

# High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets

Xiaoshan M. Shao<sup>1,2</sup>, Rohit Bhattacharya<sup>1,3</sup>, Justin Huang<sup>1,3</sup>, I.K. Ashok Sivakumar<sup>1,3,4</sup>, Collin Tokheim<sup>1,2</sup>, Lily Zheng<sup>1,5</sup>, Dylan Hirsch<sup>1,2</sup>, Benjamin Kaminow<sup>1,6</sup>, Ashton Omdahl<sup>1,2</sup>, Maria Bonsack<sup>7,8,9</sup>, Angelika B. Riemer<sup>7,8</sup>, Victor E. Velculescu<sup>1,5,10</sup>, Valsamo Anagnostou<sup>10</sup>, Kymberleigh A. Pagel<sup>1,2</sup>, and Rachel Karchin<sup>1,2,10</sup>

## ABSTRACT

Computational prediction of binding between neoantigen peptides and major histocompatibility complex (MHC) proteins can be used to predict patient response to cancer immunotherapy. Current neoantigen predictors focus on *in silico* estimation of MHC binding affinity and are limited by low predictive value for actual peptide presentation, inadequate support for rare MHC alleles, and poor scalability to high-throughput data sets. To address these limitations, we developed MHCnuggets, a deep neural network method that predicts peptide–MHC binding. MHCnuggets can predict binding for common or rare alleles of MHC class I or II with a single neural network architecture. Using a long short-term memory network (LSTM), MHCnuggets accepts peptides of variable length and is faster than other methods. When compared with methods that integrate binding affinity and MHC-bound peptide

(HLAp) data from mass spectrometry, MHCnuggets yields a 4-fold increase in positive predictive value on independent HLAp data. We applied MHCnuggets to 26 cancer types in The Cancer Genome Atlas, processing 26.3 million allele–peptide comparisons in under 2.3 hours, yielding 101,326 unique predicted immunogenic missense mutations (IMM). Predicted IMM hotspots occurred in 38 genes, including 24 driver genes. Predicted IMM load was significantly associated with increased immune cell infiltration ( $P < 2 \times 10^{-16}$ ), including CD8<sup>+</sup> T cells. Only 0.16% of predicted IMMs were observed in more than 2 patients, with 61.7% of these derived from driver mutations. Thus, we describe a method for neoantigen prediction and its performance characteristics and demonstrate its utility in data sets representing multiple human cancers.

## Introduction

The presentation of peptides bound to major histocompatibility complex (MHC) proteins on the surface of antigen-presenting cells and subsequent recognition by T-cell receptors is fundamental to the mammalian adaptive immune system. Neoantigens derived from somatic mutations are targets of immunoediting

and drive therapeutic responses in cancer patients treated with immunotherapy (1, 2). Because experimental characterization of neoantigens is both costly and time-consuming, computational methods have been developed to predict peptide–MHC binding and subsequent immune response (3, 4). Supervised neural network machine learning approaches have performed the best (5–7) and are the most widely used *in silico* methods. Despite advances in computational approaches, improvements in predictive performance have been minimal, due in part to a lack of sufficiently large sets of experimentally characterized peptide binding affinities for most MHC alleles.

Although neoantigen prediction for common MHC class I alleles is well studied (8), predictive accuracy on rare and less-characterized MHC alleles remains poor (9, 10). Class II predictors are scarce (11). Current estimates suggest that class II antigen lengths primarily range from 13 to 25 amino acids (12), and this diversity has been an obstacle to developing *in silico* neoantigen predictors (11, 13). As most neural network architectures are designed for fixed-length inputs, methods such as NetMHC (14–17) and MHCflurry (18) require preprocessing of peptide sequences or training of separate classifiers for each peptide length.

Clinical application of MHC–peptide binding predictors, to identify biomarkers for cancer immunotherapy, requires scalability to large patient cohorts and low false-positive rates (19). A cancer may contain hundreds of candidate somatically altered peptides, but few will actually bind to MHC proteins and elicit an immune response (20). For many years, most neoantigen predictors were trained primarily on quantitative peptide–MHC binding affinity data from *in vitro* experiments (21). Advances in immunopeptidomics technologies have enabled the identification of thousands of naturally presented MHC-bound peptides (HLAp) from cancer patient samples and cell lines (19, 22). Several neoantigen predictors

<sup>1</sup>Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland. <sup>2</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland. <sup>3</sup>Department of Computer Science, Johns Hopkins University, Baltimore, Maryland. <sup>4</sup>Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland. <sup>5</sup>McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland. <sup>6</sup>Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland. <sup>7</sup>Immunotherapy and Immunoprevention, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>8</sup>Molecular Vaccine Design, German Center for Infection Research (DZIF), partner site Heidelberg, Heidelberg, Germany. <sup>9</sup>Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>10</sup>The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland.

**Note:** Supplementary data for this article are available at Cancer Immunology Research Online (<http://cancerimmunolres.aacrjournals.org/>).

X.M. Shao, R. Bhattacharya, J. Huang, and I.K.A. Sivakumar contributed equally to this article.

**Corresponding Author:** Rachel Karchin, Johns Hopkins University, 316 Hackerman Hall, 3400 North Charles Street, Baltimore, MD 21204. Phone: 410-516-5578; Fax: 410-516-5294; E-mail: [karchin@jhu.edu](mailto:karchin@jhu.edu)

Cancer Immunol Res 2020;8:396–408

doi: 10.1158/2326-6066.CIR-19-0464

©2019 American Association for Cancer Research.

are trained only on HLAp data for class I, and only for a limited number of peptide lengths (21, 23). The EDGE neural network is trained primarily on multiallelic HLAp and RNA sequencing (RNA-seq) data from 74 cancer patients; ForestMHC is a random forest trained on HLAp from publicly available monoallelic and deconvoluted multiallelic cell lines. The potential to improve neoantigen predictors by integrating binding affinity and HLAp data (19) has motivated hybrid approaches (14, 18). However, most methods predict more candidate neoantigens than are actually immunogenic in patients (11, 19).

Here, we present a long short-term memory (LSTM) neural network method, MHCnuggets, a neoantigen predictor designed for MHC class I and II alleles in a single framework. The method leverages transfer learning and allele clustering to accommodate both common, well-characterized MHC alleles and rare, less-studied alleles. Although existing computational neoantigen predictors generate a ranked list of candidate peptides, maximizing the number of predictions that identify immunogenic peptides would be preferred in many applications (18). We demonstrate that MHCnuggets' predictive performance is competitive with widely used methods on binding affinity benchmark data sets. In comparison with hybrid methods that integrate binding affinity and HLAp data, MHCnuggets shows fewer false positives and increased positive predictive value (PPV) in a held-out cell line data set of ligands identified by mass spectrometry (7, 22). To demonstrate the clinical utility and applicability of MHCnuggets to large patient cohorts, we investigated candidate immunogenic mutations from 26 tumor types in The Cancer Genome Atlas (TCGA). MHCnuggets yielded 101,326 predicted immunogenic missense mutations (IMM), observed in at least 1 individual (out of 1,124,266) in less than 2.3 hours. These mutations were correlated with increased lymphocyte infiltration; however, only 0.16% were observed in more than 2 patients.

## Materials and Methods

### Implementation

MHCnuggets uses an LSTM neural network architecture (ref. 24; Fig. 1A). LSTM architectures excel at handling variable length sequence inputs and can learn long-term dependencies between noncontiguous elements, enabling an input encoding that does not require peptide shortening or splitting (Fig. 1B). LSTMs are capable of handling peptides of any length. In practice, a maximum peptide length should be selected for network training. We set maximum peptide input length of 15 for class I and 30 for class II, for computational efficiency purposes. These values cover the majority of lengths observed in naturally presented MHC-bound peptides (12). The networks were trained with transfer learning (25), which allows networks for less well-characterized alleles to leverage information from extensively studied alleles (Fig. 1C). Transfer learning was also used to train networks combining binding affinity and HLAp data sets. In addition, MHCnuggets architectures can be trained using either continuous binding affinity measurements from *in vitro* experiments (half maximal affinity or  $IC_{50}$ ) and/or immunopeptidomic (HLAp) binary labels. The former utilizes a mean-squared error (MSE) loss, whereas the latter utilizes binary cross-entropy (BCE) loss for training. For each MHC allele, we trained a neural network model consisting of an LSTM layer of 64 hidden units, a fully connected layer of 64 hidden units, and a final output layer of a single sigmoid unit (Fig. 1A).

For the 16 alleles where allele-specific HLAp training data were available (26), we trained networks on both binding affinity and HLAp data (MHCnuggets). Next, we trained networks only with binding affinity measurements (MHCnuggets without mass spectrometry data or noMS) for all MHC class I alleles. Due to the lack of allele-specific HLAp training data for class II, all MHC class II networks were trained only on binding affinity measurements. In total, we trained 148 class I and 136 class II allele-specific networks. Common alleles comprise a small fraction (<1%) of all known MHC alleles (27). To handle binding predictions for rare alleles, MHCnuggets selects a network by searching for the closest allele, based on previously published supertype clustering approaches. We prioritized approaches based on binding pocket biochemical similarity when available. Briefly, HLA-A and HLA-B alleles were clustered by MHC binding pocket amino acid residue composition (28), and HLA-C and all MHC II alleles were hierarchically clustered based upon experimental mass spectrometry and binding assay results (29, 30). For alleles with no supertype classification, the closest allele was from the same HLA gene, and allele group if available, with preference for alleles with the largest number of characterized binding peptides. All networks were implemented with the Keras Python package (TensorFlow back-end; refs. 31, 32). Open-source software is available at <https://github.com/KarchinLab/mhcnuggets>, installable via pip or Docker, and has been integrated into the PepVacSeq (33), pvactools (34), and Neoepi-scope (35) pipelines.

