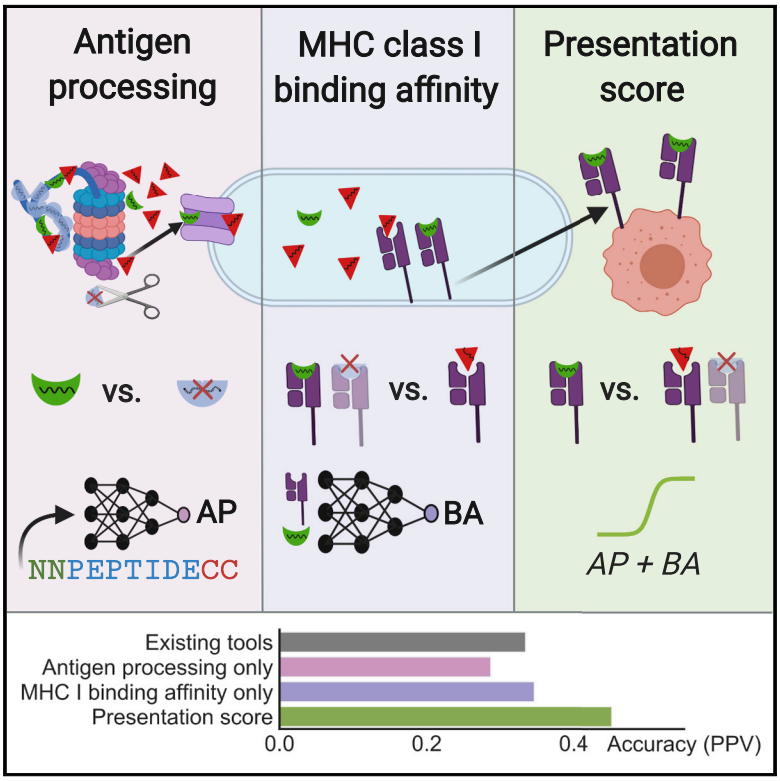
Cell Systems

Methods

MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing

Graphical Abstract



Authors

Timothy J. O'Donnell, Alex Rubinsteyn, Uri Laserson

Correspondence

tim@openvax.org

In Brief

O'Donnell et al. developed improved models for predicting the antigens available for recognition by cytotoxic T cells. Separate predictors of MHC class I binding and antigen processing were trained using published datasets of peptides naturally presented on MHC. The software is open source and readily incorporated into workflows for neoantigen discovery and vaccine design.

Highlights

- New pan-allele MHC class I binding predictor
- Antigen processing predictor trained on mass spectrometry-identified MHC ligands
- Combined model outperforms existing methods
- Open source Python package with command line and library interfaces

O'Donnell et al., 2020, Cell Systems *11*, 42–48
July 22, 2020 © 2020 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.06.010

CellPress

CellPress
OPEN ACCESS

Cell Systems

Methods

MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing

Timothy J. O'Donnell,[1,5,*] Alex Rubinsteyn,[2,3] and Uri Laserson[1,4]
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3]Department of Genetics, UNC School of Medicine, Chapel Hill, NC 27599, USA
[4]Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[5]Lead Contact
*Correspondence: tim@openvax.org
https://doi.org/10.1016/j.cels.2020.06.010

SUMMARY

Computational prediction of the peptides presented on major histocompatibility complex (MHC) class I proteins is an important tool for studying T cell immunity. The data available to develop such predictors have expanded with the use of mass spectrometry to identify naturally presented MHC ligands. In addition to elucidating binding motifs, the identified ligands also reflect the antigen processing steps that occur prior to MHC binding. Here, we developed an integrated predictor of MHC class I presentation that combines new models for MHC class I binding and antigen processing. Considering only peptides first predicted by the binding model to bind strongly to MHC, the antigen processing model is trained to discriminate published mass spectrometry-identified MHC class I ligands from unobserved peptides. The integrated model outperformed the two individual components as well as NetMHCpan 4.0 and MixMHCpred 2.0.2 on held-out mass spectrometry experiments. Our predictors are implemented in the open source MHCflurry package, version 2.0 (github.com/openvax/mhcflurry).

INTRODUCTION

Cytotoxic (CD8$^+$) T cells recognize peptides presented in complex with major histocompatibility (MHC) class I molecules on cell surfaces. These peptides are usually derived from the degradation of endogenous proteins and comprise a snapshot of the protein content of the cell, enabling T cells to distinguish healthy cells from those with viral, bacterial, or tumor-associated mutated proteins (Rossjohn et al., 2015; Schumacher and Schreiber, 2015). The repertoire of MHC class I-presented peptides is generated through a complex series of biochemical processes, beginning with cleavage of a protein into peptides in the proteasome, further cleavage (or destruction) by cytosolic peptidases, peptide transport into the endoplasmic reticulum (ER) through the transporter associated with antigen processing (TAP) complex, trimming by ER-resident aminopeptidases (ERAP), and stable association with one of the several MHC class I proteins expressed by a cell (Rock et al., 2016). The MHC class I genes (*HLA-A, HLA-B*, and *HLA-C* in humans) are the sites of dramatic allelic variation at the population level, with each of the thousands of known *HLA* alleles associated with a strict, potentially distinct peptide-binding preference. While a high-affinity interaction with MHC class I is the most selective requirement for a peptide to be presented, the other processes in the antigen presentation pathway likely exert important secondary effects.

Prediction of MHC class I-presented peptides is a critical tool in vaccine design and studies of infectious disease, autoimmunity, and cancer (Peters et al., 2020). Most predictive pipelines in use today focus only on MHC class I binding affinity (BA) prediction. While predictors fit to small datasets for individual antigen processing (AP) steps have been proposed (Bhasin et al., 2007; Keşmir et al., 2002; Nielsen et al., 2005; Peters et al., 2003) and integrated with BA predictions to give a composite score (Larsen et al., 2005; Stranzl et al., 2010; Tenzer et al., 2005), improvements in accuracy from these approaches have been modest at best (Koşaloğlu-Yalçın et al., 2018). The relatively recent accumulation of large datasets of mass spectrometry (MS)-identified MHC class I ligands provides an opportunity to revisit AP prediction using larger and potentially more biologically relevant datasets. While the AP information in these datasets likely already informs existing MHC class I binding predictors trained on MS datasets (Gfeller et al., 2018; Jurtz et al., 2017), AP remains intertwined with MHC class I binding preferences in these predictors, making it difficult to interpret the individual contributions and potentially leading to lower predictive accuracy.

In this work, we develop separate predictors for MHC allele-dependent effects (BA prediction) and allele-independent effects (AP prediction). We first trained a new pan-allele MHC class I BA predictor (referred to as MHCflurry 2.0 BA) on available MHC class I ligand data, including affinity measurements and MS

 © 2020 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

CellPress
OPEN ACCESS

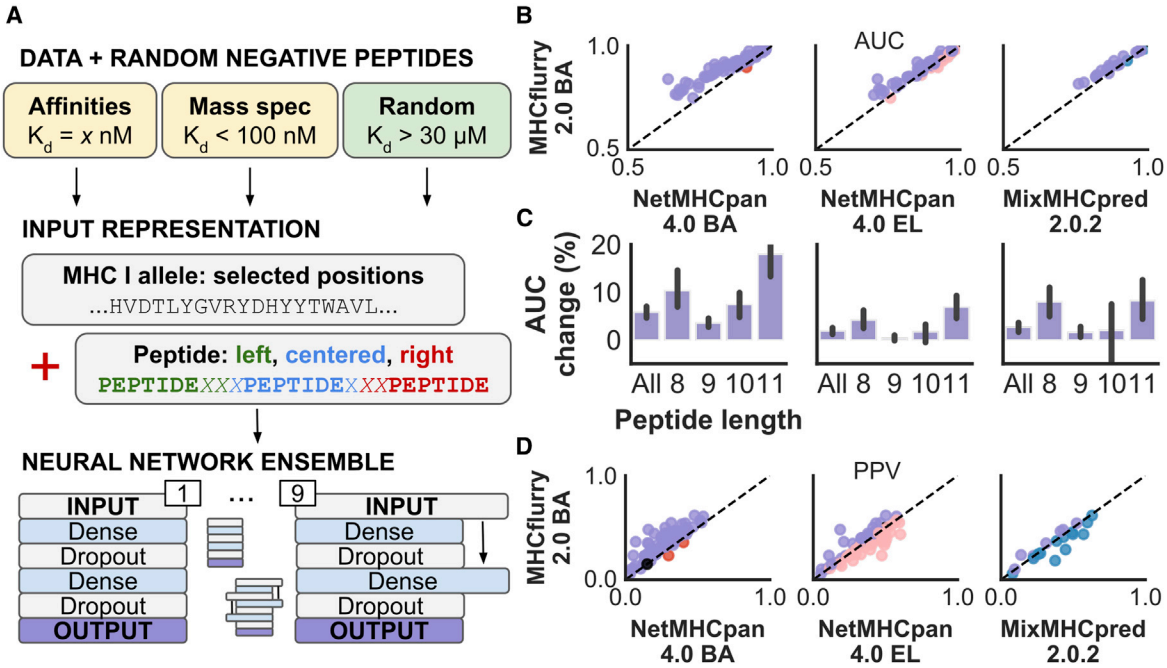


Figure 1. MHCflurry 2.0 BA Predictor Architecture and Benchmark

(A) BA predictor training data, model input representations, and neural network architectures.

