A model of antigen processing improves prediction of MHC class I-presented peptides

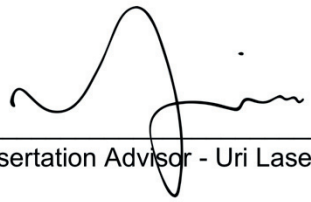
by

Timothy John O'Donnell

A dissertation submitted to the Graduate Faculty of the Graduate School of Biomedical Sciences, Biomedical Sciences Doctoral Program, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Icahn School of Medicine at Mount Sinai

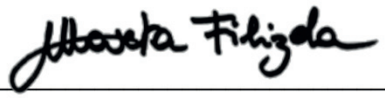
2020

This manuscript has been read and accepted by the Graduate Faculty of the Graduate School of Biomedical Sciences, in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.



August 10, 2020

Dissertation Advisor - Uri Laserson, Ph.D. Date



8/21/2020

Dean - Marta Filizola, Ph.D. Date

Committee members: Icahn School of Medicine at Mount Sinai

Amir Horowitz, Ph.D

Brad R. Rosenberg, M.D., Ph.D.

Jeremiah Faith, Ph.D.

External examiner committee member: Memorial Sloan Kettering Cancer Center

Benjamin Greenbaum, Ph.D.

Icahn School of Medicine at Mount Sinai

Abstract

A model of antigen processing improves prediction of MHC class I-presented peptides

by

Timothy John O'Donnell

Advisor: Uri Laserson, Ph.D.

T cells recognize peptides presented by major histocompatibility complex (MHC) proteins on cell surfaces. Computational prediction of MHC-presented peptides is an essential tool for epitope mapping and vaccine design. In this dissertation, I introduce improved predictors of peptide presentation on MHC class I. The predictors are fit to published datasets of MHC-presented peptides identified by mass spectrometry, as well as other sources. Separate models are developed for predicting MHC/peptide binding and the antigen processing steps that occur prior to MHC binding *in vivo*. I show that a combination model incorporating these two components achieves higher accuracy than existing methods at predicting MHC presentation, as well as neoantigens recognized by CD8+ T cells from cancer patients. The new methods are made available as an open source software package called MHCflurry (https://github.com/openvax/mhcflurry).

Preface

Work presented in this dissertation is adapted from the following publications:

O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems.* 2018.

O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved pan-allele prediction of MHC Class I-presented peptides by incorporating antigen processing. *Cell Systems.* 2020.

Some sections of Chapters 1 and 5 are based on the author's contribution to the following book chapter:

Boegel S, Castle JC, Kodysh J, O'Donnell TJ, Rubinsteyn A. Bioinformatic methods for cancer neoantigen prediction. *Progress in Molecular Biology and Translational Science.* Academic Press. 2019.

Acknowledgements

I've been beyond fortunate in the community I've come to know since I joined Mount Sinai in 2014. I can't name everyone here, but thank you, all of you, for making this work possible through your help and friendship.

Uri Laserson took the leap to be my advisor, encouraged me to think big, and was fully open to any direction that caught my interest.

Elliot Merritt gave me a patient introduction to laboratory benchwork. This experience helped ground my thinking and is a highlight of my time in graduate school. Meimei Shan was another generous teacher.

Nina Bhardwaj and her group have been an unwavering source of support and education for nearly all my time at Mount Sinai. Nina's pioneering clinical trials in cancer neoantigen vaccination are an important motivation for the work described here.

Jeff Hammerbacher convinced me to come work at Mount Sinai in 2014. The years working together were formative and a lot of fun.

Andrew Kasarskis's help at key moments made it possible for me to stay at Mount Sinai and later to enroll in graduate school.

Throughout all of this, I've somehow been able to collaborate for years with two close friends, Alex Rubinsteyn and Julia Kodysh. They've kept my head in the sky and my feet on the ground.

Finishing this work during the pandemic, I leaned especially hard on my family and friends. Thanks Mom, Dad, and Genie for your endless encouragement. And thank you, Jess, for so much happiness.

Table of Contents

List of Figures

List of tables

Chapter 1. Introduction

T cells provide the cell-mediated component of adaptive immunity. They complement and collaborate with the other component, antibodies, to direct and effect the adaptive immune response. A well-regulated T cell response can eliminate cells that harbor viruses, bacteria, or cancer-associated mutations, direct the development of high-affinity antibodies, modulate inflammation, and retain immunological memory. An aberrant T cell response can lead to autoimmunity, allergy, and lethal acute inflammatory syndromes. Central to both kinds of responses are the specific antigenic determinants, termed epitopes, recognized by T cells.

The epitopes recognized by most T cells are peptides bound to major histocompatibility complex (MHC) proteins on the surfaces of other cells. Understanding the repertoire of peptide/MHC complexes (pMHC) across tissues and disease states is critical for studying and manipulating the associated T cell responses. Important advances have recently been made in methods for detecting MHC-presented peptides using mass spectrometry (MS). However, it is far from practical to perform such experiments every time this information is needed. Predictive models enable the expanding collection of experimental evidence on pMHC presentation to be applied to new contexts.

This dissertation addresses the question of how to apply currently-available datasets to better predict the repertoire of pMHC. The focus is on MHC class I, the target recognized by cytotoxic (CD8+) T cells. I show that substantial improvements over existing approaches are possible by integrating results from MS and other kinds of experiments, accounting for variability in the lengths of peptides bound to MHC, and modeling the antigen processing steps that precede MHC binding. A software package implementing these techniques, MHCflurry, is readily applicable to tasks that require prediction of potential CD8+ T cell epitopes, such as vaccine design.

In this chapter, I first describe the cellular processes that contribute to pMHC presentation, the experimental assays used to collect information on the pMHC repertoire, and two applications of pMHC

prediction. I next review some highlights from the large body of work on predictive methods for this task. In the final sections I define and motivate the study objectives, describe key aspects of my approach, and outline the rest of the dissertation.

Background

MHC presentation determines the antigens available to T cells

In the 1970s, studies of mice responding to viral infection revealed that T cells have a dual specificity: each T cell appears to recognize both self and not-self components (Zinkernagel and Doherty 1974). Genetic mapping suggested that the self-component involved major histocompatibility (MHC) proteins, but it initially remained unclear if the viral and self-components were recognized by different T cell receptors or if it was the interaction of MHC and viral material that T cells recognize (Zinkernagel and Doherty 1997). This puzzle was definitively solved in the 1980s and 1990s by a series of crystal structures showing the MHC protein binding individual peptides in a groove clearly adapted for this purpose (Figure 1-1A) (Bjorkman et al. 1987; Brown et al. 1993).

It soon became clear that the antigens recognized by T cells are peptide / MHC (pMHC) complexes, where the MHC is a host protein and the peptide may be of host or foreign origin. T cells recognize their antigens through interactions between the T cell receptor (TCR) expressed on T cells and pMHC presented on the surface of a target cell. The dual specificity of T cells arises because the TCR makes contacts with both the MHC and its bound peptide, as shown later in crystal structures of pMHC/TCR complexes (Garcia et al. 1996; Garboczi et al. 1996).

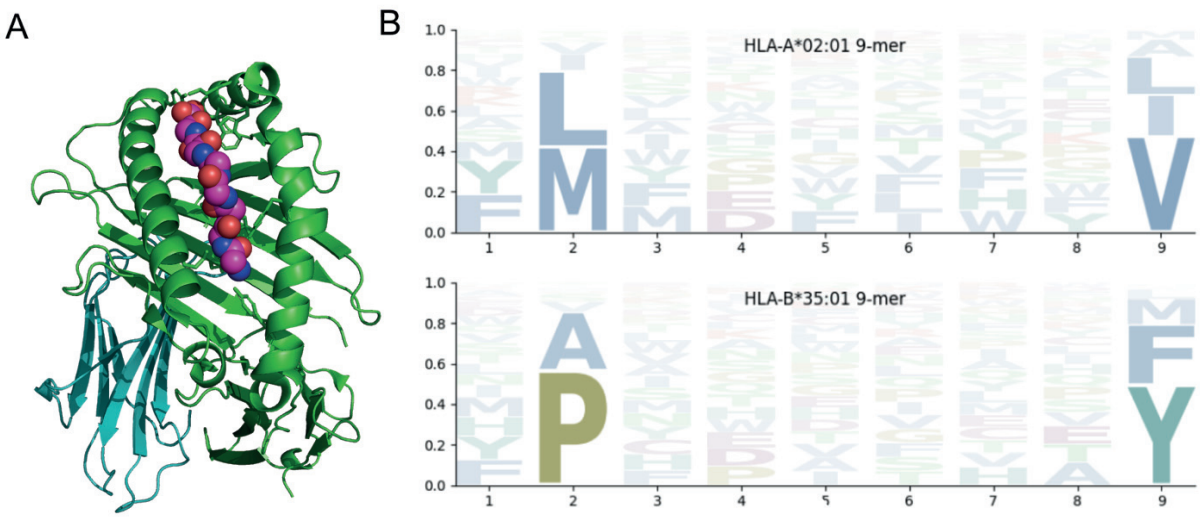


Figure 1-1. MHC structure and example binding preferences

(a) Structure of HLA-A*02:01 in complex with peptide. The MHC class I heavy chain is shown in green and β2 microglobulin in cyan. The main chain atoms of the peptide are drawn as spheres. This is not the original 1987 structure mentioned in the text, but rather a later one from the same group. PDB entry: 1DUZ (A. R. Khan et al. 2000). **(b)** Sequence logo showing binding motifs for HLA-A*02:01 (top) and HLA-B*35:01 (bottom). The logos show the binding preferences modeled by the MHCflurry 2.0 BA predictor developed in Chapter 3.

Most T cells express the alpha/beta TCR, a heterodimer of two protein chains coded by the *TRA* and *TRB* genes. These genes are formed somatically during T cell development by recombination events and the addition of non-templated nucleotides. The diversity of outcomes from this process is the basis for the unique antigen specificities of each T cell (Schatz and Ji 2011). As a general principle, each T cell expresses a single TCR, which has potential specificity for a range of pMHCs. After somatic recombination, T cells undergo two consecutive selection steps. The first, positive selection, removes T cells with non-functional TCRs. During positive selection, T cells that cannot bind host pMHC at least weakly undergo programmed cell death. The second step, negative selection, removes T cell clones that react strongly with self pMHC. The result is a diverse repertoire of naive T cells that recognize potentially foreign material presented on MHC (Klein et al. 2014; Singer, Adoro, and Park 2008). Estimates vary for the diversity of this pool but center around 10^9 to 10^{11} distinct TCRs arising from 10^6 to 10^8 beta sequences paired with fewer alpha sequences (Arstila et al. 1999; Robins et al. 2009; Qi et al. 2014).

Signals received by T cells during positive selection result in lineage commitment to express one of two accessory surface proteins, CD8 or CD4. These proteins interact with MHC in an antigen-independent manner and help stabilize the antigen-specific interaction between the pMHC and TCR. They are important surface markers for studies of T cells, as their expression divides T cells into two major functional groups. As a broad simplification, CD8+ T cells (cytotoxic T cells) can directly eliminate infected or aberrant cells, whereas CD4+ T cells (helper T cells) coordinate immune responses by interacting with other immune cells (Singer, Adoro, and Park 2008).

The antigens recognized by CD8+ T cells are peptides in complex with MHC class I, whereas CD4+ T cells recognize peptides in complex with MHC class II. MHC class I is expressed on the surface of all nucleated cells and generally presents peptides derived from a cytosolic processing pathway. MHC class II expression is restricted to specialized antigen presenting cells and certain other cell types. Peptides presented on MHC class II are generated by a pathway that occurs in lysosomes and frequently derive from proteins taken up from the extracellular environment (Rock, Reits, and Neefjes 2016). The remainder of this discussion focuses on MHC class I.

The human MHC genes and proteins are referred to as human leukocyte antigen (HLA), and in the human context I will use MHC and HLA interchangeably. There are three classical human MHC class I genes, *HLA-A*, *HLA-B*, and *HLA-C.* These genes, as well as those for MHC class II and dozens of others with roles in immunity, are found in the HLA genomic region, a 3.6 megabase stretch of chromosome 6 (Shiina et al. 2009). The HLA genes are among the most polymorphic of all human genes, and there are thousands of HLA alleles across the population. The previously mentioned early crystal structures of pMHC gave a strong hint at the function of this diversity by revealing that hotspots of variability in these genes fall predominantly in the two alpha helices and the beta sheet that form the peptide-binding groove. This suggested that the variability provides a diversification of potential interactions with the bound peptide.

Indeed, the allelic variation of the *HLA-A*, *HLA-B*, and *HLA-C* genes results in MHC proteins with distinct peptide binding specificities. For example, the *HLA-A* allele HLA-A*02:01 binds peptides with leucine or methionine at the second position and valine, isoleucine, or leucine at the C terminus (K. Falk et

al. 1991), whereas HLA-B*35:01 requires proline at the second position and phenylalanine, methionine or tyrosine at the C terminus (Figure 1-1B) (Kirsten Falk et al. 1993). The interaction with MHC class I is stabilized by hydrogen bonds with the peptide backbone and sidechain packing into polymorphic pockets on the MHC (Wieczorek et al. 2017).

The MHC class I binding groove is closed at both ends, and all human MHC class I alleles have an optimal peptide length of nine. However, peptides of different lengths, generally 8-12 but sometimes much longer (Hassan et al. 2015), can also bind, and the tolerance for other lengths varies by allele (Gfeller et al. 2018). Longer peptides typically accommodate the closed groove by bulging in the middle, although instances of overhangs have been reported (Collins, Garboczi, and Wiley 1994; Remesh et al. 2017; Guillaume et al. 2018).

The peptides presented on MHC class I are derived from a cytosolic processing pathway in which the ubiquitin/proteasome system for protein degradation is co-opted to serve a role in adaptive immunity. The proteasome is the primary protein degradation system in the cell. It cleaves proteins targeted for destruction into peptides of about 3 to 22 amino acids (Kisselev et al. 1999). Most cells express the constitutive proteasome, which can cleave after basic residues (referred to as tryptic-like cleavage), acidic residues (caspase-like), or hydrophobic residues (chymotryptic-like). These activities have implications for the presented pMHC as in most cases the peptide C-terminus is directly determined by proteasomal cleavage. Under inflammatory conditions, some subunits of the constitutive proteasome are replaced with alternate versions, resulting in the immunoproteasome, which has modified cleavage preferences (Rock, Farfán-Arribas, and Shen 2010).

Most proteasome-cleaved peptides are rapidly degraded to single amino acids by cytosolic peptidases (Reits et al. 2003). However, a fraction survive to interact with the transporter for antigen processing (TAP) complex on the endoplasmic reticulum (ER) membrane. TAP is a heterodimeric ATP-dependent transporter that moves peptides from the cytosol into the ER lumen. On its ER-luminal side, it assembles with a number of other proteins to form the peptide loading complex (PLC). The PLC is composed of TAP, an MHC class I protein without peptide bound, the tapasin/Erp57 chaperone complex,

and calreticulin. The tapasin complex stabilizes the empty MHC and "proofreads" pMHC, disrupting peptides that bind the MHC weakly (Thomas and Tampé 2017). Calreticulin helps the MHC fold and traffics it to the PLC (Raghavan et al. 2013). Before (or possibly while) associating with MHC, peptides may undergo N-terminal trimming by either of two ER-associated aminopeptidase (ERAP) proteases (Hattori and Tsujimoto 2013; Chen et al. 2016). When a stable pMHC complex is formed, it detaches from the PLC and traffics via the Golgi to the cell membrane. A remarkable structure of the PLC was obtained in 2017 using cryo-electron microscopy (Blees et al. 2017).

The peptides that survive this process and are presented on cell surface MHC are the potential epitopes available for recognition by T cells. The binding specificity of an individual's MHC alleles and the activities of the antigen processing steps therefore constitute a selective filter on what may be recognized by an individual's T cells.

Experimental methods for characterizing the MHC-presented repertoire

The identification of T cell epitopes is central to understanding and manipulating T cell responses. As one general example, once an epitope is discovered, the cognate T cells may be isolated by flow cytometry gated on MHC multimers, enabling phenotypic characterization (Kai, Sine, and Hadrup 2017; Dolton et al. 2018). There has consequently been significant interest in understanding and modeling the MHC presentation pathway that generates potential T cell epitopes. There are two major categories of experimental approaches used to characterize this system (Bjoern Peters, Nielsen, and Sette 2020).

First, the binding affinity between soluble MHC and peptide can be measured *in vitro*. The binding preferences for many MHC alleles have been determined in this way. In one common protocol, purified MHC is incubated with a high-affinity radiolabeled peptide and a range of concentrations of an unlabeled competitor peptide of interest (Sidney et al. 2013). After the incubation, the amount of MHC-bound radiolabeled peptide is quantified, and the binding affinity of the peptide of interest is approximated as the concentration needed to abolish binding of the radiolabeled peptide by 50%.

Second, naturally-presented pMHC can be purified from cells, and the peptides eluted from MHC for identification by liquid chromatography-mass spectrometry (MS) (Purcell, Ramarathinam, and Ternette 2019). While this approach has been long pursued (Hunt et al. 1992), the scale of these experiments has increased dramatically over the past five years due to improvements in chromatography, MS instrumentation, and analysis methods (Creech et al. 2018). It is now common for studies to report tens of thousands of identified MHC ligands (Bassani-Sternberg et al. 2016; Sarkizova et al. 2019; Abelin et al. 2017, 2019). MS experiments can be performed on cells expressing the natural complement of MHC proteins of up to six MHC class I alleles (Bassani-Sternberg et al. 2016) or in cell lines engineered to express a single MHC allele (Abelin et al. 2017), a tagged MHC allele (Abelin et al. 2019), or a secreted MHC construct (Hawkins et al. 2008; Trolle et al. 2016) so that only peptides bound to a single allele are identified. Throughout this dissertation, I refer to these MS experimental designs as multiallelic and monoallelic, respectively.

A potential advantage of the MS approach is that these experiments capture naturally-processed and presented pMHC. The identified peptides represent the final outcome of the MHC presentation pathway, rather than MHC binding preference alone. An important limitation, however, is that the detectability of a peptide by MS depends on peptide chemistry. The most well-known bias is that MS-identified peptides are depleted for cysteine-containing peptides (Abelin et al. 2017), but a range of effects are likely involved (Fusaro et al. 2009). The MS assays are also relatively sample- and resource-intensive.

Applications of predictive models of MHC presentation

Predictive models of the MHC class I presentation system have been pursued since the time that the essential biology first became clear (Sette et al. 1989). Most methods have focused on predicting peptide/MHC binding affinity (BA), as this is the most restrictive determinant of presentation. Overall, these methods have achieved remarkable success. Here, I describe two example applications of BA prediction.

First, predictions of MHC presentation are a powerful tool for screening for T cell epitopes given fixed budgets of time, cells, or experiments. A prioritization of potential T cell epitopes in SARS-CoV-2 using predictors (Grifoni, Sidney, et al. 2020) enabled the rational design of pools of peptides enriched for CD8+ T cell epitopes and the detection of CD8+ T cell responses in most convalescent donors (Grifoni, Weiskopf, et al. 2020b). The predictors were likely critical to the feasibility of the study, as a viral proteome containing over ten thousand 9-mer peptides was reduced to 628 predicted MHC class I binders across 12 prevalent HLA-A and HLA-B alleles.

Second, a compelling body of evidence indicates that anti-tumor T cell responses can eradicate tumors by recognizing MHC-presented mutant peptides arising from protein-changing tumor mutations, termed neoantigens (Schumacher, Scheper, and Kvistborg 2019). Personalized cancer vaccines aim to therapeutically elicit a T cell response against predicted neoantigens. In a common protocol, exome sequencing is performed to identify tumor-specific mutations, which are translated *in silico* into potential mutant peptides. Next, MHC class I binding prediction based on the individual's MHC genotype is used to select the peptides most likely to be presented on tumor MHC, which are incorporated into a personalized vaccine and delivered as peptides (Ott et al. 2017) or mRNA (Sahin et al. 2017). This modality is in clinical trials across a range of tumor types (Pan et al. 2018).

Prior work

The first computer program for predicting MHC peptide binding was developed by Sette and colleagues in 1989 (Sette et al. 1989). This method scored peptide sequences based on amino acid similarity for a 6-mer motif present in a model strong-binding peptide for two murine MHC class II alleles. This approach was extended to tens of MHC class I and II alleles as data became available over the next several years, generally with a focus on one or two highly-selective residues, termed "anchors" (Bjoern Peters, Nielsen, and Sette 2020).

A desire to make the models more empirical and quantitative and also take into account secondary contributions from non-anchor residues (Ruppert et al. 1993) led to the first generation of statistical

predictive methods. These methods were based on position specific scoring matrices (PSSMs), in which the amino acid at each position in a peptide was scored independently and the results summed to give a score for the peptide (Rammensee et al. 1999; Bui et al. 2005; Parker et al. 1995). Methods varied in terms of their training dataset, generally peptides identified by Edman degradation or quantitative affinity measurements, and the statistical approach used to fit the PSSM. In some respects, this line of work recapitulated an earlier progression from motifs to matrices for identifying functional regions in nucleic acid sequences (Stormo et al. 1982).

While the PSSM models performed remarkably well, further improvements are theoretically possible by taking into account interactions between peptide residues. Early methods taking this approach using neural networks (Gulukota et al. 1997), decision trees (Segal, Cummings, and Hubbard 2001), or regression models incorporating many interaction terms (Irini A. Doytchinova, Blythe, and Flower 2002) actually underperformed the simpler PSSM models, likely due to overfitting the small datasets available at the time (Björn Peters, Tong, et al. 2003). The stabilized matrix method (SMM), which combined PSSMs with additional terms for interactions among pairs of peptide residues and used regularization to avoid overfitting, was the first method that showed clear improvement by including interactions (Björn Peters, Tong, et al. 2003; Bjoern Peters and Sette 2005).

Accuracy on MHC class I ligand prediction was further improved with the introduction of the artificial neural network method NetMHC (Nielsen et al. 2003; Buus et al. 2003). This model employed several approaches to improve accuracy, including early stopping to avoid overfitting during training, multiple peptide encodings, and the use of an ensemble of neural networks. Only 9-mer peptides are directly modeled by NetMHC. Predictions for other peptide lengths are generated by averaging the predictions of insertions and deletions that bring the peptide to a 9-mer. Further development of NetMHC has progressed using updated training sets and a scheme for incorporation of training data for non-9-mer peptides (Lundegaard et al. 2008; Andreatta and Nielsen 2015).

A key limitation of all statistical models described so far is that, because separate models are developed for each MHC allele, these methods only support the limited number of MHC alleles for which

there is sufficient data to fit a model. For example, the latest version of NetMHC (4.0) supports only 86 human MHC class I alleles. This is a small fraction of the diversity in the human population, and the majority of individuals will have at least one of their MHC class I alleles missing from the supported list. The NetMHCpan tool was introduced to address this limitation (Hoof et al. 2009). In this method, much larger neural networks are trained using a representation of both the peptides and the MHC class I allele. The MHC allele representation was devised by identifying 34 polymorphic positions from the MHC class I binding groove that contact the peptide in peptide/MHC crystal structures. The NetMHCpan method has been updated several times to incorporate larger training sets and, as in NetMHC, to better support non-9-mer peptides (Nielsen and Andreatta 2016; V. Jurtz et al. 2017; Reynisson, Alvarez, et al. 2020).

The NetMHCpan line of work was also noteworthy for its incorporation of MS elution data in its training set, starting in version 4.0 (V. Jurtz et al. 2017). Previously, all models had been trained on binding affinity measurements. This data integration was done by adding a second output node to the neural network. Predictions are generated by the model for both binding affinity (BA) and eluted ligand probability (EL), i.e. the likelihood of detection in an MS experiment. This scheme theoretically enables the model to learn peptide features that are specific to each experimental modality while sharing information between modalities in the peptide feature representation. NetMHCpan 4.0, which I apply in the benchmarks described in chapters 2, 3 and 4, is trained on binding affinity measurements and eluted ligands from monoallelic MS experiments. Very recently, this method has been further updated to use data from multiallelic MS experiments as well (Reynisson, Alvarez, et al. 2020).

Other MHC ligand predictors have been developed based entirely on MS elution data. In the approach pursued by David Gfeller and Michal Bassani-Sternberg, peptides identified in multiallelic MS experiments across individuals with varying MHC genotypes are clustered. The co-occurrence pattern of clusters and MHC alleles is analyzed to associate MHC alleles with particular clusters, indicating the allele motifs. This approach was originally pursued for MHC class I (Gfeller et al. 2018; Bassani-Sternberg and Gfeller 2016) and later extended to class II (Racle et al. 2019). Its key advantage is that it can make use of the large sets of multiallelic MS data available, which is especially important for MHC class II. However,

large-scale efforts to collect monoallelic MS for both class I (Sarkizova et al. 2019) and class II (Abelin et al. 2019) may somewhat temper this advantage. The models also do not take into account interactions between peptide residues. This may be partially mitigated by the ability to associate multiple motifs with a single allele, however. In Chapters 2, 3, and 4 I benchmark against the MixMHCpred 2.0.2 tool, the MHC class I ligand predictor developed using these methods.

This summary of MHC binding prediction omits a large body of work on alternative approaches that have failed to show consistently improved accuracy. For example, a line of work has pursued physics-based models of the interaction between peptide and MHC, using molecular dynamics simulation (Schueler-Furman et al. 2000; Rognan et al. 1994), free energy calculations (Froloff, Windemuth, and Honig 1997), molecular docking (J. M. Khan and Ranganathan 2010), or quantitative structure-activity relationship regression of peptide chemical features on binding affinities (I. A. Doytchinova and Flower 2001). These methods have not shown improved accuracy, probably at least in part because the structural data they typically require (e.g. pMHC crystal structures) is much more scarce than the pMHC affinities or other measurements used to fit the purely statistical models (Bjoern Peters, Nielsen, and Sette 2020).

Prediction of the antigen processing steps prior to MHC binding has also been pursued. The NetChop proteasomal cleavage predictor is the most well-known antigen processing predictor (Keşmir et al. 2002; Nielsen et al. 2005). It has two variants fit to different datasets. The "20S" variant is fit to a small dataset of actual cleavage sites identified by fragments from a digested protein that were identified using MS or Edman degradation. This variant is very rarely used, however. The more popular "C-term" variant is fit to the C-terminal residues of presented MHC class I ligands. This complicates the interpretation of its predictions, as this tool is likely modeling a mixture of cleavage, TAP transport, and MHC binding. NetChop has not been updated since 2005, despite the availability of larger datasets (Wolf-Levy et al. 2018).

Models of TAP transport, as well as combination models that integrate predictions for multiple steps to give a combined presentation prediction, have also been devised (Larsen et al. 2005; Tenzer et al. 2005; Stranzl et al. 2010; Irini A. Doytchinova, Guan, and Flower 2006; Hakenberg et al. 2003; Björn Peters, Bulik, et al. 2003). The NetCTLpan package, for example, integrates an earlier version of

NetMHCpan with NetChop and a TAP transport predictor. The TAP predictor is based on a PSSM-like model that was fit to a dataset of about 500 peptide/TAP affinity measurements (Björn Peters, Bulik, et al. 2003). It focuses on the C-terminal residue and the three N-terminal residues. As TAP transport occurs before ERAP trimming, it also considers a one residue N-terminal extension of the peptide.

A key takeaway from work on antigen processing predictors is that the major sequence biases of antigen processing steps are largely redundant with MHC binding, likely a consequence of co-evolution of this pathway for efficient presentation (Nielsen et al. 2005). For example, the tryptic-like (cleavage after negatively-charged residues) and especially chymotryptic (cleavage after hydrophobic residues) activities of the proteasome generate peptides that are efficiently transported by TAP, which prefers peptides with C-terminal Phe, Tyr, Arg, or Leu (Uebel et al. 1997). These C-termini are also compatible with a range of MHC class I alleles. On the other hand, the third specificity of the proteasome, the caspase-like specificity, cleaves after acidic residues. Peptides with C-terminal acidic residues will not be transported by TAP and, even if they were, would not bind any human MHC class I allele. This gives an explanation for the observation that incorporating predictors of cleavage and TAP transport have not resulted in significant accuracy improvements for epitope prediction over MHC binding prediction alone (Koşaloğlu-Yalçın et al. 2018; Nielsen et al. 2005). On the other hand, the small datasets used to fit these methods mean that they can only account for the most dominant biases of these steps.

Rationale

This dissertation aims to improve prediction of the repertoire of peptides presented on MHC class I. The hypothesis is that the expanding data on pMHC presentation from MS studies may enable the application of more sophisticated modeling approaches than used previously. I focus on MHC class I because, at the time this project was initiated, available MS data for class I far exceeded that of class II. Class I prediction is also more frequently used. However, MHC class II prediction has more room for improvement and future work will benefit from recently-published surveys of class II presented peptides (Racle et al. 2019; Abelin et al. 2019).

As a note on terminology, I refer to the physical phenomenon measured by *in vitro* binding affinity assays as "MHC binding" and its prediction as "MHC binding affinity prediction." By contrast, "MHC presentation" refers to the display of peptides on cell-surface MHC *in vivo*, which is what MS experiments detect. I use "antigen processing" to refer to the steps that peptides undergo prior to MHC association *in vivo*.

This dissertation pursues two key improvements in MHC class I presentation prediction: more accurate peptide / MHC binding affinity prediction and the development of a model of antigen processing. The first goal is treated in Chapter 2 and the first part of Chapter 3, which introduce new allele-specific and pan-allele binding affinity predictors, respectively. The second goal, for which this dissertation is titled, is addressed in the second part of Chapter 3.

