Supplementary Material

# Supplementary Materials and Methods

## Choice of TCR framework and creation of TCR yeast display libraries

The TCR framework used throughout this study is a patient-derived A6 TCR clone that recognizes the Human T-lymphotropic virus 1 (HTLV-1)-derived *Tax* peptide (1–3). The framework sequence was expressed as a single chain TCR (scTv) as previously described (4). Briefly, the conventional pCT302 vector is implemented with an aga2 leader peptide sequence and N and C terminal expression tags, hemagglutinin tag (HA) and Flag epitope polypeptide tag, respectively (5,6).

We used Garboczi et al.’s (2) crystal structure (PDB: 1AO7) to identify residues along the CDR3α and CDR3β that make the most contacts with *Tax* peptide. The positions 99-101 in CDR3α and positions 98-100 in CDR3β were randomized simultaneously via oligonucleotides with uniform amino acid distribution, excluding cysteines that could form disulfide bonds, from Ella Biosciences. The final TCR library had the size of 196 (~47 million). The DNA constructs were created via randomized oligos inserted into the yeast display vector pCT302 and electroporated into the RJY100 yeast strain as previously described (7–10). We focus on CDR3 as its chains are the main drivers of peptide specificity, while CDR1 and CDR2 improve the MHC affinity alone (11–13).

## Verifying TCR reactivity to selected pMHC monomers

We chose a subset of TCRs enriched across selections with a range of predicted cross-reactivities to pMHC monomers. The eight TCRs described in Extended Data Figure 2C were expressed as individual clones on yeast and surface expression of the construct was assessed via flow cytometry by antibody staining the Flag epitope tag (Cell Signaling Technology, clone D6W5B). pMHC tetramers were generated by incubating 5-fold molar excess of pMHC monomer with streptavidin–Alexa Fluor 647 (SAV-647) made in-house. Yeasts were then stained with different concentrations of the pMHC tetramers ranging from 0.01nM - 200nM for 1 hour at 4 °C (we report tetramer concentration as the concentration of SAV-647). Following staining, cells were washed and fixed in 2% paraformaldehyde (PFA). Fixed cells were analyzed by flow cytometry. The tetramer binding was quantified by median fluorescence intensity (MFI) and plotted against tetramer concentration to generate titration curves.

**1.3 Selection of peptide targets and production of pMHC monomer targets**

Along with *Tax* peptide that is the cognate antigen for our A6 TCR framework, we selected additional peptide targets that serve as representative self-peptide off-targets. A previous *in vitro* cell lysis study (3) reported six peptides (HUD, BENE, phosphofructokinase, tyrosine kinase, Tax Mutant TM10, and homeobox) that are also recognized by the native (DSW/LAG) A6 TCR (Figure 1C). We ran seven sets of parallel selections against *Tax* and these six peptides.

The pMHC reagents were produced in a single chain trimer format as previously described (14). Briefly, each component of the expression system is linked via GS linkers and are in the following order, signal peptide (for excretion), peptide, human b2m, HLA:A\*0201, AviTag for biotinylation (15) and 8x His for purification. This construct was cloned into the pVL1393 insect cell expression plasmid and expressed in sf9 and hi5 insect cells for production. The resulting product was size-separated via FLPC, biotinylated, and validated via a streptavidin shift assay.

## Choice of TCR framework and creation of TCR yeast display libraries

The TCR framework used throughout this study is a patient-derived A6 TCR clone that recognizes the Human T-lymphotropic virus 1 (HTLV-1)-derived *Tax* peptide (1–3). The framework sequence was expressed as a single chain TCR (scTv) as previously described (4). Briefly, the conventional pCT302 vector is implemented with an aga2 leader peptide sequence and N and C terminal expression tags, hemagglutinin tag (HA) and Flag epitope polypeptide tag, respectively (5,6).

We used Garboczi et al.’s (2) crystal structure (PDB: 1AO7) to identify residues along the CDR3α and CDR3β that make the most contacts with *Tax* peptide. The positions 99-101 in CDR3α and positions 98-100 in CDR3β were randomized simultaneously via oligonucleotides with uniform amino acid distribution, excluding cysteines that could form disulfide bonds, from Ella Biosciences. The final TCR library had the size of 196 (~47 million). The DNA constructs were created via randomized oligos inserted into the yeast display vector pCT302 and electroporated into the RJY100 yeast strain as previously described (7–10). We focus on CDR3 as its chains are the main drivers of peptide specificity, while CDR1 and CDR2 improve the MHC affinity alone (11–13).

