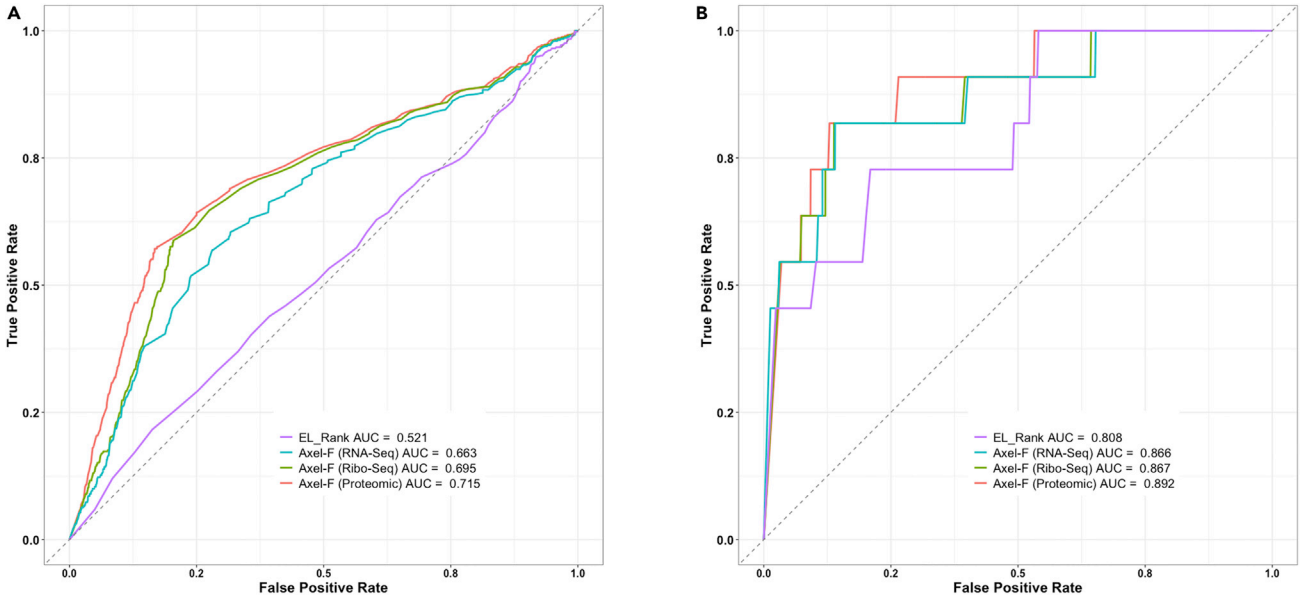


Figure 3. Performance of different predictors in identifying SARS-CoV-2 epitopes

Receiver operating characteristic (ROC) curves are displayed for NetMHCpan predictions (EL_Rank), and predictions using AXEL-F combining binding predictions with abundance measurements of viral proteins utilizing RNA-Seq, Ribo-Seq, and Proteomics.
(A and B) Tarke SARS-CoV-2 epitopes dataset; (B) Peng SARS-CoV-2 epitopes dataset.

from TCGA, the predictive performance of this estimated TPM was almost on par with NetMHCpan predictions. Importantly, neoantigens included in the NCI dataset were not prefiltered based on MHC binding or expression (Parkhurst et al., 2019), which makes it possible to accurately compare the performance of these metrics. Currently, many epitope prioritization algorithms only use expression data to filter out candidate neoantigens that do not meet a specified gene expression threshold (Garcia-Garijo et al., 2019). Our results indicate that the value of antigen expression levels has potentially been underestimated, and rather than using expression as a filtering step, it might be used in combination with HLA binding predictions to more efficiently identify neoantigens.

TPM values describe the transcript abundance of the corresponding gene normalized against the length of the gene and the total reads available in the RNA-Seq experiment. Unfortunately, TPM values were not available for the NCI neoantigen dataset, and we first used tumor_rna_alt_reads, i.e. the number of reads overlapping the mutation site and confirming the mutation, to estimate RNA expression of the neoantigen. The predictive performance of TCGA_TPM and tumor_rna_alt_reads alone was comparable, and when we used these metrics in our model, the model using TCGA_TPM clearly outperformed the model using tumor_rna_alt_reads. One reason for this might be that we trained our model using TPM values of source proteins of eluted ligands, and the parameters might have been fitted differently when trained with a neoantigen-specific dataset using tumor_rna_alt_reads. However, neoantigen datasets that provide expression details are still limited, and we are unfortunately not able to further explore this approach at this point. Another complication with using tumor_rna_alt_reads to estimate the expression of mutated transcripts is that this metric directly counts the RNA-seq reads that support the mutation and is, in contrast to TPM, not normalized considering the total number of mapped reads. As more datasets become available, we will further explore how to best specifically estimate the abundance of mutated transcripts.

Due to time or financial limitations and given the many challenges of obtaining, preserving, and sequencing RNA samples, RNA-Seq data are often not available in a clinical setting. Here, we have shown that expression data publicly available from TCGA can be used to effectively estimate patient-specific gene expression values. Our results also highlight the importance of considering tissue-specific expression patterns. Using TPM values derived from all TCGA cancer types or from mismatched cancer subtypes for neoantigen prediction was less accurate than using cancer-type-matched TCGA data. We have included TCGA expression values for 35 different cancer types in the current implementation of AXEL-F and plan to extend the available expression datasets to cell lines and different cell types.

