Conserved Biophysical Compatibility Among the Highly Variable Germline-Encoded Regions Shapes TCR-MHC Interactions

Christopher T. Boughter[1*] and Martin Meier-Schellersheim[1*]

[1]Computational Biology Section, Laboratory of Immune System Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892

[*]To whom correspondence should be addressed. Email: Christopher.Boughter@NIH.gov, mms@NIAID.NIH.gov

December 7, 2022

Abstract

T cells are critically important components of the adaptive immune system primarily responsible for identifying and responding to pathogenic challenges. This antigen recognition is based on the interaction between membrane-bound T cell receptors (TCRs) and peptides presented on major histocompatibility complex (MHC) molecules. The formation of the TCR-peptide-MHC complex (TCR-pMHC) can be broken into two types of interactions, one between the hypervariable TCR CDR3α/β loops and the presented peptide and the second between germline-encoded regions of the TCR and MHC. Experimental investigation of the latter interaction, mediated by the CDR1 and CDR2 loops of $\alpha\beta$ TCRs, has been historically difficult. To address this prominent gap in our knowledge, we analyzed every possible germline-encoded TCR-MHC contact in humans, thereby generating the first comprehensive characterization of these antigen-independent interactions. Our analysis shows that germline-encoded TCR-MHC interactions that are conserved at the sequence level are rare due to the exceptional diversity of the TCR CDR1 and CDR2 loops. Instead, binding properties such as the docking orientation are defined by regions of biophysical compatibility between these loops and the MHC surface. We further find that TCR genes which encode CDR loops with low potential to interact with MHC are over-represented in autoimmune contexts.

INTRODUCTION

The interaction between T cell receptors (TCRs) and the peptide-presenting major histocompatibility complex (pMHC) molecule is a nanometer-scale recognition event with macroscopic consequences for health and control of disease. In $\alpha\beta$ T cells, the six complementarity-determining region (CDR) loops of the TCR are responsible for scanning the pMHC surface and determining the appropriate immunological response. Since the first crystal structure of a TCR-pMHC complex was solved in 1996 [1], general "rules of engagement" between the TCR and the pMHC complex have been identified and reliably reproduced in subsequent structures [2,3]. First, the peptide is presented in the "groove" formed by the two α-helices of either the class I or class II MHC molecule [4,5]. The TCR then scans this peptide primarily using the highly diverse CDR3 loops of the TCRα and TCRβ chains, whose sequences are determined by V(D)J recombination [6, 7]. The sequences of the remaining two CDR1 and CDR2 loops of the α and β chains are entirely determined by germline TRAV and TRBV genes, respectively. In a majority of structures, these CDR1 and CDR2 loops of the TCR make a number of key contacts with well-defined regions of the MHC α-helices [8].

These contact regions define a docking orientation of the TCR over MHC molecules (Fig. 1A, B), which a majority of crystallized TCR-pMHC complex structures solved thus far adopt within a tight distribution of angles (Fig. 1C, D) despite strong variation in the CDR3 and peptide sequences involved. This orientational preference may be a consequence of thymic selection [9–11], whereby T cells are first positively selected on TCR recognition of MHC molecules and later negatively selected against TCR promiscuity towards MHC molecules loaded with organismal self-peptides [12–14]. This selection occurs in the presence of the potentially orientation-determining co-receptors CD4 or CD8. However, evidence for this hypothesis of docking determined by selection comes largely from a study utilizing a knockout system that deviates strongly from the physiological norm [9], or from the identification of rare "reverse-docking" TCRs [15,16]. Alternatively, the extent of this orientational preference may reflect a co-evolution of interacting, germline-encoded regions of TCR and MHC molecules. In support of this hypothesis, a number of mouse studies have identified key residue pairs critical for TCR-pMHC binding which were later structurally validated and found in other, distantly related organisms [17–22]. While these studies find evidence of conserved interactions for a small subset of alleles, no work thus far has found consistent conservation across the full range of possible TCR-MHC molecule interactions.

An important question is whether the germline-encoded contacts create a consistent framework for TCR-pMHC interactions and whether deviations from the canonical interaction mode contribute to errors in immune recognition. Multiple studies have found that specific germline genes, and therefore the corresponding CDR1 and CDR2 loops encoded by these genes, are strongly associated with autoimmune disorders. Specifically, a subset of TRBV alleles are associated with celiac disease [23, 24], multiple sclerosis [25], systemic lupus erythematosus [26], and type I diabetes [27]. Further, TCR-pMHC interactions with limited CDR1 and CDR2 contacts to MHC can amplify TCR cross-reactivity, and in some cases increase the probability of autoimmune responses [25,28–31]. These observations underscore the importance of a more thorough understanding of the fundamental role of CDR1 and

CDR2 in the recognition of pMHC complexes, either as sculptors of the TCR-pMHC complex, or as opportunistic reinforcements to a protein-protein interaction largely determined by CDR3-peptide binding. Due in part to the concentration of diversity at the CDR3-peptide interface, the vast majority of studies characterizing TCR-pMHC interactions have focused on describing immune recognition events in the context of this interface [18, 32–36]. This runs counter to the relative contributions of the germline-encoded regions that account for a majority (75-80%) of the binding interface [37].

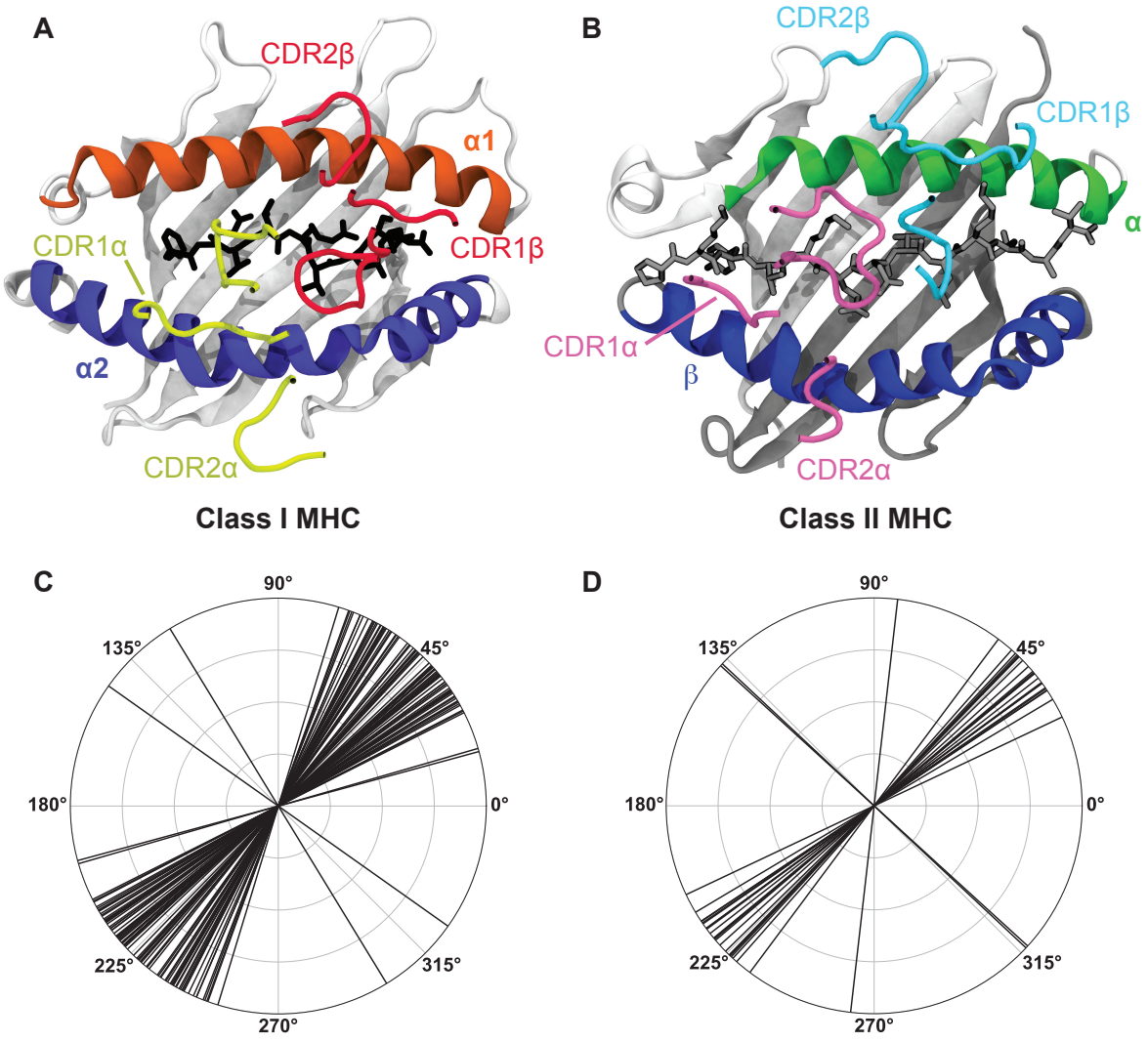

Figure 1: A breakdown of the canonical docking orientation adopted by TCRs over the pMHC complex in all solved crystal structures to date highlights the strong structural conservation of the TCR-pMHC interaction. (A, B) Example renderings of TCR-pMHC complexes for class I (A, PDB: 6MTM) and class II (B, PDB: 1J8H) MHC molecules. The CDR loops are largely representative of the placement over the MHC helices and peptide for the vast majority of TCR-pMHC complexes. (C, D) Polar coordinate plots of TCR docking angles over class I (C) and class II (D) MHC molecules (adapted from TCR3D database [38]). Note that these docking angles do not perfectly overlay on to the structures of panels A and B, as there are also slight deviations in the location of the TCR center of mass over the MHC complex.

While a majority of the 45 productive TCRα-encoding TRAV and 48 productive TCRβ-encoding TRBV genes in humans have been crystallized at least once in complex with pMHC [38], the total of 149 crystallized human TCR-pMHC complexes pales in comparison to the 2,160 possible productive TRAV-TRBV pairings. When we further consider the nearly 900 known unique amino acid sequences of class I MHC in humans (also referred to as class I human leukocyte antigen - HLA), we find that structural studies to date have covered less than 0.01% of the total possible TCR-MHC molecular interaction space. Further, the aforementioned complicated contributions of CDR3-peptide interactions to the TCR-pMHC complex stability make the study of germline-encoded contributions to TCR recognition of MHC molecules nearly intractable experimentally.

Using the Automated Immune Molecule Separator (AIMS) software [39, 40] we generated the first systematic characterization of every germline-encoded TRAV-MHC and TRBV-MHC interaction. The AIMS architecture allows us to analyze the CDR1 and CDR2 contributions to interactions with MHC molecules independent of CDR3, providing a novel, antigen-independent perspective on the TCR-MHC complex. We find that the exceptional diversity present in the CDR1 and CDR2 loops encoded by TRAV and TRBV genes in humans makes strongly conserved interactions between TCR germline-encoded residues and MHC α-helices highly unlikely. This suggests that the few strongly evolutionarily conserved interactions that have been identified previously in the literature should be considered exceptions and not the rule. Importantly, however, analyzing the biophysical compatibility for each germline-encoded CDR residue and those on the MHC α-helices, we find evidence of an evolutionary conserved region of permissible binding on the MHC molecular surface responsible for the observed canonical orientation of the TCR-MHC molecular complex. The few TRAV/TRBV genes lacking this biophysical compatibility are subsequently those most strongly associated with autoimmunity.

RESULTS

Starting from the TRAV and TRBV protein entries from the ImMunoGeneTics (IMGT) database [41, 42] and the IMGT-HLA database containing all identified HLA alleles [43], we include only productive genes and unique amino acid sequences in our final analysis dataset. This final dataset included a total of 882 unique class I MHC sequences (HLA-A, B, and C), 44 class II α chain sequences (HLA-DQα, DRα, and DPα), and 431 class II β chain sequences (HLA-DQβ, DRβ, and DPβ). Due to significant differences in the alignments and more nuanced differences in the structures of these MHC molecules, we analyzed these three groups separately and compare the results as appropriate.

Polymorphic HLA Molecules are Far Less Diverse Than Human TRAV and TRBV Genes

We first searched for potentially conserved interactions between TCR and MHC molecules that would suggest a strong coevolution between the two. Such biases in the amino acid usage of TCR and MHC molecules at specific sites can occur either in the form of full amino acid conservation, or as pairs of meaningfully co-varying amino acids found across multiple diverse sites. Without the structural data necessary to precisely identify these conserved interactions, we can characterize the diversity present in TRAV, TRBV, and HLA amino acid sequences using bioinformatic approaches and determine whether a conserved germline-encoded interaction is possible for each unique molecular combination. We quantified the diversity in these molecular subsets using concepts from information theory. Specifically, the amino acid diversity of each TRAV, TRBV, and HLA sequence encoded into an alignment matrix (Fig. 2A) was quantified by calculating the Shannon entropy [44] for each individual dataset (Fig. 2B).

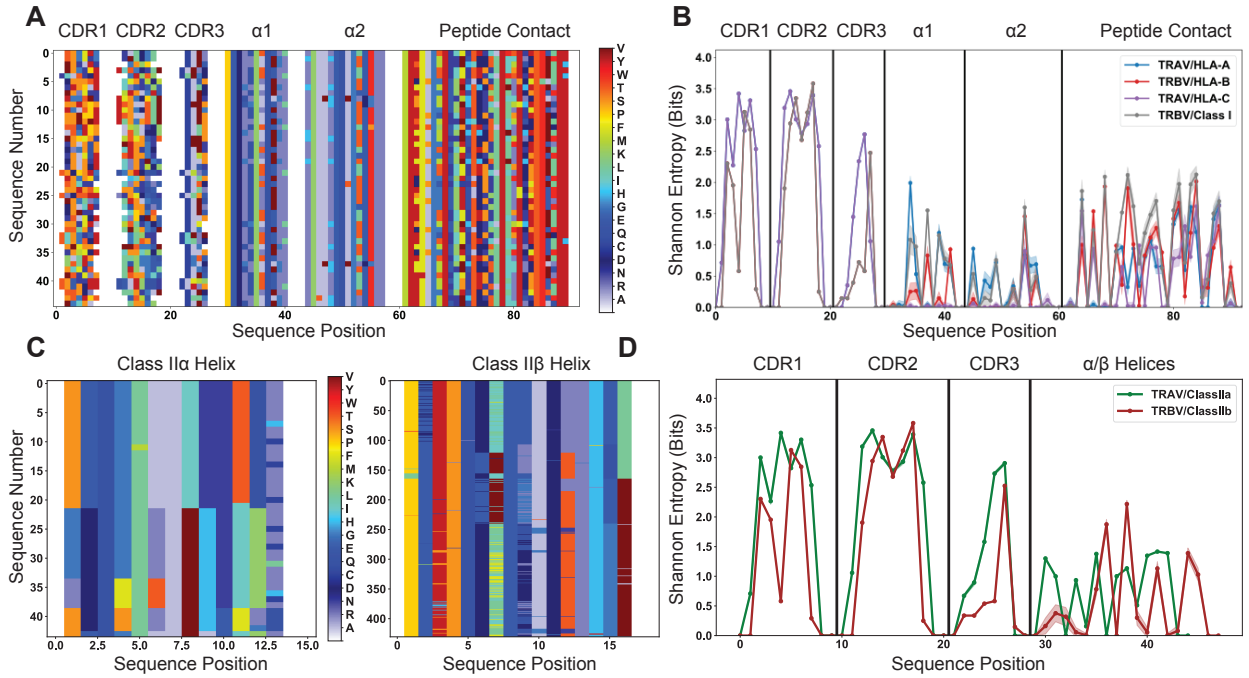

Figure 2: Quantification of diversity of TRAV, TRBV, and HLA alleles reveals limited sites for fully conserved interactions. (A) Visualization of the TRAV amino acid sequences paired with a subsampled HLA class I dataset. (B) This subsampling is repeated 1000 times to generate the average population Shannon entropy for each allelic subset as a function of position in the matrix of (A). (C) Visualization of the alignment-encoded matrices of HLA class IIa and classIIb datasets. (D) Calculated position-sensitive entropy for the HLA class IIa and class IIb datasets using the same subsampling scheme as (B). The color bar for panels (A) and (C) give the amino acid key for the matrix encoding, where each individual color represents an individual amino acid, and gaps in the matrix are colored white. In panels (B) and (D), the maximal entropy for the 20 possible amino acids at a given site is 4.3 bits, whereas an entropy of 0 bits represents a fully conserved amino acid. Any limited variation in the entropies over subsampling repetitions is represented by the standard deviation as shadowed regions around the solid line averages.

