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

Giulio Isacchini[a,b], Aleksandra M. Walczak[b,1,2], Thierry Mora[b,1,2], and Armita Nourmohammad[a,c,d,1,2]

[a]Statistical Physics of Evolving Systems, Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany; [b]Laboratoire de Physique de l'Ecole Normale Supérieure, Paris Sciences & Lettres (PSL) University, CNRS, Sorbonne Université and Université de Paris, 75005 Paris, France; [c]Department of Physics, University of Washington, Seattle, WA 98195; and [d]Herbold Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Edited by Andrew L. Ferguson, University of Chicago, Chicago, IL, and accepted by Editorial Board Member Mehran Kardar February 27, 2021 (received for review November 5, 2020)

Subclasses of lymphocytes carry different functional roles to work together and 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 cell 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.

adaptive immune repertoires | thymic selection | central tolerance | deep neural networks | statistical inference

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 of V-, D-, and J-genes, and sequence insertions and deletions at the junctions, a process known as V(D)J recombination (1, 2). Recognition of a pathogen by a T cell receptor (TCR) or B cell receptor (BCR) is mediated through molecular interactions between an immune receptor protein and a pathogenic epitope. TCRs interact with short protein fragments (peptide antigens) from the pathogen that are presented by specialized pathogen-presenting major histocompatibility complexes (MHCs) on cell surface. BCRs 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 contributes 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 twofold 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 (Tregs; CD4$^+$ FOXP3$^+$) (10), each of which can be found in the nonantigen 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 subtypes 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 both for single-chain (11–14) and paired-chain (15–17) BCRs and TCRs 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

Significance

The adaptive immune system relies on many types of B and T cells, whose functions are reflected in the distinct molecular features of their receptor sequences. Here, we introduce an inference framework, soNNia, which integrates interpretable knowledge-based models of immune receptor generation with flexible and powerful deep learning approaches to characterize sequence determinants of receptor function. Using soNNia, we characterize sequence-specific selection associated with receptors harvested from different cell types and tissues. We quantify synergetic interactions between the molecular features of the paired chains making up the receptor. Lastly, we develop a selection-based classifier to identify T cells specific to distinct pathogenic epitopes. Our approach provides a molecular understanding for how sequence determines the specific functionality of immune receptors.

Author contributions: G.I., A.M.W., T.M., and A.N. designed research; G.I., A.M.W., T.M., and A.N. performed research; G.I. analyzed data; and G.I., A.M.W., T.M., and A.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.L.F. is a guest editor invited by the Editorial Board.

[1] A.M.W., T.M., and A.N. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: aleksandra.walczak@phys.ens.fr, thierry.mora@phys.ens.fr, or armita@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2023141118/-/DCSupplemental.

Published April 1, 2021.

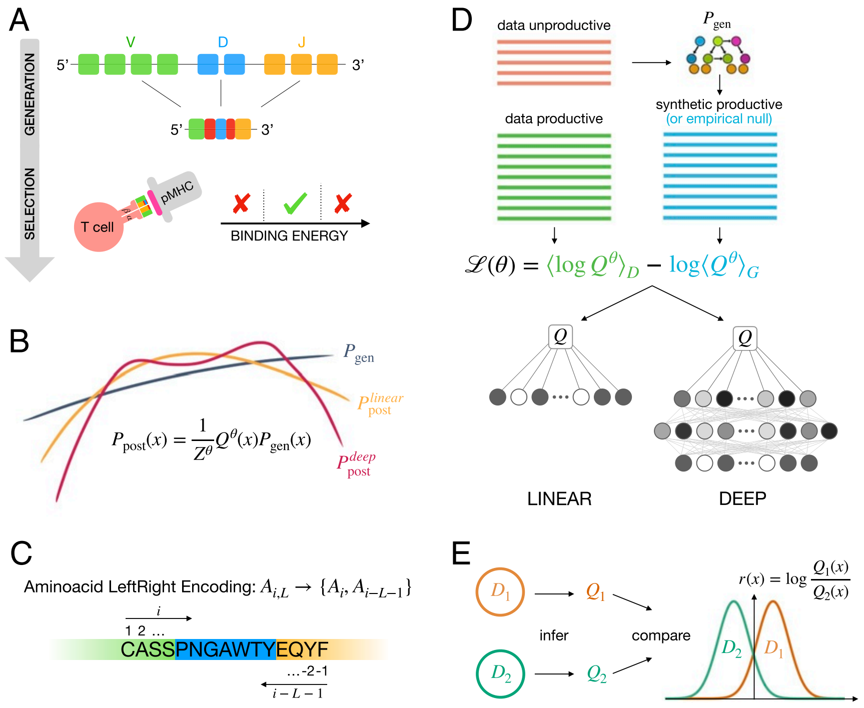


Fig. 1. Inference of functional selection models for immune receptor repertoires. (*A*) TCR α and β chains are stochastically rearranged through a process called V(D)J recombination. Successfully rearranged receptors undergo selection for binding to self-pMHCs (self peptides loaded onto major histocompatibility complexes). Receptors that bind too weakly or too strongly are rejected, while intermediately binding ones exit the thymus and enter peripheral circulation. Development of BCRs 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, $P_{\rm gen}(\sigma)$, are inferred using the IGoR software (3). $P_{\rm gen}(\sigma)$ acts as a baseline for the selection model. We then infer selection factors $\mathcal{Q}$, which act as weights that modulate the initial distribution $P_{\rm gen}(\sigma)$. We infer two types of selection weights: linear in log space [using the SONIA software (4)] and nonlinear weights using a DNN, in the soNNia software presented here. Nonlinear selection weights are more flexible than linear ones. (*C*) Pipeline of the algorithm: $P_{\rm gen}$ is inferred from unproductive sequences using IGoR. Selection factors for both the linear and nonlinear 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 (5). (*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, for example, log-likelihood ratios $r(x)$.

receptors (3–5, 18). Machine learning techniques have also been used to infer deep generative models to characterize the T cell repertoire composition as a whole (21), as well as discriminate between public and private B cell clones based on complementarity-determining region 3 (CDR3) sequence (22, 23). While biophysically informed models can still match and even outperform machine learning techniques (24), 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 biophysical tools to model what we know (e.g., receptor generation) and exploit the powerful machinery of deep neural networks (DNNs) to model what we do not know (e.g., functional selection). Using the nonlinear and flexible structure of the DNNs, we characterize the sequence properties that encode selection of the specificity of the combined chains during receptor maturation in α and β 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 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 (25–29).