### Transformation of peptide binding affinities

Predicted binding affinity can be transformed into a range of values well suited for neural network learning by selecting a logarithmic base to match the weakest binding affinity of interest (36). For most benchmarks in this work, we used the standard upper limit of 50,000 nmol/L, so that predicted binding affinity was

$$y = \max(0, 1 - \log_{50k}(IC_{50}))$$

For the Bonsack and colleagues data set (8), the upper limit was changed to 100,000 nmol/L because in their experiments, as described in O'Donnell and colleagues (18), binders were defined as peptides with  $IC_{50} < 100,000$  nmol/L. As binding affinity was determined based on *in vitro* HLA binding-competition versus a known strong binder (reported  $IC_{50} < 50$  nmol/L) experimental  $IC_{50}$  values were in the micromolar range.

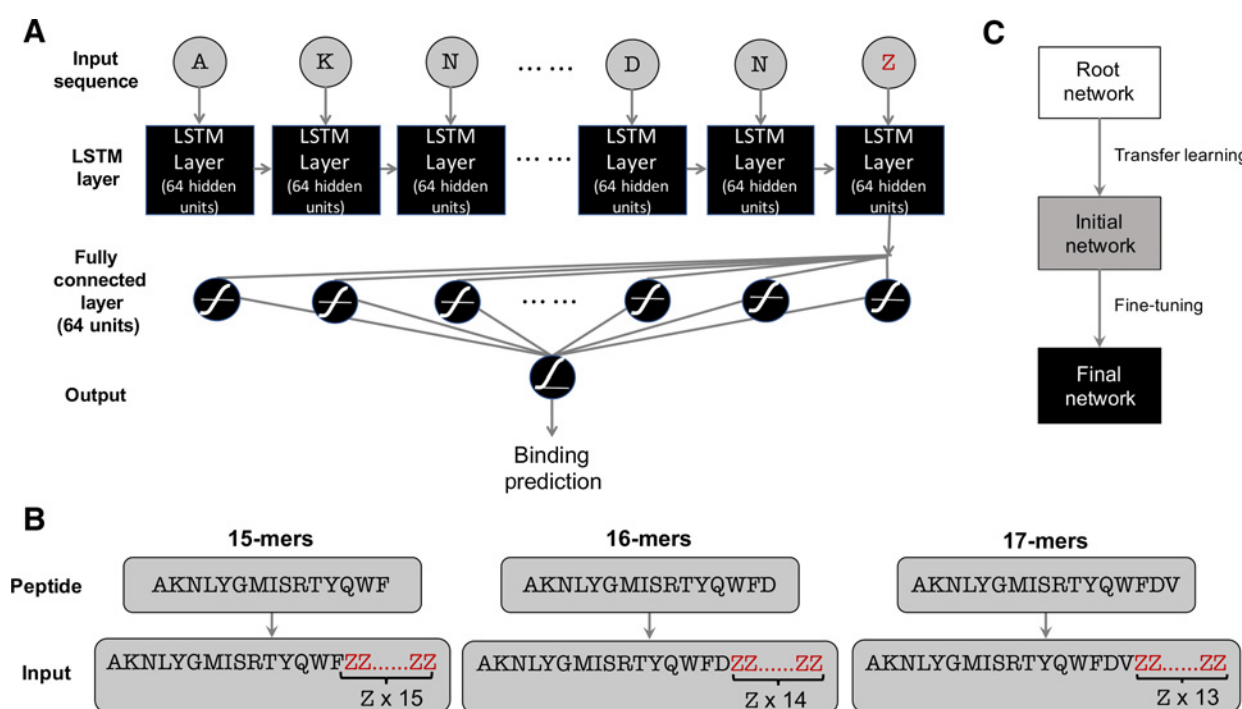
### Performance metrics

$PPV = NTP/(NTP + NFP)$ , where NTP is the number of true positives and NFP is the number of false positives. We calculated PPV with respect to the top-ranked  $n$  peptides, where  $n$  is the number of true binders in the ranked list, denoted as  $PPV_n$ . For the Bassani-Sternberg/Trolle (BST) benchmark, we also calculated PPV over the top 50- and 500-ranked peptides.

### Selection of final network weights

To minimize overfitting, network training was stopped after 100 epochs but if the best  $PPV_n$  was reached earlier, network weights from that earlier epoch were used in the final network. Notably, although we chose to optimize the networks on  $PPV_n$ , an alternative approach could optimize on area under the ROC curve (auROC), Kendall's  $\tau$ , or Pearson  $r$  correlation. For the two alleles in the Immune Epitope Database (IEDB) with the most training examples in their respective

Shao et al.

**Figure 1.**

**A**, MHCnuggets' architecture. A network is trained for each MHC allele. Each network has an LSTM layer with 64 hidden units, a fully connected layer with 64 hidden units, and a final output layer of a single sigmoid unit. **B**, Input scheme for peptides with variable lengths. MHCnuggets' architecture is capable of handling peptides of any length, but in practice, a maximum length should be selected. Peptides are extended with padding until they reach the maximum length, prior to input into the neural network. The example shows padding for class II peptides with maximum length set to 30 amino acids. **C**, Transfer learning protocol for parameter sharing among alleles. A base allele-specific network is trained for each MHC class, with an allele selected by largest number of training examples. Transfer learning is applied to train networks for the remaining alleles, with initial network weights set to final base network weights. A fine-tuning step identifies alleles that can be leveraged for a second round of transfer learning to produce a final network (Materials and Methods).

class, HLA-A\*02:01 for class I and HLA-DRB1\*01:01 for class II, training was stopped after 200 epochs.

### Network training

MSE loss  $L_{\text{MSE}}$  was used to train networks with continuous-valued binding affinity data and BCE loss  $L_{\text{BCE}}$  for binary HLAp data. For a data set with  $n$  samples,

$$L_{\text{MSE}}(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

$$L_{\text{BCE}}(\hat{y}, y) = -\frac{1}{n} \sum_{i=1}^n y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

All training used backpropagation with the Adam optimizer (37) and learning rate of 0.001. Regularization was performed with dropout and recurrent dropout (38) probabilities of 0.2. The number of hidden units, dropout rate, and number of training epochs were estimated by 3-fold cross-validation on MHC class I A\*02:01, a common allele with a large number of experimentally characterized binding peptides.

### One-hot encoding

Peptides were represented to the network as a series of amino acids; each amino acid was represented as a 21-dimensional smoothed, one-hot encoded vector (0.9 and 0.005 replace 1 and 0, respectively).

### Peptide padding

MHCnuggets' architecture is capable of handling peptides of any length, but in practice a maximum length should be selected, which in this work was 15 for class I and 30 for class II). Peptides that are less than the maximum length are padded at the end with a character ("Z" which is not in the amino acid alphabet) until they reach the maximum length.

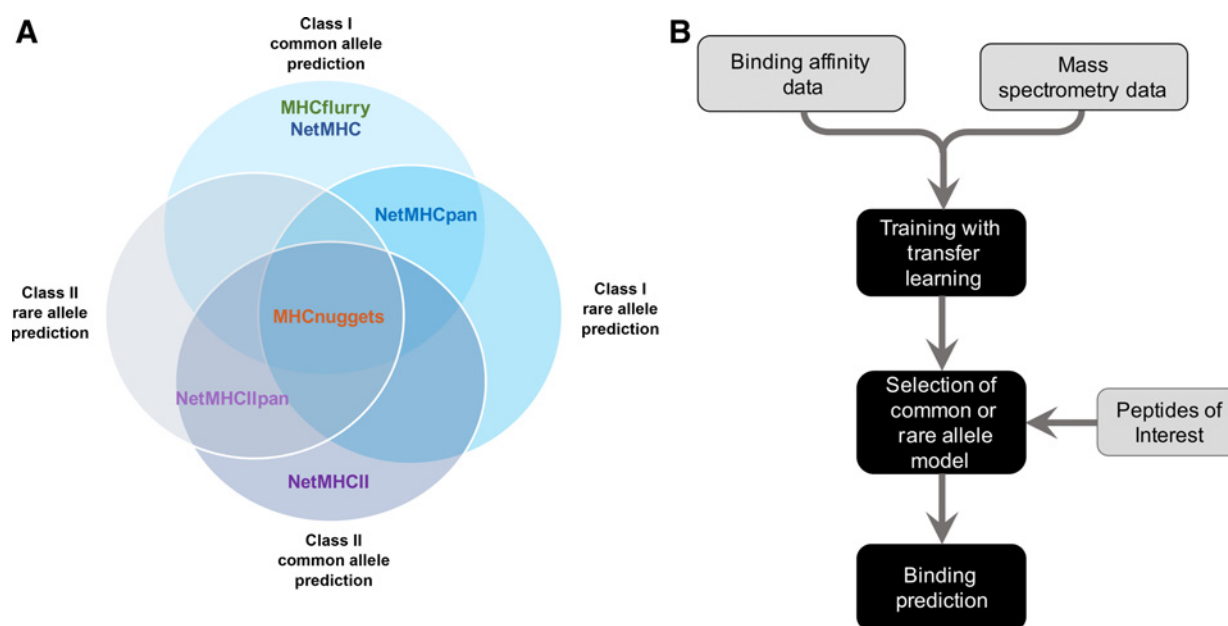
### Transfer learning protocol for binding affinity data only

We used transfer learning to improve network learning for MHC alleles with limited characterized peptides available for training. We first trained base allele-specific networks for class I and class II, using alleles with the most training examples in IEDB (HLA-A\*02:01 for class I and HLA-DRB1\*01:01 for class II). For all other alleles, the final weights of the base network for its respective class were used to initialize network training, and then an allele-specific network was trained for each allele. Next, we assessed prediction performance of each allele-specific network on the training examples for each of the alleles. For each allele, if the network that performed best was not the HLA-A\*02:01 network (for class I alleles) or HLA-DRB1\*01:01 network (for class II alleles), we did a second round of training, with the best-performing network's weights used in the initialization step.