Fig. 2B indicates that the TRAV and TRBV genes are very diverse, even more so than the polymorphic HLA alleles. Despite the low number of TRAV and TRBV genes significant diversity persists in CDR1 and CDR2, suggesting limited conservation in those residues most likely to contact the MHC surface. Further, most of the entropy in MHC class I alleles is concentrated in the peptide-binding region. In this more diverse region the entropy tops out at around 2 bits, which corresponds to at most 4 equiprobable amino acids, or a distribution of a slightly larger number of amino acids strongly skewed towards one particular amino acid. In the α-helical regions that are in contact with the germline-encoded CDR loops, the class I entropy reaches a maximum of 1.5 bits, or at most 3 equiprobable amino acids. These same trends persist for the analysis of class II HLA alleles, with sequence diversity more limited in the class II α-helices when compared to the germline-encoded CDR loops (Fig. 2C, D).

HLA alleles are polymorphic, which can be misconstrued as having high sequence diversity in MHC molecules across the human population. However, the polymorphic sites are limited to a minority of the total variable positions within the MHC nucleotide sequences [45], which suggests even more limited variability at the amino acid level. While there are many unique alleles, these alleles are largely similar, especially when compared to other, more diverse protein families that are strongly structurally conserved but sequentially divergent, as in the neuronal Dpr-DIP system in *Drosophila melanogaster* [46]. Conversely, despite the low number of TCR germline genes present in humans, the diversity is substantial, with multiple sites across the CDR loops harboring over 10 unique amino acids. This mismatched diversity between the TCR germline-encoded CDR loops and the MHC α-helices suggests that conserved germline-encoded interactions are unlikely to exist for every possible molecular combination, either within a single individual or across the human population more broadly. Mutual information calculations (see Methods) corroborate these observations by quantifying this lack of co-variation of germline-encoded TCR and MHC residues in both humans (Fig. S1) and in a wider range of mammals (Figs. S2, S3).

Biophysical Diversity in TRAV and TRBV Genes Makes Strongly Conserved Contacts Across HLA Molecules Unlikely

The diversity in TRAV and TRBV germline-encoded regions that we had found excluded the possibility of conserved contacts as key mediators across all TCR-MHC germline-encoded interactions. However, this did not rule out the possibility that within this diversity, there are conserved biophysical interactions. For example, a site on CDR2β could be highly diverse, but only allow expression of hydrophilic residues of neutral charge. Pairing this with an α-helical site on HLA molecules which likewise only permits hydrophilic residues would represent a likely site of conserved interaction. We can test for such relations by quantifying the position-sensitive amino acid biophysical properties for both the germline-encoded TCR (Fig. 3A) and HLA (Fig. 3B) residues. Looking at both the position-sensitive charge (Fig. 3) and the position-sensitive hydropathy (Fig. S4), we see that the substantial diversity in the TRAV and TRBV sequences corresponds to a similar biophysical diversity in the amino acids found across the germline-encoded CDR loops.

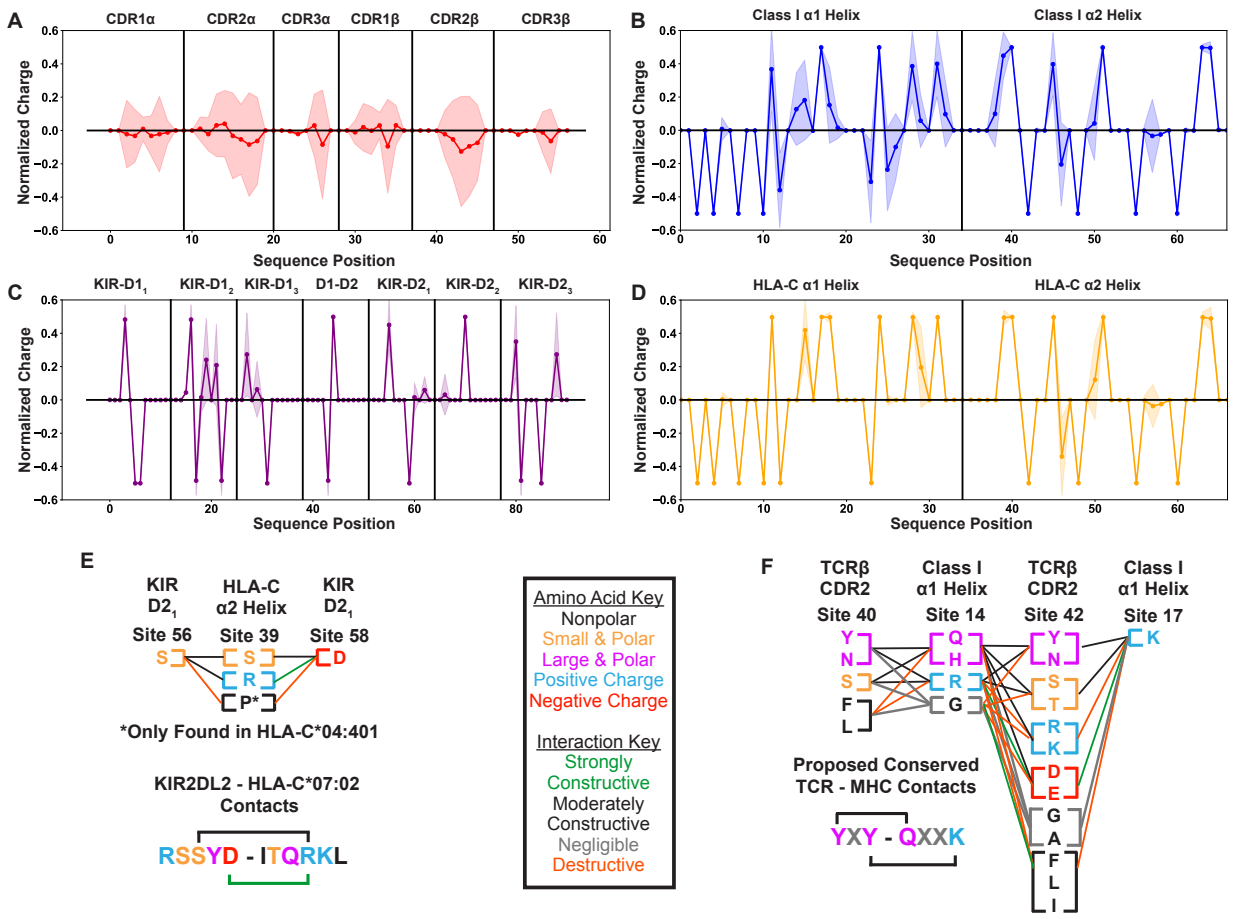

Figure 3: Position sensitive biophysical characterization of germline-encoded CDR loops highlights increased variation when compared to KIR or MHC sequences. Position sensitive amino acid charge averaged over all germline-encoded TCR CDR loops (A), MHC class I α-helices (B), KIR MHC-contacting regions (C), or HLA-C α-helices (D). Solid lines represent averages over unique amino acid sequences, while the standard deviations about these averages are given by the shaded regions. Visualization of specific interactions for KIR-MHC (E) or TCR-MHC (F) and their relative conservation across each respective dataset. Interactions are represented either in a single sequence context where a known binding network exists or as columns of residues present across polymorphisms or across TRBV genes. Key in the center gives color coding for either the biophysical properties of each amino acid (letters) or the relative contribution of these amino acids to an interaction interface (lines). Charges are normalized to a mean of 0 and a standard deviation of 1. Only polymorphisms across two-domain KIRs with well-characterized binding partners (KIR2DL1, 2DL2, 2DL3, 2DS1, 2DS2) are considered.

To put these results into perspective, we compared them to corresponding variabilities found for another class of molecules that interact with the MHC. The killer cell immunoglobulin-like receptors (KIRs) largely recognize the MHC class I subset HLA-C, subsequently relaying either inhibitory or stimulatory signals to the natural killer cell expressing these KIRs [47]. While each individual possesses a fixed subset of KIRs, with no analogous diversity introduced by a V(D)J-like process, polymorphisms in these KIRs exist across the human population [48]. The interactions between these KIRs and HLA-C molecules are more well-defined than the germline-encoded interactions between TCR and MHC molecules, with strong salt

bridge and hydrogen bond networks in place between two immunoglobulin-like domains of the KIR and each α-helix of HLA-C (Fig. S5A, B) [49, 50].

We found that from an initial dataset of 369 KIR polymorphisms compiled from the IPD-KIR database [48], only 31 productive alleles have unique amino acid sequences in the MHC molecule-binding region (Fig. S5C) with very limited diversity (Fig. S5D). Further, within the few moderately diverse sites, we found almost no variation in the biophysical properties of the amino acids contacting the MHC molecule, or in those of the amino acids on the HLA-C α-helices (Fig. 3C, D). We further contextualized these findings with specific examples from KIR-MHC and TCR-MHC molecular interactions (Fig. 3E, F). We selected a key hydrogen bonding network in the KIR2DL2-HLA-C*07:02 interface [50] and compared this to the evolutionarily conserved YXY motif of CDR2β [19, 21] found in a small subset of human TRBV genes. From this, we can see how diversity in each region supports or disrupts potentially conserved interactions at each site. Across all KIRs tested, the interacting residues are fully conserved, suggesting diversity is not tolerated at these sites. Similarly, diversity is exceptionally limited at the interacting sites on the HLA-C α2 helix.

Contrasting this with the regions involved in the YXY motif interactions, we see much higher diversity across both the TCR and MHC molecules. Considering specifically the conserved lysine on the α1 helix, we see that some of the residues utilized by TRBV-encoded CDR2 would create a strong conflict with the proposed conserved interaction site. While the YXY motif present on a few TRBV genes is clearly evolutionarily conserved [21] it is evident that substitutions either on CDR2β or on the MHC α1 helix would disrupt this interaction. These results strongly suggest that conserved interactions between germline-encoded CDR loops and MHC α-helices are deviations from the norm, and that CDR1 and CDR2 loops of TCRα and TCRβ likely adopt a more opportunistic binding strategy.

Calculation of TCR-MHC Interaction Potentials Finds Strong Differences In Germline-Encoded Recognition Strategies

While our information-theoretic and biophysical analyses suggested limited conserved contacts between TCR α/β CDR1 and CDR2 loops and MHC helices, they did not rule out coevolution at the level of biophysical compatibility over a broader interaction region. To explore the possibility of a coevolution that has not enforced genetic correlations but instead produced such compatibility among the TCR CDR loops and the MHC helices, we utilized a previously-validated [46] interaction scoring metric (see Methods). These interaction scores are calculated between all possible amino acid sequences of HLA alleles and TRAV or TRBV genes, providing a useful quantification of the relative contribution of germline-encoded CDR1 and CDR2 loops to a given TCR-MHC complex.

Fig. 4 shows the results of this scoring metric, focusing on the resultant patterns evident in both the x-axis, corresponding to the TRBV (Fig. 4A) or TRAV (Fig. 4B) genes, and the y-axis, corresponding to each HLA allele. We note that the variance in the strength of interaction with MHC class I is greater for the CDR1 and CDR2 loops of TCRα than those

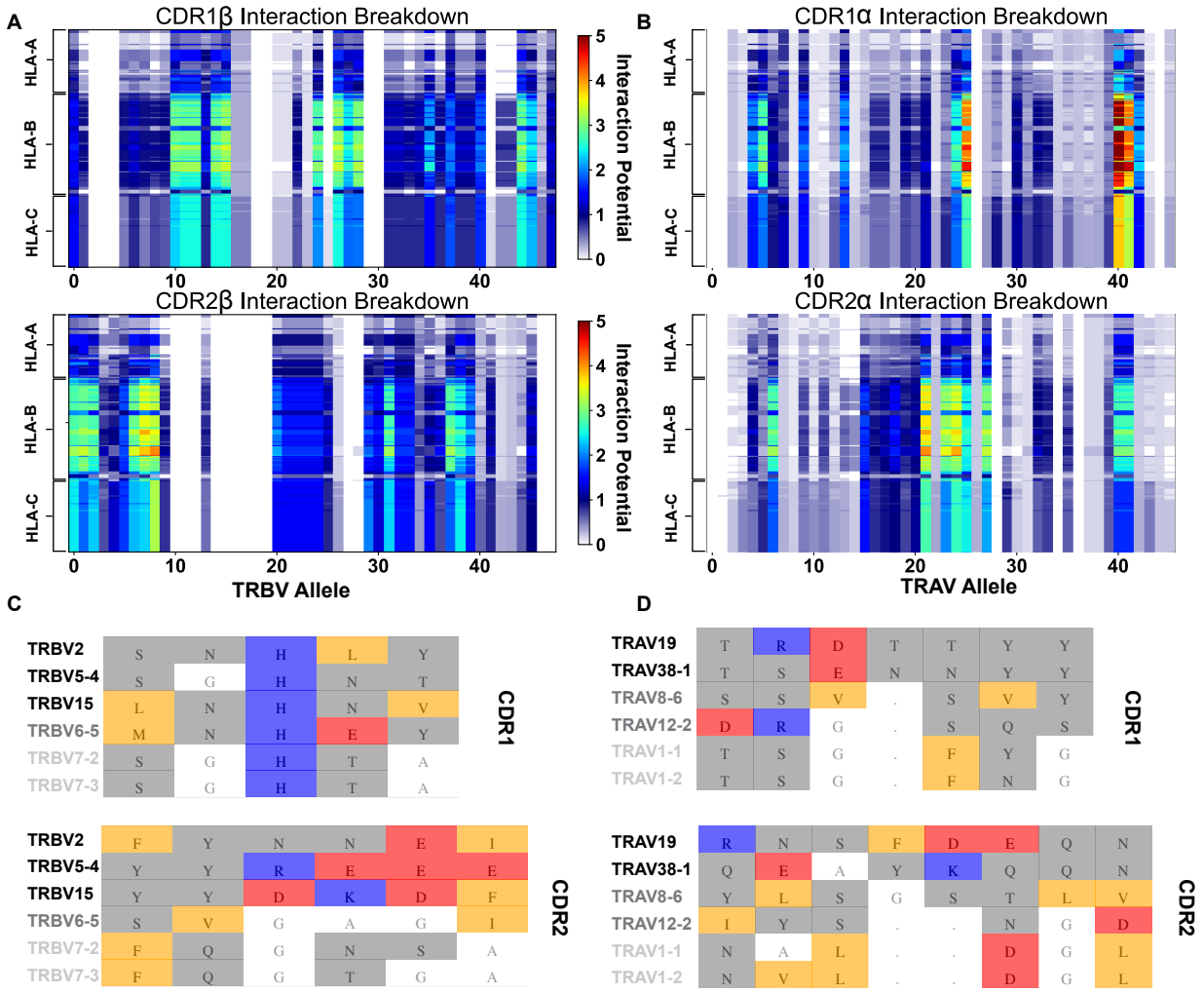

Figure 4: Interaction score between every TRBV (A) or TRAV (B) sequence and HLA allele for all four germline-encoded CDR loops. The x-axis moves across each productive germline-encoded TCR gene (see Supplemental Table S2 for key), while the y-axis is grouped by each broad HLA class I allele group. The color bar gives the interaction potential (unitless, higher potentials suggest stronger interactions). Alignments of TRBV (C) or TRAV (D) sequences highlighting genes with a range of interaction potentials, colored by biophysical property. Gene names for each sequence are colored by interaction potential: high-black, moderate-gray, and negligible-light gray. Color coding for alignment: grey - hydrophilic, blue - positive, red - negative, orange - hydrophobic, white - non-interacting. Gaps in an alignment are denoted by a dot (.).

of TCRβ (Fig. S6). Our analysis showed that TRAV8-6, 19, and 38-1 have the strongest potential interactions with class I HLA molecules of any TCR genes. Conversely, the TRBV genes 7-2 and 7-3 are predicted to have a severely limited interaction potential with HLA molecules (Fig. 4C, D). Interestingly, these alleles, TRBV7-2 and 7-3, are of significant interest in autoimmunity as they have been associated with celiac disease [23,24] and multiple sclerosis [25], respectively. Similarly, those TRBV alleles most strongly enriched in celiac patients (7-2, 20-1, and 29-1) [24] likewise all have a corresponding low AIMS interaction potential with HLA molecules.

