Tricked by Edge Cases: Can Current Approaches Lead to Accurate Prediction of T-Cell Specificity with Machine Learning?

Martin Culka[1] and Darya Orlova[2#]

[1] Department of Systems Biology, Columbia University, New York, New York, 10032, USA

[2] Cell Signaling Technology, Danvers, Massachusetts, 01915, USA

[#] To whom correspondence should be addressed: Darya Orlova, darya.orlova@cellsignal.com

Abstract

The ability to predict T-cell receptor (TCR) specificity computationally could revolutionize personalized immunotherapies, vaccine development, and the understanding of immunology and autoimmune diseases. While progress depends on obtaining training data that represent the vast range of possible TCR-ligand pairs, systematic assessment of modeling assumptions is equally important and can begin with existing data. We illustrate this by evaluating two ideas currently present in the field[1,2]: treating TCR specificity and T cell activation as distinct modeling tasks, and using unsupervised models based on sequence similarity for TCR specificity prediction. Although presented as general strategies, we argue these are exceptions rather than universally applicable principles.

Results and discussion

A few decades ago, multimer-binding technology revolutionized immunology[3] by enabling researchers to directly identify and quantify antigen-specific T cells. Public databases such as IEDB (https://www.iedb.org/) and VDJdb (https://vdjdb.cdr3.net/) are largely composed of data generated during this "multimer-binding era". While this technology remains invaluable in certain scenarios, recent insights have revealed its limitations. Over time, multimer-binding technology has biased research toward predominantly higher-affinity TCRs that are not necessarily specific. This bias is evident from two key observations: the growing number of empirical studies revealing a disconnect between multimer-binding capacity and T cell activation (Supplementary Table 1, Figure 1) and the realization that, while we still lack a clear biophysical definition of TCR specificity, it is evident that affinity alone cannot serve as that measure[4-6].

As a result, assays based on the binding strength measurements of TCR-pMHC interactions at equilibrium, when decoupled from T cell activation, cannot accurately serve as TCR specificity assays. From a machine learning perspective, training data generated from such binding assays are likely contaminated with both false positives and false negatives (Figure 1), making it challenging to separate the task of predicting TCR specificity from T cell activation. Until we achieve a clearer biophysical definition of TCR specificity, it is safer to rely on training data where binding assays are coupled with T cell functional readouts.

Initial success with multimer-binding assays in identifying antigen-specific T cells led to the hypothesis that TCRs with similar sequences, or dominant binding motifs[7], likely recognize the same pMHC. This assumption spurred the development of unsupervised machine learning models that use TCR sequence similarity as a proxy for specificity.

Recent studies[8] that claim these approaches are effective classifiers, despite reporting low discriminatory power, underscore the need for thorough evaluation. Systematic assessments of previously reported unsupervised modeling data[9], as we show (Figure 2a), suggest that the overall utility of published clustering methods for predicting TCR specificity is questionable. Across methods, only a minority of TCRs are grouped into pure clusters predominantly containing peptide-specific TCRs (see Figure 4 and SI in Leary et al.[9]).

Further evaluation (Figure 2b and Supplementary Figure 1) using datasets from the IMMREP_2022 benchmark[10] reveals that while binding motifs are present in some cases (4 out of 17 peptide-specific TCR repertoires), they are not a reliable feature for general specificity inference. Even TCRs explicitly containing the predominant motif sequence are distributed among multiple (though fewer) clusters (Figure 2b, inset). Although the presence of binding motifs can aid modeling in specific cases, it does not hold as a general principle (see Figure 6 and Supplementary Figure 11 in Simpson et al.[11]). TCRs recognizing different antigens are often more sequence-similar than those targeting the same epitope, both in the ESM2 latent space and with direct sequence similarity measures like plain BLOSUM62 pairwise similarity (Supplementary Figure 1) or the tcrdist3 distance measure[12,13] (Figure 2b). However, in simpler peptide-specific contexts, distance-based clustering performs comparably to supervised methods.[10]

This underscores the need for a clearer understanding of TCR specificity and identification of reliable features from 1D sequences and/or 3D structures (experimental or computer-predicted) that can serve as similarity measures in unsupervised settings. Until such clarity is reached, supervised models remain the preferred approach for specificity prediction. Although generalized predictions are still limited by the amount of available training data, supervised modeling has shown promise in peptide-specific contexts[10].

Materials and methods

Data overview

For the assessment of the peptide-specific predictive power of various clustering methods (Figure 2a), we used data from the Source Data table in Leary et al[9].