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 (3, 30). In brief, these models are parametrized according to the probabilities by which different V, D, J genes are used and base pairs are inserted in or deleted from the CDR3 junctions to generate a receptor sequence. We infer the parameters of these models using the IGoR software (3) 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 preselection receptors (5). These generated receptors define a baseline $\mathcal{G}$ for statistics of repertoires prior to any functional selection.

The amino acid sequence of an immune receptor protein determines its function. To identify sequence properties that are linked to function, we compare the statistics of sequence features f (e.g., V-, J-gene usage and CDR3 amino acid composition) present in a given B or T cell functional repertoire with 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 $\mathcal{Q}^\theta(\mathbf{x})$,

$$P^{\theta}_{\text{post}}(\mathbf{x}) = P_{\text{gen}}(\mathbf{x})\mathcal{Q}^{\theta}(\mathbf{x}) \equiv \frac{1}{Z^{\theta}} P_{\text{gen}}(\mathbf{x})\ Q^{\theta}(\mathbf{x}), \qquad [1]$$

where θ denotes the parameters of the selection model and Z^{θ} ensures normalization of P^{θ}_{post}. Previous work (4, 31, 32) 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. 1*B*). Selection can in general be a highly complex and nonlinear function of the underlying sequence features. Here, we introduce soNNia, a method to infer generic nonlinear selection functions, using DNNs. 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 $\mathcal{L}(\theta) = \langle \log P^{\theta}_{\text{post}} \rangle_{\mathcal{D}}$, where the probability P^{θ}_{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 (*SI Appendix*)

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

where $\langle \cdot \rangle_{\mathcal{G}}$ is the expectation over the ensemble of sequences $\mathcal{G}$ that reflect the baseline. This baseline set is often generated by sampling from a previously inferred generation model P_{gen}, using the IGoR software (3). Note that this expression becomes exact as the number of generated sequences approaches infinity.

In soNNia, we divide the sequence features f into three categories: 1) (V, J) usage; 2) CDR3 length; and 3) 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 3′ ends of the CDR3, equivalent to the left–right encoding of the SONIA model (4) (Fig. 1*D*). Inputs 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; *SI Appendix*, *SI Text* and Figs. S1, S3, and S4 have details on the architecture of the DNN.

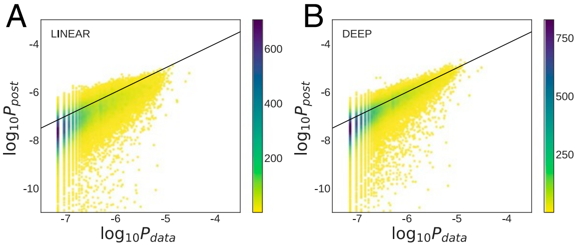
The baseline ensemble $\mathcal{G}$, which we have described as being generated from the P_{gen} model (Fig. 1*C*), can in principle be replaced by any dataset, including empirical ones, at no additional computational cost, for selection inference with soNNia. This is especially useful when the goal is to only compare the selection models associated with different subrepertoires with distinct functions. We will use this feature of soNNia to learn selection coefficients of subsets relative to an empirically constructed generic functional repertoire. In that case, the inferred selection factors Q only reflect differential selection relative to the generic baseline. Importantly, this approach enables us to infer differential selection without having to infer a common underlying generation model P_{gen} for the subrepertoires. After 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 Nonlinear Selection Model Best Describes Functional TCR Repertoire. First, we systematically compare the accuracy of the (nonlinear) soNNia model with linear SONIA (4) (Fig. 1*B*) by inferring selection on TCRβ repertoires from a large cohort of 743 individuals from ref. 33. 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 from all individuals and construct a universal donor. 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 and trained both a SONIA and a soNNia selection model on the training set (*Methods* and Fig. 1*C*). Our inference is highly stable, and the selection models are reproducible when trained on subsets of the training data (*SI Appendix*, *SI Text* and Fig. S2).

To assess the performance of our selection models, we compared their inferred probabilities $P_{\text{post}}(\mathbf{x})$ with the observed frequencies of the receptor sequences $P_{data}(\mathbf{x})$ in the test set (Fig. 2 *A* and *B*). Prediction accuracy can be quantified through the Pearson correlation between the two log frequencies or through their Kullback–Leibler divergence $\mathcal{D}_{\text{KL}}(P_{\text{data}}|P_{\text{post}})$ (*Methods* and Fig. 2*C*). 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 with the linear SONIA model with $\mathcal{D}_{KL} \simeq 1.6$ bits (and Pearson correlation $\rho^2 \simeq 0.48$) (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 with the P_{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



C

	ρ^2	$\mathcal{D}_{KL}$[bits]
P_{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. Scatterplot of observed frequency, P_{data}, vs. predicted probability P_{post} for (*A*) linear SONIA and (*B*) DNN soNNia models trained on the TCRβ repertoires of 743 individuals from ref. 33. The baseline is formed by sampling 10^7 sequences from the P_{gen} model, learned from the nonproductive sequences of the same dataset (*SI Appendix*). Color indicates number of sequences. (*C*) The soNNia model performs significantly better, as quantified by both the Kullback–Leibler divergence $\mathcal{D}_{\text{KL}}$ (*Methods*) and the Pearson correlation coefficient ρ^2, without overfitting (*SI Appendix*, Fig. S2).

inaccessible to the soNNia selection factors are likely due to the sampling of rare features, individual history of pathogenic exposures, or differences in human leukocyte antigen (HLA) complexes among individuals.

Intra- and Interchain Interactions in TCRs and BCRs. TCRs 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, BCRs 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 TCRs, with the largest contributions between V_α, V_β and J_α, V_β (34–38). In B cells, preferences for receptor features within immunoglobulin heavy (IgH) and light (λ or κ) chains have been studied separately (39, 40), but interchain correlations have not been systematically investigated.