### Transfer learning protocol for binding affinity and HLAp data

To integrate HLAp data into the class I networks, we initially trained each network with binding affinity data as described above, transferred

## High-Throughput Prediction of Neoantigens with MHCnuggets

**Figure 2.**

MHCnuggets' features. **A**, Venn diagram representation of the MHC-peptide binding prediction functions of MHCnuggets and similar tools. **B**, Training and MHC allele model selection scheme for MHCnuggets.

the final weights to a new network, and then continued training with the HLAp data as positive examples augmented with random peptide decoys as negative examples.

### Performance assessment

To accurately assess the performance of MHCnuggets on a variety of MHC-peptide binding prediction tasks, we utilized six benchmark sets: MHC class I alleles, MHC class II alleles, common alleles with a trained model (allele-specific prediction), and rare alleles (pan-allele prediction; Fig. 2; Supplementary Table S1). To compare to the HLA ligand prediction tools from the NetMHC group (NetMHC 3.0, NetMHC 4.0, NetMHCpan 2.0, NetMHCpan 4.0; refs. 16, 17), which can be trained only by their developers, as well as the open-source MHCflurry tools (18), we used multiple benchmarking strategies: (i) independent benchmark test set of peptides not included as training data for any of the methods; (ii) a previously published paired training/testing benchmark; (iii) 5-fold cross-validation benchmark; and (iv) leave-one-molecule-out (LOMO) benchmark.

We evaluated six MHC class I predictors on independent binding affinity and HLAp data sets (7, 8, 22). First, we compared MHCnuggets to several class I predictors that incorporate both binding affinity and HLAp data: MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. Each method was benchmarked using an independent set of MHC-bound peptides identified by mass spectrometry across seven cell lines for six MHC I alleles. For testing, HLAp hits were combined with random decoy peptides sampled from the human proteome in a 1:999 hit-decoy ratio, as described by Abelin and colleagues (26), totaling 23,971,000 peptides. Next, four MHC class I predictors trained only on binding affinity data [MHCnuggets (noMS) and MHCflurry (noMS), NetMHC 3.0 and NetMHCpan 2.0] were evaluated with the Kim and colleagues data set (5), in which each predictor was trained with the BD2009 data and tested on BLIND data. It was possible to compare NetMHC 3.0 and NetMHCpan

2.0 performance on Kim and colleagues, because they have previously published predicted  $IC_{50}$  values for all peptide-MHC pairs in BLIND. This allowed us to calculate their PPV<sub>n</sub>, auROC, Kendall's  $\tau$ , and Pearson  $r$  correlations.

Next, we compared MHCnuggets' class II ligand prediction performance with self-reported performance statistics of NetMHC group's MHC class II methods (39). We used the Jensen and colleagues 5-fold cross-validation benchmark to assess allele-specific MHC class II prediction of MHCnuggets and NetMHCII 2.3, for 27 alleles. NetMHCII 2.3 reported the average auROC for five-fold cross-validation, and we report MHCnuggets' PPV for each of the 27 alleles as well as the average auROC, Pearson  $r$ , and Kendall's  $\tau$  correlations.

The LOMO benchmarks are a type of cross-validation designed to estimate the performance of peptide binding prediction with respect to rare MHC alleles. Given training data for  $n$  MHC alleles, the data for a single allele are held out and networks are trained for the remaining  $n - 1$  alleles. Then for each peptide, predictions are generated by the remaining networks. We designed a LOMO benchmark to evaluate MHC class I rare allele prediction, by selecting 20 alleles with 30 to 100 characterized peptides in the IEDB (40). For class II rare allele prediction, we used the Jensen and colleagues (39) LOMO benchmark. We were unable to assess rare allele prediction for NetMHC class I methods, as no published results were available. For the NetMHC class II methods, we compared MHCnuggets to their self-reported auROCs.

### Data set collection and curation

Data sources for network training and testing, TCGA somatic mutations, TCGA tumor gene expression, and haplotype calling are shown in Supplementary Tables S1 and S2. A curated version of the IEDB 2018 (40) and the 16 class I monoallelic B-cell line immunopeptidomes (26) was provided by Tim O'Donnell (<https://data.men.deley.com/datasets/8pz43nvvxh/2>), binding affinity assays of human



Shao et al.

papillomavirus (HPV)-derived peptides were provided by Maria Bonsack and Angelika Riemer (8), BST = immunopeptidomes from six cell lines with multiallelic MHCs (26 MHC class I alleles; ref. 22) and from soluble HLA(sHLA)-transfected HeLa cells separated by allele (4 MHC class I alleles; ref. 7). Decoy random peptides were sampled from the human proteome.

### Training data sets

The networks were trained with data from a curated version of the IEDB (2018; ref. 40), containing chemical binding affinity measurements for 241,553 peptide-allele pairs covering 217 class I alleles and 96,211 peptide-allele pairs covering 135 class II alleles. Additional training data consisted of 16 class I mon-allelic B-cell line immunopeptidomes (26). The immunopeptidome data are limited to HLAp binders and were supplemented by decoy random peptides sampled from the human proteome (<https://data.mendeley.com/datasets/8pz43nvvxh/2>).

### Benchmark data sets

*Kim and colleagues:* This benchmark contained 53 MHC class I alleles and 137,654  $IC_{50}$  measurements published prior to 2009 (training set) and 53 unique MHC class I alleles with 26,888  $IC_{50}$  measurements, published from 2009 to 2013 (test set). Three alleles (HLA-B\*27:03, HLA-B\*38:01, and HLA-B\*08:03) did not contain sufficient training data, and two alleles (HLA-A\*46:01 and HLA-B\*27:03) did not contain any peptides defined as binders in this work ( $IC_{50} < 500$  nmol/L). Therefore, a total of four alleles (HLA-A\*46:01, HLA-B\*27:03, HLA-B\*38:01, and HLA-B\*08:03) were dropped from the analysis. All peptides in this benchmark set consisted of 8 to 11 amino acid residues.

*Bonsack and colleagues:* This data set contains 475 synthetic peptides derived from model protein sequences HPV16 E6 and E7 tested for binding to 7 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, and HLA-B\*15:01). Each peptide was tested in competition-based cellular binding assays with a known high-affinity fluorescein-labeled reference peptide. EBV-transformed B-lymphoblastic cells were stripped of their naturally bound peptides and mixed with serially diluted test peptides and 150 nmol/L of reference peptide. Each synthetic peptide was tested at 8 different concentrations ranging from 780 nmol/L to 100,000 nmol/L. Mixture fluorescence at each synthetic peptide concentration was measured with flow cytometry, and a nonlinear regression analysis was used to find the test peptide concentration that inhibited 50% of the reference peptide binding ( $IC_{50}$ ). Peptides were classified as binders ( $IC_{50} \leq 100,000$  nmol/L) or nonbinders ( $IC_{50} > 100,000$  nmol/L). Peptides in this independent benchmark set do not have IEDB entries.

*Bassani-Sternberg and colleagues, 2015:* This data set contains 22,598 unique peptides eluted from 6 cell lines with multiallelic MHCs. Out of the total 6 cell lines, a total of 26 alleles were reported. For each multiallelic cell line, peptide/MHC pairs were found through deconvolution, following the protocol described by (26), with the difference that we used MHCnuggets rather than NetMHCpan 2.8 (41) to predict  $IC_{50}$  values for each peptide-MHC pair. For each cell line, each peptide was initially assigned as a binder to all expressed alleles. Then, for each allele, we filtered out any peptide predicted to bind with  $IC_{50} > 1,000$  nmol/L to that allele, and with  $IC_{50} < 150$  nmol/L to any other allele. Peptides found for 6 alleles (HLA-A\*01:01, HLA-A\*02:01, HLA-A\*03:01,

HLA-A\*24:02, HLA-A\*31:01, and HLA-B\*51:01) were selected for allele-specific prediction testing. Trained networks were available for these alleles from all the methods that we compared.

*Trolle and colleagues:* This data set contains 15,524 unique peptides eluted from soluble HLA (sHLA) transfected HeLa cells, a process that allowed for separating binding peptides to a single MHC allele. This data set reports peptides for 5 MHC alleles. Peptides found for 4 alleles (HLA-A\*01:01, HLA-A\*02:01, HLA-A\*24:02, and HLA-B\*51:01) were selected for testing. Peptide lengths in this data set range from 8 to 15 amino acid residues.

*BST:* This benchmark consists of 23,971 HLAp hits for 6 alleles, from Bassani-Sternberg and colleagues (22) and Trolle and colleagues (42) plus 23,947,029 random decoy peptides sampled from the human proteome. Any peptides found to overlap with the training HLAp data (26) were removed.

*Jensen and colleagues:* This benchmark was designed to assess both allele-specific and rare MHC class II binding affinity predictors. Allele-specific prediction was tested with a 5-fold cross-validation experiment on peptides found in IEDB in 2016 but not 2013. Rare allele predictions were tested with the LOMO protocol.

