Journal Pre-proof

HLAEquity: Examining biases in pan-allele peptide-HLA binding predictors

Anja Conev, Romanos Fasoulis, Sarah Hall-Swan, Rodrigo Ferreira, Lydia E. Kavraki

PII: S2589-0042(23)02690-1

DOI: https://doi.org/10.1016/j.isci.2023.108613

Reference: ISCI 108613

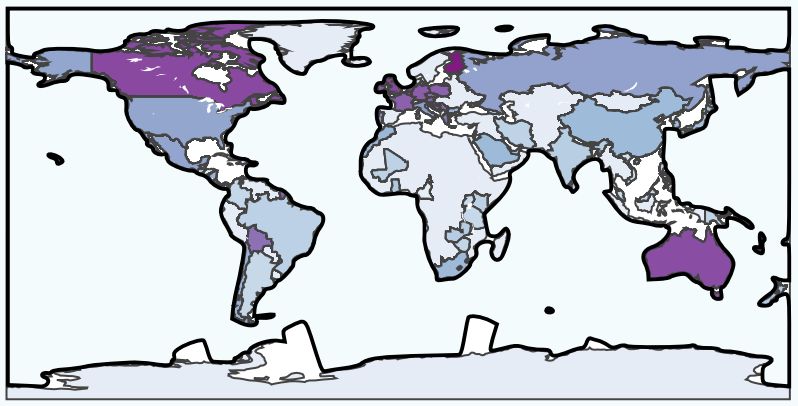
To appear in: *ISCIENCE*

Received Date: 8 November 2023

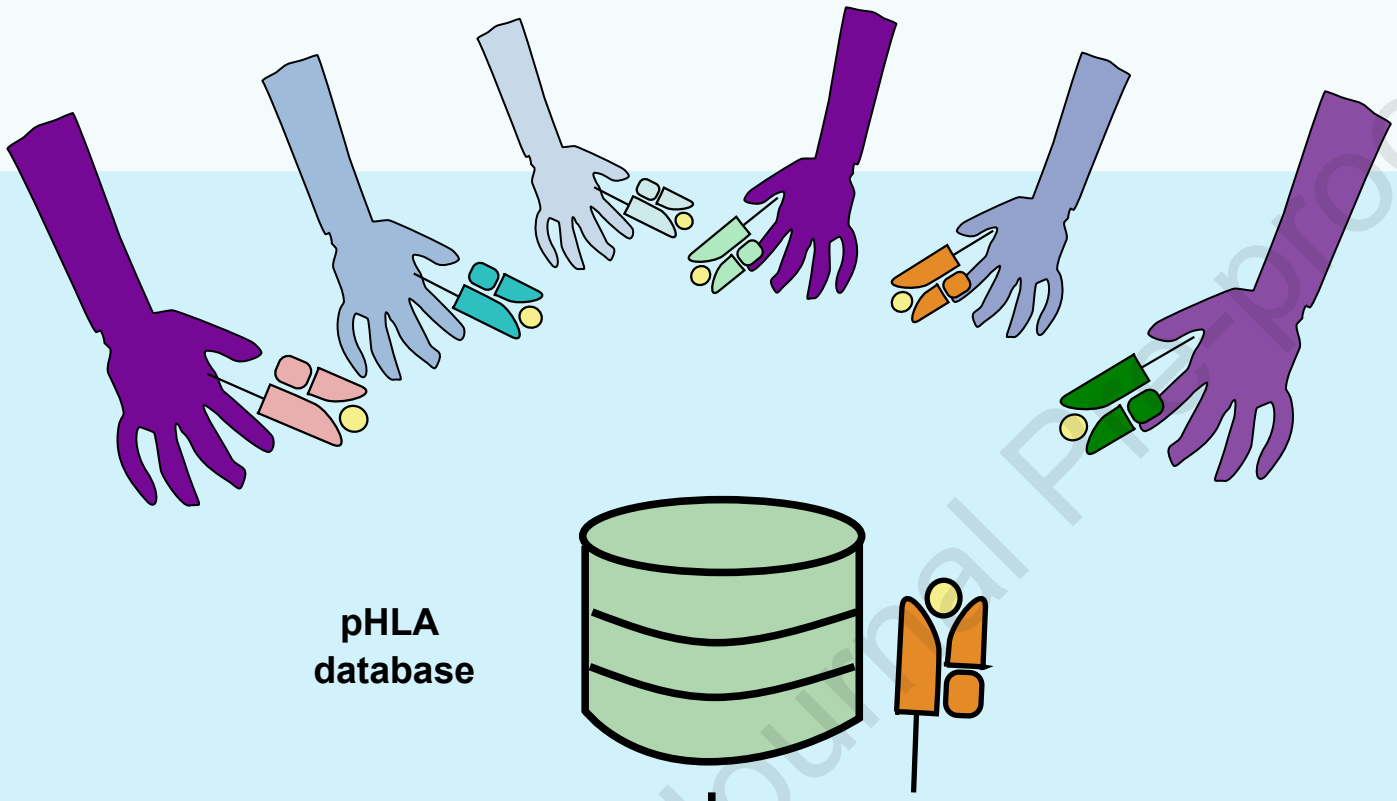
Revised Date: 13 November 2023

Accepted Date: 29 November 2023

Please cite this article as: Conev, A., Fasoulis, R., Hall-Swan, S., Ferreira, R., Kavraki, L.E., HLAEquity: Examining biases in pan-allele peptide-HLA binding predictors, *ISCIENCE* (2024), doi: https://doi.org/10.1016/j.isci.2023.108613.



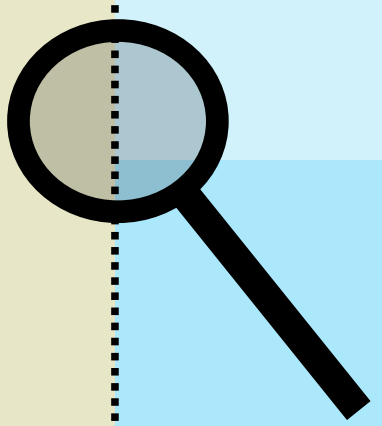
SOCIO-ECONOMIC BIAS



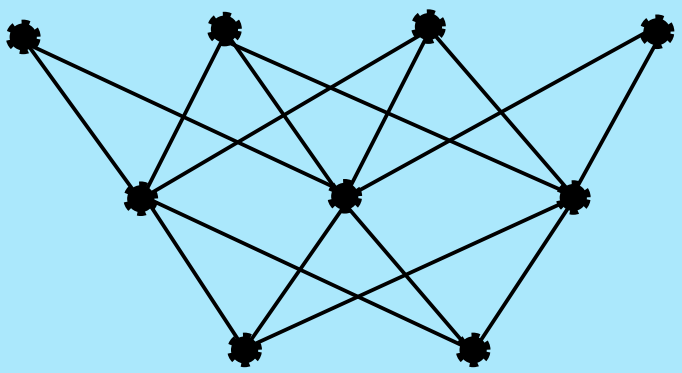
pHLA database

DATA BIAS

HLAEquity



ML-based pan-allele pHLA prediction



ALGORITHMIC BIAS

BIASED HEALTHCARE OUTCOMES

HLAEquity: Examining biases in pan-allele peptide-HLA binding predictors

Anja Conev[1,2], Romanos Fasoulis[1,2], Sarah Hall-Swan[1,2], Rodrigo Ferreira[1,*], Lydia E. Kavraki[1,3,**]

[1]Department of Computer Science, Rice University, Houston, TX, United States of America
[2]These authors contributed equally
[3]Lead Contact
[*]Correspondence: rf29@rice.edu
[**]Correspondence: kavraki@rice.edu

KEYWORDS Ethics in AI, AI and healthcare, AI bias, peptide-HLA

SUMMARY

Peptide-HLA (pHLA) binding prediction is essential in screening peptide candidates for personalized peptide vaccines. Machine Learning (ML) pHLA binding prediction tools are trained on vast amounts of data and are effective in screening peptide candidates. Most ML models report generalizing to HLA alleles unseen during training ("pan-allele" models). However, the use of datasets with imbalanced allele content raises concerns about biased model performance. First, we examine the data bias of two ML-based pan-allele pHLA binding predictors. We find that the pHLA datasets overrepresent alleles from geographic populations of high-income countries. Second, we show that the identified data bias is perpetuated within ML models, leading to algorithmic bias and subpar performance for alleles expressed in low-income geographic populations. We draw attention to the potential therapeutic consequences of this bias, and we challenge the use of the term "pan-allele" to describe models trained with currently available public datasets.

