

Supplementary Tables

Supplementary Table 1. Field descriptions and metadata categories in the CROSSTCR database. This table summarizes the standardized metadata fields used in CROSSTCR, including TCR information, peptide and MHC information, experimental evidence, source information, and cross-reactive group annotations.

	Field name	Description
TCR-related metadata	CDR3_alpha	Complementarity Determining Region 3 sequence of the alpha chain of the T-cell receptor (TCR).
	CDR3_beta	Complementarity Determining Region 3 sequence of the beta chain of the T-cell receptor (TCR).
	V_alpha	Variable (V) gene segment identifier of the TCR alpha chain.
	J_alpha	Joining (J) gene segment identifier of the TCR alpha chain.
	V_beta	Variable (V) gene segment identifier of the TCR beta chain.
	J_beta	Joining (J) gene segment identifier of the TCR beta chain.
	CDR3_alpha_Length	Length (number of amino acids) of the CDR3 region in the alpha chain of the TCR.

	CDR3_beta_Length	Length (number of amino acids) of the CDR3 region in the beta chain of the TCR.
	TCR_Species	Species from which the T-cell receptor (TCR) was derived.
Peptide-related metadata	Peptide	Amino acid sequence of the antigenic peptide recognized by the TCR.
	Peptide_Length	Length (number of amino acids) of the antigenic peptide.
	Peptide_gene	Gene encoding the antigenic peptide.
	Peptide_Species	Species from which the antigenic peptide originates.
	Peptide_Origin	Biological or pathological source (e.g., virus, cancer, self) of the antigenic peptide.
	Peptide_Similarity	Similarity measure of this peptide to other peptides, typically calculated via sequence alignment methods.
	Flank_Sequence	Amino acid sequences immediately adjacent (flanking regions) to the antigenic peptide within the source protein.
	Match_Count	Number of matches identified.
	Source_Protein	Name or description of the protein from which the antigenic peptide is derived.

	Protein_ID	Unique identifier (e.g., UniProt accession) of the source protein.
	Start / End	The start and end positions of the antigenic peptide within the source protein sequence.
	Protein_Length	Total length of the source protein in amino acids.
MHC-related metadata	MHC_Allele	Major Histocompatibility Complex (MHC) allele presenting the peptide.
	MHC_Class	Class (e.g., class I or class II) of the MHC molecule presenting the peptide.
Experimental metadata	Binding_Outcome	Outcome indicating whether the peptide was experimentally determined to bind the specified MHC allele.
	Evidence_Level	Quality or reliability of the evidence supporting the reported peptide-MHC-TCR interaction.
	Functional_Response	Description or measure of T-cell functional response (e.g., cytokine secretion, proliferation) upon interaction with peptide-MHC complex.
	Binding_Score	Numerical score quantifying the strength or affinity of the peptide-MHC interaction.

	Assay_Type	Type of experimental assay or method used to characterize or validate the TCR-peptide-MHC interaction.
	Is_Same_Experiment	Indicator specifying if the data points originate from the same experimental setup or condition.
Meta-level annotations data	Release-date	Date on which the dataset or specific data entries were released or updated.
	Clinical_Context	Clinical or experimental setting of TCR-peptide cross-reactivity, categorized into cancer immunotherapy, autoimmunity, viral infection, vaccine response, healthy individuals, synthetic/engineered TCRs, or unknown context.
	Donor	Identifier of the biological donor providing the T-cells or peptides (e.g., patient ID, demographic information).
	Tissue	Type of tissue from which the T-cells or peptides were derived.
	Reference	Bibliographic reference (e.g., PubMed ID) documenting the original experimental data.

	PDB	Protein Data Bank (PDB) identifier for available 3D structural information related to the peptide-MHC-TCR complex.
	CrossReactive_Group_ID	Identifier indicating membership in a defined cross-reactivity group, representing sets of TCRs or peptides with known cross-reactive relationships.

Supplementary Table 2. Peptide sequence identity in representative cross-reactive groups. This table reports pairwise peptide sequence identity values for representative cross-reactive TCR groups used to evaluate sequence divergence among experimentally linked peptides.