These interaction potentials suggest that not only do germline-encoded CDR loops not contact the same conserved regions of the HLA molecule, but further imply that different TRAV/TRBV gene usage encourages differential utilization of the CDR loops. For instance, TRBV6-5 has a high CDR1 interaction potential and a negligible CDR2 interaction potential. Conversely, the suggested CDR bias is reversed for TRBV11-2. These trends are similarly found for analysis of germline-encoded interactions with HLA class II molecules (Fig. S7). These results further highlight the variability across the full repertoire of germline-encoded CDR1 and CDR2 loops and the role of this variability in interactions with MHC molecules.

Structural Data Validate Interaction Scores

To test our scoring of CDR-MHC interactions, we analyzed all TCR-pMHC complexes crystallized thus far. Using the TCR3D Database [38], we were able to process 132 total human TCR-pMHC complexes and quantify germline-encoded contacts. Final interactions included in the analysis were rigorously defined, with distinctions made between "productive" and "non-productive" interactions (Methods). This comparison shows exceptional agreement between our bioinformatic results and structural analyses (Fig. 5). Of the 179 total side-chain contacts between TCR CDR loops and MHC, 174 of these are made by CDR loops predicted to have a nonzero interaction potential (Fig. 5A). This prediction accuracy drops somewhat for class II complex structures (Fig. 5B), potentially due to the low number of total complex structures deposited into the PDB. The interaction potentials also succeed in predicting TCR complexes that will not make contacts with MHC. 20 of the 22 structures predicted to have poor CDR2β binding make no contact with MHC, while the last two only make one contact with MHC (Fig. 5C). Further, all 8 structures predicted to have poor CDR1β binding make no contact with MHC. (Fig. 5C). Again, this prediction accuracy is lower for class II predictions (Fig. 5D).

Not shown in these figures are the high number of TRAV and TRBV germline-encoded CDR loops which are predicted to make strong interactions with MHC but are found to make no interactions in crystallized TCR-pMHC complexes. In these structures, it appears that the CDR3-peptide interactions dominate the interface, overruling all other trends in germline-encoded interactions. Fig. 5E shows an extreme example, looking at the germline-encoded and peptide contacts of a so-called "super-bulged" peptide bound to MHC class I [51]. In this structure the TRAV19 and TRBV6-1 encoded CDR loops, which are both identified as having a high interaction potential with HLA class I molecules, make limited germline-encoded contacts with MHC side-chains. In this example, there is only a single germline-encoded interaction between TRAV19 and the MHC α2-helix. As is evident from this structure, significant CDR3-peptide interactions are capable of completely altering the TCR-MHC interface, regardless of which TRAV and TRBV genes are used. Absent these dominating CDR3-peptide contacts, the interaction potentials serve as a strong, structurally validated quantification of the likelihood of each germline-encoded CDR loop to interact with MHC α-helices.

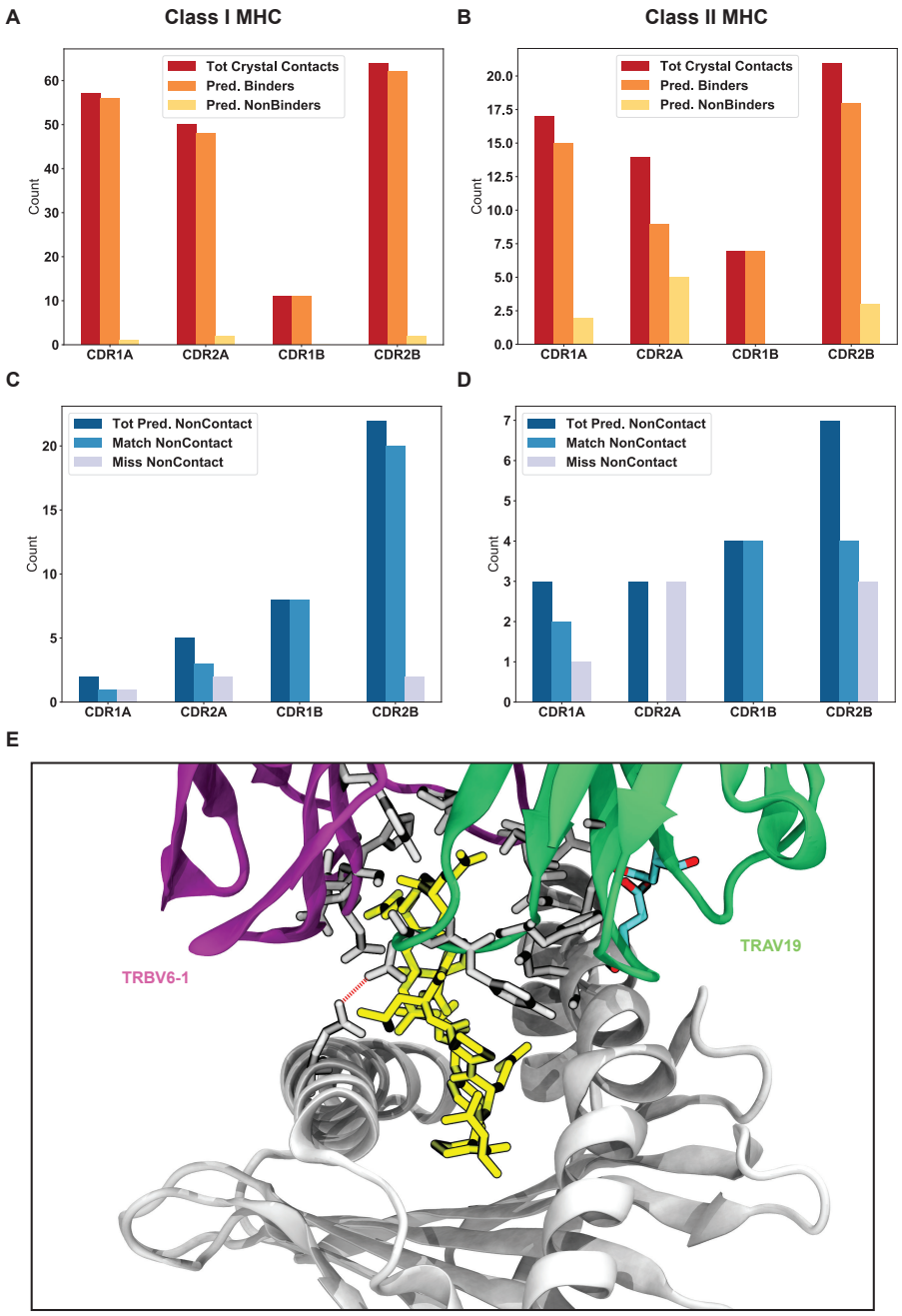

Figure 5: Comparison of interaction predictions to crystallized pMHC complexes. (A, C) Class I MHC comparisons. (B, D) Class II MHC Comparisons. Top row (A, B) Red – total number of contacts. Orange bar – total number of crystal contacts contributed by TRAV/TRBV genes that are predicted to bind. Yellow bar – total number of contacts contributed by TRAV/TRBV genes that are predicted not to bind. Bottom row (C, D) Blue bar – total number of PDBs with predicted nonbinding TRAV/TRBV. Cyan bar – total number of PDBs with TRAV/TRBV that are predicted not to bind and do not. Grey bar – total number of PDBs with TRAV/TRBV that are predicted not to bind but do bind. (E) Rendering of a TCR-MHC complex dominated by CDR3-peptide interactions (PDB:2AK4). Interactions with CDR3 or peptide are rendered in gray, germline-encoded interactions are colored cyan, and the peptide is shown in yellow.

Biophysical Constraints on the MHC Surface Create Germline-Encoded Biases in Docking Angles

Having experimentally validated the interaction matrices in Fig. 4, we explored whether there exist structural biases in the interactions between germline-encoded TCR CDR loops and MHC α-helices. First, we divided these matrices into separate datasets for each HLA group A, B, and C. We then calculated the interaction potentials across every HLA allele and TRAV or TRBV gene and identified the relative contribution of each CDR loop interacting with each HLA α-helix. From these distributions of contributions, we found that the constructive TCR β-chain interactions with the HLA α1-helix (Fig. 6A), and perhaps to a lesser extent destructive interactions between the TCR α-chain and the MHC α1-helix (Fig. S8), moderately bias the binding orientation between TCR and class I HLA molecules. This reliance on TCR β-chain interactions to guide the overall TCR-MHC interaction has been suggested previously in the literature based upon multiple early observations in mice [52]. It is also in line with previous findings that CDR1 and CDR2 of the TCR β-chain are key drivers of the TCR-MHC docking orientation [19,37]. Conversely, the α2-helix is only weakly involved in the determination of binding orientation (Fig. 6B). The exposed residues on the α2-helix of HLA class I molecules are enriched in alanine and glycine relative to the α1-helix, which are highly unlikely to be involved in a specific, orientation-altering productive interaction.

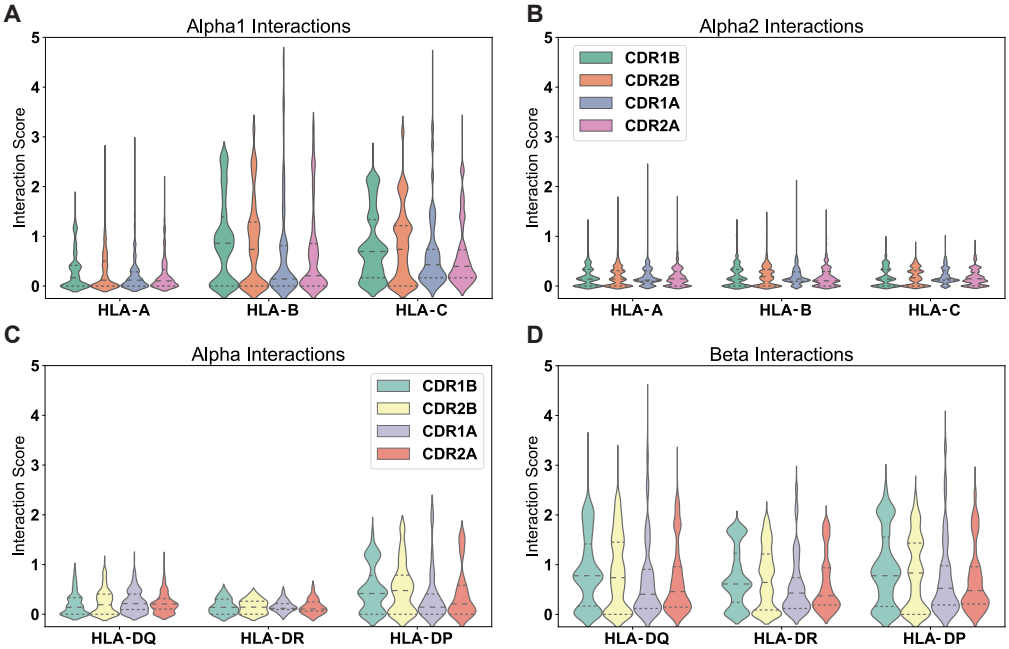

Figure 6: Positive interactions between each germline-encoded TCR CDR loop and all HLA alleles. Violin plots give distribution of interaction scores, with dashed lines separating quartiles of the distributions. The individual interaction potentials are shown for the individual CDR loop interactions with MHC Class I α1-helix (A), MHC Class I α2-helix (B), MHC Class II α-helix (C), and MHC Class II β-helix (D). Reported values are averages for each individual TRAV or TRBV gene and their interactions with each individual MHC allele over both CDR loops.

Interestingly, the TCR-HLA interaction appears flipped for class II molecules: interactions between TCR germline-encoded regions and HLA class II molecules are broadly driven instead by the β-chain helix (Fig. 6C, D). We note that the interaction potential incorrectly predicts that the TCR β-chain CDR loops should be the strongest interactors with the HLA class II β-chain helix. Every crystallized TCR-HLA class II complex solved thus far adopts the canonical docking orientation whereby the TCR β-chain binds to the HLA α-chain helix, while the TCR α-chain binds to the HLA β-chain helix. It is important to note that these interaction potentials take an unbiased approach, calculating every possible interaction between TCR and MHC residues to produce this final score. Therefore, this inconsistency between our interaction potential distributions and crystallized structures may arise from variations across a given MHC helix obscuring the interaction potentials within the canonical binding region, obviating the need for a more precise approach. To more carefully characterize the CDR loop interactions with MHC α-helices and attempt to alleviate this inconsistency, we isolated the contributions of each individual residue in the HLA class I and class II sequence alignments to the interaction potentials (Fig. 7).

Fig. 7 A,B show a structural visualization of the amino acids of interest on the class I (Fig. 7A) and class II (Fig. 7B) HLA molecules. These TCR-exposed residues are colored based approximately upon the register a TCR would adopt upon binding each subset, whereby TCRs adopting the canonical binding orientation would contact the residues colored in cyan and pink, and non-canonical binders would contact the blue and red residues. For both HLA class I (Fig. 7C, top) and class II (Fig. 7D, top), we found a conserved pattern of specific regions of peak interaction potentials surrounded by immediate drop-offs in these interaction potentials. In all but the HLA class I α1-helix, the interaction potentials adopt a distinctly bimodal distribution over the surface of the HLA molecules. In the case of class I molecules, our interaction analysis suggests that the TCR β-chain binds to the HLA class I α1-helix nearly indiscriminately, with potential to form many interactions. The TCR α-chain then contributes more significantly to the orientation of the interactions by binding to the HLA class I α2-helix in one of two discrete regions. Conversely, interactions between the $\alpha\beta$ TCR and HLA class II molecules appear to have a more defined register for binding, with both the HLA class II α-chain and β-chain interactions with TCR forming distinct bimodal distributions.

We then compared these by-residue interaction breakdowns to crystallized TCR-pMHC structures, counting the number of productive side-chain interactions between CDR loops and the solvent-exposed residues of the MHC helices. For HLA class I (Fig. 7C, bottom), we found excellent agreement between our predictions and the crystal contacts. Particularly for contacts with the α2-helix, we see that the majority of CDR1/2α contacts fall directly within the predicted regions of non-zero interaction potential. Comparisons to HLA class II (Fig. 7D, bottom) are more difficult, again due to the relative dearth of crystallized structures. Overall, these results further contextualize the seemingly inconsistent findings of Fig. 6D. While the interactions with the class II β-chain are stronger, we see that the average over all interactions with the α-chain are subdued by the very weak interaction potential of the non-canonical binding region. Within the canonical binding regions of the class II molecule, the TCR α-chain and β-chain CDR loops have similar binding potentials with either α-helix.

