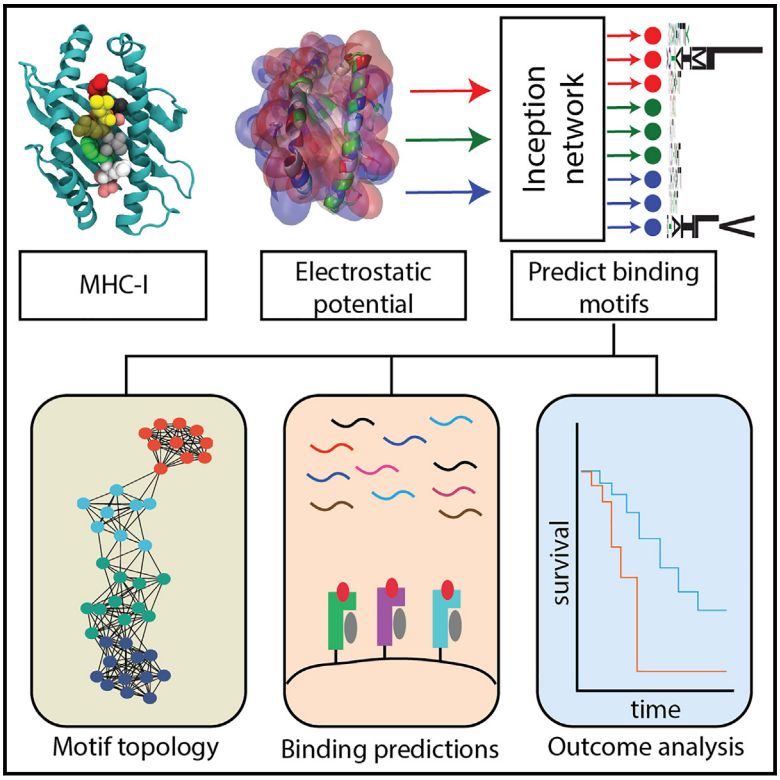
Cell Systems

Article

The electrostatic landscape of MHC-peptide binding revealed using inception networks

Graphical abstract



Highlights

- Inception networks trained on electrostatic maps predict MHC-I peptide binding motifs
- HLA-Inception predicts MHC-I binding motifs for 5,821 MHC-I alleles
- Inception network allows proteome-scale MHC-I binding predictions in seconds

Authors

Eric Wilson, John Kevin Cava, Diego Chowell, ..., Marion Curtis, Karen S. Anderson, Abhishek Singharoy

Correspondence

karen.anderson.1@asu.edu (K.S.A.), asinghar@asu.edu (A.S.)

In brief

Wilson et al. leverage inception-based convolutional neural networks to predict MHC-I binding motifs for 5,821 alleles using only protein surface electrostatic potentials. Predicted motifs were used to generate highly precise and efficient MHC-I peptide binding predictions and disease stratification.

Wilson et al., 2024, Cell Systems *15*, 1–12
April 17, 2024
https://doi.org/10.1016/j.cels.2024.03.001

CellPress

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

Cell Systems

CellPress

Article

The electrostatic landscape of MHC-peptide binding revealed using inception networks

Eric Wilson,[1,2] John Kevin Cava,[3] Diego Chowell,[2] Remya Raja,[4] Kiran K. Mangalaparthi,[5] Akhilesh Pandey,[5,6,7] Marion Curtis,[4,8,9] Karen S. Anderson,[10,*] and Abhishek Singharoy[1,11,*]

[1]School of Molecular Sciences, Arizona State University, Tempe, AZ 85207, USA
[2]The Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85207, USA
[4]Department of Immunology, Mayo Clinic, Scottsdale, AZ 85259, USA
[5]Department of Laboratory Medicine and Pathology, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA
[6]Center for Individualized Medicine, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA
[7]Manipal Academy of Higher Education, Manipal 576104, Karnataka, India
[8]College of Medicine and Science, Mayo Clinic, Scottsdale, AZ 85259, USA
[9]Department of Cancer Biology, Mayo Clinic, Scottsdale, AZ 85259, USA
[10]School of Life Sciences, Arizona State University, Tempe, AZ 85207, USA
[11]Lead contact
*Correspondence: karen.anderson.1@asu.edu (K.S.A.), asinghar@asu.edu (A.S.)
https://doi.org/10.1016/j.cels.2024.03.001

SUMMARY

Predictive modeling of macromolecular recognition and protein-protein complementarity represents one of the cornerstones of biophysical sciences. However, such models are often hindered by the combinatorial complexity of interactions at the molecular interfaces. Exemplary of this problem is peptide presentation by the highly polymorphic major histocompatibility complex class I (MHC-I) molecule, a principal component of immune recognition. We developed human leukocyte antigen (HLA)-Inception, a deep biophysical convolutional neural network, which integrates molecular electrostatics to capture non-bonded interactions for predicting peptide binding motifs across 5,821 MHC-I alleles. These predictions of generated motifs correlate strongly with experimental peptide binding and presentation data. Beyond molecular interactions, the study demonstrates the application of predicted motifs in analyzing MHC-I allele associations with HIV disease progression and patient response to immune checkpoint inhibitors. A record of this paper's transparent peer review process is included in the supplemental information.

INTRODUCTION

Major histocompatibility complex I (or MHC-I) protein plays an integral role in permitting the immune surveillance of host cells, viral clearance, and tumor rejection. The MHC-I protein, also denoted as human leukocyte antigen (HLA), drives molecular recognition by presenting endogenous peptide fragments on the cell surface for interaction with T cell receptors (TCRs) from the $CD8^+$ T cells.[1] Such processing and presentation pathways have the remarkable ability to present peptides from virtually any expressed cytosolic protein to the circulating T cells. Therefore, MHC-I is an ideal system to study molecular recognition, which accommodates thousands to millions of protein-peptide interactions, and it offers an opportunity to bridge these details to phenotypic outcomes and biomedical consequences.

Illustrated in Figure 1A, the MHC-I protein canonically binds peptides of lengths between 8 and 14 amino acids.[2] The second or third and C-terminal residues of the peptide ligand are commonly referred to as "anchor" positions. The identity of the anchor residues remains highly conserved in strong peptide binders.[3,4] The amino acid distribution at each position of the peptides binding to an MHC-I variant exhibits a consistent amino acid signature, commonly referred to as the peptide binding motif (Figure 1A).[5] Once presented on the cell surface, the peptide-loaded MHC-I complex interacts with $CD8^+$ T cell via the TCR. Stably bound peptide ligands that sufficiently diverge from the naturally presented host peptides, such as those derived from viral or mutated proteins, can trigger an immune reaction.[6]

The cell-specific nature of MHC-I-mediated immune responses is exploited by promising anticancer immunotherapies.[9,10] However, such therapies rely on the ability to identify peptide targets from an antigen of interest. While recent advances in high-throughput tumor immunopeptidome characterization techniques have enabled experimental verification of possible MHC-I targets,[11,12] the associated costs are still prohibitive for broad clinical application. A major obstacle in predicting the binding of peptides to MHC-I stems from the diversity of this protein in the human population. The protein, despite its similar structure, is encoded by one of the most polymorphic

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

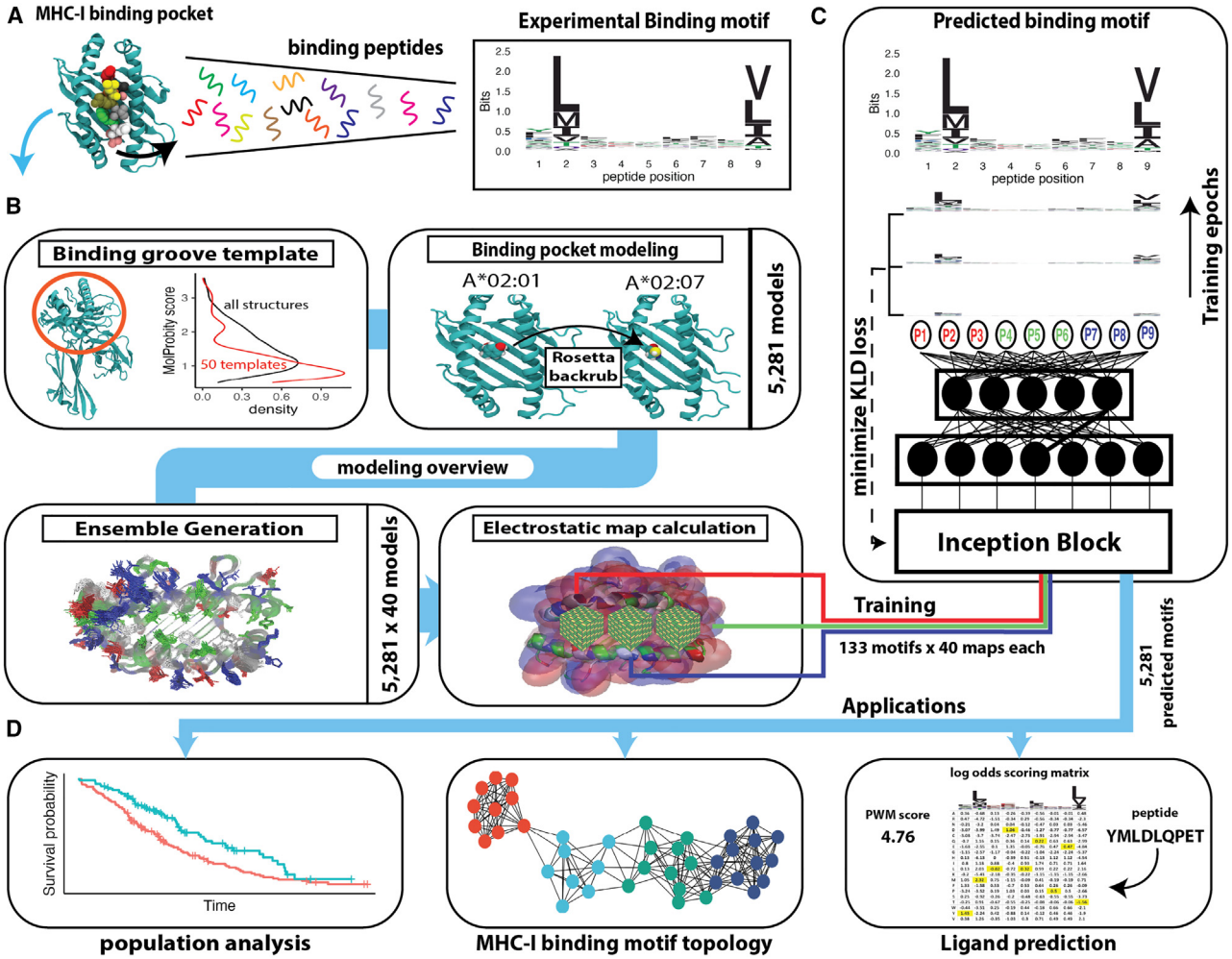


Figure 1. Overview of HLA-Inception

(A) Schematic of MHC-I complex and MHC-I binding motif.

(B) Molecular models and associated binding pocket electrostatic potentials of MHC-I proteins were constructed in a 4-step process. First, MHC-I structural templates are selected based on the Molprobity score, a measure of structure quality. Second, MHC-I variants without existing structures are modeled using the best aligning template structure. Third, an ensemble of 40 binding pockets is generated via side-chain rotamer sampling using the Rosetta simulation software.[7] Fourth, the electrostatic potential of the MHC-I binding pocket is calculated for each ensemble member using APBS.[8]

(C) A schematic representation of the inception-based CNN that was trained on MHC-I binding pocket electrostatic potentials in order to predict binding motifs.

