*Journal Title Here*, YEAR, pp. 1–17
DOI added during production
Published: Date added during production
**Paper**

# Systematic analysis of CDR contacts and pairing constraints between T cell receptor $\alpha\beta$ chains.

Martina Milighetti,[1,2] Yuta Nagano,[1,3] James Henderson,[1,4] Uri Hershberg,[5] Andreas Tiffeau-Mayer,[1,4] Anne-Florence Bitbol[6,7] and Benny Chain[1,8,*]

[1]Division of Infection and Immunity, University College London, Gower St, WC1E 6BT, London, United Kingdom
[2]Cancer Institute, University College London, Gower St, WC1E 6BT, London, United Kingdom
[3]Division of Medicine, University College London, Gower St, WC1E 6BT, London, United Kingdom
[4]Institute for the Physics of Living Systems, University College London, Gower St, WC1E 6BT, London, United Kingdom
[5]Department of Human Biology, University of Haifa, 199 Aba Khoushy Ave., Mount Carmel, 3498838, Haifa, Israel
[6]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Lausanne, Switzerland
[7]SIB Swiss Institute of Bioinformatics, CH-1015, Lausanne, Switzerland
[8]Department of Computer Science, University College London, Gower St, WC1E 6BT, London, United Kingdom

*Corresponding author. b.chain@ucl.ac.uk

## Abstract

**Motivation: The six complementarity determining regions (CDRs) of the T cell receptor (TCR) form multiple contacts with cognate peptide and major histocompatibility complex, thus determining antigen specificity. However, the contacts between the CDRs themselves are less understood. Results: Our systematic study of all available TCR crystallographic structures identified consistent patterns of intra- and inter-chain CDR contacts in both free and antigen-bound TCRs. In addition, the protein sequences of TCR$\alpha$ and TCR$\beta$ from sets of TCRs which recognise a shared antigen shared mutual information and were not independent. As a result, sequence-based models can partially predict TCR$\alpha$/TCR$\beta$ pairing de novo. The conserved patterns of CDR amino acid contacts, and the mutual sequence constraints between antigen-specific sets of TCR $\alpha$ and $\beta$ chains represent an under-appreciated element of TCR structure, which may play an important role in T cell antigen recognition. Availability and Implementation: The code and data necessary to reproduce the analyses are available at `https://github.com/mm523/TCRab-pairing`. Contact: b.chain@ucl.ac.uk. Supplementary Information: Supplementary data are available at Bioinformatics online.**

**Keywords** pMHC, TCR, TCR pairing

## Introduction

The $\alpha\beta$ T cell receptor (TCR) is a heterodimer which recognises peptides bound to major histocompatibility complex (MHC) molecules (pMHC). Each TCR chain is generated by somatic V(D)J recombination in each T cell between a Variable [V], a Joining [J] and, in the case of TCR$\beta$, a Diversity [D] segment. The recombination generates high sequence diversity, which has been estimated to be greater than $10^8$ (Mora and Walczak, 2019; Keşmir et al., 2000; Davis and Bjorkman, 1988). TCRs bind pMHC through its complementarity-determining regions (CDRs or CDR1, CDR2 and CDR3; Milighetti et al. (2021); Szeto et al. (2020)), which are the most variable regions of each TCR chain. CDR1 and CDR2 are encoded by the germline V gene, while CDR3 spans the recombination junction between V, D and J genes and is consequently the most diverse (Schwartz and Hershberg, 2013). Notably, thousands of TCRs of different sequences can bind to the same pMHC (Dash et al., 2017; Glanville et al., 2017) and a single TCR may bind to many pMHCs (Sewell, 2012).

Experimental studies have suggested that interactions between CDR loops might be important for epitope specificity, by for example affecting the inter-domain angle between the two TCR chains (McBeth et al., 2008). The interactions between CDRs may therefore constitute an important consideration in the design and optimisation of TCR-engineered T cells (Baulu et al., 2023; Campillo-Davo et al., 2020) or soluble TCRs (Robinson et al., 2021) for therapeutic applications. In this study we investigate the interaction between the TCR $\alpha$ and $\beta$ chain CDR loops. We first systematically document the intra-chain and inter-chain contacts that the CDR loops make with each other. We then quantify the extent to which the CDR sequences of each chain constrain and hence predict the sequence of the other chain. If contacts between CDR loops are a conserved feature of TCRs and impact antigen binding, we expect the sequences of the two TCR chains to be interdependent in epitope-specific but not unselected repertoires. Since TCR repertoire data is often single chain only, the extent to which information on one chain predicts the sequence of the other is an important practical question, as is quantifying the extent to which the sequence of one chain

determines specificity. Paired TCR$\alpha\beta$ have been shown to carry more information about epitope specificity than each chain independently (Carter et al., 2019; Springer et al., 2021; Mayer and Callan, 2023; Henderson et al.), although the increase from adding the paired chain varies for different epitopes. We quantify the interchain constraints using a mutual information (MI) framework, and then use models originally developed in the context of protein co-evolution to predict chain pairing.

## Methods

### Analysis of existing crystal structures

A list of crystal structures for unbound TCRs and TCR-pMHC complexes was obtained from The Structural T-Cell Receptor Database (STCRDab, Leem et al. (2018), downloaded on 27$^{\text{th}}$ February 2023, N = 918 structures). We selected all $\alpha\beta$TCRs renumbered according to the International Immunogenetics Information System (IMGT, Lefranc et al. (2003)), excluding all TCRs bound to non-canonical MHC (N = 616). We removed any sequence with missing epitope information (N = 590) and excluded non-peptide or chemically modified peptide antigens or superantigens, since their binding modalities may differ and are beyond the scope of this study (N = 527). Where multiple structures were available from a single PDB, we chose the structure with the most complete chain combination (N = 340 structures). For the analysis of bound and unbound receptors, we extracted the sequences from the resulting 340 PDB files, and retained a non-redundant set of 157 structures of TCRs bound to cognate pMHC, and 78 unbound TCRs using STRCRDab annotations.

Structures were visualised using Pymol Open Source v2.5.0 (Schrödinger, LLC, 2015).

To calculate intra-chain and inter-chain distances between two residues, we calculated the Euclidean distance between each pair of non-H atoms and retained the minimum distance across all atom pairs using the BioPDB package in Python v3.11 (Hamelryck and Manderick, 2003). Contacts were defined as a minimum distance $< 5$Å between two residues. A number of alternative thresholds were evaluated and 5Å was chosen as the average number of contacts between CDRs forms an elbow around the 5Å threshold (Figure S1).

For the epitopes analysed, we retrieved the crystal structures of pMHC complexes from the Protein DataBank (PDB), where available. Solvent-accessible surface area (SASA) for each epitope was calculated using Pymol Open Source v2.5.0 (Schrödinger, LLC (2015), command `get_area` on the peptide chain with `dot_solvent` 1 and solvent radius 1.4 Å) on a representative structure for each epitope (Table 1).

| Epitope | PDB code | SASA | Reference |
|---|---|---|---|
| ASNENMETM | 4HUX | 311 Å$^2$ | Valkenburg et al. (2013) |
| CINGVCWTV | 3MRG | 264 Å$^2$ | Reiser et al. (2014) |
| ELAGIGILTV | 1JF1 | 329 Å$^2$ | Sliz et al. (2001) |
| GILGFVFTL | 2VLL | 259 Å$^2$ | Ishizuka et al. (2008) |
| GLCTLVAML | 3MRE | 341 Å$^2$ | Reiser et al. (2014) |
| LLWNGPMAV | 5N6B | 307 Å$^2$ | Bovay et al. (2018) |
| LSLRNPILV | 3BUY | 308 Å$^2$ | La Gruta et al. (2008) |
| NLVPMVATV | 3GSO | 375 Å$^2$ | Gras et al. (2009) |
| RAKFKQLL | 3SPV | 243 Å$^2$ | *To be published* |
| SSLENFRAYV | 1WBY | 464 Å$^2$ | Meijers et al. (2005) |
| SSYRRPVGI | 1WBZ | 331 Å$^2$ | Meijers et al. (2005) |
| YLQPRTFLL | 7N6D | 460 Å$^2$ | Chaurasia et al. (2021) |
| SPRWYFYYL | 7LGD | 436 Å$^2$ | Lineburg et al. (2021) |

**Table 1 List of epitopes for which a PDB structure is available and calculated solvent-accessible surface area.**

### Paired-chain datasets

The VDJDb database (Goncharov et al., 2022) was exported on 3$^{\text{rd}}$ February 2023. It was refined to only include TCRs for which both TCR$\alpha$ and TCR$\beta$ chains are available, and to include unique TCR-epitope pairs. We removed epitopes for which fewer than 100 (N = 914) or more than 10,000 (N = 1) sequences are available, to reduce skewing due to small samples and spurious annotations and be able to compare similarly-sized sets (Figure S2A). We chose the largest cut-off size which allowed us to retain enough epitopes for analysis and we also excluded very large sets for which computational time would have been unwieldy. stitchr (Heather et al., 2022) was used to obtain full chain sequences for each V/CDR3/J annotation for both mouse and human TCRs, and ANARCI (Dunbar and Deane, 2016) was used to obtain IMGT-renumbered (Lefranc et al., 2003) CDR3 sequences from each chain. To ensure that CDR3 sequences were all the same length for MI calculations, sequences longer than 19 residues were discarded (Figure S2B), corresponding to $< 1\%$ (67/9852) of all CDR3$\alpha\beta$. Sequences shorter than 19 were padded by addition of gaps at position $\frac{L}{2}$ (where $L$ is the length of the CDR3 sequence). In order to allow within study randomisation for mutual information calculation (Appendix A2), we restricted the analysis to studies which contain at least 3 sequences. `tidytcells v1.6.0` (Nagano and Chain, 2023) was used to standardise to the V and J genes, removing allele information. We further dropped duplicates in TRAV-CDR3A-TRAJ / TRBV-CDR3B-TRBJ pairs within each epitope repertoire after gene standardisation to avoid inflating the estimated mutual information (Appendix A2, N = 41 sequences removed). The final set of sequences comprised 9,713 sequences from 22 epitopes (Table S1). 7 of the epitopes and 4 of the TCRs (defined at the level of CDR3$\alpha$-CDR3$\beta$) in this set are shared with the set of crystal structures analysed. Only one of the 9,713 TCR-epitope pairs overlaps with the TCR-pMHC complexes included in the structural analysis.

Pre-processed TCR sequencing data from Tanno et al. (2020) was retrieved and further processed as in Mayer and Callan (2023). The clonotypes from sample A1, sorted on CD4$^+$CD45RA$^+$CCR7$^+$ and classified by the authors as naïve cells were used as a control unselected repertoire. As for the VDJDb set, a length of 19 was enforced on the CDR3s.

We downloaded all human data from The Observed T Cell Receptor Space (OTS) database (Raybould et al., 2024). We used the annotations as given by the authors to remove $\gamma\delta$ T cells, thymocytes or Mucosal-Associated Invariant T (MAIT) cells, and samples that had $< 100$ sequences. As above, a length of 19 was enforced on the CDR3s.

### Calculating mutual information between TCR regions

To quantify how much restriction one region of the TCR poses on another, we calculated the mutual information (MI) between different TCR regions. For CDR3s, where the diversity is too large to calculate entropy using the full sequence as a single categorical variable, we used a pairwise approximation by calculating the MI between single CDR3 residues and either the V or J genes, or single residues in the paired CDR3. All MI estimations were corrected for small sample size bias and finite sampling, background entropy and batch effects. Full details on the MI calculation are in Appendix A2.

MI was calculated using `scikit-learn v1.1.3` (Pedregosa et al., 2012).

### Calculation of effective set size

We define effective set size as the number of distinct sequences present in a repertoire, accounting for repeated or similar sequences (Weigt et al., 2009; Bitbol et al., 2016). For each sequence ($S$) in a set of size $N$, the number of sequence neighbours within a similarity threshold are counted ($m_S$) and $S$ is assigned a weight $w_S = \frac{1}{m_S+1}$. The total effective set size can be calculated as:

$$N_{\text{eff}} = \sum_{S\in 1}^{N} w_S = \sum_{S\in 1}^{N} \frac{1}{m_S + 1} \tag{1}$$

If all sequences are distinct and have no neighbours, $N_{\text{eff}} = N$, whilst if all sequences are neighbours of all other sequences $N_{\text{eff}} = 1$. The diversity is calculated by Hamming distance on the padded CDR3 sequences, excluding the first 3 and the last 3 amino acids of the CDR3. We set the threshold of similarity to be up to Hamming distance 1 for CDR3$\alpha$ and CDR3$\beta$, and up to Hamming distance 2 for CDR3$\alpha\beta$ (calculated as a sum of the distance over the two chains). The effective set size $N_{\text{eff}}$ is normalised by total repertoire size $N$ for comparison across repertoires. The effective set size was not correlated to total repertoire size (Figure S3).

