IMMUNOLOGY

Central T cell tolerance from sparse peptide sampling

Hannah V. Meyer[1]*, Sanjoy Dasgupta[2], Amitava Banerjee[1], Yong Lin[1], Rishvanth K. Prabakar[1], Sarah R. Chapin[1], Carl Kingsford[3], Saket Navlakha[1]*

Negative selection in the thymus limits autoimmunity by eliminating T cells that react strongly to self. Individual T cells, however, are only exposed to a small fraction of all self-peptides during their "training" in the thymus, and how tolerance is generalized to the remaining "test" self-peptides across peripheral tissues in the body remains an open question. We show that this can be achieved because the immune system satisfies two conditions necessary for generalization in machine learning settings. Consequently, sparse, random sampling of only 10% of self-peptides in the thymus is sufficient to avoid reactivity to 90% of peripheral self. We support this result and validate predictions from our model with diverse experimental data. Overall, we provide a plausible answer to a long-standing question underlying adaptive immunity, and we highlight how generalization, a fundamental challenge faced by nearly every learning algorithm, is tackled by the immune system.

INTRODUCTION

Negative selection in the thymus (*1*, *2*) is a crucial process for developing a T cell immune response that can eliminate infected or malignant cells in the body while sparing their healthy counterparts. At first glance, however, this process is a numbers game that seems doomed to fail. Developing T cells express receptors (TCRs) on their surface, which are screened for reactivity against self-peptides presented by major histocompatibility complexes (MHCs) on thymic antigen-presenting cells (*3*–*5*). A T cell that is reactive to any sampled peptide-MHC complex (pMHC) is either deleted or redirected into a regulatory T cell lineage as a safeguard from autoimmune responses. Each developing T cell, however, encounters only a small fraction of the millions of possible self-peptides during its short dwell time in the thymus (*6*–*9*). How then do T cells generalize tolerance from the small thymic training sample to the remaining "unseen" self-peptides across the body to avoid autoimmune responses?

In reality, this training process is far from perfect; self-reactive T cells escape negative selection and are well known to exist in the periphery (*10*–*14*). Consequently, there has been intense focus on how peripheral tolerance mechanisms (*15*)—e.g., regulatory T cells (*16*), quorum sensing (*17*, *18*), and anergy (*19*)—help to reduce the chance that these self-reactive T cells drive autoimmunity. However, peripheral tolerance alone is not sufficient; the loss or suppression of negative selection can directly result in autoimmunity. The most evident example is autoimmune polyglandular syndrome type 1 (APS1) (*20*), a multiorgan (Fig. 1A) autoimmune disorder caused by genetic mutations in the autoimmune regulator (AIRE) gene, an important regulator of negative selection (*21*, *22*). This raises the question of how sparse sampling of self-peptides during negative selection can possibly generate a T cell population that avoids gross self-reactivity.

To address this challenge, we developed a theoretical model of negative selection that, unlike prior models (*9*, *17*, *23*–*25*), integrates experimental estimates of (i) approximate peptide abundance levels in the thymus and periphery using whole-tissue-level expression data (*26*, *27*); (ii) the number of peptides a developing T cell "sees" in the thymus using single-cell expression of antigen-presenting cells; and (iii) T cell cross-reactivity sizes using a state-of-the-art method trained on a large database of mutational scans (*28*). Together, we show that central tolerance can be achieved through sparse, random sampling.

RESULTS

The immune system satisfies two conditions necessary for generalization

From a machine learning viewpoint, there are two necessary conditions that allow a learning algorithm to generalize to correctly classify data that it has not been trained on (*29*).

Condition 1

The first condition is that the training and testing data must be correlated; i.e., the two must come from the same or a very similar underlying distribution (*29*, *30*). In our case, the training and testing data correspond to the abundance levels (weights) of self-peptides in the thymus and peripheral tissues in the body, respectively. Ideally, a highly abundant peptide in the periphery should also be highly abundant in the thymus, so that any developing T cell that binds such a peptide would, with relatively high probability, encounter the peptide during training and be deleted (*31*). Mismatches between the two distributions—e.g., peptides with high peripheral expression but low thymic expression—are known to be targets of autoimmunity (*32*–*35*).