*IEDB class I rare alleles:* This data set was designed to apply the LOMO protocol to class I alleles. It included 20 "pseudo-rare" alleles with 30 to 100 binding affinity peptide measurements in IEDB.

All data sets used in this work are available at <http://dx.doi.org/10.17632/8c26kkrfpr.2>.

### TCGA analysis pipeline

To assess candidate immunogenic somatic mutations in patients from the TCGA cohort, we developed and implemented a basic pipeline based on whole-exome and RNA-seq data (Supplementary Fig. S1). Our analysis builds upon work from the TCGA PanCancer Analysis teams for drivers (43), mutation calling (44), and cancer immune landscapes (45). We obtained somatic mutation calls for all cancer types from Multicenter Mutation Calling in Multiple Cancers (MC3; v0.2.8; 7,775 patients). Tumor-specific RNA expression values from Broad TCGA Firehose were standardized across tumor types using the RSEM Z-score (46). MHC allele calls were obtained from the TCGA cancer immune landscape publication, in which up to 6 MHC class I alleles (HLA-A, HLA-B, and HLA-C) were identified for each patient using OptiType (ref. 47; Supplementary Table S2). We included patients for which mutation calls, MHC allele calls, and RNA expression values were available from TCGA. After these considerations, the analysis included 6,613 patients from 26 TCGA tumor types. Six cancer types were not included in our analysis, because 15 or fewer patients met this requirement: lymphoid neoplasm diffuse large B-cell lymphoma, esophageal carcinoma, mesothelioma, skin cutaneous melanoma, stomach adenocarcinoma, and ovarian serous cystadenocarcinoma.

The somatic missense mutations identified in each patient were filtered to include only those with strong evidence of mutant gene RNA expression in that patient ( $Z > 1.0$ ). For each mutation that passed this filter, we used the transcript assigned by MC3 to pull flanking amino acid residues from the SwissProt database (48), yielding a 21 amino acid residue sequence fragment centered at the mutated residue. All candidate peptides of length 8, 9, 10, and 11 that included the mutated residue were extracted from each sequence fragment. Next binding

affinity predictions were generated for each mutated peptide for up to six MHC class I alleles, depending on the patient's HLA genotypes. In total, each somatic mutation was represented by 38 mutated peptides for up to six possible MHC pairings.

We applied a permissive filter to select candidate immunogenic peptides, requiring mutated peptides to have binding affinity of  $IC_{50} < 500$  nmol/L for at least one MHC allele. Somatic missense mutations that generated neoantigens meeting these criteria were considered predicted IMMs. For a given patient, if a mutation was predicted to be a predicted IMM for multiple alleles, it was counted only once using the MHC allele with the lowest predicted  $IC_{50}$ . Finally, for each patient, we counted the number of predicted IMMs found in their exome and stratified by tumor type. We then identified predicted IMMs that were harbored by more than 1 patient.

We sought to ascertain whether predicted IMMs occurred preferentially in particular gene or protein regions. Using the HotMAPS 1D algorithm v1.2.2 (49), we clustered primary amino acid residue sequence to identify regions where mutations were frequently predicted as IMM, with statistical significance ( $q < 0.01$ , Benjamini-Hochberg method; ref. 50). In this analysis, mutations were stratified by cancer type, and we considered enrichment within linear regions of 50 amino acid residues.

We considered that mutation immunogenicity might be associated with potential driver status of a mutation. Driver status was inferred by CHASMplus (51), a random forest classifier that utilizes a multifaceted feature set to predict driver missense mutations and is effective at identifying both common and rare driver mutations. For each mutation, immunogenicity was represented as a binary response variable

and driver status was used as a covariate. Mutations with CHASMplus  $q$ -value  $< 0.01$  were considered drivers (51). We modeled the relationship with univariate logistic regression (R glm package with binomial link logit function).

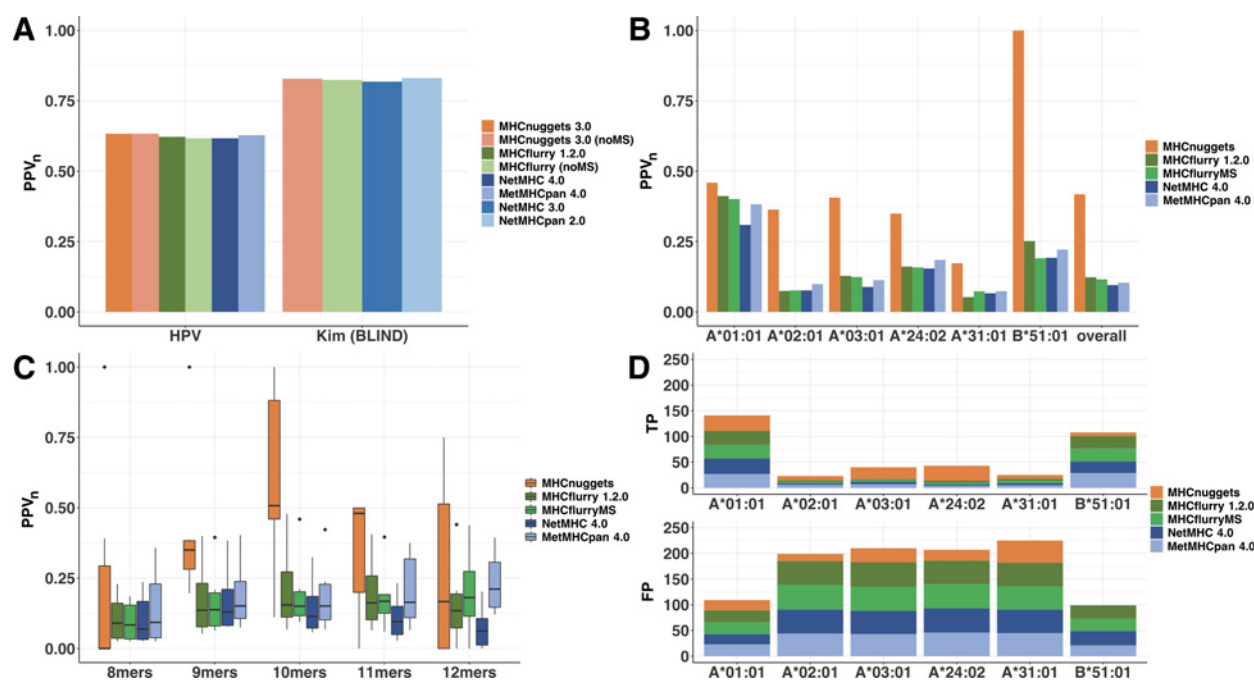
To assess whether the total number of predicted IMMs per patient was associated with changes in tumor immune infiltrates, we performed Poisson regression (R glm package with Poisson link log function). All estimates of immune infiltrates were obtained from Thorsson and colleagues (45, 52). We fit two univariate models in which the response variable was the predicted IMM count and the covariate was either total leukocyte fraction or fraction of  $CD8^+$  T cells.

### MC3 mutation filtering

MC3 TCGA somatic mutation calls were filtered for missense mutations.

### Regression models

We applied two univariate Poisson regression models. In the first model, each patient's predicted IMM load was the response variable and the independent variable  $X$  was the total leukocyte fraction. The fitted coefficient  $\beta = 0.75$  ( $P < 2 \times 10^{-16}$ , Wald test) indicated that increased predicted IMM load was significantly associated with increased leukocyte fraction in a patient's cancer. In a second model,  $X$  was the proportion of  $CD8^+$  T cells inferred by CIBERSORT (53). The fitted coefficient  $\beta = 5.9$  ( $P < 2 \times 10^{-16}$ , Wald test) indicated that increased predicted IMM load was associated with increased tumor-infiltrating  $CD8^+$  T cells. Total lymphocyte and (Aggregate3)  $CD8^+$  T-cell fractions were estimated in Thorsson and colleagues (45).



**Figure 3.**

MHC class I benchmark comparisons. **A**, PPV<sub>n</sub> for MHC class I allele-specific prediction on binding affinity test sets from Bonsack and colleagues (seven alleles) and Kim and colleagues (53 alleles; refs. 5, 8). **B**, PPV<sub>n</sub> for MHC class I allele-specific prediction on HLAp BST data set (Bassani-Sternberg and colleagues and Trolle and colleagues; refs. 7, 22), stratified by allele (six alleles). **C**, PPV<sub>n</sub> for MHC class I allele-specific prediction on HLAp BST data set (from **B**) stratified by peptide sequence length. **D**, True and false positives for each method on the top 50 ranked peptides from the HLAp BST data set. FP, false positives; PPV<sub>n</sub>, positive predictive value on the top  $n$  ranked peptides, where  $n$  is the number of true binders; TP, true positives.

## Runtime analysis

To assess the speed and scalability of the tested methods, we selected 1 million peptides sampled from the Abelin and colleagues data set (26) for class I alleles, and one million peptides sampled from the IEDB (curated data set 2018; ref. 54) for class II alleles. Sampling was done with replacement. For each method listed in **Fig. 1A**, networks for three class I MHC alleles (HLA-A\*02:01, HLA-A\*02:07, and HLA-A\*01:01) and three class II MHC alleles (HLA-DRB1\*01:01, HLA-DRB1\*11:01, and HLA-DRB1\*04:01) were used to predict binding over a range of input sample sizes ( $10^2$ ,  $10^3$ ,  $10^4$ ,  $10^5$ , and  $10^6$ ). All methods were run on a single graphics processing unit (GPU) compute node (one NVIDIA TESLA K80 GPU plus six 2.50 GHz Intel Xeon E5-2680v3 CPUs, 20 GB memory).