INTRODUCTION

The adaptive cellular immune response is a vital aspect of the human immune system, seeking to destroy infected or cancerous cells. A major component of the adaptive immune response in humans is the peptide-HLA (pHLA) complex, which consists of a class I human leukocyte antigen (HLA) receptor and a bound peptide derived from the proteasomal cleavage of intracellular proteins. Circulating T-cells recognize and respond to HLAs presenting a foreign peptide stemming from a viral or a cancer protein. Peptides that bind to HLAs are targets for therapeutics ranging from cancer immunotherapy to viral vaccines. Predicting binding affinity between target peptides and HLAs is a crucial step in developing effective therapeutics[1].

Table 1. List of pan-allele pHLA binding affinity predictors compiled by a recent comprehensive review by *Wang et al.*[2] with the addition of the number of citations (queried from the Pubmed library July 2023).

Predictor name	*Algorithm*	*Software available*	*Citations*	*Year*	*Reference*
NetMHCpan 4.1	**FFNN**	**Y**	**753**	**2020**	[3]
MHCflurry 2.0	**FFNN**	**Y**	**491**	**2018,2020**	[4,5]
NetMHCcons	Consensus	N	340	2012	[6]
MixMHCpred	Scoring function	Y	314	2017,2018	[7,8]
PickPocket	Scoring function	Y	198	2009	[9]

NetMHCstabpan	FFNN	Y	154	2016	[10]
ConvMHC	CNN	Y	94	2017	[11]
DeepHLApan	GRU+Attention	Y	72	2019	[12]
PSSMHCpan	Scoring function	Y	64	2017	[13]
MHCSeqNet	GRU	Y	59	2019	[14]
ACME	CNN	Y	55	2019	[15]
DeepSeqPan	CNN	Y	53	2019	[16]
TransPHLA	Multi-head self-attention	Y	38	2022	[17]
Anthem	AODE	Y	30	2021	[18]
MHCAttnNet	LSTM+Attention	Y	26	2020	[19]
DeepAttentionPan	CNN+Attention	Y	12	2021	[20]
MATHLA	LSTM+Attention	Y	10	2021	[21]
HLAB	XGBoost, KNN, SVM, NB, LR, DTree, Bagging	Y	9	2022	[22]
DeepNetBim	CNN+Attention	Y	8	2021	[23]
Seq2Neo	CNN	Y	5	2022	[24]

The task of predicting pHLA binding affinity is challenging. Genes encoding HLA receptors are among the most variable genes in the human genome, with over 25,000 identified alleles across the global population. Additionally, the number of potential peptide targets is large and difficult to experimentally screen. However, high-throughput mass-spectrometry brought about increasing amounts of pHLA binding data. These data opened the door for *in silico* pHLA binding affinity prediction and the development of machine-learning (ML) based tools, with the latest approaches adopting neural network architectures[5,25]. Several ML models in the current literature provide pHLA binding affinity predictions for any HLA allele, even when the allele is absent during the training process. The authors of these models refer to them as “pan-allele" models[5,25]. The promise of pan-allele predictions has great therapeutic significance, as it enables prediction for any HLA expressed in a patient. The early models were proclaimed as a technology that will enable individualized immunotherapy[26]. Today, pan-allele prediction models are a significant component in immunotherapy pipelines[27–29]. A recent survey identified 27 different methods for pHLA binding affinity prediction[2]; 20 out of 27 methods claim to be pan-allele while 17 out of 20 pan-allele methods utilize ML and neural network approaches (see also **Table 1**). The field strongly leans toward the ML-based pan-allele prediction paradigm.

In the field of ML, it is widely recognized that models can demonstrate various forms of bias. As the models are deployed in real-world applications this phenomenon can lead to disparate impacts[30]. Biased facial recognition software showed discrimination based on race and gender[31]. Decision-making algorithms deployed in crime prediction, credit lending, and hiring can perpetuate racial bias and injustice[32,33]. The same issues arise with ML applications in healthcare. A risk assessment algorithm was found to misassign sick Black patients with the same low level of risk as less sick White patients[34]. There are also issues related to ML bias in genomic-driven cancer treatments, as most of the sequenced patients in The Cancer Genome Atlas project are of European ancestry, while people with other ancestries are underrepresented[35].

Focusing on ML models in healthcare, Norori et al. categorized different perspectives of bias as human, data, and algorithmic bias[36]. Human bias refers to the individual biases, societal prejudices, and power imbalances that affect every human. Because humans create the data and the algorithms, our biases have a direct effect on what we create regardless of intent. Data bias refers to imbalanced data that may not be representative of the relevant portion of the human population. Lastly, algorithmic bias refers to how algorithms enforce the biases in the data. In healthcare, this translates to ML models that could give misguided predictions on specific geographic populations, affecting the efficacy of treatments that they might receive. Algorithmic bias includes the training criterion chosen for an ML model, as well as the way existing imbalances in the data (e.g., class imbalance) are handled during training[36].

In this work, we investigate data and algorithmic bias in current pan-allele pHLA binding affinity prediction models. First, we find bias within publicly available pHLA datasets. Using the population coverage metric, we clearly see that the available peptide-HLA datasets do not equally represent different geographic populations. Moreover, by using the four different income classification levels defined by the World Bank, we associate the inequalities found in the calculated allele population coverage with income inequalities between nations. Next, we look at the algorithmic bias in two popular pan-allele pHLA binding predictors. We discover that the algorithms perpetuate the data bias, leading to differences in model performance across alleles. Due to this algorithmic bias, populations in lower-income countries could benefit less from the ML predictions of the pan-allele models than populations in higher-income countries, in regards to therapeutic efficacy. Ultimately, we question the use of the term "pan-allele" to describe a pHLA binding predictor. Our aim is to raise consciousness about the possible impact that bias can have in pHLA binding predictors, and, ultimately, in immunoinformatics and immunotherapy research.

RESULTS

Skewed allele representation in pHLA training datasets reveals data bias disadvantaging low-income populations

First, we investigate the distribution of alleles in the training datasets of the NetMHCpan4.1[3] and MHCFlurry2.0[5] models (**Figures 1, S1, S2**). Note that mass spectrometry (MS) datasets (blue) have more data than binding affinity (BA) datasets (red). This is expected as the MS experiments have higher throughput. Overall, each dataset contains data for a limited number of alleles as compared to over 25,000 present in the human population and the allele distributions have a "long tail". In particular, there are less than 25 alleles that are represented with more than 5,000 data points in the datasets. We notice an overrepresentation of the A*02:01 and A*03:01 alleles. Previous literature[37] shows that alleles A*02:01, A*03:01 are most prevalent in Caucasian populations while A*11:01 and A33:03 are prevalent in Asian populations and A*23:01 and A*30:02 are most common in African populations. As expected, since both NetMHCpan4.1 and MHCFlurry2.0 collect their data from the Immune Epitope Database (IEDB)[38], BA datasets have similar allele content (i.e., red markers align). However, MHCFlurry2.0_MS has more data as compared to NetMHCpan4.1_MS for a few alleles (for example, A*11:01, A*34:02, B*40:02, C*12:02 among others). These alleles are outlined in bold in **Figures 1, S1, and S2** and for them, the blue lines in the plot diverge. In particular, MHCFlurry2.0_MS includes recently collected by Sarkizova et al.[37] targeting most of the human population and specifically targeting some of the previously underrepresented alleles.

Second, we quantify how the allele content of each dataset relates to the allele contents of specific geographic populations (**Figure 2**). We calculate the scaled population coverage (sPC90)[39] for each dataset across all geographic populations contained in the Allele Frequency Net Database (AFND)[40]. The higher values of sPC90 indicate a better representation of the population within the dataset. We group the results based on the income level of the country of origin (**Figure 2**). We see a clear imbalance in terms of population coverage across different income levels. We highlight the statistical significance in the distributions of sPC90 across populations of different income levels. All datasets are biased towards the countries with higher income levels and on average they have higher sPC90 coverage for those populations. Note that the difference in sPC90 across the income categories is smallest for MHCFlurry2.0_MS (the boxes are closer together and the green low-income box is higher than for other datasets). The data recently sampled by Sarkizova et al.[37] for underrepresented alleles and included in the MHCFlurry2.0_MS could be narrowing this difference down. Note that the high and the higher middle-income populations have a high deviation of the sPC90 scores. This is especially evident in countries with a high diversity of the ancestries of the populations within the country. For example, when we divide the US populations by their ancestry (**Figure S3**) we see that different ancestries are unequally represented.