(D) Three example uses for the predicted binding motifs.

genes in the human genome with over 24,000 known sequence variations.[13] The characterization of binding interactions across such diversity proves to be a formidable challenge as even single-point mutations in the binding pocket can lead to altered MHC-I peptide binding motifs.[14] *In vitro* binding assessments[15] and mass spectrometry-based profiling of cell lines[12] or tumors[16] have provided crucial data for training MHC-I peptide prediction algorithms. Still, binding preferences of the vast majority of MHC-I alleles remain unresolved, with the binding data from only 205 variants available in the public databases.[17] Consequently, computational interventions provide a critical solution for identifying potential MHC-I peptides from the antigens of interest.[12,18–20]

The diversity of the MHC-I protein was initially tackled through the definition of "MHC-I supertypes" or clusters of MHC-I alleles assumed to produce similar MHC-I binding motifs.[21] Recently, this diversity has been addressed through the development of machine learning algorithms that perform pan-allele sequence predictions.[12,18–20] An example of a very widely used pan-allele prediction algorithm is the netMHCpan suite, which is a sequence-based algorithm that supports the predictions of over 11,000 different human and non-human MHC-I alleles.[20] While such sequence-based prediction methods are useful, their binding predictions do not explicitly account for observed physical properties of the MHC-I binding pocket. Therefore, single amino acid changes that may impact the physical environment may be equally weighted with another sequence substitution that preserves the overall state. Attempts to resolve this discrepancy have involved numerically encoding amino acids with molecular properties, such as hydrophobicity

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

and the BLOSUM62 substitution matrix scores.[12,20] These augmented descriptions of the residues improve the predictive properties of the models; however, they weakly track with geometric alternations in the binding pocket topology or side-chain orientations arising from MHC-I polymorphism.[22]

Protein-peptide binding is primarily driven by long-range electrostatic interactions.[23–25] The kinetics of this process is also controlled by the shorter-range interactions, sequence, as well as shape complementarity of the binding partners at the interface.[26] Quantitative molecular simulations of such complementarity depend on the accuracy of the molecular model, treatment of protonation equilibria, high-resolution rotamer sampling, geometry optimization, and explicit modeling of the *apo* and *holo* states.[27] Despite this knowledge, repeating detailed molecular computations across the allelic diversity of the peptide-MHC-I complexes is extremely expensive, if not prohibitive. Here, we simplify this complex recognition problem into terms of sequence, geometric, and electrostatic variables to elucidate the binding patterns of typical peptide-MHC interfaces from a finite training set of known structures and bonding motifs. Thereafter, the interaction signatures are extrapolated to infer binding across the population of human MHC-I variants.

By analyzing this fundamental immunological process through the lens of molecular electrostatics, we explore an algorithm for MHC-I peptide binding motif prediction that learns their complementarity relationship with the peptide binders, monitor how MHC-I polymorphisms alter the binding motifs, and generalize computations to a diverse set of alleles. This is accomplished by training an inception-based convolutional neural network (CNN), "HLA-Inception," for predicting MHC-I binding motifs based on their correlation with the three-dimensional electrostatic potential distribution of the MHC-I binding pocket. The predicted MHC-I binding motifs are then used to compute peptide binding and stability, study the granularity and heterogeneity of MHC-I motif networks, and to analyze phenotypic outcomes to MHC-I associated disease.

RESULTS

We start by examining whether the electrostatic potential of the MHC-I binding pockets bears signatures of their peptide binding motifs or distributions. Using molecular models of 133 unique MHC-I binding pockets, HLA-Inception is then trained to predict the peptide binding motifs of 5,821 MHC variants covering most of the human population, based on their electrostatic potentials alone (Figure S1). We then applied it to predicting natural MHC-I ligands and examined disease associations.

Electrostatic potential of MHC-I protein complements peptide binding motif variations

Electrostatic features are central to protein-protein binding.[28,29] We explore these features within the highly polymorphic family of peptide-MHC complexes to seek the collective signatures of peptide binding sequences on the surface of the three-dimensional MHC-I structures. A major roadblock to performing such an analysis of peptide motif diversity is the data available on peptide-MHC interactions, which are sparse relative to the total allelic variance. There are only 133 MHC-I alleles with binding data known for 50 peptides or greater and an even more limited number with known structures (Figure S2). So, by generating molecular models of the MHC-I binding pockets, coupled with knowledge on peptide-protein interactions derived from the 133 alleles with known binding motifs, we track the covariance between pocket features and peptide configuration. This mapping between binding pockets and peptide configurations sets up an algorithm for grouping the peptide diversity into a small number of functionally relevant binding motifs.

Figure 1B presents an overview of the multiresolution $sequence \rightarrow molecularensemble \rightarrow electrostaticsmap$ pipeline, used for generating a set of 5,821 MHC-I molecular models. Molecular models of 5,281 MHC-I binding pockets were constructed using a combination of biophysical simulations and homology modeling. In short, homology models for each of the MHC-I binding pockets were created using the best aligning MHC-I binding pocket sequence with an existing crystal structure as a template. The selected template structures were then mutated using the Rosetta modeling software. The majority of MHC-I models required four or less template mutations for modeling 5,821 MHC-I variants (Figure S3A). Following mutation, each newly created MHC-I binding pocket underwent 40 individual side-chain sampling procedures to create an ensemble of 40 confirmations of the MHC-I binding pocket. The variance between pocket features and peptide diversity is then tracked using the subset of 133 three-dimensional MHC-I molecular models out of this set (Figure S3). We define a metric called the "electrostatic potential distance" (EPD) for distinguishing between the MHC-I molecular models in the space of electrostatic features (Equation 1). An EPD between two MHC-I variants represents the Euclidean distance between two sets of binding pocket potential volumes. The electrostatic potential volumes were extracted near the N- and C-terminal binding pockets, and the pairwise EPDs were computed between all the 133 alleles. Similarly, to describe the diversity between peptide motifs (with 50 or more known binders) that are already known to bind these MHC-I pockets, the inter-motif Kullback-Leibler divergence (KLD) or "binding motif KLD" was computed (Equation 4).

Correspondence between the MHC-I EPD and binding motif KLD across the reference set of models, illustrated in Figure 2A, reveals a linear correlation between MHC-I and binding peptide motif diversity (Spearman's rho = 0.32; $p < 2.2 \times 10^{-16}$). As controls, two sequence-dependent distance metrics, namely the Hamming distance and BLOSUM62 alignment, were also employed to perform inter-allele comparisons. While the Hamming distances track weakly with the binding motif KLD, the more biochemically aware BLOSUM62 alignment improved the correlation coefficient from 0.17 (Spearman's rho; $p < 2.2 \times 10^{-16}$) to 0.24 (Spearman's rho; $p < 2.2 \times 10^{-16}$). Although the estimated correlations are weak, they indicate that biophysical information provides a relative boost in tracking with molecular diversity. Taken together, the incorporation of physical volumetric data beyond sequence information better tracks MHC-I protein changes with substantive alterations to the peptide binding motif.

Building on the covariance between MHC-I EPD and binding motif KLD, we wanted to establish whether the electrostatic description of MHC-I protein can be employed to classify their peptide binding motifs. A K-means clustering performed on

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

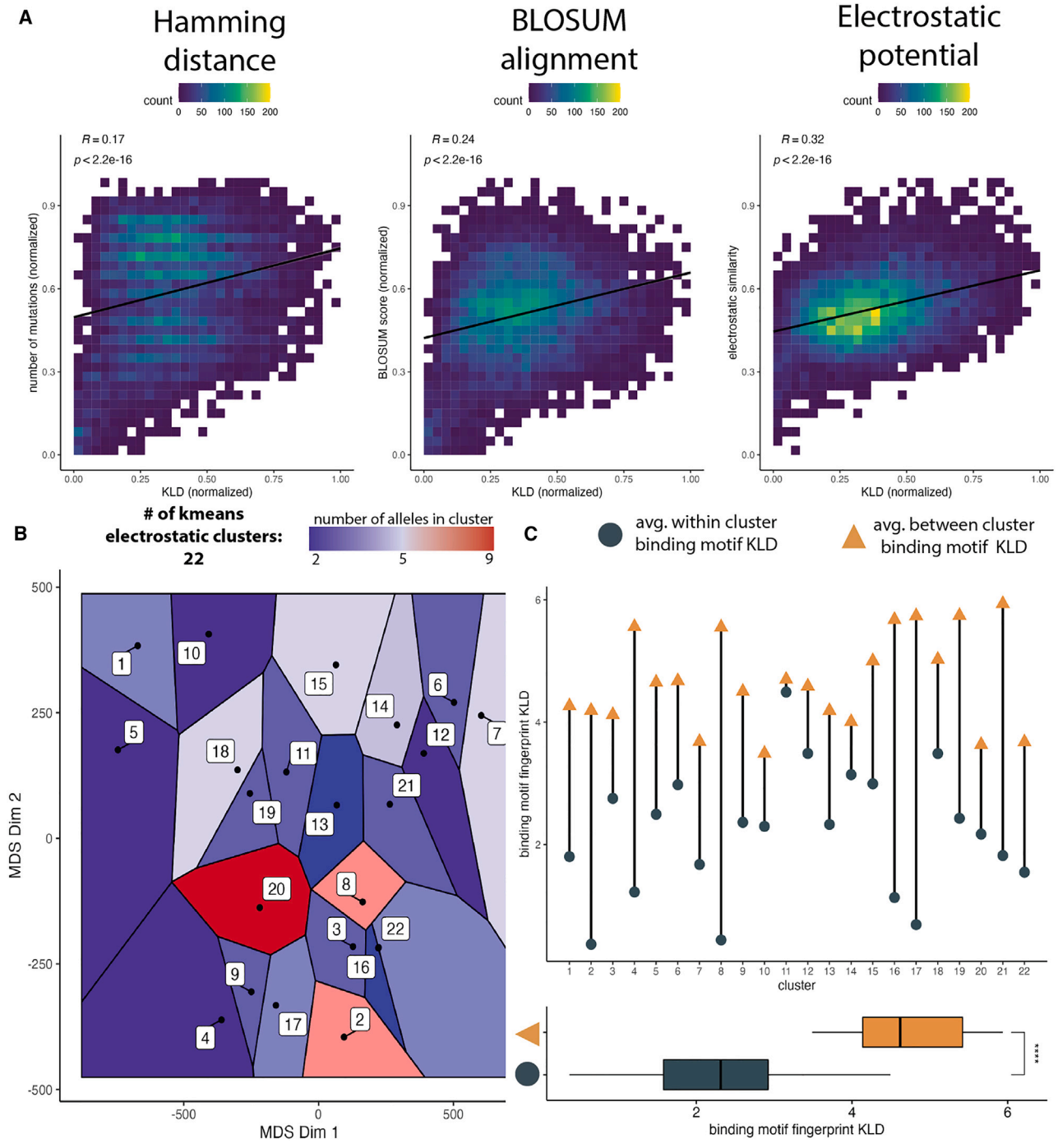


Figure 2. Electrostatic potential configurational space better captures MHC-I binding motif variation

(A) The correlation between different binding pocket variation measurement methods (indicated by header) and binding motif Kullback-Leibler divergence (KLD) for 133 MHC-I alleles with known binding motifs (n = 17,556; correlation coefficient: Spearman's rho). For visualization purposes, the pairwise data for each metric was aggregated into a 2D density plot.

(B) Spherical regions of electrostatic potential corresponding to N- and C-terminal anchor binding pockets were extracted, and the K-means clustering method was applied to find the optimal number of clusters where each cluster had at least two assigned alleles. Multidimensional scaling was performed on the concatenated electrostatic data, and the position of each cluster was defined as the arithmetic average of cluster members. The reduced space was then visualized using a Voronoi diagram. Each Voronoi cell is labeled with the respective cluster number, and the fill color indicates the number of alleles assigned to each cluster (cluster size).

