We collect the number of TCRs and pMHCs used by different TCR-pMHC methods and number of positive and negative examples. In all cases, we disregard the (usually negative) data that is artificially generated with experimentally testing TCR-pMHC binding. Note that the numbers are approximate since multiple validation methods exist for train-test split and we broadly report the numbers as counted from indicated full *training* datasets (i.e., not considering subsampling, folds etc).

1. BATMAN: NB is treated as negative instance and SB+WB are treated as positive instance. Only 1 Hamming scans, and both MHC I and II data are considered.
2. pTEAM: The data is from mutational scans: normalized activations of 46.9, 66.09, and 40 are used as thresholds for defining binders as per the paper: <https://www.cell.com/cell-genomics/fulltext/S2666-979X(24)00238-6#mmc2> (data source: datasets S1-S4 files). Unique peptides: all 153 SIINFEKL mutants + 134 neoantigen mutants + 172 NLV-mutants. Unique TCRs: (15+21) Ova-TCRs + 7 neoantigen-TCRs + 20 NLV-TCRs.
3. IEDB: paper: [10.1371/journal.pcbi.1003266](https://doi.org/10.1371/journal.pcbi.1003266). Data is found in 2 Dataset files. Since the dataset does not use any TCR information, TCR count is recorded as 0 and the same convention is followed for all other immunogenicity models below.
4. PRIME 2: paper: <https://doi.org/10.1016/j.cels.2022.12.002>,

Data source: supplementary table S4. In the data, random=1 are discarded as they were not experimentally verified (only about 6k were, corresponding to random=0).

1. ImRex: paper: <https://academic.oup.com/bib/article/22/4/bbaa318/6042663> ,

Data Source: “This dataset was reduced to 19,842 unique CDR3-epitope pairs by selecting only human TCR sequences”. (negative data from unspecific TCR seq discarded, so only positive examples remain). To know the exact TCR numbers, we downloaded an example train data from <https://github.com/pmoris/ImRex/blob/master/models/pretrained/2020-07-30_11-30-27_trbmhci-shuffle-padded-b32-lre4-reg001/iteration_no_validation/train.csv>

filtered it for y=1 (known binder, y=0 is random shuffle) and found the number of unique CDR3s and peptides.

1. pMTnet: paper: <https://www.nature.com/articles/s42256-021-00383-2#Sec1>

Data Source: “We collected a total of 32,607 pairs of TCR–pMHC bindings from a series of peer-reviewed publications…We created ten times more negative pairs by randomly mismatching the TCRs and pMHCs of these 32,607 pairs.” (so no experimentally verified negative data). We downloaded the training data from <https://github.com/tianshilu/pMTnet/blob/master/data/training_data.csv> and saw there are 29226 unique CDR3 sequences and 429 unique peptides.

1. epiTCR: They have full training data validated – no random shuffling

paper: <https://academic.oup.com/bioinformatics/article/39/5/btad284/7140137#404203429>

data source: <https://github.com/ddiem-ri-4D/epiTCR/blob/main/data/categories/full-training-with-categories.csv.zip>

1. ERGO II: paper: <https://www.frontiersin.org/articles/10.3389/fimmu.2021.664514/full> (they use randomization to generate the full negative data)

Data source: Since McPAS model performed well, we record stats of that . Following the supplementary table of thr ERGO II paper (https://www.frontiersin.org/articles/10.3389/fimmu.2021.664514/full#supplementary-material), we record the number of unique CDR3b sequences as the number of unique TCRs in the training data: 9490, and number of unique peptides=319, number of positive sample=11630

1. NetTCR-2.0: paper: <https://www.nature.com/articles/s42003-021-02610-3#Sec8>

Numbers are recorded for paired TCR which performs better: “Positive data points were taken from … 3859 unique binding pairs were identified from IEDB and 2843 from VDJdb. These provided 4598 unique CDR3α-/β-peptide interactions with 276 different peptides specific to allele HLA-A\*02:01.