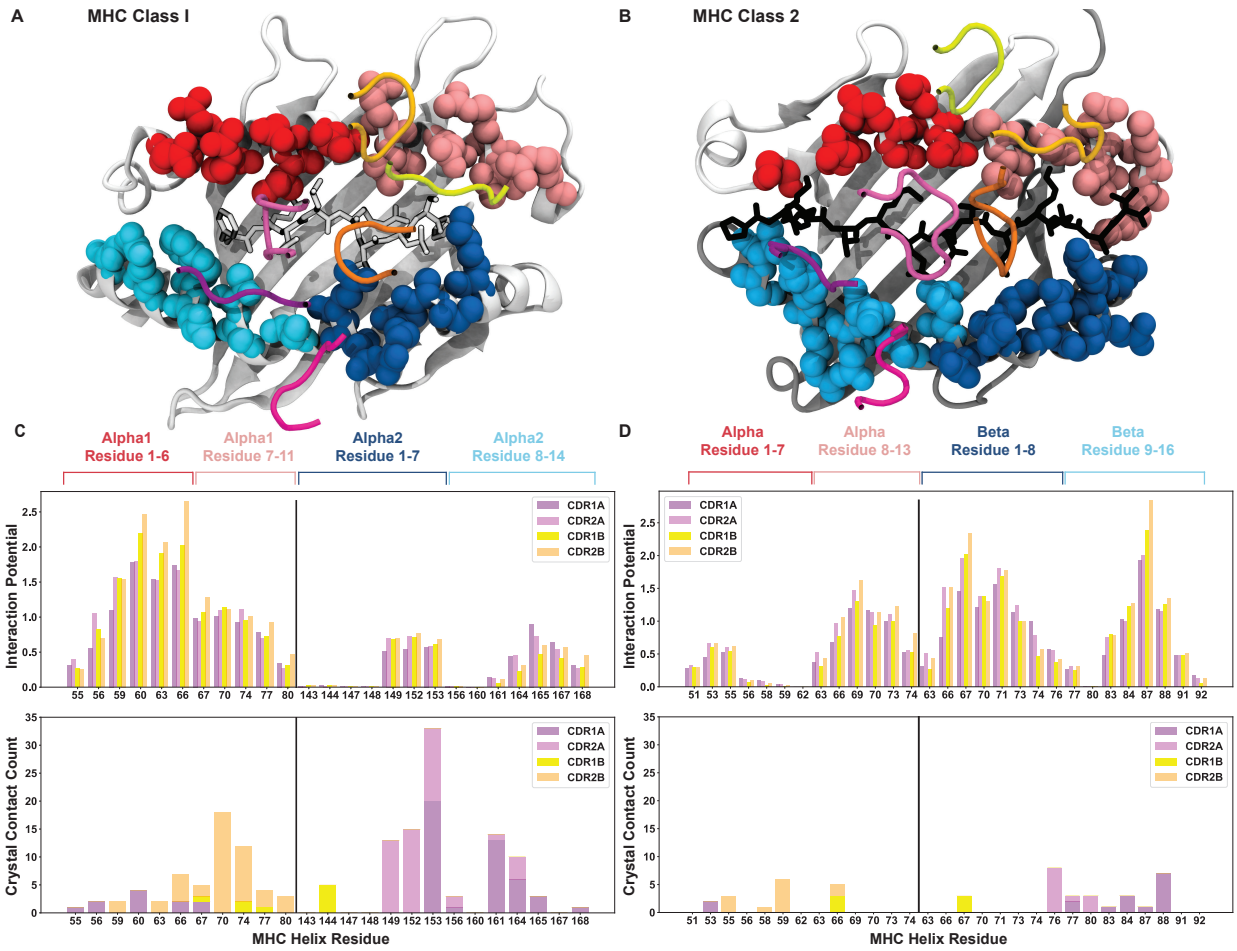

Figure 7: Breakdown of the interaction potential for every exposed residue on HLA surfaces. Color coded structures of HLA class I (A, PDB: 6MTM) and class II (B, PDB: 1J8H) provide the structural context to the bioinformatic results. The colors of the amino acid side-chains match the labels of the later panels. The averaged interaction potential across TRAV and TRBV genes with each individual MHC helix is shown in the top panels for HLA class I (C) and HLA class II (D). Lower panels of (C) and (D) compare these interaction potentials to total crystal contacts with these same residues.

The Binding Register of MHC Molecules is Well Conserved by Areas of Low Interaction Potential

The identification of these localized regions of low interaction potential contrasts with what has been postulated previously to be the root of the conserved TCR-pMHC docking orientation. Instead of confirming that conserved interactions determine the docking orientation, our results suggested that regions that are less likely to contact the TCR generate this preference. To further characterize these regions of low interaction potential flanked by regions of increased interaction potential on the MHC surface, we examined position-sensitive biophysical properties to characterize the exposed residues of the MHC α-helices. Specifically, we directly compared the position-sensitive Shannon entropy to the position-sensitive hydropathy of these residues across all MHC class I (Fig. 8A) and class II (Fig. 8B) molecules.

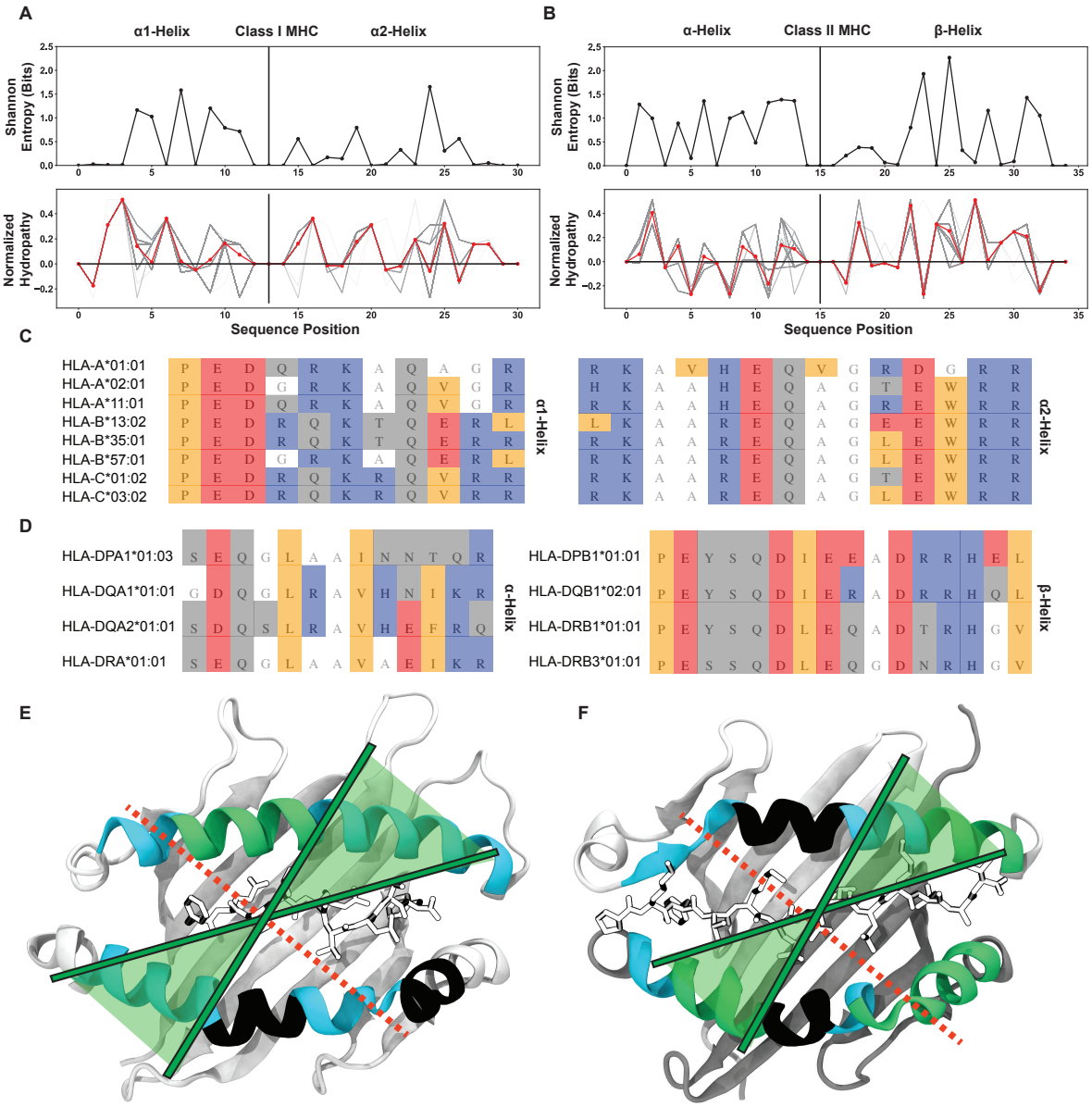

Figure 8: Identification of well conserved regions of low interaction potential finalize a working model for the root cause of canonical TCR-MHC docking orientations. Position-sensitive Shannon entropy (top) and normalized amino acid hydropathy (bottom) for class I (A) and class II (B) HLA molecules. Red lines in the hydropathy plots indicate an average over all HLA molecules, while gray lines give the position-sensitive biophysical properties of individual molecules. Alignments of class I (C) or class II (D) HLA alleles from a subsampling of parental alleles, colored by biophysical property. Color coding for alignment: grey - hydrophilic, blue - positively charged, red - negatively charged, orange - hydrophobic, white - non-interacting. Renders of class I (E, PDB: 6MTM) and class II (F, PDB: 1J8H) HLA molecules with α-helices colored by interaction potential. Green - regions of high interaction potential, cyan - regions of moderate interaction potential, black - regions of negligible interaction potential. Green lines and filled translucent space span regions of increased interaction potential, while the red dotted line gives a TCR docking orientation that may be theoretically allowed but is structurally unfavorable due to the conserved kink in the helix.

We found that in many of the regions where the hydropathy is at or near zero, meaning the region is neither hydrophilic nor hydrophobic, there is drop in the entropy to a value near zero. These residues with a hydropathy near zero likewise have an interaction potential near zero. This suggests that these residues with a lower interaction potential are well-conserved across HLA molecules. Indeed, looking again at alignments of HLA class I (Fig. 8C) and class II (Fig. 8D) sequences, we found that these regions of alanine and glycine usage are well conserved in a sampling of the so-called "parental" HLA alleles [45]. Further inspection of a matrix encoding of a broader set of mammalian MHC molecules (Fig. S9) showed that alanine and glycine are well conserved in specific regions across species in the α2-helix of MHC class I molecules, as well as in both helices of MHC class II molecules.

These observations completed our final model for the source of bias in the conserved TCR-MHC docking orientation (Fig. 8E, F). The structures of HLA class I (Fig. 8E) and HLA class II (Fig. 8F) are colored based on their average interaction potentials as shown in Fig. 7. We readily see the well-conserved areas of negligible interaction potential (black) occur near the centers of the α-helices for each MHC. For the TCRs to optimally span the peptide-MHC complex, these areas of negligible interaction define two distinct modes of allowed binding. The canonical mode of binding, swept out by the green lines and filled translucent space, spans these regions of increased interaction potential. The binding mode perpendicular to the canonical binding mode is potentially allowed based solely upon interaction potential but directly passes over the kink in the α2- or β-helices, which is not explicitly accounted for in our bioinformatic calculations, suggesting the structure of the MHC molecules does not allow this binding mode.

DISCUSSION

The concept of evolutionarily conserved interactions guiding immune recognition dates back half a century to work by Niels Jerne [53]. Despite predating the first TCR-MHC structures the prediction has withstood the test of time. Canonical TCR-MHC docking orientations have repeatedly been observed and specific instances of evolutionarily conserved contacts have been identified in a subset of structures [19–21]. However, the identification of TCRs that are capable of binding to non-MHC ligands [9] or in a reversed docking orientation [15,16] call into question the general validity of these findings of evolutionary conservation and highlight the need for systematic analyses covering the entire space of TRAV, TRBV, and MHC alleles. Even without these eccentric exceptions, evidence for evolutionarily conserved interactions have only been convincingly shown for a small subset of TRAV and TRBV genes. Here, we report the results of the first systematic study across these variable alleles.

Using AIMS, a recently developed software [39, 40] for encoding and analyzing amino acid sequences and their biophysical properties in the context of molecular interactions, we found that across all tested TRAV, TRBV, and HLA alleles, there is no evidence of strongly conserved germline-encoded interactions that exist for all possible combinations of TRAV/TRBV genes and HLA polymorphisms. The diversity in the germline-encoded TCR CDR loops alone makes such interactions highly unlikely. These findings suggest that conserved TCR-

MHC interactions spanning species identified thus far [17, 20–22] may be rare across the entirety of the immune repertoire. To characterize these TCR-MHC germline-encoded interactions outside the lens of this evolutionary context, we employed a biophysical scoring function to fully characterize the pairwise interaction potential for every TRAV/HLA and TRBV/HLA pair. These by-allele interaction potentials can further be broken down on a per-residue basis for each MHC α-helix, helping to generate a working model for the source of conserved TCR-pMHC docking orientations.

Across nearly all tested human polymorphisms, we found that class I and class II MHC helices are composed of moderately diverse regions of increased interaction potential interrupted by centrally-located, well-conserved regions of low interaction potential. These regions of low interaction potential define a conserved docking orientation, with room for slight variability in orientation about the flanking regions of higher interaction potential. Our findings are in line with the previously proposed concept of "interaction codons" whereby specific TCR-MHC pairs are predisposed to interact in multiple registers dependent upon the peptide bound to the MHC [19, 37, 54]. However, these codons as currently posited suggest the existence of multiple rigid docking modes, consistent with their ideation deriving from solved crystal structures. The results of our analyses instead suggest a broader, dynamic interpretation of the TCR-pMHC interface [55–59], with the TCR CDR loops sampling local regions of increased interaction potential on the MHC surface, utilizing a more opportunistic approach to binding.

The general TCR-MHC compatibility as suggested by our results does not rule out the possibility of reversed-docking TCRs, suggesting both evolution and selection play a strong role in the conservation of the canonical docking orientation. While it is unclear how prevalent these reversed-docking TCRs are in the pre-selected immune repertoire, it is clear that they are deficient in their ability to signal due to the distance-sensitivity of the coreceptor-associated Lck in interactions with the CD3 signaling complex [15, 16, 60]. In our analysis of the biophysical compatibility between germline-encoded TCR sequences and their potential interaction sites on the MHC molecule we found strong indications that this may be a necessary coevolutionary tradeoff. Over evolutionary history, TCR and MHC molecules have had to balance a conservation of canonical modes of contact with the need to maintain sequence variability in the face of evolutionary pressure. This pressure, exerted by pathogenic challenges, may have forced a drift away from rigid coevolution at the level of amino acid sequences, necessitating this loose compatibility, even at the cost of some subset of receptors that are destined for thymic deletion due to an inability to generate productive signal.

While many TRAV and TRBV genes have equal interaction potentials for their CDR1 and CDR2 loops, a small subset appears to favor one or the other for binding to MHC molecules. An even smaller subset of weakly interacting genes has little to no propensity for binding to the MHC α-helices. The weakest interacting of these genes may in fact be the most interesting, with weakly interacting genes such as TRBV7-2, 7-3, 20-1, and 29-1 associated with celiac disease [23, 24] and multiple sclerosis [25]. Combining this with previous observations that TCR-pMHC interactions dominated by non-germline-encoded CDR3 contacts enable TCR cross-reactivity, and in some cases autoimmunity [25, 28–31], leads to the intriguing possibility that interactions between germline-encoded TCR regions and the MHC provide

a framework for reliable peptide differentiation whereas the lack of such interactions, while granting more flexibility, may lead to higher instances of misguided T cell activation.

METHODS

The AIMS software package used to generate the analysis in this manuscript is available at https://github.com/ctboughter/AIMS, including the original Jupyter Notebooks used to generate the Figures in this manuscript as well as generalized Notebooks and a python-based GUI application for analysis of novel datasets. The AIMS software is constantly evolving, so to ensure exact recapitulation of results presented here AIMS v0.7.1 should be used. Detailed descriptions of the analysis and the instructions for use can be found at https://aims-doc.readthedocs.io.

Repertoire Analysis Using AIMS

As with all AIMS analyses, the first step is to encode each sequence into an AIMS-compatible matrix. In this encoding each amino acid in structurally conserved regions is represented as a number 1-21, with zeros padding gaps between these structurally conserved features. The encoding is straightforward for the TCR sequences, with only the germline encoded regions of the CDR loops included for each gene. For the MHC encoding, only the structurally relevant amino acids are included for optimal alignment of each unique sequence. In this case, the structurally relevant amino acids of the class I (or class II) molecules were divided into three distinct groups; the TCR-exposed residues of the $\alpha 1$ (or α) helix, the TCR-exposed residues of the $\alpha 2$ (or β) helix, and the peptide-contacting residues of the given MHC.

