Deep Neural Networks Predict MHC-I Epitope Presentation and Transfer Learn Neoepitope Immunogenicity

Benjamin Alexander Albert[1,2], Yunxiao Yang[1,2], Xiaoshan M. Shao[1], Dipika Singh[3,4], Kellie N. Smith[3,4], Valsamo Anagnostou[3,4], Rachel Karchin[1,2,3,5]

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

2 Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

3 The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

4 Bloomberg~Kimmel Institute for Cancer Immunotherapy, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

5 Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA

Corresponding Author

Rachel Karchin

Johns Hopkins University

217A Hackerman Hall, 3400 North Charles Street, Baltimore, MD 21218

Phone: 410-516-5578

Fax: 410-516-5294

karchin@jhu.edu

Abstract

In this study, we present computational methods for predicting MHC-I epitope presentation with high accuracy and improved prediction of immunogenic neoepitopes. The BigMHC method comprises a large (51 million parameters), pan-allelic deep neural network trained on peptide-MHC presentation data from mass spectrometry assays and transfer learned on data from assays of antigen-specific immune response. Compared with NetMHCpan-4.1, PRIME, MHCflurry 2.0, MixMHCpred 2.1, MHCnuggets 2.3.2, and TransPHLA, BigMHC significantly improved the accuracy of predicting presented epitopes on a test set of 946,008 peptide-MHC complexes (AUROC=0.9838, AUPRC=0.9506). After transfer learning on immunogenicity data, BigMHC predicted immunogenic neoepitopes with the highest performance on independent test sets of neoepitopes validated for immunogenicity by IFN-γ release and MANAFEST assays. Moreover, BigMHC yielded significantly higher precision in identifying 215 immunogenic neoepitopes from a large (n=45,410,552) background of non-immunogenic epitopes, making BigMHC effective in clinical settings. All the data and codes are freely available at https://github.com/KarchinLab/bigmhc.

Class I major histocompatibility complex (MHC) plays a crucial role in vertebrate adaptive immunity. The MHC region is highly polymorphic and comprises thousands of known alleles, each encoding a molecule with varying ligand specificities. Identifying non-self-antigens that elicit strong immune responses, given a patient genotype, may yield precise immunotherapies. In the present study, we present a suite of novel computational methods for predicting epitopes and neoepitopes that are likely to be present and elicit an antigen-specific immune response (immunogenic).

MHC molecules may bind peptides to form a peptide-MHC (pMHC) multimer, which may be present at the cell surface for T-cell receptor (TCR) recognition and subsequent CD8+ T-cell expansion. The set of peptides in each stage is a superset of the following; in other words, given an MHC molecule M, then $B \supset P \supset R \supset T \supset S$, where:

- B is the set of peptides that binds to M to form a pMHC complex.
- P is the set of peptides present on the cell surface.
- R is the set of peptides that form TCR–MHC complexes
- T is a set of peptides that elicits CD8+ T-cell clonal expansion.
- S is the set of peptides that elicits an immune response.

The classifiers of set B[1-3] are generally trained on binding affinity (BA) assay data. BA data may be qualitative readings or quantitative half-maximal inhibitory concentration (IC_{50}) data. Qualitative readings were mapped to IC_{50} values, and the domain of IC_{50} measurements was scaled such that the weakest binding affinity of interest, 5×10^5 nmol/L, was mapped to 0 accordingly: $f(IC_{50}) = \max(0, 1 - \log_{5\times10^5}(IC_{50}))$. Classifiers of set P are generally trained on mass spectrometry data in which pairs of peptides and MHC alleles are associated with a Boolean value indicating an eluted ligand (EL), which is bound and presented on the cell surface.

To classify set R, a recent method [4] incorporated CDR3β sequences from the TCR to predict the binding affinity between TCR and pMHC. The data for classes R, T, and S were very limited, making it difficult to train classifiers directly for these sets. Classifiers of set T[5,6] employ transfer learning from the classifiers of sets B and P. To the best of our knowledge, there are no predictors of set S.

Although predicting peptide immunogenicity without TCR information may be challenging, such methods have increased usability because only MHC molecules and peptide sequences are necessary. Because most datasets do not include TCR information, the proposed methods operate solely on MHC genotype and peptides to predict epitope presentation and immunogenicity. Achieving a high predictive capability, however, is made difficult by a host of other factors: an MHC molecule may bind peptides of varying lengths[1], data for many alleles are scarce[1], and immunogenicity data are limited[5,6]. Moreover, most peptides are not presented, and among those that are presented, even fewer elicit an immune response (termed immunogenic). Experimentally validating the immunogenicity of epitopes and neoepitopes is resource intensive; therefore, achieving high precision is important for a predictive method. We briefly overview six state-of-the-art methods to which we compare with the proposed BigMHC (b□g mæk) methods.

NetMHCpan-4.1[2] predicts sets B and P using an ensemble of 100 single-layer networks. This group introduced the idea of a pan-allelic network[7,8], which consumes a peptide and a short representation of an MHC allele of interest, thereby allowing a single model to generalize across MHC alleles. MHCflurry-2.0[3] is a pan-allelic method that predicts sets B and P using an ensemble of multilayer feed-forward networks, convolutional networks, and logistic regression. TransPHLA[9] is a pan-allelic method that predicts set P, using a transformer-based model. MHCnuggets[1] predicts set B, using allele-specific Long Short-Term Memory (LSTM) networks.

MixMHCpred[10,11] predicts set P, using a mixture model and positional weight matrices (PWM) to extract epitope motifs. PRIME[5] is an extension of MixMHCpred to predict set T and was designed to infer the mechanisms of TCR recognition of pHLA complexes. Using the procedure illustrated in Fig. 1, we developed two BigMHC models: one model, BigMHC EL, was trained to predict set P, whereas BigMHC IM was transfer-learned to predict set T. The results demonstrate that BigMHC EL is a highly accurate predictor of set P, outperforming other evaluated methods by a substantial margin. Although PRIME[5] is the best predictor of set T infectious disease epitopes, BigMHC IM is the best predictor of set T neoepitopes.

Results

Features of Compared Methods

We compared the features of BigMHC with those of six state-of-the-art methods, as shown in Table 1. The methods are listed left-to-right in descending order of publication year. The training and prediction data modalities of each method were as follows: binding affinity (BA), eluted ligand presentation (EL), and immunogenicity (IM). Relative execution speeds are noted for the method performance on machine architectures with either one or two GPUs, although only BigMHC can utilize multiple GPUs. We also included information on whether the model is retrainable, licensing, optional GPU acceleration, source language, backend libraries, minimum and maximum peptide length, webserver availability, and peptide amino acid restrictions.

Epitope Presentation Prediction

BigMHC, NetMHCpan-4.1, TransPHLA, MixMHCpred-2.1 and MHCnuggets 2.3.2 are first evaluated on a very large independent set (n=946,008) of eluted ligands (set P). This dataset is the same as that used to evaluate NetMHCpan-4.1[2], but with deduplicated instances. MHCflurry-

2.0 and PRIME were not evaluated because the MHCflurry-2.0 training dataset contained overlapping points and PRIME did not predict set P. The results, illustrated in Fig. 4, suggest that BigMHC dramatically improved EL predictive capability overall, reaching 0.9838 AUROC and 0.9506 AUPRC, whereas the best prior method was NetMHCpan-4.1, with an AUROC of 0.9469 and an AUPRC of 0.8257. BigMHC is particularly effective for peptides of length 8-10, which are the most common lengths of peptides presented by MHC-I. Although the predictive capability of BigMHC decreases as the peptide length increases, it is still at least comparable and generally superior to that of the state-of-the-art for longer peptides. BigMHC also demonstrates strong performance across HLA supertypes, whereas the performance of other methods degrades on HLA-B, and particularly HLA-C.

Immunogenicity Prediction

The vast majority of neoepitopes are not immunogenic. Furthermore, experimental validation of immunogenicity is resource intensive, so only a few dozen neoepitopes can be validated per patient in a clinical setting. It is therefore critical that predictors of immunogenic neoepitopes have high precision among their most highly ranked outputs, as only the top predictions are used in practice. To measure this precision, it is common to evaluate positive predictive value (PPV) among the top n outputs (PPVn)[1-3]. To calculate PPVn, the pMHCs are first sorted by a predictor's output. Then, PPVn is the percentage of the top n pMHCs that are actually immunogenic. In this study, we additionally calculated the area under the PPV curve generated by all valid choices of n (AUPPVn).

