PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY
IMMUNOLOGY AND INFLAMMATION OPEN ACCESS

Limits on inferring T cell specificity from partial information

James Henderson[a,b], Yuta Nagano[a,c], Martina Milighetti[a,d], and Andreas Tiffeau-Mayer[a,b,1]

Edited by Herbert Levine, Northeastern University, Boston, MA; received May 1, 2024; accepted September 3, 2024

A key challenge in molecular biology is to decipher the mapping of protein sequence to function. To perform this mapping requires the identification of sequence features most informative about function. Here, we quantify the amount of information (in bits) that T cell receptor (TCR) sequence features provide about antigen specificity. We identify informative features by their degree of conservation among antigen-specific receptors relative to null expectations. We find that TCR specificity synergistically depends on the hypervariable regions of both receptor chains, with a degree of synergy that strongly depends on the ligand. Using a coincidence-based approach to measuring information enables us to directly bound the accuracy with which TCR specificity can be predicted from partial matches to reference sequences. We anticipate that our statistical framework will be of use for developing machine learning models for TCR specificity prediction and for optimizing TCRs for cell therapies. The proposed coincidence-based information measures might find further applications in bounding the performance of pairwise classifiers in other fields.

TCR | immune repertoire | information theory | Renyi information | receptor-ligand interaction

Mapping the amino acid sequence of a particular T cell receptor (TCR) to its antigen specificity is a holy grail of systems immunology (1–3). The T cell receptor endows T cells with the ability recognize snippets of pathogenic material presented on the surface of antigen-presenting cells by major histocompatibility complexes (MHC) (4). TCRs are specific, meaning a given T cell will only activate in response to a select range of antigen stimuli. Coverage of the vast antigen space explored by evolving pathogens is enabled by immense sequence variation within the TCR (5, 6), in particular within six hypervariable loops of the heterodimeric receptor, named complementarity determining regions (CDRs).

The immense diversity of TCRs implies that many have no experimentally determined ligands (7). Emerging computational approaches predict the specificity of such orphan TCRs by their sequence similarity to annotated TCRs (1, 3, 8, 9). However, which level of partial matching is sufficient for reliable prediction has remained unclear. Moreover, there is substantial interest in understanding for which immunological questions knowledge of paired receptor chains obtainable by single-cell sequencing is worth the trade-off with the higher throughput achievable by bulk sequencing (10) and which TCR features are most informative for machine learning applications (3, 11).

Here, we address these important open questions by putting universal limits on the accuracy with which TCR specificity can be predicted from partial information. Our work takes inspiration from a long history of successful applications of information theory to the study of complex biological input–output relationships from neural coding (12–14) and transcriptional regulation (15, 16) to pattern formation during embryo development (17–19). Following recent applications of information theory to TCR repertoires by us (20) and others (21), our analysis builds on a fundamental insight from evolutionary biology: Patterns of sequence conservation in protein families provide clues about functionally relevant properties. In the immunological context, this means that TCR features that are important for specific recognition of a particular epitope will often be highly conserved among epitope-specific TCRs relative to their global diversity (Fig. 1).

In our current work, we provide the first comprehensive map of how much information each section of the paired chain TCR sequence provides about its specificity. To provide such a map, we make use of two recent datasets that have sequenced TCRs specific to a dozen viral MHC class I epitopes (1, 22). We overcome statistical limitations of prior analyses to pairs of residues (20, 21, 23) using coincidence-based measures of repertoire diversity (24). These measures can be estimated from smaller samples than traditional measures based on Shannon entropy (25–27). The information-theoretic approach naturally allowed us to identify synergies between different TCR sections in determining antigen specificity. Importantly, our quantification of coincidence

Significance

The specificity of cellular immune responses is determined by the binding of T cell receptors (TCRs) to diverse ligands, yet due to their vast diversity, most TCRs lack experimentally validated binding partners. To overcome this gap requires understanding the recognition code linking receptors and ligands. Here, we introduce an information theoretic approach to rank TCR features by their relevance to predicting specificity and bound how accurately T cell specificity can be predicted from partial information. By identifying informative features, our work provides a rational basis for prioritizing matches in TCR databases and for developing machine learning models to predict TCR–ligand interactions.

Author affiliations: [a]Division of Infection and Immunity, University College London, London WC1E 6BT, United Kingdom; [b]Institute for the Physics of Living Systems, University College London, London WC1E 6BT, United Kingdom; [c]Division of Medicine, University College London, London WC1E 6BT, United Kingdom; and [d]Cancer Institute, University College London, London WC1E 6DD, United Kingdom

Author contributions: J.H., Y.N., M.M., and A.T.-M. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: andreas.mayer@ucl.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2408696121/-/DCSupplemental.

Published October 7, 2024.

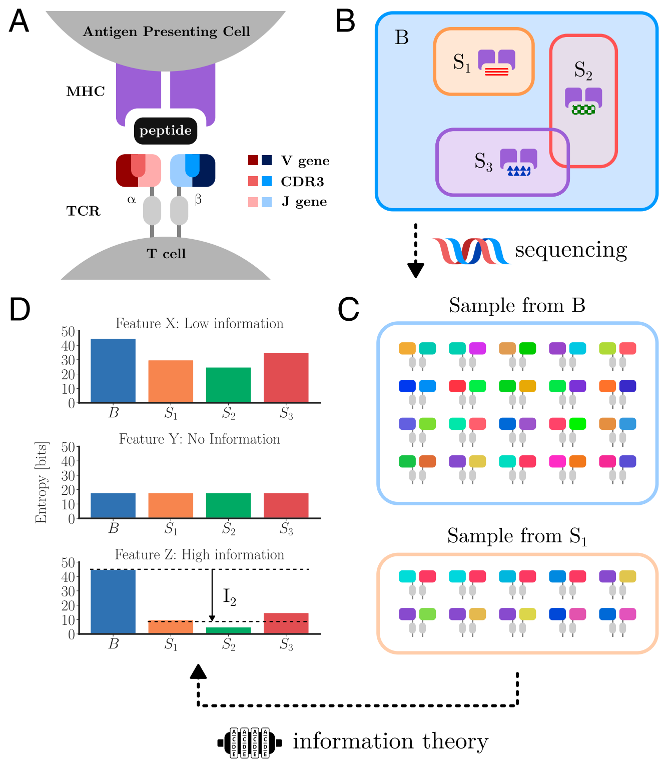

Fig. 1. Overview of analysis methodology. (*A*) Sketch of T cell receptor structure highlighting the V, CDR3, and J regions and their interaction with MHC-bound peptides. The TCR is composed of two chains, most commonly α and β chains. Each chain is generated by the process of V(D)J recombination during T cell development, which combines a V (variable), J (joining), and C (constant) gene, with the addition of a D (diversity) gene in the β chain. Within each chain, the CDR1 and CDR2 amino acid loops are coded for by the V gene while the CDR3 regions are at the V(D)J intersection, which is additionally diversified through the random insertion and deletion of nucleotides at gene template junctions. (*B*) An abstracted view of TCR sequence space. The set B includes all possible TCRs. The subsets S_i represent TCRs specific to particular ligands. (*C*) Sequencing TCR from either the whole repertoire or epitope-specific subsets gives us samples from their respective distributions. (*D*) The number of pairs which match in a particular feature may then be recorded to compute a probability of coincidence. The logarithm of the probability of coincidence gives a measure of the entropy of the feature. Our information theoretic approach quantifies the change in entropy between background TCRs and sets of specific TCRs of different features (*Top* to *Bottom*). Features which experience a large reduction in entropy (*Bottom*) are the most informative for predicting the epitope specificity of a sequence.

information is underpinned by theory that directly links achievable classification accuracy to the coincidence information gained from a partial match and prior beliefs about the prevalence of epitope-specific T cells in a repertoire.

