Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Peptide-independent deep learning with TRIOPS predicts TCR–HLA restriction and recovers locus-biased HLA loss in tumors

Vanessa Jonsson

v.jonsson@ucsc.edu

University of California, Santa Cruz

Nicholas Rose

University of California, Santa Cruz

Cynthia Ramirez

University of California, Santa Cruz

Lydia Mok

University of California, Santa Cruz

Christopher Wong

UC Santa Cruz https://orcid.org/0000-0001-6012-7001

Roberta Zappasodi

Weill Cornell Medicine https://orcid.org/0000-0002-7031-6671

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11143133/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Peptide-independent deep learning with TRIOPS predicts TCR–HLA restriction and recovers locus-biased HLA loss in tumors

Nicholas R Rose[1], Cynthia M Ramirez[2], Lydia Mok[1], Christopher K Wong[1,3], Roberta Zappasodi[4,5,6,7], Vanessa D Jonsson[1,3*]

[1]Department of Biomolecular Engineering and Bioinformatics, University of California, Santa Cruz, California, USA

[2] Department of Applied Mathematics, University of California, Santa Cruz, California, USA

[3] Genomics Institute, University of California, Santa Cruz, California, USA

[4]Department of Medicine, Division of Hematology & Medical Oncology, Weill Cornell Medical College, New York, NY 10021

[5]Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021

[6]Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medicine, New York, NY 10021

[7]Parker Institute for Cancer Immunotherapy @Weill Cornell Medicine, Weill Cornell Medicine, New York, NY 10021

[*] Corresponding author. Email: vjonsson@ucsc.edu

Abstract

T cell receptor (TCR) antigen recognition is MHC-restricted, yet the restricting HLA allele is unknown for nearly all sequenced receptors. We present TRIOPS, a dual-branch convolutional model that predicts TCR–HLA restriction from amino acid sequence alone, trained with cross-reactivity-aware sampling. TRIOPS reaches an AUC of 0.97 on held-out TCRαβ pairs, 0.92 on HLA alleles withheld from training, and 0.81 on unseen TCRs. We tested whether predicted engagement tracks allele-specific HLA loss in 8,906 TCGA tumors across 23 cancer types, assigning each clonotype to its likeliest restricting allele within the corresponding tumor genotype. At HLA-A and HLA-B, engagement was higher at the lower-expressed allele of each heterozygous pair (concordance 0.60 and 0.65; 23/23 cancers). TRIOPS resolves this locus-dependent pattern from receptor sequences and HLA genotypes alone, establishing HLA-restricted TCR engagement as a relevant measurable feature of the T cell receptor repertoire.

Introduction

T cells recognize and eliminate infected or malignant cells by engaging antigens presented on major histocompatibility complex (MHC) proteins. Recognition is mediated by the T cell receptor (TCR), whose specificity arises from V(D)J rearrangement during thymic T-cell development. The resulting theoretical diversity, estimated at 10^{15}–10^{20} unique receptors[1,2], enables broad immune surveillance but renders systematic experimental mapping of TCR–peptide MHC (pMHC) restriction intractable.

Central to this recognition is MHC restriction: a TCR engages a peptide only when that peptide is successfully presented by an MHC molecule[3]. MHC alleles are exceptionally polymorphic (>43,000 alleles[4]), each with distinct peptide binding specificities. Thymic selection produces a peripheral repertoire that is broadly tolerant to self-peptide–MHC complexes, yet retains intrinsic reactivity to the selecting MHC alleles and their presentable peptidomes[5].

Predicting a TCR's restricting HLA allele confines its possible antigens to the peptides that allele presents. Three translational applications require knowledge of the TCR, its cognate antigen, and the restricting HLA allele together. Engineered T-cell therapies, in which tumor-antigen-specific TCRs are introduced into patient T cells[6], are emerging as powerful immunotherapies mediating regression of metastatic disease. Epitope-based vaccines prioritize

candidate peptides by their predicted presentation on a defined HLA allele[7]. In autoimmunity, disease-risk HLA alleles determine which self-peptides an expanded clonotype can recognize[8]. HLA biases in experimental pairs restrict therapeutic translation to a few well-studied alleles[9].

Computational models for predicting TCR specificity restriction remain constrained in several ways. Peptide-dependent methods such as ERGO-I[10], TCRAI[11], NetTCR 2.2[12], pMTnet[13] and TAPIR[14] require a known epitope, restricting their use to settings where the antigen is already defined, and independent benchmarks show that they generalize poorly to peptides underrepresented in training[15].

Methods that predict TCR–MHC restriction without the peptide face their own limitations. CLAIRE[16] is trained on experimentally validated pairs, but its per-allele performance is nonuniform: many alleles near AUC 0.6 and few substantially above 0.8. Similarly, TAPIR predicts TCR–HLA pairs through peptide-masking augmentation, but its aggregate AUC of 0.81 obscures substantially weaker per-allele performance[14]. Neither method provides trained weights, model architecture, or source code for reproducibility.

DePTH2.0[17] encodes both partners but is trained on statistically associated TCR β-chain (TCRβ)–HLA pairs identified by co-occurrence in population repertoires[18] rather than on validated binding. Co-occurrence may reflect HLA haplotype linkage as much as direct restriction. The resulting associations transfer poorly to experimentally validated pairs, a mismatch the DePTH2.0 authors have themselves documented[17].

Existing methods generate negative data through the random shuffling of experimentally derived TCR–MHC pairs. But TCR recognition is inherently degenerate; each receptor engages an estimated ~10^6 peptides on a given MHC[19]. Within individual patients, cross-reactivity extends to peptide restriction itself. Structurally similar HLA alleles present overlapping peptide repertoires and are recognized by overlapping sets of TCRs[20]. Random shuffling ignores allele similarity, therefore introducing false negatives wherever cross-reactivity exists. The resulting label noise penalizes discrimination at the allele boundaries where it matters most.

Here we present TRIOPS (T cell Receptor–Major Histocompatibility Complex Specificity), a deep learning model that predicts TCR–MHC compatibility directly from amino acid sequence. By representing MHC alleles as pseudosequences[21], it utilizes shared features between related

alleles to generalize beyond the alleles seen in training. This is essential given that the HLA system comprises thousands of alleles[4], only a fraction of which have substantial representation in experimental datasets.

TRIOPS introduces three innovations: cross-reactivity–aware negative sampling by MHC sequence similarity, a dual-branch convolutional architecture with loss reweighting and contrastive regularization, and loci-specific False Positive Rate (FPR) calibration. Applied to 8,906 TCGA pan-cancer samples, TRIOPS predictions recovered an association between TCR engagement and allele-specific HLA loss. This mirrors an association between HLA allele loss and neoantigen presentation capacity that McGranahan et al.[22] reported using predicted neoantigen binding and matched tumor–normal exomes.

Results

TRIOPS predicts TCR-HLA restriction from amino acid sequence alone

TRIOPS predicts TCR–HLA restriction directly from amino acid sequence, producing a per-pair compatibility score and a learned joint embedding (Fig. 1a). The model was trained on 61,106 unique TCR–HLA pairs with experimentally validated binding, aggregated from five public databases and harmonized to IMGT nomenclature[23–28]. TRIOPS employs a dual-branch architecture in which TCR and HLA representations are learned independently before their interaction is modeled (Fig. 1b), trained with an asymmetric focal loss and contrastive regularization of positive-pair embeddings (Methods). TCR sequences are encoded as the concatenated CDR1–CDR2–CDR3 loops of both chains, and the model accepts paired TCRαβ and single-chain inputs (Supplementary Fig. 1).

Negatives were sampled from mismatched alleles at a pseudosequence Levenshtein distance (d) of at least 10 amino acids from a TCR's true restricting allele (Supplementary Fig. 2). The threshold follows from two independent observations. Pseudosequence similarity cluster separation plateaued at d = 10–12, where intracluster pairs remained more similar than intercluster pairs (Supplementary Fig. 2b,c), and normalized overlap among allele pairs sharing TCRs persisted at short and intermediate distances but fell to near zero at d ≥ 10 (Supplementary Fig. 2d). Mismatched pairs at or beyond this distance are therefore unlikely to be cross-reactive and can be treated as negatives. A per-allele top-off procedure, applied to

balance the negative-to-positive ratio, relaxed the threshold to d ≥ 9 for a minority of negatives (Methods).

Performance holds across HLA class I, class II and TCR single chain input

We evaluated TRIOPS by six-fold cross-validation, with all metrics computed from ensemble predictions (mean interaction score across folds). On the held-out evaluation set, per allele AUCs for paired input ranged from 0.82 to 1.00, with the majority above 0.89 (Fig. 2a). Class II alleles were predicted with comparable accuracy to class I (per-allele AUC 0.89–1.00 versus 0.82–0.99; Fig. 2a) despite the marked training-set imbalance in favor of class I (Supplementary Fig. 3).

Pooled across alleles, paired TCRαβ inputs achieved the highest ROC-AUC (0.97), followed by TCRα-only (0.95) and TCRβ-only (0.92) (Fig. 2b). Both single-chain configurations remained predictive, indicating that each chain carries information relevant to restriction.

To assess generalization to receptors not represented in training, we evaluated TRIOPS on the IMMREP 2023[29] TCRαβ benchmark, restricted to TCRs absent from every training fold and the held-out evaluation dataset. On these unseen receptors, paired αβ inputs reached a ROC-AUC of 0.81 (Fig. 2c), and a precision–recall AUC of 0.61 (Fig. 2d).

TRIOPS generalizes to HLA alleles absent from training

To assess generalization to unseen HLA alleles, 15 pseudosequences, corresponding to 16 HLA alleles, were withheld across all six folds while their associated TCRs were retained, thereby isolating MHC novelty from receptor novelty (n = 56,703 pairs, 9.3% binders; Supplementary Fig. 4). The six-fold ensemble achieved a ROC-AUC of 0.92 on this holdout. Per-allele performance ranged from 0.43 (HLA-A*25:01) to 1.00 (HLA-A*02:56 and HLA-C*04:01), and the allele-balanced macro-average AUC was 0.87.

TRIOPS exceeds DePTH2.0 predictions, including after retraining

The public implementation of DePTH2.0 accepts only TCRβ and a fixed set of 85 class I and 250 class II alleles, whereas TRIOPS represents each allele by its pseudosequence, extending

to alleles absent from training. Across the alleles that DePTH2.0 accepts, TRIOPS exceeds DePTH2.0 AUC at every HLA allele (TRIOPS 0.82–1.00 for TCRαβ and 0.68–0.96 for TCRβ-only; DePTH2.0 0.37–0.81; Fig. 2a). On the IMMREP benchmark of unseen TCRs[29], DePTH2.0 reached a ROC-AUC of 0.52 and a precision–recall AUC of 0.26 for TCRβ, against 0.81 and 0.61 for TRIOPS αβ (Fig. 2c,d). DePTH2.0 provides no paired-chain mode, so this comparison reports additional TRIOPS capability rather than a matched contrast.

Because DePTH2.0 was trained on co-occurrence-derived TCR-HLA associations rather than validated binding, these comparisons evaluate it outside its training regime. To separate architecture from training signal, we retrained DePTH2.0 on the TRIOPS training set, restricted to TCRβ, and evaluated both models on identical held-out pairs with all training–evaluation overlap removed. Retraining substantially improved DePTH2.0: its TCRβ ROC-AUC rose from 0.52 to 0.83 and its precision–recall AUC from 0.23 to 0.62 (Fig. 2c,d). A gap nonetheless persisted, with TRIOPS reaching a TCRβ ROC-AUC of 0.91 and precision–recall AUC of 0.81, exceeding retrained DePTH2.0 by 0.08 and 0.19 (Fig. 2c,d).

Cross-reactivity-aware negatives and asymmetric focal loss drive accuracy

We ablated each design component in turn to isolate its contribution to model performance: the negative-sampling strategy, the loss, the contrastive regularizer, and the cross-attention gate. Each was compared against a full model retrained in the same environment on identical cross-validation splits and seeds, at three seed replicates per arm. Replacing cross-reactivity-aware negatives with random negatives reduced AUC by 0.133 (paired-t $P < 0.001$, three seeds; a multi-factor comparison, Supplementary Fig. 5) and lowered per-allele AUC at every allele (Supplementary Fig. 5a,b). Removing the asymmetric focal loss produced the largest degradation ($\Delta AUC = -0.351$, $P < 0.001$), concentrated at HLA-A and HLA-C, negligible at HLA-B, and mixed across class II (Supplementary Fig. 5a,b). Disabling positive-pull regularization left pooled AUC unchanged ($\Delta AUC = +0.001$, not significant) but degraded the learned representation: allele-level embedding separation fell from 2.58 to 1.62 and class-level separation from 1.65 to 1.25 (Supplementary Fig. 5c), and within-allele cosine distances, tightly concentrated near zero in the full model, became broadly dispersed without the regularizer (Supplementary Fig. 5d). The cross-attention gate converged near-closed and did not contribute to accuracy (Supplementary Fig. 5e).