## Results

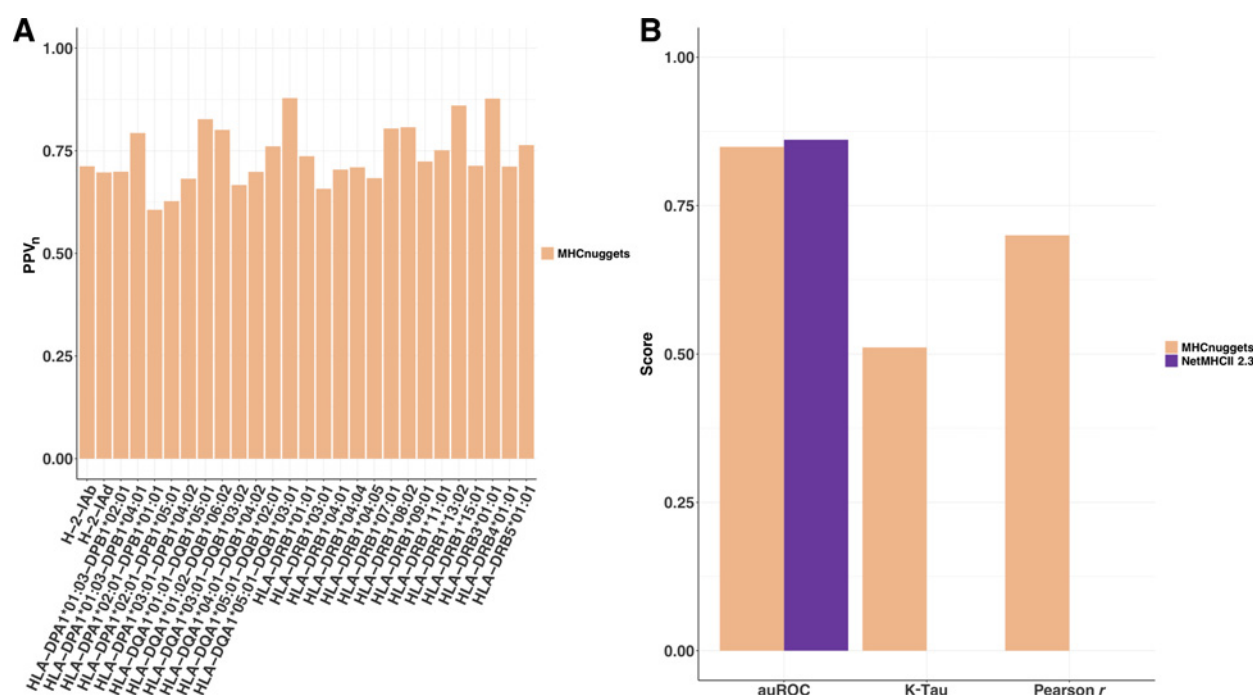
### High-throughput MHCnuggets breaks the MHC ligand prediction plateau

The MHCnuggets LSTM neural network architecture accepts peptides of variable lengths as inputs so that ligand binding prediction can be performed for both MHC class I and II alleles. To enable binding prediction for rare MHC alleles that have limited associated experimental data, we designed a method that leverages networks built for closely related common alleles with extensive data. When available, we utilize a transfer learning protocol to integrate binding affinity and HLA data in a single network model, to better represent the natural diversity of MHC binding peptides.

To assess the baseline performance of MHCnuggets allele-specific networks on binding affinity data, we compared our approach with

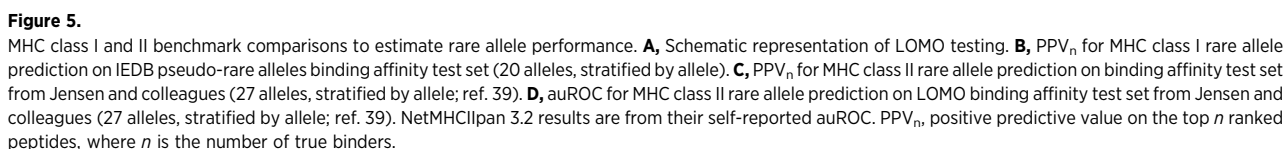
widely used MHC class I ligand prediction methods using two validation sets of binding affinity measurements (5, 8). We trained and tested MHCnuggets (noMS) and MHCflurry (noMS) using the Kim and colleagues data set (5) and evaluated the predictions provided by NetMHC 3.0 and NetMHCpan 2.0. We observed that MHCnuggets' performance ( $PPV_n = 0.829$ ,  $auROC = 0.924$ ) was comparable with these methods (**Fig. 3A**;  $PPV_n$  of all methods =  $0.825 \pm 0.005$ ,  $auROC$  of all methods =  $0.928 \pm 0.0031$ ). MHCnuggets was also comparable ( $PPV_n = 0.633$ ,  $auROC = 0.794$ ) to these methods when tested on the Bonsack and colleagues data set (ref. 8;  $PPV_n$  of all methods =  $0.625 \pm 0.008$ ,  $auROC$  of all methods =  $0.77 \pm 0.02$ ; **Fig. 3A**;  $\pm$  refers to standard deviation; Supplementary Tables S3A and S3B, and S4A and S4B).

Earlier neoantigen prediction methods focused on class I and trained on binding affinity data from IEDB (54). More recent work incorporated both binding affinity and HLA data into network training (14, 18). We compared MHCnuggets with several class I predictors that used both binding affinity and HLA data: MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. We selected the BST HLA data set (7, 22, 26) as an independent benchmark, as it was not used as training data by any of these methods. For all alleles tested, MHCnuggets achieved an overall  $PPV_n$  of 0.42 and  $auROC$  of 0.82 (**Fig. 3B**). On average, MHCnuggets'  $PPV_n$  was more than three times higher than MHCflurry 1.2.0, MHCflurry (train-MS), NetMHC 4.0, and NetMHCpan 4.0. For all alleles, MHCnuggets predicted fewer binders than other methods, resulting in fewer false-positive predictions. Stratifying by peptide length, MHCnuggets' increased  $PPV_n$  was most evident for peptides of length 9, 10, and 11 (**Fig. 3C**). The length distribution of predicted binders



**Figure 4.**

MHC class II benchmark comparisons. **A**,  $PPV_n$  for MHC class II allele-specific prediction on binding affinity test set from Jensen and colleagues (27 alleles, stratified by allele). **B**,  $auROC$ , K-Tau, and Pearson  $r$  scores for MHC class II alleles from 5-fold cross-validation. NetMHCII 2.3 performance is from their self-reported  $auROC$ . K-Tau, Kendall's  $\tau$  correlation;  $PPV_n$ , positive predictive value on the top  $n$  ranked peptides, where  $n$  is the number of true binders.

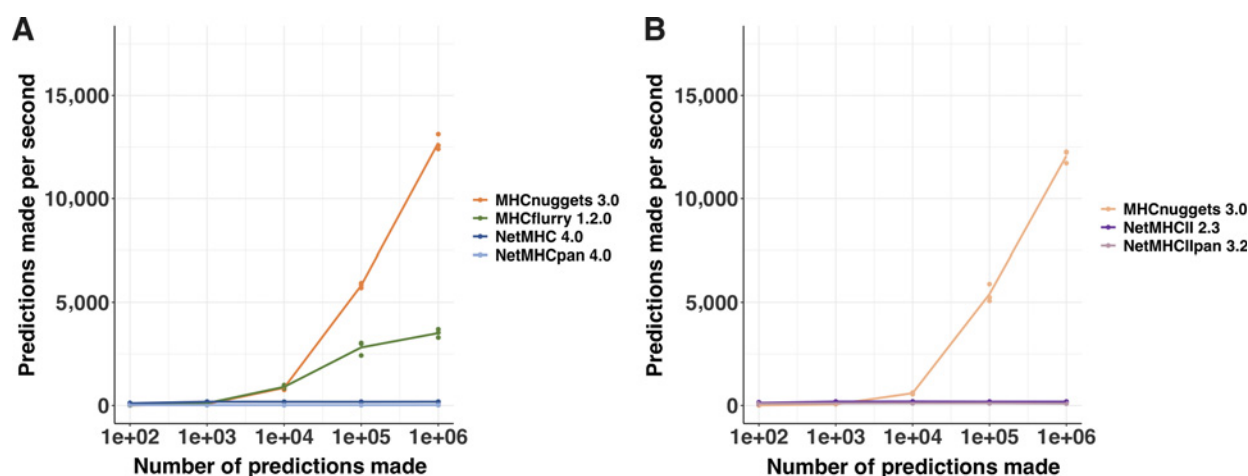


For some clinical applications, it may be desirable to minimize the number of false positives among a small number of top-scored peptides. We also compared PPV of the methods listed above on their top 50 and 500 ranked peptides from the BST data set (six MHC class I alleles; Supplementary Table S5E and S5F). MHCnuggets exhibited the highest PPV in the top 50 for all alleles except HLA-B\*51:01 and the highest PPV in the top 500 for all alleles (**Fig. 3D**).

