Intra- and inter-chain contacts determine TCR specificity: applying protein co-evolution methods to TCR$\alpha\beta$ pairing

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
[2]Cancer Institute, University College London
[3]Division of Medicine, University College London
[4]Institute for the Physics of Living Systems, University College London
[5]Department of Human Biology, University of Haifa
[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

Abstract

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 importance of contacts between the CDRs themselves remains poorly understood. With a systematic study of over 200 unique TCR structures, we identify consistent intra and inter-chain CDR contact zones. We hypothesise that these interactions may restrict TCRα/TCRβ pairing within epitope-specific repertoires. Indeed, we show that the sequences of paired TCRα and TCRβ are not independent within the repertoires of TCRs specific for most epitopes examined. We show that this sequence restriction can be quantified using a mutual information framework, can be learnt by co-evolution models without using a training set of known pairs and allows *de novo* predictions of TCRα/TCRβ pairing.

Introduction

The $\alpha\beta$ T cell receptor (TCR) is a heterodimeric membrane protein which recognises peptides bound to major histocompatibility complex (MHC) molecules (pMHC). Each chain in the dimer is generated by a process of somatic recombination between V, J and (for TCRβ) D genes, which generates an enormous diversity of TCR sequences (Davis & Bjorkman, 1988; Keşmir et al., 2000; Mora & Walczak, 2019). The complementarity-determining regions (CDRs or CDR1, CDR2 and CDR3) are the most variable region of each TCR chain, and make contact with the cognate pMHC. CDR1 and CDR2 are encoded by the germline V gene, while CDR3 contains the recombination junctions between V, J and D genes and is consequently the most diverse (Schwartz & Hershberg, 2013). TCRs have a conserved docking mode on pMHC, with CDR3 making the most contacts with the peptide, whilst CDR1 and CDR2 mostly contact the MHC

(Szeto et al., 2020; Milighetti et al., 2021). Notably, thousands of TCRs of different sequence can bind to the same pMHC (Dash et al., 2017; Glanville et al., 2017) and a single TCR may also bind to many pMHCs (Sewell, 2012).

Despite the heterodimeric nature of the receptor, and the evident contribution of both chains to binding (Szeto et al., 2020; Milighetti et al., 2021), most studies consider the TCRα and β sequences independently. Therefore, the extent to which TCR diversity is constrained by TCR$\alpha\beta$ interactions remains unclear. At a whole repertoire level, only very weak or no constraints on pairing of TCRα and TCRβ sequences have been reported (Dupic et al., 2019; Yu et al., 2019; Shcherbinin et al., 2020). The few studies of TCRα and TCRβ pairing within a set of TCRs which all share the same pMHC specificity do not provide a consistent picture. Shcherbinin et al., 2020 detected no constraints on TCR$\alpha\beta$ pairing in PBMCs or in epitope-specific repertoires, but showed that the residues at the α/β interface can affect the relative orientation of the two TCR chains. However, 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 & Callan, 2023; Henderson et al., 2024). Experimental studies have also shown that interactions between CDR loops might be important for epitope specificity. For instance, McBeth et al., 2008 showed that residues in the CDR3s comprise over 30% of the α/β interface, and affect the inter-domain angle between the two TCR chains and therefore pMHC binding. Therefore, whilst there are few or no constraints on which TCRα can pair with which TCRβ at a protein-folding level, thymic and post-thymic antigen selection may restrict α/β pairing. Moreover, CDR1 and CDR2 may also contribute to antigen binding by directing loop conformations, as single mutations in these loops can lead to loss of antigen binding (Gras et al., 2010), and the effect of mutation is dependent on the CDR3 context (Stadinski et al., 2014). Overall, these studies suggest that interactions between the CDR loops influence antigen binding. Importantly, these interactions may impact affinity for antigen or TCR cross-reactivity, and may therefore constitute an important consideration in the design and optimisation of TCR-engineered T cells (Campillo-Davo et al., 2020; Baulu et al., 2023) or soluble TCRs (Robinson et al., 2021) for therapeutic applications. Indeed, it is known that amino acids at the interface between TCRα and TCRβ can modulate expression and stability of transduced TCRs (Thomas et al., 2019), but their impact on TCR affinity and specificity remains to be resolved.

We hypothesise that, since the TCR binding surface is formed by all 6 loops coming together, a change in any one of the loops may require compensatory changes in the other loops for antigen binding to be preserved. To explore this hypothesis, we first systematically document the interactions that the hypervariable loops make with each other. We then quantify the constraint imposed by chain pairing and V gene selection by sequence similarity and mutual information. Finally, we show that the TCR$\alpha\beta$ pairing signal within epitope-specific repertoires can be learnt by co-evolution based models.

Methods

Analysis of existing crystal structures

A list of existing crystal structures for TCRs and TCR-pMHC complexes was obtained from The Structural T-Cell Receptor Database (STCRDab, Leem et al., 2018, downloaded on 27th February 2023). The dataset was curated to include only $\alpha\beta$TCRs, either unbound or bound to class I or class II MHC. Both mouse and human TCRs were included. Structures annotated as bound for which the IMGT-renumbered file available from STCRDab had missing epitope information were removed, as well as epitopes with special groups, structures that dock with reverse polarity and TCRs binding superantigens and enterotoxins. The complete set included 529 unique TCR structures. The IMGT-renumbered sequence was then extracted from each TCR, and TCRs with identical sequences for both TCRα and TCRβ chains were grouped together. This further reduced the set to 214 unique receptors.

Structures were visualised using Pymol Open Source v2.5.0 (Schrödinger, LLC, 2015). Intra-chain and inter-chain distances between Cα along the chains were calculated using the BioPDB package in Python v3.11 (Hamelryck & Manderick, 2003).

Paired-chain datasets

The entire VDJDb database (Goncharov et al., 2022) was exported on 3rd February 2023. It was refined to only include TCRs for which both TCRα and TCRβ chains are available, and to include unique TCR-epitope pairs. All epitopes for which fewer than 100 or more than 10,000 sequences are available were removed from the set. The final set of sequences was used for all analysis as described in the main text (Table S1).

Pre-processed data from Tanno et al., 2020 was obtained from the authors, and further processed as in Mayer and Callan, 2023. The clonotypes from sample A1, sorted naïve cells were used as a control of unselected repertoire.

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 & Deane, 2016) was used to obtain CDR3 sequences from each chain, renumbered to include standard IMGT gaps (Lefranc et al., 2003). To ensure that CDR3 sequences were all the same length for mutual information calculations, a length of 19 residues was enforced. Sequences longer than 19 residues were discarded (Figure S1), corresponding to $< 1\%$ (67 of 9852) of all CDR3$\alpha\beta$. Where the sequences were shorter than 19, additional gaps were added at position $\frac{L}{2}$ (where L is the length of the CDR3 sequence) when calculating mutual information.

Sequence clustering based on triplet similarity

Sequence similarity between CDR3s was assessed by normalised triplet similarity as previously described (Joshi et al., 2019). Briefly, each CDR3 is decomposed into sets of overlapping triplets. The number of triplets shared between two CDR3s is counted and normalised

by $\sqrt{(L_1-1)(L_2-1)}$, where L_1 and L_2 are the lengths of the two CDR3s. Triplet similarity is calculated by using the stringdot function (norm = TRUE) of the kernlab package (Karatzoglou et al., 2004). The similarity function is called in Python by using package rpy2 (https://rpy2.github.io/) and networks are plotted using python-igraph (https://python.igraph.org/en/stable/, Csárdi & Nepusz, 2006).

Triplet similarity was calculated on CDR3 sequences which did not include gaps. Sequence similarity graphs were defined by thresholding triplet similarity at a value of 0.76 and 0.72 for α and β respectively. The thresholds were set such that 99.99% of pairwise distances between CDR3 sequences not recognising the same epitope are below the threshold.

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 regions. MI is a statistical measure of the dependence of two random variables, i.e. of how much information each variable carries about the other. The MI between two regions where variability is moderate, such as V or J genes, can readily be estimated by using the observed frequencies of each gene in the set, together with their observed frequencies of co-occurrence. However, estimating the MI between one such region and a CDR3 sequence, or between two CDR3 sequences, is not as straight-forward. Indeed, most sequences occur only once (Table S1) and the data is thus insufficient to estimate the entropy of CDR3 sequences as a whole (and thus the MI involving them). To circumvent this issue, we focused on pairwise approximations of the MI involving CDR3 sequences, where the amino acid sequence of the CDR3 region is considered one residue at a time. The MI between a region of moderate variability and each single residue position of the CDR3 region was calculated, and then summed to obtain a pairwise approximation of the total MI between these two regions. This is an approach that has been successfully used to study protein-protein interactions and co-evolution (Dunn et al., 2008; Skerker et al., 2008; Bitbol, 2018; for an alternative approach using unbiased estimators for Renyi-Simpson entropies see Tiffeau-Mayer, 2023 and Henderson et al., 2024). Concretely, the MI between a moderately variable region X and a CDR3 sequence S was approximated by:

$$I(X;S) = \sum_{p=1}^{L} \sum_{r \in R} \sum_{x \in \Omega_X} f_{X,p}(x,r) \log \left(\frac{f_{X,p}(x,r)}{f_X(x) f_p(r)} \right) , \tag{1}$$

where f denotes a frequency observed in the data, L is the length of the CDR3 amino acid sequence of interest (we enforced $L = 19$ as described above), r is a residue at position p in sequence S, R is the set of possible residues (i.e. the 20 amino acids plus a character for a gap in alignment), and X is the random variable describing the moderately variable region of interest (e.g. the V gene), taking values x in an ensemble Ω_X. Similarly, a pairwise approximation of the MI between two CDR3 regions T and S was calculated by summing over all pairs of residue

positions involving one position in each of these two regions:

$$I(T;S) = \sum_{q=1}^{L}\sum_{p=1}^{L}\sum_{r\in R}\sum_{w\in R} f_{q,p}(w,r)\log\left(\frac{f_{q,p}(w,r)}{f_q(w)f_p(r)}\right), \tag{2}$$

where w is a residue at position q in sequence T (while r is a residue at position p in sequence S, as before).

Since plug-in estimates of MI are biased for small sample sizes (Nemenman et al., 2002; Archer et al., 2014), the 'real' value of the MI was extrapolated by taking sequential subsamples of size N [25, 35, 50, 80, 100, 150, 200, 300, 500, 1000, 1500, 2000, 2500, 3000, 5000, 10000] until sample size was reached. A regression line was then fit to the MI versus $1/N$ at each subsample and the y-intercept was then used as the MI estimate ($\lim_{N\to\infty}$) (Strong et al., 1998; Panzeri et al., 2007). The procedure was repeated for both the real set and a shuffled version of the same set (where one variable is kept constant, whilst the second variable is randomly shuffled, so that the background entropy remains the same but the relationship between the two variables is broken), to control for background, given the finite sample size. Each subsampling was repeated 10 times. Figure S2 shows the estimation for epitope GLCTLVAML as an example.

Since the CDR3 sequences were padded to obtain the alignment needed to calculate MI, we evaluated the effect of padding the sequences in the middle compared to adding padding at the end (Figure S3). Overall, the values were well correlated between the two methods (Spearman $\rho \geq 0.95$), with the exception of Ja-CDR3a and Jb-CDR3b, were the correlation was less strong. Indeed, the J gene influences the sequence of the C terminal of the CDR3, therefore a change in alignment in this region changes the estimated MI between the two.

MI was calculated using scikit-learn v1.1.3 (Pedregosa et al., 2012). To reduce the diversity in V genes and J genes, tidytcells v1.6.0 (Nagano & Chain, 2023) was used to standardise to the gene level, removing allele information.

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. It is calculated as in Weigt et al., 2009 and Bitbol et al., 2016. Briefly, for each sequence (S) in a set of size N the number of sequence neighbours within a similarity threshold are counted (m_S). S is then given a weight $w_S = \frac{1}{m_S+1}$. The total effective set size can finally be calculated as:

$$N_{\text{eff}} = \sum_{S\in 1}^{N} w_S = \sum_{S\in 1}^{N} \frac{1}{m_S+1} \tag{3}$$

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 sequences. The effective set size N_{eff} is then normalised by total repertoire size N

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.**

for comparison across repertoires. The effective set size was not correlated to total repertoire size (Figure S4).

Calculation of peptide solvent-accessible surface area

For the epitopes analysed, we retrieved the crystal structures of pMHC complexes from the Protein DataBank (PDB), when available. Solvent-accessible surface area (SASA) for each epitope was calculated using Pymol Open Source v2.5.0 (Schrödinger, LLC, 2015) using 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).

Mutual information iterative pairing algorithm (MI-IPA)

The mutual information-based iterative pairing algorithm (MI-IPA) was described in Bitbol, 2018. Briefly, the algorithm aims to correctly pair interacting proteins from two input lists based on their amino acid sequences. It does so by approximately maximising the co-evolution signal measured by mutual information. We adapted the method to pair a set of unpaired TCRα/β sequences and implemented it in Python v3.11.0. Correct pairings were known but blinded to the algorithm.

The MI-IPA is run on each epitope-specific repertoire from VDJDb, without using a training set of known pairs (as would be the case if no information about pairing was available) and inputting the CDR3 sequences only. For epitopes that have > 1000 sequences, 5 subsamples of 700 sequences were generated and paired, to reduce computational time. The original implementation of the MI-IPA used species to reduce the combinatorial number of assignments to be evaluated. As species are not available in this case, we used study and individual information as available from VDJDb as a substitute. As repeats are present both in the TCRα and TCRβ, randomness was added into the model at the assignment stage. To account for this, as well as

the influence of the initial random pairing, we ran the model 10 times in each condition.

Parameters θ (sequence diversity threshold) and λ (pseudocount) were defined as in Bitbol, 2018 and used to correct the frequency calculation for MI estimation. Step size and confidence calculation were also optimised for the dataset. We compared the Hungarian scoring method, a greedy assignment based on confidence calculation, and a 'no confidence' scenario, where the greedy assignment is performed directly on the scores assigned to each pair, without calculating a confidence score. These parameters are optimised on epitopes GLCTLVAML and YLQPRTFLL (Figure S5). We chose to run the MI-IPA with diversity threshold $\theta = 0.6$, pseudocount $\lambda = 0.6$, step size of 6 sequences and no confidence calculation. Equivalent settings but with $\lambda = 1$ were used as a negative control yielding the chance expectation (as $\lambda = 1$ prevents the algorithm from learning).

Graph alignment (GA) algorithm

The graph alignment algorithm (GA) aims to align two sequence similarity networks built on the lists of sequences to be paired (Bradde et al., 2010). First, all pairwise $\alpha - \alpha$ and $\beta - \beta$ CDR3 distances in each input list are calculated. The distances are then used to calculate the two similarity networks and the generated graphs are aligned by trying to maximise the number of overlapping edges. Gandarilla-Pérez et al., 2023 recently published an implementation of the GA, with code freely available at https://github.com/carlosgandarilla/GA-IPA. The published code was retrieved and integrated in our pipeline. No amendments were made to the code that performs the alignment. To allow integration between the GA, which was coded in Julia v1.8.5, and the existing Python v3.11.0 pipeline we used PyJulia v0.6.0 (https://pyjulia.readthedocs.io/en/latest/).

Four different distance measures were used: (1) Levenshtein (or edit) distance, (2) Weighted Levenshtein distance (where substitutions are weighted 1 and gaps are weighted $1 + \ln(4)$), (3) Triplet distance (calculated as 1-triplet similarity, as described above), (4) TCRdist (Dash et al., 2017) calculated on the CDR3 sequence only. Levenshtein and weighted Levenshtein distance are implemented in rapidfuzz (https://pypi.org/project/rapidfuzz/ Bachmann, 2024), whilst TCRdist is calculated using the implementation available in package pwseqdist (https://github.com/agartland/pwseqdist), using the parameters specified within tcrdist3 (https://tcrdist3.readthedocs.io/en/latest/, Mayer-Blackwell et al., 2021).

As for the MI-IPA, the GA was run on each epitope separately. Epitopes with > 1000 sequences were subsampled 5 times to 700 sequences to reduce computational time. The same two epitopes as for MI-IPA, GLCTLVAML and YLQPRTFLL, were used to optimise the distance metric and the number of neighbours (k, defined as in the original method) to achieve the best pairing (Figure S6). The GA was run using Levenshtein distance and $k = 20$.

Combining GA and MI-IPA (GA+MI-IPA)

We combined the GA and MI-IPA as proposed by Gandarilla-Pérez et al., 2023. In the original implementation, the GA was used to calculate a first pairing for the sequences of interest. The

stable assignments (i.e. the pairs that were consistently assigned over many iterations of the GA) were then used as a golden set for the MI-IPA (i.e. a training set of pairs that are used to initialise the MI-IPA and kept as they are throughout the IPA). Here, we explored three different ways of combining GA with MI-IPA:

(1) the most stable pairs (selected $\geq 95\%$ of the time, or the top 5 pairs, whichever is largest) from the GA are used as a golden set for the MI-IPA and removed from the set of pairs to be paired by the MI-IPA (as in Gandarilla-Pérez et al., 2023);

(2) the most stable pairs (selected $\geq 95\%$ of the time, or the top 5 pairs, whichever is largest) from the GA are used as a golden set for the MI-IPA but they are not removed from the set of pairs that the MI-IPA tries to assign. This allows the MI-IPA to correct the pairs in the final assignments;

(3) the complete consensus assignment from the GA (i.e. the TCRβ that is most often assigned for each TCRα across 100 repeats of the GA) is used as an input training set to the MI-IPA, but all α and all β are also available for re-pairing by the MI-IPA.

Note that when multiple β are chosen the same number of times for an α in the GA, the β included in the stable assignment is chosen at random. The final assignments for all pairs are then extracted and evaluated against the known pairs.

As for the MI-IPA and GA, also the GA+MI-IPA was run on each epitope separately, using 5 subsamples of 700 sequences for epitopes with > 1000 sequences. Epitopes GLCTLVAML and YLQPRTFLL were used to find the optimal combination of GA+MI-IPA (Figure S7). We chose to run the GA+MI-IPA with the combination regime described in (2) above.

Calculation of precision

To quantitatively compare pairing algorithms, we calculated the proportion of predicted pairs that are correct (precision). Note that there are repeats in the α and β chain sets which will influence this calculation (see example in Figure S8A, B). Importantly, the calculation of precision takes into account information about which individual repertoire a TCR sequence is found in. Therefore, a pair must be correct and must be found in the correct individual to be called correctly assigned. For instance, if a pair belonging to individual A but not individual B is predicted for individual B, it is marked as incorrect (Figure S8A, C).