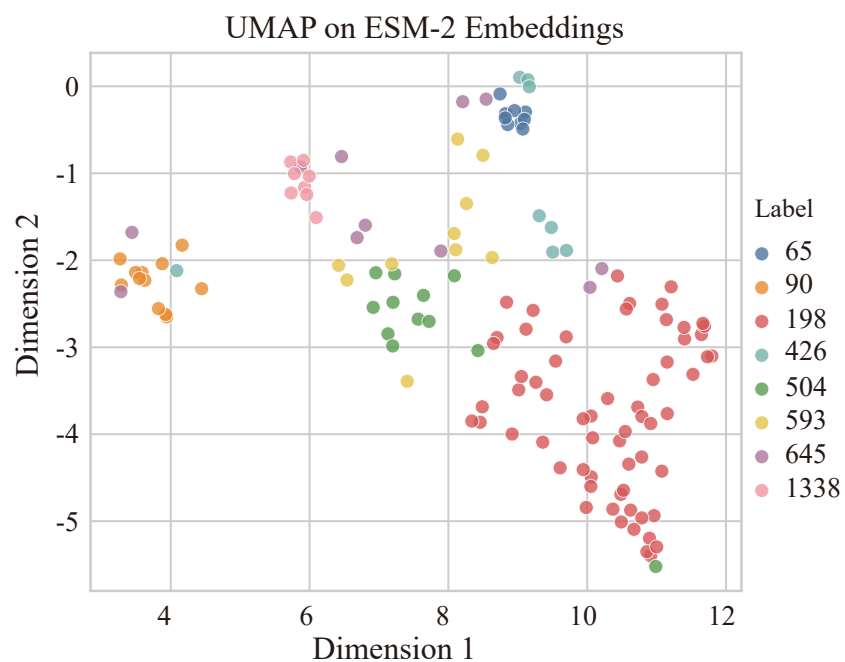
Group	n_seqs	n_pairs	mean_identity
65	9	36	0.771164
90	12	66	0.555463
198	62	1891	0.474207
426	8	28	0.401310
504	12	66	0.282560
593	10	45	0.336895
645	11	55	0.172758
1338	8	28	0.651723
Average	-	-	0.4688

Supplementary Table 3. Structural information for the PDB entries used in Figure 4B.

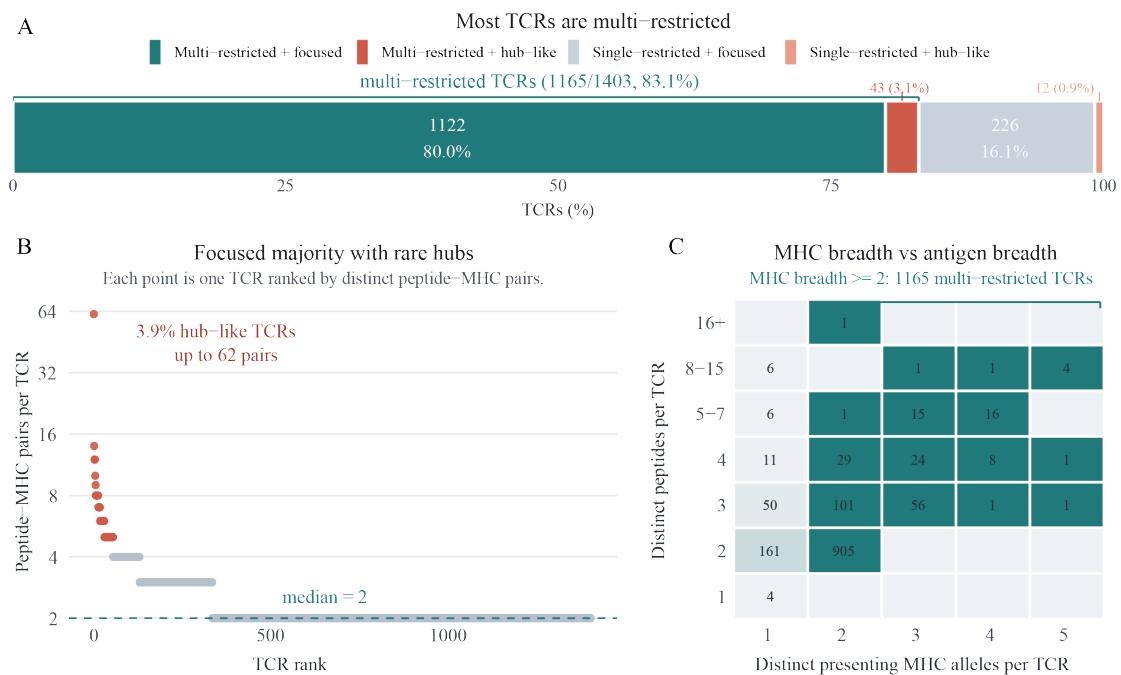
This table lists the experimentally resolved structural templates used for the representative peptide-backbone overlays shown in Figure 4B, including PDB identifiers, peptide sequences, MHC restrictions, RMSD values, and source references.