Learned representations recover HLA allelic structure and MHC contacting domains

TCR–HLA joint embeddings from positive validation pairs separated by HLA gene group. Across the 23 alleles with at least 20 representatives, embedding coherence exceeded the permutation null for every metric: kNN purity 0.989 (null 0.284), adjusted Rand index (ARI) 0.224, normalized mutual information (NMI) 0.639 (null=0.0108) and silhouette coefficient 0.357 (Fig. 3abc). The ratio of between-cluster to within-cluster distance was 2.62 at the allele level (95% CI 2.568 to 2.667) and 1.83 at the class level (95% CI 1.796 to 1.861), both exceeding the no-separation value of 1.0 (Fig. 3d). Cosine-distances increased with allelic and class separation: within-allele (median 0.016), between-allele within-class (0.049), and between-class (0.081) (Fig. 3e).

Positional saliency analysis on the TCR input data showed that, for paired inputs, CDR3α and CDR3β carried the highest average gradients, followed by CDR2α and CDR2β (Fig. 3f). Saliency for TCRβ-only inputs concentrated in CDR3β and CDR2β, and for TCRα-only inputs in CDR3α and CDR2α. HLA pseudosequence saliency was allele-specific, with HLA-A*02:01 and HLA-B*08:01 both showing a dominant peak at pseudosequence position 1 (Fig. 3g).

Predicted class I restriction recapitulates clonal dominance

We applied TRIOPS to a TCRβ repertoire dataset reconstructed with TRUST4[30] from 8,906 bulk tumor RNA-seq samples in the TCGA Pan-Cancer cohort[31] using consensus HLA genotypes[32] and allele-specific expression quantified using arcasHLA-quant[33]. Each TCRβ clonotype was scored against every HLA allele carried by the patient, calibrated by a locus-specific 20% FPR offset, and assigned to its highest-margin allele (Methods, Supplementary Fig. 6).

TRIOPS produced a 192-dimensional joint TCR·MHC embedding for each of 417,400 argmax-assigned TCR–HLA pairs. Visualized by t-SNE (50,000-pair proportional subsample), the embeddings separated by MHC class (Fig. 4a) and resolved into allele-specific clusters (Fig. 4b). HLA-A contributes the largest single-locus share (median ≈ 0.37) (Fig. 4c), a skew consistent with the detection limits of unenriched repertoire reconstruction. Predicted class I restricted clonotypes were also retained at higher confidence margins than predicted class II (median class-level margin 0.115 vs 0.089; allele-level 0.098 vs 0.032), the gap widening at the allele level (Fig. 4d).

Within-HLA class Simpson dominance was higher for predicted class I pairs than class II pairs (Cliff's δ = 0.797; Wilcoxon $P < 10^{-15}$; n = 8,900 patients, CLES = 89.8%; Fig. 4e), recapitulating the general observation of greater clonal expansion of $CD8^+$ compared to $CD4^+$ repertoires in tumors[34]. In lung adenocarcinoma, clonotype counts are correlated with CIBERSORT-estimated T cell proportions (MHCI r=0.585, MHCII r=0.603) (Supplementary Fig. 7). Per-patient assignments are shown in Fig. 4f–h.

Across cancers, allele specific HLA expression imbalance is concordant with TCR-engagement, except at HLA-C

Using the TCGA pan-cancer TRIOPS predictions, we quantified each HLA allele's TCR engagement within each patient as the read-weighted abundance of the clonotypes assigned to it, and defined the more engaged allele of each heterozygous pair as the engaged one. Homozygous loci, which present a single allele and admit no lost-versus-retained contrast, were excluded. Allelic expression was measured by allele-resolved RNA-seq B-allele frequency (BAF), a readout that captures allele-specific loss at the level of expression, encompassing both genomic loss and transcriptional silencing.

For each cancer type, we summarized the engagement–loss concordance c, the probability that the more-engaged allele of a pair is also the under-expressed one (null 0.5; Firth regression adjusting for tumor purity and read depth, one fit per cancer, aggregated across cancers on the logit scale; Methods). After filtering, 5,233 patients contributed 24,936 informative pairs across 23 cancer types. Because a tumor without functional B2M presents no surface class I and is therefore not under allele-specific class I selection, the 65 such patients were excluded from the class I estimates. Concordance was strongest at HLA-B (c = 0.65; 23/23 cancers lost-engaged), unanimous but weaker at HLA-A (c = 0.60; 23/23), and reversed at HLA-C (c = 0.37; 0/23), where the engaged allele was preferentially retained (Fig. 5a).

We next designated the lost allele of each informative pair by expression alone (BAF < 0.5), independently of engagement, and compared TCRβ engagement between the lost and retained allele within individual tumors (Fig. 5b). Engagement was higher at the lost allele in 21/23 (HLA-A) and 22/22 (HLA-B) cancers (1 tie excluded), and in 0/23 at HLA-C, a reversal that holds each pair's tumor context fixed, independent of the across-cancer aggregation. The two

HLA-A exceptions (Fig. 5b), THCA and READ, are both cancers in which little immunoediting signal is expected due to low mutational burden and hence few neoantigens: THCA carries one of the lowest mutational burdens in TCGA[35], READ is a small cohort in which highly-mutated tumors are rare[36]. Resolved by individual cancer type (Fig. 5c), the effect was broadly distributed across cancers, with per-cancer estimates reaching $q < 0.05$ (Benjamini–Hochberg) at HLA-A and HLA-B in 16 and 19 of 23 cancers respectively.

Discussion

TRIOPS predicts TCR–HLA restriction directly from the corresponding TCR and allele amino acid sequences, without peptide information. Applied to bulk tumor repertoires, it recovers a locus-biased relationship between TCR engagement and allele-specific HLA expression. On experimentally validated binding data, TRIOPS outperformed DePTH2.0 and generalized to unseen receptors. The advantage persisted after retraining DePTH2.0 on the same pairs, isolating architecture from differences in training[17]. Replacing pseudosequence-distance negatives with random negatives, a comparison that also changes the training pool (Supplementary Fig. 5), reduced AUC by 0.133, consistent with the model learning graded relationships among related alleles rather than hard allele boundaries.

Loss of HLA class I expression is an established mechanism of immune escape in cancer[37]. Allele-specific genomic loss, in particular, is widespread: across a pan-cancer cohort of 83,644 samples, Montesion et al.[38] found somatic HLA-I loss of heterozygosity in 17% of tumors. Using matched tumor-normal exomes, McGranahan et al.[22] reported enrichment of subclonal neoantigens in NSCLC predicted to bind to the lost HLA allele, which they read as evidence that immune targeting shapes allele-specific loss. Consistent with this, TRIOPS-predicted engagement at HLA-A and HLA-B was concentrated on the under-expressed allele of each heterozygous pair.

The two analyses address complementary components of the same TCR-pMHC interaction: one ranks alleles by predicted neoepitope repertoire, the other by predicted receptor engagement. TRIOPS was trained exclusively on experimentally validated TCR–HLA binding pairs, with no exposure to allelic expression, to individual HLA genotypes, or to loss of heterozygosity, so the observed association cannot be attributed to information present during training.

Across 6,319 tumors, Martínez-Jiménez et al.[39] found that focal HLA-I loss of heterozygosity preferentially eliminates the allele presenting the largest predicted neoepitope repertoire. We detect the same signal in TCGA, where samples are predominantly treatment-naive primary tumors. RNA-level measurement explains this, since it resolves editing that copy-number analysis misses: allele-specific HLA LOH is largely subclonal[22], and allele-specific expression loss can also arise without copy-number change, through transcriptional or epigenetic silencing.

At HLA-C, the engaged allele was retained (Fig. 5a,b). Three features of this locus, absent at HLA-A and HLA-B, each explain the reversal on their own. Expression varies across HLA-C alleles, in part through 3′ UTR regulation by miR-148a[40], so allelic imbalance here need not reflect somatic loss. HLA-C is also the dominant ligand for inhibitory killer-cell immunoglobulin-like receptors, and its low surface expression marks a locus whose primary role is NK tuning rather than antigen presentation[41]. DeWitt et al.[42] reported weaker TCR–HLA-C association than at other class-I loci. And because NK cells are licensed by self class-I MHC and become competent to respond when a tuned allele is lost[43,44], losing HLA-C would expose the cell to missing-self lysis. The trade-off disappears in B2M loss, where total class I loss removes both T cell recognition and the inhibitory ligands for NK licensing, and escape is achieved despite NK exposure[45].

TRIOPS outputs are predictions conditioned on the recovered TCR repertoire rather than the repertoire in vivo. The predicted engagement landscape is dominated by HLA-A, which we attribute to the detection-floor of untargeted sequence assembly. TCGA tumor RNA-seq is unenriched for TCRs (locus coverage ≈0.04×)[46]. TRUST4 recovery scales with locus expression (Spearman $r \approx 0.35$)[30], and reliable detection is confined to the ~0.1–1% repertoire-frequency range[15]. An allele binding a broader peptide repertoire therefore distributes its response across more low-frequency clonotypes and is preferentially under-recovered. HLA-B alleles bind a broader peptide repertoire than HLA-A (~430 peptides versus ~212 on average)[47,48]. The HLA-A dominance observed is therefore confounded with a recovery floor, and cannot be purely read as evidence of CD8-driven selection. The class I versus class II margin gap (Fig. 4d) is open to the same dual reading, since class II restriction is supported by fewer labeled pairs in training and less expanded clonotypes fall below the recovery floor described above (Fig. 4e).

TRIOPS predicts restriction from either TCR (α or β) chain alone, extending its application to the unpaired repertoires that constitute most existing data in cancer, autoimmunity, and vaccine

response[49]. TRIOPS supports two translational applications. TRIOPS pairs a patient's full TCR repertoire against their genotype to assign each TCR clonotype to its restricting HLA allele and identify dominant restricting alleles, information inaccessible by bulk deconvolution[50]. It can also rank all repertoire TCRs against a chosen allele, pre-filtering candidates for deorphanization against neoantigens predicted to bind that allele. The second use addresses a concrete bottleneck: ~80% of TCR-T trials target HLA-A*02[9], and extending TCR restriction prediction across diverse alleles could broaden eligibility and reduce disparities in the development of and access to TCR-based therapeutics.

Methods

Data collection and harmonization

Experimentally validated TCR–pMHC binding data were downloaded from TCR3d[23], VDJdb[24], IEDB[25], McPAS-TCR[26], and PATCRdb[27] (accessed 2024). Records were filtered to human TCRs with confirmed binding to human HLA alleles (HLA-A, -B, -C, -DR, -DP, -DQ). All sequences were harmonized to IMGT nomenclature[28]. CDR1 and CDR2 sequences were extracted from IMGT-gapped V-gene reference sequences and extended with flanking framework residues to 10-amino-acid fixed-length representations. CDR3 sequences were extracted according to IMGT definitions, truncated at the conserved F/W of the F-G-X-G motif. CDR loops were concatenated with hyphen delimiters (TCRα = CDR1α-CDR2α-CDR3α; TCRβ = CDR1β-CDR2β-CDR3β). Missing chains were encoded as “X-X-X”. MHC alleles were represented using 34-residue NetMHCpan[21] pseudosequences (Supplementary Fig. 1). After deduplication, the positive dataset comprised 61,106 positives in the training table (129,058 after chain masking).

HLA pseudosequence similarity and shared TCR usage

Pairwise Levenshtein distances were computed and alleles grouped by average-linkage hierarchical clustering. Group coherence was assessed across cut heights from 4 to 12 using a separation ratio R(d), the median intercluster distance divided by the median intracluster distance (95% confidence intervals from 200 bootstrap replicates). Shared TCR usage was measured between HLA alleles in our dataset, for the α chain, β chain, or both chains jointly. For every allele pair sharing at least one identical TCR, a normalized overlap was computed as the number of shared TCRs divided by the summed sizes of the two TCR sets, averaged across sharing pairs at each integer Levenshtein distance.

Negative data generation

Negative examples were generated by permuting the allele while retaining the TCR sequence. Candidates were rejected if the Levenshtein distance between original and replacement pseudosequences was ≤9 (retaining only distance ≥10). Sampling was stratified to approximately 65% cross-gene and 35% within-gene negatives. A top-off procedure raised each allele to at least two negatives per positive, with the Levenshtein distance threshold relaxed to ≥9; for HLA-A*02:01, which accounts for 48% of positives, the eligible candidate pool was exhausted at 1.07:1. The final dataset comprised 281,265 pairs (positive-to-negative ratio 1:3.60), or 595,141 after chain masking (1:3.61).

Data augmentation and partitioning

Data were stratified by allele, label, and chain status into a fixed held-out set, with remaining data divided into six cross-validation folds (~80/20 train/validation). Paired TCRαβ entries were then augmented by chain masking (α-only and β-only versions), applied independently within each partition so that masked copies of one pair could not straddle a train/validation boundary. Held-out pairs were disjoint from training before masking. Because deduplication preceded masking and keyed on the two chains separately, two receptors sharing one chain and restricting allele were retained as distinct records, and masking reduced them to identical single-chain representations; a masked copy of a held-out pair could therefore appear among the training folds. Generalization to entirely unseen TCRs was evaluated on the IMMREP 2023 benchmark and to unseen alleles on a pseudosequence-disjoint holdout; neither partition is affected by chain masking.

