Genome Medicine

RESEARCH

Open Access

Predicting MHC-I ligands across alleles and species: how far can we go?

Daniel M. Tadros[1,2,3,4], Julien Racle[1,2,3,4] and David Gfeller[1,2,3,4*]

Abstract

Background $CD8^+$ T-cell activation is initiated by the recognition of epitopes presented on class I major histocompatibility complex (MHC-I) molecules. Identifying such epitopes is useful for molecular understanding of cellular immune responses and can guide the development of personalized vaccines for various diseases including cancer. For a few hundred common human and mouse MHC-I alleles, large datasets of ligands are available and machine learning MHC-I ligand predictors trained on such data reach high prediction accuracy. However, for the vast majority of other MHC-I alleles, no ligand is known.

Methods We capitalize on an expanded architecture of our MHC-I ligand predictor (MixMHCpred3.0) to systematically assess the extent to which predictions of MHC-I ligands can be applied to MHC-I alleles that currently lack known ligand data.

Results Our results reveal high prediction accuracy for most MHC-I alleles in human and in laboratory mouse strains, but significantly lower accuracy in other species. Our work further outlines some of the molecular determinants of MHC-I ligand prediction accuracy across alleles and species. Robust benchmarking on external data shows that our MHC-I ligand predictor demonstrates competitive performance relative to other state-of-the-art MHC-I ligand predictors and can be used for $CD8^+$ T-cell epitope predictions.

Conclusions Our work provides a valuable tool for predicting antigen presentation across all human and mouse MHC-I alleles. MixMHCpred3.0 tool is available at https://github.com/GfellerLab/MixMHCpred.

Background

$CD8^+$ T cells play a key role in eliminating infected or malignant cells. To perform this task, $CD8^+$ T cells recognize small peptides displayed on class I major histocompatibility complex (MHC-I) molecules on the surface of the targeted cells. MHC-I ligands are considered as promising therapeutic targets and have been used in pre-clinical and clinical studies. For instance, in cancer immunotherapy, MHC-I ligands have been used as personalized vaccines to boost the immune system to recognize neo-antigens [1–5]. Additionally, viral peptides presented on MHC-I molecules have been utilized in vaccines against infectious diseases to elicit strong T-cell recognition [6].

MHC-I molecules bind short peptides (roughly 8–14 amino acids) with a general preference for 9-mers [7–10]. The binding is typically determined by primary anchor residues at the second and last positions of the peptides. Several alleles display additional anchor residues at other positions [7, 11]. In humans, MHC-I molecules are encoded by three commonly expressed genes (HLA-A, HLA-B, HLA-C) along with a few other genes (e.g., HLA-E, HLA-F, HLA-G). MHC-I

*Correspondence:
David Gfeller
david.gfeller@unil.ch
[1] Department of Oncology, Ludwig Institute for Cancer Research Lausanne, University of Lausanne, Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[3] Agora Cancer Research Centre, Lausanne 1011, Switzerland
[4] Swiss Cancer Center Leman (SCCL), Lausanne, Switzerland

genes exhibit a very high level of polymorphism, with thousands of distinct alleles [12]. MHC-I genes in various species are known to evolve rapidly and are not strongly conserved, even among relatively closely related species [13–15]. Different MHC-I alleles have different peptide-binding specificities, which include differences in binding motifs and peptide length distributions [7, 10, 16]. This results in distinct repertoires of MHC-I ligands in different individuals and different species [17–20]. Over the last decade, mass spectrometry-based MHC-I peptidomics has emerged as the leading source of information about MHC-I binding specificities. These data have enabled researchers to compute binding motifs and peptide length distributions supported by thousands of ligands for more than 100 MHC-I alleles [21–24].

Many in silico prediction tools for MHC-I ligands have been developed to narrow down the list of potential epitopes [23–28]. These tools are mainly trained on MHC-I peptidomics data and such data are currently available for a bit more than a hundred alleles [16, 24, 26, 29, 30]. These include all common alleles in humans but only a few alleles from other species. Two main classes of MHC-I ligand predictors can be distinguished: allele-specific or pan-allele predictors. Allele-specific predictors, such as NetMHC [31] or MixMHCpred2.2 [26], train a separate model for each allele with known ligands. These tools are therefore restricted to the set of MHC-I alleles with available data. Pan-allele predictors, like NetMHCpan4.1 [28], MHCflurry2.0 [27], ACME [32], MATHLA [33], DeepLigand [34], HLAthena [24], BigMHC [35], or SHERPA [23] can make predictions for a broader range of alleles. These predictors leverage shared properties of MHC-I molecules across different alleles and correlations between MHC-I binding site residues and binding specificities to make predictions even when experimental ligands are not available for a given allele. Pan-allele MHC-I ligand predictors have been successfully used in multiple studies in humans [27, 28, 32–34]. However, considering the rapid evolution of MHC-I genes and alleles across species, it remains unclear how far predictions can be expanded, especially in species without known MHC-I ligands.

In this study, we capitalized on high-quality MHC-I peptidomics data for hundreds of alleles to perform a careful benchmarking of how predictions of MHC-I ligands can be extrapolated across alleles and species. Our work provides insights into the molecular determinants underlying MHC-I ligand prediction accuracy, as well as a robust implementation of such predictions.

Methods

Collection of MHC-I ligands

Naturally presented MHC-I ligands were collected from more than 250 MHC-I peptidomics samples from human, mouse, cattle, canid, and non-human primate. These include all samples considered in [26]. We further included data from a few recent MHC-I peptidomics studies [23, 29, 30, 36–41]. All data were retrieved from the original studies to prevent having filtered data based on MHC-I ligand predictors. All samples were processed with our motif deconvolution tool (MixMHCp) to identify shared motifs across samples sharing the same allele [7]. Further information regarding this procedure and the results obtained can be found in our previous publications [7, 22, 26]. The final dataset of naturally presented MHC-I ligands comprises 511,553 peptide-MHC-I interactions with 143 different MHC-I alleles.

Building MHC-I binding motifs and peptide length distributions

For all MHC-I alleles with naturally presented ligands, Position Probability Matrices (PPMs) were constructed by computing the frequency of each amino acid at each position in the set of ligands of the given allele, including standard pseudocounts based on BLOSUM62 as detailed in Gfeller et al. [7] and Racle et al. (2019). Separate PPMs were generated for each ligand length L from 8 to 14. The Position Weight Matrices (PWMs) representing the final binding motifs were computed by normalizing the PPMs with the amino acid background frequencies of the human proteome, as outlined in [7, 42]. Binding motifs were visualized using ggseqlogo [43] and Logomaker [44].

To determine peptide length distributions, the fraction of naturally presented MHC-I ligands of each length (from 8 to 14) was computed as described in [7].

Predicting MHC-I binding motifs

Inspired by our recent work on MHC-I and MHC-II motifs [16, 45], neural networks were used to predict PPMs of MHC-I molecules without known ligands. More precisely, distinct networks were trained for each peptide length (8 to 14). The input of each neural network is the list of binding site residues from the MHC-I molecules (34 residues). This binding site was defined as in [46]. Each binding site residue was encoded as a 20-dimensional vector based on the BLOSUM62 probability matrix. The output of each network consists of a matrix of 20xL, representing the PPM at the corresponding motif length L. Each network is composed of an input layer (34×20 nodes), one fully connected hidden layer (256 nodes) followed by a dropout of 0.2, and

an additional layer that reshapes the output layer from a vector (20xL nodes) to a matrix of size 20 rows and L columns. We used a rectified linear unit (ReLU) activation function for the hidden layer and a custom softmax function for the output layer that applies the softmax activation function on each column of the matrix. We used the Kullback Leibler divergence as a loss function, and it was optimized using the Adam optimizer with a learning rate of 0.0001. These neural networks were implemented in Python (version 3.7.11), using Keras packages relying on TensorFlow (version 2.2.4-tf). 1000 epochs were set for the training process. For each allele and each peptide length (8 to 14), we then normalized by background human proteome frequencies to create the final predicted PWM.