Group ID	Peptide sequence	MHC allele/class	PDB ID	Pairwise RMSD (Å)	Ref
426	YQFGPDFPIA	HLA-A*02:01:48	5C07	0.472 Å	PMID:27183389
426	ALWGPDPAAA	HLA-A*02:01:48	3UTT		PMID:22245737
90	LGYGFVNYI	HLA-A*02:01:48	3PWP	0.456 Å	PMID:21282516
90	MLWGYLQYV	HLA-A*02:01:48	3H9S		PMID:20064447
765	NPVVHFFKNIVTPR	HLA-DQA1*01:11	3PL6	0.350 Å	PMID:21199956
765	RLLMLFAKDVVSRN	HLA-DQA1*01:11	4GRL		PMID:24136005
838	MDFARVHFISALHGSG	HLA-DRA*01:02:03	2WBJ	0.701 Å	PMID:19303388
838	ENPVVHFFKNIVTP	HLA-DRA*01:02:03	1YMM		PMID:15821740

Supplementary Figures

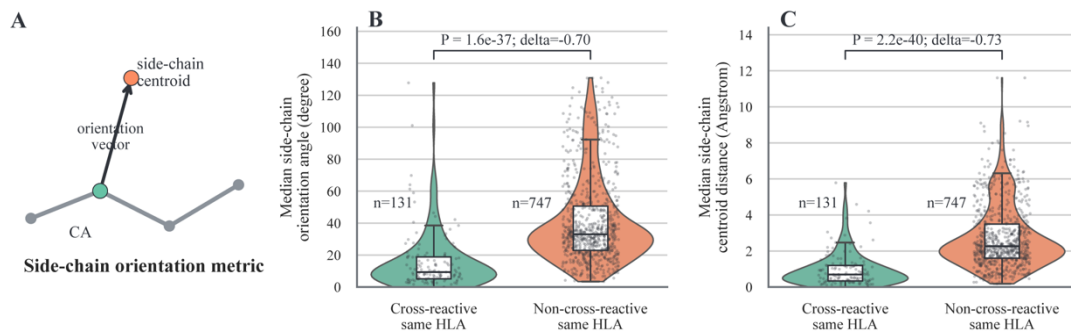


Supplementary Figure 1. UMAP visualization of peptides from the eight cross-reactive TCR groups with the largest number of TCR–pMHC pairs, based on ESM-2 embeddings.



Supplementary Figure 2. TCR-centric analysis of multi-restricted yet focused cross-reactivity in CROSSTCR. (A) Distribution of cross-reactive TCR groups stratified by MHC restriction breadth and peptide-MHC connectivity. Multi-restricted TCR groups were defined as those recognizing peptide-MHC ligands presented by more than one MHC allele or MHC class. In CROSSTCR v1.3, 83.1% of cross-reactive TCR groups were multi-restricted (1165/1403), including 80.0% multi-restricted and focused groups (≤ 4 distinct peptide-MHC pairs) and 3.1% multi-restricted hub-like groups (> 4 distinct peptide-MHC pairs). (B) Ranked distribution of peptide-MHC connectivity per TCR group. Each point represents one TCR group ordered by the number of distinct peptide-MHC pairs recognized. Most TCR groups showed limited connectivity, with a median of two peptide-MHC pairs, whereas a small hub-like subset comprising 3.9% of TCR groups showed broader connectivity, reaching up to 62 peptide-MHC pairs. (C) Heatmap showing the number of TCR groups stratified by MHC breadth and antigen-recognition breadth. The bracket highlights TCR groups with MHC breadth ≥ 2 ,

corresponding to 1165 multi-restricted groups. Together, these data indicate that multi-restriction is common among curated cross-reactive TCR groups in CROSSTCR, whereas broad antigen connectivity is restricted to a rare hub-like subset.