We first aimed to quantify dependencies between chains by reanalyzing recently published single-cell datasets: TCR$\alpha\beta$ pairs of unfractionated repertoires from ref. 37 and BCR of naive cells from ref. 41 (*Methods*). 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. 3*A*), IgHλ (Fig. 3*B*), and IgHκ (Fig. 3*C*) repertoires. Mutual information is a nonparametric measure of correlation between pairs of variables (*SI Appendix*).

Both TCRs and BCRs have intra- and interchain correlations of sequence features, with stronger empirical mutual dependencies present within chains (Fig. 3).

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:

$$P_{\text{post}}(\mathbf{x}) = \frac{1}{Z} P^a_{\text{gen}}(\mathbf{x}^a) P^b_{\text{gen}}(\mathbf{x}^b) Q(\mathbf{x}), \qquad [3]$$

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 DNNs that can account for complex correlations between features of the same chain, similar to the single-chain case (*SI Appendix*, Fig. S3). The resulting postselection distributions for both the linear and the deep-indep model factorize, $P_{\text{post}}(\mathbf{x}) = P^a_{\text{post}}(\mathbf{x}^a) P^b_{\text{post}}(\mathbf{x}^b)$, 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 (*SI Appendix*, Fig. S4).

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 information between pairs of features (Fig. 3). The preselection generation model [$Q(\mathbf{x}) = 1$; green bars] explains part but not all of the intrachain feature dependencies, for both T and B cells,

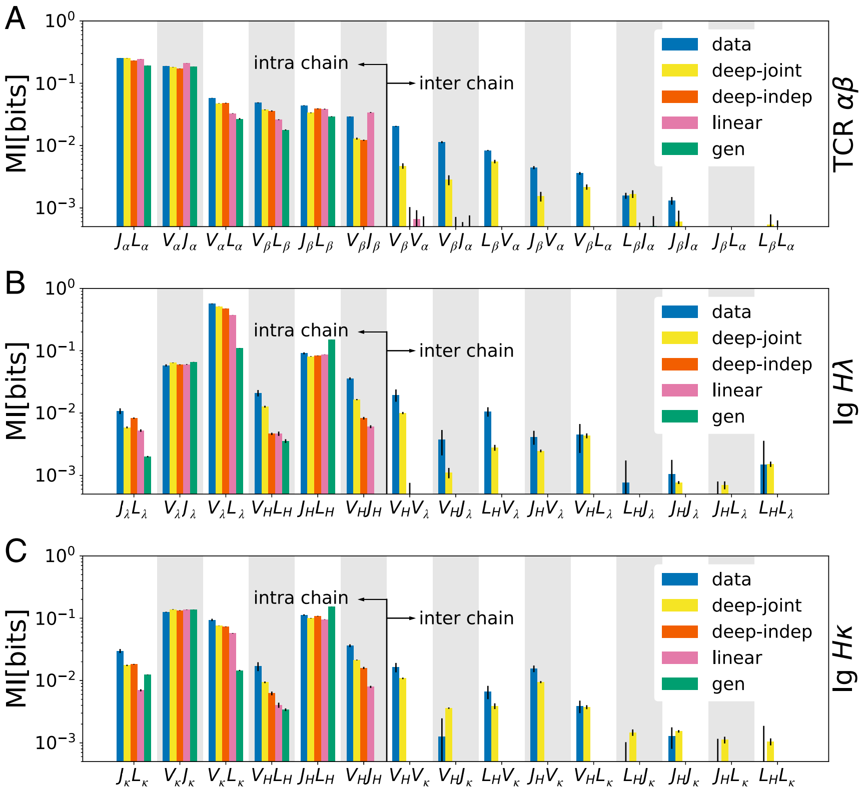


Fig. 3. Inference of selection on intra- and interchain receptor features. Mutual information between pairs of major intra- and interchain features (*V* and *J* gene choice and *L*= CDR3 length for each chain) for (*A*) TCR$\alpha\beta$, (*B*) IgHλ, and (*C*) IgHκ paired chains are shown. Mutual information 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 DNN, is able to recover interchain correlations. Mutual information is corrected for finite-size bias, and error bars are obtained by subsampling (*SI Appendix*). The diversity of the paired-chain B and T cell repertoires and the contributions of different features to this diversity are reported in *SI Appendix*, Table S1.

while the linear (purple), deep-indep (red), and deep-joint (yellow) models explain them very well (Fig. 3). By construction, the generation, linear, and deep-indep models do not allow for interchain correlations. Only the deep-joint model (yellow) is able to recover part of the interchain 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 intrachain and interchain dependencies of feature-forming immune receptors.

The inferred selection on correlated interchain receptor features is consistent with previous analyses in TCRs (34–38) and is likely due to the synergy of the two chains interacting with self-antigens presented during thymic development for TCRs and preperipheral selection (including central tolerance) for BCRs or later when recognizing antigens in the periphery. Notably, the largest interchain dependencies and synergistic selection are associated with the V-gene usages of the two chains (Fig. 3), which encode a significant portion of antigen-engaging regions in both TCRs and BCRs.

Our results show that the process of selection in BCRs is restrictive, in agreement with previous findings (7), significantly increasing interchain feature correlations. 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. However, the selection strengths inferred by our models should not be directly compared with estimates of the percentage of cells passing preperipheral selection: $\sim$ 10% for B cells vs. 3 to 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 pressures in the T cell data than in the purely naive T cells. However, previous work analyzing naive and memory TCRs separately using linear selection models did not report substantial differences between the two subsets (31).

Lastly, to quantify the diversity of immune receptor repertoires, we compared the entropy of unpaired- and paired-chain repertoires in *SI Appendix*, Table S1 (*SI Appendix*, *SI Text*). These entropy measures suggest a repertoire size (i.e., a typical number of amino acid sequences) of about 10^9 receptors for TCRβ (consistent with ref. 4), 10^7 receptors for TCRα, 10^{13} receptors for BCR heavy-chain, and 10^4 receptors for BCR light-chain sequences. The paired-chain entropy measures suggest repertoire sizes of 10^{16} for TCR$\alpha\beta$ and 10^{17} BCR IgHλ and IgHκ receptors, which are compatible with the small correlations observed between heavy and light chains in Fig. 3 and previously reported in refs. 34–38.