Predicting peptide length distributions

A neural network was developed to predict the peptide length distribution of MHC-I molecules. The input layer is the same as for the MHC-I motifs prediction (34×20 nodes), followed by one hidden layer (128 nodes) with the rectified linear unit (ReLU) activation function followed by a dropout of 0.2. The output layer is the peptide length distribution (from 8 to 14, i.e., 7 nodes) based on the softmax activation function. We used the Kullback Leibler divergence as a loss function, and it was optimized using the Adam optimizer with a learning rate of 0.0001. A maximum of 125 epochs were set for the training process with early stopping applied if no improvement in loss was observed over a span of 20 consecutive epochs.

Predicting MHC-I ligands

Following the procedure described in [26], the presentation score of a peptide X of length L with allele a is given by:

$$S^a(X) = \frac{M^{(a,L)}(X) - C^{(a,L)}}{D^{(a,L)}} \; with$$

$$M^{(a,L)}(X) = \frac{\log(\prod_{l=1}^{L} M^{(a,L)}_{X_l})}{L}$$

$M^{(a,L)}(X)$ represents the raw score of peptide $X = (X_1, \ldots, X_L)$ given by the PWM representing the motif of allele a for L-mers. The correction factor $D^{(a,L)}$ is computed so that $S^a(X)$ has a standard deviation of 1 over a set of 100′000 peptides of length L randomly selected from the human proteome. The correction factors $C^{(a,L)}$ are computed so that the length distribution of the top 0.1% of 700,000 random peptides (taken from the human proteome with uniform length distribution between 8- and 14-mers) follows exactly the peptide length distribution of allele a observed in HLA-I peptidomics data. The correction factors $C^{(a,L)}$ and $D^{(a,L)}$ are previously defined in [26]. %ranks given as output of MixMHCpred3.0 are estimated based on the distribution of scores $S^a(X)$ of a set of 700,000 random peptides (100,000 of each length from 8 to 14), as done in other MHC-I ligand predictors.

Leave-one-allele-out (LOA) cross-validation

We performed leave-one-allele-out cross-validation for ligand, binding motif, and peptide length distribution predictions using iteratively as a test set for each allele. For ligand predictions, a 99-fold excess of negatives was added randomly from the human proteome with uniform length distribution from 8 to 14. Subsequently, binding scores were predicted for each peptide in the test set, and the performance was evaluated based on the AUC values (Fig. 2A). For each length (from 8 to 14), the predicted PWMs were compared to the experimental ones by computing the Euclidean distance for each position of the PWM and averaging these distances. The lower the distance, the closer the predicted motifs are to the experimental ones (Fig. 2B shows the Euclidean distance for the 9-mers motifs). Similarly, we computed the Euclidean distance between the predicted and experimental peptide length distributions (Fig. 2C).

Binding site sequence distances

The binding site distance between two alleles was calculated as described in the following formula:

$$1 - \frac{\sum_{j=1}^{J} \text{blosum}(a_j, b_j)}{\sqrt{\sum_{j=1}^{J} \text{blosum}(a_j, a_j) \times \sum_{j=1}^{J} \text{blosum}(b_j, b_j)}}$$

in which blosum refers to the Blosum62 scoring matrix, which is used to score amino acid substitutions [47], J represents the length of the MHC-I binding site sequence (34 amino acids), a_j and b_j denote the amino acid from the two alleles being compared. The resulting score ranges from 0 to 1, where a higher score indicates a greater distance between the two MHC-I binding site sequences. For alleles without known ligands, the binding site distance to the set of alleles with known ligands is defined as the minimum of the distances to alleles with known ligand, or equivalently the distance to the closest alleles with known ligands. For simplicity, this distance to the closest allele is often referred to as the "binding site distance."

MHC-I sequences retrieval and alignment

Human MHC-I sequences were retrieved from the IPD-IMGT/HLA database [12]. MHC-I sequences from multiple other species were retrieved from the IPD-MHC database [48]. Mouse MHC-I sequences are not part of

the IPD-MHC database so they were manually retrieved from the UniProtKB database [49]. We then aligned the MHC-I sequences using the MAFFT algorithm ([50], version 7.520) and took as a reference the list of sequences of alleles with known ligands for the alignment.

Population frequencies for HLA-I alleles with known ligands

Human MHC-I allele frequencies were obtained from The Allele Frequency Net Database (AFND) [51]. Only samples with a sample size > 500 and a resolution level of ≥ 2 fields were included to ensure data reliability. The weighted average frequency of each allele across all populations was then calculated.

Benchmarking with other MHC-I ligand predictors

To assess the accuracy of the ligand predictions for MHC-I molecules and compare with the state-of-the-art methods, such as NetMHCpan, MHCflurry and BigMHC, we performed the leave-one-allele-out cross-validation, where each allele absent from the training of NetMHCpan4.1, MHCflurry2.0 and BigMHC was successively removed from the training set of MixMHCpred3.0 (30, 10, and 31 alleles, respectively). A 4-fold excess of negatives was added randomly from the human proteome with uniform length distribution from 8 to 14 (Fig. 4A, B, and C, Additional file 1: Table S1). A similar process was carried out using a 99-fold excess of negatives (Additional file 1: Fig. S6).

In the second benchmark based on full HLA-I peptidomics samples, we used HLA-I peptidomics datasets coming from 10 meningioma samples measured in [7] and 10 HLA-I peptidomics samples from [23] that were not part of the training of any version of MixMHCpred, NetMHCpan, MHCflurry nor BigMHC. To these, we added a third dataset that comprises twenty recently published HLA-I peptidomics samples from COVID-19 patients [52]. In their paper, the full HLA-I typing was not provided for each sample, so we ran our motif deconvolution tool (MixMHCp) [7] to annotate the alleles to 4 digits typing in each sample excluding sample "UPN17" due to ambiguity in HLA-I annotation. All peptides from a given sample were used together with the set of alleles describing this sample and considered as positives and we added four times more random peptides from the human proteome as negatives. The scores for all peptides across all alleles were calculated, and the best score among the alleles of a sample was retained (%Rank_bestAllele for MixMHCpred, lowest %Rank_EL for NetMHCpan, presentation_percentile for MHCflurry, highest BigMHC_EL score for BigMHC). Using the predicted scores for each peptide, the AUC and PRAUC were computed separately for each predictor and sample, providing a performance evaluation and comparison with existing state-of-the-art methods (see Fig. 4D).

Benchmarking mouse alleles

A 5-fold cross-validation was performed for each of the eight mouse MHC-I alleles by randomly removing one-fifth of the positive peptides for each length (8–14 mers) before building the binding motifs. The remaining one fifth of positives was used in the test set further adding a 4-fold excess of random negative peptides to this test set.

PRIME2.1 benchmarking

PRIME was retrained with the scores provided by MixMHCpred3.0, resulting in the updated version, PRIME2.1. The benchmarking of PRIME2.1 based on 10-fold cross-validation presented in Fig. 4E used exactly the same data as in the PRIME2.0 publication [26].

Results

MHC-I peptidomics data enables predictions of binding specificity for MHC-I alleles without known ligands

To characterize MHC-I binding motifs across multiple alleles and species, we first collected experimental data from a large compendium of MHC-I peptidomics studies [23, 26, 29, 30, 36–41]. These studies encompass MHC-I ligands from human, mouse, cattle, canid, and non-human primate MHC-I alleles. Motif deconvolution was performed on all samples to annotate ligands for the different alleles in each sample following our previously established procedure [7, 22, 26] (see the "Methods" section). This approach ultimately yielded a collection of 511,553 ligands (Fig. 1A) interacting with 143 MHC-I alleles (Fig. 1B and C, Additional file 1: Fig. S1, Additional file 2: Table S1). As expected, the vast majority of ligands and alleles came from humans (Fig. 1A, B). For each allele hundreds to thousands of ligands are available (Fig. 1C). From these data, binding motifs (mathematically represented with position weight matrices) and peptide length distributions were computed (see the "Methods" section). Distinct motifs were built for each peptide length, ranging from 8- to 14-mers (Fig. 1D), as ligands of varying lengths exhibit differences in their motifs.