HLA loss or downregulation is a major tumor escape mechanism that has been described in several cancer types (Garrido and Algarra, 2001). Our model could be further refined to incorporate HLA expression levels. Our model will, however, not be able to capture the complex interaction of the microenvironment that are known to bias antigen presentation and recognition (Murciano-Goroff et al., 2020).

Many factors during an RNA-Seq experiment can affect gene expression and thus the TPM of a specific gene. In our model, we introduced the parameter "minTPM" to avoid dropping transcripts not detected at all. This was necessary because RNA-Seq has a detection limit, so no result is ever truly "zero." Also, we are typically working with RNA-Seq data from different samples than what was used in the ligand elution experiment. Using a minimal value for TPM will also ensure that our model always has a minimal number of peptides to work with. This also allows our model to detect cases of high binding affinity-low TPM and vice-versa. In such cases, the likelihood scores will be low, reflecting the biology, as peptides with strong binding affinity and high TPM will always be more likely to be presented than the cases mentioned earlier. The optimal value we obtained for minTPM was 0.567, which implicitly states that lower measured TPM values do not add any additional confidence that the antigen was actually not expressed. Biologically speaking, there is no definitive TPM threshold that determines which gene is or is not expressed. Technically speaking, however, a TPM cutoff is often utilized to select genes that are considered "significantly" expressed. The EMBL-EBI Expression Atlas, for example, uses a default minimum expression level of 0.5 TPM (Papatheodorou et al., 2020). This value is very close to the minTPM value that our training determined and thus supports the biological relevance of this value.

We used a dataset of eluted ligands to train our model. Such data are generated by eluting ligands from MHC molecules and then sequencing the eluted peptides, typically by tandem mass spectrometry (MS/MS). Just as any quantification methods, these assays have a limit of detection, which is the lowest quantity of a substance that can be distinguished from system noise. The limit of detection can be influenced by several factors, such as the instrument background signal and noise, the analyte signal, and the signal-to-noise ratio. As we used published datasets and did not perform the MS/MS assays in-house, we did not have any of this information available, and our model does not take the limit of detection of MS/MS into account. Accordingly, we could also not investigate whether our observation that HLA ligands originated from highly expressed transcripts were biased by the limit of detection and the sensitivity of the MS/MS assays. It was, however, previously demonstrated in proteomic studies that MHC-peptide presentation is strongly correlated with mRNA expression levels, as well as with protein abundance, length, and half-life (Bassani-Sternberg et al., 2015; Abelin et al., 2017).

We demonstrated that the abundance of viral antigens also improved the prediction of viral epitopes in the case of SARS-CoV-2. Interestingly, epitope predictions became most accurate when using proteomic data, which is largely reflecting the protein content in viral particles. This could suggest that the antigen sources of peptides driving epitope recognition are not intracellular proteins expressed in an infected cell but rather uptake of viral particles by professional antigen-presenting cells. Although it is interesting to examine different measures of antigen abundance, for the vast majority of applications where epitope predictions come into play, the only measure available will be RNA-Seq data. Thus, although we encourage utilizing Ribo-Seq or proteomic data for predicting epitopes more accurately if available, in most practical and/or clinical settings, they will not be, which is why we are focusing most of our study on mRNA expression.

The performance of NetMHCpan and AXEL-F was higher in the eluted ligand datasets compared with the cancer and SARS-CoV-2 epitope datasets. As there was no dataset of "noneluted peptides" available, we used random peptides that were drawn from the proteome with a ratio of 1:10 (ligand: background). As shown in Figure 1, many of the random background peptides are not expressed or are only expressed at low levels, and the majority of the peptides are also not predicted to bind, which makes it easier for the classifiers to predict. In the cancer and SARS-CoV-2 epitope datasets, all peptides have been tested for immunogenicity, so we know for each peptide if it is a "real" positive or negative. In the epitope datasets, positives are more highly expressed than negatives and are also better binders; the difference is, however, not as drastic as it was the case when comparing eluted ligands with random background peptides (Figures 1A and S8). This makes it harder for the classifiers to distinguish positives from negatives. An added challenge for the cancer dataset was that the provided peptides were mainly 29mer peptides with the mutation in the center of the peptide. When performing predictions, we considered all contained 8–12mer peptides and all HLA class I alleles provided for the corresponding patient. For each 29mer

peptide, we then assigned the best prediction scores among all its k-mer and HLA combinations. Distinguishing a positive long peptide from a negative long peptide is intrinsically harder than distinguishing positive versus negative short peptides, as it is not known what the actual recognized minimal epitope within the long peptide is. The SARS-CoV-2 peptides were preselected for testing based on the EL predictions in the original study (Tarke et al., 2021). Hence, it was expected that a post-hoc evaluation of the predictive performance in detecting epitopes is systematically lower.

How the interplay between source antigen abundance and peptide-MCH binding can be utilized to more efficiently predict epitopes for other viruses remains to be further investigated. During infection, viruses hijack host cells to express genes necessary for virus propagation. Which genes are expressed and at what level depends on several factors (Cohen and Kobiler, 2016). The kinetics of the viral infection play an important role, as different genes are expressed during different stages of the viral infection (Assarsson et al., 2008), and importantly, many viruses subvert the MHC processing and presentation pathway at later stages of the infection. Hence, to include source antigen expression data for viral epitope prediction, it would be necessary to know the kinetic class of the antigen of interest. The genes in each kinetic class, however, are different for each viral family and are not well known for many viruses. In addition, viral gene expression varies significantly among genetically identical cells, and the source of these variations is still not well understood (Cohen and Kobiler, 2016; Cheng et al., 2017).