Supplementary Figure 3. Side-chain-level structural comparison of same-HLA peptide pairs. (A) Schematic definition of peptide side-chain orientation. After peptide backbone superposition, side-chain orientation was represented for each comparable non-glycine residue by a vector from the C α atom to the side-chain heavy-atom centroid. (B) Distribution of pair-level median side-chain orientation angles for same-HLA cross-reactive peptide pairs and same-HLA non-cross-reactive controls. Cross-reactive pairs showed smaller side-chain orientation angles than non-cross-reactive controls (median 9.37° vs. 33.02°; two-sided Wilcoxon rank-sum test, $P = 1.6 \times 10^{-37}$; Cliff's $\delta = -0.70$). (C) Distribution of pair-level median side-chain centroid distances for the same pair groups. Cross-reactive pairs showed shorter side-chain centroid distances than non-cross-reactive controls (median 0.697 Å vs. 2.262 Å; two-sided Wilcoxon rank-sum test, $P = 2.2 \times 10^{-40}$; Cliff's $\delta = -0.73$). Each point represents one structure-resolved peptide pair; boxes show the median and interquartile range. Glycine residues and residues lacking side-chain heavy atoms were excluded at the residue level.



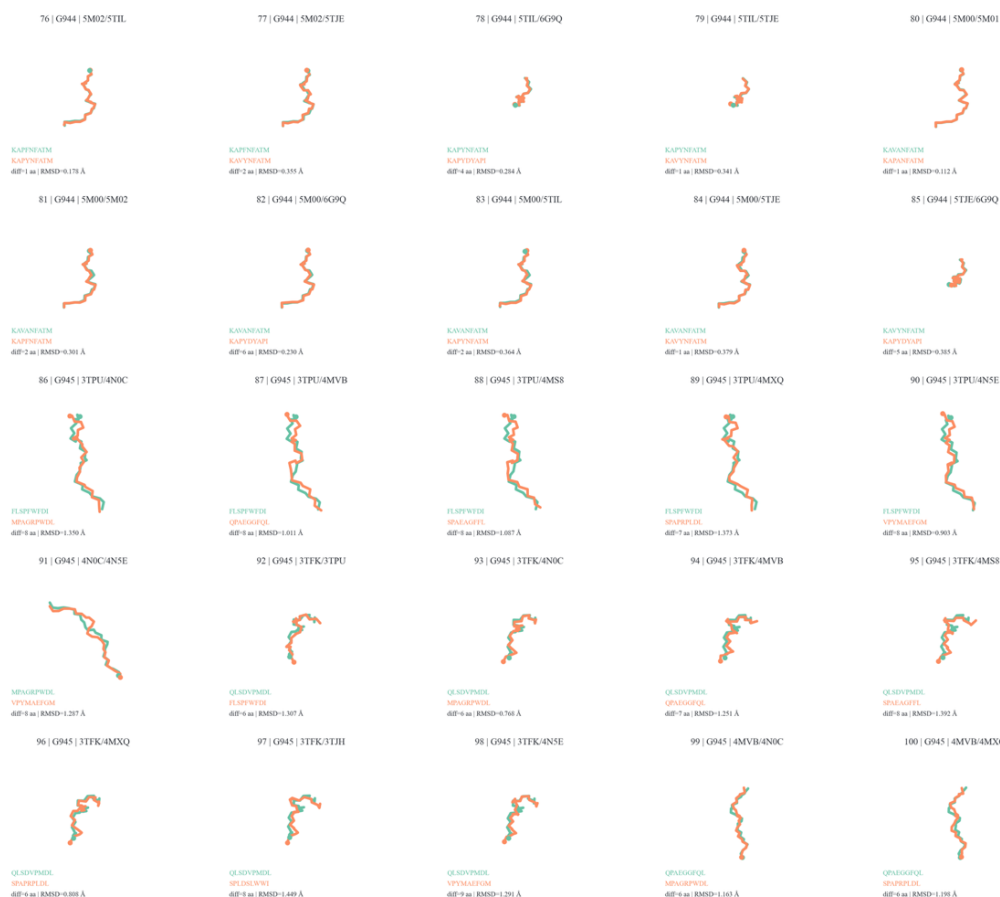
Supplementary Figure 4. Peptide-backbone overlays for structure-resolved cross-reactive peptide pairs 1–25. Each numbered item shows one cross-reactive peptide pair after peptide-backbone superposition. Green and orange traces represent peptide 1 and peptide 2, respectively, and the peptide sequences are displayed in the corresponding colors. Labels above each item indicate the pair index, CROSSTCR cross-reactive group identifier (G), and Protein Data Bank identifiers of the two structures. “diff” denotes the number of amino-acid differences among aligned residues. RMSD, root-mean-square deviation; PDB, Protein Data Bank; aa, amino acids; Å, angstrom.



