HLA Alleles Imprint Distinct Biases in the Usage Preferences of TCR Vβ segments

Leonardo V. Castorina[1], Matthew T. Noakes[2], Lorenzo Pisani[1], Julia Greissl[1], Harlan Robins[2], Haiyin Chen-Harris[2], and H. Jabran Zahid[1,*]

[1]Microsoft Research, 14820 NE 36th Street, Redmond, Washington, 98052, USA
[2]Adaptive Biotechnologies, 1165 Eastlake Ave E Seattle, Washington, 98109, USA
[*]Corresponding author: H.Zahid@microsoft.com

Abstract

T cells must co-recognize peptide and HLA, yet the extent to which this specificity is shaped by germline-encoded TCR–HLA contacts versus selection during thymic development has remained difficult to quantify. Leveraging population-scale TCRβ repertoires linked to donor HLA genotypes, we construct allele-specific Vβ usage profiles and normalize them to repertoire-wide baselines to derive interpretable HLA-TCR Vβ preference vectors. We demonstrate that different HLA alleles imprint distinct biases in the usage frequency of TCR Vβ gene segments among the public TCRs that engage those alleles; consistent with germline-encoded TCR–HLA contact preferences, certain Vβ genes are over-represented for particular HLA alleles. Similarities in HLA amino-acid sequence predict similarities in both their Vβ preferences and peptide-binding motifs; a residue-level analysis disentangles HLA positions primarily associated with TCR engagement from those associated with peptide motifs. The TCR-associated HLA positions localize to canonical TCR-facing helices, whereas peptide-associated HLA positions track binding pockets, revealing distinct molecular routes by which HLA polymorphism shapes the TCR and peptide sides of recognition. CMV exposure stratifications confirmed these patterns are not explained by a few dominant infections. Together, these data support that germline-encoded constraints set the landscape of TCR–HLA compatibility, while thymic and peptide-driven forces tune the realized repertoire. These HLA-specific Vβ biases are a biological prior that should provide a baseline for better understanding of TCR-pHLA specificity as a whole and should be accounted for in future evaluation of any TCR-pHLA specificity prediction methods.

Keywords: Human Leukocyte Antigen (HLA), HLA Polymorphism, T Cell Receptor (TCR), Germline Bias, TCR Vβ Gene Usage, TCR Repertoire, Structural Immunology.

1 Introduction

Interactions between T cell receptors (TCRs) and peptide-presenting Human Leukocyte Antigens (HLAs) are central to adaptive immunity, enabling the immune system to distinguish self from non-self

and mount targeted responses to pathogens, cancer and other immunological threats (Murphy and Weaver, 2017a). A deeper understanding of this interaction could transform our ability to predict immune responses, design vaccines, personalize immunotherapies and uncover the pathogenesis of autoimmune disorders. Achieving this will likely require a precise understanding of how TCRs engage with peptide–HLA (pHLA) complexes at the molecular level. Despite substantial progress (Davis and Bjorkman, 1988; Rudolph et al., 2006; Rossjohn et al., 2015; Murphy and Weaver, 2017b), key aspects of this molecular interaction remain incompletely understood.

Understanding TCR–pHLA interactions is complicated by their extraordinary complexity. HLAs are among the most polymorphic human genes, with tens of thousands of alleles identified across the six classical class I (HLA-A, -B, -C) and II loci (HLA-DR, -DQ and -DP) (Klein and Sato, 2000). Structurally, the HLA peptide-binding groove is defined by two helices: in Class I the helices are named $\alpha1$ and $\alpha2$ and are part of the same polypeptide chain, while for Class 2 the groove is formed by two distinct chains, $\alpha1$ and $\beta1$. In the canonical binding mode, the TCRβ interacts with around the $\alpha1$ helix in Class I and the $\beta1$ helix in Class II (See Figure5 A-D) Christopher Garcia et al. (2009); Cohn et al. (2019)).

Polymorphisms are concentrated in the peptide-binding groove, influencing both peptide presentation and TCR binding (Tiercy, 2002; Nakamura et al., 2019; Karnaukhov et al., 2022). Moreover, TCR binding is degenerate: a single TCR can recognize many pHLAs and many TCRs can recognize the same pHLA (Davis and Bjorkman, 1988; Sewell, 2012; Murphy and Weaver, 2017b). While many predictive models use sequence-based, structural or molecular dynamics approaches (Springer et al., 2021; Grazioli et al., 2023; Bradley, 2023; Jensen and Nielsen, 2024), they often fail to generalize (Drost et al., 2025), limited by interaction complexity and sparsity of high-quality data (Grazioli et al., 2022; Castorina et al., 2025). In the absence of massive training sets, uncovering mechanistic principles can provide inductive biases that improve model generalizability. This strategy mirrors the success of AlphaFold, which leveraged Multiple Sequence Alignments to identify co-evolutionary constraints and guide structural prediction (Jumper et al., 2021). Similarly, statistical frameworks like ours may uncover residue-level features that constrain TCR–pHLA binding and inform the next generation of predictive models. This perspective is especially relevant given ongoing efforts to determine whether TCR specificity arises from germline-encoded features such as CDR1 and CDR2 loops that may have evolved HLA specificity or is shaped entirely by thymic selection acting on random repertoires (La Gruta et al., 2018). Clarifying these mechanisms may guide the development of more general and interpretable models.

Recent advances in sequencing technologies have enabled statistical approaches to examine how HLA polymorphism shapes TCR gene usage. Sharon et al. (2016) proposed that if germline-encoded contacts influence TCR–HLA interactions, then TCR Vβs should show biased usage depending on HLA type. Supporting this, their trans-eQTL analysis found that polymorphisms near the HLA binding groove correlate with Vβ expression, linking HLA polymorphism to TCR repertoire composition. However, their use of bulk RNA sequencing averaged over the entire T cell compartment, potentially dilutes the HLA-specific signals. To improve resolution, we leverage a previously developed

method that uses large-scale repertoire sequencing to associate public TCRβs (i.e., those shared across individuals) with specific HLAs. This approach, introduced and validated by Zahid et al. (2025a), leverages TCR co-occurrence across thousands of subjects with known HLA genotypes to identify TCRβs statistically associated with specific HLAs. These associations generalize to diverse, unseen populations, yielding TCRβ sets with strong evidence of true immunological specificity.

Using HLA-associated TCRβ sets, we quantify Vβ gene usage for each HLA and examine how these patterns correlate with HLA polymorphism, while statistically accounting for potential mediation by shared peptide motifs. We identify HLA residues associated with Vβ usage, peptide repertoire or both. To compare patterns across HLA types, we harmonize sequence and structural data through sequence alignments and 3D structure mapping. Although our analysis is limited to the TCRβ chain, it reveals population-level trends consistent with germline influences and highlights candidate HLA residues that may shape TCR–HLA specificity through direct structural mechanisms. The scale of this analysis enables resolution of TCR–HLA associations not previously accessible, offering biologically grounded priors to inform predictive models.

2 Materials and Methods

2.1 HLAdb

Using direct HLA genotyping and 4,144 TCRβ repertoires, Zahid et al. (2025a) built sensitive and specific HLA imputation models from TCRβs associated with hundreds of common HLAs. We impute 145 of the most common HLAs for 30,000 subjects in the T-Detect COVID cohort (Zahid et al., 2025b). These imputed HLAs were used as case-control labels and TCRβs were statistically associated to HLAs using via a one-sided Fisher's Exact Test following Zahid et al. (2025a). By imputing HLAs on a set of repertoires that are nearly an order of magnitude larger in sample size than the original data set used to build the models, we are able to leverage the genotyped labels available for a smaller set of repertoires and the statistical power of sequence discovery offered by the larger cohort.