Negatives were derived from 10X. Using the same restrictions as for the positives (CDR3 length between 8 and 18 AAs, peptide length 9, and peptides specific for HLA-A\*02:01), 627,323 unique data points with 0 UMI counts to all the tested peptides were identified. These contained 33,017 unique TCRs tested against a set of 19 different peptides. In total, 17 of these overlapped with the peptides in the positive data set.”

So, total #TCRs=4598+33017=37615, 4598 positive and 627,323 negative instances, and 276+19-17= 278 peptides

1. TCR-BERT: <https://doi.org/10.1101/2021.11.18.469186> . We disregard pre-training and TCR embedding learning data, and only record the numbers of LCMV peptide specific TCR sequences, so 1 peptide, and “Overall, this results in (n=17,702) unique TRA/TRB pairs with consistent labels that we use for model training and evaluation. Among these, 13% (n=2306) are observed to have mid or high binding – we consider these “positive” examples of TRA/TRBs binding GP33.”
2. TCRdist: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5616171/> , no negative data, data summarized in Extended Data Table 1. Total number of TCRs=total number of positive examples = sum of #clones for 10 epitope-specific repertoires = 117 +305 +324 +642 +87 +158 +291 +76 +275+61
3. TITAN: all negative data generated by shuffling. “With this procedure, we build a training dataset of 46 290 examples, 50% of which are positive, encompassing 192 different epitopes.” <https://academic.oup.com/bioinformatics/article/37/Supplement_1/i237/6319659#409220337>
4. DeepImmuno: General immunogenicity prediction, no TCR info involved. <https://academic.oup.com/bib/article/22/6/bbab160/6261914#312126902> : “Specifically, 8971 data instances were retained in the final dataset, among which 4059 were positive reactive instances and the remaining 4912 were negative”
5. MixTCRpred: All negative data generated by shuffling. <https://doi.org/10.1101/2023.09.13.557561> , “ For further analysis, only pMHCs with ten or more binding TCRs were considered, resulting in a total of 17,715 αβTCRs interacting with 146 pMHCs (Figure 1C and Table S1)”. We also downloaded the full training data from <https://github.com/GfellerLab/MixTCRpred/blob/main/full_training_set_146pmhc.csv> and checked that the total number of data points is equal to 17715 as well.
6. PanPep: negative data was generated computationally. For the rest, “As a result, the base dataset contains 699 unique peptides and 29,467 unique TCRs with 32,080 related peptide–TCR binding pairs considering the cross-reactivity of peptide–TCR binding.”: <https://www.nature.com/articles/s42256-023-00619-3#Sec13>
7. BigMHC: We collect data for “Immunogenicity Training” mode. “BigMHC transfer learned only on the nonrandom pMHC data, of which 1,580 are positive and 5,293 are negative.”: <https://www.nature.com/articles/s42256-023-00694-6#Sec7>
8. catELMo: Negative data is generated randomly. For positive data, “Altogether, we obtained 150,008 unique TCR-epitope pairs known to bind to each other having 140,675 unique TCRs and 982 unique epitopes.”: <https://www.biorxiv.org/content/10.1101/2023.04.12.536635v1>
9. ATM-TCR: Negative data is generated randomly. For positive data, “This resulted in the primary dataset containing 128,142 unique TCR-epitope pairs, with 931 unique epitopes and 119,984 unique TCRs.”: <https://www.frontiersin.org/articles/10.3389/fimmu.2022.893247/full>
10. GLIPH: No negative data. “In total, the training set consisted of 2,068 unique TCRs of known specificity (Supplementary Table 1)”: <https://www.nature.com/articles/nature22976#Sec2> . Supplementary Table 1 (sheet=”Curated”) had 1973 unique CDR3s and 7 unique peptides, 2067 pairs in total.