Model architecture

TCR and MHC sequences were encoded using a 22-character vocabulary and embedded into 64-dimensional representations. Each branch processes sequences through three convolutional blocks (kernel size 3, filters 64→128→256, batch normalization, max pooling after blocks 1–2). Cross-attention projects TCR features to queries and MHC features to keys/values via 128-filter convolutions; scaled dot-product attention with dropout ($p = 0.1$) is combined with the original TCR features through a learned residual gate (init 0.2) and layer normalization. Attention pooling aggregates TCR features; GeM pooling (trainable p, init 3.5) aggregates MHC features. The 64-dimensional TCR and MHC embeddings and their element-wise product form a 192-dimensional fusion vector, processed through a 256-unit dense layer (ReLU, dropout 0.3)

and sigmoid output.

Training procedure

The model was trained with asymmetric focal loss (γ_{pos} = 0.05, γ_{neg} = 1.5) on soft labels (0.25 within-gene negatives, 0.0 cross-gene). Positives were weighted at 1.6× and negatives at 0.8×, with class-balanced allele upweighting (β = 0.999, cap 3.0×). The objective included a PositivePull regularization (λ = 0.05). The model was trained with the Adam optimizer (lr 10^{-3}, gradient clip 1.0, batch 256), up to 60 epochs with early stopping (patience 8) on validation AUC. LR reduced 0.5× on plateau (patience 4, min 10^{-5}). Six-fold cross validation was run on 90% of data; all folds share the same 10% held-out set.

Evaluation metrics

ROC-AUC and precision–recall AUC were computed from ensemble predictions (mean across six folds), stratified by allele, chain status, and gene group. 95% confidence intervals were estimated by bootstrapping (2,000 iterations). Generalization was assessed on the IMMREP 2023 benchmark, restricted to TCRs absent from all training folds and the held-out set.

Comparison with DePTH2.0

DePTH2.0[17] was run with published pretrained weights and applied on TCRβ sequences (≤27 aa, standard residues only) from the held-out set and IMMREP benchmark, using separate class I and class II models. TRIOPS was evaluated on the identical sequences using TCRβ-only and paired chain inputs. DePTH2.0 provides no paired-chain mode, so paired-input results report additional TRIOPS capability rather than a matched comparison. AUC was computed identically for both methods, including positive/negative definition, allele stratification, and 2,000-iteration bootstrap CIs, reported separately for class I and class II.

Retraining DePTH2.0 on the TRIOPS training data

To disentangle architecture from training signal, we retrained DePTH2.0 on the same experimentally validated data used for TRIOPS. The six cross-validation folds (training and validation partitions) were pooled and reduced to TCRβ inputs by discarding α-only records and rows lacking a β chain or β V-gene, then deduplicated to unique (β V-gene, CDR3β, allele) triples, with positive labels taking precedence where a triple appeared with both labels. To guarantee a leak-free comparison, every triple shared between the training pool and the frozen held-out evaluation set was removed from both, leaving disjoint (β V-gene, CDR3β, allele) sets

and an evaluation subset scored identically by both models. Alleles and β V-genes were remapped to DePTH2.0's encoders, and CDR3β sequences were restricted to at most 27 standard-residue amino acids.

The retained pairs were split 75/25 into training and validation, stratified by label and processed separately for HLA class I and class II. For each class we trained a 20-seed ensemble with DePTH2.0's default configuration (one-hot encoding, two dense layers of 64 and 16 units, dropout 0.2, learning rate 10^{-4}) and averaged the per-seed scores. TRIOPS was scored on the identical leak-free TCRβ pairs; because DePTH2.0 provides no paired-chain mode, TRIOPS paired-chain results are reported as an additional reference rather than a matched comparison.

Unseen-HLA generalization

For this analysis the data were repartitioned separately from the main split, grouping on the MHC pseudosequence (the sole input to the MHC branch) rather than on individual pairs. Whole pseudosequence groups were assigned at random to a fixed held-out set comprising ~10% of records; groups each exceeding 10% of the data were retained in training to preserve signal. The held-out set comprised 56,703 pairs (9.3% binders) spanning 15 pseudosequences and 16 HLA alleles. The remaining data was divided into six cross-validation folds (k = 6, seed 42), used for early stopping only, so inner-fold validation could share alleles with training while the external held-out set remained pseudosequence-disjoint. Architecture, encoding, masking, loss, and optimization matched the primary model. Held-out pairs were scored by the six-fold ensemble; ROC AUC was computed on pooled pairs and per allele. The unweighted macro-average AUC across alleles is reported as the allele-balanced summary.

Component ablations

To quantify the contribution of each design choice, we trained three single-component ablations alongside a matched full-model control, each following the full TRIOPS protocol (six-fold cross-validation; 60 epochs with early stopping at patience 8; batch size 256; base learning rate 10^{-3}). The ablations were: (i) random negatives, replacing the cross-reactivity-aware sampler (HLA-pseudosequence Levenshtein distance ≥ 10) with negatives drawn uniformly at random over observed alleles; (ii) no focal loss, replacing the asymmetric focal loss with plain binary cross-entropy; and (iii) no positive-pull, disabling the positive-pull contrastive regularizer ($\lambda = 0$, versus $\lambda = 0.05$ in the full model). To prevent library- and hardware-version differences from contaminating the effect estimates, each ablation was compared against a full model retrained

in the same environment on identical cross-validation splits and per-fold random seeds rather than against the released checkpoint.

Because this matched control used the fast nondeterministic GPU kernels, its pooled held-out AUC differed from the released checkpoint by 0.005 (0.946 versus 0.941); measuring every ablation effect relative to the matched control removes this residual environment difference from the reported deltas. Each arm was trained at three independent seed replicates; ablation effects on pooled AUC were reported as the mean ± s.d. of the matched-seed difference from the full model and tested by a matched-seed paired t-test, with case-level significance additionally assessed by DeLong's test on one replicate.

All arms were scored on the frozen held-out set (the 10% of pairs withheld from every fold; n = 59,055 pairs, 21.8% binders). Per-allele AUC (Supplementary Fig. 5a) was computed for alleles with ≥ 15 binding and ≥ 15 non-binding held-out pairs. To isolate the effect of engineered negatives on apparent performance, the random arm was additionally scored against random negatives over the same held-out TCRs, so that the two evaluation sets differ only in negative source.

Embedding geometry (Supplementary Fig. 5c,d) was assessed on fold-4 fusion embeddings of held-out binding pairs: from randomly sampled embedding pairs we computed allele-level separation (mean between-allele over mean within-allele cosine distance), class-level separation (mean between-class over mean within- and between-allele cosine distance), each with bootstrap 95% confidence intervals, and the full within-allele cosine-distance distribution.

To characterize the cross-attention residual gate (Supplementary Fig. 5e), the scalar gate logit was overwritten in each trained fold model at inference, with all other weights held fixed, and the six-fold ensemble was rescored on the frozen held-out set at eight values of sigmoid(g) spanning 0 to 1.

HLA gene loci-specific false positive score calibration

For each of the six HLA gene loci (A, B, C, DR, DP, DQ), an ROC curve was constructed from each fold model's predictions on the frozen held-out evaluation set (n = 59,055), and the threshold yielding a 20% false-positive rate was found by linear interpolation. We chose this operating point to equalize the per-locus false-positive rate rather than sensitivity. Because the

six gene groups occupy different score scales, their raw thresholds are not directly comparable; subtracting each locus's 20% FPR threshold rescales every score as a margin above a shared error rate, making the rectified margins comparable across loci. The HLA gene locus-specific threshold was the mean of the six per-fold thresholds. For each TCR–allele pair, the locus threshold was subtracted from the six-fold ensemble mean score and rectified, max(0, mean score − threshold), with the gene group inferred from the allele. Each clonotype was assigned to the allele with the highest positive margin among those the patient carries; clonotypes with no positive margin were left unassigned. The per-gene-group threshold is a reference operating point, not a false-positive guarantee on patient repertoires; subtracting it aligns class I and class II at a common operating point, after which the engaged allele is selected among those a patient carries.

Embedding visualization and coherence analysis

Fold-4 fusion embeddings were extracted for positive validation examples (binder score ≥ 0.5) from the joint TCR–MHC concatenation layer. Embeddings were L2-normalized and projected to two dimensions with t-SNE using a cosine metric, perplexity of 30, and PCA initialization. Embedding coherence was quantified with four metrics computed on alleles represented by at least 20 examples: kNN purity (k = 15), the adjusted Rand index (ARI) and normalized mutual information (NMI) of a 30-cluster k-means partition relative to allele identity, and the cosine silhouette score. Each metric was compared against a null distribution generated by 1,000 random permutations of allele labels. Separation ratios were defined as the mean between-group cosine distance divided by the mean within-group cosine distance, computed at both the allele and MHC-class levels, with 95% confidence intervals obtained by bootstrap resampling.

Positional saliency analysis

Input-embedding gradients were computed from the fold-4 model as the magnitude of the gradient of the predicted binding score with respect to the input embeddings, averaged across embedding dimensions and across examples. TCR saliency was computed per sequence position and stratified by MHC class (class I and class II) and by chain status (paired TCRαβ, TCRβ only, and TCRα only). MHC saliency was computed per pseudosequence position for three representative alleles, HLA-A*02:01, HLA-B*08:01, and HLA-DPA1*01:03/DPB1*04:01, with saliency profiles normalized to their per-allele maximum.

TCGA pan-cancer analysis

TCGA bulk RNA-seq data were obtained from the NCI Genomic Data Commons[51] (https://portal.gdc.cancer.gov). TCGA TCRβ repertoire reconstruction data was obtained from Dr. Li Song. All TCRβ clonotypes were paired using TRIOPS with each patient's HLA alleles, and six-fold ensemble predictions were calibrated by HLA gene loci-specific FPR thresholds, with each clonotype assigned to its argmax allele. Immune cell abundances were estimated from LUAD RNAseq data with CIBERSORT[50] using the LM22 signature matrix, applied to LUAD RNA-seq quantified as transcripts per million (TPM). TCR-HLA engagement was correlated against total T cells (CD4+ plus CD8+) by Pearson r, both unweighted and read-count-weighted; Spearman correlations yielded similar results. Clonal dominance was quantified as Simpson's D = Σp_i^2.

Statistical analysis

Pearson correlations between total T-cell proportion and predicted class I or class II clonotype counts were computed with SciPy, with 95% confidence intervals obtained by bootstrap resampling over 10,000 iterations. Group comparisons were assessed with the Wilcoxon rank-sum test, and effect sizes were quantified with Cliff's δ. The common-language effect size (CLES) was reported as the probability that a randomly chosen class I-targeting sample shows greater dominance than a randomly chosen class II-targeting sample.

TCR engagement and HLA allele-specific loss

We applied the engagement–loss analysis to the same 8,906-sample TCGA pan-cancer TCRβ repertoire used above (Fig. 4), pairing each patient's TRIOPS predictions with HLA allele-resolved RNA-seq B-allele frequency (BAF). TCR engagement at each HLA allele was defined as the summed TCRβ read count of the clonotypes assigned to that allele, where each clonotype in a patient was assigned to the single allele with the highest calibrated TRIOPS binding margin across that patient's HLA genotype (cross-locus argmax). The continuous loss measure was loss = 1 − 2e, where e is the expressed fraction (BAF) of the more-engaged HLA allele, so that loss > 0 marks under-expression of the engaged allele and loss < 0 marks its retention. Informative events were those with |loss| > 0.15 (equivalently BAF outside 0.425 to 0.575), a threshold set to exclude near-symmetric BAF noise.

Heterozygous locus selection

Because allele-specific loss is only defined when a locus presents two distinguishable alleles,

the analysis was restricted to heterozygous HLA loci: the class I genes HLA-A, HLA-B, and HLA-C, and the class II β-chain genes HLA-DRB1, HLA-DQB1, and HLA-DPB1. Homozygous HLA loci, which present a single allele and so admit no lost-versus-retained contrast, were excluded. For HLA class II, heterozygosity and BAF were assessed on the β-chain gene of each HLA dimer, the polymorphic chain that defines the expressed allele pair.

Exclusion of tumors with truncating B2M mutations

Tumors carrying a loss-of-function *B2M* mutation were excluded from the class I analyses. *B2M* encodes the invariant class I light chain, so a tumor without functional *B2M* presents no surface class I and is subject to no allele-specific selection at HLA-A, HLA-B or HLA-C. *B2M* status was called at the patient level from the MC3 public mutation set[36], counting frame-shift insertions and deletions, nonsense, splice-site, translation start site and nonstop variants passing the MC3 FILTER field, after removal of aliquots flagged Do_not_use in the PanCanAtlas sample quality annotations. Pan-cancer, 91 of 9,808 assayed samples (0.93%) carried such a variant, 65 of them among the 5,036 analysed patients. Because *B2M* has no role in class II presentation, the exclusion was restricted to class I and every patient was retained at HLA-DRB1, HLA-DQB1 and HLA-DPB1. Mutation calls do not capture *B2M* copy loss or promoter hypermethylation, so the exclusion is partial.

