HLAequity: examining biases in pan-allele peptide-HLA binding predictors

Anja Conev*
Romanos Fasoulis*
Sarah Hall-Swan*
Computer Science Department, Rice University
Houston, Texas, USA

Rodrigo Ferreira†
Lydia E. Kavraki†
kavraki@rice.edu
Computer Science Department, Rice University
Houston, Texas, USA

ABSTRACT

Harnessing the immune response to combat disease has shown great promise. Personalized peptide vaccines target an important mechanism of cellular adaptive immunity in which the class I Human Leukocyte Antigen (HLA) proteins bind intracellular peptides. Peptide-HLA (pHLA) binding prediction is one of the crucial steps involved in finding peptide targets. Machine Learning (ML) pHLA binding prediction tools are trained on vast amounts of pHLA binding data. ML predictions are effective in guiding the search for therapeutic peptide targets. Most deployed ML models utilize neural network architectures and are reported to generalize to HLA alleles unseen during training ("pan-allele" models). However, the use of datasets with imbalanced allele content raises concerns about biased performance toward certain geographic populations. We examine the bias of two ML-based pan-allele pHLA binding affinity predictors. First, we examine the data bias and find that pHLA datasets unequally represent alleles from geographic populations in high-income countries as compared to those in low-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 underrepresented geographic populations. We aim to 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.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**.

KEYWORDS

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

1 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 [11].

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[1]. 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 [9, 16]. 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 [9, 16]. 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 [13]. Today, pan-allele prediction models are a significant component in immunotherapy pipelines. A recent survey identified 27 different methods for pHLA binding affinity prediction [22]; 20 out of 27 methods claim to be pan-allele while 17 out of 20 pan-allele methods utilize ML and neural network approaches. 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 [2]. Biased facial recognition software showed discrimination based on race with detrimental impacts in applications such as policing [4]. Decision-making algorithms deployed in crime prediction, credit lending, and hiring can perpetuate racial bias and injustice [1, 12]. 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 [15]. There are also issues related to ML bias in genomic-driven cancer treatments, as the predominant majority of sequenced patients in The Cancer Genome Atlas project are of European ancestry, while people with other ancestries are underrepresented [7].

Focusing on ML models in healthcare, Norori et al. categorized different perspectives of bias as human, data, and algorithmic bias [14]. 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

*These authors contributed equally.
†Corresponding authors.

[1]IPD-IMGT/HLA database https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics/ accessed May 2023

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 [14].

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.

2 DATASETS AND METHODS

2.1 Mapping HLA alleles to geographic populations and classifying them by income

We collect the distributions of alleles in different geographic populations from the Allele Frequency Net Database (AFND) [2] [8]. 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 [8]. 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". 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 (as described in the section below) 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 particular country[3]. 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).

2.2 Examining data bias

To examine data bias, we analyze training datasets from two predictors that are widely used in the literature: MHCFlurry2.0 [16] and NetMHCpan4.1 [18]. We choose these two state of the art tools as they are most widely used and cited. A recent comprehensive study [22] curated a list of pan-allele pHLA binding affinity predictors listed in Table SS1. We extract the number of citations of each of the tools from the Pubmed library and outline the number of citations in Table SS1. It is clear that MHCFlurry2.0 and NetMHCpan4.1 are most widely cited in addition to being recent.

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, as previously reported [17].

We calculate the population coverage using the method devised by Bui et al. and implemented in the Immune Epitope Database (IEDB) tools [3]. We use the AFND frequencies (see Section 2.1.) as ground truth allele frequencies of geographic populations. Population coverage has been used to estimate a portion of a population

[2]AFND database http://www.allelefrequencies.net/ accessed May 2023

[3]WorlBank information accessed at: https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023

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 *PC*90 metric calculated by the tool. *PC*90 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 *PC*90 by the dataset size to get the scaled *PC*90 (*sPC*90). A lower *sPC*90 indicates that 90% of individuals in this geographic population are represented by a small portion of the dataset. High values of *sPC*90 indicate that 90% of individuals in this geographic population are represented by a large portion of the dataset. Ideally, the *sPC*90 of a dataset should be high and equal across different geographic populations.

