Deep generative selection models of T and B cell receptor repertoires with soNNia

Giulio Isacchini,[1,2] Aleksandra M. Walczak,[2,*] Thierry Mora,[2,*] and Armita Nourmohammad[3,4,5,*]

[1] *Max Planck Institute for Dynamics and Self-organisation, Am Faßberg 17, 37077 Göttingen, Germany*
[2] *Laboratoire de physique de l'école normale supérieure (PSL University), CNRS, Sorbonne Université, and Université de Paris, 75005 Paris, France*
[3] *Department of Physics, University of Washington, 3910 15th Ave Northeast, Seattle, WA 98195, USA*
[4] *Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany*
[5] *Fred Hutchinson Cancer Research Center, 1100 Fairview ave N, Seattle, WA 98109, USA*

(Dated: November 5, 2020)

Subclasses of lymphocytes carry different functional roles to work together to produce an immune response and lasting immunity. Additionally to these functional roles, T and B-cell lymphocytes rely on the diversity of their receptor chains to recognize different pathogens. The lymphocyte subclasses emerge from common ancestors generated with the same diversity of receptors during selection processes. Here we leverage biophysical models of receptor generation with machine learning models of selection to identify specific sequence features characteristic of functional lymphocyte repertoires and subrepertoires. Specifically using only repertoire level sequence information, we classify $CD4^+$ and $CD8^+$ T-cells, find correlations between receptor chains arising during selection and identify T-cells subsets that are targets of pathogenic epitopes. We also show examples of when simple linear classifiers do as well as more complex machine learning methods.

I. INTRODUCTION

The adaptive immune system in vertebrates consists of highly diverse B- and T-cells whose unique receptors mount specific responses against a multitude of pathogens. These diverse receptors are generated through genomic rearrangement and sequence insertions and deletions, a process known as V(D)J recombination [4, 5]. Recognition of a pathogen by a T- or B-cell receptor is mediated through molecular interactions between an immune receptor protein and a pathogenic epitope. T-cell receptor proteins interact with short protein fragments (peptide antigens) from the pathogen that are presented by specialized pathogen presenting Major Histocompatibility Complexes (MHC) on cell surface. B-cell receptors interact directly with epitopes on pathogenic surfaces. Upon an infection, cells carrying those specific receptors that recognize the infecting pathogen become activated and proliferate to control and neutralize the infection. A fraction of these selected responding cells later contribute to the memory repertoire that reacts more readily in future encounters. Unsorted immune receptors sampled from an individual reflect both the history of infections and the ongoing responses to infecting pathogens.

Before entering the periphery where their role is to recognize foreign antigens, the generated receptors undergo a two-fold selection process based on their potential to bind to the organism's own self-proteins. On one-hand, they are tested to not be strongly self-reactive (Fig. 1 A) On the other hand, they must be able to bind to some of the presented molecules, to assure minimal binding capabilities. This pathogen-unspecific selection, known as thymic selection for T-cells [6] and the process of central tolerance in B-cells [7], can prohibit over 90% of generated receptors from entering the periphery [6, 8, 9].

Additionally to receptor diversity, T and B cell-subtypes are specialized to perform different functions. B- and T-cells in the adaptive immune system are differentiated from a common cell-type, known as lymphoid progenitor. T-cells differentiate into cell subtypes identified by their surface markers, including helper T-cells ($CD4^+$), killer T-cells ($CD8^+$) [6], and regulatory T-cells or T-regs ($CD4^+$ $FOXP3^+$) [10], each of which can be found in the non-antigen primed naive or memory compartment. The memory compartment can be further divided into subtypes, such as effector, central or stem cell-like memory cells, characterized by different lifetimes and roles. B-cells develop into, among other subtypes, plasmablasts and plasma cells, which are antibody factories, and memory cells that can be used against future infections. These cell types perform distinct functions, react with different targets, and hence, experience different selection pressures. Here we ask whether these different functions and selection pressures are reflected in their receptors' sequence compositions.

Recent progress in high-throughput immune repertoire sequencing (RepSeq) both for single-chain [11–14] and paired-chain [15–18] B- and T-cell receptor has brought significant insight into the composition of immune repertoires. Based on such data, statistical inference techniques have been developed to infer biophysically informed sequence-based models for the underlying processes involved in generation and selection of immune receptors [1–3, 19–21]. Machine learning techniques have also been used to infer deep generative models to charac-

*These authors contributed equally. Correspondence should be addressed to: aleksandra.walczak@phys.ens.fr, thierry.mora@phys.ens.fr, and armita@uw.edu.

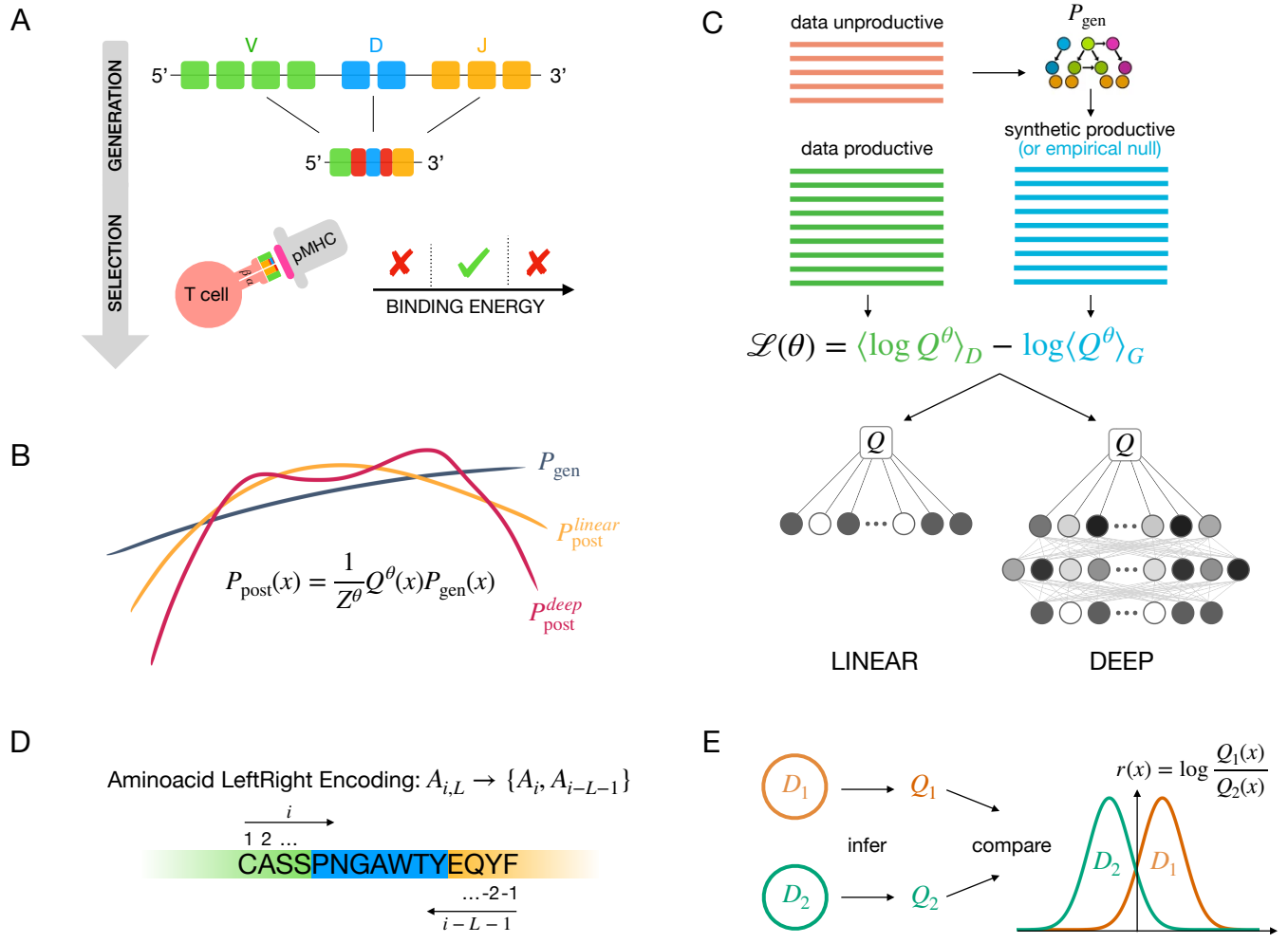


FIG. 1: **Inference of functional selection models for immune receptor repertoires. (A)** T cell receptor α and β chains are stochastically rearranged through a process called V(D)J recombination. Successfully rearranged receptors undergo selection for binding to self-pMHCs. Receptors that bind too weakly or too strongly are rejected, while intermediately binding ones exit the thymus and enter peripheral circulation. Development of B-cell receptors follows similar stages of stochastic recombination and selection.**(B)** We model these two processes independently. The statistics of the V(D)J recombination process described by the probability of generating a given receptor sequence σ, $P_{\text{gen}}(\sigma)$, are inferred using the IGOR software [1]. $P_{\text{gen}}(\sigma)$ acts as a baseline for the selection model. We then infer selection factors $\mathcal{Q}$, which act as weights that modulate the initial distribution $P_{\text{gen}}(\sigma)$. We infer two types of selection weights: linear in log space (using the SONIA software [2]) and non-linear weights using a deep neural network, in the soNNia software presented here. Non-linear selection weights are more flexible than linear ones. **(C)** Pipeline of the algorithm: P_{gen} is inferred from unproductive sequences using IGOR. Selection factors for both the linear and non-linear models are inferred from productive sequences by maximizing their log-likelihood $\mathcal{L}$, which involves a normalization term calculated by sampling unselected sequences generated by the OLGA software [3]. **(D)** In both selection models the amino acid composition of the CDR3 is encoded by its relative distance from the left and right borders (left-right encoding). **(E)** After inferring repertoire specific selection factors, repertoires are compared by computing e.g. log likelihood ratios $r(x)$.

terize the T-cell repertoire composition as a whole [22], as well as discriminate between public and private B-cell clones based on Complementarity Determining Region 3 (CDR3) sequence [23, 24]. While biophysically informed models can still match and even outperform machine-learning techniques (see e.g. [25]), deep learning models can be extremely powerful in describing functional subsets of immune repertoires, for which we lack a full biophysical understanding of the selection process.

Here, we introduce a framework that uses the strengths of both biophysical models and machine learning approaches to characterize signatures of differential selection acting on receptor sequences from subsets associated with specific function. Specifically, we leverage bio-

physical tools to model what we know (e.g. receptor generation) and exploit the powerful machinery of deep neural networks (DNN) to model what we do not know (e.g. functional selection). Using the non-linear and flexible structure of the deep neural networks, we characterize the sequence properties that encode selection of the specificity of the combined chains during receptor maturation in α andβ chains in T-cells, and heavy and light (κ and λ) chains in B-cells. We identify informative sequence features that differentiate $CD4^+$ helper T-cells, $CD8^+$ killer T-cells and regulatory T-cells. Finally, we demonstrate that that biophysical selection models can be used as simple classifiers to successfully identify T-cells specific to distinct targets of pathogenic epitopes—a problem that is of significant interest for clinical applications [26–30].

II. RESULTS

Neural network models of TCR and BCR selection

Previous work has inferred biophysically informed models of V(D)J recombination underlying the generation of TCRs and BCRs [1, 31]. We infer the parameters of these models using the IGOR software [1] from unproductive receptor sequences, which are generated, but due to a frameshift or insertion of stop codons are not expressed, and hence, are not subject to functional selection. The inferred models are used to characterize the generation probability of a receptor sequence $P_{\rm gen}$, and to synthetically generate an ensemble of pre-selection receptors [3]. These generated receptors define a baseline $\mathcal{G}$ for statistics of repertoires prior to any functional selection.

To identify sequence properties that are linked to function, we compare the statistics of sequence features f (e.g. V-, D-, J- gene usage and CDR3 amino acid composition) present in a given B- or T- cell functional repertoire to the expected baseline of receptor generation (Fig. 1 C). To do so, we encode a receptor sequence σ as a binary vector $\mathbf{x}$ whose elements $x_f \in \{0,1\}$ specify whether the feature f is present in a sequence σ. The probability $P^\theta_{\rm post}(\mathbf{x})$ for a given receptor $\mathbf{x}$ to belong to a functional repertoire is described by modulating the receptor's generation probability $P_{\rm gen}(x)$ by a selection factor $\mathcal{Q}^\theta(\mathbf{x})$,

$$P^\theta_{\rm post}(\mathbf{x}) = P_{\rm gen}(\mathbf{x})\mathcal{Q}^\theta(\mathbf{x}) \equiv \frac{1}{Z^\theta} P_{\rm gen}(\mathbf{x})\, Q^\theta(\mathbf{x}), \qquad (1)$$

where θ denotes the parameters of the selection model and Z^θ ensures normalization of $P^\theta_{\rm post}$. Previous work [2, 32, 33] inferred selection models for functional repertoires by assuming a multiplicative form of selection $Q^\theta(\mathbf{x}) = \exp(\sum_f \theta^f x_f)$, where feature-specific factors θ^f contribute independently to selection. We refer to these models as linear SONIA (Fig. 1B). Selection can in general be a highly complex and non-linear function of the underlying sequence features. Here, we introduce soNNia, a method to infer generic non-linear selection functions, using deep neural networks (DNN). To infer a selection model that best describes sequence determinants of function in a data sample $\mathcal{D}$, soNNia maximizes the mean log-likelihood of the data $\mathcal{L}(\theta) = \langle \log P^\theta_{\rm post}\rangle_{\mathcal{D}}$, where the probability $P^\theta_{\rm post}$ is defined by Eq. (1), and $\langle\cdot\rangle_{\mathcal{D}}$ denotes expectation over the set of sequences $\mathcal{D}$. This likelihood can be rewritten as (see Methods),

$$\mathcal{L}(\theta) = \langle \log P^\theta_{\rm post}\rangle_{\mathcal{D}} = \langle \log Q^\theta\rangle_{\mathcal{D}} - \log\langle Q^\theta\rangle_{\mathcal{G}} + \text{const}, \qquad (2)$$

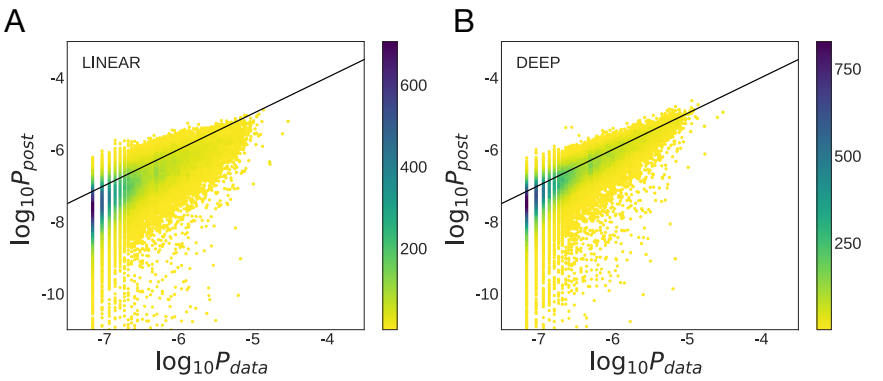
where $\langle\cdot\rangle_{\mathcal{G}}$ is the expectation over the baseline $\mathcal{G}$, which was generated from $P_{\rm gen}$. Note that this expression becomes exact as the number of generated sequences approaches infinity.