Standard AIMS analysis, including calculation of information-theoretic metrics and position-sensitive biophysical properties, is then derived from these AIMS-encoded matrices. The AIMS interaction scoring is based on a matrix that quantifies a basic pairwise interaction scheme (Table S1), whereby productive amino acid interactions (salt bridges, hydrogen bonds) are scored positively while destructive interactions (hydrophilic-hydrophobic and like-charge clashes) are scored negatively. The first version of this interaction scoring matrix has previously been used to classify interacting and non-interacting molecular partners with a distinguishability of nearly 80% [46].

In calculating the interaction score, we assume that productive contacts are only made by the side-chains of the interacting residues. This simplification does not capture all TCR-pMHC complex contacts, but here we are looking for selectivity enforced by specific TCR-MHC interactions mediated by side-chains. Further, given that any single pair of amino acids on adjacent interfaces of protein binding partners can potentially form strong interactions without being meaningful for the formation of a given complex, we require that any productive interaction include a triad of at least weakly interacting residues. Thus defining our search parameters for broad biophysical compatibility between TCR CDR loops and the MHC helices, we can generate averaged interaction potentials for every TRAV and TRBV gene with

every possible HLA allele.

Sequence Processing

All sequences used in this work are derived from the ImMunoGeneTics (IMGT) [41, 42] database of TRAV and TRBV alleles (https://www.imgt.org/IMGTrepertoire/Proteins) from the following organisms: *Bos taurus* (cow), *Capra hircus* (goat), *Aotus nancymaae* (Nancy Ma's Night Monkey), *Mus musculus* (mouse), *Macaca mulatta* (rhesus macaque), and *Ovis aries* (sheep). MHC sequences are derived from the IMGT-HLA database containing all identified HLA alleles from human and a wide range of organisms (https://www.ebi.ac.uk/ipd/imgt/hla/download) [43]. Only unique, productive sequences are included, which excludes all open reading frames and pseudogenes. Specific structural features are identified from these productive sequences for each molecular species. For TCRs, these structural features are the complementarity determining region (CDR) loops as defined by IMGT. For MHC molecules, these are the TCR-exposed residues of the MHC α-helices as defined in Bjorkman & Parham [61] for class I and as identified in visual molecular dynamics (VMD) [62] for class II. Class I identifications were likewise validated in VMD. For comparison to non-TCR contacting residues, the peptide-contacting residues were also included for class I molecules, again from identification in Bjorkman & Parham [61]. All downstream sequence processing was done from these regions of each TCR and MHC sequence.

Structural Comparison Processing Pipeline

We accessed 149 human TCR-pMHC class I structures and 44 human TCR-pMHC class II structures with PDB IDs drawn from the TCR3D database [38] and each PDB loaded into python using the mdtraj package [63]. After parsing the TCR3D database for degenerate structures and improperly deposited PDBs, we compiled a final dataset of 96 class I structures and 36 class II structures. We extracted CDR-MHC distances for all side-chain contacts, with a specific threshold set to identify "productive" interactions. The interaction cutoff is set to 3.5 Å and Oxygen-Oxygen/Nitrogen-Nitrogen are not counted as productive hydrogen bonding pairs. Further, pairwise residue context for the contact is included in the determination of an interaction. Hydrophobic Carbon-Carbon contacts are productive, but hydrophilic (say Arginine-Lysine) Carbon-Carbon contacts are not counted as a productive interaction. Through a random selection of 10 complex structures analyzed using this pipeline, we were able to validate structural distance measurements using VMD and found that 100% of the identified contacts match those in structures. All code for sequence processing and structural analysis is included in a separate repository, PRESTO (PaRsEr of Solved Tcr cOmplexes). Code for reproducing the analysis can be found at https://github.com/ctboughter/PRESTO.

Information Theoretic Calculations

Information theory, a theory classically applied to communication across noisy channels, is incredibly versatile in its applications and has been applied with success to a range of immunological problems [44, 64–68]. In this work, we utilize two powerful concepts from information theory, namely Shannon entropy and mutual information. Shannon entropy, in its simplest form, can be used as a proxy for the diversity in a given input population. This entropy, denoted as H, has the general form:

$$H(X) = -\sum_{X} p(x) \log_2 p(x) \tag{1}$$

Where $p(x)$ is the occurrence probability of a given event, and X is the set of all events. We can then calculate this entropy at every position along the CDR loops or MHC α-helices, where X is the set of all amino acids, and $p(x)$ is the probability of seeing a specific amino acid at the given position. In other words we want to determine, for a given site in a CDR loop or MHC helix, how much diversity (or entropy) is present. Given there are only 20 amino acids used in naturally derived sequences, we can calculate a theoretical maximum entropy of 4.32 bits, which assumes that every amino acid occurs at a given position with equal probability.

Importantly, from this entropy we can calculate an equally interesting property of the dataset, namely the mutual information. Mutual information is similar, but not identical to, correlation. Whereas correlations are required to be linear, if two amino acids vary in any linked way, this will be reflected as an increase in mutual information. In this work, mutual information $I(X;Y)$ is calculated by subtracting the Shannon entropy described above from the conditional Shannon entropy $H(X|Y)$ at each given position as seen in equations 2 and 3:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{2}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{3}$$

These equations for the entropy and the mutual information are used for all information theoretic calculations in this manuscript, but special consideration of the TCR and MHC sequences must be applied in order to reformulate these sequences into the classic input/output framework necessary for information theory. Keeping with the terminology of information theory, we would define the TCR-pMHC interactions as a "communication channel", and thus think of each TCR sequence as a given input. In this picture, each HLA allele would be seen as a corresponding output. If there exists some systematic relationship between the amino acids in the TCR and HLA sequences, we would see a significant influence of TCR sequence variations on HLA sequences, and this should manifest as an increase in the mutual information.

To calculate this mutual information, we start with the assumption that if the concept of evolutionary conservation of TCR-pMHC interactions is correct, then one should assume that every TCR should interact with every HLA allele. Humans largely possess the same TRAV and TRBV alleles, but each individual possesses a maximum of 12 HLA alleles. We expect that specific alleles that are unable to enforce the supposed evolutionary rules for canonical docking will not be allowed to persist in the population. Continuing from this assumption we then subsample the data and calculate the mutual information on this subsampled dataset. Each TRAV and TRBV allele (the input) is matched with a single HLA allele (the output), and the mutual information is calculated for these pairings. This process is repeated 1000 times and the average mutual information is reported. Further validation of our approach on non-human organisms was carefully formulated to only reclassify TRAV/TRBV alleles and MHC sequences as input/output sequences from the same organismal source. The mutual information calculation was then carried out across organisms, with this within-species architecture conserved.

Acknowledgements

This work was supported by the intramural program of the National Institute of Allergy and Infectious Diseases (NIAID), NIH. We would like to thank David Margulies, Pamela Schwartzberg, Alexander Brown, and Charles Dulberger for their insightful comments on initial drafts of this manuscript.

Competing Interests

The authors declare no competing interests.

Data and materials availability

All data and code used for the analysis in this manuscript are freely available online with no restrictions. All input FASTA sequences and code needed to recreate the analysis can be found via the AIMS GitHub page: github.com/ctboughter/AIMS. Specific analysis for structural comparisons between interaction potentials and TCR-pMHC complexes are found via a separate repository, called PRESTO, also hosted on GitHub: github.com/ctboughter/PRESTO. Due to the significant time required to calculate the interaction scores calculated via AIMS, the calculated scores can be found on Zenodo at the following DOI: 10.5281/zenodo.7023681 . In case of future updates to either AIMS or PRESTO, the specific versions used for this manuscript are also hosted on Zenodo, as AIMS v0.7.1 [DOI: 10.5281/zenodo.7025972] and PRESTO v0.1 [DOI: 10.5281/zenodo.7025652].

References

[1] K. C. Garcia, M. Degano, R. L. Stanfield, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton, and I. A. Wilson. An alphabeta t cell receptor structure at 2.5 a and its orientation in the tcr-mhc complex. *Science*, 274(5285):209–19, 1996.

[2] K. C. Garcia, L. Teyton, and I. A. Wilson. Structural basis of t cell recognition. *Annu Rev Immunol*, 17:369–97, 1999.

[3] L. Yin, J. Scott-Browne, J. W. Kappler, L. Gapin, and P. Marrack. T cells and their eons-old obsession with mhc. *Immunological Reviews*, 250:49–60, 2012.

[4] P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. Structure of the human class i histocompatibility antigen, hla-a2. *Nature*, 329(6139):506–12, 1987.

[5] J. H. Brown, T. S. Jardetzky, J. C. Gorga, L. J. Stern, R. G. Urban, J. L. Strominger, and D. C. Wiley. Three-dimensional structure of the human class ii histocompatibility antigen hla-dr1. *Nature*, 364(6432):33–9, 1993.

[6] L. C. Wu, D. S. Tuot, D. S. Lyons, K. C. Garcia, and M. M. Davis. Two-step binding mechanism for t-cell receptor recognition of peptide-mhc. *Nature*, 418(6897):552–556, 2002.

[7] David Jung and Frederick W. Alt. Unraveling V(D)J Recombination: Insights into Gene Regulation. *Cell*, 116(2):299–311, January 2004.

[8] S. Gras, S. R. Burrows, S. J. Turner, A. K. Sewell, J. McCluskey, and J. Rossjohn. A structural voyage toward an understanding of the mhc-i-restricted immune response: lessons learned and much to be learned. *Immunol Rev*, 250(1):61–81, 2012.

[9] F. Van Laethem, S.D. Sarafova, J.H. Park, X. Tai, L. Pobezinsky, T.I. Guinter, S. Adoro, A. Adams, S.O. Sharrow, L. Feigenbaum, and A. Singer. Deletion of cd4 and cd8 coreceptors permits generation of alphabetat cells that recognize antigens independently of the mhc. *Immunity*, 27(5):735–50, 2007.

[10] A. N. Tikhonova, F. Van Laethem, K. Hanada, J. H. Lu, L. A. Pobezinsky, C. W. Hong, T. I. Guinter, S. K. Jeurling, G. Bernhardt, J. H. Park, J. C. Yang, P. D. Sun, and A. Singer. alpha beta t cell receptors that do not undergo major histocompatibility complex-specific thymic selection possess antibody-like recognition specificities. *Immunity*, 36(1):79–91, 2012.

[11] F. Van Laethem, A. N. Tikhonova, and A. Singer. Mhc restriction is imposed on a diverse t cell receptor repertoire by cd4 and cd8 co-receptors during thymic selection. *Trends Immunol*, 33(9):437–41, 2012.

[12] M. Blackman, J. Kappler, and P. Marrack. The role of the t-cell receptor in positive and negative selection of developing t-cells. *Science*, 248(4961):1335–1341, 1990.

[13] J. Juang, P. J. R. Ebert, D. Feng, K. C. Garcia, M. Krogsgaard, and M. M. Davis. Peptide-mhc heterodimers show that thymic positive selection requires a more restricted set of self-peptides than negative selection. *Journal of Experimental Medicine*, 207(6):1223–1234, 2010.

[14] R. Germain. T-cell development and the cd4–cd8 lineage decision. *Nature Reviews Immunology*, 2002.

[15] S. Gras, J. Chadderton, C. M. Del Campo, C. Farenc, F. Wiede, T. M. Josephs, X. Y. X. Sng, M. Mirams, K. A. Watson, T. Tiganis, K. M. Quinn, J. Rossjohn, and N. L. La Gruta. Reversed t cell receptor docking on a major histocompatibility class i complex limits involvement in the immune response. *Immunity*, 45(4):749–760, 2016.

[16] P. Zareie, C. Szeto, C. Farenc, S. D. Gunasinghe, E. M. Kolawole, A. Nguyen, C. Blyth, X. Y. X. Sng, J. Li, C. M. Jones, A. J. Fulcher, J. R. Jacobs, Q. Wei, L. Wojciech, J. Petersen, N. R. J. Gascoigne, B. D. Evavold, K. Gaus, S. Gras, J. Rossjohn, and N. L. La Gruta. Canonical t cell receptor docking on peptide-mhc is essential for t cell signaling. *Science*, 372(6546), 2021.

[17] J. P. Scott-Browne, J. White, J. W. Kappler, L. Gapin, and P. Marrack. Germline-encoded amino acids in the alphabeta t-cell receptor control thymic selection. *Nature*, 458(7241):1043–6, 2009.

[18] S. H. Krovi, J. W. Kappler, P. Marrack, and L. Gapin. Inherent reactivity of unselected tcr repertoires to peptide-mhc molecules. *Proc Natl Acad Sci U S A*, 116(44):22252–22261, 2019.

[19] D. Feng, C. J. Bond, L. K. Ely, J. Maynard, and K. C. Garcia. Structural evidence for a germline-encoded t cell receptor-major histocompatibility complex interaction 'codon'. *Nat Immunol*, 8(9):975–83, 2007.

[20] S. J. Blevins, B. G. Pierce, N. K. Singh, T. P. Riley, Y. Wang, T. T. Spear, M. I. Nishimura, Z. P. Weng, and B. M. Baker. How structural adaptability exists alongside hla-a2 bias in the human alpha beta tcr repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, 113(9):E1276–E1285, 2016.

[21] J. P. Scott-Browne, F. Crawford, M. H. Young, J. W. Kappler, P. Marrack, and L. Gapin. Evolutionarily conserved features contribute to alphabeta t cell receptor specificity. *Immunity*, 35(4):526–35, 2011.

[22] S. Dai, E. S. Huseby, K. Rubtsova, J. Scott-Browne, F. Crawford, W. A. Macdonald, P. Marrack, and J. W. Kappler. Crossreactive t cells spotlight the germline rules for alpha beta t cell-receptor interactions with mhc molecules. *Immunity*, 28(3):324–334, 2008.

[23] Kristin Støen Gunnarsen, Lene Støkken Høydahl, Louise Fremgaard Risnes, Shiva Dahal-Koirala, Ralf Stefan Neumann, Elin Bergseng, Terje Frigstad, Rahel Frick, M. Fleur du Pré, Bjørn Dalhus, Knut E. A. Lundin, Shuo-Wang Qiao, Ludvig M.

Sollid, Inger Sandlie, and Geir Åge Løset. A TCRα framework–centered codon shapes a biased T cell repertoire through direct MHC and CDR3β interactions. *JCI Insight*, 2(17), September 2017.

[24] Shuo-Wang Qiao, Asbjørn Christophersen, Knut E. A. Lundin, and Ludvig M. Sollid. Biased usage and preferred pairing of α- and β-chains of TCRs specific for an immunodominant gluten epitope in coeliac disease. *International Immunology*, 26(1):13–19, January 2014.

[25] Dhruv K. Sethi, Susana Gordo, David A. Schubert, and Kai W. Wucherpfennig. Cross-reactivity of a human autoimmune TCR is dominated by a single TCR loop. *Nature Communications*, 4(1):2623, October 2013.

[26] Xiao Liu, Wei Zhang, Ming Zhao, Longfei Fu, Limin Liu, Jinghua Wu, Shuangyan Luo, Longlong Wang, Zijun Wang, Liya Lin, Yan Liu, Shiyu Wang, Yang Yang, Lihua Luo, Juqing Jiang, Xie Wang, Yixin Tan, Tao Li, Bochen Zhu, Yi Zhao, Xiaofei Gao, Ziyun Wan, Cancan Huang, Mingyan Fang, Qianwen Li, Huanhuan Peng, Xiangping Liao, Jinwei Chen, Fen Li, Guanghui Ling, Hongjun Zhao, Hui Luo, Zhongyuan Xiang, Jieyue Liao, Yu Liu, Heng Yin, Hai Long, Haijing Wu, Huanming Yang, Jian Wang, and Qianjin Lu. T cell receptor repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 78(8):1070–1078, August 2019.