We restricted analysis to HLAs with more than 5,000 associated public TCRβs. In total, 2,827,857 public TCR Vβ sequences from 41 unique public TCR Vβ genes associated with 91 HLA alleles were included in this analysis (see Supplementary Table S1). Using the set of TCRβs associated with each HLA, we generate the TCR Vβ usage pattern as the distribution of Vβ gene preferences for that particular HLA. Vβ genes are not uniformly represented in the TCRβ repertoire, thus we normalize Vβ gene usage to the background frequency measured from all TCRβs (public and private) summed across repertoires and this is represented as a vector (see Supplementary Figure S2 and Supplementary Table S1).

2.2 TCR Vβ Normalization

We start with a matrix whose rows are HLAs and columns are TCR Vβ s, containing co-occurrence counts. We also have background counts for each TCR Vβ . First, we remove ambiguous or undefined

Vβ genes. We then normalize TCRβ V-gene counts by dividing by the corresponding background counts to correct for any underlying baseline TCR Vβ frequency, for each TCR Vβ gene, the total number of times it was observed across 30,000 repertoires was used for background normalization. Any ambiguous or undefined Vβ genes were removed.

After background adjustment, we convert counts into frequencies by dividing each each N_{ij} by the total for HLA i:

- For each HLA i, compute $\sum_j N_{ij}$.
- Define $F_{ij} = \dfrac{N_{ij}}{\sum_j N_{ij}}$
  This ensures $\sum_j F_{ij} = 1$ for each HLA i.

We call the frequency vector F_i for HLA i the "Vβ gene preference".

2.3 Pairwise Distance Analysis of HLA

The TCR-pHLA interaction involves the TCR, the HLA and a peptide. We characterize HLA differences with three pairwise distance matrices:

- **HLA polymorphism**: BLOSUM62 distance between HLA sequences (See 2.3.1).
- **Vβ gene preference change**: cosine distance between Vβ preference vectors (See 2.3.2).
- **Peptide motif change**: Jensen–Shannon (JS) divergence between Position Weight Matrix (PWM) (See 2.3.3).

We then correlate these matrices to identify HLA positions whose variation is associated with TCRβ usage or peptide motifs, disentangling TCR–HLA from pHLA effects.

We analyse both **within** HLA loci and **between** loci. Within-locus analyses aim to identify binding-groove positions most significant for the peptide or the TCRβ.

Between-locus analyses aim to identify sequence positions that distinguish loci and correlate with the TCRβ or the peptide motifs.

2.3.1 HLA Polymorphism

We use the HLA sequences dataset from the TCR Dock pipeline Bradley (2023), restricted to the TCR-binding groove to ensure comparable sequence lengths and also focuses the analysis around areas of major polymorphism.

HLA sequences are aligned using Clustal ω (Sievers et al., 2011). Pairwise differences are computed using the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) whose log-odds quantify amino acids substitution likelihoods. We calculate both per-position and overall differences by summing scores across positions..

For position k in aligned sequences S_i and S_j, we calculate a per-position distance, $D_{seq}[k]$, using the exponential of the BLOSUM score:

$$D_{seq}[k] = \frac{1}{\exp(\text{BLOSUM}(S_i[k], S_j[k]))}$$

The overall sequence divergence is calculated by summing these per-position distances:

$$D_{seq}(i, j) = \sum_{k=1}^{L} D_{seq}[k]$$

To summarize:

- S_i, S_j: HLA sequences i and j.
- L: sequence length.
- $S_i[k], S_j[k]$: amino acids at position k.
- $\text{BLOSUM}(S_i[k], S_j[k])$: Raw BLOSUM62 log-odds score.
- $D_{seq}[k]$: per-position distance (inverse likelihood).

2.3.2 TCR Vβ Preference Distance

For each HLA, we calculate the Vβ gene preference F (Section 2.2). Difference between two vectors F_i and F_j are measured with cosine distance:

$$D_{VGene}(i, j) = 1 - \frac{F_i \cdot F_j}{\|F_i\|\|F_j\|}$$

where:

- F_i and F_j: Vβ preference vectors for HLAs i and j.
- $F_i \cdot F_j$: dot product.
- $\|F_i\|$ and $\|F_j\|$: vector magnitudes.

2.3.3 Peptide Motif Distance

For each HLA, aligned peptide sequences from the MHC Motif Atlas dataset (Tadros et al., 2023) are used to build a peptide motif matrix as a PWM. Distances between peptide motifs are measured with JS divergence, yielding a symmetric distance matrix.

$$D_{motif}(i, j) = \text{JS}(P_i, P_j)$$

where:

- P_i and P_j: PWMs for HLAs i and j.

The following HLAs are absent from the Atlas and excluded from peptide analysis: DPA1 × 01:03+DPB113:01, DQA101:02+DQB106:09, DQA101:03+DQB106:01, DQA101:04+DQB105:03, DQA101:05+DQB105:01, DQA103:03+DQB102:02, DQA103:03+DQB103:01, DQA104:01+DQB1 × 04:02, DQA105:05+DQB103:01.

2.3.4 Distance Correlations

To capture non-linear relationships among D_{seq}, D_{VGene}, D_{motif}, we use Spearman correlation with a stringent 1% False Discovery Rate (FDR) $N = 1000$ bootstrapping replicates to ensure robust correlation estimates.

Each HLA position is assigned a correlation coefficient and standard deviation for TCRβ usage and peptide motifs, enabling identification of positions more strongly correlated with one or the other.

To avoid redundancy, correlations are computed only on the lower triangle of the distance matrices. For any pairs considered (i, j) where $i < j$:

$$\text{Correlation}_{seq-VGene} = \rho\left(D_{seq}(i,j), D_{VGene}(i,j)\right)$$

$$\text{Correlation}_{seq-motif} = \rho\left(D_{seq}(i,j), D_{motif}(i,j)\right)$$

$$\text{Correlation}_{VGene-motif} = \rho\left(D_{VGene}(i,j), D_{motif}(i,j)\right)$$

We also report the overall correlations from the full distance matrices.

2.4 HLA Position Importance Disentanglement

For each HLA position, $\text{Correlation}_{seq-VGene}$ and $\text{Correlation}_{seq-motif}$ provide coefficients for TCRβ usage and peptide motifs. Significant position are plottedin Figure 4C (per-locus plots and per-position values in Appendix E).

These positions are highlighted in their respective 3D structure for each locus (see Appendix E.3) and overall (see Figure 4D).

2.5 Vβ Promiscuity and HLA Restriction

To understand the breadth of HLA restriction for individual TCR Vβs, we calculate the conditional probability $P(\text{HLA}_i \mid \text{V}\beta_j)$. While our previous analysis focused on HLA preferences ($P(\text{V}\beta \mid \text{HLA})$), this "V-gene centric" view quantifies how specific or promiscuous a Vβ is across the HLA landscape.

We derive this by column-normalizing the co-occurrence matrix C:

- C_{ij}: count of TCR Vβ j for HLA i.

- $\sum_k C_{kj}$: total count of Vβ j across all HLAs.
- Restriction probability:

$$G_{ji} = \frac{C_{ij}}{\sum_k C_{kj}}$$

Each row j of matrix G represents the probability distribution of HLA alleles associated with a given Vβ. To identify Vβs with shared HLA binding preferences, we perform hierarchical clustering on the columns using cosine distance.

Finally, we quantify the "promiscuity" of each Vβ using the Shannon entropy of its HLA distribution:

$$H(\mathrm{V}\beta_j) = -\sum_i G_{ji} \log_2 G_{ji}$$

To compare restriction breadth across HLA loci with varying numbers of alleles (N), we calculated the *normalized entropy* ($H_{\text{norm}} = H/\ln N$), which scales the measure between 0 and 1. We interpret Vβs with low normalized entropy as "specialists" and those with high normalized entropy as "generalists."