## Library selection, sequencing, and read processing

We performed three rounds of magnetic-activated cell sorting (MACS) selection using microbeads (16,17), generating two independent biological replicates for each of seven pMHC targets in parallel. The consecutive rounds of selection enriched for specific TCR sequences in the library that bind strongly to each pMHC target. Each sample was yeast miniprepped independently and then PCR amplified once with Illumina adapters and then PCR amplified one final time with smRNA\_F and BMC\_R barcoded primers such that all samples could be pooled together for subsequent sequencing with high confidence. The samples were then given to the MIT BioMicroCenter for sequencing on the Illumina NovaSeq 6000 machine to receive an estimated 2.5B paired end reads that are 2 x 150 bps and contain the full CDR3 alpha and beta chain sequences.

To extract CDR3 sequence from each read pair, each of CDR3 alpha and beta chain sequences is identified and extracted using the pre-defined oligonucleotide (15 bps) sequence that flank the CDR3 chain sequences. Each extracted sequence is checked to ensure that it has the correct length (9 nucleotides long) and depleted with codons corresponding to cysteine. The sequences that do not satisfy these conditions are discarded. The DNA sequences are translated into amino acid sequences according to the standard codon table. The resulting alpha and beta chain amino acid sequences are joined to form a randomized CDR3 sequence. The TCR sequences that are present in the MACS Round 3 in both replicates were considered as reproducibly target-binding TCRs and employed for the subsequent analyses described below.

**1.6 Confirmation and quantification of TCR sequence enrichment during yeast panning**

We confirmed that TCR sequences are successfully enriched during yeast panning by computing two metrics and visualizing the logos of enriched sequences. First, the number of unique TCR sequences in each panning round is computed by first extracting the CDR3 sequences as described above and counting the distinctive sequences. Second, the percentages of reads contributed by top 100 TCR sequences in each round is computed by identifying the 100 TCR sequences with the greatest number of read counts and summing their counts. This value is divided by the number of total counts of pertaining to TCR sequences.

The sequence logos were generated using WebLogo (18). As the WebLogo can take up to 10,000 sequences for logo creation, we sampled 10,000 sequences from each panning round of each replicate, where the probability of a TCR sequence being sampled is correspondent to its frequency in the given dataset.

# Supplementary Figures

**Supplementary Figure 1. Yeast panning rounds enrich for specific CDR3α/β sequences.** We confirm that specific TCR sequences are enriched during the yeast panning rounds.

(A) The number of unique TCR sequences observed in each panning round is counted. Across all target peptides and in both replicates, the number of unique TCR sequences in the panning Round 2 is smaller than that of the panning Round 1 (*p* = 1.2e-03, one-sided Wilcoxon signed-rank test). Likewise, the number of unique TCR sequences in the panning Round 3 is smaller than that of the panning Round 2 (*p* = 6.1e-05).

(B) The percentages of reads contributed by the top 100 TCR sequences are computed. Across all target peptides and in both replicates, the percentages of reads contributed by the top 100 TCR sequences in the panning Round 2 is greater than that of the panning Round 1 (*p* = 6.1e-05, Wilcoxon signed-rank test). Likewise, the percentages of reads contributed by the top 100 TCR sequences in the panning Round 3 is greater than that of the panning Round 2 (*p* = 6.1e-05).

(C) Weighted sampling of TCR sequences present in each panning round against *Tax* peptide is performed, and the logos of the sampled sequences are generated (*Methods*). N only certain amino acid preferences are enriched in position-dependent manner, but also these preferences in two replicates are convergent.

**Supplementary Figure 2. Titration with pMHC monomers recapitulate the yeast panning assays.** We validate the yeast-display selection with high-fidelity titration assay with pMHC tetramers.

(A) A subset of TCRs enriched from each set of selections were expressed as individual clones on yeast and stained with different concentrations of selected pMHC tetramers ranging from 0.01nM - 200nM. The titration curves are plotted by TCRs.

(B) The titration curves are plotted by pMHCs.