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 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. 36 and 42, 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 nine individuals (from ref. 42) sorted into three cell types (CD4$^+$ conventional T cells [Tconvs], CD4$^+$ Tregs, and CD8$^+$ T cells) and harvested from three tissues (pancreatic draining lymph nodes [pLNs], "irrelevant" nonpancreatic draining lymph nodes [iLNs], and spleen).

Training a nonlinear soNNia model (Fig. 1*C*) for each subset leads to overfitting issues due to limited data. To solve this problem, we train the model in two steps. First, we use the unfractionated data from ref. 43 to construct a shared baseline for all repertoire subsets. We then learn independent linear SONIA models for each repertoire subset so that the inferred Q factors only reflect selection relative to the baseline. We approach this problem in two ways. 1) We infer an SONIA model atop an empirical baseline set $\mathcal{G}$ constructed from the unfractionated repertoire, and 2) we use the technique of transfer learning, which consists of learning a shared nonlinear soNNia model for the unfractionated repertoire and then adding an additional linear layer (similar to standard SONIA) for each subrepertoire (*SI Appendix*, *SI Text* and Fig. S5). The subrepertoire selection factors Q inferred by these two approaches are very similar (*SI Appendix*, Fig. S5), but the former method is simpler, and we use it for our main analysis in Fig. 4. For comparison, we also used the generation model P_{gen} (trained earlier for Fig. 2) as a baseline, in which case the selection factors include selection effects that are shared among the subrepertoires. Distributions of selection factors obtained by both approaches are shown in *SI Appendix*, Fig. S6.

To quantify differential selection on subrepertoires, we use Jensen–Shannon divergence D_{JS} between the distributions of receptors P^r_{post} and $P^{r'}_{\text{post}}$ for pairs of subrepertoires $(r,\ r')$ (*Methods*). Clustering of cell types based on Jensen–Shannon divergence shows strong differential selection preferences between the CD4$^+$ and CD8$^+$ receptors, with an average $D_{\text{JS}} \simeq 0.08 \pm 0.01$ bits across respective tissues and subrepertoires (Fig. 4*A*; *SI Appendix*, Fig. S7*A* has similar results where P_{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 of the SONIA model trained atop P_{gen} as a baseline reveals the VJ (*SI Appendix*, Fig. S8) and amino acid usage features (Fig. 4*C*) that are differentially selected in the Tconv CD4$^+$ and CD8$^+$ subsets (in spleen). Linear selection models are organized 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 (*SI Appendix*). 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 (*SI Appendix*, Fig. S9). It should be noted that the baseline models P_{gen} for these subrepertoires, inferred from their unproductive receptors, are similar (*SI Appendix*, Fig. S10) and do not contribute to these differential preferences.

One key difference between CD4$^+$ and CD8$^+$ TCRs amino acid composition is their CDR3 charge preferences. We observe an overrepresentation of positively charged (Lysine, K, and Arginine, R) and suppression of negatively charged (Aspartate, D, and Glutamate, E) amino acids in CD4$^+$ TCRs compared with CD8$^+$ TCRs (Fig. 4*B*), consistent with previous observations (44). These charge preferences arise due to differential selection on the two subtypes (Fig. 4*B*), indicating broad differences between amino acid features of peptide–MHC-I and peptide–MHC-II complexes, which respectively interact with CD8$^+$ and