We divide the sequence features f into three categories: (i) (V,J) usage, (ii) CDR3 length, and (iii) CDR3 amino acid composition encoded by a 20×50 binary matrix that specifies the identity of an amino acid and its relative position within a 25 amino acid range from both the 5' and the 3' ends of the CDR3, equivalent to the left-right encoding of the SONIA model [2] (Fig. 1D). Input from each of the three categories are first propagated through their own network. Outputs from these three networks are then combined and transformed through a dense layer. This choice of architecture reduces the number of parameters in the DNN and makes the contributions of the three categories (which have different dimensions) comparable; see Methods and Figs. S1-S3 for details on the architecture of the DNN.

The baseline ensemble $\mathcal{G}$, which we have described as being generated from the $P_{\rm gen}$ model (Fig. 1 C), can in principle be replaced by any dataset, including empirical ones, at no additional computational cost. We will use this functionality of soNNia to learn selection coefficients of subsets relative to a generic functional repertoire. In that case, the inferred selection factors Q only reflect differential selection relative to the generic baseline. Once two soNNia models have been learned from two distinct datasets, their statistics may be compared by computing a sequence-dependence log-likelihood ratio $r(x) = \log Q_1(x)/Q_2(x)$ predicting the preference of a sequence for a subset over the other. This log-likelihood ratio can be used as a functional classifier for receptor repertoires (Fig. 1 E).

Deep non-linear selection model best describes functional TCR repertoire

First, we systematically compare the accuracy of the (non-linear) soNNia model with linear SONIA [2] (Fig. 1 B) by inferring selection on TCRβ repertoires from a large cohort of 743 individuals from Ref. [34]. Our goal is to characterize selection on functional receptors irrespective of their phenotype. To avoid biases caused by expansions of particular receptors in different individuals, we pool the *unique* nucleotide sequences of receptors



	ρ^2	$\mathcal{D}_{KL}$[bits]
$P_{\rm gen}$	0.43 ± 0.01	3.03 ± 0.01
linear	0.48 ± 0.01	1.58 ± 0.01
deep	$\mathbf{0.61 \pm 0.01}$	$\mathbf{1.01 \pm 0.01}$
noise	0.791 ± 0.001	0.402 ± 0.001

FIG. 2: **Performance of selection models on TCR repertoires.** Scatter plot of observed frequency, $P_{\rm data}$, versus predicted probability $P_{\rm post}$ for **(A)** linear SONIA and **(B)** deep neural network soNNia models trained on the TCRβ repertoires of 743 individuals from ref. [34]. Color indicates number of sequences. **(C)** The soNNia model performs significantly better, as quantified by both the Kullback-Leibler divergence $\mathcal{D}_{\rm KL}$ and the Pearson correlation coefficient ρ^2.

from all individuals and construct a universal donor totalling 9×10^7 sequences. Multiplicity of an amino-acid sequence in this universal donor indicates the number of independent recombination events that have led to that receptor (in different individuals, or in the same individual by convergent recombination).

We randomly split the pooled dataset into a training and a test set of equal sizes. We then subsampled the training set to 10^7 to reduce the computational cost of inference. We trained both a SONIA and a soNNia model on the training set, using 10^7 sequences sampled from $P_{\rm gen}$ (learned from the nonproductive sequences of the same dataset) as the baseline $\mathcal{G}$ (Fig. 1 C). To assess the performance of our selection models, we compare their inferred probabilities $P_{\rm post}(\mathbf{x})$ with the observed frequencies of the receptor sequences $P_{data}(\mathbf{x})$ in the test set (Fig. 2A and B). Prediction accuracy can be quantified through the Pearson correlation between the two log-frequencies, or through their Kullback-Leibler divergence (Fig. 2C)

$$\mathcal{D}_{\rm KL}(P_{\rm data}|P_{\rm post}) = \left\langle \log_2 \frac{P_{\rm data}}{P_{\rm post}} \right\rangle_{P_{\rm data}} . \qquad (3)$$

We estimate the Kullback-Leibler divergence using 10^5 receptors in the test set with multiplicity larger than two. A smaller Kullback-Leibler divergence indicates a higher accuracy of the inferred model in predicting the data. The estimated accuracy of an inferred model is limited by the correlation between the test and the training set, which provides a lower bound on the Kullback-Leibler divergence $\mathcal{D}_{KL} \simeq 0.4$ bits, and an upper bound on the Pearson correlation $\rho^2 \simeq 0.8$.

We observe a substantial improvement of selection inference for the generalized selection model soNNia with $\mathcal{D}_{KL} \simeq 1.0$ bits (and Pearson correlation $\rho^2 \simeq 0.61$) compared to the linear SONIA model with $\mathcal{D}_{KL} \simeq 1.6$ bits (and Pearson correlation $\rho^2 \simeq 0.48$); see Fig. 2. Both models show a strong effect of selection, reducing the $\mathcal{D}_{KL}$ from 3.03 bits (and increasing the correlation ρ^2 from 0.43) for the comparison of data to the $P_{\rm gen}$ model alone (Fig. 2). This result highlights the role of complex nonlinear selection factors acting on receptor features that shape a functional T-cell repertoire. The features that are still inaccessible to the soNNia selection factors are likely due to the sampling of rare features.

Intra- and inter-chain interactions in TCRs and BCRs

T-cell receptors are disulfide-linked membrane-bound proteins made of variable α and β chains, and expressed as part of a complex that interact with pathogens. Similarly, B-cell receptors and antibodies are made up of a heavy and two major groups (κ and λ) of light chains. Previous work has identified low but consistent correlations between features of $\alpha\beta$ chain pairs in T-cell receptors, with the largest contributions between V_α, V_β and J_α, V_β [35–37]. In B-cells, preferences for receptor features within heavy and light chains have been studied separately [38, 39] but inter-chain correlations have not been systematically investigated.

We first aimed to quantify dependencies between chains by re-analyzing recently published single-cell datasets: TCR $\alpha\beta$ pairs of unfractionated repertoires from ref. [40] (totalling 5×10^5 receptors), and BCR of naive cells from ref. [41] (totalling 22×10^3 and 28×10^3 receptors for the Hλ and Hκ repertoires, respectively). The blue bars of Fig. 3 show the mutual information between the V and J choices and CDR3 length of each chain, for TCR $\alpha\beta$ (Fig. 3A), Ig Hλ (Fig. 3B), and Ig Hκ (Fig. 3C) repertoires. Mutual information is a non-parametric measure of correlation between pairs of variables (see Methods).

Both TCR and BCR have intra- and inter chain correlations of sequence features, with a stronger empirical mutual dependencies present within chains. The largest

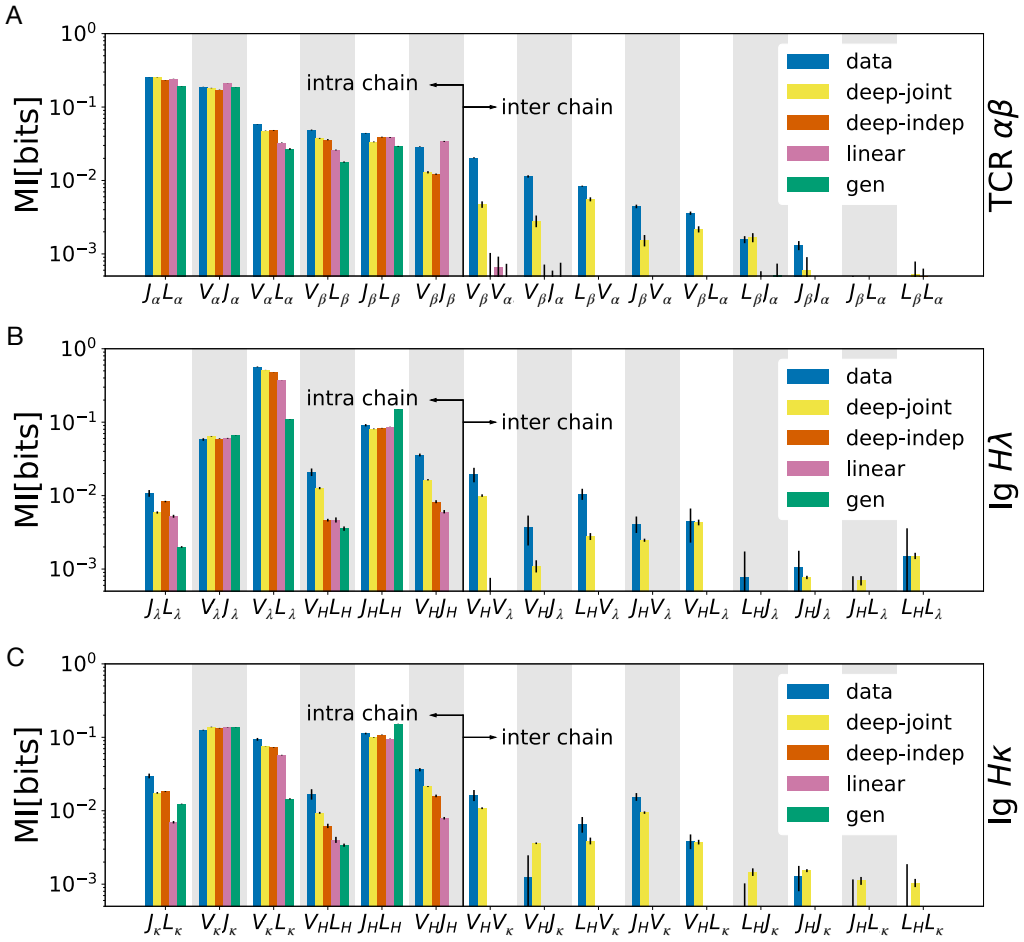


FIG. 3: **Inference of selection on intra- and inter-chain receptor features.** Mutual information between pairs of major intra- and inter- chain features (V and J gene choice and L =CDR3 length for each chain) for **(A)** TCR $\alpha\beta$, **(B)** Ig Hλ, and **(C)** Ig Hκ paired chains are shown. Mutual information is estimated directly from data (blue), and from receptors generated based on inferred models: generative baseline (green), *linear* SONIA (pink), *deep-indep* (red), and *deep-joint* (yellow). For both TCRs and BCRs, only the *deep-joint* model (yellow), which correlates the features of both chains through a deep neural network, is able to recover inter-chain correlations. Mutual informations are corrected for finite-size bias and error bars are obtained by subsampling (see Methods).

inter-chain dependencies is associated with the V-gene usages of the two chains, for both T-cells and B-cells, consistent with previous observations in T-cells [36, 40, 42].

To account for these dependencies between chains, we generalize the selection model of eq. 1 to pairs, $\mathbf{x} = (\mathbf{x}^a, \mathbf{x}^b)$, where $(a, b) = (\alpha, \beta)$ in TCRs or (H, κ) or (H, λ) in BCRs:

$$P_{\rm post}(\mathbf{x}) = \frac{1}{Z^\theta} P^a_{\rm gen}(\mathbf{x}^a) P^b_{\rm gen}(\mathbf{x}^b) Q(\mathbf{x}),$$

where we have dropped the dependence on parameters θ for ease of notation.

Analogously to single chains, we first define a *linear* selection model specified by $Q(\mathbf{x}) = \exp(\sum_f \theta_f x_f)$, where the sum now runs over features of both chains a and b. Because of its multiplicative form, selection can then be decomposed as the product of selection factors for each chain: $Q(\mathbf{x}) = Q^a(\mathbf{x}^a) Q^b(\mathbf{x}^b)$, where Q^a and Q^b are linear models. We also define a deep independent model (*deep-indep*), which has the multiplicative form $Q(\mathbf{x}) = Q^a(\mathbf{x}^a) Q^b(\mathbf{x}^b)$, but where Q^a and Q^b are each described by deep neural networks that can account for complex correlations between features of the same chain, similar to the single-chain case (Fig. S2). The resulting post-selection distributions for both the linear and the deep-indep model factorize, $P_{\rm post}(\mathbf{x}) = P^a_{\rm post}(\mathbf{x}^a) P^b_{\rm post}(\mathbf{x}^b)$, making the two chains independent. Thus, by construction neither the linear nor the deep-indep model can account for correlations between chains. Finally, we define a full soNNia model (*deep-joint*) where $Q(\mathbf{x})$ is a neural network combining and correlating the features of both chains (Fig. S3).