Quality control and covariates

A heterozygous allele pair was retained when at least one allele carried nonzero engagement and the pair passed all quality-control filters: at least 100 combined RNA-seq reads spanning the locus, a patient repertoire of at least 10 unique TCRβ clonotypes and at least 30 total TCRβ reads, and a cancer type with at least 20 evaluable patients. All models adjusted for ABSOLUTE[52] tumor purity and $\log_{10}$ total RNA-seq read depth (per cancer distributions, Supplementary Fig. 8), mean-centered within each fit and retained only where finite with nonzero variance. Adjacent normal tissue was not used as a within-patient baseline; adjacent normal RNA-seq is available for only a minority of TCGA patients[53].

TCR-HLA engagement loss concordance model

For each cancer type we fit a Jeffreys-penalized (Firth) logistic regression of engaged-allele loss on the two centered covariates: tumor purity and RNAseq read depth. Because both covariates were centered within each cancer, the per-cancer concordance c□ = invlogit(intercept) is the purity- and depth-adjusted probability, evaluated at that cancer's mean purity and depth, that the

engaged allele is the lost one; c□ > 0.5 denotes preferential loss of engaged alleles and c□ < 0.5 preferential retention. A cancer contributed c□ only with at least 15 informative events. The pan-cancer estimate aggregates each cancer once, c = invlogit(mean of the per-cancer intercepts), with a 95% confidence interval from a t-distribution on the logit scale (J − 1 degrees of freedom, J contributing cancers) back-transformed through the inverse logit. A two-sided exact binomial sign test of the number of cancers with c□ > 0.5 against an expected 0.5 provides a reproducibility test across cancers, independent of the per-cell tests. Per cancer-by-locus significance is the penalized likelihood-ratio test of the intercept (null c□ = 0.5), with Benjamini-Hochberg control at q < 0.05 across all finite per-cancer-by-locus cells; asterisks mark class I cells passing this threshold.

Within-tumor paired analysis

For each class I locus and cancer type we compared TCRβ engagement at the lost versus retained allele of these informative pairs, with the lost allele defined directly as the under-expressed member (BAF < 0.5), independent of engagement. Per-allele engagement was winsorized at the 95th percentile, log(1 + x) transformed, averaged within each patient, and summarized as the per-cancer median across patients. The cross-cancer comparison was a two-sided exact binomial sign test of the number of cancer types with higher median engagement at the lost allele, among those with a nonzero median difference.

Software and computational environment

All models were implemented in Python 3.12.13 using TensorFlow 2.20.0 with Keras 3.13.2. Sequence distance calculations used the python-Levenshtein library. Data processing and downstream analysis used NumPy 2.0.2, pandas 2.2.2, scikit-learn 1.6.1, and SciPy 1.16.3; figures were generated with Matplotlib 3.10.0 and seaborn 0.13.2. Training and inference were performed on an NVIDIA L4 GPU (24 GB, compute capability 8.9) in Google Colab, with TensorFlow built against CUDA 12.5.1 and cuDNN 9. Claude (Anthropic) was used to assist in writing portions of the analysis and model code. All code was reviewed, tested, and validated by the authors, who take full responsibility for its correctness.

Data availability

Experimentally validated TCR–pMHC binding data were obtained from publicly available databases: TCR3d (https://tcr3d.ibbr.umd.edu), VDJdb (https://vdjdb.cdr3.net), IEDB (https://www.iedb.org), McPAS-TCR (http://friedmanlab.weizmann.ac.il/McPAS-TCR/), and

PATCRdb (https://patcrdb.org). The IMMREP 2023 benchmark dataset is available at https://github.com/viragbioinfo/IMMREP_2023_TCRSpecificity. TCGA pan-cancer bulk RNA-seq data are available through the NCI Genomic Data Commons[51] (https://portal.gdc.cancer.gov). HLA genotypes were obtained from the TCGA pan-cancer consensus HLA typing benchmark dataset[32] (https://doi.org/10.1002/1878-0261.12895). Somatic mutation calls were taken from the MC3 public MAF[36], and sample quality annotations and ABSOLUTE[52] tumor purity estimates from the PanCanAtlas release (https://gdc.cancer.gov/about-data/publications/pancanatlas). TCGA TCRβ repertoires reconstructed with TRUST4 were provided by Dr. Li Song. HLA allele-resolved expression was quantified with arcasHLA-quant[33] from chromosome 6 reads in TCGA tumor RNA-seq BAM files. The curated TCR–HLA training set (61,106 positive pairs; 595,141 pairs after masking and negative sampling), the allele-resolved B-allele frequency table, and the per-clonotype TRIOPS allele assignments underlying Figs. 4 and 5 are deposited at Zenodo (DOI to be assigned).

Code availability

Source code for model training, calibration, inference, and all analyses reproducing the figures, together with trained model weights, is provided as Supplementary Software with this submission and is available at https://github.com/jonssonlab/triops and archived at Zenodo (DOI to be assigned). The code is released for academic, research and educational use; commercial use requires a license from the University of California, Santa Cruz.

Author contributions

NRR and VDJ conceived the study. NRR designed the TRIOPS framework and curated data for the study. NRR and VDJ wrote the software, developed the methodology and conducted formal analysis. CMR performed TCGA HLA genotype calls and pan-cancer allele-specific expression quantification. LM performed deconvolution analysis and statistical analyses. VDJ supervised the work. NRR and VDJ wrote the manuscript. All authors provided feedback on manuscript and figures.

Acknowledgments

We thank the UC Santa Cruz Genomics Institute for computing resources. This work was supported by a University of California Cancer Research Coordinating Committee grant and the Hellman Fellows grant to VDJ. We thank Li Song for providing the TCGA TCR data and Joshua

Stuart for valuable feedback on the manuscript.

References

1. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
2. Lythe, G., Callard, R. E., Hoare, R. L. & Molina-París, C. How many TCR clonotypes does a body maintain? *J. Theor. Biol.* **389**, 214–224 (2016).
3. Zinkernagel, R. M. & Doherty, P. C. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* **248**, 701–702 (1974).
4. Barker, D. J. *et al.* The IPD-IMGT/HLA database: recent developments in sequence submission. *Nucleic Acids Res.* **54**, D1152–D1158 (2026).
5. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
6. Tran, E. *et al.* T-cell transfer therapy targeting mutant KRAS in cancer. *N. Engl. J. Med.* **375**, 2255–2262 (2016).
7. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
8. Tran, M. T. *et al.* T cell receptor recognition of hybrid insulin peptides bound to HLA-DQ8. *Nat. Commun.* **12**, 5110 (2021).
9. Shafer, P., Kelly, L. M. & Hoyos, V. Cancer therapy with TCR-engineered T cells: Current strategies, challenges, and prospects. *Front. Immunol.* **13**, 835762 (2022).
10. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.*

11, 1803 (2020).

11. 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).
12. Jensen, M. F. & Nielsen, M. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration. *Elife* **12**, RP93934 (2024).
13. Lu, T. *et al.* Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
14. Fast, E., Dhar, M. & Chen, B. TAPIR: a T-cell receptor language model for predicting rare and novel targets. *bioRxivorg* (2023) doi:10.1101/2023.09.12.557285.
15. Peng, K. *et al.* Rigorous benchmarking of T-cell receptor repertoire profiling methods for cancer RNA sequencing. *Brief. Bioinform.* **24**, bbad220 (2023).
16. Glazer, N., Akerman, O. & Louzoun, Y. Naive and memory T cells TCR-HLA-binding prediction. *Oxf. Open Immunol.* **3**, iqac001 (2022).
17. Li, F., Liu, S. & Sun, W. Improved deep learning prediction of TCR–HLA associations. *Stat. Biosci.* https://doi.org/10.1007/s12561-025-09506-4 (2025) doi:10.1007/s12561-025-09506-4.
18. Emerson, R. O. *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
19. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol. Today* **19**, 395–404 (1998).
20. Sidney, J., Peters, B., Frahm, N., Brander, C. & Sette, A. HLA class I supertypes: a revised and updated classification. *BMC Immunol.* **9**, 1 (2008).
21. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation. *Nucleic Acids Res.* **48**, W449–W454 (2020).

22. McGranahan, N. *et al.* Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271.e11 (2017).
23. Gowthaman, R. & Pierce, B. G. TCR3d: The T cell receptor structural repertoire database. *Bioinformatics* **35**, 5323–5325 (2019).
24. Shugay, M. *et al.* VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2018).
25. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
26. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
27. Lee, Y. *et al.* PATCRdb: Database of TCRs from data mining patent documents. *medRxiv* (2023) doi:10.1101/2023.01.05.23284150.
28. Lefranc, M.-P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–22 (2015).
29. Nielsen, M. *et al.* Lessons learned from the IMMREP23 TCR-epitope prediction challenge. *Immunoinformatics (Amst.)* **16**, 100045 (2024).
30. Song, L. *et al.* TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
31. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
32. Li, X. *et al.* Benchmarking HLA genotyping and clarifying HLA impact on survival in tumor immunotherapy. *Mol. Oncol.* **15**, 1764–1782 (2021).
33. Orenbuch, R. *et al.* arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40 (2020).
34. Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory T-cell differentiation:

implications for vaccine development. *Nat. Rev. Immunol.* **2**, 251–262 (2002).

35. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
36. Ellrott, K. *et al.* Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281.e7 (2018).
37. Aptsiauri, N., Ruiz-Cabello, F. & Garrido, F. The transition from HLA-I positive to HLA-I negative primary tumors: the road to escape from T-cell responses. *Curr. Opin. Immunol.* **51**, 123–132 (2018).
38. Montesion, M. *et al.* Somatic HLA class I loss is a widespread mechanism of immune evasion which refines the use of tumor mutational burden as a biomarker of checkpoint inhibitor response. *Cancer Discov.* **11**, 282–292 (2021).
39. Martínez-Jiménez, F. *et al.* Genetic immune escape landscape in primary and metastatic cancer. *Nat. Genet.* **55**, 820–831 (2023).
40. Kulkarni, S. *et al.* Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature* **472**, 495–498 (2011).
41. Parham, P. & Moffett, A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat. Rev. Immunol.* **13**, 133–144 (2013).
42. DeWitt, W. S., 3rd *et al.* Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* **7**, e38358 (2018).
43. Kim, S. *et al.* Licensing of natural killer cells by host major histocompatibility complex class I molecules. *Nature* **436**, 709–713 (2005).
44. Anfossi, N. *et al.* Human NK cell education by inhibitory receptors for MHC class I. *Immunity* **25**, 331–342 (2006).
45. Ljunggren, H. G. & Kärre, K. In search of the ‘missing self’: MHC molecules and NK cell recognition. *Immunol. Today* **11**, 237–244 (1990).
46. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung

adenocarcinoma. *Nature* **511**, 543–550 (2014).

47. Schellens, I. M. M. *et al.* Comprehensive analysis of the naturally processed peptide repertoire: Differences between HLA-A and B in the immunopeptidome. *PLoS One* **10**, e0136417 (2015).
48. Chappell, P. *et al.* Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife* **4**, e05345 (2015).
49. Wu, T. D. *et al.* Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* **579**, 274–278 (2020).
50. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
51. Grossman, R. L. *et al.* Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
52. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
53. Zeng, W. Z. D., Glicksberg, B. S., Li, Y. & Chen, B. Selecting precise reference normal tissue samples for cancer research using a deep learning approach. *BMC Med. Genomics* **12**, 21 (2019).