Results

Interactions between CDR loops in existing TCR structures

To document the patterns of CDR interactions, we examined the TCR-pMHC crystal structures available from the The Structural T-Cell Receptor Database (Leem et al., 2018, STCRDab). Observation of 4 representative structures (Figure 1) shows that CDR loops form a continuous

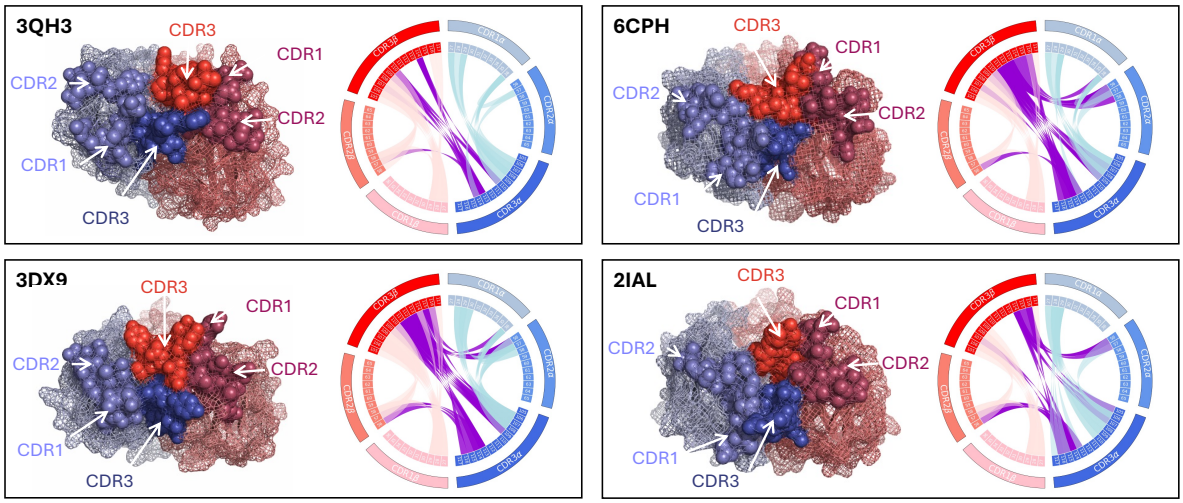

Figure 1: **CDR loops come together to form the TCR binding site in four example structures.** Positioning of CDR loops of 4 representative TCR structures (2 recognise epitope on class I MHC, 3QH3 and 3DX9; 2 recognise epitope on class II MHC, 6CPH and 2IAL). TCRα chain is in shades of blue, TCRβ is in shades of red. CDR1, 2 and 3 (brighter colour) 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 Cα distance $\leq$ 10Å). Light blue: contacts between loops on TCRα; light red: contacts between loops on TCRβ; purple: contacts between the two chains. 3QH3: Scott et al., 2011, 3DX9: Archbold et al., 2009, 6CPH: Galperin et al., 2018, 2IAL: Deng et al., 2007.

contact surface that binds pMHC. Detailed examination of the structures suggested that there are both intra- (between two residues on the same chain) and inter-chain (between one residue on the α and one residue on the β chain) contacts (right panel for each structure).

To more systematically map the interactions between TCR residues, we first measured the intra-chain distances between the Cα of each residue and the Cα on all other residues on the same chain for the four example structures (Figure 2A and B for TCRα and β, respectively). Consistently with Figure 1, we see extensive intra-chain contacts, especially between CDR1 and CDR3 and CDR1 and CDR2, both in the TCRα and in the TCRβ chain. Remarkably, the patterns are consistent across the four structures and in both chains. We extended this analysis to 214 unique TCRs (Figure 2C, D - contact is defined as distance $\leq$ 10Å between the Cα on the two residues; Duarte et al., 2010). The patterns observed in the 4 example structures are generalisable: CDR1 forms extensive contacts with CDR3, and CDR2 in turn forms contacts with CDR1. We also observe conserved contacts between the first half of CDR2 and the first half of CDR3.

Having observed interactions between the CDRs within each chain, we asked whether interactions can be observed between the CDRs in the two chains (inter-chain contacts). Figure 3A shows the inter-chain distances for the four TCRs in Figure 1. Again, the distance patterns are conserved across the four TCRs, both in the conserved framework regions and in the CDR loops. Specifically, framework region 2 (FR2), found between CDR1 and CDR2, makes

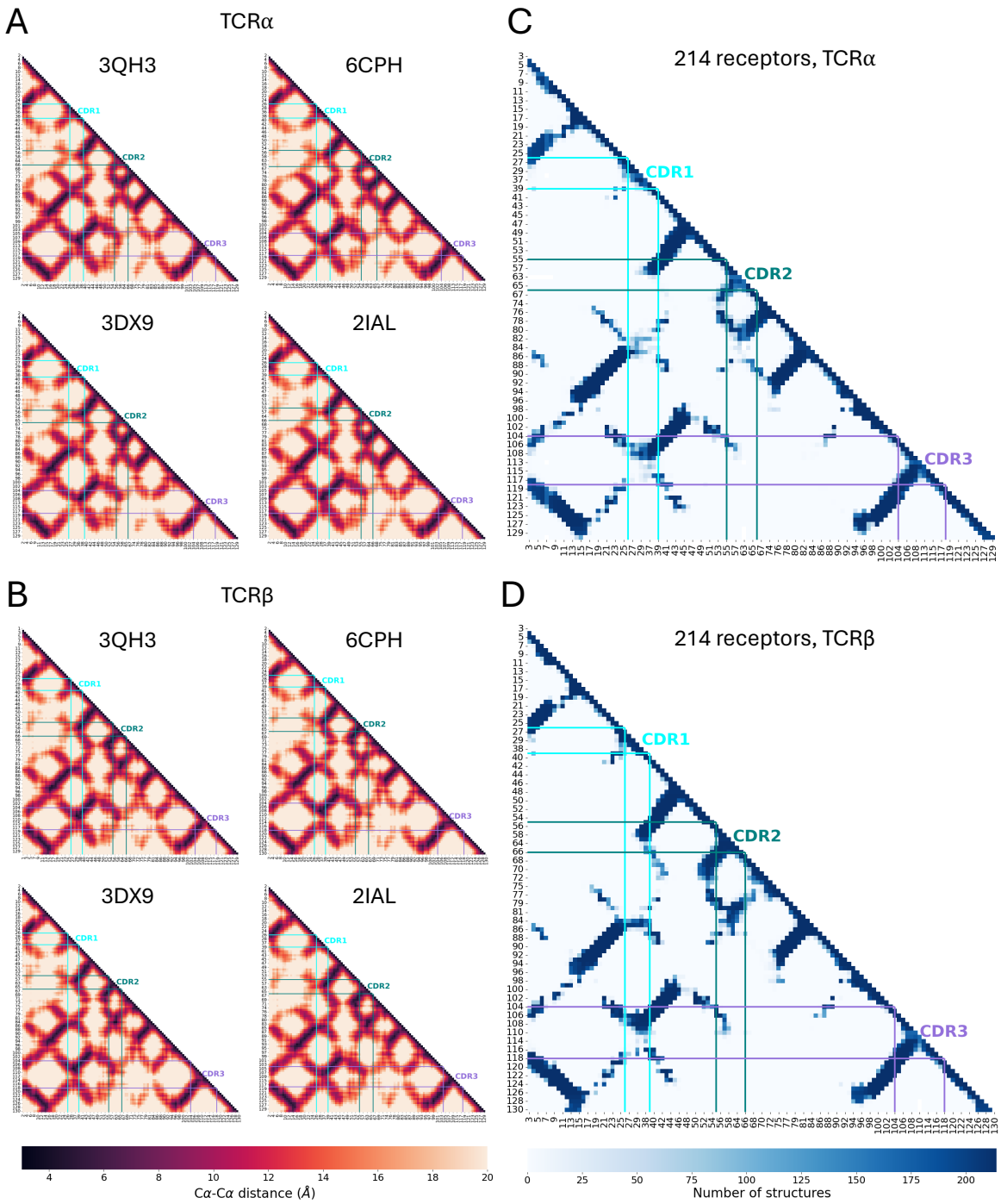

Figure 2: **Conserved intra-chain interactions are observed in TCR structures. A, B)** Intra-chain contact maps for the structures in Figure 1 for the TCR α and β chain, respectively. The distance (Å, colour bar) is calculated between Cα at each residue. **C, D)** Intra-chain contact maps were calculated for the TCR α and β chain (respectively) of 214 unique TCRs. A contact was defined as distance between C$\alpha \leq 10$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 start and end positions of CDR1, 2 and 3 are highlighted.

conserved contacts with FR2 on the other chain, as well as contacts with the beginning and end of the CDR3, where the CDR3 sequence is least variable (the diamond-shaped pattern). We speculate these are conserved structural contacts that allow proper formation of the TCR heterodimer. On top of these, inter-chain contacts are observed between CDR1 and 2 and the CDR3 on the paired chain (CDR2β and CDR3α in all four structures; CDR2α to CDR3β in all but 3QH3). Remarkably, we observe a high number of contacts between the two CDR3s, whilst no contacts are observed between the germline loops across the two chains. Notably, the two CDR3s form a X-shaped pattern: the interactions here involve the two more variable regions, but no interactions are observed between the variable regions and the conserved segments on the loop (the opposite of the diamond pattern above). We extended these observations to the whole set of 214 unique TCRs (Figure 3B). Compared to the four selected examples, the contacts between CDR2 and CDR3 across the two chains are not as conserved across structures. On the other hand, the diamond-shaped pattern between FR2 and CDR3α/β contacts is conserved across most structures, as is the X-shaped interaction between the CDR3s. Overall, these results suggest that the CDR loops come into close proximity with one another to form the antigen binding site. We thus hypothesise that binding interactions between these loops may stabilise the conformation and hence specificity of the binding surface.

A footprint of inter-chain interactions on TCRα and β paired sequences

We reasoned that the observed intra- and inter-chain interactions should impose constraints on the paired TCR$\alpha\beta$ amino acid sequences recognising a given antigen. Specifically, within an antigen-specific repertoire the sequence of one TCR chain should restrict the allowed residues on the other chain. One testable prediction of this hypothesis is that a set of similar antigen-specific TCRα should bind to a set of similar TCRβ, and vice versa. Furthermore, this constraint should be specific to antigen-specific repertoires, and very low in a background repertoire.

To test this hypothesis, we retrieved paired-chain data from the VDJDb and analysed each epitope-specific repertoire (Bagaev et al., 2020; Goncharov et al., 2022). We clustered all CDR3α and all CDR3β sequences based on triplet similarity. Figure 4A shows clustering of CDR3α and β for a single epitope (GLCTLVAML). Both chains show some clustering, but there is no one-to-one relationship between clusters of α sequences and clusters of β sequences. However, within a single cluster of CDR3α sequences, the similarity of the paired CDR3β is higher than size-matched random subsamples of CDR3β sequences from that same repertoire (Figure 4B), and similarly for CDR3β clusters. Notably, all but one α clusters (cluster 139) and 5 of 7 β clusters (clusters 0, 1, 6, 89 and 94) from epitope GLCTLVAML show higher similarity in paired sequences compared to background (β cluster 5 shows a similar trend but does not reach statistical significance). We extended the analysis to all available epitopes within the VDJDb for which there are at least 100 paired $\alpha\beta$ TCRs (Figure 4C and Figure S9). We restricted the analysis to clusters of size 3 or greater. We detected no CDR3α clusters of size 3 or greater in epitopes LLWNGPMAV, SSYRRPVGI and TTDPSFLGRY and no CDR3β clusters of size 3 or greater in epitopes RLRAEAQVK and SPRWYFYYL. Most epitopes showed increased

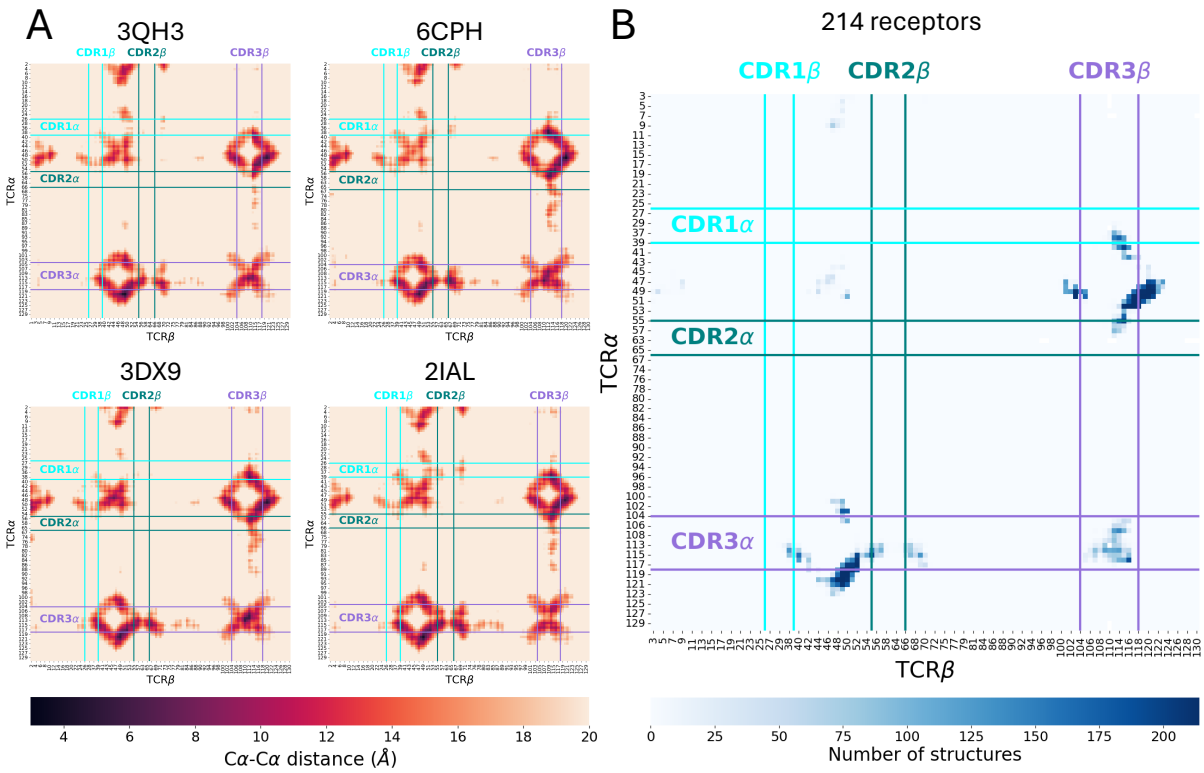

Figure 3: **Conserved inter-chain interactions are observed in TCR structures. A)** Inter-chain (TCRα to TCRβ) contact maps for the four structures in Figure 1. The distance (Å, colour bar) is calculated between Cα at each residue along the two chains. **B)** Inter-chain contact maps were calculated for 214 unique TCRs. A contact was defined as distance between C$\alpha \leq$ 10Å. 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 start and end positions of CDR1, 2 and 3 along the sequence are highlighted.

similarity in TCRα or β associated with clusters of the other chain (Figure S9). However, the increase in sequence similarity was variable, ranging from epitope GLCTLVAML where there is a 10-fold increase in median sequence similarity in CDR3α and a 100-fold increase in median similarity in CDR3β to epitope AVFDRKSDAK where no signal is detected for either chain, and the sequence similarity even decreases for CDR3β. The latter would suggest chain independence, i.e. TCRs binding to this particular epitope are not constrained in the CDR3$\alpha\beta$ pairing.

Clustering relies on selecting a fixed threshold to connect two sequences. This leads to cases where no large clusters are detected. To address this limitation, we evaluated more generally whether the pairwise similarity between two TCRα and their two paired TCRβ was correlated. We grouped the pairwise similarities calculated on TCRα sequences into bins, and plotted the similarity of the paired TCRβ sequences within each bin (Figure S10A and the reverse correlation, exchanging α and β, in S10B). This analysis also shows correlation between the similarity on the two chains for most epitopes. Overall, these results suggest that, within an epitope-specific repertoire, similar TCRα chains tend to pair with TCRβ chains that are more similar than random, i.e. TCR$\alpha\beta$ pairing is constrained.

Quantifying the interaction between CDR loops sequences using mutual information

In order to generalise our analysis beyond distance metrics, we re-formulated the problem in the context of information theory. We estimated the mutual information (MI) between TCRα and β sequences, and between V or J genes and CDR3 sequences, using a pairwise approximation and corrected for sample size and background distributions (see Methods and Figure S2). As negative control for the antigen-specific analysis, we estimated the MIs for all TCRs in the VDJDb (referred to as background), as well as for a single sample of sorted naïve T cells from Tanno et al., 2020 (referred to as Tanno::A1::naïve). The former contains information about the 22 epitopes under study pooled together (and might therefore have some residual epitope-specific signal), whilst the latter should only depend on the biology of receptor formation and thymic selection, without the influence of antigen selection.