(B) AUC of MHCflurry 2.0 BA in comparison to other predictors across the MULTIALLELIC benchmark experiments. Each point corresponds to a single experiment from the MULTIALLELIC (NetMHCpan 4.0 BA and EL) or MULTIALLELIC-RECENT (MixMHCpred 2.0.2) benchmarks.

(C) Mean improvement in AUC across experiments per peptide length. Error bars are bootstrap 95% confidence intervals of the mean.

(D) PPV of MHCflurry 2.0 BA in comparison to other predictors. Points correspond to the same experiments as in (B).

datasets. The use of *in vitro* affinity measurements in the training data, which are largely independent of AP, is one of several design choices intended to limit the BA predictor's tendency to learn AP signals. We use the BA predictor to generate a training set for a model of AP by combining MS-identified peptides (hits) with unobserved peptides (decoys), where both hits and decoys are predicted by the BA predictor to bind the relevant MHC class I alleles. The AP predictor thus models the residual allele-independent sequence properties that were not learned by the BA predictor. In support of its biological relevance, the processing predictor favored peptides consistent with established motifs for key AP steps and showed quantitative agreement with an independent dataset of proteasome-cleaved peptides (Wolf-Levy et al., 2018). We combined the BA and AP predictors in a logistic regression model, which we refer to as the presentation score (PS) (MHCflurry 2.0 PS). Using a benchmark of held-out MS datasets, we found that the PS predictor outperformed both its component models and the commonly used NetMHCpan 4.0 and MixMHCpred 2.0.2 predictors. The margin of improvement was substantial, with at least a 40% increase in positive predictive value (PPV) for all comparisons.

RESULTS

We first tested the new MHCflurry 2.0 BA predictor (Figure 1A) on affinity measurements and MS-identified MHC ligands held-out from its training data. The predictor showed good performance for most alleles, with 214 of 236 (91%) alleles having an area under the curve (AUC) score of at least 0.90 (Figure S1; Table S1). To compare against existing BA (NetMHCpan 4.0 BA) and MS ligand (NetMHCpan 4.0 eluted ligand [EL] and MixMHCpred 2.0.2) predictors, we compiled a benchmark using published datasets of MS-identified MHC ligands (Tables S2 and S3; Data S1). As the peptides identified in these experiments may have bound any of the up to six classical MHC class I alleles expressed in an individual, we refer to this as the MULTIALLELIC benchmark. MHCflurry BA performed best in terms of AUC at differentiating MS hits from decoy peptides sampled from the same proteins, outperforming NetMHCpan BA on 75 of 76 samples, NetMHCpan EL on 56 of 76, and MixMHCpred on 18 of 20 samples in the MULTIALLELIC-RECENT subset (Figure 1B). The increase in AUC relative to NetMHCpan BA was 5.8% (bootstrap 95% confidence interval 4.8–6.8), 1.8% (1.3–2.4) relative to NetMHCpan EL, and 2.6% (1.8–3.4) relative to MixMHCpred, with the greatest improvements observed for non-9-mer peptides (Figure 1C).

We expanded our evaluation to include a set of 100 MS experiments that used cell lines genetically engineered to express a single MHC class I allele (Abelin et al., 2019; Sarkizova et al., 2020), which we refer to as the MONOALLELIC benchmark (Table S2; Data S2). As this dataset is normally incorporated into the training set for the MHCflurry BA predictor (Table S4), we tested a variant of MHCflurry BA trained without this data. We again found that MHCflurry BA outperformed the other tools in terms of AUC (Figure S2A). Notably, good accuracy (AUC > 0.94) and a small improvement over existing tools was observed for the

11 alleles that had no representation in the training data (Figure S2B; Table S5). Performance on these alleles slightly exceeded the accuracy obtained by substituting predictions for similar well-characterized alleles, with a mean improvement of 0.42% (0.14–0.72) in AUC (Figure S2C). These observations suggest that MHCflurry BA has some capacity to generalize to alleles that are not experimentally characterized.

We next returned to the MULTIALLELIC benchmark and evaluated the predictors in terms of PPV, which emphasizes differences at the high end of a predictor's output. In contrast to the AUC evaluation, MHCflurry BA showed only a PPV advantage in comparison to NetMHCpan BA, with a mean improvement in PPV of 48% (37–62). Differences were within error in comparison to the MS ligand predictors, with a mean 1.9% (−3.4– +7.6) improvement over NetMHCpan EL and a trend toward underperforming MixMHCpred, with a mean difference of −4.0% (−12.6–+5.6; Figure 1D). Similar results were observed for PPV on the MONOALLELIC benchmark (Figure S2D).

We hypothesized that explicitly modeling processes that do not depend on MHC class I allele might enable accuracy improvements over MHCflurry BA alone. We therefore developed an MHC class I allele-independent model trained to distinguish hits from decoys where both the hits and decoy peptides are predicted to be tight binders (rank less than 2%) by the MHCflurry BA predictor. We refer to this model as the MHCflurry 2.0 AP predictor. Its neural network architecture is motivated by the possibility of learning peptide N- and C-terminal cleavage or processing signals (Figure 2A). We trained two versions of the AP predictor on the MONOALLELIC benchmark dataset. One predictor includes the peptide plus the five immediately upstream and downstream residues from its source protein (AP with flanks); the second predictor includes only the peptide (AP without flanks).

To understand if the MHCflurry AP predictors learned a meaningful signal, we evaluated their accuracy on the MULTIALLELIC benchmark (Figure 2B). While the AP variants underperformed the standard MHC binding predictors and MHCflurry BA, they performed better than might be expected given that they do not take MHC allele as an input. The AP without flanks and AP with flanks predictors had mean AUCs of 0.85 (bootstrap 95% CI 0.84–0.87) and 0.86 (0.84–0.87), respectively, compared with 0.91 (0.90–0.92) for the MHCflurry BA predictor (Figure 2B). This suggested that the MHCflurry AP predictors had learned a meaningful MHC class I allele-independent signal from the MONOALLELIC MS training set.

To understand what the MHCflurry AP predictors may be learning, we ranked all 9-mer peptides in the MULTIALLELIC benchmark by AP prediction, calculated position weight matrices for the top 1% highest predictions, and plotted a sequence logo (Figure 2C; Table S6). This analysis showed the AP predictor learned that hits are depleted for cysteines across the peptide, a known bias of MS (Abelin et al., 2017). It also showed depletion of prolines from the first position in the peptide (P1) extending upstream into the N-flank as well as strong but complex specificity at the C terminus of the peptide and some preferences at the downstream flanking residue. These signals suggested qualitative agreement with established signatures for TAP transport, proteasomal cleavage, and ERAP trimming (see Discussion).

As cysteine depletion is one of the strongest known MS biases, we were concerned that much of the accuracy of the AP predictor may be due to its modeling of this bias. We therefore repeated the AUC analysis on the MULTIALLELIC benchmark after removing peptides (hits and decoys) that contained cysteine. This analysis showed very similar AUC values as the earlier analysis, with a change in AUC of less than 3% for all samples (Figure 2D). Thus, while the AP predictor learns the cysteine MS bias, this effect alone is not the primary driver of its performance.