We trained these three classes of models on each of the TCR $\alpha\beta$, and BCR Hκ and Hλ paired repertoire data

described earlier. We then used these models to generate synthetic data with a depth similar to the real data, and calculated mutual informations between pairs of features (Fig. 3). The pre-selection generation model ($Q(\mathbf{x}) = 1$, green bars) explains part but not all of the intra-chain feature dependencies, for both T- and B-cells, while the linear (purple), deep-indep (red), and deep-joint (yellow) models explain them very well. Notably, the increase in correlations (difference between green and other bars) due to selection is larger in naive B-cells than in unsorted (memory and naive) T-cells. By construction, the generation, linear, and deep-indep models do not allow for inter-chain correlations. Only the deep-joint model (yellow) is able to recover part of the inter-chain dependencies observed in the data. It even overestimates some correlations in BCRs, specifically between the CDR3 length distributions of the two chains, and between the heavy-chain J and the light-chain CDR3 length. Thus, the deep structure of soNNia recapitulates both intra-chain and inter-chain dependencies of feature forming immune receptors.

Cell type and tissue-specific selection on T-cells

During maturation in the thymus, T-cells differentiate into two major cell-types: cytotoxic ($CD8^+$) and helper ($CD4^+$) T-cells. $CD8^+$ cells bind peptides presented on major histocompatibility complex (MHC) class I molecules that are expressed by all cells, whereas $CD4^+$ cells bind peptides presented on MHC-class II molecules, which are only expressed on specialized antigen presenting cells. Differences in sequence features of $CD8^+$ and $CD4^+$ T-cells should reflect the distinct recognition targets of these receptors. Although these differences have already been investigated in refs. [42, 43], we still lack an understanding as how selection contributes to the differences between $CD8^+$ and $CD4^+$ TCRs. In addition to functional differentiation at the cell-type level, T-cells also migrate and reside in different tissues, where they encounter different environments and are prone to infections by different pathogens. As a result, we expect to detect tissue-specific TCR preferences that reflect tissue-specific T-cell signatures.

To characterize differential sequence features of TCRs between cell types in different tissues, we pool unique TCRs from 9 individuals (from Ref. [43]) sorted into three cell-types ($CD4^+$ conventional T cells (Tconv), $CD4^+$ regulatory T cells (Treg) and $CD8^+$ T cells), and harvested from 3 tissues (pancreatic draining lymph nodes (pLN), "irrelevant" non-pancreatic draining lymph nodes (iLN), and spleen).

Training a deep soNNia model (see Fig. 1 C) for each subset leads to overfitting issues due to limited data. To solve this problem, we use the technique of transfer learning, which consists of learning a shared deep soNNia model for all subsets, and then add an additional linear layer for each sub-repertoire (see Fig. S4). However, an equivalent but simpler way is to train linear SONIA models atop a common baseline set $\mathcal{G}$ made of the empirical unfractionated repertoire from ref. [44] , so that the inferred Q factors only reflect selection relative to the generic set. Alternatively, we used the generative model $P_{\rm gen}$ (trained earlier for Fig. 2) as baseline, in which case the selection factors include selection effects that are shared among the sub-repertoires. Distribution of selection factors obtained by both approaches are shown in Fig. S5. We evaluate the Jensen-Shannon divergence $D_{\rm JS}(r, r')$ between the distribution of pairs (r, r') of these sub-repertoires, $P^r_{\rm post}$ and $P^{r'}_{\rm post}$,

$$D_{\rm JS}(r,r') = \frac{1}{2}\Big\langle \log_2 \frac{2\mathcal{Q}^r}{\mathcal{Q}^r + \mathcal{Q}^{r'}} \Big\rangle_r + \frac{1}{2}\Big\langle \log_2 \frac{2\mathcal{Q}^{r'}}{\mathcal{Q}^r + \mathcal{Q}^{r'}} \Big\rangle_{r'} \tag{4}$$

where $\langle\cdot\rangle_r$ denotes averages over $P^r_{\rm post}$ (see Methods for evaluation details). This divergence is symmetric and only depends on the relative differences of selection factors between functional sub-repertoires, and not on the baseline model.

Clustering of cell types based on Jensen-Shannon divergence shows strong differential selection preferences between the $CD4^+$ and $CD8^+$ receptors, with an average $D_{\rm JS} \simeq 0.08 \pm 0.01$ bits across respective tissues and sub-repertoires (Fig. 4A; see also Fig. S6A for similar results where $P_{\rm gen}$ is used as baseline). We identify differential selection between Tconv and Treg receptors within $CD4^+$ cells with $D_{JS} \simeq 0.015 \pm 0.004$. We also detect moderate tissue specificity for $CD8^+$ and Treg receptors, but no such signal can be detected for $CD4^+$ Tconv cells across different tissues.

Examining the linear selection factors θ_f of the SONIA model trained with $P_{\rm gen}$ as a baseline reveals the VJ (Fig. S7) and amino-acid usage features (Fig. S8) that are differentially selected in the Tconv $CD4^+$ and $CD8^+$ subsets (in spleen). Linear selection models are organised according to a hierarchy from the least to the most constrained model. As one adds selection factors for each feature, the Kullback-Leibler divergence between the repertoire and the baseline increases (see Methods). Decomposing in this way the divergence between $CD4^+$ Tconv and $CD8^+$ repertoires, we find that contributions to the total divergence are evenly split between amino-acid features and VJ gene usage, with only a minor contribution from CDR3 length (Fig. S9).

Decomposing unsorted repertoires using selection models

Knowing specific $P^r_{\rm post}$ models specific to sub-repertoires enables us to infer the fraction of each class r in unsorted data. Estimating the relative fraction of $CD4^+$ and $CD8^+$ sub-types in a repertoire can be informative for clinical purposes, e.g. as a probe for Tumor Infiltrating Lymphocytes (TIL), where over-abundance of $CD8^+$ cells in the sample has been associated with posi-

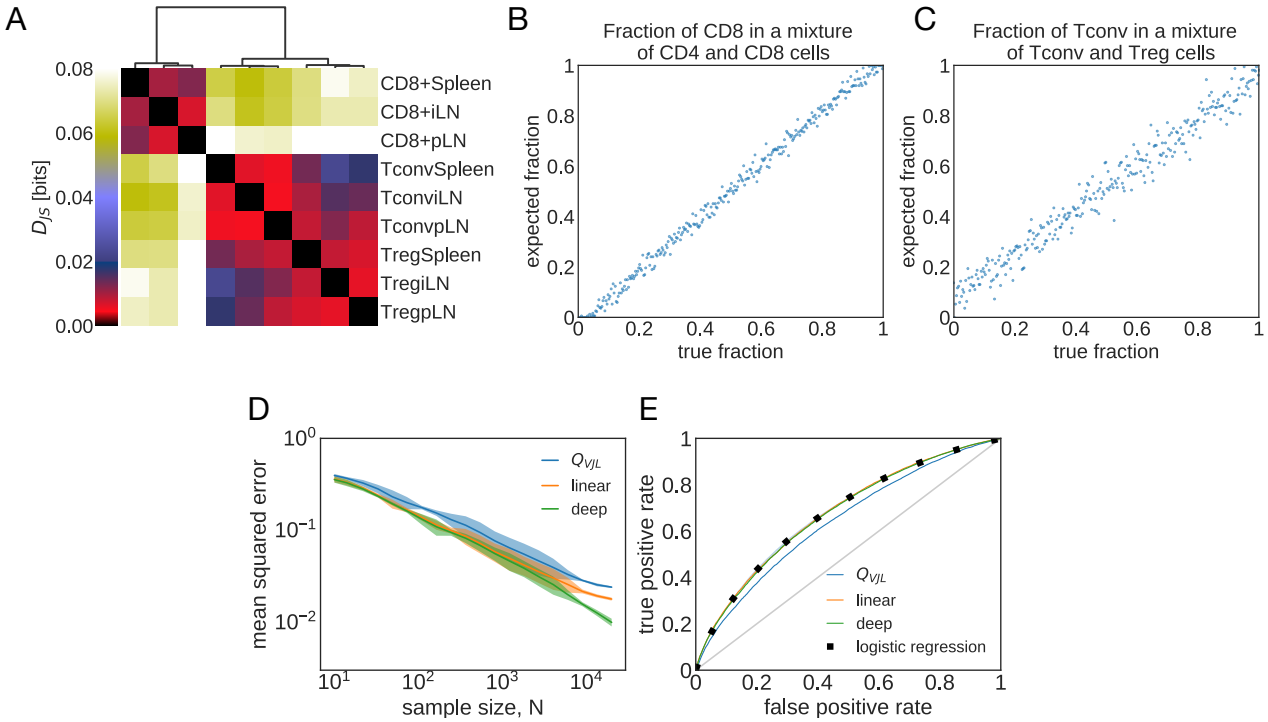


FIG. 4: **Cell type and tissue-specific selection on TCRs. (A)** Jensen-Shannon divergences (D_{JS}, see eq. 4) computed from models trained on different sub-repertoires are shown. **(B)** Maximum-likelihood inference of the fraction of CD8$^+$ TCRs in mixed repertoires of conventional CD4$^+$ T cells (Tconvs) and CD8$^+$ cells from spleen (Eq 5) is shown. Each repertoire comprises 5×10^3 unique TCRs. **(C)** Same as (B) but for a mixture of Tconv and Treg TCRs. **(D)** Mean squared error of the inferred sample fraction from (B) as a function of sample size N, averaged over all fractions, using models of increasing complexity: "Q_{VJL}" is a linear model with only features for CDR3 length and VJ usage, "linear" is linear SONIA model, "deep" is the full soNNia model (see Fig. 1 C). **(E)** Receiving-Operating Curve (ROC) for classifying individual sequences coming from CD8$^+$ cells or from CD4$^+$ Tconvs from spleen, using the log-likelihood ratios. Curves are generated by varying the threshold in eq. 6. The accuracy of the classifier is compared to a traditional logistic classifier inferred on the same set of features as our selection models. The training set for the logistic classifier has $N = 3 \times 10^5$ Tconv CD4$^+$, and $N = 8.7 \times 10^4$ CD8$^+$ TCRs, and the test set has $N = 2 \times 10^4$ CD4$^+$, and $N = 2 \times 10^4$ CD8$^+$ TCR sequences.

tive prognosis in ovarian cancer [45]. Given a repertoire composed of the mixture of two sub-repertoires r and r' in unknown proportions, we maximize the log-likelihood function $L(f)$ based on our selection models to find the fraction f of a sub-repertoire r within the mixture:

$$\begin{aligned} L(f) &= \langle \log(f P^r_{\text{post}}(\sigma) + (1-f) P^{r'}_{\text{post}}(\sigma)) \rangle_D \qquad (5) \\ &= \langle \log(f \mathcal{Q}^r(\sigma) + (1-f) \mathcal{Q}^{r'}(\sigma)) \rangle_D + \text{const}, \end{aligned}$$

where $\langle \cdot \rangle_D$ is the empirical mean over sequences in the mixture. Previous work has used differential V- and J-usage, and CDR3 length to characterize the relative fraction of CD4$^+$ and CD8$^+$ cells in an unfractionated repertoire [46]. The log-likelihood function in eq. 5 provides a principled approach for inferring cell-type composition using selection factors that capture the differential receptor features of each sub-repertoire, including but not limited to their V- and J- usage and CDR3 length and amino acid preferences.

To test the accuracy of our method, we formed a synthetic mixture of previously sorted CD4$^+$ (Tconv from spleen [43]) and CD8$^+$ (from spleen [43]) receptors with different proportions, and show that our selection-based inference can accurately recover the relative fraction of CD8$^+$ in the mix (Fig 4 B). Our method can also infer the proportion of Treg cells in a mixture of Tconv and Treg CD4$^+$ cells from spleen (Fig. 4C), which is a much harder task since these subsets are very similar (Fig. 4A). The accuracy of the inference depends on the size of the unfractionated data, with a mean expected error that falls below 1% for datasets with size 10^4 or larger for the CD8$^+$/CD4$^+$ mixture (red and orange lines in Fig. 4D).

Our method uses a theoretically grounded maximum likelihood approach, which includes all the features captured by the soNNia model. Nonetheless, a simple linear selection model with only V- and J- gene usage and CDR3 length information (blue line in Fig. 4D), analogous to the method used in ref. [46], reliably infers the composition of the mixture repertoire. Additional information about amino acid usage in the linear SONIA model results in moderate but significant improvement (orange line). The accuracy of the inference is insensitive to the choice of the baseline model for re-

ceptor repertoires: using the empirical baseline from ref. [44] (Fig. 4D) or $P_{\rm gen}$ (Fig. S6D) does not substantially change the results.

The method can be extended to the decomposition of 3 or more sub-repertoires. To illustrate this, we inferred the fractions of Tconv, Treg, and CD8$^+$ cells in synthetic unfractionated repertoires from spleen, showing an accuracy of $3 \pm 1\%$ in reconstructing all three fractions (Fig. S10) in a mixture of size 5×10^3.

Computational sorting of CD4$^+$ and CD8$^+$ TCR

Selection models are powerful in characterizing the broad statistical differences between distinct functional classes of immune receptors, including the CD4$^+$ and CD8$^+$ T-cells (Fig. 4A). A more difficult task, which we call computational sorting, is to classify *individual* receptors into functional classes based on their sequence features. We use selection models inferred for distinct sub-repertoires r and r' to estimate a log-likelihood ratio $R(\mathbf{x})$ for a given receptor $\mathbf{x}$ to belong to either of the sub-repertoires,

$$R(\mathbf{x}) = \log \frac{P^r_{\rm post}(\mathbf{x})}{P^{r'}_{\rm post}(\mathbf{x})} = \log \frac{\mathcal{Q}^r(\mathbf{x})}{\mathcal{Q}^{r'}(\mathbf{x})}. \qquad (6)$$

A larger log-likelihood ratio $R(\mathbf{x})$ indicates that the receptor is more likely to be associated with the sub-repertoire r than r'. We set a threshold R_c, to assign a receptor to r if $R(\mathbf{x}) \geq R_c$ and to r' otherwise. The sensitivity and specificity of this classification depends on the threshold value. We evaluate the accuracy of our log-likelihood classifier between sets of CD8$^+$ and Tconv CD4$^+$ receptors harvested from spleen [43]. The Receiver Operating Characteristic (ROC) curve in Fig. 4E shows that our selection-based method can classify receptors as CD8$^+$ or CD4$^+$ cells, with an area under the curve AUC $= 0.68$. Performance does not depend on the choice of the baseline model ($P_{\rm emp}$ in Fig. 4E and $P_{\rm gen}$ in Fig. S6E). Applying this classification method to all the possible pairs of sub-repertoires in Fig. 4A, we find that CD4$^+$ vs CD8$^+$ discrimination generally achieves AUC$\approx$ 0.7, while discriminating sub-repertoires within the CD4$^+$ or CD8$^+$ classes yields much poorer performance (Fig. S11).