To evaluate the precision of immunogenicity predictions, methods are evaluated on a dataset comprised of 215 immunogenic neoepitopes and 45,410,552 non-immunogenic. IFN-γ

release and MANAFEST assays contribute to both positives and negatives, whereas mass spectrometry data supplement negatives. To approximate the ratio of immunogenic neoepitopes to non-immunogenic observed in a clinical setting, we calculate precision for positive:negative ratios of $1:10^4$, $1:10^5$, and, using all data, approximately $1:2\times10^5$. For all ratios, BigMHC achieves the highest area under the PPVn curve (AUPPVn). NetMHCpan-4.1[2] achieves the second-best AUPPVn for all ratios. The improvement BigMHC offered is substantial, reaching values of AUPPVn between 2 times and 25 times greater than those of NetMHCpan-4.1 (Fig. 5).

We further compared the performances of all seven methods on infectious disease epitope immunogenicity from the Immune Epitope Database[12], neoepitope immunogenicity determined from IFN-γ release assays, and neoepitope immunogenicity from MANAFEST assays.

The immunogenicity prediction results on each dataset are shown in Extended Data Fig. 4. PRIME achieved the best performance on the infectious disease dataset (AUROC=0.6110 and AUPRC=0.8442). BigMHC achieved the best performance on both the IFN-γ release neoepitope dataset (AUROC=0.8534, 0.5481) and MANAFEST neoepitope dataset (AUROC=0.5599, AUPRC=0.3032). The baseline AUROC was 0.5, whereas the baseline AUPRC (positives/(positives+negatives)) was 0.7821 for the infectious disease epitope dataset, 0.1920 for the IFN-γ release dataset, and 0.1995 for the MANAFEST dataset. The BigMHC base model results are included in Extended Data Fig. 5 to demonstrate the results of transfer learning on the immunogenicity data.

MHC Attention

The BigMHC network architecture offers a unique attention mechanism whereby prior to sigmoidal activation, the scalar output of the network is a linear combination of the input MHC encoding. Hence, we were able to visualize interpretable attention, the amino acid residue

positions important for classification, in the form of a heatmap overlay on a modelled three-dimensional structure of an MHC molecule of interest. The attention and MHC encoding for all HLA supertypes in the EL training set are visualized in Extended Data Fig. 2, and the three examples are visualized in 3D in Fig. 3. The attention of the EL testing set is similarly visualized in Part A of Fig. 3. The proposed MHC pseudosequences are comprised of the top 30 aligned positions by information content; the most important are those that are in the binding groove. For some alleles, however, some transmembrane and intracellular residues strongly contribute to EL prediction, such as position 338 for HLA-A*24:06 or 319 for HLA-C*07:02. This suggests that the NetMHCpan pseudosequences, which capture positions only nearest to the peptide, may lose information valuable for predicting pMHC presentation. Importantly, this affects all the referenced pan-allele state-of-the-art methods as they currently adopt NetMHCpan pseudosequence MHC representations.

Discussion

We first trained BigMHC to predict peptide presentation (set P) because an enormous amount of data for MHC class I peptide presentation is publicly available, making it feasible to train a deep learning model with over 51 million parameters. BigMHC EL achieved the highest predictive capability for set P, significantly outperforming the state-of-the-art across HLA supertypes and epitope lengths. We further demonstrated several technical findings: training on pMHC binding affinity is unnecessary for predicting pMHC presentation, information content is a useful approach for deriving new MHC pseudosequence representations, and some transmembrane and intracellular MHC positions are important for presentation prediction.

While the goal of neoepitope prediction is ultimately to predict neoepitopes that induce an immune response (set S), there is limited immunogenicity data to train deep learning models.

To address the data scarcity problem, after initially training the base models on presentation data (set *P*), we applied transfer learning using immunogenicity data (set *T*) to produce BigMHC IM. We evaluated BigMHC IM and six other methods on three independent datasets: infectious disease antigen immunogenicty, IFN-γ release neoepitope immunogenicity, and MANAFEST neoepitope immunogenicity.

However, a non-random pMHC selection process in experimental workflows to assess immunogenicity acts as a confounder in these evaluations. Because researchers seek to optimize their workflow for identifying immunogenic pMHCs, it is necessary to first select a relatively small number of pMHCs from a larger set of candidate pMHCs. The selection often incorporates popular *in silico* tools such as NetMHCpan. As a result, the dataset is a biased subset rather than a randomly selected set of pMHCs; consequently, evaluating on existing immunogenicity data measures the ability of a predictor to refine the selection process, rather than the ability to perform the selection itself.

To reduce the effect of this confounder, we generated a very large neoepitope immunogenicity evaluation dataset, comprised of 215 immunogenic neoepitopes among 45,410,552 non-immunogenic. This independent test was created by aggregating the IFN-γ release and MANAFEST neoepitope immunogenicity datasets, and it was supplemented with non-presented peptides, which are inherently non-immunogenic. The resulting dataset contained a ratio of immunogenic to non-immunogenic designed to reflect a clinical environment. The model generated from transfer learning, BigMHC IM, achieved the highest precision in identifying the immunogenic neoepitopes.

Although this evaluation better captures a clinical setting, a limitation of our study is that all evaluations were retrospective. Future work will include prospective evaluation of the predictions of BigMHC with neoepitope immunogenicity assays.

Methods

Datasets

Epitope presentation training. Presentation training data spanned 150 alleles and included 17,303,334 unique pMHC instances, of which 488,599 were positive. The training dataset was compiled from the NetMHCpan-4.1[2] and MHCflurry-2.0[3] single-allelic (SA) EL datasets.

Epitope presentation evaluation. The presentation evaluation set spanned 36 HLA alleles and was comprised of 946,008 unique instances, of which 45,416 were positive. The test set is the same SA EL dataset as that used in the NetMHCpan-4.1 study. For the EL training and testing datasets, the pMHC instances that existed in both training and testing were removed from the training data.

Immunogenicity training. BigMHC IM transfer learns on PRIME[5] and TESLA[6] datasets. The training data consisted of viral antigens, cancer-testis antigens, neoepitopes, and 9-mer peptides randomly sampled from the human proteome to supplement the negative instances. The PRIME dataset contains a mix of immunogenicity assays aggregated from prior studies, whereas TESLA is validated using multimer staining.

Infectious disease antigen immunogenicity evaluation. Infectious disease antigen immunogenicity prediction was evaluated using data collected from the Immune Epitope Database (IEDB)[12]. The queries to IEDB included linear peptides, T-cell assays, MHC-I restriction, human hosts, and infectious diseases. Data was further processed to allow all prior methods to be evaluated: only peptides of length at least 8 and at most 14 were kept, peptides with dummy amino acid 'X' were removed, and pMHCs with MHC alleles incompatible with MixMHCpred[10,11] and PRIME[5] were removed. In total, 7,520 immunogenic and 2,095 non-immunogenic antigens were collected.

IFN-γ release neoepitope immunogenicity evaluation. The IFN-γ release neoepitope immunogenicity dataset was compiled using NEPdb[13] and Neopepsee[14]. NEPdb is a database of neoepitopes curated from the literature with a semi-automated pipeline, whereas Neopepsee aggregated neoepitopes from two prior sources. Both Neopepsee sources validated neoepitopes using IFN-γ production assays, and one of the sources additionally validated using dextramer assays.

MANAFEST neoepitope immunogenicity evaluation. The MANAFEST dataset is collected from 16 cancer patients, comprised of 167 immunogenic and 672 non-immunogenic neoepitopes and validated using the MANAFEST assay[15]. MANAFEST quantifies antigen-specific T-cell clonotype expansion in peptide-stimulated T-cell cultures. The MANAFEST workflow included candidate neoepitope selection using ImmunoSELECT-R from Personal Genome Diagnostics[16,17], incorporating predictions from NetMHCpan 3.0[18] and NetCTLpan[19].

Neoepitope immunogenicity evaluation. To evaluate the precision of immunogenicity predictions, methods are evaluated on a dataset comprised of: the IFN-γ release neoepitope immunogenicity dataset, the MANAFEST neoepitope immunogenicity dataset, and 45,409,678 non-presented pMHCs. These supplementary non-presented pMHCs, which are inherently non-immunogenic, are collected from the multi-allelic MHCflurry-2.0 dataset[3].

Dataset Compositions. The compositions of all datasets are illustrated in Extended Data Fig. 3.