To evaluate the extent to which the AP predictors learned a signal that is also learned by the MHCflurry BA predictor, we measured the correlation between AP and BA predictions for random peptides across the *HLA-A*, *HLA-B*, and *HLA-C* alleles in the BA training set (n = 183). The correlations were positive, significant but modest in magnitude (Pearson $r < 0.5$) for all alleles tested. For example, AP and BA predictions for the HLA-A*02:01 allele showed a Pearson correlation of 0.23. While peptides predicted to bind HLA-A*02:01 tightly tended to have higher AP scores, it was possible to find peptides with high scores for one predictor but not the other (Figure 2E). Correlations with the AP predictor were somewhat higher for *HLA-B* and *HLA-C* than for *HLA-A* alleles (Figure 2F). The alleles used to train the AP predictor showed no greater AP versus BA correlations on average than those that were not included in the AP training set; in fact the correlation was slightly lower for the alleles used in training (Figure 2G). Overall, this analysis suggested that the AP predictor is at least partially non-redundant with the MHCflurry BA predictor.

To quantitatively assess if the AP predictor captures biologically relevant effects, we tested it on an independent dataset of proteasome-cleaved peptides. We applied the AP predictor to 1,079 peptides identified by Wolf-Levy et al. using the "mass spectrometry analysis of proteolytic peptides" (MAPP) assay, in which cleaved peptides are reversibly cross linked to cellular proteasomes and identified by MS (Wolf-Levy et al., 2018). AP predictor scores were significantly higher for MAPP-identified peptides (hits) than unobserved (decoy) peptides drawn from the same genes (Mann-Whitney $p < 0.01$ for each peptide length 8–11; Figure 2H; Table S7). This indicated that the AP predictor learned a signal consistent with a key AP step.

We next asked if the AP predictor may be combined with the MHCflurry BA predictor to achieve higher performance than either alone. We trained a logistic regression model that takes two inputs: the strongest MHCflurry BA prediction across alleles for the sample (transformed to fall in the range 0.0 –1.0, with higher indicating a stronger binder), and the AP prediction (Figure 3A). We refer to this logistic regression model as the MHCflurry PS predictor. We trained MHCflurry PS predictors using either the AP with flanking (PS with flanking) or AP without flanking (PS without flanking) predictors on the MULTIALLELIC-OLD dataset and evaluated performance on the MULTIALLELIC-RECENT benchmark.

Both PS models showed significantly improved accuracy, in terms of both AUC and PPV, over all other predictors tested (Figures 3B and 3C). In terms of PPV, the PS without flanks predictor showed a 120% (65–184) improvement over NetMHCpan 4.0 BA, 56% (22–95) improvement over NetMHCpan 4.0 EL, 41% (22–64) improvement over MixMHCpred 2.0.2, and 51% (29–78) improvement over MHCflurry 2.0 BA. The use of flanking sequences made a small but consistent difference, further improving PPV by a mean 2.1% (0.14–3.9).

CellPress
OPEN ACCESS

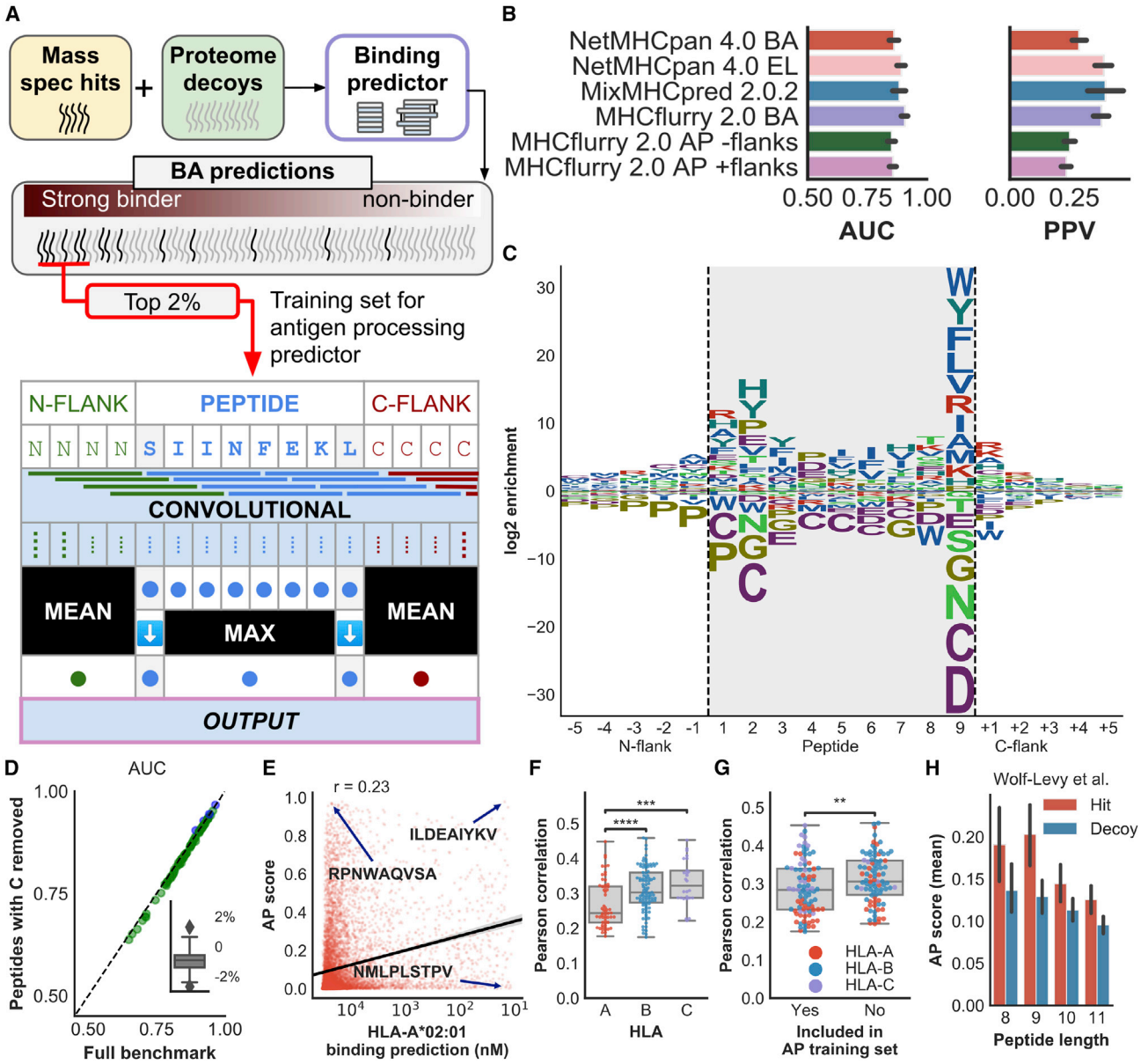


Figure 2. The MHCflurry 2.0 AP Predictor Models MHC Class I Allele-Independent Effects

(A) AP predictor training scheme and neural network architecture.

(B) Mean AUC and PPV accuracy on the multiallelic MS benchmarks for the AP predictor in comparison to other predictors. The MixMHCpred 2.0.2 tool was benchmarked on the MULTIALLELIC-RECENT subset; the other predictors were tested on the full MULTIALLELIC benchmark.

(C) Sequence logo for the motif learned by the AP predictor. Positive values (above the center line) indicate enrichments above proteome level; negative values indicate depletions.

(D) Comparison of AP predictor AUC scores on the MULTIALLELIC benchmark samples when peptides containing cysteine were removed (y axis) versus retained (x axis). The inset shows the percentage change in AUC when cysteine-containing peptides are removed.

(E) Correlation of the AP predictor with the BA prediction for HLA-A*02:01. Each red dot corresponds to a 9-mer peptide sampled from the proteome. The black line indicates the best-fit regression line, and example peptides are indicated.

(F and G) Correlation of AP predictor with the BA prediction across HLA alleles by gene (F) or by allele representation in the AP training set (G).

(H) Mean AP score for proteasome-associated peptides observed by Wolf-Levy et al. (Wolf-Levy et al., 2018) (hits) and unobserved peptides from the same genes (decoys).