MHC class I binding affinity prediction is a well-studied problem, and existing approaches achieve remarkably high accuracy for 9-mer peptides (the preferred length) on well-characterized alleles. The binding affinity prediction work in this dissertation emphasizes improving accuracy for non-9-mer peptides. This is motivated by the observation that MS datasets indicate a wider length profile for some MHC class I alleles than is captured by binding affinity measurements (Figure 1-2). This implies that binding affinity studies — and the older prediction tools designed using these studies — may have over focused on 9-mer peptides. The MS datasets in combination with appropriate modeling approaches may enable more accurate prediction for longer peptides.

Antigen processing prediction has been attempted by an earlier generation of tools that have generally failed to show improved accuracy (Larsen et al. 2005; Tenzer et al. 2005; Stranzl et al. 2010; Koşaloğlu-Yalçın et al. 2018). The rationale for pursuing it here is that the increase in MS datasets has dramatically expanded the data available to develop predictors. Instead of hundreds of peptide/TAP affinities (Björn Peters, Bulik, et al. 2003), we now have hundreds of thousands of known presented peptides. This may enable the development of a significantly improved antigen processing predictor.

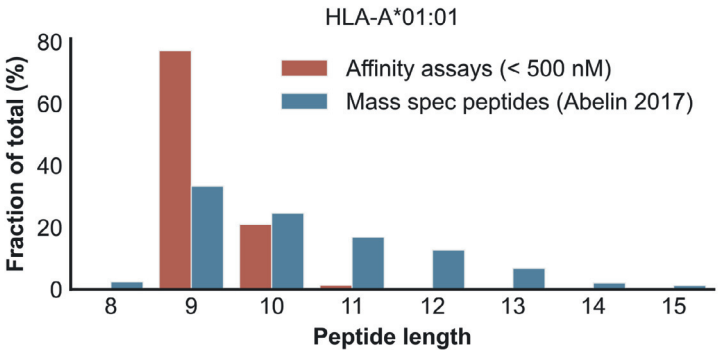


Figure 1-2. Length distribution of HLA-A*01:01-binding peptides identified by MS or affinity assays

The red bars show the length distribution of peptides measured to have affinities tighter than 500 nM for A*01:01 deposited in the Immune Epitope Database (Vita et al. 2019). The blue bars show the length distribution of peptides identified by MS for HLA*01:01 in one study (Abelin et al. 2017).

An additional motivation for this project is the opportunity to develop a software package for MHC ligand prediction that is open source, documented, and convenient to use. My earlier work in tumor neoantigen discovery has suggested that existing tools have room for improvement in these areas (T. O'Donnell et al. 2018). An accurate predictor that adheres to software development best-practices has the potential for wide use.

Approach

The predictive models introduced in this dissertation are made possible by a substantial set of available training data and a toolbox of techniques and software recently developed by the machine learning community. This section gives an overview of the key methods used to develop the MHC binding and antigen processing predictors.

Data curation and integration

The most important factor in developing a high-performing predictor is generally the size and quality of the training data. For the predictors introduced here, I frequently found that updates in training data led to larger accuracy gains than other kinds of improvements. The most recent version of the training data, which was used to fit the pan-allele MHC binding predictor in Chapter 3, consists of approximately a half million MS-

identified peptides and a quarter million affinity measurements. These derive from over 600 published studies. The curation task is greatly facilitated by the Immune Epitope Database (IEDB), which provides the majority of this set (Vita et al. 2019). I also include data from some other curation projects, such as the SysteMHC Atlas project (Shao et al. 2018), as well as MS data manually extracted from recent publications.

For the MHC binding predictors, a key question is how to integrate affinity measurements and MS-identified peptides. These are intrinsically different kinds of data: a predictor of binding affinity is solving a regression problem, whereas a predictor of MS-identified peptides is solving a binary classification problem. The method I introduce for integrating these data sources is based on the observation that most peptides identified by MS have a strong binding affinity, often 100 nanomolar or tighter. While MS does not determine the precise affinity, it is possible to train a binding affinity predictor on MS data by applying a training objective (loss function) that supports inequalities. In the framework I introduce, binding affinity measurements can be specified as exact affinities, but MS peptides are assigned an inequality, for example "100 nM or tighter." This enables the binding affinity predictors to include MS data in their training sets.

For the antigen processing predictor, which is fit to MS data only, there are two considerations. First, as MS datasets include only positively-identified peptides (referred to as hits), a method is needed to generate negative data points (referred to as decoys) to derive a classification label (hit vs. decoy). I do so by sampling decoy peptides from the same genes as the hits. This is a rudimentary way of controlling for source protein expression. Second, MHC binding exerts a much stronger influence on the presented peptide repertoire than antigen processing. If an antigen processing predictor were trained directly on MS hits and decoys, it would primarily model an average of MHC binding preferences for the most well-represented MHC alleles in the training set. I introduce a simple approach to control for MHC binding in the antigen processing predictor training set: only hits and decoys that are first predicted by the MHC binding predictor to bind tightly to the relevant MHC allele are used to train the antigen processing predictor. In this way, the antigen processing predictor models the residual sequence properties beyond MHC binding that distinguish hits from decoys.

Machine learning models

Three new predictive models are developed in this dissertation, two MHC binding predictors and an antigen processing predictor. The simpler of two MHC predictors, introduced in Chapter 2, uses separate models for each MHC allele. These allele-specific models take a peptide and predict the binding affinity for the relevant allele. This approach has the disadvantage that it only works for alleles with substantial training data, as no information is shared between alleles. In Chapter 3, I extend this work to pan-allele models, which take both a peptide and a representation of the MHC allele. These models can generate predictions even for alleles with little or no training data, although accuracy is still best for those with the most data. The model of antigen processing (Chapter 3) takes as input a peptide plus adjacent sequence from its source protein, and predicts an antigen processing score. This score is intended to capture effects other than MHC binding that affect peptide presentation.

All three predictors are implemented using neural networks. For the purposes of this introduction, a neural network can be considered a stack of simple transformations, referred to as hidden layers (Goodfellow, Bengio, and Courville 2016). Each hidden layer transforms an input vector to an output vector by multiplying by a matrix and applying a nonlinear transformation. The output vector from one layer is the input to the next. The transform matrices are determined by numerical optimization to maximize accuracy on the training data. While each layer in a neural network computes a simple function, the stacking of layers allows the network to model complex feature interactions and decision boundaries. The number of layers, their sizes, and other properties that are selected prior to model fitting are called the neural network architecture. Networks with deeper (more layers) and wider (larger hidden layers) architectures can model more complex functions, but may require more training data and/or sophisticated model fitting procedures.

The predictors developed in this dissertation use multiple neural networks with varying architectures. These are collectively called an ensemble. Predictions are averaged across the networks in the ensemble to obtain a consensus prediction, which is generally more reliable than that of any individual model. The ensemble is selected by a model selection procedure that identifies the best-performing

networks from a large number of possibilities using held-out data. As an example, for the MHC binding affinity predictor developed in Chapter 3, an initial set of 140 networks was used to select 10 for the final ensemble.

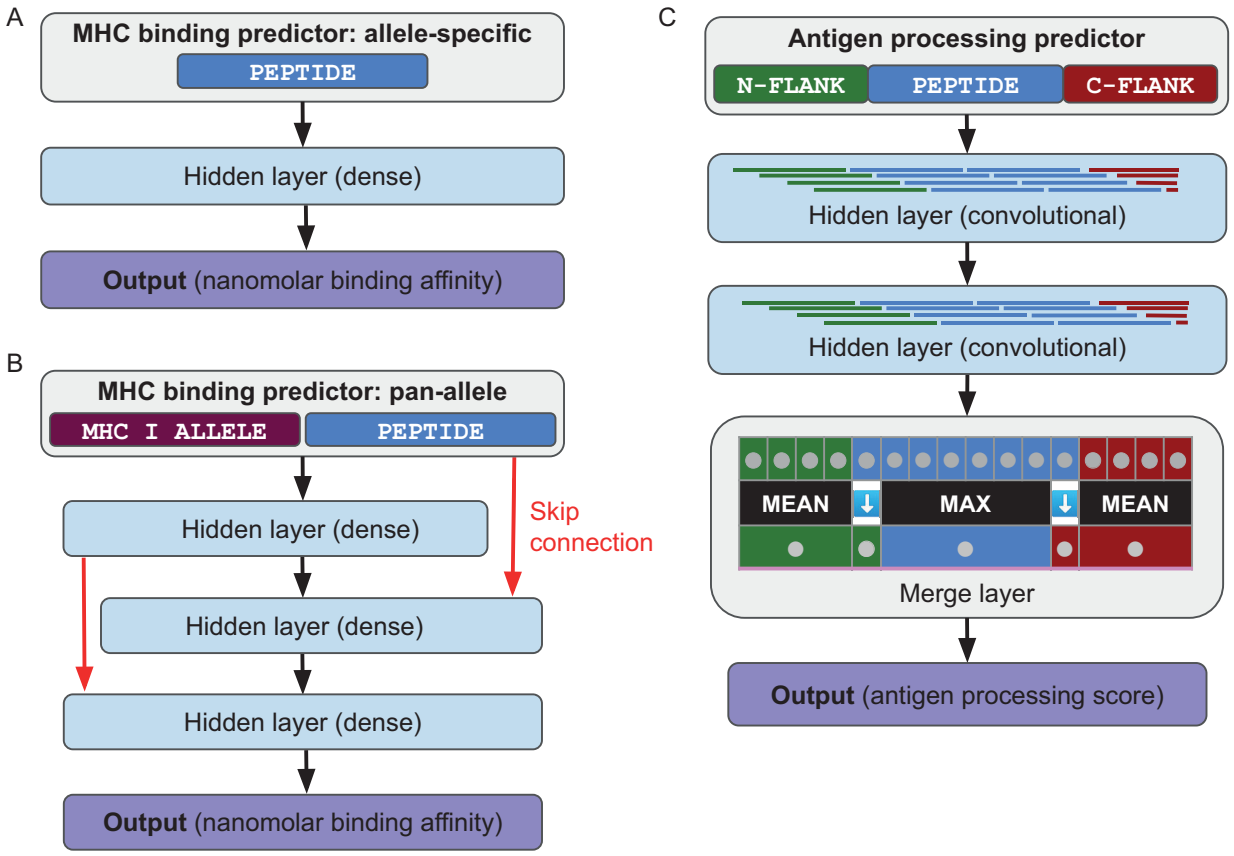


Figure 1-3. Neural network architectures for prediction of MHC binding and antigen processing

(a) Architecture for allele-specific MHC binding prediction. In this approach, separate models are trained for each MHC allele. **(b)** Architecture for pan-allele MHC binding prediction. An individual model can generate predictions for any MHC allele. **(c)** Architecture for antigen processing prediction. For all three predictors, the architectures shown are examples from a range of architectures used in an ensemble.

The MHC binding prediction models consist of one to three hidden layers for the allele-specific predictor (Chapter 2) and two to three hidden layers for the pan-allele predictor (Chapter 3). Example allele-specific and pan-allele architectures are shown in Figure 1-3A and B, respectively. The architectures used for pan-allele binding prediction are capable of modeling more complex functions than the single-layer neural networks used in other tools such as NetMHCpan. However, they are also more difficult to train as

the increase in depth makes them more susceptible to issues such as overfitting and failure to find a viable solution. I apply several techniques to overcome these issues. These include skip connections, in which some layers are connected to the two preceding layers, as well as various forms of regularization (Srivastava et al. 2014; Krogh and Hertz 1992). I also employ a pre-training step, in which millions of synthetic binding affinities are generated using the allele-specific predictors. These synthetic measurements are used to train the pan-allele predictors prior to training on real data. This allows the training on real data to begin from a viable position in the parameter space.

The antigen processing predictor introduced in Chapter 3 uses a neural network architecture motivated by proteasomal cleavage (Figure 1-3C). The intuition for this architecture is that presented peptides must be cleaved at their termini but not at internal residues. The initial layers of the model predict cleavage favorability at each position in the peptide and flanking sequence. This is implemented using convolutional layers, which apply the same learned transformations independently to contiguous patches (i.e. subsequences) of the input. In a subsequent layer, the predictions at the peptide termini (i.e. the actual cleavage sites) and the maximum prediction across the internal residues of the peptide are extracted. The output layer uses these values to make an overall prediction that favors peptides likely to be cleaved at their termini but not at internal residues.

Closely related to neural network architecture is the method of input representation to the neural network. For MHC binding prediction, a fundamental issue is that these neural networks require a fixed-size vector of inputs, but peptides vary in length. The standard approaches for representing variable-length inputs to neural networks, convolutional and recurrent layers, did not perform well in preliminary testing, leading me to develop custom approaches for peptide representation. In Chapter 2, I use a scheme that maps the ends of the peptide to fixed positions in the representation, and allows the middle of the peptide to fill in according to the length. In Chapter 3, this scheme is replaced with an approach in which the peptide is represented redundantly in three different ways: left aligned, centered, and right aligned. Both methods work by allowing the network to access residues at either end of the peptide at fixed positions in the input

representation. For example, the second and last peptide positions, which are often anchor residues, always map to the same fixed positions in the input representations, regardless of peptide length.

In addition to the peptide, the pan-allele binding predictors also require a representation of the MHC allele. Here, I use the same approach as NetMHCpan, in which the MHC residues at a fixed set of polymorphic positions are given as input to the network. These positions were selected by the NetMHCpan authors on the basis of high variability and close proximity to the bound peptide in crystal structures. The predictors developed in Chpapter 3 extend the 34 positions used by NetMHCpan slightly, to a total of 37 polymorphic MHC positions.

Benchmarking

Benchmark design is a crucial factor for identifying model improvements and validating the final predictors. A recurring challenge is that, as essentially all publicly-available MHC binding and presentation data is integrated to develop predictors such as NetMHCpan, novel sources of validation data must be found in order to compare new approaches to standard tools. In Chapter 2, this is accomplished using unpublished binding affinity measurements provided by a collaborator. I also use published MS-identified peptides to compare predictor variants trained without MS datasets. In Chapter 3, I take advantage of the fact that the predictors under evaluation are trained on only monoallelic MS experiments – i.e., those that identify peptides presented by a single MHC allele – which enables me to use published multiallelic MS datasets for benchmarking.

In the MS benchmarks, decoy peptides are sampled from the same genes as the hits, similarly to the approach used to train predictors. The predictors are scored on their ability to distinguish hits from decoys, quantified using two accuracy metrics.

A very commonly-used metric for machine learning classifiers is the area under the receiver operating characteristic curve (AUC). The AUC can be interpreted as the probability that a randomly-selected hit is scored higher by a model than a randomly-selected decoy. A perfect predictor would score

1.0 in AUC, and a random predictor would score 0.5. The AUC is reported throughout this dissertation because it is standard and easily interpreted.

The most common use for predictors such as those developed here is to select a small subset of highly-ranked peptides from a large universe of possibilities. This scenario is not captured well by AUC, which focuses on comparisons between pairs of peptides. For a more realistic measure of accuracy on MS-identified peptides, Abelin and coworkers introduced a metric they refer to as positive predictive value (PPV) (Abelin et al. 2017). To calculate a predictor's PPV on an MS experiment, the *n* MS-identified hits (for a typical example, n=1000) are matched to *99n* unobserved decoy peptides. The *100n* hits and decoys are sorted by prediction, and the fraction of the top n that are hits is calculated (Figure 1-4). A perfect predictor would rank all hits above all decoys, resulting in a PPV of 1.0. A random predictor would score 0.01 in PPV. The PPV emphasizes the high-end of a predictor's output, whereas the AUC weights all regions equally. In some cases, the two metrics can lead to different conclusions, as seen in Chapter 3.

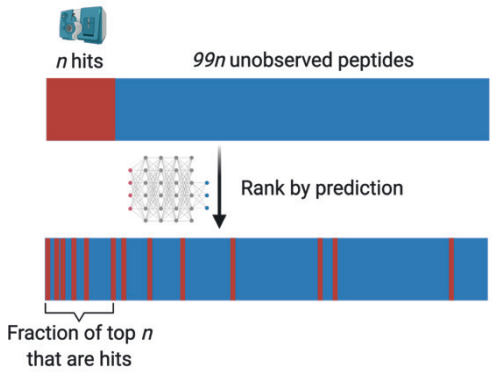


Figure 1-4. Positive predictive value (PPV) calculation

The *n* MS hits (red bars) and *99n* unobserved decoys (blue) are sorted by prediction. The PPV is defined as the fraction of the top *n* predictions that are hits.

In Chapter 4, I apply a benchmark of T cell epitopes to compare predictive approaches. The number of T cell epitopes is small in these benchmarks. For example, the tumor neoantigen benchmark has only 52 T cell epitopes out of 2,841 tested peptides. In this setting, the PPV has very high statistical uncertainty. I instead rank all sequences by prediction and report the position in the ranking that captures a specified fraction (e.g. 50%) of the T cell epitopes.

Software implementation

The predictors described here are implemented in Python using the tensorflow deep learning library (Abadi et al. 2016). This library provides a very large set of functionality for developing sophisticated predictive models. For example, it makes it straightforward to use graphics processing units (GPUs) to accelerate model fitting. This is essential for training the complex models described in Chapter 3.

The predictors introduced in this dissertation are available in the open source MHCflurry software package (https://github.com/openvax/mhcflurry). MHCflurry aims to be a practical tool. It is licensed under a permissive open source license (Apache 2), documented (https://pypi.org/project/mhcflurry/), and may be installed from the Python package index. In addition to the predictors themselves, MHCflurry includes the scripts and training data used to fit them, a requirement for precise accuracy comparisons against new approaches developed in the future. A step-by-step guide to fitting the predictors developed in Chapter 2 is available as a book chapter (T. O'Donnell and Rubinsteyn 2020). My hope is that MHCflurry will be a useful tool for bioinformaticians as well as a reference implementation that helps others to develop further improvements.

Outline

The remaining chapters of this dissertation are organized as follows.

In Chapter 2, I introduce the first version of the MHCflurry peptide/MHC class I binding affinity predictor. The predictor applies a new approach for integrating MS-identified peptides with binding affinity measurements. It also introduces a technique for representing peptides that accounts for the way longer peptides are typically accommodated in the MHC binding groove. I show that a model applying these techniques has improved accuracy over existing approaches, especially for peptides that deviate from the canonical nine amino acid length.

In Chapter 3, I first extend the simple predictor developed in Chapter 2, which supports only 112 MHC alleles, to support over 14,000 alleles. I next introduce a scheme for fitting a model of antigen processing from pMHC detected by MS. I find that the resulting model recapitulates known sequence biases

of key antigen processing steps. A combination model incorporating both the MHC binding predictor and the antigen processing model outperforms the individual models alone as well as existing tools at predicting presented pMHC.

In Chapter 4, I benchmark the new models as well as existing tools at predicting neoantigens and viral epitopes. I find that the combination model introduced in Chapter 3 outperforms the others on the neoantigen benchmark, but not the benchmark of viral epitopes. I discuss possible reasons for these divergent results, which may relate to the way the epitopes used in the benchmarks were identified. I also perform a correlative analysis of the antigen processing predictor from Chapter 3 with earlier tools for antigen processing prediction. I find that it differs from these tools largely due to sensitivity to a wider range of positions in the peptide.

Finally, Chapter 5 summarizes potential future directions for the work described here and for the related problems of MHC class II ligand prediction and T cell epitope prediction.

Chapter 2. MHCflurry: open-source class I MHC binding affinity prediction

This chapter is adapted from the following article:

O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Systems.* 2018.

Summary

Predicting the binding affinity of major histocompatibility complex I (MHC class I) proteins and their peptide ligands is important for vaccine design. We introduce an open-source package for MHC class I binding prediction, MHCflurry. The software implements allele-specific neural networks that use a novel architecture and peptide encoding scheme. When trained on affinity measurements, MHCflurry outperformed the standard predictors *NetMHC 4.0* and *NetMHCpan 3.0* overall and particularly on non-9-mer peptides in a benchmark of ligands identified by mass spectrometry. The released predictor, *MHCflurry 1.2.0,* uses mass spectrometry datasets for model selection and showed competitive accuracy with standard tools, including the recently released *NetMHCpan 4.0,* on a small benchmark of affinity measurements. On a single CPU, MHCflurry's prediction speed exceeded 7,000 predictions per second, 396 times faster than *NetMHCpan 4.0*. MHCflurry is freely available to use, retrain, or extend, includes Python library and command line interfaces, may be installed using package managers, and applies software development best practices.

Introduction

Adaptive immunity depends on T cell recognition of peptides bound to major histocompatibility complex I (MHC class I) proteins on cell surfaces. There are thousands of MHC class I alleles in the human population, each with specificity for binding a distinct set of peptides, which, when displayed by MHC, can be the target of an immune response. Computational prediction of the binding affinity between a specified

peptide and MHC allele has found wide application in infectious diseases, autoimmunity, vaccine design, and cancer immunotherapy (Lundegaard et al. 2007; Ott et al. 2017).

The NetMHC and NetMHCpan tools are considered the state-of-the-art predictive models for this task (Trolle et al. 2015). NetMHC uses an "allele-specific" approach, in which separate predictors are trained for each MHC allele; the input to the model is the peptide of interest (Andreatta and Nielsen 2015). NetMHCpan uses a "pan-allele" approach, in which a single model takes as input both the peptide and a representation of the MHC allele (Nielsen and Andreatta 2016). Both *NetMHC 4.0* and *NetMHCpan 3.0* are ensembles of shallow neural networks trained on affinity measurements deposited in the immune epitope database (IEDB) (Vita et al. 2015). The recently released *NetMHCpan 4.0* additionally includes peptides eluted from MHC and identified by mass spectrometry (MS) in its training set, generating separate predictions for binding affinity and likelihood of MS identification using two-output neural networks (V. Jurtz et al. 2017).

We describe and benchmark a package of allele-specific class I MHC binding predictors, *MHCflurry 1.2.0*. MHCflurry predictors show competitive accuracy with the *NetMHC* tools and a significant speed improvement while addressing a number of limitations. In particular, MHCflurry is open source, publishes the data and workflow used to train models, exposes library and command line interfaces, may be installed using the Python package manager, and applies software development best practices, including unit testing and code documentation.

Implementation

Each supported MHC allele is associated with an ensemble of 8-16 neural networks trained on affinity measurements from IEDB and other sources. In the default MHCflurry predictor (*MHCflurry 1.2.0*), MS data and held-out affinity measurements are used to select 8-16 models for each allele. The software also includes two experimental predictors: *MHCflurry (no MS)*, which does not use MS datasets, and *MHCflurry (train-MS)*, which uses MS datasets for both training and model selection (Figure 2-1A).

MHCflurry introduces a peptide representation in which peptides of length 8-15 are encoded as fixed-length vectors. The motivation for the encoding is to preserve the positionality of the residues that make the most important stabilizing contacts with the MHC molecule. These "anchor positions" occur toward the beginning or end of the peptide for most alleles. MHCflurry represents peptides as length 15 sequences, in which missing residues are filled with an *X* character, effectively a 21st amino acid (Figure 2-1B). The first four and last four residues in the peptide map to the first four and last four positions in the representation. The middle seven residues are filled as needed: an 8-mer leaves all middle positions as an *X* whereas a 15-mer fills all positions. In this way, peptides of length 8-15 are supported, and the positions most likely to contain anchor residues are consistently mapped to the same positions in the representation.

The MHCflurry models are feedforward neural networks composed of the following layers: the length-15 peptide representation with each residue encoded by its vector in the BLOSUM62 substitution matrix (Henikoff and Henikoff 1992), zero, one, or two locally connected layers, a fully connected layer, and a sigmoidal output (Figure 2-1C). Locally connected layers are one dimensional convolutional layers without weight sharing; each neuron receives a neighborhood of adjacent points.

For alleles with fewer than 1,000 training measurements, a pre-training step is used to set the initial model weights. In this step, the training data is augmented with measurements from similar alleles to make a set of at least 1,000 measurements, where similarity is in terms of the BLOSUM62 similarity score of the binding-core residues of the allele. The model is first trained using the augmented training set, then re-trained on the non-augmented set starting from the weights learned in the first step.

To train and select MHCflurry models, we assembled a dataset of 230,735 affinity measurements across 130 alleles from IEDB and a published benchmark (Kim et al. 2014). Ten percent of this data plus 226,684 MS-identified ligands deposited in IEDB, SysteMHC Atlas (Shao et al. 2018), or published elsewhere (Abelin et al. 2017) were used to perform model selection. Of 130 alleles for which preliminary full ensembles were trained, 112 showed sufficient performance on the model selection dataset to include in the final predictor. For these 112 alleles, the model selection dataset was used to select 8-16 of the 320 total models trained per allele using a forward stepwise selection procedure (Caruana et al. 2004).

MHCflurry predicts quantitative binding affinities, but one fourth of the entries (57,828 of 230,735) in the affinity dataset are qualitative, represented as *positive*, *positive-high*, *positive-intermediate*, *positive-low*, or *negative*. To use these measurements, the MHCflurry models are trained using a modification to the mean square error (MSE) loss function, in which measurements may be associated with an inequality, *(>)* or *(<)*, and contribute to the loss only when the inequality is violated. For example, we assigned measurements represented as *positive-high* the value "< 100 nM", as such peptides are likely to have binding affinities tighter than (i.e. less than) 100 nM. During training, these peptides contribute to the loss only when their predictions are greater than 100 nM. For the *MHCflurry (train-MS)* predictor, this approach is used to include MS-identified ligands, which are assigned a "< 500 nM" value.

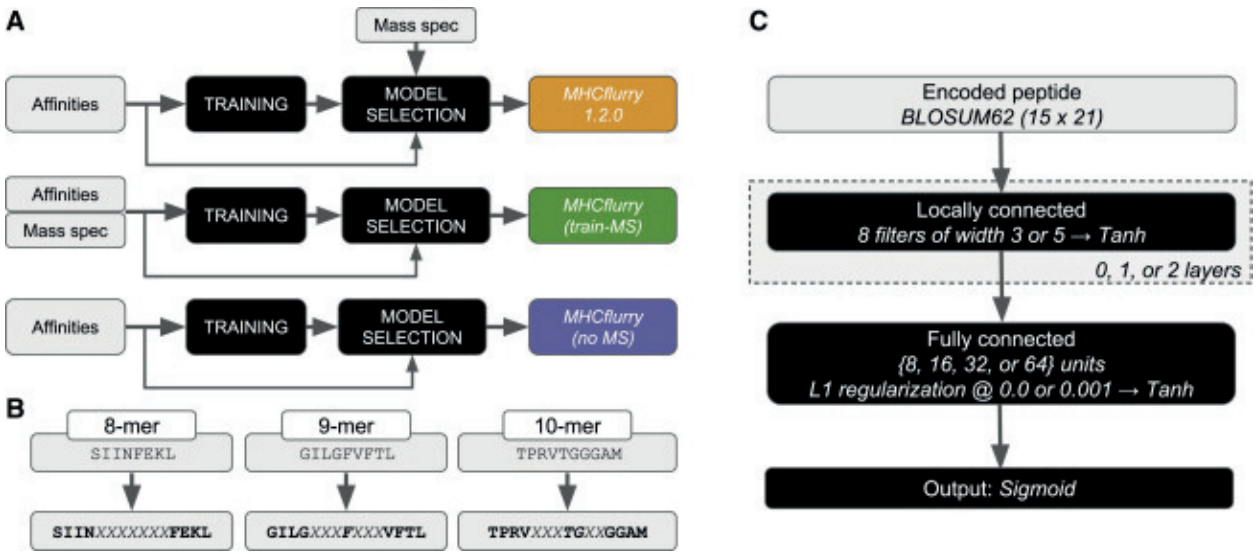


Figure 2-1. MHCflurry variants, peptide representation, and neural network architectures

(a) Training and model selection schemes for the three MHCflurry variants. **(b)** Encodings for three example peptides. **(c)** Neural network architectures. Layers with trained weights are shown in black.

Results

As the predictors considered have similar accuracy, power to identify differences requires a large held-out validation dataset. However, nearly all published measurements are potentially included in the training data for the NetMHC tools. We therefore adopted a strategy based on two benchmarks, in which first a large

recently published MS dataset is used to benchmark the *MHCflurry (no MS)* predictor against versions of the NetMHC tools trained without MS, *NetMHC 4.0* and *NetMHCpan 3.0*. In a second benchmark, we apply a smaller dataset of unpublished affinity measurements to assess accuracy across all tools, including the default *MHCflurry 1.2.0* predictor and *NetMHCpan 4.0*, which include MS datasets in their training pipelines. In the first benchmark, we find that MHCflurry outperforms the other predictors due to improved accuracy on non-9-mer peptides. In the smaller benchmark, *MHCflurry 1.2.0* narrowly outperformed *NetMHC 4.0* and *MHCflurry (no MS)*, but performed equivalently within statistical uncertainty to *MHCflurry (train-MS)*, *NetMHCpan 3.0,* and *NetMHCpan 4.0.*

The MS benchmark consists of a published set of 23,651 MHC ligands (Abelin et al. 2017). We scored predictors by their positive predictive value (PPV) at differentiating MS-identified peptides from decoys sampled from the same transcripts (Methods). *MHCflurry (no MS)* showed improved accuracy over *NetMHC 4.0* and *NetMHCpan 3.0,* outperforming these predictors on 15/15 and 12/15 alleles tested, respectively (binomial test $p < 0.02$ for both tests; Figure 2-2A and Supplemental Table 2-1). Across alleles, MHCflurry showed a median 16.1% (range 3.8 – 110.5) higher PPV than *NetMHC 4.0* and 5.0% (-5.4 – +18.0) higher PPV than *NetMHCpan 3.0*.