2.3 Examining algorithmic bias

2.3.1 Independent test set for assessing 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 [17], 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), as previously proposed [16, 17].

2.3.2 Algorithm evaluation 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) [16, 17]. 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.

3 RESULTS

3.1 Results of assessing data bias

First, we investigate the distribution of alleles in each of the training datasets (Fig S1, Fig S2, Fig S3). Note that MS datasets (blue) have more data than 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 have more than 5,000 data points in the datasets. For example, there is an overrepresentation of the alleles A*02:01 and A*03:01. Previous literature [20] 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 same source (IEDB), 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 Fig S1, Fig S2, and Fig S3 and for them, the blue lines in the plot diverge. In particular, MHCFlurry2.0_MS includes recently collected by Sarkizova et al. [20] targeting most of the human population and specifically some of the previously underrepresented alleles.

Next, we quantify how the allele content of each dataset relates to the allele contents of specific geographic populations (Fig 1). We calculate the scaled population coverage (*sPC*90) for each dataset across all geographic populations contained in AFND. We group the results based on the income level of the country of origin (Fig 1). The higher values of *sPC*90 indicate a better representation of the population within the dataset. We see a clear imbalance in terms of population coverage across different income levels. All datasets are biased towards the countries with higher income levels and on average they have higher *sPC*90 coverage for those populations. Note that the difference in *sPC*90 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. [20] 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 very 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 (Fig S4) we see that different ancestries are represented unequally.

3.2 Results of assessing algorithmic bias

We assess whether the notion of algorithmic bias, as defined by [14], exists in popular pan-allele pHLA binding prediction tools.

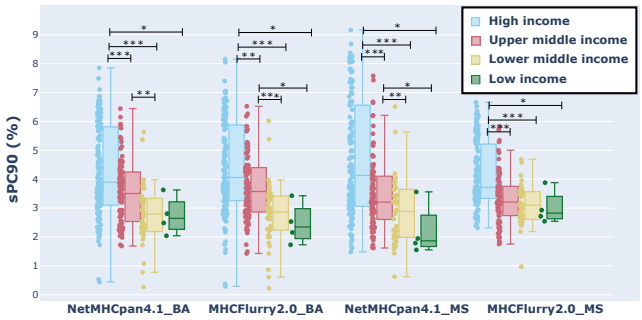

Figure 1: Scaled population coverage (sPC90) indicated on the y-axis of training datasets indicated on the x-axis. Each point corresponds to a particular geographic population and points are grouped based on the income level of the country of the population.

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 [17] (see Methods). Ideally, we would like to see NetMHCpan4.1 and MHCFlurry2.0 performing equally well on all HLAs (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 2. 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 [17], 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 [20], 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 Isreal 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.

4 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 (Figure S1). 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. This finding is quantified with the population coverage metric (sPC90) in Figure 1 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.

Note that we focus our analysis on the two highly regarded state-of-the-art prediction tools (NetMHCpan4.1 and MHCFlurry2.0). The rationale behind this selection is supported by Table S1 where these two tools emerge as the most widely cited among a range of other pan-allele predictors. 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. Nevertheless, 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 [22]. 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 Figure S1, Figure S2, Figure S3) while the MS datasets show an overlap across most of the alleles (see blue markers in Figure S1, Figure S2, Figure S3). The reported overlap enhances the representativeness of our analysis. Our findings are significant for the future development of medical treatments based on these datasets 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