We assessed the baseline performance of MHCnuggets class II allele-specific networks on binding affinity data. To enable comparison with the class II methods from the NetMHC group, we used a 5-fold cross-validation benchmark derived from IEDB that was included in the publication describing NetMHCII 2.3 and

We estimated performance for those class I and II MHC alleles for which we were unable to train allele-specific networks, using LOMO cross-validation (39). In this LOMO protocol, MHC-peptide binding is assessed for a well-characterized allele that has been held out from training, to approximate prediction performance for a rare allele (**Fig. 5A**). For the 20 class I alleles, the mean PPV<sub>n</sub> was 0.65, and the mean auROC was 0.671. For the 27 class II alleles, the mean PPV<sub>n</sub> was 0.65 and the mean auROC was 0.792. In comparison, the class II mean auROC of NetMHCIIpan 3.2 was 0.781 (**Fig. 5B–D**; Supplementary



**Figure 6.**

Timing and scalability. Runtime benchmark of tested methods using versions available on October 1, 2019, over a range of inputs (up to 1 million peptides). **A**, MHC class I prediction. **B**, MHC class II prediction.

Tables S7, and S8A and S8B). NetMHCpan rare allele predictor LOMO test results for class I are not publicly available; therefore, we were unable to compare with them.

### Fast and scalable computation

When run on a GPU architecture, MHCnuggets was faster and scaled more efficiently than MHC ligand predictors from the NetMHC family and MHCflurry. Given an input of one million peptides randomly selected from Abelin and colleagues (26), MHCnuggets runtime was 3.62, 69.7, and 624.5 times faster than MHCflurry 1.2.0, NetMHC 4.0, NetMHCpan 4.0, respectively (Fig. 6A). The improvement was similar for class II peptides, for which an input of one million peptides to MHCnuggets ran 65.6 times and 126 times faster than NetMHCII 2.3 and NetMHCIIpan 3.2, respectively (Fig. 6B). As the total number of input peptides was increased from 0 to 1 million, the runtime per peptide plateaued for other methods but decreased exponentially for MHCnuggets.

### Predicted MHC class I IMMs in TCGA patients

To illustrate the utility of MHCnuggets' improvements in scalability and PPV for the analysis of very large patient cohorts, we predicted class I IMMs in patients whose exomes were sequenced by the TCGA consortium (Materials and Methods). In our analysis pipeline, patient exomes were split into 21 amino acid residue sequence fragments, centered on each somatic missense mutation. For each sequence fragment, MHCnuggets predicted the MHC binding for all possible 8-, 9-, 10-, and 11-length peptide windows. Peptides that passed filters of predicted  $IC_{50}$  threshold ( $<500$  nmol/L) and gene expression ( $Z > 1.0$ ; Materials and Methods) for at least one patient-specific MHC allele were classified as predicted IMMs (Supplementary Table S9A). Finally, we characterized driver status and positional hotspot propensity of the predicted IMMs.

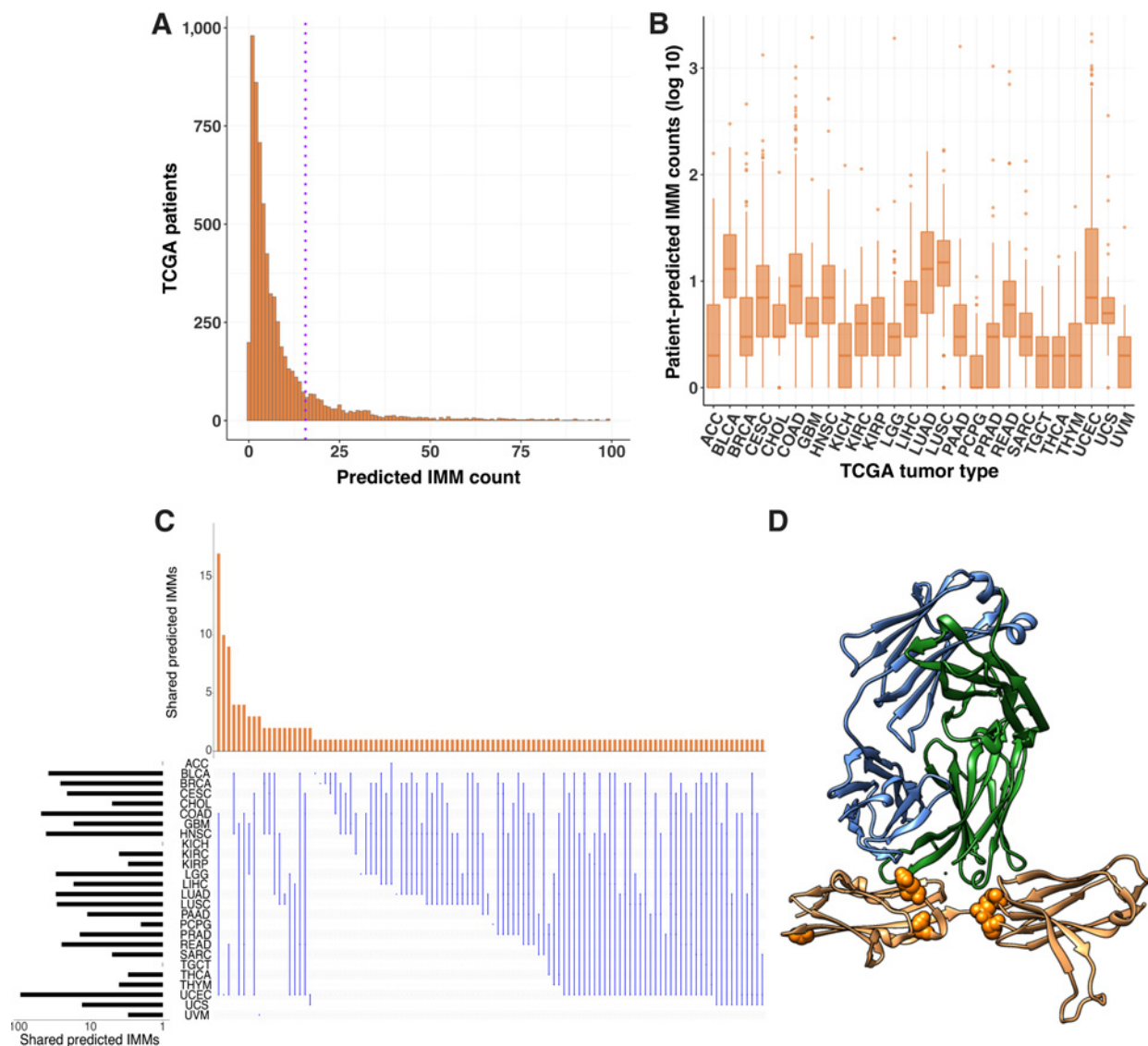
Total processing time for 26,284,638 allele-peptide comparisons supported by RNA-seq expression was under 2.3 hours. First, we sought to ascertain the extent of variability in predicted IMM count among individuals with different cancer types. Next, we identified predicted IMMs and protein regions enriched for predicted IMMs that were shared across patients, because these might be informative

for neoantigen-based therapeutic applications. Then we considered whether predicted IMMs were more or less likely to be driver mutations. Finally, we assessed the associations between predicted patient IMM load and computationally estimated immune cell infiltrates.

After applying a gene-expression filter, we identified 101,326 unique predicted IMMs in 26 TCGA cancer types, with a mean of 15.6 per patient. We found that the majority of patients harbored fewer than six predicted IMMs, and 197 patients had none. Seventy-two percent of patients had from one to 10 predicted IMMs, compared with 1.9% of patients with more than 100, and 9 patients with more than 1,000 (Fig. 7A). Cancer types with the highest number of predicted IMMs were uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD), and lung adenocarcinoma (LUAD), all three of which are known for high mutation burden and immunogenicity (45). UCEC and COAD are also known to have a high frequency of microsatellite-unstable (MSI) tumors. The lowest number were found in uveal melanoma (UVM), paraganglioma and pheochromocytoma (PCPG), and testicular germ cell tumors (TGCT; Fig. 7B; Supplementary Table S9B).

Across all cancer types, we identified 1,393 predicted IMMs harbored by 2 or more patients, of which 167 were identified in 3 or more patients. Of these, 167 only 11.5% occurred exclusively in a single cancer type (Fig. 7C). The predicted IMMs identified in the largest number of patients were *IDH1* R132H (62), *FGFR3* S249C (24), *PIK3CA* E545K (23), *KRAS* G12D (18), *PIK3CA* E542K (18), *TP53* R175H (18), *TP53* R248Q (18), *TP53* R273C (17), and *KRAS* G12V (16), which are known recurrent oncogenic driver mutations (55, 56). Of the 1,071 genes harboring predicted IMMs in 2 or more patients, the ones containing the most included *TP53* (68), *CTNNB1* (18), *PIK3CA* (16), *HRAS* (8), *KRAS* (7), *PTEN* (7), *FBXW7* (6), *EGFR* (5), *MDN1* (5), *POLE* (5), *TRRAP* (5), and *VPS13C* (5; Supplementary Table S9C). Six missense mutations harbored by patients in the TCGA cohorts were previously validated by CD8<sup>+</sup> T-cell response assays (57, 58, 59). Of the six missense mutations, *TP53*<sup>R248Q</sup>, *TP53*<sup>Y220C</sup>, *TP53*<sup>R175H</sup>, *TP53*<sup>R248W</sup>, and *KRAS*<sup>G12D</sup> were predicted to be IMMs by our MHCnuggets pipeline and were shared by 3 or more of the TCGA patients.