AXEL-F combining NetMHCpan HLA binding predictions and RNA expression, i.e. AXEL-F (IC50), outperformed binding predictions alone in discriminating eluted ligands from random background peptides in all tested ligand elution and epitope datasets. We observed similar results when we combined MHCflurry affinity predictions and RNA expression, highlighting the general benefit of integrating antigen abundance for different HLA binding prediction tools. We furthermore wanted to improve NetMHCpan EL predictions by combining them with antigen abundance measures. EL predictors are trained with eluted ligand data and were proven to perform better than affinity predictors that are usually trained on HLA-peptide binding data. AXEL-F (EL_to_IC50) outperformed NetMHCpan EL_Rank alone in all tested epitope sets and in two of the four tested ligand elution datasets. For the two ligand elution sets Pyke Cell Line and Pyke Tissue, AXEL-F did not improve EL_Rank predictions, and AXEL-F improved MHCflurry predictions only in the Pyke Tissue set and not in the Pyke Cell Line set. More peptidomics datasets from both cell lines and tissue samples, ideally together with matched expression data, will be necessary to further validate and investigate the effect of integrating antigen abundance to better predict eluted ligands.

There have been other publications of epitope prediction tools that consider RNA expression levels and that showed significant improvements in predicting eluted ligands and neoantigens (Sarkizova et al., 2020; Bulik-Sullivan et al., 2018). Those tools, however, use complex machine learning methods that do not reveal how antigen abundance impacts ligand presentation or epitope recognition. AXEL-F, in contrast, only includes two features, HLA binding and antigen abundance. In this study, we also analyzed the performance of HLAthena, a publicly available tool that combines HLA binding prediction and RNA expression levels. Sarkizova et al. used a large dataset of ligands eluted from mono-allelic cells to train a neural network predictor (Sarkizova et al., 2020). The final model includes several features, including HLA binding, transcript expression, peptide cleavability, and gene presentation bias. The performance of AXEL-F and HLAthena was mostly comparable, with AXEL-F outperforming HLAthena on the eluted ligand datasets and the Tarke SARS-CoV-2 epitope dataset using RNA-Seq, Ribo-Seq, and Proteomic data. In contrast, HLAthena outperformed AXEL-F on the neoantigen datasets and the Peng SARS-CoV-2 epitope dataset using Proteomic data, whereas performance was lower when using RNA-Seq or Ribo-Seq on this dataset. More datasets will be necessary to further evaluate these findings.

Finally, we did not address the prediction of MHC class II-restricted epitopes presented to $CD4^+$ T cells, which play an important role in autoimmunity and antitumor immunity. Although MHC class I binding peptides are mainly derived from endogenous proteins, peptides binding MHC class II are mainly derived from extracellular proteins. Our model needs to be adjusted to describe the MHC class II antigen presentation pathway, and the cellular location of the source antigen might be one variable to consider. Unfortunately, the quality of eluted ligands from MHC class II is still lacking as ligand elution experiments from MHC class II are more complex when compared with MHC class I. Due to the open binding groove of MHC class II, ligands are variable in length, and it is challenging to deconvolute multiallelic ligand data. Recently, more computational methods to accurately deconvolute multiallelic ligand data are becoming available (Racle

et al., 2019; Alvarez et al., 2019; Reynisson et al., 2020a; 2020b) and also more ligands eluted from mono-allelic cell lines are being published. We will take advantage of these experimental and computational advances to retrieve quality datasets that we can use to train a model for MHC class II presentation.

Taken together, we have, to our knowledge, for the first time developed a biophysically motivated model to combine peptide-MHC binding and abundance of the peptide's source protein and showed that RNA-Seq, as well as Ribo-Seq and proteomics data, can be used to measure antigen abundance and improve epitope predictions. AXEL-F is freely available and should be useful for predicting and selecting epitopes more efficiently.

Limitations of the study

Our model does not take the limit of detection of MS/MS into account, and we could not investigate whether our observation that HLA ligands originated from highly expressed transcripts were biased by the limit of detection and the sensitivity of the MS/MS assays.

We utilized the interplay between source antigen abundance and peptide-MCH binding to more efficiently predict eluted ligands, neoantigens, and epitopes from SARS-CoV-2; we did, however, not investigate any other viruses. We did also not address the prediction of MHC-class II-restricted epitopes presented to $CD4^+$ T cells

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Training dataset of eluted ligands
  - Validation dataset of eluted ligands
  - Background data generation
  - Dataset of validated immunogenic neoantigens
  - TCGA expression data analysis
  - Dataset of validated SARS-CoV-2 $CD8^+$ epitopes
  - SARS-CoV-2 antigen abundance datasets
  - HLA class I binding predictions
  - Model development using the Boltzmann formalism
  - Transforming EL_Rank to IC50 values
  - Performance evaluation
  - Model training
  - Visualizations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103850.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number U24CA248138, by the National Institute of Dental & Craniofacial Research of the National Institutes of Health under award numbers U01DE028227 and R01DE027749, and by the National Institute of Allergy and Infectious Diseases (NIAID) under award number 75N93019C00001.

AUTHOR CONTRIBUTIONS

Study concept and design, Z.K.Y. and B.P.; Acquisition of Data, J.L., J.G., and Z.K.Y.; Data Analysis and Interpretation, Z.K.Y., B.P., J.L., and M.N.; Writing of the Manuscript, Z.K.Y., and J.L.; Writing—Review & Editing, B.P., A.S., S.P.S., A.M., M.N., Y.J.K.; Study Oversight, B.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 9, 2021
Revised: October 19, 2021
Accepted: January 26, 2022
Published: February 18, 2022

REFERENCES

Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells Enables more accurate epitope prediction. Immunity *46*, 315–326.

Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M., and Nielsen, M. (2019). NNAlign_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions. Mol. Cell Proteomics *18*, 2459–2477.

Assarsson, E., Greenbaum, J.A., Sundstrom, M., Schaffer, L., Hammond, J.A., Pasquetto, V., Oseroff, C., Hendrickson, R.C., Lefkowitz, E.J., Tscharke, D.C., et al. (2008). Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. Proc. Natl. Acad. Sci. U S A *105*, 2140–2145.

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P.O., Kandalaft, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput. Biol. *13*, e1005725.

Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol. Cell Proteomics *14*, 658–673.

Bezstarosti, K., Lamers, M.M., Haagmans, B.L., and Demmers, J.A.A. (2020). Targeted proteomics for the detection of SARS-CoV-2 proteins. bioRxiv.

Bhasin, M., and Raghava, G.P. (2004). Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci. *13*, 596–607.

Bjerregaard, A.M., Nielsen, M., Jurtz, V., Barra, C.M., Hadrup, S.R., Szallasi, Z., and Eklund, A.C. (2017). An analysis of natural T cell responses to predicted tumor neoepitopes. Front Immunol. *8*, 1566.

Bojkova D., Klann K., Koch B., Widera M., Krause D., Ciesek S., Cinatl J., Munch C., Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. (2020) Nature, 469-472

Bulik-Sullivan, B., Busby, J., Palmer, C.D., Davis, M.J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2018). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. Nat. Biotechnol. *37*, 55–63.

Cantarella, S., Carnevali, D., Morselli, M., Conti, A., Pellegrini, M., Montanini, B., and Dieci, G. (2019). Alu RNA modulates the expression of cell cycle genes in human fibroblasts. Int. J. Mol. Sci. *20*, 3315.

Chen, B., Khodadoust, M.S., Olsson, N., Wagar, L.E., Fast, E., Liu, C.L., Muftuoglu, Y., Sworder, B.J., Diehn, M., Levy, R., Davis, M.M., et al. (2019). Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol. 1332–1343.

Cheng, S., Caviness, K., Buehler, J., Smithey, M., Nikolich-Zugich, J., and Goodrum, F. (2017). Transcriptome-wide characterization of human cytomegalovirus in natural infection and experimental latency. Proc. Natl. Acad. Sci. U S A *114*, E10586–E10595.

Cohen, E.M., and Kobiler, O. (2016). Gene expression correlates with the number of herpes viral genomes initiating infection in single cells. PLoS Pathog. *12*, e1006082.

Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carrol, M.W., Heesom, K.J., Zambon, M., Ellis, J., Lewis, P.A., Hiscox, J.A., and Matthews, D.A. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Med. *68*.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837–845.

Dhanda, S.K., Mahajan, S., Paul, S., Yan, Z., Kim, H., Jespersen, M.C., Jurtz, V., Andreatta, M., Greenbaum, J.A., Marcatili, P., et al. (2019). IEDB-AR: immune epitope database-analysis resource in 2019. Nucleic Acids Res. *W502-W506*.

Eggers, M., Boes-Fabian, B., Ruppert, T., Kloetzel, P.M., and Koszinowski, U.H. (1995). The cleavage preference of the proteasome governs the yield of antigenic peptides. J. Exp. Med. *182*, 1865–1870.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., et al. (2021). The coding capacity of SARS-CoV-2. Nature *589*, 125–130.

Fortier, M.H., Caron, E., Hardy, M.P., Voisin, G., Lemieux, S., Perreault, C., and Thibault, P. (2008). The MHC class I peptide repertoire is molded by the transcriptome. J. Exp. Med. *205*, 595–610.

Garcia-Garijo, A., Fajardo, C.A., and Gros, A. (2019). Determinants for neoantigen identification. Front Immunol. *10*, 1392.

Garrido, F., and Algarra, I. (2001). MHC antigens and tumor escape from immune surveillance. Adv. Cancer Res. *83*, 117–158.

Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., Mcdonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the cancer cell line Encyclopedia. Nature *569*, 503–508.

Goldman, M.J., Craft, B., Hastie, M., Repecka, K., Mcdade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol. 675–678.

Hickman, H.D., Luis, A.D., Buchli, R., Few, S.R., Sathiamurthy, M., Vangundy, R.S., Giberson, C.F., and Hildebrand, W.H. (2004). Toward a definition of self: proteomic evaluation of the class I peptide repertoire. J. Immunol. *172*, 2944–2952.

Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. *15*, 205–213.

Jorgensen, K.W., Rasmussen, M., Buus, S., and Nielsen, M. (2014). NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. Immunology *141*, 18–26.

Juncker, A.S., Larsen, M.V., Weinhold, N., Nielsen, M., Brunak, S., and Lund, O. (2009). Systematic characterisation of cellular localisation and expression profiles of proteins containing MHC ligands. PLoS ONE *4*, e7448.

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.

Kim, Y., Sidney, J., Buus, S., Sette, A., Nielsen, M., and Peters, B. (2014). Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC Bioinformatics *15*, 241.

Kosaloglu-Yalcin, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premlal, A., Sidney, J., Vaughan, K., Greenbaum, J., Robbins, P., Gartner, J., Sette, A., and Peters, B. (2018). Predicting T cell recognition of MHC class I restricted neoepitopes. Oncoimmunology *7*, e1492508.