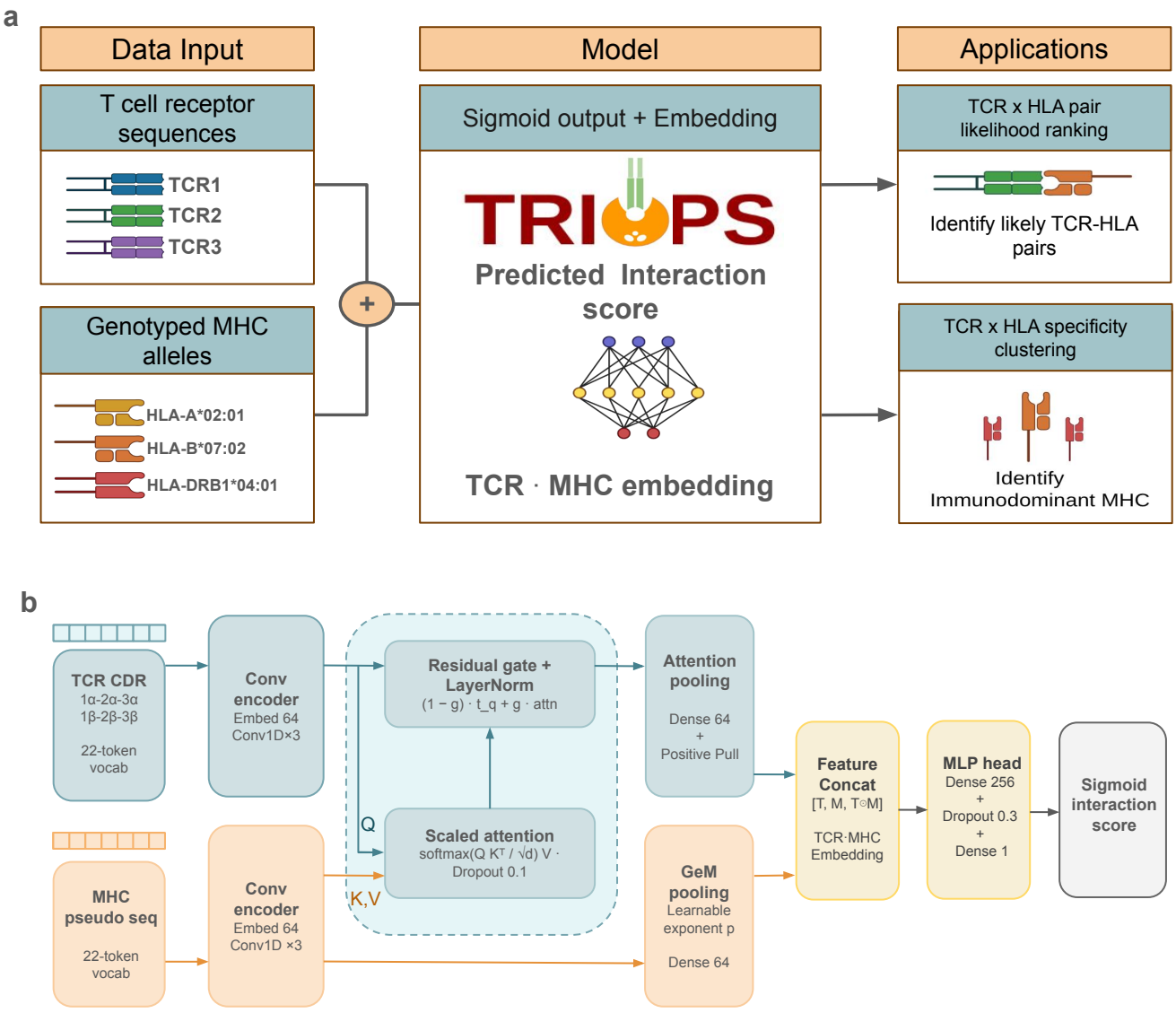


Figure 1 | Overview and architecture of the TRIOPS framework for prediction of TCR–MHC specificity. a, Conceptual overview. T cell receptor sequences and genotyped MHC alleles are paired inputs to a model that predicts binding compatibility, returning an interaction score and a joint TCR·MHC embedding. Applications include TCR × HLA pair likelihood ranking and specificity clustering to identify immunodominant MHC. b, Neural network architecture. The six TCR CDR loops and the MHC pseudosequence are tokenized, embedded, and encoded through three Conv1D blocks each. In scaled cross-attention, the TCR supplies the query (Q) and the MHC the keys (K) and values (V), so each TCR position draws on its most relevant MHC features; a learned gate then merges this output with the original TCR representation. Attention pooling (TCR) and GeM pooling (MHC) yield 64-dimensional representations that are concatenated with their element-wise product into a 192-dimensional joint embedding and passed through an MLP head with sigmoid activation to produce the interaction score.

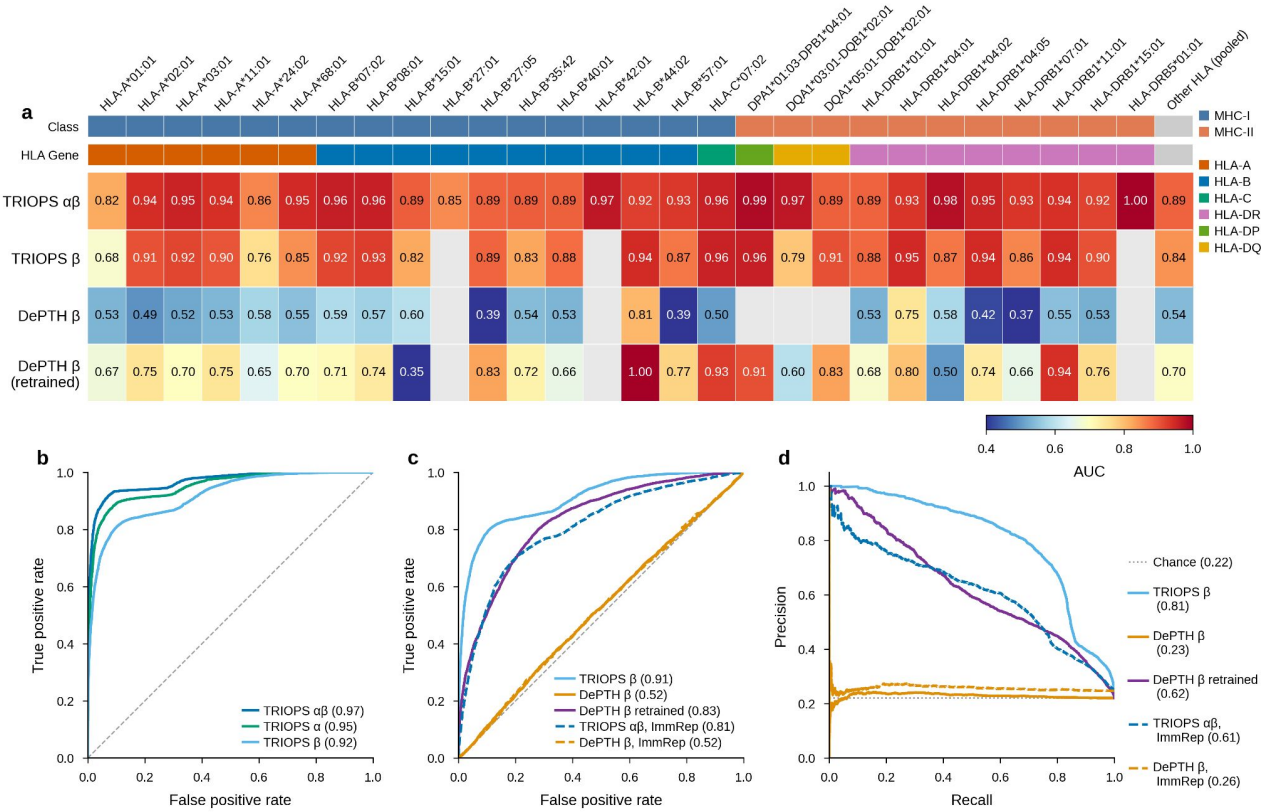


Figure 2 | Predictive performance of TRIOPS and comparison with DePTH2.0. a, Per-allele AUC heatmap for TRIOPS (paired TCRαβ and TCRβ-only), DePTH2.0 as published (TCRβ-only), and DePTH2.0 retrained on the TRIOPS training data (TCRβ-only, 1 split, 20-seed ensemble). Alleles ordered by gene group; colour strips indicate MHC class and gene. Grey cells: alleles with insufficient evaluable pairs. **b**, ROC curves by TRIOPS chain configuration, evaluated on all held-out pairs for each configuration (paired, AUC = 0.97; α-only, 0.95; β-only, 0.92). **c**, Head-to-head ROC on identical held-out pairs (solid): TRIOPS β, 0.91; DePTH β as published, 0.52; DePTH β retrained, 0.83. Dashed, IMMREP: TRIOPS αβ, 0.81; DePTH β, 0.52. The three β-only models are compared on the common subset of pairs scored by all methods with training–evaluation overlap removed; this subset differs slightly from the full β-only set in **b**, so TRIOPS β is 0.91 here versus 0.92 in **b** (same model, different evaluation pairs). **d,** Precision–recall curves for the same comparisons. Solid, held-out: TRIOPS β, AP = 0.81; DePTH β as published, 0.23; DePTH β retrained, 0.62; dotted, chance = 0.22. Dashed, IMMREP: TRIOPS αβ, 0.61; DePTH β, 0.26.

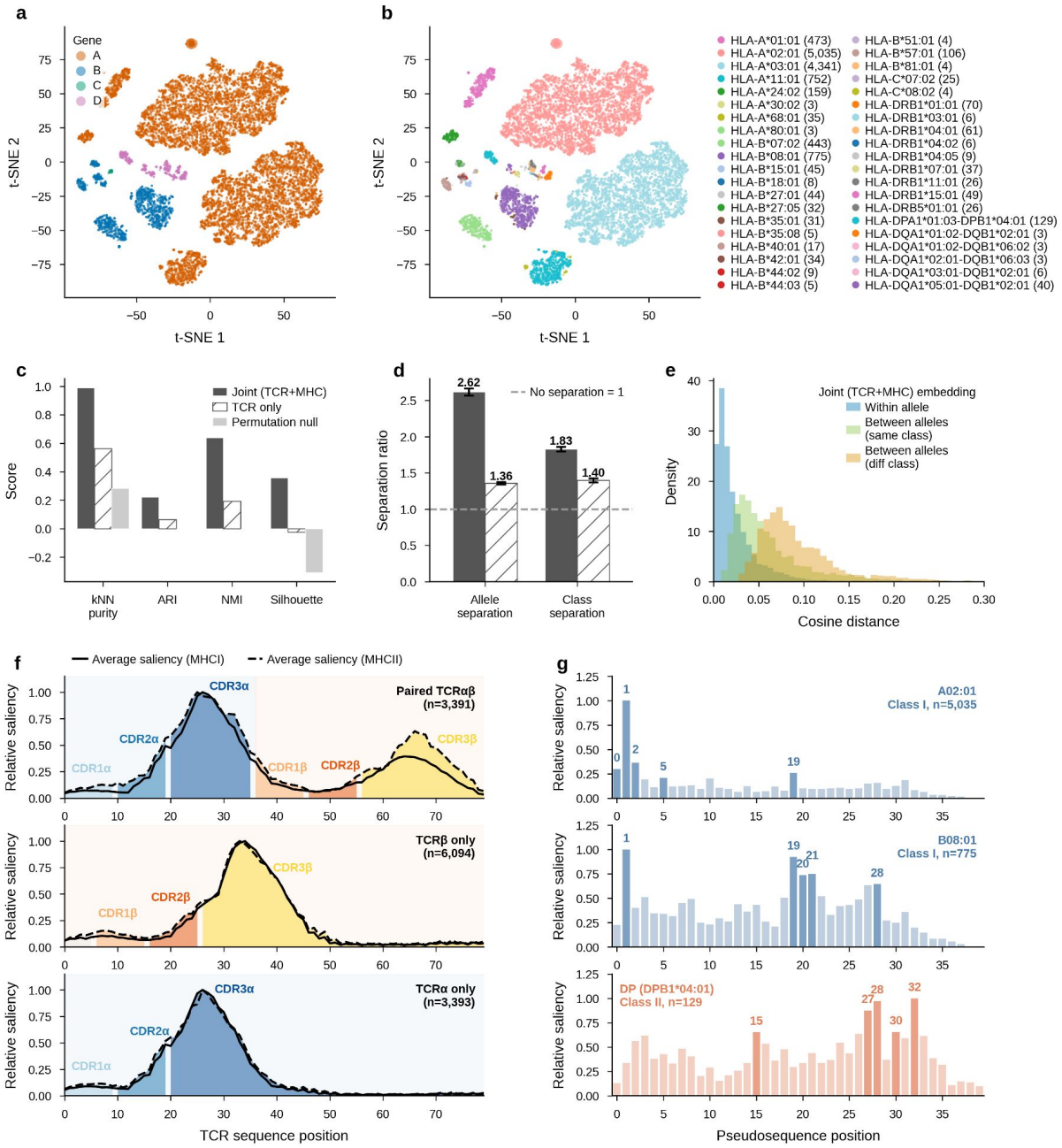


Figure 3 | TRIOPS embedding structure, coherence analysis, and positional saliency. All panels use the fold-4 model. a, t-SNE of fusion embeddings colored by HLA gene group. b, Same projection colored by individual allele (top 40 by frequency). c, Coherence metrics (kNN purity, ARI, NMI, silhouette) for the joint TCR·MHC embedding and a TCR-only embedding, each compared against the permutation null. d, Separation ratios (between- over within-group cosine distance) at the allele and class level for the joint embedding (2.62 and 1.83) and the TCR-only embedding (1.36 and 1.40); dashed line marks no separation (ratio = 1). e, TCR·MHC cosine distance distributions for within-allele, between-allele same-class, and between-class pairs. f, TCR positional saliency stratified by MHC class. Top, paired (n = 3,391); middle, β-only (n = 6,094); bottom, α-only (n = 3,393). g, MHC pseudosequence saliency for A*02:01 (n = 5,035), B*08:01 (n = 775), and DPA1*01:03-DPB1*04:01 (n = 129).