DISCUSSION

Our MHC class I ligand prediction method consists of two neural network models: the MHCflurry BA predictor and the MHC class I allele-independent MHCflurry AP predictor. The AP predictor is trained to learn what the BA predictor missed, i.e., residual sequence properties that distinguish MS hits from decoys among peptides predicted to bind MHC class I tightly. Both

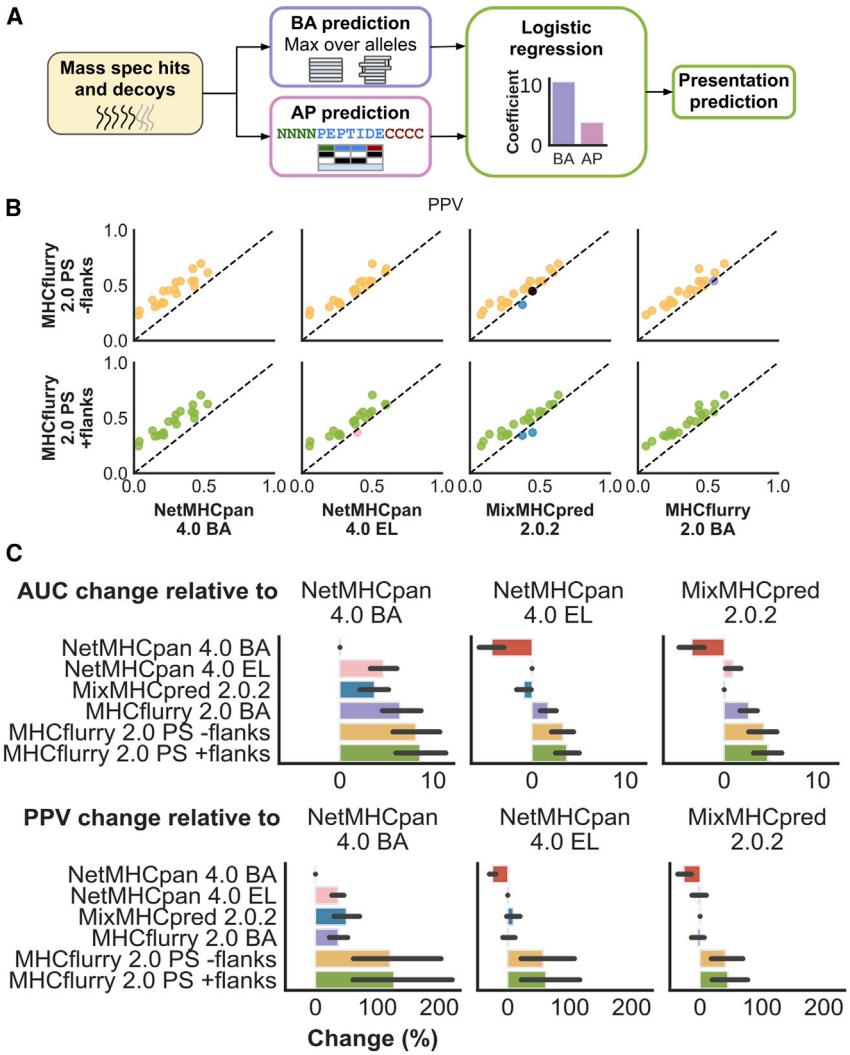


Figure 3. The MHCflurry 2.0 PS Model Combines BA and AP Prediction

(A) The PS model is a two-input logistic regression model that integrates BA and AP predictions to give a composite score. It is trained on multiallelic MS hits and decoys.

(B) Comparison of PPV scores of the PS models with other predictors. The top row shows the performance of the PS model variant that uses the AP without flanks predictor, which considers only the peptide and not the upstream and downstream flanking sequences. The bottom row corresponds to the variant that also considers the flanking sequences.

(C) Mean percent change in AUC and PPV for the indicated predictors (y axis) relative to each of the three existing predictors (columns). Error bars indicate 95% confidence intervals for the mean change.

predictors are trained on monoallelic MS datasets (plus affinity measurements for the BA predictor), and their results combined using a logistic regression model fit to multiallelic MS datasets. When evaluated on held-out multiallelic MS experiments, the combined predictor, referred to as MHCflurry PS, outperformed the individual components and standard tools. Although inclusion of flanking sequences, i.e., the adjacent residues in the peptide's source protein, provided a small additional accuracy boost, the overall improvement over standard tools was also evident when only the peptide was provided to the AP predictor.

In comparison to NetMHCpan 4.0, the MHCflurry BA predictor uses larger neural networks with two or three hidden layers and over one million trained parameters. It also benefits from additional training data that have been published since the release of NetMHCpan. In preliminary experiments, we found that the deeper networks could consistently outperform a shallow (one hidden layer) version of MHCflurry BA, but that differences were modest in comparison to the improvement from incorporating additional training data, especially when such data expanded the number of represented alleles. We therefore expect that most of the improvement in accuracy of MHCflurry 2.0 BA is due to incorporation of recently published datasets, such as that of Sarkizova et al. (Sarkizova et al., 2020), rather than the deeper neutral network architecture.

Our work builds on the approach by Abelin et al., who developed a proteasomal cleavage predictor fit to MS-identified MHC class I ligands and observed an increase in accuracy when it was included along with other features in a logistic regression model (Abelin et al., 2017). Their work used a different method to control for MHC class I binding signals in the cleavage predictor training set: decoys were selected to match the first two and last two positions of the hit peptide, which encompass the anchor positions for most alleles. This is expected to disrupt the cleavage model's ability to learn features at these positions, which are also the positions where we observed the strongest preferences.

The AP predictor sequence motif (Figure 2C) shows similarities with the established preferences of key AP steps. Deconvolution of effects, however, is complicated by the overlapping specificity of these steps, potentially a consequence of coevolution of the AP machinery (Nielsen et al., 2005). In particular, the AP predictor's preferences at the C terminus of the peptide may reflect TAP binding and/or proteasomal cleavage. Work by Tampé and colleagues (Uebel et al., 1997) found that TAP favors peptides with a C-terminal Phe, Tyr, Arg, or Leu and disfavors Asp, Glu, Asn, and Ser, all recapitulated by the AP predictor. The AP motif is also consistent with the effects of the chymotryptic-like activity of the proteasome (cleavage after Phe, Tyr, Leu, Trp but not Gly) as well as tryptic-like (cleavage after Arg and Lys), but not caspase-like specificity(Harris et al., 2001; Nussbaum et al., 1998). Some agreement with proteasomal processing is also apparent at the interior residues of the peptide, such as a depletion of proline at P8 and a slight enrichment for proline at P6 (referred to as P2 and P4, respectively, in cleavage studies). At the first flanking residue downstream of the peptide, which is free from the effects of TAP and MHC binding, the

CellPress
OPEN ACCESS

enrichments for Arg, Lys, Ala, Ser, and Gly are consistent with proteasomal cleavage preferences for the position after the cut site (Nussbaum et al., 1998; Wolf-Levy et al., 2018), although the enrichment for Arg and Lys could also indicate tryptic-like specificity working from the C terminus of the protein (Abelin et al., 2017). The striking depletion of prolines in the N-flanking residues up to and including the first position of the peptide (P1) and the enrichment for proline at P2 is consistent with trimming by ERAP (Serwold et al., 2002), although again we cannot exclude a contribution from proteasomal cleavage. These observations and the higher AP scores for proteasome-cleaved peptides identified by MAPP (Figure 2H) suggest that the AP predictor has learned certain AP signals, although a detailed deconvolution of effects remains future work.

An important limitation of this work is that we apply datasets of MHC class I ligands detected by MS both to train and to benchmark our predictors. Assay biases, which we expect are modeled by the AP predictor, have the potential to erroneously inflate our accuracy scores. Although the main known bias, depletion of cysteine, does not seem to have a dramatic effect on AP predictive accuracy, we cannot rule out the contributions of other kinds of bias. Our work also only addresses the steps contributing to MHC class I ligand presentation, not T cell recognition of presented epitopes. Future work will need to assess whether the predictors described here enable improved prediction of T cell epitopes.

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
  - MS Benchmark Construction and Approach
  - MHC Class I Binding Affinity Predictor
  - Antigen Processing Predictor
  - Presentation Score Model
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Accuracy Metrics
  - Statistical Significance

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.06.010.

ACKNOWLEDGMENTS

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. The graphical abstract was created with BioRender.com.

AUTHOR CONTRIBUTIONS

T.J.O. designed research, performed research, and wrote the paper. A.R. and U.L. advised on data interpretation. All authors critically reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 16, 2020
Revised: June 16, 2020
Accepted: June 19, 2020
Published: July 14, 2020; corrected online September 15, 2020

REFERENCES

Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. Immunity *51*, 766–779.e17.

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.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Barry, K.C., Hsu, J., Broz, M.L., Cueto, F.J., Binnewies, M., Combes, A.J., Nelson, A.E., Loo, K., Kumar, R., Rosenblum, M.D., et al. (2018). A natural killer-dendritic cell axis defines checkpoint therapy-responsive tumor microenvironments. Nat. Med. *24*, 1178–1191.

Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., et al. (2016). Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat. Commun. *7*, 13404.

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.

Bhasin, M., Lata, S., and Raghava, G.P.S. (2007). TAPPred prediction of TAP-binding peptides in antigens. Methods Mol. Biol. *409*, 381–386.

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.

Gloger, A., Ritz, D., Fugmann, T., and Neri, D. (2016). Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. Cancer Immunol. Immunother. *65*, 1377–1393.

Harris, J.L., Alper, P.B., Li, J., Rechsteiner, M., and Backes, B.J. (2001). Substrate specificity of the human proteasome. Chem. Biol. *8*, 1131–1141.

Hassan, C., Kester, M.G.D., de Ru, A.H., Hombrink, P., Drijfhout, J.W., Nijveen, H., Leunissen, J.A., Heemskerk, M.H.M., Falkenburg, J.H.F., and van Veelen, P.A. (2013). The human leukocyte antigen-presented ligandome of B lymphocytes. Mol. Cell. Proteomics *12*, 1829–1843.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA *89*, 10915–10919.

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.

Keşmir, C., Nussbaum, A.K., Schild, H., Detours, V., and Brunak, S. (2002). Prediction of proteasome cleavage motifs by neural networks. Protein Eng *15*, 287–296.

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.

Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. arXiv arxiv.org/abs/1412.6980.

Koşaloğlu-Yalçın, 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.

Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur. J. Immunol. *35*, 2295–2303.

Maccari, G., Robinson, J., Ballingall, K., Guethlein, L.A., Grimholt, U., Kaufman, J., Ho, C.S., de Groot, N.G., Flicek, P., Bontrop, R.E., et al. (2017). IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res *45*, D860–D864.

Mishkin, D., and Matas, J. (2015). All you need is a good init. arXiv arxiv.org/abs/1511.06422.

Mommen, G.P.M., Frese, C.K., Meiring, H.D., van Gaans-van den Brink, J., de Jong, A.P.J.M., van Els, C.A.C.M., and Heck, A.J.R. (2014). Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). Proc. Natl. Acad. Sci. USA *111*, 4507–4512.

Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, 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.

Nussbaum, A.K., Dick, T.P., Keilholz, W., Schirle, M., Stevanović, S., Dietz, K., Heinemeyer, W., Groll, M., Wolf, D.H., Huber, R., et al. (1998). Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. Proc. Natl. Acad. Sci. USA *95*, 12504–12509.

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

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

Pearson, H., Daouda, T., Granados, D.P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J.P., Côté, C., Mader, S., et al. (2016). MHC class I-associated peptides derive from selective regions of the human genome. J. Clin. Invest. *126*, 4690–4701.

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

Peters, B., Bulik, S., Tampe, R., Van Endert, P.M., and Holzhütter, 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.

Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., and Fugmann, T. (2016). High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. Proteomics *16*, 1570–1580.

Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G.E. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. *43*, D423–D431.

Rock, K.L., Reits, E., and Neefjes, J. (2016). Present yourself! By MHC class I and MHC class II molecules. Trends Immunol. *37*, 724–737.

Rossjohn, J., Gras, S., Miles, J.J., Turner, S.J., Godfrey, D.I., and McCluskey, J. (2015). T cell antigen receptor recognition of antigen-presenting molecules. Annu. Rev. Immunol. *33*, 169–200.

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., and Schreiber, R.D. (2015). Neoantigens in cancer immunotherapy. Science *348*, 69–74.

Serwold, T., Gonzalez, F., Kim, J., Jacob, R., and Shastri, N. (2002). ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. Nature *419*, 480–483.

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

Shraibman, B., Barnea, E., Kadosh, D.M., Haimovich, Y., Slobodin, G., Rosner, I., López-Larrea, C., Hilf, N., Kuttruff, S., Song, C., et al. (2019). Identification of tumor antigens Among the HLA peptidomes of glioblastoma tumors and plasma. Mol. Cell. Proteomics *18*, 1255–1268.

Shraibman, B., Kadosh, D.M., Barnea, E., and Admon, A. (2016). Human leukocyte antigen (HLA) peptides derived from tumor antigens induced by inhibition of DNA methylation for development of drug-facilitated immunotherapy. Mol. Cell. Proteomics *15*, 3058–3070.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal Omega. Mol. Syst. Biol. *7*, 539.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout : a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Stranzl, T., Larsen, M.V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics *62*, 357–368.

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M.M., Kloetzel, P.M., Rammensee, H.G., Schild, H., and Holzhütter, H.G. (2005). Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. Cell. Mol. Life Sci. *62*, 1025–1037.

Uebel, S., Kraas, W., Kienle, S., Wiesmüller, K.H., Jung, G., and Tampé, R. (1997). Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. Proc. Natl. Acad. Sci. USA *94*, 8976–8981.

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

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

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.

Wolf-Levy, H., Javitt, A., Eisenberg-Lerner, A., Kacen, A., Ulman, A., Sheban, D., Dassa, B., Fishbain-Yoskovitz, V., Carmona-Rivera, C., Kramer, M.P., et al. (2018). Revealing the cellular degradome by mass spectrometry analysis of proteasome-cleaved peptides. Nat. Biotechnol. https://doi.org/10.1038/nbt.4279.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
MULTIALLELIC benchmark dataset with predictions (CSV)	This paper; and Mendeley Data	Data S1; 10.17632/zx3kjzc3yx.3
MONOALLELIC benchmark dataset with predictions (CSV).	This paper; and Mendeley Data	Data S2; 10.17632/zx3kjzc3yx.3
Training data for MHCflurry 2.0 binding affinity (BA) predictor (CSV).	This paper; and Mendeley Data	Data S3; 10.17632/zx3kjzc3yx.3
Training data for the variant of MHCflurry BA evaluated on the MONOALLELIC benchmark in Figure S2 (CSV).	This paper; and Mendeley Data	Data S4; 10.17632/zx3kjzc3yx.3
Training data for MHCflurry 2.0 antigen processing (AP) predictors (CSV).	This paper; and Mendeley Data	Data S5; 10.17632/zx3kjzc3yx.3
Training data for MHCflurry 2.0 presentation score (PS) predictors (CSV).	This paper; and Mendeley Data	Data S6; 10.17632/zx3kjzc3yx.3
IEDB MHC ligand data	(Vita et al., 2019)	http://www.iedb.org/doc/mhc_ligand_full.zip
BD2013 affinity dataset	(Kim et al., 2014); http://tools.iedb.org/main/datasets/	http://tools.iedb.org/static/main/benchmark_mhci_reliability.tar.gz
SysteMHC Atlas	(Shao et al., 2018)	https://systemhcatlas.org/Builds_for_download/180409_master_final.tgz
RNA-seq: Human Protein Atlas (Cell Lines)	(Uhlén et al., 2015)	https://www.proteinatlas.org/download/rna_celline.tsv.zip
RNA-seq: Human Protein Atlas (Blood)	(Uhlén et al., 2015)	https://www.proteinatlas.org/download/rna_blood_cell_sample_tpm_m.tsv.zip
RNA-seq: Human Protein Atlas (GTEx)	(Uhlén et al., 2015)	https://www.proteinatlas.org/download/rna_tissue_gtex.tsv.zip
RNA-seq: Expression Atlas (CCLE)	(Barretina et al., 2012; Papatheodorou et al., 2020)	https://www.ebi.ac.uk/gxa/experiments-content/E-MTAB-2770/resources/ExperimentDownloadSupplier.RnaSeqBaseline/tpms.tsv
RNA-seq: Melanoma Tumors	(Barry et al., 2018)	GEO: GSE113126; https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE113126&format=file
HLA-A protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/A_prot.fasta
HLA-B protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/B_prot.fasta
HLA-C protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/C_prot.fasta
HLA-E protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/E_prot.fasta
HLA-F protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/F_prot.fasta
HLA-G protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/fasta/G_prot.fasta
H-2 Db protein sequence	UniProt	P01899
H-2 Dd protein sequence	UniProt	P01900
H-2 Dp protein sequence	UniProt	P14427