We then used the binding motifs and peptide length distributions to train a pan-allele predictor of MHC-I ligands, referred to as MixMHCpred3.0. To this end, we first trained neural networks to predict binding motifs for each peptide length, as well as peptide length distributions (Fig. 1D, see the "Methods" section). These networks take as input the MHC-I binding site sequence (see the "Methods" section). In a second step, we integrated the output from these neural networks and computed a final presentation score and %rank of a peptide (Fig. 1D,

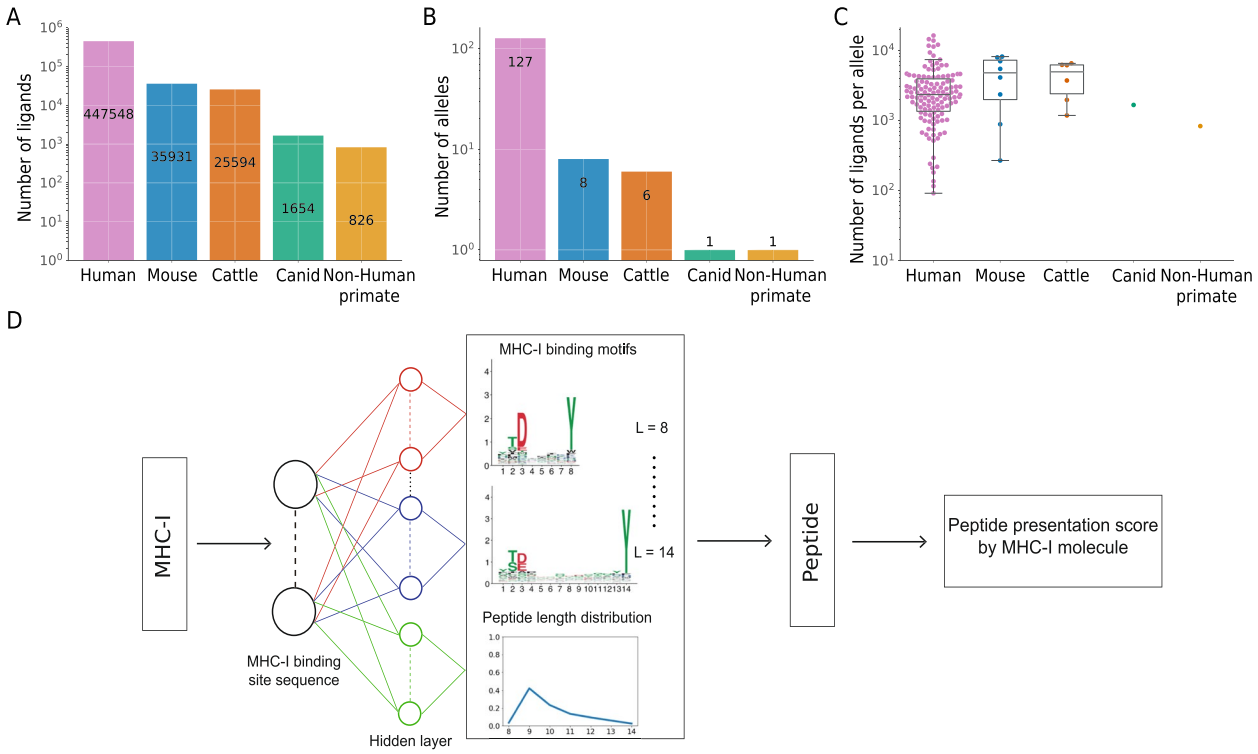

Fig. 1 Overview of the MHC-I peptidomics data used to predict MHC-I specificity and ligands for alleles without known ligands. **A** Number of MHC-I ligands for MHC-I molecules from different species collected in this work. **B** Number of MHC-I alleles with known ligands from different species. **C** Number of ligands per MHC-I allele across different species. **D** Description of the proposed architecture for predicting binding motifs and peptide length distribution (middle) and predicting peptide presentation scores (right). For the binding motif predictor, distinct neural networks were trained for each peptide length (from 8 to 14). An additional neural network was trained to predict peptide length distributions for MHC-I molecules without experimental ligands. The outputs of both predictors are combined in a final step to predict peptide presentation score and derive a %rank

see the "Methods" section). This framework enables us to predict ligands for any MHC-I allele. In addition, it can accommodate either predicted motifs or motifs directly computed from experimental ligands for alleles with such data. For such alleles, predictions MixMHCpred3.0 are therefore basically identical to those of the allele-specific predictor MixMHCpred2.2 [26], with the only differences coming from the inclusion of a few additional recent immunopeptidomic datasets not part of the training data of MixMHCpred2.2.

To test how reliably MHC-I ligand predictions can be extrapolated to MHC-I alleles without known ligands, we performed an extensive leave-one-allele-out (LOA) cross-validation. All ligands for each allele were iteratively excluded from the training of our model and used as a test set, and the model was trained on the motifs and length distributions from the remaining alleles (see the "Methods" section). A 99-fold excess of random peptides from the human proteome was used as negatives to compute the area under the curve (AUC) of the receiver operating characteristic curve (ROC). These AUC values serve as an indicator of the model's predictive power, with a value of 1 for a perfect predictor and 0.5 in the case of random predictions. Overall, the predictions were much better than random for all alleles (Fig. 2A). Predictions for human alleles also outperformed predictions for alleles in other species.

We then checked the accuracy of the predicted motifs and predicted length distributions, as an alternative to AUC for benchmarking. To this end, we computed the Euclidean distance between the predicted and actual 9-mer motifs (Fig. 2B, see the "Methods" section). For most human alleles, the predicted binding motifs were highly similar to the actual ones. In other species, the distances between predicted and actual motifs were higher. Similar observations were made when comparing the Euclidean distance between predicted and experimental peptide length distributions (Fig. 2C). These results demonstrate that both binding motifs and peptide length distributions for MHC-I alleles without known ligands can be accurately predicted in humans, and less so in other species, thereby providing a rational explanation for the LOA AUC values in Fig. 2A. To further visualize the quality of the model predictions and better interpret