BigMHC Network Architecture

The BigMHC network architecture (Fig. 2), comprised of about 51 million parameters, was designed to capture recurrent patterns via two dense, multilayered, bidirectional LSTMs (BiLSTM) and pMHC binding information via an anchor block. The BiLSTM cells are preceded by self-attention modules; these units are equivalent to transformer multi-headed attention modules[20] where the number of heads is set to 1. The two parallel BiLSTM modules had varying window sizes: one and eight epitope residues. Each BiLSTM cell unroll consumes the entire MHC representation while recurrently processing the variable-length epitope, as illustrated in the Extended Data Fig. 1. Although this imposes a minimum epitope length of 8, a few peptides of length 7 or less are binders. The anchor block is comprised of two dense[21] linear layers with tanh activations, followed by Dropout[22] units with a probability of 0.5. This anchor block consumes the first and last four residues of the peptide, along with the full MHC representation to focus on the anchor site residues. The outputs of the two BiLSTMs and the anchor block are concatenated before being consumed by a condenser, which is also comprised of two dense linear layers with tanh activations, proceeded by Dropout with a probability of 0.5. The output is projected to the same size as the MHC one-hot encoding and passed through tanh activation to attend to MHC

encoding. This attention vector can then be superimposed onto a three-dimensional structure of an MHC allele of interest to identify important amino acid residue positions for a given pMHC. Moreover, because the final output of the model is a linear combination of the MHC one-hot encoding, the scalar output is interpretable, each MHC position is assigned a weight that contributes in favor of, or against, presentation, and their sum is the model output prior to sigmoid activation.

BigMHC Models

Two BigMHC models were developed and applied in this study. One model, BigMHC EL, was trained on mass spectrometry data to predict eluted ligands (set P). The BigMHC IM model uses BigMHC EL as a base model for transfer learning to predict immunogenic epitopes and neoepitopes (set T). Transfer learning was performed by retraining the final and penultimate fully-connected layers of the base model.

MHC Pseudosequence

We introduced a new MHC pseudosequence representation. Prior pan-allele methods[3,9] adopted the NetMHCpan[2] pseudosequences, which represent the MHC molecule based on residues estimated to be closest to the peptide. By contrast, BigMHC uses multiple sequence alignments to identify positions with high information content. In total, 18,589 MHC-I sequences across 11 species[23,24] were aligned using the buildmodel and align2model from the SAM suite[25-27] version 3.5 with default parameters, yielding 364 aligned residues per sequence. The top 30 positions by information content were identified using makelogo from the SAM suite and were selected to

represent the MHC sequences, which can be one-hot encoded with 388 binary variables. These new pseudosequences are provided in our public Git repository.

Compared Methods

NetMHCpan-4.1[2] is a popular tool for simultaneously predicting BA and EL. This method consists of an ensemble of 100 single-layer networks, each of which consumes a peptide 9-mer binding core and a subsequence of the MHC molecule. The 9-mer core is extracted by the model, whereas the MHC representation, termed a "pseudosequence," is a predetermined 34-mer core extracted from the MHC molecule sequence. The 34-mer residues were selected based on the estimated proximity to bound peptides so that only residues within 4 Å were included.

MHCflurry-2.0[3] is an ensemble of neural networks that predicts BA and EL for MHC-I. BA prediction is the output of a neural network ensemble, where each member is a 2 or 3 layer feed-forward neural network. Then, an antigen processing (AP) convolutional network is trained on a subset of the BA predictions, along with N-terminus and C-terminus flanking regions, to capture antigen processing information that is missed by the BA predictor. EL prediction is the result of logistically regressing BA and AP outputs. The MHC representation was adopted from NetMHCpan pseudosequences.

TransPHLA[9] is a transformer-based model that adopts NetMHCpan pseudosequences for MHC encoding. This model encodes peptides and MHC pseudosequences using the original transformer encoding procedure[20] before inferring the encodings via ReLU-activated fully-connected layers. TransPHLA was trained on the BA and EL data, although the BA data were binarized as binding or non-binding instances. We found that the final softmax activation of TransPHLA forces many outputs to precisely 0 or 1, which prevents the calculation of metrics,

such as AUROC and AUPRC. Therefore, we removed the final softmax activation from TransPHLA to increase model output granularity. Because softmax is monotonic and none of the evaluation metrics rely on arbitrarily thresholding model outputs, removing the final softmax activation does not affect the evaluation metrics used in this study.

MHCnuggets[1] is comprised of many allele-specific Long Short-Term Memory (LSTM) networks to handle arbitrary-length peptides for MHC-I and MHC-II. Transfer learning was used across the alleles to address data scarcity. MHCnuggets trained on qualitative BA, quantitative BA, and EL data. MHCnuggets trained on up to two orders of magnitude fewer data than the other methods.

MixMHCpred[10,11] is built using positional weight matrices (PWM) to extract epitope motifs for each allele in their training data for peptides of lengths 8 to 14. PRIME[5] builds off MixMHCpred, training directly on immunogenicity, and was designed to infer the mechanisms of TCR recognition of pHLA complexes. MixMHCpred and PRIME cannot make predictions for alleles on which they have not trained.

Evaluation

Duplicate pMHCs were removed from all datasets. Furthermore, all testing instances that exist in the training sets were removed. All methods were evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC). To evaluate precision on the large neoepitope immunogenicity dataset, all methods were evaluated using Positive Predictive Value in the top n predictions (PPVn) and Area Under the PPVn curve (AUPPVn) for all valid choices of n. We chose not to include metrics such as accuracy, Matthew's Correlation Coefficient, or F_1-score because these metrics rely on an arbitrary thresholding model output, which introduces bias to the evaluation.

All the methods that were not trained on the testing datasets were evaluated. As the EL test dataset was large, the results could be stratified by HLA supertype and peptide length, in addition to the overall AUROC and AUPRC curves. The HLA supertype stratification is comprised of AUROC and AUPRC results per allele and is aggregated by supertype in violin plots.

Settings

BigMHC was developed using Python 3.9.12, PyTorch 1.13[28], and CUDA 11.7 on an AMD EPYC 7443P CPU with 256 GB of DDR4, and NVIDIA RTX 3090 with 24 GB of GDDR6X. The maximum batch size was set to 6144 to maximally occupy the available GPU global memory. Each batch is comprised only of peptides of equal length. The Adam optimizer[29] is used with $\gamma = 10^{-4}, \lambda = 10^{-5}, \beta_1 = 0.900, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$ to minimize binary cross entropy loss. BigMHC EL training and BigMHC IM transfer learning were conducted for 50 epochs on the respective datasets. Transfer learning was performed with a batch size of 1024. All dropout modules are set to have a dropout rate of 0.5 to maximize regularization. Each model had 512 hidden neurons for all layers except the final attention layer, which had 388 neurons to match the MHC sequence encoding length.

Ethics Declaration

Under a license agreement between Genentech and the Johns Hopkins University, X.M.S., and R.K., and the University are entitled to royalty distributions related to MHCnuggets technology discussed in this publication. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict-of-interest policies. V.A. receives research funding to her institution from Astra Zeneca and has received research funding to her institution

from Bristol Myers Squibb over the past five years. The remaining authors have declared no conflicts of interest.

Data Availability

All data except MANAFEST data were collected from publicly available sources[2,3,5,6,12-14]. All data, including model outputs, are provided in our public Mendeley repository (https://doi.org/10.17632/dvmz6pkzvb.1).

Code Availability

All code used in this study is provided in our public GitHub repository under the Apache v2.0 license: https://github.com/KarchinLab/bigmhc. In addition, all trained models are provided in this repository.

Contributions

B.A.A. and R.K. conceived the study and performed the experiments; Y.Y. contributed to 3D visualizations and model ideas; X.M.S. curated the MANAFEST data; D.S. and K.N.S. collected the MANAFEST dataset; B.A.A. and R.K. wrote the draft manuscript; B.A.A., V.A. and R.K. revised the manuscript; R.K. supervised the research.

References

1 Xiaoshan, S. M. *et al*. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunology Research* **8**, 396-408 (2020).

2 Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research* **W1**, 48 (2020).

3 O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Systems* **11**, 42-48 (2020).

4 Lu, T. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence* **3**, 864-875 (2021).
5 Schmidt, J. *et al.* Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Reports Medicine* **2**, 100194 (2021).
6 Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818-834 (2020).
7 Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1-13, doi:10.1007/s00251-008-0341-z (2009).
8 Nielsen, M. *et al.* NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLOS ONE* **2**, e796, doi:10.1371/journal.pone.0000796 (2007).
9 Chu, Y. *et al.* A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence* **4**, 300–311 (2022).
10 Gfeller, D. *et al.* The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *The Journal of Immunology* **201**, 3705-3716 (2018).
11 Bassani-Sternberg, M. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS computational biology* **13**, e1005725 (2017).
12 Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* **47**, D339-D343 (2018).
13 Xia, J. *et al.* NEPdb: a database of T-cell experimentally-validated neoantigens and pan-cancer predicted neoepitopes for cancer immunotherapy. *Frontiers in Immunology* **12**, 644637 (2021).
14 Kim, S. *et al.* Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology* **29**, 1030-1036 (2018).
15 Danilova, L. *et al.* The Mutation-Associated Neoantigen Functional Expansion of Specific T Cells (MANAFEST) Assay: A Sensitive Platform for Monitoring Antitumor Immunity. *Cancer Immunology Research* **6**, 888-899 (2018).
16 Anagnostou, V. *et al.* Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non–Small Cell Lung Cancer. *Cancer Discovery* **7**, 264–276 (2017).
17 Jones, S. *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Science Translational Medicine* **7**, 283ra253 (2015).
18 Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine* **8**, 1-9 (2016).
19 Stranzl, T., Larsen, M. V., Lundegaard, C. & Nielsen, M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* **62**, 357–368 (2010).
20 Vaswani, A. *et al.* in *Advances in neural information processing systems 30 (NeurIPS)*.
21 Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4700-4708.
22 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929-1958 (2014).
23 Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Research* **48**, D948–D955 (2020).
24 Maccari, G. *et al.* IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. *Nucleic Acids Research* **45**, D860–D864 (2017).
25 Hughey, R. & Krogh, A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics* **12**, 95-107 (1996).
26 Karplus, K. *et al.* What is the value added by human intervention in protein structure prediction? *Proteins: Structure, Function, and Bioinformatics* **45**, 86-91 (2001).
27 Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of Molecular Biology* **235**, 1501-1531 (1994).
28 Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32 (NeurIPS)*.
29 Kingma, D. P. & Ba, J. L. in *3rd International Conference for Learning Representations (ICLR)*.
30 Rego, N. & Koes, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* **31**, 1322–1324 (2015).
31 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Display Items