To evaluate condition 1, we approximated peptide abundance levels from bulk gene expression data of 29 tissues in the human body [(GTEx) (*27*)], as well as of human medullary thymic epithelial cells (mTECs) (*26*) (Fig. 1, A and B; table S1; and the "Datasets" section in Materials and Methods), which represent the major source of training peptides. To go from gene expression to peptide abundance (see the "Peptide abundance" section in Materials and Methods), we mapped each gene to its corresponding protein using the human reference proteome and then mapped each protein to all of the peptides that comprise it using a sliding window. Each peptide's abundance was taken as the sum of gene expression values of all genes that map to the peptide. We focused on peptide sequences that are nine amino acids long (9-mers), which are the most abundant peptides bound on MHC class I molecules and serve as the sampling space for cytotoxic CD8 T cells. Consequently, we only considered peptides that are visible to these T cells, i.e., peptides that bind to MHC class I

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [2]Computer Science and Engineering Department, University of California San Diego, La Jolla, CA USA. [3]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA USA.
*Corresponding author. Email: hmeyer@cshl.edu (H.V.M.); navlakha@cshl.edu (S.N.)

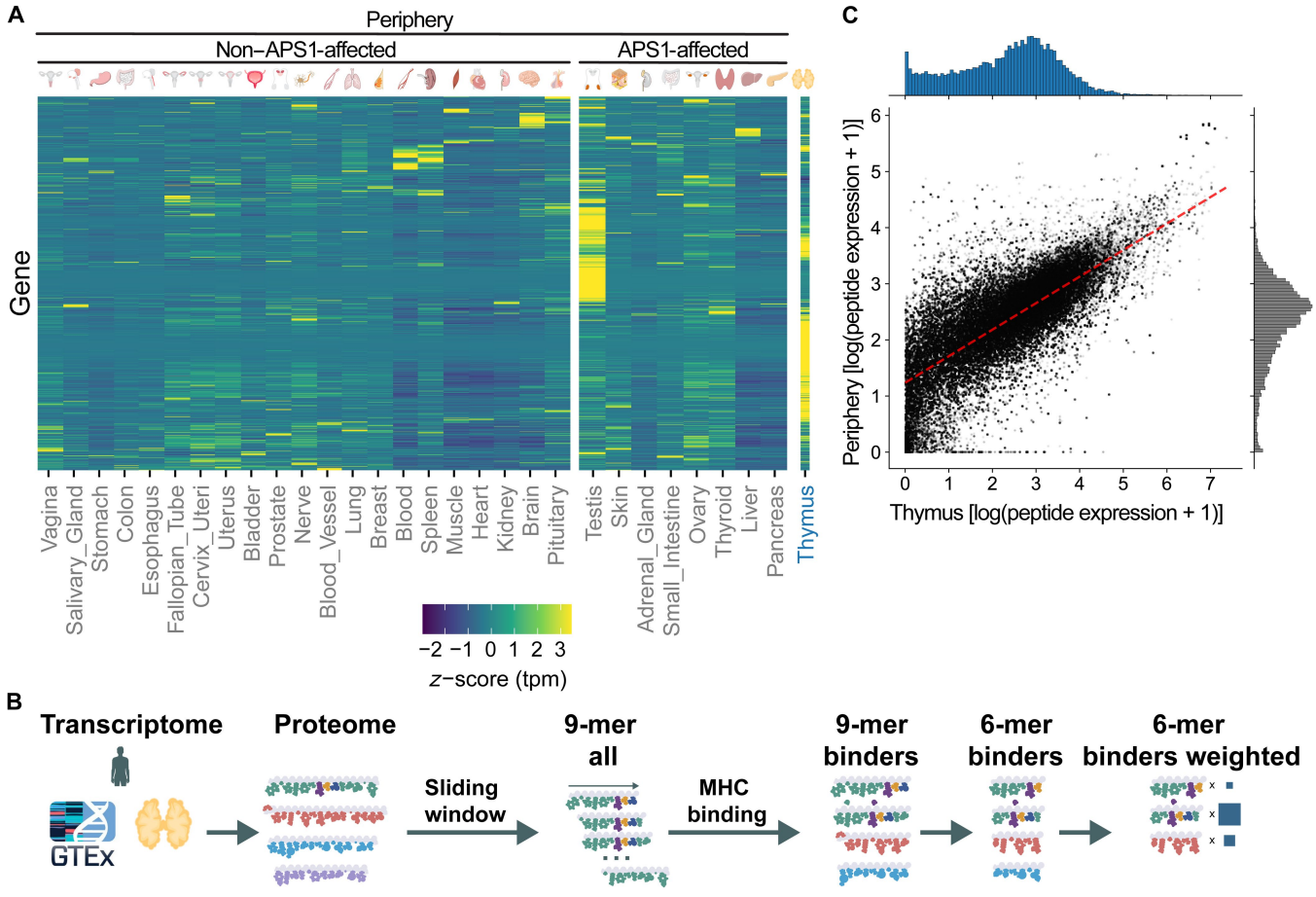

Fig. 1. Peptide abundance levels in the thymus and periphery are highly correlated. (**A**) Gene expression as *z*-score of transcript per million (tpm) from 29 tissues in GTEx, as well as thymus (mTECs) (*26*). For visualization, the *z*-score maximum is capped at the absolute value of the minimum, yielding a symmetric scale. Tissues affected and not affected in APS1 are grouped separately and are used to test whether our model can predict APS1-associated tissues (see the "Identifying vulnerabilities in generalization" section). (**B**) Pipeline to approximate MHC-binding weighted 6-mer peptides starting from gene expression data. Peptide weights are indicated as blue boxes in the last step of the pipeline. (**C**) Correlation of normalized thymus (*x* axis) and peripheral (*y* axis) expression of 6-mer MHC-binding peptides, with Pearson's $r = 0.76$ (red dashed line). For visualization, we plot in log space.