1. An Information-Theoretic Approach to T Cell Specificity

1.1. Coincidence Analysis for Features. We have recently introduced a coincidence-based statistical framework to measure antigen-driven selection in TCR repertoires (24). The main idea of this work was to quantify clonal convergence by counting how often pairs of independently recombined clonal lineages in a sample have TCRs that are more similar to each other than some threshold level. Here, we pursue a conceptually related but unique approach that considers near-coincidences as coincidences on the level of coarse-grained TCR features. A feature may be a gene segment choice at a given locus, an amino acid at a particular residue, or a physical property of a hypervariable loop such as its charge or length. Features may also contain other features such as the α chain containing the Vα, Jα, and the CDR3α as component features.

Mathematically, a feature is a random variable that maps the sample space of all TCR sequences to a discrete set of possible categories. We denote the measure on the feature set for randomly drawn TCRs from a repertoire by $P(X)$. The probability that two independent draws return the same outcome, i.e., the probability of coincidence of X, is then defined by

$$p_C[X] = \sum_x P(x)^2, \qquad [1]$$

where $P(x)$ represents $P(X = x)$ and the sum runs over all possible outcomes of X. We recall that in ecology, $p_C[X]$ is referred to as the Simpson's diversity index of X with $D_2[X] = 1/p_C[X]$ being an effective number of distinct species in a population (28).

Intuitively, we expect the most informative features to be those whose diversity is most reduced among TCRs specific to the same epitope when compared with background TCRs. In the following, we will make this intuition mathematically precise using a coincidence-based formulation of information theory.

We note that feature importance in this information-theoretic sense is not necessarily synonymous with the biophysical importance of a feature for the receptor–ligand interaction, as there are often multiple binding solutions for a given ligand (1, 24): TCR properties involved in binding but variable across solutions might not be globally informative in the way considered here. We will revisit the impact of the multiplicity of binding solutions on our information-theoretic measures in Section 3.4.

1.2. Coincidence Entropy. A central quantity in information theory is entropy. The entropy of a probability distribution $P(X)$ is given in its form proposed by Shannon in 1948 as (29)

$$H[X] = \sum_x P(x) \log P(x). \qquad [2]$$

Entropy represents the average amount of information lacking about the outcome of a measurement of discrete random variable X. It is usually calculated with the logarithm taken to base 2 such that its units are in bits and all logarithms in the following should be understood as logarithms taken with respect to this base. In 1961, Renyi showed that by relaxing one of the Shannon–Khinchin axioms from which the mathematical form of entropy is uniquely derived (strong additivity), a more general expression for entropy may be obtained (30, 31)

$$H_\alpha[X] = \frac{1}{1-\alpha} \log \left(\sum_x P(x)^\alpha \right), \qquad [3]$$

where α is referred to as the order of the Renyi entropy. The family of Renyi entropies include Shannon's entropy measure as the limit of $\alpha \to 1$.

We may note that the probability of coincidence introduced in the previous subsection provides a measure for the Renyi entropy of order $\alpha = 2$

$$H_2[X] = -\log p_C[X]. \qquad [4]$$

The Renyi entropy of order 2 is known as collision entropy in cryptography and may also be motivated from an optimal code length perspective with nonlinearly weighted length penalties (32). Here, we use the term coincidence entropy to stress its relation to coincidence-counting among sample pairs. We focus on this entropy measure in the following as we will show that it relates directly to pairwise classification. For a generalization to higher-order Renyi entropies, see *SI Appendix*, Text 4.

1.3. Coincidence Mutual Information. We have previously used the coincidence ratio $p_C[X|\Pi]/p_C[X]$ between specific and background TCRs as a measure of antigen-driven selection (24), where $p_C[X|\Pi]$ is the probability of coincidence among epitope-specific TCRs averaged over a distribution of epitopes, $P(\Pi)$, and $p_C[X]$ the probability of coincidence among background TCRs. Different definitions of conditional Renyi entropy for $\alpha \neq 1$ have been proposed. Here, we follow refs. 33 and 34 and define

$$H_2[X|Y] = -\log p_C[X|Y], \qquad \textbf{[5]}$$

where $p_C[X|Y]$ is an average of $p_C[X|y]$ over all outcomes y of Y

$$p_C[X|Y] = \sum_y \rho_2(y) p_C[X|y], \qquad \textbf{[6]}$$

with weighting factors

$$\rho_2(y) = \frac{P(y)^2}{\sum_y P(y)^2}. \qquad \textbf{[7]}$$

Detailed justification for these definitions is provided in *SI Appendix*, Text 1. This definition allows us to express the coincidence probability ratio in terms of coincidence entropies

$$\log\left(\frac{p_C[X|\Pi]}{p_C[X]}\right) = H_2[X] - H_2[X|\Pi]. \qquad \textbf{[8]}$$

We note that for Shannon entropy this difference defines the mutual information between X and Π (29), which motivates the following definition of *coincidence mutual information*

$$I_2(X, \Pi) = \log\left(\frac{p_C[X|\Pi]}{p_C[X]}\right). \qquad \textbf{[9]}$$

Importantly, our definition of conditional entropy maintains additivity $H_2[X, Y] = H_2[X] + H_2[Y|X]$, where $H_2[X, Y]$ is the coincidence entropy of $P(X, Y)$, the joint distribution of the random variables X and Y. As a correlate it follows that coincidence mutual information is symmetric, $I_2(X, Y) = I_2(Y, X)$—as is its Shannon counterpart—so it tells us not only how much information we gain about sequence features upon learning their epitope specificity, but also, by symmetry, how much information a sequence feature provides about its epitope specificity. Coincidence mutual information thus provides a natural way to score the importance of a TCR feature in predicting specificity, which we will refer to as the feature relevancy.

1.4. Describing the Interactions between Features with Redundancy and Synergy. The connection between coincidence analysis and information theory naturally allows us to apply additional notions from information theory (35, 36) to describe how multiple features work in tandem to provide antigen specificity. First, conditional mutual information

$$I_2(X, \Pi|Y) = H_2[X|Y] - H_2[X|\Pi, Y], \qquad \textbf{[10]}$$

describes the remaining information provided by feature X given that the value of a second feature Y is already known. Here, $H_2[X|\Pi, Y]$ indicates conditioning on both epitope specificity and feature Y. If $I_2(X, \Pi|Y) = 0$, then we refer to X as a fully redundant feature in the context of Y. As a trivial example, knowledge of the complete primary sequence of the full α chain makes any information provided by CDR3α redundant, and so on.

Second, interaction information

$$I_{2,\text{int}}(X, Y|\Pi) = I_2([X, Y], \Pi) - I_2(X, \Pi) - I_2(Y, \Pi) \qquad \textbf{[11]}$$

describes how much additional information both features provide in conjunction (*SI Appendix*, Text 2). Here, $I_2([X, Y], \Pi)$ is the relevancy of the feature produced by combining the two features X and Y. If $I_{2,\text{int}}(X, Y|\Pi) > 0$, then there is synergy between the two features.

2. Bounding Classification Accuracy of Partial TCR Matches

2.1. Pairwise Classification Odds. There are well-known connections between information measures and achievable classification errors both in the Shannon (37) and Renyi case (38, 39). In the following, we derive how TCR classification accuracy using partial feature matches with a reference sequence is bounded when only partial information is available. We consider a classification setting where the task is to identify spiked-in TCR sequences specific to a particular epitope π in an otherwise naive repertoire. We will derive how posterior classification odds depend on feature relevancy and prior beliefs, i.e., the fraction of spiked-in sequences $P(\pi)$. Mathematically, in this setting, the presence of a TCR sequence σ is due to either of two generative processes:

$$P(\sigma) = P(\pi)P(\sigma|\pi) + P(B)P(\sigma|B), \qquad \textbf{[12]}$$

where $P(B) = (1 - P(\pi))$ and where $P(\sigma|\pi)$ is the probability of drawing σ from the distribution of TCR sequences specific to epitope π, $P(\Sigma|\pi)$, and $P(\sigma|B)$ the probability of drawing σ from the distribution of background TCR sequences according to V(D)J recombination, $P(\Sigma|B)$. In practice, we used a computational model to determine the probability of generation of TCR sequences $P(\Sigma|B)$ (40). This choice of background yields a distribution without the imprints of thymic or peripheral selection that determine TCR coincidence probabilities in naive and memory repertoires (24).