## High-Throughput Prediction of Neoantigens with MHCnuggets

**Figure 7.**

MHC class I IMMs in TCGA patients. **A**, Number of predicted IMMs identified in 6,613 TCGA patients. Dotted line, mean predicted IMMs per patient (15.6). Note that 123 patients had >100 predicted IMMs but are not included for visual clarity. **B**, Number of predicted IMMs by cancer type. **C**, Predicted IMMs shared by 3 or more patients and the cancer types in which they occurred. Each row represents a cancer type, and each column illustrates the overlap of predicted IMMs seen in a single cancer type or multiple cancer types. For example, the first column shows the number of predicted IMMs shared among patients with colorectal adenocarcinoma and uterine corpus endometrial carcinoma. Bars to the left show the total number of unique predicted IMMs in each cancer type. Bar heights reflect the count of unique shared predicted IMMs, not the total number of patients in which the predicted IMM was observed. Image generated with UpSetR. **D**, Fibroblast growth factor receptor (FGFR3) predicted IMM hot region identified by HotMAPs in bladder cancer (BLCA). Predicted IMMs shown and number of BLCA patients with the predicted IMM: p.E216K (1), p.D222N (1), p.G235D (1), p.R248C (3), and p.S249C (24). Except for p.G235D, these predicted IMMs are proximal to the interface of FGFR3 protein and the light and heavy chains of an antibody fragment designed for therapeutic application in bladder cancer (PDB ID: 3GRW; ref. 61). ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangioadenocarcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; TGCT, testicular germ cell tumor; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma; UCS, uterine carcinosarcoma; UVM, uveal melanoma.

Furthermore, 61.7% of the 167 predicted IMMs shared by 3 or more patients were classified as driver missense mutations by CHASMplus ( $q < 0.01$ ). This percentage is significantly higher than the number of predicted drivers among all TCGA missense mutations (9,821 out of

791,637 or 1.2%). Although many shared IMMs were predicted to be driver missense mutations, the percentage of predicted IMMs predicted to be drivers was ~0.1% of total predicted IMMs in our study. When compared with the OncoKB database of experimentally

confirmed driver mutations (60), 53.9% of the shared predicted IMMs identified as “oncogenic” or “likely oncogenic” driver mutations. The percentage is lower (25.7%) if “likely oncogenic” mutations are excluded.

Although we observed a limited number of shared predicted IMMs, we reasoned that protein regions enriched for predicted IMMs could present a therapeutic opportunity in certain cancer types. Using HotMAPS 1D, we identified clusters of residues within protein regions having statistically significant enrichment of predicted IMMs ( $q < 0.01$ ). These included *CIC* in low-grade glioma (LGG); *NFE2L2* and *FGFR3* (Fig. 7D) in bladder cancer (BLCA; ref. 61); *KRAS* in pancreatic adenocarcinoma (PAAD); *KIT* in TGCTs; *HRAS* in head and neck squamous carcinoma (HNSC); *PTEN*, *POLE*, and *PPP2R1A* in UCEC; and *GNAQ* and *SF3B1* in UVM. Three genes harbored predicted immunogenic regions in more than one cancer type: *P53* in BLCA, BRCA, HNSC, LGG, and UCEC; *PIK3CA* in HNSC and cervical squamous cell carcinoma (CESC); and *CTNNB1* in liver hepatocellular carcinoma (LIHC) and UCEC (Supplementary Table S9D).

We explored the relationship between mutation driver status predicted by CHASMplus and predicted IMM status using logistic regression. The log odds of being a predicted IMM was significantly decreased for drivers ( $\beta = -0.66$ , Wald test  $P < 2e-16$ ), which is consistent with previous work suggesting that negative evolutionary selection eliminates MHC class I immunogenic oncogenic mutations early in tumor development (62).

Finally, we considered whether a patient's predicted IMM load was associated with changes in immune cell infiltrates as estimated from RNA-seq of bulk cancer tissue. Predicted IMM load was significantly associated with increased total leukocyte fraction ( $\beta = 0.75$ , Wald test  $P < 2 \times 10^{-16}$ ) and with increased CD8<sup>+</sup> T-cell fraction ( $\beta = 5.9$ , Wald test  $P < 2 \times 10^{-16}$ ; Supplementary Table S9E).

These findings suggest that IMMs drive tumor immunoediting and may be informative for the interpretation of clinical responses to immunotherapy.

## Discussion

MHCnuggets provides a flexible open-source platform for MHC-peptide binding prediction that can handle common MHC class I and II alleles as well as rare alleles of both classes. The LSTM network architecture can handle peptide sequences of arbitrary length, without shortening or splitting. The single neural network architecture requires fewer hyperparameters than more complex architectures and simplifies network training. In addition, our neural network transfer learning protocols allow for parameter sharing among allele-specific, binding affinity- and HLAp-trained networks. When trained on binding affinity data, MHCnuggets performs as well as other current methods. When trained on both binding affinity and HLAp data, we demonstrate improved PPV<sub>n</sub> on an independent HLAp test set, with respect to other methods that use both binding affinity and HLAp data. Although PPV<sub>n</sub> was lowest for the independent HLAp test set for all methods, this result is likely due to systematic differences between training HLAp data (monoallelic B-cell lines; ref. 26) and the test data comprised of seven multiallelic cell lines (HeLa, HTC116, JY, fibroblasts, SupB15, HCC1937, and HCC1143; refs. 7, 22), yielding a more challenging prediction problem. We attribute MHCnuggets' improvement on the independent test set with respect to other methods to (1) optimization of PPV<sub>n</sub> in our network training protocol and (2) our implementation of transfer learning to integrate information from binding affinity and HLAp measurements. The

performance of all methods is generally highest when both training and test data come from similar binding affinity experiments, but performance improvement on HLAp data is more biologically relevant (63).

We demonstrate improved scalability by comparing the runtime of MHCnuggets on 1 million peptides to comparable methods, and further by processing over 26 million expressed peptide-allele pairs across TCGA samples in under 2.3 hours. We identified 101,326 unique IMMs harbored by patients using 26 cancer types sequenced by the TCGA, based on transcriptional abundance and differential binding affinity compared with reference peptides. These results contrast with a previous report of neoantigens in TCGA patients in several respects. Rech and colleagues (64) applied a minimum expression threshold of 1 RNA-seq read count, an IEDB-recommended combination of neoantigen predictors derived primarily from different versions of NetMHC, and IC<sub>50</sub> threshold of 50 nmol/L to identify strong MHC binders. Their approach yielded 495,793 predicted class I classically defined neoantigen peptides (each harboring a single immunogenic mutation) from 6,324 patients in 26 cancer types. As in our study, high variability in neoantigen burden across cancer types was observed. The difference between predicted IMM and neoantigen burden in the two studies is likely due to differences in RNA expression threshold and the low false-positive rate of MHCnuggets compared with IEDB-recommended tools.

Based on our conservative thresholds, predicted IMMs were almost exclusively private to individual TCGA patients, with only 1,393 predicted IMMs observed in more than 1 patient. Although more than 61% of predicted IMMs shared by more than 2 patients were predicted to be driver mutations, the overall log odds of immunogenicity decreased for predicted driver mutations, indicating immunogenicity might shape the driver mutation landscape. Patient predicted IMM counts were also associated with increase in total leukocyte fraction and fraction of CD8<sup>+</sup> T cells, suggesting that they may be relevant to immune system response to cancer.

This work has several limitations. First, our analyses are limited to missense mutations, which, although numerous, cannot account for the various somatic gene fusions, frameshift indels, splice variants, etc., in tumors that also generate neoantigens. Although MHCnuggets can handle peptide sequences regardless of their mutational origins, we prioritized missense mutations in this study. Indeed, the context of a peptide sequence, such as what sequences are flanking, its source protein and the expression of the source protein, is informative for MHC ligand prediction (21, 26). This type of information is available only for a limited number of HLAp data sets, which were unavailable to us for training purposes. As more well-characterized HLAp data sets become available, we will further develop MHCnuggets to include these features. We did not address T-cell receptor binding to bound peptide-MHC complexes or T-cell activation upon complex binding. Although we are pursuing this more complex modeling problem, we believe that improved prediction of peptide binding to MHC is also therapeutically relevant (21). Finally, we are unable to directly compare performance to the MHC class II prediction methods from the NetMHC group, except for self-reported auROC. Although we are not able to do a rigorous comparison of MHCnuggets class II prediction, our benchmark comparisons suggested that MHCnuggets was competitive with NetMHCII 2.3 and that MHCnuggets class II rare allele performance was competitive with NetMHCIIpan 3.2, suggesting that further work in this area is warranted.

In summary, we present MHCnuggets, an open-source software package for MHC ligand prediction that performs better than existing

## High-Throughput Prediction of Neoantigens with MHCnuggets

methods with respect to PPV by leveraging transfer learning to integrate binding affinity and HLA data. In contrast to previous methods, MHCnuggets handles both MHC class I and II ligand prediction and both common and rare HLA alleles, all within a single framework. We demonstrated the utility of MHCnuggets as a basic pipeline to analyze mutation immunogenicity, shared predicted IMMs, and the relationship between mutation immunogenicity, driver potential, and immune infiltrates from large-scale cancer patient sequencing data from TCGA.

## Disclosure of Potential Conflicts of Interest

V.E. Velculescu is founder and on the board of directors of Personal Genome Diagnostics, is a scientific advisory board member for Takeda, and has ownership interest (including patents) in Personal Genome Diagnostics. V. Anagnostou reports receiving a commercial research grant from Bristol-Meyers Squibb. No potential conflicts of interest were disclosed by the other authors.