(C) The average binding motif KLD between alleles within the same cluster (circle) identified in (B) was compared with the pairwise binding motif KLDs of every allele outside that cluster (triangle). Significance was determined using Wilcoxon rank-sum test.

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

the electrostatic potential grids extracted for EPD calculation revealed 22 unique clusters (Figures 2B and S4). Notably, the MHC-I variants with comparable peptide binding motifs are grouped together, with the average separation within their clusters being less than those between the clusters (Figure 2C). Therefore, the electrostatic potential features analyzed across a set of MHC-I proteins can be indicative of the signature distributions of amino acids that compose a peptide binding motif.

Inception model trained on electrostatic features detects the heterogeneity of binding motifs

While the EPD values successfully tracked with variations in MHC-I binding motifs, such analysis is inherently limited by the paucity of experimental data. To address this limitation, we developed a deep learning model, named HLA-Inception. This CNN model is trained on three-dimensional representation of the electrostatic potential within the MHC-I binding pocket (Figure 1C) to reproduce known peptide binding motifs. Following training, the neural network was applied to the generated structures to estimate binding motifs for MHC-I structures with unknown peptide interaction signatures.

Our approach works by segmenting an MHC-I binding pocket electrostatic potential grid into a number of volumes. Here, we have chosen three equal-sized volume segments of the peptide binding pocket—the region corresponding to the N-terminal binding pocket, the TCR contact region, and the C-terminal binding pocket. The N-terminal segment was used to predict amino acid distribution for peptide positions 1–3, the TCR contact segment monitors positions 4–6, and the C-terminal segment probes positions 7–9. The segmented electrostatic volumes of the MHC-I pocket at each of these peptide positions are passed through the CNN, using an output layer of amino acid distributions at the respective peptide positions (P1–P9). Using the training set of 133 MHC-I binding pocket models with experimentally verified peptide binding motifs, each CNN was trained to replicate the known amino acid distribution at a specific peptide position. Due to the overall importance of positions 2 and 9 to peptide binding, subsequent hyperparameter tuning was focused on these positions (Figure S5).

Following model training with 5,320 unique maps (ensemble of 40 conformations/allele × 133 alleles with known binders), the HLA-Inception model was applied to the average electrostatic potential maps from all generated MHC-I structures (Figures 1B, 1C, and S1). As an initial check of quality, we found that the binding motifs predicted from the average electrostatic maps of the 133 training alleles were in close agreement with the experimentally determined peptide binding motifs (Figure S6A). Thereafter, using the binding motifs data from only these 133 variants, our model assigned all the 5,821 MHC-I alleles.

The macroscopic arrangement and structure of the MHC-I binding motif space was investigated using a force-directed graph, where each node represents a different MHC-I supertype,[21] and an edge represents the minimum KLD between the binding motifs of the corresponding supertypes (Figures 3A and S7).

Supplementing the traditional supertype classifications (945 alleles), we assigned each of the 5,821 MHC-I alleles to an MHC-I supertype, based on the minimum KLD between the HLA-Inception-predicted binding motif and that of a supertype node in Figure 3A. Following graph generation, the heterogeneity of submotifs within each supertype was explored through community structure analysis.[30] MHC-I supertypes marked by high submotif heterogeneity indicate a larger number of distinct binding submotifs, whereas low heterogeneity supertypes suggest fewer and more homogeneous submotifs. Notably, the population or size of an MHC-I supertype does not always imply its heterogeneity (Figures 3B and S8A). While populated supertypes like B07 and B44 represent heterogeneous communities of motifs, a similar-sized A02 node offers a much more homogeneous distribution (Figure 3C). Similarly, the more populous C06 augmented supertype is in fact less heterogeneous than B07 and B44, and a smaller-sized A01 supertype is more heterogeneous than the larger A02.

MHC-I supertypes are important for generalized vaccine development.[31] Based on our classification, the supertypes marked by high heterogeneity (e.g., B44) exhibit a broad and more distinct range of peptide binding submotifs, while the homogeneous supertypes (e.g., A02) have a sharper distribution across similar numbers of member MHC-I alleles (Figures S8B–S8D). Importantly, within supertypes of high heterogeneity, there is a significantly larger loss of average predicted binding affinity when performing intra-supertype cross-allele binding predictions that is not observed in more homogeneous supertypes (Figure S8C). This observation of supertype heterogeneity will have important implications for supertype-level peptide vaccine design, as therapeutic peptides targeted to bind to supertypes with high heterogeneity might not equally cover all member alleles. Taken together, the electrostatic augmentations to classical supertype via HLA-Inception bring to light unforeseen topological details of peptide-MHC complexes.

Integration of electrostatics with sequence information offers precise pan-allele MHC-I peptide ligand prediction

The extreme polymorphism of the MHC-I protein typically results in the need to identify peptide targets for MHC-I alleles without experimentally resolved peptides. This need has brought forth pan-allele prediction algorithms, which leverage information from alleles with known binders to extrapolate to the unknown ones.[12,18–20,32] Here, we perform pan-allele peptide prediction by employing the peptide binding motifs derived from HLA-Inception to define a position-weighted matrix (PWM) scoring system.[33]

PWM score utilizes log odds ratios of observing an amino acid at a particular position to determine how well a peptide fits a binding motif for a given allele. Therefore, peptides characterized with a high PWM score are implied to have a high probability of stable binding. To assess PWM as a peptide binding metric, peptides with quantitative binding estimates, namely peptide binding affinity (IC50) and MHC-I stability (minutes), were ranked based on the PWM score. We found that PWM scores were associated with binding estimates (Figure 4A). PWM scores had an absolute correlation coefficient of 0.62 (Spearman's rho; $p < 2.2 \times 10^{-16}$) with MHC-I stability data and a -0.62 (Spearman's rho; $p < 2.2 \times 10^{-16}$) correlation with MHC-I affinity. These results suggest that the most probable binders determined from our algorithm are also found to be strong binders,

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

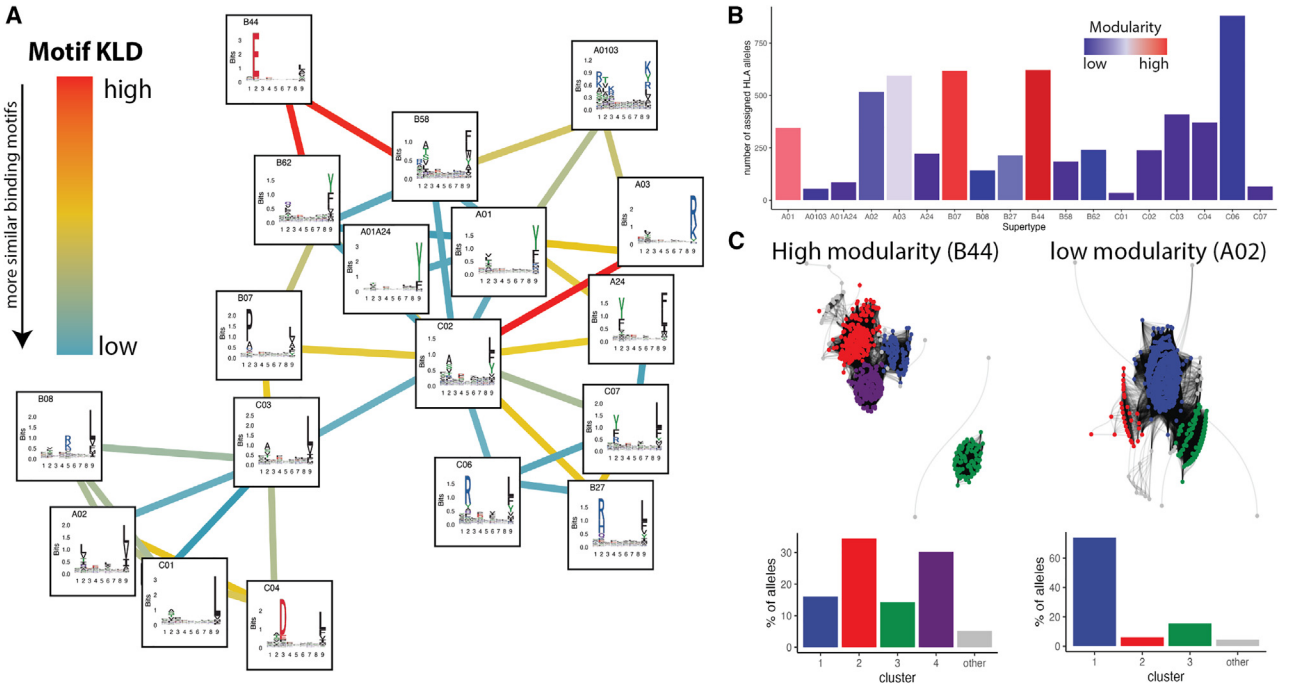


Figure 3. Exploration of predicted binding motifs

(A) A force-directed network of supertype binding motifs was constructed. Each node indicates a different MHC-I supertype with the binding motif indicated by the inset logo plot. Each node was then connected to the three most similar MHC-I supertype binding motifs as determined by binding motif KLD with any edge defining a binding motif KLD of ≥ 4 also being removed. Warm colors indicate dissimilar binding motifs, while cooler colors indicate more similar motifs (arbitrary units).

(B) The bar plot indicates the number of the 5,821 MHC-I alleles (y axis) assigned to each MHC-I supertype (x axis). All alleles assigned to a given supertype were then placed into a network, and communities within the network were determined using the infomap algorithm (STAR Methods). The modularity of each supertype graph, a measure of connection density within each community, is indicated by the color of the bar (arbitrary units).

(C) Two supertypes, A02 and B44, with similar graph size but a significant difference in modularity, are shown. Large clusters are colored as indicated by the bar plot below.

even though no quantitative protein-peptide interaction data were used to train the model. The 62% correlation indicates that the inception network has learned to capture the strengths of their molecular interaction with the MHC-I binding pocket, leveraging only information about the MHC-I binding pocket environment and overall binding motif of the peptides.

The precise identification of peptides that are likely to be naturally presented on the cell surface has high clinical value for the development of T cell-based immunotherapies. However, this task necessitates the consideration of many potential peptides where the number of presented peptides is greatly outnumbered by the number of possible peptides, often leading to high false positive rates.[34] So, we examined the precision of HLA-Inception on the recovery of naturally presented peptide ligands. The target peptides used for this analysis were extracted from a dataset of MHC-I peptides determined by mass spectrometry to be naturally presented[35] and were combined with an excess of decoy peptides that were not detected by mass spectrometry, resulting in a final target-decoy ratio of 1:100. The performance of HLA-Inception-based predictions was compared with several of the current state-of-the-art MHC-I peptide prediction algorithms: MHCflurry2.0,[18] NetMHCpan-4.1,[20] and MixMHCpred2.2.[32] Peptide selection was based on two commonly used score thresholds, 0.5% and 2%, where peptides with this score or lower can be considered to bind stronger than 99.5% and 98% of all possible peptides, respectively. We found that HLA-Inception achieved a median precision of 0.46 and 0.234 for the 0.5% and 2% threshold, respectively (Figure 4B). While the other algorithms achieved better recall (Figures S9A and S10C), they had significantly lower median precision values. This result is important, as the training sets for other algorithms potentially contained explicit representations of individual peptides overlapping with this dataset, whereas HLA-Inception predictions are only based on an overall peptide binding motif, utilizing no information about individual peptide identities. Furthermore, a similar result was achieved when testing each algorithm against completely novel peptides eluted from ovarian tumors (Figure S11), supporting that the observed results were not overly biased to existing data.