[27] Yin Tong, Zhoufang Li, Hua Zhang, Ligang Xia, Meng Zhang, Ying Xu, Zhanhui Wang, Michael W. Deem, Xiaojuan Sun, and Jiankui He. T Cell Repertoire Diversity Is Decreased in Type 1 Diabetes Patients. *Genomics, Proteomics & Bioinformatics*, 14(6):338–348, December 2016.

[28] Laura Ciacchi, Carine Farenc, Shiva Dahal-Koirala, Jan Petersen, Ludvig M. Sollid, Hugh H. Reid, and Jamie Rossjohn. Structural basis of T cell receptor specificity and cross-reactivity of two HLA-DQ2.5-restricted gluten epitopes in celiac disease. *Journal of Biological Chemistry*, 298(3):101619, March 2022.

[29] Jan Petersen, Veronica Montserrat, Jorge R. Mujico, Khai Lee Loh, Dennis X. Beringer, Menno van Lummel, Allan Thompson, M. Luisa Mearin, Joachim Schweizer, Yvonne Kooy-Winkelaar, Jeroen van Bergen, Jan W. Drijfhout, Wan-Ting Kan, Nicole L. La Gruta, Robert P. Anderson, Hugh H. Reid, Frits Koning, and Jamie Rossjohn. T-cell receptor recognition of HLA-DQ2–gliadin complexes associated with celiac disease. *Nature Structural & Molecular Biology*, 21(5):480–488, May 2014.

[30] Jan Petersen, Laura Ciacchi, Mai T. Tran, Khai Lee Loh, Yvonne Kooy-Winkelaar, Nathan P. Croft, Melinda Y. Hardy, Zhenjun Chen, James McCluskey, Robert P. Anderson, Anthony W. Purcell, Jason A. Tye-Din, Frits Koning, Hugh H. Reid, and Jamie Rossjohn. T cell receptor cross-reactivity between gliadin and bacterial peptides in celiac disease. *Nature Structural & Molecular Biology*, 27(1):49–61, January 2020.

[31] M. Hahn, M. J. Nicholson, J. Pyrdol, and K. W. Wucherpfennig. Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune t cell receptor. *Nat Immunol*, 6(5):490–6, 2005.

[32] M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Ozkan, M. M. Davis, K. W. Wucherpfennig, and K. C. Garcia. Deconstructing the peptide-mhc specificity of t cell recognition. *Cell*, 157(5):1073–1087, 2014.

[33] E. Sharon, L. V. Sibener, A. Battle, H. B. Fraser, K. C. Garcia, and J. K. Pritchard. Genetic variation in mhc proteins is associated with t cell receptor expression biases. *Nature Genetics*, 48(9):995–+, 2016.

[34] Leah V. Sibener, Ricardo A. Fernandes, Elizabeth M. Kolawole, Catherine B. Carbone, Fan Liu, Darren McAffee, Michael E. Birnbaum, Xinbo Yang, Laura F. Su, Wong Yu, Shen Dong, Marvin H. Gee, Kevin M. Jude, Mark M. Davis, Jay T. Groves, William A. Goddard, James R. Heath, Brian D. Evavold, Ronald D. Vale, and K. Christopher Garcia. Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. *Cell*, 2018.

[35] S. R. Burrows, Z. J. Chen, J. K. Archbold, F. E. Tynan, T. Beddoe, L. Kjer-Nielsen, J. J. Miles, R. Khanna, D. J. Moss, Y. C. Liu, S. Gras, L. Kostenko, R. M. Brennan, C. S. Clements, A. G. Brooks, A. W. Purcell, J. McCluskey, and J. Rossjohn. Hard wiring of t cell receptor specificity for the major histocompatibility complex is underpinned by tcr adaptability. *Proceedings of the National Academy of Sciences of the United States of America*, 107(23):10608–10613, 2010.

[36] K. Ishigaki, K. Lagattuta, Y. Luo, E. James, J. Buckner, and S. Raychaudhuri. Hla autoimmune risk alleles restrict the hypervariable region of t cell receptors. *Nature Genetics*, 54(4):393–+, 2022.

[37] K. C. Garcia, J. J. Adams, D. Feng, and L. K. Ely. The molecular basis of tcr germline bias for mhc is surprisingly simple. *Nat Immunol*, 10(2):143–7, 2009.

[38] R. Gowthaman and B. G. Pierce. Tcr3d: The t cell receptor structural repertoire database. *Bioinformatics*, 35(24):5323–5325, 2019.

[39] Christopher T. Boughter, Marta T. Borowska, Jenna J. Guthmiller, Albert Bendelac, Patrick C. Wilson, Benoit Roux, and Erin J. Adams. Biochemical patterns of antibody polyreactivity revealed through a bioinformatics-based analysis of CDR loops. *eLife*, 2020.

[40] Christoper T. Boughter and Martin Meier-Schellersheim. An integrated approach to the characterization of immune repertoires using aims: An automated immune molecule separator. *biorxiv*, 2022.

[41] Xavier Brochet, Marie Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic acids research*, 2008.

[42] Marie Paule Lefranc. Immunoglobulins: 25 years of immunoinformatics and imgt-ontology. *Biomolecules*, 2014.

[43] S. J. Mack, P. Cano, J. A. Hollenbach, J. He, C. K. Hurley, D. Middleton, M. E. Moraes, S. E. Pereira, J. H. Kempenich, E. F. Reed, M. Setterholm, A. G. Smith, M. G. Tilanus, M. Torres, M. D. Varney, C. E. Voorter, G. F. Fischer, K. Fleischhauer, D. Goodridge, W. Klitz, A. M. Little, M. Maiers, S. G. Marsh, C. R. Muller, H. Noreen, E. H. Rozemuller, A. Sanchez-Mazas, D. Senitzer, E. Trachtenberg, and M. Fernandez-Vina. Common and well-documented hla alleles: 2012 update to the cwd catalogue. *Tissue Antigens*, 81(4):194–203, 2013.

[44] Claude E. Shannon. The Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948.

[45] J. Robinson, L. A. Guethlein, N. Cereb, S. Y. Yang, P. J. Norman, S. G. E. Marsh, and P. Parham. Distinguishing functional polymorphism from random variation in the sequences of > 10,000 hla-a, -b and -c alleles. *Plos Genetics*, 13(6), 2017.

[46] P. Nandigrami, F. Szczepaniak, C.T. Boughter, F. Dehez, C. Chipot, and B. Roux. Computational assessment of protein-protein binding specificity within a family of synaptic surface receptors. *Journal of Physical Chemistry B*, 2022.

[47] Ariane Thielens, Eric Vivier, and François Romagné. NK cell MHC class I specific receptors (KIR): from biology to clinical intervention. *Current Opinion in Immunology*, 24(2):239–245, April 2012.

[48] James Robinson, Kavita Mistry, Hamish McWilliam, Rodrigo Lopez, and Steven G. E. Marsh. IPD—the Immuno Polymorphism Database. *Nucleic Acids Research*, 38(Database issue):D863–D869, January 2010.

[49] Shoeib Moradi, Richard Berry, Phillip Pymm, Corinne Hitchen, Simone A. Beckham, Matthew C. J. Wilce, Nicholas G. Walpole, Craig S. Clements, Hugh H. Reid, Matthew A. Perugini, Andrew G. Brooks, Jamie Rossjohn, and Julian P. Vivian. The Structure of the Atypical Killer Cell Immunoglobulin-like Receptor, KIR2DL4. *The Journal of Biological Chemistry*, 290(16):10460–10471, April 2015.

[50] Shoeib Moradi, Sanda Stankovic, Geraldine M. O'Connor, Phillip Pymm, Bruce J. MacLachlan, Camilla Faoro, Christelle Retière, Lucy C. Sullivan, Philippa M. Saunders, Jacqueline Widjaja, Shea Cox-Livingstone, Jamie Rossjohn, Andrew G. Brooks, and Julian P. Vivian. Structural plasticity of KIR2DL2 and KIR2DL3 enables altered docking geometries atop HLA-C. *Nature Communications*, 12(1):2173, April 2021.

[51] F.E. Tynan, S.R. Burrows, A.M. Buckle, C.S. Clements, N.A. Borg, J.J. Miles, T. Beddoe, J.C. Whisstock, M.C. Wilce, S.L. Silins, J.M. Burrows, L. Kjer-Nielson, L. Kostenko, A.W. Purcell, J. McCluskey, and J. Rossjohn. T cell receptor recognition of a 'super-bulged' major histocompatibility complex class i–bound peptide. *Nature Immunology*, 2005.

[52] P. Marrack and J. Kappler. The t cell receptor. *Science*, 238(4830):1073–9, 1987.

[53] N.K. Jerne. The somatic generation of immune recognition. *European Journal of Immunology*, 1971.

[54] J. J. Adams, S. Narayanan, M. E. Birnbaum, S. S. Sidhu, S. J. Blevins, M. H. Gee, L. V. Sibener, B. M. Baker, D. M. Kranz, and K. C. Garcia. Structural interplay between germline interactions and adaptive recognition determines the bandwidth of tcr-peptide-mhc cross-reactivity. *Nature Immunology*, 17(1):87–+, 2016.

[55] B. M. Baker, D. R. Scott, S. J. Blevins, and W. F. Hawse. Structural and dynamic control of t-cell receptor specificity, cross-reactivity, and binding mechanism. *Immunological Reviews*, 250:10–31, 2012.

[56] Jason R. Devlin, Jesus A. Alonso, Cory M. Ayres, Grant L. J. Keller, Sara Bobisse, Craig W. Vander Kooi, George Coukos, David Gfeller, Alexandre Harari, and Brian M. Baker. Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nature Chemical Biology*, 16(11):1269–1276, November 2020.

[57] Angela R. Smith, Jesus A. Alonso, Cory M. Ayres, Nishant K. Singh, Lance M. Hellman, and Brian M. Baker. Structurally silent peptide anchor modifications allosterically modulate T cell recognition in a receptor-dependent manner. *Proceedings of the National Academy of Sciences*, 118(4), January 2021.

[58] Oleg Y. Borbulevych, Kurt H. Piepenbrink, Brian E. Gloor, Daniel R. Scott, Ruth F. Sommese, David K. Cole, Andrew K. Sewell, and Brian M. Baker. T Cell Receptor Cross-reactivity Directed by Antigen-Dependent Tuning of Peptide-MHC Molecular Flexibility. *Immunity*, 31(6):885–896, December 2009.

[59] Daniel R. Scott, Oleg Y. Borbulevych, Kurt H. Piepenbrink, Steven A. Corcelli, and Brian M. Baker. Disparate Degrees of Hypervariable Loop Flexibility Control T-Cell Receptor Cross-Reactivity, Specificity, and Binding Mechanism. *Journal of Molecular Biology*, 414(3):385–400, December 2011.

[60] J. J. Adams, S. Narayanan, B. Y. Liu, M. E. Birnbaum, A. C. Kruse, N. A. Bowerman, W. Chen, A. M. Levin, J. M. Connolly, C. Zhu, D. M. Kranz, and K. C. Garcia. T cell receptor signaling is limited by docking geometry to peptide-major histocompatibility complex. *Immunity*, 35(5):681–693, 2011.

[61] P. J. Bjorkman and P. Parham. Structure, function, and diversity of class-i major histocompatibility complex-molecules. *Annual Review of Biochemistry*, 59:253–288, 1990.

[62] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 1996.

[63] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 2015.

[64] Ramón Román-Roldán, Pedro Bernaola-Galván, and José L. Oliver. Application of information theory to DNA sequence analysis: A review. *Pattern Recognition*, 1996.

[65] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information transduction capacity of noisy biochemical signaling networks. *Science*, 2011.

[66] Susana Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 2014.

[67] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 2010.

[68] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

Supplemental Figures

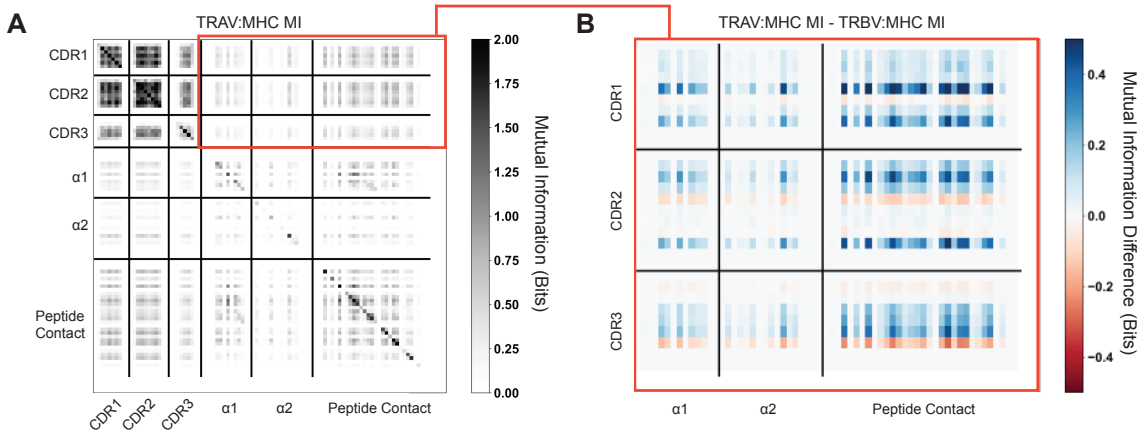

Figure S1: Quantification of mutual information (MI) between germline encoded TCR regions and MHC α-helices show no strong signal. Peptide-contacting residues of the MHC platform domain are included as a reference. (A) Mutual information calculated between TRAV germline encoded regions and specific structural features of class I HLA molecules. The boxed red region highlights the TCR-MHC mutual information, which is then further characterized in (B). Within-molecule mutual information (i.e. TCR-TCR and MHC-MHC) shows clear signals of co-variation over each dataset, contrasting strongly with the TCR-MHC mutual information. (B) Difference calculated between TRAV-MHC mutual information (MI) and TRBV-MHC mutual information (MI) to search for patterns in co-varying residues that differ between the two datasets. Positive values signify increased TRAV-MHC MI in that region, whereas negative values signify increased TRBV-MHC MI in that region. We would expect that TRBV-MHC MI should be higher (blue) in the α1-helix region, whereas TRAV-MHC MI should be higher (red) in the α2-helix region. Neither trend is seen strongly here.