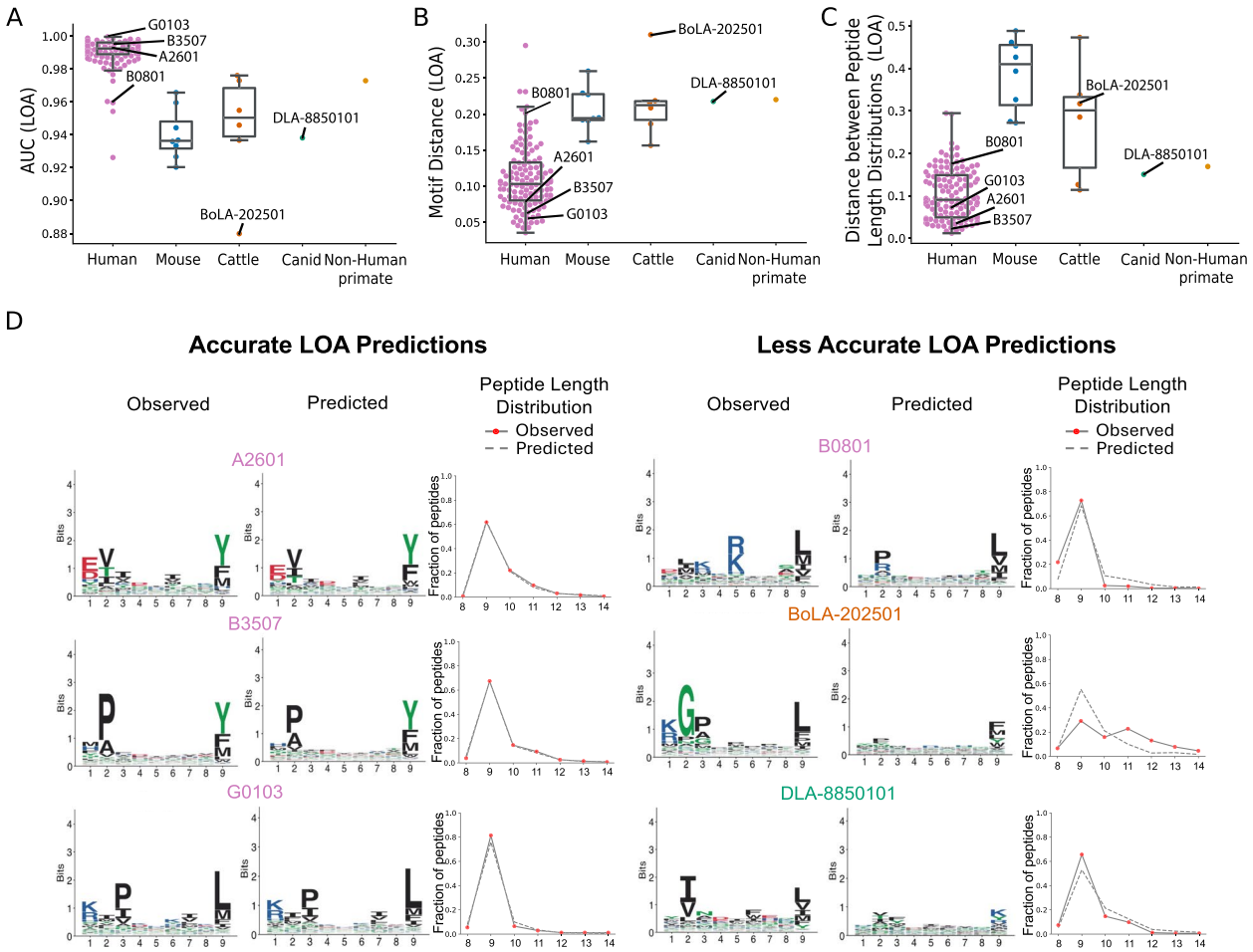

Fig. 2 Leave-one-allele-out benchmarking of MHC-I binding specificity predictions. **A** AUC for predictions of peptides presented by MHC-I molecules from different species, obtained in the leave-one-allele-out (LOA) cross-validation. **B** Euclidean distance between the predicted and experimental 9-mer motifs in the LOA cross-validation. **C** Euclidean distance between the predicted and experimental peptide length distributions in the LOA cross-validation. **D** Examples of predicted MHC-I binding motifs and peptide length distributions in a LOA context. Examples of 3 alleles with accurate predictions are shown on the left, and 3 alleles with less accurate predictions are shown on the right

binding motif and peptide length distribution distances, we selected three alleles with very high LOA AUC values and three alleles with much lower LOA AUC values (Fig. 2A). Figure 2D shows the comparison between the actual and the predicted 9-mer binding motifs and peptide length distributions. For alleles displaying very high LOA AUC (>0.98), we observed that binding motifs and peptide length distributions were indeed accurately predicted. For alleles with lower LOA AUC (0.88 to 0.95), we observed that the predicted binding motifs were less well predicted (Euclidean distances of 0.2 to 0.3) and much less specific, capturing mainly a weakly conserved specificity for hydrophobic residues at the last position (Fig. 2D). Similar observations could be made for peptide length distributions where the predicted peptide length distributions for alleles with lower LOA AUC were less accurate, while still capturing the preference for 9-mers (Fig. 2D). These results also indicate that fairly high LOA AUC values of up to 0.95 can be obtained with relatively unspecific motifs. This likely reflects the fact that many MHC-I alleles share some similarity (e.g., preference for 9-mers, preference for hydrophobic residues at the last position of their ligands), which leads to some predictive power even when failing to capture the actual specificity of each allele. Accurately predicted motifs and peptide length distributions corresponded to cases with LOA AUC values around 0.98 or higher.

Binding site similarity determines MHC-I ligand prediction accuracy

To explore the determinants of MHC-I ligand prediction accuracy for an allele without known ligands, we investigated the binding site similarity for alleles with known ligands. This binding site similarity was computed as a

binding site sequence distance (see Methods, Additional file 3: Table S1). For each allele, we identified the closest allele with known ligands in terms of binding site distance and referred to this distance to the closest allele with known ligands as the "binding site distance," for simplicity. We observed a strong inverse correlation between binding site distances and AUC values computed in the LOA cross-validation (Fig. 3A). As a rule of thumb, our data suggest that accurate MHC-I ligand predictions can be achieved when an allele shows a binding site distance smaller than 0.1 with another allele with known ligands. For bigger distances, predictions will generally be of lower accuracy. We further explored the relationship between the AUC computed in LOA cross-validation and the distance to the closest human allele with known ligands based on binding-site distance (Additional file 1: Fig. S2). As expected, this analysis reveals that the single MHC-I allele from non-human primates shows higher binding site similarity with human alleles than most mouse and cattle alleles. Considering the large bias towards human alleles in our data, this may explain the higher AUC values observed for this allele.

To explore whether predictions in the leave-one-allele-out setting are influenced by allele frequency, we compared our AUC values with the population frequencies for all alleles (Additional file 1: Fig. S3, Additional file 4: Table S1, see the "Methods" section). Overall, we observed a slight negative correlation. This can be explained by the fact that some common alleles (e.g., HLA-B*08:01, Fig. 3A) show low similarity with other alleles with known ligands, and therefore lower LOA AUC.

We then investigated to which extent predictions with MixMHCpred3.0 could be applied with high confidence across all known MHC-I alleles. To this end, we collected over 19,000 MHC-I protein sequences from humans and other species, extracted the binding site sequence for each allele, and computed its binding site distance with the closest allele with known ligands (see the "Methods" section, Additional file 5: Table S1). We then calculated for various species the fraction of alleles with a binding site distance lower than 0.1 (Fig. 3B). We observed that more than 97% of human HLA-I alleles passed this threshold and are therefore expected to be accurately predicted. The few alleles with binding site distances > 0.5 corresponded to HLA-F alleles, for which we did not have reliable ligands in our data. This suggests that the 127 human HLA-I alleles with available ligands provide a very good coverage of the specificity space of human MHC-I alleles, including most HLA-A, HLA-B, and HLA-C alleles. For mouse alleles, we observed that 48% of the MHC-I alleles met the threshold and these include all alleles from laboratory mouse strains, for which ligands are available. For other species, our data show that most alleles without MHC-I ligands do not pass the threshold on the binding site distance, suggesting that predictions of MHC-I ligands and motifs will be of lower accuracy.

To enhance our understanding of these limitations in predicting MHC-I ligands in non-human species, we explored two different scenarios underlying binding site distances larger than 0.1 (Fig. 3C). In the first scenario ("new amino acids"), some binding site positions display amino acids that are never found among alleles with known ligands (Additional file 1: Fig. S4). In these cases, predictions are complicated, since the training set of MixMHCpred3.0 does not contain information about these "unseen" amino acids. In the second scenario ("new arrangements"), we considered cases where all amino acids in the binding site are found in alleles with known ligands, but not in a single allele. Figure 3C shows examples of these two scenarios. In the first case

(See figure on next page.)