## Authors' Contributions

**Conception and design:** X.M. Shao, R. Bhattacharya, V.E. Velculescu, R. Karchin  
**Development of methodology:** X.M. Shao, R. Bhattacharya, J. Huang, I.K.A. Sivakumar, R. Karchin

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** X.M. Shao, R. Bhattacharya, J. Huang, I.K.A. Sivakumar, C. Tokheim, L. Zheng, B. Kaminow, A. Omdahl, V.E. Velculescu, V. Anagnostou, R. Karchin

## References

1. Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J, et al. Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov* 2017;7:264–76.
2. Yarchoan M, Johnson BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* 2017;17:209–22.
3. Lundegaard C, Lund O, Buus S, Nielsen M. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology* 2010;130:309–18.
4. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 2016;32:511–7.
5. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics* 2014;15:241.
6. Kim Y, Sidney J, Pinilla C, Sette A, Peters B. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 2009;10:394.
7. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The length distribution of class I-Restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J Immunol* 2016;196:1480–7.
8. Bonsack M, Hoppe S, Winter J, Tichy D, Zeller C, Kupper MD, et al. Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC-peptide binding data set. *Cancer Immunol Res* 2019;7:719–36.
9. Gfeller D, Bassani-Sternberg M, Schmidt J, Luescher IF. Current tools for predicting cancer-specific T cell immunity. *Oncoimmunology* 2016;5:e1177691–e.
10. Liu XS, Mardis ER. Applications of immunogenomics to cancer. *Cell* 2017;168:600–12.
11. The problem with neoantigen prediction. *Nat Biotechnol* 2017;35:97.
12. Wiczorek M, Abualrous ET, Sticht J, Alvaro-Benito M, Stolzenberg S, Noe F, et al. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front Immunol* 2017;8:292.
13. Lu YC, Robbins PF. Targeting neoantigens for cancer immunotherapy. *Int Immunol* 2016;28:365–70.
14. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;199:3360–8.
15. Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method

**Writing, review, and/or revision of the manuscript:** X.M. Shao, R. Bhattacharya, J. Huang, I.K.A. Sivakumar, C. Tokheim, L. Zheng, D. Hirsch, A. Omdahl, M. Bonsack, A.B. Riemer, V.E. Velculescu, V. Anagnostou, K.A. Pagel, R. Karchin  
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** J. Huang, D. Hirsch, M. Bonsack, A.B. Riemer  
**Study supervision:** V.E. Velculescu, R. Karchin

## Acknowledgments

Part of this research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC). This work was supported in part by a William R. Brody Faculty Scholarship to R. Karchin, Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up To Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415), the Commonwealth Foundation, the U.S. NIH (grants CA121113, CA006973, and CA180950), the V Foundation, and LUNGevity to V.E. Velculescu. Stand Up To Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 20, 2019; revised October 8, 2019; accepted December 20, 2019; published first December 23, 2019.

- od including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 2013;65:711–24.
16. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res* 2008;36:509–12.
17. Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2007;2:1–10.
18. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;7:129–32.
19. Bassani-Sternberg M, Coukos G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr Opin Immunol* 2016;41:9–17.
20. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* 2015;348:69–74.
21. Bulik-Sullivan B, Busby J, Palmer CD, Davis MJ, Murphy T, Clark A, et al. Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol* 2019;37:55–63.
22. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, Mann M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol Cell Proteomics* 2015;14:658–73.
23. Boehm KM, Bhinder B, Raja VJ, Dephoure N, Elemento O. Predicting peptide presentation by major histocompatibility complex class I: an improved machine learning approach to the immunopeptidome. *BMC Bioinformatics* 2019;20:7.
24. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computat* 1997;9:1735–80.
25. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I, editors. *Artificial neural networks and machine learning – ICANN 2018. Proceedings, Part III, of the 27th International Conference on Artificial Neural Networks*; 2018 Oct 4–7; Rhodes, Greece. Cham (Switzerland): Springer; 2018. p. 270–9.
26. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 2017;46:315–26.
27. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT(R), the international ImmunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* 2015;43:D413–22.



28. Sidney J, Assarsson E, Moore C, Ngo S, Pinilla C, Sette A, et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res* 2008;4:2.
29. Gfeller D, Guillaume P, Michaux J, Pak HS, Daniel RT, Racle J, et al. The length distribution and multiple specificity of naturally presented HLA-I ligands. *J Immunol* 2018;201:3705–16.
30. Greenbaum J, Sidney J, Chung J, Brander C, Peters B, Sette A. Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across super-types. *Immunogenetics* 2011;63:325–35.
31. Abadi Mn, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *CoRR* 2016; abs/1603.04467. Available from: <https://www.tensorflow.org/>.
32. Chollet F, others. Keras. GitHub; 2015. Available from: <https://github.com/keras-team/keras>.
33. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med* 2016;8:11.
34. Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res* 2020;8:409–20.
35. Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K, et al. Population-level distribution and putative immunogenicity of cancer neoepitopes. *BMC Cancer* 2018;18:414.
36. Nielsen M, Lundegaard C, Warming P, Lauemoller SL, Lamberth K, Buus S, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 2003;12:1007–17.
37. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: *Proceedings of the Third International Conference on Learning Representations, ICLR 2015; 2015 May 7–9; San Diego, CA*.
38. Gal YA, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. In: *Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R. Advances in Neural Information Processing Systems 29 (NIPS 2016); 2016 Dec 5–10; Barcelona, Spain. San Diego (CA): Neural Information Processing Systems; 2016*.
39. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 2018;154:394–406.
40. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;47: D339–D43.
41. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 2009;61:1–13.
42. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 2015; 31:2174–.
43. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018;173:371–85.
44. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 2018;6:271–81.
45. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. *Immunity* 2018;48:812–30.
46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.
47. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014;30:3310–6.
48. UniProt Consortium. Activities at the universal protein resource (UniProt). *Nucleic Acids Res* 2014;42:D191–8.
49. Tokheim C, Bhattacharya R, Niknafs N, Gyax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* 2016;76:3719–31.
50. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc: Ser B* 1995;57: 289–300.
51. Tokheim C, Karchin R. CHASMPplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst* 2019;9:9–23.
52. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015; 12:453–7.
53. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol* 2018; 171:243–59.
54. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Res* 2015;43: D405–D.
55. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA oncogene in human cancers. *Br J Cancer* 2006;94:455–9.
56. Tomlinson DC, Hurst CD, Knowles MA. Knockdown by shRNA identifies S249C mutant FGFR3 as a potential therapeutic target in bladder cancer. *Oncogene* 2007;26:5889–99.
57. Tran E, Ahmadvadeh M, Lu YC, Gros A, Turcotte S, Robbins PF, et al. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 2015;350:1387–90.
58. Malekzadeh P, Pasetto A, Robbins PF, Parkhurst MR, Paria BC, Jia L, et al. Neoantigen screening identifies broad TP53 mutant immunogenicity in patients with epithelial cancers. *J Clin Invest* 2019;129: 1109–14.
59. Parkhurst MR, Robbins PF, Tran E, Prickett TD, Gartner JJ, Jia L, et al. Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. *Cancer Discov* 2019;9:1022–35.
60. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017 Jul;2017. doi: 10.1200/PO.17.00011.
61. Qing J, Du X, Chen Y, Chan P, Li H, Wu P, et al. Antibody-based targeting of FGFR3 in bladder carcinoma and t(4;14)-positive multiple myeloma in mice. *J Clin Invest* 2009;119:1216–29.
62. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, et al. MHC-I genotype restricts the oncogenic mutational landscape. *Cell* 2017;171: 1272–83.
63. Bassani-Sternberg M, Chong C, Guillaume P, Solleder M, Pak H, Gannon PO, et al. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput Biol* 2017;13:e1005725.
64. Rech AJ, Balli D, Mantero A, Ishwaran H, Nathanson KL, Stanger BZ, et al. Tumor immunity and survival as a function of alternative neopeptides in human cancer. *Cancer Immunol Res* 2018;6:276–87.

# Cancer Immunology Research

## High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets

Xiaoshan M. Shao, Rohit Bhattacharya, Justin Huang, et al.

*Cancer Immunol Res* 2020;8:396-408. Published OnlineFirst December 23, 2019.

<b>Updated version</b>	Access the most recent version of this article at: doi: <a href="https://doi.org/10.1158/2326-6066.CIR-19-0464">10.1158/2326-6066.CIR-19-0464</a>
<b>Supplementary Material</b>	Access the most recent supplemental material at: <a href="http://cancerimmunolres.aacrjournals.org/content/suppl/2019/12/21/2326-6066.CIR-19-0464.DC1">http://cancerimmunolres.aacrjournals.org/content/suppl/2019/12/21/2326-6066.CIR-19-0464.DC1</a>

<b>Cited articles</b>	This article cites 58 articles, 12 of which you can access for free at: <a href="http://cancerimmunolres.aacrjournals.org/content/8/3/396.full#ref-list-1">http://cancerimmunolres.aacrjournals.org/content/8/3/396.full#ref-list-1</a>
<b>Citing articles</b>	This article has been cited by 14 HighWire-hosted articles. Access the articles at: <a href="http://cancerimmunolres.aacrjournals.org/content/8/3/396.full#related-urls">http://cancerimmunolres.aacrjournals.org/content/8/3/396.full#related-urls</a>

<b>E-mail alerts</b>	<a href="#">Sign up to receive free email-alerts</a> related to this article or journal.
<b>Reprints and Subscriptions</b>	To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at <a href="mailto:pubs@aacr.org">pubs@aacr.org</a> .
<b>Permissions</b>	To request permission to re-use all or part of this article, use this link <a href="http://cancerimmunolres.aacrjournals.org/content/8/3/396">http://cancerimmunolres.aacrjournals.org/content/8/3/396</a> . Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.