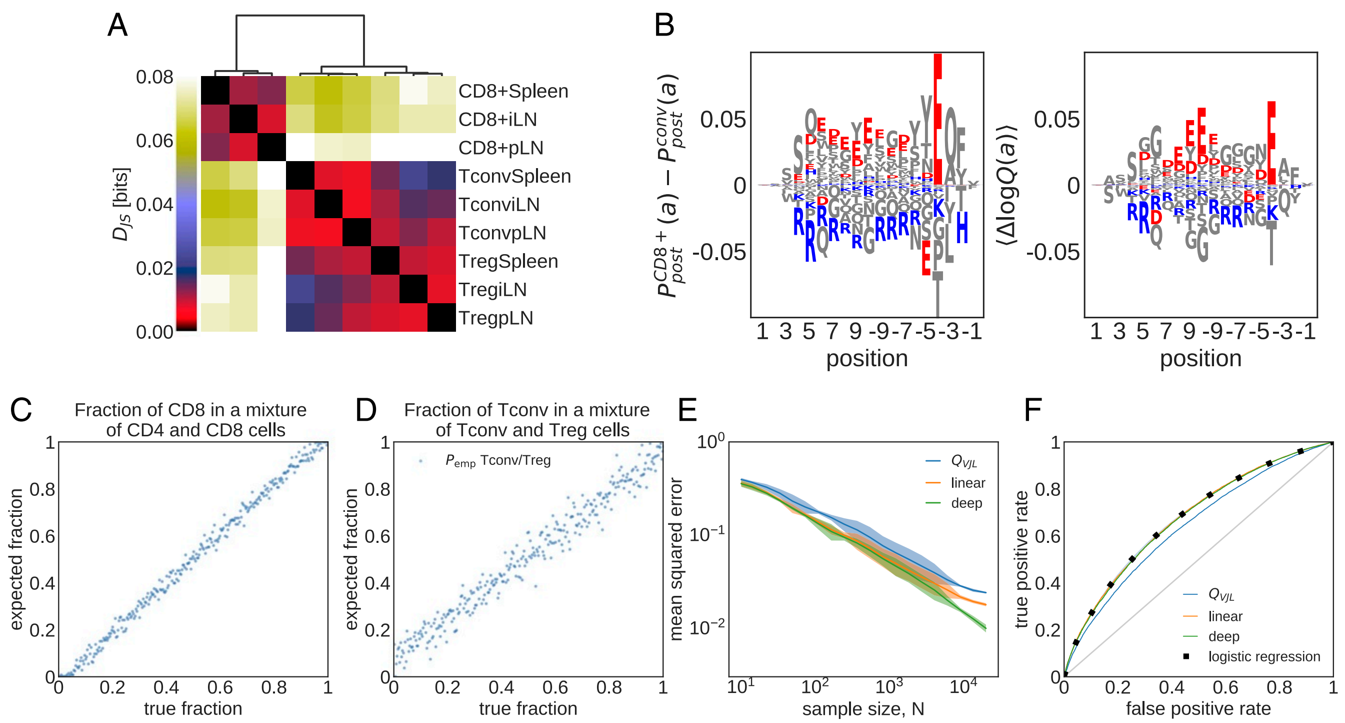


Fig. 4. Cell type- and tissue-specific selection on TCRs. (*A*) Jensen–Shannon divergences (D_{JS}) (Eq. **8**) computed from models trained on different subrepertoires are shown. (*B*) Difference in the marginal probability for amino acid composition along the CDR3, $P_{\text{post}}^{\text{CD8}}(a) - P_{\text{post}}^{\text{CD4}}(a)$, between CD8$^+$ and CD4$^+$ Tconv (*Left*) and the mean difference in the corresponding log-selection factors for amino acid usage $\Delta \log Q = \log Q^{\text{CD8}} - \log Q^{\text{CD4}}$ (*Right*) are shown (the mean is taken over the distribution $(P_{\text{post}}^{\text{CD8}} + P_{\text{post}}^{\text{CD4}})/2$). The negatively charged amino acids (Aspartate, D, and Glutamate, E) and the positively charged amino acids (Lysine, K, and Arginine, R) are indicated in red and blue, respectively. Other amino acids are shown in gray. (*C*) Maximum likelihood inference of the fraction of CD8$^+$ TCRs in mixed repertoires of conventional CD4$^+$ T cells (Tconvs) and CD8$^+$ cells from spleen (Eq. **4**) is shown. Each repertoire comprises 5×10^3 unique TCRs. (*D*) Same as *C* but for a mixture of Tconv and Treg TCRs. (*E*) Mean squared error of the inferred sample fraction from *C* 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, and "deep" is the full soNNia model (Fig. 1*C*). (*F*) 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. **5**. The accuracy of the classifier is compared with 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.

CD4$^+$ TCRs. For example, a statistical survey of peptides presented by different MHC classes shows that MHC-I molecules tend to present more positively charged peptides compared with MHC-II molecules—a bias that is complementary to the charge preferences of the respective TCR subtypes (45).

Decomposing Unsorted Repertoires Using Selection Models. Knowing P_{post}^{r} models specific to subrepertoires enables us to infer the fraction of each class r in unsorted data. Estimating the relative fraction of CD4$^+$ and CD8$^+$ subtypes in a repertoire can be informative for clinical purposes (e.g., as a probe for tumor-infiltrating lymphocytes), where overabundance of CD8$^+$ cells in the sample has been associated with positive prognosis in ovarian cancer (46). Given a repertoire composed of the mixture of two subrepertoires 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 subrepertoire r within the mixture:

$$\begin{aligned} L(f) &= \langle \log(f P_{\text{post}}^{r}(\sigma) + (1-f) P_{\text{post}}^{r'}(\sigma)) \rangle_D \qquad \textbf{[4]} \\ &= \langle \log(f \mathcal{Q}^{r}(\sigma) + (1-f)\mathcal{Q}^{r'}(\sigma)) \rangle_D + \text{const}, \end{aligned}$$

where $\langle \cdot \rangle_D$ is the empirical mean over sequences in the mixture.

Previous work has used differential V- and J-gene usage and CDR3 length to characterize the relative fraction of CD4$^+$ and CD8$^+$ cells in an unfractionated repertoire (47). The log-likelihood function in Eq. **4** provides a principled approach for inferring cell-type composition using selection factors that capture the differential receptor features of each subrepertoire, 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 (42)] and CD8$^+$ [from spleen (42)] receptors with different proportions and show that our selection-based inference can accurately recover the relative fraction of CD8$^+$ in the mix (Fig 4*C*). Our method can also infer the proportion of Treg cells in a mixture of Tconv and Treg CD4$^+$ cells from spleen (Fig. 4*D*), which is a much harder task since these subsets are very similar (Fig. 4*A*). 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. 4*E*).

Our method uses a theoretically grounded maximum likelihood approach, which includes all of 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. 4*E*), analogous to the method used in ref. 47, 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 in Fig. 4*E*). The accuracy of the inference is insensitive to the choice of the baseline model for receptor repertoires: Using the empirical baseline from ref. 43 (Fig. 4*E*) or P_{gen} (*SI Appendix*, Fig. S7*D*) does not substantially change the results.

The method can be extended to the decomposition of three or more subrepertoires. 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 (*SI Appendix*, Fig. S11) in a mixture of size 5×10^3.

Computational Sorting of CD4$^+$ and CD8$^+$ TCRs. Selection models are powerful in characterizing the broad statistical differences between distinct functional subsets of immune receptors, including the CD4$^+$ and CD8$^+$ TCRs (Fig. 4*A*). A more difficult task, which we call computational sorting, is to classify individual receptors into functional classes based on their sequence features. In other words, how accurately can one classify a given receptor as a member of a functional subset (e.g., CD4$^+$ or CD8$^+$ TCRs)?

We use selection models inferred for distinct subrepertoires r and r' to estimate a log-likelihood ratio $R(\mathbf{x})$ for a given receptor $\mathbf{x}$ to belong to either of the subrepertoires,

$$R(\mathbf{x}) = \log \frac{P^{r}_{\text{post}}(\mathbf{x})}{P^{r'}_{\text{post}}(\mathbf{x})} = \log \frac{Q^{r}(\mathbf{x})}{Q^{r'}(\mathbf{x})}. \quad \textbf{[5]}$$

A larger log-likelihood ratio $R(\mathbf{x})$ indicates that the receptor is more likely to be associated with the subrepertoire 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 depend on the threshold value. We evaluate the accuracy of our log-likelihood classifier between sets of CD8$^+$ and Tconv CD4$^+$ receptors harvested from spleen (42). The receiver operating characteristic (ROC) curve in Fig. 4*F* 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_{emp} in Fig. 4*F* and P_{gen} in *SI Appendix*, Fig. S7*E*). Applying this classification method to all of the possible pairs of subrepertoires in Fig. 4*A*, we find that CD4$^+$ vs. CD8$^+$ discrimination generally achieves AUC ≈ 0.7, while discriminating subrepertoires within the CD4$^+$ or CD8$^+$ classes yields much poorer performance (*SI Appendix*, Fig. S12).