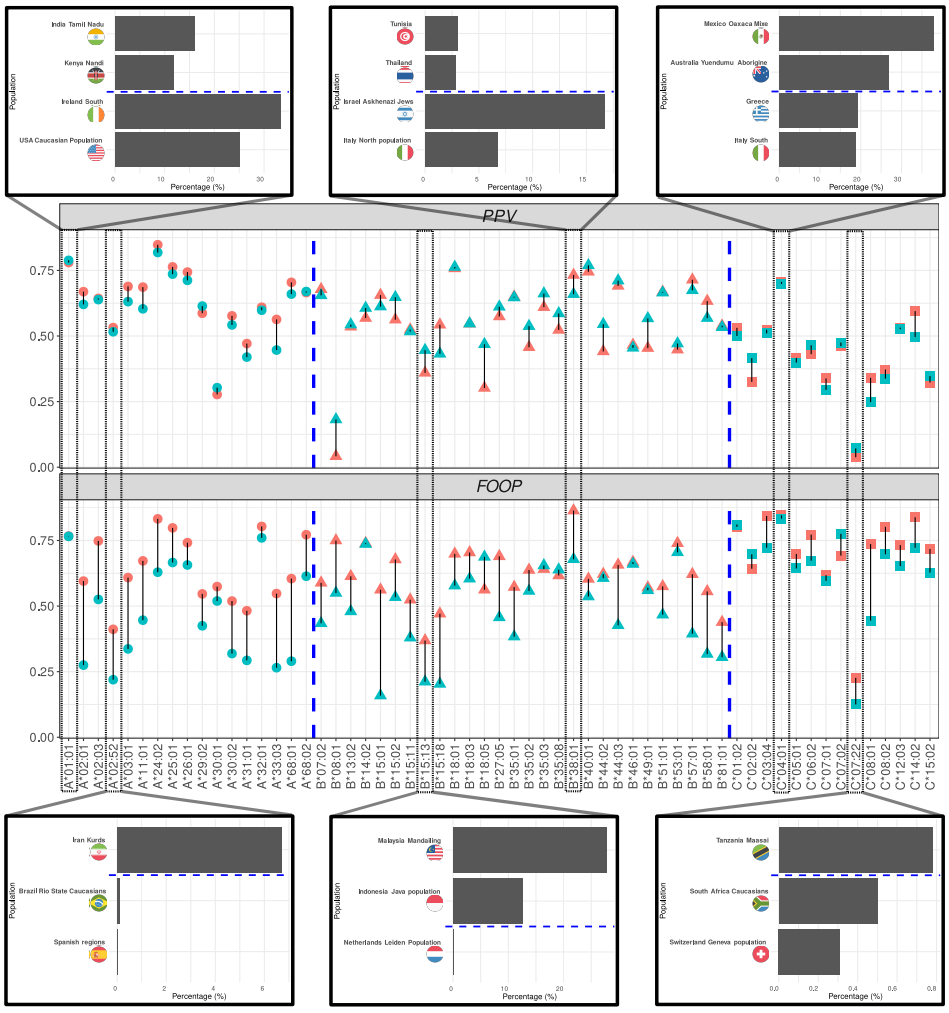

Figure 2: PPV (first row) and FOOP (second row) results for MHCFlurry2.0 (points in red) and NetMHCpan4.1 (points in blue). The x-axis corresponds to different alleles in the dataset. The y-axis corresponds to either the computed PPV value or the computed FOOP value. Different point shapes correspond 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.)

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 [4]. In the case of the HLA datasets, associated metadata shows that the collected samples were collected primarily from countries with higher levels of income. Figure S5 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.

There is an opportunity for further research, not only on differences between geographic populations in accordance to their income level but on differences within a single geographic population in accordance to 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 in section 2.1., it is composed of different *ethnogeographic* populations, such as USA European, USA Hispanic, USA African American, and USA Asian. As Figure S4 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 [5], as predicated on practices of colonialism and ethnic segregation [19].

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 [14]; 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 [21] and AFND [8], among others, are sharing pHLA data effectively. New pHLA binding prediction approaches are adopting explainability modules moving away from the black-box ML paradigm [6]. Another interesting avenue for allele bias mitigation is to train personalized, per-patient models [10]. However, more work can be done in regard to data collection, where very few studies sample binding affinities for alleles from different geographic populations [17, 20].

5 CONCLUSION

In this work, we identify the data and algorithmic bias of popular pan-allele pHLA binding prediction tools. The pan-allele model paradigm does not make up for this bias. We find that the models have unequal accuracy across alleles expressed in different geographic populations. We are worried that the uncritical use of the models will lead to perpetuating disparities in healthcare. It is important that we don't lose sight of the identified bias, and desist from using language that obscures differences in performance of the ML methods (i.e., "pan-allele" language).