Differences in normalized entropy distributions between loci were assessed using pairwise Mann-Whitney U tests. To account for multiple comparisons across all locus pairs, p-values were adjusted using the Bonferroni correction.

Class	**Locus**	**# Alleles**	**HLA-TCR**		**HLA-Peptide**	
			Spearman $\rho \pm$ SD	**p-Value**	**Spearman $\rho \pm$ SD**	**p-Value**
1	all	34	**0.30 ± 0.04**	**4.43E-13**	**0.25 ± 0.04**	**1.08E-09**
2	all	57	**0.60 ± 0.02**	**1.07E-155**	**0.52 ± 0.02**	**1.04E-74**
1	A	12	**0.35 ± 0.11**	**4.17E-03**	**0.57 ± 0.09**	**4.31E-07**
1	B	17	**0.29 ± 0.08**	**7.78E-04**	**0.33 ± 0.08**	**9.76E-05**
1	C	5	-0.18 ± 0.38	6.51E-01	0.57 ± 0.28	8.97E-02
2	DP	12	0.26 ± 0.12	3.22E-02	**0.80 ± 0.06**	**2.22E-13**
2	DQ	16	**0.27 ± 0.09**	**3.19E-03**	**0.48 ± 0.15**	**8.31E-03**
2	DR	29	**0.48 ± 0.04**	**1.00E-24**	**0.36 ± 0.05**	**5.25E-14**

Table 1: Spearman ρ coefficients and bootstrap standard deviation (SD) of HLA sequence polymorphism measured with BLOSUM Distance, comparing TCR Vβ gene Preference with cosine distance and Peptide Motifs with JS Divergence for all loci and classes. A significance threshold of 0.01 was used and p-Values below this threshold are in bold. See details about TCR-HLA in Appendix Section C and HLA-Peptide in Appendix Section D.

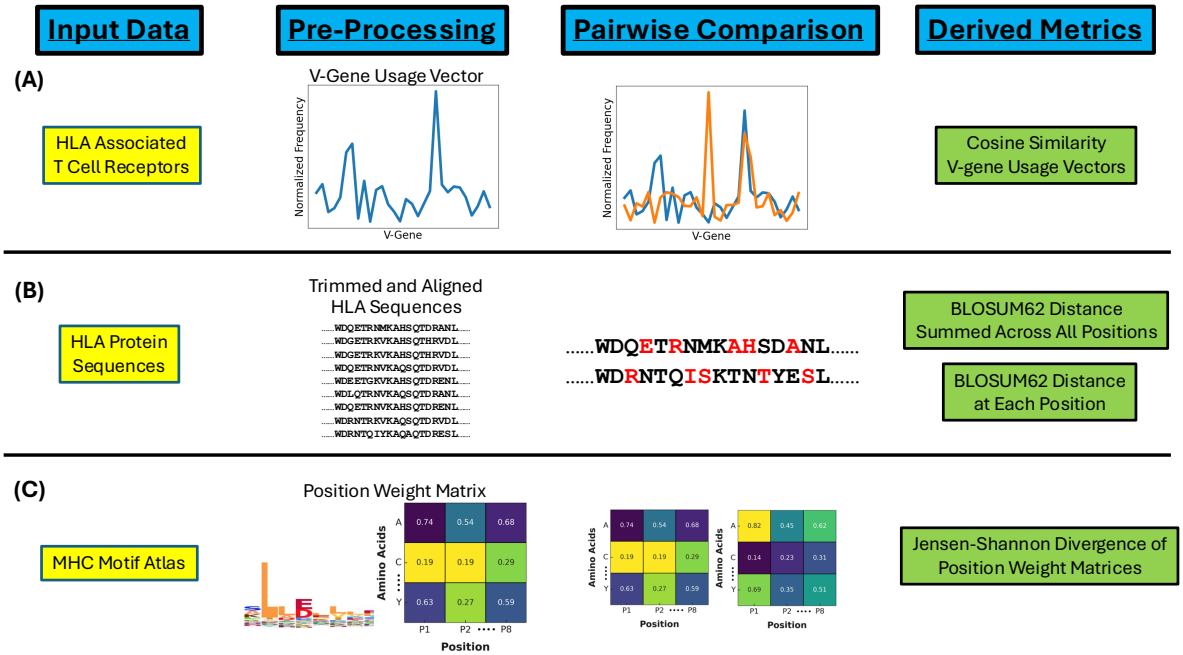

Figure 1: **Summary of our analysis workflow. (A)** Starting from sets of TCRβs associated with specific HLAs, we generate normalized Vβ gene usage vectors for each HLA. We quantify TCR Vβ usage similarity in a pairwise manner, calculating the cosine distance between the Vβ gene usage vectors of two HLAs.
(B) Starting from HLA protein sequences, HLA sequences are trimmed to the TCR-binding surface and aligned via Multiple Sequence Alignment (MSA) (Bradley, 2023). We quantify both overall and position-dependent HLA polymorphism by comparing HLA sequences using the BLOSUM62 substitution matrix.
(C) We use MHC Motif Atlas (Tadros et al., 2023) to generate Position Weight Matrices (PWMs) of the peptide repertoire for a given HLA. PWMs encode the same information as logo plots, indicating the frequency of amino acid use at any given position. Using PWMs, we calculate the Jensen-Shannon Divergence to quantify similarity of the peptide repertoire of any two HLAs.

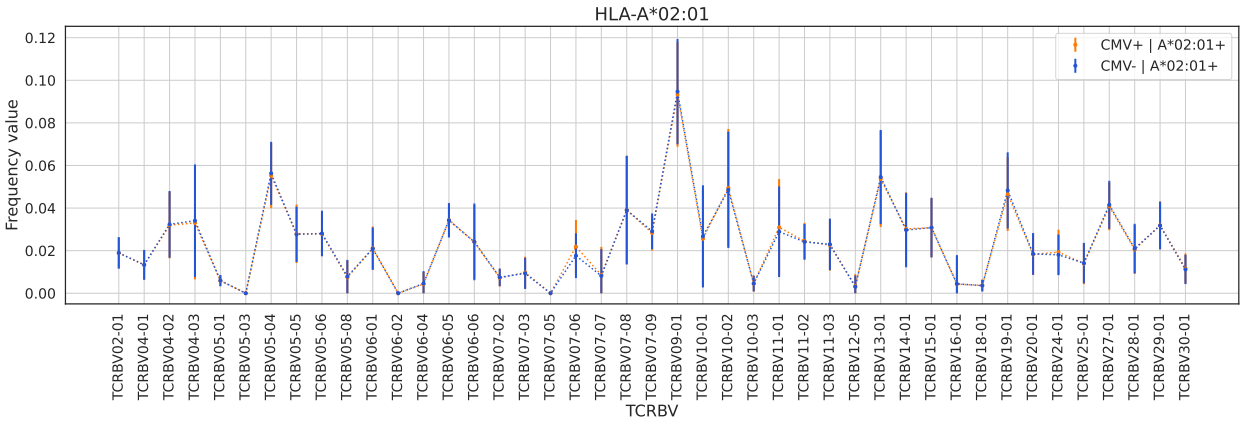

Figure 2: **TCR Vβ usage patterns among HLA-A*02:01-associated sequences are unaffected by CMV exposure.** Background-normalized Vβ gene usage across HLA-A*02:01-positive donors is shown for CMV-positive (orange) and CMV-negative (blue) cohorts. Error bars indicate the standard deviation across individuals within each group. Usage patterns are consistent between CMV-positive and CMV-negative subjects, suggesting that CMV exposure does not influence HLA-associated Vβ preferences.