To recapitulate the empirical procedure of matching TCR sequences to a database of known binders, we consider the following one-shot classification strategy: We classify a query sequence as having been generated from $P(\Sigma|\pi)$, if it matches in a feature X with a reference sequence randomly drawn from $P(\Sigma|\pi)$. Using the odds formulation of Bayes' theorem, we may express the posterior odds of correct classification as

$$\frac{P(\pi|x = x')}{P(B|x = x')} = \frac{P(x = x'|\pi)}{P(x = x'|B)}\frac{P(\pi)}{P(B)}. \qquad \textbf{[13]}$$

Here, $P(x = x'|\pi) = p_C[X|\pi]$ is the probability of a match in feature X if both sequences were truly drawn from distribution $P(\Sigma|\pi)$, while $P(x = x'|B)$ is the probability of a match in feature X for a query drawn from $P(\Sigma|B)$ and a reference drawn from $P(\Sigma|\pi)$. Under the assumption that the propensity of a TCR for specific binding is independent of its recombination probability (24), one can show that $P(x = x'|B) = p_C[X]$ (*SI Appendix*, Text 3.A). Therefore,

$$\frac{P(\pi|x=x')}{P(B|x=x')} = \frac{p_C[X|\pi]}{p_C[X]}\frac{P(\pi)}{P(B)} \qquad [14]$$

This expression can be generalized for mixtures of multiple epitope groups, in which case the average odds over epitopes (*SI Appendix*, Text 3.B) can be expressed as

$$\left\langle \frac{P(\pi|x=x')}{P(B|x=x')} \right\rangle = \frac{p_C[X|\Pi]}{p_C[X]} \left\langle \frac{P(\pi)}{P(B)} \right\rangle, \qquad [15]$$

where $p_C[X|\Pi]$ is the conditional probability of coincidence defined previously and the averages for the odds are taken over $P(\pi)/(1-P(B))$. By the definition of coincidence mutual information (Eq. **9**), we can rewrite the last equation as

$$\mathbb{O}_{\text{post}} = 2^{I_2(X,\Pi)}\,\mathbb{O}_{\text{prior}}, \qquad [16]$$

which links the average posterior odds $\mathbb{O}_{\text{post}}$ to average prior odds $\mathbb{O}_{\text{prior}}$ via coincidence mutual information. Each bit of coincidence mutual information between X and Π corresponds to a two-fold gain in posterior odds.

2.2. When Is Partial Information Sufficient? Eq. **16** captures an important Bayesian intuition about classification: Correct classification depends not only on how much information we have available but also on our prior belief. Here, our prior belief about the likelihood that any particular sequence is specific should reflect the total fraction of spiked-in sequences. If we are searching for a needle in a haystack, this is when $\mathbb{O}_{\text{prior}}$ is small, we need to use more highly informative features for correct classification. Mathematically, a minimal prior odds of $2^{-I_2(X,\Pi)}\,T$ is needed to ensure that the average posterior odds exceeds a threshold value T. Expressed in terms of prior probabilities

$$P_{\text{prior}}(I_2) \geq \frac{T\,2^{-I_2}}{1+T\,2^{-I_2}}, \qquad [17]$$

is needed if only I_2 bits are available for classification. To illustrate this result, we performed in silico simulations with a toy model of TCR specificity (*SI Appendix*, Text 3.D). These simulations showed close agreement between predicted values for $P_{\text{prior}}(I_2)$ and those obtained through numerical simulation (*SI Appendix*, Fig. S1).

Note that sequences drawn from $P(\Sigma|B)$ may also be specific to π. Therefore, $P(\pi)$ and $P(\pi|x=x')$ are not exactly equal to the fraction of sequences specific to π and the posterior probability of specificity, respectively. However, as shown in *SI Appendix*, Text 3.C in most cases of practical interest, where $P(\pi)$ exceeds the background frequency of sequences specific to a given epitope, this distinction is irrelevant.

3. Application of the Methodology to TCR Sequence Data

To illustrate how our framework can be applied, we curated a dataset of multimer-sorted TCRs from CD8$^+$ T cells with specificity to viral antigens (*SI Appendix*, Text 7). We restricted our dataset to epitopes where at least one full TCR coincidence was observed to allow computation of coincidence information for the full TCR. Remarkably, such coincidences are observed in many epitope-specific repertoires: Here, we combined nine SARS-CoV-2-specific repertoires with such coincidences studied by Minervina et al. (22) with three repertoires specific to other viral epitopes from Dash et al. (1). To obtain background TCRs, we randomly paired TCRα and TCRβ sequences generated by a computational model of V(D)J recombination (40).

3.1. A Decomposition of TCR Specificity into Its Component Parts. To provide a top–down decomposition of the information content of the TCR, we computed the relevancy of different sections of the TCR for its specificity, as well as their combinations (Fig. 2). We first analyzed the information provided by the α and β chains alone which recapitulated the expected greater relevancy of the β chain (19 bits) than the α chain (12 bits). By Eq. **17**, the information provided by each chain bounds prior probabilities needed for accurate classification using single chain matches. A β chain match requires a prior probability $P_{\text{prior}} \geq 3\cdot 10^{-5}$ for a 95% posterior confidence. In contrast, an α chain match allows reliable classification only for prior probabilities $P_{\text{prior}} \geq 3\cdot 10^{-3}$. We then broke down the two chains further into their component V and J gene segments and CDR3 amino acid sequence. A CDR3β match provides 16 bits of information (corresponding to $P_{\text{prior}} \geq 4\cdot 10^{-4}$) while a CDR3$\alpha$ match provides only 10 bits of information (corresponding to $P_{\text{prior}} \geq 1\cdot 10^{-2}$).

In addition to features such as the CDR3, V, and J regions, our definition of a feature also extends to physical properties of the TCR such as the length of the CDR3 loop and its net charge (*SI Appendix*, Fig. S3). Our results confirm that these properties, which have been described in the literature as being important for epitope specificity (1, 41), have some relevancy in determining TCR specificity. For instance, CDR3β net charge is roughly as informative as Jβ choice. However, neither property captures a substantial proportion of CDR3 information demonstrating the contribution of higher-order sequence features to specific binding (*SI Appendix*, Text 6).

