The pitfalls of negative data bias for the T-cell epitope specificity challenge

Reference:
Dens Ceder, Laukens Kris, Bittremieux Wout, Meysman Pieter.- The pitfalls of negative data bias for the T-cell epitope specificity challenge
Nature Machine Intelligence - ISSN 2522-5839 - 5:10(2023), p. 1060-1062
Full text (Publisher's DOI): https://doi.org/10.1038/S42256-023-00727-0
To cite this reference: https://hdl.handle.net/10067/1994840151162165141

The pitfalls of negative data bias for the T-cell epitope specificity challenge

Ceder Dens[a,b], ceder.dens@uantwerpen.be,
Kris Laukens[a,b], kris.laukens@uantwerpen.be
Wout Bittremieux[a*,], wout.bittremieux@uantwerpen.be
Pieter Meysman[a,b*] *(corresponding author)*, pieter.meysman@uantwerpen.be

[a]Adrem Data Lab, Department of Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium
[b]AUDACIS consortium, University of Antwerp, Middelheimlaan 1, 2020 Antwerpen, Belgium
*These authors contributed equally.

ARISING FROM: Gao, Y., Gao, Y., Fan, Y. et al. Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. Nat Mach Intell 5, 236–249 (2023). https://doi.org/10.1038/s42256-023-00619-3

Recently, Gao et al.[1] introduced a combination of meta-learning and the neural Turing machine to tackle a very important but yet unsolved problem in immunology: the TCR–epitope binding prediction challenge for novel epitopes. All high-performing machine learning models can have problems when deployed in a real-world setting if the data used to train and test the model contains biases. In this article, we describe how the technique used to create negative data for the TCR–epitope interaction prediction task can lead to a strong bias and makes that the performance drops to random when tested in a more realistic scenario.

Unexpected or unknown biases within machine learning datasets are a common issue that has hindered many well-designed approaches from translating to real-world applications, despite seemingly generalizable performance achieved during model development and evaluation. A well-known example of this issue is a classifier that was trained to identify malignant skin lesions, but ended up relying on the presence of a measuring ruler in the images due to the bias present in the training data[2]. However, the presence of data bias is not always obvious. Multiple cases have been reported where specific demographics, such as gender, skin type, ethnicity, or socio-economic status, were underrepresented in the data, leading to unexpected performance differences between different subpopulations and potentially delaying access to care[3]. Indeed, as algorithmic approaches become increasingly more advanced and datasets grow larger and are necessarily compiled using less curation, these issues are becoming more and more commonplace. Even small biases within a dataset often suffice for a machine learning model to overfit on bogus data characteristics and drive its predictive behavior. Crucially, if the same bias persists in any held-out test data, this issue will remain undetected. One such bias, as will be described in this article, is caused by a confounding factor linked to the input data and prediction label, causing shortcut learning[4–6].

The T-cell epitope prediction challenge, as recently tackled by Gao et al.[1], involves computationally identifying the target epitope of T-cells using their T-cell receptor (TCR) sequence. T-cells are a critical part of the adaptive immune system, as they recognize intruders from self, induce immune responses, and retain memory. When antigen-presenting cells display short peptides (called epitopes) from pathogens or malignant cells, such as cancer cells, on their cell surface, this TCR is able to bind with them in a specific manner, upon which the T-cell will be activated and the immune response will be triggered.

If we would be able to annotate TCR sequences with their targets, this would unlock myriad applications, ranging from vaccine design and cancer treatments to diagnostics. However, the number of possible TCR sequences is incredibly large, with a conservative estimate in the range of 10^{15} unique sequences[7]. Consequently, the epitope targets of the vast majority of TCRs are unknown. On the other hand, it is known that the specificity of a T-cell is fully driven by its TCR and its static co-receptors[8]. Therefore, the entire recognition event must be encoded within the TCR sequence and is seemingly a straight-forward prediction problem where the right TCR has to be matched with the right target.