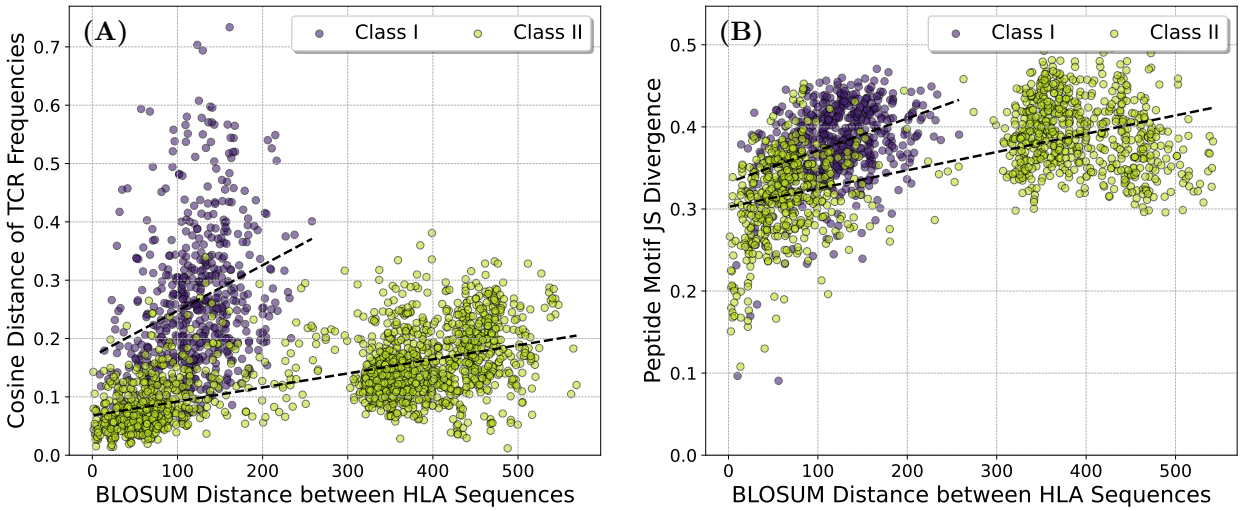

Figure 3: **A pairwise comparison showing the relationship between HLA polymorphism, Vβ gene usage and the peptide repertoire. (A)** TCR Vβ usage similarity plotted against HLA polymorphism. HLAs with similar sequences tend to have similar Vβ gene preferences. Dotted lines illustrate correlations which we quantify using partial Spearman's ρ (class I: ρ 0.30 $\pm$ 0.04, class II: ρ 0.60 $\pm$ 0.02).
(B) Peptide repertoire similarity plotted against HLA polymorphism. HLAs with similar sequences tend to have similar peptide repertoires. Dotted lines illustrate correlations which we quantify using partial Spearman's ρ (class I: ρ 0.25 $\pm$ 0.04, class II: ρ 0.52 $\pm$ 0.02). See Table 1 for per locus Spearman's ρ and p values and Appendix Section C for TCR-HLA and Appendix Section D.1 for HLA-peptide.

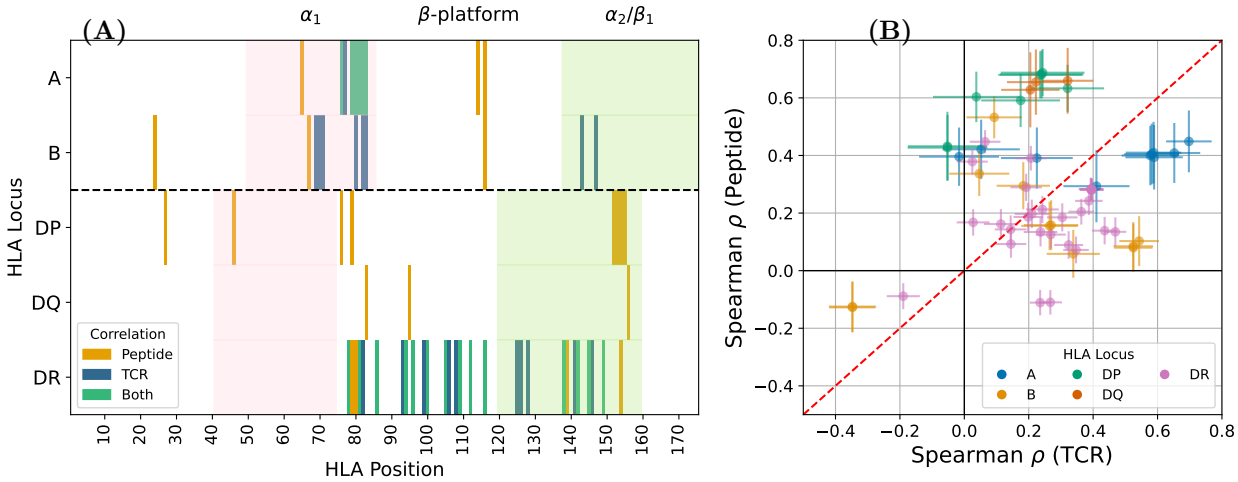

Figure 4: **Disentangling HLA positions correlated with Vβ gene usage vectors, the peptide repertoire or both. (A)** Across all loci, HLA polymorphic positions significantly correlated (p-value ≤ 0.01) with similarity in Vβ gene usage (blue), peptide repertoire similarity (yellow) and both (green). Shaded are the rough positions of the HLA binding groove helices. To allow positional comparison, for Class II we combined the $\alpha1$ and $\beta1$ sequence. In light red is the $\alpha1$ helix region (closest to TCRβ). In light green is the $\alpha2$ helix (or $\beta1$ for Class II) (closest to TCRα). See also per-locus plots in Appendix C (TCRβ) and D (peptide).
(B) Scatter plot of Spearman's ρ values and bootstrapped standard deviation for the correlation of HLA polymorphism with peptide motif similarity versus its correlation with Vβ gene usage similarity. Only HLA positions where at least one significant correlation is derived are plotted. See per-locus plots and per-position values in Appendix E.