Pan-allele algorithms produce less accurate predictions for alleles from low-income populations

We assess whether the notion of algorithmic bias, as defined by Norori et al.[36], exists in popular pan-allele pHLA binding prediction tools. Algorithmic bias could be caused by training bias or the imbalance that occurs by having an uneven number of data points corresponding to each allele. In the pHLA binding prediction task, the algorithmic bias would translate to having vastly unequal prediction performance for alleles that are expressed in different geographic populations. We tested both NetMHCpan4.1 and MHCFlurry2.0, two widely used pan-allele neural network-based pMHC binding predictors, on the dataset from Pyke et al.[41] (see Methods). Ideally, we would like to see NetMHCpan4.1 and MHCFlurry2.0 performing equally well on all HLA alleles (with both PPV and FOOP being high). This would ensure that predictions of these models are accurate and can be used in downstream applications and therapeutics, independently of a patient's geographic origin or allele expression.

Performance for MHCFlurry2.0 and NetMHCPan4.1 can be seen in **Figure 3**. Both MHCFlurry2.0 and NetMHCPan4.1 perform differently across different alleles. Moreover, we see that the fluctuations in performance mostly follow the same pattern for both MHCFlurry2.0 and NetMHCpan4.1, indicating that the two tools mostly succeed on the same alleles and fail on the same alleles too. Nevertheless, both methods fail to perform equally well on all alleles, given the big fluctuations in per-allele performance.

Furthermore, we identify alleles that both MHCFlurry2.0 and NetMHCpan4.1 succeed or underperform in terms of PPV and FOOP. The allele HLA-A*02:52, for which the models are underperforming, especially in terms of FOOP, was previously identified to be prominent in Iranian Kurdish populations[41], and at the same time, it is not prominent in higher-income countries or populations. It is also an allele that did not exist in the datasets that were used by NetMHCpan4.1 and MHCFlurry2.0 for training, indicating that pan-allele ML models may well underperform for alleles not previously seen. On the contrary, the allele HLA-A*01:01, previously found to be expressed in high percentage in European and North American populations[37], performs very well, both in terms of PPV and FOOP. Similarly, the allele HLA-B*15:13 is a low-performing allele for both pHLA binding prediction tools and is mostly expressed in upper/lower middle-income countries like Malaysia or Indonesia, but it is non-existent in higher-income countries and populations. On the contrary, allele HLA-B*38:01 is much more prominent in high-income countries

(examples here are Israel and Italy) than in countries of low/middle income (examples here are Tunisia and Thailand). Similar patterns arise when examining other high/low performing allele pairs, with very few notable exceptions, such as the HLA-B*08:01, an allele expressed mostly in higher-income populations, but with remarkably low PPV.

DISCUSSION

In this study, we inspect data and algorithmic bias in the pHLA binding prediction pipeline. We examine the content of different training datasets and identify the lack of alleles corresponding to populations in lower-income countries. For example, there are many data points associated with the alleles prevalent in European populations (i.e., A*02:01), while there are fewer points for the alleles prevalent in the African (i.e., A*23:01) or Asian (i.e., A*11:01) populations (**Figure 1**). This finding is quantified with the population coverage metric (sPC90) in **Figure 2** and it is clear that the populations in higher-income countries are better represented by the datasets. We showcase how the pan-allele algorithms accumulate and perpetuate identified data biases. We specifically show that state-of-the-art pHLA binding predictors underperform on alleles expressed in populations in lower-income countries (i.e., Iranian Kurds, Malaysia Mandailing, Tanzania Masai populations). Ultimately, because these algorithms do not perform well on all alleles, they should not be described as pan-allele, as the term falsely implies that they will provide good predictions for all alleles. Our findings are significant for the future development of medical treatments on at least two levels.

At one level, we can take these results to highlight potential disparate impacts when it comes to the usage of these datasets and models for developing medical treatments for different geographic populations. Through our analysis, we find that MHCFlurry2.0 and NetMHCpan4.1 perform poorly on some alleles while performing well on others. More importantly, the models have superior performance for populations from high-income countries for which alleles are highly represented in the datasets, as compared to populations from low-income countries that are not well represented by the alleles in the datasets. When a tool does not perform well on certain alleles, the therapeutics that are developed using that tool may not perform well on individuals who have those alleles. Therefore, there is a danger of developing sub-par peptide vaccines or T-cell-based immunotherapy protocols for certain populations from lower-income countries. This distinction would only help exacerbate the long history of inequity that has existed when it comes to medical treatment for groups in higher-income countries than for groups in lower-income countries.

At a second level, our investigation invites additional research not only on differences in performance across different economic levels but also on the relationship between existing economic differences and the social and historical circumstances that have helped make way for these differences and that appear inconspicuously in other ways around the data. Researchers have shown that biases in ML are not always grounded on arbitrary circumstances or statistical inaccuracies but are at times predicated on historical social practices and institutions[31]. In the case of the pHLA datasets, associated metadata shows that the samples were collected primarily from countries with higher levels of income. **Figure 4** summarizes the country of origin for the institutions that conducted the experimental essays curated by the IEDB. More than a quarter of studies originate from institutions within the United States, followed by more than 11% from Germany, 9% from Australia, and around 8% from China. This information gives reason to speculate about human bias driving data bias: as institutions in higher-income countries are able to collect more data, differences between allele representations become present in the database, ultimately leading to algorithmic bias as seen in the difference in performance between populations with lower and higher

incomes. More information on the practices on how data is collected worldwide and how more opportunities for populations in lower-income countries can be granted, could in turn help shape strategies for including further data from these populations.

The ultimate aim of this study is to highlight the issues of bias in the pHLA binding prediction workflow and to highlight that this bias relates to the inherent systemic and historical patterns against geographic populations of a certain economic status. From dataset collection to algorithm development the identified bias is perpetuated by pan-allele models. We hope that our work leads the community in continuing to recognize sources of bias such as those we identify in this study. Once the sources of bias are acknowledged, they can be mitigated. For instance, Norori et al. propose addressing existing bias in healthcare applications through open science[36]; this can be achieved through data sharing, setting proper data standards, defining proper evaluation metrics that are common among studies, and promoting AI explainability. Many of these propositions have already been established in the immunoinformatics community. Databases like IEDB[40] and AFND[40], among others, share pHLA data effectively. New pHLA binding prediction approaches are adopting explainability modules moving away from the black-box ML paradigm[17]. Another interesting avenue for allele bias mitigation is to train personalized, per-patient models[42]. However, more work can be done regarding data collection, where very few studies sample binding affinities for alleles from different geographic populations[37,41].

Limitations of the study

Note that we focus our analysis on the two highly regarded state-of-the-art prediction tools (NetMHCpan4.1 and MHCFlurry2.0). We chose these two tools because they are the most widely cited among a range of other pan-allele predictors. It has been reported that most of the pan-allele tools rely on the same source of IEDB curated experimental data for training and that their training datasets have a large overlap of content[2]. We see this overlap in our analysis between the NetMHCpan4.1 and MHCFlurry2.0 datasets. In particular, BA datasets of the two tools almost entirely align (see red markers in **Figures 1, S1, S2**) while the MS datasets show an overlap across most of the alleles (see blue markers in **Figures 1, S1, S2**). The reported overlap enhances the representativeness of our analysis. Nevertheless, we acknowledge that many other pan-allele predictors lie beyond the scope of our current analysis, presenting an exciting avenue for future work. In addition, we make our evaluation pipeline open source. Authors of future tools can test the population coverage of their training datasets prior to training.

There is an opportunity for further research, not only of the differences between geographic populations in accordance with their income level but of the differences within a single geographic population in accordance with different "ancestries" in that population. Based on the World Bank's classification table, we have considered geographic populations in the USA as "high income". However, the geographic population of the USA does not have a homogeneous "ancestry." Instead, as pointed out previously, it is composed of different *ethnogeographic* populations, such as USA European, USA Hispanic, USA African American, and USA Asian. As **Figure 4** shows, the datasets cover the USA European population more than they cover populations with other ancestries. These differences in turn correlate to differences in levels of income between different ethnic populations in the USA[43], as predicated on practices of colonialism and ethnic segregation[44].