		BigMHC (Proposed)	TransPHLA	PRIME	MHCflurry-2.0	NetMHCpan-4.1	MHCnuggets-2.3.2	MixMHCpred-2.1
Year Released		2022	2022	2021	2021	2020	2020	2020
Training	BA		✓		✓	✓	✓	✓
	EL	✓	✓		✓	✓	✓	✓
	IM	✓		✓				
Prediction	BA				✓	✓		
	EL	✓	✓		✓	✓	✓	✓
	IM	✓		✓				
1x GPU	Seconds	42.49	74.31	34.15	54.67	14645.10	184.81	24.62
	Speed	1	0.57	1.24	0.78	0.0029	0.23	1.73
2x GPU	Seconds	27.04	74.31	34.15	54.67	14645.10	184.81	24.62
	Speed	1	0.36	0.79	0.49	0.0018	0.15	1.10
Retrainable		✓					✓	
Open Source		✓	✓	✓	✓		✓	✓
Pan-Allele		✓	✓		✓	✓		
GPU-Accelerated		✓	✓		✓		✓	
Optional Multi-GPU		✓						
Source Language		Python 3	Python 3	C++ / Perl	Python 3	Unknown	Python 2	C++ / Perl
Backend		PyTorch	PyTorch	None	TensorFlow	None	TensorFlow	None
Has Webserver			✓			✓		
Min peptide length		8	8	8	5	8	None	8
Max peptide length		None	15	14	15	None (14)	None	14
Allows WT AA ‘X’		✓	✓			✓	✓	

Table 1 Method and feature comparison of BigMHC and prior works. Training rows indicate the type of data on which models trained, whereas Prediction rows indicate what type of peptides the model explicitly predicts. BA indicates binding affinity, EL indicates eluted ligand (presentation), and IM indicates immunogenicity. Models that are provided with executables or source code for retraining on new data are considered retrainable. Pan-allele methods are ones that encode the MHC sequence to generalize predictions across alleles rather than employing multiple allele-specific models. Execution speed is evaluated on an AMD EPYC 7443P with one and two RTX 3090 GPUs. Relative speed is normalized to BigMHC on the EL test dataset; 2 indicates twice as fast as BigMHC and 0.5 indicates half as fast. NetMHCpan does not have a maximum peptide length when consuming pMHC pairs, but caps peptide lengths to 14 when consuming FASTA data. Models that can consume wild-type amino acids, such as ‘X’, are indicated as such in the final row.

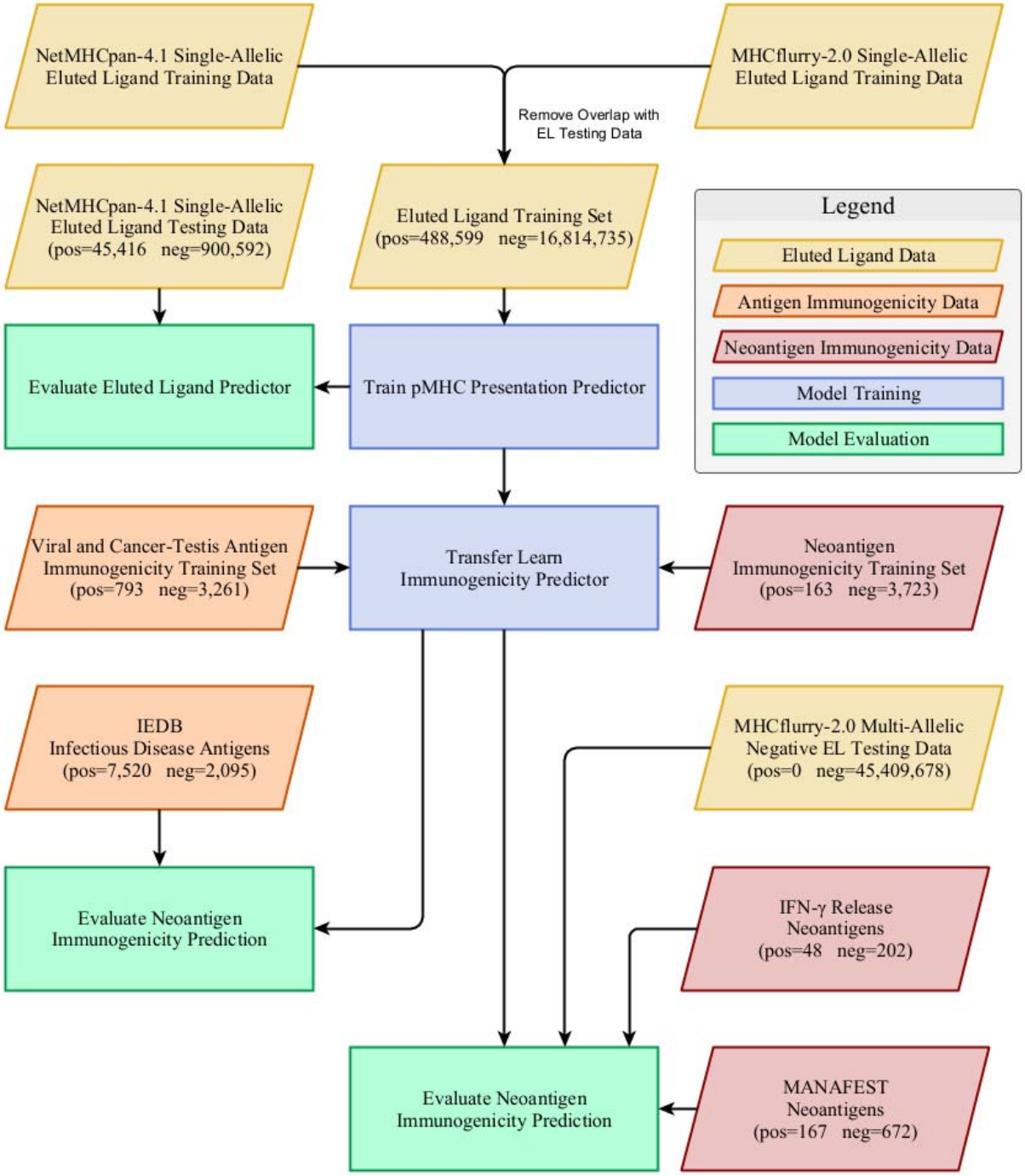


Fig. 1 Experiment Procedure. The procedure includes eluted ligand training, immunogenicity transfer learning, and independent evaluation on multiple datasets.