Several methods have shown significant potential in extrapolating from a set of TCRs known to bind a specific epitope, to other TCRs targeting the same epitope[9]. However, the number of epitopes with known TCRs is counted in the hundreds, which is just a drop in the ocean of possible TCR targets. Therefore, zero-shot TCR–epitope annotation—i.e. predicting TCR–epitope binding for novel, unseen epitopes—is currently seen as the 'holy grail' of immunology[7]. This requires machine learning methods to actually learn the underlying recognition code of the TCRs, which has turned out to be a substantially harder problem. We can define the unseen epitope–TCR prediction task as: predict the probability that a given TCR sequence will recognize a given epitope sequence, with the condition that the epitope sequence was not yet seen by the prediction model.

An important issue that complicates this challenge is the lack of high-quality negative data. While the experimental methods to determine TCR–epitope pairs have a high specificity, they are hindered by a low sensitivity with a high false negative rate[10]. As a result, the number of true negative pairs in TCR–epitope databases is a small fraction of the known positive pairs. Consequently, this is often dealt with as a positive and unlabeled data learning problem,[11] where presumed non-positive instances are generated by artificially pairing TCR and epitope sequences as a stand-in for true negative data.

There are two approaches commonly used for generating negative data in the context of TCR–epitope annotation (Fig. 1). The first is shuffling the known positive pairs, where each TCR is matched with an epitope to create random combinations that differ from those in the positive data. This relies on the principle that a TCR known to be specific for one epitope is unlikely to be specific for another unrelated epitope. However, because of the limited number of epitopes with known TCRs, it is complex to design a held-out negative dataset using this approach. The second strategy, applied by Gao et al.[1], is using background TCR data. In this case, epitopes from known positive samples are paired with random TCRs from a background set, which is often obtained from a broad sequencing experiment without epitope specificity. These strategies for generating negative data are a poor approximation of the real-world scenario, as they both have the potential to create false negative pairs. For the first strategy, this can be caused by cross-reactive TCRs, which bind to more than one epitope. For the second strategy, the background TCRs might bind the epitope it is paired with, which is a substantial risk as many epitopes with known TCRs are immunodominant with prevalent high frequency clones. Irrespective of how the training dataset is generated, any model claiming to capture TCR–epitope recognition rules should be performant on test data generated by either strategy, otherwise it could be guilty of shortcut learning[6].

Indeed, multiple studies have shown that shortcut learning is an inherent risk with the second approach. It introduces an artificial confounding factor between the TCR sequence and the prediction label because positive and negative samples have TCR sequences drawn from a different background distribution, irrespective of the target epitope[4,5]. One study used a decoy dataset that removed any chance of true binding[12]. When generating negatives with the second strategy, a performance better than random was achieved for the decoys, demonstrating that the background TCR data contained a bias that caused a difference between positive and negative CDR3 sequences independent of the paired epitope. Similarly, a second study showed that using a background data set to generate negatives leads to sequence memorization and making predictions only based on the CDR3 sequence, without considering the epitope[13]. The cause of these problems is that the negative pairs and positive pairs are derived from different experiments, performed by different labs, and often even in a different part of the world with different subject ethnicities. Any high-performance machine learning method will exploit this dataset shortcut and utilize it to differentiate between positive and negative samples, and consequently suffer from unexpected generalization failures[6].