### Pairing models

We adapted two previously-published pairing algorithms to pair TCR$\alpha\beta$ chains: the mutual information iterative pairing algorithm (MI-IPA, Bitbol (2018)) and the graph alignment algorithm (GA, Bradde et al. (2010)). The two methods perform pairing by either maximising the residue-level MI, or the alignment of similarity graphs of CDR3$\alpha$ and CDR3$\beta$ sequences, respectively. These models were originally developed to leverage two distinct signals in co-evolving protein families: the co-evolution of specific residue positions (detected by MI) and the similarity in the phylogenetic trees of co-evolving protein families (detected in the graph alignment). We also evaluated a combination of the two algorithms (GA+MI-IPA, Gandarilla-Pérez et al. (2023)). Full details on the implementation of each algorithm are in Appendix A3.

The MI-IPA, GA and GA+MI-IPA models for each epitope were run on a computer with Intel Xeon W-2295 × 36 processor with 128GB RAM. All models were run on each epitope-specific repertoire from VDJDb separately, without using a training set of known pairs, as would be the case if no information about pairing was available. We inputted the CDR3 sequences only. To equalize epitope-specific TCR set sizes and reduce computational load, we subsampled the three largest epitope sets to 700 sequences, matching the size of the fourth-largest epitope-specific set. The set used for each epitope is summarised in Table S1. For all models, epitopes GLCTLVAML (EBV) and YLQPRTFLL (SARS-CoV-2) were used to find the optimal parameters.

To quantitatively compare pairing algorithms, we calculated the proportion of predicted pairs that are correct, i.e. they are paired as observed in the downloaded VDJDb set (precision). Importantly, there are repeats in the $\alpha$ and $\beta$ chain sets which will influence this calculation (Figure S4).

### Benchmarking the pairing models

To benchmark the performance of MI-IPA, GA and GA+MI-IPA, we predicted pairing using TULIP (Meynard-Piganeau et al., 2024) and SCEPTR (Nagano et al., 2025). TULIP prediction was performed using the available pre-trained model. For each epitope, we generated all possible CDR3$\alpha$-CDR3$\beta$ pairs available from each individual and predicted the probability of binding of each pair against the epitope. We then assigned $\alpha\beta$ pairs with a greedy algorithm similar to that used for the MI-IPA. Briefly, for each individual, we selected the most likely pair (lowest score). We then removed the $\alpha$ and the $\beta$ sequences from the set, assigned the best remaining pair and continued iteratively until all pairs were assigned. We calculated precision as above.

To assign pairing with SCEPTR, we embedded the $\alpha$ and the $\beta$ chains separately using the pre-trained default model. Since the training is performed by masking sections of the TCR, we expect that $\alpha\beta$ paired chains may have similar embeddings. Therefore, we calculated the Euclidean distance between all $\alpha$ and all $\beta$ chains from one individual for each epitope and then assigned pairs with the same greedy algorithm as above (starting with the pair with the smallest distance) and calculated precision as previously defined. SCEPTR was run both with its default model (trained on paired data) and with a model trained on synthetic data, in which the pairing in the training set was random (SCEPTR - synthetic).

## Results

### Contacts between CDR loops in existing TCR structures

One type of interaction between CDR loops may be mediated by molecular contacts between amino acids. We therefore examined 340 curated TCR-pMHC structures available from the The Structural T-Cell Receptor Database (STCRDab, Leem et al., 2018). We defined contact as minimum distance between two non-H atoms $< 5$ Å. Observation of 4 representative TCRs (Figure 1) shows that CDR loops form a continuous contact surface that binds pMHC. There are both intra- (between two CDR residues on the same chain) and inter-chain (between one residue on the $\alpha$ and one residue on the $\beta$ chain) contacts observed both in the bound (right panel, indicated with a *) and in the unbound (middle panel) TCR structures, although the number of contacts varies between the ligated and free receptors.

Consistent with Figure 1, the intra-chain distances for the four example unligated structures (Figures 2A and B for TCR$\alpha$ and $\beta$, respectively) show intra-chain contacts between CDR1 and CDR3 and CDR1 and CDR2, both in the TCR$\alpha$ and in the TCR$\beta$ chain. The patterns are consistent across the four structures and in both chains. The same analysis was extended to 78 unique unbound receptors (Figure S5A), 157 unique pMHC bound TCRs (Figures 2C, D and S5B, C) and TCRs that bind MHC with reversed polarity (Figure S5D). We also calculated the number of contacts between CDR loops, defined as minimum distance between two non-H atoms $< 5$ Å, for all 340 available structures (Figure S6A). The patterns observed in the 4 example structures are generalisable and observed in TCRs both in the unligated and ligated form. Strikingly, all 340 PDB files analysed show contacts between CDR1 and CDR3 on both chains (Figure S6A); median of 13 contacts [range 5-23] for TCR$\alpha$ and median of 14 [range 8 - 22] contacts for TCR$\beta$). CDR2 in turn forms contacts with CDR1 (median of 5 contacts for both TCR$\alpha$ and TCR$\beta$, ranges of 0-12 and 0-7, respectively). We also observe

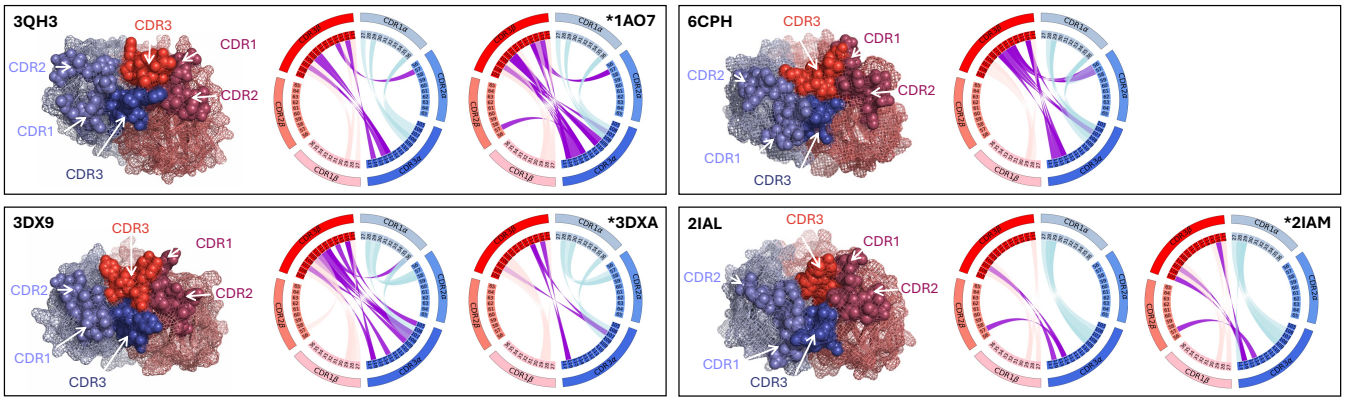


**Figure 1 CDR loops come together to form the TCR binding site in four example TCRs.** Positioning of CDR loops of 4 representative TCRs (2 recognise epitope on class I MHC, 3QH3 and 3DX9; 2 recognise epitope on class II MHC, 6CPH and 2IAL). The TCR$\alpha$ chain is in shades of blue, whilst the TCR$\beta$ is in shades of red. CDR1, 2 and 3 for each chain are represented with spheres and annotated on the figure. Mesh shows the rest of the TCR structure. On the right of each structure, the circle plot shows contacts between the CDR loops (defined as minimum distance between two non-H atoms on two residues $< 5$Å) for the same TCR in both the unbound (left) and bound (right, indicated with a $*$) structure. Light blue: contacts between loops on TCR$\alpha$; light red: contacts between loops on TCR$\beta$; purple: contacts between the two chains. 3QH3: Scott et al. (2011), 1AO7: Garboczi et al. (1996), 3DX9 and 3DXA: Archbold et al. (2009), 6CPH: Galperin et al. (2018), 2IAL and 2IAM: Deng et al. (2007). Structures are plotted using Pymol Open Source (Schrödinger, LLC, 2015). Circos plots are plotted using Python package pyCirclize. CDR = Complementarity-Determining Region.

contacts between CDR2 and CDR3 (range 0-3 contacts in TCR$\alpha$ and 0-4 contacts in TCR$\beta$, Figure S6A).

Inter-chain distances for the four unligated TCRs in Figure 1 also show contacts both in the conserved framework regions and in the CDR loops (Figure 3A). These interactions are also observed in most studied structures (Figures 3B and S7A-C; and quantified in Figure S6B-C). Specifically, framework region 2 (FR2), found between CDR1 and CDR2, makes contacts with FR2 on the other chain, as well as with the edges of the CDR3, where the CDR3 sequence is least variable (the diamond-shaped pattern, median 9 contacts [range 0-17] between FR2$\alpha$-CDR3$\beta$ and median 7 contacts [range 1-16] between FR2$\beta$-CDR3$\alpha$, Figure S6B; contacts with edges and middle of the CDR3 are quantified in Figure S6C: median of 3 [range 0-9] central contacts for FR2$\alpha$-CDR3$\beta$ and of 2 [range 0-8] for FR2$\beta$-CDR3$\alpha$ compared to median of 6 [range 0-12] edge contacts for FR2$\alpha$-CDR3$\beta$ and of 5 [range 0-10] for FR2$\beta$-CDR3$\alpha$). Inter-chain contacts are also observed between CDR1 and 2 and the CDR3 on the paired chain (Figure S6B). We observed contacts between the two CDR3s (median of 7 contacts, range 1-22), whilst no contacts are observed between the germline loops CDR1 and CDR2 on the two chains in any of the 340 PDB files analysed (Figure S6B). The pattern of contacts between the two CDR3s is X-shaped: the interactions involve the variable junctional regions, but fewer interactions are observed between the central variable regions and the conserved segments on the loop (quantified in Figure S6C: median of 3 [range 0-13] interactions with the central segment, compared to median of 2 [range 0-6] for the edges). Notably, these patterns can also be observed in TCRs which bind pMHC with reverse orientation (Figure S7D). Overall, these results suggest that the CDR loops form multiple intra- and inter-chain contacts independently of pMHC binding.

## Quantifying sequence inter-dependence between TCR$\alpha$ and $\beta$ in epitope-specific repertoires using MI

We examined whether, in the context of sets of TCRs recognising a shared pMHC, the sequences of the TCR $\alpha$ and $\beta$ were determined independently by the requirement to bind the specific antigen, or whether the $\alpha$ and $\beta$ CDR sequences were dependent on each other. We initially observed increased sequence similarity in TCR$\beta$ paired to similar TCR$\alpha$, and vice versa (Appendix A1, median increase of 0.56 and 0.15, respectively). We quantified interchain constraints in a more generalisable way using mutual information (MI) to measure the extent to which a change in one chain is accompanied by a complementary change in the paired chain. We estimated the MI between TCR$\alpha$ and $\beta$ regions of epitope-specific repertoires using a pairwise approximation and corrected for sample size and background distributions. As negative controls, we estimated the MIs for all TCRs in VDJDb ("background"), and for a single sample of sorted CD4+ naïve T cells from Tanno et al. (2020) ("Tanno::A1::naïve"). The former contains information about the 22 epitopes under study pooled together and might have some residual epitope-specific signal. The latter depends only on the biology of receptor recombination and thymic selection, without the influence of antigen selection.

Figure 4A shows the estimated MIs for each epitope compared to VDJdb background. The naïve repertoire showed no increase in MI compared to the pooled VDJDb repertoire. In contrast, most individual epitopes showed an increase in MI compared to the background. A few epitopes, such as AVFDRKSDAK showed low or no signal. The MIs for all epitopes are summarised in Figure 4B and the MIs for all pairs in all epitopes are in Table S2.