The accuracy advantage of MHCflurry over the NetMHC tools was due to better performance on non-9-mer peptides, which offset nominally lower accuracy on 9-mers relative to *NetMHCpan 3.0* (Figure 2-2B). On non-9-mers (i.e. peptides of lengths 8 and 10-15), MHCflurry outperformed *NetMHC 4.0* on 15/15 alleles (median PPV improvement 34.1%; range 7.1 - 224.4; $p < 0.0001$) and *NetMHCpan 3.0* on 13/15 alleles (median 10.2%; -2.8 - +41.9; $p=0.004$). On 9-mer peptides, no tool showed significant performance improvement over the others: *MHCflurry (no MS)* outperformed *NetMHC 4.0* on 9/15 alleles ($p=0.3$), with a median difference in PPV of 0.5% (range -6.4 - +9.0), but underperformed *NetMHCpan 3.0* on 10/15 alleles ($p=0.15$) with a median difference in PPV of -0.5% (range -8.3 - +6.9).

The second benchmark uses an unpublished set of affinity measurements for 475 human papillomavirus 16 (HPV16)-derived peptides in cell-based competitive binding assays across seven MHC alleles. Several accuracy metrics were computed for each predictor (Figure 2-2C; Methods). While the

small size of the dataset led to high uncertainties in absolute scores, an analysis of the differences in score relative to *MHCflurry 1.2.0* within bootstrap resamples indicated that *MHCflurry 1.2.0* outperformed *NetMHC 4.0* and *MHCflurry (no MS)* on some metrics (Supplemental Table 2-2). However, no differences were detected between *MHCflurry 1.2.0* and the *NetMHCpan 4.0*, *NetMHCpan 3.0*, or *MHCflurry (train-MS)* predictors.

When running a large numbers of predictions, MHCflurry was substantially faster than the other predictors (Figure 2-2D). Using only the CPU, *MHCflurry 1.2.0* approached 7,500 predictions / second, which is 396 times faster than *NetMHCpan 4.0* and 12 times faster than *NetMHC 4.0*. Use of a graphics processing unit (GPU) did not substantially improve MHCflurry prediction performance.

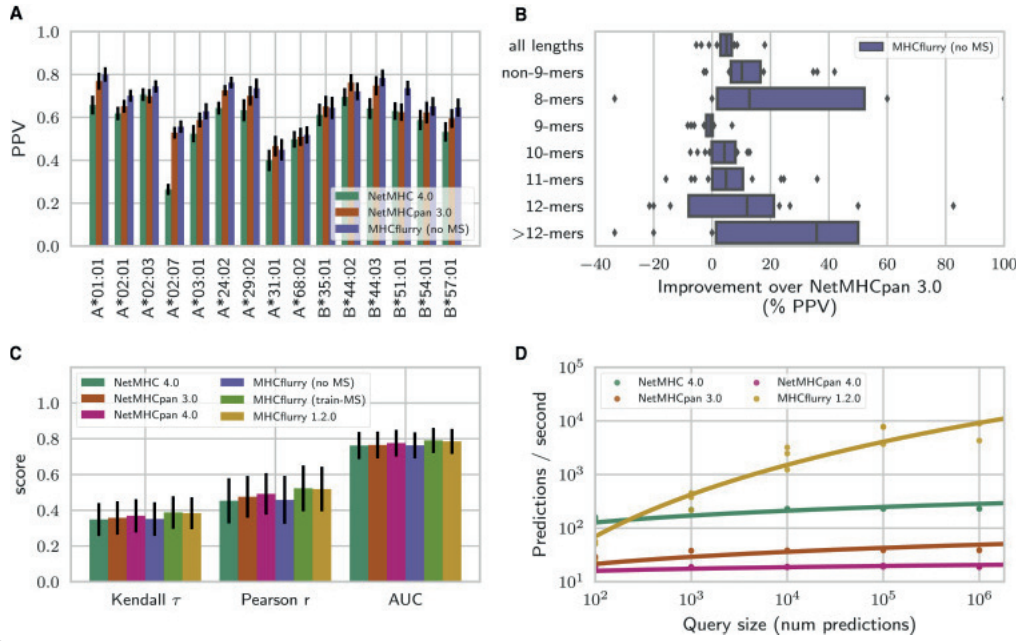


Figure 2-2. Benchmark results

(a) Positive predictive value (PPV) of *NetMHC 4.0*, *NetMHCpan 3.0*, and the *MHCflurry (no MS)* predictors on the mass spec (MS) benchmark. The *NetMHCpan 4.0*, *MHCflurry (train-MS)*, and *MHCflurry 1.2.0* predictors are excluded from this evaluation because their training or model selection datasets include mass spec data. **(b)** *MHCflurry (no MS)* accuracy difference with respect to *NetMHCpan 3.0* for each peptide length across alleles in the MS benchmark. The median line is indicated, boxes show quartiles, and points indicate alleles outside the interquartile region. The >12-mers category includes 13-, 14-, and 15-mers. **(c)** Kendall rank correlation coefficient, Pearson correlation over log affinities, and area under the receiver operating characteristic curve (AUC) on the HPV dataset. **(d)** Prediction speed on subsets of the MS dataset consisting of a single allele and a varying number of peptides. Error bars in (a) and (c) indicate bootstrap 95% confidence intervals.

Discussion

MHCflurry is an open source package for MHC class I affinity prediction with a fast and documented implementation. On the MS benchmark, *MHCflurry (no MS)* outperformed the *NetMHC 4.0* and *NetMHCpan 3.0* tools overall and particularly on non-9-mer peptides. While part of this advantage may be due to improvements such as explicit support for variable length peptides, differences in training datasets between the tools, which are difficult to assess as those for the NetMHC tools are not released, likely also contribute.

As MHCflurry is an allele-specific predictor, only a fixed set of alleles are supported. Pan-allele predictors such as NetMHCpan remain the best option for alleles with little data. However, the data required to fit a MHCflurry model can be modest. For example, the A*02:07 allele has 126 measurements in the affinity measurement set, the fewest of any allele tested in the MS benchmark. As expected, *NetMHC 4.0* performs poorly, with a PPV of 0.26 on this benchmark. The *MHCflurry (no MS)* predictor, however, performs respectably (PPV=0.56), in fact narrowly outperforming *NetMHCpan 3.0* (PPV=0.53). This is largely due to the pre-training step used in MHCflurry, which enables models to borrow information from similar alleles.

The standard *MHCflurry 1.2.0* predictor is trained on affinity measurements and uses MS ligands only for model selection. This is a conservative choice that minimizes the potential impact of any biases associated with ligands identified by MS, such as depletion of cysteines (Abelin et al. 2017). However, the *MHCflurry (train-MS)* predictor, which includes MS in its training set, showed good performance in the HPV benchmark and may eventually become the default MHCflurry predictor.

Acknowledgements

We thank Mike Rooney for helpful discussions. Generation of the HPV dataset was supported by intramural funding of the German Cancer Research Center (DKFZ), grant number TTU 07.706 by the German Center

for Infection Research (DZIF) to A.B.R, and a PhD scholarship to M.B. by the Helmholtz International Graduate School of the DKFZ. This work was supported by the Parker Institute for Cancer Immunotherapy.

Author contributions

Timothy O'Donnell and Alex Rubinsteyn (Mount Sinai) developed MHCflurry. Timothy O'Donnell benchmarked the software and wrote the manuscript. Maria Bonsack and Angelika B. Riemer (German Cancer Research Center, Heidelberg, Germany) performed the HPV peptide binding experiments and advised on benchmarking approaches. Uri Laserson and Jeff Hammerbacher (Mount Sinai) provided computational resources and advice.

Declaration of Interests

The authors declare no competing interests.

Methods

Table 2-1. Key Resources for Chapter 2

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
8-11-mer peptides derived from HPV type 16 Protein E6	DKFZ Genomics and Proteomics Core Facility – Peptide Synthesis	UniProtKB - P03126
8-11-mer peptides derived from HPV type 16 Protein E7	DKFZ Genomics and Proteomics Core Facility – Peptide Synthesis	UniProtKB - P03129
β2-microglobulin	MP Biomedicals EMEA, Illkirch, France	Cat#153903.5 MDL#MFCD00130615
Deposited Data		
IEDB affinity data	Vita et al. 2015	http://www.iedb.org/doc/mhc_ligand_full.zip Downloaded on Dec. 1, 2017

Table 2-1. Key Resources for Chapter 2 *Continued from previous page.*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BD2013	Kim et al. 2014 http://tools.iedb.org/main/datasets/	http://tools.iedb.org/static/main/benchmark_mhci_reliability.tar.gz
Peptides eluted from MHC class I and identified by mass spec	Abelin et al. 2017	Table S1
MHCflurry 1.2.0 models	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.1
MHCflurry (no MS) models	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.1
MHCflurry (train-MS) models	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.1
Curated training and model selection dataset	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.1
MS benchmark dataset	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.1
Experimental Models: Cell Lines		
1341-8346	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW01060
BSM	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09032
E481324	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09011
EA	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09081

Table 2-1. Key Resources for Chapter 2 *Continued from previous page.*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FH8	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09382
LKT3	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09107
WT100BIS	International Histocompatibility Working Group Cell Bank, Seattle, WA, USA	IHW09006
Software and Algorithms		
MHCflurry 1.2.0	This paper	https://github.com/openvax/mhcflurry
NetMHC 4.0	Andreatta & Nielsen 2015	http://www.cbs.dtu.dk/cgi-bin/sw_request?netMHC
NetMHCpan 3.0	Nielsen & Andreatta 2016	http://www.cbs.dtu.dk/cgi-bin/sw_request?netMHCpan+3.0
NetMHCpan 4.0	Jurtz et al. 2017	http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?netMHCpan
Keras 2.1.2	Github	https://github.com/keras-team/keras
Other		
HPV dataset	This paper; and Mendeley Data	http://dx.doi.org/10.17632/8pz43nvvxh.3

Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Timothy O'Donnell (timothy.odonnell@icahn.mssm.edu).

Software implementation

MHCflurry is implemented in Python using the Keras neural network library (https://github.com/keras-team/keras). In our experiments we used the tensorflow backend (https://www.tensorflow.org).

Model inputs and outputs. Each allele is associated with an ensemble of 8-16 neural networks selected from 320 models. The predicted nanomolar affinity is taken to be the geometric mean of the individual model outputs. The variance of the predictions across the ensemble gives an indication of uncertainty and is made available to users, as is the quantile of each prediction among a large set of random peptides (100,000 for each length 8-15) pre-computed for each allele.

As in the NetMHC tools, MHCflurry internally transforms binding affinities to values between 0.0 and 1.0, where 0.0 is a non-binder and 1.0 is a strong binder. The neural networks are trained using the transformed values and the inverse transform is used to return prediction results as nanomolar affinities. The transform is given by $1 - log_{50000}(x)$ where x is the nanomolar affinity. Affinities are capped at 50,000 nM.

The input to each MHCflurry model is a peptide encoded as a 15 x 21 matrix. Rows give the BLOSUM62 encoding of each residue in the peptide, after transforming it to a length-15 sequence as described. As BLOSUM62 is a substitution matrix, amino acids are thus represented by their similarity to the other amino acids (Henikoff and Henikoff 1992). We define the special "no residue" *X* character used in the 15-mer peptide representation to have 0.0 similarity to all other amino acids and 1.0 similarity to itself. No representation of the MHC allele, such as its amino acid sequence, is provided to the network.

Training. Models are drawn from 40 architectures trained on either (1) all affinity data for an allele or (2) only quantitative affinity data. For each of these 80 possibilities, four replicates are trained, for a total of 320 models per allele.

Models are trained for all alleles with at least 25 affinity measurements. Ten percent of the training data is set aside for model selection. Each neural network is trained on a different 90% sample of the remaining data, with the other 10% used as a test set for early stopping. Training proceeds with the RMSprop optimizer using a minibatch size of 128 until the accuracy on the test set has not improved for 20 epochs. At each epoch, 25 synthetic negative peptides for each length 8-15 are randomly generated. These random negative peptides are sampled so as to have the same amino acid distribution as the training peptides and are assigned affinities >20,000 nM. For the *MHCflurry (train-MS)* variant, the number of random peptides for each length is $0.2n + 25$where n is the number of training peptides.

A modified mean squared error (MSE) loss function that supports data with inequalities is used for both the training loss and test set accuracy metric. For this loss function, measurements are associated with an inequality: *(<)*, *(>)*, or *(=)*. The loss *L* is defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{n} \sum_{i}^{n} l(\hat{y}_i, y_i)$$

$$l(\hat{y}_i, y_i) =$$

$(max(\hat{y}_i - y_i, 0))^2$ *if inequality for measurement i is (<)*

$(max(y_i - \hat{y}_i, 0))^2$ *if inequality for measurement i is (>)*

$(\hat{y}_i - y_i)^2$ *if inequality for measurement i is (=)*

where n is the total number of measurements and $\hat{y}_i$ and y_i are the predicted and measured values for measurement i, respectively. Quantitative affinity data is associated with an inequality of *(=)*. For qualitative affinity data, we assigned the following inequalities and measurement values: *positive-high*, < 100 nM; *positive*, < 500 nM, *positive-intermediate*, < 1,000 nM; *positive-low*, < 5,000 nM; *negative* > 5,000 nM. In the *MHCflurry (train-MS)* variant, MS-identified ligands are assigned the value "< 500 nM."

Model selection. Model selection for *MHCflurry 1.2.0* uses the 10% of affinity measurements held out during training augmented by ligands identified in MS elution experiments. Ensembles are evaluated using the variant of mean squared error previously described for affinity data and positive predictive value (PPV) on MS data. The decoy set used for the MS evaluation consists of all other peptides identified by MS for an allele other than the one in question. The final score is the average of the MSE and PPV scores weighted by the number of observations contributing to each.

As the first step in model selection, the full ensembles (320 models per allele) are evaluated to identify alleles for which training failed to give an acceptable predictor, which is generally due to insufficient data. This is done using a permutation test, requiring that the score of the actual predictions of the full ensemble fall above the 95th percentile among the scores obtained by randomly permuting the measurement labels. This can be interpreted as rejecting the null hypothesis that the ensemble is a random predictor. Alleles for which we cannot reject this null hypothesis are excluded from model selection and are unsupported by the standard MHCflurry predictor. From the 130 alleles for which training was attempted, this filter reduced the supported alleles to 112.

Ensembles are selected using a forward stepwise selection procedure (Caruana et al. 2004). Starting with an empty ensemble, models are added to maximize the score of the ensemble at each step, stopping when no model's addition improves the score. We require that at least 8 and no more than 16 models are selected per allele. To help reduce the noise associated with model selection on alleles with limited data, we additionally include in the combined score the consistency with the predictions of the full ensemble of 320 models (in terms of Kendall τ) on a large number of random peptides, weighting it be equivalent to 10 affinity measurements or MS hits.

Construction of training and validation datasets

Affinity measurements used for training and model selection. The affinity measurement dataset used for training and model selection was assembled from a snapshot of the Immune Epitope Database (IEDB) MHC ligands downloaded on Dec. 1, 2017 augmented with the BD2013 dataset (Kim et al. 2014). IEDB

entries with non-class I, non-specific, mutant, or unparseable allele names were dropped, as were those with peptides identified by MS or containing post-translational modifications or noncanonical amino acids. This yielded an IEDB dataset of 143,898 quantitative and 43,978 qualitative affinity measurements. Of 179,692 measurements in the BD2013 dataset (Kim et al. 2014), 57,506 were not also present in the IEDB dataset. After selecting peptides of length 8-15 and dropping alleles with fewer than 25 measurements, the combined dataset consists of 230,735 measurements across 130 alleles.

MS model selection dataset. The MS model selection dataset consists of 226,684 ligands formed by combining 186,415 ligands from IEDB with 39,741 additional ligands from SysteMHC Atlas (Shao et al. 2018) and 530 additional ligands from ref. (Abelin et al. 2017). The unprocessed SysteMHC and Abelin et al. datasets are much larger but most entries are already present in IEDB, are duplicates, or report on alleles for which training was not attempted. The ligands from SysteMHC Atlas were first filtered to remove entries with low confidence ($prob < 0.99$). Of the 112 alleles supported by the predictor, 57 had at least 100 MS ligands available for model selection (Supplemental Table 2-1).

MS benchmark dataset. The MS benchmark was derived from 23,651 sequences of MHC-displayed ligands eluted from a B cell line expressing a single MHC class I allele (Abelin et al. 2017). We excluded one allele (HLA-A*02:04) not supported by MHCflurry or NetMHC due to insufficient representation in the training dataset (fewer than 25 measurements) and discarded peptides with post-translational modifications or lengths outside the supported range (8-15 residues). We sampled unobserved sequences (decoys) from the protein-coding transcripts that contained the identified peptides (hits) based on protein sequences in the UCSC hg19 proteome and transcript quantification from RNA sequencing of the relevant cell line (B721.221) downloaded from the Gene Expression Omnibus (accession GSE93315). For an allele with n hits, we sampled $100n$ decoys, weighting transcripts by the number of hits and sampling an equal number of decoys of each length 8-15. After removing any entries also present in the training dataset, this yielded a benchmark of 23,651 hits and 2,377,042 decoys.

HPV affinity measurement benchmark dataset. The HPV benchmark dataset consists of affinity measurements for 475 peptides of length 8-11 across seven MHC alleles (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*11:01, HLA-A*24:02, HLA-B*07:02, HLA-B*15:01).

The binding affinity of HPV16 E6 and E7 derived peptides to selected MHC class I molecules was tested in competition-based binding assays as described in (Kessler et al. 2003, 2004). Briefly, test peptides in 1:2 serial dilutions (final concentrations from 100 – 0.78 μM) compete with 150mM fluorescein-labeled reference peptide with a known high affinity for binding to the MHC class I molecule of interest on B-LCL cells (cell lines used: 1341-8346, BSM, E481324, EA, FH8, LKT3, WT100BIS), which were previously stripped from natural bound peptides and β2-microglobulin using ice cold citric acid buffer. After stripping, the cells were washed with culture medium and dissolved in culture medium containing 2μg/mL β2-microglobulin (MP Biomedicals) to reconstitute the MHC class I complex. B-LCL cells were diluted to $6x10^4$ cells/100μl per test peptide concentration and pipetted to a well-plate containing the mixes of test and reference peptide. After 24h incubation at 4°C the cells were washed, fixed in 1% PFA and suspended in 0.5% BSA in 1x PBS. The mean fluorescence intensity *Fmix* at each test peptide concentration was measured by flow cytometry (Accuri C6, BD Biosciences). The binding of each test peptide was calculated as the percent inhibition of reference peptide binding relative to the minimal response (without reference; *Fmin*) and the maximal response (reference only; *Fmax*) as:

$$Inhibition\,(\%) \;=\; (1 \;-\; \frac{Fmix \;-\; Fmin}{Fmax \;-\; Fmin}) \;*\; 100$$

The binding affinity of the test peptide was determined by non-linear regression analysis as the concentration that inhibits 50% binding of the fluorescein-labeled reference peptide (IC50).

Peptides with an experimental IC50 below 5 μM (5,000 nM) were defined as strong binders, 5-15 μM (5,000-15,000 nM) as intermediate binders, 15-100 μM (15,000-100,000 nM) weak binders, and peptides above 100μM (100,000 nM) were defined as non-binders, as outlined in (Kessler et al. 2003, 2004). These rather high nM-values are explained by the fact that the experimental assay uses very high affinity reference peptides, thus high concentrations of test peptides are needed to reach the IC50. For confirmation

and statistical significance, the assay was performed at least three times for binders and twice for non-binders.

Speed benchmarking. Experiments were performed on a machine with twelve Intel Core(TM) i7-5930K CPUs at 3.50GHz, four NVIDIA GeForce GTX TITAN X GPUs, and 64GB memory using the MHCtools Python interface to the MHCflurry and NetMHC tools (https://github.com/openvax/mhctools) with parallelization and GPUs disabled. We measured the time to generate various numbers of predictions (10^2, 10^3, 10^4, 10^5, and 10^6) for a single allele using peptides sampled from the MS benchmark. We repeated the experiment three times using different alleles (HLA-A*02:01, HLA-A*02:07, HLA-A*01:01). Rates and speedups reported in the main text are averages for the three alleles at the maximum number of peptides (10^6).

Training the *MHCflurry 1.2.0* full ensembles (320 models for each of 130 alleles, for 41,600 models total) took 1,049 minutes using all GPUs and CPUs on the machine. Model selection took 299 minutes, and computing the histogram of predicted affinities for each allele took 15 minutes.

Quantification and statistical analysis

Accuracy on the MS benchmark was assessed in terms of the positive predictive value (PPV) for the predicted affinity to differentiate MS hits from decoys. This was calculated separately for each allele. To compute PPV for an allele with n hits, we ranked the $n + 100n$ hits and decoys from tightest to weakest predicted binding affinity and calculated the fraction of the top n peptides that were hits. Scores relative to NetMHC or NetMHCpan were calculated by dividing the difference in PPV between MHCflurry and the other predictor by the other predictor's PPV and representing the result as a percent. We applied a two-sided binomial test (α=0.05) to determine if one predictor outperformed another on more alleles than expected by chance.

We assessed accuracy on the HPV affinity measurement dataset using three metrics: Kendall rank correlation coefficient (Kendall τ), Pearson correlation taken over the log of the predicted and measured

affinities, and the area under the receiver operator characteristic curve (AUC) at differentiating binders from non-binders. As the HPV dataset is small, these metrics were computed across all predictions, not separately for each allele. Kendall τ measures the correlation in rank when peptides are sorted by measured or predicted affinity. Kendall τ and Pearson correlation were calculated using the scipy package (https://www.scipy.org). The AUC estimates the probability that a binding peptide will have a stronger predicted affinity than a non-binding peptide. It was calculated by the scikit-learn package (http://scikit-learn.org). For the purpose of AUC, we defined a peptide to be a binder if it had any detectable binding in our assay (Figure 2-2C). AUCs calculated using more restrictive IC50 thresholds are indicated in Supplemental Table 2-2. Each predictor was compared to *NetMHCpan 4.0* on each metric by computing the difference of the two predictors' scores within bootstrap resamples of the dataset, with a result considered significant if the 95% confidence interval for the difference excludes 0.

Data and software availability

MHCflurry is available under the Apache License 2.0. It may be installed from the Python package index. Source code is maintained at https://github.com/openvax/mhcflurry. All data and scripts used to train the models are available in this repository. The training and MS datasets, including all predictions, may be downloaded at https://github.com/openvax/mhcflurry/releases/tag/1.2.0.

Chapter 3. MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing

This chapter is adapted from the following article:

> O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved pan-allele prediction of MHC Class I-presented peptides by incorporating antigen processing. *Cell Systems.* 2020.

Summary

Computational prediction of the peptides presented on MHC class I proteins is an important tool for studying T cell immunity. The data available to develop such predictors has expanded with the use of mass spec 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 spec-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 spec experiments. Our predictors are implemented in the open source MHCflurry package, version 2.0 (https://github.com/openvax/mhcflurry).

Introduction

Cytotoxic (CD8+) T cells recognize peptides presented in complex with major histocompatibility class I (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, Reits, and Neefjes 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(Bjoern Peters, Nielsen, and Sette 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 steps have been proposed(Björn Peters, Bulik, et al. 2003; Keşmir et al. 2002; Nielsen et al. 2005; Bhasin, Lata, and Raghava 2007), and integrated with binding affinity predictions to give a composite score(Larsen et al. 2005; Tenzer et al. 2005; Stranzl et al. 2010), 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 spec (MS)-identified MHC class I ligands provides an opportunity to revisit antigen processing prediction using larger and potentially more biologically-relevant datasets. While the antigen processing information in these datasets likely already informs existing MHC class I binding predictors trained on MS datasets(V. Jurtz et al. 2017; Gfeller et al. 2018), antigen processing 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 (MHC class I binding affinity; "BA") and allele-independent effects (antigen processing; "AP"). We first trained a new

pan-allele MHC class I binding affinity predictor (referred to as MHCflurry 2.0 BA) on available MHC class I ligand data, including affinity measurements and MS datasets. The use of *in vitro* affinity measurements in the training data, which are largely independent of antigen processing, is one of several design choices intended to limit the BA predictor's tendency to learn antigen processing signals. We use the BA predictor to generate a training set for a model of antigen processing 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 antigen processing 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 antigen processing steps and showed quantitative agreement with an independent dataset of proteasome-cleaved peptides(Wolf-Levy et al. 2018). We combined the BA and AP models in a logistic regression model, which we refer to as the presentation score (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 binding affinity (BA) predictor (Figure 3-1A) 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 (Supplemental Figure 3-1, Supplemental Table 3-1). To compare against existing binding affinity (NetMHCpan 4.0 BA) and MS ligand (NetMHCpan 4.0 EL and MixMHCpred 2.0.2) predictors, we compiled a benchmark using published datasets of MS-identified MHC ligands (Supplemental Table 3-2, Supplemental Table 3-3). 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 3-1B). 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 3-1C).

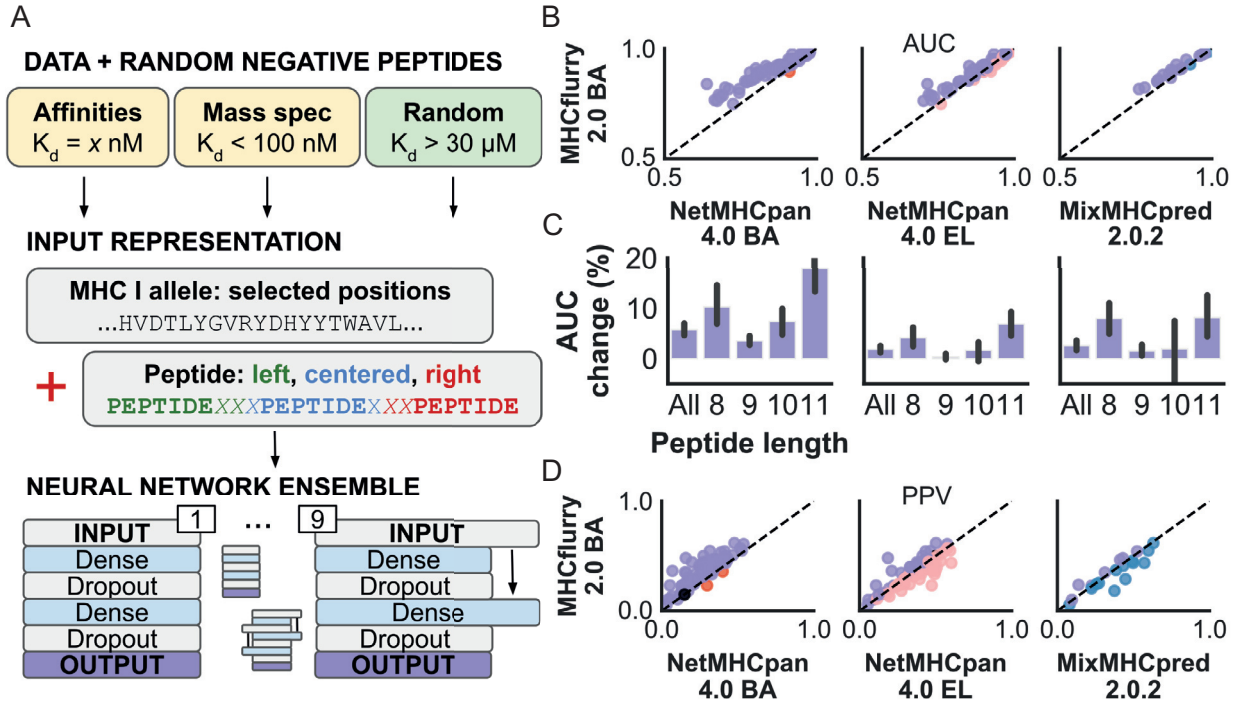


Figure 3-1. MHCflurry 2.0 binding affinity (BA) predictor architecture and benchmark

(a) BA predictor training data, model input representations, and neural network architectures. **(b)** Area under the curve (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)** Positive predictive value (PPV) of MHCflurry 2.0 BA in comparison to other predictors. Points correspond to the same experiments as in (b).

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(Sarkizova et al. 2019; Abelin et al. 2019), which we refer to as the MONOALLELIC benchmark. As this dataset is normally incorporated into the training set for the MHCflurry BA predictor, we tested a variant of MHCflurry BA trained without this data (Supplemental Table 3-4). We again found that MHCflurry BA outperformed the other tools in terms of AUC

(Supplemental Figure 3-2A). Notably, good accuracy (AUC > 0.94) and a modest improvement over existing tools was observed for the 11 alleles that had no representation in the training data (Supplemental Figure 3-2B, Supplemental Table 3-5). 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 (Supplemental Figure 3-2C). 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 positive predictive value (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 3-1D). Similar results were observed for PPV on the MONOALLELIC benchmark (Supplemental Figure 3-2D).

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 antigen processing (AP) predictor. Its neural network architecture is motivated by the possibility of learning peptide N- and C-terminal cleavage or processing signals (Figure 3-2A). 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).

