**TITLE**

Discordant results among MHC binding affinity prediction tools

**AUTHORS**

Austin Nguyen (1, 2), Abhinav Nellore* (1, 2, 3), Reid F. Thompson* (1, 2, 4, 5, 6)

**AFFILIATIONS**

1. Computational Biology Program; Oregon Health & Science University; Portland, OR, 97239; USA
2. Department of Biomedical Engineering; Oregon Health & Science University; Portland, OR, 97239; USA
3. Department of Surgery; Oregon Health & Science University; Portland, OR, 97239; USA
4. Department of Radiation Medicine; Oregon Health & Science University; Portland, OR, 97239; USA
5. Department of Medical Informatics and Clinical Epidemiology; Oregon Health & Science University; Portland, OR, 97239; USA
6. Division of Hospital and Specialty Medicine; VA Portland Healthcare System; Portland, OR, 97239; USA

* co-corresponding authors [nellore@ohsu.edu, thompsre@ohsu.edu]

## ABSTRACT

A large number of machine learning-based Major Histocompatibility Complex (MHC) binding affinity (BA) prediction tools have been developed and are widely used for both investigational and therapeutic applications, so it is important to explore differences in tool outputs. We examined predictions of four popular tools (netMHCpan, HLAthena, MHCflurry, and MHCnuggets) across a range of possible peptide sources (human, viral, and randomly generated) and MHC class I alleles. We uncovered inconsistencies in predictions of BA, allele promiscuity and the relationship between physical properties of peptides by source and BA predictions, as well as quality of training data. Our work raises fundamental questions about the fidelity of peptide-MHC binding prediction tools and their real-world implications.

## INTRODUCTION

Human Leukocyte Antigen (HLA) alleles are critical components of the immune system's ability to recognize and eliminate tumors and infections (1). Infectious diseases in particular are thought to be a major source of selective pressure on the Major Histocompatibility Complex (MHC) region which encodes HLA alleles and is one of the most diverse regions of the human genome (2–8). There is large diversity in the antigenic peptide sequences which individual HLA alleles can recognize and ultimately present to the adaptive immune system (9), with a positive correlation between increased sequence diversity recognition and fitness (10).

Tools that can predict the extent to which a given HLA allele may have an affinity for a given peptide have critical implications for our ability to understand and translationally leverage antigen-specific immune response pathways. For instance, MHC binding affinity predictors have been – or otherwise have the potential to be – used to evaluate an individual or population's susceptibility to viral infection (11), to develop an understanding of specific autoimmune conditions (12), to improve transplantation technologies (13), or even to assist in the development of personalized cancer vaccines (14–18). Numerous peptide-MHC binding prediction tools exist, and are key components in broader antigen prediction methodologies (19–22).

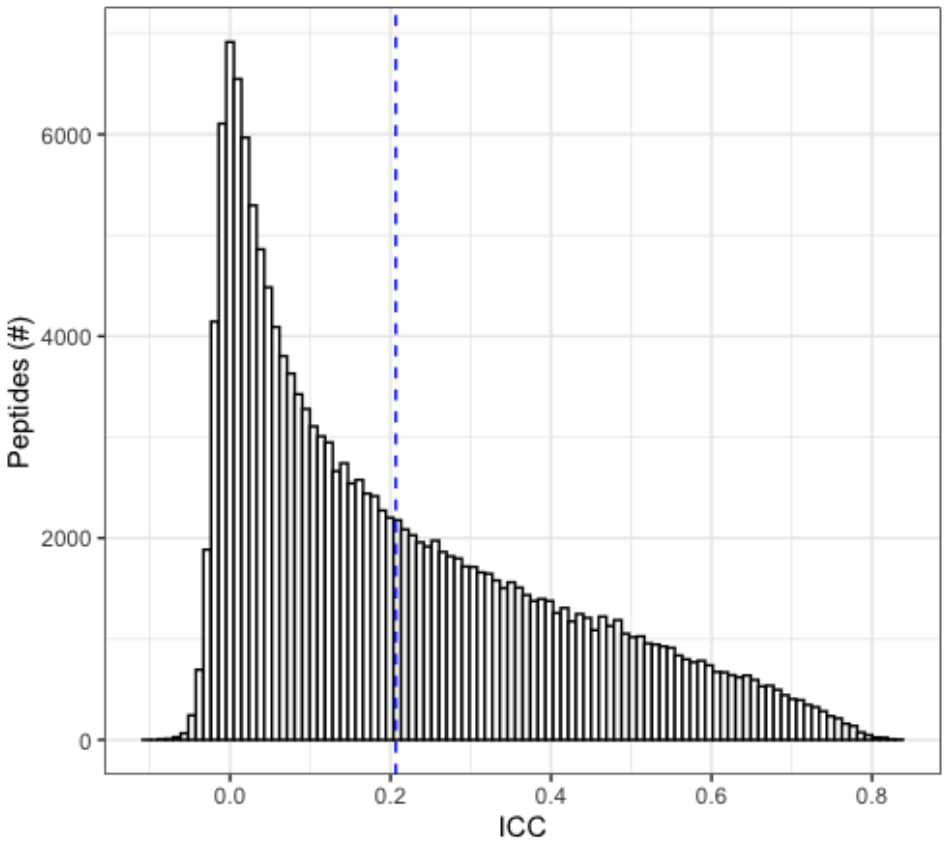
The most widely adopted MHC binding prediction tools rely on neural network models trained on binding affinity (BA) and/or eluted ligand (EL) data. The most commonly cited tool, netMHCpan (23,24), uses both BA and EL data in a neural network architecture with a single hidden layer to predict allele-specific binding affinities. MHCflurry (25) attempts to improve upon netMHCpan by increasing the number of hidden layers and augmenting BA and EL training data with unobserved decoys. MHCnuggets (26) again trains on BA and EL data but uses a different architecture, with a long short-term memory layer and a fully connected layer to improve its predictions further across different peptide lengths. Lastly, HLAthena (27), while most similar in architecture to netMHCpan, relies on independently generated EL data from mono-allelic cell lines for training.

We sought to better characterize the outputs of these tools over a large and diverse set of peptides, across different tools and HLA alleles, as well as quantify the stability of these predictions. We also sought to measure allelic binding preferences and whether they may enrich for foreign v. self peptides. In this study, we performed a comprehensive *in silico* analysis of peptides from multiple viral proteomes, the human proteome, and randomly generated peptides across HLA class I alleles.

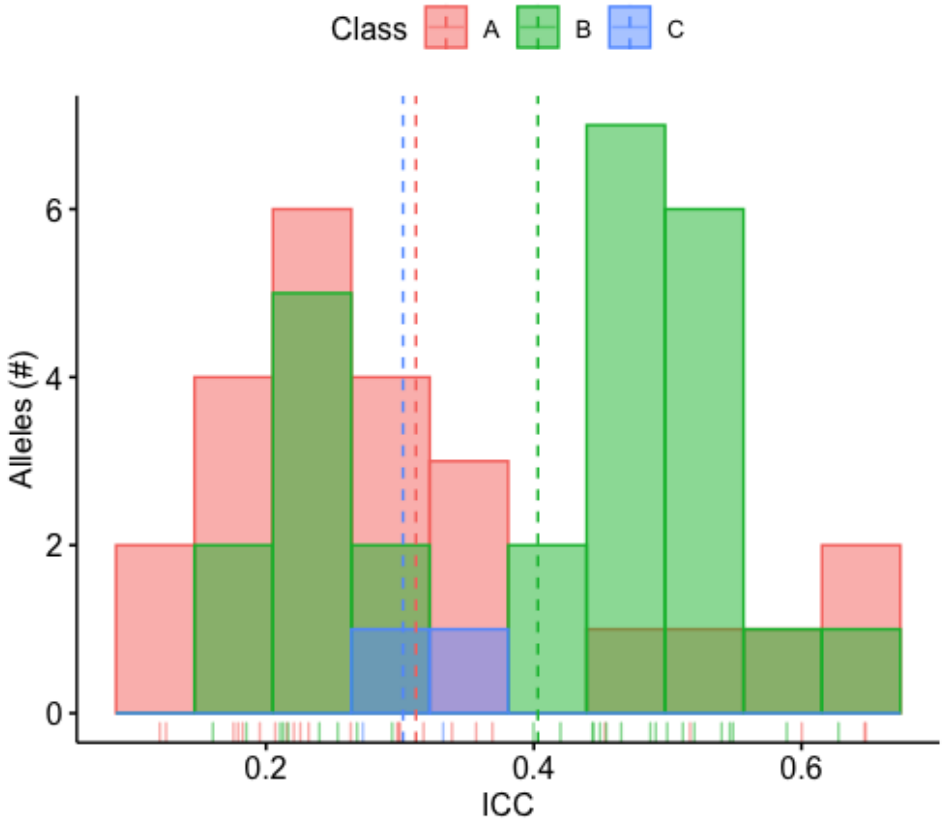
## RESULTS

### Peptide predictions are inconsistent across tools

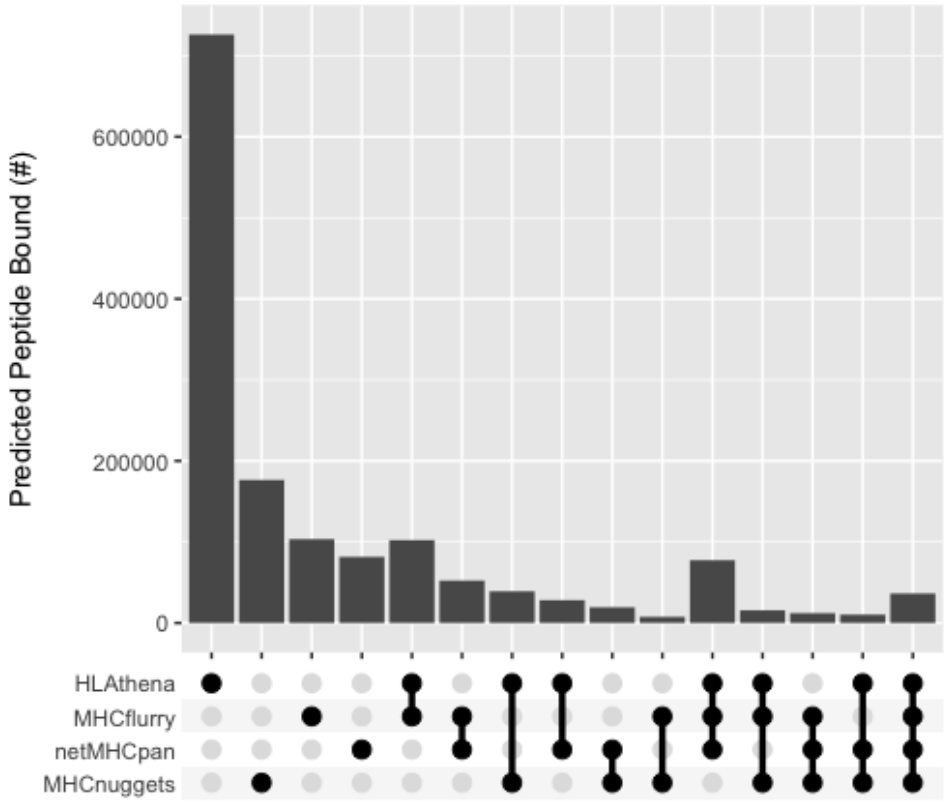
We first assessed the consistency of peptide-specific MHC I binding affinity predictions across four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan) and 52 different HLA alleles. We found substantial disagreement in peptide-specific predictions between each tool, independent of allele (Figure 1A), with median intraclass correlation coefficient (ICC) of 0.207 and only 0.48% of peptides having ICC > 0.75. On a per-allele basis, we found a wide range in consistency of predictions across tools, with a mean intraclass correlation as low as 0.12 for A02:07 and as high as 0.64 for A23:01 (Figure 1B). Among all of the peptides predicted by at least one tool to bind to at least one allele, only 7.9% were consistently predicted across all tools to bind to the same allele (Figure 1C).




**A**




**B**




**C**

*Figure 1. Inconsistency of peptide predictions across tools. A) Histogram of intraclass correlation coefficients (ICC) calculated for a set of 1 million random peptides across four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan), with ICC calculated as the overall correlation among tools across 52 HLA alleles. The dotted vertical line indicates the median ICC value (0.207) across all peptides. B) Histogram of ICCs for 52 HLA alleles between four tools (MHCnuggets, MHCflurry, HLAthena, netMHCpan). The number of alleles is shown on the y-axis and the ICC is shown on the x-axis. The dotted lines show the mean ICC for alleles belonging to each HLA class. Red, green, and blue colors represent data from -A, -B, and -C alleles, respectively. C) Detailed comparison of the complete set of random peptides predicted to bind (binding score >=0.5) to HLA alleles according to each of four tools. Patterns of agreement or disagreement among groups of peptides predicted by different combinations of tools across 1 million random peptides are shown along each column (e.g. the first column corresponds to peptides predicted by HLAthena while the final column corresponds to peptides predicted by all tools). Each row indicates the predictions associated with the indicated tool. The number of peptides in each column (vertical bars) corresponds to the size of the subset predicted by the indicated combination of tools.*

We next investigated aggregate peptide binding predictions across different HLA alleles according to each tool. As others have noted differential HLA allelic promiscuity in peptide presentation (28–31), we too found a wide range in the proportion of peptides a given allele was predicted to bind (Supplementary Figure 1). We

uncovered significant inconsistencies in these predictions between tools (Figure 2). Note that this phenomenon is independent of binding affinity threshold (Supplementary Figure 2).