Figure 5A shows the estimated MIs for each epitope. The only pairs of sequence ensembles where MI reveals statistical dependence in the naïve set are intra-chain pairs between V or J gene and CDR3 sequence (bottom right heatmap, top left and bottom right quadrant). These signals are expected due to the overlap between V/J sequences and CDR3 sequence, as well as the restrictions on CDR3 sequence imposed by V/J gene selection during V(D)J recombination. Most epitopes show some level of statistical dependence beyond the naïve repertoire, with epitope AVFDRKSDAK again showing low to no signal. Some epitopes (ASNENMETM, ATDALMTGF, GLCTLVAML, HGIRNASFI, KSKRTPMGF, SSPPMFRV and YLQPRTFLL) show moderate levels of statistical dependence between CDR3β and all other TCR components on both chains. The MIs for all epitopes are summarised in Figure 5B. The most marked increase comparing background and epitope-specific repertoires is in the CDR3α-CDR3β MI, but other inter-chain relationships (Vβ-CDR3α and Vα-CDR3β in particular), and some of the

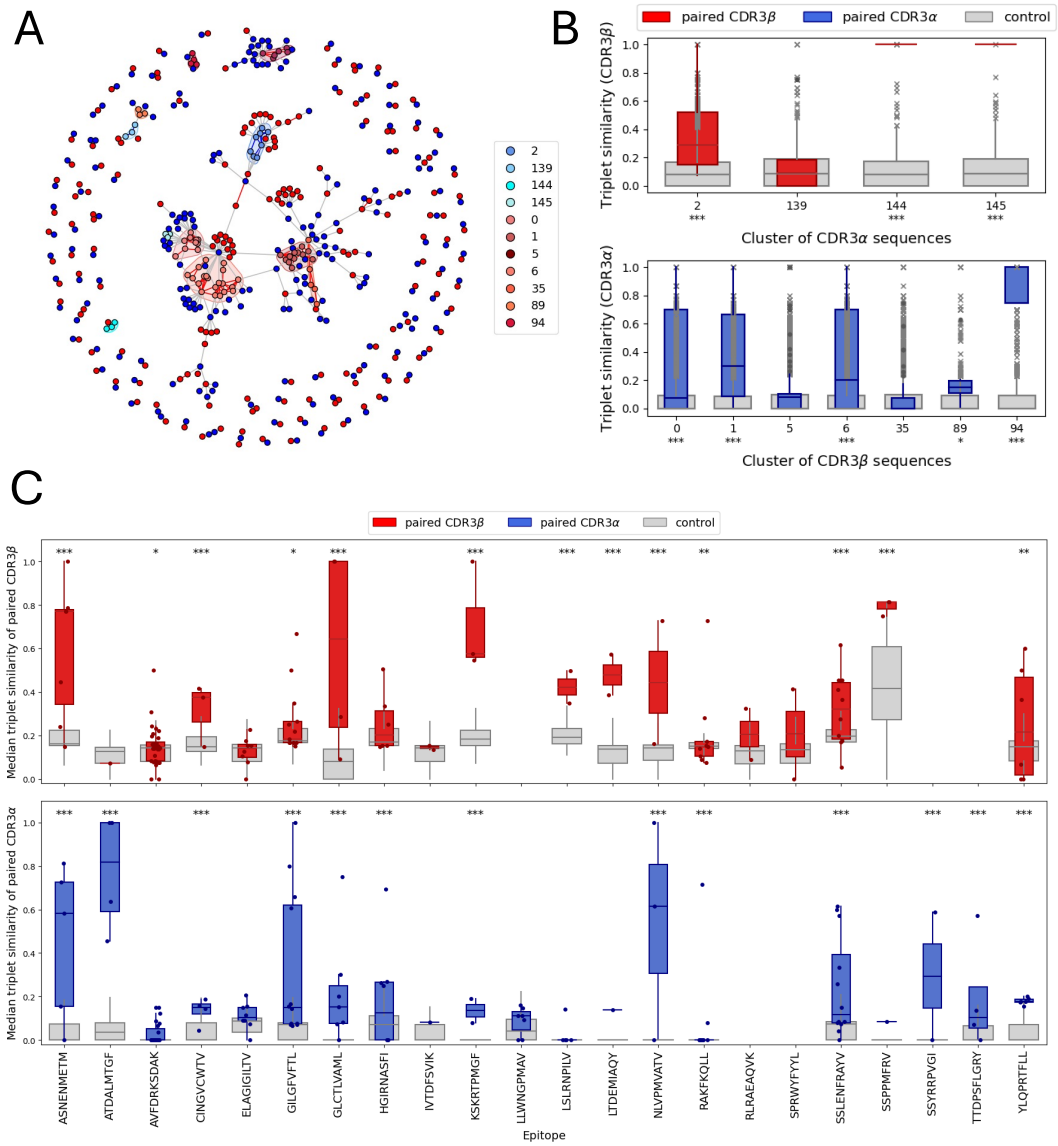

Figure 4: **Co-clustering of CDR3α and β sequences. A)** Each circle is a unique CDR3 sequence (α in blue, β in red). Similar α sequences are linked by blue edges, similar β sequences by red edges. Each α is linked to its paired β by a grey edge. The clusters analysed in **B** are coloured in shades of blue (α clusters) or red (β clusters) and identified by a shaded area around them. Each CDR3 sequence is represented only once, therefore the same CDR3β may be linked to multiple α (and vice versa), if it is not unique in the epitope-specific repertoire. **B)** Similarity of paired CDR3β when α are clustered (top, or vice versa - bottom) for the clusters identified in **A**. For each cluster, all pairwise similarities between the paired sequences are calculated (coloured boxplots). 100 equal-size random samples are taken as controls from the epitope-specific repertoire (grey boxplots). Only clusters of size 3 or greater are included. **C)** The analysis in **B** was repeated for 22 epitopes. Each dot represents the median for a cluster of size 3 or greater. Median similarity was also calculated for 100 controls for each cluster (grey boxes, random sequences drawn from the same epitope-specific repertoire). Columns are empty when no clusters of size 3 or larger are found. The median of each boxplot is shown as a line. P-values (one-tailed t-test, comparing real to control): * < 0.05; **<0.01; ***<0.0001.

intra-chain relationships (Vβ-CDR3β and Vα-CDR3α in particular) are also higher in epitope-specific repertoires.

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 6A). As CDR3 positions 104 and 118 on both chains are constant, the MI is 0. Residues 105-107 on CDR3α are strongly dependent on the V gene, which reflects the overlap between V region and the start of the CDR3, and can be observed in both VDJDb and naïve background. A similar pattern can be observed with the first few residues on the CDR3β. However, we observed statistical dependence between the V gene and the central residues of the CDR3 for most epitopes, as well as the VDJDb background. In contrast, no statistical dependence was observed for these positions in the naïve repertoire (bottom row of the heatmap). Antigen-dependent TCR selection therefore commonly induces statistical dependence between the CDR3 sequence and V gene, which we speculate may reflect the observed structural interactions between the V region and the CDR3 region (Figure 2).

Finally, we examined the MI between the CDR3α and CDR3β (Figure 6B). Little or no MI is detected in the background repertoires suggesting weak restriction of CDR3 pairing at the repertoire level. In contrast, most epitopes show statistical dependence in inter-chain CDR3 interactions, with an epitope-specific signal pattern. As in all the previous analysis, epitope AVFDRKSDAK shows very low signal. Antigen-dependent TCR selection therefore commonly generates significant statistical dependence between the CDR3α and β sequences, which we hypothesise reflects restrictions in chain pairing due to structural interactions.

Examining MI variation across epitopes

A striking feature of both the intra-chain and particularly the inter-chain statistical dependence was the variability between epitopes, akin to what observed when clustering TCRs from different epitopes (see for instance Dash et al., 2017), or estimating their repertoire diversity with unbiased estimators (Tiffeau-Mayer, 2023; Henderson et al., 2024). Understanding the factors which drive this epitope dependence may have important implications for our ability to predict antigen specificity.

We first measured the correlation between estimated MI and epitope repertoire size, to check that our results were not caused by finite size effects. Reassuringly, we did not detect a significant correlation for most of the relationships investigated (Figures 7 and S11A).

Certain epitope-specific responses are characterised by usage of a specific public TCR chain which pairs with different partner chains (see for instance Zhong et al., 2007; Pogorelyy et al., 2022), or differing levels of sequence similarity (see for instance Dash et al., 2017). To evaluate the impact of similar or public TCR chains in MI estimation, we looked at the correlation between MI and the effective set size. We calculated the effective set size of CDR3α, CDR3β and CDR3$\alpha\beta$ in each repertoire. We set the threshold of similarity to be up to Hamming distance 1 for CDR3α and CDR3β, and up to Hamming distance 2 for CDR3$\alpha\beta$ (concatenated

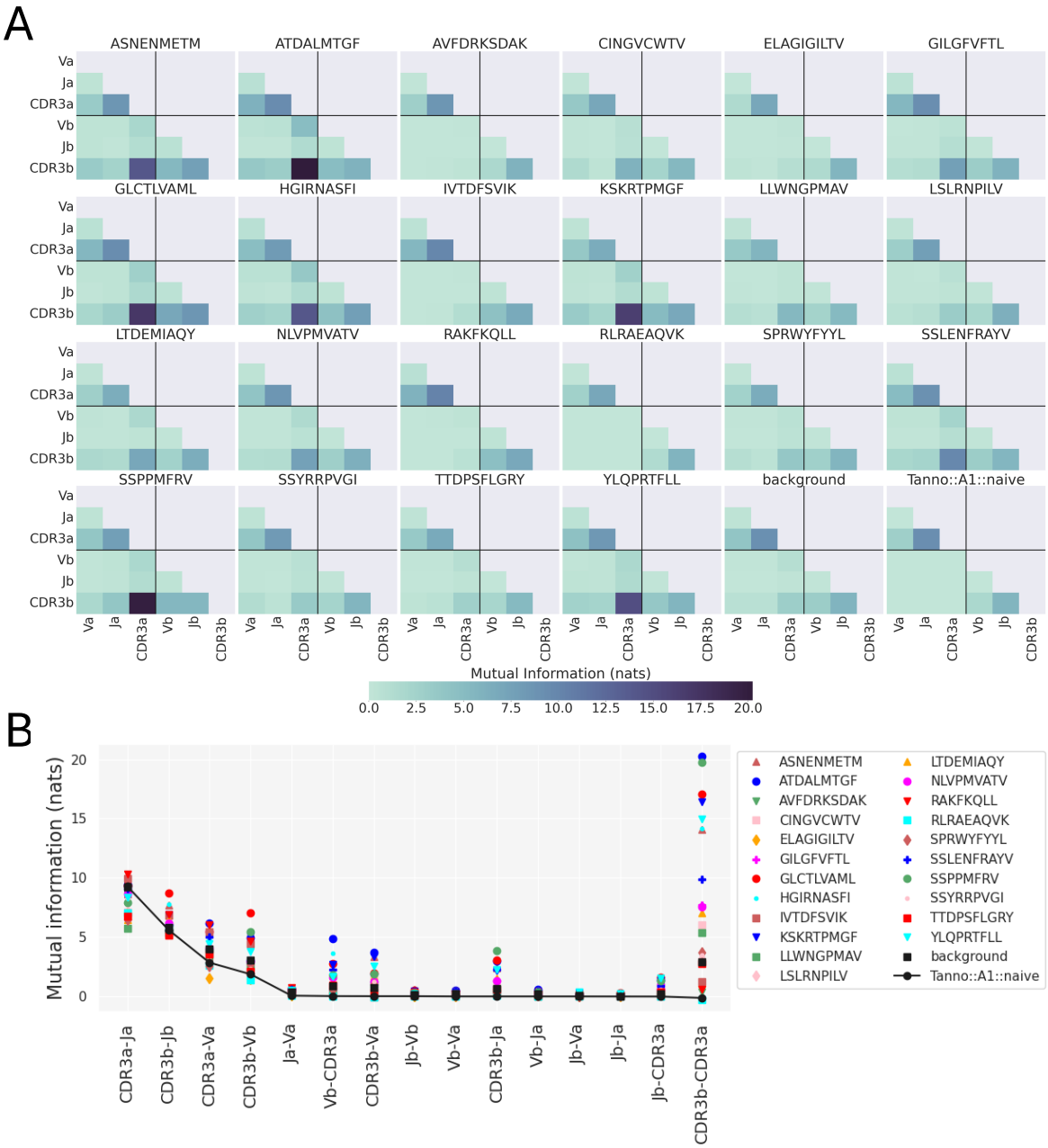

Figure 5: **Mutual information between TCR components in epitope-specific repertoires. A)** Mutual information (MI) was calculated between all different components of the TCR (V genes, J genes, CDR3 sequences) for each epitope-specific repertoire. In each heatmap, the top-left and bottom-right quadrants contain intra-chain MI (α in top-left and β in bottom-right). The bottom-left quadrant shows the inter-chain MI. The MI was estimated for various subsamples for each epitope, and an estimate correcting for finite size effects was found for both the real set and a shuffle (Figure S2). The MI plotted here corresponds to the difference between the estimated MI for real and shuffle. 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 MI in the naïve control repertoire (Tanno::A1::naïve), from largest to smallest (black circles joined by a solid line). Each epitope is identified by a combination of colour and shape.

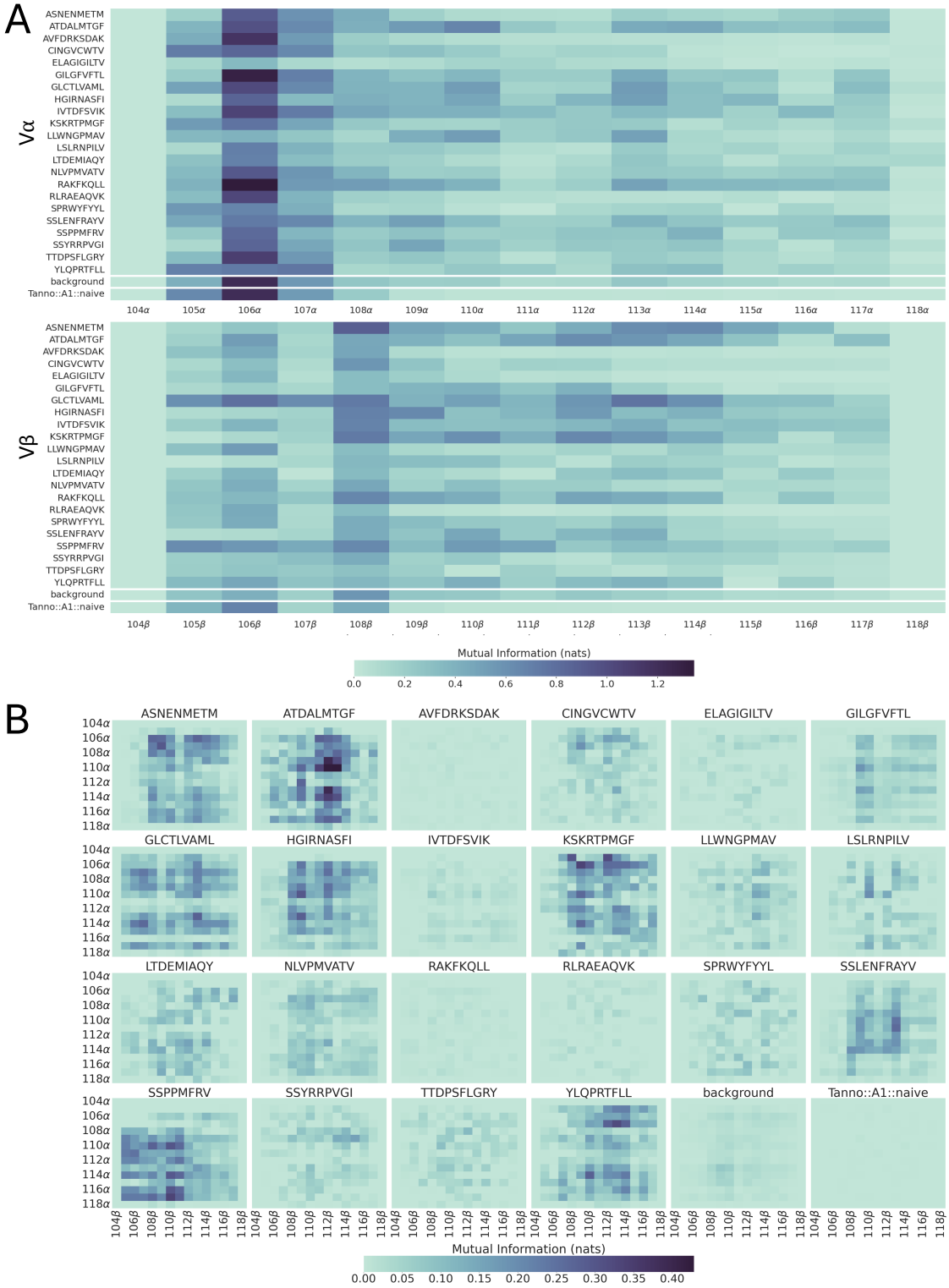

Figure 6: **Mutual information between CDR3 and associated V gene and CDR3α-CDR3β. A)** Mutual information (MI) between V genes and each residue position on the CDR3. Top: Vα to CDR3α; bottom: Vβ to CDR3β. **B)** MI between each pairs of positions on CDR3α and CDR3β. The MI was calculated with pairwise 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.

as a single sequence). We then correlated each of these measures with the MI for that repertoire (Figure 7A and Figure S11B). We observe that repertoires dominated by a single sequence or cluster of similar sequences (a smaller set size) tend to have more MI.

We also examined whether the MI may depend on the feature-ness of the epitope within the MHC (i.e. how many features the peptide residues offer for the TCR to bind), which has been proposed to correlate with repertoire diversity (Turner et al., 2005; Turner et al., 2006). Structures of peptide-MHC for the same set of epitopes were downloaded where available and the feature-ness of a peptide was approximated by the solvent-accessible surface area of the peptide (SASA, Table 1). Epitopes that have a relatively featureless surface (low SASA), or which are bulged out of the MHC (high SASA) tend to generate a less diverse TCR repertoires (Turner et al., 2005; Turner et al., 2006). Therefore, we expect the most MI to be contained in epitopes that are either very featured or featureless (very high or very low SASA). To test for this, we determined whether a relationship between SASA and MI exists by fitting a parabola with ordinary least squares (Figure 7B). No significant relationship was detected, but structures are available for only 13 out of 22 studied epitopes, which limits the statistical power of the analysis.

Using co-evolution methods for TCR$\alpha\beta$ pairing

The correlations of sequence similarity and MI across chains (Figures 4 and 5) resemble those of co-evolving protein families (De Juan et al., 2013). We therefore wondered whether we could use methods developed for co-evolving proteins to select the most likely TCRβ partner for each TCRα within epitope-specific repertoires, when the pairing information is withheld. We adapted two methods: a MI-based method (MI-IPA, Bitbol, 2018) and a graph alignment method (GA, Bradde et al., 2010), as well as a combination of the two (Gandarilla-Pérez et al., 2023, GA+MI-IPA,).

We optimised the MI-IPA and GA by comparing performance on two different epitopes, GLCTLVAML (EBV) and YLQPRTFLL (SARS-CoV-2), both presented on HLA-A*02 (Figures S5 and S6, respectively). These epitopes have comparable number of annotated TCR$\alpha\beta$ in VDJDb (345 and 333, respectively, Table S1), but GLCTLVAML shows a stronger MI signal than YLQPRTFLL (Figure 5). We then ran the optimised models on all epitopes. To reduce computational time, we subsampled epitopes with > 1000 TCRs to 5 subsamples of 700.