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
H-2 Dk protein sequence	UniProt	P14426
H-2 Dq protein sequence	UniProt	Q31145
H-2 Kb protein sequence	UniProt	P01901
H-2 Kd protein sequence	UniProt	P01902
H-2 Kk protein sequence	UniProt	P04223
H-2 Kq protein sequence	UniProt	P14428
H-2 Ld protein sequence	UniProt	P01897
H-2 Lq protein sequence	UniProt	Q31151
Other species MHC protein sequences	IMGT (Robinson et al., 2015)	ftp://ftp.ebi.ac.uk/pub/databases/ipd/mhc/MHC_prot.fasta
Software and Algorithms		
MHCflurry 2.0	This paper	https://github.com/openvax/mhcflurry
NetMHCpan 4.0	(Jurtz et al., 2017)	http://www.cbs.dtu.dk/services/NetMHCpan-4.0
MixMHCpred 2.0.2	(Gfeller et al., 2018)	https://github.com/GfellerLab/MixMHCpred
Other		
MS-identified MHC ligands	(Hassan et al., 2013)	Table S1
MS-identified MHC ligands	(Mommen et al., 2014)	Dataset S01
MS-identified MHC ligands	(Bassani-Sternberg et al., 2015)	Table S1
MS-identified MHC ligands	(Ritz et al., 2016)	Supplemental Information
MS-identified MHC ligands	(Shraibman et al., 2016)	Table S1
MS-identified MHC ligands	(Gloger et al., 2016)	Tables S1, S2, S3, S4, and S5
MS-identified MHC ligands	(Bassani-Sternberg et al., 2016)	Data S2
MS-identified MHC ligands	(Bassani-Sternberg et al., 2016)	S1 Dataset
MS-identified MHC ligands	(Pearson et al., 2016)	Used as reprocessed by (Bassani-Sternberg et al., 2017): S2 Dataset
MS-identified MHC ligands	(Shraibman et al., 2019)	Tables S1 and S2
MS-identified MHC ligands	(Abelin et al., 2019)	Data S1
MS-identified MHC ligands	(Sarkizova et al., 2020)	Data S1 and S2
Proteosome-cleaved peptides identified by "mass spectrometry analysis of proteolytic peptides" (MAPP)	(Wolf-Levy et al., 2018)	Data S2

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Timothy O'Donnell (tim@openvax.org).

Materials Availability

This study did not generate new materials.

Data and Code Availability

MHCflurry source code, training data, and trained models are available at https://github.com/openvax/mhcflurry. Datasets used to train and benchmark the predictors described in this work have been deposited in Mendeley Data under DOI:10.17632/zx3kjzc3yx.3, available at https://data.mendeley.com/datasets/zx3kjzc3yx.

CellPress
OPEN ACCESS

METHOD DETAILS

MS Benchmark Construction and Approach

Dataset Curation

To benchmark the binding predictors developed here and elsewhere, we collected datasets from 11 studies that identified MHC class I-bound peptides using MS. We included only samples with known four-digit MHC class I genotypes. Two of the studies included experiments that used cell lines engineered to express a single MHC class I allele. We refer to these as the MONOALLELIC samples, which comprise 92 samples from the recent publication by Sarkizova et al(Sarkizova et al., 2020) and 8 samples from Abelin et al(Abelin et al., 2019). We refer to the other samples, in which exact MHC class I restrictions were not experimentally determined, as the MULTIALLELIC samples. We divided these into two groups: MUTLIALLELIC-OLD, comprised of 56 experiments from eight studies published through 2018(Bassani-Sternberg et al., 2015, 2016, 2017; Gloger et al., 2016; Hassan et al., 2013; Mommen et al., 2014; Ritz et al., 2016; Shraibman et al., 2016), and MULTIALLELIC-RECENT, comprised of 20 experiments from two studies published in 2019(Sarkizova et al., 2020; Shraibman et al., 2019). Curated samples are listed in Table S2, the sample groups used for each benchmark experiment are given in Table S3, and Data S1 and S2 give the full datasets for the MULTIALLELIC and MONOALLELIC benchmarks, respectively.

Transcript Expression

Each curated mass spec experiment was associated with a publicly-available bulk RNA-seq expression dataset (or mixture of several) that approximately matched its cell type or tissue of origin. This was used during decoy selection to disambiguate MS-identified peptides found in multiple proteins by associating each peptide with its most highly-expressed possible source protein. We used three sources of expression data: (1) the Human Protein Atlas (Uhlén et al., 2015) analysis of tissues from the GTEx project, cell lines, and whole blood, (2) the Expression Atlas (Papatheodorou et al., 2020) reanalysis of RNA-seq from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), and (3) published RNA-seq of 11 metastatic melanoma samples from patient tumors (Barry et al., 2018). Data files giving Ensembl gene-level transcripts per million (TPM) expression quantifications were obtained from their respective sources. Table S2 indicates the expression dataset used for each sample.

Decoy Selection

We generated accuracy benchmarks from the curated datasets by sampling unobserved peptides (decoys) from the same proteins as the observed peptides (hits) for each sample. Each MS-identified peptide was searched in the Uniprot human reference proteome (UP000005640_9606). In the case of multiple matches, we used the RNA-seq for each sample to select a single protein with the highest mRNA expression. For each sample, we randomly selected *99n* decoy peptides, where *n* is the number of hits. Equal numbers of decoy peptides of each length (8, 9, 10, 11) were sampled. We excluded from this procedure all peptides (hits and decoys) that were present in the training data for any of a sample's alleles for the MHCflurry predictor under evaluation, as well as hits that contained noncanonical amino acids or could not be matched to a Uniprot protein and Ensembl gene (Ensembl release 98). Decoy generation for benchmarks was separate from decoy generation used for predictor training.

Comparison to Existing Tools

We included two existing tools in our benchmarks, NetMHCpan 4.0 (Jurtz et al., 2017) and MixMHCpred 2.0.2 (Bassani-Sternberg et al., 2017; Gfeller et al., 2018). The NetMHCpan 4.0 tool is an ensemble of neural networks trained on peptide-MHC class I affinity measurements plus peptides from monoallelic MS experiments. It gives separate predictions for each kind of data, i.e. a BA prediction and an EL prediction, which we evaluated as separate predictors in our benchmarks. MixMHCpred is trained on peptides identified in multiallelic MS experiments, which it deconvolves into clusters that are subsequently associated with MHC class I alleles. As both the MHCflurry binding affinity predictor and NetMHCpan 4.0 do not incorporate multiallelic MS in their training sets, we evaluated them on the full MULTIALLELIC benchmark. As MixMHCpred is trained on multiallelic MS, we tested it only on samples published recently (after it was trained), i.e. the MULTIALLELIC-RECENT benchmark. The two studies in the MONOALLELIC dataset were published after NetMHCpan 4.0 and MixMHCpred 2.0.2 were released, so it is also an appropriate test set for them, although in some cases (21 of 100 samples in the MONOALLELIC benchmark) the sample's allele was unsupported by MixMHCpred and we omitted it when evaluating MixMHCpred. As the released MHCflurry 2.0 binding predictor incorporates the MONOALLELIC dataset as training data, in evaluations of this benchmark we used a variant of MHCflurry re-trained without these datasets. Since the logistic regression model used in the MHCflurry 2.0 PS predictor was fit to the MULTIALLELIC-OLD subset, we benchmarked it only on the MULTIALLELIC-RECENT samples. The benchmarks used for each analysis are summarized in Table S3.

MHC Class I Binding Affinity Predictor

The MHCflurry 2.0 BA predictor is a new pan-allele MHC class I binding affinity predictor that supports variable-length peptides up to 15-mers. It extends an earlier version of MHCflurry(O'Donnell et al., 2018) —in which separate neural networks were trained for each of 112 supported alleles— to now support 14,993 MHC class I alleles using a single neural network ensemble. The input to each neural network consists of (1) an encoding of the peptide amino acid sequence and (2) an encoding of the amino acids at 37 selected positions from a multiple sequence alignment of a large number of MHC class I alleles across species. The neural networks each output a nanomolar binding affinity (transformed to a 0-1 scale using the formula $1 - log_{50000}(nM\ affinity)$). The mean of the transformed predictions over the ensemble gives the overall prediction. The predictor is trained on affinity measurements and MS datasets from cell lines monoallelic for a single MHC class I allele. MS hits are assigned a "<100 nM" binding affinity. The models are trained using a variant of the mean squared loss (MSE) that supports such inequalities, as described previously(O'Donnell et al., 2018).