To understand the statistical dependence between V gene and CDR3 in more detail, we examined the estimated intra-chain MI between the V gene and each CDR3 residue (Figure 5A). As CDR3 positions 104 and 118 on both chains are conserved (C for position 104, F or W most commonly for position 118), the MI is $\sim 0$. The initial residues (105-108) on CDR3$\alpha$ and $\beta$ show a higher MI with the V gene than the rest of the CDR3 for all epitopes and in the background, since the beginning of the CDR3 is usually encoded by the end of the V gene. However, we also observed MI between the V gene and the central residues of the CDR3 for most epitopes but not in the naïve repertoire. Overall, V-CDR3 MI is

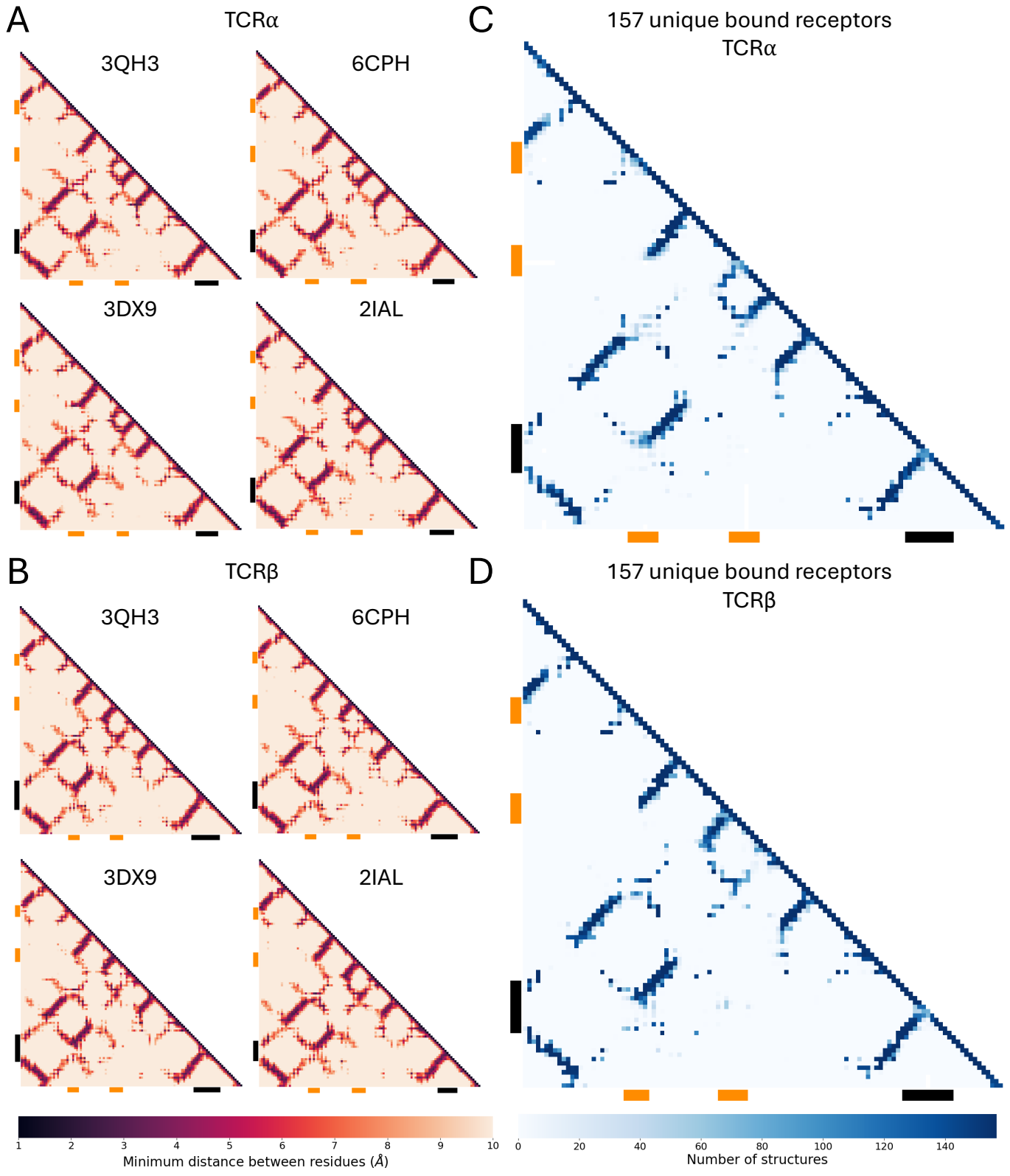


**Figure 2 Intra-chain interactions are observed in bound and unbound TCR structures. A, B)** Intra-chain contact maps for the unligated structures in Figure 1 for the TCR $\alpha$ and $\beta$ chain, respectively. The distance (Å, colour bar) is calculated between all pairs of atoms in the two residues, and the minimum distance is shown. **C, D)** Intra-chain contact maps were calculated for the TCR $\alpha$ and $\beta$ chain (respectively) of 157 unique bound TCRs. A contact is defined as minimum distance between two non-H atoms on two residues $< 5$ Å. The heatmap shows the number of structures that are found to make contact at each pair of positions. Only positions present in $> 75\%$ of structures according to the IMGT numbering scheme are shown. Only the bottom half of each distance matrix is shown as it is symmetrical across the diagonal. In each heatmap, the positions of CDR1, 2 (orange) and 3 (black) are highlighted.

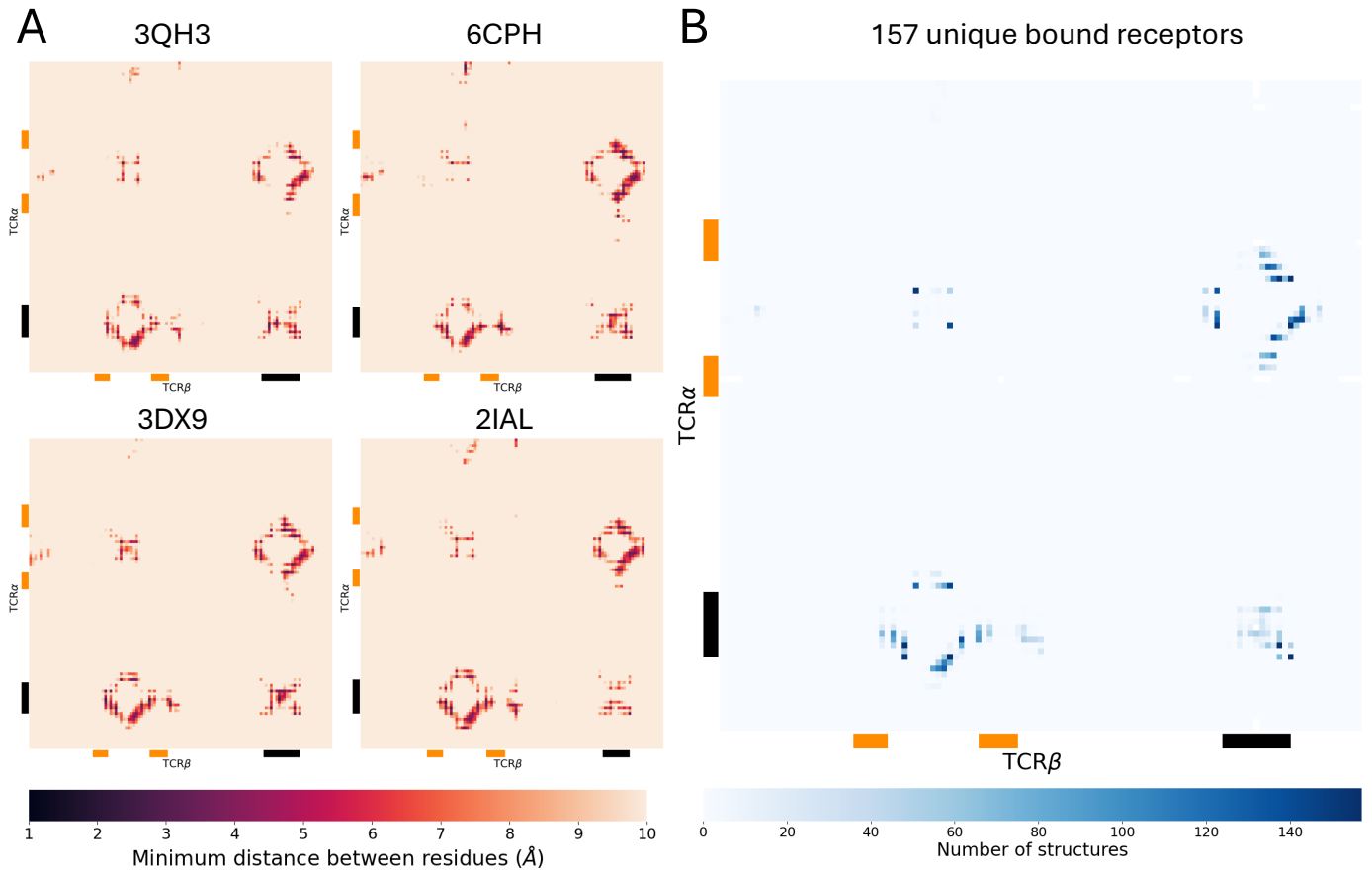


**Figure 3 Inter-chain interactions are observed in bound and unbound TCR structures. A)** Inter-chain (TCR$\alpha$ to TCR$\beta$) contact maps for the four structures in Figure 1. The distance (Å, colour bar) is calculated between all pairs of atoms in the two residues, and the minimum distance is shown. **B)** Inter-chain contact maps were calculated for 157 unique bound TCRs. A contact is defined as minimum distance between two non-H atoms on two residues $< 5$ Å. The heatmap shows the number of structures that are found to make contact at each pair of positions. Only positions present in $> 75\%$ of structures according to the IMGT numbering scheme are shown. In each heatmap, the positions of CDR1, 2 (orange) and 3 (black) along the sequence are shown as coloured bars along the x and y axes.

higher than background in 10/22 and 9/22 epitopes and higher than naïve repertoire in 18/22 and 19/22 epitopes for V$\alpha$-CDR3$\alpha$ and V$\beta$-CDR3$\beta$, respectively.

We also examined the MI between the CDR3$\alpha$ and CDR3$\beta$ (Figure 5B). Many epitopes but not background or naïve repertoire showed inter-dependence between $\alpha$ and $\beta$ CDR3 sequences, with an epitope-specific MI footprint (CDR3$\alpha$-CDR3$\beta$ MI is greater than background in 17/22 epitopes). Antigen-dependent TCR selection therefore imposes mutual sequence constraints between the two chains of the TCR. As in previous analyses, epitope AVFDRKSDAK showed very low inter-CDR3 MI.

In order to align the sequences to calculate MI, the CDR3 sequences were padded. We thus investigated the impact of padding on MI. As expected, residue positions where padding occurs more often have lower MI (Figure S8, slope $< 0$ and p-value $< 0.05$ for all relationships studied). To get a better sense of the impact of padding on MI estimation, we also compared two padding methods: padding in the middle of the sequence (as in IMGT numbering, used in all analyses) and padding at the C terminus of the sequence. MI estimated from CDR3 sequences with either method (Figure S9) were correlated for most MI pairs (Spearman $\rho > 0.95$ in 7/9 pairs, CI and p-values for each comparison reported in the figure). The correlation between estimated MIs with the two padding strategies for J$\alpha$-CDR3$\alpha$ and J$\beta$-CDR3$\beta$ was lower (Spearman $\rho = 0.68$, 95% CI = [0.39-0.85], p-value $= 2.26 \times 10^{-4}$ and $\rho = 0.66$, 95% CI = [0.35-0.84], p-value $= 4.49 \times 10^{-4}$, respectively). This reduction is expected since the J gene partially determines the sequence of the C terminal of the CDR3 and padding in the middle will better preserve the alignment at the C terminus, while padding at the end will disrupt it and therefore reduce the MI. Overall, the MI calculation was robust to the padding strategy, which did not alter the relative signal of epitopes and background.

Finally, we extended the MI estimation to poly-epitope repertoires. We downloaded TCR repertoires from the OTS database (Raybould et al., 2024) and calculated TCR $\alpha\beta$ MI for all studies (Figure S10, Table S3). Sorted memory and AIM+ cells, which are enriched for antigen-specific groups of TCRs, have higher median CDR3$\alpha$-CDR3$\beta$ MI (Table S3), compared to unsorted and unstimulated T cells.