Leone, P., Shin, E.C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. J. Natl. Cancer Inst. *105*, 1172–1187.

Milner, E., Barnea, E., Beer, I., and Admon, A. (2006). The turnover kinetics of major histocompatibility complex peptides of human cancer cells. Mol. Cell Proteomics *5*, 357–365.

Moore, W.J. (1972). Phsical Chemistry (Longman Groip Limited).

Murciano-Goroff, Y.R., Warner, A.B., and Wolchok, J.D. (2020). The future of cancer immunotherapy: microenvironment-targeting combinations. Cell Res *30*, 507–519.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. Comput. J. 308–313.

Nielsen, M., Andreatta, M., Peters, B., and Buus, S. (2020). Immunoinformatics: predicting peptide–MHC binding. Annu. Rev. Biomed. Data Sci. *3*. https://doi.org/10.1146/annurev-biodatasci-021920-100259.

Nielsen, M., Lundegaard, C., Lund, O., and Kesmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics *57*, 33–41.

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

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A.M., George, N., Fexova, S., Fonseca, N.A., Fullgrabe, A., Green, M., Huang, N., et al. (2020). Expression Atlas update: from tissues to single cells. Nucleic Acids Res. *48*, D77–D83.

Parkhurst, M.R., Robbins, P.F., Tran, E., Prickett, T.D., Gartner, J.J., Jia, L., Ivey, G., Li, Y.F., El-Gamil, M., Lalani, A., et al. (2019). Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. Cancer Discov. *9*, 1022–1035.

Patronov, A., and Doytchinova, I. (2013). T-cell epitope vaccine design by immunoinformatics. Open Biol. *3*, 120139.

Paul, S., Croft, N.P., Purcell, A.W., Tscharke, D.C., Sette, A., Nielsen, M., and Peters, B. (2020). Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLoS Comput. Biol. *16*, e1007757.

Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T., Supasa, P., Liu, C., et al. (2020). Broad and strong memory CD4(+) and CD8(+) T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat. Immunol. *21*, 1336–1345.

Peters, B., Bulik, S., Tampe, R., Van Endert, P.M., and Holzhutter, H.G. (2003). Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. J. Immunol. *171*, 1741–1749.

Peters, B., Nielsen, M., and Sette, A. (2020). T cell epitope predictions. Annu. Rev. Immunol. *38*, 123–145.

Poran, A., Harjanto, D., Malloy, M., Arieta, C.M., Rothenberg, D.A., Lenkala, D., Van Buuren, M.M., Addona, T.A., Rooney, M.S., Srinivasan, L., and Gaynor, R.B. (2020). Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes. Genome Med. *12*, 70.

Pyke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., Mcclory, R., et al. (2021). Precision neoantigen discovery using large-scale immunopeptidomes and composite modeling of MHC peptide presentation. Mol. Cell Proteomics *20*, 100111.

Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., et al. (2019). Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat. Biotechnol. *37*, 1283–1286.

Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A.B., Nielsen, I.K., Nielsen, M., and Buus, S. (2016). Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. J. Immunol. *197*, 1517–1524.

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

Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209.

Schumacher, T.N., Scheper, W., and Kvistborg, P. (2019). Cancer neoantigens. Annu. Rev. Immunol. *37*, 173–200.

Sevcik, C. (2017). Caveat on the Boltzmann distribution function use in biology. Prog. Biophys. Mol. Biol. 33–42.

Shao, W., Pedrioli, P.G.A., Wolski, W., Scurtescu, C., Schmid, E., Vizcaino, J.A., Courcelles, M., Schuster, H., Kowalewski, D., Marino, F., et al. (2018). The SysteMHC Atlas project. Nucleic Acids Res. *46*, D1237–D1247.

Soria-Guerra, R.E., Nieto-Gomez, R., Govea-Alonso, D.O., and Rosales-Mendoza, S. (2015). An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J. Biomed. Inform. *53*, 405–414.

Tarke, A., Sidney, J., Kidd, C.K., Dan, J.M., Ramirez, S.I., Yu, E.D., Mateus, J., Da Silva Antunes, R., Moore, E., Rubiro, P., et al. (2021). Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases. Cell Rep Med *2*, 100204.

Trolle, T., Mcmurtrey, C.P., Sidney, J., Bardet, W., Osborn, S.C., Kaever, T., Sette, A., Hildebrand, W.H., Nielsen, M., and Peters, B. (2016). The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. J. Immunol. *196*, 1480–1487.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Uniprot, C. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *D506-D515*.

Vaughan, K., Xu, X., Caron, E., Peters, B., and Sette, A. (2017). Deciphering the MHC-associated peptidome: a review of naturally processed ligand data. Expert Rev. Proteomics *14*, 729–736.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

Gfeller, D., Guillaume, P., Michaux, J., Pak, H.S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018). The length distribution and multiple specificity of naturally presented HLA-I ligands. J. Immunol. *201*, 3705–3716.

Kosaloglu-Yalcin, Z., Blazeska, N., Carter, H., Nielsen, M., Cohen, E., Kufe, D., Conejo-Garcia, J., Robbins, P., Schoenberger, S.P., Peters, B., and Sette, A. (2021). The cancer epitope database and analysis resource: a blueprint for the establishment of a new bioinformatics resource for use by the cancer immunology community. Front Immunol. *12*, 735609.

Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020b). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 265–269.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Trolle Ligand Dataset	IEDB (Vita et al., 2019)	http://www.iedb.org/subID/1000685
HeLa RNA-Seq	GEO	GSM3899456
Abelin Ligand Dataset	Abelin et al. (Abelin et al., 2017)	Table S1
Abelin RNA-Seq	GEO	GSE93315
Pyke Ligand and Neoantigen Datasets	Pyke et al. (Pyke et al., 2021)	Tables S1 and S5
K562 RNA-Seq	Cancer Cell Line Encyclopedia (Ghandi et al., 2019)	CCLE_RNAseq_rsem_genes_tpm_20180929.txt.gz https://depmap.org/portal/download/api/download?file_name=ccle%2Fccle_2019%2FCCLE_RNAseq_rsem_genes_tpm_20180929.txt.gz&bucket=depmap-external-downloads
TCGA PANCAN RNA-Seq Dataset	TCGA	https://tcga-pancan-atlas-hub.s3.us-east-1.amazonaws.com/download/EB%2B%2BAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.xena.gz
NCI Dataset of Neoantigens	Parkhurst et al. (Parkhurst et al., 2019)	Table S3
Literature Dataset of Neoantigens	IEDB (Vita et al., 2019; Kosaloglu-Yalcin et al., 2021)	using the following filters: Epitope Structure: Linear Sequence, Included Related Structures: Only neoepitopes, Include Positive Assays, Include Negative Assays, No B cell assays, No MHC assays, MHC Restriction Type: Class I, Host: *Homo sapiens* (human), Organism: *Homo sapiens* (human) (ID:9606, human)
Tarke Dataset of SARS-CoV-2 Epitopes	Tarke et al. (Tarke et al., 2021)	Table S5
Peng Dataset of SARS-CoV-2 Epitopes	Peng et al. (Peng et al., 2020)	Table S2
SARS-CoV-2 Proteome	UniProt	UP000464024
SARS-CoV-2 RNA-Seq and Ribo-Seq data	GEO	GSE149973
SARS-CoV-2 Proteomic Dataset	Poran et al. (Poran et al., 2020)	Table S10
Software and algorithms		
NetMHCpan 4.1	Reynisson et al. (Reynisson et al., 2020b)	http://tools.iedb.org/mhci/
IEDB	Vita et al. (Vita et al., 2019)	http://tools.iedb.org/mhci/
MHCFlurry	O'Donnell et al. (O'Donnell et al., 2020)	https://openvax.github.io/mhcflurry/intro.html
MixMHCPred	Gfeller et al. (Gfeller et al., 2018)	https://github.com/GfellerLab/MixMHCpred
HLAathena	Sarkizova et al. (Sarkizova et al., 2020)	http://hlathena.tools
R	R	https://www.r-project.org
Python	Python	https://www.python.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Dr. Bjoern Peters (bpeters@lji.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. This paper does not report original code. The formula to integrate antigen abundance estimates with binding predictions from any tool is described in STAR Methods. An implementation of AXEL-F integrating NetMHCpan 4.1 predictions with antigen abundance estimates is freely available under http://axelf-beta.iedb.org/axelf. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Training dataset of eluted ligands

For our initial analysis and as the training set for model development, we used a previously published dataset of 15,090 HLA class I ligands eluted from five different HLA class I alleles: HLA-A*01:01, HLA-A*02:01, HLA-A*24:02, HLA-B*07:02, and HLA-B*51:01 (Trolle et al., 2016). We downloaded this dataset from the IEDB under the accession number 1000685 (http://www.iedb.org/subID/1000685). The length of the ligands in this set ranged from 8 to 14 residues. Eluted ligands were retrieved from 4,831 different source proteins for which UniProt identifiers were also provided.

As expression data was not included in the Trolle study, we retrieved expression data of HeLa cells from another previously published study (Cantarella et al., 2019). We downloaded raw read data from the Gene Expression Omnibus database under accession number GEO: GSM3899456 and used an in-house pipeline to process the raw RNA-Seq data and calculate gene expression as transcripts per million (TPM).

Validation dataset of eluted ligands

We retrieved a second dataset of eluted ligands to validate our findings (Abelin et al., 2017). The dataset was provided as Supplementary tables and contained 26,089 eluted ligands from 16 HLA class alleles: HLA-A*01:01, HLA-A*02:01, HLA-A*02:03, HLA-A*02:04, HLA-A*02:07, HLA-A*03:01, HLA-A*24:02, HLA-A*29:02, HLA-A*31:01, HLA-A*68:02, HLA-B*35:01, HLA-B*44:02, HLA-B*44:03, HLA-B*51:01, HLA-B*54:01, and HLA-B*57:01. Abelin et al. also performed RNA-Seq and provided the data in the form of TPM values from four cell lines (GSE93315). We averaged TPM values from those four cell lines.

We obtained two additional datasets from a recently published study (Pyke et al., 2021). Pyke et al. used mono-allelic cell lines to elute ligands from 25 different HLA. As no expression data was provided for these cell lines, we obtained pre-calculated TPM values for the corresponding cell line K562 from the Cancer Cell Line Encyclopedia (Ghandi et al., 2019). Pyke et al. also eluted ligands from 12 tissue samples of colorectal and lung cancer patients. As it was not indicated which tissue sample corresponded to which cancer type, we used the TCGA pan-cancer expression dataset to estimate antigen abundance in this set. When performing predictions, for each peptide, predictions were performed for all 6 HLA class I of the corresponding patient and the best prediction was selected for the peptide.