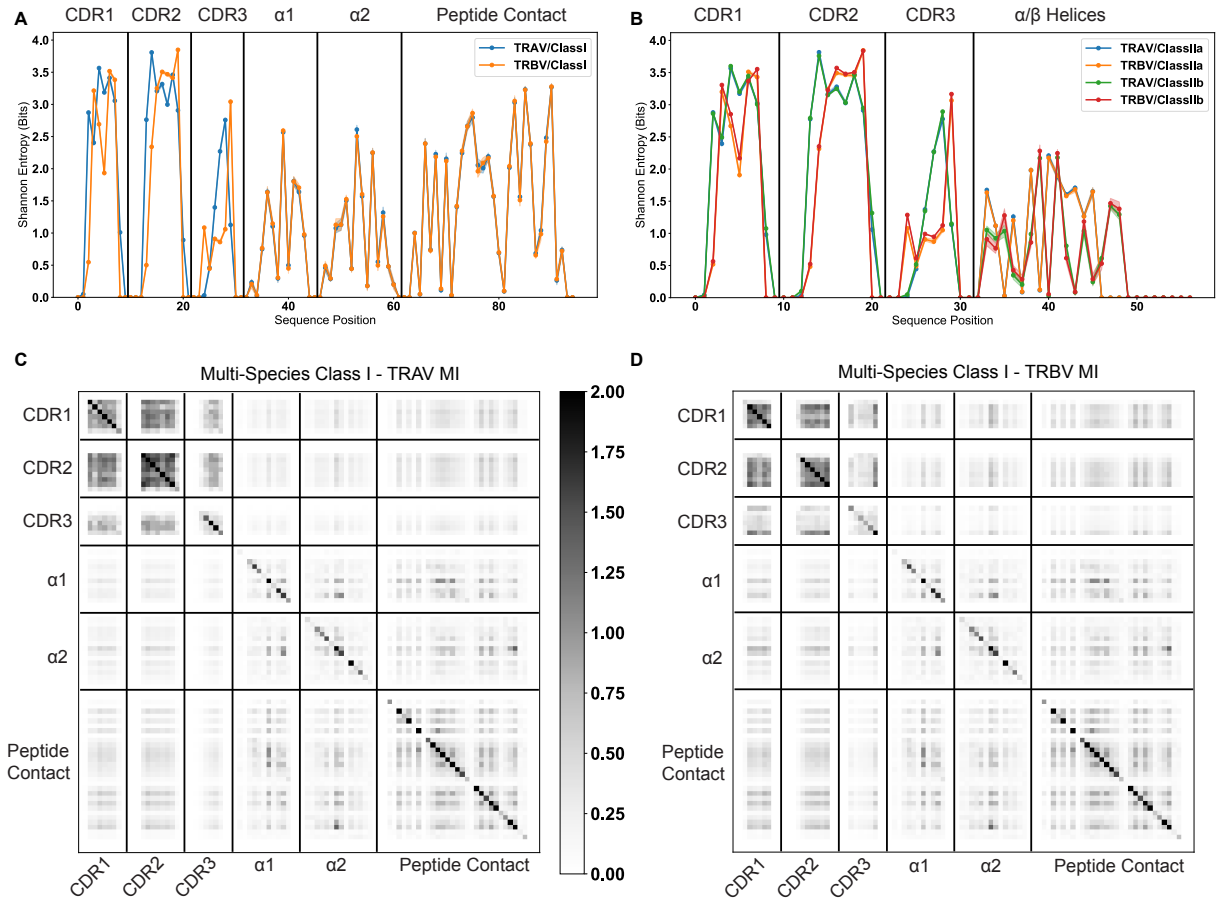

Figure S2: Information theoretic analysis of TCR-MHC pairs from a range of mammalian species shows that strongly co-varying residues between TCR and MHC germline sequences are rare. TCR-MHC pairs are matched for each bootstrapped sample, as discussed in the above methods. (A) Average bootstrapped Shannon entropy of CDR1, CDR2, and CDR3 for TRAV and TRBV amino acid sequences coupled with class I sequences. (B) Average bootstrapped Shannon entropy of CDR1, CDR2, and CDR3 for TRAV and TRBV amino acid sequences coupled with class II sequences. Standard deviation given by width of lines. Site-wise mutual information is calculated from the entropies in panel (A), with class I - TRAV MI calculated in (C) and class I - TRBV MI in (D).

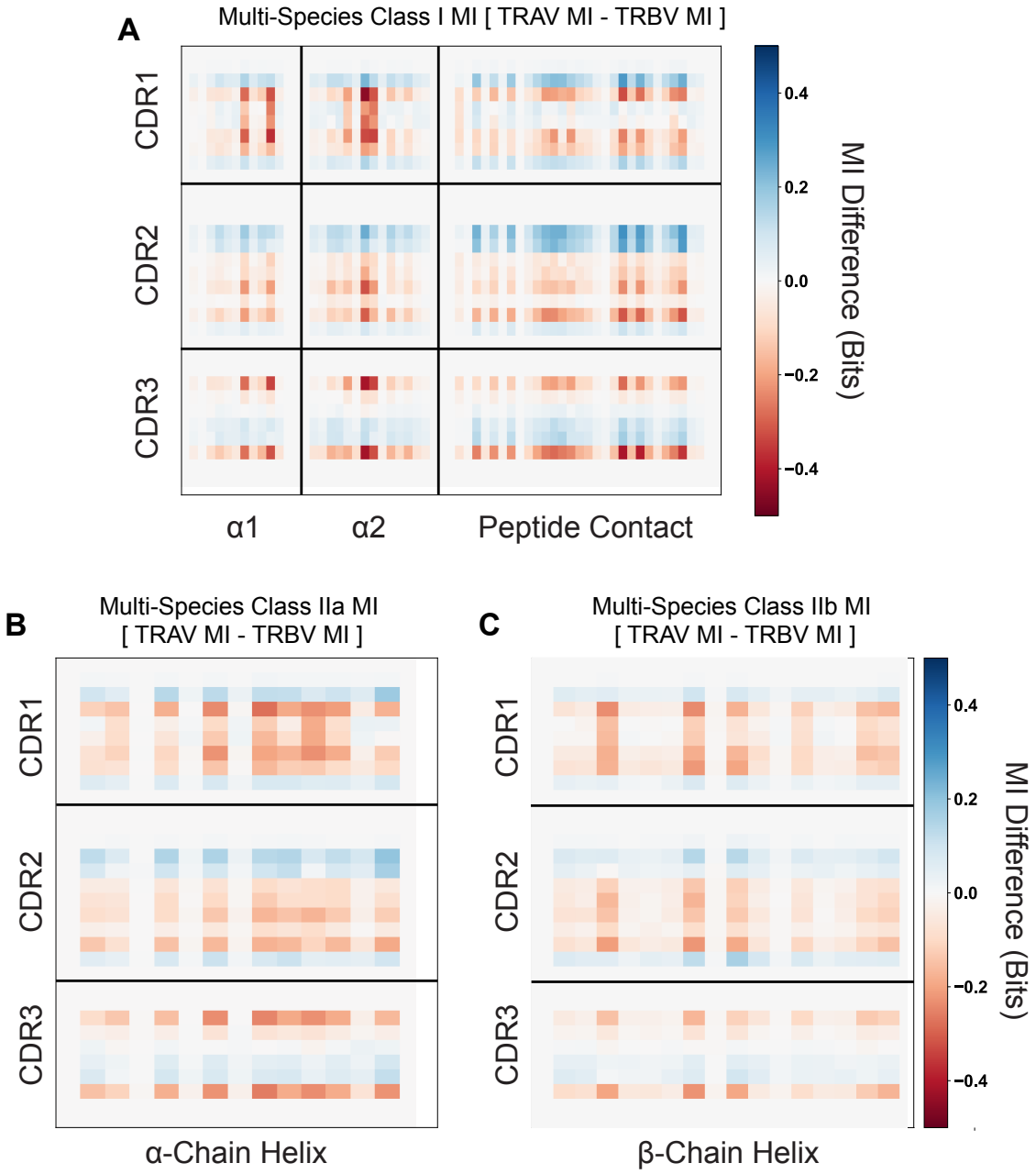

Figure S3: Mutual information differences between calculated TRAV and TRBV quantities from Supplemental Figure S2C and S2D (A), or from class IIa (B), or class IIb (C) quantities. Positions shaded blue represent a stronger MHC-TRAV signal, whereas red positions represent stronger MHC-TRBV signals.

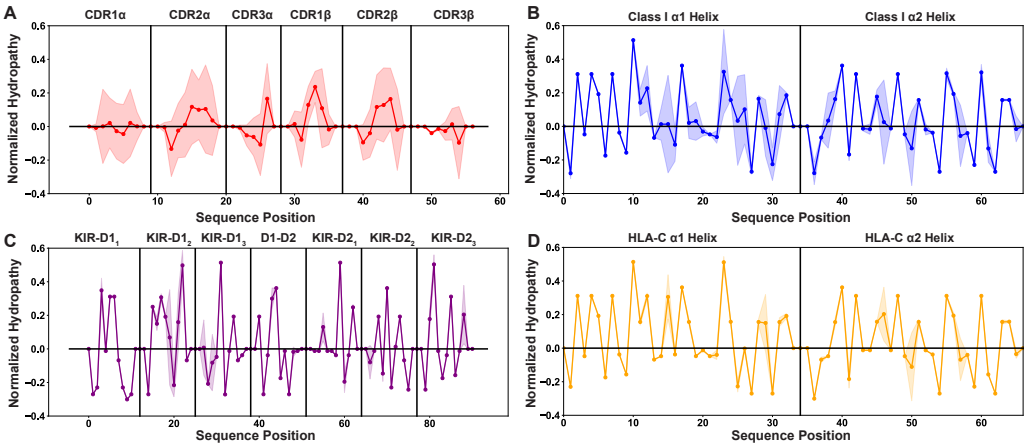

Figure S4: Position-sensitive hydropathy shows similar trends to the position-sensitive charge, showing higher variability in the germline-encoded TCR CDR loops than all other tested datasets. Position sensitive amino acid hydropathy averaged over all germline-encoded TCR CDR loops (A), MHC class I α-helices (B), KIR MHC-contacting regions (C), or HLA-C α-helices (D). Solid lines represent averages over unique amino acid sequences, while the standard deviations about these averages are given by the shaded regions. A positive hydropathy score corresponds to a hydrophilic residue, while a negative score corresponds to a hydrophobic residue. Scores are normalized to a mean of 0 and a standard deviation of 1.

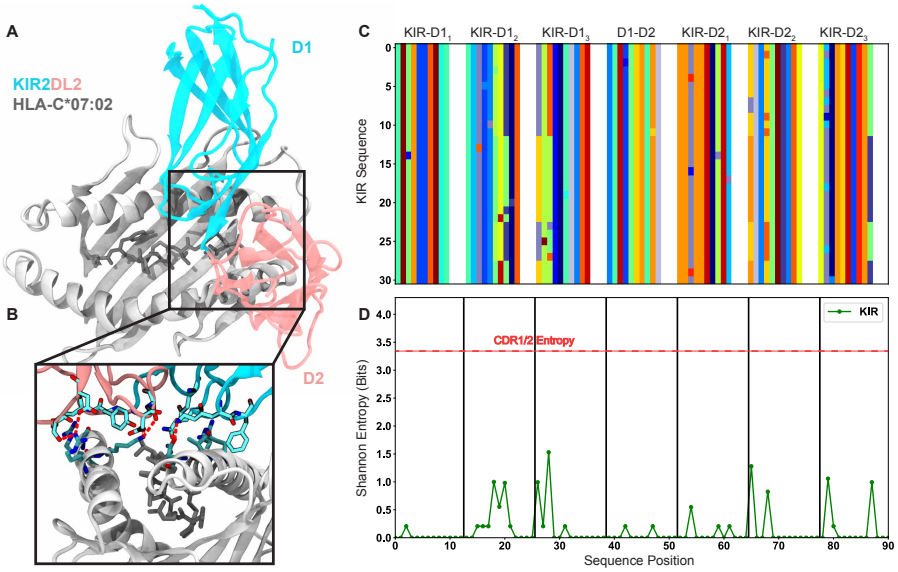

Figure S5: The killer cell immunoglobulin-like receptors (KIRs) and their recognition of HLA-C represent a suitable comparison to the germline-encoded CDR loops and their recognition of MHC. (A) Structure of KIR2DL2 in complex with HLA-C*07:02 [PDB ID: 6PA1] highlights the conserved binding mode of KIR. (B) Inset highlighting the salt bridge and hydrogen bonding network in the KIR-MHC interface shown in (A). Red lines represent electrostatic interactions. (C) Matrix encoding and alignment of the MHC-contacting regions of the KIR. Each color represents a unique amino acid. (D) Position-sensitive Shannon entropy from the matrix in (C) highlighting the exceptionally low diversity of the KIRs in the MHC-contacting regions.

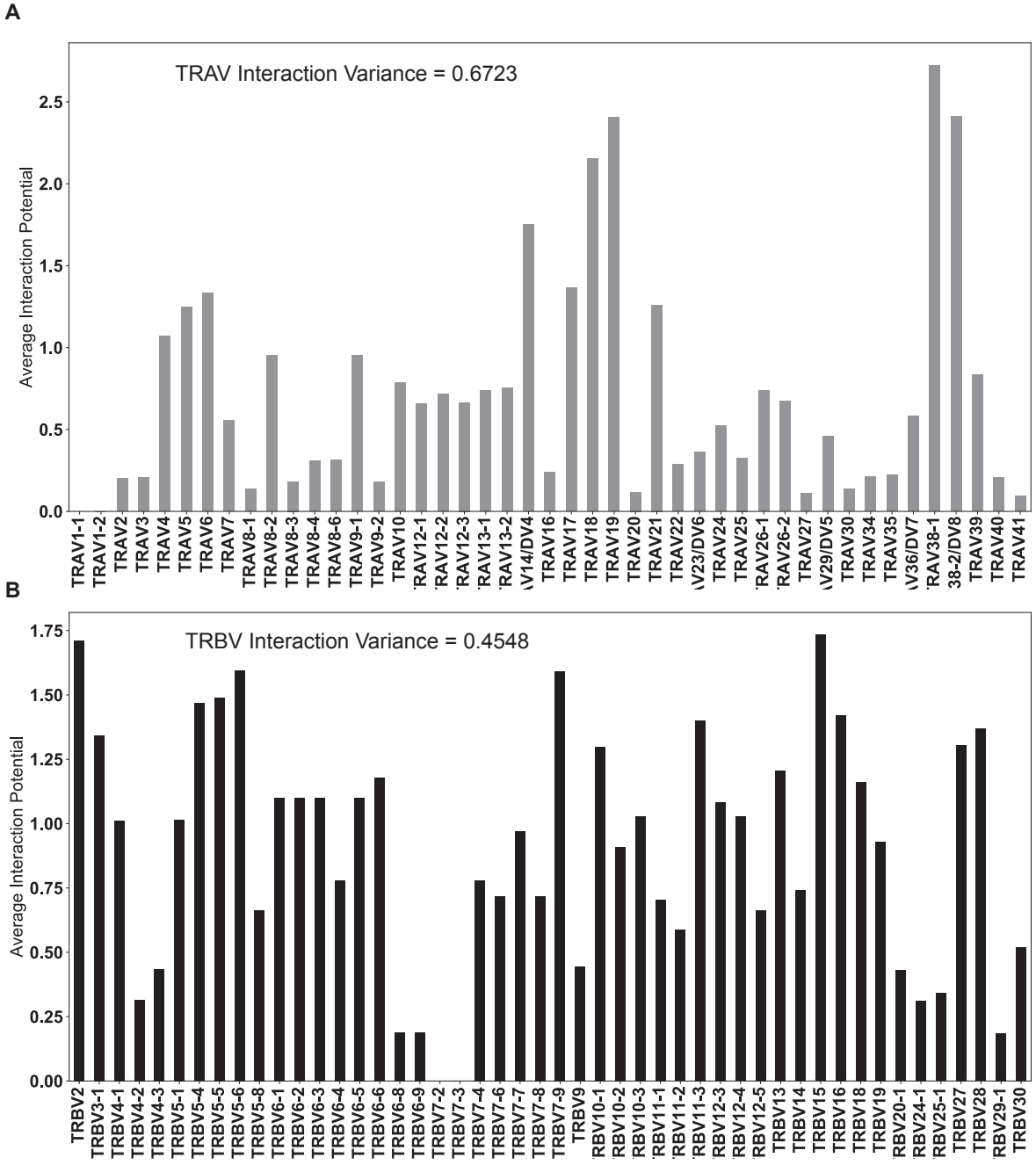

Figure S6: Per-gene interaction potentials averaged over all class I HLA molecules for TRAV genes (A) and TRBV genes (B). The variance across all of these interaction potentials, in other words over all genes, is reported within each plot. Reported values give interaction potential averaged over both HLA helices, not just the helix each TCR chain "canonically" interacts with, and over both CDR1 and CDR2.