## Examining MI variation across epitopes

A striking feature of the TCR chain sequence inter-dependence was the variability observed between epitopes. Understanding the factors

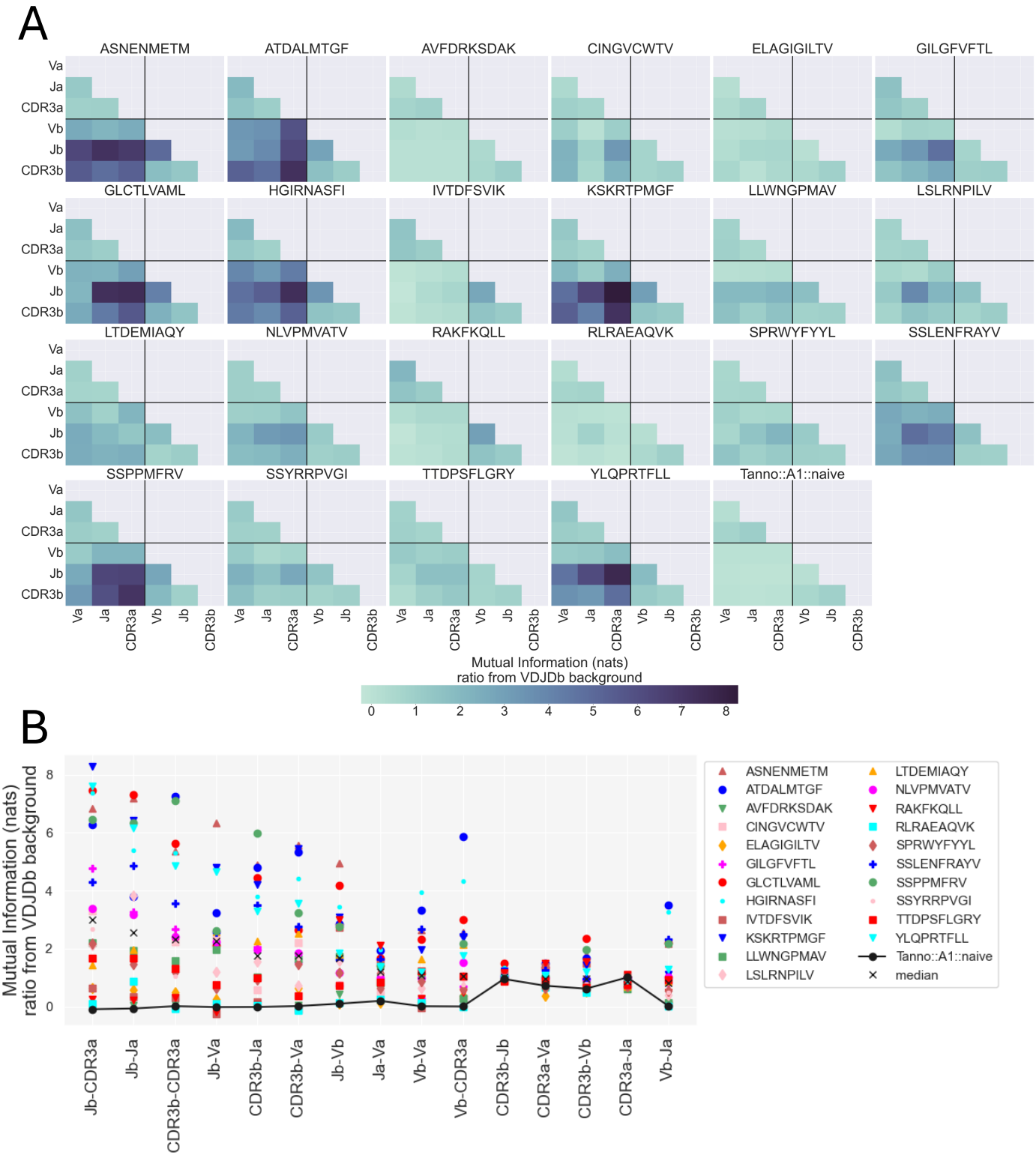


**Figure 4 Mutual information between TCR regions in epitope-specific repertoires. A)** Mutual information (MI) between different regions of the TCR (V genes, J genes, CDR3 ) was calculated for each epitope-specific repertoire. In each heatmap, the top-left and bottom-right quadrants contain intra-chain MI ($\alpha$ in top-left and $\beta$ in bottom-right). The bottom-left quadrant shows the inter-chain MI. The MI was estimated for various subsamples for each epitope, and corrected for finite size effects (Figure A5). The MI is shown as fold change from VDJDb background to normalise across MI pairs with different alphabet sizes. A pairwise approximation is used in cases involving CDR3 (see Methods). **B)** Summary of the plots in **A**. The x-axis is ordered according to the median MI across all epitope repertoires (indicated with an X). Each epitope is identified by a unique combination of colour and shape.

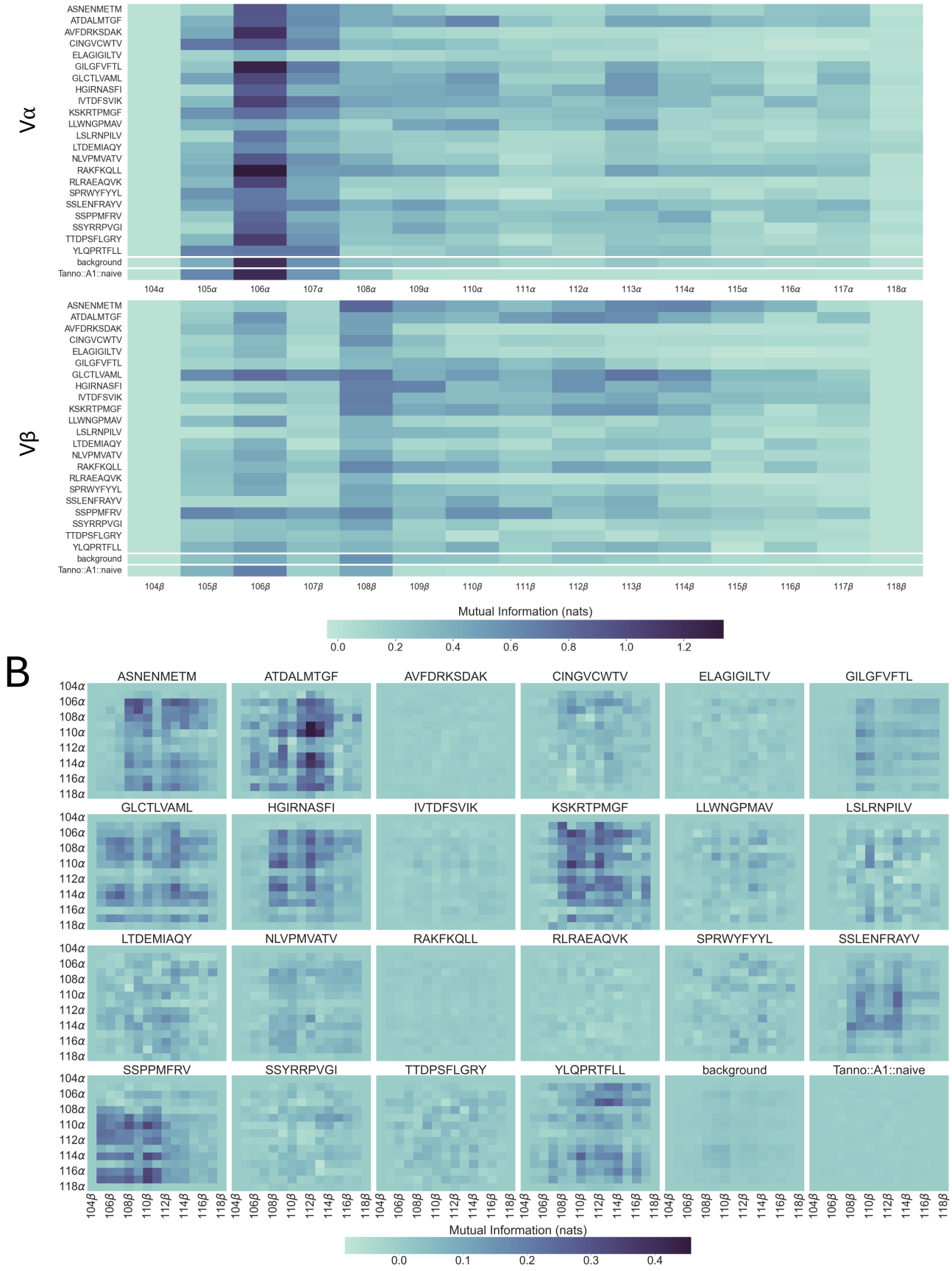


**Figure 5 Mutual information between CDR3 and associated V gene and CDR3$\alpha$-CDR3$\beta$. A)** Mutual information (MI) between V genes and each residue position on the CDR3. Top: V$\alpha$ to CDR3$\alpha$; bottom: V$\beta$ to CDR3$\beta$. **B)** MI between each pairs of positions on CDR3$\alpha$ and CDR3$\beta$. The MI was calculated with pairwise approximation, extrapolation and correction for background (for both **A** and **B**, see Methods). IMGT positions between residues 111 and 112 are not shown as they are not present in most TCRs analysed.

which drive this epitope dependence may have important implications for predicting the antigen specificity of an orphan TCR.

To validate that the observed results were not simply an artefact of finite size effects, we correlated estimated MI and epitope repertoire size ($N$). We did not detect significant correlations for most of the relationships investigated (Figures 6A, left-most column and S12A).

Epitope-specific responses are characterised by differing levels of sequence similarity (see for instance Dash et al., 2017) and occasionally public TCR chains (see for instance Zhong et al., 2007; Pogorelyy et al., 2022). To evaluate the impact of similar or public TCR chains in MI estimation, we measured the correlation between MI and the effective set size of CDR3$\alpha$, CDR3$\beta$ and CDR3$\alpha\beta$, which gives a measure of the number of distinct sequence clusters in a repertoire (Figures 6A and S12B). $\beta$ and $\alpha\beta$ effective set sizes were negatively correlated with MI calculated from CDR3$\alpha$-CDR3$\beta$ (slope $= -18.2$, 95% CI $= [-31.2, -5.3]$ and slope $= -34.7$, 95% CI $= [-50.1, -19.3]$, respectively), V$\alpha$-CDR3$\alpha$ (slopes $= -3.2$ $[-5.3, -1.1]$ and $-3.7$ $[-7.0, -0.4]$, respectively) and V$\beta$-CDR3$\beta$ (slopes $= -4.8$ $[-7.3, -2.2]$ and $-6.6$ $[-10.4, -2.9]$, respectively). $\alpha\beta$ effective set size was also correlated to the MI from many other evaluated pairs (Figure S12B). Effective $\alpha$ set size was negatively correlated with V$\alpha$-CDR3$\alpha$ (slope $= -4.2$ $[-7.1, -1.3]$) and V$\beta$-CDR3$\beta$ (slope $= -4.6$ $[-8.6, -0.6]$), but not CDR3$\alpha$-CDR3$\beta$. Overall, these significant negative correlations suggest repertoire characteristics can explain $\sim 20 - 67\%$ of the variability in estimated MI across epitopes and that repertoires dominated by a single sequence or cluster of similar sequences (smaller set size) are associated with a higher MI.

We also examined whether the MI depends on how many features the peptide offers for the TCR to bind. We approximated the "featureness" of a peptide by its solvent-accessible surface area (SASA) for the 13 epitopes where a structure was available (Table 1). Epitopes that have a relatively featureless surface (low SASA), or which are bulged out of the MHC (high SASA) tend to generate less diverse TCR repertoires (Turner et al., 2006, 2005) and might therefore have higher MI. A parabola fitted to the relationship between SASA and MI did not show any significant relationship (Figure 6B). However, the small number of epitopes examined limited the statistical power of the analysis to only detect large effect sizes, and moderate to small effects were below the threshold of detection with 80% power and p-value $< 0.05$ (Figure S11).

## Predicting TCR$\alpha\beta$ pairing *de novo*

Given the observed sequence signals in TCR pairing, we tested whether TCR$\alpha\beta$ pairing could be predicted *de novo*. Sequence-based methods have been very valuable in studying the evolutionary constraints imposed on the sequences of co-evolving protein families (De Juan et al., 2013). Epitope-specific TCR pairing signals do not arise from the evolutionary process at play in interacting protein families, but rather from antigen selection of TCRs produced by V(D)J recombination. The methods used to study protein co-evolution, however, are generalisable to studying the constraints imposed by a selection force on any sets of pairs of variable interacting proteins, and hence to predict pairing. We therefore adapted two methods: a MI-based algorithm (MI-IPA, Bitbol, 2018), which exploits the MI between co-evolving residue pairs (Figure 5), and a graph alignment algorithm (GA, Bradde et al., 2010), which exploits the similarity in phylogenetic trees, or in our use-case, similarity graphs, Figures A3 and A4). We also tested a combination of the two algorithms (GA+MI-IPA, Gandarilla-Pérez et al., 2023).

In 6/22 epitopes, the MI-IPA performs significantly better than chance expectation (median fold change from chance expectation across all 22 epitopes = 1.29, range = 0.51-2.70; Figure 7). The GA shows significantly better performance than chance expectation in 14/22 epitopes (median fold change from chance expectation across all 22 epitopes = 1.49, range = 0.81-3.11; we consider it successful for large epitopes based on whether it is significant in at least 3/5 subsamples; Figure 7). As pairing assignments were repeated 10 times by the MI-IPA and 100 times by the GA for each epitope to account for stochasticity, we explored whether the stability of the assignments between repeats could provide a confidence score for each pair (as in Bitbol et al., 2016). For most epitopes, the majority of pairs are not assigned stably across repeats (Figures S13A and S14A). Notably, stable pairs have a larger proportion of correct pairs in 23/27 epitope sets with stable pairs in the MI-IPA (Figure S13C) and 22/23 epitope sets with stable pairs in the GA (Figure S14C). Consistently, epitopes which show better performance show a larger proportion of stable pairs (Figures S13D and S14D, Spearman $\rho = 0.796$, 95% CI $= [0.627, 0.894]$, p-value $= 1.8 \times 10^{-8}$ for MI-IPA and Spearman $\rho = 0.699$, 95% CI $= [0.472, 0.839]$, p-value $= 4.3 \times 10^{-6}$ for GA).