To assess variations in feature relevancy across epitopes, we defined local relevancy, $i_2(X,\pi) = \log(p_C[X|\pi]/p_C[X])$, as the information gain for a specific epitope π. Local relevancy scores revealed a broadly consistent hierarchy of feature relevancy across epitopes (*SI Appendix*, Figs. S4–S8). The analysis also identified variability in local relevancy of features between

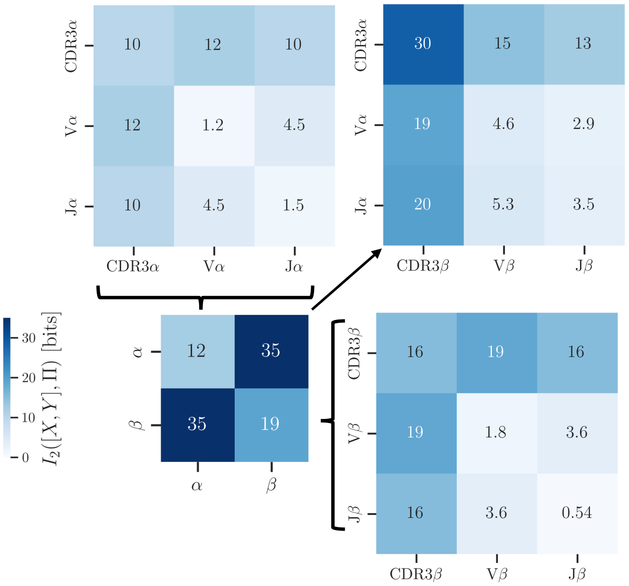

Fig. 2. Coincidence mutual information between TCR sections and antigen specificity. Relevancy scores of various sections of the T cell receptor sequence. The off-diagonal values indicate the amount of coincidence information that combinations of features provide. The *Top Right* hand grid shows the relevancy of combination of features where one is from the α chain and the other the β chain. Interaction information between features can be computed by taking the difference between the off-diagonals and the sum of the corresponding diagonal values (Fig. 3).

different epitopes not explained by finite sampling deviations alone in line with our prior findings on a subset of the studied epitopes (27). We will analyze this variability in more detail in Section 3.4.

3.2. Synergy and Redundancy between TCR Features. Comparing relevancy scores for individual and combined features revealed the pervasiveness of interactions between TCR sections (Fig. 2), where their combined information differed from the sum of their individual relevancies. Fig. 3 summarizes the interaction information between important TCR features.

Our analysis identified substantial synergy between the α and β chains (4 bits). This synergy implies that there are pairing restriction between α and β chains in specific TCRs, which make each chain more informative when considered in its full paired chain sequence context (*SI Appendix*, Text 2). These results broaden our prior findings (24) to a broader set of epitopes, and add to a growing literature investigating TCR α-β pairing rules (20, 24, 41–44). Pairing restrictions imply that the diversity of TCRs responding to a given epitope is lower than the product of the diversities of responsive α and β chains.

We also analyzed the interaction information between the CDR3 of each chain and the corresponding V segment choice. We again identified substantial synergy, presumably reflecting spatial constraints between V-gene encoded framework and CDR1/2 variability and CDR3 choice. In contrast, the interaction information between the CDR3 and J gene is negative. This is expected as the sequence variability provided by the J gene is contained in the CDR3 region (45) but demonstrates how our framework can identify redundant features without such a prior knowledge.

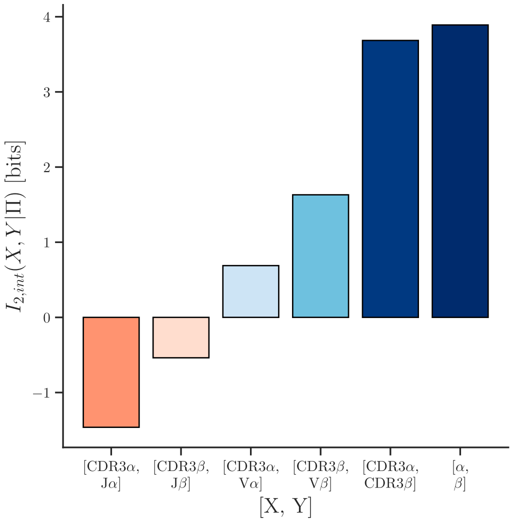

Fig. 3. Synergistic and redundant TCR sequence features. Interaction information scores for combinations of features. Positive interaction information indicates that two features become more informative in the context of one another and hence have synergy while negative interaction information suggests redundancy between the features.

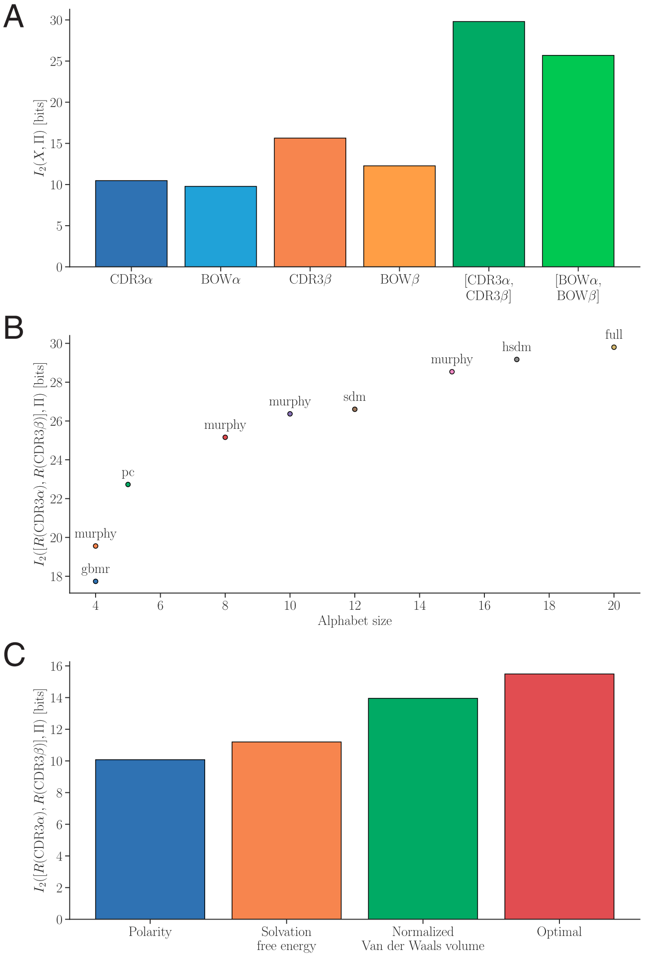

Fig. 4. Information preserved by different CDR3 compression schemes. (*A*) Relevancy of bag-of-words (BOW) representations for the CDR3α, CDR3β, and both CDR3 chains. CDR3 bag-of-words representations are vectors of dimension 20 with each entry representing the number of occurrences of a particular amino acid. (*B*) Information retained when compressing CDR3s using reduced amino acid alphabets described in *SI Appendix*, Text 7.B. R(CDR3) denotes the CDR3 remapped to the reduced alphabet. (*C*) Information retained by different two letter alphabets for both CDR3 chains. Polarity, solvation free energy, and normalized Van der Waals volume alphabets are obtained by hierarchical clustering of amino acids with respect to these biophysical properties. (For further properties and different alphabet sizes, see *SI Appendix*, Table S3.) The optimal alphabet is obtained by a greedy search algorithm described in *SI Appendix*, Text 7.D.

3.3. CDR3 Compression and Information Loss. We next sought to determine how much information about specificity is lost when compressing the CDR3 amino acid sequence into reduced representations of different complexities. We took conceptual inspiration from the information bottleneck method, which posits a trade-off between compression and preserved information (46). As good compression schemes retain relevant features such analyses can provide insights into which properties of the CDR3 sequence matter for its specificity.