In addition to precision, fast computational speed was a chief guiding design principle behind HLA-Inception. Due to the relatively simple mathematical operations underlying PWM scoring and the highly parallelized implementation, HLA-Inception-based peptide prediction enable proteome-scale binding prediction in a matter of seconds (Table 1). Furthermore, when compared with other prediction algorithms in a real world example, HLA-Inception is orders of magnitude faster (Figure S12). Hence, the primary purpose of HLA-Inception lies in efficiently screening the most confident

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

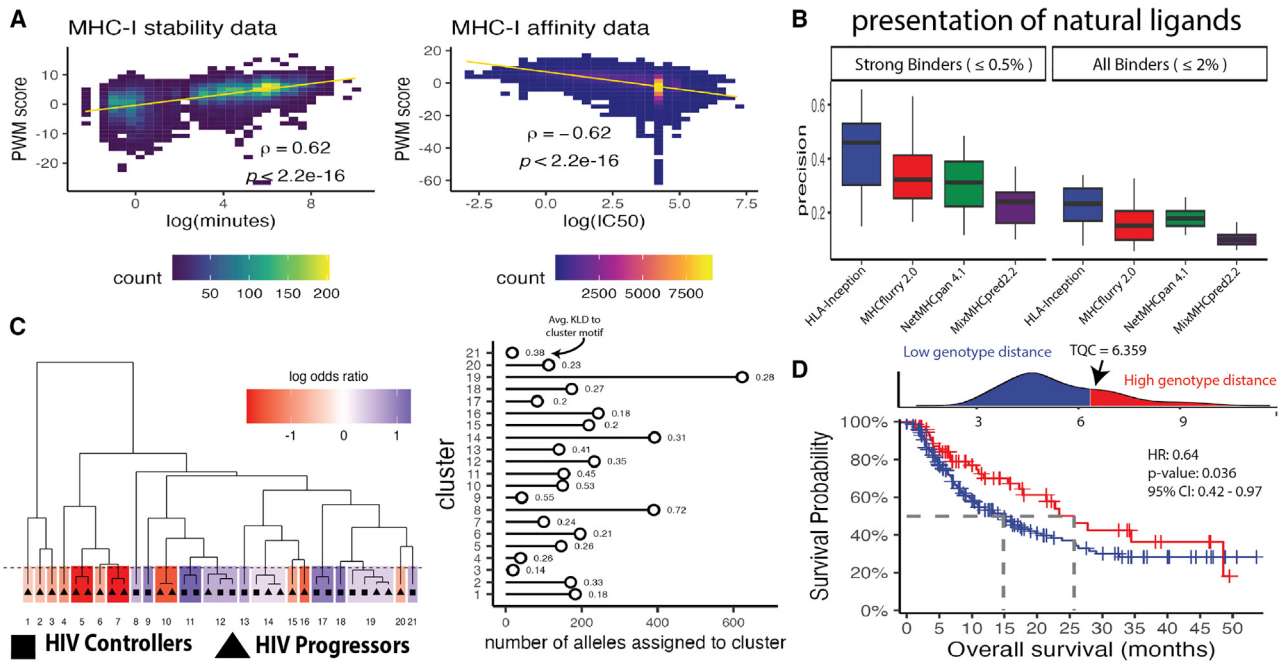


Figure 4. Pan-allele peptide prediction with HLA-Inception

(A) The scatterplots show the correlation between PWM score (y axis) and MHC-I stability (x axis; left; n = 9,702) or MHC-I binding affinity (x axis; left; n = 128,498). Correlation coefficients were determined using Spearman's rho.

(B) The precision (y axis) of HLA-Inception (blue), MHCflurry2.0 (red), netMHCpan-4.1 (green), and MixMHCpred2.2 (purple) in the recovery of naturally presented MHC-I peptides is shown by a boxplot (n = 50 single allele datasets). Precision was measured relative to 0.5% and 2% binding percentile cutoff thresholds, which select for strongly binding and all binding peptides, respectively.

(C) The predicted binding motifs for MHC-I alleles associated with HIV control (squares; n = 14) or progression (triangle; n = 19) were hierarchically clustered. Clusters were then determined by cutting the tree at a binding motif KLD of 1. The colored bars under the tree indicate each cluster, with color indicating the average log odds ratio of being an HIV controller based on alleles assigned to that cluster (arbitrary units). The lollipop plot to the right indicates the number of alleles assigned to each cluster, based on binding motif KLD distance (STAR Methods), with the color indicating the average log odds ratio.

(D) Genotype distances were calculated for patients receiving immune checkpoint inhibitors. Patients exhibiting genotype distances within the top quartile of all patients were designated as high genotype distance individuals (red; n =79), while the remaining patients were assigned to the low genotype distance group (blue; n = 235). Kaplan-Meier plots were then constructed for both groups. Statistical significance was determined using a Cox proportional-hazards model.

peptide list from an extreme excess of potential candidate ones to enable further ranking of the short list of peptides by sequence-based methods—a common scenario arising in numerous MHC-I binding studies.

Ultimately, the true value of any pan-allele prediction lies in its ability to be extrapolated to unseen MHC-I alleles, a setup in machine learning referred to as "zero-shot prediction," which can be quantified using "leave-one-out" cross-validation analysis.[36] In this analysis, data corresponding to a target allele are removed from the training set, and then the remaining data are used to train the algorithm. The accuracy of the algorithm is determined by testing the withheld data. However, such validation approaches fail to account for the existence of highly homologous MHC-I alleles still contained within the training set. The inclusion of homologous alleles has the potential to artificially boost algorithm performance. A more rigorous test of pan-allele predictive properties can be performed by ensuring that highly homologous alleles are removed prior to training, and these allele groups are collectively tested to assess the generality of the algorithm or "leave-one-cluster-out" analysis. To this end, the binding pocket sequences for the 133-allele set were clustered using BLOSUM62 alignments, where each allele was assigned to a cluster of alleles with similar amino acid sequences (Figure S13A). In order to appropriately benchmark the performance of HLA-Inception peptide predictions with a sequence-based approach, a neural network trained on these BLOSUM62-encoded peptides and key binding pocket residues was built. Using the MHC-I binding pocket sequence clusters, leave-one-cluster-out analysis was performed using both neural networks (Figure S13B). We found that electrostatics-based binding classification predictions produced a median Matthews correlation coefficient (MCC) of 0.72 (interquartile range [IQR]: 0.59–0.79), while the sequence-based model produced a median MCC of 0.52 (IQR: 0.38–0.68), suggesting a 38% improvement over a sequence-based prediction method. This improvement indicates that the incorporation of electrostatics improves universal peptide prediction. To verify that the network has indeed learned the electrostatic signals for predicting the complementary peptide motifs, all the 5,320 maps were recomputed at a higher salt condition, wherein the map features are washed out (Figure S13B). Indeed, for the majority of the peptide motifs, the MCC decreased, but it was still outperforming sequence-only predictions, even when using state-of-the-art sequence-based models (Figure S14A). Furthermore, inception-based models trained on all maps available for each allele were shown to be the most effective at

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

Table 1. Prediction timing

Proteome	Proteins	Total 9mer peptides	Binding threshold (%)	Predicted binders	Prediction time (s)
Human	81,837	29,049,213	99.5	148,892	5.73
			98	590,201	9.31
Mouse	55,286	22,844,078	99.5	121,228	4.64
			98	481,594	7.59
Virus	3,383	1,385,612	99.5	7,678	1.95
			98	30,714	2.15

HLA-Inception was used to identify binding peptides from the complete human and mouse reference proteomes, as indicated in the first column. All predictions were done with respect to nonameric HLA-A02:01 peptides that were predicted to binding in the top 0.5% percentile.

learning the underlying physical forces when compared with other machine learning models or the use of less maps per allele (Figures S14B and S14C).

Molecular fingerprints enable the modeling of disease associations

Patient MHC-I binding repertoire has been implicated in outcomes for several viral and cancer-based diseases.[34,37–40] To establish that the predicted binding motifs capture some of these high-level phenotypic relationships, we assessed whether distances between MHC-I alleles and genotypes capture known trends in the MHC-I disease associations. At the individual allele level, we determined whether MHC-I binding motifs could be used to infer HIV viral control.[38] Past work has suggested that observed MHC-I allele associations with HIV viral control are based on the capacity of individual MHC-I alleles to present structurally relevant regions of HIV proteins.[37] Following this rationale, alleles with similar outcomes are expected to have comparable binding preferences. To test this relationship, alleles with known associations to HIV disease progression were hierarchically clustered based on pairwise binding motif KLD distances (Figure 4C). We find that once the clusters are defined as alleles that feasibly share a binding motif (i.e., bind similar peptides), the ones with comparable disease outcome are grouped together. This grouping indicates that binding motifs predicted on their electrostatic properties preserve known allelic associations with HIV outcome at the individual MHC-I allele level. To extrapolate beyond alleles with known associations, we selected all alleles that demonstrated a binding motif KLD distance ≤ 1 to an existing HIV cluster motif and then assigned each of the remaining alleles to the nearest cluster. This extrapolation resulted in 66% of alleles being assigned (3,827). Clusters were generally compact with the median intra-cluster KLD being 0.2692. While there were marginally more allele clusters associated with HIV progression (11 vs. 10), 59.7% of alleles were assigned to clusters with a bias to control HIV disease progression. Despite this observation, the overall average odds ratio (OR), weighted by cluster size and calculated across all assigned alleles, was found to be close to 1 (OR = 0.91). To quantitatively test that groupings preserved known disease outcomes, alleles assigned to clusters with an average OR of less than 1 were labeled as HIV progressing alleles, while alleles assigned to clusters with an average OR ≥ 1 were labeled as HIV controlling alleles. A Fischer exact test performed between these cluster-based labels and the ground truth labels revealed a significant association ($p = 1.453 \times 10^{-5}$). In essence, we demonstrated that predicted peptide binding signatures could be used to group peptides with known outcome to HIV.

Evolutionary protein divergence of patient MHC-I genotypes has been strongly associated with outcomes to immune checkpoint inhibitors (ICIs).[39] However, evolutionary distance is a metric defined solely by analysis of the MHC-I protein sequence. To investigate whether the observed effect could also be due to variance in MHC-I binding motifs, we solved the MHC-I genotype distance (arithmetic average of all pairwise binding motif KLDs for a given genotype) for patients who were treated with ICIs, using the predicted motifs from HLA-Inception. As a baseline comparison, patients within the same cohort were also stratified based on supertype zygosity of the HLA-A and HLA-B locus.[21] However, such a stratification failed to produce a statistically significant separation between cohorts (Figure S15). Conversely, using patient genotype distance, patients were stratified, with patients in the top quartile being labeled as having high genotype distance and the rest of the patients being labeled as having low genotype distance (Figure 4D). We found that patients with a higher level of MHC-I binding motif diversity survived longer when treated with ICIs (hazard ratio = 0.64; p = 0,036; 95% CI: 0.42–0.97). Together, these data suggest that patients with a more diverse peptide repertoire likely benefit from checkpoint blockade.

DISCUSSION

In this study, a physics-based inception network is employed to probe the signatures of molecular recognition of the MHC-I protein system, an area traditionally dominated by sequence-based analyses. Capitalizing on these traditional approaches, we find that the inception networks trained on the three-dimensional electrostatic potentials of the MHC-I binding pocket combined with limited binding peptide sequence information could be leveraged to predict MHC-I peptide binding motifs across a range of diverse MHC-I alleles. By using the binding motifs from HLA-Inception, we were able to assign all MHC-I proteins to an MHC-I supertype. We found that the heterogeneity of MHC-I binding submotifs within a given supertype varied, carrying implications for the continued use of MHC-I supertypes to design broad peptide-based vaccines. We show that the predicted binding motifs can be utilized to perform pan-allele peptide binding prediction at a high level of precision and speed. Furthermore, the comparison of predicted MHC-I binding motifs was shown to recapitulate known disease associations, namely HIV and ICI response.