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 subrepertoire 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:

$$\mathcal{L}_c(\mathbf{w}, b) = \sum_{i=1}^{N} [y_i \log \hat{y}(\mathbf{x}_i) + (1 - y_i) \log(1 - \hat{y}(\mathbf{x}_i))], \quad \textbf{[6]}$$

where y_i labels each TCR by the subrepertoire (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. **5**) also reduces to a linear form—the only difference being how the linear coefficients are learned. This optimized logistic classifier (Eq. **6**) performs equally well compared with the selection-based classifier, with the same AUC = 0.68 (points in Fig. 4*F*). These AUCs are comparable with those found in ref. 36, 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 (36). For example, the relative importance of different sequence features (i.e., CDR3 length, V/J gene identity, and amino acid composition) for CD4$^+$ vs. CD8$^+$ classification is shown in *SI Appendix*, Fig. S9.

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 an MHC molecule on the T cell surface. Recent growth of data on paired TCRs and their target epitopes (26, 48) has led to the development of machine learning methods for TCR–epitope mapping (25–29). 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. **5**) to address this problem. We determine the target ensemble P^{r}_{post} from the training set of TCRs associated with a given epitope (positive data) and the alternative ensemble $P^{r'}_{\text{post}}$ from a set of generic unfractionated TCRs (negative data). For comparison, we also perform the classification task using the linear logistic regression approach (Eq. **6**) and the state-of-the-art TCRex algorithm (28), which uses a random forest model for classification.

We performed classification for the following CD8$^+$-specific epitopes, presented on HLA-A*02 molecules: 1) the influenza GILGFVFTL epitope (with $N = 3{,}107$ associated TCRs), 2) the cytomegalovirus (CMV) NLVPMVATV epitope ($N = 4{,}812$), and 3) the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) YLQPRTFLL epitope ($N = 315$). The first two epitopes have the most abundant associated TCR sets in VDJdb (26, 48), and the latter is relevant for the ongoing COVID-19 pandemic. For consistency with TCRex (28), we used the pooled data from ref. 43 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 fivefold cross-validation procedure. Due to the scarcity of data, we limit our selection inference to the linear SONIA model (Fig. 1*C*). The ROC curves show comparable performances for the three classification methods on the three epitope-specific TCR sets (Fig. 5 *A*–*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. 5 *D*–*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 it may be trained on smaller datasets. Nonetheless, we see a strong correlation between the selection-based log-likelihood ratio $R(x)$ (Eq. **5**) and the estimator of the logistic classifier $\hat{y}$ (Eq. **6**), shown for positive set (red points) and the negative set (black points) in Fig. 5 *G*–*I* for the three epitopes. This result indicates that the separation hyperplane identified by

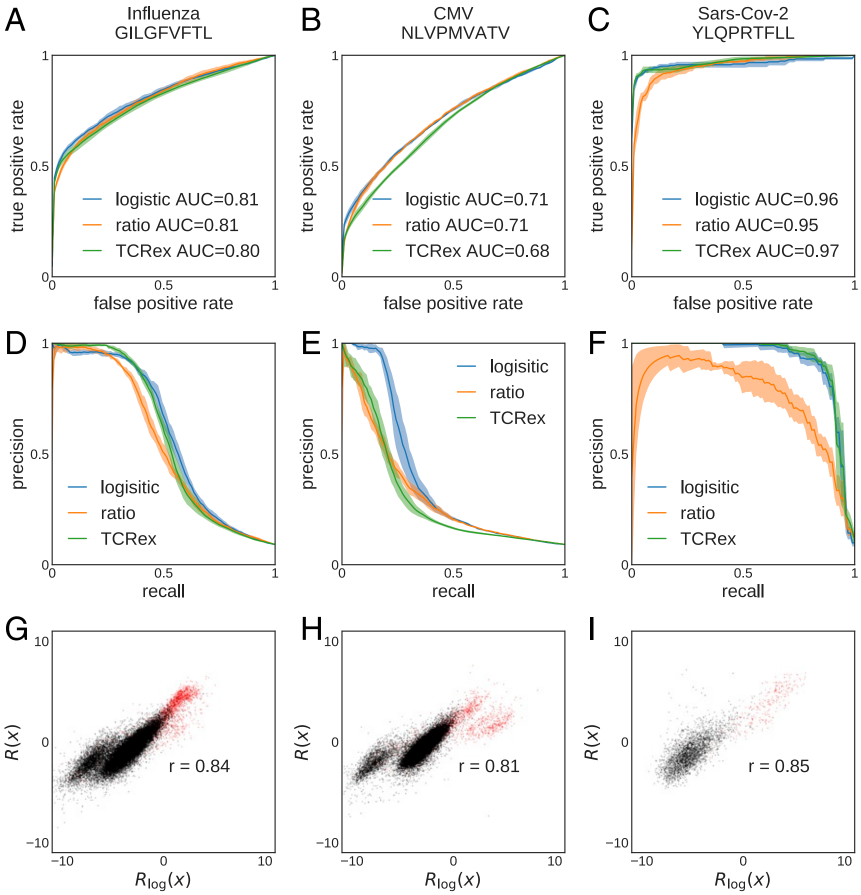


Fig. 5. Selection-based prediction of epitope specificity for TCR. TCRs are classified based on their reactivity to three pathogenic epitopes (columns), using three classification methods: TCRex, log-likelihood ratio (Eq. **5**), and linear logistic regression (Eq. **6**). (*A*–*C*) ROC curves and (*D*–*F*) precision-recall curves for (*A* and *D*) influenza epitope GILGFVFTL ($N = 3{,}107$ TCR), (*B* and *E*) CMV epitope NLVPMVATV ($N = 4{,}812$), and (*C* and *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), and black points are TCRs from bulk sequencing (negative set). r is Pearson's correlation. For all panels, we used pooled data from ref. 43 as the negative set. We used 10 times more negative data than positive data for training. Performance was quantified using fivefold cross-validation.

the logistic classifier aligns well along the effective coordinates of selection that represent sequence features relevant for function in each epitope class.