We then combined the two methods by using stable pairs from the GA assignments as training set for the MI-IPA (Gandarilla-Pérez et al., 2023), but allowed the MI-IPA to re-pair sequences in the training set to potentially correct any mistakes in the GA classification (Table S4). The combination of the two methods occasionally marginally improved on either method alone and performs significantly better than chance expectation in 15/22 epitopes (median fold change from chance expectation across all 22 epitopes = 1.43, range = 0.37-3.59, Figure 7). However, unlike the MI-IPA and GA, the GA+MI-IPA shows a larger proportion of stable pairs with 12/34 epitope sets having $> 95\%$ stable pairs (S15D). However, there is a lower enrichment for correct pairs in the stable pairs (Figure S15C: 8/34 epitopes have the same or lower percentage of correct pairs among the stable pairs compared to the overall percentage of correct pairs, instead of 4/27 and 1/23 for MI-IPA and GA, respectively, Figures S13C and S14C). We attribute this to the use of a training set in the MI-IPA, as the trajectory of the MI-IPA is less random when a training set is provided.

The limitation in model prediction accuracy can arise in the method (i.e. the way the algorithm selects pairs does not find the optimal solution) or in the data (i.e. the signal in the data is too low to achieve pairing). To distinguish these two limiting factors, we calculated the theoretical best performance for the MI-IPA. We initialised the model with all correctly-paired $\alpha\beta$ for each epitope and ran a single iteration of the MI-IPA to pair all available $\alpha/\beta$ from that epitope repertoire again, effectively teaching the model the pairing rules *a priori* and over-fitting (Figure 7). In all epitopes, the pairing algorithms under-perform compared to the theoretical limit but the theoretical limit remains well below 1 for all epitope repertoires, suggesting that both method and data limitations are present.

Finally, we tested whether TCR sequence embedding methods or models trained to predict TCR-epitope binding could capture the pairing signal. Chain pairing was predicted within epitope-specific repertoires using TULIP (Meynard-Piganeau et al., 2024), a method developed for binding prediction, and SCEPTR (Nagano et al., 2025), a method for TCR sequence embedding. We found that pairing

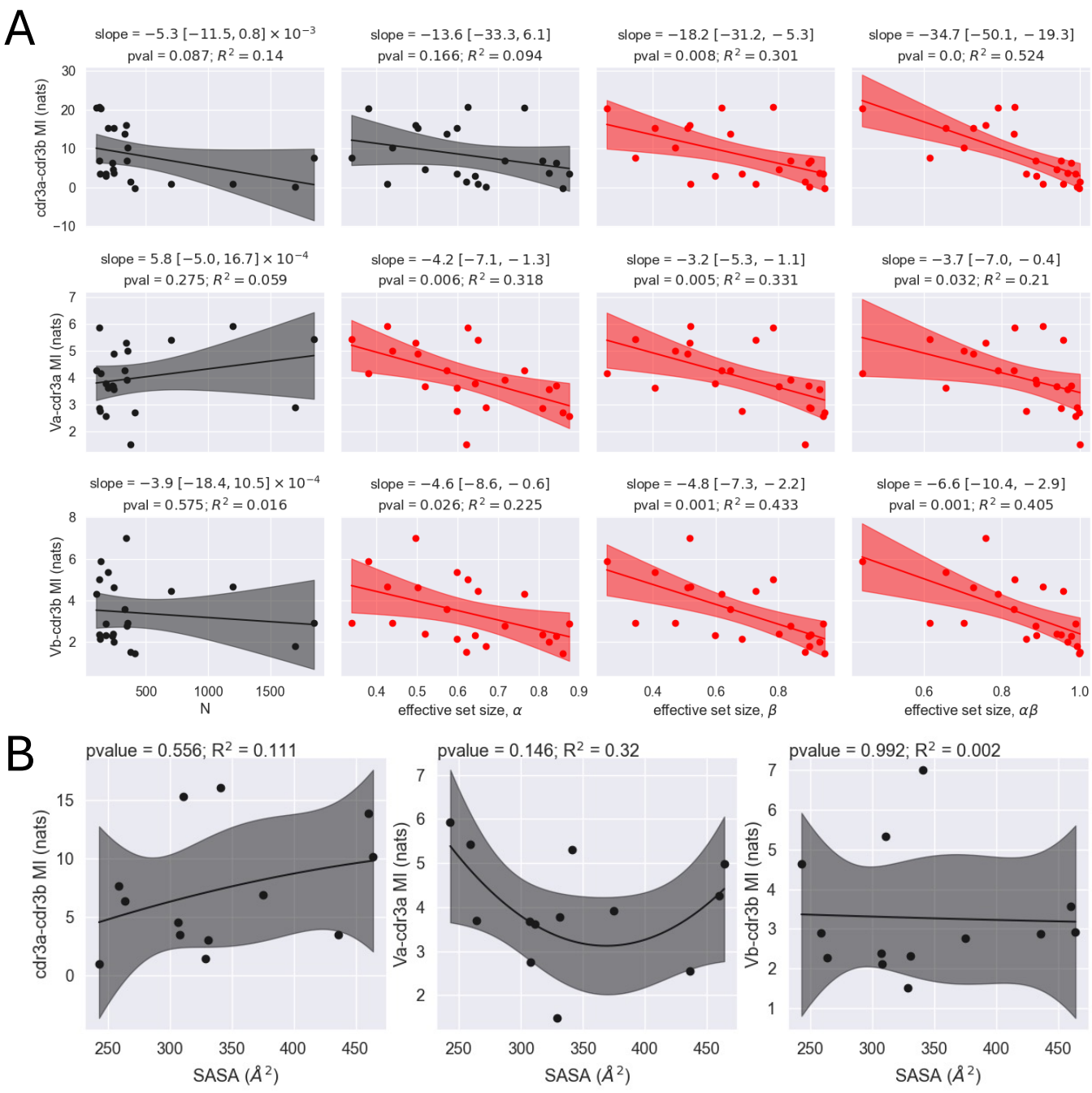


**Figure 6 Mutual information correlates to effective set size but not peptide solvent-accessible area. A)** Correlation of mutual information (MI) and effective set size. Effective set size was calculated by grouping all sequences with Hamming distance of $< 2$ on each chain and $< 3$ for the paired chains, normalised by total repertoire size (N). Correlation with N is also shown as a control, and no significant correlation was found in this case. All correlations with effective set size and N are shown in Figure S12. **B)** Correlation of MI and solvent-accessible surface area (SASA) for the available epitopes (Table 1). Correlations were calculated by fitting a linear regression. The slope of the linear fits and $R^2$ of all fits are reported. 95% CIs for the slope are indicated in square brackets. P-value for the F-test and $R^2$ were calculated for each fit. Fits that have p-value$< 0.05$ are highlighted in red. The shaded area represents the 95% CI of the fit.

models consistently perform better than both TULIP and SCEPTR for TCR $\alpha\beta$ pairing (Figure S16), suggesting that interchain MI is often overlooked in other specificity-prediction methods. However, both TULIP and SCEPTR seem to be able to learn some of the pairing signal in their training, as both perform marginally better than chance expectation for a few epitopes.

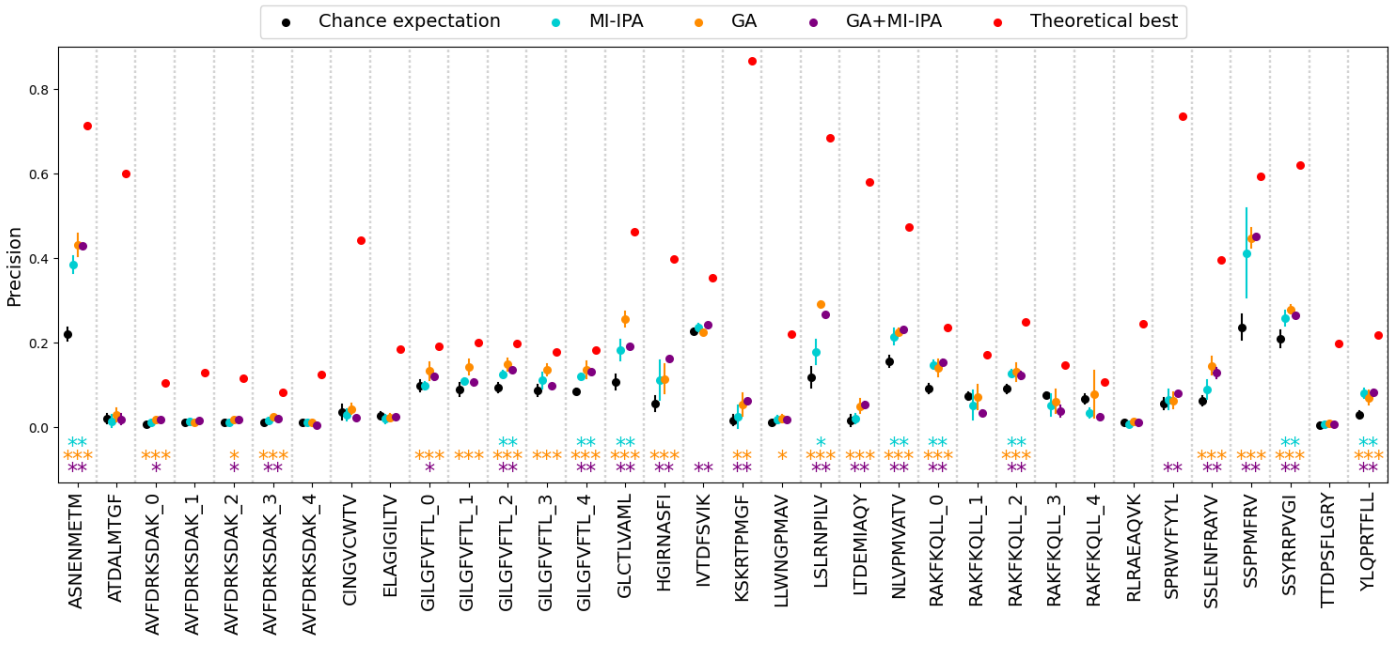


**Figure 7 Pairing models perform better than random at pairing TCR$\alpha\beta$ in epitope-specific repertoires.** Precision is calculated over 10 repeats of the MI-IPA or GA+MI-IPA (cyan and magenta, respectively) or 100 repeats of the GA (orange) for all epitopes. For epitopes with $> 1000$ sequences, 5 subsamples of 700 sequences are shown. The IPA is run both with $\lambda = 1$ (chance expectation, black) and $\lambda = 0.6$ (cyan). Error bars correspond to the standard deviation over the repeats. The red circles correspond to the theoretical best performance, i.e. the MI-IPA initialised on all correct pairs and run once to pair all sequences. One-sided Mann-Whitney U-test is calculated for each epitope for the MI-IPA, GA or GA+MI-IPA using the alternative hypothesis that the model performs better than chance expectation. Adjusted p-values with Bonferroni correction are indicated by asterisks in the corresponding colour: ***$< 0.001$; $0.001 \leq$**$< 0.01$; $0.01 \leq$*$< 0.05$.

## Correlates of pairing performance and repertoire characteristics

The pairing prediction performance varied considerably between epitopes (Figure 7). For each model, we identified factors that could impact performance and correlated them with the observed pairing performance. For all models, we evaluated the total size of the repertoire and the size of the individual repertoires, as they increase the combinatorial number of pairs to be evaluated (Bitbol, 2018). For the MI-IPA, the predictive power may also depend on the MI available in each epitope-specific repertoire (Figure 4), whilst the GA might be sensitive to the topology of the similarity networks generated from each of these epitopes. The theoretical maximal, but not the actual MI-IPA performance (calculated on the consensus assignment over 10 repeats) was inversely correlated to individual repertoire size (slope of the linear fit $= -1.1 \times 10^{-3}$, 95% CI $= [-1.7 \times 10^{-3}, -0.6 \times 10^{-3}]$) and total epitope repertoire size (slope $= -7.7 \times 10^{-4}$, 95% CI $= [-11.2 \times 10^{-4}, -4.2 \times 10^{-4}]$), and positively correlated to MI (slope $= 1.6 \times 10^{-2}$, 95% CI $= [0.3 \times 10^{-2}, 2.8 \times 10^{-2}]$; Figure 8A). These 3 variables together can explain 59.1% of the variance observed across epitopes in the theoretical best performance, and 34.7% in the learning scenario (Table S5). GA performance (calculated on the consensus assignment over 100 repeats) correlated with the extent of sequence clustering (higher average degree on the TCR$\beta$ similarity graph [slope of the linear fit $= 5.1 \times 10^{-3}$, 95% CI $= [1.8 \times 10^{-3}, 8.4 \times 10^{-3}]$], smaller proportion of singlets calculated on the graph generated using pairwise Levenshtein distance with distance threshold $< 3$ [slope for both $\alpha$ and $\beta = -4.1 \times 10^{-1}$, 95% CI $\alpha = [-7.5 \times 10^{-1}, -0.7 \times 10^{-1}]$ and 95% CI $\beta = [-6.2 \times 10^{-1}, -1.9 \times 10^{-1}]$]; Figure 8B). The model built with these variables can explain 79.9% of the variance between epitopes (Table S6).