First, we ran the MI-IPA on each epitope-specific repertoire. The MI-IPA exploits the residue-level signal that is available from TCR$\alpha\beta$ pairs to perform the pairing. The results are summarised in Figure 8. In 14/22 epitopes, the MI-IPA can perform significantly better than random guessing (we consider it successful for epitopes GILGFVTL and RAKFKQLL but not AVFDRKSDAK based on whether it is significant in at least 3/5 subsamples). As the model was run 10 times, we explored whether the stability of the assignments between repeats could provide a confidence score for each pair. Overall, for most epitopes the majority of pairs are not assigned stably across the repeats (Figure S12A), but a few $\alpha\beta$ are always paired together. Interestingly, epitopes such as ATDALMFTGF and RLRAEQVK which show no signal of learning in Figure 8

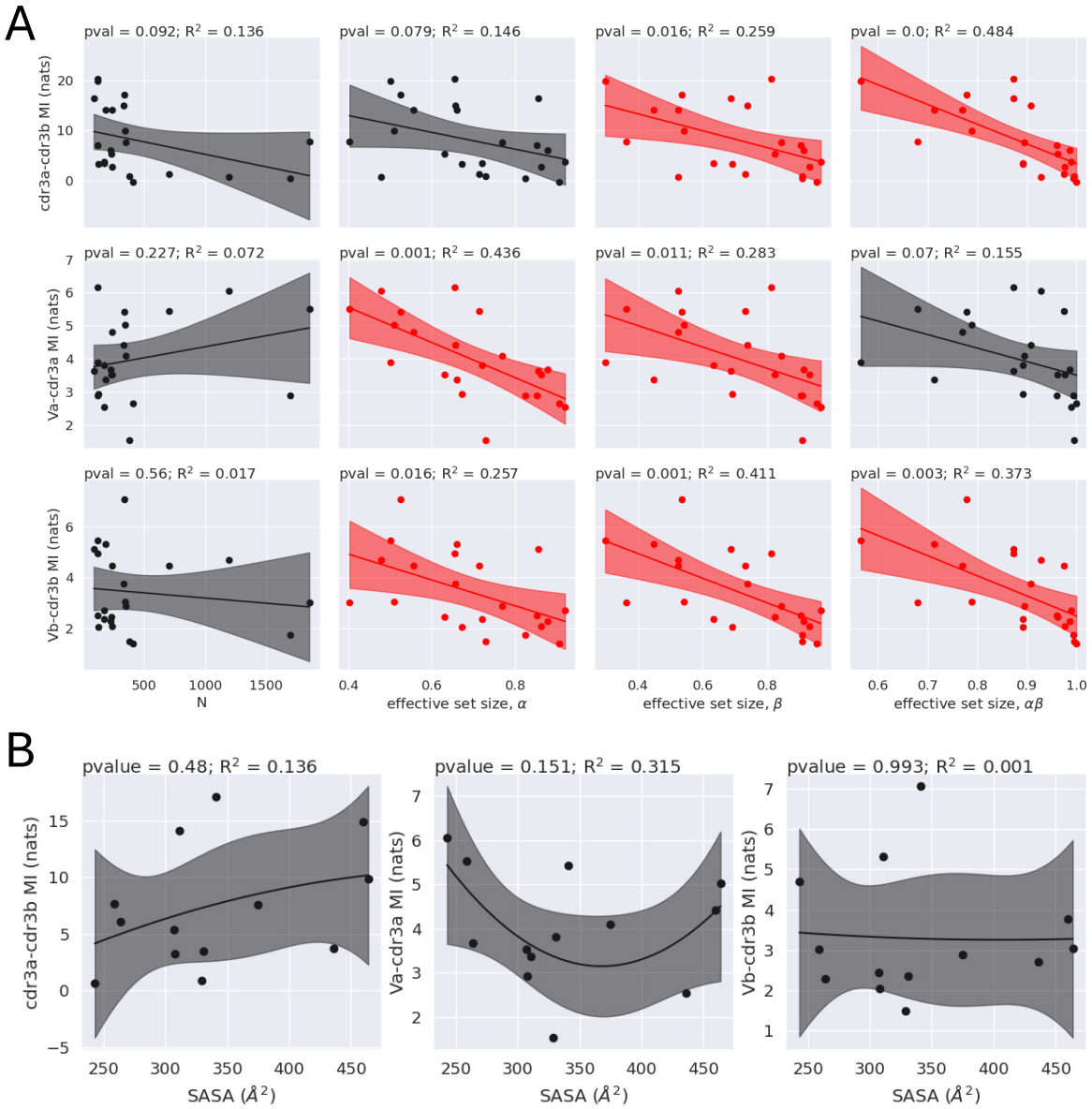

Figure 7: **Mutual information correlates to effective set size but not peptide solvent-accessible area. A)** Correlation of mutual information (MI) and effective set size. Effective set size was calculated by grouping all sequences with Hamming distance of < 2 on each chain and < 3 for the paired chains, 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 S11. **B)** Correlation of MI and solvent-accessible surface area (SASA) for the available epitopes (Table 1). Correlations were calculated by fitting a linear regression. P-value for the F-test and R^2 was 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.

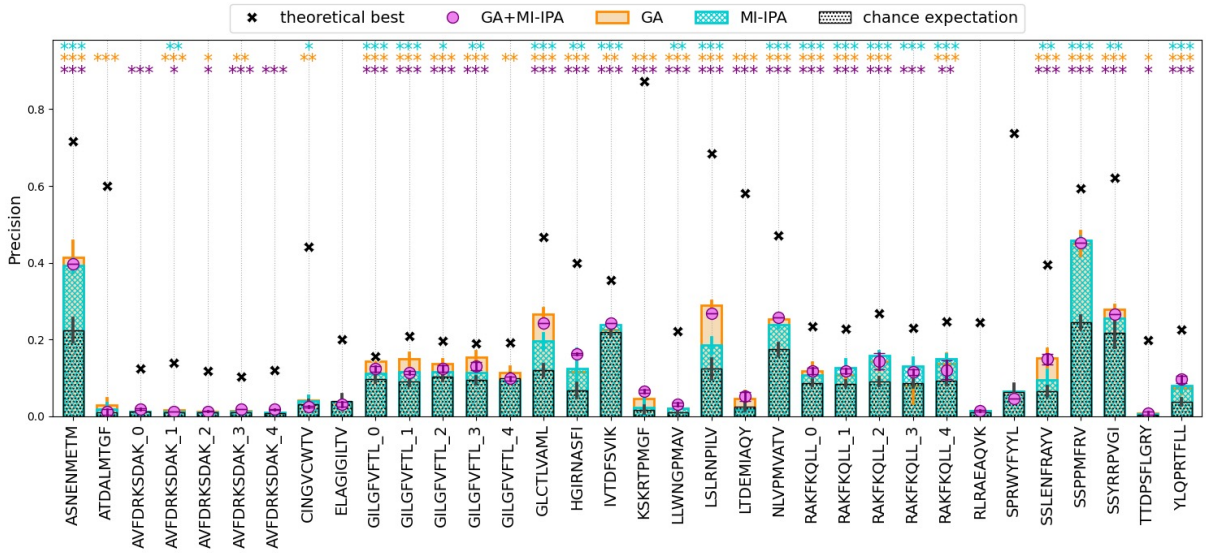

Figure 8: **Co-evolution based models perform better than random at pairing TCR$\alpha\beta$ in epitope-specific repertoires.** Precision is calculated over 10 repeats of the MI-IPA or GA+MI-IPA (cyan and magenta, respectively) or 100 repeats of the GA (orange) for all epitopes. For epitopes with > 1000 sequences, 5 subsamples of 700 sequences are shown. The IPA is run both with $\lambda = 1$ (chance expectation, black bars) and $\lambda = 0.6$ (cyan). The black crosses 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 Student's t-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. P-values are indicated by asterisks in the corresponding colour: ***< 0.001; $0.001 \leq$**< 0.01; $0.01 \leq$*< 0.05.

also seem to have unstable assignments, whilst epitopes such as ASNENMETM and SSPPMFRV, for which the MI-IPA can make predictions, show a large proportion of stable pairs. Notably, stable pairs are enriched for correct pairs for all epitopes tested (Figure S12B). Thus stability under repeat model building may be useful to assess the reliability of the pairing assignment algorithm (as in Bitbol et al., 2016).

We then implemented the GA algorithm, which attempts to optimally align edges between the two nearest neighbour graphs of TCRα and TCRβ sequences, thus exploiting the observed correlation of sequence similarity (Figure S10). We ran the GA algorithm to pair TCR$\alpha\beta$ from all available epitope repertoires (Figure 8). Overall, the GA shows significantly better performance than chance expectation in 19/22 epitopes. Some epitopes, such as LSLRNPILV, showed better precision by using the GA algorithm, whilst for others the GA achieves similar performance to the MI-IPA. Epitopes that perform well have more stable assignments than epitopes for which the GA cannot achieve good performance (Figure S13A). Moreover, stable pairs are enriched for correct pairs (Figure S13B).

Finally, we combined the two methods as suggested in Gandarilla-Pérez et al., 2023. Briefly, stable pairs from the GA assignments provide the training set for the MI-IPA. We opted to retain these sequences in the testing classification task, so as to allow the MI-IPA to correct any

mistakes in the GA classification (Figure S7). The results for the GA+MI-IPA are shown in Figure 8. The combination of the two methods can occasionally marginally improve on either method alone. We then investigated how stable the selection of a CDR3β for each CDR3α is in the GA+MI-IPA (Figure S14). Unlike the stability plots for MI-IPA and GA, the GA+MI-IPA shows a larger proportion of stable pairs. We attribute this to the use of a training set in the MI-IPA when combined to the GA as the trajectory of the MI-IPA is less random when a training set is provided. Unfortunately, this has repercussions on the ability to use stability to enrich for correct pairs: for epitopes such as ASNENMETM, GLCTLVAML or LSLRNPILV the precision in the most stable pairs is similar to the overall precision (Figure S14B).

Overall, the pairing algorithms showed a modest but significant ability to identify true pairs of $\alpha\beta$ sequences. To understand whether the limitation was in the learning step (i.e. the way the algorithm selects pairs does not allow it to find the optimal solution) or in the data (i.e. the signal in the data is too low to achieve pairing), we calculated the theoretical best performance for the MI-IPA. To achieve this, we initialised the MI-IPA with all correctly-paired $\alpha\beta$ for each epitope. This allows the MI-IPA to immediately see all the rules that govern the pairing at once (the complete set of statistical dependencies). We then ran a single iteration of the MI-IPA to pair all the available α/β from that epitope repertoire again. We could thus assess whether the MI-IPA correctly pairs all available sequences knowing the pairing rules *a priori* (Figure 8). In all epitopes, the pairing algorithms under-perform compared to the theoretical limit, suggesting that the learning process is not able to extrapolate all the rules that can be learnt from these pairs, i.e. it converges to a different solution. However, the theoretical maximum is well below 1 for all epitope repertoires, suggesting that the signal available from the data is limited, and can recapitulate only some of the rules of TCR$\alpha\beta$ pairing. We posit that dataset size and difficulty in aligning these sequences may be some of the limiting factors in this approach.

Correlates of pairing performance and repertoire characteristics

We sought to understand which factors drive the large range in precision observed in Figure 8 for each model. The MI-IPA is a data-thirsty method, and the total size of the repertoire, as well as the size of the individual repertoires can highly influence performance (Bitbol, 2018). Indeed, larger individual repertoire sizes will make the task significantly harder for the pairing algorithms, as more combinatorial pairs need to be evaluated. Moreover, the final result might be influenced by the MI available in each epitope-specific repertoire (defined as in Figure 5). We therefore correlated MI-IPA performance (calculated on the consensus assignment over 10 repeats, i.e. by taking the modal β for each α) with each of these factors (Figure 9A). The theoretical maximal, but not the actual performance was inversely correlated to individual repertoire size and total epitope repertoire size, and positively correlated to MI. These 3 variables together can explain over 50% of the variance observed across epitopes in the theoretical best performance, and almost 30% in the learning scenario (multivariate linear regression, Table S2).

To understand which factors influence GA performance, we correlated the performance of

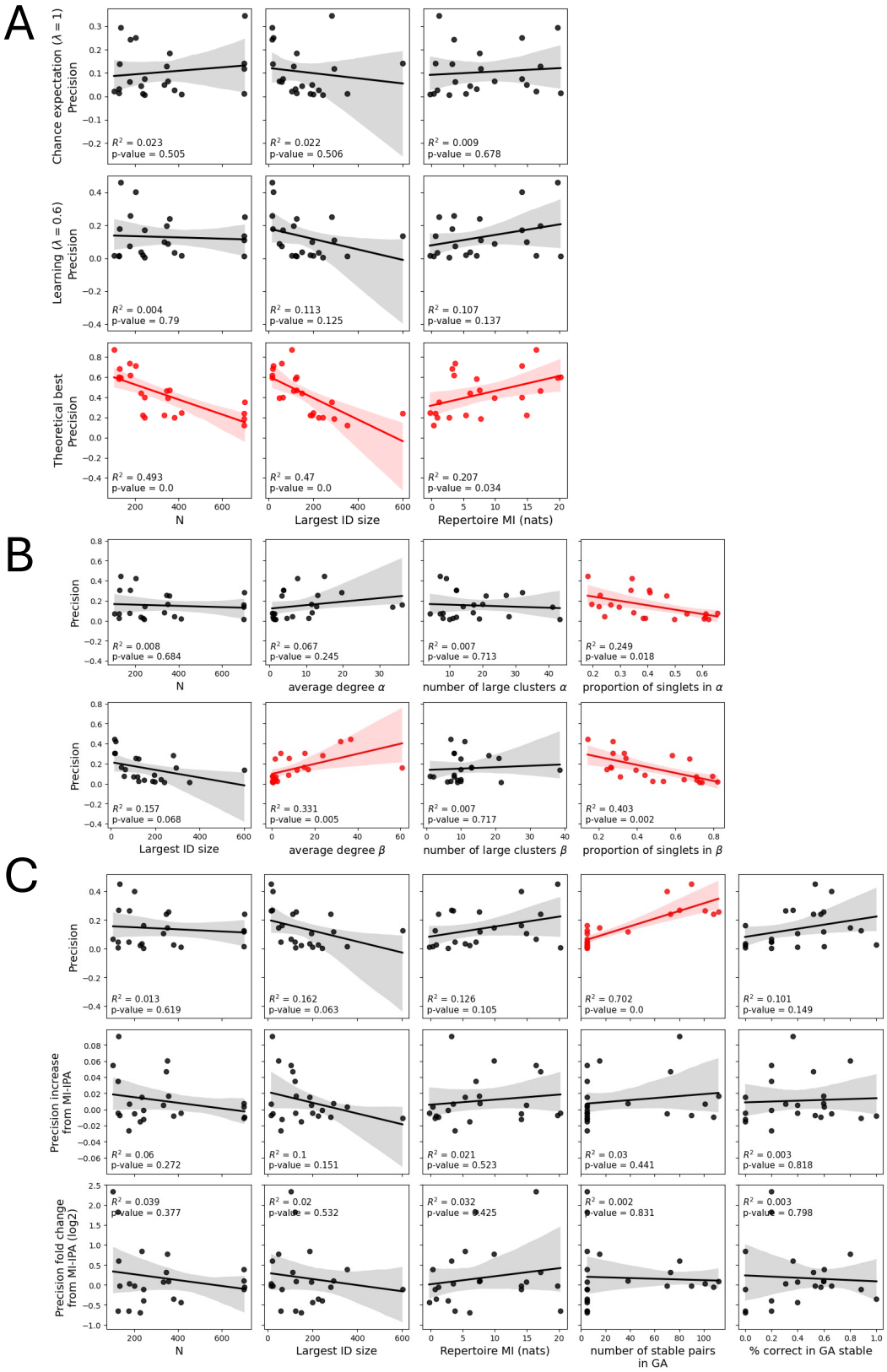

Figure 9: **Correlation of model performance and repertoire characteristics.** *Caption next page.*

Figure 9: **Correlation of model performance and repertoire characteristics.** (*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 as in Figure 5 between CDR3α and CDR3β) 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 ≥ 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 CDR3α and CDR3β 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 β for each α over multiple iterations of the model. Since multiple CDR3β may be selected the same number of times for one CDR3α, 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 R^2 for the regression is shown and 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.

the GA (calculated on the consensus assignment over 100 repeats) with repertoire size, largest individual size, as well as characteristics describing the sequence similarity networks that can be generated from each of these epitopes (calculated using pairwise Levenshtein distance, threshold < 3). The results are shown in Figure 9B. Interestingly, performance correlates with the properties of the sequence network: the more sequence clustering is observed (higher average degree, smaller proportion of singlets), the better the GA is at predicting outcome. The model built with these variables can explain almost 80% of the variance between epitopes (Table S3).

Finally, we looked at factors that might impact the performance of the GA+MI-IPA. As for the MI-IPA, we calculated the effect of repertoire size, largest individual repertoire size and repertoire MI. Moreover, we extracted the number of stable pairs from the GA and the percentage of the GA stable set that is correct. We correlated these factors with both the precision, increase in precision from using MI-IPA on its own, as well as the fold change between the precision of the MI-IPA and the precision of the GA+MI-IPA, to see if we could explain why some epitopes benefit from the initial GA step and some do not. Figure 9C shows the results for each variable independently. The number of stable pairs in the GA is significantly correlated with precision of the GA+MI-IPA. Interestingly, none of these factors can explain why certain epitopes improve on the MI-IPA and others do not. When combining these variables in a multivariate regression, they can explain over 80% of the observed variance in precision of the GA+MI-IPA, but they are unable to explain the variance in improvement compared to MI-IPA alone (Table S4).

Discussion

We report the most systematic analysis to date of interactions between TCR CDR sequences. We observed conserved intra- and inter-chain contacts between the CDRs from over 200 unique TCRs. Specifically, we observed intra-chain interactions between CDR1 and CDR3, CDR1 and CDR2, and CDR2 and CDR3 on each chain. We further observed inter-chain interactions between the conserved framework region 2 on each chain, and at the interface between the two hypervariable CDR3s. We did not detect conserved interactions between the germline-encoded loops on the two chains, consistent with the idea that there are no significant constraints in TRAV/TRBV pairing at the repertoire level (Dupic et al., 2019; Shcherbinin et al., 2020). The observed intra- and inter-chain interactions suggest that the TCR binding site for antigen is shaped by interactions between the six CDR loops.