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

There are several profound advantages to approaching the prediction of MHC-I binding motifs, and subsequently peptide ligands, from an electrostatic lens. First, we employ one of the biophysical rules that dictate peptide binding. Sequence-centric MHC-I prediction methods do not explicitly access the underlying forces that drive peptide binding, offering only amino acid configurations that lead to a particular binding motif. This understandably makes such approaches highly biased by variations in binding pocket sequences, which is problematic given the polymorphic nature of the MHC-I protein. In contrast, by training our model directly on the underlying forces, HLA-Inception is able to learn a measurable physical quantity that is ubiquitous to peptide binding and simultaneously tracks with the variations of the pocket sequences and heterogeneity of the motifs. This physics formulation enhances interpretability of binding predictions, which is evident by the results of the leave-one-cluster-out analysis. Another advantage of the shift to electrostatic modeling is the reduction of the experimental search space. Because electrostatic potential is a degenerative property, as many different sequence configurations can produce similar local electrostatic environments, the number of MHC-I alleles with unknown binding motifs that require experimental validation is diminished. This makes universal coverage of all human MHC-I binding motifs an experimentally tractable goal, and therefore, it opens new doors to broadened applications of T cell-based immunotherapies. Finally, HLA-Inception embodies a methodological advance in computational immunology. Our approach is able to perform MHC-I peptide binding prediction without information on non-binding peptides, a common challenge in machine learning (ML)-based peptide binding classification. This precision of predicting motifs and communities, coupled with the ability to infer binding affinities, makes HLA-Inception a natural complement for high-throughput MHC-I ligand identification techniques such as mass spectrometry.

There are some caveats to the current implementation stemming from the compositional bias of the training set, the use of nonameric peptide binding motifs, the representation of the data in the model with a single biophysical descriptor, and the limitations of using PWM for prediction. Like most machine learning models, predictions are biased by the composition of the training set. In cases where the training set provides a good sampling of the total input space, predictions have a high likelihood of accuracy. Conversely, in cases where isolated populations, not captured by the training set, exist, then predictions are unlikely to be accurate for these groups. The immense number of MHC-I variants makes this a valid concern for any machine learning approach to MHC-I ligand prediction and is not specific to our model. However, as outlined above, we expect that our approach will be less affected by this problem due to the learning of the underlying physical nature of peptide binding. To combat this problem, future work can be focused on the experimental resolution of MHC-I alleles with predicted electrostatic environments that fall outside the currently observed distribution. Next, predictions were only done with respect to nonameric peptide binding motifs. This decision was due to the majority of observed peptides being 9 amino acids in length. This translated into high-resolution binding motifs. However, there is a smaller but relevant population of peptides at different lengths. We used an approach analogous to NN-align[20] to extend the nonameric motifs to peptides of lengths 8–11. We observed high accuracy for peptides of these lengths, which cover 95% of all observed MHC-I peptides (Figures S9B and S9C). HLA-Inception does not implicitly account for other critical physical forces such as van der Waals interactions. It was observed that a majority of MHC-I sequence groups are described by electrostatic interactions; however, a subset of alleles is optimally described using maps of van der Waals potentials (Figure S13). It is possible that the overall shape of the binding pocket may play a larger role than the innate electrostatic forces for these groups of alleles. Future work will be focused on optimally combining multiple physical descriptors into one model. Finally, there are limits to the use of PWMs for peptide prediction. Precision-recall curves on held-out data show that sequence-based methods provide a slightly higher performance, as measured as by precision-recall curves (Figure S10). This is likely due to inherent biases in second-order relationships in amino acid usage, which would not be captured using PWM scoring.

HLA-Inception and future physics-based peptide prediction methods provide tools for the efficient predictions of MHC-I peptides. However, despite providing higher precision at several orders of magnitude better efficiency (Figures 4, S10C, S11, and S12) and showing improved generalization to unseen alleles (Figures S13 and S14), sequence-based methods maintained superior recall at a level that produced nominally better overall performance (Figures S9 and S10). This is likely due to inherent limitations in the use of PWM scoring for peptide selection. Therefore, sequence-based methods remain the preferred approach when recall is critical to the objective, as is the case when ranking a short list of peptides based on binding affinity. Nevertheless, the primary purpose of HLA-Inception lies in efficiently screening the most confident peptide list from an extreme excess of potential candidate ones—a common scenario arising in numerous MHC-I binding studies. This complementary application aligns with the core design philosophy of HLA-Inception, which aims to curate a tractable list of binding peptides when deployed at the whole-proteome scale. Ultimately, future implementations will likely need to leverage both biophysical and peptide sequence information to achieve optimal performance.

In summary, our inception models enable the discovery of biological design principles, the underlying physics of which can extend beyond the system of interest, and predictions of testable phenotypic properties across a broad range of physical conditions. Going forward, the method of learning the electrostatic environment to perform motif prediction is readily applicable to numerous applications known for high sequence variability, including MHC-II, protein-protein binding, and TCR-MHC binding.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2024.03.001.

ACKNOWLEDGMENTS

The authors acknowledge Research Computing at Arizona State University for providing HPC resources that have contributed to the research results reported within this paper. A.S., E.W., and J.K.C. acknowledge the CAREER award from NSF (MCB-1942763), National Institute of Neurological Disorders and Stroke (R01NS119505), and AstraZeneca. This material is based on work supported by the National Defense Education Program (NDEP) for Science, Technology, Engineering, and Mathematics (STEM) Education, Outreach, and Workforce Initiative Programs under grant no. HQ0034-21-S-F001. A.S. also acknowledges the ASU-Mayo Alliance Fellowship. M.C. would like to acknowledge Mayo Clinic Center of Individualized Medicine. A.P. would like to acknowledge grants from NCI (U01CA271410 and P30CA15083).

AUTHOR CONTRIBUTIONS

E.W., A.S., and K.S.A. conceived, designed, and supervised the project and research, with key input from D.C. E.W. and J.K.C. collected, developed, and analyzed the machine learning model and related outputs. R.R., K.K.M., A.P., and M.C. collected and analyzed the ovarian mass spectrometry data. The manuscript was primarily written by E.W. and A.S. with feedback from all other authors.

DECLARATION OF INTERESTS

E.W., A.S., J.K.C., and K.S.A. have submitted a patent application concerning the prediction methods outlined in this paper.

Received: April 4, 2023
Revised: November 24, 2023
Accepted: March 5, 2024
Published: March 29, 2024

REFERENCES

1. Rock, K.L., Reits, E., and Neefjes, J. (2016). Present yourself! by MHC class I and MHC class II molecules. Trends Immunol. *37*, 724–737.
2. Trolle, T., McMurtrey, C.P., Sidney, J., Bardet, W., Osborn, S.C., Kaever, T., Sette, A., Hildebrand, W.H., Nielsen, M., and Peters, B. (2016). The length distribution of class I–Restricted T cell epitopes is determined by both peptide supply and MHC Allele–Specific binding preference. J. Immunol. *196*, 1480–1487.
3. Garrett, T.P., Saper, M.A., Bjorkman, P.J., Strominger, J.L., and Wiley, D.C. (1989). Specificity pockets for the side chains of peptide antigens in HLA-Aw68. Nature *342*, 692–696.
4. Nguyen, A.T., Szeto, C., and Gras, S. (2021). The pockets guide to HLA class I molecules. Biochem. Soc. Trans. *49*, 2319–2331.
5. Rammensee, H.-G., Friede, T., and Stevanoviíc, S. (1995). Mhc ligands and peptide motifs: first listing. Immunogenetics *41*, 178–228.
6. Sundberg, E.J., Deng, L., and Mariuzza, R.A. (2007). TCR recognition of peptide/MHC class II complexes and superantigens. Semin. Immunol. *19*, 262–271.
7. Nivón, L.G., Moretti, R., and Baker, D. (2013). A pareto-optimal refinement method for protein design scaffolds. PLoS One *8*, e59004.
8. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L.E., Brookes, D.H., Wilson, L., Chen, J., Liles, K., et al. (2018). Improvements to the APBS biomolecular solvation software suite. Protein Sci. *27*, 112–128.
9. Leidner, R., Sanjuan Silva, N.Sanjuan, Huang, Huayu, Sprott, D., Zheng, C., Shih, Y.-P., Leung, A., Payne, R., Sutcliffe, K., Cramer, J., et al. (2022). Neoantigen T-Cell receptor gene therapy in pancreatic cancer. N. Engl. J. Med. *386*, 2112–2119.
10. Zacharakis, N., Chinnasamy, H., Black, M., Xu, H., Lu, Y.-C., Zheng, Z., Pasetto, A., Langhan, M., Shelton, T., Prickett, T., et al. (2018). Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. Nat. Med. *24*, 724–730.
11. Chong, C., Marino, F., Pak, Huisong, Racle, J., Daniel, R.T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2018). High-throughput and sensitive immunopeptidomics platform reveals profound Interferonγ-Mediated remodeling of the human leukocyte antigen (HLA) ligandome. Mol. Cell. Proteomics *17*, 533–548.
12. Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, R.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209.
13. Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., and Marsh, S.G.E. (2020). IPD-IMGT/HLA database. Nucleic Acids Res. *48*, D948–D955.
14. Illing, P.T., Pymm, P., Croft, N.P., Hilton, H.G., Jojic, V., Han, A.S., Mendoza, J.L., Mifsud, N.A., Dudek, N.L., McCluskey, J., et al. (2018). HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. Nat. Commun. *9*, 4693.
15. Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., et al. (2006). A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Comput. Biol. *2*, e65.
16. Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., Sinitcyn, P., Audehm, S., Straub, M., Weber, J., Slotta-Huspenina, J., Specht, K., et al. (2016). Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat. Commun. *7*, 13404.

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

18. O'Donnell, T.J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: Improved Pan-Allele prediction of MHC class I-Presented peptides by incorporating antigen processing. Cell Syst. *11*, 418–419.

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

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

21. Sidney, J., Peters, B., Frahm, N., Brander, C., and Sette, A. (2008). HLA class I supertypes: a revised and updated classification. BMC Immunol. *9*, 1.

22. Pucci, F., Schwersensky, M., and Rooman, M. (2022). Artificial intelligence challenges for predicting the impact of mutations on protein stability. Curr. Opin. Struct. Biol. *72*, 161–168.

23. McCoy, A.J., Chandana Epa, V.C., and Colman, P.M. (1997). Electrostatic complementarity at protein/protein interfaces. J. Mol. Biol. *268*, 570–584.

24. Sheinerman, F.B., and Honig, B. (2002). On the role of electrostatic interactions in the design of protein–protein interfaces. J. Mol. Biol. *318*, 161–177.

25. Singharoy, A., Barragan, A.M., Thangapandian, S., Tajkhorshid, E., and Schulten, K. (2016). Binding site recognition and docking dynamics of a single electron transport protein: Cytochrome c2. J. Am. Chem. Soc. *138*, 12077–12089.

26. Li, Y., Zhang, X., and Cao, D. (2013). The role of shape complementarity in the protein-protein interactions. Sci. Rep. *3*, 3271.

27. Gilson, M.K., and Radford, S.E. (2011). Protein folding and binding: from biology to physics and back again. Curr. Opin. Struct. Biol. *21*, 1–3.

28. Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M.M., and Correia, B.E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat. Methods *17*, 184–192.

29. Tubiana, J., Schneidman-Duhovny, D., and Wolfson, H.J. (2022). Scannet: An interpretable geometric deep learning model for structure-based protein binding site prediction. Nat. Methods *19*, 730–739.

30. Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA *105*, 1118–1123.

31. Mehla, K., and Ramana, J. (2016). Identification of epitope-based peptide vaccine candidates against enterotoxigenic escherichia coli: a comparative genomics and immunoinformatics approach. Mol. Biosyst. *12*, 890–901.

32. Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2022). Predictions of immunogenicity reveal potent sars-cov-2 cd8+ t-cell epitopes. Preprint at bioRxiv. https://www.biorxiv.org/content/10.1101/2022.05.23.492800v1.

33. Rammensee, H., Bachmann, J., Emmerich, N.P., Bachor, O.A., and Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics *50*, 213–219.

34. Wilson, E.A., Hirneise, G., Singharoy, A., and Anderson, K.S. (2021). Total predicted mhc-i epitope load is inversely associated with population mortality from sars-cov-2. Cell Rep. Med. *2*, 100221.

35. Marcu, A., Bichmann, L., Kuchenbecker, L., Kowalewski, D.J., Freudenmann, L.K., Backert, L., Mühlenbruch, L., Szolek, A., Lübke, M., Wagner, P., et al. (2021). Hla ligand atlas: a benign reference of hla-presented peptides to improve t-cell-based cancer immunotherapy. J. Immunother. Cancer *9*, e002071.

36. Zhao, W., and Sher, X. (2018). Systematically benchmarking peptide-mhc binding predictors: From synthetic to naturally processed epitopes. PLoS Comput. Biol. *14*, e1006457.

37. Gaiha, G.D., Rossin, E.J., Urbach, J., Landeros, C., Collins, D.R., Nwonu, C., Muzhingi, I., Anahtar, M.N., Waring, O.M., Piechocka-Trocha, A., et al. (2019). Structural topology defines protective cd8+ t cell epitopes in the hiv proteome. Science *364*, 480–484.

38. International HIV Controllers Study (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science *330*, 1551–1557.

39. Chowell, D., Krishna, C., Pierini, F., Makarov, V., Rizvi, N.A., Kuo, Fengshen, Morris, L.G.T., Riaz, N., Lenz, T.L., and Chan, T.A. (2019). Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. Nat. Med. *25*, 1715–1720.

40. Ferguson, A.L., Mann, J.K., Omarjee, S., Ndung'u, T., Walker, B.D., and Chakraborty, A.K. (2013). Translating hiv sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity *38*, 606–617.

41. Smith, C.A., and Kortemme, T. (July 2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. *380*, 742–756.

42. R Development Core Team (2023). R: a language and environment for statistical computing (R Foundation for Statistical Computing). https://www.R-project.org/.

43. Van Rossum, G., and Drake, F.L., Jr. (1995). Python Tutorial (Centrum voor Wiskunde en Informatica).

44. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.

45. Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Di Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. *45*, D271–D281.

46. Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

47. Park, H., Bradley, P., Greisen, P., Jr., Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D., and DiMaio, F. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J. Chem. Theory Comput. *12*, 6201–6212.

48. Henikoff, S., and Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. Proteins *17*, 49–61.

49. Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings. R package version. https://bioconductor.org/packages/Biostrings.

50. R Development Core Team (2020). R: a language and environment for statistical computing. Version 4.0. 2 (R Foundation for Statistical Computing).

51. Garrett, R.C., Nar, A., Fisher, T.J., and Maurer, K. (2018). ggvoronoi: Voronoi diagrams and heatmaps with ggplot2. J. Open Source Softw. *3*, 1096.

52. Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika *53*, 325–338.

53. Szegedy, C., Jia, S., Yangqing, Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

54. Kamada, T., and Kawai, S. (1989). An algorithm for drawing general undirected graphs. Inf. Process. Lett. *31*, 7–15.

55. Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., et al. (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One *2*, e796.

56. Kassambara, A., and Mundt, F. (2017). Package 'factoextra': Extract and Visualize the Results of Multivariate Data Analyses. R Stats. https://cloud.r-project.org/web/packages/factoextra/index.html.

57. Chicco, D., and Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics *21*, 6.

58. UniProt Consortium (2015). Uniprot: a hub for protein information. Nucleic Acids Res. *43*, D204–D212.

59. Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schäffer, A.A., and Brister, J.R. (2017). Virus variation resource – improved response to emergent viral outbreaks. Nucleic Acids Res. *45*, D482–D490.

60. McQuitty, L.L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. Educ. Psychol. Meas. *26*, 825–831.

61. Sjoberg, D.D., Baillie, M., Haesendonckx, S., and Treis, T. (2022). ggsurvfit: Flexible time-to-event figures. R package version 0.2.0. https://CRAN.R-project.org/package=ggsurvfit.

62. Therneau, T.M., and Grambsch, P.M. (2000). Modeling Survival Data: Extending the Cox Model (Springer).

63. Mangalaparthi, K.K., Madugundu, A.K., Ryan, Z.C., Garapati, K., Peterson, J.A., Dey, G., Prakash, A., and Pandey, A. (2021). Digging deeper into the immunopeptidome: characterization of post-translationally modified peptides presented by mhc i. J. Proteins Proteom. *12*, 151–160.

64. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., and Nesvizhskii, A.I. (2017). Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat. Methods *14*, 513–520.

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
MHC-I peptide data	This study	https://zenodo.org/doi/10.5281/zenodo.10516430
Data used for main figures	This study	https://zenodo.org/doi/10.5281/zenodo.10516430
Software and algorithms		
HLA-Inception	this paper	https://github.com/eawilson-CompBio/HLA-Inception
HLA-Inception	this paper	https://zenodo.org/doi/10.5281/zenodo.10516430
APBS	Jurrus et al.[8]	3.0.0
Rosetta backrub	Smith et al.[41]	2021.16
R	R Development Core Team[42]	4.3.0
Python	Van Rossum and Drake[43]	3.9.7
Tensorflow	Abadi et al.[44]	2.6.0
MHCflurry-2.0	O'Donnell et al.[18]	2.0
netMHCpan-4.1	Reynisson et al.[20]	4.1
MixMHCpred2.2	Gfeller et al.[32]	2.2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Abhishek Singharoy (asinghar@asu.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Presented Ovarian peptides are available upon request
- HLA-Inception (https://zenodo.org/doi/10.5281/zenodo.10516430) is freely available on github at https://github.com/eawilson-CompBio/HLA-Inception.
- Any additional information required to reproduce this work is available from the lead contact upon request.

METHOD DETAILS

MHC-I binding pocket modeling

All the MHC-I sequences used in this study were obtained from the IGMT-HLA database[13](accessed 7/2021). MHC-I protein sequences were filtered for those described at the complete canonical lengths (HLA-A: 265 amino acids, HLA-B: 362 amino acids, HLA-C: 366 amino acids). This resulted in the consideration of 5,821 sequences from which the peptide binding pocket residues were extracted (residues 25–210).

Modeling of all 5,821 MHC-I alleles began with the selection of templates for homology modeling. 606 potential templates for MHC-I binding pocket modeling were initially identified using the IEDB database[17] and downloaded from the RCSB Protein Data Bank.[45] These 606 structures were then scored using the Molprobity software,[46] an algorithm that ranks a structure based on several stereochemical metrics with lower scores indicating higher quality structures. Of the 606 potential structures, the ones with the lowest Molprobity scores for each unique MHC-I allele were selected, resulting in a total of 50 MHC-I templates. The peptide was removed from each template model, and the protein structure was truncated to only include the peptide binding pocket (reside 25-210 of the amino acid sequence). The templates were then minimized using the default Rosetta score function.[47]

Selected templates were used to model 5,821 MHC-I binding pockets via the following 3-step protocol. First, the peptide pocket (residue 25–210) of each of the 5,821 alleles were aligned by amino acid sequence to all 50 template sequences, and the template structure showing the best alignment (lowest number of necessary mutations) were selected for that allele. Second, computational

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

peptide binding pocket models were generated by mutating each assigned template structure to match the target allele sequence using the Rosetta backrub application with default parameters.[41] Third, an ensemble of 40 structures were generated by selecting the lowest energy models from 40 separate iterations of the Rosetta relax application.[7] Completion of these three steps resulted in a total of 232,840 unique structures (5,281 alleles x 40 ensemble members/allele).

Electrostatic map generation

The electrostatic environment of the binding pockets were determined using the APBS software.[8] Each of the 40 ensemble members for a given MHC-I binding pocket were converted into PQR files using the *pdb2pqr30* function. The electrostatic potential was then calculated using the default parameters for APBS. The grid dimensions were set to 129 Å x 161 Å x 129 Å with the fine grid extending to 24 Å beyond the boundaries of the binding pocket and the coarse grid extending to 12 Å beyond the fine grid. Using these parameters, the electrostatic potential was determined at 1 Å resolution, discretizing the three-dimensional space into voxels each containing approximately $1Å^3$ of volumetric features. The potentials were calculated using a linearized Poisson-Boltzmann equation with a protein dielectric of 2, a solvent dielectric of 78.54, and an ion concentration of 0.15M. The resulting electrostatic and van der waals maps were then separately saved in the*.dx* file format. To study the role of charge screening in detecting complementary motifs, the previous set-up was repeated exactly with the exception of the salt concentration being set to 1.5M. The correlation between full electrostatic maps was defined as the element-wise Pearson correlation between two vectorized maps.

MHC-I binding pocket electrostatic map segmentation

To increase attention on important binding pocket features during inception network training, the electrostatic maps for each MHC-I binding pocket was split into three segments: (1) an N-terminal binding pocket region, (2) a TCR contact region, and (3) C-terminal binding pocket region. Each segment has the dimensions of 12 x 6 x 12 voxels with their center coinciding on an evenly spaced vector that runs the length of the binding pocket (Figure S1A). Features from these segments were employed to predict the amino acid distribution of binding peptide residues that are most likely to reside within these regions. The N-terminal region covered the area that would likely interact with positions 1–3 of binding peptides; the TCR contact region covered the region approximately below positions 4–6 of binding peptide; the C-terminal regions covered the region that would likely interact with positions 7–9 of binding peptides. The segments were then transformed into 3 tensors with each tensor having the dimensions of 5,320 x 12 x 6 x 12. These electrostatic map tensors were then paired a response tensor of the same length containing the amino acid distributions for a single binding peptide residue assigned to that segment. This resulted in 9 training data sets, with a specific training set for each peptide position (Figure S1B). For example, when training the model to predict the amino acid distribution of position 2 of a binding peptide, the training set would correspond to the tensor of all of the N-terminal segments (5,320 x 12 x 6 x 12) with a response tensor of the frequency of all 20 amino acids observed at position 2 (5,320 x 20).

MHC molecule distance metrics

Variation between MHC-I alleles were calculated by employing three metrics, namely *Hamming distance*, *BLOSUM alignment*, and *Electrostatic Potential Distance*, where higher values indicate more divergent alleles. *Hamming distance* determines the distance between two equal length sequences as the number of mismatches between the two. For example, when comparing the peptides "YMLDLQPET" and "YMLAAQPET", the number of mismatches (colored in red) are two. Therefore, the *hamming distance* between the peptides would be two. *BLOSUM62 alignment* is the sum of the log odds ratios of a particular amino acid substitution given the background frequency of that amino acid.[48] These alignments were calculated with respect to the binding pocket residues within 6 Å of the MHC-I N- and C-terminal anchor residues using the *stringDist* function in the *Biostrings* R package.[49] *Electrostatic Potential Distance* or EPD represents the similarity between electrostatic environments near primary (N- and C- terminal) MHC-I anchor positions. To compute this quantity, a pair of spherical volumes of electrostatic potentials were extracted from the complete binding pocket electrostatic potential environment, one from each terminus. The centers of these spherical volumes were determined using the coordinates of complementary sites on the nonameric peptide ligands. Specifically, after aligning peptides from known MHC-I bound X-ray structures, the centers of the spheres represented the average sidechain centers of mass at peptide positions 2 and 9. A cutoff radius of 6 Å was chosen for defining the volume, as it produces non-overlapping spheres and captures key electrostatic features within one hydration layer. Integrating information from both these anchors, all the n number of three-dimensional voxels embedded within each of the two spherical volumes were concatenated into a single one-dimensional vector. The EPD between any pairwise combinations of such electrostatic vectors (one for each allele) is defined as the Euclidean distance:

$$EPD_{xy} = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}, \quad \text{(Equation 1)}$$

where the i^{th} iso-volumetric voxel occupies the same points in space from two different electrostatic maps corresponding to distinct MHC-I alleles x and y. For the computations in Figure 2, $n = 2\times4/3\pi6^3 \approx 900$ for voxel dimensions of 1 $Å^3$.