Background data generation

To compare the set of eluted ligands against, we sampled sets of random background peptides from the human proteome. It is common practice to utilize such decoy ligands/epitopes with unknown recognition status and expression level to test the performance of models in detecting ligands/epitopes (Bulik-Sullivan et al., 2018; Chen et al., 2019; Racle et al., 2019; Sarkizova et al., 2020). For each peptide in the training or validation dataset, 10 peptides were randomly picked from the human proteome. The lengths of the random peptides and the assignment of HLA class I alleles were chosen in a way that the total number of background peptides was uniformly distributed across all alleles and peptide lengths.

For the 15,090 ligands in the Trolle dataset, a total of 15,090 * 10 = 150,900 random background peptides needed to be generated. To be uniformly distributed over the 7 length options 8-14, 150,900 / 7 = 21,557 peptides of each length were generated. To be uniformly distributed over the 5 alleles, 150,900 / 5 = 30,180 peptides were assigned to each allele. As a result, for each length:allele combination, 150,900 / 7 / 5 = 4,312 random background peptides were generated. While picking random peptides, we ensured that no known ligands were selected. The same procedure was applied to generated random background peptides for the Abelin dataset.

Dataset of validated immunogenic neoantigens

We used data from a previously published study by Parkhurst et al. (Parkhurst et al., 2019) that reported immunogenicity screening results of neoantigens from 75 patients with various gastrointestinal cancers. The group performed whole-exome sequencing to detect somatic mutations, transfected autologous dendritic cells with tandem minigenes encoding these mutations, and determined which neoantigens were recognized by tumor-infiltrating-lymphocyte cultures (Parkhurst et al., 2019). The results were provided as a supplemental table to the study, listing all tested neoantigens and corresponding screening results ($CD4^+$ and/or $CD8^+$ or negative, (Supplementary Table S3 in original publication (Parkhurst et al., 2019)).

The peptides provided in this dataset were mainly 29mer peptides with the mutated residue located in the center of the peptide. When performing predictions, we considered all contained 8-14mer peptides and all HLA class I alleles provided for the corresponding patient. For each peptide, we then assigned the best prediction scores among all its k-mer and HLA combinations.

We retrieved an additional set of validated neoantigens from the literature using the IEDB (Kosaloglu-Yalcin et al., 2021; Vita et al., 2019). We queried the database in July 2021 using the following filters: Epitope Structure: Linear Sequence, Included Related Structures: Only neoepitopes, Include Positive Assays, Include Negative Assays, No B cell assays, No MHC assays, MHC Restriction Type: Class I, Host: *Homo sapiens* (human), Organism: *Homo sapiens* (human) (ID:9606, human). After downloading the results, we only retained peptides for which the cancer type and the HLA restriction was known. This dataset is provided as Supplementary Table S7.

TCGA expression data analysis

We downloaded pre-calculated TPM values for the TCGA Pan-cancer cohort from UCSC Xena data pages (Goldman et al., 2020). For each of the 35 cancer types included, we calculated the median expression for each gene across all samples. We utilized in-house RNA-Seq data of 25 patients with 9 different cancer to analyze how well patient-specific gene expression can be estimated by using the gene expression data obtained from TCGA. For each patient from our in-house cohort, we matched the TPM from in-house RNA-Seq with the calculated cancer type-specific median TPM values from TCGA. We then analyzed, how well TCGA median TPM values correlated with patient-specific TPM values.

Dataset of validated SARS-CoV-2 $CD8^+$ epitopes

We used data from a previously published study by Tarke et al. (Tarke et al., 2021) that reported a comprehensive map of epitopes recognized by $CD4^+$ and $CD8^+$ T cell responses across the entire SARS-CoV-2 viral proteome. Tarke et al. used NetMHCpan to predict binding for the most common 28 HLA class I and synthesized the top 200 predicted binders for each allele for experimental validation. As a result, the group reported 523 HLA class I epitopes that elicited $CD8^+$ responses (Table S5 in original publication). We retrieved the list of 523 epitopes (Table S8 in the original publication) to be used as an additional validation dataset. This dataset is provided in Table S7.

As a second dataset, we used data from another previously published study (Peng et al., 2020). Peng et al. tested 18-mer peptides spanning the SARS-CoV-2 proteome and identified 11 unique peptide-HLA with confirmed $CD8^+$ T cell responses (Supplementary Table S2 in original publication). We used these 11 peptide-HLA as positives and the remaining 18-mer peptides spanning the SARS-CoV-2 proteome together with the 36 HLA from the cohort used in the study as negatives. We downloaded the SARS-CoV-2 proteome (UniProt: UP000464024) from the UniProt database (UniProt, 2019; Wu et al., 2020) and generated all overlapping 18-mer peptides. When performing predictions, we considered all contained 8-14mer peptides and all 36 HLA class I alleles provided for the cohort. For each 18-mer peptide we then assigned the best prediction score among all its k-mer and HLA combinations.

SARS-CoV-2 antigen abundance datasets

We retrieved SARS-CoV-2 expression data from another previously published study. Finkel et al. (Finkel et al., 2021) transfected Vero E6 cells with SARS-CoV-2 and performed RNA-Seq as well as ribosome profiling (Ribo-Seq). Ribo-seq libraries were prepared from cells treated with the translation elongation inhibitor cycloheximide (CHX) and provide a snapshot of actively translating ribosomes across the body of the translated viral ORFs. We downloaded raw read data from the Gene Expression Omnibus database

under accession number GEO: GSE149973 and used an in-house pipeline to process the raw RNA-Seq and Ribo-Seq data and calculate gene expression as transcripts per million (TPM). The genomic sequence of SARS-CoV-2 (RefSeq: NC_045512.2) was used as the reference.