Finally, we looked at factors that impact the performance of the GA+MI-IPA, as well as the increase and fold-change in precision of the GA+MI-IPA compared to the MI-IPA alone. On top of the repertoire characteristics, we extracted the number of stable pairs from the GA and the percentage of the GA stable set that is correct. The number of stable pairs in the GA and percentage of them that are correct are positively correlated with precision of the GA+MI-IPA (slope $= 2.8 \times 10^{-3}$, 95% CI $= [2.0 \times 10^{-3}, 3.6 \times 10^{-3}]$ and slope $= 2.5 \times 10^{-1}$, 95% CI $= [0.4 \times 10^{-1}, 4.5 \times 10^{-1}]$, respectively), and precision is negatively correlated to the largest ID size (slope $= -4.4 \times 10^{-4}$, 95% CI $= [-8.4 \times 10^{-4}, -0.5 \times 10^{-4}]$). No factor explained why certain epitopes showed improvement using the combined models and others did not. Together, these variables explain 81.8% of the observed variance in precision of the GA+MI-IPA, but explain less $< 50\%$ of the variance in improvement compared to MI-IPA alone (Table S7).

## Discussion

In contrast to the wealth of studies that examine the interactions between TCR CDRs and pMHC, very few have examined the interactions between the $\alpha$ and $\beta$ chain CDRs. In this study we address two related but distinct aims. The first aim is to comprehensively document the amino acid contacts observed between TCR CDR loops in published TCR crystallographic structures. We

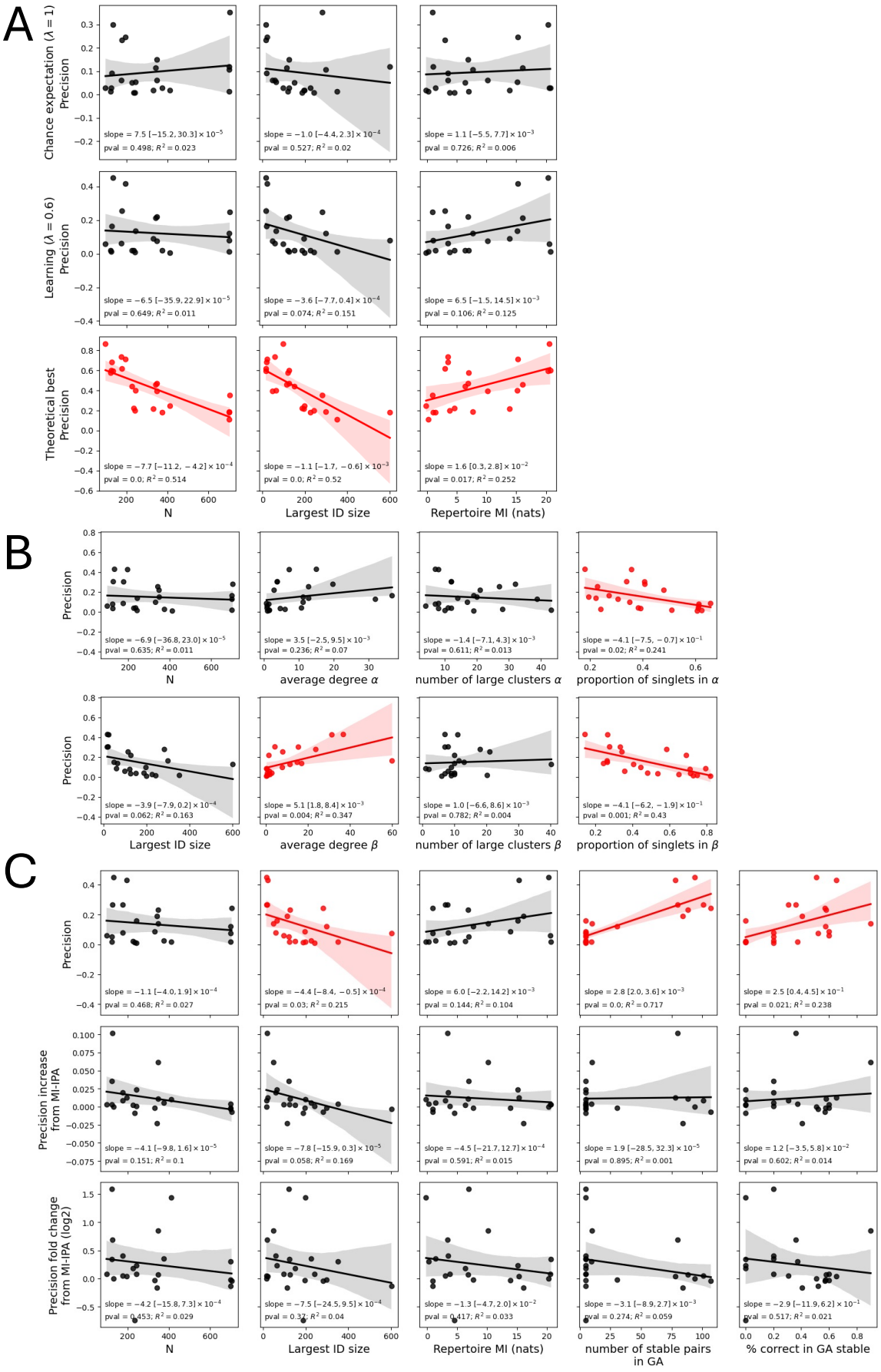


**Figure 8 Correlation of model performance and repertoire characteristics.** *Caption next page.*

**Figure 8 Correlation between model pairing prediction performance, repertoire size and MI.** (*Caption continued.*) **A)** A linear regression was calculated to evaluate the effect of repertoire size (N), largest individual repertoire size (Largest ID size) and total mutual information (Repertoire MI, calculated between CDR3$\alpha$ and CDR3$\beta$ as in Figure 4, without fold change from background) on the precision for MI-IPA (no learning: [confidence = none; $\lambda = 1.0$; $\theta = 0.6$]; learning: [confidence = none; $\lambda = 0.6$; $\theta = 0.6$]; theoretical best: [confidence = none; $\lambda = 0.6$; $\theta = 0.6$; correct pairs as training set]). Subsamples from large repertoires are included as an average, and the repertoire MI is calculated on the complete sample. **B)** A linear regression was calculated to evaluate the effect of N, Largest ID size and similarity network characteristics on the precision for GA. The graph characteristics measured are: average degree of nodes, number of clusters of size $\geq 3$ and proportion of sequences that do not have any neighbours (singlets). These are calculated on the graph built by drawing edges between sequences at Levenshtein distance $< 3$ on CDR3$\alpha$ and CDR3$\beta$ separately. Subsamples from large repertoires are included as an average, and the graph properties are calculated on each subsample separately. **C)** A linear regression was calculated to evaluate the effect of N, Largest ID size, Repertoire MI, number of stable GA pairs and proportion of the GA training set that is correct (% correct in GA stable) on the precision for GA+MI-IPA. Subsamples from large repertoires are included as an average, the repertoire MI is calculated on the complete sample and the GA stability results are calculated for each subsample separately and averaged. In each panel, the precision is calculated using the modal $\beta$ for each $\alpha$ over multiple iterations of the model. Since multiple CDR3$\beta$ may be selected the same number of times for one CDR3$\alpha$, one is selected at random. To mitigate for variation, the mode selection is run 100 times and the average precision is used. In each panel, the slope of the linear fits, the $R^2$ and the p-value for the regression are shown. 95% CIs for the slope are indicated in square brackets. The scatterplots are red when p-value for the F-statistics of the regression is $< 0.05$. The solid line shows the calculated regression and the shaded area the 95% confidence interval of the prediction.

observed a conserved pattern of intra-chain interactions between CDR1 and both CDR2 and CDR3 loops in both chains, and more limited contacts between CDR2 and CDR3 (Figures 1, 2 and S6A). We also quantified inter-chain interactions between the conserved FR2 regions on each chain, and at the interface between the two CDR3s (Figures 3 and S6B-C). These contacts are observed in both pMHC-bound and free TCR structures, although the exact topology of the contact can change upon binding. Contacts between CDR loops are therefore an important intrinsic feature of the TCR which are likely to affect overall receptor shape, antigen binding and stability. Further structural and functional studies will be required to quantify the exact impact of these contacts on TCR-antigen binding. In contrast, conserved interactions between CDR1 and/or CDR2 on one chain, and CDR1 and/or CDR2 on the other chain were not observed, consistent with the absence of constraints in TRAV/TRBV pairing in the overall repertoire (Shcherbinin et al., 2020; Dupic et al., 2019).

The observed close packing of the CDR regions of the two chains in the folded TCR structure discussed above suggested the hypothesis that the contribution of each chain to pMHC binding is not independent and consequently that the sequences of the two chains in sets of TCRs with shared pMHC specificity might be interdependent. Indeed, CDR3$\alpha$ and CDR3$\beta$ sequences within sets of epitope-specific TCRs were shown to consistently share MI (Figures 4 and 5). These observations of chain sequence interdependence are reminiscent of the 'light chain coherence' described for antibodies (Jaffe et al., 2022). Although Shcherbinin et al. (2020) did not find evidence for restriction of TCR chain pairing within epitope-specific repertoires, their study excluded CDR3 sequences that provide the majority of the information content determining epitope binding specificity (Henderson et al.).

Mutual sequence interdependence between two sets of proteins allows, at least in principle, prediction of pairing between members of the two sets. A number of sequence-based approaches have already been developed to predict pairing between co-evolving interacting protein families (Fryxell, 1996; Ochoa and Pazos, 2014; Korber et al., 1993; Dunn et al., 2008; De Juan et al., 2013). Although TCRs do not evolve, but are rather produced independently by V(D)J recombination and then selected by antigen, the mutual interdependence of TCR $\alpha\beta$ sequences suggested that similar methods could be applied to predict pairing between the $\alpha\beta$ sequences of sets of pMHC-specific TCRs. Prediction of pairing would be of significant value in analysis of TCR repertoires, since bulk TCR sequencing usually produces lists of unpaired TCR $\alpha$ and $\beta$ sequences. In order to apply the available methods, multiple TCR$\alpha$ and TCR$\beta$ chains which bind to the same epitope are treated analogously to 'paralogs' from interacting protein families, while similar TCR$\alpha$ or $\beta$ sequences that bind the same epitope but are found in different individuals are 'orthologs'. Three co-evolution based methods (Bitbol, 2018; Bradde et al., 2010; Gandarilla-Pérez et al., 2023) were adapted to predict TCR$\alpha\beta$ pairing in lists of TCR$\alpha$ and TCR$\beta$ sequences annotated to a single epitope. Each model was evaluated for its success in making *de novo* pairing predictions, in the absence of any training data. A significant improvement in pairing performance over random assignment was detected in 16/22 tested epitopes using at least one of the three methods (Figure 7).

In general, these methods show better performance in the pairing task than existing models for epitope prediction (Meynard-Piganeau et al., 2024) or TCR sequence embedding (Nagano et al., 2025), suggesting that they are learning an alternative signal for prediction. However, pairing performance remains poor, limiting practical applications for TCR pairing. The addition of more pairs for each epitope is likely to increase performance, as information-based methods are data-thirsty (Bitbol, 2018). The current algorithms are also computationally intensive, limiting their application to large epitope-specific TCR sets. Faster methods have been developed (see for instance, Lupo et al., 2025) and can be tested on larger datasets in the future. TCR pairing may also be improved by including additional information such as sequence abundance, as $\alpha$ and $\beta$ chains from the same TCR will vary across different samples in a correlated manner. Finally, the precision calculation assumes all unobserved pairs in the data to be non-binders, but this assumption may be incorrect.

Epitopes differed widely in the amount of MI shared between $\alpha$ and $\beta$ CDR sequences and in the performance achieved by the pairing algorithms (Figures 6 and 8). This variation may arise from intrinsic features of the pMHCs, and features of the TCR repertoire associated with each pMHC. The feature-ness of a peptide (how much of the peptide extends outside the MHC binding groove), for example, has been shown to impact the diversity of the associated TCR repertoire (Turner et al., 2006, 2005). However, no correlation was detected between feature-ness and MI in this study (Figure 6). In contrast, we found that pMHCs with smaller effective set size of associated TCRs have higher inter-chain MI and better pairing prediction (Figure 8). The presence of fewer, larger clusters of similar TCR sequences in these sets may allow better prediction of $\alpha\beta$ binding.