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
- METHODS DETAILS
  - Mapping HLAs to geographic populations
  - Classifying geographic populations by income
  - Examining data bias
  - Examining algorithmic bias
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Accuracy Metrics
  - Statistical Significance

SUPPLEMENTARY INFORMATION

Supplemental information can be found online at *iScience*.

ACKNOWLEDGEMENTS

We thank KavrakiLab members for many insightful discussions that helped guide this work. We are grateful to our classmates Ria Stevens, Thomas Herring and Dr. Wil Thomason for motivating this work with their in-class participation and discussions. We are especially grateful to our collaborators Dr. Dinler Amaral Antunes and Dr. Mauricio Menegatti Rigo for their guidance in immunoinformatics projects that initially inspired this work.

Funding: Work on this project has been supported by National Institutes of Health NIH [U01CA258512] and Rice University funds.

AUTHOR CONTRIBUTIONS

Concept and Design, A.C., R.Fa., S.H.S., R.Fe.; Data Bias Analysis, A.C.; Algorithm Bias Analysis R.Fa.; Writing – Original draft, A.C., R.Fa., S.H.S.; Writing – Review and Editing, A.C., R.Fa., S.H.S., R.Fe., L.K.; Funding, Resources and Supervision, R.Fe., L.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research. One or more of the authors of this paper self-

identifies as a member of the LGBTQIA+ community. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

FIGURE TITLES AND LEGENDS

Figure 1. HLA-A allele frequencies in each of the training datasets (i.e., MHCFlurry2.0_BA, NetMHCpan4.1_BA, MHCFlurry2.0_MS, NetMHCpan4.1_MS). Allele codes are indicated on the x-axis while the number of points in the dataset for each allele is indicated on the y-axis. Allele codes are bolded if the respective number of data points in MHCFlurry2.0_MS is higher than the number of data points in NetMHCpan4.1_MS.

Figure 2. Scaled population coverage (sPC90) indicated on the y-axis of training datasets indicated on the x-axis. Each point corresponds to a particular geographic population and points are grouped into boxplots based on the income level of the population's country. The difference between the sPC90 distributions is evaluated with one-way ANOVA and the one-sided KS test. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Figure 3. PPV (first row) and FOOP (second row) results for MHCFlurry2.0 (points in red) and NetMHCpan4.1 (points in blue). The x-axis corresponds to different alleles in the dataset. The y-axis corresponds to either the computed PPV value or the computed FOOP value. Different point shapes correspond to different HLA loci (A, B, C), separated by blue dashed lines. The top and bottom barplots correspond to alleles that have very good or very bad PPV and FOOP scores on average, respectively. For each allele, we plot low-income populations (above the blue dashed line) and higher-income populations (below the blue dashed line) that express this allele the most.)

Figure 4. Country of origin of work curated by the IEDB.

REFERENCES

1. Lizée, G., Overwijk, W.W., Radvanyi, L., Gao, J., Sharma, P., and Hwu, P. (2013). Harnessing the Power of the Immune System to Target Cancer. Annual Review of Medicine *64*, 71–90. 10.1146/annurev-med-112311-083918.

2. Wang, M., Kurgan, L., and Li, M. (2023). A comprehensive assessment and comparison of tools for HLA class I peptide-binding prediction. Briefings in Bioinformatics, bbad150. 10.1093/bib/bbad150.

3. 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. 10.1093/nar/gkaa379.

4. O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Systems *7*, 129-132.e4. 10.1016/j.cels.2018.05.014.

5. 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*, 42-48.e7. 10.1016/j.cels.2020.06.010.

6. Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics *64*, 177–186. 10.1007/s00251-011-0579-8.

7. Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P.O., Kandalaft, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. PLoS Comput Biol *13*, e1005725. 10.1371/journal.pcbi.1005725.

8. Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R.T., Racle, J., Coukos, G., and Bassani-Sternberg, M. (2018). The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. J Immunol *201*, 3705–3716. 10.4049/jimmunol.1800914.

9. Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinformatics *25*, 1293–1299. 10.1093/bioinformatics/btp137.

10. Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A.B., Nielsen, I.K., Nielsen, M., and Buus, S. (2016). Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. J Immunol *197*, 1517–1524. 10.4049/jimmunol.1600582.

11. Han, Y., and Kim, D. (2017). Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. BMC Bioinformatics *18*, 585. 10.1186/s12859-017-1997-x.

12. Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., Gu, X., Wu, J., Zhou, Z., and Chen, S. (2019). DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. Front Immunol *10*, 2559. 10.3389/fimmu.2019.02559.

13. Liu, G., Li, D., Li, Z., Qiu, S., Li, W., Chao, C.-C., Yang, N., Li, H., Cheng, Z., Song, X., et al. (2017). PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. Gigascience *6*, 1–11. 10.1093/gigascience/gix017.

14. Phloyphisut, P., Pornputtapong, N., Sriswasdi, S., and Chuangsuwanich, E. (2019). MHCSeqNet: a deep neural network model for universal MHC binding prediction. BMC Bioinformatics *20*, 270. 10.1186/s12859-019-2892-4.

15. Hu, Y., Wang, Z., Hu, H., Wan, F., Chen, L., Xiong, Y., Wang, X., Zhao, D., Huang, W., and Zeng, J. (2019). ACME: pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. Bioinformatics *35*, 4946–4954. 10.1093/bioinformatics/btz427.

16. Liu, Z., Cui, Y., Xiong, Z., Nasiri, A., Zhang, A., and Hu, J. (2019). DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. Sci Rep *9*, 794. 10.1038/s41598-018-37214-1.

17. Chu, Y., Zhang, Y., Wang, Q., Zhang, L., Wang, X., Wang, Y., Salahub, D.R., Xu, Q., Wang, J., Jiang, X., et al. (2022). A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. Nat Mach Intell *4*, 300–311. 10.1038/s42256-022-00459-7.

18. Mei, S., Li, F., Xiang, D., Ayala, R., Faridi, P., Webb, G.I., Illing, P.T., Rossjohn, J., Akutsu, T., Croft, N.P., et al. (2021). Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. Brief Bioinform *22*, bbaa415. 10.1093/bib/bbaa415.

19. Venkatesh, G., Grover, A., Srinivasaraghavan, G., and Rao, S. (2020). MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. Bioinformatics *36*, i399–i406. 10.1093/bioinformatics/btaa479.

20. Jin, J., Liu, Z., Nasiri, A., Cui, Y., Louis, S.-Y., Zhang, A., Zhao, Y., and Hu, J. (2021). Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. Proteins *89*, 866–883. 10.1002/prot.26065.

21. Ye, Y., Wang, J., Xu, Y., Wang, Y., Pan, Y., Song, Q., Liu, X., and Wan, J. (2021). MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. BMC Bioinformatics *22*, 7. 10.1186/s12859-020-03946-z.

22. Zhang, Y., Zhu, G., Li, K., Li, F., Huang, L., Duan, M., and Zhou, F. (2022). HLAB: learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction. Brief Bioinform *23*, bbac173. 10.1093/bib/bbac173.

23. Yang, X., Zhao, L., Wei, F., and Li, J. (2021). DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. BMC Bioinformatics *22*, 231. 10.1186/s12859-021-04155-y.

24. Diao, K., Chen, J., Wu, T., Wang, X., Wang, G., Sun, X., Zhao, X., Wu, C., Wang, J., Yao, H., et al. (2022). Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. Int J Mol Sci *23*, 11624. 10.3390/ijms231911624.

25. Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., Buus, S., and Nielsen, M. (2009). NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics *61*, 1–13. 10.1007/s00251-008-0341-z.

26. 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. 10.1371/journal.pone.0000796.

27. Antunes, D.A., Abella, J.R., Hall-Swan, S., Devaurs, D., Conev, A., Moll, M., Lizée, G., and Kavraki, L.E. (2020). HLA-Arena: A Customizable Environment for the Structural Modeling and Analysis of Peptide-HLA Complexes for Cancer Immunotherapy. JCO clinical cancer informatics *4*. 10.1200/CCI.19.00123.

28. Hundal, J., Kiwala, S., McMichael, J., Miller, C.A., Xia, H., Wollam, A.T., Liu, C.J., Zhao, S., Feng, Y.-Y., Graubert, A.P., et al. (2020). pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. Cancer Immunology Research *8*, 409–420. 10.1158/2326-6066.CIR-19-0401.

29. Rigo, M.M., Fasoulis, R., Conev, A., Hall-Swan, S., Antunes, D.A., and Kavraki, L.E. (2022). SARS-Arena: Sequence and Structure-Guided Selection of Conserved Peptides from SARS-related Coronaviruses for Novel Vaccine Development. Frontiers in Immunology *13*.