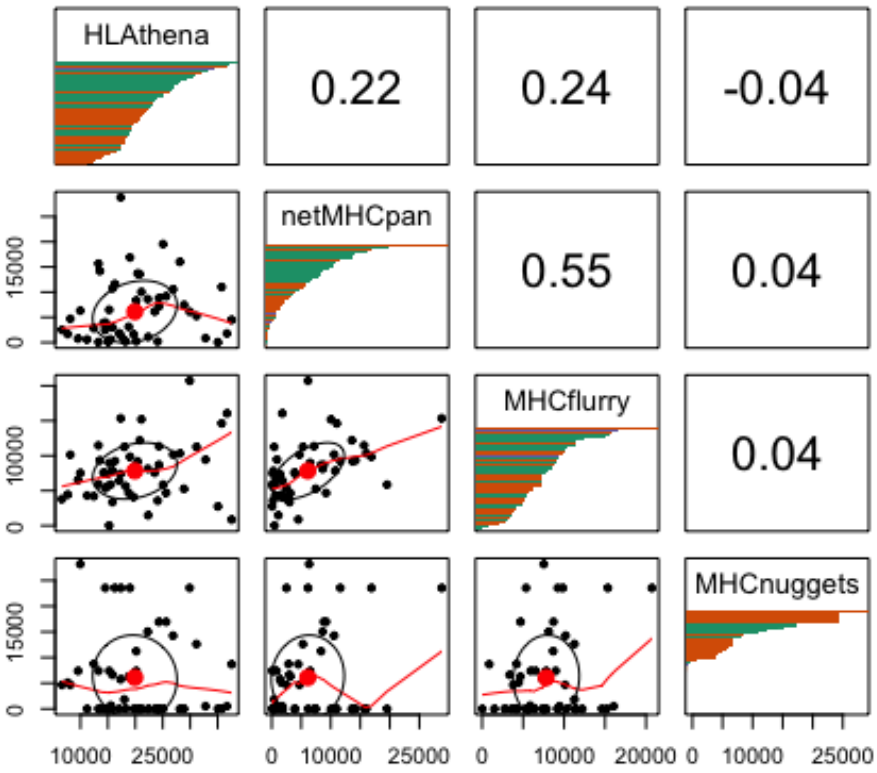
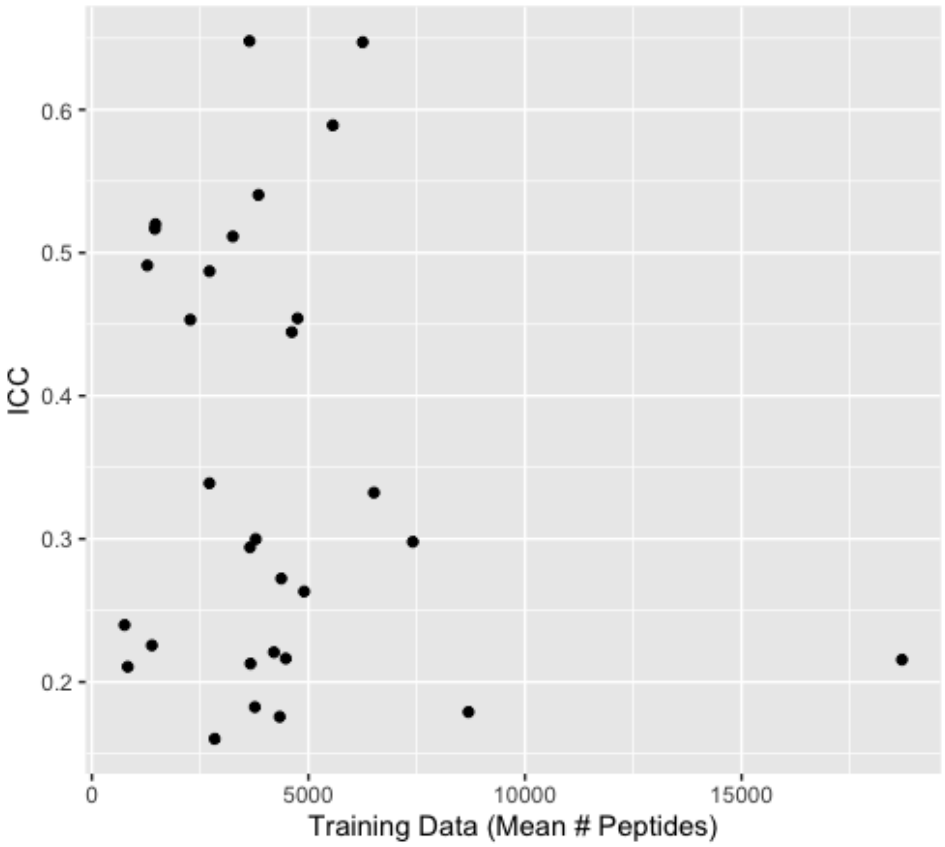



*Figure 2: The correlation of HLA allelic presentation of 8-11mers from the random proteome between tools. The lower left grouping of plots displays scatter plots of peptides predicted to bind (>= 0.5 binding probability score) between 2 tools with each point representing the number of predicted binders for each HLA allele. The upper right grouping represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.*

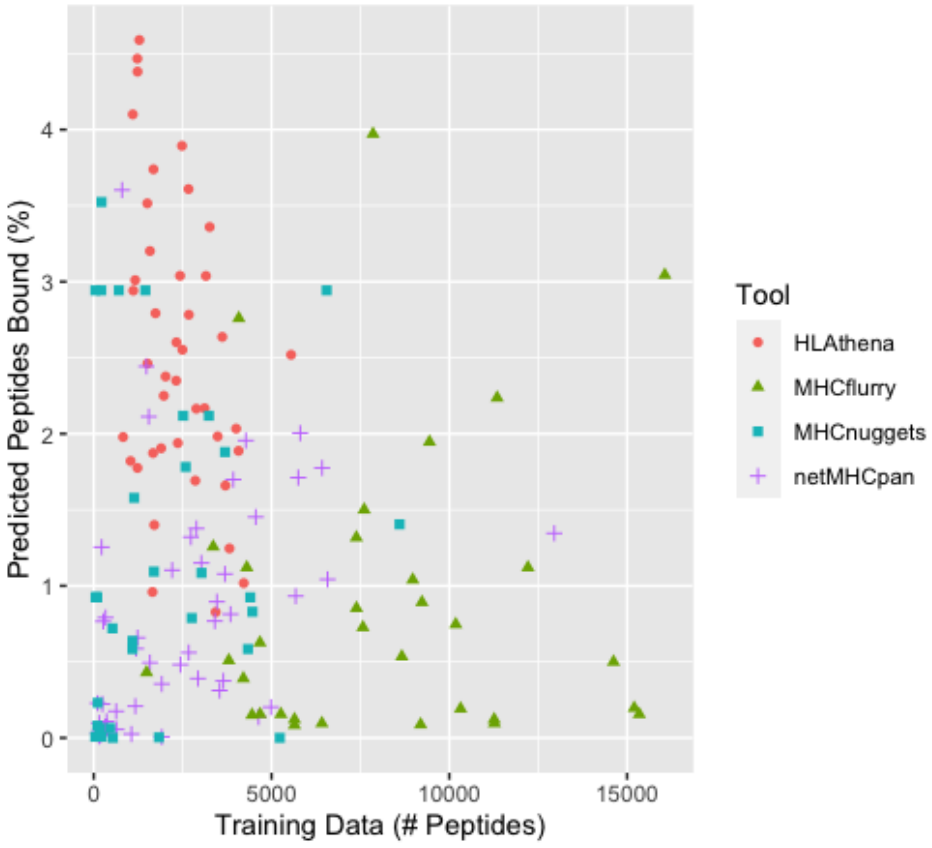
**Amount of training data does not explain inconsistencies between tools**

As each allele has a different amount of training data, we were next interested in exploring to what extent the quantity and quality of training data available to each tool might influence its allele-specific predictions. Indeed,

some netMHCpan predictive models for some alleles are based on as few as 101 peptides, while others from MHCflurry are based on as many as 31,775 peptides (Supplementary Table 1). Note that we excluded from consideration the ~95% of alleles (4108) that were available for prediction but had no underlying allele-specific training data available (Supplementary Table 2). Ultimately, we found that the amount of training data available was not significantly related to the consistency of binding predictions between tools (Figure 3a), nor was it clearly related to the quantity of binding peptides predicted by tools (Figure 3b).




**A**




**B**

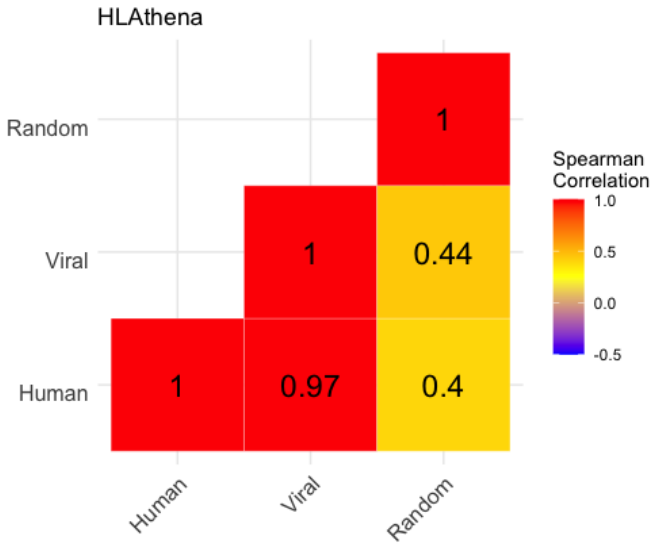
*Figure 3. The relationship between training data and consistency of predictions. A) Scatterplot of ICC vs mean training data across 4 tools with each point representing data for a single HLA allele. The mean number of training peptides is shown on the x-axis while the ICC score is shown on the y-axis. B) Scatterplot of the relationship between training data and predicted peptide binding. The number of peptides used as training data for an allele is shown on the x-axis whereas the number of peptides predicted to bind for the same allele is shown on the y-axis. Each dot is a single allele with each color representing a different tool: red circles (HLAthena), green triangles (MHCflurry), blue squares (MHCnuggets), purple plus signs (netMHCpan). We note that netMHCpan does not make all of their training data available, thus the depicted quantity of training data represents an estimate.*

**Predicted binding quantities are similar between human and viral proteomes**

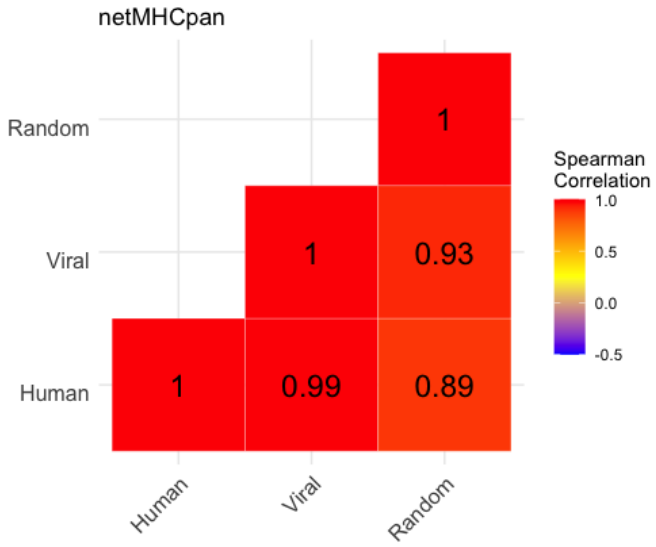
According to the pathogen driven selection theory of MHC evolution, different HLA alleles are anticipated to be particularly attuned to foreign as opposed to self antigens (3,8,32–35). We therefore sought to compare the predicted capacity of different HLA alleles to present different viral vs. self antigens. Further, we wished to establish which specific alleles had the propensity to bind a larger fraction of peptides in general (allele promiscuity) by observing the relationship between an allele's ability to bind random peptides versus peptides from a viral or human proteome.

We examined distribution of predicted allelic promiscuity across alleles for 9 sets of peptides of viral, human, and random origin (See Methods). Confining attention to human and viral proteomes, we again found a wide range in the proportion of peptides a given allele was predicted to bind and also significant inconsistencies between tools (Supplementary Figure 3).

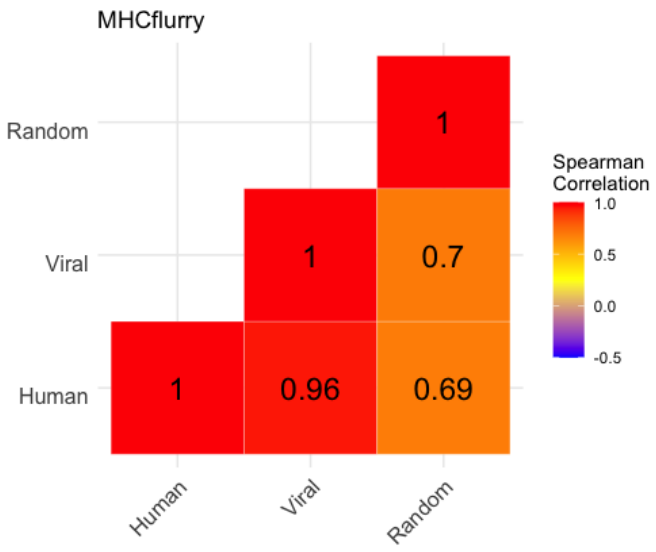
We found that the alleles with highest mean binding percentage for human and viral peptides were B15:03 (2.68%) and B15:02 (2.36%) and the allele lowest mean binding percentage were B18:01 (0.24%) and A01:01 (0.33%) (Supplementary Table 3). No alleles were predicted by any tool to preferentially present either viral or human peptides. Further, the distribution of predicted allelic promiscuity across alleles was highly consistent between human and viral proteomes, but not when applied to a set of random peptides (Figure 4). We noted that this phenomenon holds for closely related viruses across all tools and to a lesser extent for more distantly related viruses (Supplementary Figure 4).