Fig. 3 Binding site similarity determines MHC-I ligand prediction accuracy. **A** Relationship between the accuracy of MHC-I ligand predictions (AUC in the LOA cross-validation) and binding site distance to the closest allele with known ligands. Regression line and Pearson correlation coefficients were added to the plot. **B** Boxplots of binding site distances to the closest allele with known ligands for all known alleles in different species groups. The numbers above each boxplot show the percentage of MHC-I sequences in each group reaching a binding site distance lower than 0.1 (the blue dashed line). Numbers in parentheses indicate the total number of MHC-I alleles with available sequences in each species group. Cyan dots indicate MHC-I alleles with known ligands. **C** Examples of different scenarios characterizing alleles with binding site distances larger than 0.1. The amino acid frequency for the binding site positions for alleles with known ligands is shown in the middle. An example of an allele without known ligands and having new amino acids (i.e., unseen among alleles with known ligands, written in light blue) in its binding site is shown above. An example of an allele without ligands and having a different arrangement of amino acids is shown below (with amino acids non-conserved in its closest allele with known ligands indicated in green). B-pocket positions are marked in dark blue and F-pocket positions in green. **D** Stacked barplots showing the percentage of alleles with binding site distance < 0.1 (orange), alleles with binding site distance ≥ 0.1 and new amino acids at some binding site positions (light blue), and alleles with binding site distance ≥ 0.1 and new arrangements of amino acids in their binding site (green). **E** Frequency of the new amino acids in species where all MHC-I alleles have new amino acids compared to MHC-I alleles with known ligands (i.e., Salmonids, Gallus, and Suids). **F** Representative 3D structure of the MHC-I binding site (HLA-A*01:01 in gray in complex with EADPTGHSY in yellow, PDB: 1W72), highlighting several positions with low conservation across species (see panel **E**); black dashed lines are the distances between these positions on the MHC structure and their closest residue on the peptide

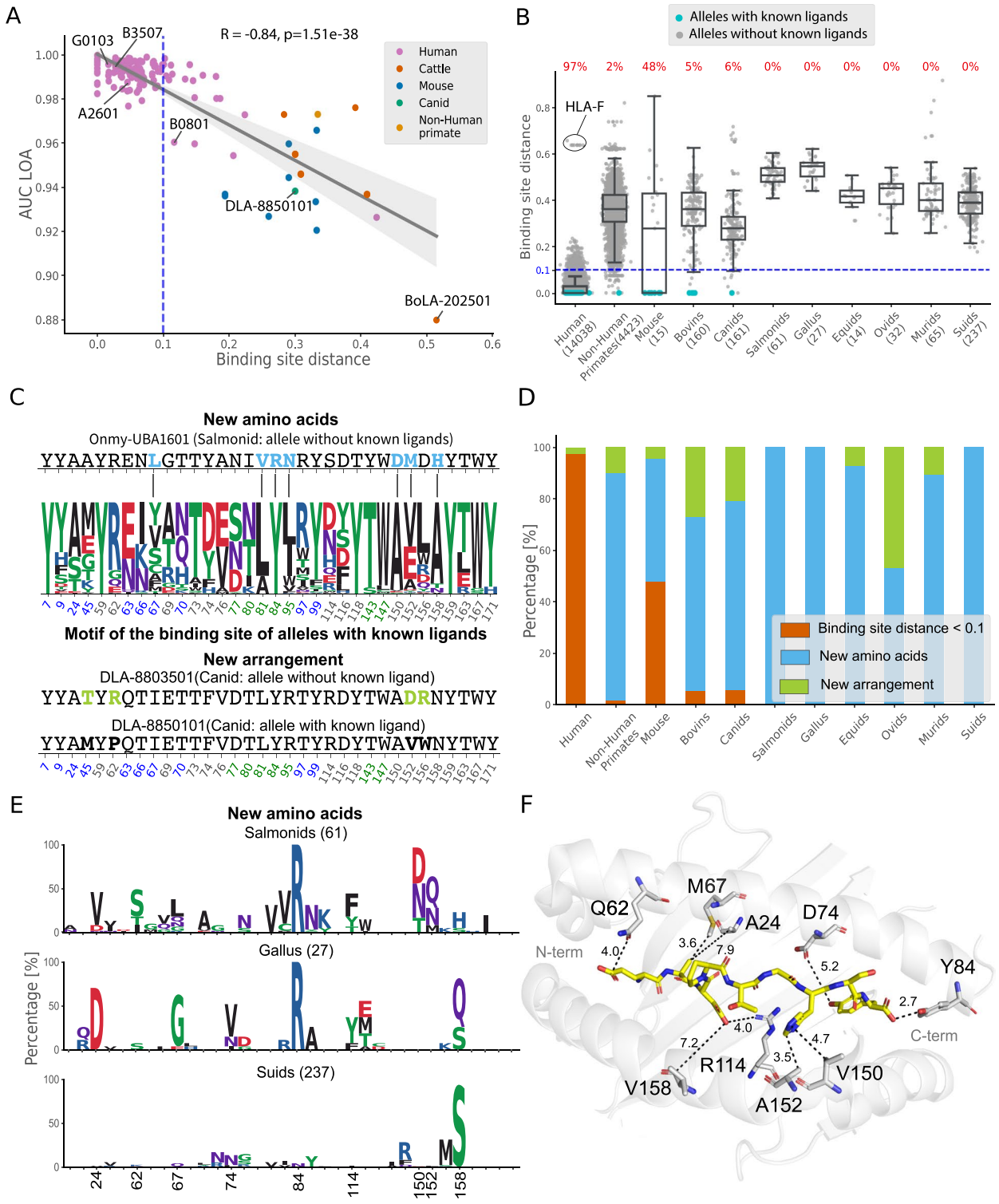

Fig. 3 (See legend on previous page.)

(Onmy-UBA1601, from Salmonids), we observed that several amino acids in the binding site (i.e., L67, V81, R84, N95, D150, M152, and H158) were not found among the alleles with known ligands. In the second case, (DLA-8803501 from canids), we observed that, even if all amino acids in the binding site are observed in