30. Barocas, S., and Selbst, A.D. (2016). Big Data's Disparate Impact. California Law Review *104*, 671–732.

31. Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR), pp. 77–91.

32. Lum, K., and Isaac, W. (2016). To predict and serve? Significance *13*, 14–19. 10.1111/j.1740-9713.2016.00960.x.

33. Ajunwa, I. (2019). The Paradox of Automation as Anti-Bias Intervention. Cardozo L. Rev. *41*, 1671–1742.

34. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453. 10.1126/science.aax2342.

35. Dankwa-Mullan, I., and Weeraratne, D. (2022). Artificial Intelligence and Machine Learning Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity. Cancer Discovery *12*, 1423–1427. 10.1158/2159-8290.CD-22-0373.

36. Norori, N., Hu, Q., Marcelle Aellen, F., Dalia Faraci, F., and Tzovara (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns *2*, 100347. 10.1016/j.patter.2021.100347.

37. Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.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. 10.1038/s41587-019-0322-9.

38. 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. 10.1093/nar/gky1006.

39. Bui, H.-H., Sidney, J., Dinh, K., Southwood, S., Newman, M.J., and Sette, A. (2006). Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics *7*, 153. 10.1186/1471-2105-7-153.

40. Gonzalez-Galarza, F.F., McCabe, A., Santos, E.J.M. dos, Jones, J., Takeshita, L., Ortega-Rivera, N.D., Cid-Pavon, G.M.D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., et al. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acids Research *48*, D783–D788. 10.1093/nar/gkz1029.

41. Pyke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., McClory, R., et al. (2021). Precision Neoantigen Discovery Using Large-scale Immunopeptidomes and Composite Modeling of MHC Peptide Presentation. Molecular & Cellular Proteomics *20*. 10.1016/j.mcpro.2021.100111.

42. Liang, S., Jiang, X., Chiu, Y., Xu, H., Kim, K.H., Lizee, G., and Chen, K. (2023). An interpretable ML model to characterize patient-specific HLA-I antigen presentation. bioRxiv, 2023.03.12.532264. 10.1101/2023.03.12.532264.

43. Center, P.R. (2016). 1. Demographic trends and economic well-being. Pew Research Center's Social & Demographic Trends Project. https://www.pewresearch.org/social-trends/2016/06/27/1-demographic-trends-and-economic-well-being/.

44. The Color of Law: A Forgotten History of How Our Government Segregated America Economic Policy Institute. https://www.epi.org/publication/the-color-of-law-a-forgotten-history-of-how-our-government-segregated-america/.

45. Barker, D.J., Maccari, G., Georgiou, X., Cooper, M.A., Flicek, P., Robinson, J., and Marsh, S.G.E. (2023). The IPD-IMGT/HLA Database. Nucleic Acids Research *51*, D1053–D1060. 10.1093/nar/gkac1011.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
NetMHCpan-4.1 training dataset	Reynisson et al.[3]	https://services.healthtech.dtu.dk/suppl/immunology/NAR_NetMHCpan_NetMHCIIpan/
MHCFlurry2.0 BA training dataset	O'Donnell et al.[5]	https://data.mendeley.com/datasets/zx3kjzc3yx/3 (**Data S3**)
MHCFlurry2.0 AP training dataset	O'Donnell et al.[5]	https://data.mendeley.com/datasets/zx3kjzc3yx/3 (**Data S5**)
MHCFlurry2.0 PS training dataset	O'Donnell et al.[5]	https://data.mendeley.com/datasets/zx3kjzc3yx/3 (**Data S6**)
Pyke et al. evaluation dataset	Pyke et al.[41]	**Table S1** and **Table S5** (in original source)
AFND Database	Gonzalez-Galarza et al.[40]	http://www.allelefrequencies.net/
Software and algorithms		
R	R	https://www.r-project.org/
Python	Python Software Foundation	https://www.python.org
NetMHCpan-4.1	Reynisson et al.[3]	https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/
MHCFlurry2.0	O'Donnell et al.[5]	https://github.com/openvax/mhcflurry
IEDB Population Coverage tool	Bui et al.[39]	http://tools.iedb.org/population/download/
Analysis scripts	Original code	https://github.com/KavrakiLab/HLAequity
Other		
World Bank's 2022-2023 "Country and Lending Groups" income classification table	https://datacatalogfiles.worldbank.org/ddh-published/0037712/DR0090755/CLASS.xlsx	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Dr. Lydia E. Kavraki (kavraki@rice.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. All of these datasets are exhaustively referenced in the Key Resources Table. In addition, the processed data is available at: *https://github.com/KavrakiLab/HLAequity*
- All the related code that is needed to reproduce the results of the study is available at: *https://github.com/KavrakiLab/HLAequity*
- Any additional information required to re-analyze the reported data can be provided by the lead contact upon request.

METHOD DETAILS

Mapping HLA alleles to geographic populations

We collect the distributions of alleles in different geographic populations from the Allele Frequency Net Database (AFND)[40] (accessed May 2023). AFND collects data on the genetic variation of highly variable immune-related genes, including HLA genes. This type of data comes from more widely conducted population studies that are not specific to the pHLA binding prediction tasks[40]. AFND has collected and curated data from more than 10 million people and classified them into more than 1600 population groups. Note that the AFND label of "population" contains both a geographic designation (the current country in which that population is found) and an ethnic designation (the "ancestry" of that population). For example, population labels for the USA appear in the AFND as USA Hispanic, USA Caucasian, USA Asian, USA African American, etc. However, for some populations the ethnic designation is missing or vague and the AFND states that the ethnic group designations are under revision and will be improved in the near future. For that reason, we focus our analysis on the geographic designation label as opposed to ethnic or ancestry-based labels. We refer to the population labels as "geographic populations". The detailed description of the steps we took to download and clean the AFND data are described in the following paragraphs.

We query the AFND database for allele frequencies using the API http://www.allelefrequencies.net/hla6006a.asp for different loci (i.e., A, B and C). We format the allele names in the standard nomenclature as described by the IPD-IMGT[45] keeping the information about the gene, the allele and the specific protein (i.e., format of HLA-A*01:01). We map the population labels given by AFND to specific countries. We remove the duplicate entries for "population"-"allele" pairs. We perform sanity checks on the data to make sure that the allele frequencies per population add up to 1. As a result, we gather the cleaned population frequencies of different alleles across in csv files (one file per loci). We unify the per-loci population frequencies into a single larger file. We clean this data keeping only the populations that have information for each of the selected loci (i.e., A, B and C). This preprocessed AFND data is used as the background population frequency for calculating the population coverage of different datasets.

Classifying geographic populations by income

To better convey our findings on the existence of data and algorithmic bias in pHLA binding predictors, we group the geographic populations according to the income levels of the countries, and we perform a nation-based economic analysis. We acknowledge the relationship between current international and international economic differences and historical forms of ethnic segregation and oppression. As we explain in detail in the Discussion section, by shedding light on existing economic differences between relevant geographic populations, we can then think more critically about these economic differences in relation to their historical complexities, including specifically on the history of colonization. In discussing

the limitations of our study, our effort is precisely to invite more research that can help make these historical relationships clearer.

To classify the geographic populations based on their income level, we first identified the "country" appearing in each group's label and then used the World Bank's 2022-2023 "Country and Lending Groups" classification table to determine the income level for that country (World Bank information accessed at: https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023). The World Bank's 2022-2023 "Country and Lending Groups" table classifies 217 countries around the world along four income levels (as defined by gross national income per capita in 2021). The four levels of income are low-income (\$1085 or less), lower-middle-income (\$1,086 to \$4,255), upper-middle-income (\$4,256 to \$13,205), and high-income (\$13,205 or more).

Examining data bias

To examine data bias, we analyze training datasets from two predictors that are widely used in the literature: MHCFlurry2.0[5] and NetMHCpan4.1[3]. We choose these two state-of-the-art tools as they are most widely used and cited. A recent comprehensive study[2] curated a list of pan-allele pHLA binding affinity predictors. We extract the number of citations for each of the tools from the Pubmed library and outline the number of citations in the following table. MHCFlurry2.0 and NetMHCpan4.1 are most widely cited in addition to being recent (see **Table 1**).