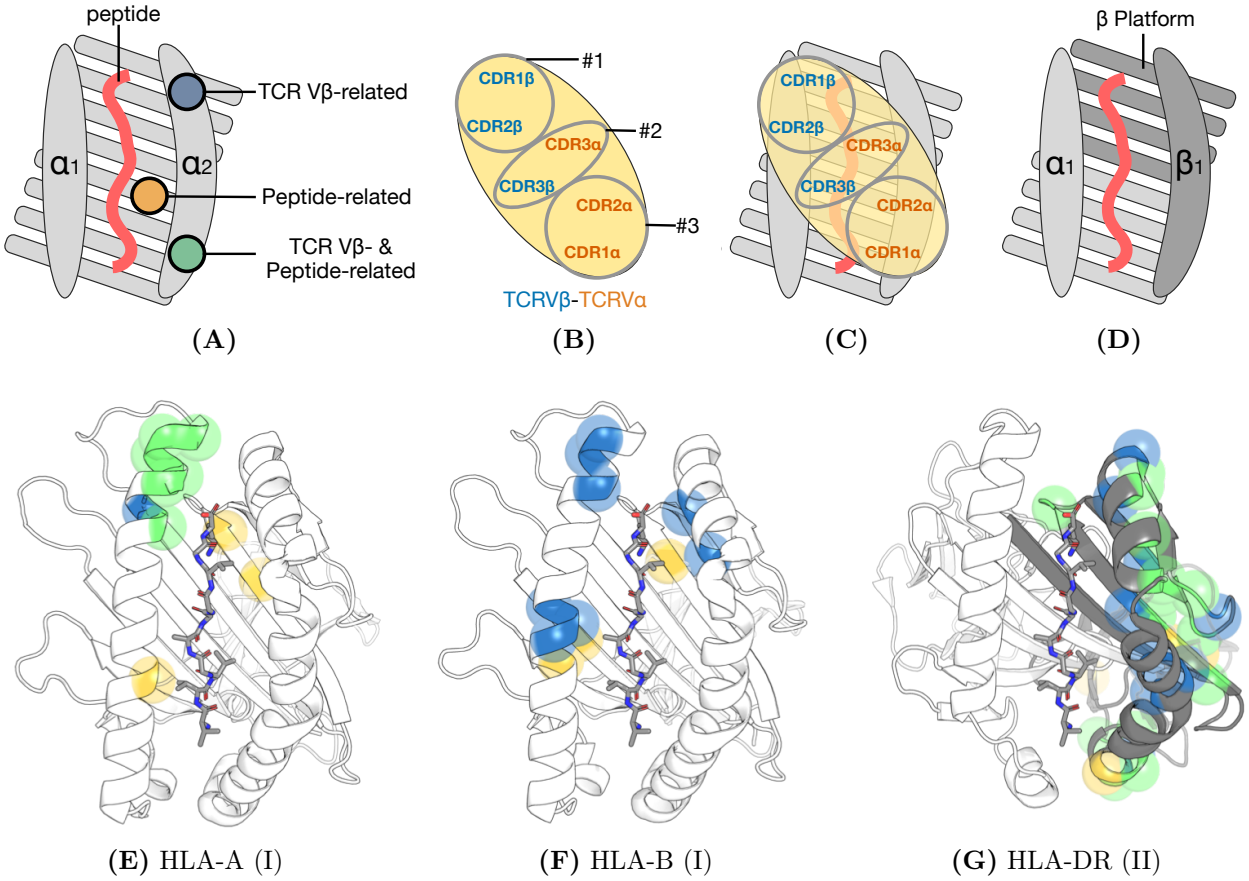

Figure 5: **Structural logic of TCR–pHLA engagement and distribution of significant residues.**
(A) Schematic of the peptide-binding groove in Class I HLAs ($\alpha1$ and $\alpha2$ helices), illustrating three categories of polymorphic residues: those related to TCR Vβ (blue), peptide (yellow) or both (green).
(B) Schematic of the TCR binding interface (footprint), showing the arrangement of CDR loops categorized into three functional 'topes' (Cohn et al., 2019): Tope #1 (TCRβ CDR1/2) and Tope #3 (TCRα CDR1/2) engage the HLA helices, while Tope #2 (hypervariable CDR3s) primarily engages the peptide. The TCR Vβ domain (blue text, corresponding to Tope #1) typically contacts the $\alpha1$ helix in the canonical diagonal orientation.
(C) Overlay of the TCR footprint onto the Class I binding groove, illustrating the proximity of Vβ CDRs to the $\alpha1$ helix.
(D) Schematic of the Class II binding groove, formed by the $\alpha1$ and $\beta1$ chains. Note that unlike Class I, for Class II HLAs, the β platform is encoded by the α chain (light grey) and β chain (dark grey).
(E–G) Correlated positions highlighted on HLA-A (E), HLA-B (F) and HLA-DR (G). Residues are colored by their statistical association: TCR Vβ (blue), peptide (yellow) or both (green). These positions are identified in Fig. 4 (See details in Supplementary Section D).

3 Results

TCR recognition of pHLA complexes is primarily mediated by three complementarity-determining regions (CDRs). CDR1 and CDR2 are encoded by the Vβ gene, while the hypervariable CDR3 region

is generated by V(D)J recombination (Davis and Bjorkman, 1988; Murphy and Weaver, 2017b). To better understand how germline-encoded TCR features and HLA polymorphisms shape pHLA recognition, we examine how HLA polymorphism correlates with TCR Vβ gene usage, focusing on CDR1 and CDR2 regions and distinguishing direct HLA effects from peptide-mediated influences. Our goal is to disentangle the contributions of direct TCR–HLA interactions from those arising indirectly through variation in peptide presentation.

Our analysis draws on large-scale repertoire sequencing data linking TCRβs with HLAs across a cohort of approximately 30,000 individuals (see Methods Section 2.1). Vβ gene usage patterns for common HLAs are derived from TCRβs statistically associated with each HLA and aggregated into vectors of normalized frequencies (Supplementary Figure S2), which we interpret as Vβ gene preferences of each HLA. To validate that these patterns are stable and not confounded by individual immune exposure, we compare repertoires from Cytomegalovirus (CMV) -positive and CMV-negative subjects, restricting analysis to HLA-associated TCRs observed in individuals carrying the relevant HLA allele. We use CMV exposure as a test case as it typically has a strong impact on the T cell repertoire (Klenerman and Oxenius, 2016) and can be inferred directly from the TCRβ repertoire (Emerson et al., 2017; May et al., 2024). This makes CMV a strong test case: if a single dominant exposure perturbs Vβ usage patterns, we should detect it in CMV-positive subjects. As shown in Figure 2, Vβ usage is consistent between CMV-positive and CMV-negative subjects, supporting the conclusion that Vβ gene preferences reflect HLA-driven selection rather than exposure-dependent biases. Using the T-Detect COVID cohort (Zahid et al., 2025b) of approximately 30,000 individuals, we infer both CMV exposure status (Emerson et al., 2017) and HLA genotype (Zahid et al., 2025a). We focus on individuals carrying HLA-A02:01, comparing Vβ usage among CMV-positive (N = 5,449) and CMV-negative (N = 8,136) subjects. For each person, we quantify Vβ usage across all HLA-A02:01–associated TCRβ sequences in their repertoire (45,737 unique sequences). We observe no significant difference in Vβ usage patterns between the CMV-positive and CMV-negative groups (see Figure 2). Restricting the analysis to other HLAs yields consistent results.

Although this analysis does not rule out contributions from other antigen exposures, it shows that even a strong immune stimulus like CMV does not measurably influence HLA-associated Vβ usage patterns. This supports the interpretation that the observed associations reflect generalizable, population-level trends.

Next, we analyzed how HLA polymorphism influences TCRβ recognition and peptide binding. We represent HLA polymorphism using aligned sequences trimmed to the TCR-facing surface, including peptide-binding domains (Bradley, 2023). Peptide repertoires for each HLA are obtained from the MHC Motif Atlas (Tadros et al., 2023) and represented as Position Weight Matrices (PWMs), which capture amino acid frequencies across positions (Parker et al., 1994) (Supplementary Table S1). Then, we define three pairwise distance metrics between HLA alleles (e.g., *A02:01* vs. *A02:02*): (i) HLA sequence divergence, calculated from BLOSUM62-derived substitution scores (Henikoff and Henikoff, 1992), either summed across aligned residues to yield a global similarity score or evaluated at individual positions for residue-level analysis; (ii) TCR Vβ usage similarity, computed as the

cosine distance between normalized Vβ gene frequency vectors; and (iii) peptide motif divergence, measured as Jensen-Shannon (JS) divergence between PWMs. This framework enables us to assess whether specific HLA polymorphisms are more strongly associated with the peptide repertoire or with the set of TCRβs likely to recognize a given HLA. A summary of the analysis workflow is shown in Figure 1.

Before examining how HLA sequence variation relates to similarity in Vβ gene usage across alleles, we first characterize the overall breadth of Vβ restriction exhibited by different HLA loci.

3.1 Variation in Vβ Preference Breadth Across HLA Loci

To assess how HLA polymorphism affects the breadth of the T cell repertoire, we quantified how "exclusive" the Vβ gene usage was for each HLA gene using Shannon entropy (H_{norm}) on the conditional probability distribution $P(\text{HLA} \mid \text{V}\beta)$. In the case of HLA-associated TCRs restricted to single Vβ gene, the entropy score would be close to 0. Conversely, uniform usage of all Vβ genes would yield a score closer to 1.