For the assessment of peptide specificity assignment using agglomerative clustering (Figure 2b), based on various distance metrics, we used the IMMREP_2022 benchmark data set (Meysman et al.[10], available at https://github.com/viragbioinfo/IMMREP_2022_TCRSpecificity), which contains 17

peptide-specific data buckets. For our clustering analysis, we used concatenated training set data, organized epitope-wise.

Data analysis

For Figure 2a, we plotted a subset of an already published analysis (Leary et al.[9]). Specifically, the data can be found in *fig4cd_sup_fig_1to7* tab of the Source Data table. We selected a subset of data points with minimal cluster size 3, without a spike of irrelevant data (*spike_x* = 0), *tcrvalid* reference labels, and clustering based on TCRβ (*TRB*). For the methods where it is defined (all except GLIPH and clustcr), we selected data points for the distance parameter ε=0.5 (or the closest value available). For each method, we plotted the value in the c-CSI column (clustering Critical Success Index).

For Figure 2b and Supplementary Figure 1, we clustered aggregated peptide-specific TCRs using agglomerative (hierarchical) clustering with the *scikit-learn*[14] Python library. We used distance matrices based on the tcrdist3[12,13] package (TCRβ only), Euclidean distance in the latent space of the ESM2 language model[15] (*esm2_t36_3B_UR50D* variant, using mean representations of CDR3β with 2560 dimensions) and negative BLOSUM62 pairwise sequence similarity (using *pairwise2.align.globaldx* in BioPython package[16]). We then divided the data by epitope and plotted the portions corresponding to each cluster within each epitope-specific data bucket. The subset of motif-containing peptides was selected based on sequence logos (created using logomaker[17]) of CDR3β sequences, aligned based in IMGT numbering using ANARCI[18].

Competing interests

D.O. is an employee of Cell Signaling Technology

References

1. 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*. **23**, 511-521 (2023).

2. Chronister, W. et al. TCRMatch: Predicting T-cell receptor specificity based on sequence similarity to previously characterized receptors. *Front. Immunol.* **12**, 640725 (2021).

3. Davis, M., Altman, J.D. & Newell, E.W. Interrogating the repertoire: broadening the scope of peptide–MHC multimer analysis. *Nat. Rev Immunol*. **11**, 551–558 (2011).

4. Irving, M. et al. Interplay between T cell receptor binding kinetics and the level of cognate peptide presented by major histocompatibility complexes governs CD8+ T cell responsiveness. *JBC*. **287**, 23068-23078 (2012).

5. Sibener, L.V. et al. Isolation of a structural mechanism for uncoupling T cell receptor signaling from peptide-MHC binding. *Cell*. **174**, 672-687.e27 (2018).

6. Zhao, X. et al. Tuning T cell receptor sensitivity through catch bond engineering. *Science*. **376**, eabl5282 (2022).

7. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature*. **547**, 94–98 (2017).

8. Hudson, D., Lubbockb, A., Bashamb, M. & Koohy, H. A comparison of clustering models for inference of T cell receptor antigen specificity. *ImmunoInformatics*. **13**, 100033 (2024).

9. Leary, A.Y., et al. Designing meaningful continuous representations of T cell receptor sequences with deep generative models. *Nature Communications*. **15**, 4271 (2024).

10. Meysman, P. et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics*. **9**, 100024 (2023).

11. Simpson, C. et al. Lifting the curse from high dimensional data: Automated projection pursuit clustering for the variety of biological data modalities. Preprint at https://www.biorxiv.org/content/10.1101/2024.04.18.589981v1 (2024).

12. Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature.* **547**, 89–93 (2017).

13. Mayer-Blackwell, K. et al. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife*. **10**, e68605 (2021).

14. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. **12**, 2825–2830 (2011).

15. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. **379**, 1123-1130 (2023).

16. Cock, P.J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. **25**, 1422–1423 (2009).

17. Tareen, A. and Kinney, J.B. Logomaker: beautiful sequence logos in Python. *Bioinformatics*. **36**, 2272–2274 (2020).