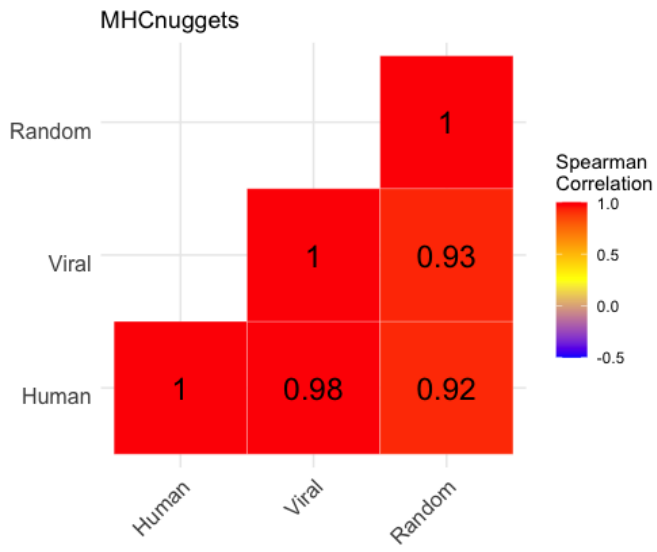

A




B




C




D

*Figure 4. The correlation between peptide sources of predicted allelic promiscuity across alleles. A) Heatmap of spearman correlation between peptide sources for HLAthena-based predictions for human peptides, viral peptides, and randomly generated peptides. Numbers show Spearman correlation coefficients between each pair respectively, while color reflects the Spearman correlation with red approaching a Spearman correlation of 1. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.*

Confining attention to the 9 alleles whose predictive models were likely most robust (based on a minimum of 2000 training peptides for every tool), we again found that the distribution of predicted allelic promiscuity across alleles was consistent between closely related viruses and to a lesser extent between more distantly related viruses (Supplementary Figure 5).

**Peptide physical properties are associated with allele-specific binding predictions**

Reasoning that differences in peptide characteristics were the likeliest explanation for predicted differences in binding affinity between different alleles and peptide sources, we next studied the distribution of physical properties among different peptide sets. Human, viral, and random peptide sets all exhibited the same range of physical properties, but were differentially enriched among different physical properties (Supplementary Figure 6). Between individual peptide sets, the differential enrichment ranged from 10% (CMV v. human) to 63% (BK v. random) of peptides (Supplementary Figure 7).

We next sought to discover the relationship between the peptide similarity in physical property space and distribution of predicted allelic promiscuity across alleles. Across all tools, there was a positive relationship between similarity in physical property space and distribution of predicted allelic promiscuity across alleles as

evidenced by the negative correlation between peptide set difference and Spearman correlation coefficient (Figure 6).

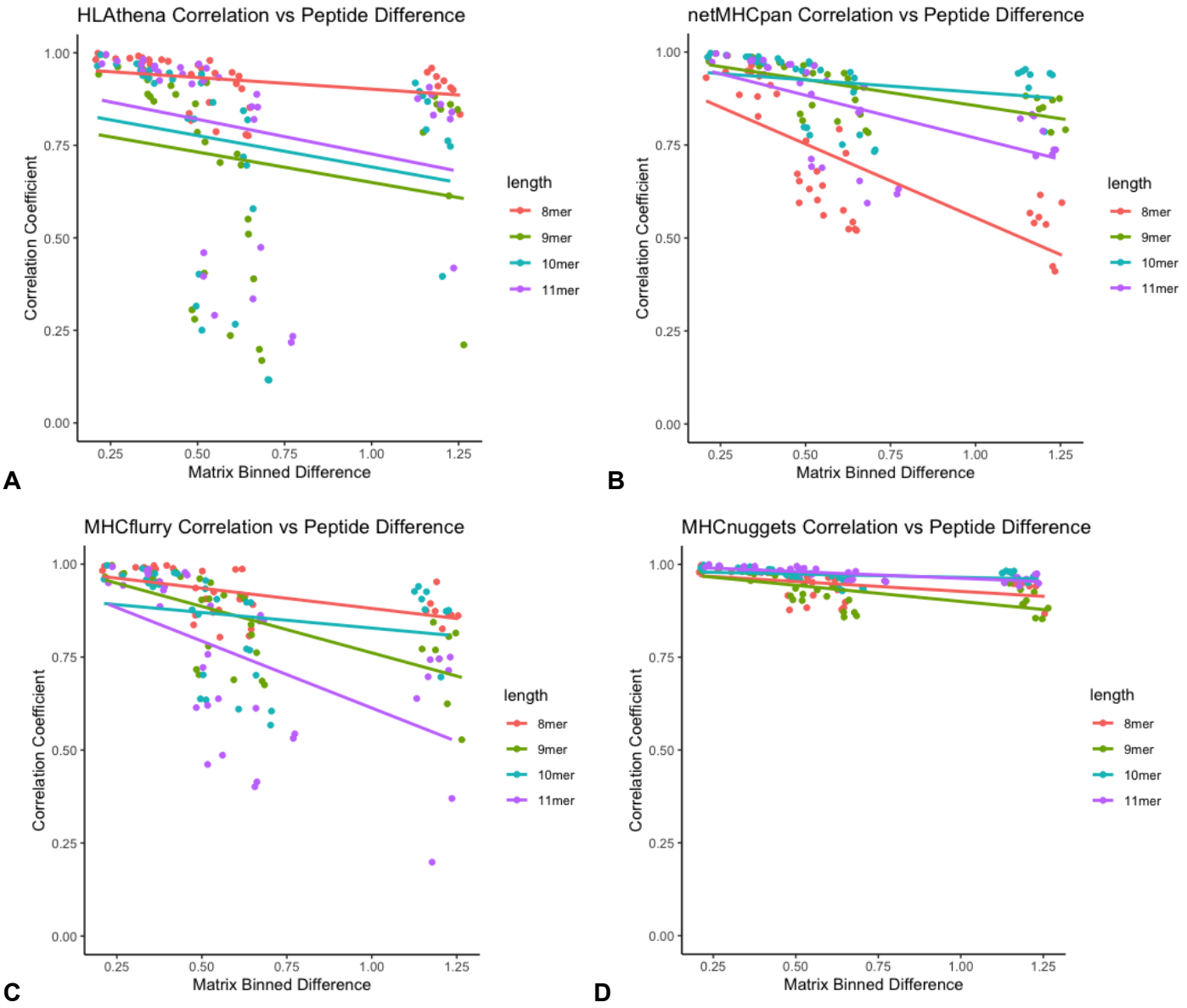



*Figure 6. The relationship between physical property similarity vs peptide binding similarity. A) Scatterplot for HLAthena-based predictions, where each point represents predictions for a species vs species pair. Peptide dissimilarity is shown on the x-axis, whereas Spearman correlation coefficients of predicted allelic promiscuity across alleles. Color represents the length of peptide, with 8-, 9-, 10-, and 11-mers shown in red, green, blue and purple, respectively. Analogous data is shown for netMHCpan, MHCflurr, and MHCnuggets in panels B, C, and D, respectively.*

Next, we found that each allele has distinct preferences for different peptide physical properties, independent of peptide length (Figure 7A, Supplementary Figure 8). Some alleles (e.g. A01:01 and B08:01) have stronger preference for certain physical properties (Figure 7B,C), while others (B45:01) do not have as clear of a preference (Figure 7D).

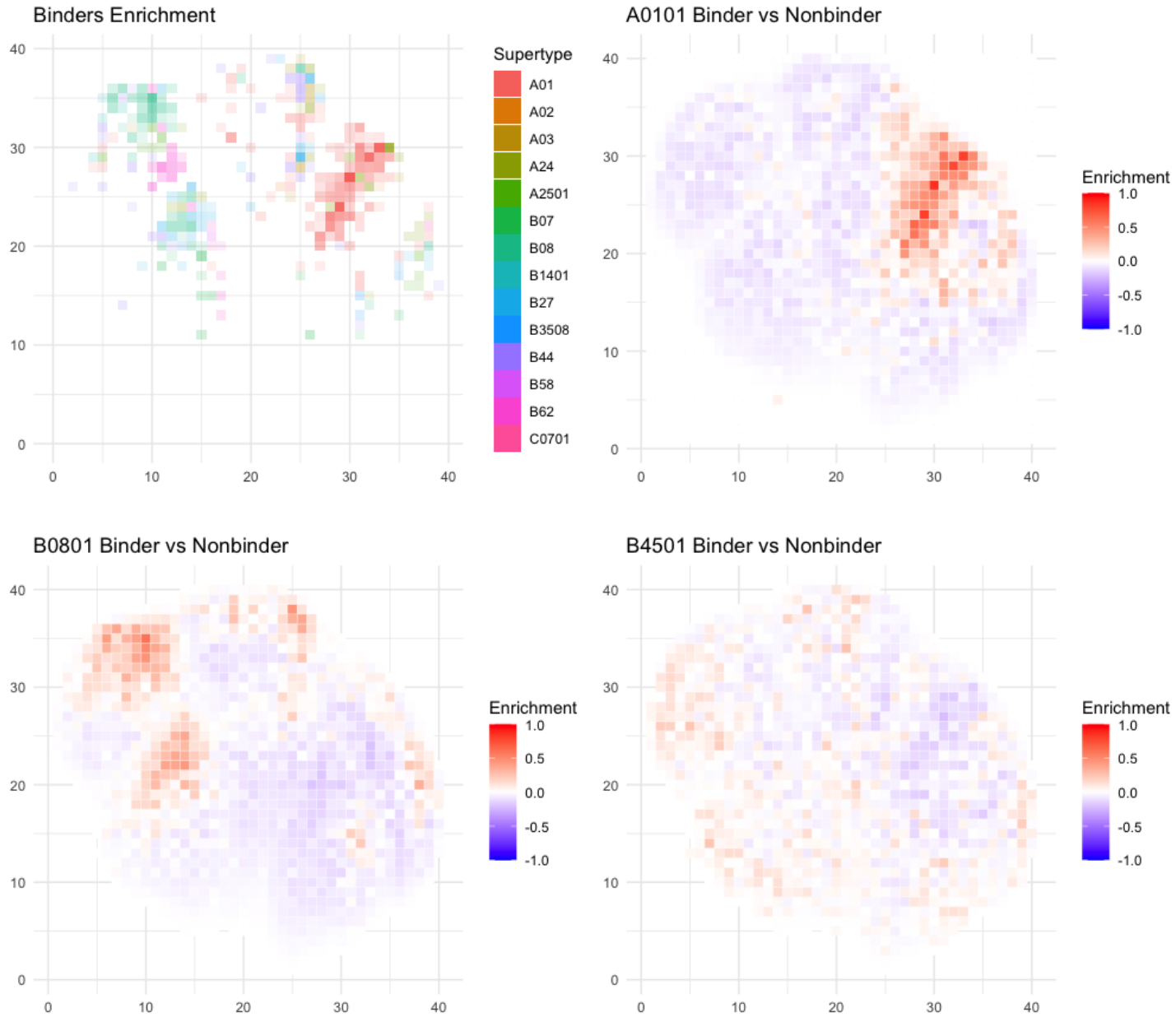



*Figure 7. Differential distributions of physical properties for 9-mer peptides predicted to bind to HLA alleles. A) The plotting coordinates represent the first two dimensions of a UMAP transform of peptide physical properties, which is divided into 1600 (40x40) equivalently-sized square bins (see Methods). For each bin where there is at least one HLA allele with >0.2% difference in proportion of all peptides predicted to bind v. non-binders, the identity of the most enriched allele is shaded in the color corresponding to that allele's supertype as*

*corresponding to the legend. B-D) Example plots of three different alleles (A01:01, B08:01, and B45:01) with different distributions of binders. Each box represents enrichment as the percent peptide difference between predicted binders and non-binders for the given allele. The color scale shows the percent of peptides difference in the given box, with red meaning a larger number of predicted binders and blue meaning a larger number of predicted non-binders.*

## DISCUSSION

To the best of our knowledge, this is the first study to examine the consistency of predictions of peptide-MHC binding across different tools, and to explore the quality and quantity of training data in this context. We note several limitations to this work. Firstly, we confined attention to MHC class I peptides and did not include predictions for MHC class II (36), of which there are numerous alleles. We also excluded from consideration any potential contributions of proteasomal cleavage or other antigen processing machinery to MHC binding (37–39). We did not seek to comprehensively assess all available tools for peptide-MHC binding affinity prediction, but rather confined our attention to four of the most widely used tools. The majority of our randomly generated peptides are not known to be found in nature and may not represent the optimal background distribution for measuring allele promiscuity or interrater reliability between tools primarily used for human and pathogenic peptides. While our analysis of peptides leveraged four essential and well-described amino acid physical properties, there may exist unassessed latent features that could capture additional variance and improve dimensionally-reduced comparisons. We did not assess the extent to which mass spectrometry biases in the training datasets might affect peptide-MHC predictions (40–43). Lastly, we did not evaluate individual tool performance based on known epitopes as this has been previously reported (23–27,44–48).