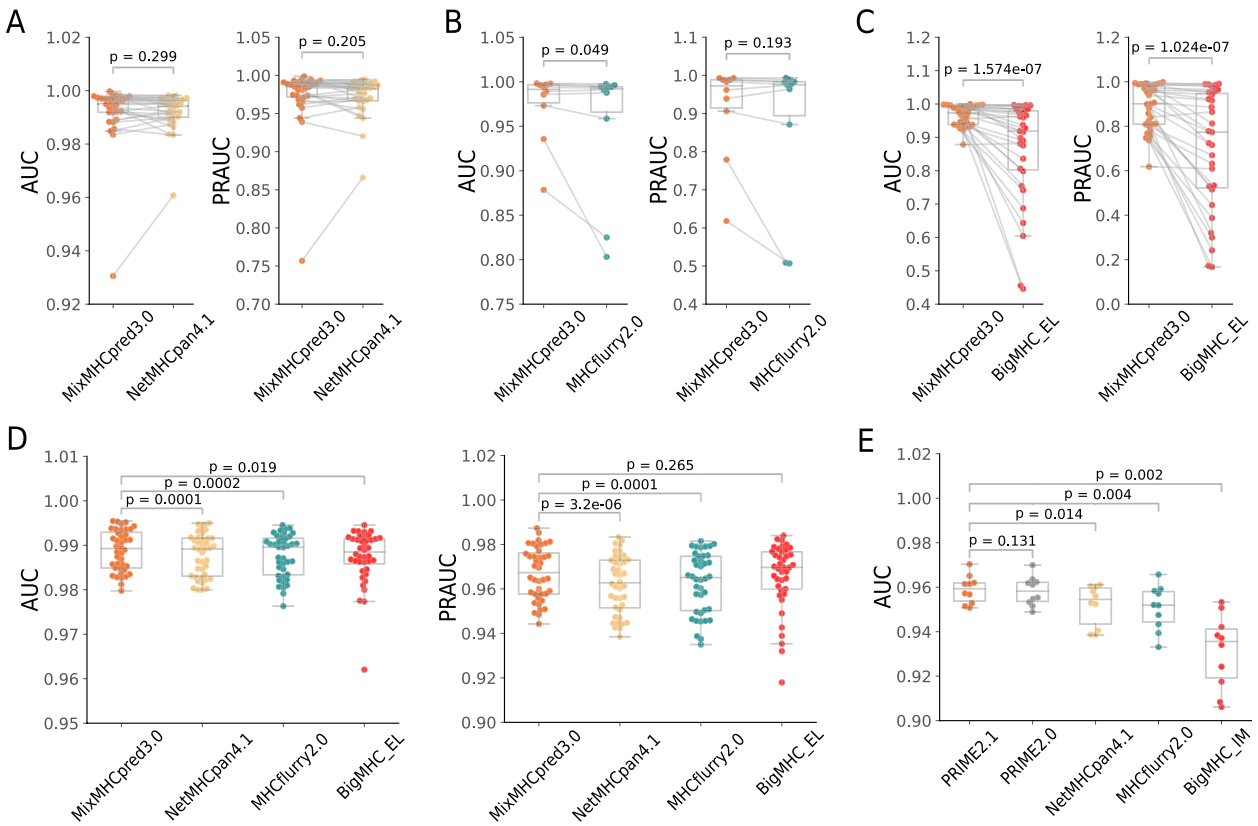

Fig. 4 MixMHCpred3.0 leads to high-quality MHC-I ligand predictions. **A**–**C** LOA cross-validation AUC and PRAUC values for the predictions of peptides presented by MHC-I molecules for alleles that are not part of the training set of **A** NetMHCpan, **B** MHCflurry or **C** BigMHC. **D** AUC and PRAUC values for the predictions of peptides presented by MHC-I coming from 40 samples in 3 different studies not included in the training set of any of the four predictors. **E** Benchmarking of PRIME2.1 on CD8 + T-cell epitope samples based on 10-fold cross-validation. *P*-values from a paired two-sided Wilcoxon signed rank test are indicated

alleles with known ligands, the closest allele with known ligands had different amino acids at 4 binding site positions (i.e., 45, 62, 152, 156). In general, we observed that most alleles with binding site distances larger than 0.1 displayed new amino acids in their binding site across all non-human species (blue bars in Fig. 3D), which likely explains why their binding motifs are difficult to predict. Figure 3E shows the frequency of the new amino acids in the MHC-I binding site in three species—Salmonids, Gallus, and Suids—where all alleles showed some amino acids absent in alleles with known ligands. This analysis shows for instance that all MHC-I alleles in Salmonids and Gallus have R84, while all alleles with known ligands in our training set had Y84. Figure 3F shows the structural location in the MHC-I binding site of these less conserved positions and suggests that new amino acids at these positions could alter MHC-I binding specificity. The potential impact of novel amino acids on MHC-I binding specificity is further supported by X-ray structures of MHC-I alleles for Salmonids or birds (Additional file 1: Fig. S5). Overall, these analyses suggest a molecular basis for the lower prediction accuracy observed in non-human species.

MixMHCpred3.0 leads to high-quality MHC-I ligand predictions

We then benchmarked our predictions for alleles without known ligands with two widely used pan-allele methods, NetMHCpan4.1 [28] and MHCflurry2.0 [27], and the recently introduced predictor BigMHC [35]. To this end, we first retrieved all alleles absent from the training sets of NetMHCpan (30 alleles in total), MHCflurry (10 alleles in total), or BigMHC (31 alleles in total) (Additional file 1: Table S1). We then retrained MixMHCpred3.0 excluding iteratively each of these alleles and tested the accuracy of the predictions on the ligands of the left-out alleles. A 4-fold excess of random peptides from the human proteome was used as negatives (Additional file 2: Table S1) to compute the AUC and Precision-Recall AUC (PRAUC) values. Our results indicate similar performance with NetMHCpan and MHCflurry (Fig. 4A and B), and improved performance compared to BigMHC

(Fig. 4C). Similar observations were obtained when using 99-fold excess of random negatives (Additional file 1: Fig. S6). In addition, when we performed a 5-fold cross-validation for each of the eight mouse MHC-I alleles (see the "Methods" section), we observed improved predictions for MixMHCpred3.0 compared with other predictors (Additional file 1: Fig. S7, Additional file 6: Table S1), which supports the model's strong performance on mouse alleles.

To further benchmark MixMHCpred3.0 on more realistic data, where most peptides come from alleles with known ligands, we employed three external datasets. The first dataset consists of ten HLA-I peptidomics samples from meningioma [7], the second one includes ten HLA-I peptidomics samples [23], and the third one comprises twenty recently published HLA-I peptidomics samples from COVID-19 [52] (Additional file 7: Table S1). To our knowledge, none of these datasets were used in the training of any predictor considered in this study. We employed a 4-fold excess of random peptides from the human proteome as negatives to compute AUC and PRAUC values (see Methods). MixMHCpred3.0 achieved significantly higher AUC than NetMHCpan4.1, MHCflurry2.0, and BigMHC, and significantly higher PRAUC than NetMHCpan4.1 and MHCflurry2.0 but not BigMHC (Fig. 4D). This demonstrates that MixMHCpred3.0 represents a state-of-the-art pan-allele MHC-I ligand predictor.

MixMHCpred3.0 accurately predicts CD8 + T-cell epitopes

To ensure compatibility between MixMHCpred3.0 and our immunogenicity predictor PRIME [26], we retrained PRIME with the scores provided by MixMHCpred3.0, resulting in the retrained version (PRIME2.1). To assess the impact of this integration on PRIME performance, we conducted a comprehensive benchmarking analysis. This analysis was designed to mirror the original validation methods described previously in [26] where we use CD8 + T-cell epitopes for benchmarking (Additional file 8: Table S1). Overall, we observed equal or better predictions compared to other tools (Fig. 4E). These findings confirm the successful integration of our updated method.

Discussion

$CD8^+$ T-cell recognition of peptides displayed on MHC-I molecules plays a central role in the immune recognition of infected or malignant cells. In this work, we capitalized on naturally presented MHC-I ligands derived from a diverse range of species, including human, mouse, cattle, canid, and non-human primate to explore how MHC-I ligand predictions can be extrapolated across alleles and species.

Our results show that predictions of MHC-I ligands can be accurately expanded to MHC-I alleles with low binding site distance with respect to alleles with known ligands. These cases encompass the vast majority of human MHC-I alleles, indicating that pan-allele predictors are likely to work well even across individuals from diverse genetic backgrounds. The main exception consists of HLA-F alleles for which a consensus on their motifs has not been reached [53, 54]. In other species, and especially in species with few known MHC-I ligands, predictions showed lower accuracy, and we expect also low accuracy for most species without documented MHC-I ligands. This results from lower accuracy in predictions of both MHC-I binding motifs and peptide length distributions. We can attribute this limitation to the lower MHC-I binding site conservation in these species. In particular, the binding sites of MHC-I alleles from several species included amino acids which were never seen in alleles with known ligands. These observations have implications for designing experiments aimed at improving the coverage of MHC-I alleles for which predictions of ligands can be made. In particular, we anticipate that MHC-I peptidomics profiling of alleles with binding site sequences including D24, S62, G67, R84, D/N150, or S/Q158 could reveal novel MHC-I binding motifs and expand our ability to accurately predict ligands for alleles across a broader range of species. Our work also shows that LOA AUC values are not positively correlated with allele frequencies. However, it is important to emphasize that this observation applies only in the leave-one-allele-out setting, where the task was to predict binding motifs and peptide length distributions when masking the actual ligands.

When performing our LOA cross-validation, we observed that AUC values up to 0.95 could be obtained even with relatively unspecific motifs (see examples in Fig. 2D). This suggests that using only AUC > 0.5 as a performance metric or success criteria does not guarantee that a pan-allele MHC-I ligand predictor has accurately learned the specificities of each allele. It also indicates that binding motifs and peptide length distributions, which are at the core of the pan-allele architecture of MixMHCpred3.0, provide a useful quality control to evaluate the ability of a pan-allele MHC-I ligand predictor to learn the actual specificity of different alleles.