For comparison, we also used a common approach for categorical classification and optimized a linear logistic classifier that takes receptor features (similar to the selection model) as input, and classifies receptors into CD8$^+$ or CD4$^+$ cells. The model predicts the probability that sequence $\mathbf{x}$ belongs to sub-repertoire r (rather than r') as $\hat{y}(\mathbf{x}) = \zeta(R_{\log}(\mathbf{x}))$, with $R_{\log}(\mathbf{x}) = \sum_f w_f x_f + b$ and $\zeta(x) = e^x/(1+e^x)$. We learn the model parameters w_f and b by maximizing the log-likelihood of the training set:

$$\mathcal{L}_c(\mathbf{w}, b) = \sum_{i=1}^{N} \Big[y_i \log \hat{y}(\mathbf{x}) + (1 - y_i) \log(1 - \hat{y}(\mathbf{x}_i)) \Big] \qquad (7)$$

where y_i labels each TCR by their sub-repertoire, e.g. $y_i = 1$ for CD8$^+$, and $y_i = 0$ for CD4$^+$. Note that when selection models are linear, the log-likelihood ratio (eq. 6) also reduces to a linear form—the only difference being how the linear coefficients are learned. This optimized logistic classifier (eq. 7) performs equally well compared to the selection-based classifier (eq. 4), with the same AUC=0.68 (points in Fig. 4E). These AUCs are comparable to those found in ref. [42], which has addressed the same issue using black-box machine learning approaches.

It should be emphasized that despite comparable performances, our fully linear selection-based method provides a biologically interpretable basis for subtype classification, in contrast to black box approaches [42]. Specifically, selection factors offer the possibility to interpret differences between cell types in terms of amino acid frequencies (Fig. S8 A and B) or V- and J- gene usage (Fig. S7). CD4$^+$ TCRs are more adverse to having a cysteine in the middle of their CDR3 relative to CD8$^+$ sequences. CD4$^+$ TCRs were reported to be more often associated with positively charged (lysine and arginine) amino acid, whereas CD8$^+$ TCRs with negatively charged (aspartic acid) amino acids [47, 48]. We calculated selection factors for positively and negatively charged amino acids and see no such correlation between selection factors and charge (Fig. S8C).

Classification of TCRs targeting distinct antigenic epitopes

Recognition of a pathogenic epitope by a TCR is mediated through molecular interactions between the two proteins. The strength of this interaction depends on the complementarity of a TCR against an antigen presented by a MHC molecule on the T-cell surface. Recent growth of data on paired TCRs and their target epitopes [27, 49] has led to the development of machine learning methods for TCR-epitope mapping [26–30]. A TCR-epitope map is a classification problem that determines whether a TCR binds to a specific epitope. We use our selection-based classifier (eq. 6) to address this problem. We determine the target ensemble $P^r_{\rm post}$ from the training set of TCRs associated with a given epitope (positive data), and the alternative ensemble $P^{r'}_{\rm post}$ from a set of generic unfractionated TCRs (negative data). For comparison, we also perform the classification task using the linear logistic regression approach (eq. 7), and the state of the art TCRex algorithm [29], which uses a random forest model for classification.

We performed classification for the following CD8$^+$-specific epitopes, presented on HLA-A*02 molecules: (i) the influenza GILGFVFTL epitope (with $N = 3107$ associated TCRs), (ii) the Cytomegalovirus (CMV) NLVPMVATV epitope ($N = 4812$), and (iii) the SARS-CoV-2 YLQPRTFLL epitope ($N = 315$). The first two epitopes have the most abundant associated TCR sets in VDJdb [27, 49], and the latter is relevant for the ongoing

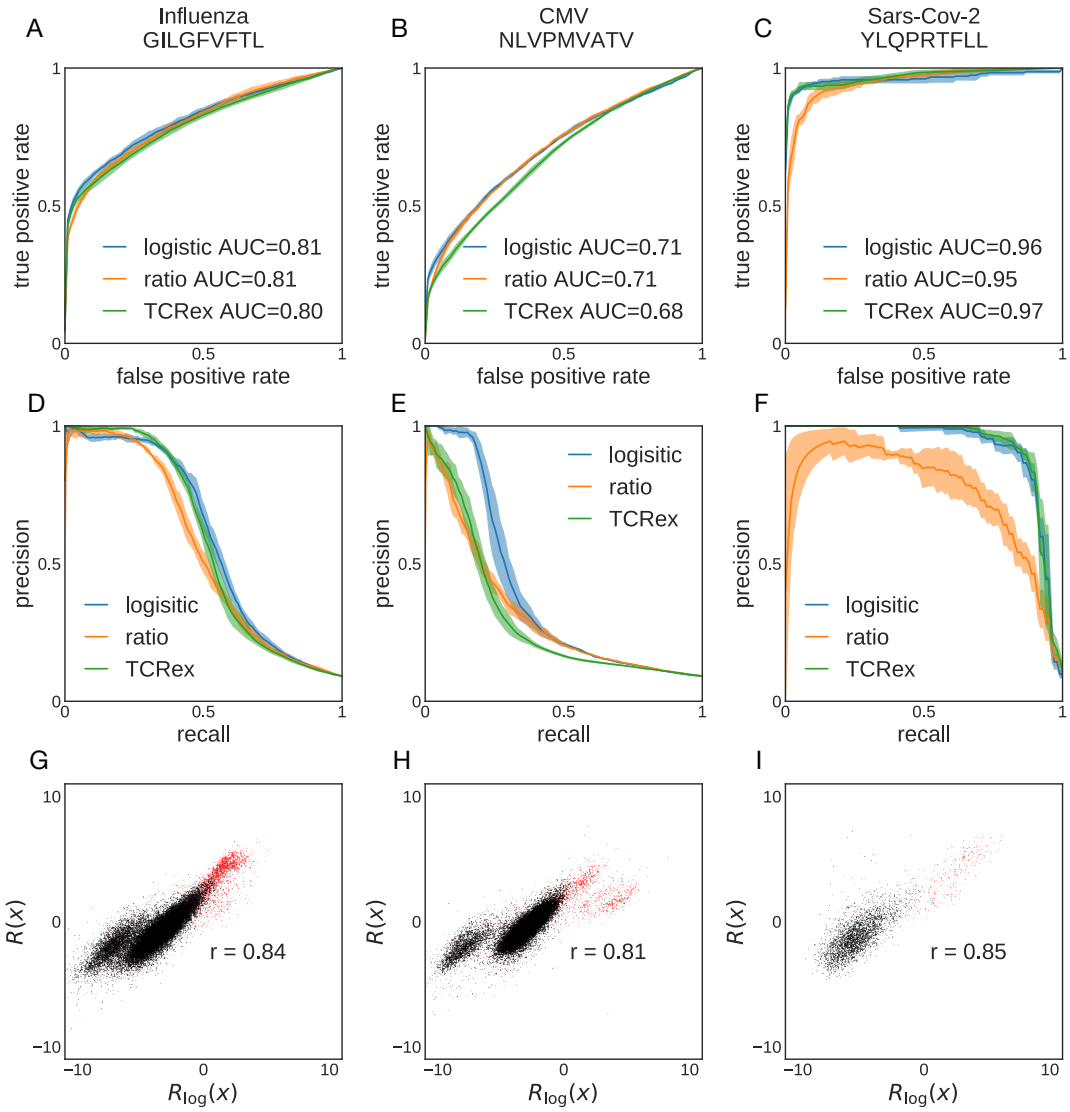


FIG. 5: **Selection-based prediction of epitope specificity for TCR.** TCRs are classified based on their reactivity to three pathogenic epitopes (columns), using three classification methods: TCRex, log-likelihood ratio (Eq. 6), and linear logistic regression (Eq. 7). **(A-C)** ROC curves, and **(D-F)** precision-recall curves for **(A,D)** influenza epitope GILGFVFTL ($N = 3107$ TCR), **(B,E)** CMV epitope NLVPMVATV ($N = 4812$), and **(C,F)** SARS-CoV-2 epitope YLQPRTFLL ($N = 315$) are shown. **(G-I)** Comparison between log-likelihood scores $R(x)$ and logistic regression scores $R_{\log}(x)$, for the three epitopes. Red points are TCRs that bind the specific epitope (positive set), black points are TCRs from bulk sequencing (negative set). r is Pearson's correlation. For all panels we used pooled data from Ref. [44] as the negative set. We used 10 times more negative data than positive data for training. Performance was quantified using 5-fold cross-validation.

COVD-19 pandemic. For consistency with TCRex [29], we used the pooled data from ref. [44] as the negative set, and used 10 times more negative data than positive data for training. To quantify performance of each classifier, we performed a 5-fold cross validation procedure. Due to the scarcity of data, we limit our selection inference to the linear SONIA model (see Fig. 1C). The ROC curves show comparable performances for the three classification methods on the three epitope-specific TCR sets (Fig. 5A-C).

The TCR-epitope mapping is a highly unbalanced classification problem, where reactive receptors against a specific epitope comprise a very small fraction of the repertoire (less than 10^{-5} [6]). Precision-recall curves are best suited to evaluate the performance of classification for imbalanced problems. In this case, a classifier should show a large precision (fraction of true predicted positives among all predicted positives) for a broad range of

recall or sensitivity (fraction of true predicted positives among positives = true positives + false negatives). The precision-recall curves in Fig. 5D-F show that TCRex and the logistic classifier can equally well classify the data, and moderately outperform the selection-based classifier. While both the logistic classifier and TCRex are optimized for classification tasks, the selection-based classifier is a *generative* model trained to infer the receptor distribution of interest (positive set) and identify its distinguishing features from the baseline (negative set). As a result, selection-based classification underperforms in the low-data regime, for which fitting a reliable distribution is difficult (e.g. for the SARS-CoV-2 epitope model, with only $N = 315$ positive examples). By contrast, the logistic classifier finds a hyperplane that best separates the two sets, and therefore, is better suited for classification tasks, and may be trained on smaller datasets. Nonetheless, we see a strong correlation between the selection-based log-likelihood ratio $R(x)$ (eq. 6) and the estimator of the logistic classifier $\hat{y}$ (eq. 7), shown for positive set (red points) and the negative set (black points) in Fig. 5G-I for the three epitopes. This result indicates that the separation hyperplane identified by the logistic classifier aligns well along the effective coordinates of selection that represent sequence features relevant for function in each epitope class.

III. DISCUSSION

Previous work has developed linear selection models to characterize the distribution of productive T cell receptors [2]. Here, we generalized on these methods by using deep neural networks implemented in the soNNia algorithm to account for nonlinearities in feature space, and have improved the statistical characterization of TCR repertoires in a large cohort of individuals [34].

Using this method, we modelled the selective pressure on paired chains of T- and B- cell receptors, and found that the observed cross-chain correlations, even if limited, could be partially reproduced with our model. These observed inter-chain correlations are consistent with previous analyses in TCRs [42, 50] and are likely due to the synergy of the two chains interacting with self-antigens presented during thymic development for TCRs and pre-peripheral selection (including central tolerance) for BCRs, or later when recognizing antigens in the periphery.

Our results show that the process of selection in BCRs is restrictive, in agreement with previous findings [7], significantly increasing inter-chain feature correlations. The selection strengths inferred by our models should not be directly compared to estimates of the percentage of cells passing pre-peripheral selection, $\sim 10\%$ for B cells versus $3 - 5\%$ for T cells [6]. Our models identify features under selection without making reference to the number of cells carrying these features. Since the T-cell pool in our analysis is a mixture of naive and memory cells, we can expect stronger selection pressure in the T-cell data than in the purely naive T cells. However, previous work analysing naive and memory TCRs separately using linear selection models did not report substantial differences between the two subsets [32].

We systematically compared T cell subsets and showed that our method identifies differential selection on $CD8^+$ T-cells, $CD4^+$ conventional T-cells, and $CD4^+$ regulatory T-cells. TCRs belonging to families with more closely related developmental paths (i.e., $CD4^+$ regulatory or conventional cells) have more similar selection features, which differentiate them from cells that diverged earlier ($CD8^+$). Cells with similar functions in different tissues are in general similar, with the exception of spleen $CD8^+$ that stands out from lymph node $CD8^+$.

One application of the soNNia method is to utilize our selection models to infer ratios of cell subsets in unsorted mixtures, following the proposal of Emerson et al. [46]. Consistently with previous results, we find that the estimated ratio of $CD4^+/CD8^+$ cells in unsorted mixtures achieves precision of the order of 1% with as few as 10^4 unique receptors. Emerson et al. validated their computational sorting based on sequence identity on data from in-vitro assays and flow cytometry, which gives us confidence that our results would also pass an experimental validation procedure.

As a harder task, we were also able to decompose the fraction of regulatory versus conventional $CD4^+$ T-cells, showing that receptor composition encodes not just signatures of shared developmental history— receptors of these two $CD4^+$ subtypes are still much more similar to each other than to $CD8^+$ receptors— but also function: Tregs down-regulate effector T-cells and curb an immune response creating tolerance to self-antigens and preventing autoimmune diseases [10], whereas Tconvs assist other lymphocytes including activation of differentiation of B-cells. Since our analysis is performed on fully differentiated peripheral cells, we cannot say at what point in their development these $CD4^+$ T-cells are differentially selected. Data from regulatory and conventional T-cells at different stages of thymic development could identify how their receptor composition is shaped over time.