18. Dunbar, J. and Deane, C.M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics.* **32**, 298–300 (2016).

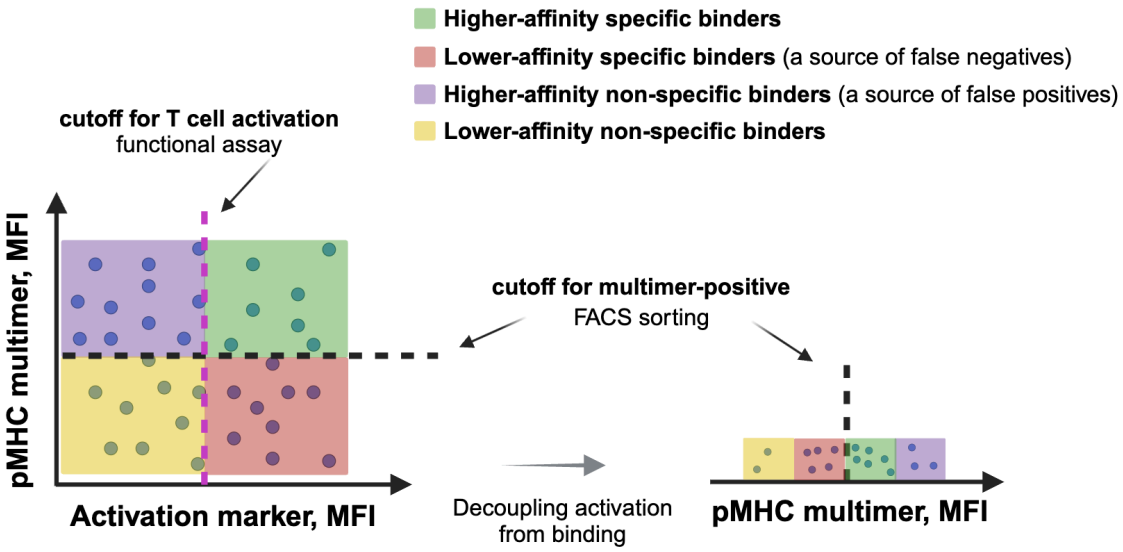


Figure 1. Current multimer-based methods for assessing TCR specificity do not support the separation of specificity and activation prediction into two independent tasks. Omitting functional readouts from multimer-binding assays is effectively equivalent to removing an essential dimension that helps differentiate between specific and non-specific TCR populations.