6 DATA AVAILABILITY

The data and scripts used in our analysis are publicly available at https://github.com/KavrakiLab/HLAEquity .

REFERENCES

[1] I. Ajunwa. The Paradox of Automation as Anti-Bias Intervention. *Cardozo Law Review*, 41(5):1671–1742, 2019.

[2] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104(3):671–732, 2016. Publisher: California Law Review, Inc.

[3] H.-H. Bui, J. Sidney, K. Dinh, S. Southwood, M. J. Newman, and A. Sette. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 7(1):153, Mar. 2006.

[4] J. Buolamwini and T. Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, Jan. 2018. ISSN: 2640-3498.

[5] P. R. Center. On views of race and inequality, blacks and whites are worlds apart: Demographic trends and economic well-being. https://www.pewresearch.org/social-trends/2016/06/27/1-demographic-trends-and-economic-well-being/. Pew Research Center.

[6] Y. Chu, Y. Zhang, Q. Wang, L. Zhang, X. Wang, Y. Wang, D. R. Salahub, Q. Xu, J. Wang, X. Jiang, Y. Xiong, and D.-Q. Wei. A transformer-based model to predict peptide-HLA class i binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence*, 4(3):300–311, Mar. 2022.

[7] I. Dankwa-Mullan and D. Weeraratne. Artificial Intelligence and Machine Learning Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity. *Cancer Discovery*, 12(6):1423–1427, June 2022.

[8] F. F. Gonzalez-Galarza, A. McCabe, E. J. M. d. Santos, J. Jones, L. Takeshita, N. D. Ortega-Rivera, G. M. D. Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, D. Middleton, and A. R. Jones. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, 48(D1):D783–D788, Jan. 2020.

[9] I. Hoof, B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, S. Buus, and M. Nielsen. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, Jan. 2009.

[10] S. Liang, X. Jiang, Y. Chiu, H. Xu, K. H. Kim, G. Lizee, and K. Chen. An interpretable ML model to characterize patient-specific HLA-i antigen presentation. *bioRxiv*, Mar. 2023.

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

[12] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, Oct. 2016. Publisher: Oxford University Press / USA.

[13] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, and S. Buus. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLOS ONE*, 2(8):e796, Aug. 2007. Publisher: Public Library of Science.

[14] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2:100347, 10 2021.

[15] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, Oct. 2019. Publisher: American Association for the Advancement of Science.

[16] T. J. O'Donnell, A. Rubinsteyn, and U. Laserson. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Systems*, 11(1):42–48.e7, July 2020.

[17] R. M. Pyke, D. Mellacheruvu, S. Dea, C. Abbott, S. V. Zhang, N. A. Phillips, J. Harris, G. Bartha, S. Desai, R. McClory, J. West, M. P. Snyder, R. Chen, and S. M. Boyle. Precision neoantigen discovery using large-scale immunopeptidomes and composite modeling of mhc peptide presentation. *Molecular & Cellular Proteomics*, 22(4):100506, 2023.

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

[19] R. Rothstein. *The color of law: a forgotten history of how our government segregated America*. Liveright Publishing Corporation, a division of W.W. Norton & Company, New York ; London, 2017.

[20] S. Sarkizova, S. Klaeger, P. M. Le, L. W. Li, G. Oliveira, H. Keshishian, C. R. Hartigan, W. Zhang, D. A. Braun, K. L. Ligon, P. Bachireddy, I. K. Zervantonakis, J. M. Rosenbluth, T. Ouspenskaia, T. Law, S. Justesen, J. Stevens, W. J. Lane, T. Eisenhaure, G. Lan Zhang, K. R. Clauser, N. Hacohen, S. A. Carr, C. J. Wu, and D. B. Keskin. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nature Biotechnology*, 38(2):199–209, Feb. 2020.

[21] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 10 2018.