As shown in Supplementary Figure S18, there is a significant difference in restriction breadth across loci. Class II loci HLA-DR and HLA-DP displayed normalized entropy values approaching the theoretical maximum ($H_{\text{norm}} \approx 0.99$), corresponding to a nearly uniform distribution of Vβ usage across alleles. In contrast, HLA-C exhibited a significantly lower mean entropy and higher variance compared to HLA-DR (see Supplementary Table S10). HLA-C shows the lowest entropy and highest variance across loci, indicating a high degree of restriction due to specific Vβ genes being associated with a limited subset of alleles. HLA-A, -B and -DQ display intermediate entropy profiles significantly distinct from both the HLA-C and HLA-DR/DP extremes (Supplementary Table S10)

These differences highlight substantial locus specific variation in the overall permissiveness of Vβ gene usage. With this baseline heterogeneity in mind, we next examine how differences in HLA sequence relate to similarity in Vβ gene preferences and peptide binding across alleles.

3.2 Global Correlations Across HLAs

To test whether HLA polymorphism systematically shapes TCR and/or peptide features, we analyze correlations between global HLA sequence similarity (summed across all aligned positions), Vβ gene usage and peptide motif similarity within each HLA class. We find that overall HLA sequence similarity correlates with both TCR Vβ gene preference (Figure 3A) and peptide motif similarity (Figure 3B; Table 1), even when using partial Spearman correlation to control for HLA locus structures. These results suggest that variation in HLA sequence alone is sufficient to predict meaningful trends in Vβ gene usage, highlighting the potential for sequence-based features to inform models of TCR–HLA specificity.

At the locus level, HLA sequence similarity correlates with both Vβ gene and peptide features for loci A, B, DQ and DR (see Table 1 and Figure 4). DR shows the strongest association with Vβ gene preference, while DP shows the strongest correlation with peptide motifs but not with Vβ gene features. Locus C does not show significant correlations in any analysis, likely due to its lower

surface expression and resulting in limited statistical power (Neefjes and Ploegh, 1988). Notably, we observe that Vβ gene usage frequencies are more variable among Class I alleles compared to Class II, which exhibit more uniform usage patterns across alleles (Supplementary Figure S3).. These findings indicate that the relationship between HLA polymorphism, TCRβ engagement and peptide binding is both locus- and class-dependent, with distinct correlation patterns emerging across class I and class II molecules.

To further probe the structural basis of these patterns, we restrict our analysis to HLA residues known to contact either the peptide or the TCR, based on solved pHLA–TCR structures. This restriction strengthens the observed correlations (see Supplementary Figure S1), suggesting the correlation is driven by a subset of positions and thus motivating a focused, residue-level investigation of TCR–HLA specificity.

3.3 Residue-Level Correlation Analysis

We hypothesize that polymorphic HLA residues influence TCR recognition through distinct mechanisms, with some acting directly on TCR engagement while others act indirectly via influence on peptide binding. To test this, we analyze residue-specific HLA polymorphisms, which are largely concentrated in the peptide-binding groove, while more distal regions remain conserved (Supplementary Figure S5). This pattern of polymorphism further suggests that only a subset of residues drive correlations with Vβ gene usage and peptide motifs. We quantify each residue's contribution by correlating per-residue polymorphism with both Vβ gene preference and peptide motif similarity (see Supplementary Figures S8 and S11). This analysis identifies HLA residues statistically associated with TCR recognition, peptide binding or both.

We visualize the spatial distribution of residues significantly associated with Vβ gene usage or peptide motifs across class I and class II loci (Figure 4; see Supplementary Sections B–D for details). Peptide-associated residues are broadly distributed across the binding groove, though in HLA-A they cluster near the α1, closest to the TCRβ. Residues associated with Vβ gene usage appear primarily in loci A, B and DR, clustering near the TCR-facing helices in class I and spanning the polymorphic DRB chain in class II. As expected, significant positions in DR are restricted to the polymorphic DRB chain, as the invariant DRA chain does not contribute variation in our polymorphism-based analysis. These findings suggest that residues correlated with peptide and TCR features occupy overlapping but structurally distinct regions, with class-specific patterns across loci.

To compare the relative influence of each HLA position on TCR Vβ usage and peptide presentation, we compute residue-specific Spearman correlation coefficients for both signals. For each polymorphic residue, we calculate its correlation with Vβ gene preference and peptide motif similarity. We then plot all positions with at least one significant correlation in Figure 4B. Points above the identity line indicate stronger correlations with peptide motifs, while those below indicate stronger correlations with Vβ gene usage. Positions in loci A and B are fairly balanced around the line, suggesting a mix of peptide- and TCR-associated effects. In contrast, most significant positions in loci DP and DQ are peptide-associated, while positions in DR cluster near the identity line with a modest bias toward

TCR association. These results highlight locus-specific differences in how HLA residues influence TCR and peptide features, with distinct correlation patterns across class I and class II loci.

Surprisingly, several positions exhibit negative correlations, indicating that sequence differences at these sites are associated with greater similarity in peptide motifs or Vβ gene usage (see Supplementary Table S4 to Table S8). Although the reasons underlying this effect remain unclear, we hypothesize that such positions may reflect compensatory or co-evolutionary structural effects, where variation at one site may alter the HLA conformation in a way that preserves peptide presentation and/or TCR binding despite sequence divergence. Further structural and mutational studies are needed to test this hypothesis and clarify how these residues contribute to pHLA recognition.

3.4 Structural Distribution of Significant Positions

We visualize the significant positions on the 3D structure of an HLA molecule (Figure 5). Most fall within the peptide-binding groove, the most polymorphic region (Supplementary Figure S5). Residues more strongly correlated with peptide motifs (yellow) are primarily located on the β-sheets beneath the peptide and on the peptide-facing α-helices. In contrast, residues associated with TCRβ features or with both TCRβ and peptide features are predominantly found on the α-helices. In class I HLAs, these TCR- and dual-associated positions cluster near the region of the α1 helix (closest to the TCRβ). In class II HLAs, they are more broadly distributed across the binding groove, particularly along the polymorphic DRB chain, consistent with the invariant nature of the DRA chain.

Together, these analyses show that HLA polymorphisms shape TCR recognition and peptide presentation by partially overlapping but distinct sets of HLA polymorphic contacts. Some residues directly influence TCR engagement, others act indirectly via peptide binding and a subset contribute to both. The balance of these effects varies across loci, with clear differences between class I and class II HLAs. This residue-level map of functional variation offers a valuable foundation for refining models of TCR–pHLA specificity and guiding future structural and computational work in immunogenetics.

4 Discussion

Our analysis leverages a large-scale statistical framework, using 30,000 TCRβ repertoires to associate specific TCRβs with individual HLAs. By averaging across thousands of individuals and diverse but commonly shared antigenic exposures, we derive normalized Vβ gene preferences for each HLA, enabling a robust, population-level assessment of TCRβ-HLA preferences. This approach yields sets of TCRβs that are statistically enriched for HLA specificity, providing statistical power to link TCRβs to their restricting HLAs despite the polygenic nature of HLA phenotypes.

To understand how HLA polymorphism shapes peptide binding and TCR engagement, we analyze correlations between HLA sequence similarity, peptide motifs and Vβ gene usage across the cohort. Our goal is to distinguish direct TCR–HLA associations from peptide-mediated effects. We

find that HLA sequence similarity correlates with both Vβ gene usage and peptide motifs. These associations map to specific polymorphic residues linked to either Vβ preference, peptide features or both. Structurally, key positions cluster around the α1 helix in class I and the β1 helix in class II, with particularly strong signals in the DR locus. These results suggest that TCR–HLA specificity is shaped by a combination of germline-encoded contacts and indirect peptide-driven effects, including potential allosteric modulation of pHLA conformation. Together, these patterns highlight distinct structural contributions to recognition across HLA classes. Notably, the clustering of TCR-associated residues on the Class I α1 helix structurally confirms the germline contact maps proposed by Marrack et al. (2008).