The study has a number of limitations. The total number of crystallographic structures is still only a tiny fraction of the number

of possible TCRs. Similarly, the number of pMHC annotated TCRs in the VDJDb database is only a tiny fraction of possible TCR/pMHC pairings. Furthermore, whilst the analysis was not deliberately limited to specific MHC genes, the VDJDb dataset is known to be skewed and only a small fraction of all MHC alleles and peptide antigens are represented Hudson et al. (2023). Moreover, the MHC information as reported by VDJDb Goncharov et al. (2022); Bagaev et al. (2020) occasionally lacks allele variant information. As detailed above, the number of pMHC/TCR sets was further limited because of the inclusion criteria used for the analyses (more than 100 but less than 10,000 TCRs associated with an epitope) to limit the bias in MI calculations for very small TCR sets, and the computational limits in analysis of very large datasets. These criteria restricted the analysed HLA alleles to few alleles from HLA-A and HLA-B and mouse MHCs (Table S1). While we expect this process to be relevant also for CD4 T cell binding to class II MHC, as exemplified by the signal in CD4+AIM+ populations from OTS (Figure S10B), the epitope-specific data from VDJdb has only allowed us to test our hypothesis on CD8 T cells. VDJDb is also known to include erroneously-annotated sequences, which will add noise to the MI calculation (Messemaker et al.). The estimation of inter-sequence MI is also subject to some technical limitations, such as finite size effects, padding artefacts, heterogeneity in sequencing platforms and other batch effects between experiments. Although every effort was made to control for these factors, some residual effect cannot be excluded. Finally, MI may not capture higher-order interactions involving epistatic interactions between multiple amino acid positions.

In conclusion, our study documents extensive conserved interactions between the CDR sequences which form the binding surface of a TCR. Inter-chain contacts can stabilise the conformational flexibility of the loops, reducing the entropy of the free TCR, and increasing affinity and specificity at the expense of reduced cross-reactivity. In addition, we quantified broader sequence constraints on TCR$\alpha\beta$ pairing in the context of antigen-specific TCR sets. These constraints may be imposed directly by the interchain contacts, or through indirect structural constraints imposed on chain interactions in the context of binding to a pMHC. Our study provides a bioinformatic foundation on which to develop functional and structural studies investigating the role of inter-CDR interactions in determining the kinetics and thermodynamics of TCR-antigen binding. Tuning inter-CDR interactions may represent an under-appreciated mechanism to optimise the design of TCRs and TCR analogues for therapeutic purposes, and may also have broader implications for predicting TCR/pMHC binding.

## Conflicts of interest

The authors declare that they have no competing interests.

## Funding

This work was supported by Cancer Research UK through a Non-Clinical Training Award to M.M. [A29287] and by grants from the Rosetrees Foundation and the UCLH Biomedical Research Centre to B.C.. Y.N. was supported by the Cancer Research UK City of London Centre (grant number BCCG1C8R). U.H. was supported by NIAID program project grant P01 AI106697 and the European Union's Horizon 2020 Research and Innovation Program under grant agreement 825821 and by Israel Science Foundation (ISF) grant 1327/22. The work of A. T.-M. was supported in parts by the Royal Free Charity. A.-F. B. thanks the European Research Council (ERC) for funding under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851173, to A.-F. B.).

## Data availability

No new data was generated with this paper. The data analysed can be downloaded from VDJDb (`https://vdjdb.cdr3.net/`), the Structural T Cell Receptor Database (STCRDab, `https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/`) and the Observed TCR Space database (OTS, `https://opig.stats.ox.ac.uk/webapps/ots/`). Processed data from Tanno et al. (2020) is available from: `https://zenodo.org/records/14288119`. The code needed to reproduce the analysis in this manuscript is available at `https://github.com/mm523/TCRab-pairing` and is deposited on Zenodo `https://doi.org/10.5281/zenodo.21390616`.

## Author contributions statement

M.M. and B.C. conceived the study. M.M., Y.N., J.H., U.H., A. T.-M., A.-F. B. and B.C. developed the methodology. M.M. wrote the code with contributions from Y.N.. M.M. carried out the analyses with contributions from Y.N., J.H., U.H., A. T.-M., A.-F. B. and B.C.. M.M., Y.N., J.H., U.H., A. T.-M., A.-F. B. and B.C. interpreted the results. M.M. and B.C. wrote the first version of the manuscript, and M.M., Y.N., J.H., U.H., A. T.-M., A.-F. B. and B.C. reviewed and edited it.

## Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. A version of the figures and results in this work were previously reported in M.M.'s PhD thesis (Milighetti, 2023). The authors acknowledge the use of the UCL Myriad High Performance Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

## References

J. K. Archbold, W. A. Macdonald, S. Gras, L. K. Ely, J. J. Miles, M. J. Bell, R. M. Brennan, T. Beddoe, M. C. Wilce, C. S. Clements, A. W. Purcell, J. McCluskey, S. R. Burrows, and J. Rossjohn. Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. *Journal of Experimental Medicine*, 206(1):209–219, Jan. 2009. ISSN 0022-1007. doi: 10.1084/JEM.20082136. Publisher: The Rockefeller University Press.

D. V. Bagaev, R. M. Vroomans, J. Samir, U. Stervbo, C. Rius, G. Dolton, A. Greenshields-Watson, M. Attaf, E. S. Egorov, I. V. Zvyagin, N. Babel, D. K. Cole, A. J. Godkin, A. K. Sewell, C. Kesmir, D. M. Chudakov, F. Luciani, and M. Shugay. VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062, Jan. 2020. ISSN 13624962. doi: 10.1093/nar/gkz874. Publisher: Oxford University Press.

E. Baulu, C. Gardet, N. Chuvin, and S. Depil. TCR-engineered T cell therapy in solid tumors: State of the art and perspectives. *Science Advances*, 9(7):eadf3700, Feb. 2023. doi: 10.1126/sciadv.adf3700. Publisher: American Association for the Advancement of Science.

A.-F. Bitbol. Inferring interaction partners from protein sequences using mutual information. *PLOS Computational Biology*, 14 (11):e1006401, Nov. 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006401.

A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43):12180–12185, Apr. 2016. ISSN 10916490. doi: 10.1073/pnas.1606762113. arXiv: 1604.08354 Publisher: National Academy of Sciences.

A. Bovay, V. Zoete, G. Dolton, A. M. Bulek, D. K. Cole, P. J. Rizkallah, A. Fuller, K. Beck, O. Michielin, D. E. Speiser, A. K. Sewell, and S. A. Fuertes Marraco. T cell receptor alpha variable 12-2 bias in the immunodominant response to Yellow fever virus. *European journal of immunology*, 48(2):258–272, Feb. 2018. ISSN 1521-4141. doi: 10.1002/eji.201747082. Publisher: Wiley-VCH Verlag.

S. Bradde, A. Braunstein, H. Mahmoudi, F. Tria, M. Weigt, and R. Zecchina. Aligning graphs and finding substructures by a cavity approach. *EPL (Europhysics Letters)*, 89(3):37009, Feb. 2010. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/89/37009. URL `http://arxiv.org/abs/0905.1893`. arXiv:0905.1893 [cond-mat, q-bio].

D. Campillo-Davo, D. Flumens, and E. Lion. The Quest for the Best: How TCR Affinity, Avidity, and Functional Avidity Affect TCR-Engineered T-Cell Antitumor Responses. *Cells*, 9(7):1720, July 2020. ISSN 2073-4409. doi: 10.3390/cells9071720.

J. A. Carter, J. B. Preall, K. Grigaityte, S. J. Goldfless, E. Jeffery, A. W. Briggs, F. Vigneault, and G. S. Atwal. Single T Cell Sequencing Demonstrates the Functional Role of $\alpha\beta$ TCR Pairing in Cell Lineage and Antigen Specificity. *Frontiers in Immunology*, 10:1516, July 2019. ISSN 1664-3224. doi: 10.3389/fimmu.2019.01516. Publisher: Frontiers.

P. Chaurasia, T. H. Nguyen, L. C. Rowntree, J. A. Juno, A. K. Wheatley, S. J. Kent, K. Kedzierska, J. Rossjohn, and J. Petersen. Structural basis of biased T cell receptor recognition of an immunodominant HLA-A2 epitope of the SARS-CoV-2 spike protein. *Journal of Biological Chemistry*, 297(3):101065, Sept. 2021. ISSN 00219258. doi: 10.1016/j.jbc.2021.101065. Publisher: Elsevier.

P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, and P. G. Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661):89–93, July 2017. ISSN 0028-0836. doi: 10.1038/nature22383. Publisher: Nature Publishing Group.

M. M. Davis and P. J. Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402, Aug. 1988. ISSN 1476-4687. doi: 10.1038/334395a0. Number: 6181 Publisher: Nature Publishing Group.

D. De Juan, F. Pazos, and A. Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, Apr. 2013. ISSN 14710056. doi: 10.1038/nrg3414.

L. Deng, R. J. Langley, P. H. Brown, G. Xu, L. Teng, Q. Wang, M. I. Gonzales, G. G. Callender, M. I. Nishimura, S. L. Topalian, and R. A. Mariuzza. Structural basis for the recognition of mutant self by a tumor-specific, MHC class II–restricted T cell receptor. *Nature Immunology 2007 8:4*, 8 (4):398–408, Mar. 2007. ISSN 1529-2916. doi: 10.1038/ni1447. Publisher: Nature Publishing Group.

J. Dunbar and C. M. Deane. ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics*, 32 (2):298–300, Jan. 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btv552. Publisher: Oxford University Press.

S. Dunn, L. Wahl, and G. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb. 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm604.

T. Dupic, Q. Marcou, A. M. Walczak, and T. Mora. Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Computational Biology*, 15(3): e1006874, Mar. 2019. ISSN 15537358. doi: 10.1371/journal.pcbi.1006874. Publisher: Public Library of Science.

K. J. Fryxell. The coevolution of gene family trees. *Trends in Genetics*, 12(9):364–369, Sept. 1996. ISSN 0168-9525. doi: 10.1016/S0168-9525(96)80020-5.

M. Galperin, C. Farenc, M. Mukhopadhyay, D. Jayasinghe, A. Decroos, D. Benati, L. L. Tan, L. Ciacchi, H. H. Reid, J. Rossjohn, L. A. Chakrabarti, and S. Gras. CD4+ T cell–mediated HLA class II cross-restriction in HIV controllers. *Science Immunology*, 3(24):687, June 2018. ISSN 24709468. doi: 10.1126/SCIIMMUNOL.AAT0687. Publisher: American Association for the Advancement of Science.

C. A. Gandarilla-Pérez, S. Pinilla, A.-F. Bitbol, and M. Weigt. Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins. *PLOS Computational Biology*, 19(3):e1011010, Mar. 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011010. Publisher: Public Library of Science.

D. N. Garboczi, P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison, and D. C. Wiley. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature*, 384(6605): 134–141, Nov. 1996. ISSN 00280836. doi: 10.1038/384134a0.

J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, and M. M. Davis. Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661):94–98, July 2017. ISSN 0028-0836. doi: 10.1038/nature22976. Publisher: Nature Publishing Group.

M. Goncharov, D. Bagaev, D. Shcherbinin, I. Zvyagin, D. Bolotin, P. G. Thomas, A. A. Minervina, M. V. Pogorelyy, K. Ladell, J. E. McLaren, D. A. Price, T. H. Nguyen, L. C. Rowntree, E. B. Clemens, K. Kedzierska, G. Dolton, C. R. Rius, A. Sewell, J. Samir, F. Luciani, K. V. Zornikova, A. A. Khmelevskaya, S. A. Sheetikov, G. A. Efimov, D. Chudakov, and M. Shugay. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nature Methods 2022 19:9*, 19 (9):1017–1019, Aug. 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01578-0. Publisher: Nature Publishing Group.

S. Gras, X. Saulquin, J.-B. Reiser, E. Debeaupuis, K. Echassérieau, A. Kissenpfennig, F. Legoux, A. Chouquet, M. Le Gorrec, P. Machillot, B. Neveu, N. Thielens, B. Malissen, M. Bonneville, and D. Housset. Structural Bases for the Affinity-Driven Selection of a Public TCR against a

Dominant Human Cytomegalovirus Epitope. *The Journal of Immunology*, 183(1):430–437, July 2009. ISSN 0022-1767. doi: 10.4049/jimmunol.0900556. Publisher: American Association of Immunologists.

T. Hamelryck and B. Manderick. PDB file parser and structure class implemented in Python. *Bioinformatics*, 19(17):2308–2310, Nov. 2003. ISSN 13674803. doi: 10.1093/bioinformatics/btg299. Publisher: Oxford Academic.

J. M. Heather, M. J. Spindler, M. H. Alonso, Y. I. Shui, D. G. Millar, D. S. Johnson, M. Cobbold, and A. N. Hata. Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. *Nucleic Acids Research*, 50(12):e68, July 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac190.

J. Henderson, Y. Nagano, M. Milighetti, and A. Tiffeau-Mayer. Limits on inferring t cell specificity from partial information. *Proceedings of the National Academy of Sciences*, 121(42):e2408696121. doi: 10.1073/pnas.2408696121. Publisher: Proceedings of the National Academy of Sciences.

D. Hudson, R. A. Fernandes, M. Basham, G. Ogg, and H. Koohy. Can we predict T cell specificity with digital biology and machine learning? *Nature Reviews Immunology*, pages 1–11, Feb. 2023. ISSN 1474-1741. doi: 10.1038/s41577-023-00835-3. Publisher: Nature Publishing Group.

J. Ishizuka, G. B. Stewart-Jones, A. van der Merwe, J. I. Bell, A. J. McMichael, and E. Y. Jones. The Structural Dynamics and Energetics of an Immunodominant T Cell Receptor Are Programmed by Its V$\beta$ Domain. *Immunity*, 28(2):171–182, Feb. 2008. ISSN 10747613. doi: 10.1016/j.immuni.2007.12.018.