MHC peptide binding motif KLD

The similarity of two MHC-I peptide binding motifs, which we call the *binding motif KLD*, was quantified as the total sum the of observed Kullback–Leibler divergence (KLD) between distributions of amino acids at each position (P1-P9). KLD is a statistical

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

distance measurement that quantifies the divergence of two probability distributions. Kullback–Leibler divergence is calculated using the following equation:

$$D_{KL}(P\|Q) = \sum_{i=1}^{k} P_i \log \frac{P_i}{Q_i}. \quad \text{(Equation 2)}$$

In the above equation, P and Q are discrete probability distributions with k number of bins. Due to the non-symmetry of KLD calculation, i.e. $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$, the KLDs reported in this paper represent a symmetrized KLD computed by finding the average of the KLD values calculated in both directions. The equation is as follows:

$$KLD = \frac{D_{KL}(P\|Q)+D_{KL}(Q\|P)}{2}. \quad \text{(Equation 3)}$$

The *Binding motif KLD* was then calculated with the following equation

$$Binding\ Motif\ KLD = \sum_{i=1}^{9} KLD_i. \quad \text{(Equation 4)}$$

where i indicates a position of a nonameric peptide and KLD_i indicates the observed KLD of the i^{th} position between the two considered MHC-I binding motifs.

K-means clustering of electrostatic potential

K-means clustering of the N- and C- terminal electrostatic potentials at MHC-I anchor positions were performed using the *kmeans* function from the stats R package.[50] To find the optimal number of clusters, cluster centers were randomly instantiated and added until a cluster containing only a single MHC-I allele was formed. The entire kmeans clustering process, i.e., starting with one cluster and iteratively adding new clusters until termination, was repeated 1000 times with different random seeds. Following the 1,000 iterations, the optimal number of clusters was determined by finding the most probable termination point across all 1,000 iterations. The overall distribution of termination points across each clustering iteration can be seen in Figure S4 with the median being 22 clusters.

The terminal potentials were grouped into 22 clusters and were graphically represented using a Voronoi diagram.[51] The center of each Voronoi cell was defined as the average xy-coordinate of each cluster following multidimensional scaling or MDS transformation.[52] In short, the concatenated terminal potential vectors were first transformed into two-dimensional Cartesian space using MDS. The center was then determined by averaging over the reduced MDS coordinates with respect to each cluster. This resulted in a single average xy-coordinate for each cluster, signifying the center of the cluster in the xy-plane.

The compactness of clusters, in terms of the similarity of associated MHC-I binding motifs, was assessed by measuring the intra-cluster and inter-cluster average peptide binding motif KLDs. Intra-cluster average binding motif KLD corresponds to the average of binding motif KLDs of all pairwise combinations of alleles within a cluster. The inter-cluster average binding motif KLD represents the average binding motif KLD of all pairwise of combinations of alleles within a cluster to all alleles not contained within the cluster. A large difference between intra-cluster and inter-cluster average KLDs indicates a compact and non-redundant group.

Training set of MHC-I peptide binding motifs

The training set of MHC-I binding motifs were determined using MHC-I peptides with experimentally known binding affinities.[17] To ensure these binding motifs have comprehensive amino acid representation at each position, initially only alleles with at least 50 experimentally validated binders were selected. Thereafter, the peptide data was filtered to select for only binding peptides that were nine amino acids in length and were assigned to an allele at four-digit resolution (Figure S2). Following the constraints of peptide length, resolution, and binding affinity, we could train the neural networks on 133 out of 5,821 MHC-I alleles. For each of these 133 alleles, all assigned peptides were then aligned, and the amino frequency at each position was calculated. Finally, these position-specific amino frequencies were used to define the peptide binding motif for a given MHC-I allele.

Model architecture and training

HLA-Inception is inspired by the inception v1 architecture developed by google.[53] The convolutional aspect of HLA-Inception contains of one inception block consisting of two inceptions modules followed by four densely connected layers, each separated by dropout layers. The output layer returns a one-dimensional vector of length 20 with the loss being calculated using a KLD loss function and optimized using the ADAM algorithm for stochastic gradient dissent. The 'relu' activation function was used for each layer with the exception of the final output layer which utilized a 'softmax' activation function. A graphical representation of the HLA-Inception architecture can be seen in Figure S17. Overall, HLA-Inception consists of a collection of nine individual models, each corresponding to a different position of the peptide binding motif. The default model used a learning rate of 1e-4 and was trained for 500 epochs. A hyperparameter search was performed to identify the best number of epochs and learning rate. Due to the general importance of position 2 and position 9, a grid search was performed on these positions covering epochs 50, 75, and 100 and learning rates 1e-2, 1e-3, and 1e-4. 100 epochs and a learning rate of 1e-3 were identified as the most optimal and were used to train all 9 models when performing 10-fold cross-validation (Figure S5) and subsequent LOC analysis.

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

Motif prediction

Using the optimal parameters, HLA-Inception was trained on available structure and binding data from 133 alleles and used to predict binding motifs for across all 5,821 alleles. In order to provide enhanced context for predictions, maps corresponding to each of the 40-structure ensembles were averaged to produce allele-specific potential representations. The averaged maps were then segmented, as previously described, and used as inputs to the trained model for predicting their associated binding motifs. Full binding motifs were then generated by combining the predictions from all nine HLA-Inception models (Figure S1).

MHC-I supertypes

Classical MHC-I supertypes were defined as alleles described as reference panel alleles in Sidney et al.[21] However, the classical supertypes failed to incorporate most HLA-C alleles. To avoid a large number of unmapped MHC-I binding motifs, six new HLA-C allele-specific supertypes were generated, namely C01, C02, C03, C04, C06, C07. The new HLA-C supertypes were assigned by inspecting logo plots for all HLA-C alleles with at least 100 known binders, and grouping alleles with visually similar plots. A table of reference alleles for each HLA supertype can be seen in Figure S16 and representative logo plots for each supertype can be seen in Figure S7.

MHC-I supertype graph

A graphical representation describing the topology of the MHC-I supertype network was initiated using an unconnected network, where each node was defined as a different MHC-I supertype binding motif. These MHC-I supertype binding motifs were created by averaging the predicted binding motifs for a set of reference alleles assigned to a given supertype (Figure S16; Figure S7). The connections or edges between the nodes were then added by measuring the pairwise binding motif KLDs (Equation 4) between all supertypes binding motifs; longer edges imply higher inter-motif divergence. In order to generate a more informative graph, where similar motifs would spatially cluster together, the edges of the fully connected network were trimmed. This trimming procedure consisted of connecting all nodes to the top three most similar nodes (not including itself) as determined by minimizing binding motif KLD. Extreme edges (*binding motif KLD* ≥ 4) were also removed. A KLD of 4 was selected as this was the smallest number that would maintain a fully connected graph. It is important to note that some nodes will have more than three connections if that node is within the top-three lowest binding motif KLDs of multiple supertypes. The trimmed network was then visualized using the Kamada-Kawai force-directed algorithm.[54] Following the generation of the MHC-I supertype graph, binding motif KLDs were remeasured for pairwise combinations between supertype binding motifs and HLA-inception-predicted individual allele binding motifs. Each individual binding motif was then assigned to the supertype that produced the lowest binding motif KLD.

Supertype subgraph generation and analysis

Subgraphs consisting of all the MHC-I alleles assigned to a particular supertype were generated (Figures 3C, S8D, and S8E). In these supertype-specific subgraphs, each node represented a different individual allele assigned to that supertype, and the edges between alleles was defined as the binding motif KLD between alleles. Similar to the supertype graph, edges between nodes in the subgraph were trimmed. In this case, edges were trimmed to remove any connection between alleles that are unlikely to bind the same peptides (*Binding motif KLD* ≥ 1). For more information about the selection of the KLD thresholds, see the next section. Peptide binding submotifs and heterogeneity (modularity) within each supertype subgraph were detected and calculated using the infomap algorithm.[30] Modularity is a measure of connectiveness within a network. Networks with high modularity indicate high connectedness between nodes within a module (cluster) while having poor connections between modules (clusters). This indicates the presence of more distinct clusters (i.e. more heterogenous submotifs).

Cross-allele binding predictions

The relationship between binding motif KLD and quantitative changes in binding affinity was assessed by performing cross-allele binding predictions (Figures S8B and S8C). Cross-allele binding predictions involve the process of taking peptides with known affinity to one allele and predicting the affinity of those peptides to a different allele. In this study, cross-allele binding predictions were used in two contexts: Assessing the impact of MHC-I supertype heterogeneity and determining a binding motif KLD threshold for shared binding.

In the context of MHC-I supertype heterogeneity, two different MHC-I alleles classically assigned to a given supertype (e.g. A02 or B44) were selected, and the binding motif KLD between those two alleles were calculated using their predicted motifs from HLA-inception. Next, 1,000 peptides with experimentally known binding affinities were randomly sampled from each of the two alleles to create a peptide set of 2,000 total peptides. NetMHCpan-4.1 was then used to predict the affinity of all peptides experimentally validated to bind to one allele on the other allele and *vice versa*. The average binding affinity was calculated with respect to all NetMHCpan-4.1 predictions, representing the expected cross-allele binding affinity. To determine the relative change in expected binding affinity, the calculated cross-allele affinity values were subtracted from the average observed affinity of peptide ligands assigned to that allele (Figure S8C).

A binding motif KLD threshold where alleles with that value or lower are expected to bind similar peptides is important for grouping alleles and estimating the potential clinical associations. To find this binding motif KLD threshold, a linear model was fit to all allele pairs with a binding motif KLD of less than 3 and compared to the average cross-allele binding predictions (Figure S8B). A binding motif KLD of 1 was shown to indicate an average affinity of 5,000 nM, and was selected as the binding motif KLD threshold.

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

Sequence-based model

A deep sequence-based model, inspired by Nielsen et al.,[55] was constructed for comparison to HLA-Inception (Figure S13). The input to this model was a BLOSUM-encoded vector of key positions within the MHC-I binding pocket, and the model was trained on a balanced data set consisting of 315,512 experimentally resolved MHC-I peptides paired with randomly generated decoy peptides. The model consisted of 3 densely connected layers separated by dropouts, offering a comparable architecture to HLA-Inception. The output of the model was the probability of the given peptide being a binder.

Position-weighted matrix score

Position-weighted matrix (PWM) score is a measure of how strongly a peptide adheres to the probability distribution underlying a given binding motif. The PWM score is calculated by the sum of the log-odd ratios of observing an amino acid at a particular position, given the background frequency of that amino acid within the motif. The equation to calculate PWM is as follows,

$$PWM\ score(pep) = \sum_{i=1}^{9} \log_2 \frac{p_{ij}}{q_j}, \quad \text{(Equation 5)}$$

where *pep* is a peptide being scored, i is the residue number being considered, p_{ij} is the probability of the i-th residue of *pep* at the i-th position according to the binding motif, and q_j is the background frequency of the i-th residue of *pep*. A higher PWM score indicates a higher probability of a peptide binding to a target allele. Allele-specific score thresholds were determined by calculating the PWM scores for all nonameric peptides in the human protein and generating a *cdf* of that distribution.