First, we removed positional information by representing the CDR3 sequence as an unordered collection of its individual amino acids, referred to as bag-of-words representations in natural language processing (47). Such compression loses 4 bits of information for the paired chain receptor (Fig. 4*A*), highlighting the importance of the ordering of amino acids within the TCR. We next determined the relevancy of single dimensions of the bag-of-word vector, this is the number of times individual amino acids occur in the sequence. TCR amino acid contents provided less than a single bit of information about specificity, with arginine, proline, and glycine being most informative when considering both chains (*SI Appendix*, Fig. S9). Interestingly, glycine and proline content have been previously described as important for determining TCR specificity (41, 48), and both are determinants of protein flexibility

Second, we compared different reduced amino acid alphabets (49–53), which map amino acids to a smaller number of groups (Fig. 4*B*). By quantifying the compression-information tradeoff of different reduced alphabets, our results can help guide the choice of reduced alphabets for TCR applications, for instance by identifying Pareto optimal alphabets at a given alphabet size. We next determined how much information is preserved by reduced alphabets obtained via hierarchical clustering with respect to single biophysical properties (54, 55). Our analyses (described in detail in *SI Appendix*, Text 7.3) revealed major differences between the informativeness of such alphabets (Fig. 4*D* and *SI Appendix*, Table S3). For instance, steric properties such as radius of gyration of side chain or accessible surface area in a tripeptide resulted in particularly informative reduced alphabets across alphabet sizes. Finally, we implemented a greedy search algorithm to find the best two-letter alphabet (*SI Appendix*, Text 7.4). This algorithm identified a maximally informative two letter alphabet, which retains 15.5 bits of information (Fig. 4*D*), providing proof-of-concept for data-driven identification of an optimal coarse-graining strategy.

3.4. Variability in Interaction Information across Epitopes Is Explained by Mixture Models. To better understand potential sources of variability of TCR sequence restriction across epitopes we defined additional measures of local sequence variation: Local conditional mutual information $i_2(X,\pi|Y) = H_2[X|Y] - H_2[X|\pi, Y]$ and local interaction information $i_{2,\mathrm{int}}(X, Y|\pi) = i_2([X, Y], \pi) - i_2(X, \pi) - i_2(Y, \pi)$. We then analyzed dependencies across four variables (Fig. 5): Interaction information $i_{2,\mathrm{int}}(\alpha, \beta|\pi)$, α-chain relevancy $i_2(\alpha, \pi)$, β-chain relevancy $i_2(\beta, \pi)$ and paired chain relevancy $i_2([\alpha, \beta], \pi)$. These analyses highlighted strong dependencies between the variables. The more informative an α chain or β chain is for a given epitope, the less α-β interaction information contributes to global diversity restriction (Fig. 5 *A* and *B*). Moreover, epitopes with more informative α chains also have more informative β chains (Fig. 5*C*) and more informative full TCR sequences (Fig. 5*D*).

Unexpectedly, all variables were highly correlated with each other and well fitted by linear regressions, suggesting the existence of a single underlying degree of freedom that drives the observed variability across epitopes. Based on the clustering of epitope-specific TCRs, we had previously proposed mixture of motif models (24), in which epitope-specific TCRs are composed of a number of distinct binding solutions (binding modes or motifs). We asked whether variability in the number of such motifs across epitopes might provide the common degree of freedom explaining the observed correlations. Deriving the expected theoretical relationships between variables (*SI Appendix*, Text 6), we found an increased local interaction information for epitopes with more binding modes and a decrease in individual feature relevance. Across all variable pairs studied in Fig. 5, the mixture model predicted linear relations with slopes of ±1, in good agreement with the best fit lines to the empirical data. Intuitively, if an epitope has multiple binding solutions, more α and β chains will be able to bind it, given the right complementary chain (thus lowering the information from each individual chain). At the same time, where many solutions exist a high degree of α–β pairing is expected as most α chains from one binding solution would not be valid with β chains from another solution (thus increasing the observed synergy between the two chains).

Another consequence of the existence of multiple binding solutions is a potential loss of relevance of features with variable restriction across TCR clusters, as we have discussed when introducing our information-theoretic definition of feature relevance. In analyzing the relevancy of CDR3 length and charges as features (*SI Appendix*, Text 5), we found evidence for this phenomenon: The relevancy of length and charge increases substantially when

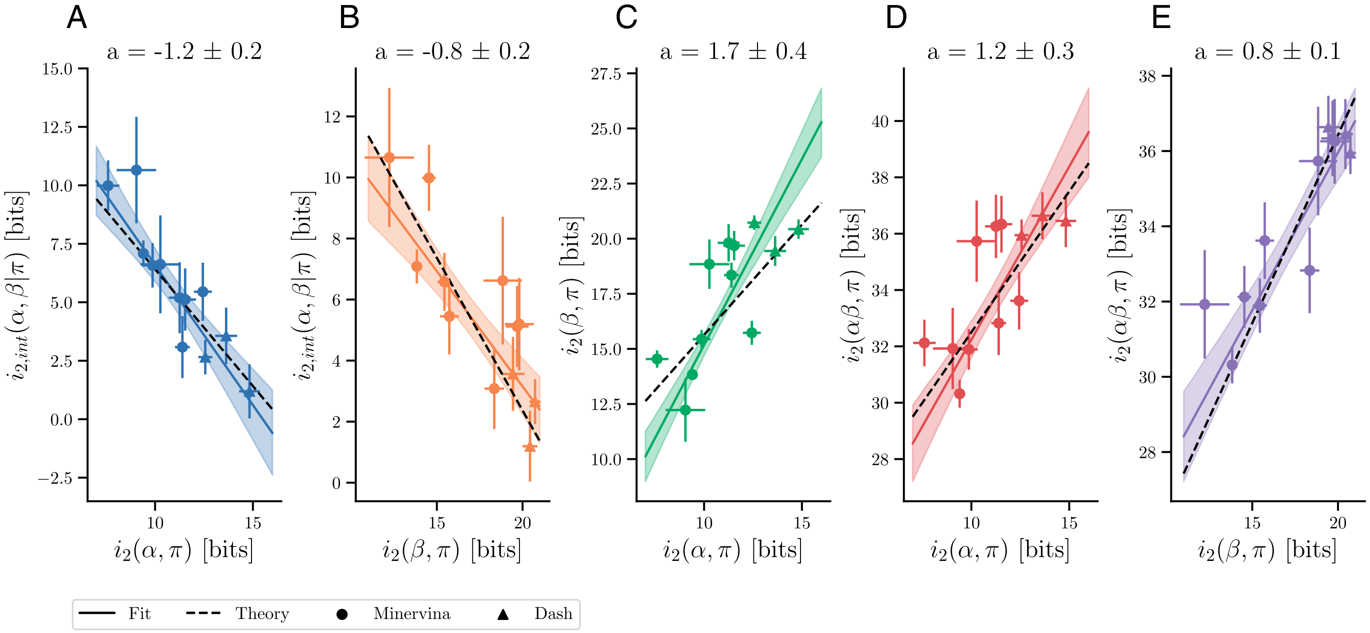

Fig. 5. Correlation between α-β interaction information and per-chain information across epitopes. Local interaction information and single-chain information across epitopes. Weighted linear fits (solid lines) obtained using orthogonal distance regression were used to quantify the dependence between variables, with regression slopes *a* displayed above each panel. Epitope-specific interaction information depends negatively on the local informational value of the (*A*) α chain and (*B*) β. We furthermore find that the (*C*) per-chain relevancies are positively correlated with each other as is (*D* and *E*) total information with both single chain relevancies. The observed dependencies between variables agree well with theoretical expectations from a mixture model (dashed lines), in which epitopes differ in the number of distinct binding solutions or contain false positives.

conditioning on V and J gene usage, which acts as a simple proxy for distinguishing TCR clusters (*SI Appendix*, Fig. S2). These results further support modeling epitope-specific repertoires as mixtures.

Given the low prevalence of epitope-specific TCRs in a repertoire, we additionally expect the dataset to be a mixture containing some false positive TCRs with no or low affinity to the epitope of interest even if sorting has high specificity. As we show in *SI Appendix*, Text 6 variations in the proportion of false positives across epitopes can also explain the observed dependencies among variables, with high interaction information for epitopes with many false positives. Both models share the common underlying insight that epitope-specific repertoires are mixtures rather than draws from a unimodal distribution—future research might elucidate the contributions of the different underlying mechanisms to the observed variability.