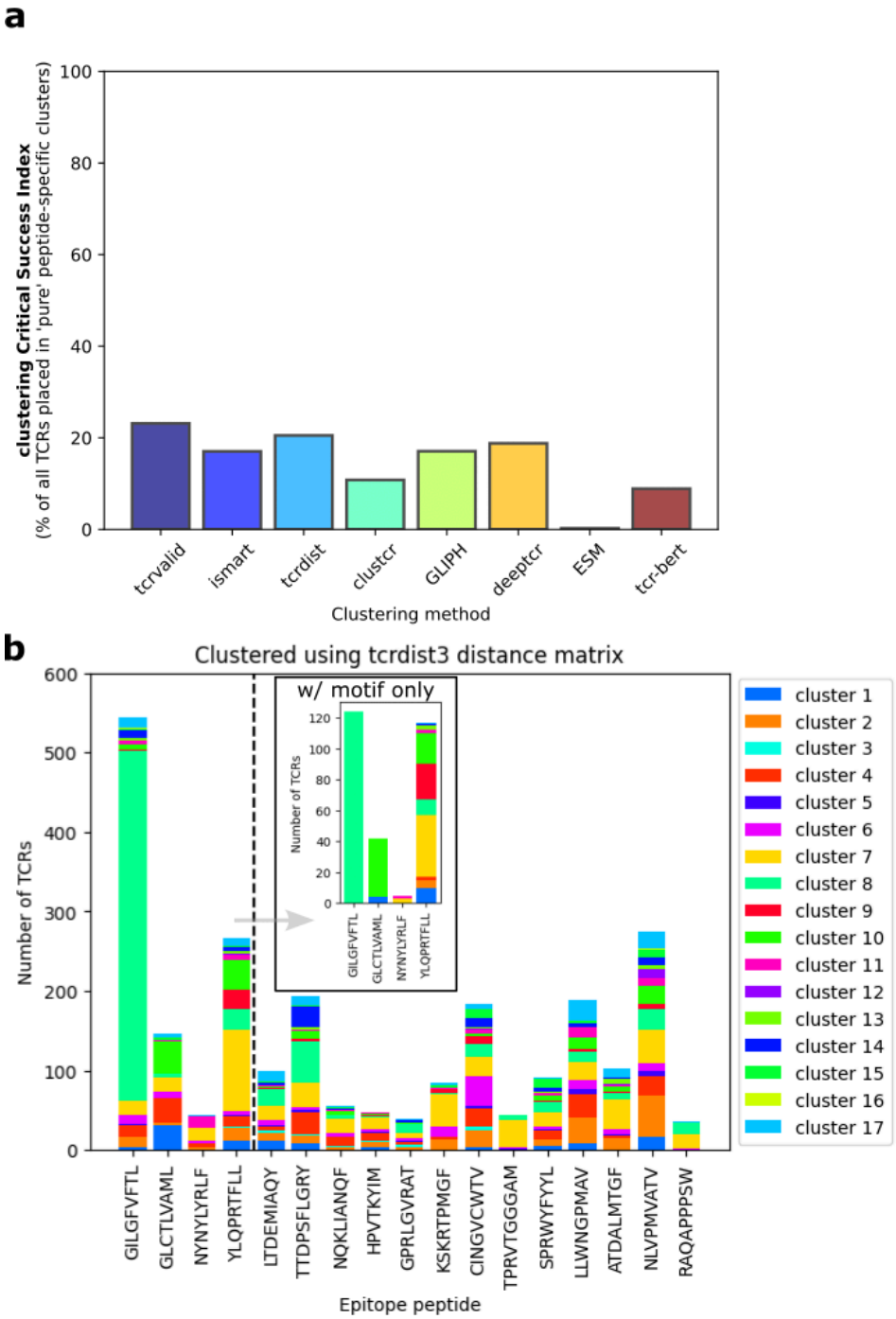


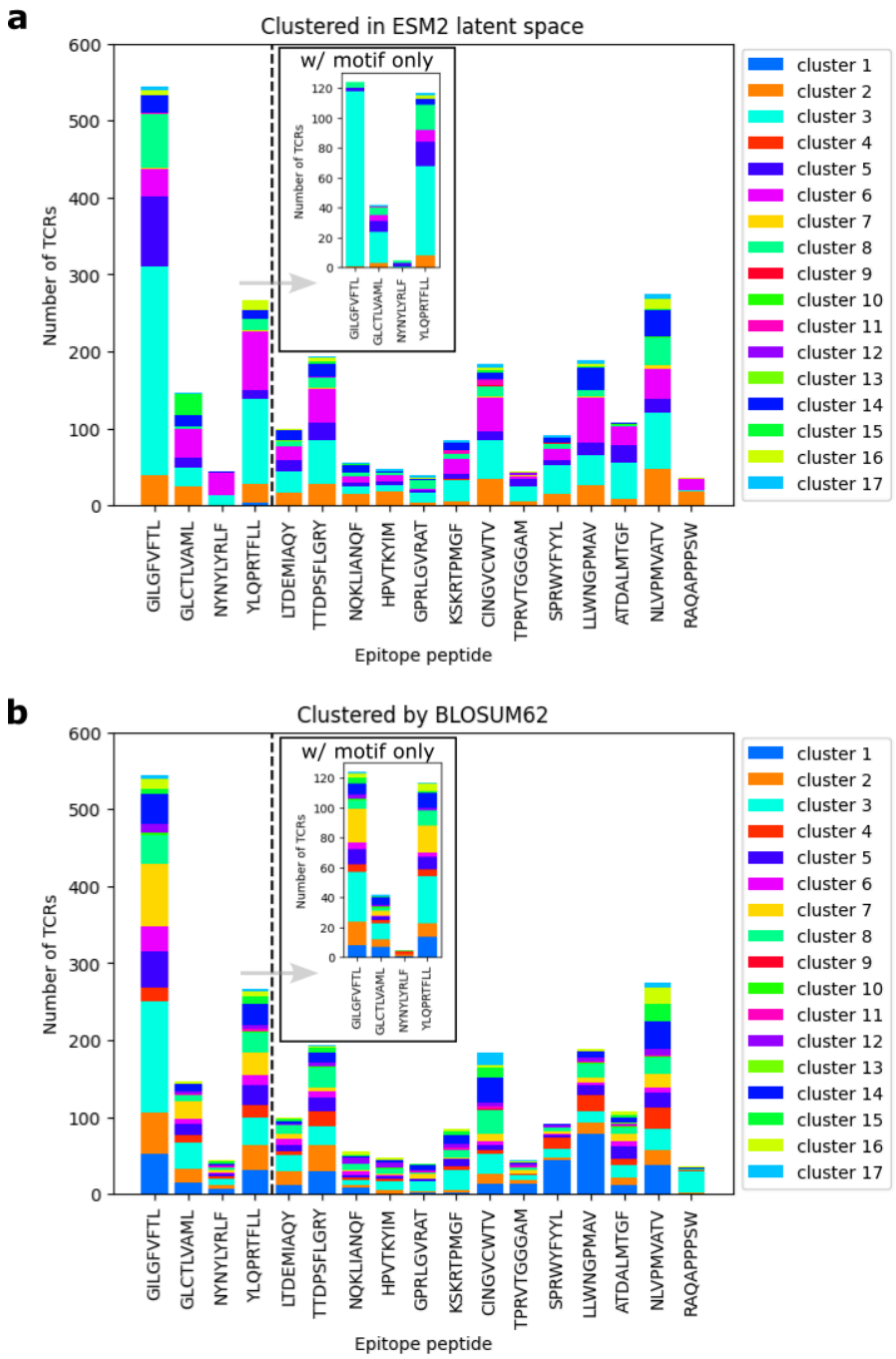
Figure 2. Unsupervised models applied to TCR specificity data group TCRs independently of their epitope specificity. (a) As was shown by *Leary et al.*[9] the current unsupervised modeling methods fail to sort TCRs into pure (antigen specific) clusters in >70 % of the cases (showing clustering by TCRβ with broadest method coverage, but results are similar for TCRα and combined TCR$\alpha\beta$, data taken from *Leary et al.*[9]). **(b)** Distribution of TCRs from 17 peptide-specific datasets from IMMREP_2022 benchmark[10] in 17 clusters acquired with hierarchical clustering based on TCRβ