Pan allele and length peptide prediction

Predictions for HLA-Inception were extended to a larger list of 15,470 alleles through the use of sequence-based alignments. In short, the binding pockets of alleles not covered by the set of 5,281 homology models were individually aligned to sequences of all modeled binding pockets. Unknown alleles were then assigned the predicted motif of the best aligning allele. Binding predictions of peptides with lengths other than 9 was accomplished by adding or removing amino acids, depending on the length of the peptide, until a peptide of length 9 was constructed. For peptides of length 8, a place holder amino with a PWM score of 7th amino acid was inserted at the n-1 position to create a peptide of length 9. Conversely, for peptides longer than 9 amino acids, only peptide positions 1 – 8 and the C terminal of the peptide were scored. This decision was made due to the low contribution of non-anchor residues to the overall PWM score as well as the general lack of secondary anchors residue near the C terminal of the peptide. Peptide length benchmarking was done using a subset of data from the MONOALLELIC benchmark dataset described in O'Donnell et al.[18]

Correlation with quantitative peptide binding scores

Peptides with quantitative values for binding affinity, IC50 or complex stability, were extracted from the IEDB database.[17] Peptide were selected according to the same criteria as peptide selection for the generation of experimental peptide binding motifs. Correlations between quantitative binding values and PWM score were calculated using Spearman's rank correlation coefficient.

Presentation prediction of natural ligands

For analysis centered on the recovery of naturally presented MHC-I ligands, 9mer MHC-I peptides with experimental evidence of presentation for 50 different alleles were obtained from the HLA Ligand Atlas.[35] For each allele, 1,000 peptides experimentally determined to be naturally presented were randomly sampled and combined with 99,000 decoy peptides (not naturally presented) extracted from the human proteome. In cases where alleles had less then 1,000 experimentally described peptides, all peptides were used and decoys peptides were sampled to maintain a ratio of 1:100 target-decoy ratio.

Leave one cluster out analysis

Leave one cluster out analysis is defined as the process of using a cluster of alleles, defined by similar binding pocket sequences, to test the generalizability of different models on unseen data. MHC-I sequences were clustered with respect to BLOSUM-encoded key positions from the MHC-I binding pocket (described in Nielsen et al.[55]). The optimal number of clusters, 11, was determined using the average silhouette width method implemented in the *fviz_nbclust* function in the *factoextra* R package.[56] Following clustering, models would then be tested on each identified cluster by withholding all alleles assigned to a given cluster from the training set, and then testing model performance on those withheld alleles. The Leave one cluster out analysis was performed on all clusters. Model accuracy was reported as the individual matthew's correlation coefficients[57] for peptide prediction of each allele within the cluster. Due to the requirement for a binary outcome to calculate MCC and the fact that the ratio of decoys and target peptides were largely equal, the median binding percentile for each allele within each cluster was designated as the cutoff value for that allele, with all peptides showing better binding percentiles being labeled as binders and all those with worse binding percentiles being labeled as non-binders. To verify that the inception network was optimal for the given data representation, LOC analysis was performed on 3 alternative machine learning architectures: a Random Forest, A 3-dimensional convolutional neural network (CNN), and a multilayer perception (MLP) (Figure S14C). All three models were trained and tested identically to the inception network, except for the model architecture itself. The random forest was trained on a one-dimensional vector that consisted of flattened representation of the 3-dimensional electrostatic potentials. The random forest was trained on a one-dimensional vector that consisted of a flattened

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

representation of the 3D electrostatic potentials. It returned a multi-class output consisting of the probabilities of each amino acid at a given position. The 3D CNN was trained on the 3D electrostatic grids. It consisted of three convolutional layers culminating in a 1D output vector of length 20, corresponding to the probabilities of each amino acid at each position. The MLP was trained on 1D vectors similar to the random forest. It consisted of three dense layers with a 1D output vector of length 20, corresponding to the probabilities of each amino acid at each position. Both alternative deep learning methods used MSE as the loss function. The leave one out cluster analysis was also performed for HLA-Inception models trained on high salt concentration electrostatic maps and van der Waals maps (Figure S13), Physics-based inception models compared to MHCflurry (Figure S14A), and Electrostatics-based inceptions models trained on only one map per allele (Figure S14B). LOC analysis was performed using a subsets of the IEDB database[18] and the ''Curated MHC I mass spec datasets'' described in O'Donnell et al.[18]

Performance benchmarking

Precision-recall curves for HLA-Inception, netMHCpan-4.1, and MHCflurry-2.0 were calculated for the 36 single allele benchmark datasets described in Reynisson et al.[20]. To account for potential overlaps with training sets for sequence-based models, MHCflurry-2.0 was retrained to excluded peptides in the benchmarking set. Threshold specific precision and recall was determined for each algorithm by calculating both metrics using cutoff threshold ranging from 0 to 99.5%. Performance efficiency for each algorithm was determined by predicting 8-11mer peptides for a full MHC-I genotype (HLA-A26:01, HLA-B07:02, HLA-C12:03, HLA-A24:02, HLA-B38:01, HLA-C07:02) of human proteomes ranging from 1 to 82,427 proteins. If supported by the algorithm, results were filtered to only included peptides that fell within a 2% binding threshold. Each prediction was computed on a single CPU core for a maximum of 24 hours. In cases where computations failed to complete in 24 hours, the overall run time was predicted using a generalized linear model fit to completed benchmarking runs for that algorithm.

Prediction of ovarian MHC-I immunopeptidomes

Peptides were initially filtered to remove any overlap with the IEDB database, the MHCFlurry training set, and the 36 single allele benchmarking data set.[17,18,20] Peptides for each patient were then combined with a 10-fold excessive of random human peptides, of equal length distributions, not observed in the data set. Binding predictions were then performed using each algorithm. Precision was measured using the binder and strong binder cutoff threshold which correspond to 2% and 0.5% respectively.

Proteome-scale peptide predictions

Full Human (UP000005640) and Mouse (UP000000589) proteomes were extracted from the uniprot database,[58] and all reference protein sequences for viruses (Virus) with *homo sapiens* listed as a host was extracted from the ncbi virus database.[59] All predictions were performed using an apple M1 pro chip with 10 cpu cores.

Hierarchical clustering of alleles associated with HIV outcome

MHC-I alleles showing statistically significant associations with HIV disease progression[38] were hierarchically clustered based on pairwise binding motif KLD calculations using the WPGMA method.[60] Clusters of alleles were identified by cutting the tree at a KLD threshold of 1. The overall log odds ratio of a cluster was defined as the average of individual log odds ratios of alleles contained within a cluster. Predicted binding motifs were then assigned to an HIV outcome associated allele cluster by calculating the distance between a given allele to all cluster motifs and assigning the allele based on minimal binding motif KLD distance.

Analysis of immune checkpoint data

A dataset describing survival following immune checkpoint blockade of 314 melanoma and non-small cell lung cancer patients was obtained from Chowell et al.[39] Each patient in the cohort was assigned a *genotype distance* value that corresponded to the average of all pairwise binding motif KLDs for that patents' MHC-I genotype. Patients with *genotype distances* in the top 75% were labeled as having high genotype distance while all other patients were labeled as having low genotype distance. A Kaplan-Meier plot describing the survival rates of both groups was generated using ggsurvfit.[61] A hazard ratio between the two groups was determined by fitting a cox proportional hazards regression model to the data.[62] As a baseline method, patient Supertype zygosity[21] was used as a stratification method. Due to the lack of supertypes for HLA-C, the analysis was restricted to the HLA-A and HLA-B locus. For each patient, the number of HLA locus in which there was a supertype mismatch (i.e. A01 vs A02) were counted. A value of 2 indicates that the patient in question had no HLA alleles that shared a supertype while conversely a value of 0 would indicate that both A and B alleles belonged to the same respective supertypes. To simplify the analysis, patients with A or B alleles that were not set within a defined supertype were excluded from the analysis. This resulted in a slight reduction in overall cohort size from 314 patients to 280 patients. Patients were then stratified into two groups: high and low. The high strata indicated patients with no shared HLA supertypes, a value of 2, while the low strata were all other patients, a value of 0 or 1. Overall, there were 180 patients sorted in the high strata and 100 sorted into the low strata.

Isolation of HLA ligands

HLA class I peptides were isolated using standard immunoaffinity purification as described.[63] In brief, snap-frozen ovarian tissue samples (n=5) were homogenized in liquid nitrogen followed by protein extraction using a lysis buffer containing (0.25% sodium deoxycholate, 0.2 mM indole acetic acid, 1 mM EDTA, 1 mM PMSF, 1% Octyl-B-glucopyranoside, 1:200 protease inhibitor cocktail) for

Please cite this article in press as: Wilson et al., The electrostatic landscape of MHC-peptide binding revealed using inception networks, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.03.001

1 h on ice. Following centrifugation, the supernatant was loaded onto a protein A sepharose column for pre-clearing. Subsequently, the eluate was passed through a pan MHC class I-specific antibody (W6/32 clone) crosslinked affinity column. The columns were sequentially washed with 150 mM NaCl in 20 mM Tris HCl pH 8.0, 400 mM NaCl in 20 mM Tris HCl pH 8.0,150 mM NaCl in 20 mM Tris HCl pH 8.0- and finally 20-mM Tris HCl pH 8.0). MHC-bound peptide complexes were eluted using 1% TFA and peptides in the eluate were purified using C18 stage tips prior to LC-MS/MS analysis.

LC-MS/MS analysis

LC–MS/MS analysis was carried out on an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Scientific, San Jose, CA) connected online to a Dionex RSLC3000 liquid chromatography system (Thermo Scientific, San Jose, CA). Survey MS scan was acquired in Orbitrap mass analyzer with 50 ms injection time, 60,000 resolution, and 4×e5 AGC target. MS/MS analysis was performed separately for charge states 2-4 (scan priority 1) and charge state 1 in the mass range of 700-1400 m/z (scan priority 2). Precursor ions (z=2–4) were fragmented with 28% HCD normalized collision energy and acquired in an Orbitrap mass analyzer with 30,000 resolution, 150 ms injection time and 1×e5 AGC target. Precursor ions (z=1) with a mass range of 700–1400 m/z were fragmented with 32% HCD normalized collision energy and analyzed in the Orbitrap analyzer. Dynamic exclusion was enabled for 30 s. Additional filters included monoisotopic precursor selection and an intensity threshold of 2.5×10^4.

Database searching

Mass spectrometry raw data files were analyzed using MSFragger in FragPipe v17 to identify peptides.[64] Database searching was performed against sample-specific personalized protein databases and search parameters included 7–25 amino acids peptide length with no enzyme specificity. Dynamic modifications included oxidation (M), cysteinylation (C), and protein N-terminus in the MSFragger search engine. Precursor ion tolerance of 20 ppm for both precursors and fragment ions was used in MSFragger. The false discovery rate (FDR) estimation was performed using Percolator and set to a peptide spectrum match level of less than 3%.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed within the R platform for statistical computing. Unless otherwise specified all correlation coeffients were estimated using Spearman's rank correlation coefficient. All pairwise comparisions were performed using Wilcoxon rank sum test. Hazard ratios were estimated using the Cox proportional hazard regression model. Statistical significance in HIV allele grouping was determined using a Fischer exact test.

ADDITIONAL RESOURCES

We also provide a web browser-based implementation of the algorithm at http://hlainception.asu.edu:3000/.