We thus hypothesised that if the intra- and inter-chain contacts are important for shaping epitope specificity, the sequences of the two TCR chains would restrict each other's diversity in the context of sets of TCRs with shared pMHC specificity. We found several strands of evidence supporting this prediction. Firstly, using a measure of pairwise sequence similarity, we find that epitope-specific TCRs with similar α sequences have paired β chains with restricted sequence diversity. The reverse relationship also holds. Secondly, we can measure significant MI between the V gene and the CDR3 sequences within an epitope-specific repertoire. This observation is consistent with the idea that CDR1 and CDR2 interact with CDR3, although we note that for this analysis we defined V genes categorically by their gene names, which loses some sequence-level information. Thirdly, we detected consistent MI between the CDR3α and the CDR3β sequences within sets of epitope-specific TCRs. Interestingly, the mutual restriction seen between TCRα and β sequence diversity in the context of shared specificity is reminiscent of the 'light chain coherence', recently described in antibodies (Jaffe et al., 2022). Our results differ from those in Shcherbinin et al., 2020, as they did not find evidence for restriction of TCR chain pairing even within epitope-specific repertoires. This discrepancy can likely be attributed to the inclusion of the CDR3 sequence in our analysis. Indeed, this is consistent with recent results showing that the CDR3α and CDR3β carry synergistic information about epitope binding, greater than the synergy measured between Vα and Vβ genes (Henderson et al., 2024). All three examples of $\alpha\beta$ diversity restriction discussed above were consistent across most epitopes tested. However, there was considerable variance in the interaction strength detected when comparing TCRs binding different pMHCs, discussed in more detail below.

A key unsolved practical problem in the field of TCR repertoire analysis is the ability to predict correct pairing from TCRα and TCRβ chains sequenced independently. A method able to make such a prediction would allow to quickly find candidates for, for example, TCR-engineered T cell therapies from bulk TCR sequencing of patient samples, thus expediting the existing experimental pipelines. We reasoned that the mutual sequence constraints imposed on TCR$\alpha\beta$ are somewhat analogous to constraints found between co-evolving protein families, which share similarity in their phylogenetic trees, and show sequence correlations at the residue

level (Korber et al., 1993; Fryxell, 1996; Dunn et al., 2008; De Juan et al., 2013; Ochoa & Pazos, 2014). A number of approaches have already been explored to predict pairing between co-evolving interacting protein families. To reframe the TCR$\alpha\beta$ pairing problem as a co-evolution problem, we consider the multiple TCRα and TCRβ chains which bind to the same epitope as 'paralogs' from interacting protein families, while similar TCRα (or β) sequences that bind the same epitope but are found in different individuals are 'orthologs'. We adapted three co-evolution based methods to try and predict TCR$\alpha\beta$ pairing (Bradde et al., 2010; Bitbol, 2018; Gandarilla-Pérez et al., 2023). Importantly, we have applied these models in the simplest possible scenario: by using lists of TCRα and TCRβ which have been previously annotated for a single epitope. Except for the theoretical best scenario, a training set of know TCR$\alpha\beta$ pairs was never provided to the pairing algorithms, so that we could evaluate each model's ability to make *de novo* pairing predictions. We were able to detect a significant improvement in pairing performance over random assignment in all epitopes we tested except three (ELAGIGILTV, RLRAEAQVK and SPRWYFYYL) using at least one of the three methods. Consistent with this, previous studies have suggested that the TCR recognition of epitope ELAGIGILTV is dominated by the TCRα chain (Trautmann et al., 2002; Dietrich et al., 2003), and may be largely independent of TCRβ.

The performance of the algorithms provides proof-of-principle that the MI detected between TCRα and β CDR sequences can translate into pairing information. However, the performance remains poor, precluding their current use for practical application in clinical settings. Two technical improvements might be possible: a better integration of GA and MI-IPA and the integration of TCRα and TCRβ abundance information. Some experiments combining the GA and MI-IPA were carried out with scarce improvement on each method singularly. Future studies may explore better ways to use the results from the GA to initialise the MI-IPA. On the other hand, we might be able to boost performance by including information other than TCR sequence, such as abundance. We expect chains that come from the same clone to have similar abundance in the same sample, and vary across samples in a correlated manner. This information is readily available from TCR sequencing experiments (see, for instance, Oakes et al., 2017). Therefore, it might be possible to integrate this information to achieve better predictions.

The optimisation of the algorithms may provide performance gains. However, more fundamental factors may also limit performance. In particular, in the co-evolutionary context, interacting proteins evolve over time, and thus existing examples presumably represent optimal (or optimised) solutions to the interaction. In the TCR context, paired $\alpha\beta$ do not co-evolve, but rather get selected because they are available in the repertoire and are functional. As such, these pairs may not represent the optimal solution to the interaction, and instead they may be a functional (and likely sub-optimal) solution that was selected. Indeed, our set includes repeated sequences in both the CDR3β and CDR3α, which suggests that multiple functional solutions are possible for the same chain (Table S1).

Different epitopes show different levels of MI, different patterns of where most of the MI is contained, and different performance of the pairing algorithms. We hypothesise that this vari-

ance may be explained by both intrinsic and extrinsic features. Intrinsic features reflect physical properties of the the TCR/pMHC complex (e.g. more or less contacts, binding constraints etc.). For example, the feature-ness of a peptide can impact the diversity of the associated repertoire (Turner et al., 2005; Turner et al., 2006). However, we were unable to detect any correlation between feature-ness and MI. Extrinsic features include size of the epitope-specific repertoires, as well as sequence publicity and similarity within a repertoire. We found significant correlations between the effective set size and the MI of the repertoire: the smaller the effective set size, the higher the MI. We speculate this is due to the presence of fewer, larger clusters, which allow us to better recapitulate the binding rules for each in the MI. This is consistent with the positive correlation between the average degree of the CDR3 similarity networks and the performance of the GA algorithm. Overall, we still do not have a full understanding of the factors which determine the relationship between TCRα and β interaction and pMHC. A much larger selection of pMHC than the 22 examples available and studied here will need to be analysed to answer these questions.

In conclusion, we identify extensive, and in some cases highly conserved interactions between the CDR sequences which form the binding surface of a TCR. These interactions impose a constraint on TCR$\alpha\beta$ pairing in the context of antigen specific TCR sets. We provide some initial indications of how this may be used to correctly pair α and β sequences when such paring is not available experimentally. The interactions between CDRs also have fundamental implications in the context of the biophysics of TCR binding. In particular, they may stabilise the conformation of the CDR loops, and reduce their ability to move freely. In this way, the interactions may reduce the entropy of the TCR prior to binding, and thus increase TCR affinity, on-rate and specificity at the expense of reduced breadth. Further experiments to measure the dynamics of TCR, before and after binding, will be required to validate these predictions. This information will be an important factor in future design of TCR-based therapeutics.

Acknowledgements

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. 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.).

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

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. 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.

References

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

Archer, E., Park, I. M., & Pillow, J. (2014). Bayesian Entropy Estimation for Countable Discrete Distributions [arXiv:1302.0328 [cs, math]]. https://doi.org/10.48550/arXiv.1302.0328

Bachmann, M. (2024). Rapidfuzz/RapidFuzz: Release 3.8.1. https://doi.org/10.5281/zenodo.10938887

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

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

Bitbol, A.-F. (2018). Inferring interaction partners from protein sequences using mutual information (C. O. Wilke, Ed.). *PLOS Computational Biology*, *14*(11), e1006401. https://doi.org/10.1371/journal.pcbi.1006401

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

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

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

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

Carter, J. A., Preall, J. B., Grigaityte, K., Goldfless, S. J., Jeffery, E., Briggs, A. W., Vigneault, F., & Atwal, G. S. (2019). Single T Cell Sequencing Demonstrates the Functional Role of TCR Pairing in Cell Lineage and Antigen Specificity [Publisher: Frontiers]. *Frontiers in Immunology*, *10*, 1516. https://doi.org/10.3389/fimmu.2019.01516

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

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, Complex Systems. Retrieved May 2, 2023, from https://igraph.org

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

Davis, M. M., & Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition [Number: 6181 Publisher: Nature Publishing Group]. *Nature*, *334*(6181), 395–402. https://doi.org/10.1038/334395a0

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

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

Dietrich, P.-Y., Le Gal, F.-A., Dutoit, V., Pittet, M. J., Trautman, L., Zippelius, A., Cognet, I., Widmer, V., Walker, P. R., Michielin, O., Guillaume, P., Connerotte, T., Jotereau,

F., Coulie, P. G., Romero, P., Cerottini, J.-C., Bonneville, M., & Valmori, D. (2003). Prevalent Role of TCR -Chain in the Selection of the Preimmune Repertoire Specific for a Human Tumor-Associated Self-Antigen1. *The Journal of Immunology*, *170*(10), 5103–5109. https://doi.org/10.4049/jimmunol.170.10.5103

Duarte, J. M., Sathyapriya, R., Stehr, H., Filippis, I., & Lappe, M. (2010). Optimal contact definition for reconstruction of Contact Maps. *BMC Bioinformatics*, *11*(1), 283. https://doi.org/10.1186/1471-2105-11-283

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

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

Dupic, T., Marcou, Q., Walczak, A. M., & Mora, T. (2019). Genesis of the T-cell receptor [Publisher: Public Library of Science]. *PLoS Computational Biology*, *15*(3), e1006874. https://doi.org/10.1371/journal.pcbi.1006874

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

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

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

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

Goncharov, M., Bagaev, D., Shcherbinin, D., Zvyagin, I., Bolotin, D., Thomas, P. G., Minervina, A. A., Pogorelyy, M. V., Ladell, K., McLaren, J. E., Price, D. A., Nguyen, T. H., Rowntree, L. C., Clemens, E. B., Kedzierska, K., Dolton, G., Rius, C. R., Sewell, A., Samir, J., … Shugay, M. (2022). VDJdb in the pandemic era: A compendium of T cell receptors specific for SARS-CoV-2 [Publisher: Nature Publishing Group]. *Nature Methods 2022 19:9*, *19*(9), 1017–1019. https://doi.org/10.1038/s41592-022-01578-0

Gras, S., Chen, Z., Miles, J. J., Liu, Y. C., Bell, M. J., Sullivan, L. C., Kjer-Nielsen, L., Brennan, R. M., Burrows, J. M., Neller, M. A., Khanna, R., Purcell, A. W., Brooks, A. G., McCluskey, J., Rossjohn, J., & Burrows, S. R. (2010). Allelic polymorphism in the T

cell receptor and its impact on immune responses. *Journal of Experimental Medicine*, *207*(7), 1555–1567. https://doi.org/10.1084/jem.20100603

Gras, S., Saulquin, X., Reiser, J.-B., Debeaupuis, E., Echasserieau, K., Kissenpfennig, A., Legoux, F., Chouquet, A., Le Gorrec, M., Machillot, P., Neveu, B., Thielens, N., Malissen, B., Bonneville, M., & Housset, D. (2009). Structural Bases for the Affinity-Driven Selection of a Public TCR against a Dominant Human Cytomegalovirus Epitope [Publisher: American Association of Immunologists]. *The Journal of Immunology*, *183*(1), 430–437. https://doi.org/10.4049/jimmunol.0900556

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

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

Henderson, J., Nagano, Y., Milighetti, M., & Tiffeau-Mayer, A. (2024). Limits on Inferring T-cell Specificity from Partial Information [arXiv:2404.12565 [cond-mat, q-bio]]. https://doi.org/10.48550/arXiv.2404.12565

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

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

Joshi, K., Robert de Massy, M., Ismail, M., Reading, J. L., Uddin, I., Woolston, A., Hatipoglu, E., Oakes, T., Rosenthal, R., Peacock, T., Ronel, T., Noursadeghi, M., Turati, V., Furness, A. J. S., Georgiou, A., Wong, Y. N. S., Ben Aissa, A., Werner Sunderland, M., Jamal-Hanjani, M., … Chain, B. (2019). Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nature Medicine*, *25*(1549), 1559. https://doi.org/10.1038/s41591-019-0592-2

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab - An S4 package for Kernal Methods in R [ISBN: 9783037851579]. *Journal of Statistical Software*, *11*(9). https://doi.org/10.4028/www.scientific.net/AMR.271-273.389

Keşmir, C., Borghans, J. A., & de Boer, R. J. (2000). Diversity of Human T Cell Receptors [Publisher: American Association for the Advancement of Science]. *Science*, *288*(5469), 1135–1135. https://doi.org/10.1126/science.288.5469.1135a

Korber, B. T., Farber, R. M., Wolpert, D. H., & Lapedes, A. S. (1993). 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. https://doi.org/10.1073/pnas.90.15.7176

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

Leem, J., de Oliveira, S. H. P., Krawczyk, K., & Deane, C. M. (2018). STCRDab: The structural T-cell receptor database. *Nucleic Acids Research*, *46*(D1), D406–D412. https://doi.org/10.1093/nar/gkx971

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

Lineburg, K. E., Grant, E. J., Swaminathan, S., Chatzileontiadou, D. S., Szeto, C., Sloane, H., Panikkar, A., Raju, J., Crooks, P., Rehan, S., Nguyen, A. T., Lekieffre, L., Neller, M. A., Tong, Z. W. M., Jayasinghe, D., Chew, K. Y., Lobos, C. A., Halim, H., Burrows, J. M., … Gras, S. (2021). CD8+ T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope cross-react with selective seasonal coronaviruses [Publisher: Cell Press]. *Immunity*, *54*(5), 1055–1065.e5. https://doi.org/10.1016/j.immuni.2021.04.006

Mayer, A., & Callan, C. G. (2023). Measures of epitope binding degeneracy from T cell receptor repertoires [Publisher: National Academy of Sciences ISBN: 2213264120]. *Proceedings of the National Academy of Sciences*, *120*(4), e2213264120. https://doi.org/10.1073/pnas.2213264120

Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J. C., Souquette, A., Gaevert, J. A., Hertz, T., Thomas, P. G., Bradley, P., & Fiore-Gartland, A. (2021). TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs (B. Chain, A. M. Walczak, B. Chain, & T. Ronel, Eds.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *10*, e68605. https://doi.org/10.7554/eLife.68605

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

Meijers, R., Lai, C.-C., Yang, Y., Liu, J.-h., Zhong, W., Wang, J.-h., & Reinherz, E. L. (2005). 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 [Publisher: Academic Press]. *Journal of Molecular Biology*, *345*(5), 1099–1110. https://doi.org/10.1016/j.jmb.2004.11.023

Milighetti, M. (2023). *Analysis of T cell receptor sequence and structure to understand the drivers of antigen specificity* (PhD thesis). UCL (University College London). Retrieved January 30, 2024, from https://discovery.ucl.ac.uk/id/eprint/10181469/

Milighetti, M., Shawe-Taylor, J., & Chain, B. (2021). 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. https://doi.org/10.3389/fphys.2021.730908

Mora, T., & Walczak, A. M. (2019). How many different clonotypes do immune repertoires contain? [arXiv: 1907.08230 Publisher: Elsevier]. *Current Opinion in Systems Biology*, *18*, 104–110. https://doi.org/10.1016/J.COISB.2019.10.001

Nagano, Y., & Chain, B. (2023). Tidytcells: Standardizer for TR/MH nomenclature. *Frontiers in Immunology*, *14*. Retrieved December 9, 2023, from https://www.frontiersin.org/articles/10.3389/fimmu.2023.1276106

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited [arXiv:physics/0108025]. https://doi.org/10.48550/arXiv.physics/0108025

Oakes, T., Heather, J. M., Best, K., Byng-Maddick, R., Husovsky, C., Ismail, M., Joshi, K., Maxwell, G., Noursadeghi, M., Riddell, N., Ruehl, T., Turner, C. T., Uddin, I., & Chain, B. (2017). Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile [Publisher: Frontiers Media S.A.]. *Frontiers in Immunology*, *8*(OCT), 1267. https://doi.org/10.3389/fimmu.2017.01267

Ochoa, D., & Pazos, F. (2014). Practical aspects of protein co-evolution. *Frontiers in Cell and Developmental Biology*, *2*. Retrieved May 19, 2023, from https://www.frontiersin.org/articles/10.3389/fcell.2014.00014

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, *98*(3), 1064–1072. https://doi.org/10.1152/JN.00559.2007

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python [arXiv: 1201.0490]. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved October 13, 2020, from http://scikit-learn.sourceforge.net.

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

Reiser, J.-B., Legoux, F., Gras, S., Trudel, E., Chouquet, A., Léger, A., Le Gorrec, M., Machillot, P., Bonneville, M., Saulquin, X., & Housset, D. (2014). Analysis of Relationships between Peptide/MHC Structural Features and Naive T Cell Frequency in Humans [Publisher:

American Association of Immunologists]. *The Journal of Immunology*, *193*(12), 5816–5826. https://doi.org/10.4049/JIMMUNOL.1303084

Robinson, R. A., McMurran, C., McCully, M. L., & Cole, D. K. (2021). Engineering soluble T-cell receptors for therapy [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/febs.15780]. *The FEBS Journal*, *288*(21), 6159–6173. https://doi.org/10.1111/febs.15780

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 2.5.*

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

Scott, D. R., Borbulevych, O. Y., Piepenbrink, K. H., Corcelli, S. A., & Baker, B. M. (2011). Disparate Degrees of Hypervariable Loop Flexibility Control T-Cell Receptor Cross-Reactivity, Specificity, and Binding Mechanism [Publisher: Academic Press]. *Journal of Molecular Biology*, *414*(3), 385–400. https://doi.org/10.1016/J.JMB.2011.10.006

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

Shcherbinin, D. S., Belousov, V. A., & Shugay, M. (2020). Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCR complex (F. A. Matsen, Ed.) [Publisher: Public Library of Science]. *PLOS Computational Biology*, *16*(3), e1007714. https://doi.org/10.1371/journal.pcbi.1007714

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the Specificity of Two-Component Signal Transduction Systems [Publisher: Elsevier]. *Cell*, *133*(6), 1043–1054. https://doi.org/10.1016/j.cell.2008.04.040

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

Springer, I., Tickotsky, N., & Louzoun, Y. (2021). Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction [Publisher: Frontiers]. *Frontiers in Immunology*, *0*, 1436. https://doi.org/10.3389/FIMMU.2021.664514

Stadinski, B. D., Trenh, P., Duke, B., Huseby, P. G., Li, G., Stern, L. J., & Huseby, E. S. (2014). Effect of CDR3 Sequences and Distal V Gene Residues in Regulating TCR–MHC Contacts and Ligand Specificity [Publisher: American Association of Immunologists]. *The Journal of Immunology*, *192*(12), 6071–6082. https://doi.org/10.4049/jimmunol.1303209

Strong, S. P., Koberle, R., De Ruyter Van Steveninck, R. R., & Bialek, W. (1998). Entropy and Information in Neural Spike Trains.