(C) Titration results overlaid with Round 3 yeast selection read counts averaged over two replicates. Yeast-display selection is more likely to yield read counts at defined thresholds if a TCR-pMHC pair is “binding” according to titration assay rather than “subtle binding or no binding”, demonstrating that yeast-display selection successfully enriches for binders.

(D) *P*-values are computed to test the hypotheses that the replicate average yeast-display read count is greater than R when the titration assay reads “binding” (Fisher’s exact test, one-sided). The test achieves statistical significance (*p* < 0.05) for all values of R except 1. When R = 1, the test achieves statistical significance (*p* = 0.013 < 0.05) if the minimum, not the average (mean), read count over two replicates is taken, showing that our strategy of requiring both replicates to have non-zero read counts to define binders adds robustness.

**Supplementary Figure 3. MLP models with shared internal weights outperform logistic regression classifiers and permuted models.** We benchmarked our MLP models against logistic regression models receiving BLOSUM50-encoded TCR sequences. Unlike the MLP models whose internal weights are shared in predicting the binding of all seven target peptides, the logistic regression classifiers were fit per each target. In addition, we conducted permutation test by shuffling the relationships between TCR sequences and peptide targets of the training datasets for each of the 10 cross-validation folds (Fig 3A), whilst keeping the test sets intact, and re-training the MLP models.

(A) MLP models outperformed the logistic regression and permuted models in terms of both F1-scores (*p*LR= 6.0e-06; *p*PM= 3.4e-08; Mann-Whitney *U*-test) and AUPRC (*p*LR= 6.0e-06; *p*PM= 3.4e-08).

(B) MLP models outperformed the logistic regression models in binding prediction for all cross-reactivity strata (∀strata, *p* = 0.0020, Wilcoxon signed-rank test). It is notable that the greatest performance improvement for MLP over logistic regression was achieved for 7-cross-reactive stratum, suggesting that the pattern of cross-reactivity implicitly learnt via the shared internal weights was useful.

(C) MLP models outperform the logistic regression models in binding prediction for all seven peptide targets (∀targets, *p*= 0.0020, Wilcoxon signed-rank test).

**Supplementary Figure 4. VHSE8 embedding improves TCR binding classification, albeit to a less degree than BLOSUM50.**

(A) Models trained with BLOSUM50 embedding outperformed those trained with VHSE8 embedding, which in turn outperformed one-hot (∀metrics, replicates *p* < 0.001, Kruskal-Wallis test followed by Dunn's post-hoc test with FDR correction). Contrastingly, the performances of the models trained with the same embedding did not differ significantly (∀metrics, embeddings *p* > 0.05, Kruskal-Wallis test followed by Dunn's post-hoc test with FDR correction).

(B) The degrees of performance improvement were significantly different for different cross-reactivity strata (*p* < 0.0001 (Kruskal-Wallis test), *ε2*= 0.73).

(C) Further, the degrees of improvement were significantly different for different pMHC targets (*p* < 0.0001 (Kruskal-Wallis test), *ε2* = 0.73).

**Supplementary Figure 5. Observed hierarchy of examined amino acid positions are invariant to the choices of sequence embeddings and performance metrics.** We performed sequence-masked training, where different residue positions are deleted.

(A) Masking of any residue or any CDR3 chain causes significant performance degradation regardless of sequence embedding (BLOSUM50, VHSE8, one-hot) and performance measures (F1 score, AUPRC)(p < 0.001, 1-sample T-test). Note that the rank order of marginal importance for different residue positions (a99 - a101 - b98 - a100 - b100 - b99) remain consistent across different sequence embeddings and performance measures, except for the switch between a101 and b98 positions within the margin of error in the case of BLOSUM50/AUPRC combination. Further, masking alpha chain caused much greater deterioration than masking beta chain across different sequence embeddings (p < 0.001 (Wilcoxon signed-rank test)).

(B) We performed computational alanine substitution and used FlexPepDock to infer the changes in interface energy (*ΔI\_sc,* lower more stable) following alanine substitution. We found that *ΔI\_sc* explains the ML performance deterioration, as measured by *ΔAUPRC* (less reduction, better model performance), following residue and chain masking.