Supplementary Figure 5. Peptide-backbone overlays for structure-resolved cross-reactive peptide pairs 26–50. Each numbered item shows one cross-reactive peptide pair after peptide-backbone superposition. Green and orange traces represent peptide 1 and peptide 2, respectively, and the peptide sequences are displayed in the corresponding colors. Labels above each item indicate the pair index, CROSSTCR cross-reactive group identifier (G), and Protein Data Bank identifiers of the two structures. “diff” denotes the number of amino-acid differences among aligned residues. RMSD, root-mean-square deviation; PDB, Protein Data Bank; aa, amino acids; Å, angstrom.



Supplementary Figure 6. Peptide-backbone overlays for structure-resolved cross-reactive peptide pairs 51–75. Each numbered item shows one cross-reactive peptide pair after peptide-backbone superposition. Green and orange traces represent peptide 1 and peptide 2, respectively, and the peptide sequences are displayed in the corresponding colors. Labels above each item indicate the pair index, CROSSTCR cross-reactive group identifier (G), and Protein Data Bank identifiers of the two structures. “diff” denotes the number of amino-acid differences among aligned residues. RMSD, root-mean-square deviation; PDB, Protein Data Bank; aa, amino acids; Å, angstrom.



Supplementary Figure 7. Peptide-backbone overlays for structure-resolved cross-reactive peptide pairs 76–100. Each numbered item shows one cross-reactive peptide pair after peptide-backbone superposition. Green and orange traces represent peptide 1 and peptide 2, respectively, and the peptide sequences are displayed in the corresponding colors. Labels above each item indicate the pair index, CROSSTCR cross-reactive group identifier (G), and Protein Data Bank identifiers of the two structures. “diff” denotes the number of amino-acid differences among aligned residues. RMSD, root-mean-square deviation; PDB, Protein Data Bank; aa, amino acids; Å, angstrom.



Supplementary Figure 8. Peptide-backbone overlays for structure-resolved cross-reactive peptide pairs 101–131. Each numbered item shows one cross-reactive peptide pair after peptide-backbone superposition. Green and orange traces represent peptide 1 and peptide 2, respectively, and the peptide sequences are displayed in the corresponding colors. Labels above each item indicate the pair index, CROSSTCR cross-reactive group identifier (G), and Protein Data Bank identifiers of the two

structures. “diff” denotes the number of amino-acid differences among aligned residues.

RMSD, root-mean-square deviation; PDB, Protein Data Bank; aa, amino acids; Å,

angstrom.

Supplementary Data

Supplementary Data 1. Pair-level metadata for expanded same-HLA structural comparison. This large table is provided separately as an Excel file. It contains pair-level metadata for the 131 same-HLA cross-reactive peptide pairs and 747 same-HLA non-cross-reactive control pairs used in the expanded peptide-backbone RMSD comparison, including group identifiers, peptide sequences, MHC annotations, PDB identifiers, chain assignments, matching information, comparable residues, backbone atoms used for RMSD calculation, and peptide-backbone RMSD values.

Supplementary Data 2. Pair-level side-chain structural similarity metrics. This large table is provided separately as an Excel file. It contains pair-level side-chain orientation and side-chain centroid-distance metrics for the same-HLA cross-reactive peptide pairs and same-HLA non-cross-reactive controls used in the side-chain-level structural comparison.