A longstanding question in immunology is whether TCR–pHLA specificity is governed more by inherited structural constraints or by developmental selection, reflecting broader questions about how evolutionary forces (Feng et al., 2007; Marrack et al., 2008; Christopher Garcia et al., 2009; Yin et al., 2012) and thymic processes (Van Laethem et al., 2007, 2012; Lu et al., 2019) shape the TCR repertoire. Evidence for intrinsic specificity includes recognition of pHLAs by randomly generated T cells (Krovi et al., 2019), QTL associations between HLAs and TCR Vβ usage (Sharon et al., 2016; Gao et al., 2019) and alloreactivity (Felix and Allen, 2007), where TCRs selected on self-HLAs cross-react with non-self HLAs. Our findings support this view. Namely, polymorphic residues, especially in the α1 helix of class I and the β1 helix of class II, correlate with Vβ gene preference but not peptide motifs, consistent with germline-encoded structural bias and allele-specific recognition (Cohn et al., 2019). We also identify residues associated only with peptide motifs and those associated with both peptide motifs and Vβ genes, indicating multiple mechanisms through which HLA polymorphism shapes TCR engagement. A parsimonious interpretation is that TCR–HLA interactions reflect both co-evolution and selection (Rangarajan and Mariuzza, 2014; La Gruta et al., 2018), with germline contacts **priming** initial binding and thymic selection, subject to these germline constraints, **sculpting** the final repertoire. This layered model helps reconcile competing hypotheses and offers a unified framework for understanding repertoire architecture.

Our analysis suggests that Vβ restriction follows a continuum rather than a simple Class I versus Class II split. HLA-DR and HLA-DP appear to be "generalists" with high entropy and broad compatibility. On the other hand, HLA-C seems to act as a "specialist" with lower entropy compared to other HLAs. This complements Gao et al. (2019)'s observation that HLA-C has fewer TCR associations than other Class I loci. While they attribute this to low surface expression, our data implies that these associations are more specific for HLA-C than for other loci. Therefore, the weak HLA-C signal could result from a combination of low abundance and structural selectivity. This divergence in part reflects the functional needs of CD8+ T cells to have precise recognition to avoid autoimmunity, whereas CD4+ T cells benefit from broader scanning capabilities.

Several findings raise compelling questions for future investigation. In class I HLAs, Vβ gene associations cluster near the α1 helix, which directly contacts the TCRβ chain. In contrast, HLA-DR associations are exclusively driven by the polymorphic β chain (DRB), since the α chain (DRA) remains constant. This provides a residue-level mechanism for the *HLA-DRB1* expression QTLs

reported by Sharon et al. (2016). Given that in Class II HLAs, the TCRβ primarily contacts the α1 helix, it suggests that DRB polymorphisms may influence TCR binding indirectly, for example by changing pHLA conformation or stability, rather than through direct contact with the Vβ domain. The absence of similar correlations in HLA-DQ and -DP, where both chains are polymorphic, underscores the unique role of the invariant DRA chain in shaping recognition. Additionally, we identify HLA positions with negative correlations to TCRβ or peptide similarity, potentially reflecting compensatory or co-evolutionary effects that preserve recognition despite sequence divergence. These patterns suggest novel structural mechanisms of immune recognition and point to the value of follow-up studies using mutagenesis, structural modeling or biophysical assays.

Despite the statistical power and scale of our framework, several limitations should guide interpretation. Our analysis focuses exclusively on the TCRβ chain and does not account for potential contributions from the TCRα chain, which also engages the pHLA surface and may exhibit its own germline biases. Likewise, our emphasis on HLA polymorphism may overlook conserved germline-encoded residues that play critical roles in TCR recognition. While we focus on Vβ gene–encoded CDR1 and CDR2 regions, we do not account for the hypervariable CDR3 loop, which is central to peptide recognition and also influences HLA binding. In addition, peptide repertoires used to define motif similarity are drawn from public datasets often biased toward disease-associated antigens, potentially limiting their generalizability. Furthermore, some observed correlations may arise from linkage disequilibrium between HLA alleles, rather than from direct structural contacts. Conversely, the absence of correlation does not imply a lack of interaction but may reflect limited power, resolution or data sparsity. Structural modeling and molecular dynamics simulations could help resolve these ambiguities by assessing the functional impact of specific residues. Finally, as high-throughput technologies improve, future studies incorporating paired TCR$\alpha\beta$ repertoires and more representative peptide libraries will be essential to validate and expand these insights.

A key assumption of our analysis is that the structural associations identified in the public repertoire reflect fundamental biophysical constraints applicable to all TCRs. Under this assumption, the Vβ–HLA associations we identify provide important constraints for modeling TCR-pHLA interactions. The $P(V\beta \mid \text{HLA})$ and $P(\text{HLA} \mid V\beta)$ matrices provide simple priors that can bias a model toward germline-consistent TCRs when assigning HLAs or ranking candidate receptors. This is particularly useful in low-data settings (e.g. rare HLAs or small clinical cohorts), where these priors could help prevent models from overfitting to noise. These priors can also be incorporated into benchmarking via a simple prior-weighted accuracy metric.

Our findings support a model of TCR–HLA specificity shaped by both inherited constraints and adaptive processes. Germline-encoded biases appear to define the baseline potential for Vβ engagement with HLA, while thymic selection and immune experience further sculpt the repertoire. By leveraging tens of thousands of repertoires to derive Vβ usage patterns and mapping these onto HLA sequence and structure, we perform a residue-level analysis that provides insights into the molecular basis of TCR–HLA interactions. Our approach demonstrates the power of population-scale, structure-aware statistical analysis to uncover fundamental principles of immune recognition.

These insights offer not only a clearer understanding of repertoire formation but also a scalable and interpretable framework to guide future efforts in TCR modeling, repertoire engineering and immunotherapy design.

Conflict of Interest Statement

HJ Zahid, L Pisani, J Greissl have employment and equity ownership with Microsoft. M Noakes, HS Robins have employment and equity ownership with Adaptive Biotechnologies. LV Castorina has employment with AstraZeneca. The authors declare no other competing interests.

Author Contributions

LVC, MTN, HCH, and HJZ were involved in the writing of the paper. LVC worked on the analyses and code. LVC, MTN, HJZ created the visualizations. LP adviced on code structure. JG, HJZ, MTN, and HCH mentored LVC throughout the stages of the work. All authors provided feedback on the drafts of the paper.

Funding

The work was funded by Microsoft Corporation and Adaptive Biotechnologies.

Acknowledgments

We thank Haiyin Chen for her persistent encouragement to pursue this problem, which ultimately motivated the work presented here.

Supplemental Data

Supplementary Material should be uploaded separately on submission, if there are Supplementary Figures, please include the caption in the same file as the figure. LaTeX Supplementary Material templates can be found in the Frontiers LaTeX folder.

Data Availability Statement

All data and code are available at https://github.com/microsoft/tcr-mhc.