Note that both MHCFlurry2.0 and NetMHCpan4.1 gather data for training by querying the Immune Epitope Database (IEDB), where they find curated experimental data. We examine both the binding affinity (BA) portions (MHCFlurry2.0_BA, NetMHCpan4.1_BA) and the mass-spectrometry (MS) portions (MHCFlurry2.0_MS and NetMHCpan4.1_MS) of the training datasets. The MS data can be either mono-allelic or multi-allelic. We refer to mono-allelic data as a definite peptide-HLA pair, while, in multi-allelic data, each peptide can potentially bind to up to six alleles. Deconvolution of the multi-allelic data is necessary in order to define the allele to which each peptide binds. To deconvolute the multi-allelic data, we used a binding affinity predictor (NetMHCpan4.1 or MHCFlurry2.0), and, for each peptide, we choose the allele to which the peptide has the strongest predicted binding affinity (out of six potential ones), thus converting multi-allelic data to mono-allelic data. All peptide pairs with a predicted binding rank of ≥ 0.5 are excluded, to remove peptides that do not bind to any of the designated alleles[41].

We calculate the population coverage using the method devised by Bui et al. and implemented in the Immune Epitope Database (IEDB) tools[39]. We use the AFND frequencies (see Section “Mapping HLA alleles to geographic populations”) as ground truth allele frequencies of geographic populations. Population coverage has been used to estimate a portion of a population protected ("covered") by a proposed peptide vaccine. In our study, instead of evaluating a quality of a vaccine, we are estimating the quality of the dataset. The inputs to the population coverage tool are peptide-allele pairs present in a dataset. We extract the PC90 metric calculated by the tool. PC90 corresponds to the number of data points in the dataset that covers 90% of the geographic population. To adequately compare differently sized datasets, we divide PC90 by the dataset size to get the scaled PC90 (sPC90). A lower sPC90 indicates that 90% of individuals in this geographic population are represented by a small portion of the dataset. High values of sPC90 indicate that 90% of individuals in this geographic population are represented by a large portion of the dataset. Ideally, the sPC90 of a dataset should be high and equal across different geographic populations.

Examining algorithmic bias

To test whether state-of-the-art binding prediction tools perform equally well among different alleles, we collected an independent dataset, found in Pyke et al.[41], which is not used in the training of state-of-the-art pHLA binding affinity predictors. This dataset is particularly valuable because it contains data on the alleles previously unseen in publicly available datasets. For example, HLA-A*02:52, unique to this dataset, exhibits high frequency, (around 7%) in the Iranian Kurdish geographic population. The fact that these alleles were completely missing in the training phase of state-of-the-art pHLA binding affinity predictors mimics the case of testing the performance of a patient with a rare or unseen allele. The dataset consists of both mono-allelic and multi-allelic data points. We deconvolute the multi-allelic data as we did with the training datasets. Finally, as mono-allelic and multi-allelic data only contain binders, many non-binder peptides (decoys) need to be generated to evaluate binding prediction tools. We generate 500,000 decoys that are randomly selected from the human proteome for each peptide length found in the dataset (8-mers to 11-mers)[5,41].