[22] M. Wang, L. Kurgan, and M. Li. A comprehensive assessment and comparison of tools for HLA class I peptide-binding prediction. *Briefings in Bioinformatics*, 24(3):bbad150, 04 2023.

A SUPPLEMENTARY MATERIAL

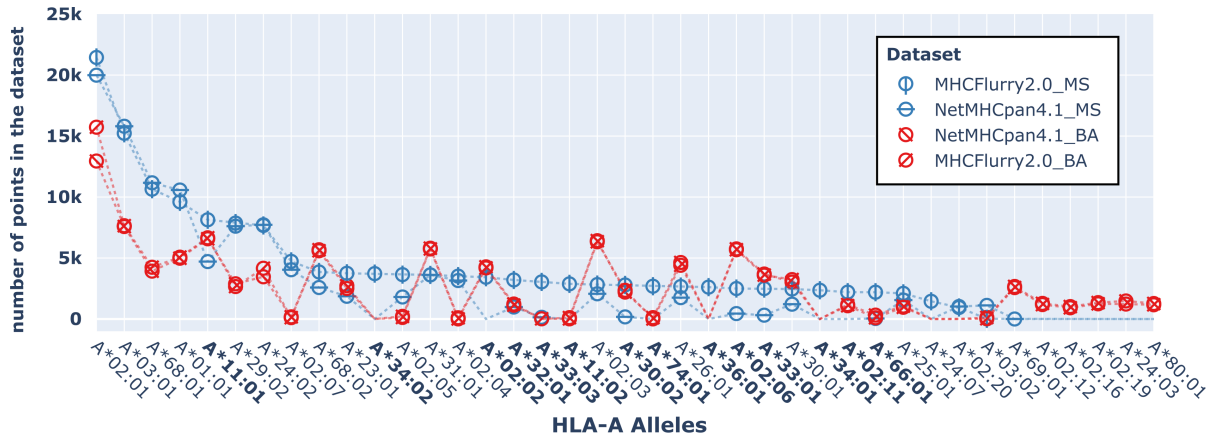

Figure S1: Analysis of the datasets. HLA-A allele frequencies in each of the training datasets (i.e., MHCFlurry2.0_BA, NetMHCpan4.1_BA, MHCFlurry2.0_MS, NetMHCpan4.1_MS). Allele codes are indicated on the x-axis while the number of points in the dataset for each allele is indicated on the y-axis. Allele 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.

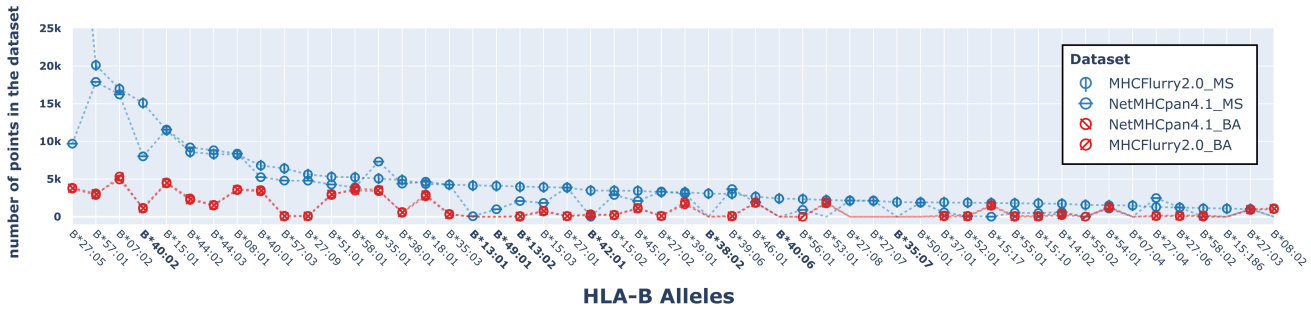

Figure S2: Analysis of the datasets. HLA-B allele frequencies in each of the training datasets (i.e., MHCFlurry2.0_BA, NetMHCpan4.1_BA, MHCFlurry2.0_MS, NetMHCpan4.1_MS). Allele codes are indicated on the x-axis while the number of points in the dataset for each allele is indicated on the y-axis. Allele 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.