Szeto, C., Lobos, C. A., Nguyen, A. T., & Gras, S. (2020). TCR Recognition of Peptide–MHC-I: Rule Makers and Breakers [Publisher: Multidisciplinary Digital Publishing Institute

(MDPI)]. *International Journal of Molecular Sciences*, *22*(1), 68. https://doi.org/10.3390/ijms22010068

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

Thomas, S., Mohammed, F., Reijmers, R. M., Woolston, A., Stauss, T., Kennedy, A., Stirling, D., Holler, A., Green, L., Jones, D., Matthews, K. K., Price, D. A., Chain, B. M., Heemskerk, M. H., Morris, E. C., Willcox, B. E., & Stauss, H. J. (2019). Framework engineering to produce dominant T cell receptors with enhanced antigen-specific function. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12441-w

Tiffeau-Mayer, A. (2023). Unbiased estimation of sampling variance for Simpson's diversity index [arXiv:2310.03439 [cond-mat, q-bio]]. https://doi.org/10.48550/arXiv.2310.03439

Trautmann, L., Labarrière, N., Jotereau, F., Karanikas, V., Gervois, N., Connerotte, T., Coulie, P., & Bonneville, M. (2002). Dominant TCR V usage by virus and tumor-reactive T cells with wide affinity ranges for their specific antigens. *European Journal of Immunology*, *32*(11), 3181–3190. https://doi.org/10.1002/1521-4141(200211)32:11<3181::AID-IMMU3181>3.0.CO;2-2

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

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

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

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

Yu, K., Shi, J., Lu, D., & Yang, Q. (2019). Comparative analysis of CDR3 regions in paired human CD8 T cells [Publisher: Wiley Blackwell]. *FEBS Open Bio*, *9*(8), 1450–1459. https://doi.org/10.1002/2211-5463.12690

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

Supplementary material

Epitope	Number of entries	Unique CDR3α	Unique CDR3β	Number of individuals
ASNENMETM	201	157	111	28
ATDALMTGF	125	99	111	1
AVFDRKSDAK	1699	1552	1584	7
subsample 0	700	661	670	7
subsample 1	700	663	671	6
subsample 2	700	657	662	5
subsample 3	700	653	679	6
subsample 4	700	657	672	6
CINGVCWTV	226	214	218	7
ELAGIGILTV	380	348	370	14
GILGFVFTL	1894	1082	980	56
subsample 0	700	477	442	48
subsample 1	700	487	460	50
subsample 2	700	488	427	50
subsample 3	700	464	436	50
subsample 4	700	475	423	44
GLCTLVAML	345	228	239	20
HGIRNASFI	243	187	180	13
IVTDFSVIK	704	526	517	7
KSKRTPMGF	103	85	70	1
LLWNGPMAV	235	202	220	2
LSLRNPILV	127	109	111	12
LTDEMIAQY	124	115	119	2
NLVPMVATV	357	305	318	48
RAKFKQLL	1200	649	650	3
subsample 0	700	411	407	3
subsample 1	700	417	405	3
subsample 2	700	410	411	3
subsample 3	700	402	420	3
subsample 4	700	419	406	3
RLRAEAQVK	412	389	394	4
SSLENFRAYV	350	240	286	19
SPRWYFYYL	175	169	174	10
SSPPMFRV	133	100	58	14
SSYRRPVGI	177	148	153	32
TTDPSFLGRY	242	231	232	1
YLQPRTFLL	333	284	304	10

Table S1: **Summary of VDJDb set used for the pairing algorithms.** For large epitopes that were subsampled for pairing, composition of each subsample is shown.

	N	Largest ID size	Repertoire MI (nats)	R^2	R^2_{adj}	F-score p-value
No-learning background: confidence = none; $\lambda = 1.0$; $\theta = 0.6$	0.1236 (*)	-0.0845	0.0140	0.232	0.104	0.180
Learning model: confidence = none; $\lambda = 0.6$; $\theta = 0.6$	0.1317	-0.1081 (*)	0.0445	0.294	0.177	0.092
Theoretical best: confidence = none; $\lambda = 0.6$; $\theta = 0.6$	-0.1362	-0.0789	0.0292	0.545	0.469	0.002

Table S2: **Factors affecting MI-IPA performance.** A multivariate linear regression was calculated to evaluate the effect repertoire size (N), largest individual (Largest ID size) and total mutual information (Repertoire MI, calculated as in Figure 5 between CDR3α and CDR3β) on the precision. Subsamples from large repertoires are included as an average, and the repertoire MI is calculated on the complete sample. Each independent variable was normalised by subtracting the mean and dividing by the mean to make the derived coefficients more comparable ($x_{norm} = (x - \bar{x})/\bar{x}$). The R^2 and adjusted R^2 (R^2_{adj}) for the regression are shown, as well as the F-score p-value. P-values associated with each coefficient are shown as asterisks: ***< 0.001; $0.001 \leq$**< 0.01; $0.01 \leq$*< 0.05.

	N	Largest ID size	Average degree of node	Number of large clusters	Proportion of singlets	R^2	R^2_{adj}	F-score p-value
GA: distance=lev; k=20	-0.3507 (*)	-0.1282	α: 0.0608 β: 0.1062 (**)	α: 0.3761 β: 0.0414	α: 0.2471 β: -0.0251	0.799	0.675	0.002

Table S3: **Factors affecting GA performance.** A multivariate linear regression was calculated to evaluate the effect repertoire size (N), largest individual (Largest ID size) and similarity network characteristics on the precision for each model. The graph characteristics measured are: average degree of nodes, number of clusters of size ≥ 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 both CDR3α and CDR3β. Epitopes for which 5 subsamples are available are averaged across subsamples, and each of the metrics is calculated on the specific subsample. Each independent variable was normalised by subtracting the mean and dividing by the mean to make the derived coefficients more comparable ($x_{norm} = (x - \bar{x})/\bar{x}$). The R^2 and adjusted R^2 (R^2_{adj}) for the regression are shown, as well as the F-score p-value. P-values associated with each coefficient are shown as asterisks: ***< 0.001; $0.001 \leq$**< 0.01; $0.01 \leq$*< 0.05.

	N	Largest ID size	Repertoire MI (nats)	Number of stable pairs in GA	% correct in GA stable	R^2	R^2_{adj}	F-score p-value
GA+MI-IPA model precision: distance=lev; k=20; GA-thresh=0.95; re-pair=True	-0.0302	-0.0027	0.0447 (*)	0.0802 (***)	0.0520 (*)	0.833	0.780	10^{-5}
precision increase from MI-IPA: precision of GA+MI-IPA − precision of MI-IPA	-0.0095	-0.0081	-0.0004	0.0010	0.0100	0.143	-0.125	0.748
precision ratio between GA+MI-IPA and MI-IPA: $\log_2(\frac{\text{precision of GA+MI-IPA}}{\text{precision of MI-IPA}})$	-0.2914	0.0458	0.1081	-0.0409	0.1188	0.060	-0.234	0.957

Table S4: **Factors affecting GA+MI-IPA performance.** A multivariate linear regression was calculated to evaluate the effect repertoire size (N), largest individual (Largest ID size), total mutual information (Repertoire MI, calculated as in Figure 5 between CDR3α and CDR3β), number of stable GA pairs and proportion of the GA training set that is correct (% correct in GA stable) on the precision. Each independent variable was normalised by subtracting the mean and dividing by the mean to make the derived coefficients more comparable ($x_{norm} = (x - \bar{x})/\bar{x}$). The R^2 and adjusted R^2 (R^2_{adj}) for the regression are shown, as well as the F-score p-value. P-values associated with each coefficient are shown as asterisks: ***$<$ 0.001; 0.001 $\leq$**$<$ 0.01; 0.01 $\leq$*$<$ 0.05.

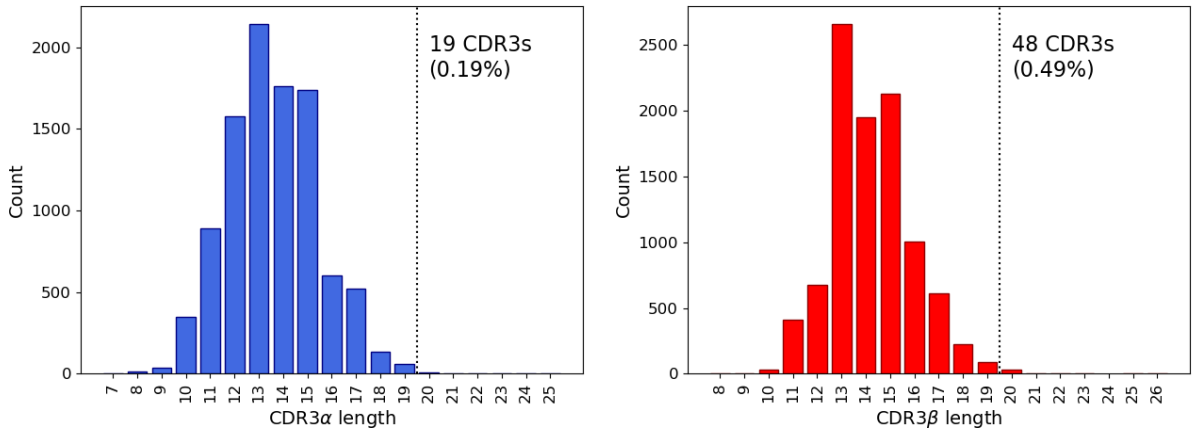

Figure S1: **Length of CDR3α and CDR3β sequences in VDJDb dataset.** Distribution of CDR3 lengths (without padding) in the α (left) and β (right) chains. The dotted line shows the length threshold chosen as maximum allowed length, and the text indicates how many CDR3s were removed and what percentage they correspond to.

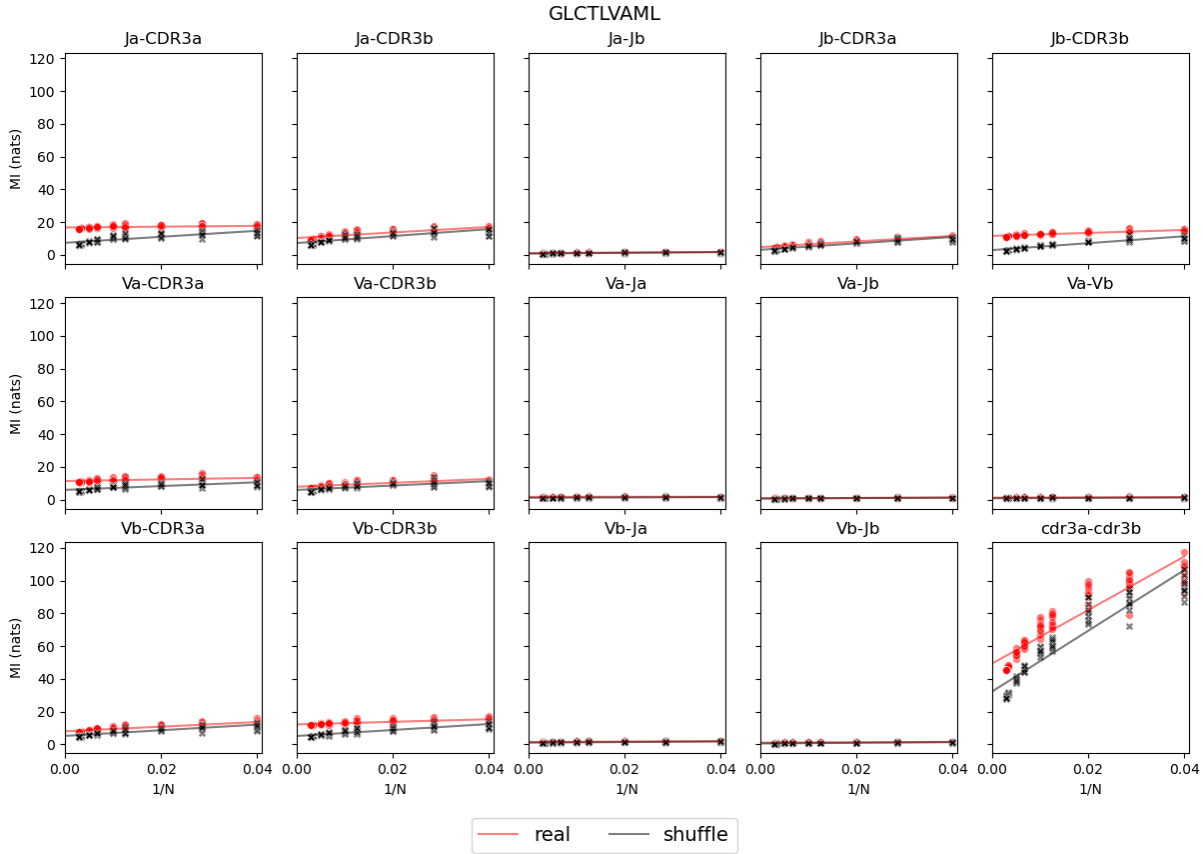

Figure S2: **Estimation of mutual information for epitope GLCTLVAML.** For each pair of TCR components (indicated in the title of each plot), mutual information (MI) was calculated at multiple subsamples for both the real set (red) and a shuffle (black). Each subsample was repeated 10 times. A line was fit through the point and MI was estimated as the y-intercept for both real set and shuffle.

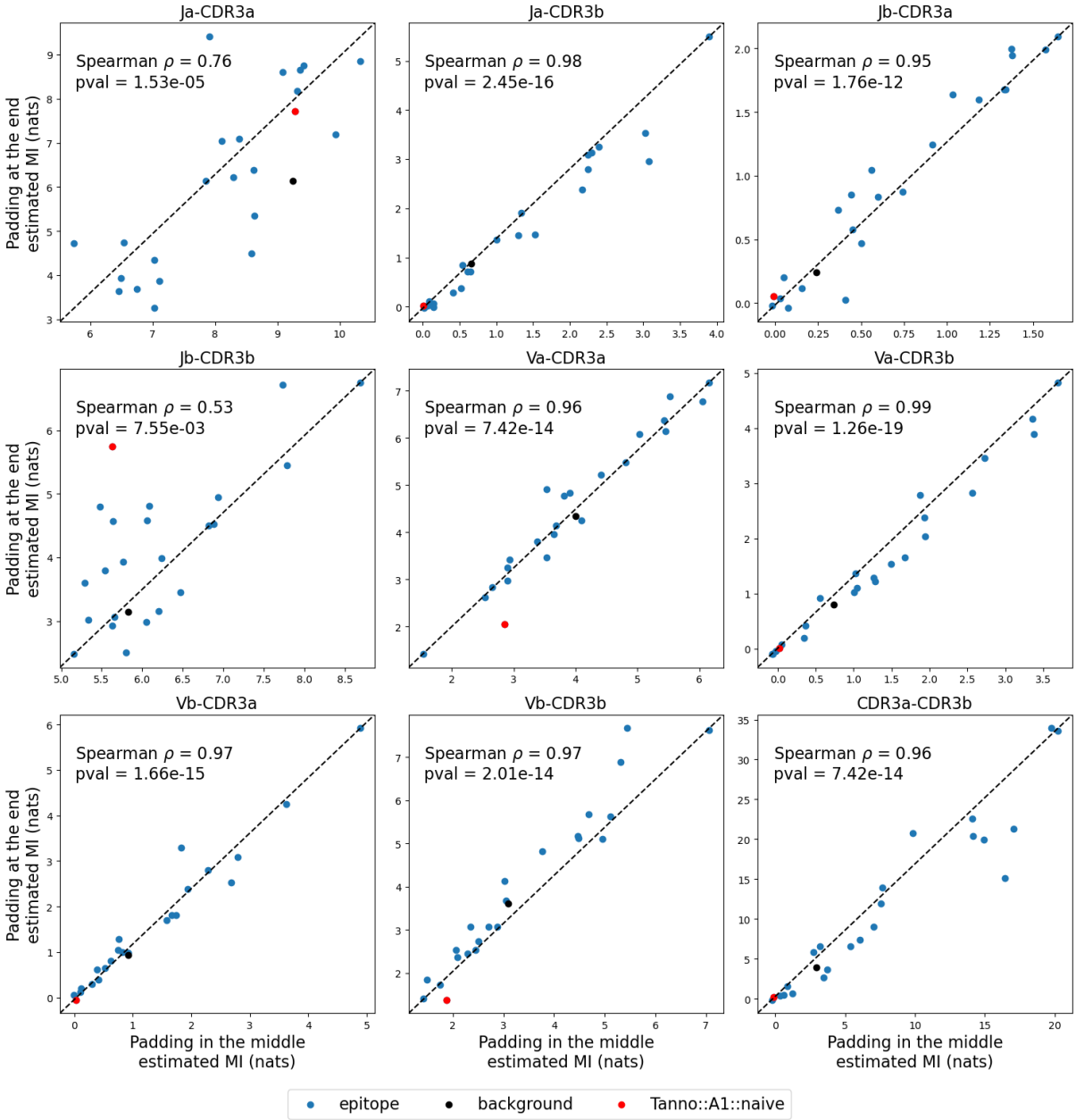

Figure S3: **Comparison of estimated MI when different padding methods are used.** For each pair of variables, the comparison of the MI estimated on CDR3s with padding in the middle (position $\frac{L}{2}$ where L is the length of the CDR3 sequence) or at the end of each sequence is shown. The $y = x$ diagonal is shown with a dashed line. Only pairs of variables including the CDR3 are shown in the plot, as V and J MI estimation is not affected by the padding strategy.