11. Repitope: Even though they consider both TCR and peptide sequences, they are not matched in the training data. So it is like other immunogenicity prediction tools. We include the stats from the data in supplementary data sheet 2. We neglect immunogenicity contradictions in data and focus on the immunogenicity column only, to count positives and negatives and add them up. <https://www.frontiersin.org/articles/10.3389/fimmu.2019.00827/full>
12. TCRAI: paper: <https://www.science.org/doi/full/10.1126/sciadv.abf5835> . We download the data from <https://github.com/regeneron-mpds/TCRAI/blob/main/data/public_TCRs.csv> and <https://github.com/regeneron-mpds/TCRAI/blob/main/data/CNN-prediction-with-REGN-pilot-version-2.csv> and record the number of unique TCRs, peptides, and pairs using a python script.
13. DeepTCR (<https://www.nature.com/articles/s41467-021-21879-w#Sec1> ) has a TCR encoding mode, trained on antigen-specific TCR sequences (positive examples), and a UMI regression mode on 10X data. For the former case, “we collected data for tetramer-sorted antigen-specific cells for nine murine (Db-F2, Db-M45, Db-NP, Db-PA, Db-PB1, Kb-M38, Kb-SIY, Kb-TRP2, Kb-m139) and seven human (A1-CTELKLSDY, A1-VTEHDTLLY, A2-GILGFVFTL, A2-GLCTLVAML, A2-NLVPMVATV, B7-LPRRSGAAGA, B7-TPRVTGGGAM) antigens where the ground truth label corresponds to a particular antigen specificity for an individual sequence”. This is all positive data, we looked into the TCR sequences in folders <https://github.com/sidhomj/DeepTCR/tree/master/Data/Murine_Antigens> and <https://github.com/sidhomj/DeepTCR/tree/master/Data/Human_Antigens> and count their total numbers (117+291+305+324+642+158+(776+655+456+600)+(410+435+347)+87 = 5603 murine and 25+271+(275+588)+(76+739)+(61+166)+214+65 = 2480 human TCRs and an equal number of positive TCR-MHC binding). In the latter case, “a second single-cell dataset published by 10x Genomics where the binding to cognate T cells of 57,229 unique α/β pairs to 44 specific pMHC multimers and 6 negative controls was characterized”. This dataset was downloaded from <https://github.com/sidhomj/DeepTCR/tree/master/Data/10x_Data> and has continuous count information on 57228\*44 TCR-pMHC pairs. Out of them, we designate negative as count=0 (n= 2562701) and positive as count>0 (n= 298699). Total 9 murine and 7+44+6-5 (common) =52 human antigens. In reporting, we add positives from two cases (5603+2480+298699=306782). Total number of TCRs: 5603+2480+57228=65311.
14. TEINet: No experimental negative data. “At last, we constructed a large dataset with 44 682 pairs of TCRs and epitopes, among which 41 610 TCRs are linked to 180 epitopes”: <https://academic.oup.com/bib/article/24/2/bbad086/7076118>
15. SwarmTCR: No experimental negative data. “In total, our default SC dataset comprised 1447 TCRs (for complete dataset counts see Additional file 1: Table 4)”: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04335-w#Sec9> . Table 4 has 83 epitopes.
16. TCRGP: No experimental negative data. “The VDJdb data and the Dash data have some overlap for TCRs specific to three epitopes: In the VDJdb data 34 (27 unique) of the 413 (242 unique) TCRs for pp65495-503, 30 (27) of 299 (152) for BMLF1280-288, and 74 (61) of 239 (138) for M158-66 can also be found from the Dash data.”: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008814#pcbi.1008814.s012> The data sources are VDJdb and Dash data. We looked into <https://github.com/emmijokinen/TCRGP/tree/master/training_data/paper> folder for data (files names marked with “unique”) and found 138+474+120+234+286+542+526+586+1168+174=4248 unique Dash TCRs (“three epitopes from humans and for seven epitopes from mice”) and 278+78+120+156+190+276+304+106+92+118+116+208+130+282+396+484+100+298+58+80+240+104+610 = 4824 unique VDJdb TCRs (23 epitopes), total 4248+4824-27-27-61= 8957 unique TCRs, combining two datasets, and 10+23-3=30 epitopes.