Peptide Representation

Peptides of 15 amino acids or shorter are supported. As our feedforward networks require fixed-length inputs, variable-length peptides are transformed to a 45-mer representation by concatenating three representations: left aligned, centered, and right aligned. Unused positions are represented using a special X symbol, treated as a 21st amino acid. For example, the peptide "GILGFVFTL" is represented by concatenating "GILGFVFTLXXXXXX" (left aligned), "XXXGILGFVFTLXXX" (centered), and "XXXXXXGILGFVFTL" (right aligned), resulting in "GILGFVFTLXXXXXXXXXGILGFVFTLXXXXXXXXXGILGFVFTL". Thus, each residue in the peptide is represented at three positions in the encoding. The motivation for this approach is to ensure that the termini of the peptide along with the central residues are available to the network at a fixed position in the encoding, regardless of peptide length. For example, the N-terminal residue will always map to encoded position 1 (as well as to two other encoded positions, which do depend on peptide length), the central residue is available at position 23, and the C-terminal residue is available at position 45. Each amino acid in this sequence is transformed to a 21-dimensional vector using the BLOSUM62 substitution matrix(Henikoff and Henikoff, 1992), extended to include the X placeholder, which is assigned similarity 1 to itself and 0 to all amino acids.

Allele Representation

MHC class I alleles are represented to the neural network by the amino acids at 37 positions from a global multiple sequence alignment. This representation is referred to as a "pseudosequence" by the NetMHCpan authors. We use the 34 peptide-contacting positions included in the NetMHCpan 4.0 pseudosequence (re-derived using a new alignment), plus 3 additional positions. The additional positions were selected to differentiate several pairs of alleles (A*23:01/A*24:13, A*29:01/A*29:02, B*44:02/B*44:27, C*03:03/C*03:04) that shared identical 34-mer NetMHCpan pseudosequences and had at least 50 entries in the training dataset. An earlier version of the training dataset (MHCflurry version 1.4) was used for this step. The three additional positions were selected as the most parsimonious set for distinguishing these alleles, not by consideration of protein structure or by comparison of predictive accuracy. Using HLA-A*01:01 protein residue numbering (IMGT accession HLA00001) starting from 1, the full set of selected positions were 31, 33, 48, 69, 83, 86, 87, 90, 91, 93, 94, 97, 98, 100, 101, 104, 105, 108, **115**, 119, 121, 123, **126**, 138, 140, 142, 167, 171, 174, 176, 180, 182, 183, 187, 191, 195, **223** (the three additional selected positions are shown in bold). Similar to the peptide encoding, each amino acid was transformed to a 21-dimensional vector using the BLOSUM62 substitution matrix, where a placeholder X character represents global alignment positions with no residue (i.e. a deletion) for a particular MHC allele.

Allele Multiple Sequence Alignment

Full-length HLA-A, -B, -C, -*E*, -*F*, and -*G* amino acid sequences were downloaded from the IMGT/HLA(Robinson et al., 2015) project database, mouse H-2 sequences from UniProt (UniProt Consortium, 2019) (H-2 Db from accession P01899; H-2 Dd, P01900; H-2 Dp, P14427; H-2 Dk, P14426; H-2 Dq, Q31145; H-2 Kb, P01901; H-2 Kd, P01902; H-2 Kk, P04223; H-2 Kq, P14428; H-2 Ld, P01897; H-2 Lq, Q31151), and MHC class I sequences for additional species from the IPD-MHC database(Maccari et al., 2017). After filtering to MHC class I alleles with names parsable by the mhcnames package (https://github.com/openvax/mhcnames), the sequences were aligned using Clustal Omega 1.2.1 (ref. (Sievers et al., 2011)) with the command "clustalo -i class1.fasta -o class1.aligned.fasta." Positions from the resulting alignment were selected that best recapitulated the NetMHCpan pseudosequences(Jurtz et al., 2017), then extended to fully differentiate the 171 alleles with at least 50 entries in the training dataset. The final sequences included with MHCflurry encompass 14,993 MHC class I alleles.

Training Data

MHC ligand entries were downloaded from the Immune Epitope Database (IEDB)(Vita et al., 2019) on April 27, 2020. Entries with parsable MHC class I allele names and peptides of length 8-15 with no post-translational modifications were retained. The MS data from IEDB was extended to include 46,880 additional MS hits from the SysteMHC Atlas project(Shao et al., 2018) (downloaded on May 6, 2019) plus hits from the MONOALLELIC benchmark: 136,742 hits from Sarakizova et al(Sarkizova et al., 2020) and 4,808 hits from Abelin et al (Abelin et al., 2019). The training set for the BA predictor is available as Data S3. For testing on the MONOALLELIC benchmark, a predictor variant trained without the MONOALLELIC datasets was used (Data S4). The set of IEDB affinity measurements was augmented with the "BD2013" dataset from Kim et al.(Kim et al., 2014). The final training set consisted of 493,473 MS entries and 219,596 affinity measurements. A summary of the training datasets used for all predictors introduced in this work is given in Table S4.

Considerations Related to Use of MS Training Data

In preliminary experiments, we found that the inclusion of MS data in the BA predictor training set was critical for learning a high-performance predictor. This is not surprising as without MS data the training set size and allele coverage is greatly diminished. Inclusion of MS data represents a tradeoff, however, as the approach we describe to train the AP predictor requires that the BA predictor does not fully model the antigen processing signals available in its MS training datasets. While the BA predictor likely includes a contribution from antigen processing, we note that several technical choices may have blunted its ability to learn these effects. The BA predictor's training set includes 69% mass spec-identified ligands and 31% peptide-MHC class I affinity measurements. The predictor therefore models a compromise between antigen processing-sensitive (MS ligand) and -insensitive (affinity measurement) training data. The influence of MS data is expected to be further diluted at the extreme strong-binder segment of the BA predictor's output because MS hits are assigned an IC_{50} of "< 100 nM" in the training set, i.e. any assigned affinity tighter than 100 nM results in zero contribution to the training loss. This means that MS training data does not guide the relative ranking (exact IC_{50}) of strong binders, potentially the regime where antigen processing signals may make the greatest difference. This effect likely also plays a role in the relatively lackluster performance of the BA predictor when assessed by the PPV metric – which emphasizes the rank-order of peptides at the extreme high-end of the predictor's output – despite good performance on AUC. Finally, an important difference between the BA and AP predictors is that the AP predictor uses peptides from the proteome as decoys (unobserved non-binders), whereas the

BA predictor uses random sequences sampled according to the same amino acid distribution as the hits. The AP predictor strategy is expected to be more realistic and informative, at the cost of giving the AP predictor more opportunity to learn MS biases.

Neural Network Architectures

Each neural network in the ensemble corresponds to one of 35 possible model architecture variants. The overall design in all cases is similar: the allele and peptide representations are concatenated, flattened, and passed through a series of two or three dense layers, whose size ranges from 256 to 1024. Each dense layer is followed by a Dropout layer(Srivastava et al., 2014) with the dropout rate set to 50%. Architectures vary in terms of the size and number of the dense layers, the amount of regularization applied to dense layer weights (L1 penalty), and whether skip connections are used to give each dense layer direct access to the two preceding layers.

Model Training

The training data was sampled four times to generate four training subsets. Each subset excluded one quarter or 100 of the training points for each MHC class I allele, whichever was less for each allele. Models corresponding to the 35 architectures were fit to each of the four training subsets, for 140 trained models in total. Initial weights were selected using layer-sequential unit-variance initialization(Mishkin and Matas, 2015). This was followed by a pre-training step, in which the network was fit to synthetic measurements generated by a previous version of MHCflurry (version 1.2.0, using allele-specific models trained without MS datasets). The synthetic data consists of affinity predictions for random peptides across 99 alleles. While pre-training on synthetic data, the training data was used as a test set for early stopping, i.e. pre-training was halted once the mean square error (MSE) was no longer improving on the actual (non-synthetic) training data. Typically, several hundred million synthetic measurements were used. Model fitting then proceeded using the training data, keeping 10% held-out for early stopping. The training dataset was augmented to include random peptides set to have a very weak affinity (>30,000 nM). The lengths of the random peptides were selected to equalize the number of non-binder data points across peptide lengths for each allele. As in NetMHC, the sequences of the random negative peptides were resampled after each epoch.

Model Selection

From the 140 trained models, the model selection procedure selected ten to use in the final predictor ensemble. This was done by selecting a set of models from each of the four training subsets independently, using MSE on the held-out points as the accuracy metric. A forward stepwise procedure was used, in which models are added until the ensemble accuracy no longer improves. The final ensemble was the union of the models selected across training subsets.