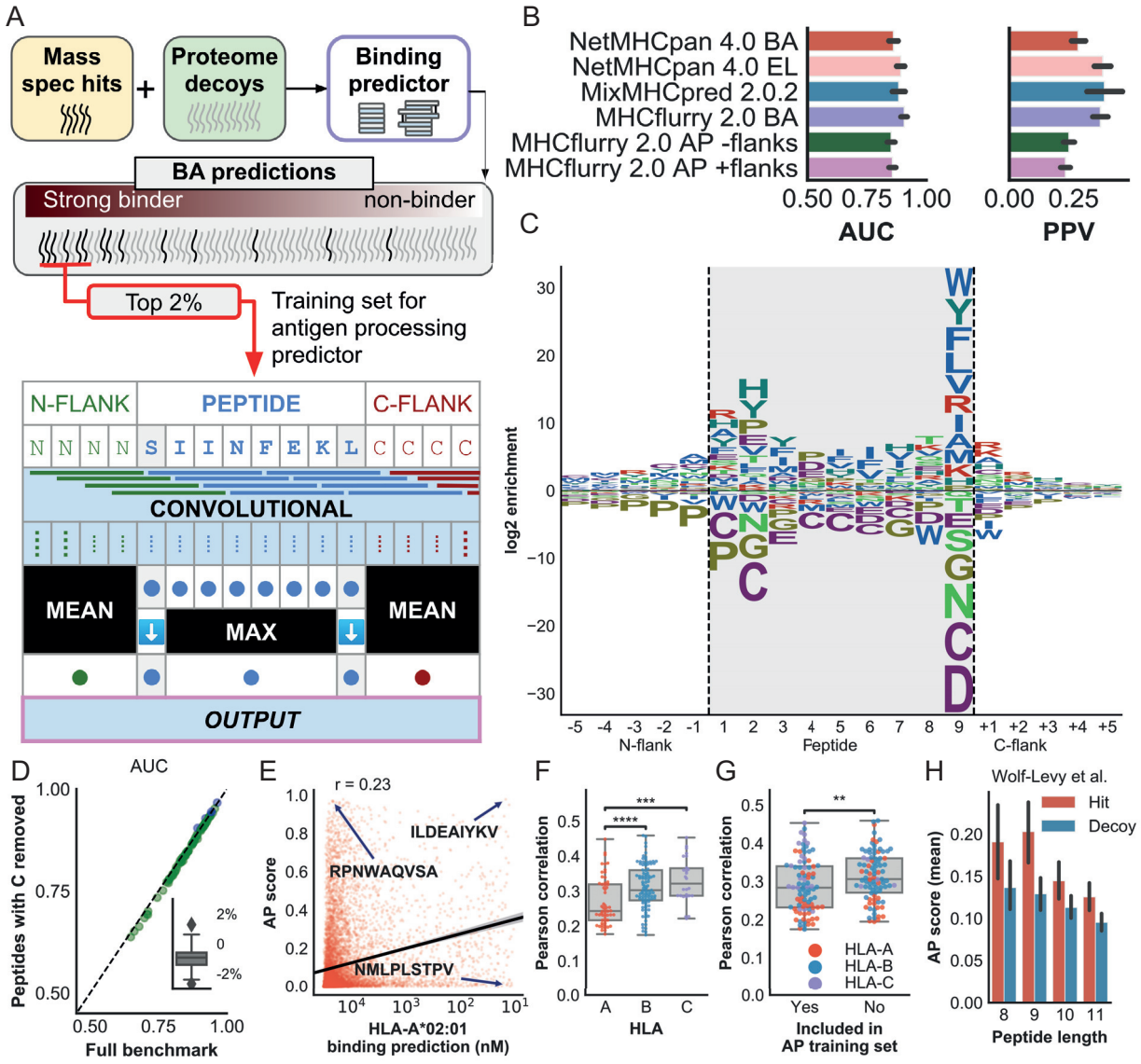


Figure 3-2. The MHCflurry 2.0 antigen processing predictor models MHC class I allele-independent effects

(a) AP predictor training scheme and neural network architecture. **(b)** Mean AUC and PPV accuracy on the multiallelic mass spec 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) vs. 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, 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 (hits) and unobserved peptides from the same genes (decoys).

To understand if the MHCflurry AP predictors learned a meaningful signal, we evaluated their accuracy on the MULTIALLELIC benchmark (Figure 3-2B). 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 to 0.91 (0.90 – 0.92) for the MHCflurry BA predictor (Figure 3-2B). 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 3-2C and Supplemental Table 3-6). 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 3-2D). 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 3-2E). Correlations with the AP predictor were somewhat higher for *HLA-B* and *HLA-C* than for *HLA-A* alleles (Figure 3-2F). The alleles used to train the AP predictor showed no greater AP vs. BA correlations on average than those that were not included in the AP training set; in fact the correlation was somewhat lower for the alleles used in training (Figure 3-2G). 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 3,017 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 3-2H and Supplemental Table 3-7). This indicated that the AP predictor learned a signal consistent with a key antigen processing 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 3-3A). We refer to this logistic regression model as the MHCflurry presentation score (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 (Figure 3-3B and C). 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).

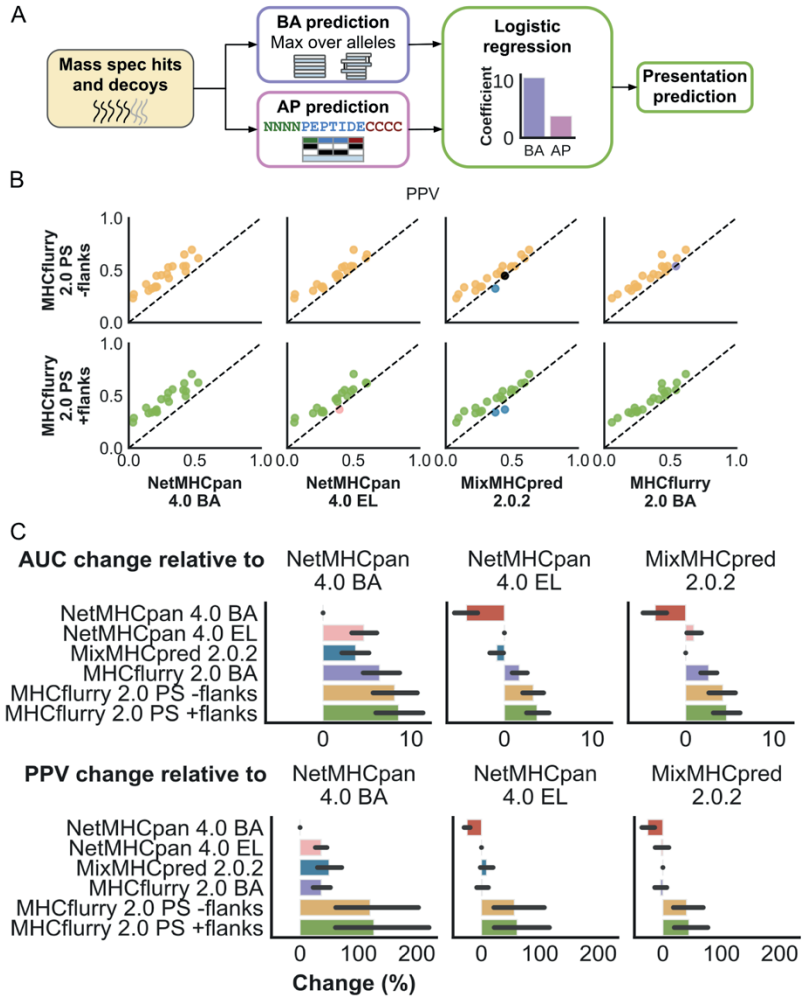


Figure 3-3. The MHCflurry 2.0 presentation score combines binding affinity and antigen processing prediction

(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 mass spec 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.

Discussion

Our MHC class I ligand prediction method consists of two neural network models: the MHCflurry BA predictor and the MHC allele-independent MHCflurry AP predictor. The AP predictor is trained to learn what the BA predictor missed, i.e. residual sequence properties that distinguish hits from decoys among peptides predicted to bind MHC tightly. Both 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 presentation score (PS), outperformed the individual components and standard tools. While 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 has 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. 2019), 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 3-2C) shows similarities with the established preferences of key antigen processing steps. Deconvolution of effects, however, is complicated by the overlapping specificity of these steps, potentially a consequence of coevolution of the antigen processing 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(Nussbaum et al. 1998; Harris et al. 2001). Some agreement with proteasomal processing is also apparent at the interior residues of the peptide, such as a depletion of proline at P8 and 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 enrichments for Arg, Ala, Lys, Ser, and Gly are consistent with proteasomal cleavage preferences for the position after the cut site(Wolf-Levy et al. 2018; Nussbaum et al. 1998), 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 3-2H) suggest that the AP predictor has learned certain antigen processing 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. While 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.

Data and software availability

The predictors described here are available under an Apache 2 open source license in the MHCflurry package (https://github.com/openvax/mhcflurry). The datasets used to train and benchmark the predictors are deposited in Mendeley Data at https://data.mendeley.com/datasets/zx3kjzc3yx.

Acknowledgements

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.

Author contributions

Timothy O'Donnell designed research, performed research, and wrote the manuscript. Alex Rubinsteyn (University of North Carolina at Chapel Hill, Chapel Hill, NC) and Uri Laserson (Mount Sinai) advised on data interpretation. All authors critically reviewed the manuscript.

Declaration of interests

The authors declare no competing interests.

Methods

Table 3-1. Key Resources for Chapter 3

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) predictors (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 Supplemental Figure 3-2 (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_cell line.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

Table 3-1. Key Resources for Chapter 3 *Continued from previous page*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNA-seq: Expression Atlas (CCLE)	(Papatheodorou et al. 2020; Barretina et al. 2012)	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
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

Table 3-1. Key Resources for Chapter 3 *Continued from previous page*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
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	(V. 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)	Supporting Information
MS-identified MHC ligands	(Shraibman et al. 2016)	Supplementary Table 1
MS-identified MHC ligands	(Gloger et al. 2016)	Supplementary Tables 1-5

Table 3-1. Key Resources for Chapter 3 *Continued from previous page*

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MS-identified MHC ligands	(Bassani-Sternberg et al. 2016)	Supplementary Data 2
MS-identified MHC ligands	(Bassani-Sternberg et al. 2017)	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)	Supplementary Tables 1 and 2
MS-identified MHC ligands	(Abelin et al. 2019)	Data S1
MS-identified MHC ligands	(Sarkizova et al. 2019)	Supplementary Data 1 and 2
Proteosome-cleaved peptides identified by "mass spectrometry analysis of proteolytic peptides" (MAPP)	(Wolf-Levy et al. 2018)	Supplementary Data 2

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. These are as follows:

Data S1. MULTIALLELIC benchmark dataset with predictions (CSV, gzipped).

Data S2. MONOALLELIC benchmark dataset with predictions (CSV, gzipped).

Data S3. Training data for MHCflurry 2.0 binding affinity (BA) predictors (CSV).

Data S4. Training data for the variant of MHCflurry BA evaluated on the MONOALLELIC benchmark in Supplemental Figure 3-2 (CSV).

Data S5. Training data for MHCflurry 2.0 antigen processing (AP) predictors (CSV).

Data S6. Training data for MHCflurry 2.0 presentation score (PS) predictors (CSV).

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 allele. We refer to these as the MONOALLELIC samples, which comprise 92 samples from the recent publication by Sarkizova et al(Sarkizova et al. 2019) 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. 2017; Shraibman et al. 2016; Bassani-Sternberg et al. 2015; Ritz et al. 2016; Gloger et al. 2016; Bassani-Sternberg et al. 2016; Hassan et al. 2013; Mommen et al. 2014), and MULTIALLELIC-RECENT, comprised of 20 experiments from two studies published in 2019(Shraibman et al. 2019; Sarkizova et al. 2019). Curated samples are listed in Supplemental Table 3-2, the sample groups used for each benchmark experiment are given in Supplemental Table 3-3, and Data S1 and S2 give the full datasets for the MULTIALLELIC and MONOALLELIC benchmarks, respectively.

Transcript expression. Each curated MS 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. Supplemental Table 3-2 gives 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 (V. 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 eluted ligand (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 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.

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 parseable by the mhcnames package (https://github.com/openvax/mhcnames), the sequences were aligned using Clustal Omega 1.2.1 (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(V. 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 parseable 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 (Sarkizova et al. 2019) and 4,808 hits from (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 Supplemental Table 3-4.

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 (Supplemental Figure 3-1, Supplemental Table 3-1), 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, and 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 (Supplemental Figure 3-2, Supplemental Table 3-5), we compared the AUC for these samples to that obtained using predictions for closely-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. Allele similarity was defined as the edit distance of 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, breaking ties by taking the similar allele 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), a length-matched unobserved decoy peptide was randomly sampled from the same gene.

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.

Statistical significance. To compare predictors (Figure 3-1C, Figure 3-3C, and Supplemental Figure 3-2) or benchmark variants (Figure 3-2D), 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 3-2B) or average prediction (Figure 3-2H), 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 (Figure 3-2F and G), we used the Mann-Whitney U test.

Chapter 4. Application to epitope prediction and comparison of antigen processing predictors

Summary

The previous chapter introduced models that improve over existing tools for the prediction of peptides presented on MHC class I. This chapter addresses whether this improvement in fact enables more accurate prediction of T cell epitopes. Applying a published unbiased screen of tumor mutations, I find that, indeed, the MHCflurry 2.0 presentation score (PS) predictor outperforms existing tools at neoantigen prediction. However, in a benchmark based on viral epitopes identified across a range of studies, I detected no advantage from the incorporation of antigen processing prediction and instead found that MHCflurry binding affinity (BA) performed best. This may be due to biases present in this benchmark, as binding predictors were typically used to select peptides for immunogenicity testing. In a second set of analyses, I show that the MHCflurry antigen processing predictor differs in important ways from an earlier generation of predictors for proteasomal cleavage and TAP transport, largely due to sensitivity to a wider set of positions in the peptide. These results support the continued evaluation and application of the new predictors and highlight the considerable remaining room for improvement in CD8+ T cell epitope prediction.

Introduction

The MHC class I presentation predictor (MHCflury PS) developed in Chapter 3 integrates two components, a pan-allele MHC class I binding affinity predictor (MHCflurry BA) and an antigen processing predictor (MHCflury AP). I showed that MHCflurry PS outperforms current tools at predicting MHC class I-presented peptides. The question remains, however, whether this improved performance actually translates to better prediction of T cell epitopes. This is more difficult to answer, as assays to identify the targets of T cells responses operate at much lower throughput than those for identifying MHC ligands. While high-

throughput T cell epitope mapping may soon be practical (Amalie Kai Bentzen et al. 2016; Li et al. 2019; Joglekar et al. 2019), the epitope datasets available at present are generally small or subject to substantial observational biases. These issues notwithstanding, in this chapter I take a first look at the performance of MHCflurry and other tools on two epitope prediction benchmarks.

The first benchmark is from a series of screens for tumor neoantigens by Steve Rosenberg's group at the National Cancer Institute. In these studies, patient T cells were tested for response to autologous antigen presenting cells engineered to display 25-mer sequences, each centered on a somatic mutation detected by exome sequencing of the patient's tumor. A tandem minigene system was used for a first pass followed by deconvolution to individual 25-mer sequences. Responding T cells were detected by ELISPOT or upregulation of activation induced markers. The origin of the T cells for this screen varied, and included tumor infiltrating lymphocytes in gastrointestinal cancers (Tran et al. 2015) as well the peripheral blood of melanoma patients (Gros et al. 2016). Critically, the mutations to screen were selected without use of MHC binding prediction, making this dataset a rare unbiased source of validation data. This benchmark is small, however. Only 52 CD8+ T cell responses were detected across 2,844 screened sequences.

The second benchmark uses viral epitopes. As I am aware of no study that has published a suitable unbiased screen for viral immunogenicity in humans, I collected 1,380 CD8+ T cell epitopes across 21 viral proteins deposited in the Immune Epitope Database (IEDB). Most epitopes were identified by tetramer sorting, chromium release, ELISPOT, or intracellular cytokine staining after viral exposure or vaccination. I also collected 527 peptides for which immunogenicity was assayed but the results were negative in all individuals tested, as well as the full sequences for the epitope source proteins.

Earlier work on antigen processing prediction focused on predictors for specific antigen processing steps fit to small datasets, which generally failed to significantly improve accuracy at predicting epitopes (Bjoern Peters, Nielsen, and Sette 2020). To better understand what the MHCflurry AP predictor learned and how it may differ from these tools, in the second part of this chapter I compare the sequence preferences of MHCflurry 2.0 AP with those of existing tools for proteasomal cleavage and TAP transport.

Results

Benchmark on neoantigens recognized by tumor infiltrating lymphocytes

I first considered a collection of 2,844 mutation-containing 25-mer sequences that were tested for recognition by tumor-infiltrating or peripheral blood lymphocytes in cancer patients. As mentioned, peptides in this set were selected for assay without use of binding predictors, making it an unbiased source of validation data. T cell responses were detected for only 56 (2%) of the peptides, 52 by CD8+ and four by CD4+ (not analyzed here).

In terms of area under the curve (AUC), all predictors performed significantly better than random (i.e. AUC of 0.5), but had lower accuracy than typically seen for binding prediction of a single minimal epitope to a single MHC class I allele, where AUCs greater than 0.90 are common (Figure 4-1A). The best-performing predictor was MHCflurry 2.0 PS, with an AUC of 0.85 (bootstrap 95% confidence interval 0.81 - 0.88). This was within statistical error from the MHCflurry 2.0 BA affinity output (AUC 0.82), but significantly better than the NetMHCpan 4.0 BA affinity and percentile outputs (0.78 and 0.80, respectively), NetMHCpan 4.0 EL score and percentile outputs (also 0.78 and 0.80), and MixMHCpred score and percentile outputs (both 0.78).

The MHCflurry AP predictor had an AUC of 0.71 (0.65 - 0.77), recapitulating the finding from the previous chapter that it has some value for predicting good MHC ligands (and in this case, T cell epitopes) without knowledge of MHC class I allele. To put this score in context, I also considered the accuracy of MHCflurry BA using a set of twelve high-prevalence MHC alleles instead of individual-specific MHC genotypes. This performed comparably to the AP predictor, with an AUC score of 0.71 (0.65 - 0.77). A trend toward a small additional boost was observed for the MHCflurry 2.0 PS predictor on population alleles (AUC of 0.73, 95% CI 0.67 - 0.78).

For an evaluation that more closely reflects how these tools are used, I ranked all peptides by prediction and considered the positions in the ranking that capture various fractions of the epitopes. This is interpretable as the number of immunogenicity assays needed to identify the specified epitopes, if the

experiments are guided by prediction. Similar results were observed as in the AUC evaluation (Figure 4-1B). The MHCflurry 2.0 PS predictor performed best at most thresholds tested, although the differences did not always reach significance. For example, the MHCflurry PS predictor would require 1,043 tests (95% CI 820 - 1,314) to identify 90% of the 52 epitopes, compared to 1,286 for MHCflurry BA, 1,355 for NetMHCpan4 BA, 1,372 for NetMHCpan 4.0 EL, and 1,710 for MixMHCpred 2.0.2, using the percentile outputs for each tool.

Benchmark on viral epitopes

I next tested predictor accuracy at recovering viral CD8+ T cell epitopes identified in a range of studies. I first compared predictions between the 1,380 epitopes with confirmed immunogenicity and 527 peptides from the same proteins that were assayed but had no response. Interestingly, the only predictor with a detectable difference in mean score between these groups was the MHCflurry AP predictor, although the difference was small (Figure 4-2A). The absence of signal for the other predictors reflects the fact that predictions are used to select which peptides to assay. Since only predicted MHC class I binders are typically tested, binding affinity is effectively controlled-for in this dataset. This result highlights the bias present in this dataset.

To address the question facing a researcher looking to identify epitopes in an uncharacterized virus, I collected the full sequences for the source proteins of the curated epitopes and scanned the predictors over all 8-11-mer peptides in the virus proteins. For these analyses, the peptides that were not assayed were considered non-epitopes, and I tested each predictor's ability to pick out the epitopes from this large universe of peptides.

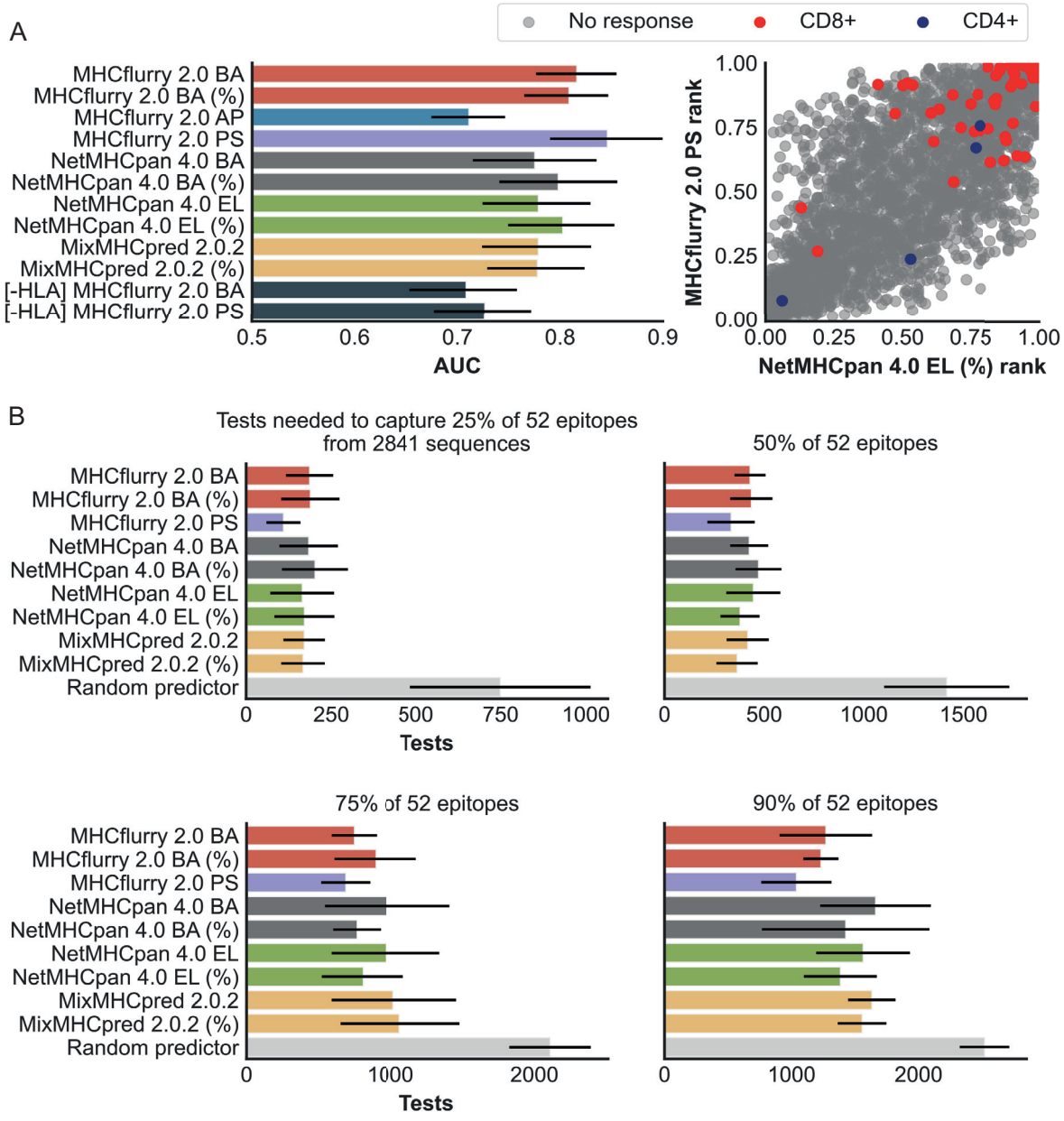


Figure 4-1. Benchmark on neoantigens targeted by tumor-infiltrating T cells

(a) Left: accuracy measured as the area under the receiver operating characteristic curve (AUC) at distinguishing 25-mer peptides detected to be the target of an intratumoral CD8+ T cell response vs. those with no detected response. Higher values indicate better accuracy. For each tool, the strongest prediction was taken across all peptide subsequences of length 8-11 across the individual's six MHC class I alleles. The predictors indicated with a (%) were evaluated using their percentile rank output rather than raw binding affinity or score. The variants marked "[-HLA]" show results taken using twelve prevalent alleles rather than individual-specific MHC class I alleles. Right: predictions for MHCflurry PS and the best-performing NetMHCpan predictor on this dataset, NetMHCpan EL (%). For each predictor, all tested 25-mer sequences were ranked by prediction on a 0.0 - 1.0 scale, with 1.0 corresponding to the peptide with the tightest affinity or highest presentation score. **(b)** The number of hypothetical experiments that would need to be performed to identify the indicated fractions of the CD8+ neoantigens. Lower values indicate better accuracy. Error bars indicate bootstrap 95% confidence intervals.

I first considered predictor sensitivity and precision at identifying epitopes using two common thresholds: 500 nM binding affinity and 2.0 percentile rank. MHCflurry 2.0 BA showed slightly higher sensitivity than NetMHCpan 4.0 BA, recovering 70% of epitopes at 500 nM and 78% of epitopes at 2% rank, compared to 65% and 71% for NetMHCpan 4.0 BA (Figure 4-2B and C). This came at the cost of lower precision, however. Only 3.4% of peptides predicted by MHCflurry BA to have affinity tighter than 500 nM were identified T cell epitopes, compared to 4.6% for NetMHCpan 4.0 BA (Figure 4-2D). MixMHCpred, at the opposite extreme of MHCflurry BA, showed the lowest sensitivity and highest precision. The MHCflurry PS predictor at the 2% threshold showed an intermediate profile that was similar to the NetMHCpan 4.0 tools.

All predictors showed good ability to rank the epitopes highly among all peptides from the viral proteome (Figure 4-2E). For example, an experimentalist guided by MHCflurry 2.0 BA would need to perform 6,833 assays to identify half of the 1,380 epitopes. This corresponds to 0.46% of the 1.5 million possibilities under consideration. However, this increases to nearly 40,000 experiments (2.6% of total) to detect 75% of epitopes.

In the ranking analysis, MHCflurry BA modestly outperformed the other predictors at identifying viral epitopes. In particular, MHCflurry BA outperformed MHCflurry PS, which was unexpected given previous results. To understand if a different way of combining BA and AP predictions might give a better composite score, I re-fit the MHCflurry PS logistic regression model to the viral epitope data and evaluated its accuracy using cross validation. The resulting model, referred to as MHCflurry PS2, was less sensitive to AP score than the original PS model (Supplemental Figure 4-1). Although MHCflurry PS2 performed better than the original MHCflurry PS model, it still showed no advantage over MHCflurry BA (Figure 4-2E).

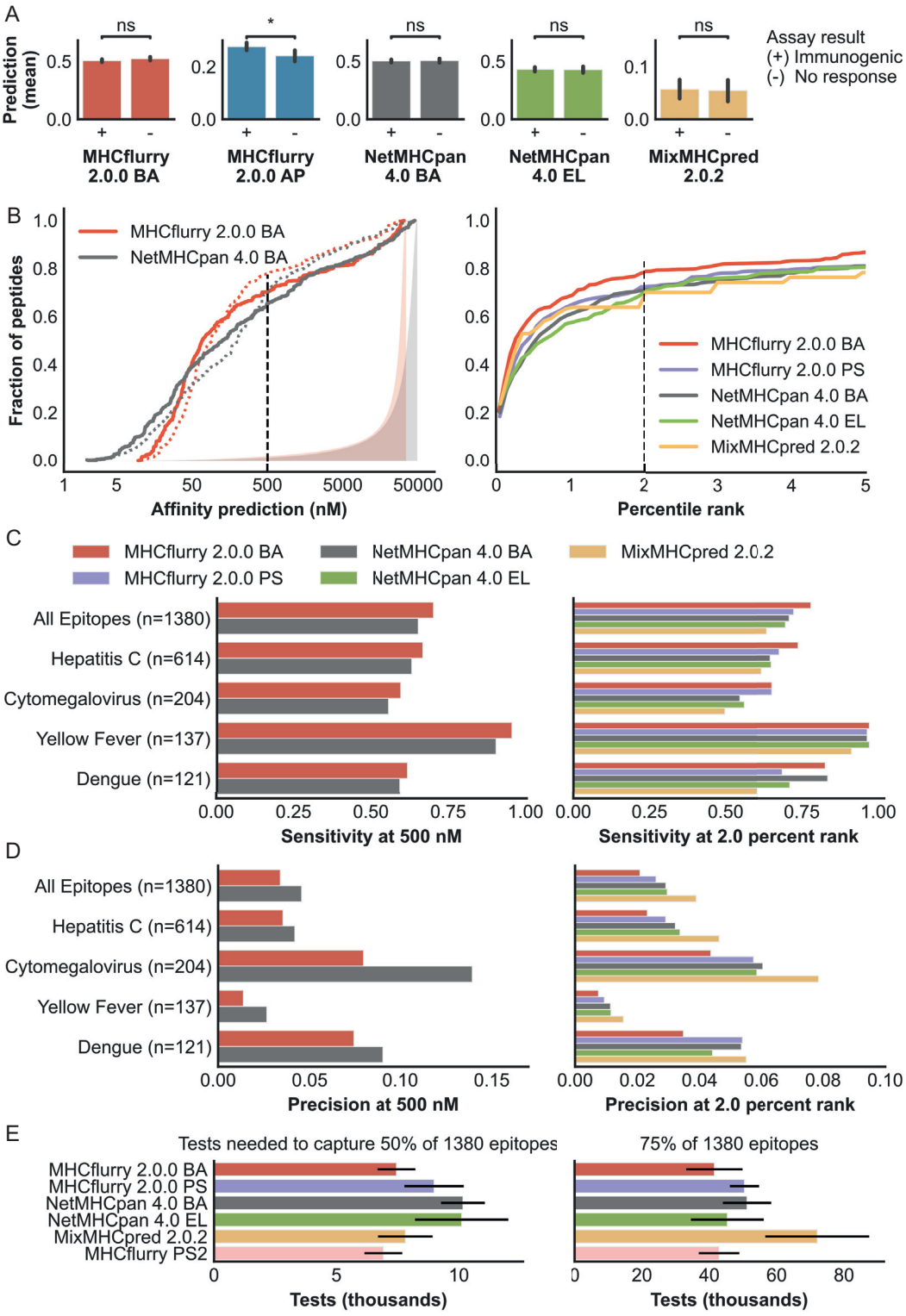


Figure 4-2. Evaluation on viral epitopes deposited in the IEDB

Legend on next page.