During thymic selection cells first rearrange a β receptor and then an α receptor is added concurrently with positive selection. Negative selection follows positive selection and overlaps with CD4/CD8 differentiation. We found that the Jensen-Shannon divergence between $CD8^+$ and $CD4^+$ cells to be very small (0.1 bit) compared to the divergence between functional and generated repertoires (ranging from 0.8 to 0.9 bits). This result suggests that the selection factors captured by our model mainly act during positive selection, which is partly shared between $CD4^+$ and $CD8^+$ cells, rather than during cell type differentiation and negative selection, which is distinct for each type. Additionally to showing statistical differences in sub-repertoires, we classified cells into $CD4^+$ and $CD8^+$ subclasses with likelihood ratios of

selection models and recovered similar results achieved using pure machine learning approaches [42], but in a fully linear and interpretable setting.

In recent years multiple machine learning methods have been proposed in order to predict antigen specificity of TCRs: TCRex [29, 51], DeepTCR [52], netTCR [53], ERGO [54], TCRGP [55] and TcellMatch [56]. All these methods have explored the question in slightly different ways, and made comparisons with each other. However, with the sole exception of TcellMatch [56], none of the above methods compared their performance to a simple linear classifier. TcellMatch [56] does not explicitly compare to other existing methods, but implicitly compares various neural network architectures. We thus directly compared a representative of the above group of machine learning models, TCRex, to a linear logistic classifier, and to the log-likelihood ratio obtained by training two SONIA models on the same set of features. We found that the three models performed similarly (Fig. 5), consistent with the view that amino acids from the CDR3 loop interact with the antigenic peptide in an additive way. This result complements similar results in Ref. [56], where a linear classifier gave comparable results to deep neural network architectures.

The linear classifier based on likelihood ratios achieves state-of-the art performance both in discriminating $CD4^+$ from $CD8^+$ cells (Fig. 4 D), and in predicting epitope specificity (Fig. 5). But unlike other classifiers, its engine can be used to generate positive and negative samples. Thus characterizing the distributions of positive and negative examples is more data demanding than mere classification. For this reason pure classifiers are generally expected to perform better, but lack the ability to sample new data. Our analysis complements the collection of proposed classifiers by adding a generative alternative that is grounded on the biophysical process of T-cell generation and selection. This model is simple and interpretable, and performs well with large amounts of data.

The epitope discrimination task discussed here and in previous work focuses on predicting TCR specificity to one specific epitope. A long-term goal would be to predict the affinity of any TCR-epitope pair. However, currently available databases [27, 49] do not contain sufficiently diverse epitopes to train models that would generalize to unseen epitopes [56]. A further complication is that multiple TCR specificity motifs may co-exist even for a single epitope [30, 57], which cannot be captured by linear models [58]. Progress will be made possible by a combination of high-throughput experiments assaying many TCR-epitope pairs [59], and machine learning based techniques such as soNNia.

In summary, we show that nonlinear features captured by soNNia capture more information about the initial and peripheral selection process than linear models. However, deep neural network methods such as soNNia suffer from the drawback of being data hungry, and show their limitations in practical applications where data are scarce. In a more general context, soNNia is a way to integrate more basic but interpretable knowledge-based models and more flexible but less interpretable deep-learning approaches within the same framework.

IV. METHODS

SoNNia

SoNNia is python software which extends the functionality of the SONIA package. It expands the choice of selection models to infer, by adding non linear single-chain models and (non-)linear paired-chain models. The pre-processing pipeline implemented in this paper is also included in the package as a separate class. The software is available on GitHub at `https://github.com/statbiophys/soNNia`.

Pre-processing steps

The standard pre-processing pipeline, which is implemented in the soNNia package and is applied to all datasets, consists of the following steps:

1. Select species and chain type
2. Verify sequences are written as V gene, CDR3 sequence, J gene and remove sequences with unknown genes and pseudogenes
3. Filter productive CDR3 sequences (lack of stop codons and nucleotide sequence length is a multiple of 3)
4. Filter sequences starting with a cysteine
5. Filter sequences with CDR3 amino acid length smaller than a maximum value (set to 30 in this paper)
6. Remove sequences with small read counts (optional).

For the analysis of Fig. 2 we analysed data from [34]. We first applied the standard pipeline. In addtion we excluded TCRs with gene TRBJ2-5 which is badly annotated by the Adaptive pipeline [22] and removed a cluster of artefact sequences, which was previously identified in [60] and corresponds to the consensus sequence CFFKQKTAYEQYF.

For the analysis of Fig. 3 we analysed data from [40] and [41]. Dataset from [40] was obtained already pre-processed directly from the authors, while pre-processed dataset from [41] is part of the supplementary material of the corresponding paper. The soNNia standard pipeline is then applied to both datasets, independently for each chain, and a pair is accepted only if it passes both filtering steps. For α TCR datasets, sequences carrying the

following rare genes were removed due to their rarity in the out-of-frame dataset: TRAJ33, TRAJ38, TRAJ24, TRAV19.

For the analysis of Figs. 4 and 5 we analysed data from [43] and [44], to which we applied our standard preprocessing pipeline.

Generation model

The generation model relies on previously published models described in [1, 3, 31]. Briefly, the model is defined by the probability distributions of the various events involved in the VDJ recombination process: V, D, and J gene usage, and number of deletions and insertions at each junction. The model is learned from nonproductive sequences using the IGoR software [1]. For BCR, only a few nonproductive sequences were available, and so we instead started from the default IGoR models learned elsewhere [1], and re-inferred only the V gene usage distribution for the heavy chain, and VJ joint gene distribution for light chains, keeping all other parameters fixed.

Amino-acid sequence probability computation and generation is done with the OLGA software, which relies on a dynamic programming approach. The process is applied to all α, β, IgH and Igκ/λ chains. We focus on naive B cells and ignore somatic hypermutations. Since it was shown that individual variability in generation was only small [2], for each locus we used a single universal model.

Neural network architectures

We describe the architecture of the soNNia neural network. The input of our network is a vector $\mathbf{x}$ where $x_f = 1$ (otherwise 0) if sequence x has feature f. A dense layer is a map $\mathbf{L}(\mathbf{x}) = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$ with $\mathbf{x}$ the input vector, $\mathbf{W}$ the matrix of weights, $\mathbf{b}$ the vector bias, and where the tanh function is applied to each element of the input vector. The model architecture of the neural network is shown in Supplementary Fig. S1.The input is first subdivided into 3 sub-vectors: the $\mathbf{x}_L$ subset of features associated with CDR3 length, the $\mathbf{x}_A$ subset of features associated with the CDR3 amino acid composition and the $\mathbf{x}_{VJ}$ subset of features associated with V and J gene usage. We applied a dense layer individually to $\mathbf{x}_L$ and $\mathbf{x}_{VJ}$. In parallel we performed an amino acid embedding of $\mathbf{x}_A$: we first reshape the vector to a $2K \times 20$ matrix $\mathbf{A}$ (the set of features associated with amino acid usage is $2 \times K \times 20$ long, where $K = 25$ the maximum distance from the left and right ends that we encode, and 20 is the number of amino acids) and apply a linear embedding trough $\mathcal{M}(\mathbf{A}) = \mathbf{A}\mathbf{M}$ with $\mathbf{M}$ a $20 \times n$ matrix with n the size of the amino acid encoding. We then flatten the matrix to an array and apply a dense layer. We merged the three transformed subsets into a vector and then applied a dense layer. We finally applied a last dense layer without non-linearity to produce the output value, $\log Q$(see Fig S1).

The model for paired chains focuses on combining the $\mathbf{x}_L$ and $\mathbf{x}_{VJ}$ inputs of the two chains. First the $\mathbf{x}_L$ and $\mathbf{x}_{VJ}$ inputs within each chain are merged and processed with a dense layer. Subsequently a Batch Normalizing Transform is applied to each encoded vector to enforce a comparable contribution of each chain once the vectors are merged and processed through a dense layer (this last step is skipped in the deep-indep model). A Batch Normalizing Transform [61] is a differentiable operator which is normally used to improve performance, speed and stability of a Neural Network. Given a batch of data, it normalizes the input of a layer such that it will have mean output activation 0 and standard deviation of 1. In parallel the aminoacid inputs are embedded as described before. Finally all the vectors are merged together and a dense layer without activation outputs the $\log Q$ (see Fig S2-3).

soNNia model inference

Given a sample of data sequences $\mathcal{D} = \{\mathbf{x}^i\}_{i=1}^{N_D}$ and a baseline $\mathcal{G} = \{\mathbf{x}'^i\}_{i=1}^{N_G}$ we want to maximize the average log-likelihood:

$$\begin{aligned}\mathcal{L}(\theta) &= \langle \log P^\theta_{\text{post}} \rangle_{\mathcal{D}} = \frac{1}{N_D}\sum_{i=1}^{N_D} \log P^\theta_{\text{post}}(\mathbf{x}^i) \\ &= \frac{1}{N_D}\sum_{i=1}^{N_D} [\log Q^\theta(\mathbf{x}^i) + \log P_{\text{gen}}(\mathbf{x}^i)] - \log Z_\theta \\ &= \langle \log Q^\theta \rangle_D + \langle \log P_{\text{gen}} \rangle_D - \log \langle Q^\theta \rangle_G,\end{aligned} \tag{8}$$

where $Z_\theta = \langle Q^\theta \rangle_{\mathcal{G}} = N_G^{-1} \sum_{i=1}^{N_G} Q^\theta(\mathbf{x}'^i)$. The P_{gen} term in the last equation is parameter independent and can thus be discarded in the inference. When an empirical baseline is used, P_{gen} is replaced by $P_{\text{emp}}(\mathbf{x}) = N_G^{-1} \sum_{i=1}^{N_G} \delta_{\mathbf{x},x'_i}$.

The above likelihood is implemented in the soNNia inference procedure (linear and non-linear case) with the Keras [62] package. The model is invariant with respect to the transformation $Q(\mathbf{x}) \to cQ(\mathbf{x})$ and $Z \to Z/c$, where c is an arbitrary constant, so we fix dynamically the gauge $Z = 1$. We lift this degeneracy by adding the penalty $\Gamma(\theta) = (Z_\theta - 1)^2$, and minimize $-\mathcal{L}_{sonia}(\theta) + \gamma\Gamma(\theta)$ with $\gamma = 1$ as a loss function.

In our implementation batch sizes between $10^3 - 10^4$ sequences produced a reliable inference. L2 and L1 regularization on kernel weights are also applied. Hyperparameters were chosen using a validation dataset of size 10 % of training data.

To learn the Q_{VJL} model of Fig. 4, we used a linear SONIA model where features f where restricted to V, J and CDR3 length features. One major difference

with the approach of Ref. [46] is that, unlike the likelihood they use, we do not double-count the distribution of length (through $P(L|V)P(L|J)$). However, our results show that that error does not affect model performance substantially.

Hierarchy of models in linear SONIA

The linear SONIA model,

$$Q^{\theta}(\mathbf{x}) = e^{\sum_f \theta_f x_f}, \tag{9}$$

may be rationalized using the principle of minimum discriminatory information. In this scheme, we look for the distribution $P_{\rm post}$ that is most similar to our prior, described by the baseline set $P_{\rm gen}$ (or empirical set $\mathcal{G}$, replacing $P_{\rm gen}$ by $P_{\rm emp}(\mathbf{x}) = N_G^{-1}\sum_{i=1}^{N_G}\delta_{\mathbf{x},\mathbf{x}^i}$), but that still reproduces the marginal probabilities in the data. This translates to the minimization of the functional:

$$\mathcal{F}(P_{\rm post}) = D_{\rm KL}(P_{\rm post}\|P_{\rm gen}) - \eta_0\Big(\sum_{\mathbf{x}} P_{\rm post}(\mathbf{x}) - 1\Big) - \sum_f \theta_f\Big(P_{\rm post}(f) - P_{\rm data}(f)\Big), \tag{10}$$

where

$$D_{\rm KL}(P_{\rm post}\|P_{\rm gen}) \doteq \sum_{\mathbf{x}} P_{\rm post}(\mathbf{x})\log\frac{P_{\rm post}(\mathbf{x})}{P_{\rm gen}(\mathbf{x})}. \tag{11}$$

The second term on the right-hand side imposes the normalization of $P_{\rm post}$ and the last term imposes the constraint that the marginal probabilities of the selected set of features f should match those in the data through the set of Lagrange multipliers θ_f. This scheme reduces to the maximum entropy principle when $\mathcal{G}$ is uniformly distributed. Minimization of Eq. 10 results in:

$$P_{\rm post}(\mathbf{x}) = \frac{e^{\sum_f \theta_f x_f}}{Z_\theta} P_{\rm gen}(\mathbf{x}), \tag{12}$$

where $Z_\theta = e^{1-\eta_0}$, which is equivalent to Eq. 9.

Because of the principle of Kullback-Leibler divergence minimization, adding new constraints on the features to the optimization necessary increases $D_{\rm KL}$. This allows us to define a hiearchy of models as we add new constraints.

To evaluate the relative contributions of each feature to the difference between CD4 and CD8 TCR, we define different models based on a baseline set $\mathcal{G}$ defined as empirical sequences, with (1) only CDR3 length features; (2) CDR3 length and amino acid features; (3) CDR3 length and VJ features; and (4) all features. We denote the corresponding KL divergences (Eq. 11) $D^r_{\rm KL}(L)$, $D^r_{\rm KL}(A)$, $D^r_{\rm KL}(VJ)$, and $D^r_{\rm KL}$(full) for each subrepertoire r =CD4 or CD8, with $D^r_{\rm KL}(\text{full}) \geq D^r_{\rm KL}(A), D^r_{\rm KL}(VJ) \geq D^r_{\rm KL}(L)$. In Fig. S11 each of these divergences are then combined to get a "fractional Jensen-Shannon" divergence $D^f_{\rm JS} = fD^{\rm CD4}_{\rm KL} + (1-f)D^{\rm CD8}_{\rm KL}$, where f is the fraction of CD4 cells.