4. Distance Metrics and Near-Coincidence Entropy

4.1. Generalization of Coincidence Mutual Information to Fuzzy Matches. As exact matches are rare for complex features, it is of interest to also quantify the information provided by fuzzy feature matches. As previously explored in ref. 24, we are not limited to computing the probability of exact coincidences between features but can also consider near-coincidences according to some distance metric. Given a feature X distributed according to $P(X)$ and a distance metric $d(x, x')$ between outcomes x and x', the probability that two draws from $P(X)$ are at distance $d(x, x') = \Delta$ can be defined as

$$p_C[X](\Delta) = \sum_{x,x'} P(x)P(x')\delta_{d(x,x'),\Delta}, \qquad [18]$$

where $\delta_{d(x,x'),\Delta}$ is the Kronecker delta. We use this measure to propose a near-coincidence entropy $H_2[X](\Delta) = -\log p_C[X](\Delta)$, and a near-coincidence conditional entropy $H_2[X|Y](\Delta) = -\log p_C[X|Y](\Delta)$, where $p_C[X|Y](\Delta)$ once again is an average of $p_C[X|y](\Delta)$ over outcomes of Y using the $\rho_2(y)$ weighting factor. We define a near-coincidence mutual information

$$I_2\left(\Delta_{X,X'}, Y\right) = H_2[X](\Delta) - H_2[X|Y](\Delta), \qquad [19]$$

where $\Delta_{X,X'}$ denotes that this information is computed for near-coincidences in feature X at distance Δ. As this measure deals with pairs of instances of a random variable rather than single instances, this quantity cannot be defined straightforwardly for Shannon entropy but is motivated naturally when using coincidence entropy.

4.2. Pairwise Classification Using Fuzzy Matches. To obtain an interpretation of near-coincidence entropy, we turn once again to pairwise classification. We consider the same classification procedure as previously but based on a fuzzy match where the sequence with unknown specificity is distance Δ from the sequence with known specificity such that $d(x, x') = \Delta$. Similarly to our prior derivations, we find (*SI Appendix*, Text 3.E)

$$\mathbb{O}_{\text{post}} = 2^{I_2\left(\Delta_{X,X'},\Pi\right)}\, \mathbb{O}_{\text{prior}}. \qquad [20]$$

One bit of near-coincidence mutual information again corresponds to an average two-fold increase in posterior classification odds. As in the case of exact matching, Eq. **20** defines regimes in which fuzzy matches at a given distance are expected to succeed or fail. In Fig. 6, we provide an example of this by computing the required prior fraction of specific sequences to obtain a posterior probability of 0.95 with fuzzy CDR3 matches at a certain Levenshtein distance. Inversely, at a given prior odds ratio and target posterior odds ratio we can use these results to compute a critical distance beyond which classification becomes unreliable.

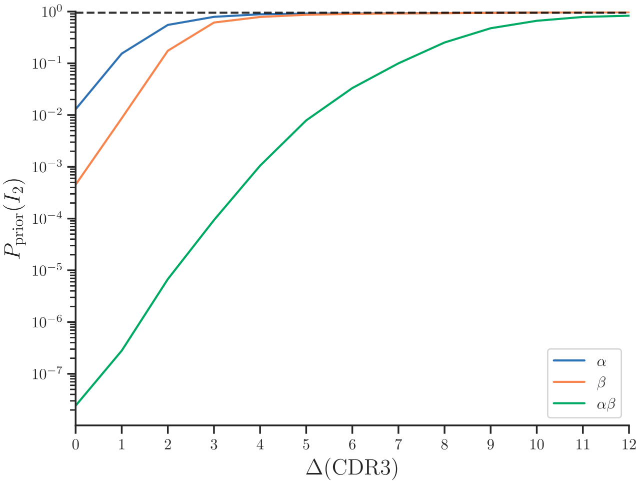

Fig. 6. Information theoretic analysis of fuzzy CDR3 matches. Critical prior probabilities for 95% confidence in classification using a fuzzy match with a Levenshtein distance Δ. Distances are CDR3α and CDR3β amino acid edit distances as well as the sum of CDR3α+CDR3β edit distance. Levenshtein distances are defined as the minimum number of insertions/deletions and substitutions required to turn one sequence into another.

5. Discussion

The ubiquity of information theory lies in its ability to describe complex relationships between data points using a simple quantitative vocabulary. As shown by Shannon (29), entropy provides the most natural measure of uncertainty and hence changes in entropy directly capture how knowledge of one event increases understanding of another. The application of information theory to the problem of immune receptor specificity has proved highly fruitful in the past. In particular, estimates of residue Shannon entropy aided in identifying potential complementary determining regions of the TCR and immunoglobulin and highlighted that TCRs were the more diverse of these two antigen receptors (56). Other, more recent studies have employed concepts from information theory such as mutual information to quantify interactions between various sections of the TCR sequence (20, 21). These previous studies have taken a "bottom–up" approach, computing an upper bound on sequence diversity by summing up the entropy of each constituent amino acid residue or pairs of residues. In part, this "bottom–up" approach has been required due to biases in estimating Shannon entropy in small samples. Although there exist methods for reducing bias in Shannon entropy estimation, these still require resolving higher-order distribution moments or essentially resort to coincidence counting and use Renyi entropy to approximate Shannon's (57, 58). In this work, we have proposed a "top–down" approach to decomposing TCR specificity firmly rooted in second-order Renyi entropy.

Our methodology provides a general framework to assess the role of individual TCR sequence features in determining antigen specificity as well as combinations of features by introducing the concepts of relevancy, redundancy, and synergy. We first compute the entropy of the full TCR sequence, divide this into its two constituent amino acid chains and then further subdivide these into their V, J, and CDR3 regions. Our results identify the β chain as the most informative of the heterodimeric TCR's chains and the CDR3 regions to be the most informative regions of each chain. However, we also find that the information these constituent parts provide is far smaller than that of the full TCR sequence. Although these results are unsurprising, with previous work highlighting the higher contribution of the β chain in epitope binding predictions and the importance of paired chain data (23, 59, 60), we provide the first full quantification of the information contained within these regions and, as our methodology has its foundations in coincidence-based statistics, we are able to directly interpret information measures in terms of achievable pairwise classification accuracy. Our work thus paves the way for the development of principled Bayesian methods for interpreting partial sequence matches.

Our results provide clear guides for when a limited amount of TCR sequence information, such as a single chain, is on average enough to solve an epitope specificity classification problem and when this loss of information may seriously impact predictive performance. We expect these insights to be important for experimental design, to decide whether the time and cost tradeoff of single cell sequencing over bulk are worth the increase in information paired chain information might provide.

We have also shown how the vocabulary of information theory can be applied to TCR near-coincidence analysis, which we have introduced in recent work (24). Our framework predicts pairwise classification performance when using fuzzy matches at a given threshold TCR distance. This approach may be used to define relevant data regimes in which current or future distance metrics (1, 61) may be usefully applied and allows setting critical distances for classifying or clustering sequences (62).

Our "top–down" approach allows us to compute interaction information, which describes synergistic and redundant relationships between TCR sequence features. We observe positive interaction information, synergy, between the α and β chain as well as the CDR3 and V regions, while knowledge of CDR3 regions makes their associated J regions redundant. We furthermore show how the relationship between interaction and single chain information across epitopes is compatible with a model in which epitopes vary in the number of distinct binding solutions (or possibly in the rate of false positives). With the steady accumulation of data on more epitopes, we envisage that our approach will help decipher principles underlying sequence space organization of responding TCRs.

Our framework can also be used to assess how much information is retained by compressed representations of the TCR such as bag-of-words vectors or reduced amino acid alphabets. We provide proof-of-concept for how our information scores may be used to construct reduced alphabets optimized for preserving epitope specific information and to discover biophysically salient measures of amino acid similarity.