Figure 4-2, continued

(a) Mean prediction scores for epitopes (+) in comparison to assayed peptides that showed no response (-). **(b)** Cumulative distribution of predicted nanomolar binding affinities (left) and percentile ranks (right). Solid lines show viral epitopes (n=1,380); dotted lines indicate peptides that were tested experimentally but showed no response (n=527). Shaded regions show peptides from the same proteins that were not experimentally tested (n=1,497,331). The common thresholds 500 nM and 2% rank are highlighted. **(c)** Sensitivity (fraction of epitopes recovered) at 500nM (left) and 2 percentile rank (right) by virus. **(d)** Precision (fraction of recovered peptides that are epitopes) at the same thresholds. Only the four viruses with the most experimentally-determined epitopes are shown separately in (c) and (d). **(e)** The number of hypothetical experiments that would need to be performed to identify 50% (left) or 75% (right) of the curated T cell epitopes if guided by each predictor. This was calculated by ranking the peptides by prediction and finding the position of the epitope at the specified quantile in the rank order. The MHCflurry PS2 predictor is a PS model that was re-fit to viral epitopes. Error bars indicate bootstrap 95% confidence intervals.

Correlative analysis of antigen processing predictors

The results for the neoantigen analysis suggested that inclusion of the AP predictor may improve T cell epitope prediction in some circumstances. This generally has not been the experience for an earlier generation of processing predictors (Bjoern Peters, Nielsen, and Sette 2020). To better understand how the AP predictor relates to these earlier efforts, I performed a correlative analysis of predictions from AP and other tools.

The AP predictor showed a positive correlation with each of the other predictors (Figure 4-3A and B). However, the strength of the relationships were modest, with many peptides scored high by one predictor but not the other. The strongest relationship, which was with NetChop C-term, had a Pearson correlation of 0.45 (Figure 4-3C left). This was near the upper range to that observed between the AP predictor and the MHC class I binding affinity (BA) predictor across many alleles (Figure 3-2F in Chapter 3). It was also less than the correlation of 0.65 observed between NetChop 3.0 and the TAP predictor (Figure 4-3C right). Thus, two predictors intended to model different biological processes were in fact more similar to each other than either was to the AP predictor.

To put these differences in additional context, I assessed the correlation between the NetCTLpan composite score and MHCflurry PS. For MHC binding predictions, I used the A*01:01 allele. The MHC binding predictors showed good agreement (Pearson correlation 0.82; Figure 4-3D left). This is expected, as MHC binding specificities are relatively strict and A*01:01 is a well-characterized allele. Despite the

agreement on MHC binding, the NetCTLpan composite score and MHCflurry PS showed only a modest correlation (0.57; Figure 4-3D right), indicating that the disagreement between the predictors on the processing steps were substantial enough to meaningfully impact the overall predictions from the composite models.

I next probed what may be causing the differences in antigen processing predictions by considering the positional amino acid preferences for each tool. I first considered the overall sensitivity of the predictors to each position in 9-mer peptides and adjacent residues in the source protein. For all tools, the greatest sensitivity was observed to the amino acid at the peptide C-terminus, which I refer to as position P9 (Figure 4-4A). The NetCTLpan TAP predictor showed the most sensitivity at this position and very little sensitivity elsewhere. The AP predictor, in contrast, showed comparatively lower sensitivity at P9 and higher sensitivity than the others at positions near the N terminus, especially, P1 and P2. The two NetChop variants showed an intermediate profile, with the NetChop 20S notably exhibiting preferences at P6, P7, and P8, and the downstream flanking residue C1.

I lastly considered the specific amino acid preferences for each tool at P1, P2, and P9 (Figure 4-4B). While there was some agreement between the AP predictor and NetChop C-term at P2, for the most part the non-AP predictors showed a flat set of preferences at P1 and P2, as expected from the positional importance analysis. At P9, in contrast, the predictors showed strong and broadly similar preferences. All predictors tended to assign high scores to peptides terminating in the large hydrophobic amino acids tyrosine, tryptophan, and phenylalanine, which arise from the chymotryptic activity of the proteasome. Some evidence for tryptic-like preferences (cleavage after arginine and lysine) were also apparent, with the exception of NetChop 20S, which favored arginine but disfavored lysine at P9. The chymotryptic and tryptic proteasome specificities generate peptides with good TAP transport efficiencies, so it is not surprising that the NetCTLpan TAP predictor showed similar preferences for these amino acids as the other tools.

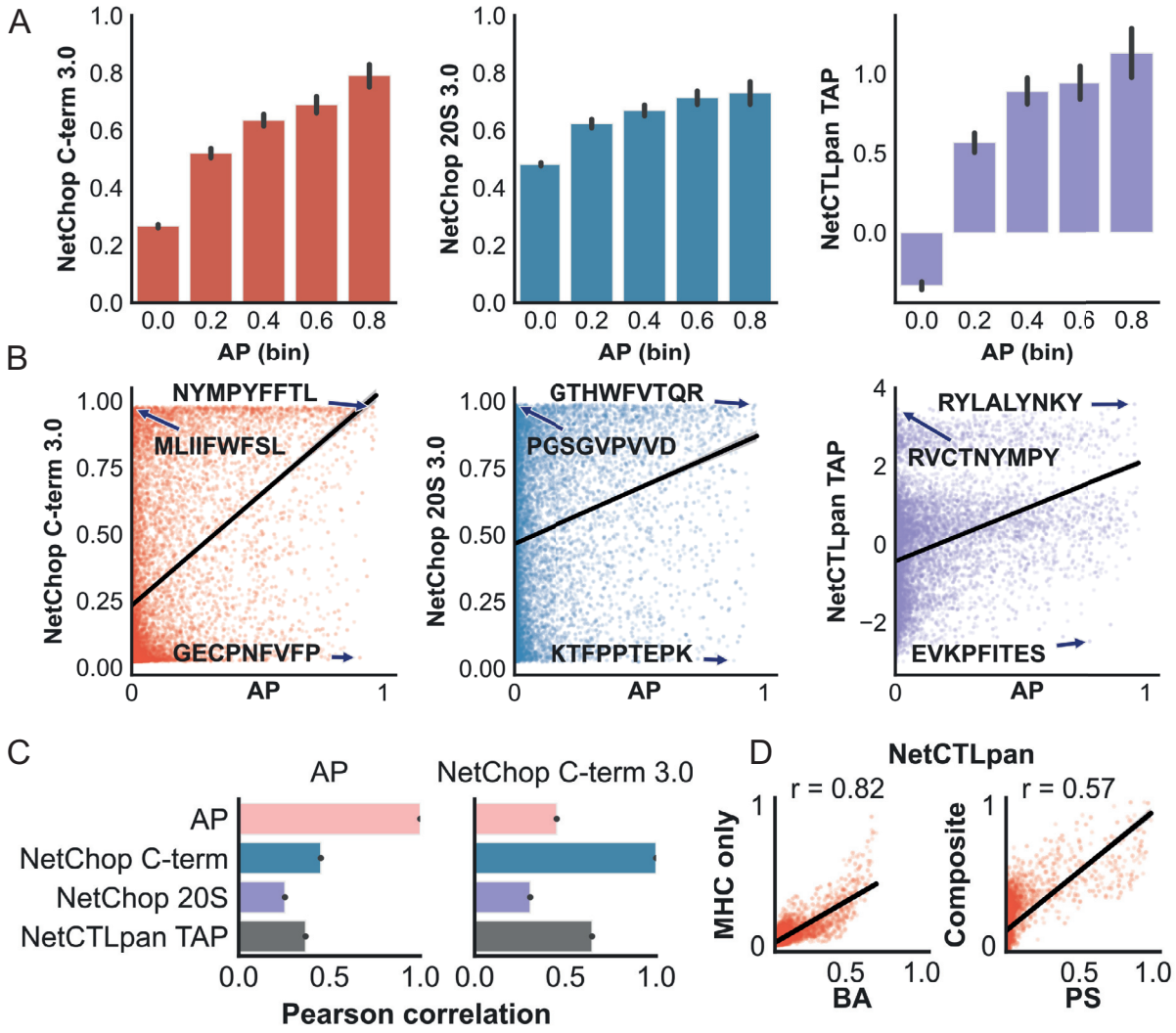


Figure 4-3. Correlation of AP predictor and predictors for proteasomal cleavage and TAP transport

(a) Peptides were grouped by AP score, and the mean score for the other predictors for the peptides in each bin was plotted. Bins are labeled by left edge (i.e. the first bar includes AP predictions between 0.0 and 0.2). **(b)** Regression fit for existing predictors against AP prediction. Example peptides in each corner are indicated. **(c)** Pearson correlation for each predictor with the AP predictor (left) or the NetChop C-term 3.0 predictor (right). **(d)** Regression fit for the MHC class I binding predictor included in NetCTLpan against the BA predictor (left) and for the NetCTLpan composite score against the PS predictor (right) using the A*01:01 allele. Pearson correlation (r) is indicated. Error bars indicate bootstrap 95% confidence intervals.

On the other hand, the preferences for the acidic residues aspartate and glutamate at P9 were consistent with the differences in training data between the tools. Peptides terminating in these amino acids arise from the caspase-like activity of the proteasome and are incompatible with TAP transport. Consistent with this understanding, the AP, NetChop 3.0 C-term, and NetCTLpan TAP predictors, all of which are trained on TAP-transported peptides, disfavored aspartate and glutamate at P9 (Figure 4-4B). In contrast,

NetChop 20S, the only predictor expected to be independent of TAP transport effects, showed a weak preference for aspartate and neither favored nor disfavored glutamate.

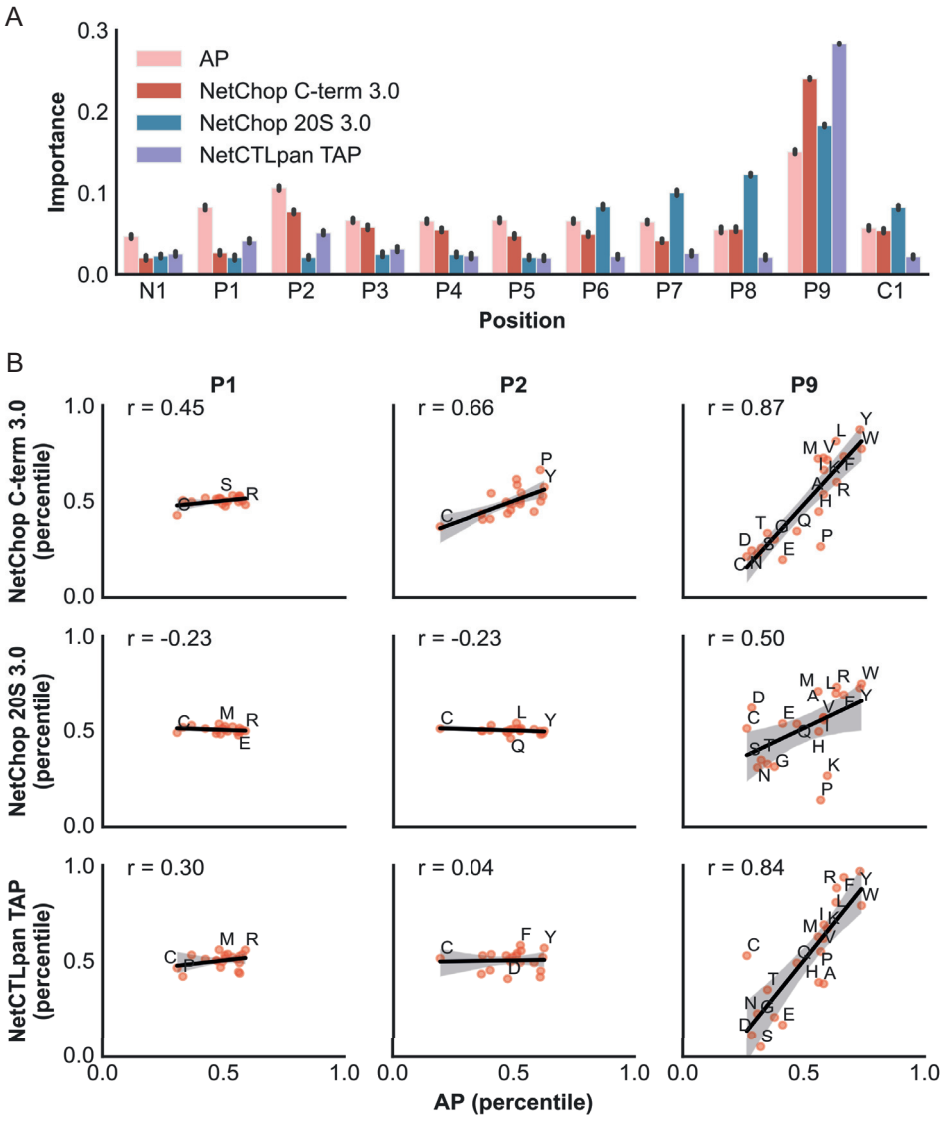


Figure 4-4. Positional preferences for AP predictor and predictors for proteasomal cleavage and TAP transport

(a) Feature importances for the upstream flanking residue (N1), each position in the peptide (P1 to P9) and the downstream flanking residue (C1) for each processing predictor. Feature importances are calculated as the standard deviation of the mean prediction percentile across amino acids. **(b)** Amino acid preferences at selected positions (P1, P2, P9) for NetChop C-term and the NetCTLpan TAP predictor in comparison to the AP predictor. Higher percentiles indicate greater favorability. Selected amino acids are labeled at P1 and P2; all amino acids are labeled at P9.

Discussion

In this chapter, I found evidence that the MHCflurry 2.0 PS predictor improves upon existing methods at predicting neoantigens. I did not observe a similar advantage in a benchmark of viral epitopes, however, even when the PS logistic regression model was re-fit to the viral data. I also showed that the MHCflurry AP predictor has learned a signal that is meaningfully distinct from that of other antigen processing tools, predominantly as a result of considering a broader sequence footprint. While further verification on additional datasets is required, these results support the application of the MHCflurry tools for epitope prediction.

Several factors may contribute to the discordance between the neoantigen and viral epitope benchmarks. First, as mentioned, the viral dataset is biased, as only peptides predicted to be strong MHC class I binders (usually by NetMHCpan) are typically tested for immunogenicity. Peptides with weaker affinity that might still be ranked highly by the PS model due to favorable AP scores were likely not assayed, and thus assumed to be non-epitopes in the benchmark. Second, in the neoantigen experiments, tumor mutations were centered in 25-mers and expressed in autologous cells. This may be ideal for the AP predictor, as the assay itself confirms that the epitope is naturally processed from its flanking sequence. The viral benchmark used a mix of assays, which in many cases were independent of antigen processing (e.g. tetramer screening). Finally, biological differences may play a role. The AP predictor is trained on peptides from unmutated human proteins, which may more closely match the processing of neoantigens than viral proteins, for example due to amino acid usage differences. Also, viral infection is expected to activate an interferon response, leading to expression of immunoproteasome subunits and a different cleavage pattern (chymotryptic instead of caspase-like cleavage). However, recent data suggests the effect on the presented repertoire is modest (Abelin et al. 2017), and the motifs observed for the AP predictor are actually closer to those of the immunoproteasome than the constitutive proteasome, probably because it also models TAP transport. Direct viral manipulation of antigen processing steps such as TAP transport could also conceivably play a role (Tomazin et al. 1996; Kyritsis et al. 2001; Ressing et al. 2005). Overall, a larger

and less biased viral epitope set is needed to investigate these issues. In response to the ongoing pandemic, several groups are pursuing unbiased surveys of the T cell response to the SARS-CoV-2 virus, which will likely prove a useful source of data for a future analysis (Grifoni, Weiskopf, et al. 2020a; Bert et al. 2020).

In the correlative analysis of antigen processing predictors, all predictors focused predominantly on the C-terminal position in the peptide. This is expected as proteasomal cleavage, TAP transport, and MHC class I binding for many alleles all converge with strong preferences on this position. The NetChop C-term and NetCTLpan TAP predictors showed good agreement with the AP predictor at this position, and NetChop 20S showed modest agreement. What appears to differentiate the AP predictor from the others is its relatively higher sensitivity to most other positions of the peptide, especially at P1. This could reflect additional biological processes that are modeled by the AP predictor, such as trimming by ER-resident aminopeptidases (ERAP). In Chapter 3, it was observed that the AP predictor modeled a depletion of proline at N1 and P1 and an enrichment at P2, which are hallmarks of trimming by ERAP (Serwold et al. 2002).

Overall, these results highlight that current MHC class I ligand predictors are sensitive but not precise at epitope prediction: most epitopes are predicted to be presented on MHC class I, but many peptides with equally-favorable predictions are not targeted by a T cell response. While the AP predictor may in some cases contribute a useful additional signal, its inclusion does not qualitatively change this overall picture. Understanding the factors beyond MHC presentation that distinguish T cell antigens remains a key challenge for the field.

Methods

Benchmark predictors. This chapter benchmarked predictions from MHCflurry 2.0 BA, AP, and PS (T. J. O'Donnell, Rubinsteyn, and Laserson 2020), NetMHCpan 4.0 BA and EL (V. Jurtz et al. 2017), and MixMHCpred 2.0.2 (Gfeller et al. 2018). These are the same predictors evaluated in Chapter 3. For the predictors that provide a percentile output to rescale scores by an allele-specific distribution, I benchmarked both the unadjusted result (nanomolar affinity or score) and the percentile rank result separately. In the neoantigen benchmark, I also considered MHCflurry BA and PS using prevalent alleles in the population

rather than patient-specific alleles. The prevalent alleles used were A*01:01, A*02:01, A*03:01, A*11:01, A*23:01, A*24:02, B*07:02, B*08:01, B*35:01, B*40:01, B*44:02, and B*44:03, as in (Weiskopf et al. 2020). For the viral benchmark, I additionally evaluated a variant of MHCflurry PS that was re-fit to viral epitopes. This predictor, referred to as MHCflurry PS2, used a logistic regression model integrating BA and AP scores, as in the original MHCflurry PS. For additional flexibility beyond that of MHCflurry PS model, I also included interaction terms and powers of the AP and BA predictions. All products of AP and BA up to polynomial degree three were included in the model. The PS2 mode was trained and evaluated on the viral epitope data using two-fold cross validation.

Neoantigens. The neoantigen benchmark was derived from the "NCI dataset" published by Bjoern Peters's group (Koşaloğlu-Yalçın et al. 2018). This dataset combines neoantigen screens previously published by Steve Rosenberg's group with additional data (Tran et al. 2015; Gros et al. 2016). For each sequence, I generated predictions for all 8-11-mer peptides that contained the mutation and took the strongest prediction over the peptides and MHC class I alleles as the prediction for the overall sequence.

Viral epitopes. T cell epitopes were downloaded from the Immune Epitope Database (IEDB) on April 27, 2020 (tcell_full_v3.zip from the IEDB database export page). I used a permissive definition of immunogenicity, requiring only a detected T cell response in at latest one experiment. Only epitopes annotated with a single four digit HLA class I restriction and an unambiguous identifier for the source protein (a non-null "Antigen Source Molecule IRI" column) were considered. I further filtered to source proteins that contained at least 10 epitopes. This resulted in a set of 1,380 epitopes spread across 65 HLA class I alleles. The A*02:01 allele accounted for over half the epitopes, with the rest split more evenly among alleles. Only four alleles had 50 or more epitopes: HLA-A*02:01 (707 epitopes), A*24:02 (160), B*35:01 (95), and B*07:02 (83). The full amino acid sequence of each epitope's source protein was downloaded and all peptides of length 8-11 from these proteins not in the epitope set were considered non-epitopes.

Antigen processing predictors. I compared the MHCflurry AP tool to the NetChop 3.0 and the NetCTLpan TAP predictors. The NetChop 3.0 tool (Nielsen et al. 2005), has two variants fit to different datasets. The "C-term" variant is trained on the C-terminal residues of MHC class I ligands, meaning that, although it is described as a cleavage predictor, it may be expected to model a mix of cleavage, MHC binding preferences, and TAP transport effects. The NetChop "20S" variant is fit to a small dataset of proteasomal cleavage sites identified definitively by mass spec or Edman degradation. The NetCTLpan TAP transport predictor is based on an earlier predictor fit to a dataset of peptide/TAP binding affinities (Björn Peters, Bulik, et al. 2003); TAP transport efficiency is closely related to peptide/TAP binding affinity. It considers the C-terminal residue and the first three N-terminal residues of the peptide. As TAP transport occurs prior to trimming by ERAP, N-terminal extensions are also considered, with the final prediction defined as the mean prediction for the original peptide and its extension by one N-terminal residue. The NetCTLpan predictor also includes an MHC class I binding predictor based on an early version of NetMHCpan.

Correlative analysis. To compare the sequence preferences of MHCflurry 2.0 AP to those of NetChop C-term 3.0, NetChop 20S 3.0, and the NetCTLpan TAP predictor, I generated predictions for all tools across 9-mer peptides from an arbitrarily-selected viral proteome (SARS-Cov-2; UniProt sequences P0DTC2, P0DTC9 P0DTC1, P0DTC7, P0DTD2, P0DTD1, P0DTC8, P0DTC6, P0DTD8, P0DTC3, P0DTC4, P0DTC5, P0DTD3). To generate predictions for the existing tools, I used the IEDB's predictor interface (http://tools.iedb.org/netchop/) to run NetCTLpan using the HLA-A*01:01 allele. This gave scores for NetChop C-term 3.0, NetCTLpan TAP, NetMHCpan, as well as the NetCTLpan composite score. I additionally used the same interface to run NetChop 20S 3.0.

Positional importance analysis. To quantify the importance predictors placed on each position in the peptide, I transformed the predictor's scores to percentiles and calculated the mean percentile score for each of the 20 possible amino acids at every position. The standard deviation of these 20 values for a position,

which captures the extent to which some amino acids deviate from the expected percentile (0.50), was used as the importance score for the position. The mean percentiles for each amino acid at a position were used to examine each predictor's specific preferences.

Quantification and statistical analysis. Error estimates were computed using bootstrap resampling. As the metric that measures the number of tests needed to capture a given fraction of epitopes is highly sensitive to the total number of epitopes, to estimate the uncertainty in these analyses I used a modified bootstrap procedure that held constant the number of epitopes and non-epitopes in each bootstrap sample.

Chapter 5. Future directions

Summary

As seen in the previous chapter, there is still substantial room for improvement in T cell epitope prediction. In this chapter, I discuss avenues for improving predictive accuracy on this problem. The first two sections describe potential approaches for improving prediction of peptides presented on MHC class I and II, respectively. The final section outlines ongoing work that moves beyond MHC presentation to predict immunodominant T cell epitopes.

Prediction of peptides presented on MHC class I

For MHCflurry BA and other MHC class I binding predictors, suboptimal accuracy on poorly-characterized alleles, especially those that are distant in sequence from well-characterized alleles, remains a persistent issue. For example, three human alleles (HLA-A*02:08, HLA-B*41:05, HLA-B*45:06), all of which had minimal training data, showed MHCflurry 2.0 BA area under the curve (AUC) significantly lower than 0.90 in Chapter 3 (Supplemental Figure 3-1). An additional source of data that may enable improvements for alleles such as these is multiallelic mass spec elution experiments, which are not used to fit MHCflurry 2.0 BA or NetMHCpan 4.0. Multiallelic data was recently incorporated into the NetMHCpan 4.1 training set using a scheme that iteratively annotates each peptide in a multiallelic experiment with the allele for which it is most likely to bind, given the current model weights (Reynisson, Alvarez, et al. 2020). A similar approach is a possible direction for improving MHCflurry BA accuracy on alleles that are only characterized by multiallelic mass spec.

In addition to learning sequence preferences, the MHCflurry 2.0 binding affinity (BA) predictor also implicitly models the peptide length preferences of each MHC class I allele. There is at least one instance where this learned length distribution is suboptimal: the mouse H-2 Kb allele. This allele is unusual in that it prefers 8-mer rather than 9-mer peptides (K. Falk et al. 1991). The MHCflurry BA sequence motif

for this allele (Figure 5-1A) is consistent with experimental data, showing a preference for F or Y at position five and L at position eight (K. Falk et al. 1991). However, the model incorrectly predicts a preference for 9-mer peptides (Figure 5-1B). The most likely cause is that too much information is shared from other (9-mer-preferring) alleles. Similar results are also observed for the macaque allele Mamu-A2*05:01, which is also thought to prefer 8-mers, although more experimental data is required (de Groot et al. 2017). A possible fix for such cases is to directly push the model to learn the correct length distribution, for example by including an additional term in the loss function to penalize differences between the learned length distribution and that of peptides identified by mass spec.

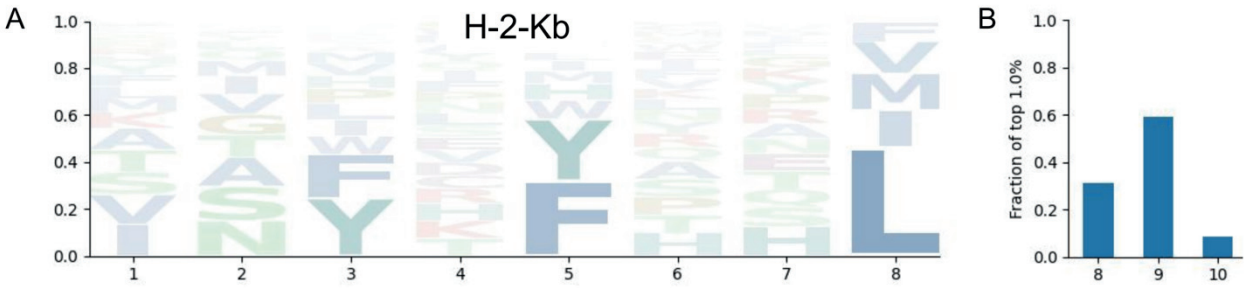


Figure 5-1. Predicted binding motif and peptide length distribution for H-2 Kb
(a) Predicted binding motif for H-2 Kb 8-mer peptides. **(b)** Predicted length distribution for H-2 Kb.

Another category of follow-up for the BA predictor is model interpretation. A better understanding of what the predictor learned has the potential to illuminate unappreciated structural determinants of the peptide / MHC interaction and suggest additional avenues for improving the models. One direction to explore is cases where the models incorporate interactions between peptide residues and other nonlinear effects. These have been identified in individual cases, but have not been systematically explored using models trained on recent datasets (Björn Peters, Tong, et al. 2003). A similarly useful direction would be to better understand the impact of each MHC residue as predicted by the model. As an example of this kind of analysis, the alleles B*39:83 and B*39:109 differ (in terms of positions used in neural network input) only at position 67, having a phenylalanine and a cysteine at this position, respectively. These are rare alleles, with no representation in the training set. The binding motifs for these alleles are predicted to differ

at peptide position 2, with B*39:83 strongly preferring proline (Figure 5-2A) and B*39:109 preferring histidine (Figure 5-2B). This effect of position 67 learned by the model is plausible given the physical proximity of MHC residue 67 and peptide position 2 (Figure 5-2C), and in fact has been suggested by structural analyses (Yagüe, Vázquez, and López de Castro 1998; Smith et al. 1996). It would be an interesting follow-up to systematically extend analyses such as this to other positions, whose interactions are not well-established.

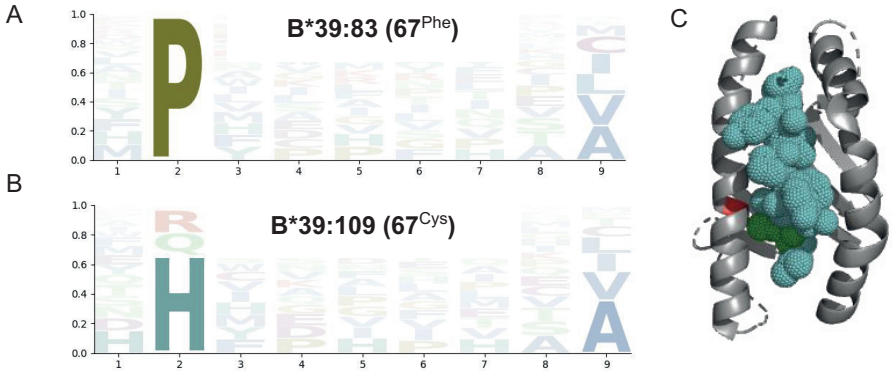


Figure 5-2. MHC position 67 is predicted to impact preferred residues at peptide position 2

(a) Sequence logo for B*39:83, which has a Phe at position 67. **(b)** Sequence logo for B*39:109, which has a cysteine at position 67. **(c)** Crystal structure of MHC class I binding groove. MHC position 67 is highlighted in red and peptide position 2 in green. The rest of the peptide is shown in cyan. The structure is A*02:01 in complex with LLFGYPVYV peptide, PDB ID 1duz (A. R. Khan et al. 2000).