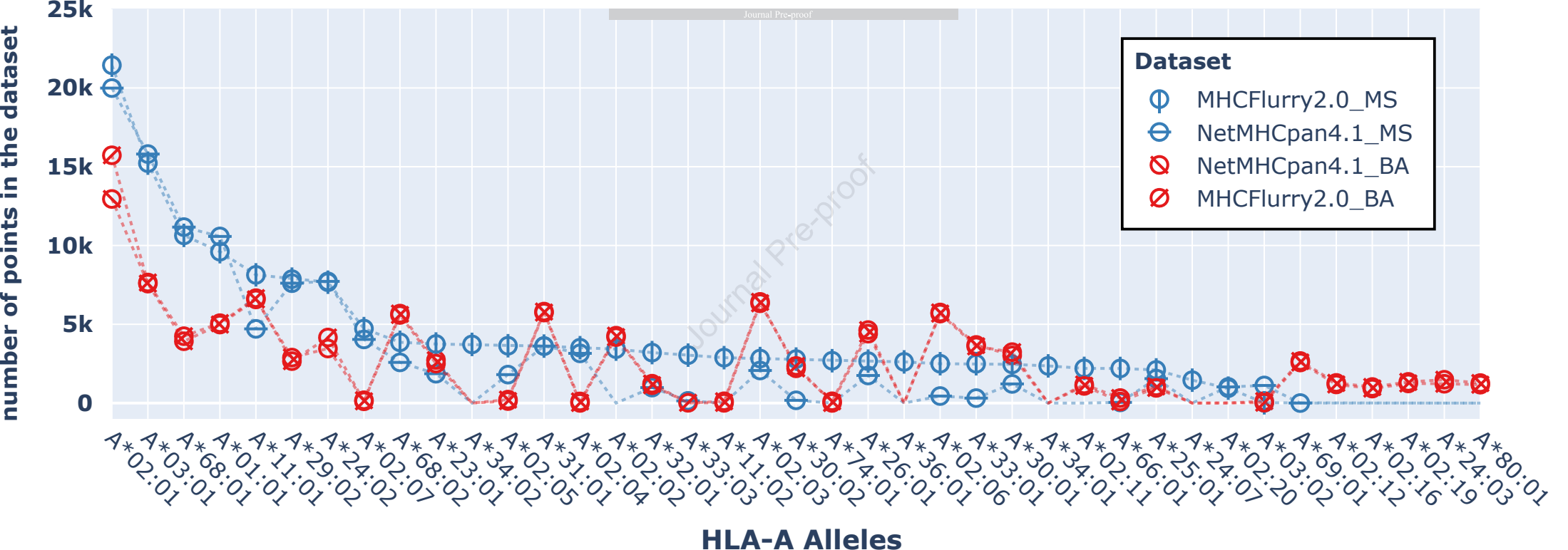
QUANTIFICATION AND STATISTICAL ANALYSIS

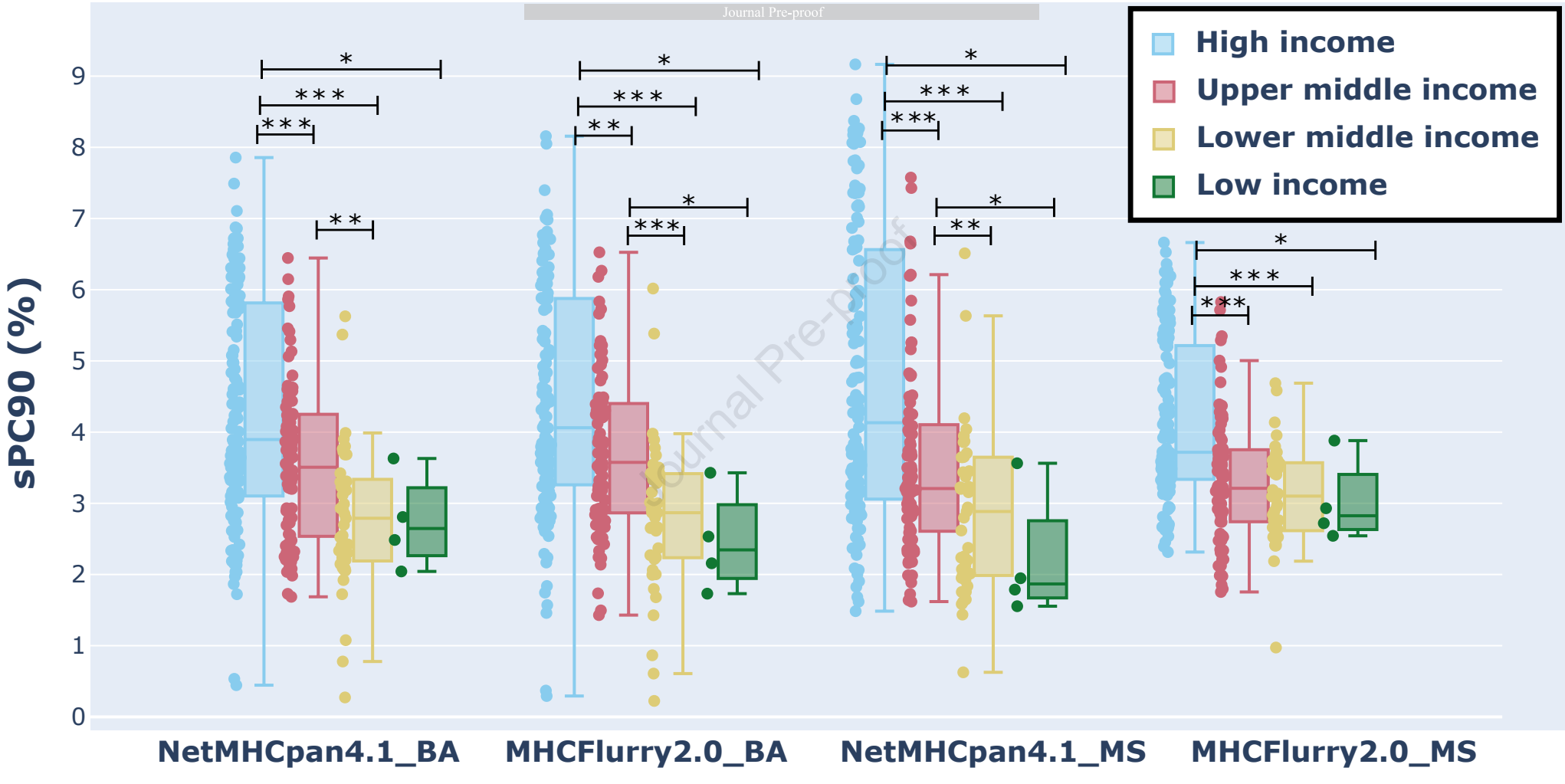
Accuracy Metrics

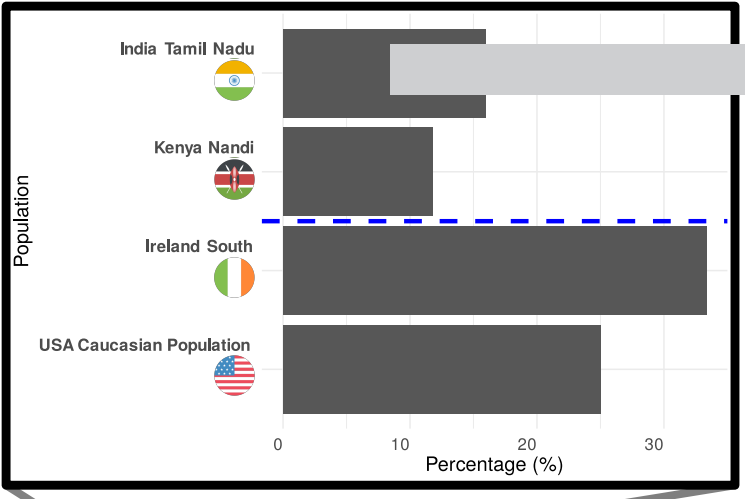
To evaluate per-allele performance for state-of-the-art pHLA binding prediction tools, we employ the commonly used metrics Positive Predicted Value (PPV) and the Fraction Of Observed Peptides (FOOP)[5,41]. PPV for each allele is calculated by predicting binding scores for all positive peptide binders and for all the decoys generated from the human proteome. These predictions are then concatenated and ranked by order of strong to weak binding. We calculate the number of positives for each allele, n_a, and we take the top n_a peptides from the ordering. The PPV for each allele, PPV_a is equal to $\frac{\#hits\ from\ top\ n_a\ peptides}{n_a}$, taking a value between 0 and 1. The maximum PPV_a value is equal to 1 when all top n_a peptides in the ranking are binders, while the minimum PPV_a value is 0 when all top n_a peptides in the ranking are decoys. In short, PPV shows the likelihood that a pHLA with a high predicted binding affinity is truly a strong binder. For FOOP, we calculate the predicted rank of binder peptides within the 500000 negative sampled decoys. The binding affinity is predicted for the whole dataset and the position is each binder is noted as its rank. As an example, a rank of $\leq 0.1\%$ is given to a peptide that is ranked within the first 500 decoys (0.1% of decoys), meaning that the peptide is a positive binder, and it is observed. FOOP is defined as the fraction of the positive pHLA instances that are predicted to bind in the top $\leq 0.1\%$ of all the 500000 decoys (percentile rank $\leq 0.1\%$). A higher number, closer to 1, means that the number of strong binding peptides that are observed is much higher, showing the robustness of the model in identifying those strong binding peptides and separating them from the rest of the decoys.

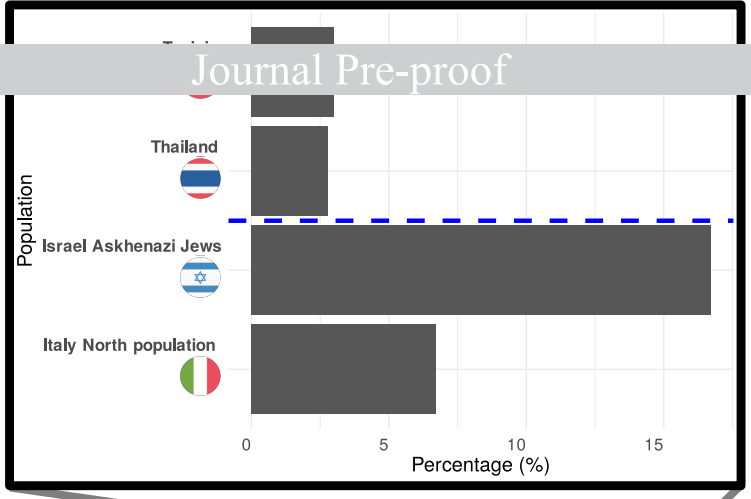
Statistical Significance

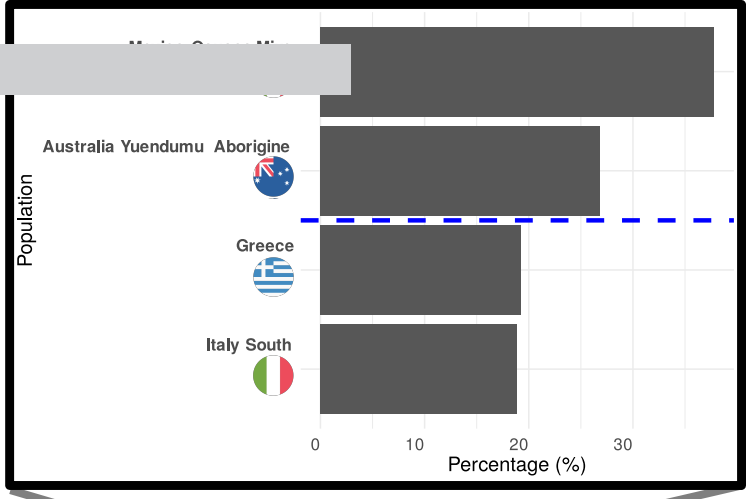
To evaluate the statistical significance of the differences of sPC90 coverage across population groups of different income levels (Figure 2) we use one-way ANOVA followed by the one-sided Kolmogorov-Smirnov (KS) test. We evaluate the KS statistic and the associated p values for all pairs of income levels (high, upper-middle, lower-middle, low) across all evaluated datasets (NetMHCpan4.1-BS).