The next steps for applying our theoretical approach are numerous. On the practical side, we propose completing the "top–down" approach and performing an analysis of the informational value of the CDR3 sequences residue by residue. This may allow for the identification of informationaly dense regions of the CDR3s and for a quantification of more complex allosteric interactions present across the receptor structure. Such analyses could complement work on structure-based prediction of TCR–pMHC interactions (63) and prediction with biophysical interaction energy models using contact maps derived from solved structures (64, 65). Further extensions of our framework could account for the hierarchy of selective processes shaping the TCR repertoire by varying the background used to compute background entropy. For example, to bound the performance of multiclass classification between a set of known epitopes, it may be more appropriate to quantify the entropy of TCRs across the chosen epitope-specific groups. Likewise, sequence statistics in a naive T cell repertoire could be used as background to account for the imprint of thymic selection. Our information theoretical tools may also be used on problems other than epitope specificity. For example, one may apply them to the study of TCR–MHC associations (66, 67) or TCR sequence to phenotype relationships (44, 68–70).

Linking feature information to classification isn't a problem unique to the field of protein function nor is the task of class prediction from pairwise comparisons. Transformer neural networks, the architecture underlying the current rise of large language models, embed data in high dimensional vector spaces (71) and may be trained in a pairwise contrastive manner, such that items from the same class are closer together than items from different classes (72–75). More generally, metric and representation learning commonly utilize pairwise measures to tackle problems ranging from sentence embeddings to facial recognition (76–80). Our pairwise coincidence information measure may be applicable for identifying interpretable informative features in these applications.

To conclude, we have introduced a theoretical framework for mapping the information content of the T cell receptor sequence with regard to its antigen specificity. Our results confirm prior insights from more limited structural studies regarding the relative importance of the α and β chains (2, 5, 81–84) but also highlight unexpected variability in the synergy between chains across epitopes. As dataset sizes continue to increase, the proposed framework should be able to guide the training of protein language models for predicting TCR–pMHC specificity (85–87) by information-content-driven masking strategies and will provide a tool to find interpretable physical features learned by such models.

Data, Materials, and Software Availability. Detailed source code and pre-processed data necessary to reproducing all results reported in this manuscript are available online at https://doi.org/10.5281/zenodo.13760163 (88). Antigen-specific TCR sequences were obtained from Dash et al. (1) as deposited in VDJdb (89), and from the *SI Appendix* of Minervina et al. (2).

ACKNOWLEDGMENTS. We thank Benny Chain, Curtis Callan, Lisa Wagar, Naresha Saligrama, Andrew Pyo, Matthew Cowley, and Sankalan Bhattacharyya for useful discussions over the course of this work. The work of J.H. and A.T.-M. was supported in parts by funding by the Wellcome Leap HOPE program (Human Organs Physiology and Engineering) and the Royal Free Charity. Y.N. and M.M. were supported by Cancer Research UK studentships under grants BCCG1C8R and A29287, respectively.

1. P. Dash *et al.*, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
2. J. Glanville *et al.*, Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
3. D. Hudson, R. A. Fernandes, M. Basham, G. Ogg, H. Koohy, Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**, 1–11 (2023).
4. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).