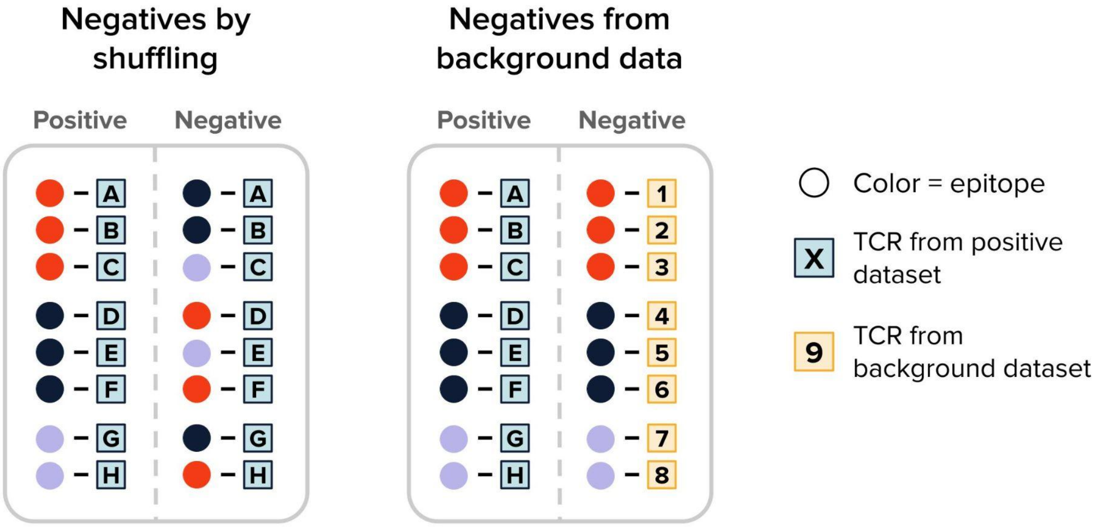

Figure 1. Schematic overview of the two approaches commonly used for generating negative TCR–epitope data. When generating negatives by shuffling (left), the same epitopes and TCR are reused but each TCR is paired with a different epitope. When generating negatives from a background dataset (right), new TCR sequences are paired with the epitopes.

To determine the potential impact of the negative set, we first tested the zero-shot predictions of PanPep using five-fold cross-validation with data generated using the shuffled epitope approach instead of the background TCR approach[12]. PanPep achieved an area under the receiver operating characteristic curve (ROC-AUC) of 54.1% ± 6.4% (mean ± standard deviation) (Fig. 2a), similar to the previously reported ROC-AUC of 54.1% ± 1.9% on this dataset[12]. Note, however, that we did not filter the data to exclude samples or epitope sequences already present in the PanPep training data. Consequently, 57.7% of the positive test samples were part of PanPeps training data and only 3.1% of the test samples had an epitope not seen during training (see Supplementary Material). As such, although this should have been a relatively easy test, the performance on data with negatives generated by shuffling significantly underperforms the zero-shot ROC-AUC of 70.8% reported originally.

Second, we tested PanPep in a true zero-shot setting by using the PanPep zero-shot positive data and generating negative data by shuffling the TCR sequences of these samples. The result is a test dataset that does not contain any samples and epitope sequences already included in the training dataset. On this test dataset, PanPep achieves a ROC-AUC of 49.2% (Fig. 2b), failing to make predictions better than random.