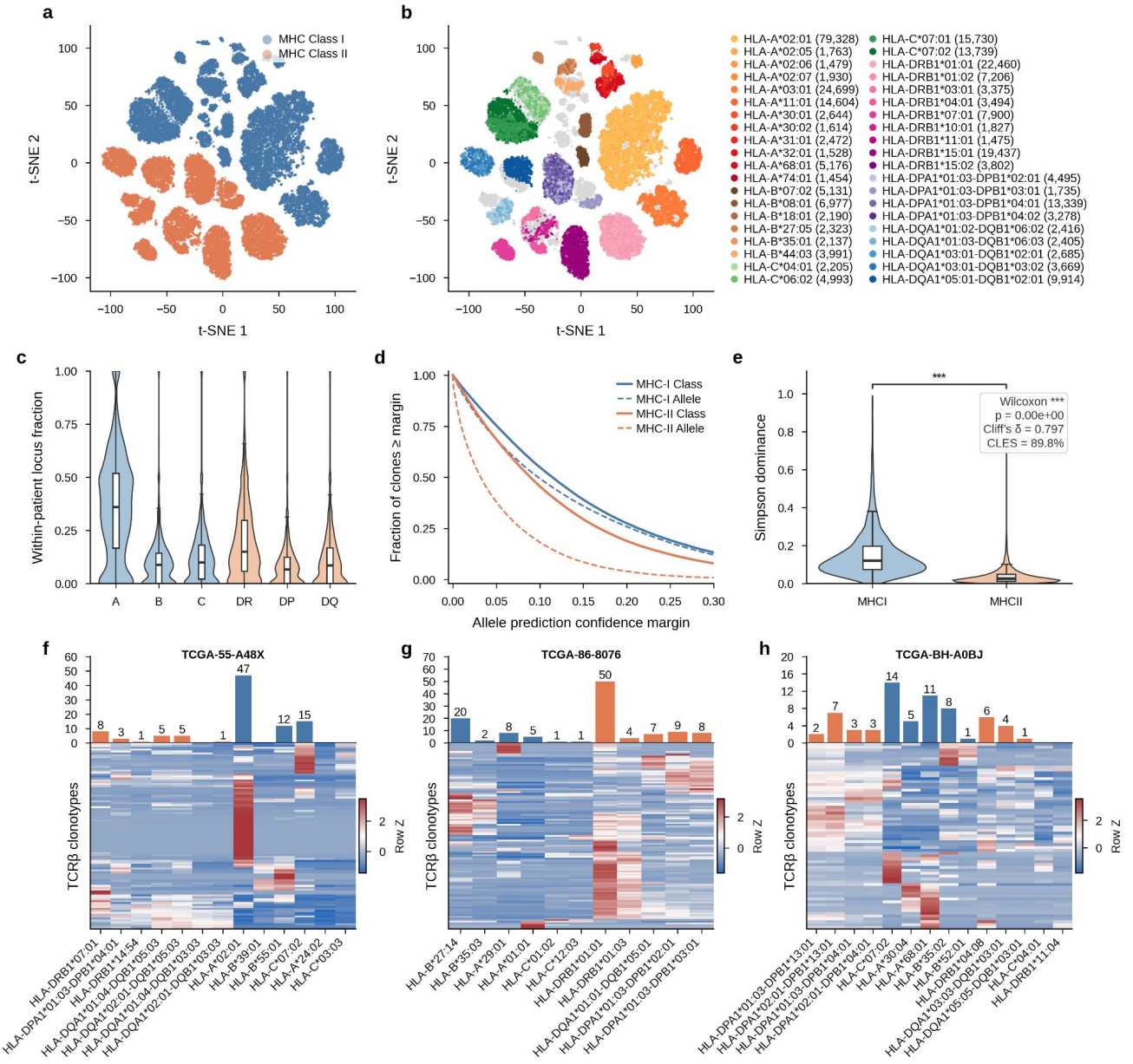


Figure 4 | TRIOPS predictions in TCGA reveal patient-specific immune specificity landscapes and recapitulate T cell biology. a, t-SNE of argmax TCRβ–HLA assignments (n ≈ 417,400 clonotype assignments across 8,900 patients) colored by MHC class. b, Same projection colored by allele, with gene-family shading. c, Within-patient locus fraction by HLA locus across n = 7,515 patients; HLA-A contributes the largest single-locus share (median ≈ 0.36). d, Fraction of assigned clonotypes whose prediction confidence margin exceeds a given threshold, for MHC-I and MHC-II assignments. Class curves (solid) give the margin between the top-ranked allele and the best allele of the opposing class; allele curves (dashed) give the margin between the top two alleles within the assigned class. Class I assignments retain higher margins than class II at both levels. e, Within-class Simpson dominance for class I– versus class II–targeting TCRβ clonotypes (Wilcoxon $P < 2.2 \times 10^{-16}$; Cliff's δ = 0.797; CLES = 89.8%). f, FPR-calibrated TRIOPS score heatmap (rows z-scored, hierarchically clustered) with allele-level clonotype counts above, for TCGA-55-A48X, a predominantly class I–restricted repertoire. g, Same for TCGA-86-8076, a predominantly class II–restricted repertoire. h, Same for TCGA-BH-A0BJ, spanning both classes.

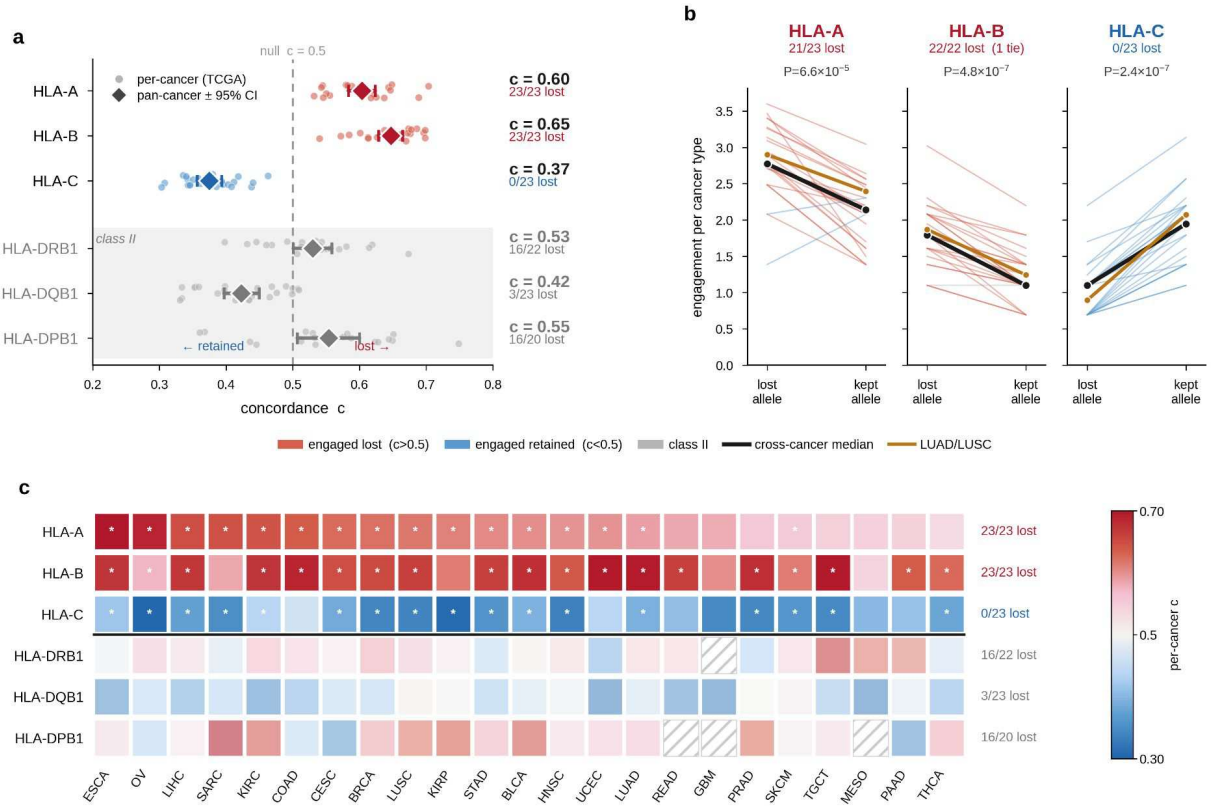
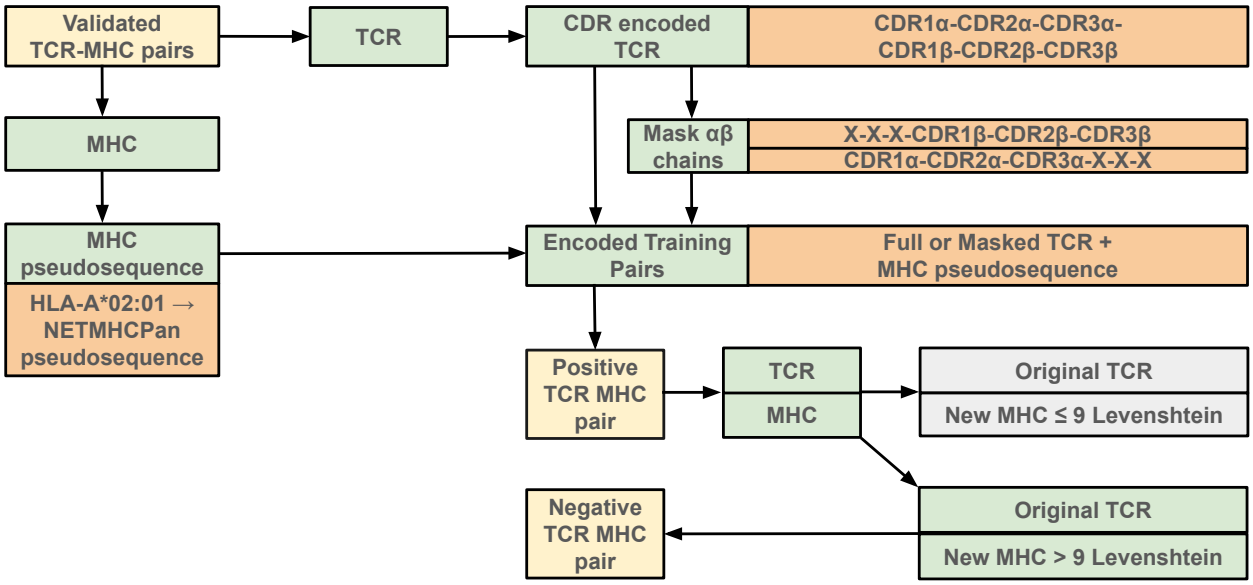
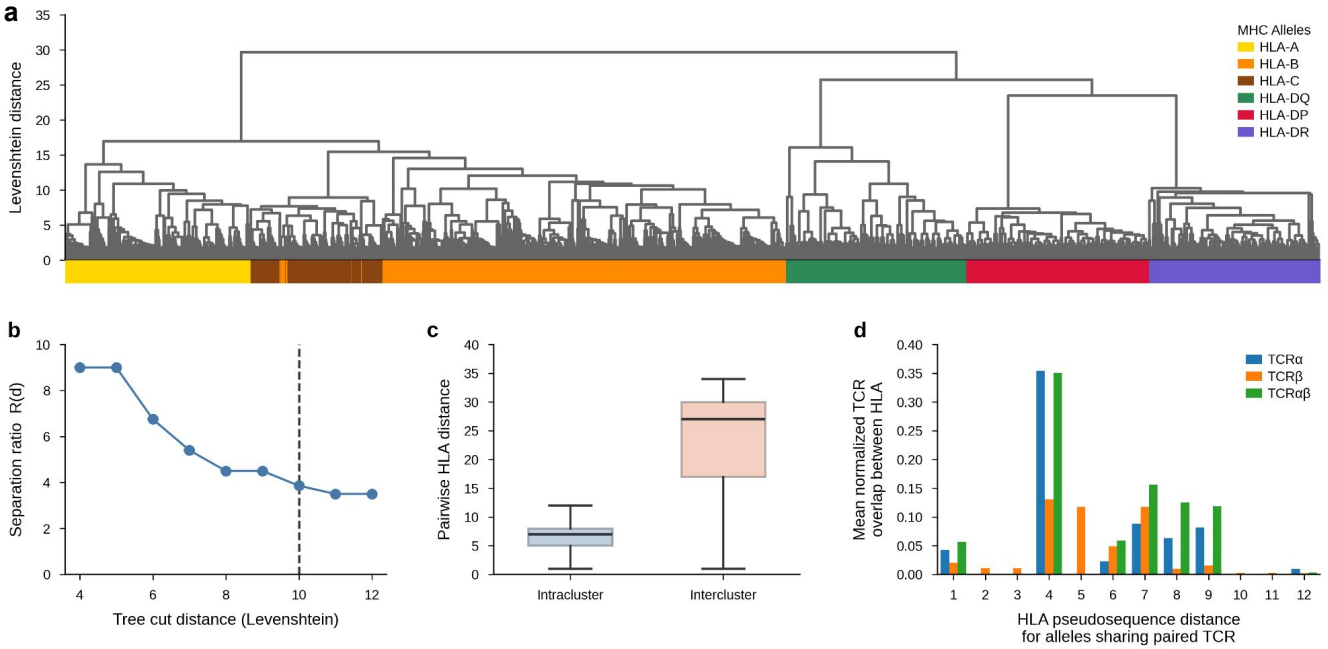