When predicting class I epitopes, the number of peptides that are being scored is typically much higher than the number of actual epitopes. Therefore, very high AUC (e.g., > 0.98) are desirable to have enough true positives among the top predicted peptides which are typically considered for experimental validation. As an alternative to AUC values, PRAUC in the presence of a large (and therefore more realistic) proportion of negatives provides a

useful measure of the quality of the predictions, with more emphasis on the top predicted peptides. Across our different benchmarks, our observations were consistent when using AUC and PRAUC, and when using different fractions of negatives (4-fold and 99-fold excess from positives).

In most practical applications for epitope discovery, MHC-I ligand predictions are performed on human samples where the majority of MHC-I alleles have known ligands in existing databases. Consistent with previous results obtained with MixMHCpred2.2, the high-quality predictions of MixMHCpred3.0 in such cases indicate that MixMHCpred provides a state-of-the-art solution with high computational efficiency and direct interpretability in terms of binding motifs and peptide length distributions for most human MHC-I allele. Moreover, the compatibility with the PRIME framework [26] ensures that MixMHCpred3.0 can be used for $CD8^+$ T-cell epitope discovery.

Conclusions

Altogether, our work shows that very accurate predictions of MHC-I motifs and ligands can be reached for the vast majority of human MHC-I alleles, as well as MHC-I alleles of laboratory mouse strains, and reveals molecular determinants of prediction accuracy in other species. The pan-allele version of MixMHCpred developed to perform these analyses shows improved predictions compared to other tools and is available at https://github.com/GfellerLab/MixMHCpred.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01450-8.

Additional file 1: Table S1. Alleles with known ligands absent from the training sets of NetMHCpan, MHCflurry or BigMHC. Fig. S1. Curation of MHC-I peptidomics data reveals binding specificities for 143 MHC-I alleles. Fig.S2. Binding site sequence similarity to the closest human alleles. Fig.S3. Relationship between the HLA-I ligand predictions and the population frequencies for alleles with known ligands. Fig.S4. New amino acids for alleles without known ligands. Fig.S5. Examples of putative binding specificity changes. Fig.S6. Leave-one-allele-out benchmarking with 99-fold excess of random negatives. Fig.S7. Mean AUC and PRAUC values for 5-fold cross-validation of each mouse MHC-I allele.

Additional file 2: Table S1. MixMHCpred3.0 training data.

Additional file 3: Table S1. Leave-one-allele-out cross-validation predictions and binding site distance.

Additional file 4: Table S1. Leave-one-allele-out cross-validation predictions vs sample size.

Additional file 5: Table S1. Binding site similarity determines MHC-I ligand prediction accuracy.

Additional file 6: Table S1. 5-fold cross-validation benchmarking of mouse MHC-I alleles.

Additional file 7: Table S1. MixMHCpred3.0 benchmarking.

Additional file 8: Table S1. PRIME2.1 benchmarking.

Acknowledgements
We thank Matei Teleman and Aurelie AG Gabriel for testing the MixMHCpred tool.

Authors' contributions
D.T. performed the new methodological developments, wrote the manuscript and prepared the figures. J.R. provided feedback for the project and the manuscript. D.G. designed the project, supervised the work and provided feedback for the manuscript. All authors read and approved the final manuscript.

Funding
Open access funding provided by University of Lausanne The project was supported by the Swiss Cancer Research Foundation (KFS-4961-02-2020).

Data Availability
• MixMHCpred3.0 is available at https://github.com/GfellerLab/MixMHCpred and PRIME2.1 is available at https://github.com/GfellerLab/PRIME. • All data used inthis study is included in Additional Files 2 to 8. • Any additional information required to reproduce this work is available from the Lead Contact upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 13 June 2024 Accepted: 10 March 2025
Published online: 20 March 2025

References

1. Migliorini D, Dutoit V, Allard M, Grandjean Hallez N, Marinari E, Widmer V, et al. Phase I/II trial testing safety and immunogenicity of the multipeptide IMA950/poly-ICLC vaccine in newly diagnosed adult malignant astrocytoma patients. Neuro-Oncology. 2019;21(7):923–33. Available from: https://academic.oup.com/neuro-oncology/article/21/7/923/5316222 . Cited 2023 Mar 27.
2. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547(7662):217–21. Available from: http://www.nature.com/articles/nature22991 . Cited 2023 Mar 27.
3. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science. 2015;348(6236):803–8. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aaa3828 . Cited 2023 Mar 27.
4. Leidner R, Sanjuan Silva N, Huang H, Sprott D, Zheng C, Shih YP, et al. Neoantigen T-Cell Receptor Gene Therapy in Pancreatic Cancer. N Engl J Med. 2022;386(22):2112–9. Available from: http://www.nejm.org/doi/10.1056/NEJMoa2119662 . Cited 2023 Apr 5.
5. Tran E, Turcotte S, Gros A, Robbins PF, Lu YC, Dudley ME, et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science. 2014;344(6184):641–5. Available from: https://www.science.org/doi/10.1126/science.1251102 . Cited 2023 Apr 5.
6. Heitmann JS, Bilich T, Tandler C, Nelde A, Maringer Y, Marconato M, et al. A COVID-19 peptide vaccine for the induction of SARS-CoV-2 T cell immunity. Nature. 2022;601(7894):617–22. Available from: https://www.nature.com/articles/s41586-021-04232-5 . Cited 2023 Apr 5.
7. 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(12):3705–16.