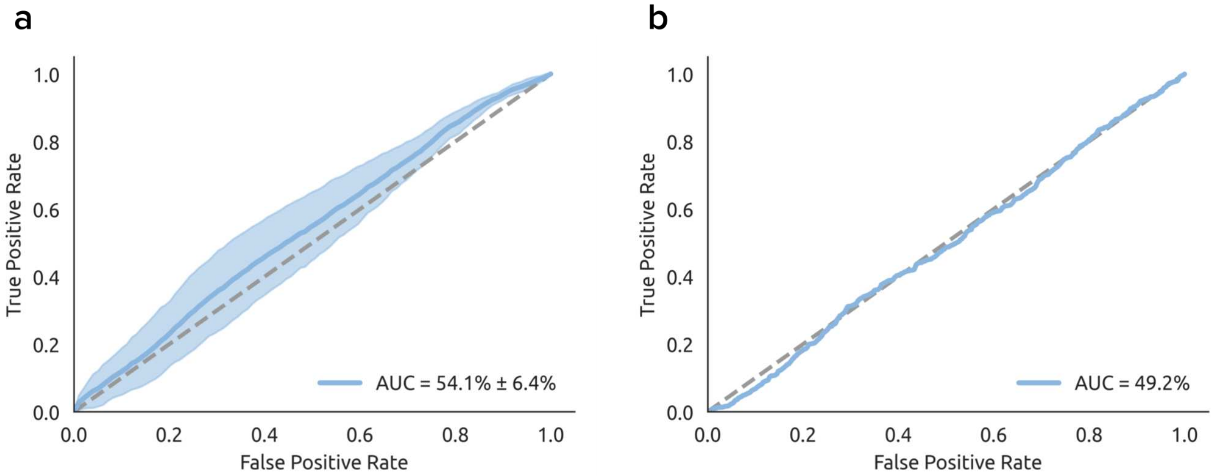

Figure 2. ROC curves of PanPep tested on shuffled negative data. (a) Mean ROC curve and standard deviation of PanPep from five-fold cross-validation with data generated through the shuffled epitope approach. The data was not filtered to exclude samples or epitope sequences already present in the PanPep training data. **(b)** ROC curve of PanPep on zero-shot data with negatives generated by shuffling.

A lack of unbiased labeled data is not unique to the TCR–epitope prediction problem. Similar issues exist within many other fields, such as for anomaly detection, where rare events by definition only occur infrequently[14], and for a broadly used benchmarking dataset of protein–ligand binding prediction that contains a bias in the negative data which makes it easy to distinguish between decoys and binding pairs[15].

In conclusion, biased data can and will lead to inaccurate and untrustworthy predictions for any machine learning task. This is also the case for TCR–epitope prediction tools trained on biased negative data, where unrealistic performances are achieved due to shortcut learning, which would not occur in a more realistic setting. Given the potential advances in healthcare that would arise from accurate TCR–epitope binding prediction tools, we argue that more effort needs to go towards this problem. More data and an unbiased benchmarking dataset are a necessary next step towards prediction models that are reliable in real-world scenarios.

Competing Interests

KL and PM hold shares in ImmuneWatch BV, an immunoinformatics company.

Author Contributions

CD performed the study. CD and PM wrote the manuscript. WB, KL, and PM conceived and supervised the study. WB, PM and KL revised the manuscript. All authors read and approved the final manuscript.

Data Availability

The data used to obtain the results is available on GitHub at https://github.com/PigeonMark/PanPep-Shuffled-Negatives and on Zenodo at https://doi.org/10.5281/zenodo.7798691.

Code Availability

All scripts used to obtain the results are available on GitHub at https://github.com/PigeonMark/PanPep-Shuffled-Negatives and on Zenodo at https://doi.org/10.5281/zenodo.7798691.

Materials & Correspondence

Correspondence to Pieter Meysman.

References

1. Gao, Y. *et al.* Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nat. Mach. Intell.* 1–14 (2023) doi:10.1038/s42256-023-00619-3.
2. Narla, A., Kuprel, B., Sarin, K., Novoa, R. & Ko, J. Automated Classification of Skin Lesions: From Pixels to Practice. *J. Invest. Dermatol.* **138**, 2108–2110 (2018).
3. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M.

Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).

4. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 3673 (2020).
5. Pavlović, M. *et al.* Improving generalization of machine learning-identified biomarkers with causal modeling: an investigation into immune receptor diagnostics. Preprint at https://doi.org/10.48550/arXiv.2204.09291 (2023).
6. Geirhos, R. *et al.* Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
7. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* 1–11 (2023) doi:10.1038/s41577-023-00835-3.
8. Krogsgaard, M. & Davis, M. M. How T cells 'see' antigen. *Nat. Immunol.* **6**, 239–245 (2005).
9. Meysman, P. *et al.* Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics* **9**, (2023).
10. Zhang, W. *et al.* A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Sci. Adv.* **7**, eabf5835 (2021).
11. Bekker, J. & Davis, J. Learning from positive and unlabeled data: a survey. *Mach. Learn.* **109**, 719–760 (2020).
12. Moris, P. *et al.* Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief. Bioinform.* **22**, bbaa318 (2021).
13. Grazioli, F. *et al.* On TCR binding predictors failing to generalize to unseen peptides. *Front. Immunol.* **13**, (2022).
14. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **41**, 15:1-15:58 (2009).
15. Chen, L. *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of

deep learning in structure-based virtual screening. *PLOS ONE* **14**, e0220113 (2019).