References

Bradley, P. (2023). Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife* 12, e82813. doi:10.7554/eLife.82813

Castorina, L. V., Grazioli, F., Machart, P., Mösch, A., and Errica, F. (2025). Assessing the generalization capabilities of tcr binding predictors via peptide distance analysis. *PLOS One* 20, e0324011. doi:10.1371/journal.pone.0324011

Christopher Garcia, K., Adams, J. J., Feng, D., and Ely, L. K. (2009). The molecular basis of tcr germline bias for mhc is surprisingly simple. *Nature immunology* 10, 143–147

Cohn, M., Anderson, C. C., and Dembic, Z. (2019). The case for allele-specific recognition by the tcr. *Scandinavian Journal of Immunology* 90. doi:10.1111/sji.12790

Davis, M. M. and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature* 334, 395–402. doi:10.1038/334395a0

Drost, F., Chernysheva, A., Albahah, M., Kocher, K., Schober, K., and Schubert, B. (2025). Benchmarking of t cell receptor-epitope predictors with epytope-tcr. *Cell Genomics* 5

Emerson, R. O., DeWitt, W. S., Vignali, M., Gravley, J., Hu, J. K., Osborne, E. J., et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire. *Nature genetics* 49, 659–665

Felix, N. J. and Allen, P. M. (2007). Specificity of t-cell alloreactivity. *Nature Reviews Immunology* 7, 942–953

Feng, D., Bond, C. J., Ely, L. K., Maynard, J., and Garcia, K. C. (2007). Structural evidence for a germline-encoded t cell receptor–major histocompatibility complex interaction'codon'. *Nature immunology* 8, 975–983

Gao, K., Chen, L., Zhang, Y., Zhao, Y., Wan, Z., Wu, J., et al. (2019). Germline-encoded tcr-mhc contacts promote tcr v gene bias in umbilical cord blood t cell repertoire. *Frontiers in Immunology* 10. doi:10.3389/fimmu.2019.02064

Grazioli, F., Machart, P., Mösch, A., Li, K., Castorina, L. V., Pfeifer, N., et al. (2023). Attentive Variational Information Bottleneck for TCR–peptide interaction prediction. *Bioinformatics* 39, btac820. doi:10.1093/bioinformatics/btac820

Grazioli, F., Mösch, A., Machart, P., Li, K., Alqassem, I., O'Donnell, T. J., et al. (2022). On TCR binding predictors failing to generalize to unseen peptides. *Frontiers in Immunology* 13, 1014256. doi:10.3389/fimmu.2022.1014256

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 10915–10919. doi:10.1073/pnas.89.22.10915

[Dataset] Jensen, M. F. and Nielsen, M. (2024). NetTCR 2.2 - Improved TCR specificity predictions by combining pan- and peptide-specific training strategies, loss-scaling and integration of sequence similarity. doi:10.7554/eLife.93934.2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature* 596, 583–589

Karnaukhov, V., Paes, W., Woodhouse, I. B., Partridge, T., Nicastri, A., Brackenridge, S., et al. (2022). HLA variants have different preferences to present proteins with specific molecular functions which are complemented in frequent haplotypes. *Frontiers in Immunology* 13, 1067463. doi:10.3389/fimmu.2022.1067463

Klein, J. and Sato, A. (2000). The hla system. *New England Journal of Medicine* 343, 702–709

Klenerman, P. and Oxenius, A. (2016). T cell responses to cytomegalovirus. *Nature Reviews Immunology* 16, 367–377

Krovi, S. H., Kappler, J. W., Marrack, P., and Gapin, L. (2019). Inherent reactivity of unselected tcr repertoires to peptide-mhc molecules. *Proceedings of the National Academy of Sciences* 116, 22252–22261. doi:10.1073/pnas.1909504116

La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G., and Rossjohn, J. (2018). Understanding the drivers of MHC restriction of T cell receptors. *Nature Reviews Immunology* 18, 467–478. doi:10.1038/s41577-018-0007-5

Lu, J., Van Laethem, F., Bhattacharya, A., Craveiro, M., Saba, I., Chu, J., et al. (2019). Molecular constraints on cdr3 for thymic selection of mhc-restricted tcrs from a random pre-selection repertoire. *Nature communications* 10, 1019

Marrack, P., Scott-Browne, J. P., Dai, S., Gapin, L., and Kappler, J. W. (2008). Evolutionarily conserved amino acids that control tcr-mhc interaction. *Annu. Rev. Immunol.* 26, 171–203

May, D. H., Woodhouse, S., Zahid, H. J., Elyanow, R., Doroschak, K., Noakes, M. T., et al. (2024). Identifying immune signatures of common exposures through co-occurrence of t-cell receptors in tens of thousands of donors. *BioRxiv* , 2024–03

Murphy, K. and Weaver, C. (2017a). *Janeway's Immunobiology* (Garland Science, Taylor & Francis Group), 9th edn.

Murphy, K. M. and Weaver, C. (2017b). *Janeway's Immunobiology* (New York London: GS, Garland Science, Taylor & Francis Group), 9th edition edn.

Nakamura, T., Shirouzu, T., Nakata, K., Yoshimura, N., and Ushigome, H. (2019). The Role of Major Histocompatibility Complex in Organ Transplantation- Donor Specific Anti-Major Histocompatibility Complex Antibodies Analysis Goes to the Next Stage -. *International Journal of Molecular Sciences* 20, 4544. doi:10.3390/ijms20184544

Neefjes, J. J. and Ploegh, H. L. (1988). Allele and locus-specific differences in cell surface expression and the association of hla class i heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class i subunit association. *European journal of immunology* 18, 801–810

Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994). Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *Journal of immunology (Baltimore, Md.: 1950)* 152, 163–175

Rangarajan, S. and Mariuzza, R. A. (2014). T cell receptor bias for mhc: co-evolution or co-receptors? *Cellular and molecular life sciences* 71, 3059–3068

Rossjohn, J., Gras, S., Miles, J. J., Turner, S. J., Godfrey, D. I., and McCluskey, J. (2015). T cell antigen receptor recognition of antigen-presenting molecules. *Annual review of immunology* 33, 169–200

Rudolph, M. G., Stanfield, R. L., and Wilson, I. A. (2006). How tcrs bind mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.* 24, 419–466

Sewell, A. K. (2012). Why must t cells be cross-reactive? *Nature Reviews Immunology* 12, 669–677. doi:10.1038/nri3279

Sharon, E., Sibener, L. V., Battle, A., Fraser, H. B., Garcia, K. C., and Pritchard, J. K. (2016). Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature Genetics* 48, 995–1002. doi:10.1038/ng.3625

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7, 539. doi:10.1038/msb.2011.75

Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology* 12, 664514. doi:10.3389/fimmu.2021.664514

Tadros, D. M., Eggenschwiler, S., Racle, J., and Gfeller, D. (2023). The MHC Motif Atlas: A database of MHC binding specificities and ligands. *Nucleic Acids Research* 51, D428–D437. doi:10.1093/nar/gkac965

Tiercy, J.-M. (2002). Molecular basis of HLA polymorphism: Implications in clinical transplantation. *Transplant Immunology* 9, 173–180. doi:10.1016/S0966-3274(02)00007-2

Van Laethem, F., Sarafova, S. D., Park, J.-H., Tai, X., Pobezinsky, L., Guinter, T. I., et al. (2007). Deletion of cd4 and cd8 coreceptors permits generation of $\alpha\beta$t cells that recognize antigens independently of the mhc. *Immunity* 27, 735–750

Van Laethem, F., Tikhonova, A. N., and Singer, A. (2012). Mhc restriction is imposed on a diverse t cell receptor repertoire by cd4 and cd8 co-receptors during thymic selection. *Trends in immunology* 33, 437–441

Yin, L., Scott-Browne, J., Kappler, J. W., Gapin, L., and Marrack, P. (2012). T cells and their eons-old obsession with mhc. *Immunological reviews* 250, 49–60

Zahid, H. J., Taniguchi, R., Ebert, P., Chow, I.-T., Gooley, C., Lv, J., et al. (2025a). Large-scale statistical mapping of t-cell receptor β sequences to human leukocyte antigens. *Frontiers in Immunology* Volume 16 - 2025. doi:10.3389/fimmu.2025.1603730

Zahid, H. J., Taniguchi, R., Noceda, M. G., Robins, H., and Greissl, J. (2025b). T cell receptor diversity, cancer and sex: Insights from 30,000 tcrβ repertoires. *bioRxiv* doi:10.1101/2024.10.21.619470