Estimation of information theoretic quantities

Given two random variables X and Y with joint distribution $p(x,y)$, the mutual information is:

$$I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{13}$$

and $P(x)$ and $P(y)$ are the respective marginal distributions of $p(x,y)$. $I(X,Y)$ can be naively estimated from data through the empirical histogram (x,y). The estimated mutual information $\hat{I}$ on a finite sample of data is affected by a systematic error [63]. We estimated the finite sample systematic error $I_0(X,Y)$ by destroying the correlations in the data through randomization. We implemented the randomization by mismatching CDR3-length, V and J assignment within the set. This mismatching procedure leads to the same marginals, $P(V)$ or $P(J)$, but destroys correlations, $P(V,J)-P(V)P(J) \simeq 0$. Errors on the Kullback divergences $D_{\rm KL}$ and Jensen-Shannon divergences, defined as

$$D_{\rm JS}(P,Q) = \frac{1}{2}D_{\rm KL}(P\|(P+Q)/2) + \frac{1}{2}D_{\rm KL}(Q\|(P+Q)/2), \tag{14}$$

are evaluated by computing the standard deviation of the above quantities using subsampled datasets of size one fifth of the original data.

V. ACKNOWLEDGEMENTS

AN and GI have been supported by the DFG grant (SFB1310) for Predictability in Evolution and the MPRG funding through the Max Planck Society. The work of TM and AMW was supported in part by grant ERCCOG n. 724208. The authors have no conflicts of interest.

[1] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* 9:561.

[2] Sethna Z, et al. (2020) Population variability in the generation and thymic selection of T-cell repertoires. *arXiv:2001.02843*.

[3] Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T (2019) OLGA: fast computation of generation probabili-

ties of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* 35:2974–2981.

[4] Tonegawa S (1983) Somatic generation of antibody diversity. *Nature* 302:575–581.

[5] Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and t-cell recognition. *Nature* 334:395–402.

[6] Yates AJ (2014) Theories and quantification of thymic selection. *Front. Immunol.* 5:13.

[7] Nemazee D (2017) Mechanisms of central tolerance for b cells. *Nat. Rev. Immunol.* 17:281–294.

[8] Murphy K, et al. (2008) *Janeway's Immunobiology*, Janeway's Immunobiology (Garland Science) No. v. 978,Num. 0-4129.

[9] Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the t cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* 14:377–391.

[10] Wing K, Sakaguchi S (2010) Regulatory T cells exert checks and balances on self tolerance and autoimmunity. *Nat. Immunol.* 11:7–13.

[11] Hou XL, Wang L, Ding YL, Xie Q, Diao HY (2016) Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes Immun.* 17:153–164.

[12] Georgiou G, et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32:158–68.

[13] Bolotin DA, et al. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12:380–381.

[14] Mcdaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G (2016) Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* 11:429–442.

[15] DeKosky BJ, et al. (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31:166–9.

[16] Turchaninova Ma, et al. (2013) Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* 43:2507–2515.

[17] Mcdaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G (2016) Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* 11:429–442.

[18] Dekosky BJ, et al. (2014) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21:1–8.

[19] Ndifon W, et al. (2012) Chromatin conformation governs T-cell receptor J gene segment usage. *Proc. Natl. Acad. Sci.* 109:15865–15870.

[20] Ralph DK, Matsen FA (2016) Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLoS Comput. Biol.* 12:1–25.

[21] Munshaw S, Kepler TB (2010) SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* 26:867–72.

[22] Davidsen K, et al. (2019) Deep generative models for T cell receptor protein sequences. *eLife* 8:e46935.

[23] Greiff V, et al. (2017) Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immun.* 199:2985–2997.

[24] Miho E, Roškar R, Greiff V, Reddy ST (2019) Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* 10:1321.

[25] Isacchini G, et al. (2020) Generative models of t-cell receptor sequences. *Phys. Rev. E* 101:062414.

[26] Glanville J, et al. (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature* 547:94–98.

[27] Shugay M, et al. (2018) VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 46:D419–D427.

[28] Jokinen E, Heinonen M, Huuhtanen J, Mustjoki S, Harri L (2019) TCRGP : Determining epitope specificity of T cell receptors. *Bioarchive* pp 4–12.

[29] Gielis S, et al. (2019) Detection of enriched t cell epitope specificity in full t cell receptor sequence repertoires. *Front. Immunol.* 10:2820.

[30] Dash P, et al. (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547:89–93.

[31] Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* 109:16161–16166.

[32] Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci.* 111:9875–9880.

[33] Elhanati Y, et al. (2015) Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140243.

[34] Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* 49:659–665.

[35] Grigaityte K, et al. (2017) Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the t cell repertoire. *bioRxiv:213462*.

[36] Dupic T, Marcou Q, Walczak AM, Mora T (2019) Genesis of the $\alpha\beta$ t-cell receptor. *PLoS Comput. Biol* 15:1–19.

[37] Shcherbinin DS, Belousov VA, Shugay M (2020) Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in tcr$\alpha\beta$ complex. *PLoS Comput. Biol* 16:1–17.

[38] Larimore K, McCormick MW, Robins HS, Greenberg PD (2012) Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* 189:3221–30.

[39] Glanville J, et al. (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci.* 106:20216–20221.

[40] Tanno H, et al. (2020) Determinants governing t cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci.* 117:532–540.

[41] DeKosky BJ, et al. (2016) Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl. Acad. Sci.* 113:E2636–E2645.

[42] Carter JA, et al. (2019) Single t cell sequencing demonstrates the functional role of $\alpha\beta$ tcr pairing in cell lineage and antigen specificity. *Front. Immunol.* 10:1516.

[43] Seay HR, et al. (2016) Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* 1:1–19.

[44] Dean J, et al. (2015) Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Medicine* 7:123.

[45] Sato E, et al. (2005) Intraepithelial cd8+ tumor-

infiltrating lymphocytes and a high cd8+/regulatory t cell ratio are associated with favorable prognosis in ovarian cancer. *Proc. Natl. Acad. Sci.* 102:18538–18543.

[46] Emerson R, et al. (2013) Estimating the ratio of cd4+ to cd8+ t cells using high-throughput sequence data. *Journal of Immunological Methods* 391:14 – 21.

[47] Li B, et al. (2016) Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* 48:725–732.

[48] Li HM, et al. (2016) Tcrβ repertoire of cd4+ and cd8+ t cells is distinct in richness, distribution, and cdr3 amino acid composition. *Journal of leukocyte biology* 99:505–513 26394815[pmid].

[49] Bagaev DV, et al. (2019) VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* pp 1–6.

[50] Dupic T, Marcou Q, Walczak AM, Mora T (2019) Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput. Biol.* 15:e1006874.

[51] De Neuter N, et al. (2018) On the feasibility of mining cd8+ t cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* 70:159–168.

[52] Sidhom JW, et al. (2019) Deeptcr: a deep learning framework for understanding t-cell receptor sequence signatures within complex t-cell repertoires. *bioRxiv:464107*.

[53] Jurtz VI, et al. (2018) Nettcr: sequence-based prediction of tcr binding to peptide-mhc complexes using convolutional neural networks. *bioRxiv:433706*.

[54] Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y (2020) Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Front. Immunol.* 11:1803.

[55] Jokinen E, Heinonen M, Huuhtanen J, Mustjoki S, Lähdesmäki H (2019) Tcrgp: Determining epitope specificity of t cell receptors. *bioRxiv:542332*.

[56] Fischer DS, Wu Y, Schubert B, Theis FJ (2020) Predicting antigen specificity of single t cells based on tcr cdr3 regions. *Molecular Systems Biology* 16:e9416.

[57] Minervina AA, et al. (2020) Primary and secondary antiviral response captured by the dynamics and phenotype of individual t cell clones. *eLife* 9:e53704.

[58] Bravi B, et al. (2020) *In preparation*.

[59] Klinger M, et al. (2015) Multiplex identification of antigen-specific t cell receptors using a combination of immune assays and immune receptor sequencing. *PLOS ONE* 10:1–21.

[60] DeWitt WS, et al. (2018) Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* 7:1–39.

[61] Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR* abs/1502.03167.

[62] Chollet F, et al. (2015) Keras. (`https://keras.io`).

[63] Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18:S231–S240.

[64] Tubiana J, Cocco S, Monasson R (2019) Learning protein constitutive motifs from sequence data. *eLife* 8:e39397.

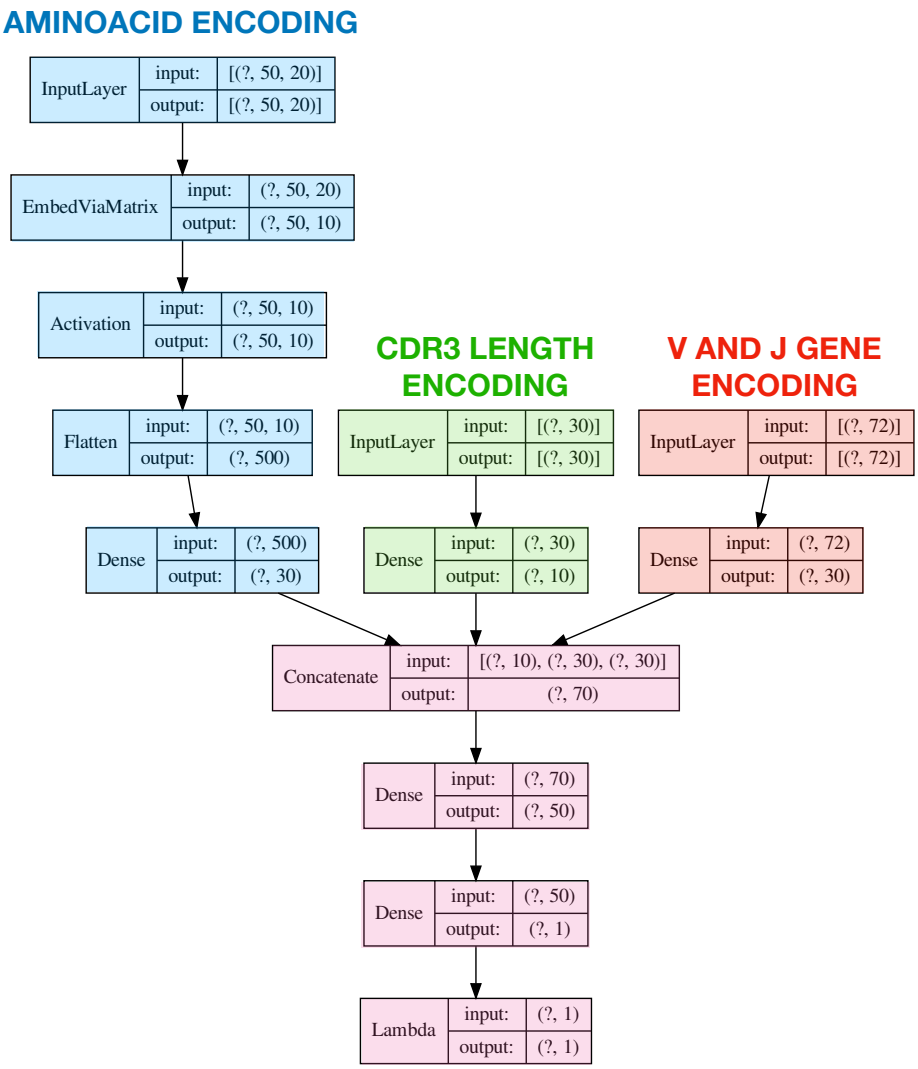


FIG. S1: Neural network structure of the deep soNNia model for the single chain case. There are three inputs, from left to right: first the encoded aminoacid composition of the CDR3 using the left-right encoding scheme, then the length of the CDR3, finally the independent V and J gene usage information. The aminoacid input is encoded using an embedding layer, called EmbedViaMatrix and then processed by a tanh non-linearity, called Activation layer. The Flatten layer turns the encoded matrix in the corresponding flattened array where each row of the matrix is concatenated to the successive one. A dense layer is then applied to reduce its dimensionality. The other two inputs are also processed through a dense feed-forward layer to reduce their corresponding dimensionality. The three groups of encoded inputs are then concatenated and two dense feed forward layers are applied to output $\log Q$. Finally $\log Q$ is clipped to avoid diverging values using the Lambda layer.

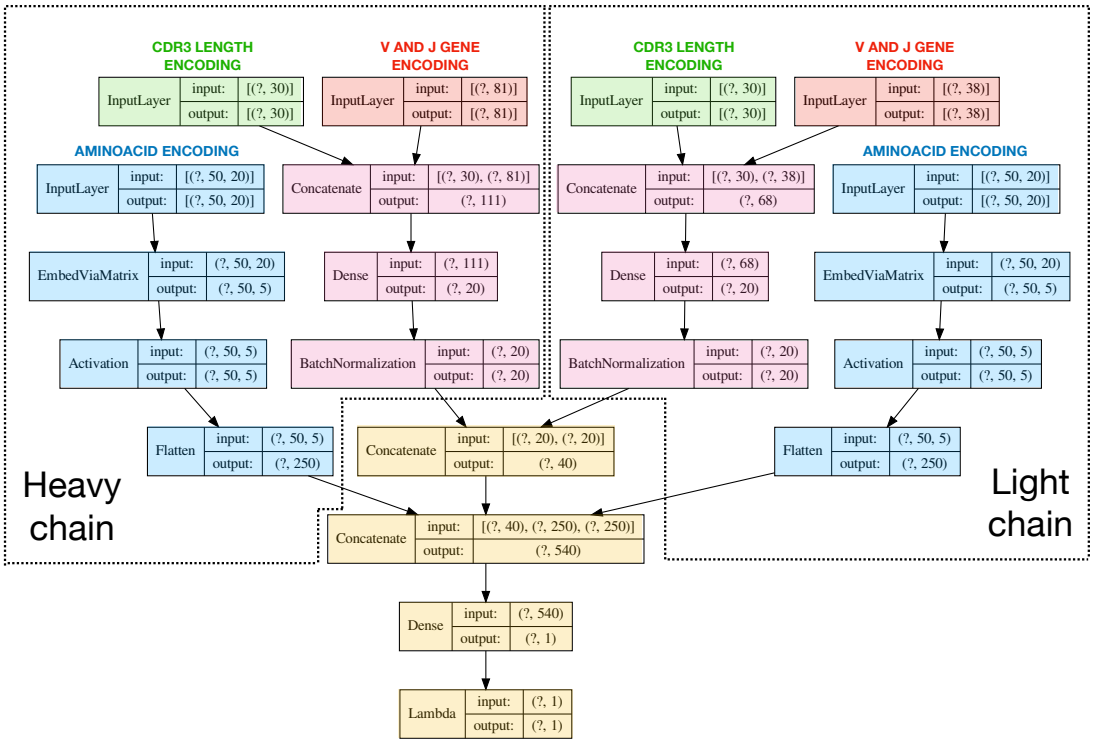


FIG. S2: Neural Network structure of the deep-indep model for paired chains. See Fig. S1 for details on what each layer does.