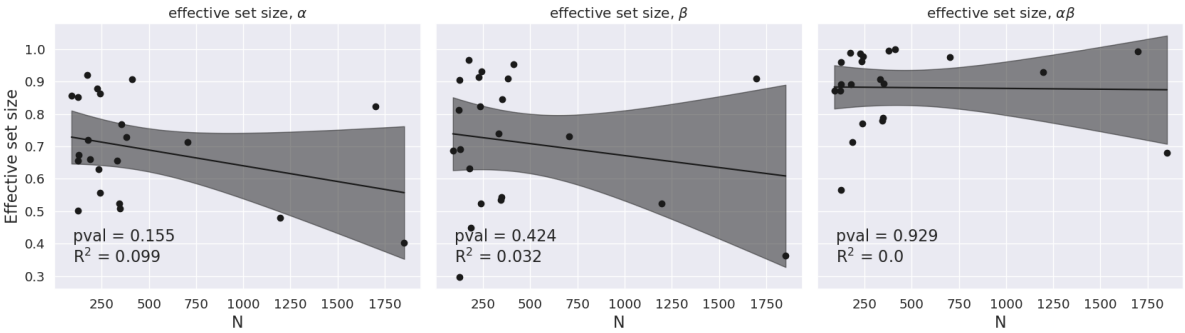

Figure S4: **Correlation of repertoire size and effective set size.** For each epitope, CDR3α, CDR3β and CDR3$\alpha\beta$ effective set size is correlated with total repertoire size (N). Effective set size was calculated by grouping all sequences with Hamming distance of < 2 on each chain or < 3 for the paired chains, normalised by N. Correlation was calculated by fitting a linear regression. p-value for the F-test and R^2 was calculated for each fit. The shaded area represents the 95% CI of the fit.

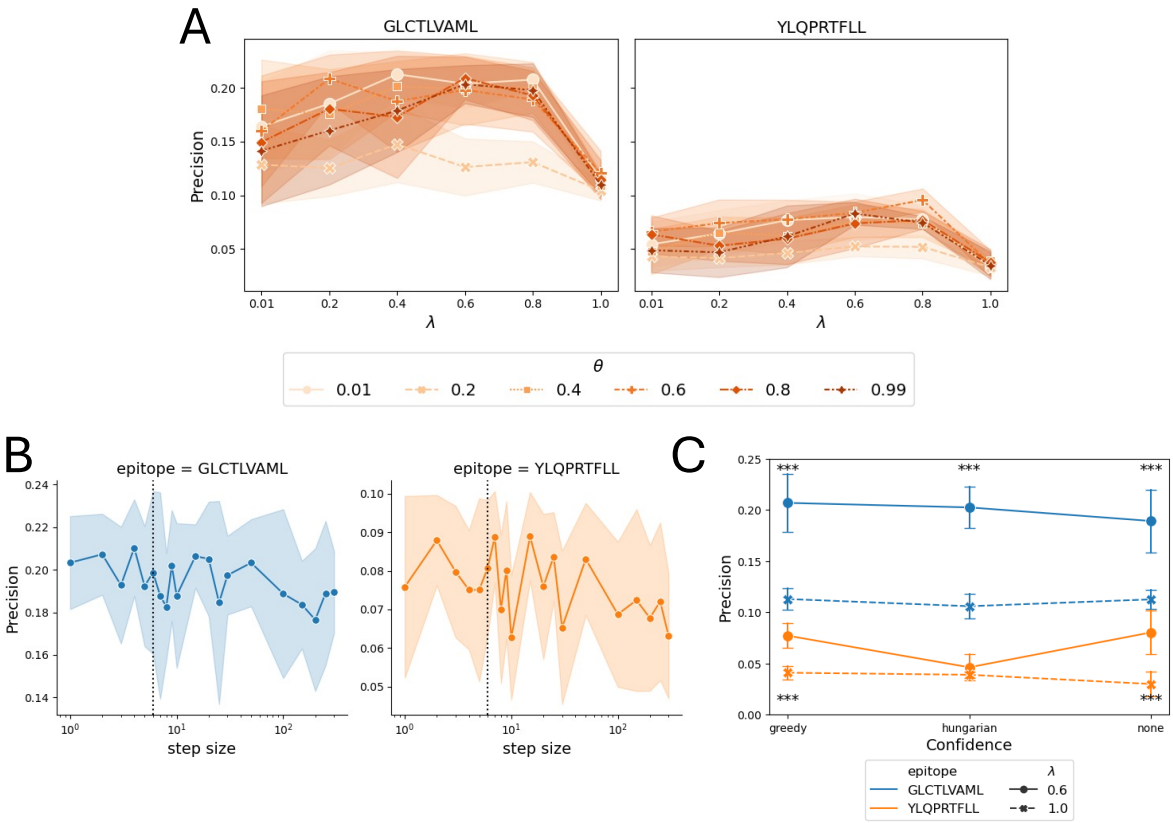

Figure S5: **Optimisation of parameters for MI-IPA.** All parameters are as in Bitbol, 2018, except that we add a scenario where no 'confidence' is calculated. This is a greedy assignment (as in "greedy"), but no confidence score is calculated for the pairing. Rather, the scoring for each pairing is used raw to rank the pairs. Shaded areas or error bars represent the standard deviation around the mean (solid line) for 10 repeats of each model. **A)**: θ and λ (controlling sampling correction and effective set size calculation, respectively); **B)**: step size; **C)**: confidence calculation algorithms. One-sided t-tests are calculated to compare performance of each model to its no-learning scenario (achieved by setting $\lambda = 1$). pvalues: ***< 0.001; $0.001 \leq$**< 0.01; $0.01 \leq$*< 0.05.

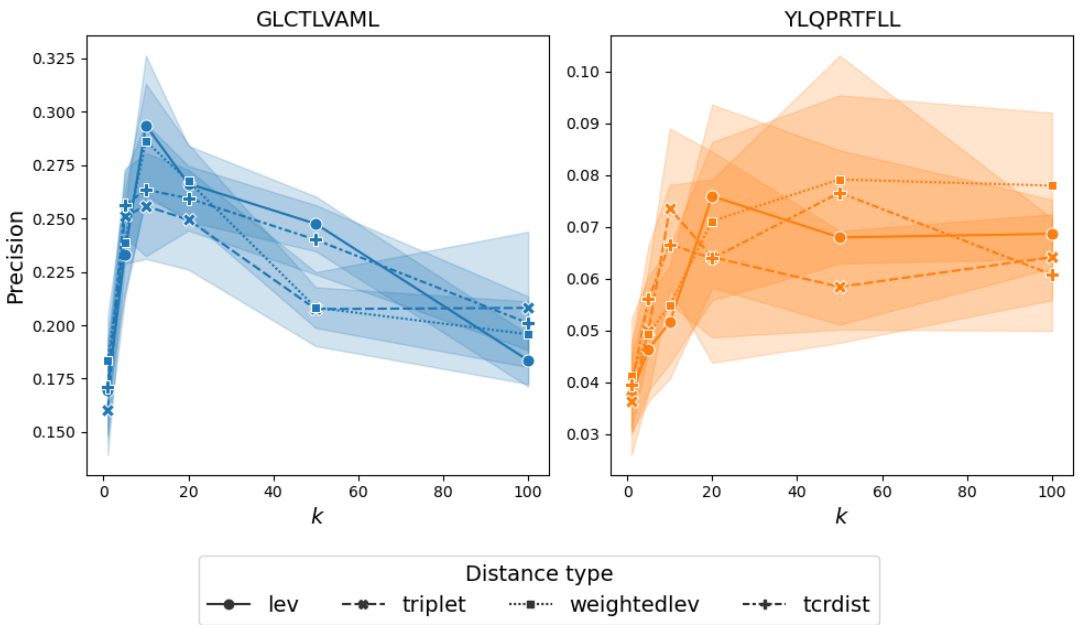

Figure S6: **Impact of k and distance function choice on GA pairing.** The GA was run with a range of k and 4 different distance metrics to evaluate their impact on the two benchmarking epitopes. Precision was calculated for each setting. The shaded areas show the standard deviation around the mean (solid line) for 100 repeats of the model.

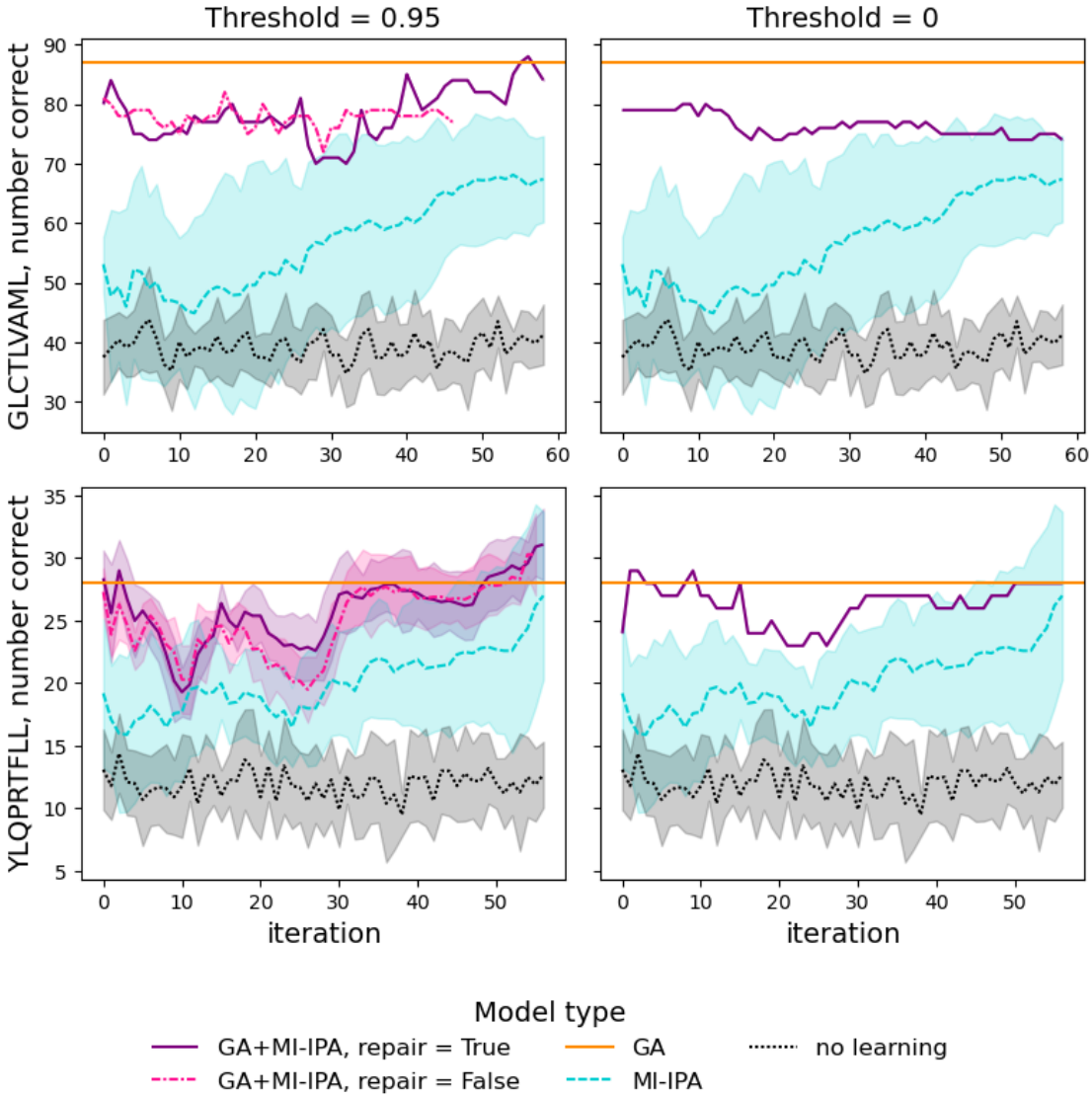

Figure S7: **Learning of the GA+MI-IPA over iterations with different settings.** Performance of combining the GA (Levenshtein distance, $k = 20$) with MI-IPA ($\theta = 0.6$, no confidence, step size = 6) is shown for the two benchmarking epitopes. The orange horizontal line shows the number of correct pairs in the GA consensus assignment which was used to select the training sets. The GA assignment is used as training for the MI-IPA, either by taking the complete assignment (threshold = 0, right column), or only the stable pairs (threshold = 0.95, left column). The MI-IPA is then allowed to pair all remaining β to all remaining α (pink dotted line with fewer iterations) or pair all α and β, including the ones in the training set (purple line). Performance of the MI-IPA on its own is also shown, with $\lambda = 0.6$ (cyan line) or $\lambda = 1$ (no learning, black dotted line). Shaded areas show the standard deviation around the mean (solid line).

A. Correct $\alpha\beta$ pairs with ID

α	β	ID
CASSALASLNEQFF	CASSS--LNTQYF	A
CASSFG-VEDEQYF	CASSS--LNTQYF	B
CASSALASLNEQFF	CASSIRSTDTQYF	B

B. Predicted $\alpha\beta$ pairs without ID

α	β	correct
CASSALASLNEQFF	CASSS--LNTQYF	True
CASSALASLNEQFF	CASSS--LNTQYF	True
CASSFG-VEDEQYF	CASSIRSTDTQYF	False

C. Predicted $\alpha\beta$ pairs with ID

α	β	ID	correct
CASSALASLNEQFF	CASSS--LNTQYF	A	True
CASSALASLNEQFF	CASSS--LNTQYF	B	False
CASSFG-VEDEQYF	CASSIRSTDTQYF	B	False

Figure S8: **Example assignment with repeat sequences and ID. A)** Correct pairs to be found. The blue α appears twice in the two individuals, but paired with two different β. The red β also appears twice, once paired with the blue α. **B)** Pairs assigned by the algorithm, disregarding individual information. Here, the blue α/red β pair appears twice. In the precision calculation, this will be counted twice as correct, giving a precision of 0.66. **C)** Pairs assigned by the algorithm, including individual information. Here, the blue α/red β pair appears twice. Because the blue α/red β pair does not appear in individual B, the pair is considered correct only in individual A.

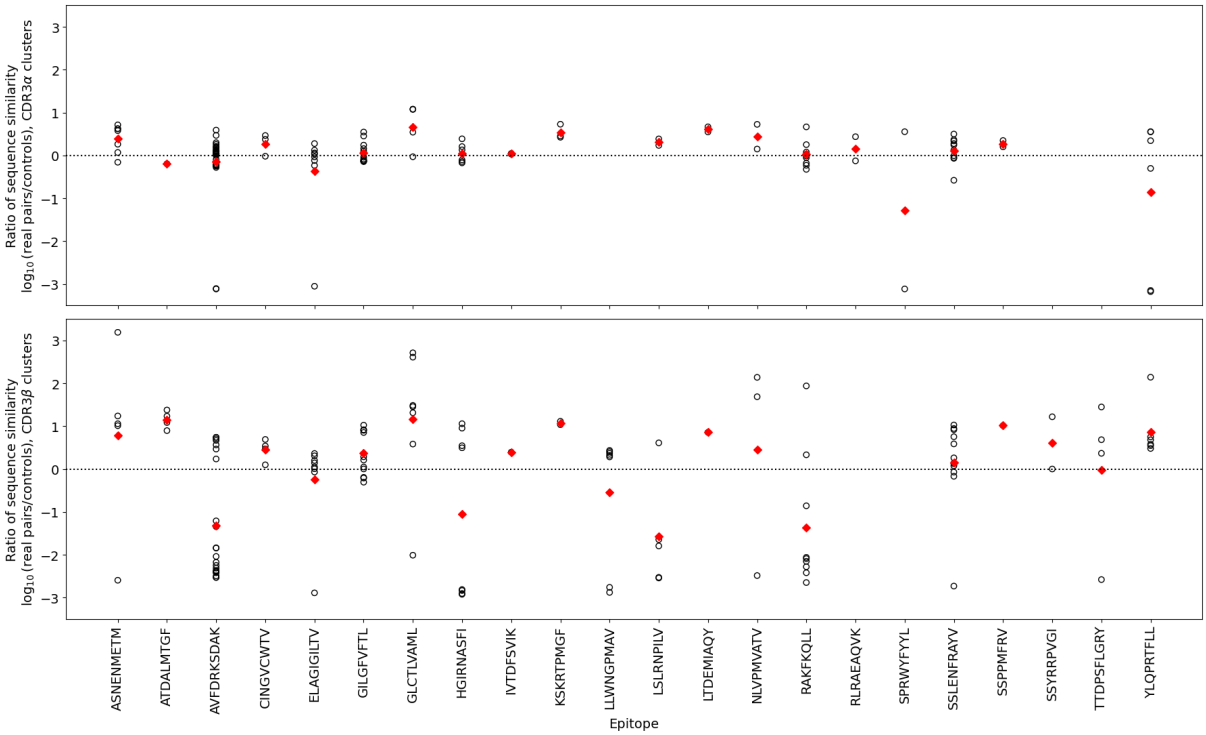

Figure S9: **Ratio of sequence similarity between sequences paired to real clusters and random samples from epitope repertoires.** For each cluster (of CDR3α, top and of CDR3β, bottom) in each epitope repertoire, the similarity between the paired CDR3s to the sequences in each cluster is calculated. 100 random samples of the same size from the same epitope-specific repertoire are also taken and their similarity calculated. The ratio between the median sequence similarity of the real pairs and of the random controls is shown (black circles, average median control similarity taken over all controls, ratio shown as $\log_{10}$, 0s imputed as 10^{-4} to be able to calculate the ratio and the $\log_{10}$ of the ratio). The red diamonds show the average across all clusters for that epitope.