D. B. Jaffe, P. Shahi, B. A. Adams, A. M. Chrisman, P. M. Finnegan, N. Raman, A. E. Royall, F. Tsai, T. Vollbrecht, D. S. Reyes, N. L. Hepler, and W. J. McDonnell. Functional antibodies exhibit light chain coherence. *Nature*, 611(7935):352–357, Nov. 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-05371-z. Publisher: Nature Publishing Group.

C. Keşmir, J. A. Borghans, and R. J. de Boer. Diversity of Human $\alpha\beta$ T Cell Receptors. *Science*, 288(5469):1135–1135, May 2000. doi: 10.1126/science.288.5469.1135a. Publisher: American Association for the Advancement of Science.

B. T. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 90(15):7176–7180, Aug. 1993. ISSN 1091-6490. doi: 10.1073/pnas.90.15.7176.

N. L. La Gruta, P. G. Thomas, A. I. Webb, M. A. Dunstone, T. Cukalac, P. C. Doherty, A. W. Purcell, J. Rossjohn, and S. J. Turner. Epitope-specific TCRbeta repertoire diversity imparts no functional advantage on the CD8+ T cell response to cognate viral peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):2034–9, Feb. 2008. ISSN 1091-6490. doi: 10.1073/pnas.0711682102. Publisher: National Academy of Sciences.

J. Leem, S. H. P. de Oliveira, K. Krawczyk, and C. M. Deane. STCRDab: the structural T-cell receptor database. *Nucleic Acids Research*, 46(D1):D406–D412, Jan. 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx971.

M.-P. Lefranc, C. Pommié, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental & Comparative Immunology*, 27(1):55–77, Jan. 2003. ISSN 0145305X. doi: 10.1016/S0145-305X(02)00039-3. Publisher: Pergamon.

K. E. Lineburg, E. J. Grant, S. Swaminathan, D. S. Chatzileontiadou, C. Szeto, H. Sloane, A. Panikkar, J. Raju, P. Crooks, S. Rehan, A. T. Nguyen, L. Lekieffre, M. A. Neller, Z. W. M. Tong, D. Jayasinghe, K. Y. Chew, C. A. Lobos, H. Halim, J. M. Burrows, A. Riboldi-Tunnicliffe, W. Chen, L. D'Orsogna, R. Khanna, K. R. Short, C. Smith, and S. Gras. CD8+ T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope cross-react with selective seasonal coronaviruses. *Immunity*, 54(5):1055–1065.e5, May 2021. ISSN 10747613. doi: 10.1016/j.immuni.2021.04.006. Publisher: Cell Press.

U. Lupo, D. Sgarbossa, M. Milighetti, and A.-F. Bitbol. DiffPaSS—high-performance differentiable pairing of protein sequences using soft scores. *Bioinformatics*, 41(1):btae738, Jan. 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae738. URL `https://doi.org/10.1093/bioinformatics/btae738`.

A. Mayer and C. G. Callan. Measures of epitope binding degeneracy from T cell receptor repertoires. *Proceedings of the National Academy of Sciences*, 120(4):e2213264120, Jan. 2023. ISSN 0027-8424. doi: 10.1073/pnas.2213264120. Publisher: National Academy of Sciences ISBN: 2213264120.

C. McBeth, A. Seamons, J. C. Pizarro, S. J. Fleishman, D. Baker, T. Kortemme, J. M. Goverman, and R. K. Strong. A new twist in TCR diversity revealed by a forbidden alphabeta TCR. *Journal of molecular biology*, 375(5):1306–19, Feb. 2008. ISSN 1089-8638. doi: 10.1016/j.jmb.2007.11.020. Publisher: Academic Press.

R. Meijers, C.-C. Lai, Y. Yang, J.-h. Liu, W. Zhong, J.-h. Wang, and E. L. Reinherz. Crystal Structures of Murine MHC Class I H-2 Db and Kb Molecules in Complex with CTL Epitopes from Influenza A Virus: Implications for TCR Repertoire Selection and Immunodominance. *Journal of Molecular Biology*, 345(5):1099–1110, Feb. 2005. ISSN 00222836. doi: 10.1016/j.jmb.2004.11.023. Publisher: Academic Press.

M. Messemaker, B. P. Y. Kwee, v. Moravec, D. Álvarez Salmoral, J. Urbanus, S. d. Paauw, J. Geerligs, R. Voogd, B. Morris, A. Guislain, M. Mußmann, Y. Winkler, M. Steinmetz, M. Iras, E. Marcus, J. Teuwen, A. Perrakis, R. L. Beijersbergen, W. Scheper, and T. N. Schumacher. A functionally validated TCR-pMHC database for TCR specificity model development. URL `https://www.biorxiv.org/content/10.1101/2025.04.28.651095v3`. Pages: 2025.04.28.651095 Section: New Results.

B. Meynard-Piganeau, C. Feinauer, M. Weigt, A. M. Walczak, and T. Mora. TULIP: A transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proceedings of the National Academy of Sciences*, 121(24):e2316401121, June 2024. doi: 10.1073/pnas.2316401121. Publisher: Proceedings of the National Academy of Sciences.

M. Milighetti. *Analysis of T cell receptor sequence and structure to understand the drivers of antigen specificity*. PhD thesis, UCL (University College London), Nov. 2023.

M. Milighetti, J. Shawe-Taylor, and B. Chain. Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-Major Histocompatibility Complexes. *Frontiers in Physiology*, 12:2021.05.19.444843, Sept. 2021. ISSN 1664-042X. doi: 10.3389/

fphys.2021.730908.

T. Mora and A. M. Walczak. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology*, 18:104–110, Dec. 2019. ISSN 2452-3100. doi: 10.1016/J.COISB.2019.10.001. arXiv: 1907.08230 Publisher: Elsevier.

Y. Nagano and B. Chain. tidytcells: standardizer for TR/MH nomenclature. *Frontiers in Immunology*, 14, 2023. ISSN 1664-3224.

Y. Nagano, A. G. T. Pyo, M. Milighetti, J. Henderson, J. Shawe-Taylor, B. Chain, and A. Tiffeau-Mayer. Contrastive learning of T cell receptor representations. *Cell Systems*, 16(1):101165, Jan. 2025. ISSN 2405-4712. doi: 10.1016/j.cels.2024.12.006.

D. Ochoa and F. Pazos. Practical aspects of protein co-evolution. *Frontiers in Cell and Developmental Biology*, 2, 2014. ISSN 2296-634X.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, Jan. 2012. ISSN 1533-7928. arXiv: 1201.0490.

M. V. Pogorelyy, E. Rosati, A. A. Minervina, R. C. Mettelman, A. Scheffold, A. Franke, P. Bacher, and P. G. Thomas. Resolving SARS-CoV-2 CD4+ T cell specificity via reverse epitope discovery. *Cell Reports Medicine*, 3(8):100697, Aug. 2022. ISSN 26663791. doi: 10.1016/j.xcrm.2022.100697. Publisher: Elsevier.

M. I. J. Raybould, A. Greenshields-Watson, P. Agarwal, B. Aguilar-Sanjuan, T. H. Olsen, O. M. Turnbull, N. P. Quast, and C. M. Deane. The Observed T Cell Receptor Space database enables paired-chain repertoire mining, coherence analysis, and language modeling. *Cell Reports*, 43(9):114704, Sept. 2024. ISSN 2211-1247. doi: 10.1016/j.celrep.2024.114704.

J.-B. Reiser, F. Legoux, S. Gras, E. Trudel, A. Chouquet, A. Léger, M. Le Gorrec, P. Machillot, M. Bonneville, X. Saulquin, and D. Housset. Analysis of Relationships between Peptide/MHC Structural Features and Naive T Cell Frequency in Humans. *The Journal of Immunology*, 193(12):5816–5826, Dec. 2014. ISSN 0022-1767. doi: 10.4049/JIMMUNOL.1303084. Publisher: American Association of Immunologists.

R. A. Robinson, C. McMurran, M. L. McCully, and D. K. Cole. Engineering soluble T-cell receptors for therapy. *The FEBS Journal*, 288(21):6159–6173, 2021. ISSN 1742-4658. doi: 10.1111/febs.15780.

Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.5. Nov. 2015.

G. W. Schwartz and U. Hershberg. Conserved variation: Identifying patterns of stability and variability in BCR and TCR v genes with different diversity and richness metrics. *Physical Biology*, 10(3), June 2013. ISSN 14783967. doi: 10.1088/1478-3975/10/3/035005.

D. R. Scott, O. Y. Borbulevych, K. H. Piepenbrink, S. A. Corcelli, and B. 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, Dec. 2011. ISSN 0022-2836. doi: 10.1016/J.JMB.2011.10.006. Publisher: Academic Press.

A. K. Sewell. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, Sept. 2012. ISSN 1474-1741. doi: 10.1038/nri3279. Number: 9 Publisher: Nature Publishing Group.

D. S. Shcherbinin, V. A. Belousov, and M. Shugay. Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCR$\alpha\beta$ complex. *PLOS Computational Biology*, 16(3):e1007714, Mar. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007714. Publisher: Public Library of Science.

P. Sliz, O. Michielin, J.-C. Cerottini, I. Luescher, P. Romero, M. Karplus, and D. C. Wiley. Crystal Structures of Two Closely Related but Antigenically Distinct HLA-A2/Melanocyte-Melanoma Tumor-Antigen Peptide Complexes. *The Journal of Immunology*, 167(6):3276–3284, Sept. 2001. ISSN 0022-1767. doi: 10.4049/JIMMUNOL.167.6.3276. Publisher: American Association of Immunologists.

I. Springer, N. Tickotsky, and Y. Louzoun. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, 0: 1436, Apr. 2021. ISSN 1664-3224. doi: 10.3389/FIMMU.2021.664514. Publisher: Frontiers.

C. Szeto, C. A. Lobos, A. T. Nguyen, and S. Gras. TCR Recognition of Peptide–MHC-I: Rule Makers and Breakers. *International Journal of Molecular Sciences*, 22(1):68, Dec. 2020. ISSN 1422-0067. doi: 10.3390/ijms22010068. Publisher: Multidisciplinary Digital Publishing Institute (MDPI).

H. Tanno, T. M. Gould, J. R. McDaniel, W. Cao, Y. Tanno, R. E. Durrett, D. Park, S. J. Cate, W. H. Hildebrand, C. L. Dekker, L. Tian, C. M. Weyand, G. Georgiou, and J. J. Goronzy. Determinants governing T cell receptor $\alpha/\beta$-chain pairing in repertoire formation of identical twins. *Proceedings of the National Academy of Sciences*, 117(1):532–540, Jan. 2020. ISSN 0027-8424. doi: 10.1073/pnas.1915008117. Publisher: National Academy of Sciences.

S. J. Turner, K. Kedzierska, H. Komodromou, N. L. La Gruta, M. A. Dunstone, A. I. Webb, R. Webby, H. Walden, W. Xie, J. McCluskey, A. W. Purcell, J. Rossjohn, and P. C. Doherty. Lack of prominent peptide–major histocompatibility complex features limits repertoire diversity in virus-specific CD8+ T cell populations. *Nature Immunology*, 6(4):382–389, Apr. 2005. ISSN 1529-2908. doi: 10.1038/ni1175. Publisher: Nature Publishing Group.

S. J. Turner, P. C. Doherty, J. McCluskey, and J. Rossjohn. Structural determinants of T-cell receptor bias in immunity. *Nature Reviews Immunology*, 6(12):883–894, Dec. 2006. ISSN 1474-1733. doi: 10.1038/nri1977. Publisher: Nature Publishing Group.

S. A. Valkenburg, S. Gras, C. Guillonneau, L. A. Hatton, N. A. Bird, K. A. Twist, H. Halim, D. C. Jackson, A. W. Purcell, S. J. Turner, P. C. Doherty, J. Rossjohn, and K. Kedzierska. Preemptive priming readily overcomes structure-based mechanisms of virus escape. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5570–5575, Apr. 2013. ISSN 00278424. doi: 10.1073/PNAS.1302935110. Publisher: National Academy of Sciences.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, Jan. 2009. doi: 10.1073/pnas.0805923106. Publisher: Proceedings of the National Academy of Sciences.

W. Zhong, S. B. Dixit, R. J. Mallis, H. Arthanari, A. A. Lugovskoy, D. L. Beveridge, G. Wagner, and E. L. Reinherz. CTL Recognition of a Protective Immunodominant Influenza A Virus Nucleoprotein Epitope Utilizes a Highly Restricted V$\beta$ but Diverse V$\alpha$ Repertoire: Functional and Structural Implications. *Journal of Molecular Biology*, 372(2):535–548, Sept. 2007. ISSN 0022-2836. doi: 10.1016/J.JMB.2007.06.057. Publisher: Academic Press.