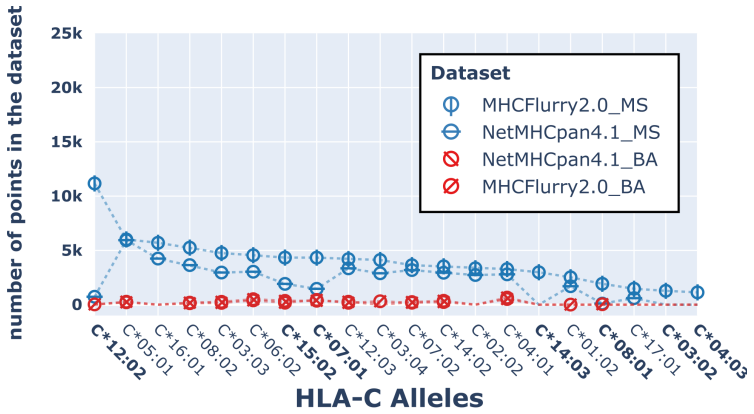

Figure S3: Analysis of the datasets. HLA-C allele frequencies in each of the training datasets (i.e., MHCFlurry2.0_BA, NetMHCpan4.1_BA, MHCFlurry2.0_MS, NetMHCpan4.1_MS). Allele codes are indicated on the x-axis while the number of points in the dataset for each allele is indicated on the y-axis. Allele 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.

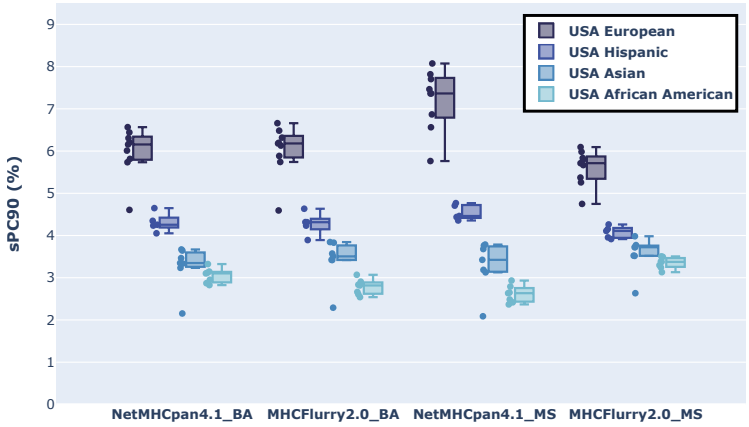

Figure S4: Scaled population coverage (sPC90) for the populations within the US of MHCFlurry2.0_BA, NetMHCpan4.1_BA, MHCFlurry2.0_MS and NetMHCpan4.1_MS training datasets. Datasets are indicated on the x-axis. sPC90 values are indicated on the y-axis for each US population across the datasets. Each point corresponds to a particular population within the US and points are grouped based on the population ancestry.