Discussion

Previous work has developed linear selection models to characterize the distribution of productive TCRs (4). Here, we generalized on these methods by using DNNs 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 (33).

Using this method, we modeled the selective pressure on paired chains of T and BCRs and found that the observed cross-chain correlations, even if limited, could be partially reproduced with our model (Fig. 3). These observed interchain correlations are likely due to the synergy of the two chains interacting with self- and nonself-antigens, which determine the selection pressure that shapes the functional TCR and BCR repertoires.

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$^+$. These differences capture broad differential preferences of CD8$^+$ and CD4$^+$ TCRs, which can arise from their distinct structural features complementary to their different targets (i.e., peptide–HLAI and peptide–HLAII complexes). A next step would be to uncover more fine-grained differential features, associated with the distinct pathogenic history or HLA composition of different individuals.

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. (47). 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. (47) 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 vs. 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 with 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 subrepertoires, we classified cells into CD4$^+$ and CD8$^+$ subclasses with likelihood ratios of selection models and recovered similar results achieved using pure machine learning approaches (36) 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 (28, 49), DeepTCR (50), netTCR (51), ERGO (52), TCRGP (27) and TcellMatch (53). All these methods have explored the question in slightly different ways and made comparisons with each other. However, with the sole exception of TcellMatch (53), none of the above methods compared their performance with a simple linear classifier. TcellMatch (53) does not explicitly compare with 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, with a linear logistic classifier and with 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. 53, where a linear classifier gave comparable results to DNN architectures.

The linear classifier based on likelihood ratios achieves state-of-the-art performance both in discriminating CD4$^+$ from CD8$^+$ cells (Fig. 4*F*) and in predicting epitope specificity (Fig. 5). However, 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 (26, 48) do not contain sufficiently diverse epitopes to train models that would generalize to unseen epitopes (53). A further complication is that multiple TCR specificity motifs may coexist even for a single epitope (29, 54), which cannot be captured by linear models (55). Progress will be made possible by a combination of high-throughput experiments assaying many TCR–epitope pairs (56) 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 processes than linear models. However, DNN methods such as soNNia suffer from the drawback of being data hungry and show their limitations in practical applications where data are scarce. Nonetheless, with the rapid growth of functionally annotated datasets, we expect soNNia to be more readily used for inference of nonlinear selection on immune receptor sequences. Such nonlinearity is expected as it would reflect the ubiquitous epistatic interactions between residues of a receptor protein that encode for a specific function. 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.

Methods

Data Description. In this work, we used different datasets to evaluate selection on TCR and BCR features.

1) To quantify the accuracy of the soNNia model (Fig. 2), we used the TCRβ repertoires from a large cohort of 743 individuals from ref. 33. We pool the unique nucleotide sequences of receptors from all individuals and construct a universal donor totaling 9×10^7 sequences. We randomly split the pooled dataset into a training set and a test set of equal sizes. We then subsampled the training set to 10^7 to reduce the computational cost of inference.
2) To characterize selection on paired-chain receptors (Fig. 3), we analyzed TCR$\alpha\beta$ pairs of unfractionated repertoires from ref. 37 (totaling 5×10^5 receptors) and BCR of naive cells from ref. 41 (totaling 22×10^3 and 28×10^3 receptors for the Hλ and Hκ repertoires, respectively).
3) To characterize differential sequence features of TCRs between cell types in different tissues (Fig. 4), we pooled unique TCRs from nine healthy individuals from ref. 42, sorted into CD4$^+$ Tconvs, CD4$^+$ Tregs, and CD8$^+$ T cells, harvested from three tissues: pLNs (2.3×10^5 Tconvs, 2.9×10^5 Tregs, 2.5×10^5 CD8s), iLNs (2.0×10^5 Tconvs, 9.0×10^4 Tregs, 1.0×10^5 CD8s), and spleen (3.2×10^5 Tconvs, 1.1×10^5 Tregs, 1.1×10^5 CD8s). We used the unfractionated data from ref. 43, composed of 2.2×10^6 receptor sequences, to construct a baseline model for this analysis.

Quantifying Accuracy of Selection Models. To assess the performance of our selection models, we compare their inferred probabilities $P_{\text{post}}(\mathbf{x})$ with the observed frequencies of the receptor sequences $P_{\text{data}}(\mathbf{x})$ in the test set. Prediction accuracy can be quantified through the Pearson correlation between the two log frequencies or the Kullback–Leibler divergence between the data and the distribution predicted by the selection model P_{post}:

$$\mathcal{D}_{\text{KL}}(P_{\text{data}}|P_{\text{post}}) = \left\langle \log_2 \frac{P_{\text{data}}}{P_{\text{post}}} \right\rangle_{P_{\text{data}}} . \qquad [7]$$

A smaller Kullback–Leibler divergence indicates a higher accuracy of the inferred model in predicting the data. In Fig. 2, we estimate the Kullback–Leibler divergence using 10^5 receptors in the test set with multiplicity larger than two.

Comparing Selection on Different Subrepertoires. To characterize differences in subrepertoires due to selection, we evaluate the Jensen–Shannon divergence $D_{\text{JS}}(r, r')$ between the distribution of pairs (r, r') of subrepertoires P^r_{post} and $P^{r'}_{\text{post}}$,

$$D_{\text{JS}}(r, r') = \frac{1}{2}\left\langle \log_2 \frac{2\mathcal{Q}^r}{\mathcal{Q}^r + \mathcal{Q}^{r'}} \right\rangle_r + \frac{1}{2}\left\langle \log_2 \frac{2\mathcal{Q}^{r'}}{\mathcal{Q}^r + \mathcal{Q}^{r'}} \right\rangle_{r'} , \qquad [8]$$

where $\langle \cdot \rangle_r$ denotes averages over P^r_{post} (*SI Appendix* has evaluation details). This divergence is symmetric and only depends on the relative differences of selection factors between functional subrepertoires, and not on the baseline model.