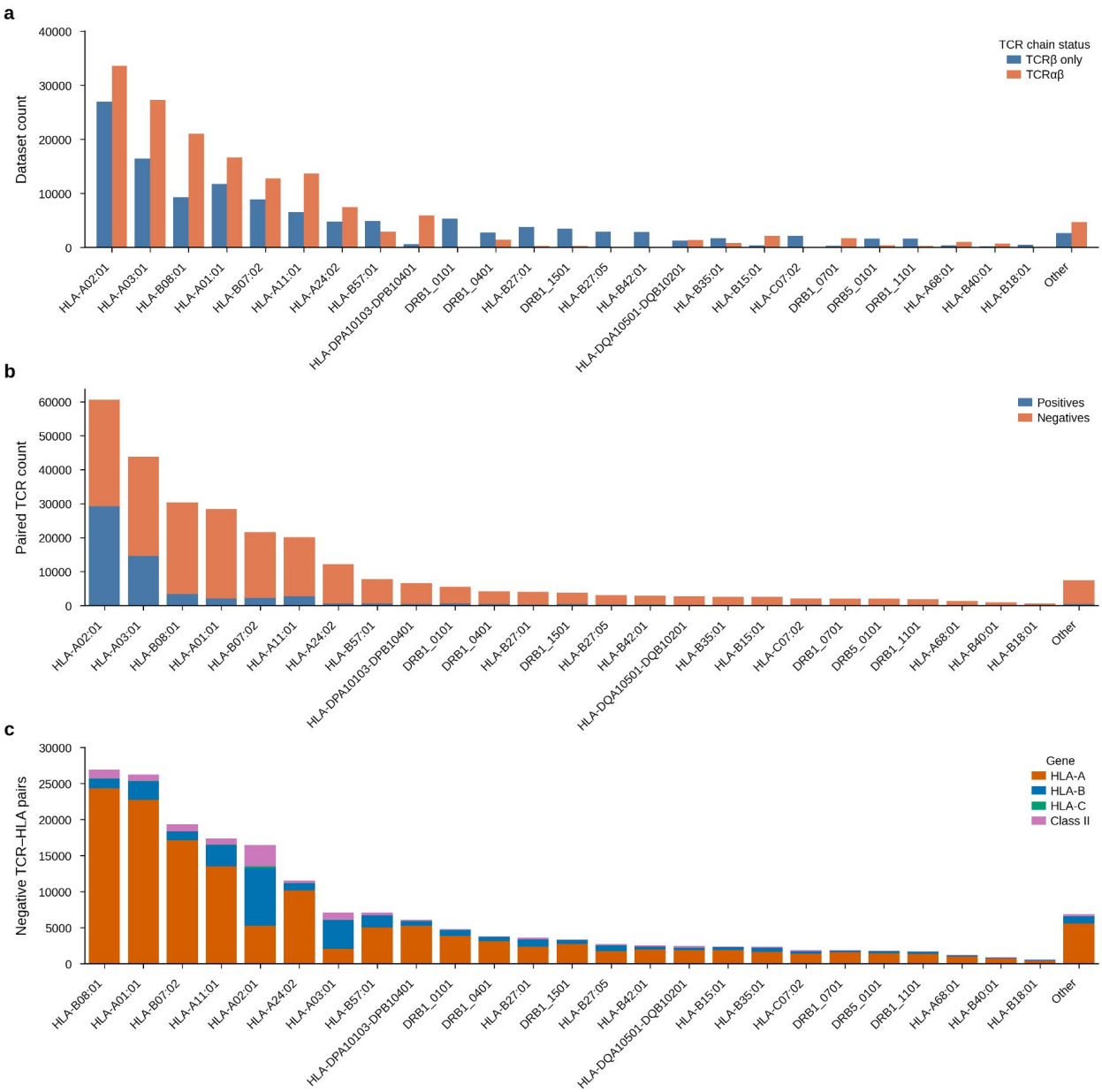
Figure 5 | TCR engagement is associated with allele-specific HLA loss across cancers. a, Covariate-adjusted concordance c by HLA locus, where c is the probability that the more-engaged allele is the under-expressed (lost) member of a heterozygous pair (null c = 0.5; purity and depth-adjusted Firth logistic regression). Class I estimates exclude tumors with a truncating B2M mutation; class II estimates retain all patients (Methods). Diamonds, pan-cancer estimate with 95% CI; points, individual cancer types; red, engaged allele preferentially lost (c > 0.5); blue, preferentially retained (c < 0.5). Class II loci (gray) are exploratory. At right, the number of cancer types in the lost-engaged direction over the number testable, and P (two-sided exact binomial sign test across cancer types). b, Within-tumor paired engagement at the lost (BAF < 0.5) versus retained allele for the three class I loci. Each thin line is one cancer type's median engagement; black, cross-cancer median; gold, LUAD/LUSC. Downward slopes at HLA-A and HLA-B indicate higher engagement at the lost allele; the upward slope at HLA-C indicates higher engagement at the retained allele. P, two-sided exact binomial sign test across cancer types. The two HLA-A exceptions are THCA and READ. c, Per-cancer concordance c for each locus × cancer type, estimated without pooling; class I loci above the rule, class II below. Color encodes direction and magnitude (red, engaged-lost; blue, engaged-retained; intensity, |c − 0.5|; range 0.30 to 0.70); asterisks mark cells significant at q < 0.05 (Benjamini–Hochberg across all finite locus × cancer cells). At right, cancer-type count in the consistent direction over the number testable, and sign-test P.



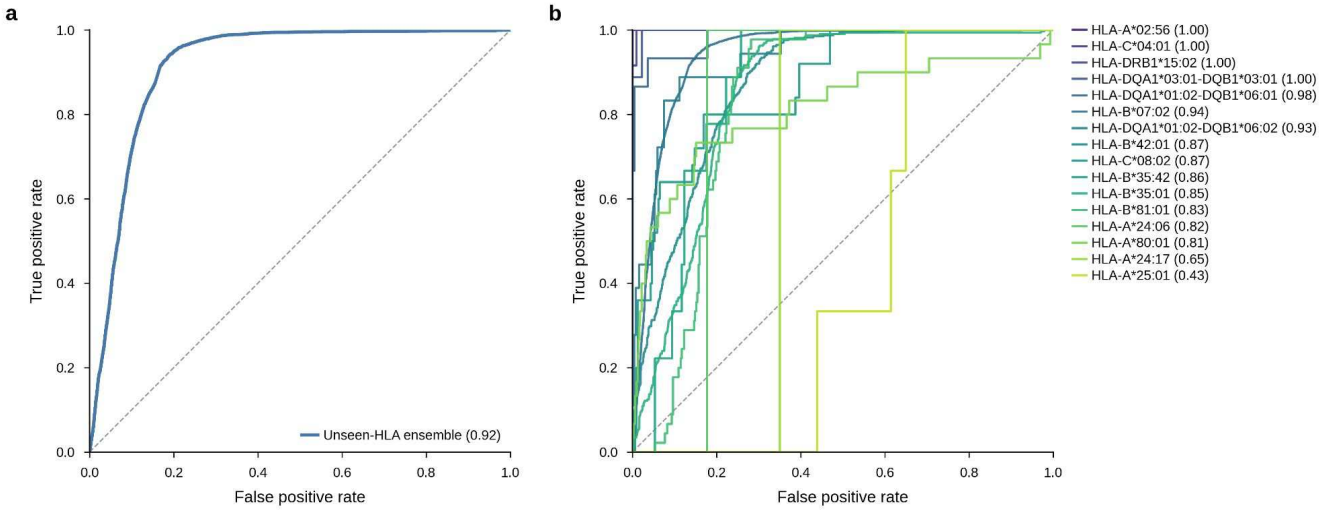
Supplementary Figure 1 | Training data generation pipeline. Schematic of CDR encoding, chain masking, MHC pseudosequence representation, and cross-reactivity–aware negative sampling. A candidate negative was accepted only when the replacement allele's pseudosequence differed from that of the TCR's true restricting allele by a Levenshtein distance of at least 10, excluding closely related alleles likely to share TCRs. A per-allele top-off step additionally admitted candidates at d = 9 where an allele had fewer than two negatives per positive (Methods).



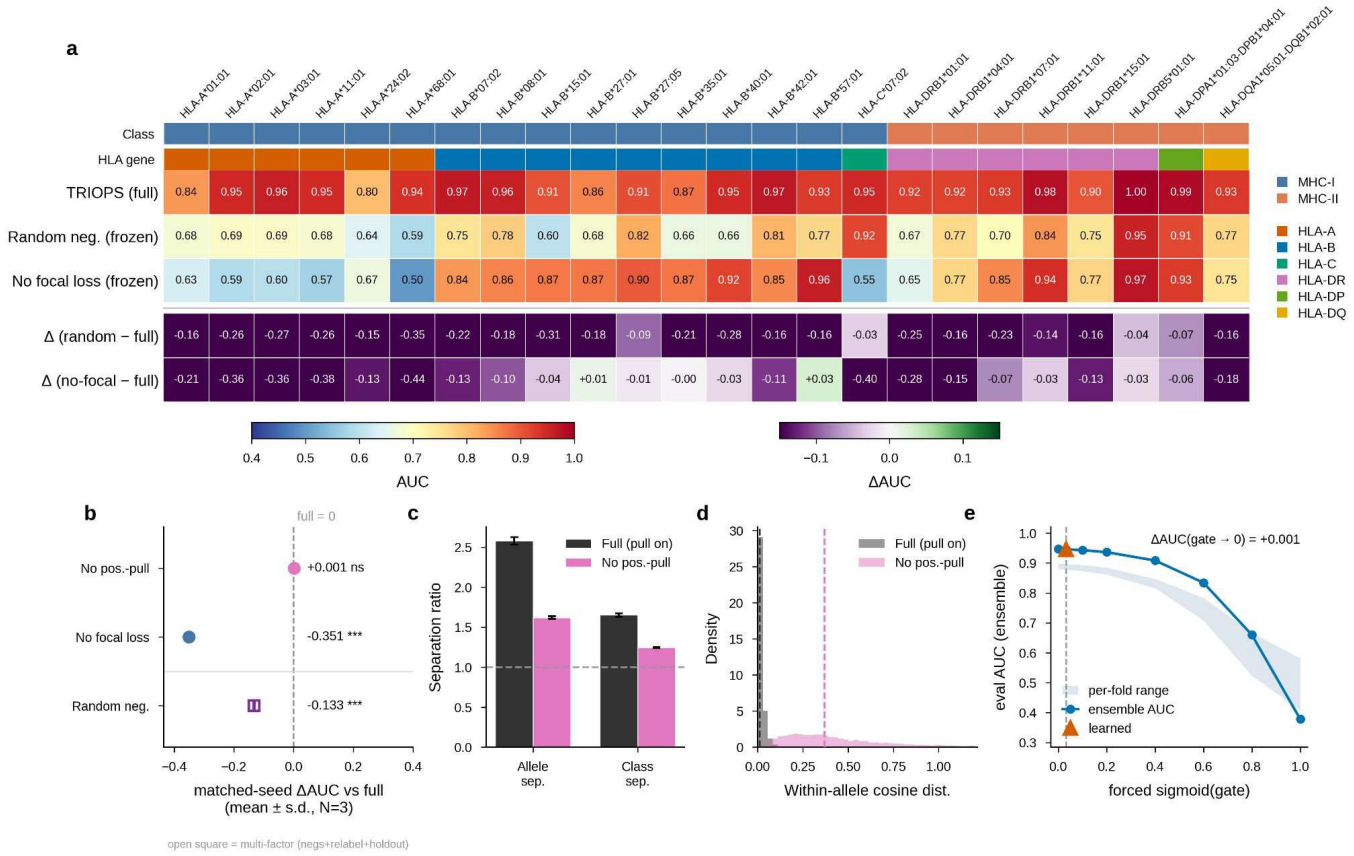
Supplementary Figure 2 | HLA pseudosequence similarity and shared TCR recognition. a, Hierarchical clustering of MHC alleles by Levenshtein distance of NetMHCpan pseudosequences (average linkage). Leaves colored by gene group. b, Cluster separation ratio R(d); plateau at d = 10 (dashed lines). c, Pairwise distances for intracluster and intercluster pairs at cut height d = 10. d, Mean normalized TCR overlap between alleles as a function of pseudosequence distance for TCRα, TCRβ, and paired TCRαβ. Monotonic decrease validates the cross-reactivity–aware sampling threshold.



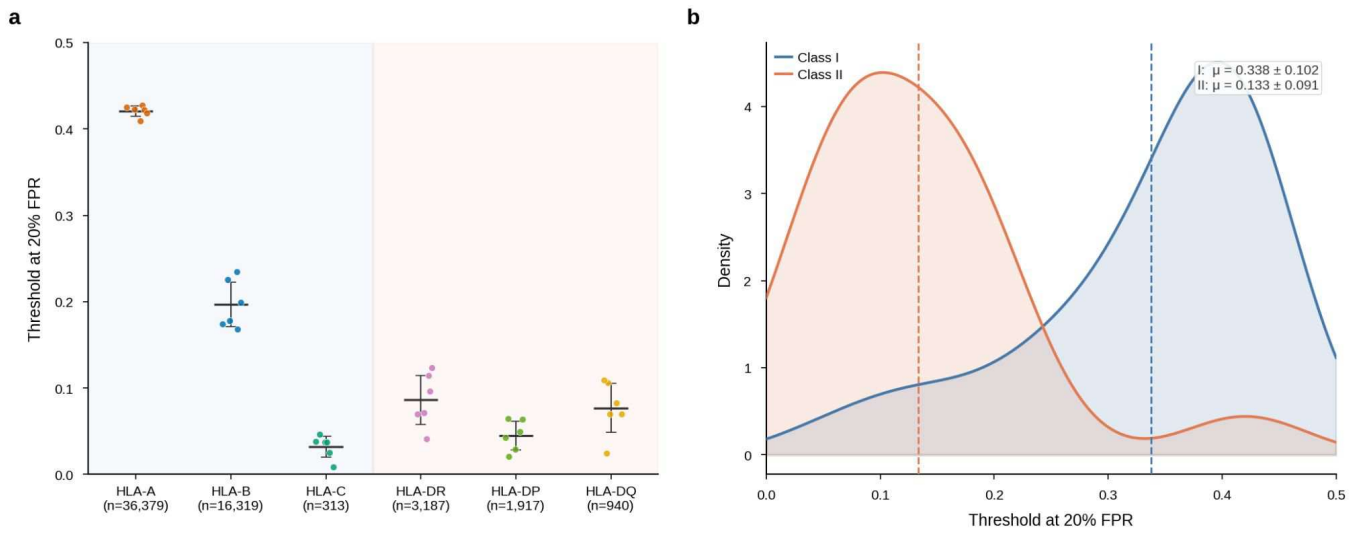
Supplementary Figure 3 | TRIOPS training dataset composition. a, TCR–HLA pair counts by allele, stratified by chain status (TCRαβ vs. TCRβ-only). b, Positive and negative pair counts per allele. c, Negative pair distribution by original HLA gene group, colored by assigned negative allele gene group.