distance calculated using tcrdist3[12,13]. First four peptides separated by the dashed line contain a significant CDR3β binding motifs. Distribution of motif-containing-only data points in the 17 clusters is shown in the inset plot (motifs detected using sequence logos).

Supplementary materials

Type of an observation	Assay type	Conclusion	Reference
Binding or activation following a multimer stain or incubation with peptides is not an indicator of whether a T cell is specific to an endogenously presented epitope	Peptide-based MIRA	Multimer staining or incubation with peptides is not an indicator of T cell specificity against endogenously presented epitopes	Nolan et al[1]
Up to 72% of TCR clones against C virus NS3 variant peptides as well as the HY self-antigen were tetramer positive yet did not produce an antigen specific response	Tetramer staining and functional response by CD107a	pMHC binding does not uniformly predict T cell activity	Sibener et al[2] Yu et al[3]
pMHC class I tetramers and dextramers can fail to detect relevant functional T cell clonotypes and underestimate antigen-reactive T cell populations	Multimer staining and functional assays including cytokine release and cytotoxic assays Dodecamers vs dextramers vs tetramers	Many previous attempts to enumerate antigen specific T cell populations by pMHC multimer staining will have underestimated the size of these populations	Rius et al[4] Tungatt et al[5] Huang et al[6]

Supplementary Table 1. Representative studies underscore how multimer-binding assays significantly bias our perception of TCR specificity.



Supplementary Figure 1. Distribution of TCRs from 17 peptide-specific data sets from IMMREP_2022 benchmark in 17 clusters acquired with hierarchical clustering based on TCRβ distance calculated using mean ESM2 (3B variant - 2560 dimensions) latent representations of TCR CDR3β sequences (**a**) and BLOSUM62 global alignment score among TCR CDR3β sequences (**b**). First four peptides separated by the dashed line contain a significant CDR3β binding motif. Distribution of motif-containing-only data points in the 17 clusters is shown in the inset plots (motifs detected using sequence logos).

References

1. Nolan, S. et al. A large-scale database of T-cell receptor beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. Preprint at https://www.researchsquare.com/article/rs-51964/v1 (2020).

2. Sibener, L.V. et al. Isolation of a structural mechanism for uncoupling T cell receptor signaling from peptide-MHC binding. *Cell*. **174**, 672-687.e27 (2018).

3. Yu, W. et al. Clonal deletion prunes but does not eliminate self-specific αβ CD8+ T lymphocytes. *Immunity*. **42**, 929-941 (2015).

4. Rius, C. et al. Peptide–MHC class I tetramers can fail to detect relevant functional T cell clonotypes and underestimate antigen-reactive T cell populations. *JI*. **200**, 2263–2279 (2018).

5. Tungatt, K. et al. Antibody stabilization of peptide–MHC multimers reveals functional T cells bearing extremely low-affinity TCRs. *JI*. **194**, 463–474 (2015).

6. Huang, J. et al. Detection, phenotyping, and quantification of antigen-specific T cells using a peptide-MHC dodecamer. *PNAS*. **113**, E1890-E1897 (2016).