Our work raises fundamental questions about the fidelity of peptide-MHC binding prediction tools. Why, for instance, can predictions be so discordant among tools for which training datasets are otherwise so similar? We especially worry about the real-world use of these prediction tools for alleles without any direct basis in training data. Why is the predicted range of allele promiscuity so substantial, and yet not demonstrative of any meaningful differences in enrichment between potential foreign versus self antigens? Moreover, is this differential promiscuity a universal biological phenomenon, with certain alleles being generally poor functional presenters of antigen? If this is the case, what selective advantage might have evolutionarily maintained these alleles in the population? Evaluating more viruses – as well as bacteria, fungi, and other pathogens – and linking these analyses with metrics such as evolutionary distance may give greater insight into the relationship between HLA evolution and disease.

## METHODS

### Sequence retrieval, peptide filtering, and kmerization

FASTA-formatted protein sequence data was retrieved from the National Center of Biotechnology Information (NCBI) (49,50) using RefSeq as of 1-31-22 for BK, SARS-CoV-2, HHV-5, HHV-6, HSV-1, HSV-2, HSV-4, and Human. Protein sequence data was inputted into netchop v3.0 “C-term” model with a cleavage threshold of 0.1 to remove peptides that were not predicted to undergo canonical MHC class I antigen processing via proteasomal cleavage (of the peptide’s C-terminus). The results from netchop v3.0 were then kmerized sequentially into 8- to 12-mers. Code used for kmerization and netchop filtering can be found at: https://github.com/Boeinco/peptide-MHCassess. We additionally generated a set of 1 million random peptides of length 8-12 drawn uniformly at random. Peptide sets had negligible overlap (<1% shared between human vs viral vs random peptides).

### Peptide-MHC class I binding affinity predictions

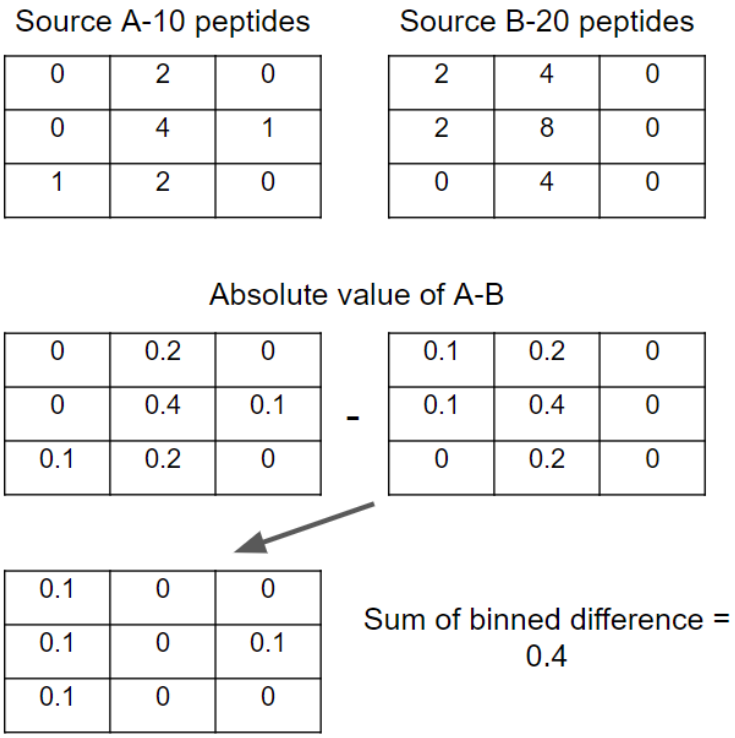
MHC class I binding affinity predictions were performed for the peptides generated from the kmerization process above using 4 tools: netMHCpan v4.1 (23), HLAthena v1.0 (27), MHCflurry v2.0 (25), and MHCnuggets v2.3 (51). netMHCpan was run with default options with the ‘-l’ option to specify peptides of lengths 8-12. MHCflurry was run with default options. MHCnuggets was run with default options. HLAthena was run using the dockerized version of HLAthena with default options, which predicts peptides of length 8-11. MHC class I binding affinity predictions were performed for each of 24, 26, and 2, HLA-A, -B, and -C alleles, respectively. Only alleles that were in common between all 4 tools were used (52 total alleles in common between 2489 possible alleles). Binding affinity values were converted to binding probability values for MHCflurry and MHCnuggets using 1- log(binding affinity) / log(50000) in order to match HLAthena and netMHCpan binding probability predictions. Alleles were grouped into supertypes when applicable using the HLA class I revised classification (29).

### Dimensional reduction and binning analysis

Peptides were converted into physical property matrices using amino acid sequence mapping into a 4*kmer length matrix containing each amino acid’s properties in sequence. The following physical properties of the amino acids were encoded: side chain polarity was recorded as its isoelectric point (pI) (52), the molecular volume of each side chain was recorded as its partial molar volume at 37°C (53), the hydrophobicity of each side chain was characterized by its simulated contact angle with nanodroplets of water (54) and conformational entropy was derived from peptide bond angular observations among protein sequences without observed secondary structure (55).

Each dimensional reduction was performed on the pooled set of k-mers. UMAP dimensionality was performed using uwot UMAP R implementation v0.1.11. PCA was performed using default `prcomp()` functions in base R v4.1.3.

For each peptide source, binned matrices were computed using the `bin2()` function with 40x40 (1600) bins from the `Ash` v1.0.15 package (56) in R v4.1.3. Bin values were then divided by the total number of peptides to create bins with the % of total peptides. In order to compare between 2 peptide sources, a matrix, called the difference matrix, is created by subtracting one matrix of a peptide source from another. Taking the absolute value of each bin in the difference matrix, then summing the values together, results in a single metric ranging from 0-2 measuring the difference in binned density between 2 peptide sources, the value 2 indicating that no peptides were shared between bins and the value 0 indicating the same percentage of peptides in every bin (Methods Figure 1).




Methods Figure 1.

Allele ordering similarity

For each allele-peptide source combination, the percentage of peptides predicted to bind with a binding probability score of 0.5 or greater was calculated for all processed peptides. 0.5 binding score is estimated to be equivalent to 250-300nM depending on the tool used. For each peptide source, alleles were ranked from best to worst binders (most to least peptides >= 0.5 score) t. In order to compute allele ordering similarity

between 2 peptide sources for a single tool, Spearman's Rank Correlation Coefficient was calculated between the 2 sets of allele ranks.

For the random group 1 vs random group 2 analysis, we conducted 100 replicates of dividing the randomly generated peptides into 2 random groups and performed a Spearman rank test of allele ordering between these groups for each of the tools.

Interrater reliability

Intraclass correlation coefficients (ICCs) were calculated using the `ICC()` function from the `IRR` v0.84.1 R package (57). Binding prediction scores for all 1 million randomly generated peptides were separated by tool and HLA allele, and an ICC was calculated as the interrater reliability metric between the 4 tools for each allele. ICC was also between the 4 tools on a per peptide basis, each peptide receiving a score across 4 tools using predictions separated by tool and peptide.

## REFERENCES

1. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018 May;18(5):325–39.

2. Blackwell JM, Jamieson SE, Burgner D. HLA and Infectious Diseases. Clin Microbiol Rev. 2009 Apr;22(2):370–85.

3. Meyer D, C. Aguiar VR, Bitarello BD, C. Brandt DY, Nunes K. A genomic perspective on HLA evolution. Immunogenetics. 2018;70(1):5–27.

4. Zernich D, Purcell AW, Macdonald WA, Kjer-Nielsen L, Ely LK, Laham N, et al. Natural HLA Class I Polymorphism Controls the Pathway of Antigen Presentation and Susceptibility to Viral Evasion. J Exp Med. 2004 Jun 28;200(1):13–24.

5. Bihl F, Frahm N, Giammarino LD, Sidney J, John M, Yusim K, et al. Impact of HLA-B Alleles, Epitope Binding Affinity, Functional Avidity, and Viral Coinfection on the Immunodominance of Virus-Specific CTL Responses. J Immunol. 2006 Apr 1;176(7):4094–101.

6. Berger CT, Carlson JM, Brumme CJ, Hartman KL, Brumme ZL, Henry LM, et al. Viral adaptation to immune selection pressure by HLA class I–restricted CTL responses targeting epitopes in HIV frameshift sequences. J Exp Med. 2010 Jan 18;207(1):61–75.

7. Schellens IM, Meiring HD, Hoof I, Spijkers SN, Poelen MCM, van Gaans-van den Brink JAM, et al. Measles Virus Epitope Presentation by HLA: Novel Insights into Epitope Selection, Dominance, and Microvariation. Front Immunol [Internet]. 2015 [cited 2019 Nov 15];6. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2015.00546/full

8. Kaufman J. Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. Trends Immunol. 2018 May 1;39(5):367–79.

9. Barbosa CRR, Barton J, Shepherd AJ, Mishto M. Mechanistic diversity in MHC class I antigen recognition. Biochem J. 2021 Dec 23;478(24):4187–202.

10. Slade JWG, Watson MJ, MacDougall-Shackleton EA. “Balancing” balancing selection? Assortative mating at the major histocompatibility complex despite molecular signatures of balancing selection. Ecol Evol. 2019 Apr 13;9(9):5146–57.

11. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. J Virol [Internet]. 2020 Apr 17 [cited 2022 Jul 19]; Available from: https://journals.asm.org/doi/10.1128/JVI.00510-20

12. Mishto M, Mansurkhodzhaev A, Rodriguez-Calvo T, Liepe J. Potential Mimicry of Viral and Pancreatic β Cell Antigens Through Non-Spliced and cis-Spliced Zwitter Epitope Candidates in Type 1 Diabetes. Front Immunol [Internet]. 2021 [cited 2022 Sep 29];12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.656451

13. Geneugelijk K, Thus KA, Spierings E. Predicting Alloreactivity in Transplantation. J Immunol Res. 2014 Apr 28;2014:e159479.

14. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat Rev Immunol. 2018 Mar;18(3):168–82.

15. Blass E, Ott PA. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. Nat Rev Clin Oncol. 2021 Apr;18(4):215–29.

16. Nelde A, Maringer Y, Bilich T, Salih HR, Roerden M, Heitmann JS, et al. Immunopeptidomics-Guided Warehouse Design for Peptide-Based Immunotherapy in Chronic Lymphocytic Leukemia. Front Immunol [Internet]. 2021 [cited 2022 Sep 30];12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.705974

17. Terasaki M, Shibui S, Narita Y, Fujimaki T, Aoki T, Kajiwara K, et al. Phase I trial of a personalized peptide vaccine for patients positive for human leukocyte antigen--A24 with recurrent or progressive glioblastoma multiforme. J Clin Oncol Off J Am Soc Clin Oncol. 2011 Jan 20;29(3):337–44.

18. Kibe S, Yutani S, Motoyama S, Nomura T, Tanaka N, Kawahara A, et al. Phase II study of personalized peptide vaccination for previously treated advanced colorectal cancer. Cancer Immunol Res. 2014 Dec;2(12):1154–62.

19. Bjerregaard AM, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunol Immunother CII. 2017 Sep;66(9):1123–30.

20. Wood MA, Nguyen A, Struck AJ, Ellrott K, Nellore A, Thompson RF. neoepiscope improves neoepitope prediction with multivariant phasing. Bioinformatics. 2020 Feb 1;36(3):713–20.

21. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med. 2016 Jan 29;8(1):11.

22. Bais P, Namburi S, Gatti DM, Zhang X, Chuang JH. CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. Bioinforma Oxf Engl. 2017 Oct 1;33(19):3110–2.

23. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020 Jul 2;48(W1):W449–54.

24. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017 Nov 1;199(9):3360–8.

25. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. Cell Syst. 2020 Jul;11(1):42-48.e7.

26. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. Cancer Immunol Res. 2020;8:396–408.

27. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat Biotechnol. 2020 Feb;38(2):199–209.

28. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J Immunol Baltim Md 1950. 2013 Dec 15;191(12):5831–9.

29. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. BMC Immunol. 2008 Jan 22;9(1):1.

30. Pavlos R, McKinnon EJ, Ostrov DA, Peters B, Buus S, Koelle D, et al. Shared peptide binding of HLA Class I and II alleles associate with cutaneous nevirapine hypersensitivity and identify novel risk alleles. Sci Rep. 2017 Aug 17;7(1):8653.

31. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity. 2017 Feb;46(2):315–26.

32. Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc R Soc B Biol Sci. 2010 Apr 7;277(1684):979–88.

33. Manczinger M, Boross G, Kemény L, Müller V, Lenz TL, Papp B, et al. Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. PLoS Biol. 2019 Jan 31;17(1):e3000131.

34. White CF, Pellis L, Keeling MJ, Penman BS. Detecting HLA-infectious disease associations for multi-strain pathogens. Infect Genet Evol. 2020 Sep 1;83:104344.

35. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Curr Biol. 2005 Jun 7;15(11):1022–7.

36. Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen processing and presentation. Nat Rev Immunol. 2015 Apr;15(4):203–16.

37. pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification | Bioinformatics | Oxford Academic [Internet]. [cited 2022 Oct 2]. Available from: https://academic.oup.com/bioinformatics/article/37/21/3723/6363787

38. Ritz U, Seliger B. The Transporter Associated With Antigen Processing (TAP): Structural Integrity, Expression, Function, and Its Clinical Relevance. Mol Med. 2001 Mar;7(3):149–58.

39. López de Castro JA. How ERAP1 and ERAP2 Shape the Peptidomes of Disease-Associated MHC-I Proteins. Front Immunol [Internet]. 2018 [cited 2022 Oct 2];9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02463

40. Dincer AB, Lu Y, Schweppe DK, Oh S, Noble WS. Reducing Peptide Sequence Bias in Quantitative Mass Spectrometry Data with Machine Learning. J Proteome Res. 2022 Jul 1;21(7):1771–82.

41. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database

compression. Mol Syst Biol. 2007 Jan;3(1):102.

42. Prakash A, Piening B, Whiteaker J, Zhang H, Shaffer SA, Martin D, et al. Assessing Bias in Experiment Design for Large Scale Mass Spectrometry-based Quantitative Proteomics*. Mol Cell Proteomics. 2007 Oct 1;6(10):1741–8.

43. Timp W, Timp G. Beyond mass spectrometry, the next step in proteomics. Sci Adv. 2020 Jan 10;6(2):eaax8978.

44. Paul S, Croft NP, Purcell AW, Tscharke DC, Sette A, Nielsen M, et al. Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. PLOS Comput Biol. 2020 May 26;16(5):e1007757.

45. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. Bioinformatics. 2015 Jul 1;31(13):2174–81.

46. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLOS Comput Biol. 2018 Nov 8;14(11):e1006457.

47. Venkatesh G, Grover A, Srinivasaraghavan G, Rao S. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. Bioinformatics. 2020 Jul;36(Suppl 1):i399–406.

48. Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, et al. Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins. bioRxiv. 2017 Jul 27;154757.

49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745.

50. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015 Jan;43(Database issue):D571-577.

51. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets. Cancer Immunol Res. 2019 Dec 23;canimm.0464.2019.

52. Lide D. CRC handbook of chemistry and physics, 1992-1993 : a ready-reference book of chemical and physical data [Internet]. 1992 [cited 2022 Sep 4]. Available from: https://www.worldcat.org/title/crc-handbook-of-chemistry-and-physics-1992-1993-a-ready-reference-book-of-chemical-and-physical-data/oclc/758080758

53. A new set of peptide-based group heat capacities for use in protein stability calculations - ScienceDirect [Internet]. [cited 2022 Sep 4]. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0022283699929522

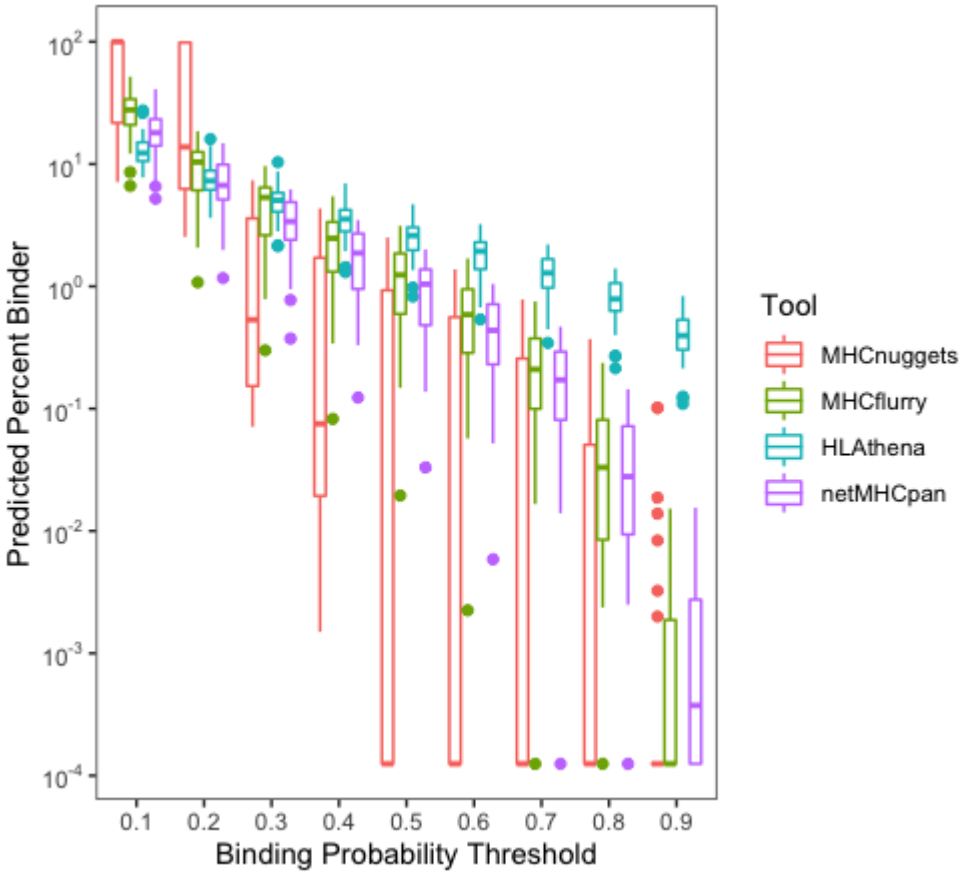
54. Zhu C, Gao Y, Li H, Meng S, Li L, Francisco JS, et al. Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network. Proc Natl Acad Sci U S A. 2016 Nov 15;113(46):12946–51.

55. Fogolari F, Corazza A, Fortuna S, Soler MA, VanSchouwen B, Brancolini G, et al. Distance-Based Configurational Entropy of Proteins from Molecular Dynamics Simulations. PloS One. 2015;10(7):e0132356.

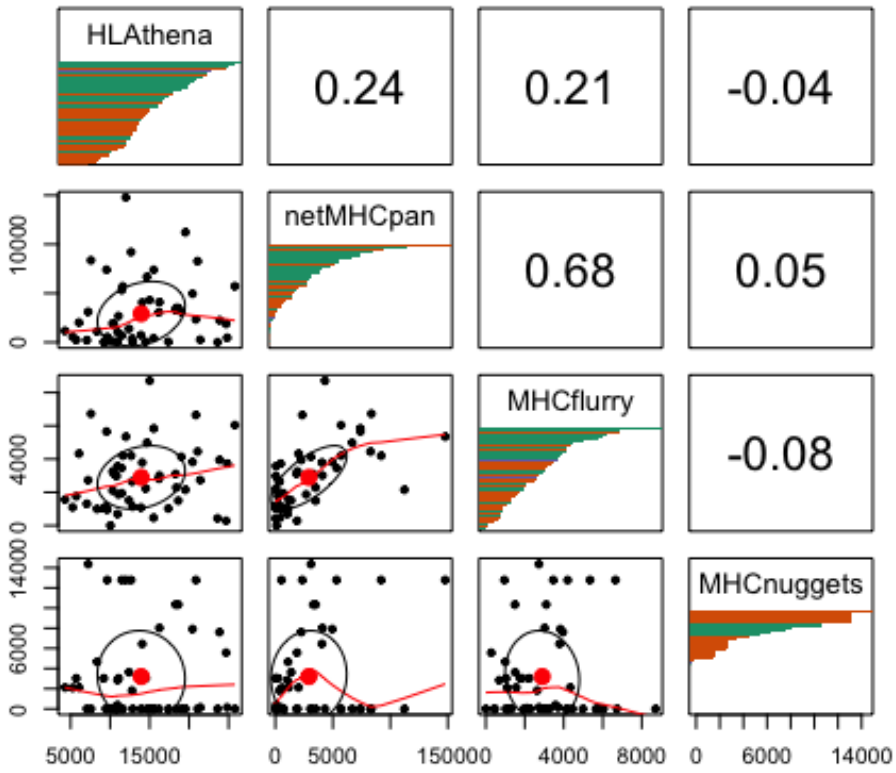
56. Kaluzny S original by DWSR port by AG adopted to recent SP by S. ash: David Scott's ASH Routines [Internet]. 2015 [cited 2022 Jul 11]. Available from: https://CRAN.R-project.org/package=ash

57. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.

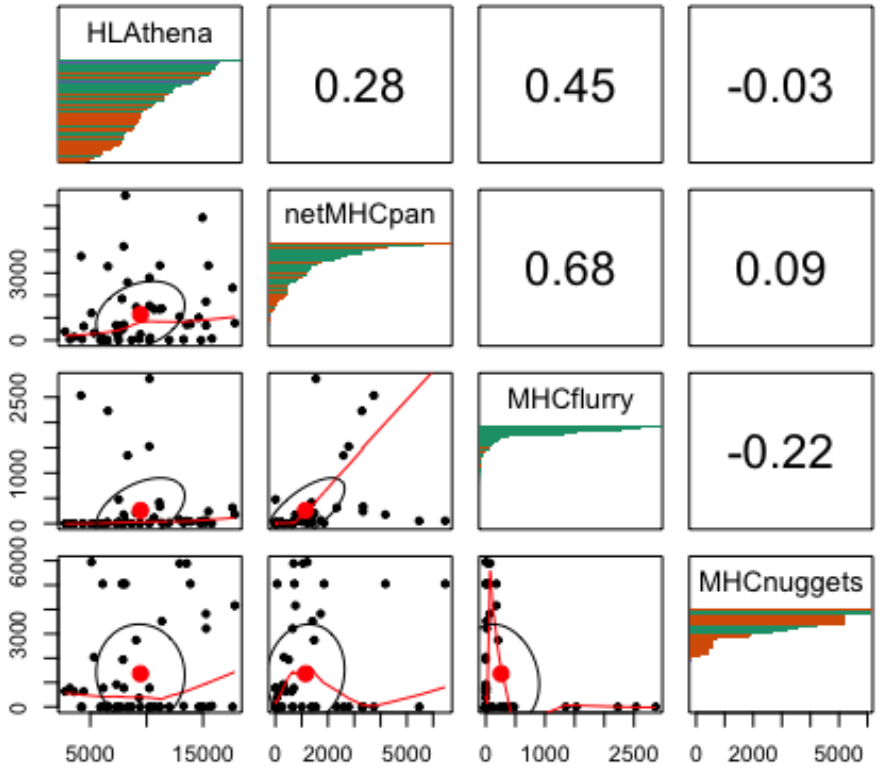
**SUPPLEMENTARY MATERIAL**




*Sup Figure 1: Boxplots of the relationship between predicted binding and the threshold used to determine binding for random peptides. Each color represents a different tool with each boxplot representing the IQR of predicted percent peptides to bind for the given threshold.*