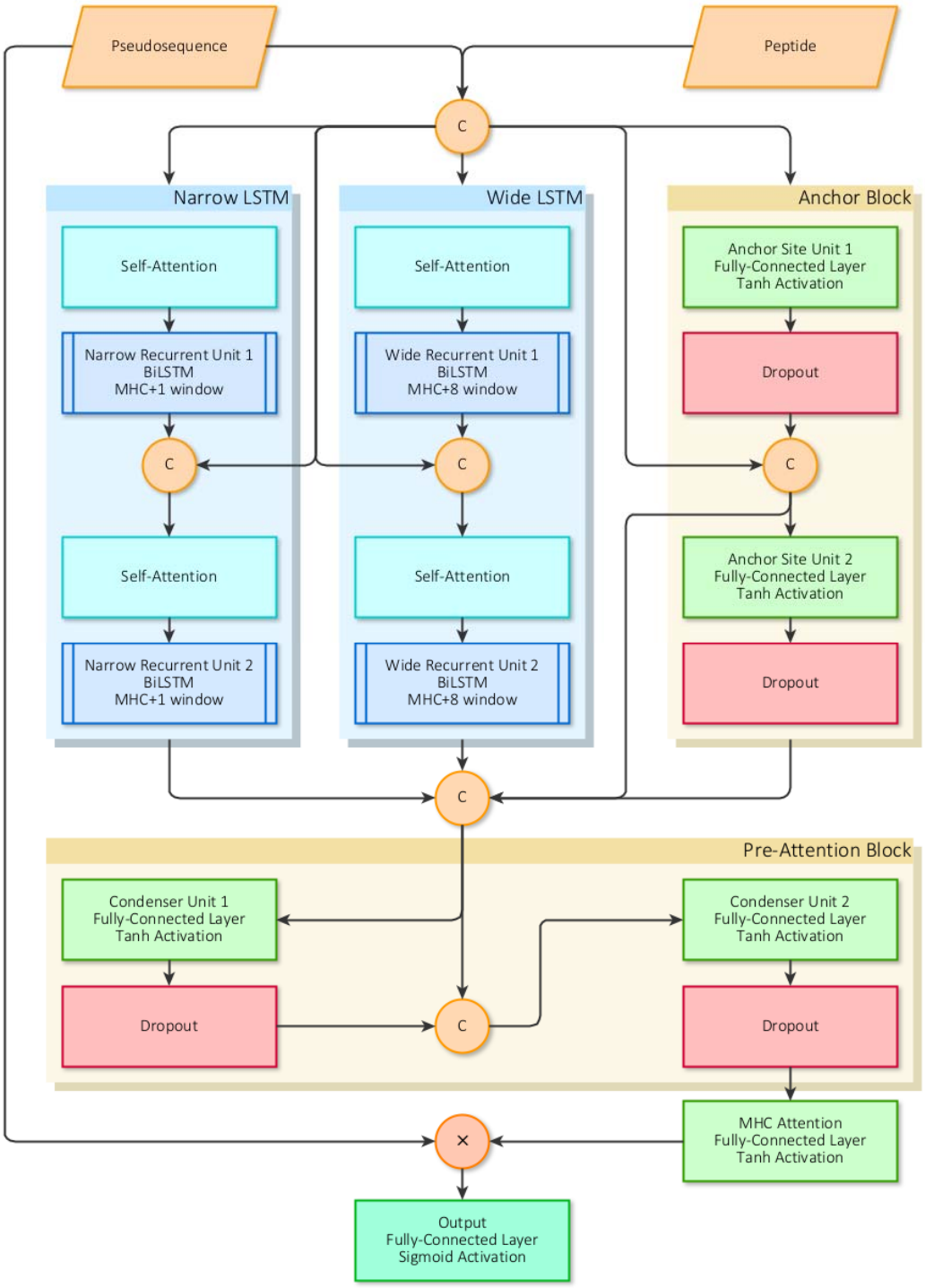


Fig. 2 BigMHC network architecture. The pseudosequences and peptides are one-hot encoded prior to feeding them into the model. The circles labelled 'C' indicate concatenation and the circle labelled '×' denotes elementwise multiplication. The self-attention modules are single-headed attention units, which is analogous to setting the number of heads of a standard multi-headed Transformer attention module to 1. Prior to the final sigmoid activation, the output of the model is a weighted sum of the MHC pseudosequence onehot encoding; the weights are referred to as attention. Because all connections except internal BiLSTM cell connections are dense, data are not bottlenecked until the MHC attention node maps the pre-attention block output to a tensor of the same shape as the onehot-encoded MHC pseudosequences.

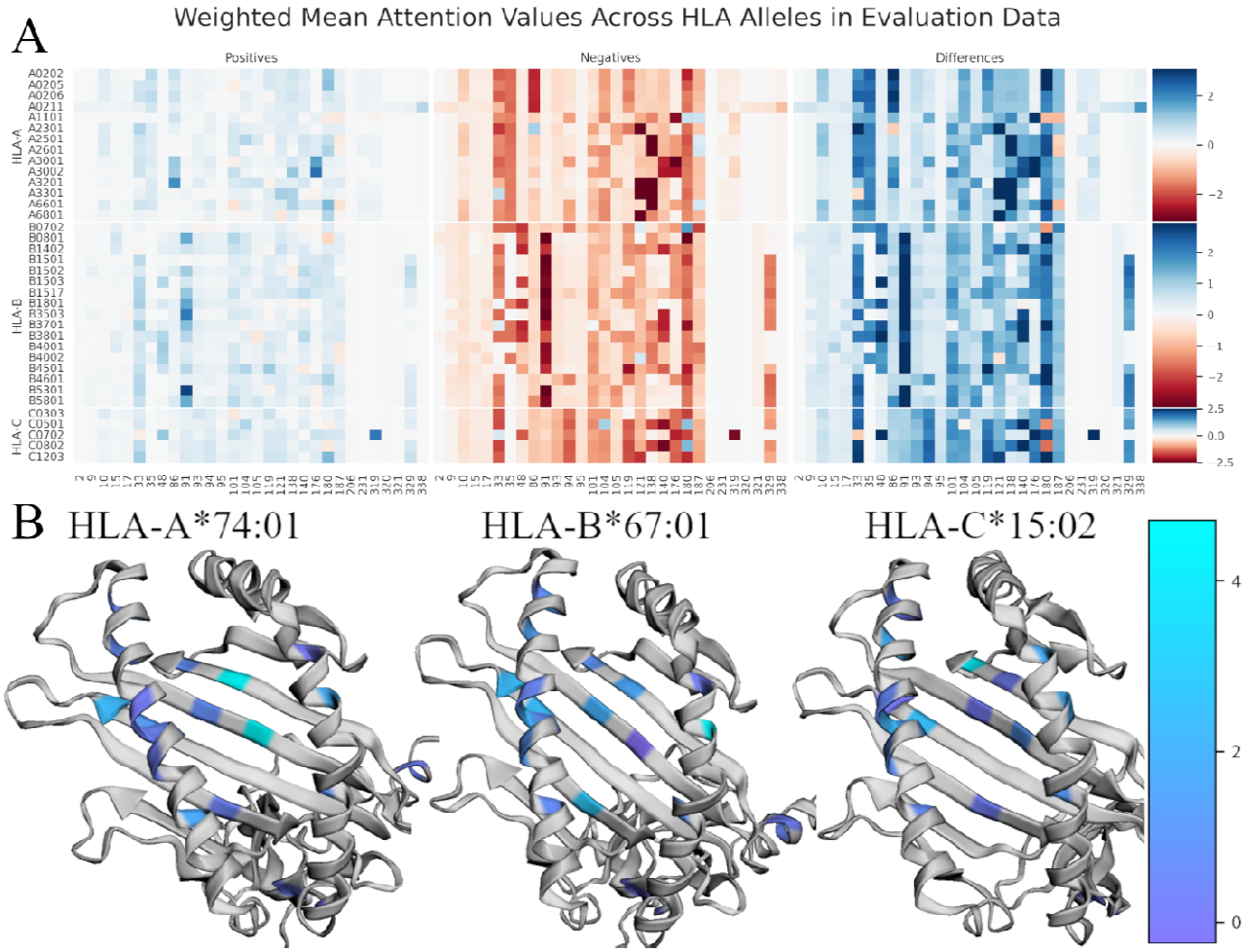


Fig. 3 Visualization of BigMHC attention to MHC encodings. A-B, **A** Heatmap visualization of the average MHC attention values stratified by MHC allele and EL target class on the EL testing dataset. Darker values indicate MHC positions that are more influential on the final model output. The column of Differences depicts the Negatives values subtracted from the Positives values; thus, darker blue colors are most correctly discriminative whereas darker red attention values in this column highlight erroneous inferences. **B** Overlays of the Differences column from the training dataset on the MHC molecule using py3Dmol[30]. HLA crystal structures are retrieved from the AlphaFold Protein Structure Database[31]. The AlphaFold database contains one of each HLA molecule supertype, but the training dataset does not contain HLA-B*67:01; as such, the Differences column for HLA-B*42:01 is used to overlay on HLA-B*67:01 as these pseudosequences differ by only 3 out of 30 pseudosequence residues.