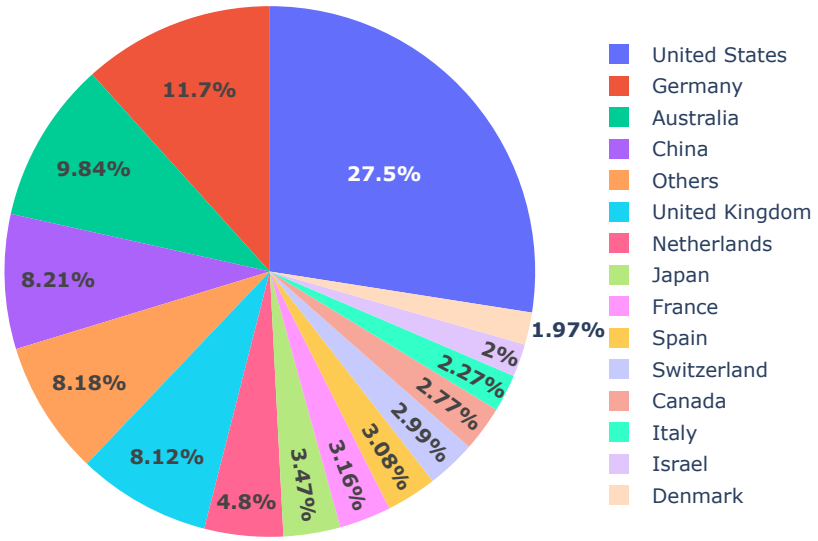

Figure S5: Country of origin of work curated by the IEDB

Predictor name	*Algorithm*	*Software available*	*Citations*	*Year*	*Reference*
NetMHCpan 4.1	**FFNN**	**Y**	**753**	**2020**	[1]
MHCflurry 2.0	**FFNN**	**Y**	**491**	**2018,2020**	[2, 3]
NetMHCcons	Consensus	N	340	2012	[4]
MixMHCpred	Scoring function	Y	314	2017,2018	[5, 6]
PickPocket	Scoring function	Y	198	2009	[7]
netMHCstabpan	FFNN	Y	154	2016	[8]
ConvMHC	CNN	Y	94	2017	[9]
DeepHLApan	GRU+Attention	Y	72	2019	[10]
PSSMHCpan	Scoring function	Y	64	2017	[11]
MHCSeqNet	GRU	Y	59	2019	[12]
ACME	CNN	Y	55	2019	[13]
DeepSeqPan	CNN	Y	53	2019	[14]
TransPHLA	Multi-head self-attention	Y	38	2022	[15]
Anthem	AODE	Y	30	2021	[16]
MHCAttnNet	LSTM+Attention	Y	26	2020	[17]
DeepAttentionPan	CNN+Attention	Y	12	2021	[18]
MATHLA	LSTM+Attention	Y	10	2021	[19]
HLAB	XGBoost, KNN, SVM, NB, LR, DTree, Bagging	Y	9	2022	[20]
DeepNetBim	CNN+Attention	Y	8	2021	[21]
Seq2Neo	CNN	Y	5	2022	[22]

Table S1: A comprehensive list of pan-allele pHLA binding affinity predictors curated by Wang et al. Number of citations is queried from the Pubmed library (July 2023).

SUPPLEMENTARY REFERENCES

[1] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, May 2020.

[2] Timothy J. O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B. Riemer, Uri Laserson, and Jeff Hammerbacher. MHCflurry: Open-source class i MHC binding affinity prediction. *Cell Systems*, 7(1):129–132.e4, July 2018.

[3] Timothy J. O'Donnell, Alex Rubinsteyn, and Uri Laserson. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Systems*, 11(1):42–48.e7, July 2020.

[4] Edita Karosiene, Claus Lundegaard, Ole Lund, and Morten Nielsen. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, 64(3):177–186, March 2012.

[5] Michal Bassani-Sternberg, Chloé Chong, Philippe Guillaume, Marthe Solleder, HuiSong Pak, Philippe O. Gannon, Lana E. Kandalaft, George Coukos, and David Gfeller. Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS computational biology*, 13(8):e1005725, August 2017.

[6] David Gfeller, Philippe Guillaume, Justine Michaux, Hui-Song Pak, Roy T. Daniel, Julien Racle, George Coukos, and Michal Bassani-Sternberg. The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands. *Journal of Immunology (Baltimore, Md.: 1950)*, 201(12):3705–3716, December 2018.

[7] Hao Zhang, Ole Lund, and Morten Nielsen. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics (Oxford, England)*, 25(10):1293–1299, May 2009.

[8] Michael Rasmussen, Emilio Fenoy, Mikkel Harndahl, Anne Bregnballe Kristensen, Ida Kallehauge Nielsen, Morten Nielsen, and Søren Buus. Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *Journal of Immunology (Baltimore, Md.: 1950)*, 197(4):1517–1524, August 2016.

[9] Youngmahn Han and Dongsup Kim. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC bioinformatics*, 18(1):585, December 2017.