Another useful step would be to investigate if including additional MHC residues in the input representation (pseudosequence) can improve accuracy. The 37-mer MHC pseudosequence used in MHCflurry 2.0 BA is essentially the same as the 34-mer representation used by NetMHCpan, differing only by the addition of three residues to differentiate a few pairs of rare alleles. This representation excludes many polymorphic positions that do not directly contact the peptide in the available crystal structures. Some of these positions may still contribute to peptide binding by affecting the MHC conformation or by modulating interactions with chaperones such as tapasin. A simple approach to investigate this question would be to exhaustively test the accuracy impact of including of each polymorphic residue separately

using cross validation. This has the potential to identify an improved MHC representation and generate hypotheses on unappreciated structural interactions.

For the antigen processing (AP) predictor, a clearer understanding of the biological processes and assay biases learned would help explain in what contexts it may be useful and potentially suggest refinements to improve accuracy. One useful step would be determining the extent to which MHCflurry AP models the effects of ERAP. I have made an initial attempt to investigate this based on published MS elution data from rats with or without the *Erap1* gene (Barnea et al. 2017) and have observed that eluted 9-mer peptides from *Erap1* knockout animals have modestly lower AP scores than those from wildtype animals. However, the reverse relationship is observed for longer peptides. It is unclear how to interpret this result. It would also be helpful to test if the AP predictor has learned signals of cleavage by cytosolic peptidases, for example thimet oligopeptidase (York et al. 2003; Lázaro, Gamarra, and Del Val 2015). Development of a new AP predictor that adjusts for – rather than learns – mass spec biases is also an important next step. Here it may be helpful to include other biases besides cysteine depletion (Fusaro et al. 2009). Finally, a number of cases have been reported where flanking residues are critical to the presentation of an immunogenic epitope (Del Val et al. 1991; Yellen-Shaw et al. 1997; Shastri, Serwold, and Gonzalez 1995; Beekman et al. 2000; Draenert et al. 2004). Verifying if these individual instances are predicted by MHCflurry AP could further test the relevance of its predictions.

Accounting for peptide abundance is a relatively simple next step for improving prediction of presented pMHC. As expected from a thermodynamic model, there appears to be a multiplicative interaction between peptide abundance and binding affinity in determining the number of a particular cell-surface pMHC (Abelin et al. 2017). A model integrating peptide abundance and affinity would likely be especially applicable to tumor neoantigen prediction, as tumor RNA sequencing is frequently available and each tumor has a potentially unique expression profile. While existing neoantigen prediction pipelines account for peptide source gene expression using heuristics, I am aware of no tools that apply a data-driven model to integrate expression and binding affinity (Hundal et al. 2016; Bjerregaard et al. 2017; Kodysh and Rubinsteyn 2020). One issue, however, is that mass spec elution data is generally not quantitative, making

it difficult to derive a training set. The imperfect relationship between transcript expression and protein abundance is another complication (Y. Liu, Beyer, and Aebersold 2016).

Further improvements are likely possible by taking into account variation in MHC protein abundance. *HLA-C* is typically expressed at less than one tenth the level of *HLA-A* or *HLA-B* (Apps et al. 2015), and there is also within-gene variation due to differences in methylation and microRNA regulation across alleles (O'huigin et al. 2011; Ramsuran et al. 2015). For example, HLA-A*24 is expressed at nearly four times the level of HLA-A*03 (Ramsuran et al. 2018). A full accounting of these differences will require more data linking MHC alleles with surface abundance. Quantification of surface MHC by flow cytometry is the most straightforward approach but is limited by the availability of antibodies specific to particular MHC alleles. It may instead be possible to analyze multiallelic mass spec elution datasets to understand variations in MHC expression, as peptides with motifs for high-abundance MHC alleles would be expected to be more prevalent in these datasets. This too is difficult, however, as there is also variation in repertoire breadth across alleles, and peptides for some binding motifs are likely more easily detected by mass spec than others. My preliminary attempts to perform such analyses have not given results consistent with other sources of MHC expression information.

Predicting the stability (off-rate) rather than the affinity (off-rate divided by on-rate) of the peptide / MHC complex may also improve predictions of the steady-state cell-surface pMHC repertoire (Harndahl et al. 2012; Strønen et al. 2016), as the off-rate is likely the main driver of the overall level of a pMHC on the surface of the cell. Predictors for pMHC stability have been developed (Jørgensen et al. 2014), but were limited by a lack of data. The recent development of high-throughput assays for pMHC stability may create an opportunity to revisit stability prediction (Blaha et al. 2019). Mass spec elution data may also turn out to be a readout of stability more than affinity. This question would benefit from investigation.

An additional category of potential improvements relates to peptides with post translational modifications, especially phosphorylation. Phosphopeptides are presented by MHC class I and II (Zarling et al. 2006; Depontieu et al. 2009) and are especially attractive for cancer immunotherapy, as they may arise from oncogenic aberrations in signaling (Cobbold, De La Peña, and Norris 2013). Phospho-

neoantigens may even be shared across patients with a common signaling aberration, making them highly attractive targets for cancer vaccines. Phosphorylated peptide residues can interact with the MHC binding pocket to stabilize or abrogate peptide binding (Mohammed et al. 2008; Alpízar et al. 2017) as well as determine TCR reactivity (Mohammed et al. 2017; Cobbold, De La Peña, and Norris 2013). Until recently, only hundreds of presented phosphopeptides were known, but work by David Gfeller's group has now identified over two thousand phosphopeptides by mass spec, enough for a first attempt at a phosphorylation-aware MHC ligand predictor (Solleder et al. 2020). The extension of pan-allele predictors like MHCflurry 2.0 BA or NetMHCpan to phosphopeptides is a promising direction as data accumulates.

A final consideration is that elicitation of a CD8+ T cell response requires priming by dendritic cells (DCs) cross-presenting the pMHC. The mass spec elution data used to develop models such as MHCflurry are overwhelmingly capturing direct-presentation, and the two processes may have different biases. For example, it was shown that a peptide derived from the N-terminal signal sequence of a secreted protein construct underwent efficient direct-presentation but not cross-presentation (Wolkers et al. 2004). Similarly, a strong dependence on flanking sequence for cross-presentation but not direct-presentation has been demonstrated for the model SIINFEKL epitope (Ma et al. 2009). It is likely that effects such as these could be modeled given enough data, resulting in improved prediction of the cross-presented pMHC repertoire.

Prediction of peptides presented on MHC class II

A CD4+ (helper) T cell response is required for high-quality antibody generation against proteins and can also potentiate the response by other immune compartments. CD4+ T cells recognize MHC class II-restricted antigens, typically presented by professional antigen presenting cells. Predicting presentation on MHC class II is of great interest for vaccine development and other applications, but has lagged behind MHC class I ligand prediction in accuracy. Modeling MHC class II binding is more difficult for several reasons. The peptide-binding groove of MHC class II is open at both ends, allowing long peptides of variable length to bind (Brown et al. 1993). Predictive models must identify the part of an MHC class II-

binding peptide that lies in the binding groove not only to generate predictions, but also during training. The binding also appears somewhat less specific for MHC class II, as much of the stabilization of the peptide comes from hydrogen bonding between MHC residues and the peptide backbone, rather than the interactions with the peptide side chains that determine MHC class II allelic motifs (X. Liu et al. 2002). The fact that both chains are polymorphic for the MHC class II heterodimer and can pair in multiple ways is an additional complicating factor.

The main reason for the lower performance of MHC class II ligand prediction, however, is that there is less available data. As of July 2020, there were about half a million MHC class I ligands or affinity measurements deposited in the Immune Epitope Database (IEDB), but only 93,000 entries for MHC class II. As with class I, however, recent large-scale mass spec surveys are addressing the data limitation, enabling the development of a new generation of MHC class II ligand predictors (Abelin et al. 2019; Racle et al. 2019; Reynisson, Barra, et al. 2020). The combination of these larger datasets will likely enable the use of more sophisticated machine learning approaches than are currently applied for this problem, such as the use of deeper neural networks, which may drive additional accuracy improvements in the near future.

Antigen processing is likely a key determinant of the peptide repertoire presented on MHC class II. Interestingly, it appears that taking into account the flanking sequences around MHC class II eluted ligands improves prediction of presented peptides but not actual CD4+ T cell epitopes (Paul et al. 2018; Reynisson, Barra, et al. 2020). One possibility is that the mass spec elution experiments, which typically involve cells presenting self-peptides derived from an autophagy pathway, may be a misleading source of information on the endocytic MHC class II presentation pathway that dendritic cells use to prime T cell responses. If this is the case, one approach, in some respects similar to that taken in Chapter 3, would be to use the standard mass spec elution experiments to develop an MHC class II binding affinity predictor and small datasets of endocytosed antigens displayed on MHC class II to develop a processing predictor. Potential data sources for such an antigen processing model could include the endocytosed antigen presented on dendritic cells (DCs) that were fed tumor cells in the work by Abelin et al. (Abelin et al. 2019) or bacteria-derived peptides presented by mouse DCs from Graham et al. (Graham et al. 2018).

As an additional possible direction for MHC class II prediction, recently it was shown that the open binding groove of MHC class II enables the interrogation of MHC class II binding motifs using a peptide array (Osterbye et al. 2020). Applying this assay to additional alleles could generate a dataset large enough to develop a new MHC class II ligand predictor, potentially integrating affinity measurements, mass spec eluted ligands, and peptide array results. Exploring alternatives to peptide arrays that allow even larger libraries, such as programmable phage display, is another possible direction (Larman et al. 2011).

Predicting immune recognition

Only a small fraction of the peptide content of a virus is targeted by an immune response. For example, about half of the CD8+ T cell response to vaccinia virus in C57BL/6 mice is directed to just five epitopes, out of about 175,000 peptides in the viral proteome (Tscharke et al. 2005; Moutaftsi et al. 2006). The vast majority (> 90%) of the remaining T cells are directed toward about 50 other epitopes. This organization into dominant and sub-dominant responses is not stochastic: the same epitopes will reproducibly elicit the same level of response among C57BL/6 mice. The term “immunodominance” has been coined to describe this phenomenon (Yewdell and Bennink 1999).

Predicting or experimentally determining MHC presentation narrows the search space but is not sufficient on its own to identify immunodominant epitopes. Using the vaccinia virus example, on the order of 170 vaccinia peptides are presented on MHC class I, at least at levels sufficient for detection by mass spectrometry (Croft et al. 2019). While this is a great reduction from the full viral proteome, refining this list to the five dominant epitopes is a predictive problem that has so far eluded substantial progress. The reason for this difficulty is that immunodominance is not the result of a single pathway but rather the final outcome of competing genetic, molecular, biological, and ecological processes operating at widely disparate scales of space and time (Yewdell 2006). It is not clear what factors should be incorporated in a predictive model of immunodominance. I describe three overlapping approaches that are under investigation.

First, some physical properties of the peptide or pMHC complex do seem to contribute to recognition, both by CD8+ (Calis et al. 2013; Chowell et al. 2015) and CD4+ T cells (Dhanda et al. 2018). The general observation has been that large, hydrophobic residues toward the center of the peptide promote immunogenicity. This effect is only mildly predictive, however. Similarly, it has been shown that tumor neoantigens tend to have similarities with epitopes from infectious diseases (Łuksza et al. 2017), presumably due to common sequence properties that promote TCR interaction. The accumulation of larger datasets of T cell epitopes using high-throughput technologies will likely enable the extension and integration of signals like these (Sharma, Rive, and Holt 2019).

Second, immunodominance models can take into account biases specific to the disease context. For example, there is evidence that CD8+ responses against viral proteins preferentially target proteins expressed early in the life cycle (Oseroff et al. 2005; Jing et al. 2007; Croft et al. 2019), whereas CD4+ responses target structural proteins expressed late (Moutaftsi et al. 2007). The prior exposure history of an individual is likely important in determining the dominant epitopes in a new challenge by a related pathogen by shaping the memory T cell repertoire, as seen in SARS-exposed individuals tested for reactivity to SARS-CoV-2 (Le Bert et al. 2020). In cancer, one study found that clonal mutated pMHC present across all cancer cells in a tumor are more likely to be targeted than those present only in a subclone (McGranahan et al. 2016). It has also been suggested that somatic frameshift mutations may be more immunogenic than missense mutations due to lower similarity to self, although this awaits clear validation (Roudko, Greenbaum, and Bhardwaj 2020). While it is not always obvious how best to quantitatively integrate biases like these into predictive models, improved immunogenicity predictions are likely possible for key disease contexts.

Finally, the repertoire of naive T cells plays a central role in determining immunodominance. Modeling the size of the naive pool available to recognize a given pMHC, while currently out of reach, may ultimately be the key step forward in accurate immunodominance prediction. This will require a number of major methodological advances, however. Most importantly, an accurate model of the interaction strength

between an arbitrary TCR and pMHC is needed. There has been only limited early work on this problem so far (V. I. Jurtz et al. 2018) due to a lack of training data. Recent methods to read out interacting pairs of pMHC and TCRs may soon begin to address this issue (Amalie K. Bentzen et al. 2018; Li et al. 2019; Joglekar et al. 2019). We would also need a model of the naive TCR repertoire. While modeling TCR rearrangement appears tractable (Murugan et al. 2012; Marcou, Mora, and Walczak 2018), any useful model of the naive repertoire would probably also have to take into account positive and negative selection. For example, it has been shown that positive selection can have critical effects on immunodominance (Lo et al. 2014). While these challenges are substantial, a convergence of high-throughput technologies for TCR/pMHC identification and repertoire analyses may soon deliver the datasets needed to make headway.

Conclusion

Despite great progress in the past two decades, there is still much work ahead to achieve accurate prediction of T cell epitopes. Methods for improving the prediction of MHC class I-presented peptides will likely focus on the incorporation of additional input signals or extension to novel ligands, such as phosphopeptides. Improvements to class II peptide presentation will likely come from an accelerating accumulation of mass spectroscopy elution experiments and refinement of machine learning models. In the longer term, models that predict the pMHC / TCR interaction, combined with knowledge of the naive TCR repertoire, may eventually allow a direct prediction of immunodominance.

Appendix 1. Supplemental Figures and Tables

Chapter 2

Supplemental Table 2-1. Training, model selection, and validation dataset sizes and performance for all predictors by allele

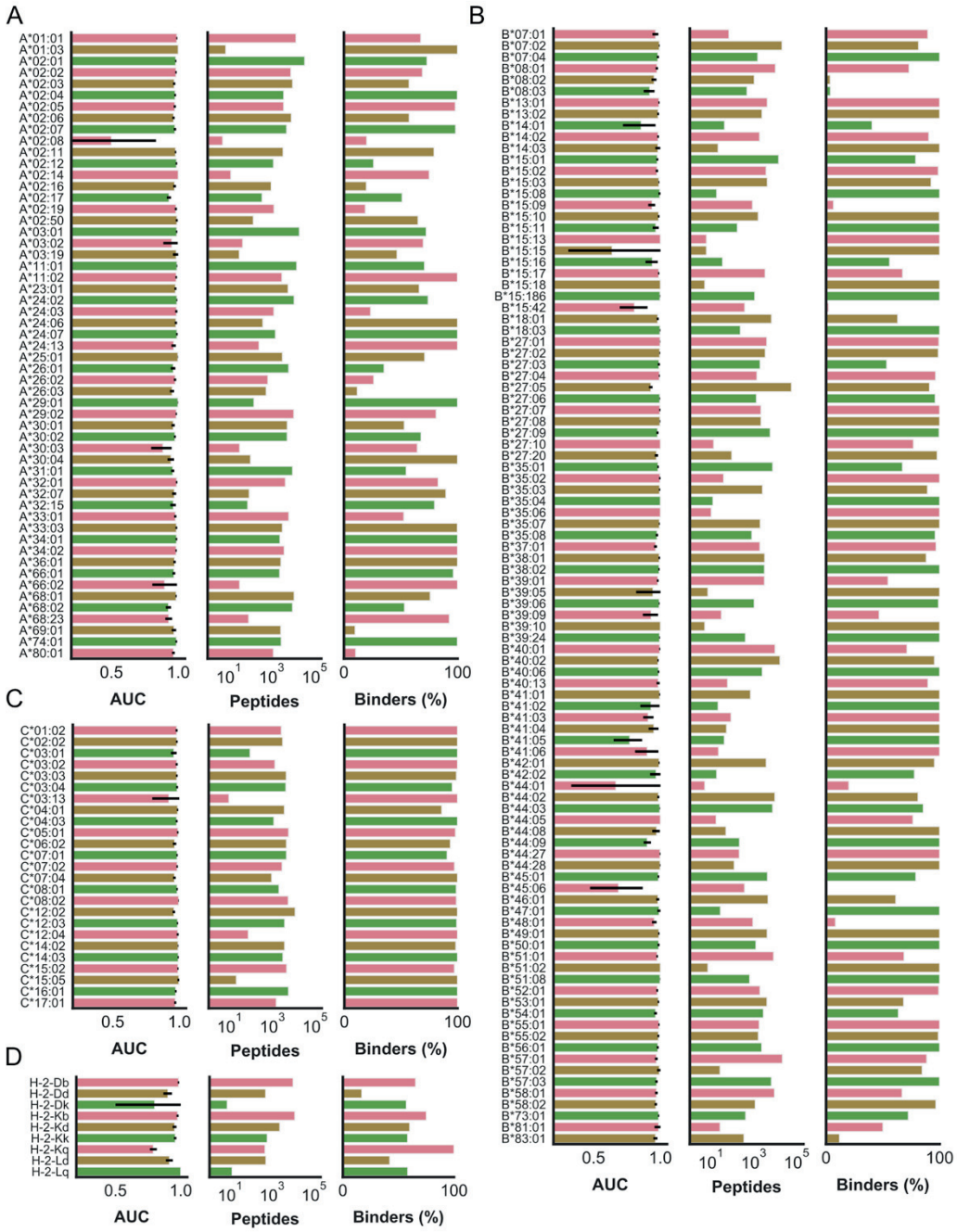
The scores marked "pre-model selection" indicate performance of the full ensemble (320 models for each allele) on the data held out for model selection.

Supplemental Table 2-1 is provided separately.

Supplemental Table 2-2. HPV benchmark scores

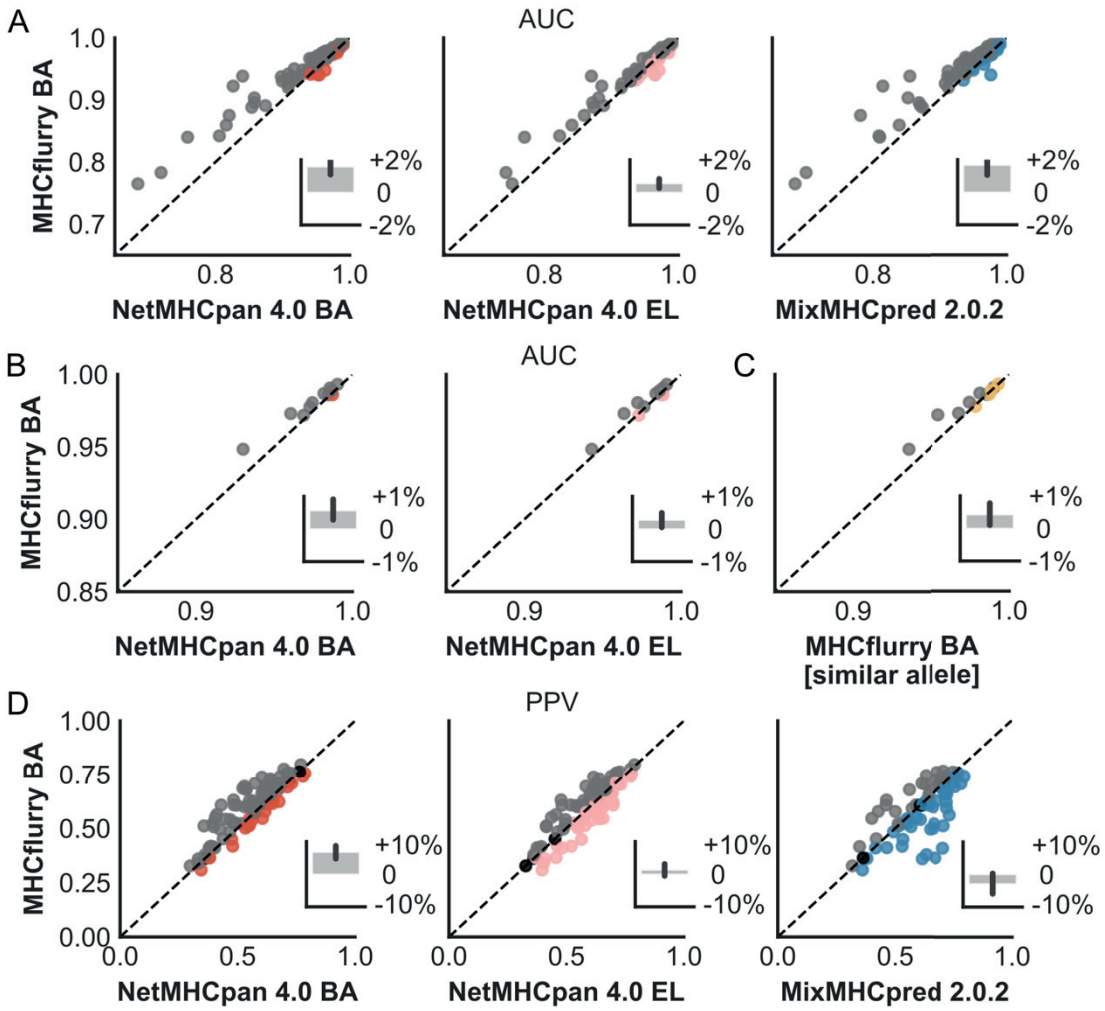
Each cell indicates a score, the bootstrap 95% confidence interval for the score, and the 95% confidence interval for the score's difference with respect to MHCflurry 1.2.0 (positive values indicate higher performance than MHCflurry 1.2.0; intervals that exclude the value 0 are highlighted with an asterisk). The nominally best-performing predictor for each metric is shown in bold. The AUC(any) column indicates the area under the receiver operator characteristic curve (AUC) considering peptides with any level of binding detectable by the assay (approximately 100,000 nM or less) to be binders. The other AUC scores are calculated using the indicated nM thresholds to distinguish binders from non-binders. In the column labels for the AUC scores, the number of peptides considered a binder at the given threshold (of 475 total peptides) are shown in parenthesis.

	Kendall τ	Pearson r	AUC(any) (180)	AUC@500 nM (7)	AUC@5000 nM (52)	AUC@15000 nM (99)
NetMHC 4.0	0.348 (0.30-0.39) -0.08 - +0.008	0.453 (0.39-0.51) -0.1 - -0.01*	0.762 (0.73-0.80) -0.06 - +0.01	0.826 (0.71-0.93) -0.2 - -0.02*	0.746 (0.70-0.79) -0.08 - +0.02	0.799 (0.76-0.84) -0.03 - +0.05
NetMHCpan 3.0	0.357 (0.31-0.40) -0.07 - +0.01	0.475 (0.42-0.53) -0.1 - +0.01	0.765 (0.73-0.80) -0.06 - +0.01	0.874 (0.81-0.94) -0.1 - +0.02	0.763 (0.72-0.81) -0.05 - +0.04	0.802 (0.76-0.84) -0.03 - +0.05
NetMHCpan 4.0	0.369 (0.33-0.41) -0.06 - +0.03	0.492 (0.44-0.55) -0.08 - +0.02	0.776 (0.74-0.81) -0.05 - +0.02	0.883 (0.82-0.94) -0.07 - +0.007	**0.776** **(0.73-0.82)** **-0.04 - +0.05**	**0.808** **(0.77-0.85)** **-0.02 - +0.06**
MHCflurry (no MS)	0.351 (0.30-0.40) -0.06 - -0.007*	0.458 (0.40-0.53) -0.1 - -0.02*	0.763 (0.73-0.80) -0.04 - -0.0009*	0.764 (0.59-0.93) -0.3 - -0.03*	0.756 (0.70-0.81) -0.05 - +0.02	0.770 (0.73-0.82) -0.04 - +0.005
MHCflurry (train-MS)	**0.388** **(0.34-0.43)** **-0.01 - +0.02**	**0.523** **(0.46-0.59)** **-0.02 - +0.03**	**0.791** **(0.75-0.83)** **-0.006 - +0.02**	**0.922** **(0.86-0.98)** **-0.02 - +0.04**	0.775 (0.72-0.82) -0.01 - +0.02	0.787 (0.74-0.83) -0.01 - +0.01
MHCflurry 1.2.0	0.384 (0.34-0.43)	0.519 (0.46-0.59)	0.785 (0.75-0.82)	0.912 (0.84-0.97)	0.770 (0.72-0.82)	0.788 (0.75-0.83)



Supplemental Figure 3-1. MHCflurry BA performance on data held-out from model fitting

HLA-A **(a)**, HLA-B **(b)**, HLA-C **(c)** and murine **(d)** MHC class I alleles are shown here; alleles for additional species are included in Supplemental Table 3-1. For each allele, the accuracy (AUC), the number of unique peptides in the training set, and the fraction of the training set that are binders (defined as MS hits or affinity measurements tighter than 500 nM) are indicated. Colors are for legibility.



Supplemental Figure 3-2. Comparison of predictive accuracy on the monoallelic benchmark

(a) AUC across all MONOALLELIC samples. The insets show the mean percent differences, with positive values indicating higher performance by MHCflurry BA. Error bars indicate 95% confidence intervals of the mean. A variant of MHCflurry BA trained without the MONOALLELIC benchmark datasets was used. Alleles that were unsupported by MixMHCpred are omitted from its evaluation. **(b)** AUC on the 11 samples in the MONOALLELIC benchmark that had no entries in the training data. MixMHCpred is not shown here because it does not support these alleles. **(c)** AUC of MHCflurry BA for same samples shown in (b) compared to that obtained from substituting predictions for related, well-characterized alleles. Predictions were performed using MHCflurry BA for the most closely-related allele (in terms of edit distance of the 37-mer pseudosequence) with at least 50 entries in the training data. **(d)** PPV across all MONOALLELIC samples.

Supplemental Table 3-1. MHCflurry BA performance on data held out from its training set

Supplemental Table 3-1 *is provided separately.*

Supplemental Table 3-2. Curated MHC class I mass spec datasets and accuracy scores

The expression column indicates the publicly-available RNA-seq source(s), using the following abbreviations: HPA-CELLS, Human Protein Atlas (Cell Lines dataset); HPA-BLOOD, Human Protein Atlas (Blood dataset), HPA-GTEX, Human Protein Atlas (GTEX dataset), CCLE, Expression Atlas (CCLE dataset), NATURE-MED, Metastatic melanoma expression from(Barry et al. 2018).

Supplemental Table 3-2 is provided separately.

Supplemental Table 3-3. Sample groups used to benchmark each predictor

The MULTIALLELIC set is the union of the MULTIALLELIC-OLD and MULTIALLELIC-RECENT samples.

Experiment	Dataset	Figures
Benchmark of NetMHCpan 4.0	MULTIALLELIC	Figure 3-1 and Figure 3-2
Benchmark of MixMHCpred 2.0.2	MULTIALLELIC-RECENT	Figure 3-1 and Figure 3-2
Benchmark of MHCflurry 2.0 BA	MULTIALLELIC	Figure 3-1 and Figure 3-2
Benchmark of MHCflurry BA variant trained without MONOALLELIC samples	MONOALLELIC	Supplemental Figure 3-2
Benchmark of MHCflurry 2.0 AP	MULTIALLELIC	Figure 3-2
Benchmark of MHCflurry 2.0 PS and other predictors	MULTIALLELIC-RECENT	Figure 3-3

Supplemental Table 3-4. Summary of predictor training datasets.

Predictor	Train data type	Training dataset	Figures
MHCflurry 2.0 BA	Binding affinities and MS	MONOALLELIC, IEDB(Vita et al. 2019), SysteMHC(Shao et al. 2018), Kim et al(Kim et al. 2014)	Figure 3-1, Figure 3-2, and Figure 3-3
MHCflurry BA variant	Binding affinities and MS	IEDB(Vita et al. 2019), SysteMHC(Shao et al. 2018), Kim et al(Kim et al. 2014)	Supplemental Figure 3-2
MHCflurry 2.0 AP	MS	MONOALLELIC	Figure 3-2
MHCflurry 2.0 PS (logistic regression weights)	MS	MULTIALLELIC-OLD	Figure 3-3

Supplemental Table 3-5. Performance on alleles not present in the training data

A predictor trained without the MONOALLELIC benchmark data was evaluated on the MONOALLELIC benchmark. The highest AUC for each allele is shown in bold. To compute the AUC shown in the last column, the most similar allele (in terms of edit distance of its 37-mer representation used as input to the predictor) with at least 50 measurements in the training data was selected, and the MHCflurry BA predictions for this allele were used instead.