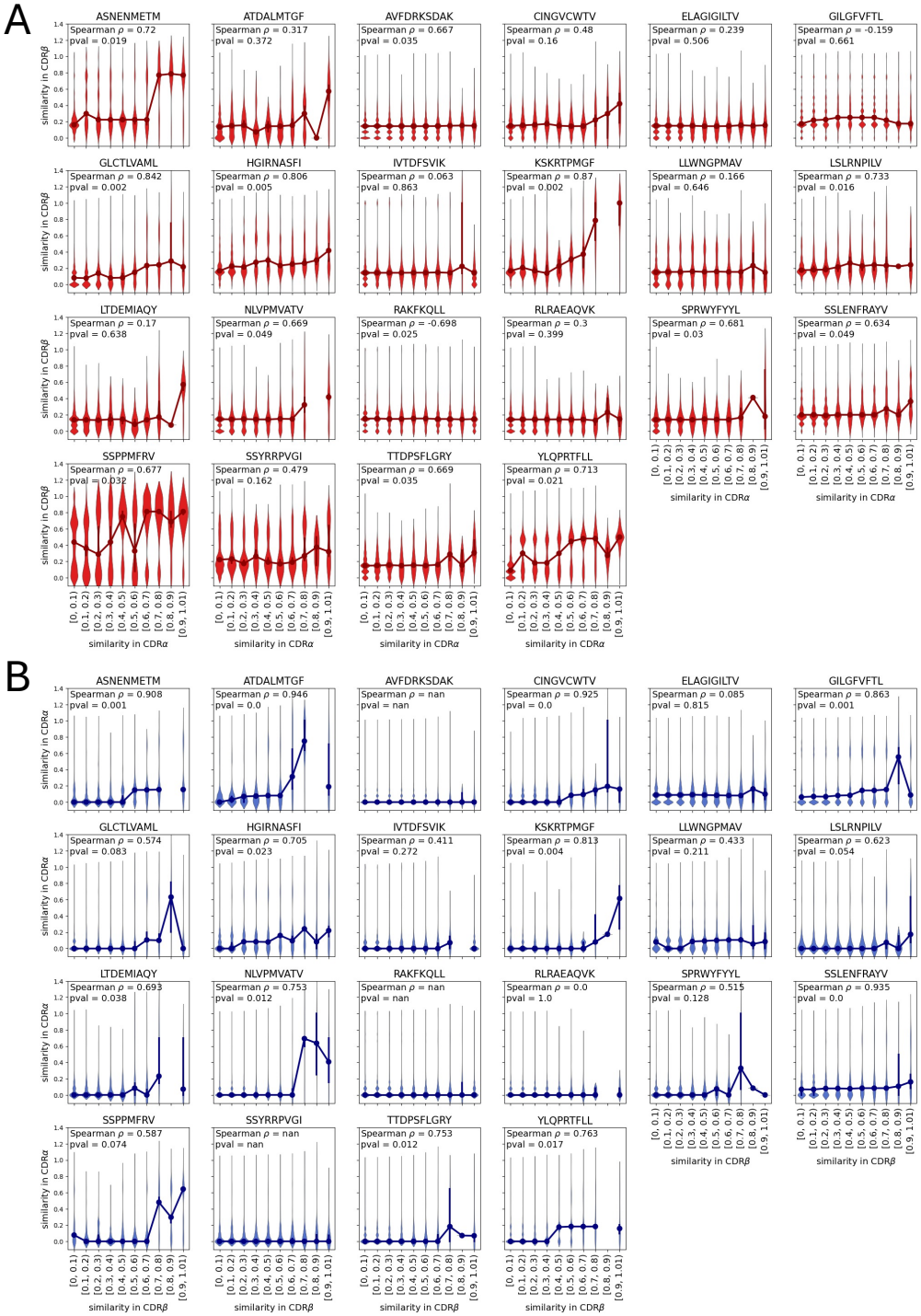

Figure S10: **Correlation of CDR3 pairwise sequence similarity between the two chains.** Pairwise similarity was calculated between all CDR3α and all CDR3β annotated for each epitope (including duplicate sequences). The pairwise similarity on the α chains was then binned and the similarity of the paired CDR3β examined (**A**, vice versa in **B**). The large circles joined by a solid line represent the median for each distribution, whilst the violin plots show the distribution of the underlying distances. Spearman correlation was calculated using the mid-point of each bin, and correlating with the median of the values for that bin.

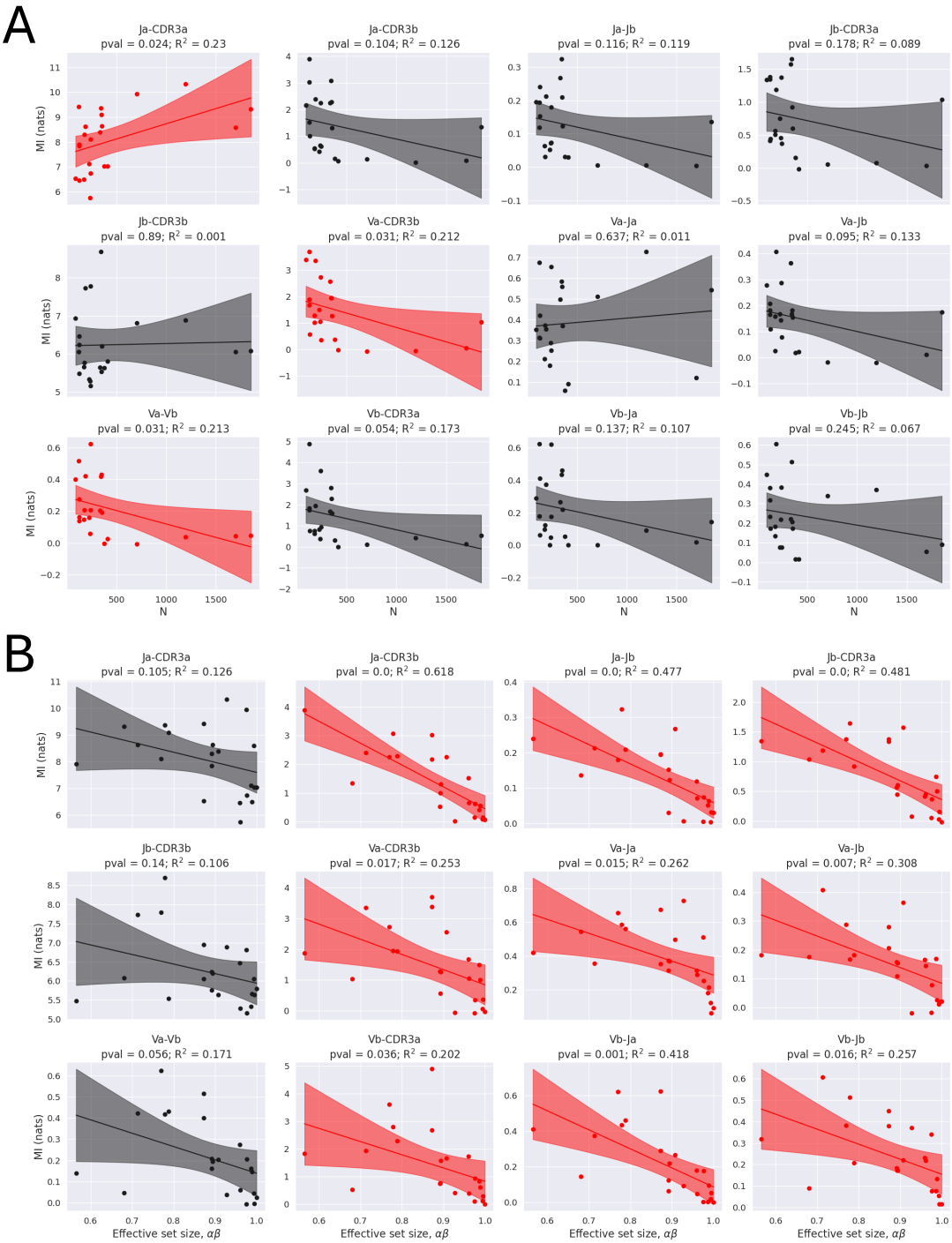

Figure S11: **Correlation between mutual information and epitope repertoire size and mutual information and effective set size.** For each epitope and each pair of variables the mutual information (MI) was estimated, correcting for finite size effects (see Methods). The correlation between MI and epitope repertoire size (N, **A**) or $\alpha\beta$ effective set size (**B**) is shown for each pair of variables. Effective set size was calculated by grouping all sequences with Hamming distance < 3 on the paired chain, normalised by N. Correlation was calculated by fitting a linear regression. p-value for the F-test and R^2 was 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.

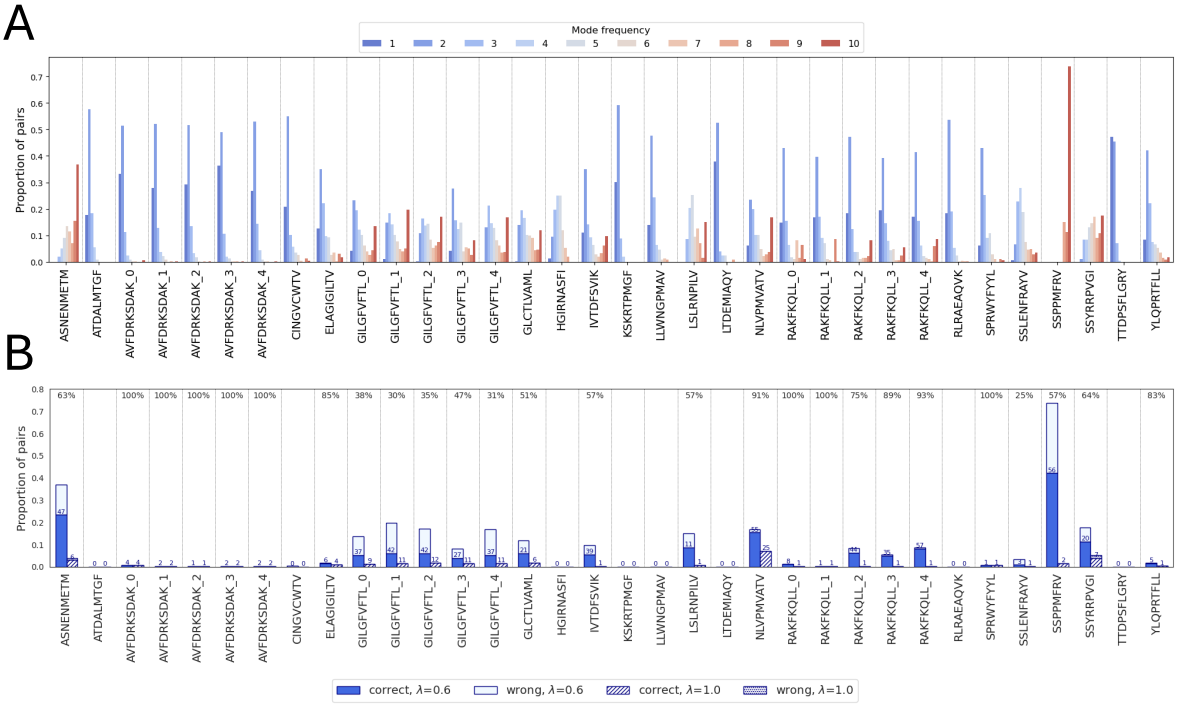

Figure S12: **Stability of the pairs assigned by the MI-IPA. A)** The bar chart shows the stability of the assignments made for each epitope. The x-axis shows the number of times an assignment is made across 10 repeats (i.e. the frequency of the mode, also colour-coded) and the y-axis shows the proportion of assignments that have that mode. The bar graph is repeated for each epitope, and for large epitopes all five subsamples are shown. The results shown correspond to pairing with $\lambda = 0.6$. **B)** The right-most bar for each epitope in **A** is broken down to show the proportion of stable assignments that are correct (with $\lambda = 0.6$ or $\lambda = 1$). The number on top of each bar indicates the number of correct pairs that were selected 10 times. The number at the top of each section indicates the percentage of correctly assigned pairs among pairs that are stably selected across repeats. Of note, some epitopes show stable pairs also in the chance expectation ($\lambda = 1$) scenario. These arise from individuals of size 1, i.e. examples where only a single α and a single β are available from an individual, which will always be assigned correctly independently of how well the model is learning.

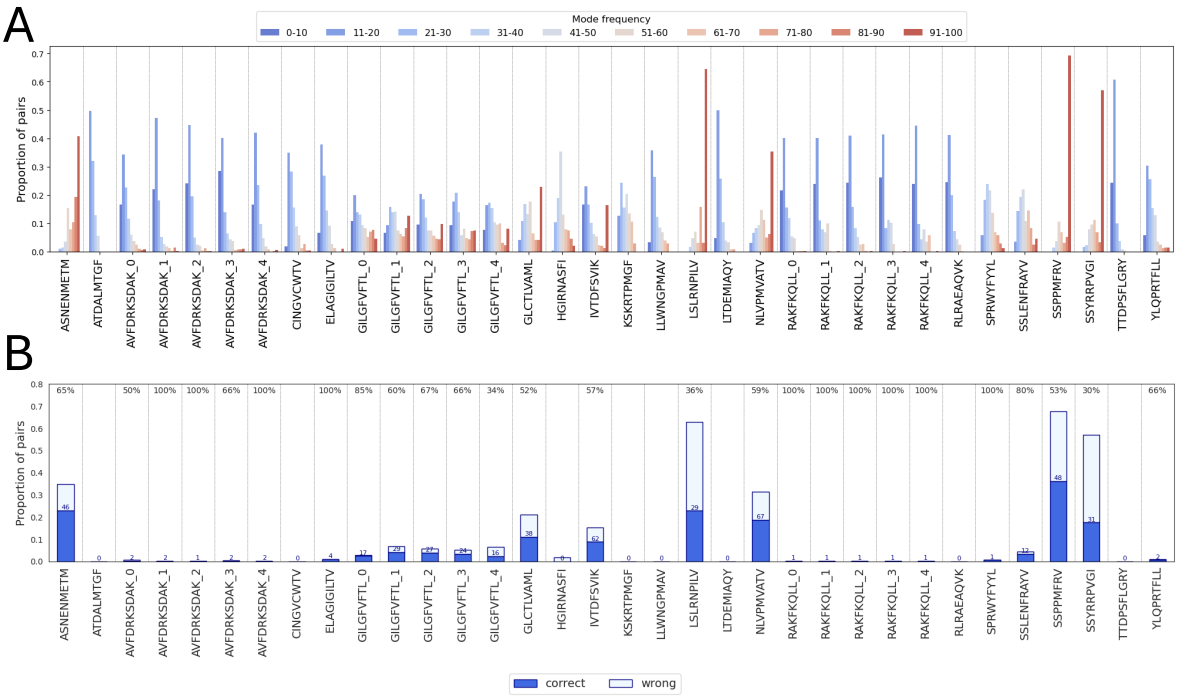

Figure S13: **Stability of the pairs assigned by the GA. A)** The bar chart shows the stability of the assignments made for each epitope. The x-axis shows the number of times an assignment is made across 100 repeats (i.e. the frequency of the mode, also colour-coded) and the y-axis shows the proportion of assignments that have that mode. The bar graph is repeated for each epitope, and for large epitopes all five subsamples are shown. The results shown correspond to GA run with $k = 20$ on Levenshtein distance. **B)** The right-most bar for each epitope in **A** is broken down to show the proportion of stable assignments that are correct. The number on top of each bar indicates the number of correct pairs that were selected ≥ 95 times. The number at the top of each section indicates the percentage of correctly assigned pairs among pairs that are stably selected across repeats.

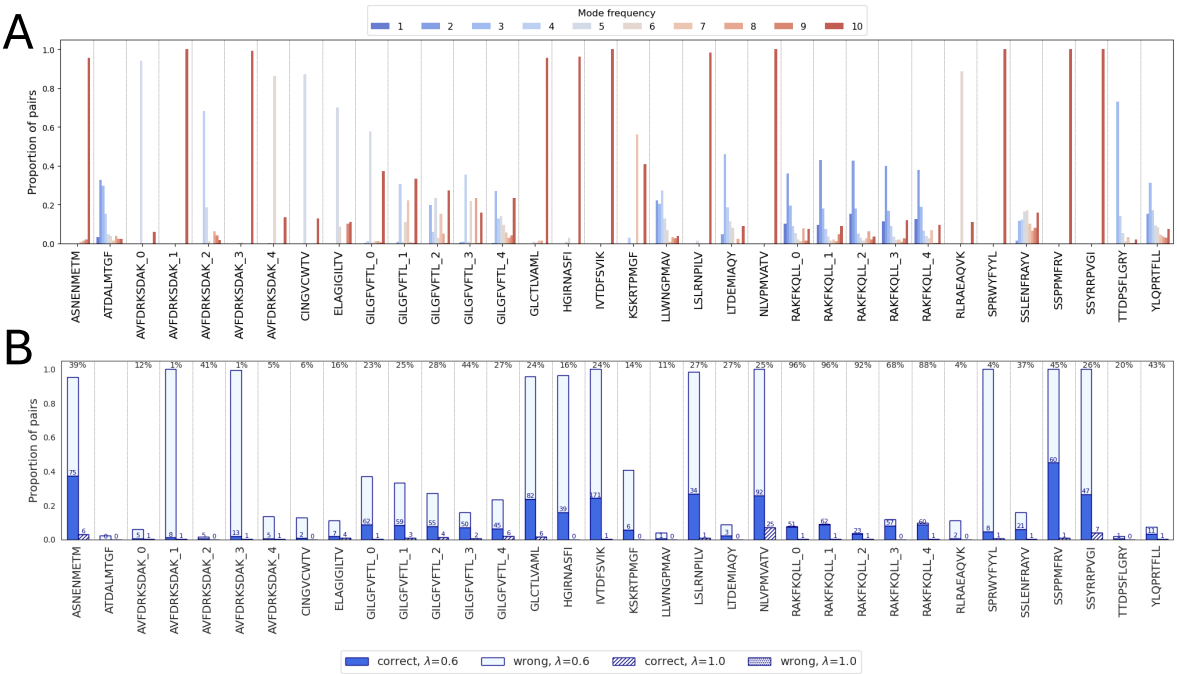

Figure S14: **Stability of the pairs assigned by the GA+MI-IPA. A)** The bar chart shows the stability of the assignments made for each epitope. The x-axis shows the number of times an assignment is made across 10 repeats (i.e. the frequency of the mode, also colour-coded) and the y-axis shows the proportion of assignments that have that mode. The bar graph is repeated for each epitope, and for large epitopes all five subsamples are shown ($\lambda = 0.6$, golden set from the GA, re-pairing allowed). **B)** The right-most bar for each epitope in **A** is broken down to show the proportion of stable assignments that are correct (MI-IPA run with $\lambda = 0.6$ or $\lambda = 1$, golden set from the GA, re-pairing allowed). The number on top of each bar indicates the number of correct pairs that were selected 10 times. The number at the top of each section indicates the percentage of correctly assigned pairs among pairs that are stably selected. Of note, some epitopes show stable pairs also in the chance expectation ($\lambda = 1$) scenario. These arise from individuals of size 1, i.e. examples where only a single α and a single β are available from an individual, which will always be assigned correctly independently of how well the model is learning.