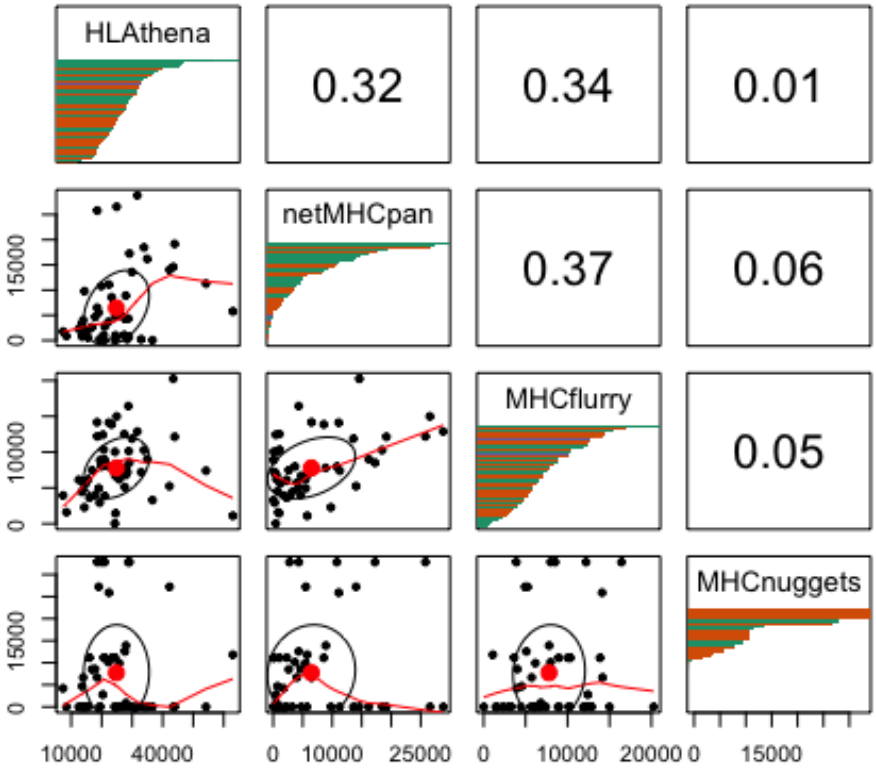



**A**

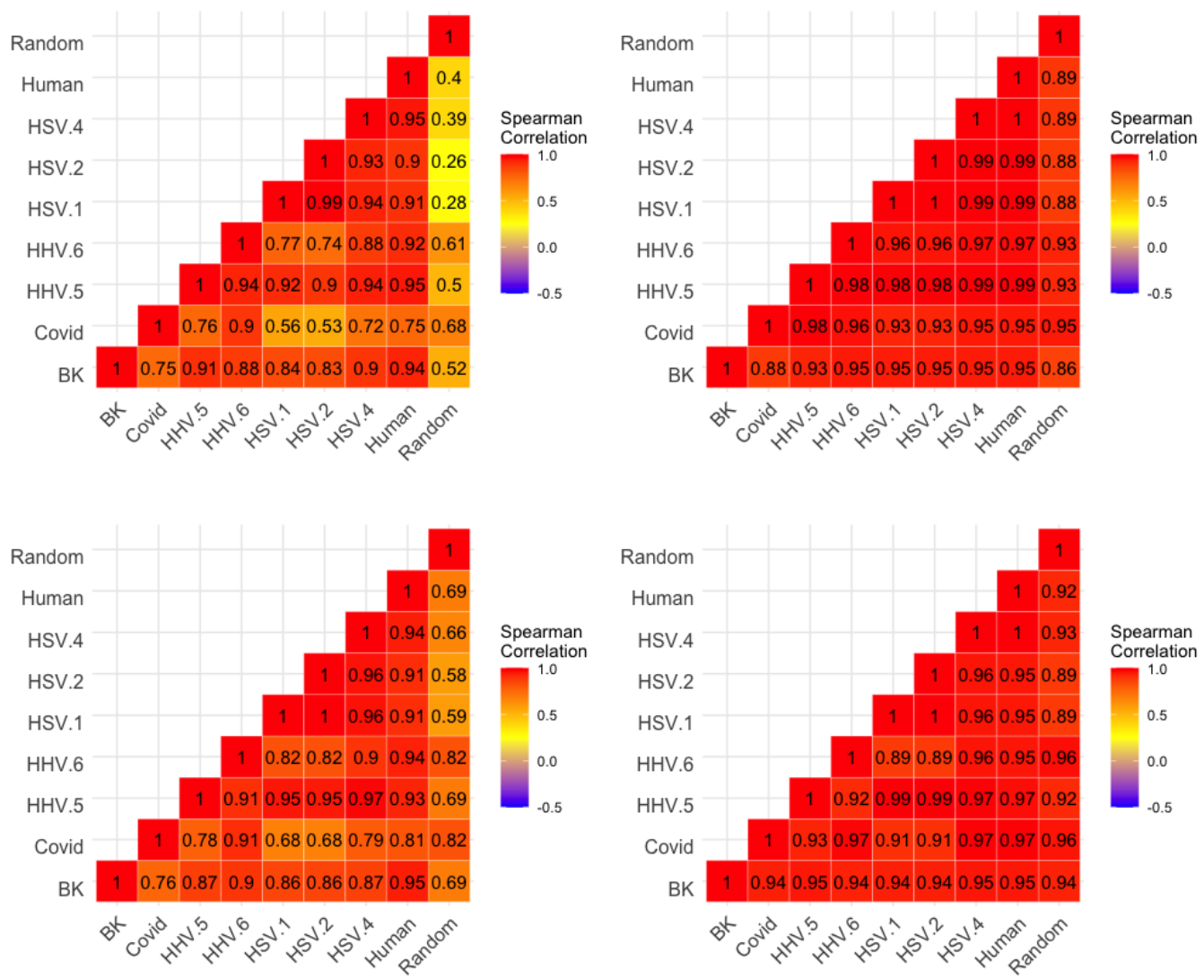



**B**

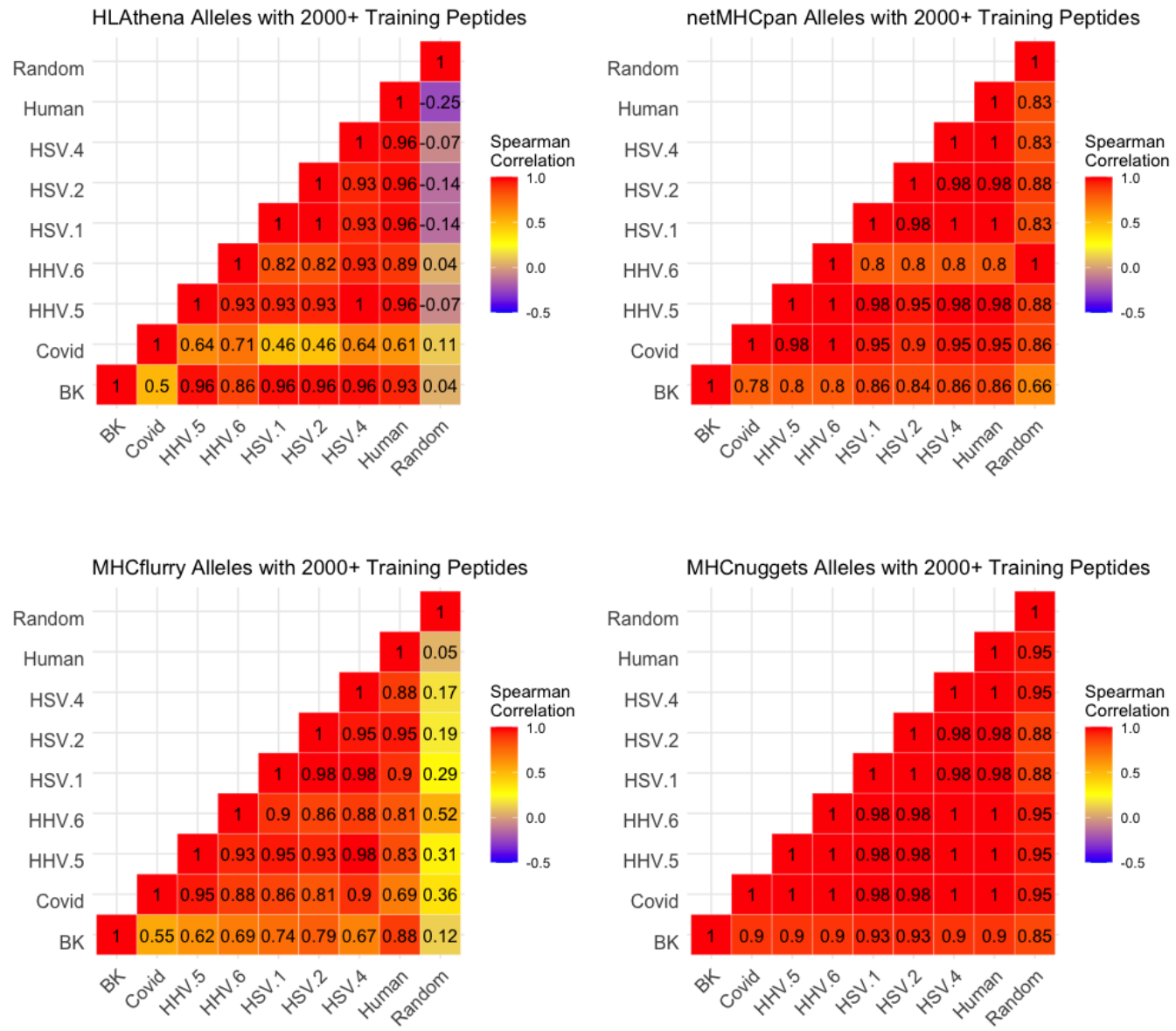
*Sup Figure 2. Pairplot of HLA allelic presentation of 8-11mers from the random proteome. The lower left triangle displays scatter plots of peptides predicted to bind using 0.6 (A) and 0.7 (B) as cutoffs respectively between 2 tools with each point representing an HLA allele. The upper right triangle represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.*




*Sup Figure 3: Pairplot of HLA allelic presentation of 8-11mers from the human and viral proteome. The lower left triangle displays scatter plots of peptides predicted to bind (>= 0.5 binding probability score) between 2 tools with each point representing an HLA allele. The upper right triangle represents the Spearman correlation of the number of peptides predicted to bind to all alleles between tools. Note that MHCnuggets has a number of alleles with 0 random peptides predicted to bind. The diagonal panels show distribution of HLA allelic presentation from the random proteome for each tool. The number of peptides that putatively bind to each of the HLA alleles is shown along the x-axis as a series of horizontal bars with green, orange, and purple colors representing HLA-A, -B, and -C alleles, respectively, sorted in order of decreasing quantity of binders.*

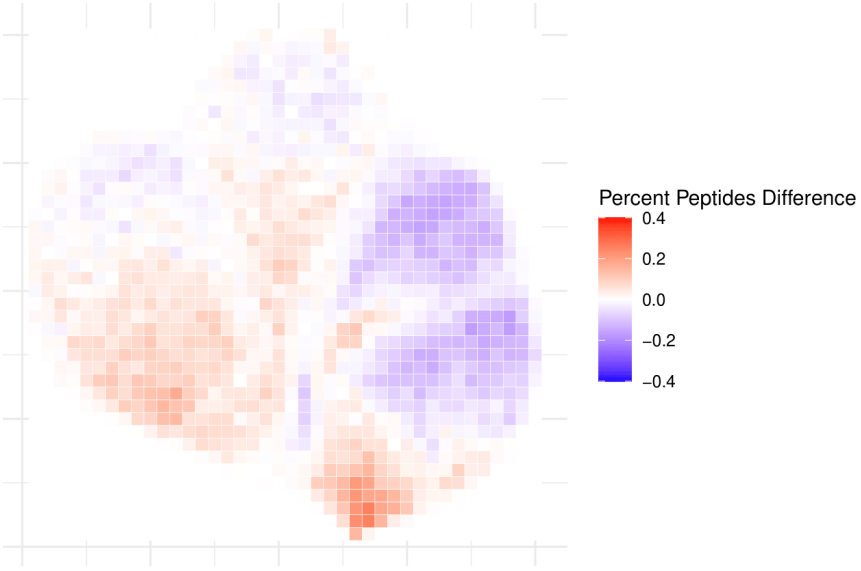



*Sup Figure 4. Heatmaps of correlation between peptides for each species of predicted allelic promiscuity across alleles. A) Spearman correlation is shown between peptide sources for HLAthena-based predictions. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.*

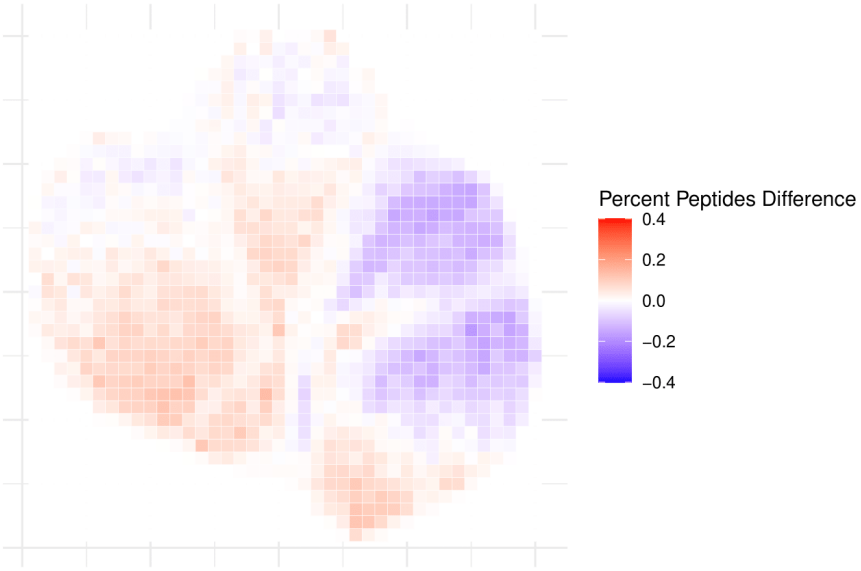



*Sup Figure 5. Heatmaps of correlation between peptides for each species of predicted allelic promiscuity across alleles for which there was a minimum of 2000 peptides of training data. A) Spearman correlation is shown between peptide sources for HLAthena-based predictions. Analogous data is shown for netMHCpan, MHCflurry, and MHCnuggets in panels B, C, and D, respectively.*