[10] Jingcheng Wu, Wenzhe Wang, Jiucheng Zhang, Binbin Zhou, Wenyi Zhao, Zhixi Su, Xun Gu, Jian Wu, Zhan Zhou, and Shuqing Chen. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Frontiers in Immunology*, 10:2559, 2019.

[11] Geng Liu, Dongli Li, Zhang Li, Si Qiu, Wenhui Li, Cheng-Chi Chao, Naibo Yang, Handong Li, Zhen Cheng, Xin Song, Le Cheng, Xiuqing Zhang, Jian Wang, Huanming Yang, Kun Ma, Yong Hou, and Bo Li. PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *GigaScience*, 6(5):1–11, May 2017.

[12] Poomarin Phloyphisut, Natapol Pornputtapong, Sira Sriswasdi, and Ekapol Chuangsuwanich. MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC bioinformatics*, 20(1):270, May 2019.

[13] Yan Hu, Ziqiang Wang, Hailin Hu, Fangping Wan, Lin Chen, Yuanpeng Xiong, Xiaoxia Wang, Dan Zhao, Weiren Huang, and Jianyang Zeng. ACME: pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics (Oxford, England)*, 35(23):4946–4954, December 2019.

[14] Zhonghao Liu, Yuxin Cui, Zheng Xiong, Alierza Nasiri, Ansi Zhang, and Jianjun Hu. DeepSeqPan, a novel deep convolutional neural network model for pan-specific class I HLA-peptide binding affinity prediction. *Scientific Reports*, 9(1):794, January 2019.

[15] Yanyi Chu, Yan Zhang, Qiankun Wang, Lingfeng Zhang, Xuhong Wang, Yanjing Wang, Dennis Russell Salahub, Qin Xu, Jianmin Wang, Xue Jiang, Yi Xiong, and Dong-Qing Wei. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nature Machine Intelligence*, 4(3):300–311, March 2022. Number: 3 Publisher: Nature Publishing Group.

[16] Shutao Mei, Fuyi Li, Dongxu Xiang, Rochelle Ayala, Pouya Faridi, Geoffrey I. Webb, Patricia T. Illing, Jamie Rossjohn, Tatsuya Akutsu, Nathan P. Croft, Anthony W. Purcell, and Jiangning Song. Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Briefings in Bioinformatics*, 22(5):bbaa415, September 2021.

[17] Gopalakrishnan Venkatesh, Aayush Grover, G. Srinivasaraghavan, and Shrisha Rao. MHCAttnNet: predicting MHC-peptide bindings for MHC alleles classes I and II using an attention-based deep neural model. *Bioinformatics (Oxford, England)*, 36(Suppl_1):i399–i406, July 2020.

[18] Jing Jin, Zhonghao Liu, Alireza Nasiri, Yuxin Cui, Stephen-Yves Louis, Ansi Zhang, Yong Zhao, and Jianjun Hu. Deep learning pan-specific model for interpretable MHC-I peptide binding prediction with improved attention mechanism. *Proteins*, 89(7):866–883, July 2021.

[19] Yilin Ye, Jian Wang, Yunwan Xu, Yi Wang, Youdong Pan, Qi Song, Xing Liu, and Ji Wan. MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. *BMC bioinformatics*, 22(1):7, January 2021.

[20] Yaqi Zhang, Gancheng Zhu, Kewei Li, Fei Li, Lan Huang, Meiyu Duan, and Fengfeng Zhou. HLAB: learning the BiLSTM features from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction. *Briefings in Bioinformatics*, 23(5):bbac173, September 2022.

[21] Xiaoyun Yang, Liyuan Zhao, Fang Wei, and Jing Li. DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. *BMC bioinformatics*, 22(1):231, May 2021.

[22] Kaixuan Diao, Jing Chen, Tao Wu, Xuan Wang, Guangshuai Wang, Xiaoqin Sun, Xiangyu Zhao, Chenxu Wu, Jinyu Wang, Huizi Yao, Casimiro Gerarduzzi, and Xue-Song Liu. Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. *International Journal of Molecular Sciences*, 23(19):11624, October 2022.