**Supplementary Figure 6. BLOSUM-50 (BLOcks SUbstitution Matrix) matrix of amino acids.**

The BLOSUM-50 substitution matrix is shown (19,20).

**Supplementary Figure 7. VHSE-8 (Vectors of Hydrophobic, Steric, and Electronic properties-8) matrix of amino acids.**

The VHSE-8 matrix of amino acid descriptor is shown (21).

**3 Supplementary References**

1. Ding YH, Smith KJ, Garboczi DN, Utz U, Biddison WE, Wiley DC. Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. Immunity. 1998 Apr;8(4):403–11.

2. Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature. 1996 Nov 14;384(6605):134–41.

3. Hausmann S, Biddison WE, Smith KJ, Ding YH, Garboczi DN, Utz U, et al. Peptide recognition by two HLA-A2/Tax11-19-specific T cell clones in relationship to their MHC/peptide/TCR crystal structures. J Immunol Baltim Md 1950. 1999 May 1;162(9):5389–97.

4. Aggen DH, Chervin AS, Insaidoo FK, Piepenbrink KH, Baker BM, Kranz DM. Identification and engineering of human variable regions that allow expression of stable single-chain T cell receptors. Protein Eng Des Sel PEDS. 2011 Apr;24(4):361–72.

5. Zhang L, Hernan R, Brizzard B. Multiple tandem epitope tagging for enhanced detection of protein expressed in mammalian cells. Mol Biotechnol. 2001 Nov;19(3):313–21.

6. Hopp TP, Prickett KS, Price VL, Libby RT, March CJ, Pat Cerretti D, et al. A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. Bio/Technology. 1988 Oct;6(10):1204–10.

7. Van Deventer JA, Kelly RL, Rajan S, Wittrup KD, Sidhu SS. A switchable yeast display/secretion system. Protein Eng Des Sel PEDS. 2015 Oct;28(10):317–25.

8. Van Deventer JA, Wittrup KD. Yeast surface display for antibody isolation: library construction, library screening, and affinity maturation. Methods Mol Biol Clifton NJ. 2014;1131:151–81.

9. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. Nat Protoc. 2006;1(2):755–68.

10. Adams JJ, Narayanan S, Liu B, Birnbaum ME, Kruse A, Bowerman NA, et al. T cell receptor signaling is limited by docking geometry to peptide-Major Histocompatibility Complex. Immunity. 2011 Nov 23;35(5):681–93.

11. Smith SN, Wang Y, Baylon JL, Singh NK, Baker BM, Tajkhorshid E, et al. Changing the peptide specificity of a human T-cell receptor by directed evolution. Nat Commun. 2014 Nov 7;5:5223.

12. Fischer DS, Wu Y, Schubert B, Theis FJ. Predicting antigen specificity of single T cells based on TCR CDR3 regions. Mol Syst Biol. 2020 Aug;16(8):e9416.

13. Springer I, Tickotsky N, Louzoun Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. Front Immunol. 2021;12:664514.

14. Hansen TH, Connolly JM, Gould KG, Fremont DH. Basic and translational applications of engineered MHC class I proteins. Trends Immunol. 2010 Oct;31(10):363–9.

15. Fairhead M, Howarth M. Site-specific biotinylation of purified proteins using BirA. Methods Mol Biol Clifton NJ. 2015;1266:171–84.

16. Schmitz J, Petrasch S, van Lunzen J, Racz P, Kleine HD, Hufert F, et al. Optimizing follicular dendritic cell isolation by discontinuous gradient centrifugation and use of the magnetic cell sorter (MACS). J Immunol Methods. 1993 Feb 26;159(1–2):189–96.

17. Gupta VK, McConnell I, Dalziel RG, Hopkins J. Two B cell subpopulations have distinct recirculation characteristics. Eur J Immunol. 1998 May;28(5):1597–603.

18. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004 June;14(6):1188–90.

19. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915–9.

20. Isa MN, Ahmad MI, Murad SAZ, Ismail RC, Benkrid K. Biological sequence alignments: A review of hardware accelerators and a new PE computing strategy. In: 2014 IEEE REGION 10 SYMPOSIUM [Internet]. 2014 [cited 2025 Aug 22]. p. 39–44. Available from: https://ieeexplore.ieee.org/document/6862994

21. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers. 2005;80(6):775–86.