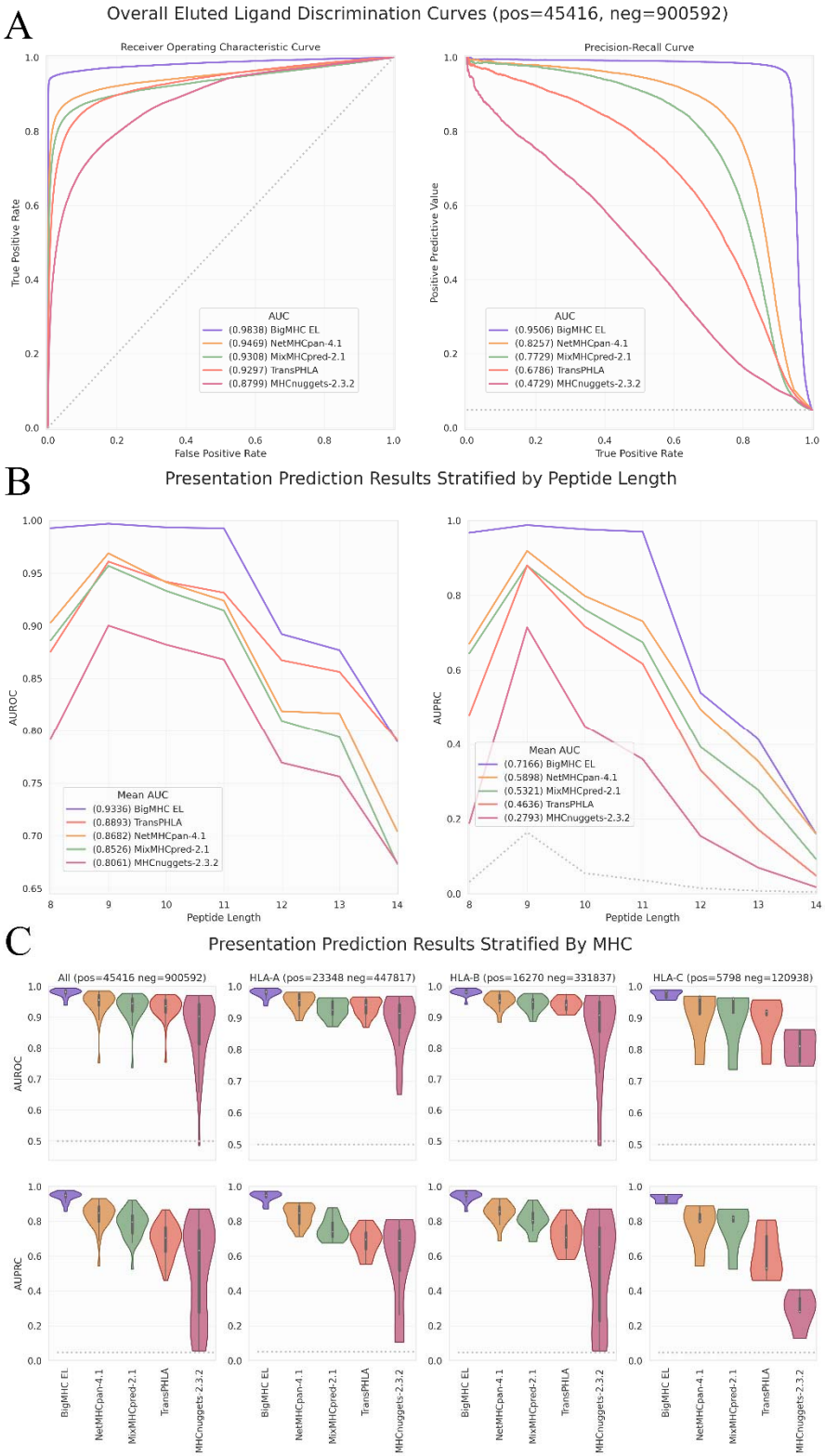


Fig. 4 Eluted Ligand prediction results. A-C, **A** Overall Receiver Operating Characteristic (ROC) and Precision-Recall (PRC) curves for the eluted ligand (EL) testing dataset. **B** Area under the ROC (AUROC) and PRC (AUPRC) curves on the