8. Lundegaard C, Lund O, Nielsen M. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. Bioinformatics. 2008;24(11):1397–8. Available from: https://academic.oup.com/bioinformatics/article/24/11/1397/191077 . Cited 2024 May 2.
9. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 2016;8(1):33. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0288-x . Cited 2024 May 2.
10. 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(4):1480–7. Available from: https://journals.aai.org/jimmunol/article/196/4/1480/43223/The-Length-Distribution-of-Class-I-Restricted-T . Cited 2023 Apr 28.
11. Gfeller D, Bassani-Sternberg M. Predicting antigen presentation—what could we learn from a million peptides? Front Immunol. 2018;9:1716. Available from: https://www.frontiersin.org/article/10.3389/fimmu.2018.01716/full . Cited 2023 Mar 27.
12. Barker DJ, Maccari G, Georgiou X, Cooper MA, Flicek P, Robinson J, et al. The IPD-IMGT/HLA database. Nucleic Acids Res. 2023;51(D1):D1053-60. Available from: https://academic.oup.com/nar/article/51/D1/D1053/6814448 . Cited 2023 Apr 4.
13. Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol. 2003;16(3):363–77. Available from: https://onlinelibrary.wiley.com/doi/10.1046/j.1420-9101.2003.00531.x . Cited 2023 Apr 28.
14. Piertney SB, Oliver MK. The evolutionary ecology of the major histocompatibility complex. Heredity. 2006;96(1):7–21. Available from: https://www.nature.com/articles/6800724 . Cited 2023 Apr 28.
15. Sommer S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool. 2005;2(1):16. Available from: https://frontiersinzoology.biomedcentral.com/articles/10.1186/1742-9994-2-16 . Cited 2023 Apr 28.
16. Tadros DM, Eggenschwiler S, Racle J, Gfeller D. The MHC Motif Atlas: a database of MHC binding specificities and ligands. Nucleic Acids Res. 2023;51(D1):D428-37. Available from: https://academic.oup.com/nar/article/51/D1/D428/6786193 . Cited 2023 Mar 27.
17. Gfeller D, Liu Y, Racle J. Contemplating immunopeptidomes to better predict them. Semin Immunol. 2023;66:101708. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044532322001269 . Cited 2023 Mar 27.
18. Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. Nat Rev Immunol. 2011;11(12):823–36. Available from: http://www.nature.com/articles/nri3084 . Cited 2023 Mar 27.
19. Thibault P, Perreault C. Immunopeptidomics: reading the immune signal that defines self from nonself. Mol Cell Proteomics. 2022;21(6):100234. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1535947622000421 . Cited 2024 Mar 6.
20. Vyas JM, Van Der Veen AG, Ploegh HL. The known unknowns of antigen processing and presentation. Nat Rev Immunol. 2008;8(8):607–18. Available from: https://www.nature.com/articles/nri2368 . Cited 2024 Mar 6.
21. 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(2):315–26. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1074761317300420 . Cited 2023 Apr 12.
22. 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 allostery regulating HLA specificity. PLoS Comput Biol. 2017;13(8):e1005725.
23. Pyke RM, Mellacheruvu D, Dea S, Abbott CW, Zhang SV, Phillips NA, et al. Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation. Mol Cell Proteomics. 2021;12(20):100111.
24. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol. 2020;38(2):199–209. Available from: http://www.nature.com/articles/s41587-019-0322-9 . Cited 2023 Apr 28.
25. 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(1):55–63. Available from: https://www.nature.com/articles/nbt.4313 . Cited 2023 Apr 28.
26. Gfeller D, Schmidt J, Croce G, Guillaume P, Bobisse S, Genolet R, et al. Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. Cell Syst. 2023;14(1):72-83.e5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471222004707 . Cited 2023 Mar 27.
27. 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 Syst. 2020;11(1):42-48.e7.
28. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;48(W1):W449-54.
29. Nielsen M, Connelley T, Ternette N. Improved prediction of bovine leucocyte antigens (BoLA) presented ligands by use of mass-spectrometry-determined ligand and in vitro binding data. J Proteome Res. 2018;17(1):559–67. Available from: https://pubs.acs.org/doi/10.1021/acs.jproteome.7b00675 . Cited 2023 Mar 30.
30. 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(D1):D339-43. Available from: https://academic.oup.com/nar/article/47/D1/D339/5144151 . Cited 2023 Mar 30.
31. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics. 2016;32(4):511–7. Available from: https://academic.oup.com/bioinformatics/article/32/4/511/1744469 . Cited 2023 Apr 28.
32. Hu Y, Wang Z, Hu H, Wan F, Chen L, Xiong Y, et al. ACME: pan-specific peptide–MHC class I binding prediction through attention-based deep neural networks. Cowen L, editor. Bioinformatics. 2019;35(23):4946–54. Available from: https://academic.oup.com/bioinformatics/article/35/23/4946/5497763.
33. Ye Y, Wang J, Xu Y, Wang Y, Pan Y, Song Q, et al. MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. BMC Bioinformatics. 2021;22(1):7. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03946-z . Cited 2023 Apr 28.
34. Zeng H, Gifford DK. DeepLigand: accurate prediction of MHC class I ligands using peptide embedding. Bioinformatics. 2019;35(14):i278-83. Available from: https://academic.oup.com/bioinformatics/article/35/14/i278/5529131 . Cited 2023 Apr 28.
35. Albert BA, Yang Y, Shao XM, Singh D, Smith KN, Anagnostou V, et al. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. Nat Mach Intell. 2023;5(8):861–72. Available from: https://www.nature.com/articles/s42256-023-00694-6 . Cited 2024 Feb 21.
36. DeVette CI, Andreatta M, Bardet W, Cate SJ, Jurtz VI, Jackson KW, et al. NetH2pan: a computational tool to guide MHC peptide prediction on murine tumors. Cancer Immunol Res. 2018;6(6):636–44. Available from: https://aacrjournals.org/cancerimmunolres/article/6/6/636/468885/NetH2pan-A-Computational-Tool-to-Guide-MHC-Peptide . Cited 2022 Jul 27.
37. Ebrahimi-Nik H, Michaux J, Corwin WL, Keller GLJ, Shcheglova T, Pak H, et al. Mass spectrometry–driven exploration reveals nuances of neoepitope-driven tumor rejection. JCI Insight. 2019;4(14):e129152. Available from: https://insight.jci.org/articles/view/129152 . Cited 2022 Jul 27.
38. Faridi P, Woods K, Ostrouska S, Deceneux C, Aranha R, Duscharla D, et al. Spliced peptides and cytokine-driven changes in the immunopeptidome of melanoma. Cancer Immunol Res. 2020;8(10):1322–34. Available from: https://aacrjournals.org/cancerimmunolres/article/8/10/1322/466927/Spliced-Peptides-and-Cytokine-Driven-Changes-in . Cited 2024 Apr 26.
39. Lampen MH, Hassan C, Sluijter M, Geluk A, Dijkman K, Tjon JM, et al. Alternative peptide repertoire of HLA-E reveals a binding motif that is strikingly similar to HLA-A2. Mol Immunol. 2013;53(1–2):126–31. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0161589012003562 . Cited 2022 Jul 27.

40. Marcu A, Bichmann L, Kuchenbecker L, Kowalewski DJ, Freudenmann LK, Backert L, et al. HLA Ligand Atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. J Immunother Cancer. 2021;9(4):e002071.
41. Murphy JP, Yu Q, Konda P, Paulo JA, Jedrychowski MP, Kowalewski DJ, et al. Multiplexed Relative Quantitation with Isobaric Tagging Mass Spectrometry Reveals Class I Major Histocompatibility Complex Ligand Dynamics in Response to Doxorubicin. Anal Chem [Internet]. 2019 Apr 16 [cited 2025];91(8):5106–15. Available from: https://pubs.acs.org/doi/10.1021/acs.analchem.8b05616.
42. Racle J, Michaux J, Rockinger GA, Arnaud M, Bobisse S, Chong C, et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat Biotechnol. 2019;37(11):1283–6.
43. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics. 2017;33(22):3645–7.
44. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Valencia A, editor. Bioinformatics. 2020;36(7):2272–4. Available from: https://academic.oup.com/bioinformatics/article/36/7/2272/5671693. Cited 2024 Mar 12
45. Racle J, Guillaume P, Schmidt J, Michaux J, Larabi A, Lau K, et al. Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. Immunity. 2023;S1074761323001292. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1074761323001292. Cited 2023 Apr 28.
46. 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):1–13.
47. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89(22):10915–9. Available from: https://pnas.org/doi/full/10.1073/pnas.89.22.10915 . Cited 2023 Apr 3.
48. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J, et al. IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res. 2017;45(D1):D860-4. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1050 . Cited 2023 Apr 4.
49. The UniProt Consortium, Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9. Available from: https://academic.oup.com/nar/article/49/D1/D480/6006196. Cited 2023 Apr 4.
50. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf436 . Cited 2024 Feb 14.
51. Gonzalez-Galarza FF, McCabe A, Santos EJMD, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Res. 2019;gkz1029. Available from: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz1029/5624967. Cited 2025 Jan 24.
52. Nelde A, Rieth J, Roerden M, Dubbelaar ML, Hoenisch Gravel N, Bauer J, et al. Increased soluble HLA in COVID-19 present a disease-related, diverse immunopeptidome associated with T cell immunity. iScience. 2022;25(12):105643. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2589004222019150. Cited 2023 Mar 28.
53. Dulberger CL, McMurtrey CP, Hölzemer A, Neu KE, Liu V, Steinbach AM, et al. Human leukocyte antigen F presents peptides and regulates immunity through interactions with NK cell receptors. Immunity. 2017;46(6):1018-1029.e7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1074761317302297 . Cited 2024 Jun 4.
54. Hò GGT, Heinen FJ, Blasczyk R, Pich A, Bade-Doeding C. HLA-F allele-specific peptide restriction represents an exceptional proteomic footprint. IJMS. 2019;20(22):5572. Available from: https://www.mdpi.com/1422-0067/20/22/5572 . Cited 2024 Jun 4.

Publisher's Note