		AUC			Well-characterized allele		
Allele	**Sample**	**NetMHCpan 4.0 BA**	**NetMHCpan 4.0 EL**	**MHCflurry BA**	**Allele**	**Distance**	**AUC**
A*24:07	KESKIN_A2407	0.986	0.989	**0.991**	A*24:02	1	0.990
A*34:01	KESKIN_A3401	0.973	0.976	**0.978**	A*66:01	2	0.979
A*34:02	KESKIN_A3402	0.969	**0.973**	0.972	A*03:01	3	0.955
A*36:01	KESKIN_A3601	0.930	0.943	**0.948**	A*01:01	2	0.936
B*07:04	KESKIN_B0704	0.985	**0.989**	0.986	B*07:02	1	0.986
B*35:07	KESKIN_B3507	0.991	0.991	**0.993**	B*35:01	0	0.993
B*38:02	KESKIN_B3802	**0.987**	0.987	0.986	B*38:01	1	0.987
B*40:06	KESKIN_B4006	0.982	0.984	**0.987**	B*40:02	2	0.981
C*03:02	KESKIN_C0302	0.974	0.972	**0.981**	C*03:04	2	0.975
C*04:03	KESKIN_C0403	0.961	0.964	**0.973**	C*04:01	1	0.968
C*14:03	KESKIN_C1403	0.987	0.986	**0.989**	C*14:02	0	0.989

Supplemental Table 3-6. Position weight matrices for the MHCflurry antigen processing predictors

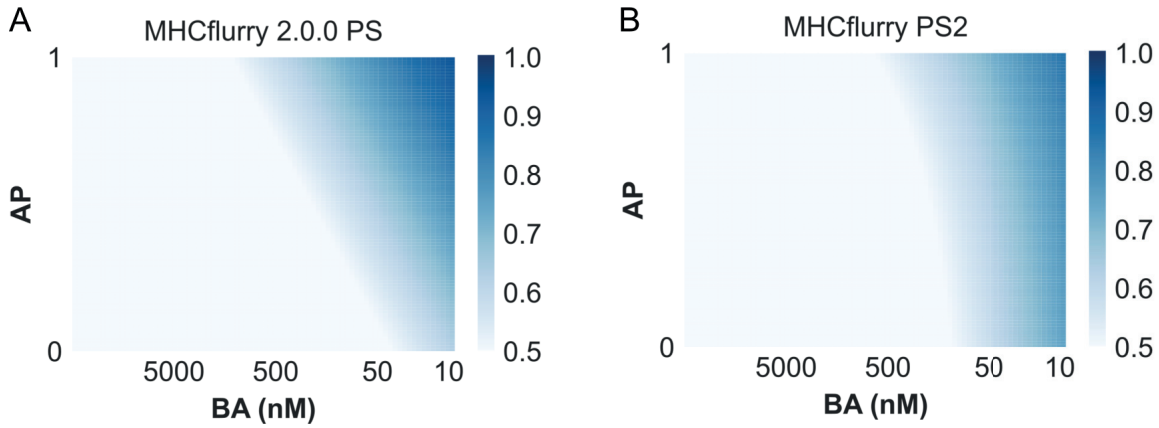
Supplemental Table 3-6 is provided separately.

Supplemental Table 3-7. Antigen processing predictions for proteasome-cleaved peptides

AP predictions are shown for the peptides identified using "mass spectrometry analysis of proteolytic peptides" (MAPP) by Wolf-Levy et al and decoy peptides drawn from the same genes.

Supplemental Table 3-7 is provided separately.

Chapter 4



Supplemental Figure 4-1. Predictions for the MHCflurry PS models across BA and AP values

(a) Predictions for the standard PS model. **(b)** Predictions for the PS variant that was fit to viral epitopes. Values were normalized to quantiles (ranks), with 1.0 corresponding to the highest prediction.

Bibliography

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." *arXiv [cs.DC]*. arXiv. http://arxiv.org/abs/1603.04467.

Abelin, Jennifer G., Dewi Harjanto, Matthew Malloy, Prerna Suri, Tyler Colson, Scott P. Goulding, Amanda L. Creech, et al. 2019. "Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction." *Immunity* 51 (4): 766–79.e17.

Abelin, Jennifer G., Derin B. Keskin, Siranush Sarkizova, Christina R. Hartigan, Wandi Zhang, John Sidney, Jonathan Stevens, et al. 2017. "Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-Allelic Cells Enables More Accurate Epitope Prediction." *Immunity* 46 (2): 315–26.

Alpízar, Adán, Fabio Marino, Antonio Ramos-Fernández, Manuel Lombardía, Anita Jeko, Florencio Pazos, Alberto Paradela, César Santiago, Albert J. R. Heck, and Miguel Marcilla. 2017. "A Molecular Basis for the Presentation of Phosphorylated Peptides by HLA-B Antigens." *Molecular & Cellular Proteomics: MCP* 16 (2): 181–93.

Andreatta, Massimo, and Morten Nielsen. 2015. "Gapped Sequence Alignment Using Artificial Neural Networks: Application to the MHC Class I System." *Bioinformatics* 32 (4): 511–17.

Apps, Richard, Zhaojing Meng, Gregory Q. Del Prete, Jeffrey D. Lifson, Ming Zhou, and Mary Carrington. 2015. "Relative Expression Levels of the HLA Class-I Proteins in Normal and HIV-Infected Cells." *Journal of Immunology* 194 (8): 3594–3600.

Arstila, T. P., A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky. 1999. "A Direct Estimate of the Human Alphabeta T Cell Receptor Diversity." *Science* 286 (5441): 958–61.

Barnea, Eilon, Dganit Melamed Kadosh, Yael Haimovich, Nimman Satumtira, Martha L. Dorris, Mylinh T. Nguyen, Robert E. Hammer, et al. 2017. "The Human Leukocyte Antigen (HLA)-B27 Peptidomein Vivo, in Spondyloarthritis-Susceptible HLA-B27 Transgenic Rats and the Effect of Erap1 Deletion." *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.m116.066241.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7.

Barry, Kevin C., Joy Hsu, Miranda L. Broz, Francisco J. Cueto, Mikhail Binnewies, Alexis J. Combes, Amanda E. Nelson, et al. 2018. "A Natural Killer-Dendritic Cell Axis Defines Checkpoint Therapy-Responsive Tumor Microenvironments." *Nature Medicine* 24 (8): 1178–91.

Bassani-Sternberg, Michal, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, et al. 2016. "Direct Identification of Clinically Relevant Neoepitopes Presented on Native Human Melanoma Tissue by Mass Spectrometry." *Nature Communications* 7 (May): 13404.

Bassani-Sternberg, Michal, Chloé Chong, Philippe Guillaume, Marthe Solleder, Huisong Pak, Philippe O. Gannon, Lana E. Kandalaft, George Coukos, and David Gfeller. 2017. "Deciphering HLA-I Motifs across HLA Peptidomes Improves Neo-Antigen Predictions and Identifies Allostery Regulating

HLA Specificity." *PLoS Computational Biology* 13 (8): e1005725.

Bassani-Sternberg, Michal, and David Gfeller. 2016. "Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions." *The Journal of Immunology* 197 (6): 2492–99.

Bassani-Sternberg, Michal, Sune Pletscher-Frankild, Lars Juhl Jensen, and Matthias Mann. 2015. "Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation." *Molecular & Cellular Proteomics: MCP* 14 (3): 658–73.

Beekman, N. J., P. A. van Veelen, T. van Hall, A. Neisig, A. Sijts, M. Camps, P. M. Kloetzel, J. J. Neefjes, C. J. Melief, and F. Ossendorp. 2000. "Abrogation of CTL Epitope Processing by Single Amino Acid Substitution Flanking the C-Terminal Proteasome Cleavage Site." *Journal of Immunology* 164 (4): 1898–1905.

Bentzen, Amalie Kai, Andrea Marion Marquard, Rikke Lyngaa, Sunil Kumar Saini, Sofie Ramskov, Marco Donia, Lina Such, et al. 2016. "Large-Scale Detection of Antigen-Specific T Cells Using Peptide-MHC-I Multimers Labeled with DNA Barcodes." *Nature Biotechnology* 34 (August). https://doi.org/10.1038/nbt.3662.

Bentzen, Amalie K., Lina Such, Kamilla K. Jensen, Andrea M. Marquard, Leon E. Jessen, Natalie J. Miller, Candice D. Church, et al. 2018. "T Cell Receptor Fingerprinting Enables in-Depth Characterization of the Interactions Governing Recognition of peptide–MHC Complexes." *Nature Biotechnology* 36 (12): 1191–96.

Bert, Nina Le, Nina Le Bert, Anthony T. Tan, Kamini Kunasegaran, Christine Y. L. Tham, Morteza Hafezi, Adeline Chia, et al. 2020. "SARS-CoV-2-Specific T Cell Immunity in Cases of COVID-19 and SARS, and Uninfected Controls." *Nature*. https://doi.org/10.1038/s41586-020-2550-z.

Bhasin, Manoj, Sneh Lata, and G. P. S. Raghava. 2007. "TAPPred Prediction of TAP-Binding Peptides in Antigens." *Methods in Molecular Biology* 409: 381–86.

Bjerregaard, Anne-Mette, Morten Nielsen, Sine Reker Hadrup, Zoltan Szallasi, and Aron Charles Eklund. 2017. "MuPeXI: Prediction of Neo-Epitopes from Tumor Sequencing Data." *Cancer Immunology, Immunotherapy: CII* 66 (9): 1123–30.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987. "The Foreign Antigen Binding Site and T Cell Recognition Regions of Class I Histocompatibility Antigens." *Nature*. https://doi.org/10.1038/329512a0.

Blaha, Dylan T., Scott D. Anderson, Daniel M. Yoakum, Marlies V. Hager, Yuanyuan Zha, Thomas F. Gajewski, and David M. Kranz. 2019. "High-Throughput Stability Screening of Neoantigen/HLA Complexes Improves Immunogenicity Predictions." *Cancer Immunology Research* 7 (1): 50–61.

Blees, Andreas, Dovile Januliene, Tommy Hofmann, Nicole Koller, Carla Schmidt, Simon Trowitzsch, Arne Moeller, and Robert Tampé. 2017. "Structure of the Human MHC-I Peptide-Loading Complex." *Nature* 551 (7681): 525–28.

Brown, J. H., T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger, and D. C. Wiley. 1993. "Three-Dimensional Structure of the Human Class II Histocompatibility Antigen HLA-DR1."

Nature 364 (6432): 33–39.

Bui, Huynh-Hoa, John Sidney, Bjoern Peters, Muthuraman Sathiamurthy, Asabe Sinichi, Kelly-Anne Purton, Bianca R. Mothé, Francis V. Chisari, David I. Watkins, and Alessandro Sette. 2005. "Automated Generation and Evaluation of Specific MHC Binding Predictive Tools: ARB Matrix Applications." *Immunogenetics* 57 (5): 304–14.

Buus, S., S. L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak. 2003. "Sensitive Quantitative Predictions of Peptide-MHC Binding by a 'Query by Committee' Artificial Neural Network Approach." *Tissue Antigens*. https://doi.org/10.1034/j.1399-0039.2003.00112.x.

Calis, Jorg J. A., Matt Maybeno, Jason A. Greenbaum, Daniela Weiskopf, Aruna D. De Silva, Alessandro Sette, Can Keşmir, and Bjoern Peters. 2013. "Properties of MHC Class I Presented Peptides That Enhance Immunogenicity." *PLoS Computational Biology* 9 (10): e1003266.

Caruana, Rich, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. "Ensemble Selection from Libraries of Models." In *Twenty-First International Conference on Machine Learning - ICML '04*. https://doi.org/10.1145/1015330.1015432.

Chen, Hanna, Lenong Li, Mirjana Weimershaus, Irini Evnouchidou, Peter van Endert, and Marlene Bouvier. 2016. "ERAP1-ERAP2 Dimers Trim MHC I-Bound Precursor Peptides; Implications for Understanding Peptide Editing." *Scientific Reports* 6 (August): 28902.

Chowell, Diego, Sri Krishna, Pablo D. Becker, Clément Cocita, Jack Shu, Xuefang Tan, Philip D. Greenberg, Linda S. Klavinskis, Joseph N. Blattman, and Karen S. Anderson. 2015. "TCR Contact Residue Hydrophobicity Is a Hallmark of Immunogenic CD8+ T Cell Epitopes." *Proceedings of the National Academy of Sciences of the United States of America* 112 (14): E1754–62.

Cobbold, M., H. De La Peña, and A. Norris. 2013. "MHC Class I–associated Phosphopeptides Are the Targets of Memory-like Immunity in Leukemia." *Science Translational Medicine*. https://stm.sciencemag.org/content/5/203/203ra125.short?casa_token=zj3Mo8phRwgAAAAA:0gVUITnMMHSgpFPXW86qqkERtlYiZG2us8krWZAALa9Jz2HIRIWov94hwddymWx8wpgm0wf-oi_d-Aw.

Collins, E. J., D. N. Garboczi, and D. C. Wiley. 1994. "Three-Dimensional Structure of a Peptide Extending from One End of a Class I MHC Binding Site." *Nature* 371 (6498): 626–29.

Creech, Amanda L., Ying S. Ting, Scott P. Goulding, John F. K. Sauld, Dominik Barthelme, Michael S. Rooney, Terri A. Addona, and Jennifer G. Abelin. 2018. "The Role of Mass Spectrometry and Proteogenomics in the Advancement of HLA Epitope Prediction." *Proteomics* 18 (12): e1700259.

Croft, Nathan P., Stewart A. Smith, Jana Pickering, John Sidney, Bjoern Peters, Pouya Faridi, Matthew J. Witney, et al. 2019. "Most Viral Peptides Displayed by Class I MHC on Infected Cells Are Immunogenic." *Proceedings of the National Academy of Sciences of the United States of America* 116 (8): 3112–17.

Del Val, M., H. J. Schlicht, T. Ruppert, M. J. Reddehase, and U. H. Koszinowski. 1991. "Efficient Processing of an Antigenic Sequence for Presentation by MHC Class I Molecules Depends on Its Neighboring Residues in the Protein." *Cell* 66 (6): 1145–53.

Depontieu, Florence R., Jie Qian, Angela L. Zarling, Tracee L. McMiller, Theresa M. Salay, Andrew Norris, A. Michelle English, et al. 2009. “Identification of Tumor-Associated, MHC Class II-Restricted Phosphopeptides as Targets for Immunotherapy.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (29): 12073–78.

Dhanda, Sandeep Kumar, Edita Karosiene, Lindy Edwards, Alba Grifoni, Sinu Paul, Massimo Andreatta, Daniela Weiskopf, et al. 2018. “Predicting HLA CD4 Immunogenicity in Human Populations.” *Frontiers in Immunology* 9 (June): 1369.

Dolton, Garry, Efthalia Zervoudi, Cristina Rius, Aaron Wall, Hannah L. Thomas, Anna Fuller, Lorraine Yeo, et al. 2018. “Optimized Peptide–MHC Multimer Protocols for Detection and Isolation of Autoimmune T-Cells.” *Frontiers in Immunology* 9: 1378.

Doytchinova, I. A., and D. R. Flower. 2001. “Toward the Quantitative Prediction of T-Cell Epitopes: coMFA and coMSIA Studies of Peptides with Affinity for the Class I MHC Molecule HLA-A*0201.” *Journal of Medicinal Chemistry* 44 (22): 3572–81.

Doytchinova, Irini A., Martin J. Blythe, and Darren R. Flower. 2002. “Additive Method for the Prediction of Protein−Peptide Binding Affinity. Application to the MHC Class I Molecule HLA-A*0201.” *Journal of Proteome Research*. https://doi.org/10.1021/pr015513z.

Doytchinova, Irini A., Pingping Guan, and Darren R. Flower. 2006. “EpiJen: A Server for Multistep T Cell Epitope Prediction.” *BMC Bioinformatics* 7 (March): 131.

Draenert, Rika, Sylvie Le Gall, Katja J. Pfafferott, Alasdair J. Leslie, Polan Chetty, Christian Brander, Edward C. Holmes, et al. 2004. “Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection.” *The Journal of Experimental Medicine* 199 (7): 905–15.

Falk, Kirsten, Olaf Rötzschke, Blazenka Grahovac, Dolores Schendel, Stefan Stevanović, Günther Jung, and Hans-Georg Rammensee. 1993. “Peptide Motifs of HLA-B35 and-B37 Molecules.” *Immunogenetics*. https://doi.org/10.1007/bf00190906.

Falk, K., O. Rötzschke, S. Stevanović, G. Jung, and H. G. Rammensee. 1991. “Allele-Specific Motifs Revealed by Sequencing of Self-Peptides Eluted from MHC Molecules.” *Nature* 351 (6324): 290–96.

Froloff, Nicolas, Andreas Windemuth, and Barry Honig. 1997. “On the Calculation of Binding Free Energies Using Continuum Methods: Application to MHC Class I Protein-Peptide Interactions.” *Protein Science: A Publication of the Protein Society* 6 (6): 1293–1301.

Fusaro, Vincent A., D. R. Mani, Jill P. Mesirov, and Steven A. Carr. 2009. “Prediction of High-Responding Peptides for Targeted Protein Assays by Mass Spectrometry.” *Nature Biotechnology* 27 (2): 190–98.

Garboczi, D. N., P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison, and D. C. Wiley. 1996. “Structure of the Complex between Human T-Cell Receptor, Viral Peptide and HLA-A2.” *Nature*.

Garcia, K. C., M. Degano, R. L. Stanfield, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton, and I. A. Wilson. 1996. “An Alpha Beta T Cell Receptor Structure at 2.5 A and Its Orientation in the TCR-MHC Complex.” *Science*. https://doi.org/10.1126/science.274.5285.209.

Gfeller, David, Philippe Guillaume, Justine Michaux, Hui-Song Pak, Roy T. Daniel, Julien Racle, George Coukos, and Michal Bassani-Sternberg. 2018. “The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands.” *Journal of Immunology* 201 (12): 3705–16.

Gloger, Andreas, Danilo Ritz, Tim Fugmann, and Dario Neri. 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 Immunology, Immunotherapy: CII* 65 (11): 1377–93.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Graham, Daniel B., Chengwei Luo, Daniel J. O’Connell, Ariel Lefkovith, Eric M. Brown, Moran Yassour, Mukund Varma, et al. 2018. “Antigen Discovery and Specification of Immunodominance Hierarchies for MHCII-Restricted Epitopes.” *Nature Medicine* 24 (11): 1762–72.

Grifoni, Alba, John Sidney, Yun Zhang, Richard H. Scheuermann, Bjoern Peters, and Alessandro Sette. 2020. “A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2.” *Cell Host & Microbe* 27 (4): 671–80.e2.

Grifoni, Alba, Daniela Weiskopf, Sydney I. Ramirez, Jose Mateus, Jennifer M. Dan, Carolyn Rydyznski Moderbacher, Stephen A. Rawlings, et al. 2020a. “Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals.” *Cell*. https://doi.org/10.1016/j.cell.2020.05.015.

———. 2020b. “Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals.” *Cell* 181 (7): 1489–1501.e15.

Groot, Natasja G. de, Corrine M. C. Heijmans, Arnoud H. de Ru, George M. C. Janssen, Jan W. Drijfhout, Nel Otting, Christelle Vangenot, et al. 2017. “A Specialist Macaque MHC Class I Molecule with HLA-B*27-like Peptide-Binding Characteristics.” *Journal of Immunology* 199 (10): 3679–90.

Gros, Alena, Maria R. Parkhurst, Eric Tran, Anna Pasetto, Paul F. Robbins, Sadia Ilyas, Todd D. Prickett, et al. 2016. “Prospective Identification of Neoantigen-Specific Lymphocytes in the Peripheral Blood of Melanoma Patients.” *Nature Medicine* 22 (4): 433–38.

Guillaume, Philippe, Sarah Picaud, Petra Baumgaertner, Nicole Montandon, Julien Schmidt, Daniel E. Speiser, George Coukos, Michal Bassani-Sternberg, Panagis Filippakopoulos, and David Gfeller. 2018. “The C-Terminal Extension Landscape of Naturally Presented HLA-I Ligands.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (20): 5083–88.

Gulukota, K., J. Sidney, A. Sette, and C. DeLisi. 1997. “Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules.” *Journal of Molecular Biology* 267 (5): 1258–67.

Hakenberg, Jörg, Alexander K. Nussbaum, Hansjörg Schild, Hans-Georg Rammensee, Christina Kuttler, Hermann-Georg Holzhütter, Peter-M Kloetzel, Stefan H. E. Kaufmann, and Hans-Joachim Mollenkopf. 2003. “MAPPP: MHC Class I Antigenic Peptide Processing Prediction.” *Applied Bioinformatics* 2 (3): 155–58.

Harndahl, Mikkel, Michael Rasmussen, Gustav Roder, Ida Dalgaard Pedersen, Mikael Sørensen, Morten

Nielsen, and Søren Buus. 2012. “Peptide-MHC Class I Stability Is a Better Predictor than Peptide Affinity of CTL Immunogenicity.” *European Journal of Immunology* 42 (6): 1405–16.

Harris, J. L., P. B. Alper, J. Li, M. Rechsteiner, and B. J. Backes. 2001. “Substrate Specificity of the Human Proteasome.” *Chemistry & Biology* 8 (12): 1131–41.

Hassan, Chopie, Eric Chabrol, Lorenz Jahn, Michel G. D. Kester, Arnoud H. de Ru, Jan W. Drijfhout, Jamie Rossjohn, et al. 2015. “Naturally Processed Non-Canonical HLA-A*02:01 Presented Peptides.” *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.m114.607028.

Hassan, Chopie, Michel G. D. Kester, Arnoud H. de Ru, Pleun Hombrink, Jan Wouter Drijfhout, Harm Nijveen, Jack a. M. Leunissen, Mirjam H. M. Heemskerk, J. H. Frederik Falkenburg, and Peter a. van Veelen. 2013. “The Human Leukocyte Antigen-Presented Ligandome of B Lymphocytes.” *Molecular & Cellular Proteomics: MCP* 12 (7): 1829–43.

Hattori, Akira, and Masafumi Tsujimoto. 2013. “Endoplasmic Reticulum Aminopeptidases: Biochemistry, Physiology and Pathology.” *Journal of Biochemistry* 154 (3): 219–28.

Hawkins, Oriana E., Rodney S. VanGundy, Annette M. Eckerd, Wilfried Bardet, Rico Buchli, Jon A. Weidanz, and William H. Hildebrand. 2008. “Identification of Breast Cancer Peptide Epitopes Presented by HLA-A* 0201.” *Journal of Proteome Research* 7 (4): 1445–57.

Henikoff, S., and J. G. Henikoff. 1992. “Amino Acid Substitution Matrices from Protein Blocks.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (22): 10915–19.

Hoof, Ilka, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. 2009. “NetMHCpan, a Method for MHC Class I Binding Prediction beyond Humans.” *Immunogenetics* 61 (1): 1–13.

Hundal, Jasreet, Beatriz M. Carreno, Allegra A. Petti, Gerald P. Linette, Obi L. Griffith, Elaine R. Mardis, and Malachi Griffith. 2016. “pVAC-Seq: A Genome-Guided in Silico Approach to Identifying Tumor Neoantigens.” *Genome Medicine* 8 (1): 11.

Hunt, D., R. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. Cox, E. Appella, and V. Engelhard. 1992. “Characterization of Peptides Bound to the Class I MHC Molecule HLA-A2.1 by Mass Spectrometry.” *Science*. https://doi.org/10.1126/science.1546328.

Jing, L., T. M. Chong, C. L. McClurkan, J. Huang, B. T. Story, and D. M. Koelle. 2007. “Diversity in the Acute CD8 T Cell Response to Vaccinia Virus in Humans.” *The Journal of Immunology*. https://doi.org/10.4049/jimmunol.179.4.2658.

Joglekar, Alok V., Michael T. Leonard, John D. Jeppson, Margaret Swift, Guideng Li, Stephanie Wong, Songming Peng, et al. 2019. “T Cell Antigen Discovery via Signaling and Antigen-Presenting Bifunctional Receptors.” *Nature Methods* 16 (2): 191–98.

Jørgensen, Kasper W., Michael Rasmussen, Søren Buus, and Morten Nielsen. 2014. “NetMHCstab - Predicting Stability of Peptide-MHC-I Complexes; Impacts for Cytotoxic T Lymphocyte Epitope Discovery.” *Immunology* 141 (1): 18–26.

Jurtz, Vanessa Isabell, Leon Eyrich Jessen, Amalie Kai Bentzen, Martin Closter Jespersen, Swapnil

Mahajan, Randi Vita, Kamilla Kjærgaard Jensen, et al. 2018. “NetTCR: Sequence-Based Prediction of TCR Binding to Peptide-MHC Complexes Using Convolutional Neural Networks.” *bioRxiv*. https://doi.org/10.1101/433706.

Jurtz, Vanessa, Sinu Paul, Massimo Andreatta, Paolo Marcatili, Bjoern Peters, and Morten Nielsen. 2017. “NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data.” *The Journal of Immunology* 199 (9): 3360–68.

Kai, Amalie, Bentzen Sine, and Reker Hadrup. 2017. “Evolution of MHC-Based Technologies Used for Detection of Antigen-Responsive T Cells.” *Cancer Immunology, Immunotherapy: CII* 66 (5): 657–66.

Keşmir, Can, Alexander K. Nussbaum, Hansjörg Schild, Vincent Detours, and Søren Brunak. 2002. “Prediction of Proteasome Cleavage Motifs by Neural Networks.” *Protein Engineering* 15 (4): 287–96.

Kessler, Jan H., Willemien E. Benckhuijsen, Tuna Mutis, Cornelis J. M. Melief, Sjoerd H. van der Burg, and Jan W. Drijfhout. 2004. “Competition-Based Cellular Peptide Binding Assay for HLA Class I.” *Current Protocols in Immunology / Edited by John E. Coligan ... [et Al.]*, 18–12.

Kessler, Jan H., Bregje Mommaas, Tuna Mutis, Ivo Huijbers, Debby Vissers, Willemien E. Benckhuijsen, Geziena M. Th Schreuder, et al. 2003. “Competition-Based Cellular Peptide Binding Assays for 13 Prevalent HLA Class I Alleles Using Fluorescein-Labeled Synthetic Peptides.” *Human Immunology* 64 (2): 245–55.

Khan, A. R., B. M. Baker, P. Ghosh, W. E. Biddison, and D. C. Wiley. 2000. “HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN (HLA-A 0201) IN COMPLEX WITH A NONAMERIC PEPTIDE FROM HTLV-1 TAX PROTEIN.” https://doi.org/10.2210/pdb1duz/pdb.

Khan, Javed Mohammed, and Shoba Ranganathan. 2010. “pDOCK: A New Technique for Rapid and Accurate Docking of Peptide Ligands to Major Histocompatibility Complexes.” *Immunome Research* 6 Suppl 1 (September): S2.

Kim, Yohan, John Sidney, Søren Buus, Alessandro Sette, Morten Nielsen, and Bjoern Peters. 2014. “Dataset Size and Composition Impact the Reliability of Performance Benchmarks for Peptide-MHC Binding Predictions.” *BMC Bioinformatics* 15 (July): 241.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980.

Kisselev, A. F., T. N. Akopian, K. M. Woo, and A. L. Goldberg. 1999. “The Sizes of Peptides Generated from Protein by Mammalian 26 and 20 S Proteasomes. Implications for Understanding the Degradative Mechanism and Antigen Presentation.” *The Journal of Biological Chemistry* 274 (6): 3363–71.

Klein, Ludger, Bruno Kyewski, Paul M. Allen, and Kristin A. Hogquist. 2014. “Positive and Negative Selection of the T Cell Repertoire: What Thymocytes See (and Don’t See).” *Nature Reviews. Immunology* 14 (6): 377–91.

Kodysh, Julia, and Alex Rubinsteyn. 2020. “OpenVax: An Open-Source Computational Pipeline for Cancer Neoantigen Prediction.” *Methods in Molecular Biology* 2120: 147–60.

Koşaloğlu-Yalçın, Zeynep, Manasa Lanka, Angela Frentzen, Ashmitaa Logandha Ramamoorthy Premlal, John Sidney, Kerrie Vaughan, Jason Greenbaum, et al. 2018. "Predicting T Cell Recognition of MHC Class I Restricted Neoepitopes." *Oncoimmunology* 7 (11): e1492508.

Krogh, Anders, and John A. Hertz. 1992. "A Simple Weight Decay Can Improve Generalization." In *Advances in Neural Information Processing Systems 4*, edited by J. E. Moody, S. J. Hanson, and R. P. Lippmann, 950–57. Morgan-Kaufmann.

Kyritsis, C., S. Gorbulev, S. Hutschenreiter, K. Pawlitschko, R. Abele, and R. Tampé. 2001. "Molecular Mechanism and Structural Aspects of Transporter Associated with Antigen Processing Inhibition by the Cytomegalovirus Protein US6." *The Journal of Biological Chemistry* 276 (51): 48031–39.

Larman, H. Benjamin, Zhenming Zhao, Uri Laserson, Mamie Z. Li, Alberto Ciccia, M. Angelica Martinez Gakidis, George M. Church, et al. 2011. "Autoantigen Discovery with a Synthetic Human Peptidome." *Nature Biotechnology* 29 (6): 535–41.

Larsen, Mette Voldby, Claus Lundegaard, Kasper Lamberth, Søren Buus, Søren Brunak, Ole Lund, and Morten Nielsen. 2005. "An Integrative Approach to CTL Epitope Prediction: A Combined Algorithm Integrating MHC Class I Binding, TAP Transport Efficiency, and Proteasomal Cleavage Predictions." *European Journal of Immunology* 35 (8): 2295–2303.

Lázaro, Silvia, David Gamarra, and Margarita Del Val. 2015. "Proteolytic Enzymes Involved in MHC Class I Antigen Processing: A Guerrilla Army That Partners with the Proteasome." *Molecular Immunology* 68 (2 Pt A): 72–76.

Le Bert, Nina, Anthony T. Tan, Kamini Kunasegaran, Christine Y. L. Tham, Morteza Hafezi, Adeline Chia, Melissa Hui Yen Chng, et al. 2020. "SARS-CoV-2-Specific T Cell Immunity in Cases of COVID-19 and SARS, and Uninfected Controls." *Nature*, July. https://doi.org/10.1038/s41586-020-2550-z.

Li, Guideng, Michael T. Bethune, Stephanie Wong, Alok V. Joglekar, Michael T. Leonard, Jessica K. Wang, Jocelyn T. Kim, et al. 2019. "T Cell Antigen Discovery via Trogocytosis." *Nature Methods* 16 (2): 183–90.