EL testing dataset stratified by epitope length, along with macro-averaged results. **C** AUROC and AUPRC on the EL testing dataset stratified by HLA supertype.

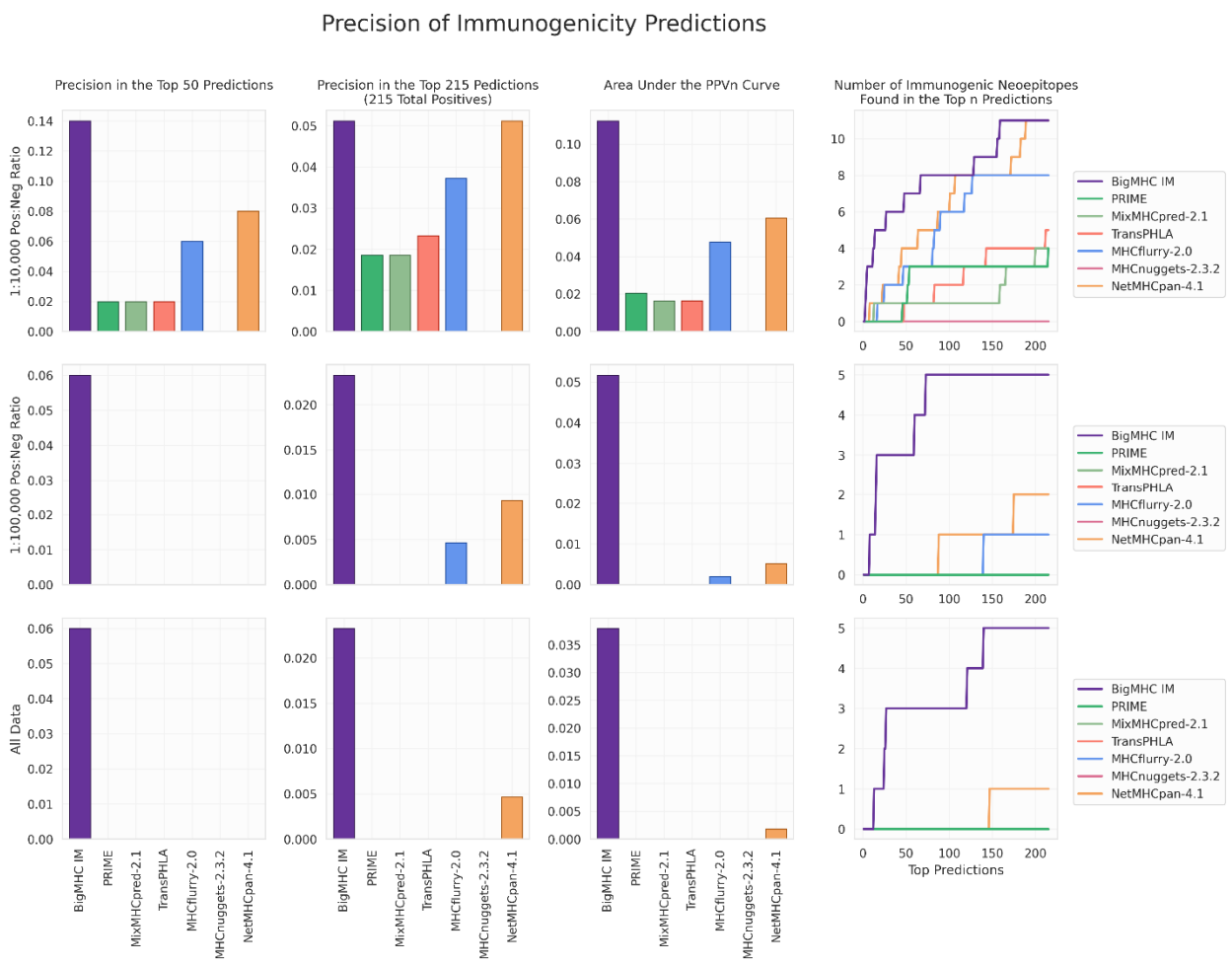
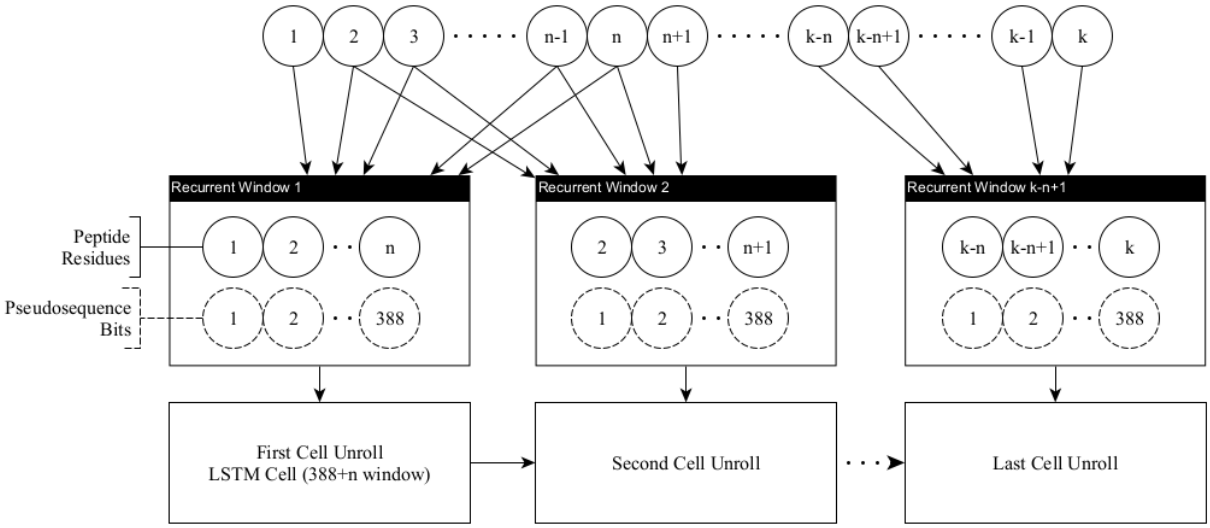
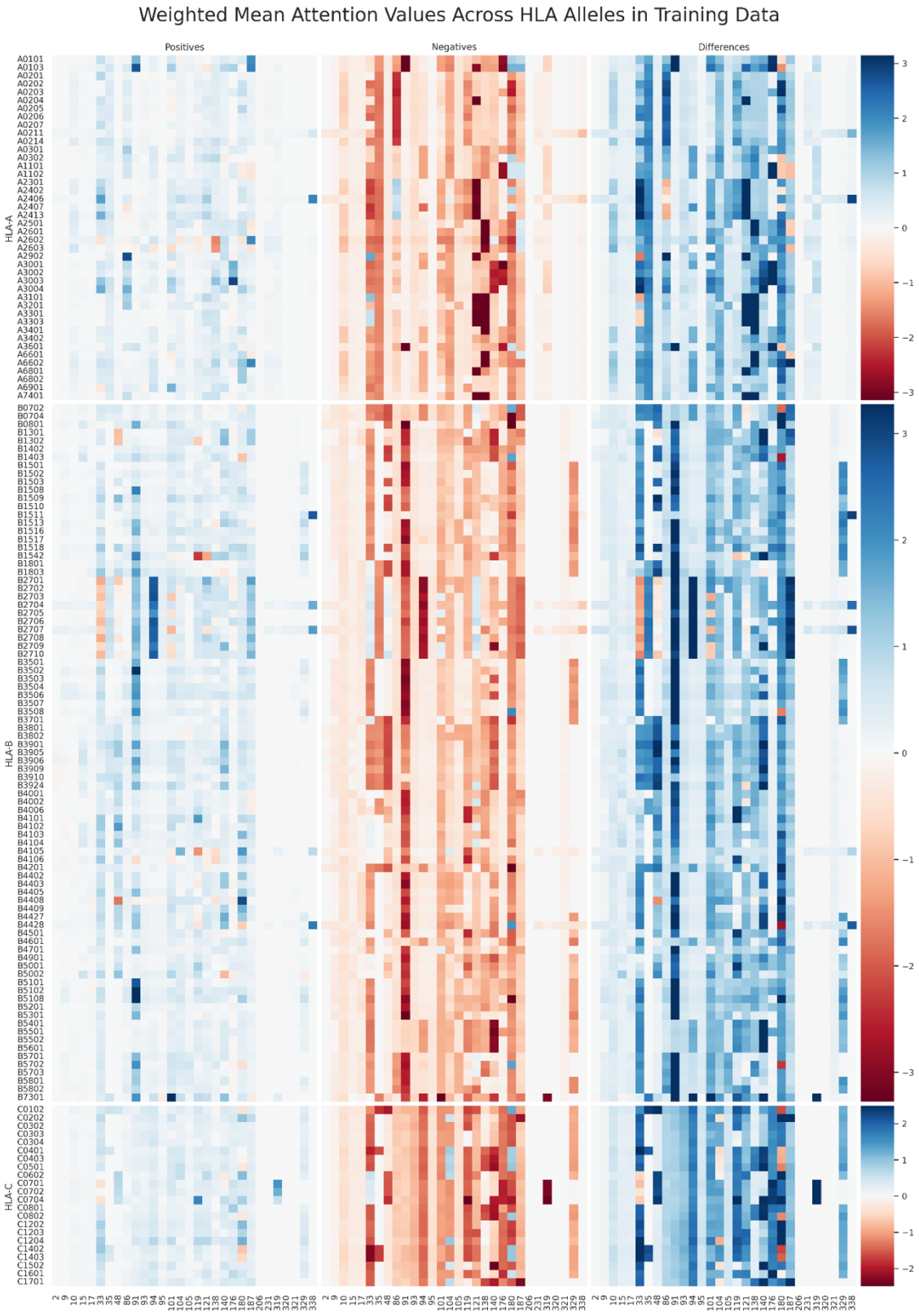


