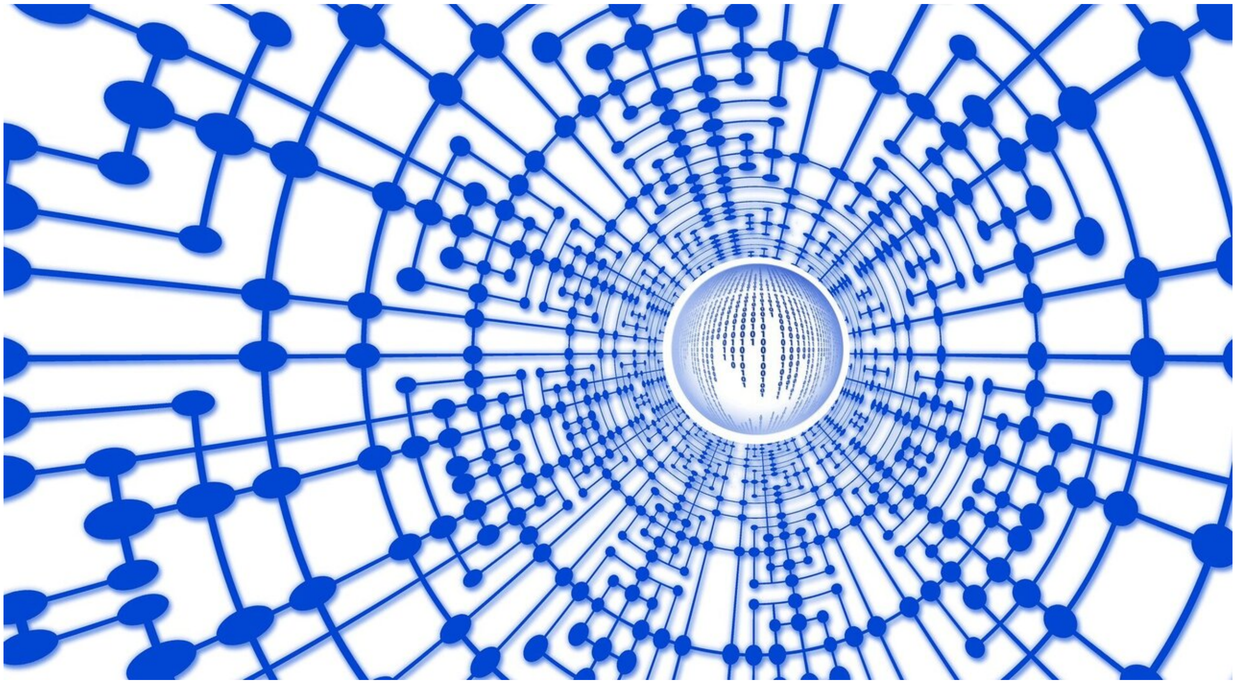
PHYS.ORG

Widely used machine learning models reproduce dataset bias: Study

February 18 2024, by John Bogna



Credit: CC0 Public Domain

Rice University computer science researchers have found bias in widely used machine learning tools used for immunotherapy research.

Ph.D. students Anja Conev, Romanos Fasoulis and Sarah Hall-Swan, working with computer science faculty members Rodrigo Ferreira and Lydia Kavraki, reviewed publicly available peptide-HLA (pHLA) binding prediction data and found it to be skewed toward higher-income communities. Their paper examines the way that biased data input affects the algorithmic recommendations being used in important immunotherapy research.

Peptide-HLA binding prediction, machine learning and immunotherapy

HLA is a gene in all humans that encodes proteins working as part of our immune response. Those proteins bind with protein chunks called peptides in our cells and mark our infected cells for the body's immune system, so it can respond and, ideally, eliminate the threat.

Different people have slightly different variants in genes, called alleles. Current immunotherapy research is exploring ways to identify peptides that can more effectively bind with the HLA alleles of the patient.

The end result, eventually, could be custom and highly effective immunotherapies. That is why one of the most critical steps is to accurately predict which peptides will bind with which alleles. The greater the accuracy, the better the potential efficacy of the therapy.

But calculating how effectively a peptide will bind to the HLA allele takes a lot of work, which is why machine learning tools are being used to predict binding. This is where Rice's team found a problem: The data used to train those models appears to geographically favor higher-income communities.

Why is this an issue? Without being able to account for genetic data from lower-income communities, future immunotherapies developed for them may not be as effective.

"Each and every one of us has different HLAs that they express, and those HLAs vary between different populations," Fasoulis said. "Given that machine learning is used to identify potential peptide candidates for immunotherapies, if you basically have biased machine models, then those therapeutics won't work equally for everyone in every population."

Redefining 'pan-allele' binding predictors

Regardless of the application, machine learning models are only as good as the data you feed them. A bias in the data, even an unconscious one, can affect the conclusions made by the algorithm.

Machine learning models currently being used for pHLA binding prediction assert that they can extrapolate for allele data not present in the dataset those models were trained on, calling themselves "pan-allele" or "all-allele." The Rice team's findings call that into question.

"What we are trying to show here and kind of debunk is the idea of the 'pan-allele' machine learning predictors," Conev said. "We wanted to see if they really worked for the data that is not in the datasets, which is the data from lower-income populations."

Fasoulis' and Conev's group tested publicly available data on pHLA binding prediction, and their findings supported their hypothesis that a bias in the data was creating an accompanying bias in the algorithm. The team hopes that by bringing this discrepancy to the attention of the research community, a truly pan-allele method of predicting pHLA binding can be developed.

Ferreira, faculty advisor and paper co-author, explained that the problem of bias in machine learning can't be addressed unless researchers think about their data in a social context. From a certain perspective, datasets may appear as simply "incomplete," but making connections between what is or what is not represented in the dataset and underlying historical and economic factors affecting the populations from which data was collected is key to identifying bias.

"Researchers using machine learning models sometimes innocently assume that these models may appropriately represent a global population," Ferreira said, "but our research points to the significance of when this is not the case." He added that "even though the databases we studied contain information from people in multiple regions of the world, that does not make them universal. What our research found was a correlation between the socioeconomic standing of certain populations and how well they were represented in the databases or not."

Professor Kavraki echoed this sentiment, emphasizing how important it is that tools used in clinical work be accurate and honest about any shortcomings they may have.

"Our study of pHLA binding is in the context of personalized immunotherapies for cancer—a project done in collaboration with MD Anderson," Kavraki said. "The tools developed eventually make their way to clinical pipelines. We need to understand the biases that may exist in these tools. Our work also aims to alert the research community on the difficulties of obtaining unbiased datasets."

Conev noted that, though biased, the fact that the data was publicly available for her team to review was a good start. The team is hoping its findings will lead new research in a positive direction—one that includes and helps people across demographic lines.

The paper is published in the journal *iScience*.

More information: Anja Conev et al, HLAEquity: Examining biases in pan-allele peptide-HLA binding predictors, *iScience* (2023). DOI:

10.1016/j.isci.2023.108613

Provided by Rice University

Citation: Widely used machine learning models reproduce dataset bias: Study (2024, February 18) retrieved 29 April 2024 from https://phys.org/news/2024-02-widely-machine-dataset-bias.html