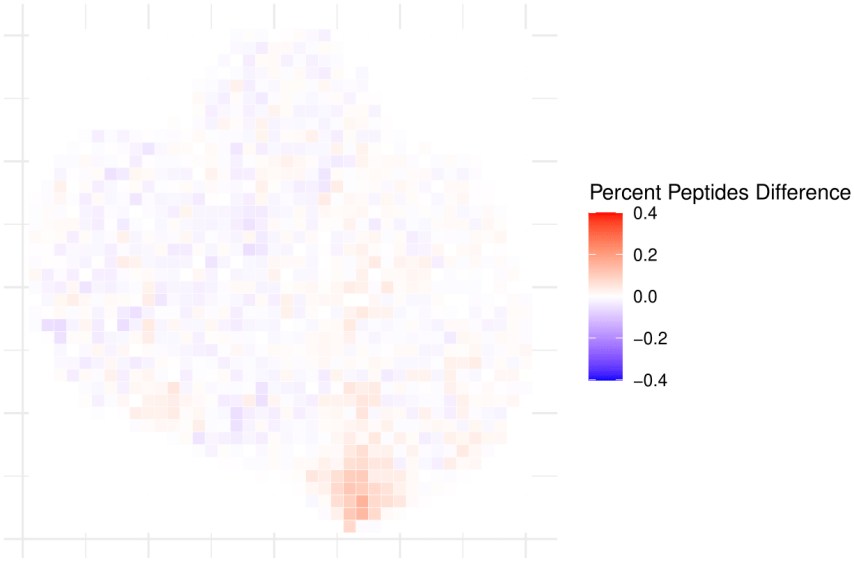
Viral vs Random 8mers



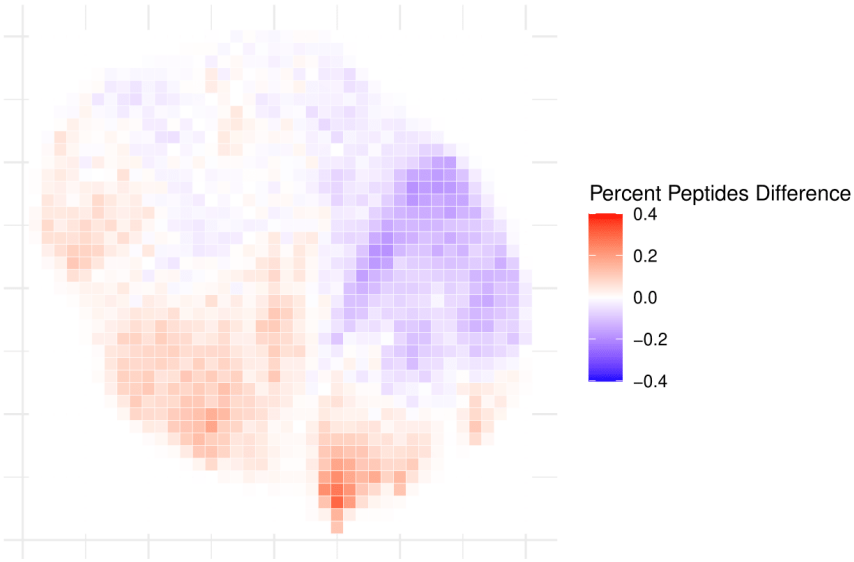

Human vs Random 8mers



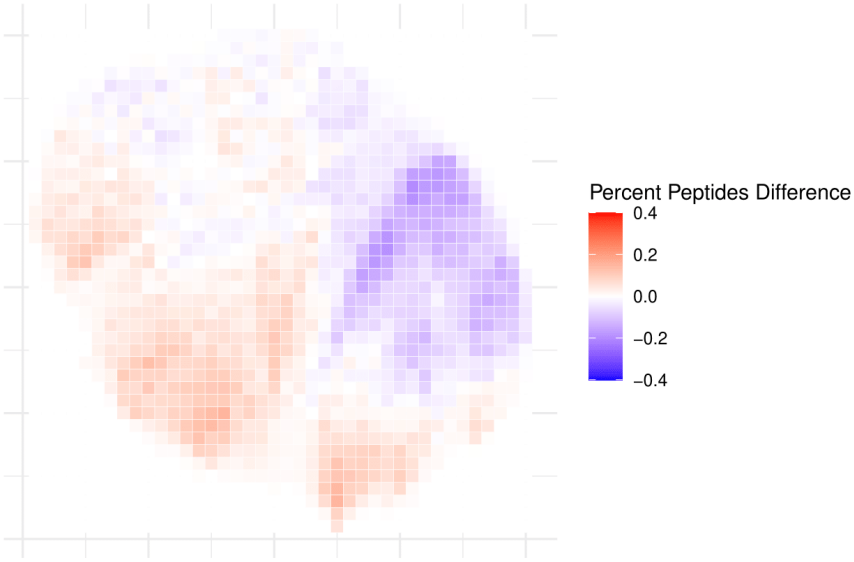
Viral vs Human 8mers



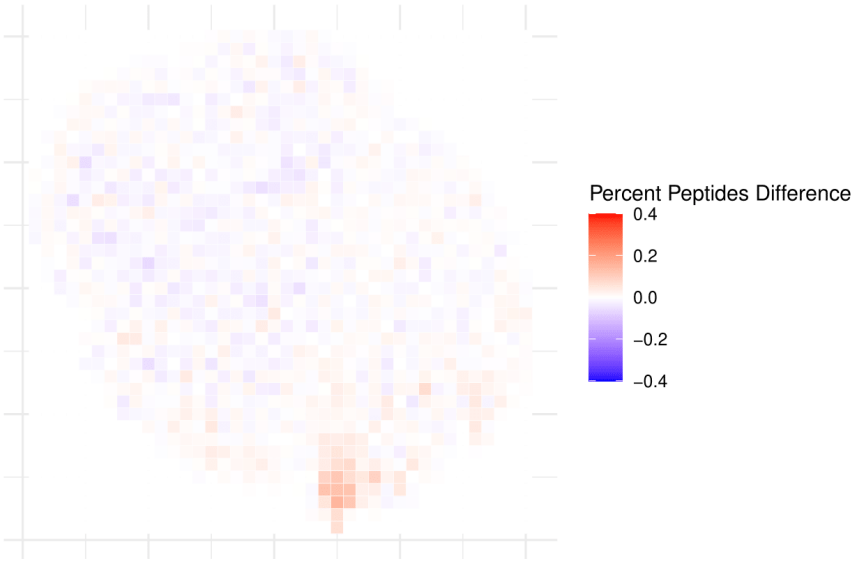

Viral vs Random 9mers



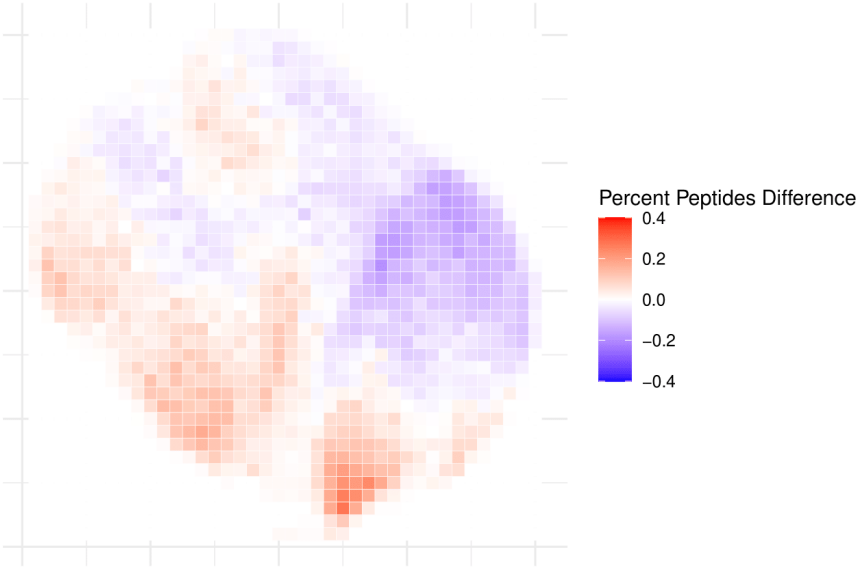
Human vs Random 9mers



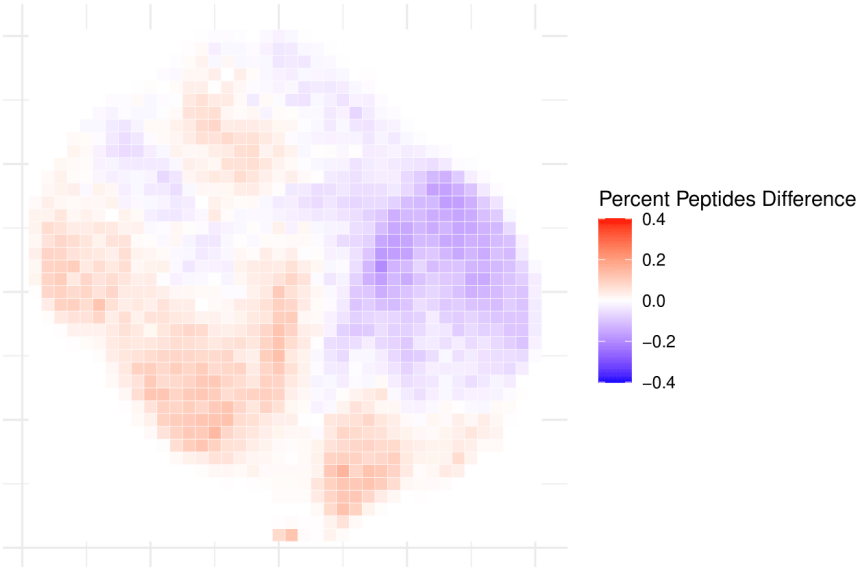

Viral vs Human 9mers



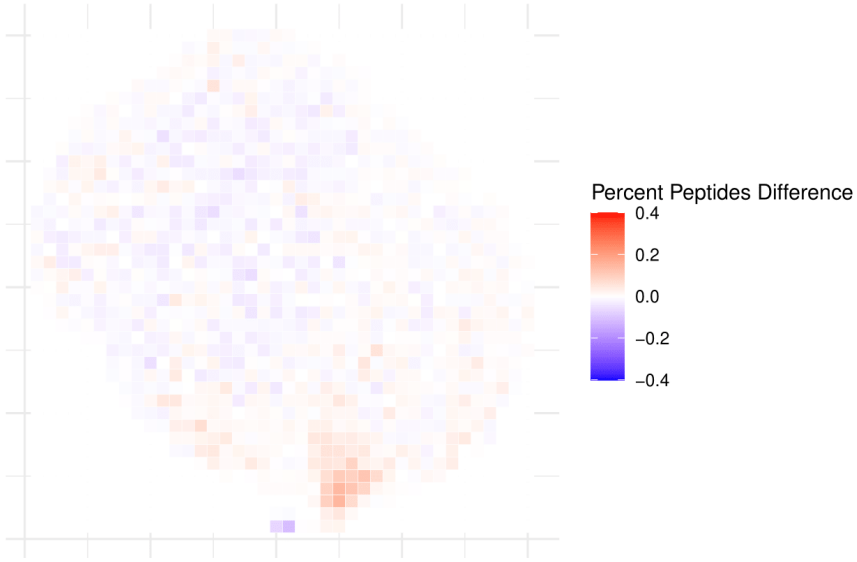
Viral vs Random 10mers



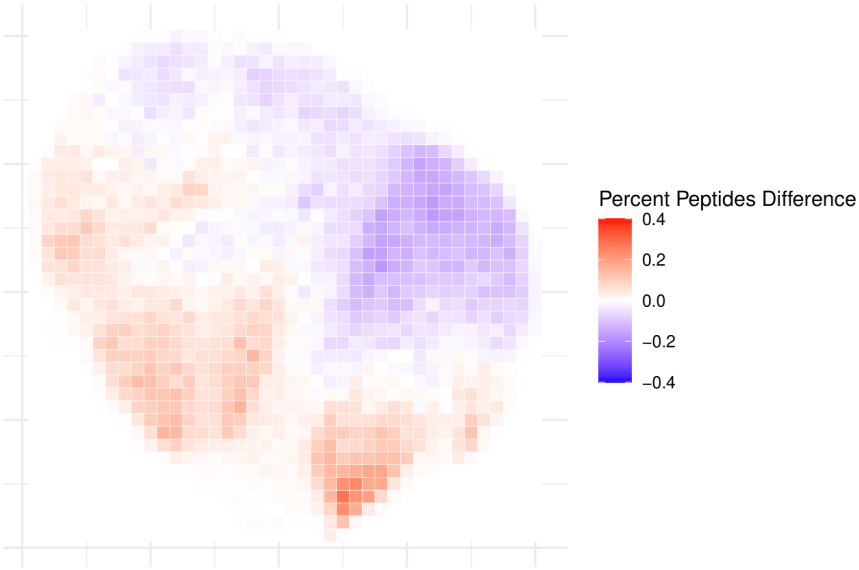

Human vs Random 10mers

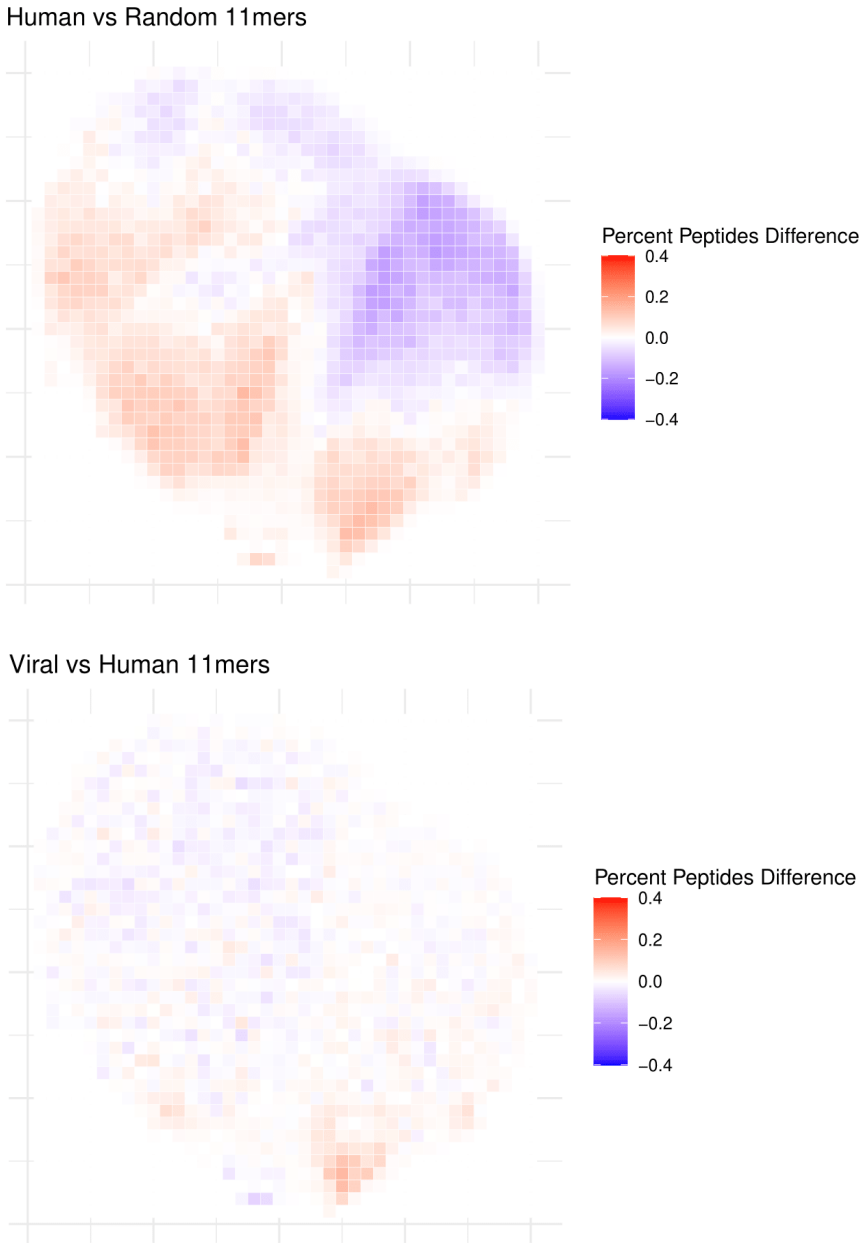


Viral vs Human 10mers

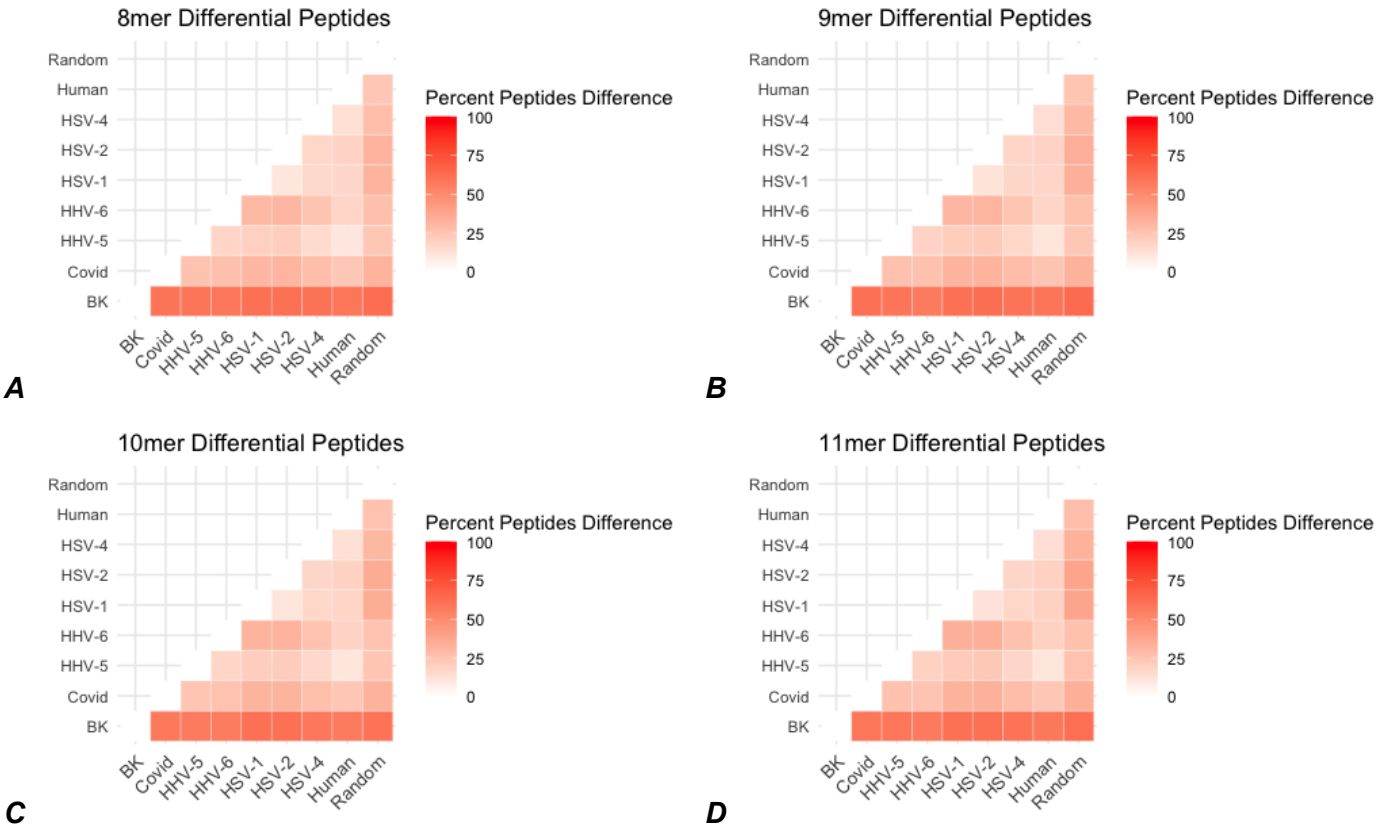



Viral vs Random 11mers

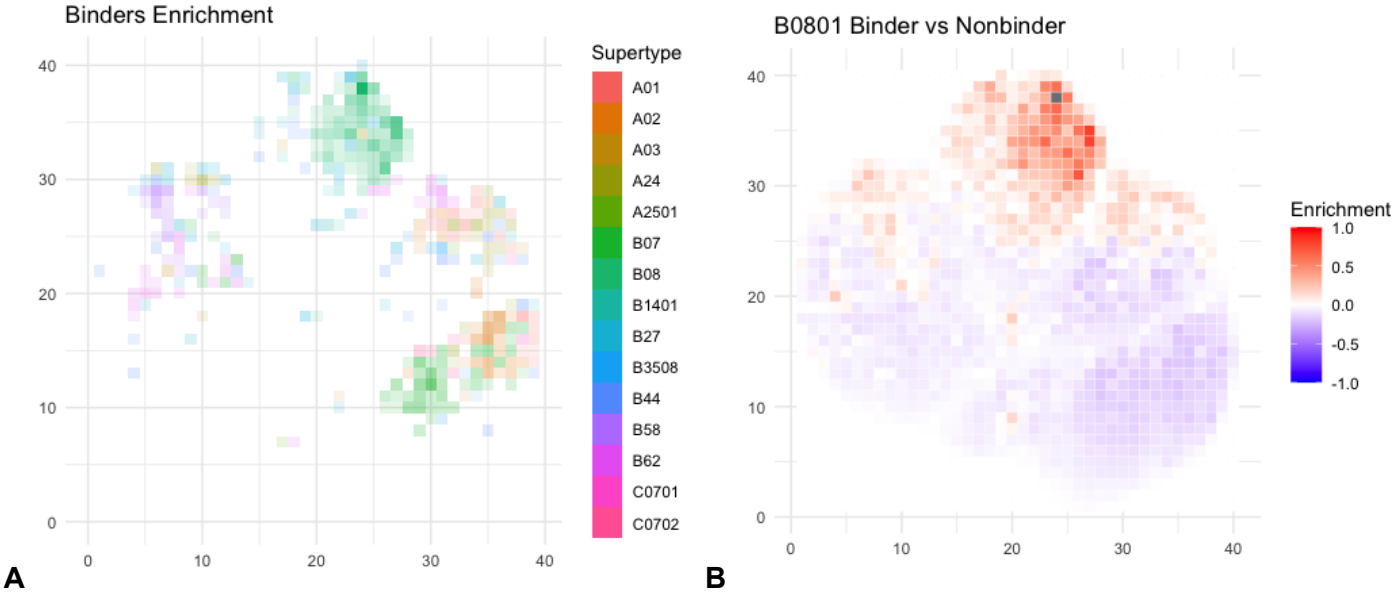


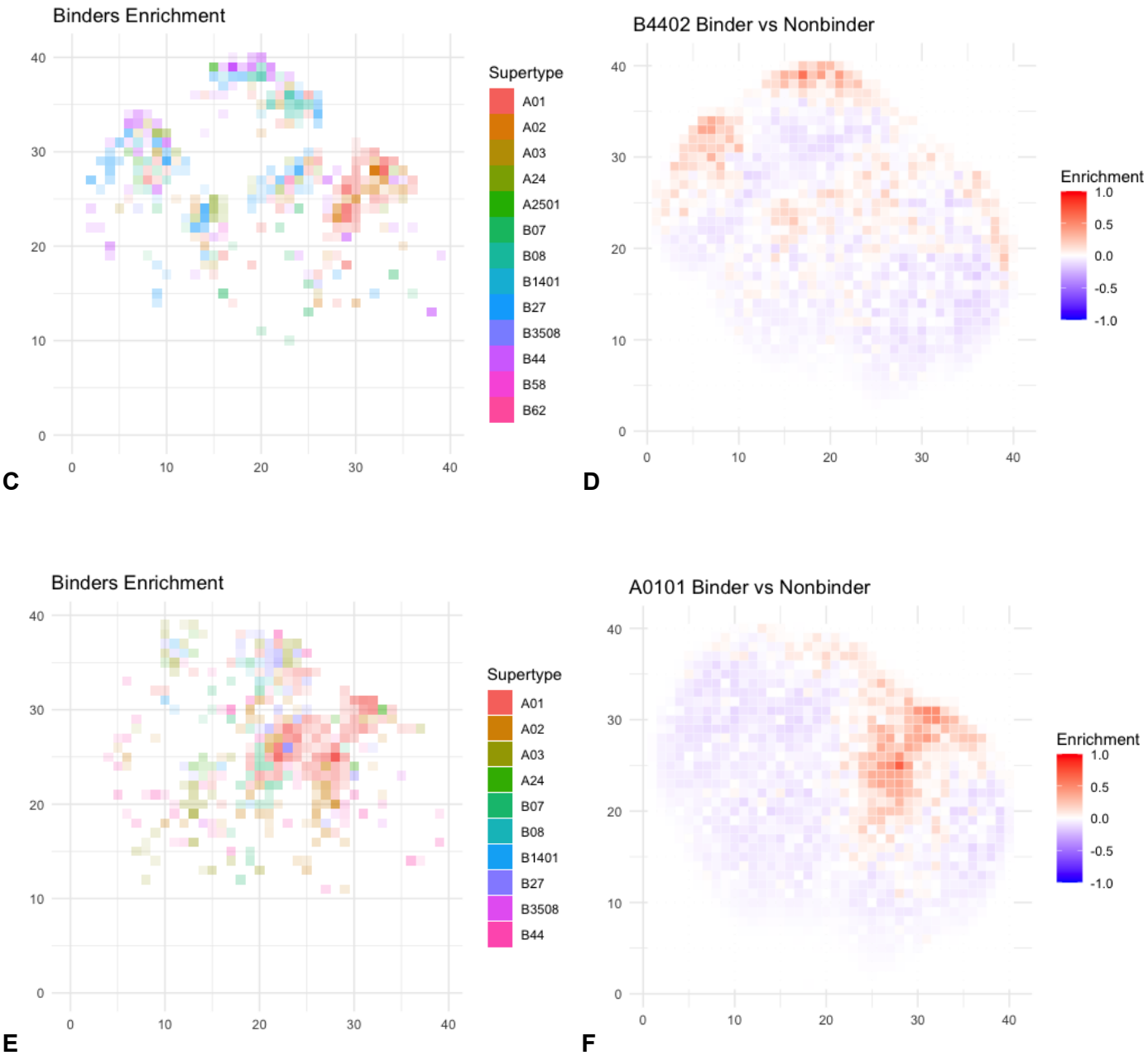



*Sup Figure 6. Peptide physical property differences between different peptide sources. Each tile plot is composed of 1600 tiles, with each tile colored by the percent peptide difference between the 2 peptide sources in that particular tile. Red indicates an enrichment of the first label (e.g. viral vs human, viral enrichment will be red) while blue indicates enrichment of the second label.*




*Sup Figure 7. Peptide physical property difference by k-mer length. Each heatmap is the pairwise percent difference metric between each pair of peptide sets. The redder the value, the more difference in the percent difference metric.*






*Sup Figure 8. Differential distributions of physical properties for 8,10, and 11-mer peptides predicted to bind to HLA alleles. A,C,E) Tile plots highlighting binders enrichment 8, 10, and 11-mers respectively. The plotting coordinates represent the first two dimensions of a UMAP transform of peptide physical properties, which is divided into 1600 (40x40) equivalently-sized square bins (see Methods). For each bin where there is at least one HLA allele with >0.2% difference in proportion of all peptides predicted to bind v. non-binders, the identity of the most enriched allele is shaded in the color corresponding to that allele's supertype as corresponding to the legend. B,D,F) Example plots of alleles with different distributions of binders for 8, 10, and 11-mers respectively. Each box represents enrichment as the percent peptide difference between predicted binders and non-binders for the given allele. The color scale shows the percent of peptides difference in the given box, with red meaning a larger number of predicted binders and blue meaning a larger number of predicted non-binders.*