Liu, Xinqi, Shaodong Dai, Frances Crawford, Rachel Fruge, Philippa Marrack, and John Kappler. 2002. "Alternate Interactions Define the Binding of Peptides to the MHC Molecule IA(b)." *Proceedings of the National Academy of Sciences of the United States of America* 99 (13): 8820–25.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on mRNA Abundance." *Cell* 165 (3): 535–50.

Lo, Wan-Lin, Benjamin D. Solomon, David L. Donermeyer, Chyi-Song Hsieh, and Paul M. Allen. 2014. "T Cell Immunodominance Is Dictated by the Positively Selecting Self-Peptide." *eLife* 3 (January): e01457.

Łuksza, Marta, Nadeem Riaz, Vladimir Makarov, Vinod P. Balachandran, Matthew D. Hellmann, Alexander Solovyov, Naiyer A. Rizvi, et al. 2017. "A Neoantigen Fitness Model Predicts Tumour Response to Checkpoint Blockade Immunotherapy." *Nature* 551 (7681): 517–20.

Lundegaard, Claus, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. 2008. "NetMHC-3.0: Accurate Web Accessible Predictions of Human, Mouse and Monkey MHC

Class I Affinities for Peptides of Length 8–11." *Nucleic Acids Research* 36 (suppl_2): W509–12.

Lundegaard, Claus, Ole Lund, Can Keşmir, Søren Brunak, and Morten Nielsen. 2007. "Modeling the Adaptive Immune System: Predictions and Simulations." *Bioinformatics* 23 (24): 3265–75.

Maccari, Giuseppe, James Robinson, Keith Ballingall, Lisbeth A. Guethlein, Unni Grimholt, Jim Kaufman, Chak-Sum Ho, et al. 2017. "IPD-MHC 2.0: An Improved Inter-Species Database for the Study of the Major Histocompatibility Complex." *Nucleic Acids Research* 45 (D1): D860–64.

Madden, D. R., D. N. Garboczi, and D. C. Wiley. 1993. "The Antigenic Identity of Peptide-MHC Complexes: A Comparison of the Conformations of Five Viral Peptides Presented by HLA-A2." *Cell* 75 (4): 693–708.

Marcou, Quentin, Thierry Mora, and Aleksandra M. Walczak. 2018. "High-Throughput Immune Repertoire Analysis with IGoR." *Nature Communications* 9 (1): 561.

Ma, Xueying, Amparo Serna, Ren-Huan Xu, and Luis J. Sigal. 2009. "The Amino Acid Sequences Flanking an Antigenic Determinant Can Strongly Affect MHC Class I Cross-Presentation without Altering Direct Presentation." *Journal of Immunology* 182 (8): 4601–7.

McGranahan, Nicholas, Andrew J. S. Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, et al. 2016. "Clonal Neoantigens Elicit T Cell Immunoreactivity and Sensitivity to Immune Checkpoint Blockade." *Science* Accepted f (March): 1–11.

Mishkin, Dmytro, and Jiri Matas. 2015. "All You Need Is a Good Init." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1511.06422.

Mohammed, Fiyaz, Mark Cobbold, Angela L. Zarling, Mahboob Salim, Gregory A. Barrett-Wilt, Jeffrey Shabanowitz, Donald F. Hunt, Victor H. Engelhard, and Benjamin E. Willcox. 2008. "Phosphorylation-Dependent Interaction between Antigenic Peptides and MHC Class I: A Molecular Basis for the Presentation of Transformed Self." *Nature Immunology* 9 (11): 1236–43.

Mohammed, Fiyaz, Daniel H. Stones, Angela L. Zarling, Carrie R. Willcox, Jeffrey Shabanowitz, Kara L. Cummings, Donald F. Hunt, Mark Cobbold, Victor H. Engelhard, and Benjamin E. Willcox. 2017. "The Antigenic Identity of Human Class I MHC Phosphopeptides Is Critically Dependent upon Phosphorylation Status." *Oncotarget*. https://doi.org/10.18632/oncotarget.16952.

Mommen, Geert P. M., Christian K. Frese, Hugo D. Meiring, Jacqueline van Gaans-van den Brink, Ad P. J. M. de Jong, Cécile A. C. M. van Els, and Albert J. R. Heck. 2014. "Expanding the Detectable HLA Peptide Repertoire Using Electron-Transfer/higher-Energy Collision Dissociation (EThcD)." *Proceedings of the National Academy of Sciences of the United States of America* 111 (12): 4507–12.

Moutaftsi, Magdalini, Huynh-Hoa Bui, Bjoern Peters, John Sidney, Shahram Salek-Ardakani, Carla Oseroff, Valerie Pasquetto, et al. 2007. "Vaccinia Virus-Specific CD4+ T Cell Responses Target a Set of Antigens Largely Distinct from Those Targeted by CD8+ T Cell Responses." *Journal of Immunology* 178 (11): 6814–20.

Moutaftsi, Magdalini, Bjoern Peters, Valerie Pasquetto, David C. Tscharke, John Sidney, Huynh-Hoa Bui, Howard Grey, and Alessandro Sette. 2006. "A Consensus Epitope Prediction Approach

Identifies the Breadth of Murine T CD8+ -Cell Responses to Vaccinia Virus." *Nature Biotechnology* 24 (7): 817–19.

Murugan, Anand, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan Jr. 2012. "Statistical Inference of the Generation Probability of T-Cell Receptors from Sequence Repertoires." *Proceedings of the National Academy of Sciences of the United States of America* 109 (40): 16161–66.

Nielsen, Morten, and Massimo Andreatta. 2016. "NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets." *Genome Medicine* 8 (1). https://doi.org/10.1186/s13073-016-0288-x.

Nielsen, Morten, Claus Lundegaard, Ole Lund, and Can Keşmir. 2005. "The Role of the Proteasome in Generating Cytotoxic T-Cell Epitopes: Insights Obtained from Improved Predictions of Proteasomal Cleavage." *Immunogenetics* 57 (1-2): 33–41.

Nielsen, Morten, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. 2003. "Reliable Prediction of T-Cell Epitopes Using Neural Networks with Novel Sequence Representations." *Protein Science: A Publication of the Protein Society* 12 (5): 1007–17.

Nussbaum, Alexander K., Tobias P. Dick, Wieland Keilholz, Markus Schirle, Stefan Stevanović, Klaus Dietz, Wolfgang Heinemeyer, et al. 1998. "Cleavage Motifs of the Yeast 20S Proteasome β Subunits Deduced from Digests of Enolase 1." *Proceedings of the National Academy of Sciences of the United States of America* 95 (21): 12504–9.

O'Donnell, Timothy, Elizabeth L. Christie, Arun Ahuja, Jacqueline Buros, B. Arman Aksoy, David D. L. Bowtell, Alexandra Snyder, and Jeff Hammerbacher. 2018. "Chemotherapy Weakly Contributes to Predicted Neoantigen Expression in Ovarian Cancer." *BMC Cancer* 18 (1): 87.

O'Donnell, Timothy J., Alex Rubinsteyn, Maria Bonsack, Angelika B. Riemer, Uri Laserson, and Jeff Hammerbacher. 2018. "MHCflurry: Open-Source Class I MHC Binding Affinity Prediction." *Cell Systems*, 1–4.

O'Donnell, Timothy J., Alex Rubinsteyn, and Uri Laserson. 2020. "MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing." *Cell Systems*, July. https://doi.org/10.1016/j.cels.2020.06.010.

O'Donnell, Timothy, and Alex Rubinsteyn. 2020. "High-Throughput MHC I Ligand Prediction Using MHCflurry." In *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*, edited by Sebastian Boegel, 113–27. New York, NY: Springer US.

O'huigin, Colm, Smita Kulkarni, Yunping Xu, Zhihui Deng, Judith Kidd, Kenneth Kidd, Xiaojiang Gao, and Mary Carrington. 2011. "The Molecular Origin and Consequences of Escape from miRNA Regulation by HLA-C Alleles." *American Journal of Human Genetics* 89 (3): 424–31.

Oseroff, Carla, Ferdynand Kos, Huynh-Hoa Bui, Bjoern Peters, Valerie Pasquetto, Jean Glenn, Tara Palmore, et al. 2005. "HLA Class I-Restricted Responses to Vaccinia Recognize a Broad Array of Proteins Mainly Involved in Virulence and Viral Gene Regulation." *Proceedings of the National Academy of Sciences of the United States of America* 102 (39): 13980–85.

Osterbye, Thomas, Morten Nielsen, Nadine L. Dudek, Sri H. Ramarathinam, Anthony W. Purcell, Claus Schafer-Nielsen, and Soren Buus. 2020. "HLA Class II Specificity Assessed by High-Density Peptide Microarray Interactions." *Journal of Immunology* 205 (1): 290–99.

Ott, Patrick A., Zhuting Hu, Derin B. Keskin, Sachet A. Shukla, Jing Sun, David J. Bozym, Wandi Zhang, et al. 2017. "An Immunogenic Personal Neoantigen Vaccine for Patients with Melanoma." *Nature* 547 (7662): 217–21.

Pan, Ren-You, Wen-Hung Chung, Mu-Tzu Chu, Shu-Jen Chen, Hua-Chien Chen, Lei Zheng, and Shuen-Iu Hung. 2018. "Recent Development and Clinical Application of Cancer Vaccine: Targeting Neoantigens." *Journal of Immunology Research* 2018 (December): 4325874.

Papatheodorou, Irene, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A. Fonseca, et al. 2020. "Expression Atlas Update: From Tissues to Single Cells." *Nucleic Acids Research* 48 (D1): D77–83.

Parker, Kenneth C., Michael Shields, Marianne DiBrino, Andrew Brooks, and John E. Coligan. 1995. "Peptide Binding to MHC Class I Molecules: Implications for Antigenic Peptide Prediction." *Immunologic Research*. https://doi.org/10.1007/bf02918496.

Paul, Sinu, Edita Karosiene, Sandeep Kumar Dhanda, Vanessa Jurtz, Lindy Edwards, Morten Nielsen, Alessandro Sette, and Bjoern Peters. 2018. "Determination of a Predictive Cleavage Motif for Eluted Major Histocompatibility Complex Class II Ligands." *Frontiers in Immunology* 9 (August): 1795.

Pearson, Hillary, Tariq Daouda, Diana Paola Granados, Chantal Durette, Eric Bonneil, Mathieu Courcelles, Anja Rodenbrock, et al. 2016. "MHC Class I-Associated Peptides Derive from Selective Regions of the Human Genome." *The Journal of Clinical Investigation* 126 (12): 4690–4701.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.

Peters, Bjoern, Morten Nielsen, and Alessandro Sette. 2020. "T Cell Epitope Predictions." *Annual Review of Immunology*, February. https://doi.org/10.1146/annurev-immunol-082119-124838.

Peters, Bjoern, and Alessandro Sette. 2005. "Generating Quantitative Models Describing the Sequence Specificity of Biological Processes with the Stabilized Matrix Method." *BMC Bioinformatics* 6 (May): 132.

Peters, Björn, Sascha Bulik, Robert Tampe, Peter M. Van Endert, and Hermann-Georg Holzhütter. 2003. "Identifying MHC Class I Epitopes by Predicting the TAP Transport Efficiency of Epitope Precursors." *Journal of Immunology* 171 (4): 1741–49.

Peters, Björn, Weiwei Tong, John Sidney, Alessandro Sette, and Zhiping Weng. 2003. "Examining the Independent Binding Assumption for Binding of Peptide Epitopes to MHC-I Molecules." *Bioinformatics* 19 (14): 1765–72.

Purcell, Anthony W., Sri H. Ramarathinam, and Nicola Ternette. 2019. "Mass Spectrometry-Based Identification of MHC-Bound Peptides for Immunopeptidomics." *Nature Protocols* 14 (6): 1687–1707.

Qi, Qian, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A. Olshen, Cornelia M. Weyand, Scott D. Boyd, and Jörg J. Goronzy. 2014. "Diversity and Clonal Selection in the Human T-Cell Repertoire." *Proceedings of the National Academy of Sciences of the United States of America* 111 (36): 13139–44.

Racle, Julien, Justine Michaux, Georg Alexander Rockinger, Marion Arnaud, Sara Bobisse, Chloe Chong, Philippe Guillaume, et al. 2019. "Robust Prediction of HLA Class II Epitopes by Deep Motif Deconvolution of Immunopeptidomes." *Nature Biotechnology* 37 (11): 1283–86.

Raghavan, Malini, Sanjeeva J. Wijeyesakere, Larry Robert Peters, and Natasha Del Cid. 2013. "Calreticulin in the Immune System: Ins and Outs." *Trends in Immunology* 34 (1): 13–21.

Rammensee, H., J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. 1999. "SYFPEITHI: Database for MHC Ligands and Peptide Motifs." *Immunogenetics* 50 (3-4): 213–19.

Ramsuran, Veron, Smita Kulkarni, Colm O'huigin, Yuko Yuki, Danillo G. Augusto, Xiaojiang Gao, and Mary Carrington. 2015. "Epigenetic Regulation of Differential HLA-A Allelic Expression Levels." *Human Molecular Genetics* 24 (15): 4268–75.

Ramsuran, Veron, Vivek Naranbhai, Amir Horowitz, Ying Qi, Maureen P. Martin, Yuko Yuki, Xiaojiang Gao, et al. 2018. "Elevated HLA-A Expression Impairs HIV Control through Inhibition of NKG2A-Expressing Cells." *Science* 359 (6371): 86–90.

Reits, Eric, Alexander Griekspoor, Joost Neijssen, Tom Groothuis, Kees Jalink, Peter van Veelen, Hans Janssen, Jero Calafat, Jan Wouter Drijfhout, and Jacques Neefjes. 2003. "Peptide Diffusion, Protection, and Degradation in Nuclear and Cytoplasmic Compartments before Antigen Presentation by MHC Class I." *Immunity* 18 (1): 97–108.

Remesh, Soumya G., Massimo Andreatta, Ge Ying, Thomas Kaever, Morten Nielsen, Curtis McMurtrey, William Hildebrand, Bjoern Peters, and Dirk M. Zajonc. 2017. "Unconventional Peptide Presentation by Major Histocompatibility Complex (MHC) Class I Allele HLA-A*02:01." *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.m117.776542.

Ressing, Maaike E., Sinéad E. Keating, Daphne van Leeuwen, Danijela Koppers-Lalic, Isabel Y. Pappworth, Emmanuel J. H. J. Wiertz, and Martin Rowe. 2005. "Impaired Transporter Associated with Antigen Processing-Dependent Peptide Transport during Productive EBV Infection." *Journal of Immunology* 174 (11): 6829–38.

Reynisson, Birkir, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. 2020. "NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data." *Nucleic Acids Research*. https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa379/5837056.

Reynisson, Birkir, Carolina Barra, Saghar Kaabinejadian, William H. Hildebrand, Bjoern Peters, and Morten Nielsen. 2020. "Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data." *Journal of Proteome Research* 19 (6): 2304–15.

Ritz, Danilo, Andreas Gloger, Benjamin Weide, Claus Garbe, Dario Neri, and Tim Fugmann. 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 (10): 1570–80.

Robins, Harlan S., Paulo V. Campregher, Santosh K. Srivastava, Abigail Wacher, Cameron J. Turtle, Orsalem Kahsai, Stanley R. Riddell, Edus H. Warren, and Christopher S. Carlson. 2009. "Comprehensive Assessment of T-Cell Receptor β-Chain Diversity in αβ T Cells." *Blood, The Journal of the American Society of Hematology* 114 (19): 4099–4107.

Robinson, James, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G. E. Marsh. 2015. "The IPD and IMGT/HLA Database: Allele Variant Databases." *Nucleic Acids Research* 43 (Database issue): D423–31.

Rock, Kenneth L., Diego J. Farfán-Arribas, and Lianjun Shen. 2010. "Proteases in MHC Class I Presentation and Cross-Presentation." *Journal of Immunology* 184 (1): 9–15.

Rock, Kenneth L., Eric Reits, and Jacques Neefjes. 2016. "Present Yourself! By MHC Class I and MHC Class II Molecules." *Trends in Immunology* 37 (11): 724–37.

Rognan, D., L. Scapozza, G. Folkers, and A. Daser. 1994. "Molecular Dynamics Simulation of MHC-Peptide Complexes as a Tool for Predicting Potential T Cell Epitopes." *Biochemistry* 33 (38): 11476–85.

Rossjohn, Jamie, Stephanie Gras, John J. Miles, Stephen J. Turner, Dale I. Godfrey, and James McCluskey. 2015. "T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules." *Annual Review of Immunology* 33: 169–200.

Roudko, Vladimir, Benjamin Greenbaum, and Nina Bhardwaj. 2020. "Computational Prediction and Validation of Tumor-Associated Neoantigens." *Frontiers in Immunology* 11 (January): 27.

Ruppert, Jörg, John Sidney, Esteban Celis, Ralph T. Kubo, Howard M. Grey, and Alessandro Sette. 1993. "Prominent Role of Secondary Anchor Residues in Peptide Binding to HLA-A2.1 Molecules." *Cell*. https://doi.org/10.1016/0092-8674(93)90472-3.

Sahin, Ugur, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, et al. 2017. "Personalized RNA Mutanome Vaccines Mobilize Poly-Specific Therapeutic Immunity against Cancer." *Nature* 547 (7662): 222–26.

Sarkizova, Siranush, Susan Klaeger, Phuong M. Le, Letitia W. Li, Giacomo Oliveira, Hasmik Keshishian, Christina R. Hartigan, et al. 2019. "A Large Peptidome Dataset Improves HLA Class I Epitope Prediction across Most of the Human Population." *Nature Biotechnology*, December. https://doi.org/10.1038/s41587-019-0322-9.

Schatz, David G., and Yanhong Ji. 2011. "Recombination Centres and the Orchestration of V(D)J Recombination." *Nature Reviews. Immunology* 11 (4): 251–63.

Schueler-Furman, O., Y. Altuvia, A. Sette, and H. Margalit. 2000. "Structure-Based Prediction of Binding Peptides to MHC Class I Molecules: Application to a Broad Range of MHC Alleles." *Protein Science: A Publication of the Protein Society* 9 (9): 1838–46.

Schumacher, Ton N., Wouter Scheper, and Pia Kvistborg. 2019. "Cancer Neoantigens." *Annual Review of Immunology* 37 (April): 173–200.

Schumacher, Ton N., and Robert D. Schreiber. 2015. "Neoantigens in Cancer Immunotherapy." *Science* 348 (6230): 69–74.

Segal, Mark R., Michael P. Cummings, and Alan E. Hubbard. 2001. “Relating Amino Acid Sequence to Phenotype: Analysis of Peptide-Binding Data.” *Biometrics* 57 (2): 632–43.

Serwold, Thomas, Federico Gonzalez, Jennifer Kim, Richard Jacob, and Nilabh Shastri. 2002. “ERAAP Customizes Peptides for MHC Class I Molecules in the Endoplasmic Reticulum.” *Nature*. https://doi.org/10.1038/nature01074.

Sette, A., S. Buus, E. Appella, J. A. Smith, R. Chesnut, C. Miles, S. M. Colon, and H. M. Grey. 1989. “Prediction of Major Histocompatibility Complex Binding Regions of Protein Antigens by Sequence Pattern Analysis.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (9): 3296–3300.

Shao, Wenguang, Patrick G. A. Pedrioli, Witold Wolski, Cristian Scurtescu, Emanuel Schmid, Juan A. Vizcaíno, Mathieu Courcelles, et al. 2018. “The SysteMHC Atlas Project.” *Nucleic Acids Research* 46 (D1): D1237–47.

Sharma, Govinda, Craig M. Rive, and Robert A. Holt. 2019. “Rapid Selection and Identification of Functional CD8+ T Cell Epitopes from Large Peptide-Coding Libraries.” *Nature Communications* 10 (1): 4553.

Shastri, N., T. Serwold, and F. Gonzalez. 1995. “Presentation of Endogenous peptide/MHC Class I Complexes Is Profoundly Influenced by Specific C-Terminal Flanking Residues.” *Journal of Immunology* 155 (9): 4339–46.

Shiina, Takashi, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K. Kulski. 2009. “The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease.” *Journal of Human Genetics* 54 (1): 15–39.

Shraibman, Bracha, Eilon Barnea, Dganit Melamed Kadosh, Yael Haimovich, Gleb Slobodin, Itzhak Rosner, Carlos López-Larrea, et al. 2019. “Identification of Tumor Antigens Among the HLA Peptidomes of Glioblastoma Tumors and Plasma.” *Molecular & Cellular Proteomics: MCP* 18 (6): 1255–68.

Shraibman, Bracha, Dganit Melamed Kadosh, Eilon Barnea, and Arie Admon. 2016. “Human Leukocyte Antigen (HLA) Peptides Derived from Tumor Antigens Induced by Inhibition of DNA Methylation for Development of Drug-Facilitated Immunotherapy.” *Molecular & Cellular Proteomics: MCP* 15 (9): 3058–70.

Sidney, John, Scott Southwood, Carrie Moore, Carla Oseroff, Clemencia Pinilla, Howard M. Grey, and Alessandro Sette. 2013. “Measurement of MHC/peptide Interactions by Gel Filtration or Monoclonal Antibody Capture.” *Current Protocols in Immunology / Edited by John E. Coligan ... [et Al.]* Chapter 18 (February): Unit 18.3.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (1). https://www.embopress.org/doi/abs/10.1038/msb.2011.75.

Singer, Alfred, Stanley Adoro, and Jung-Hyun Park. 2008. “Lineage Fate and Intense Debate: Myths, Models and Mechanisms of CD4- versus CD8-Lineage Choice.” *Nature Reviews. Immunology* 8 (10): 788–801.

Smith, K. J., S. W. Reid, D. I. Stuart, A. J. McMichael, E. Y. Jones, and J. I. Bell. 1996. "An Altered Position of the Alpha 2 Helix of MHC Class I Is Revealed by the Crystal Structure of HLA-B*3501." *Immunity* 4 (3): 203–13.

Solleder, Marthe, Philippe Guillaume, Julien Racle, Justine Michaux, Hui-Song Pak, Markus Müller, George Coukos, Michal Bassani-Sternberg, and David Gfeller. 2020. "Mass Spectrometry Based Immunopeptidomics Leads to Robust Predictions of Phosphorylated HLA Class I Ligands." *Molecular & Cellular Proteomics: MCP* 19 (2): 390–404.

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout : A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research: JMLR* 15: 1929–58.

Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht. 1982. "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. Coli." *Nucleic Acids Research* 10 (9): 2997–3011.

Stranzl, Thomas, Mette Voldby Larsen, Claus Lundegaard, and Morten Nielsen. 2010. "NetCTLpan: Pan-Specific MHC Class I Pathway Epitope Predictions." *Immunogenetics* 62 (6): 357–68.

Strønen, Erlend, Mireille Toebes, Sander Kelderman, Marit M. van Buuren, Weiwen Yang, Nienke van Rooij, Marco Donia, et al. 2016. "Targeting of Cancer Neoantigens with Donor-Derived T Cell Receptor Repertoires." *Science* 352 (6291): 1337–41.

Tenzer, S., B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P-M Kloetzel, H-G Rammensee, H. Schild, and H-G Holzhütter. 2005. "Modeling the MHC Class I Pathway by Combining Predictions of Proteasomal Cleavage, TAP Transport and MHC Class I Binding." *Cellular and Molecular Life Sciences: CMLS* 62 (9): 1025–37.

Thomas, Christoph, and Robert Tampé. 2017. "Proofreading of Peptide-MHC Complexes through Dynamic Multivalent Interactions." *Frontiers in Immunology* 8 (February): 65.

Tomazin, R., A. B. Hill, P. Jugovic, I. York, P. van Endert, H. L. Ploegh, D. W. Andrews, and D. C. Johnson. 1996. "Stable Binding of the Herpes Simplex Virus ICP47 Protein to the Peptide Binding Site of TAP." *The EMBO Journal* 15 (13): 3256–66.

Tran, Eric, Mojgan Ahmadzadeh, Yong-Chen Lu, Alena Gros, Simon Turcotte, Paul F. Robbins, Jared J. Gartner, et al. 2015. "Immunogenicity of Somatic Mutations in Human Gastrointestinal Cancers." *Science*, no. October: 1–9.

Trolle, Thomas, Curtis P. McMurtrey, John Sidney, Wilfried Bardet, Sean C. Osborn, Thomas Kaever, Alessandro Sette, William H. Hildebrand, Morten Nielsen, and Bjoern Peters. 2016. "The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference." *Journal of Immunology* 196 (4): 1480–87.

Trolle, Thomas, Imir G. Metushi, Jason A. Greenbaum, Yohan Kim, John Sidney, Ole Lund, Alessandro Sette, Bjoern Peters, and Morten Nielsen. 2015. "Automated Benchmarking of Peptide-MHC Class I Binding Predictions." *Bioinformatics* 31 (13): 2174–81.

Tscharke, David C., Gunasegaran Karupiah, Jie Zhou, Tara Palmore, Kari R. Irvine, S. M. Mansour Haeryfar, Shanicka Williams, et al. 2005. "Identification of Poxvirus CD8 T Cell Determinants to Enable Rational Design and Characterization of Smallpox Vaccines." *Journal of Experimental*

Medicine. https://doi.org/10.1084/jem.20041912.

Uebel, S., W. Kraas, S. Kienle, K. H. Wiesmüller, G. Jung, and R. Tampé. 1997. "Recognition Principle of the TAP Transporter Disclosed by Combinatorial Peptide Libraries." *Proceedings of the National Academy of Sciences of the United States of America* 94 (17): 8976–81.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419.

UniProt Consortium. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47 (D1): D506–15.

Vita, Randi, Swapnil Mahajan, James A. Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R. Cantrell, Daniel K. Wheeler, Alessandro Sette, and Bjoern Peters. 2019. "The Immune Epitope Database (IEDB): 2018 Update." *Nucleic Acids Research* 47 (D1): D339–43.

Vita, Randi, James A. Overton, Jason A. Greenbaum, Julia Ponomarenko, Jason D. Clark, Jason R. Cantrell, Daniel K. Wheeler, et al. 2015. "The Immune Epitope Database (IEDB) 3.0." *Nucleic Acids Research* 43 (Database issue): D405–12.

Weiskopf, Daniela, Katharina S. Schmitz, Matthijs P. Raadsen, Alba Grifoni, Nisreen M. A. Okba, Henrik Endeman, Johannes P. C. van den Akker, et al. 2020. "Phenotype and Kinetics of SARS-CoV-2-Specific T Cells in COVID-19 Patients with Acute Respiratory Distress Syndrome." *Science Immunology*. https://doi.org/10.1126/sciimmunol.abd2071.

Wieczorek, Marek, Esam T. Abualrous, Jana Sticht, Miguel Álvaro-Benito, Sebastian Stolzenberg, Frank Noé, and Christian Freund. 2017. "Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation." *Frontiers in Immunology* 8 (March): 292.

Wolf-Levy, Hila, Aaron Javitt, Avital Eisenberg-Lerner, Assaf Kacen, Adi Ulman, Daoud Sheban, Bareket Dassa, et al. 2018. "Revealing the Cellular Degradome by Mass Spectrometry Analysis of Proteasome-Cleaved Peptides." *Nature Biotechnology*, October. https://doi.org/10.1038/nbt.4279.

Wolkers, Monika C., Nathalie Brouwenstijn, Arnold H. Bakker, Mireille Toebes, and Ton N. M. Schumacher. 2004. "Antigen Bias in T Cell Cross-Priming." *Science* 304 (5675): 1314–17.

Yagüe, J., J. Vázquez, and J. A. López de Castro. 1998. "A Single Amino Acid Change Makes the Peptide Specificity of B*3910 Unrelated to B*3901 and Closer to a Group of HLA-B Proteins Including the Malaria-Protecting Allotype HLA-B53." *Tissue Antigens* 52 (5): 416–21.

Yellen-Shaw, A. J., E. J. Wherry, G. C. Dubois, and L. C. Eisenlohr. 1997. "Point Mutation Flanking a CTL Epitope Ablates in Vitro and in Vivo Recognition of a Full-Length Viral Protein." *Journal of Immunology* 158 (7): 3227–34.

Yewdell, Jonathan W. 2006. "Confronting Complexity: Real-World Immunodominance in Antiviral CD8+ T Cell Responses." *Immunity* 25 (4): 533–43.

Yewdell, Jonathan W., and Jack R. Bennink. 1999. "Immunodominance in Major Histocompatibility Complex Class I--Restricted T Lymphocyte Responses." *Annual Review of Immunology* 17 (1): 51–

88.

York, Ian A., Annie X. Y. Mo, Kristen Lemerise, Wanyong Zeng, Yuelei Shen, Carmela R. Abraham, Tomo Saric, Alfred L. Goldberg, and Kenneth L. Rock. 2003. “The Cytosolic Endopeptidase, Thimet Oligopeptidase, Destroys Antigenic Peptides and Limits the Extent of MHC Class I Antigen Presentation.” *Immunity* 18 (3): 429–40.

Zarling, Angela L., Joy M. Polefrone, Anne M. Evans, Leann M. Mikesh, Jeffrey Shabanowitz, Sarah T. Lewis, Victor H. Engelhard, and Donald F. Hunt. 2006. “Identification of Class I MHC-Associated Phosphopeptides as Targets for Cancer Immunotherapy.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (40): 14889–94.

Zinkernagel, R. M., and P. C. Doherty. 1974. “Restriction of in Vitro T Cell-Mediated Cytotoxicity in Lymphocytic Choriomeningitis within a Syngeneic or Semiallogeneic System.” *Nature* 248 (5450): 701–2.

———. 1997. “The Discovery of MHC Restriction.” *Immunology Today* 18 (1): 14–17.