Data Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank Phil Bradley for helpful discussions about differential charge preferences of CD4$^+$ and CD8$^+$ TCRs. This work was supported by Deutsche Forschungsgemeinschaft Grant SFB1310 for Predictability in Evolution (G.I., A.M.W., and A.N.), Max Planck Research Group funding through the Max Planck Society (G.I. and A.N.), the Royalty Research Fund from the University of Washington (A.N.), European Research Council Consolidator Grant 724208 (G.I., A.M.W., and T.M.), and Agence Nationale de la Recherche Grant ANR-19-CE45-0018 (G.I., A.M.W., and T.M.).

1. S. Tonegawa, Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
2. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
3. Q. Marcou, T. Mora, A. M. Walczak, High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
4. Z. Sethna *et al.*, Population variability in the generation and selection of T-cell repertoires. *PLoS Comput. Biol.* **16**, e1008394 (2020).
5. Z. Sethna, Y. Elhanati, C. G. Callan, 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).
6. A. J. Yates, Theories and quantification of thymic selection. *Front. Immunol.* **5**, 13 (2014).
7. D. Nemazee, Mechanisms of central tolerance for B cells. *Nat. Rev. Immunol.* **17**, 281–294 (2017).
8. K. Murphy *et al.*, *Janeway's Immunobiology* (Garland Science, 2008).
9. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
10. K. Wing, S. Sakaguchi, Regulatory T cells exert checks and balances on self tolerance and autoimmunity. *Nat. Immunol.* **11**, 7–13 (2010).
11. X. L. Hou, L. Wang, Y. L. Ding, Q. Xie, H. Y. Diao, Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Gene Immun.* **17**, 153–164 (2016).
12. G. Georgiou, et al., The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–68 (2014).
13. D. A. Bolotin *et al.*, MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
14. J. R. Mcdaniel, B. J. DeKosky, H. Tanno, A. D. Ellington, G. Georgiou, Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* **11**, 429–442 (2016).
15. B. J. DeKosky *et al.*, High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166–9 (2013).
16. M. Turchaninova *et al.*, Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).
17. B. J. Dekosky *et al.*, In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2014).
18. W. Ndifon *et al.*, Chromatin conformation governs T-cell receptor J gene segment usage. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15865–15870 (2012).
19. D. K. Ralph, F. A. Matsen, Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.* **12**, 1–25 (2016).
20. S. Munshaw, T. B. Kepler, SoDA2: A hidden Markov model approach for identification of immunoglobulin rearrangements. *Bioinformatics* **26**, 867–72 (2010).
21. K. Davidsen *et al.*, Deep generative models for T cell receptor protein sequences. *eLife* **8**, e46935 (2019).
22. V. Greiff *et al.*, Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immun.* **199**, 2985–2997 (2017).
23. E. Miho, R. Roškar, V. Greiff, S. T. Reddy, Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
24. G. Isacchini *et al.*, Generative models of T-cell receptor sequences. *Phys. Rev. E* **101**, 062414 (2020).
25. J. Glanville *et al.*, Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
26. M. Shugay *et al.*, VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res..* **46**, D419–D427 (2018).
27. E. Jokinen, J. Huuhtanen, S. Mustjoki, M. Heinonen, H.Lähdesmäki, Determining epitope specificity of T cell receptors with TCRGP. BioRxiv [Preprint] (2019). https://doi.org/10.1101/542332 (Accessed 26 March 2021).
28. S. Gielis *et al.*, Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* **10**, 2820 (2019).
29. P. Dash *et al.*, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
30. A. Murugan, T. Mora, A. M. Walczak, C. G. Callan, Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16161–16166 (2012).
31. Y. Elhanati, A. Murugan, C. G. Callan, T. Mora, A. M. Walczak, Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9875–9880 (2014).
32. Y. Elhanati *et al.*, Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140243 (2015).
33. R. O. Emerson *et al.*, Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
34. K. Grigaityte *et al.*, Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire. BioRxiv [Preprint] (2017). https://doi.org/10.1101/213462 (Accessed 26 March 2021).
35. T. Dupic, Q. Marcou, A. M. Walczak, T. Mora, Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput. Biol.* **15**, 1–19 (2019).
36. 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).
37. H. Tanno *et al.*, Determinants governing T cell receptor $\alpha\beta$-chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 532–540 (2020).
38. 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**, 1–17 (2020).
39. K. Larimore, M. W. McCormick, H. S. Robins, P. D. Greenberg, Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–30 (2012).
40. J. Glanville *et al.*, Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 20216–20221 (2009).
41. B. J. DeKosky *et al.*, Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2636–E2645 (2016).
42. H. R. Seay *et al.*, Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, 1–19 (2016).
43. J. Dean *et al.*, Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Med.* **7**, 123 (2015).
44. H. M. Li *et al.*, TCRβ repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *J. Leukoc. Biol.* **99**, 505–513 (2016).
45. N. Rapin, I. Hoof, O. Lund, M. Nielsen, The MHC motif viewer: A visualization tool for MHC binding motifs. *Curr. Protoc. Im.* **88**, 18.17.1–18.17.13 (2010).
46. E. Sato *et al.*, 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. U.S.A.* **102**, 18538–18543 (2005).
47. R. Emerson *et al.*, Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. *J. Immunol. Methods* **391**, 14–21 (2013).
48. D. V. Bagaev *et al.*, VDJdb in 2019: Database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.* **8**, D1057–D1062 (2020).
49. N. De Neuter *et al.*, On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* **70**, 159–168 (2018).
50. J.-W. Sidhom *et al.*, DeepTCR: A deep learning framework for understanding T-cell receptor sequence signatures within complex T-cell repertoires. BioRxiv [Preprint] (2019). https://doi.org/10.1101/464107 (Accessed 26 March 2021).
51. V. I. Jurtz *et al.*, NetTCR: Sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. BioRxiv [Preprint] (2018). https://doi.org/10.1101/433706 (Accessed 26 March 2021).
52. I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, Y. Louzoun, Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* **11**, 1803 (2020).
53. D. S. Fischer, Y. Wu, B. Schubert, F. J. Theis, Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).
54. A. A. Minervina *et al.*, Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife* **9**, e53704 (2020).
55. B. Bravi, *et al.*, Probing T-cell response by sequence-based probabilistic modeling. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.12.17.423283 (Accessed 26 March 2021).
56. M. Klinger *et al.*, Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PloS One* **10**, e0141561 (2015).