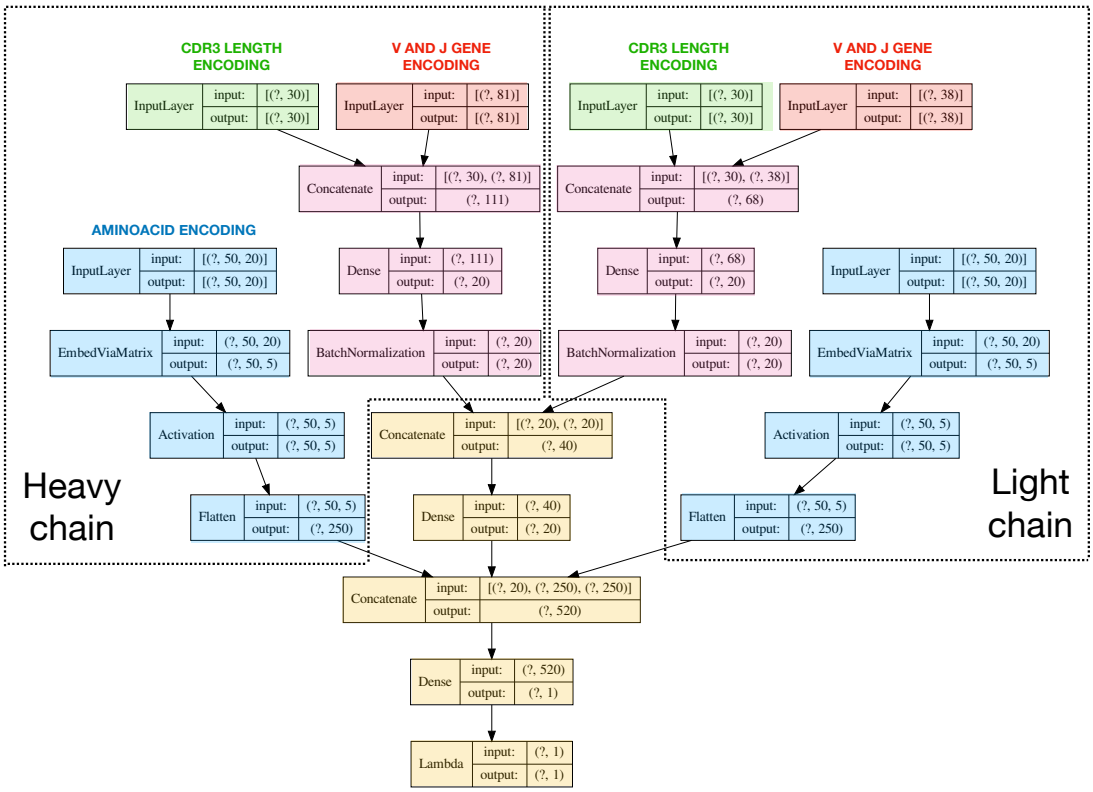


FIG. S3: Neural Network structure of the deep-joint model for paired chains. See Fig. S1 for details on what each layer does.

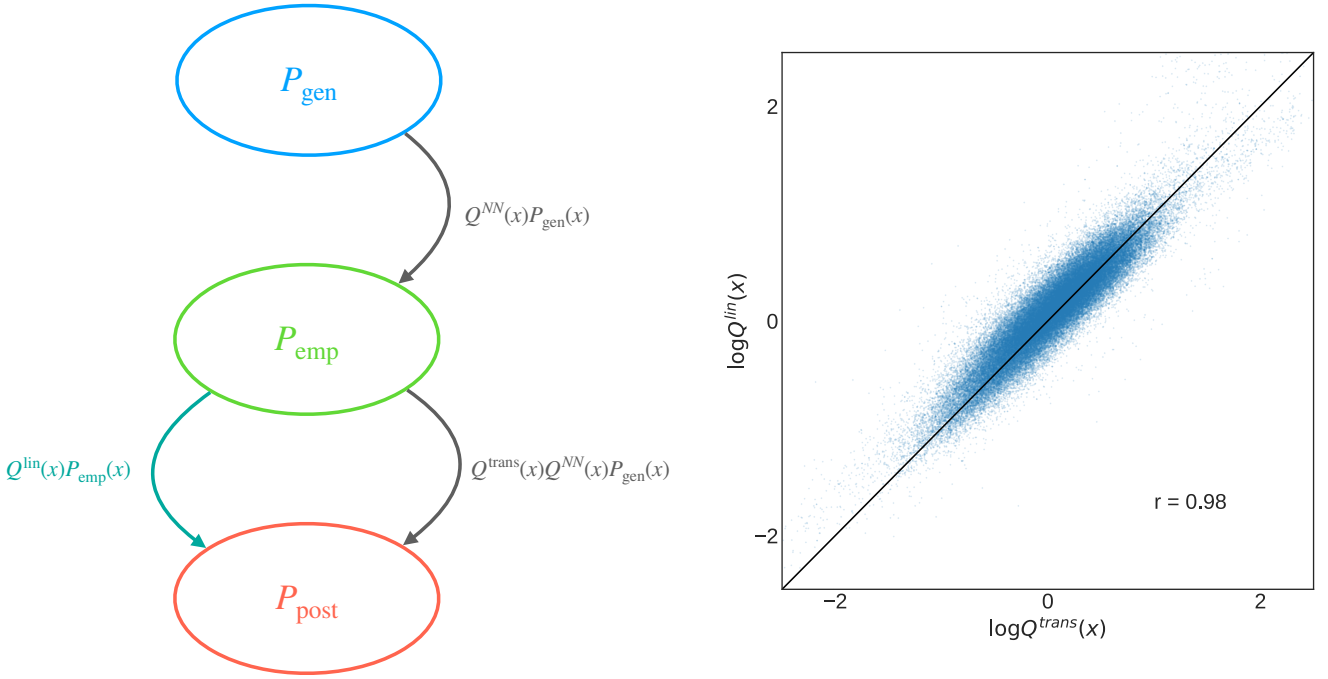


FIG. S4: Transfer learning consists generally in a 2-step inference: in the first step we infer a deep neural network on a bigger data set $\mathcal{G}$, in the second second step we re-infer a subsection of the network, or an additional layer on a smaller dataset, which is the real target. In our specific application, the big data $\mathcal{G}$ is unfractionated repertoire from [44] ($P_{\rm emp}(\mathbf{x}) = N_G^{-1}\sum_{i=1}^{N_G}\delta_{\mathbf{x},x_i'}$), and the real targets are the cells subsets harvested from different tissues [43]. We can infer a deep selection model to characterize well $P_{\rm emp}$ and then correct with one additional linear model. This procedure is equivalent to using $P_{\rm emp}$ as null distribution in the inference of a linear selection factor, as it can be seen by the high correlation between selection factors inferred with the two different methodologies. In this work we do not use transfer learning but instead work directly with the empirical baseline $\mathcal{G}$.

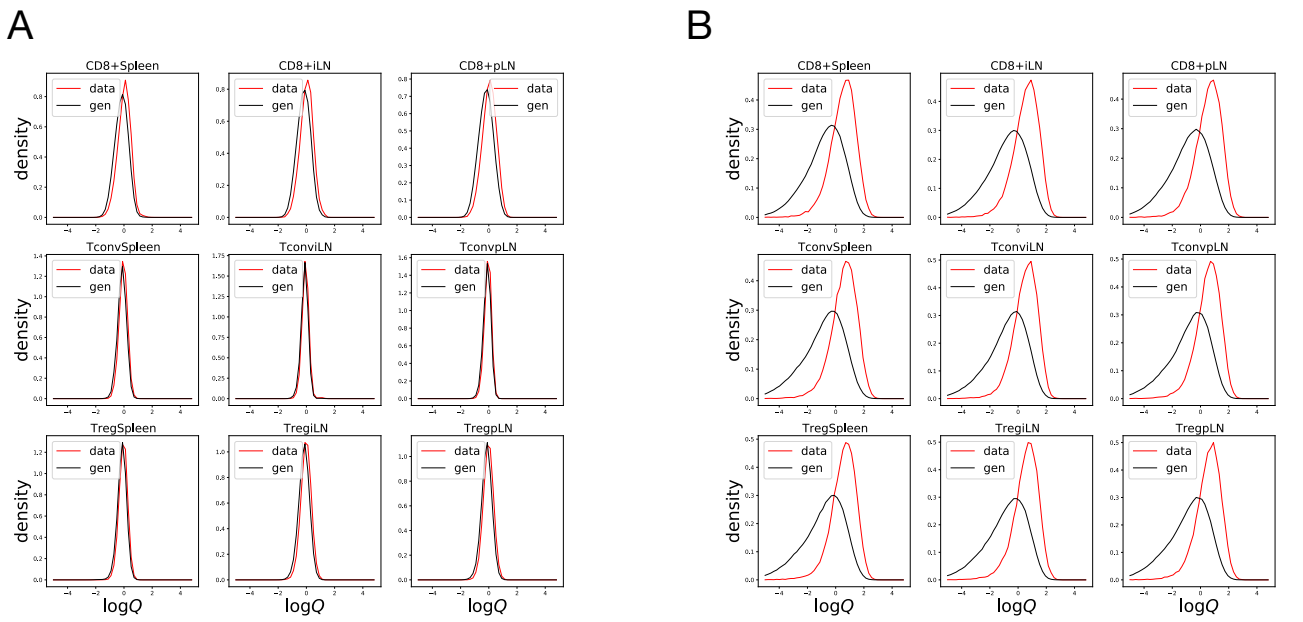


FIG. S5: (A) Distribution of $\log \mathcal{Q}$ of inferred models starting from a empirical baseline $\mathcal{G}$ and (B) Distribution of $\log \mathcal{Q}$ of inferred models starting from P_{gen} as a baseline.

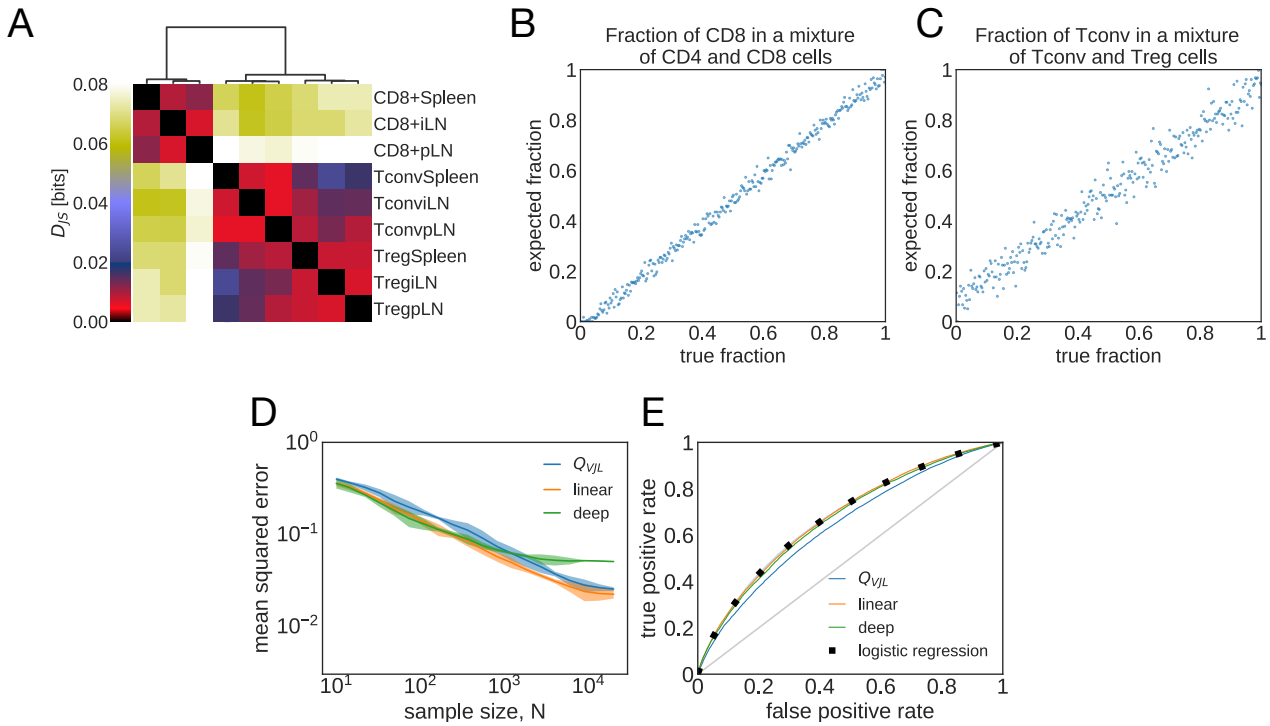


FIG. S6: Analogous to Fig 4 in main text but with P_{gen} as null model. (A) Jensen-Shannon divergences (D_{JS}, see Eq. 4) computed from models trained on different subrepertoires. (B) Maximum-likelihood inference of the fraction of CD8 TCR in mixed repertoires of Tconv and CD8 from spleen (Eq 5). Each repertoires comprises 5×10^3 unique TCR. (C) Same as (B) but for a mixture of Tconv and Treg TCR. (D) Mean squared error of the inferred sample fraction from (B) as a function of sample size N, averaged over all fractions, using models of increasing complexity: 'Q_{VJL}' is a linear model with only features for CDR3 length and VJ usage, 'linear' is linear SONIA model, 'deep' is the full soNNia model (see Fig. 1C). (E) Receiving-Operating Curve (ROC) for classifying individual sequences as coming from CD8 cells or from CD4+ conventional T cells from spleen, using the log-likelihood ratios. Curves are generated by varying the threshold in Eq. 6. The accuracy of the classifier is compared to a traditional logistic classifier inferred on the same set of features as our selection models. The training set for the logistic classifier has $N = 3 \times 10^5$ Tconv CD4, $N = 8.7 \times 10^4$ CD8 TCRs, and the test set has $N = 2 \times 10^4$ CD4, $N = 2 \times 10^4$ CD8 TCR sequences.