Fig. 5 Precision of immunogenicity predictions for all methods. Precision is calculated for each method's top 50 predictions and top 215 predictions for this dataset containing 215 immunogenic neoepitopes and 45,410,552 non-immunogenic. Additionally, the area under the PPVn curve is shown along with the curve illustrating the number of immunogenic neoepitopes found in the top n predictions.

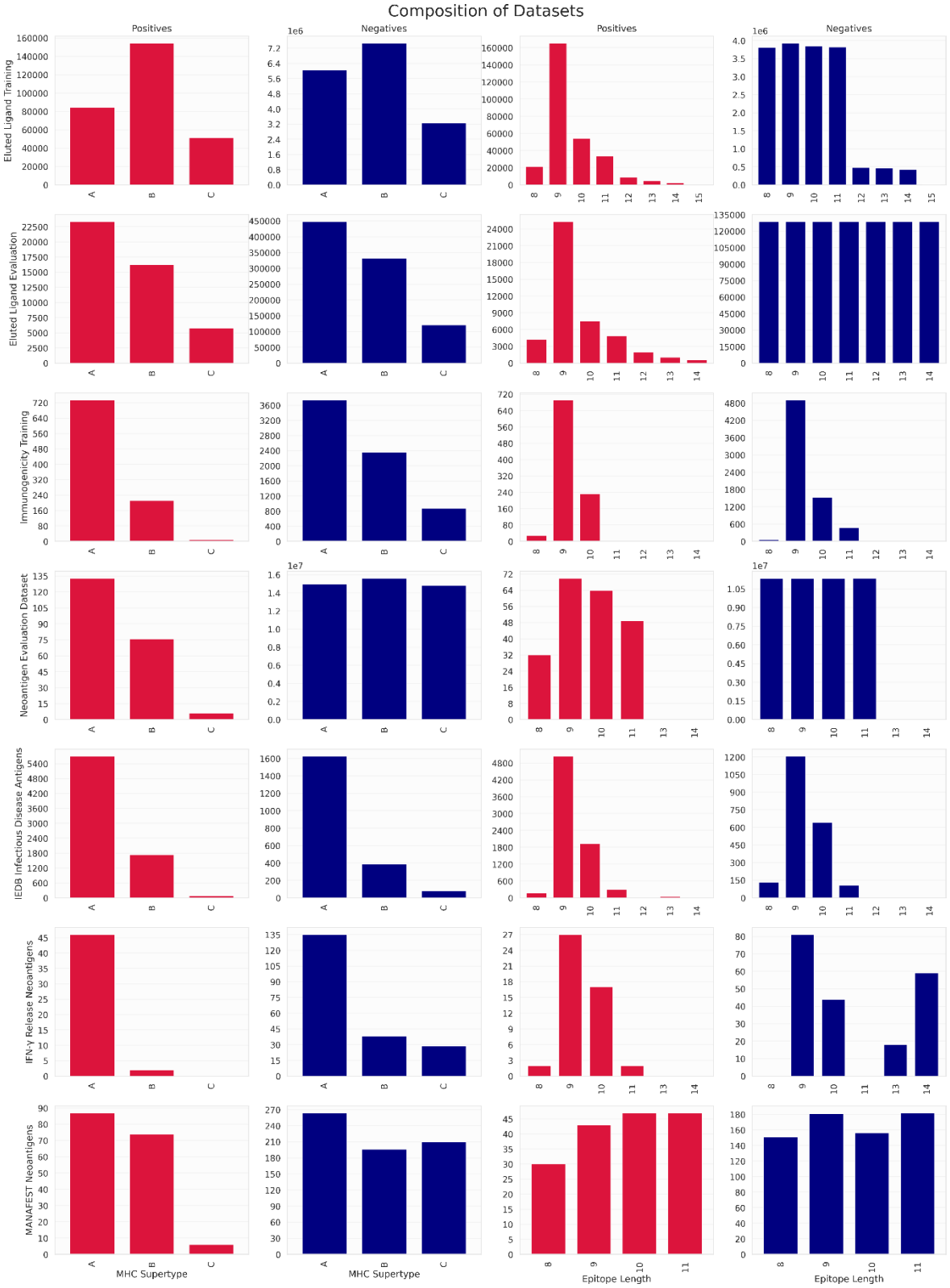
Extended Figures



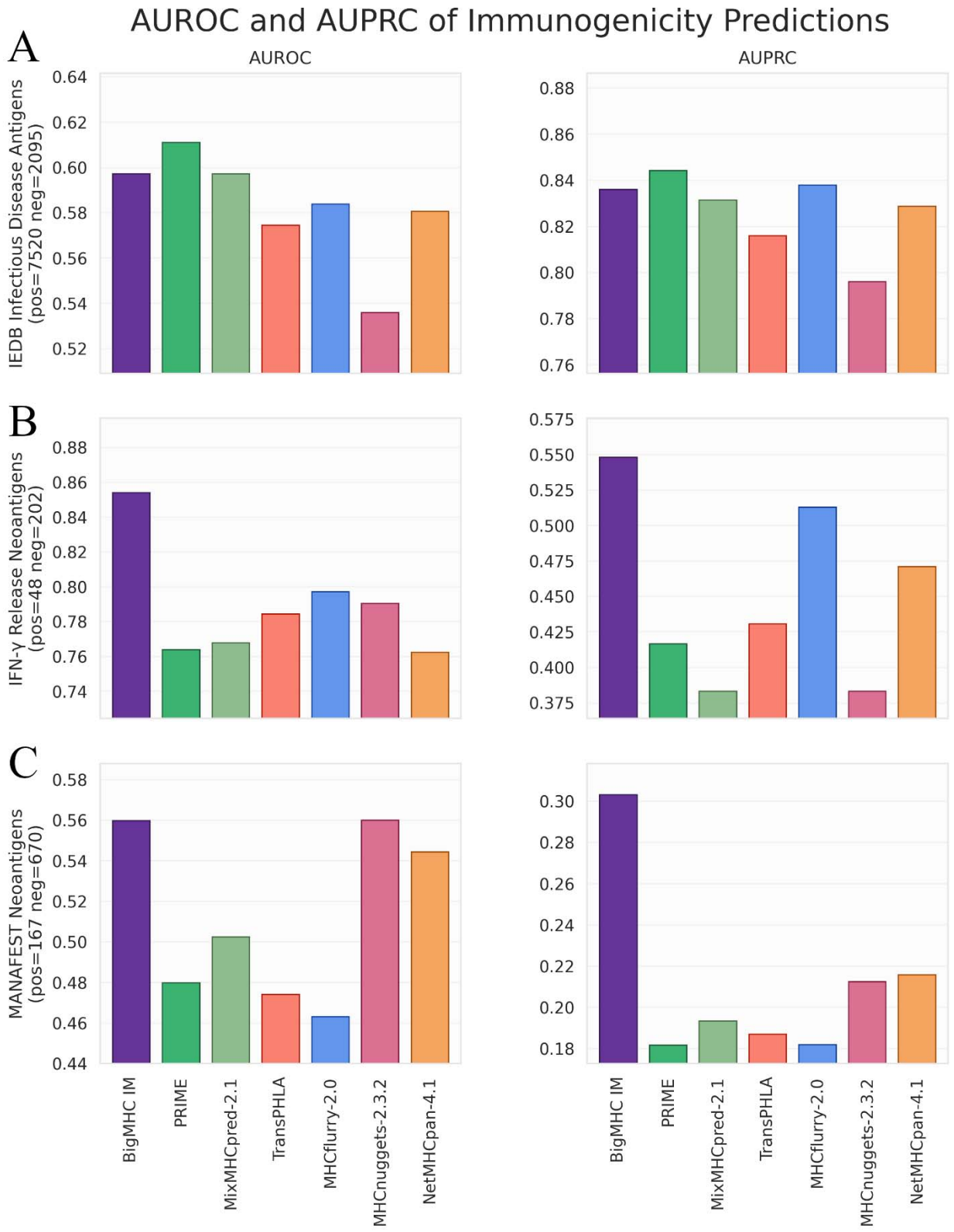
Extended Data Fig. 1 LSTM with a variable-length peptide window. Each cell unroll processes the entire MHC pseudosequence but only a fixed-length window of the peptide. The BigMHC Narrow LSTM uses a window length of 1, whereas the BigMHC Wide LSTM uses a window length of 8.



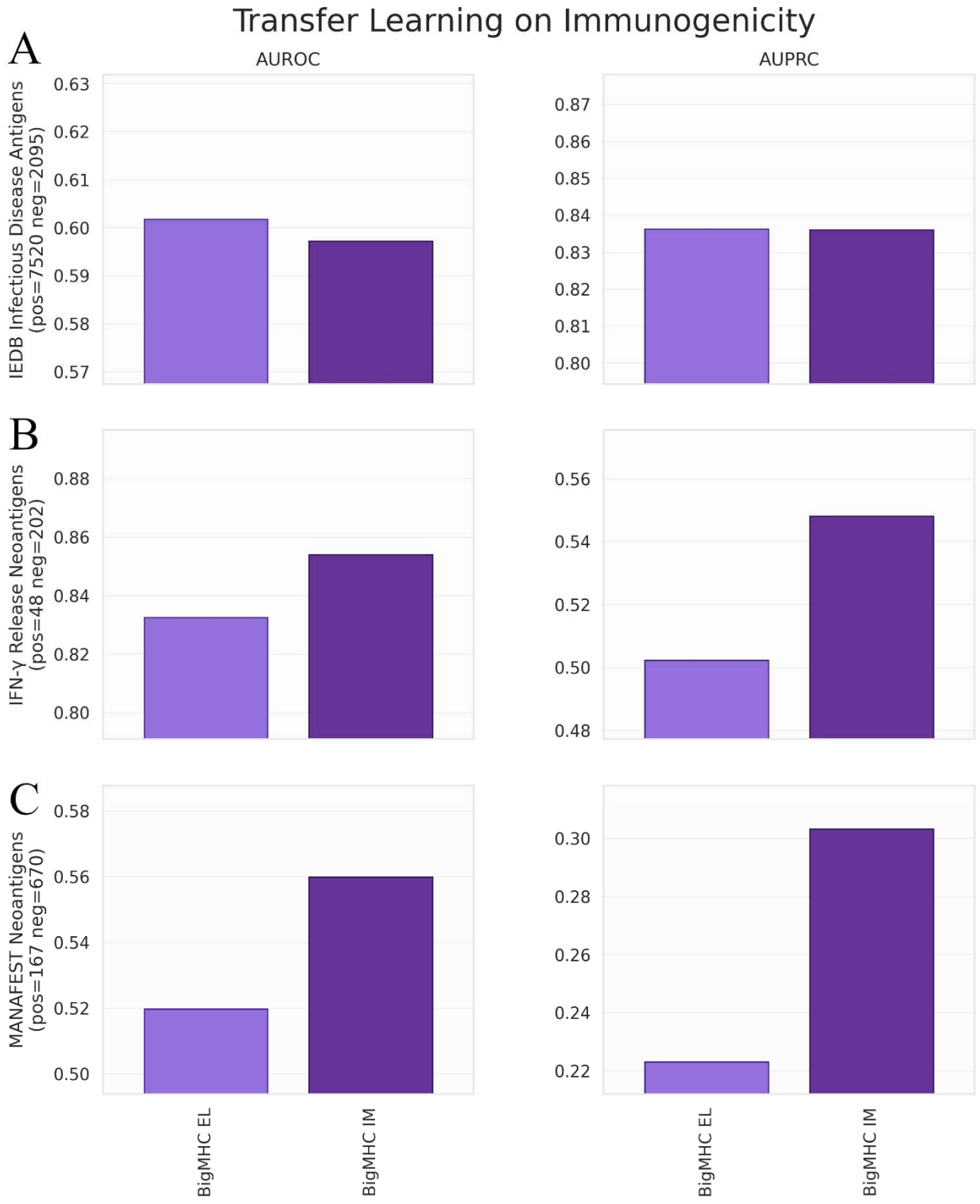
Extended Data Fig. 2 Heatmap visualization of the average MHC attention values stratified by MHC allele and EL target class on the EL training dataset, analogous to the visualization of the EL testing set in Fig. 3.



Extended Data Fig. 3 Composition of all training and evaluation datasets. Positive and negative instances were stratified by HLA supertype in the first two columns and by epitope length in the latter two columns.



Extended Data Fig. 4 Immunogenicity AUROC and AUPRC. A-C, **A** Infectious disease antigen predictions. **B** IFN-γ Release neoepitope predictions. **C** MANAFEST neoepitope predictions.



Extended Data Fig. 5 Results of BigMHC EL transfer learning on immunogenicity data. A-C, **A** Infectious disease antigen predictions. **B** IFN-γ release neoepitope predictions. **C** MANAFEST neoepitope prediction.