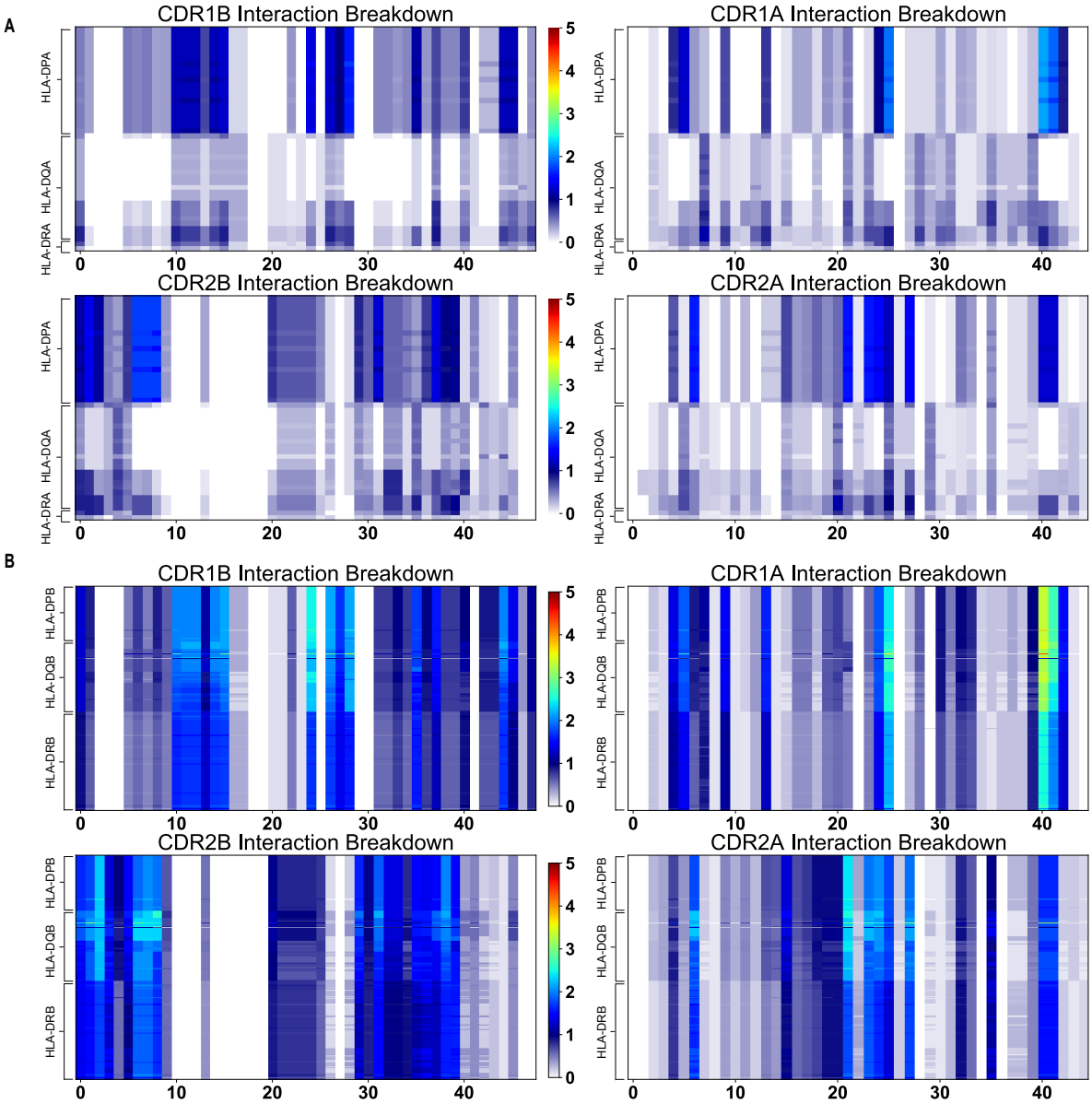

Figure S7: AIMS interaction potentials for CDR1 and CDR2 β and α chains with HLA class II molecules. These calculations are made separately for class IIa (A) and class IIb (B) alleles. The interaction potential scale is given in the center of each plot. The key for which TRAV or TRBV gene corresponds to which number on the x-axis of each plot can be found in Supplemental Table S2.

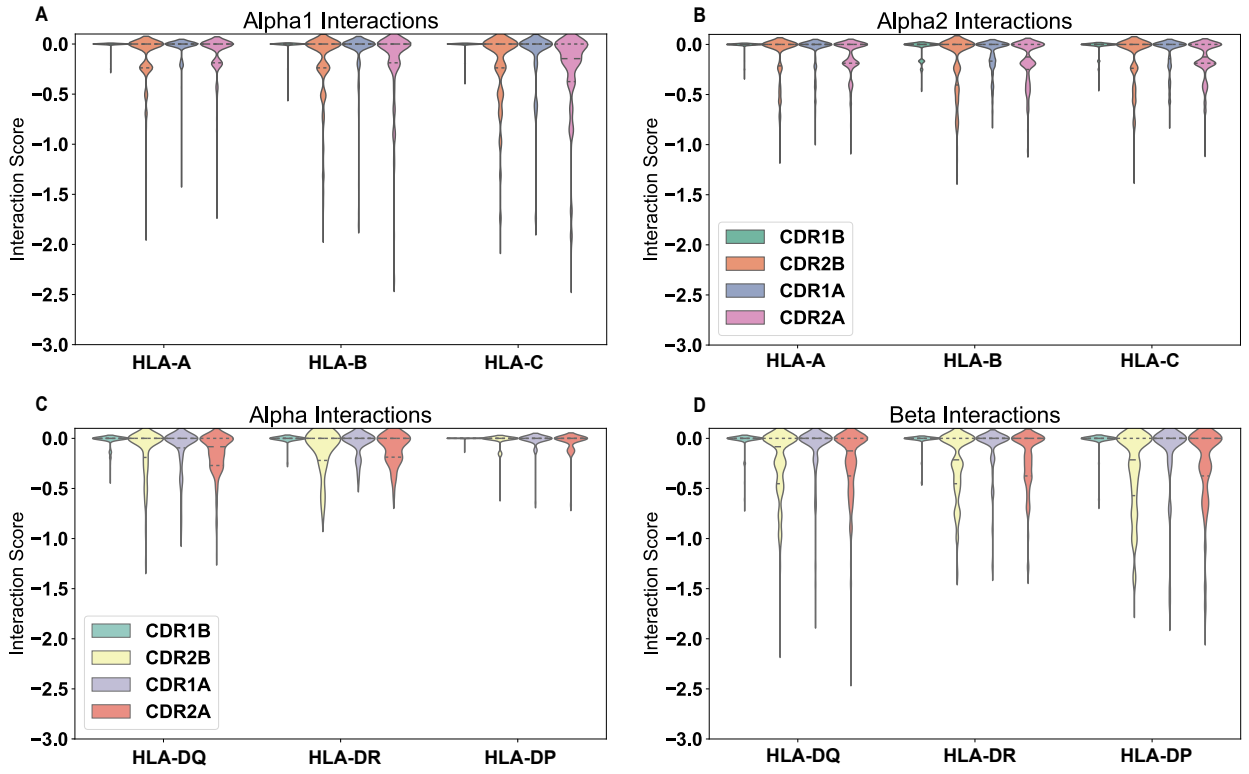

Figure S8: AIMS clash potentials calculated for all possible CDR-MHC helix interactions. Clashes are estimated from single sites in the CDR and MHC that should form unfavorable interactions (as shown in table S1). Calculations are shown for HLA class I interactions with the $\alpha 1$ helix (A), the $\alpha 2$ helix (B), and for HLA class II interactions with the α-helix (C) and the β helix (D). Reported values are averages for each individual TRAV or TRBV gene and their proposed possible clashes with each individual MHC polymorphism. Clashes are, on the whole, rare between germline-encoded CDR loops and MHC α-helices.

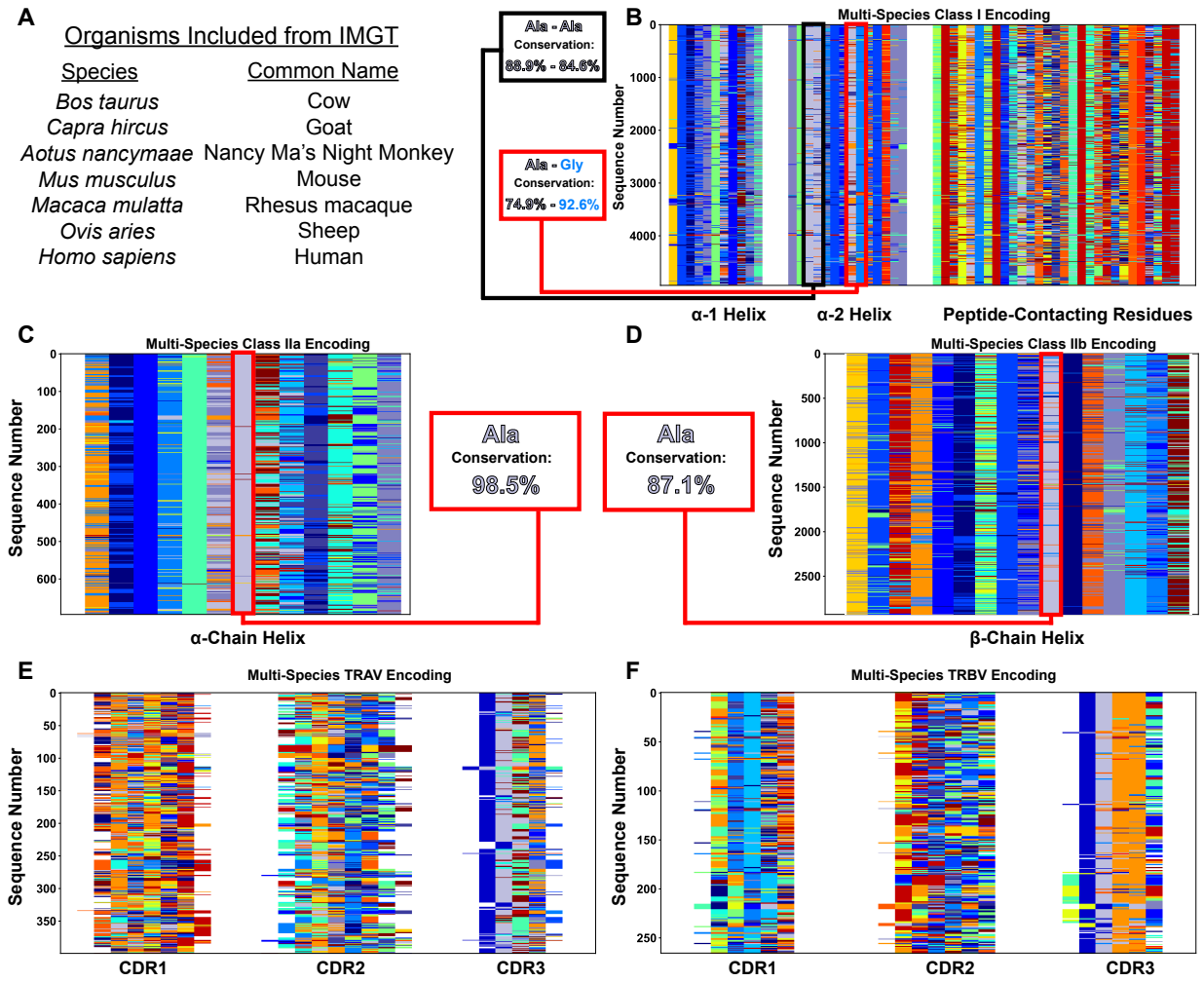

Figure S9: Across a range of mammalian species, the regions of low interaction potential on class I and class II MHC α-helices are very well conserved. Species used in this analysis highlighted in (A) as found in the IPD-MHC Database. Color-coded matrix alignments of MHC class I (B), MHC class IIa (C), MHC class IIb (D), TRAV (E), and TRBV (F) highlight the extent of this conservation. Specifically for the MHC alignments, the regions of conserved low interaction potential have the extent of conservation quantified, along with the amino acid identity at these sites. Colors of the named amino acids match the colors in the matrices. By comparison, we can see by eye that outside of the TRBV CASS motif and the TRAV AV motif in the germline-encoded region of CDR3, there is almost no clear conservation in the germline-encoded CDR loops.

Table S1: Table used for the second version of the AIMS scoring of pairwise amino acid interactions. The table attempts to recapitulate the interactions between amino acids at the level of an introductory biochemistry course.

Res/Res	A	G	L	M	F	W	K	Q	E	S	P	V	I	C	Y	H	R	N	D	T
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	1	1	1	-2	-1	-2	-0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
M	0	0	1	1	1	1	-2	-1	-2	-0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
F	0	0	1	1	1	1	-2	-1	-2	-0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
W	0	0	1	1	1	1	-2	-1	-2	-0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
K	0	0	-2	-2	-2	-2	-2	1	2	0.5	-1	-2	-2	0.5	1	1	-2	1	2	0.5
Q	0	0	-1	-1	-1	-1	1	1	1	0.5	-1	-1	-1	0.5	1	1	1	1	1	0.5
E	0	0	-2	-2	-2	-2	2	1	-2	0.5	-1	-2	-2	0.5	1	1	2	1	-2	0.5
S	0	0	-0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
P	0	0	1	1	1	1	-1	-1	-1	-0.5	0	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
V	0	0	1	1	1	1	-2	-1	-2	-0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
I	0	0	1	1	1	1	-2	-1	-2	-.0.5	1	1	1	-0.5	-0.5	-0.5	-2	-1	-2	-0.5
C	0	0	-0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	2	0.5	0.5	0.5	0.5	0.5	0.5
Y	0	0	-0.5	-0.5	-0.5	-0.5	1	1	1	0.5	-0.5	-0.5	-0.5	0.5	1	1	1	1	1	0.5
H	0	0	-0.5	-0.5	-0.5	-0.5	1	1	1	0.5	-0.5	-0.5	-0.5	0.5	1	1	1	1	1	0.5
R	0	0	-2	-2	-2	-2	-2	1	2	0.5	-2	-2	-2	0.5	1	1	-2	1	2	0.5
N	0	0	-1	-1	-1	-1	1	1	1	0.5	-1	-1	-1	0.5	1	1	1	1	1	0.5
D	0	0	-2	-2	-2	-2	2	1	-2	0.5	-2	-2	-2	0.5	1	1	2	1	-2	0.5
T	0	0	-0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table S2: Key for Figures 4 and S7 relating the numbers on the X-axis of each plot to the corresponding TRAV or TRBV gene. Note, the pairing of TRAV and TRBV genes to a specific X-axis number has no meaningful relation. Genes are listed in the same order as found on IMGT, with pseudogenes not included.

X-Axis #	TRAV Gene	TRBV Gene	X-Axis #	TRAV Gene	TRBV Gene
0	TRAV1-1	TRBV2	24	TRAV18	TRBV7-9
1	TRAV1-2	TRBV3-1	25	TRAV19	TRBV9
2	TRAV2	TRBV4-1	26	TRAV20	TRBV10-1
3	TRAV3	TRBV4-2	27	TRAV21	TRBV10-2
4	TRAV4	TRBV4-3	28	TRAV22	TRBV10-3
5	TRAV5	TRBV5-1	29	TRAV23/DV6	TRBV11-1
6	TRAV6	TRBV5-4	30	TRAV24	TRBV11-2
7	TRAV7	TRBV5-5	31	TRAV25	TRBV11-3
8	TRAV8-1	TRBV5-6	32	TRAV26-1	TRBV12-3
9	TRAV8-2	TRBV5-8	33	TRAV26-2	TRBV12-4
10	TRAV8-3	TRBV6-1	34	TRAV27	TRBV12-5
11	TRAV8-4	TRBV6-2	35	TRAV29/DV5	TRBV13
12	TRAV8-6	TRBV6-3	36	TRAV30	TRBV14
13	TRAV9-1	TRBV6-4	37	TRAV34	TRBV15
14	TRAV9-2	TRBV6-5	38	TRAV35	TRBV16
15	TRAV10	TRBV6-6	39	TRAV36/DV7	TRBV18
16	TRAV12-1	TRBV6-8	40	TRAV38-1	TRBV19
17	TRAV12-2	TRBV6-9	41	TRAV38-2/DV8	TRBV20-1
18	TRAV12-3	TRBV7-2	42	TRAV39	TRBV24-1
19	TRAV13-1	TRBV7-3	43	TRAV40	TRBV25-1
20	TRAV13-2	TRBV7-4	44	TRAV41	TRBV27
21	TRAV14/DV4	TRBV7-6	45		TRBV28
22	TRAV16	TRBV7-7	46		TRBV29-1
23	TRAV17	TRBV7-8	47		TRBV30