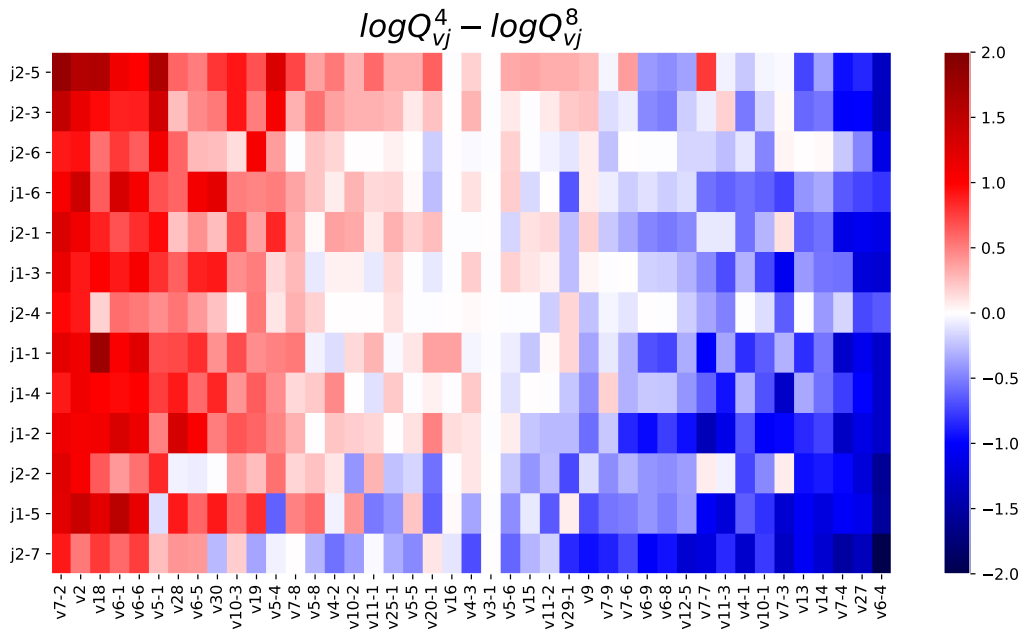


FIG. S7: Differential selection in V and J gene usage between $CD4^+$ and $CD8^+$ models

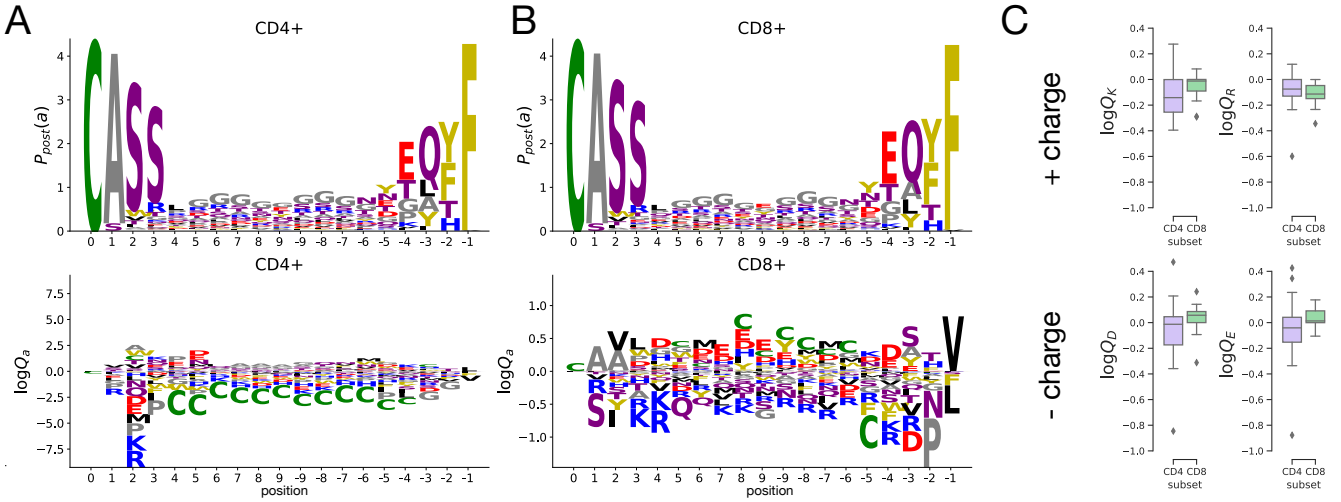


FIG. S8: Sequence motifs of selection factors associated to aminoacid composition for (A) $CD4^+$ and (B) $CD8^+$ cells. Logo plots produced with the code developed in [64] (C) Average selection factor for specific charged aminoacids (K–lysine, R–arginine, D–aspartate, E–glutamate), in $CD4^+$ and $CD8^+$ models. $Q_K = \exp(\theta_f)_{f\sim K}$ corresponds to selection factor associated to features that involve the presence of lysine (K) at any position, and likewise for R, D, and E aminoacids.

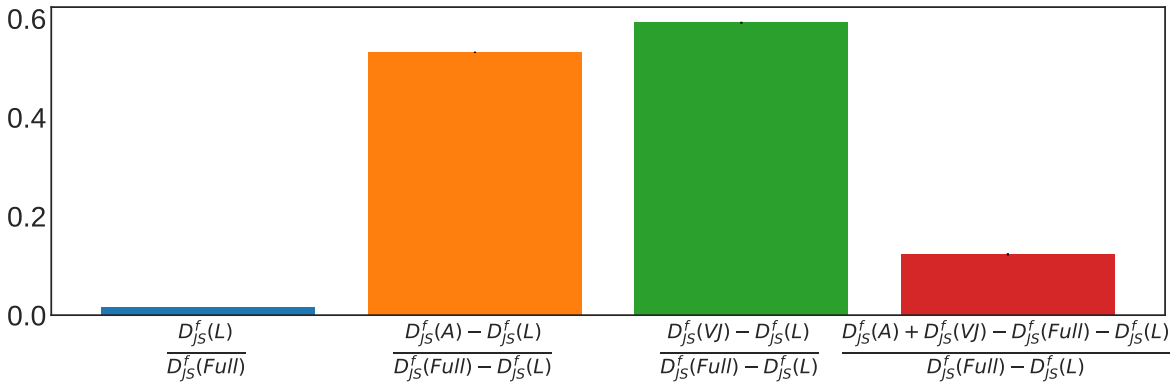


FIG. S9: Decomposition of contribution from different features to the fractional Jensen Shannon divergence between the CD4 and CD8 subpertoire statistics, $D^f_{\rm JS}(L) \leq D^f_{\rm JS}(A), D^f_{\rm JS}(VJ) \leq D^f_{\rm JS}(\text{full})$. The blue bar is the contribution of CDR3 length; orange and green bars are the relative contributions from the amino-acid composition and VJ usage, respectively. Red bar is the fraction that's redundant between VJ and amino acid usage. Contributions are balanced between amino acid and VJ usage, with moderate redundancy between the two.

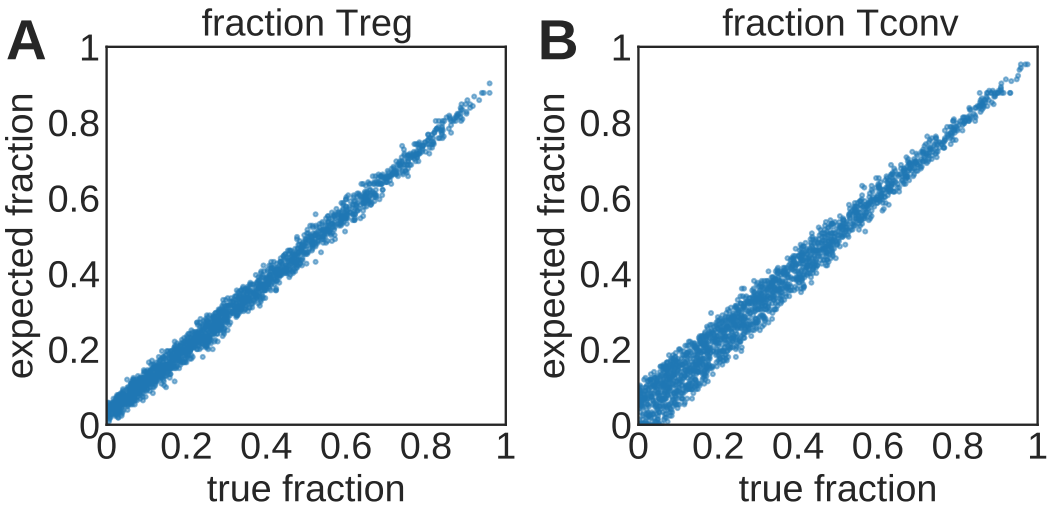


FIG. S10: Joint Inference of fraction of TCR belonging to different subclasses in a mixture of 3 repertoires: CD8+,CD4+ Tconv and CD4+ Treg cells. We optimize the likelihood $L(f_1, f_2) = \sum_i (f_1 Q_{\rm conv}(x_i) + f_2 Q_{\rm reg}(x_i) + (1 - f_1 - f_2) Q_{CD8}(x_i))$ to infer jointly the two fractions f_1 and f_2 in a chosen mixture of 3×10^4 TCRs x_i, built by combining repertoires of purified subsets harvested from spleen [43]. Each point corresponds to a mixture with f_1 and f_2 sampled uniformly 2000 times in the simplex $f_1 \geq 0$, $f_2 \geq 0$, $f_1 + f_2 \leq 1$.

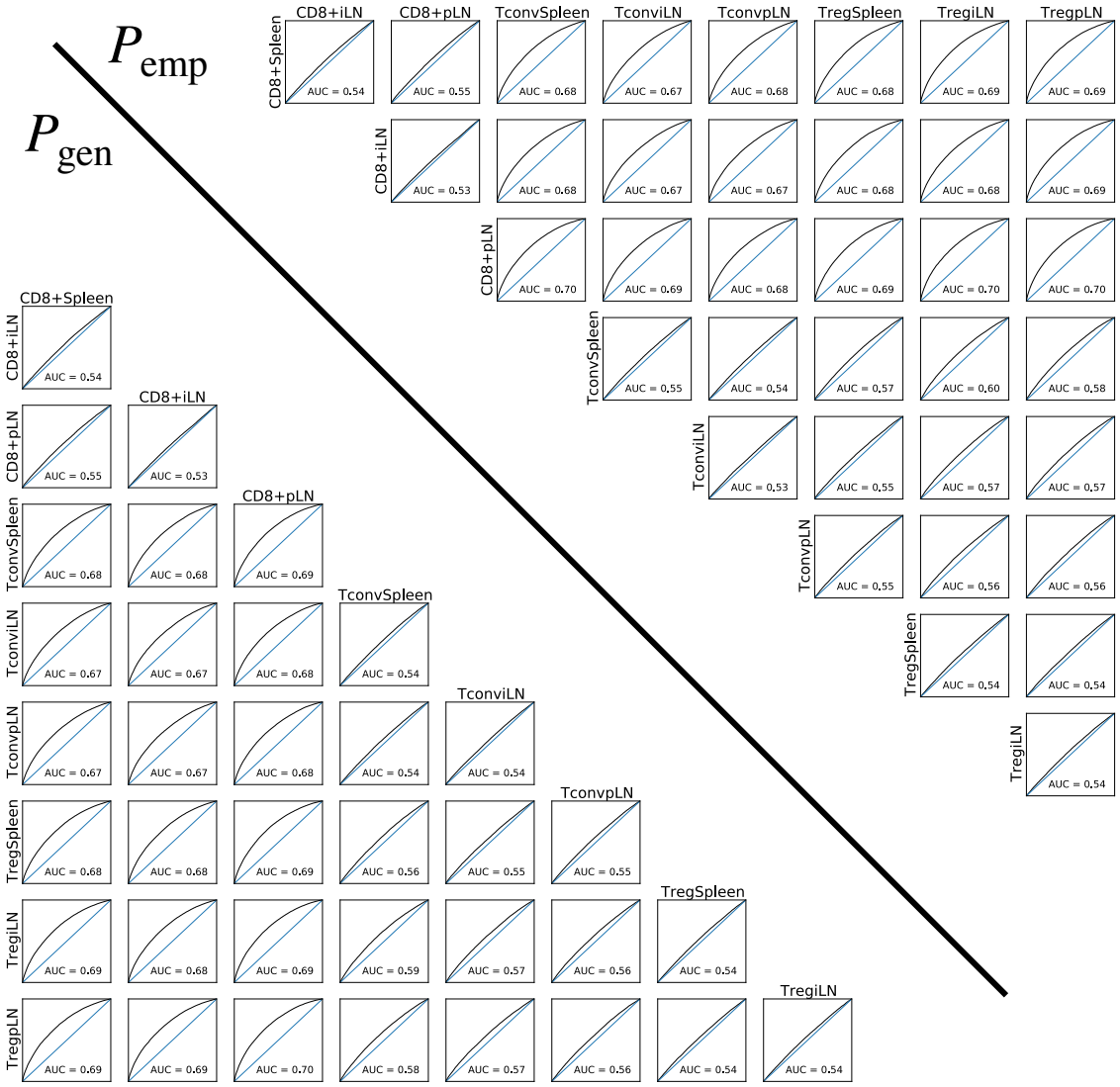


FIG. S11: ROC curve between all subsets based on the log ratio $R(x)$ defined on main text, where the selection factors are inferred starting from the empirical baseline $\mathcal{G}$ ($P_{\text{emp}}(\mathbf{x}) = N_G^{-1} \sum_{i=1}^{N_G} \delta_{\mathbf{x},x_i'}$, above diagonal) or P_{gen} (below diagonal).