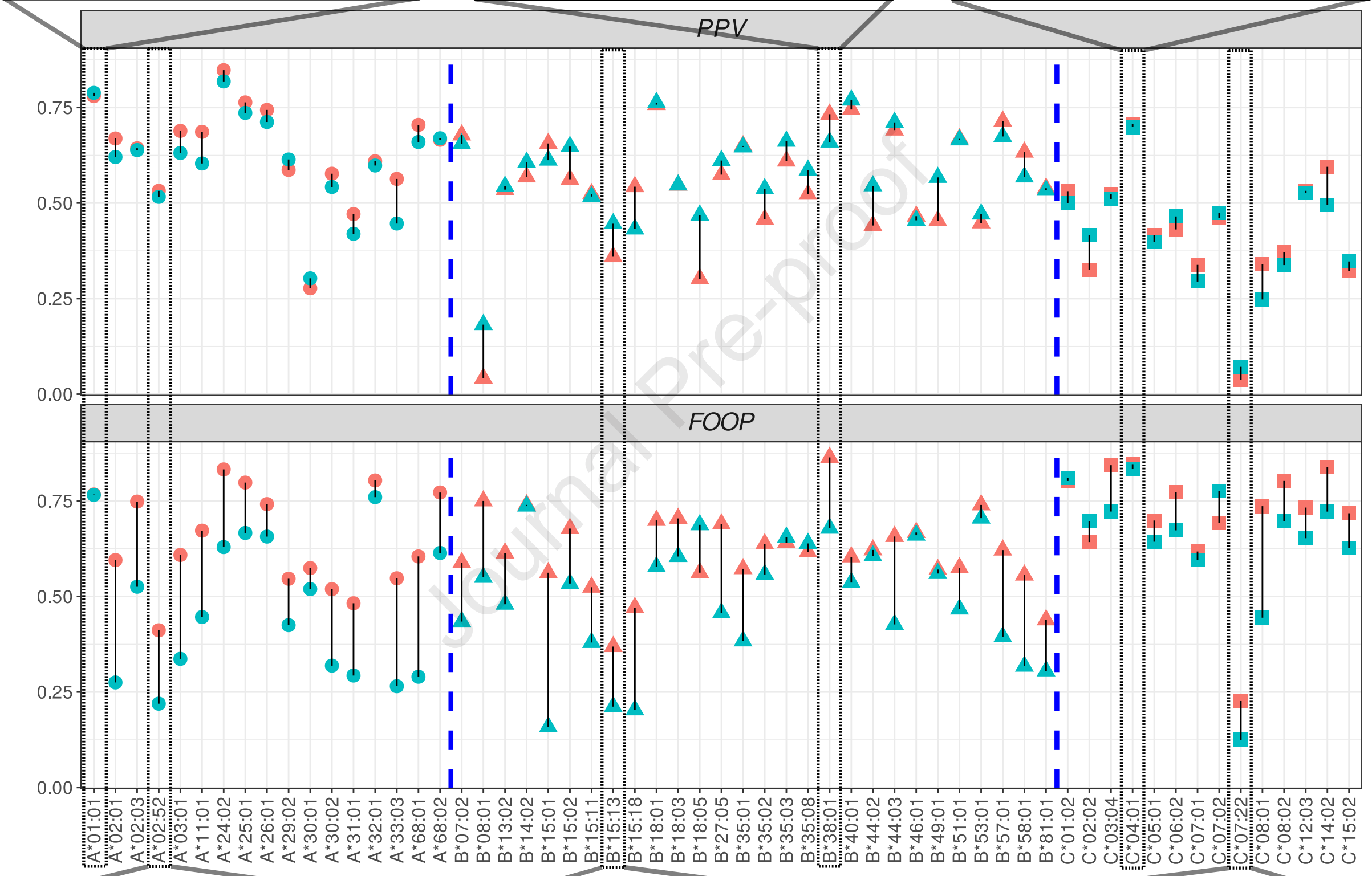

Journal Pre-proof
number of points in the dataset
25k
20k
15k
10k
5k
0
Dataset
MHCFlurry2.0_MS
NetMHCpan4.1_MS
NetMHCpan4.1_BA
MHCFlurry2.0_BA
Journal Pre-proof
A*02:01
A*03:01
A*68:01
A*01:01
A*11:01
A*29:02
A*24:02
A*02:07
A*68:02
A*23:01
A*34:02
A*02:05
A*31:01
A*02:04
A*02:02
A*32:01
A*33:03
A*11:02
A*02:03
A*30:02
A*74:01
A*26:01
A*36:01
A*02:06
A*33:01
A*30:01
A*34:01
A*02:11
A*66:01
A*25:01
A*24:07
A*02:20
A*03:02
A*69:01
A*02:12
A*02:16
A*02:19
A*24:03
A*80:01
HLA-A Alleles

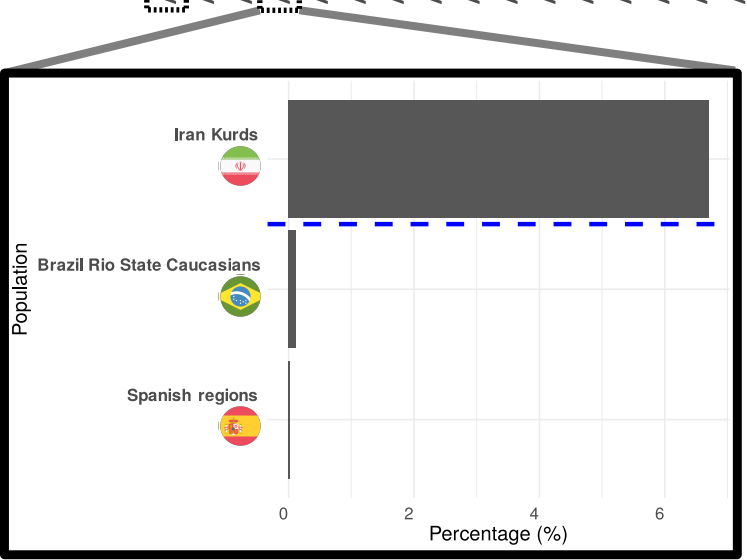


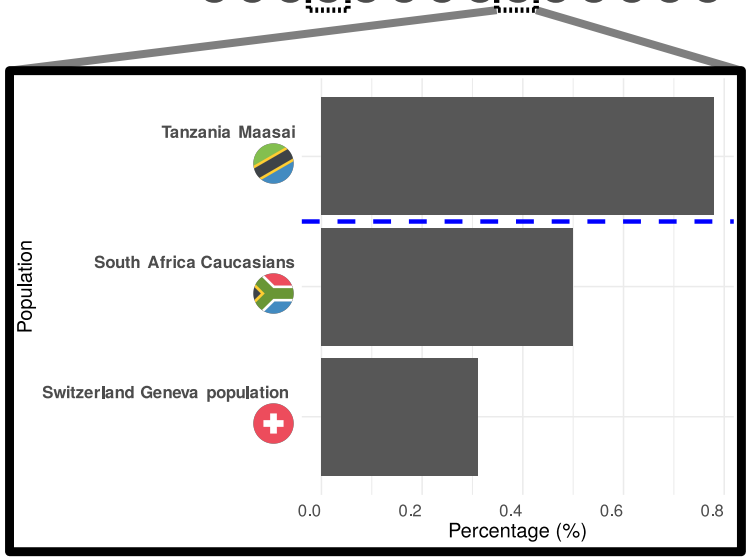


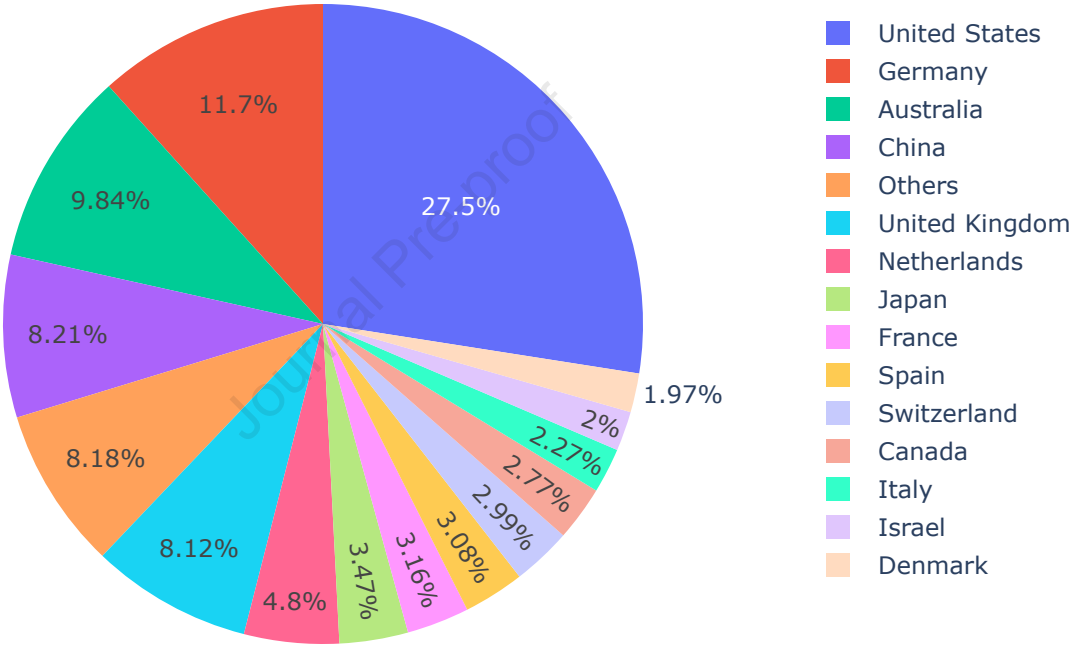












HIGHLIGHTS

- pHLA binding data has HLAs more common in high-income populations than low-income ones
- HLA bias in training data affects the pan-allele models' performance
- Pan-allele predictors have lower accuracy for HLAs not found in training datasets
- Pan-allele predictors have lower accuracy for HLAs expressed in lower-income populations