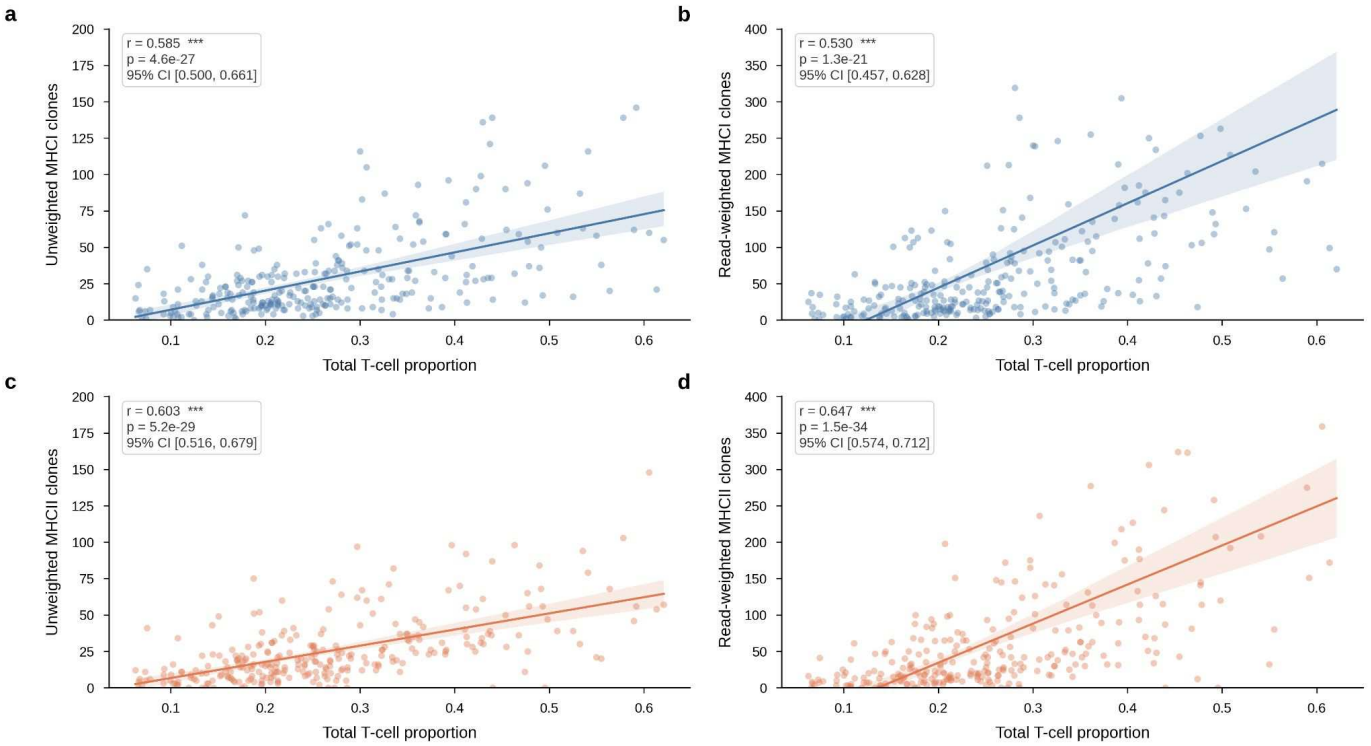
Supplementary Figure 4 | Generalization to HLA alleles unseen during training. 15 pseudosequences (16 HLA alleles) were withheld from training across six cross-validation folds (n = 56,703 TCR–MHC pairs, 9.3% binders). TCRs were not held out, isolating MHC novelty. a, Receiver operating characteristic (ROC) curve for the six-fold ensemble on the holdout (AUC = 0.92; dashed line, chance). b, Per-allele ROC for the 16 held-out alleles, ordered by descending AUC. Per-allele AUC ranged from 0.43 (HLA-A*25:01) to 1.00 (HLA-A*02:56 and HLA-C*04:01); macro-average AUC = 0.87. HLA-B*07:02 accounted for 83% of pairs, so the macro-average provides the allele-balanced summary. Line color denotes allele and follows the AUC ordering.



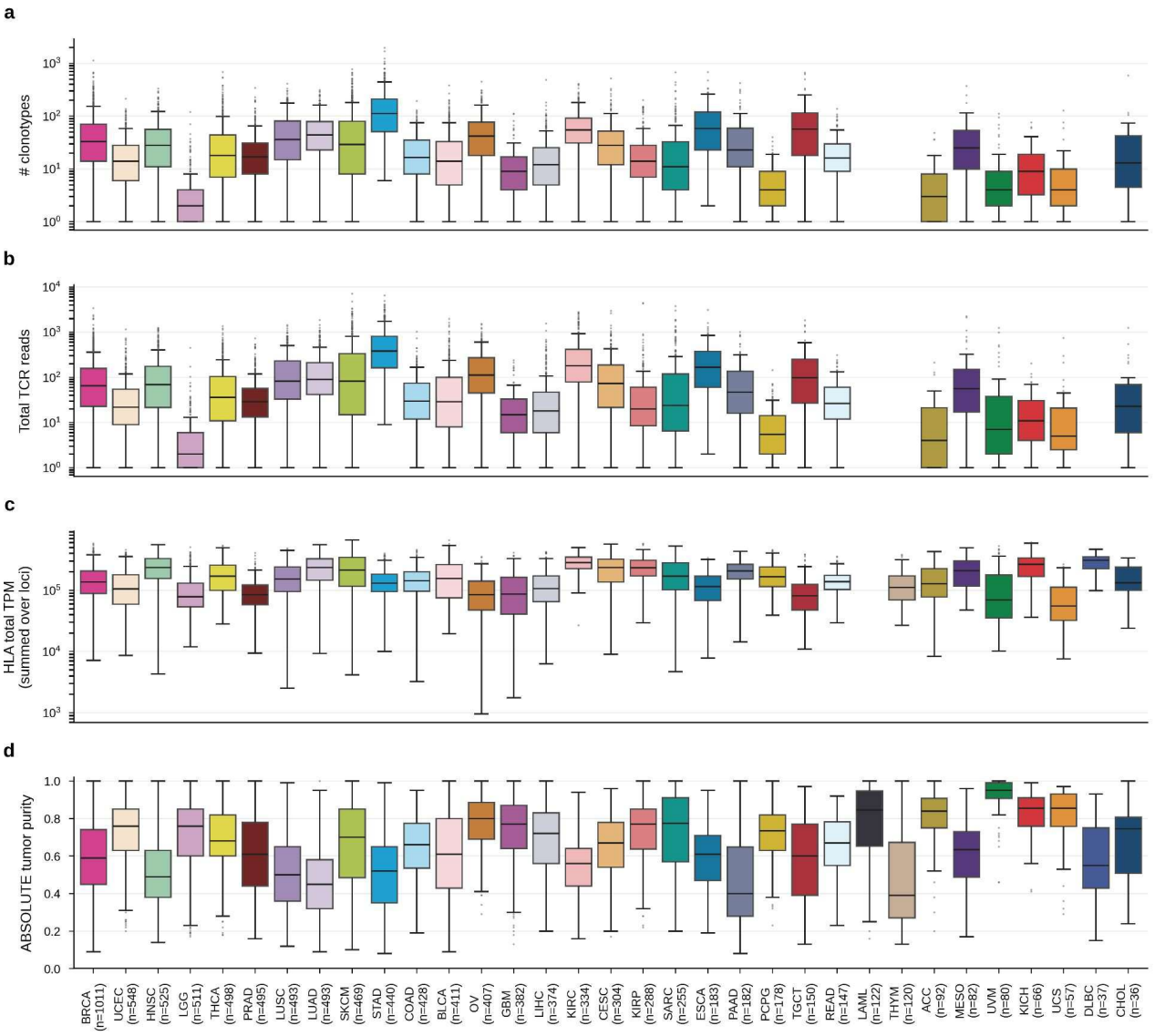
Supplementary Figure 5 | Component ablations. The full TRIOPS model and three ablation variants (random negatives replacing cross-reactivity-aware negative sampling, plain cross-entropy replacing asymmetric focal loss, and positive-pull disabled, λ = 0) were each trained by six-fold cross-validation at three seed replicates and scored on the frozen held-out set (n = 59,055 pairs). Effects are measured against the full model retrained on identical splits and seeds. **a**, Per-allele AUC for the full model and the random-negative and no-focal arms (top three rows, AUC scale) with each ablation's difference from full (bottom two rows, ΔAUC scale), for the 24 alleles with ≥ 15 binding and ≥ 15 non-binding held-out pairs, ordered by MHC class then HLA gene (colour strips). **b**, Matched-seed change in pooled AUC versus full (mean ± s.d., three seeds; full = 0): no focal loss, −0.351; random negatives, −0.133; no positive-pull, +0.001 (matched-seed paired t-test across three seed replicates; ***P < 0.001; ns, not significant). The random arm (open square) is a multi-factor perturbation, not a single-knob ablation. **c**, Allele- and class-level embedding separation (between- over within-group cosine distance; ratio = 1, no separation) for full versus no-positive-pull, from fold-4 embeddings of held-out binders; error bars, bootstrap 95% CIs. **d**, Within-allele cosine-distance distributions for the same embeddings (dashed lines, medians): tight near zero with positive-pull, dispersed without it. **e**, Cross-attention gate sweep. The residual gate returns (1 − g)·t_q + g·attn, so sigmoid(g) = 0 bypasses cross-attention and sigmoid(g) = 1 forwards it alone. The gate was initialized at sigmoid(g) = 0.55 and every fold converged below 0.08 (mean 0.043 ± 0.026 s.d.; range 0.014 to 0.077), indicating consistent suppression rather than an initialization artifact. Clamping the gate shut changed AUC by +0.001. Forcing it open reduced AUC to approximately 0.38.



Supplementary Figure 6 | FPR calibration stability across folds. a, Per-locus 20% FPR thresholds across six cross-validation folds for all HLA gene groups (per-locus sample sizes on the x-axis). Points, individual folds; bars, mean ± s.d. Low inter-fold variance indicates robust calibration. b, Threshold distributions by MHC class. Class I, μ = 0.338 ± 0.102; class II, μ = 0.133 ± 0.091 (dashed lines, class means).



Supplementary Figure 7 | TRIOPS-predicted MHC-targeting clone counts correlate with CIBERSORT-estimated T cell proportions. a, Unweighted class I clone counts versus total T cell proportion ($r = 0.585$, $P = 4.6 \times 10^{-27}$). b, Read-weighted class I clone counts versus total T cell proportion ($r = 0.530$, $P = 1.3 \times 10^{-21}$). c, Unweighted class II clone counts versus total T cell proportion ($r = 0.603$, $P = 5.2 \times 10^{-29}$). d, Read-weighted class II clone counts versus total T cell proportion ($r = 0.647$, $P = 1.5 \times 10^{-34}$). Read-count weighting modestly strengthens the class II correlation (c versus d), and modestly attenuates the class I correlation (a versus b). Shaded regions, 95% confidence intervals; n = 279 TCGA-LUAD patients with available CIBERSORT estimates.



Supplementary Figure 8 | Pre-filter distributions of immune and tumor covariates across TCGA cancer types. Per-cancer distributions of the quantities entering the engagement–loss analysis, shown before quality-control filtering (no minimum TCRβ read, clonotype, locus-read, or per-cancer patient thresholds). Each box is one cancer type (sample sizes in parentheses on the x-axis), showing the median and interquartile range with whiskers at 1.5× IQR and outliers as points; cancers are ordered by sample count. a, Number of recovered TCRβ clonotypes per sample (log scale). b, Total TCRβ read count per sample (log scale). c, Total HLA expression per sample, summed across loci (TPM, log scale). d, ABSOLUTE tumor purity per sample (linear scale, tumor samples only).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TRIOPSsupplementarysoftware.zip