5. J. Rossjohn *et al*., T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
6. A. Mayer, V. Balasubramanian, T. Mora, A. M. Walczak, How a well-adapted immune system is organized. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5950–5955 (2015).
7. M. Goncharov *et al*., VDJdb in the pandemic era: A compendium of T cell receptors specific for SARS-CoV-2. *Nat. Methods* **19**, 1017–1019 (2022).
8. M. V. Pogorelyy, M. Shugay, A framework for annotation of antigen specificities in high-throughput T-cell repertoire sequencing studies considerations for experimental design. *Front. Immunol.* **10**, 1–9 (2019).
9. W. D. Chronister *et al*., TCRMatch: Predicting T-cell receptor specificity based on sequence similarity to previously characterized receptors. *Front. Immunol.* **12**, 640725 (2021).
10. S. Valkiers *et al*., Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics* **5**, 100009 (2022).
11. Z. S. Ghoreyshi, J. T. George, Quantitative approaches for decoding the specificity of the human T cell repertoire. *Front. Immunol.* **14**, 1228873 (2023).
12. S. Laughlin, A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* **36**, 910–912 (1981).
13. N. Brenner, S. P. Strong, R. Koberle, W. Bialek, RRdRv steveninck, synergy in a neural code. *Neural Comput.* **12**, 1531–1552 (2000).
14. S. E. Palmer, O. Marre, M. J. Berry, W. Bialek, Predictive information in a sensory population. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6908–6913 (2015).
15. J. B. Kinney, A. Murugan, C. G. Callan Jr., E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9158–9163 (2010).
16. N. M. Belliveau *et al*., Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4796–E4805 (2018).
17. J. O. Dubuis, G. Tkačik, E. F. Wieschaus, T. Gregor, W. Bialek, Positional information, in bits. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16301–16308 (2013).
18. M. D. Petkova, G. Tkačik, W. Bialek, E. F. Wieschaus, T. Gregor, Optimal decoding of cellular identities in a genetic network. *Cell* **176**, 844–855 (2019).
19. L. McGough *et al*., Finding the last bits of positional information. *PRX Life* **2**, 013016 (2024).
20. M. Milighetti *et al*., Intra-and inter-chain contacts determine TCR specificity: Applying protein co-evolution methods to TCR$\alpha\beta$ pairing. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.05.24.595718. Accessed 29 May 2024.
21. A. M. Xu *et al*., Entropic analysis of antigen-specific CDR3 domains identifies essential binding motifs shared by CDR3s with different antigen specificities. *Cell Syst.* **14**, 273–284 (2023).
22. A. A. Minervina *et al*., SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. *Nat. Immunol.* **23**, 781–790 (2022).
23. C. T. Boughter, M. Meier-Schellersheim, An integrated approach to the characterization of immune repertoires using aims: An automated immune molecule separator. *PLoS Comput. Biol.* **19**, e1011577 (2023).
24. A. Mayer, C. G. Callan Jr., Measures of epitope binding degeneracy from T cell receptor repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2213264120 (2023).
25. S. Ma, Calculation of entropy from data of motion. *J. Stat. Phys.* **26**, 221–240 (1981).
26. I. Nemenman, Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy* **13**, 2013–2023 (2011).
27. A. Tiffeau-Mayer, Unbiased estimation of sampling variance for simpson's diversity index. *Phys. Rev. E* **109**, 064411 (2024).
28. E. H. Simpson, Measurement of diversity. *Nature* **163**, 688–688 (1949).
29. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
30. A. Rényi, "On measures of entropy and information" in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, J. Neyman, Ed. (University of California Press, 1961), vol. 4, pp. 547–562.
31. A. Y. Khinchin, *Mathematical Foundations of Information Theory* (Courier Corporation, 1957).
32. L. L. Campbell, A coding theorem and rényi's entropy. *Inf. Control* **8**, 423–429 (1965).
33. P. Jizba, T. Arimitsu, The world according to rényi: Thermodynamics of multifractal systems. *Ann. Phys.* **312**, 17–59 (2004).
34. V. M. Ilić, M. S. Stanković, Generalized shannon-khinchin axioms and uniqueness theorem for pseudo-additive entropies. *Phys. A Stat. Mech. Appl.* **411**, 138–145 (2014).
35. J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24**, 175–186 (2014).
36. P. L. Williams, R. D. Beer, Nonnegative decomposition of multivariate information. arXiv [Preprint] (2010). https://arxiv.org/abs/1004.2515. Accessed 17 April 2024.
37. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, Hoboken, NJ, 2005).
38. M. Ben-Bassat, J. Raviv, Renyi's entropy and the probability of error. *IEEE Trans. Inf. Theory* **24**, 324–331 (1978).
39. I. Csiszár, Generalized cutoff rates and rényi's information measures. *IEEE Trans. Inf. Theory* **41**, 26–34 (1995).
40. Z. Sethna, Y. Elhanati, C. G. Callan Jr., A. M. Walczak, T. Mora, Olga: Fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).
41. K. Yu, J. Shi, D. Lu, Q. Yang, Comparative analysis of CDR 3 regions in paired human $\alpha\beta$ CD8 T cells. *FEBS Open Bio* **9**, 1450–1459 (2019).
42. D. S. Shcherbinin, V. A. Belousov, M. Shugay, Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCR$\alpha\beta$ complex. *PLoS Comput. Biol.* **16**, e1007714 (2020).
43. T. Dupic, Q. Marcou, A. M. Walczak, T. Mora, Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput. Biol.* **15**, e1006874 (2019).
44. J. A. Carter *et al*., Single T cell sequencing demonstrates the functional role of $\alpha\beta$ TCR pairing in cell lineage and antigen specificity. *Front. Immunol.* **10**, 1516 (2019).
45. M. P. Lefranc *et al*., IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* **27**, 55–77 (2003).
46. N. Tishby, F. C. Pereira, W. Bialek, The information bottleneck method. arXiv [Preprint]. https://doi.org/10.48550/arXiv.physics/0004057. Accessed 17 April 2024.
47. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv [Preprint] (2013). https://arxiv.org/abs/1301.3781. Accessed 17 April 2024.
48. S. C. Li, N. K. Goto, K. A. Williams, C. M. Deber, Alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6676–6681 (1996).
49. E. L. Peterson, J. Kondev, J. A. Theriot, R. Phillips, Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics* **25**, 1356–1362 (2009).
50. P. J. Cock *et al*., Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).
51. A. D. Solis, S. Rackovsky, Optimized representations and maximal information in proteins. *Proteins Struct. Funct. Bioinf.* **38**, 149–164 (2000).
52. L. R. Murphy, A. Wallqvist, R. M. Levy, Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* **13**, 149–152 (2000).
53. A. Prlić, F. S. Domingues, M. J. Sippl, Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.* **13**, 545–550 (2000).
54. S. Kawashima *et al*., Aaindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–D205 (2007).
55. H. Mei, Z. H. Liao, Y. Zhou, S. Z. Li, A new set of amino acid descriptors and its application in peptide QSARs. *Pept. Sci. Orig. Res. Biomol.* **80**, 775–786 (2005).
56. J. J. Stewart *et al*., A shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.* **34**, 1067–1082 (1997).
57. A. Chao, Y. Wang, L. Jost, Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* **4**, 1091–1100 (2013).
58. I. Nemenman, F. Shafee, W. Bialek, Entropy and inference, revisited. *Adv. Neural Inf. Process. Syst.* **14**, 471–478 (2001).
59. I. Springer, N. Tickotsky, Y. Louzoun, Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. *Front. Immunol.* **12**, 664514 (2021).
60. P. Meysman *et al*., Benchmarking solutions to the T-cell receptor epitope prediction problem: Immrep22 workshop report. *ImmunoInformatics* **9**, 100024 (2023).
61. R. Ehrlich *et al*., Swarmtcr: A computational approach to predict the specificity of T cell receptors. *BMC bioinforma.* **22**, 1–14 (2021).
62. K. Mayer-Blackwell *et al*., TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife* **10**, e68605 (2021).
63. P. Bradley, Structure-based prediction of T cell receptor: Peptide-MHC interactions. *eLife* **12**, e82813 (2023).
64. X. Lin *et al*., Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat. Comput. Sci.* **1**, 362–373 (2021).
65. A. Wang *et al*., Racer-m leverages structural features for sparse T cell specificity prediction. *Sci. Adv.* **10**, eadl0161 (2024).
66. M. R. Ortega *et al*., Learning predictive signatures of HLA type from T-cell repertoires. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.01.25.577228. Accessed 17 April 2024.
67. H. J. Zahid *et al*., Large-scale statistical mapping of T-cell receptor β sequences to human leukocyte antigens. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.04.01.587617. Accessed 17 April 2024.
68. S. A. Schattgen *et al*., Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (conga). *Nat. Biotechnol.* **40**, 54–63 (2022).
69. N. Ceglia *et al*., TCRi: Information theoretic metrics for single cell RNA and TCR sequencing in cancer. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.10.01.510457. Accessed 17 April 2024.
70. J. Textor *et al*., Machine learning analysis of the T cell receptor repertoire identifies sequence features of self-reactivity. *Cell Syst.* **14**, 1059–1073 (2023).
71. A. Vaswani *et al*., Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
72. C. Sun, F. Baradel, K. Murphy, C. Schmid, Learning video representations using contrastive bidirectional transformer. arXiv [Preprint] (2019). https://arxiv.org/abs/1906.05743. Accessed 17 April 2024.
73. R. Singh, S. Sledzieski, B. Bryson, L. Cowen, B. Berger, Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2220778120 (2023).
74. T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings. arXiv [Preprint] (2021). https://arxiv.org/abs/2104.08821. Accessed 17 April 2024.
75. A. Neelakantan *et al*., Text and code embeddings by contrastive pre-training. arXiv [Preprint] (2022). https://arxiv.org/abs/2201.10005. Accessed 17 April 2024.
76. K. Musgrave, S. Belongie, S. N. Lim, "A metric learning reality check" in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm, Eds. (Springer, 2020), pp. 681–699.
77. A. Bellet, A. Habrard, M. Sebban, A survey on metric learning for feature vectors and structured data. arXiv [Preprint] (2013). https://arxiv.org/abs/1306.6709. Accessed 17 April 2024.
78. X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, L. Davis, P. Torr, S.-C. Zhu, Eds. (IEEE, 2019), pp. 5022–5030. Accessed 17 April 2024.
79. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, "A simple framework for contrastive learning of visual representations" in *International Conference on Machine Learning*, H. Daumé, A. Singh, Eds. (PMLR, 2020), pp. 1597–1607.
80. T. Wang, P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere" in *International Conference on Machine Learning*, H. Daumé, A. Singh, Eds. (PMLR, 2020), pp. 9929–9939.
81. K. C. Garcia *et al*., An $alpha\eta$ T cell receptor structure at 2.5 å and its orientation in the TCR-MHC complex. *Science* **274**, 209–219 (1996).
82. J. B. Reiser *et al*., CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nat. Immunol.* **4**, 241–247 (2003).
83. Y. He *et al*., Peptide-MHC binding reveals conserved allosteric sites in MHC class I-and class II-restricted T cell receptors (TCRs). *J. Mol. Biol.* **432**, 166697 (2020).
84. M. Milighetti, J. Shawe-Taylor, B. Chain, Predicting T cell receptor antigen specificity from structural features derived from homology models of receptor-peptide-major histocompatibility complexes. *Front. Physiol.* **12**, 730908 (2021).

85. B. Meynard-Piganeau, C. Feinauer, M. Weigt, A. M. Walczak, T. Mora, Tulip: A transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2316401121 (2024).
86. B. P. Kwee *et al*., Stapler: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.04.25.538237. Accessed 17 April 2024.
87. Y. Nagano *et al*., Contrastive learning of T cell receptor representations. arXiv [Preprint] (2024). https://arxiv.org/abs/2406.06397. Accessed 10 June 2024.
88. J. Henderson, A. Tiffeau-Mayer. qimmuno/paper_tcrinfo: TCRinfo V1. Zenodo. https://doi.org/10.5281/zenodo.13760163. Deposited 13 September 2024.
89. M. Shugay, antigenomics/vdjdb-db. Github. https://github.com/antigenomics/vdjdb-db/issues/195. Deposited 13 June 2017.