Evaluation on Held-Out Data

To estimate MHCflurry BA performance across MHC class I alleles (Figure S1; Table S1), we evaluated the full ensemble of trained models (i.e. prior to model selection) on the held-out model selection data. For each model selection data point, only the models in the ensemble that were not trained on that point were used to generate a prediction. The results are expected to slightly underestimate true predictor accuracy, as models that were not selected by the model selection routine due to poor performance are still included in the evaluation. For each allele, the evaluation dataset consists of three sources: (1) held-out affinity measurements, which we labelled as binders or non-binders using a 500 nM threshold, (2) held-out MS-identified peptides (binders), and (3) peptides randomly sampled from the proteome (non-binders). The number of synthetic non-binder peptides sampled from the proteome was set to be four times the number of binders, split evenly across lengths 8, 9, 10, 11. We computed AUC at distinguishing binders from non-binders for each allele, using bootstrap resampling to derive a 95% confidence interval.

Performance Evaluation on Alleles with No Representation in the Training Data Set

To better understand MHCflurry BA accuracy on alleles in the MONOALLELIC benchmark that had no entries in the training data (Figure S2C; Table S5), we compared the AUC for these samples to that obtained using predictions for related alleles with substantial training data. As in all evaluations on the MONOALLELIC benchmark, a variant of MHCflurry BA trained without this dataset was used. Distance between alleles was defined as the number of differing positions in the 37-mer sequences provided as input to the predictor. Each allele was matched to the most-similar allele with at least 50 entries in the training set. Two alleles had more than one equally-similar allele using this approach (A*34:02 matched A*03:01, A*68:01, and A*68:23; C*03:02 matched C*03:04 and B*46:01), and we broke ties by selecting the similar with the most training data. The MHCflurry BA predictor was evaluated on the relevant samples in the MONOALLELIC benchmark using the corresponding selected similar alleles.

Antigen Processing Predictor

The MHCflurry 2.0 AP predictor is trained to model the MHC allele-independent effects that were not captured by the BA predictor. We designed it with proteasomal cleavage prediction in mind, although the resulting models are expected to capture a range of effects. Since the efficiency of cleavage by proteasomes and peptidases are affected by the residues both before and after the cut site, we experimented with the inclusion of the flanking sequence on either side of the peptide from its source protein. To understand the importance of flanking sequences, we trained two variants of the MHCflurry 2.0 AP predictor: one takes as input a peptide plus the five amino acids on either side in its source protein (AP with flanks) and one takes only a peptide (AP without flanks). Both models are independent of MHC allele.

Training Data

The AP predictor is trained on the MS hits from the MONOALLELIC benchmark samples (Data S5). Hits of length 8-11 were matched 1:100 to decoys of the same length and from the same proteins. The resulting peptides for each sample were sorted by MHCflurry 2.0 BA prediction for the relevant allele, and the top 2% strongest binding peptides (hits and decoys) were selected for inclusion in the AP predictor's training set. This resulted in a training set of 399,392 entries (encompassing 297,548 distinct peptides, as duplicate

peptides were sometimes selected from different MS samples), of which 175,836 (44%) were hits. The median BA prediction for the peptides in AP training set was 73 nM for decoys and 38 nM for hits.

Neural Network Architectures

AP models were trained using 128 neural network architectures with a similar overall design but varying in layer sizes, activation function, level of L1 weight regularization, and dropout rate. The input to each network consists of peptides and flanking sequences (if used) as they occur in the source protein (N-flank - peptide - C-flank), with each amino acid encoded as a 21-dimensional vector using the BLOSUM62 substitution matrix extended to include an X symbol, as in the BA predictor. The sequence is right-padded with X characters to generate fixed-length inputs of length 25 (AP with flanks) or 15 (AP without flanks). The first layer is convolutional, with a kernel size of 11-17 and 256 or 512 filters, tanh or relu activation, and dropout. This transforms the input sequence into a new sequence of the same length but with up to 512 channels instead of 21. From this representation, two parallel convolutional layers with a kernel size of 1 (i.e. each position is considered independently) are applied to predict the favorability of an "N-terminal cut" and a "C-terminal cut" at each position in the sequence. The cut site predictors are implemented using two stacked convolutional layers with a kernel size of 1, and thus can be thought of as 2-layer dense networks that consider a single position in the learned representation of the input sequence. For example, in one model architecture the "C-terminal cut" predictor takes a 512-vector corresponding to a position in the input sequence and applies a dense layer with 8 outputs followed by a single-output dense layer. The final layer in the AP architecture is a dense layer that integrates these cut-site predictor results. It takes as input: (1) the N-terminal cut site prediction at the peptide N-terminus, (2) the max of the N-terminal cut site predictions across the rest of the peptide, (3) the C-terminal cut site prediction at the peptide C terminus, and (4) the max of the C-terminal cut site predictions across the rest of the peptide. This design is motivated by the intuition that presented MHC class I ligands must have favorable cleavability at their termini but avoid cleavage at interior residues. To give the model further opportunity to consider average properties of the flanking sequences (e.g. secondary structure), two additional inputs are given to the final layer: (5) the result from a dense layer applied to the per-channel average of the initial convolutional layer across the N-flank, (6) a similar result for the C terminus. The final layer in the AP predictor is thus intended to model the tradeoff between cleavability at the peptide termini, cleavability at interior positions in the peptide, and overall favorability of cleavage given the flanking sequences.

Model Training and Selection

The training data was sampled to generate four subsets, where each training subset held out 10 randomly selected MS experiments. Models corresponding to the 128 neural network architectural variants were fit to each training subset (512 models total). Models were trained using the Adam optimizer(Kingma and Ba, 2014) and binary cross entropy loss with a random 10% of each training subset used for early stopping. Model selection was performed for each training subset separately using AUC on the held-out data as the accuracy metric. The final selected ensemble included 8 models.

Evaluation on "Mass Spectrometry Analysis of Proteolytic Peptides" (MAPP) Data

Proteasome-cleaved peptides of lengths 8-11 were extracted from Wolf-Levy et al., Supplementary Data 2(Wolf-Levy et al., 2018). Peptides identified in any experiment (untreated or TNF/IFN-treated) were included. For each observed peptide (hit), two length-matched unobserved decoy peptides were randomly sampled from the same gene. To avoid any contribution from the cysteine detectability bias, all peptides (hits and decoys) containing cysteine were dropped.

Presentation Score Model

The MHCflurry 2.0 PS predictor is a two-input logistic regression model that integrates a BA prediction (tightest predicted binding affinity over the MHC class I alleles for a sample) with an AP prediction to give a composite prediction, referred to as the presentation score. It has just three learned parameters: two coefficients and an intercept. In contrast to the BA and AP predictors, which use only monoallelic MS, the PS model is fit to multiallelic MS datasets. Training data for the PS model was generated using the MULTIALLELIC-OLD set of samples by sampling length-matched decoys (two decoys per hit) from the same proteins as the hits and sampling 10% of the resulting dataset for efficiency. The training set included 75,378 entries (of which 24,983 were hits) from 56 samples (Data S6). The model was fit using the logistic regression implementation in scikit-learn(Pedregosa et al., 2011) with the LBFGS solver and default parameters. Two variants of the PS model were generated: one that made use of the AP with flanks predictor (i.e. upstream and downstream amino acids included) and another that used the AP without flanks predictor. The PS model was benchmarked on the MULTIALLELIC-RECENT samples.

QUANTIFICATION AND STATISTICAL ANALYSIS

Accuracy Metrics

We assessed predictor accuracy at distinguishing MS hits from decoys using two scores, area under the curve (AUC) and positive predictive value (PPV). The AUC is a standard accuracy metric for classification tasks, interpretable as the probability that a randomly selected hit will be scored higher by a predictor than a randomly selected decoy. The PPV, as defined here, focuses on a predictor's ability to rank hits far above the decoys. To calculate PPV, for each sample we sorted the n hits and $99n$ decoys by their predictions, and calculated the fraction of the top n peptides that are hits. A random predictor would score 0.01 in PPV and 0.5 in AUC.

CellPress
OPEN ACCESS

Statistical Significance

To compare predictors (Figures 1C, 3C, and S2) or benchmark variants (Figure 2D), we computed the percent difference in accuracy score (AUC or PPV) for each sample in the benchmark and calculated the mean difference across samples as well as its 95% confidence interval by bootstrap resampling of the benchmark samples. Differences were deemed significant if the confidence interval excluded the value 0.0. To quantify average accuracy (Figure 2B) or average prediction (Figure 2H), we calculated the mean value and its 95% confidence interval using bootstrap resampling. To test for differences in correlation between AP prediction and BA prediction across groups of alleles (Figures 2F and 2G), we used the Mann-Whitney U test.