In addition to RNA-Seq and Ribo-Seq, we additionally wanted to evaluate SARS-CoV-2 proteomic datasets to measure the abundance of viral antigens. Poran et al. (Poran et al., 2020), obtained three publicly available proteomic datasets (Bezstarosti et al., 2020; Bojkova et al., 2020; Davidson et al., 2020) and used a custom pipeline to re-analyze the datasets. Datasets were searched against the SARS-CoV-2 proteome and peptide-spectrum matches (PSMs) for SARS-CoV-2 proteins were provided (Table S10 in original publication). We used the 'PSMs / Length – Relative' value provided in that table as a measure for SARS-CoV-2 antigen abundance.

HLA class I binding predictions

NetMHCpan version 4.1 as hosted on the IEDB Analysis Resource (IEDB-AR) was used to perform binding predictions (Dhanda et al., 2019; Jurtz et al., 2017). We also used MHCFlurry 2.0 (O'Donnell et al., 2020), MixMHCpred 2.0 (Bassani-Sternberg et al., 2017; Gfeller et al., 2018) and HLAthena (Sarkizova et al., 2020) to perform predictions.

Model development using the Boltzmann formalism

We developed a model to capture the effect of quantitative expression differences and combine them with peptide-HLA binding using the Boltzmann distribution, which is often used to describe biophysical systems (Moore, 1972). The Boltzmann distribution is a probability distribution that predicts, in an ensemble of particles, the proportion of particles that will be in a certain state with a specific energy (Sevcik, 2017). This function can be adapted to describe peptide presentation, as we want to detect, among all available peptides, the ones that are in a state bound to HLA with a specific binding free energy. We adapted the Boltzmann formalism by empirically adding parameters to describe peptide presentation.

In this context, the number of all available peptides of a certain species was considered proportional to the RNA expression values of its source protein (TPM), and the binding free energy can be inferred from binding affinities (IC50). Combining these considerations with the Boltzmann distribution function yields:

$$\#\text{peptides} = \alpha * \text{TPM} * e^{-\log(IC50)/kT}$$

Where α is a scaling factor for TPM and kT is a scale that mimics the product of the Boltzmann's constant k and the thermodynamic temperature T, as adapted from the original Boltzmann distribution function. To account for the detection limit of RNA-Seq, we additionally introduced a parameter minTPM and modified the function to select for the higher value between minTPM and the input TPM value:

$$\#\text{peptides} = \alpha * \max(\text{minTPM},\ \text{TPM}) * e^{-\log(IC50)/kT}$$

This function will estimate the number of peptides for a given species that are bound to MHC. We want to know the likelihood of finding at least one of these peptides when performing a mass spectrometry experiment and/or when a T cell scans a cell:

$$P(p>0 \mid \#\text{peptides}) = 1 - e^{-\ \#\text{peptides}}$$

Our final model estimates, for a given peptide with IC50 and TPM value, its likelihood of being presented on HLA and being an epitope. We named this model AXEL-F, standing for **A**ntigen e**X**pression based **E**pitope **L**ikelihood-**F**unction.

Transforming EL_Rank to IC50 values

EL_Score and EL_Rank are output values of the neural networks the NetMHCpan method consists of. These values are abstract and cannot be directly used in the biological context of our model like IC50. We therefore translated the EL_Rank values to IC50 values by comparing the percentile ranks of the two metrics in the Trolle dataset. To do so, we first calculated the global percentile rank of each IC50 value within the Trolle set. We then defined an interpolation function that maps each of these percentile ranks to the corresponding IC50 value. This interpolation function was then used to map each EL_Rank to IC50 values to obtain our new metric EL_to_IC50. The same interpolation function based in the Trolle dataset was used to

calculate AXEL-F scores for the validation datasets Abelin and NCI and the function is implemented as part of the AXEL-F method.

Performance evaluation

We used AUCs as well as partial AUCs (pAUC) to measure performance. We calculated pAUC at false positive rate (FPR) of 10% as this would be considered an acceptable FPRs in subsequent experimental validations of predicted peptides. The R package pROC was used for performing ROC analysis and calculating AUC and pAUC values, packages plotROC and ggplot 2 were used to plot ROC curves. DeLong's test was used to compare ROC curves (DeLong et al., 1988).

Model training

We used the R function optim that implements an optimization method based on Nelder–Mead (Nelder and Mead, 1965). The parameters α, minTPM, and kT were fitted concurrently to maximize the AUC value for predicting eluted ligands in the Trolle dataset. To avoid overfitting, we performed 5-fold-cross-validation: the dataset was split randomly into 5 parts using R package caret, the optimization was performed for all parameters concurrently on 4/5 of the data and tested on the remaining 1/5 of the data by calculating AUC. This was done 10 times and the median of the fitted parameters was used for α, minTPM, and kT.

The parameters obtained in this way fell into a consistent range between the five subsets corresponding to the five alleles in the Trolle set (Table S3).

Visualizations

All figures in the results and supplemental sections were generated using R and Python software. The graphical abstract was created using BioRender.com.

QUANTIFICATION AND STATISTICAL ANALYSIS

R software was used to perform data and statistical analyses. Statistical details are provided in the respective figure legends. Statistical analyses were performed using Wilcoxon tests and DeLong's tests for comparing ROC curves. Details pertaining to significance are also noted in the respective figure legends.