molecules. We predicted binding to the well-studied human leukocyte antigen-A02:01 (HLA-A02:01) allele, which yielded 434,276 9-mers of the total 11,136,576 reference proteome peptides. Last, to home-in on the positions most relevant for T cell binding, we reduced this set of 9-mers to 6-mers by removing positions 1, 2, and 9 from each self-peptide (*24*). The resulting 6-mer peptides comprised positions 3 to 8 of the original 9-mer peptides. Positions 3 to 8 show large sequence diversity, reflective of the large diversity of TCRs that bind them. In contrast, positions 2 and 9 are tightly constrained because they are typically MHC anchor residues—mutating them prevents peptides from being presented to T cells in the first place (*25*, *36*–*38*)—and mutations in position 1 tend to have marginal effects on T cell binding (*24*, *37*, *38*). This reduction from 9-mers to 6-mers also allowed our model to exhaustively explore peptide space (20^6 = 64 million possible peptides), without subsampling within the much larger 9-mer space (20^9 = 512 billion peptides). After this reduction, there were a total of $N = 426,316$ self-peptides.

The abundance levels of self-peptides in the thymus and periphery were highly correlated (Pearson $r = 0.76$, Spearman $\rho = 0.75$; Fig. 1C). Correlations of peptide abundances between the thymus and individual peripheral tissues closely distributed around this average (fig. S1). The distributions were also highly nonuniform with abundance levels varying by over seven orders of magnitude. Moreover, only 0.6% (2494 of the 426,316) peripheral self-peptides were not present (i.e., had zero abundance) in the thymus. This suggests that at the peptide identity level, the thymus contains nearly everything that a T cell could encounter in the periphery (although as we will see, individual T cells only sample a fraction of these peptides during negative selection). We observed similar correlations between abundance levels in the thymus and periphery when considering peptides that bind to other MHC alleles (Spearman ρ ranging from 0.74 to 0.76; table S2), as well as when considering all peptides without MHC restriction ($\rho = 0.77$).

Overall, this means (i) that under the simplest assumption of random, independent and identically distributed (IID) sampling in the thymus (*39*, *40*), T cells will likely interact with highly abundant peptides; and (ii) that these highly abundant peptides are also highly abundant in the periphery and therefore will likely be protected against. That is, the "training" ground of the thymus well approximates the "testing" ground of the periphery, such that interactions

experienced in the thymus closely reflect those that will be experienced in the periphery, satisfying the first condition.

Condition 2

The second condition is that there must be a notion of similarity between data points, such that highly similar points are likely to have similar outcomes. In our case, the data points of interest are pMHC molecules and the outcome of interest is the probability that a TCR binds a given pMHC molecule.

Each TCR can bind many pMHCs (referred to simply as a peptide hereafter), a property called cross-reactivity (*41*–*44*). This property means that a T cell not only binds its index peptide (i.e., the peptide that its receptor most strongly recognizes) but also binds other "similar" peptides at a sufficient strength to trigger an immune response (Fig. 2A). We address a TCR based on its index peptide and define the set of peptides it can bind as its cross-reactivity ball (*45*). That is, we assume there are 64 million possible TCRs (one per index peptide). Consequently, a self-reactive T cell (i.e., a T cell with at least one self-peptide in its cross-reactivity ball) will escape negative selection if it misses sampling all of the self-peptides in its ball during negative selection (*17*). Equivalently, the T cell will be correctly deleted if it samples any self-peptide in its ball; crucially, it does not need to encounter most or even many of them. Thus, cross-reactivity provides a mechanism by which T cells can "generalize" (*24*) information about whether it is self-reactive without having to explicitly interact with its index peptide.

Modeling T cell cross-reactivity requires a notion of similarity between a given TCR's index peptide and a candidate peptide. We used a computational method, called BATMAN (*28*), which recently achieved state-of-the-art performance on a TCR cross-reactivity prediction benchmark containing 25 distinct index peptides (see the "Modeling T cell cross-reactivity using BATMAN" section in Materials and Methods). Given the index peptide of a TCR and a mutated

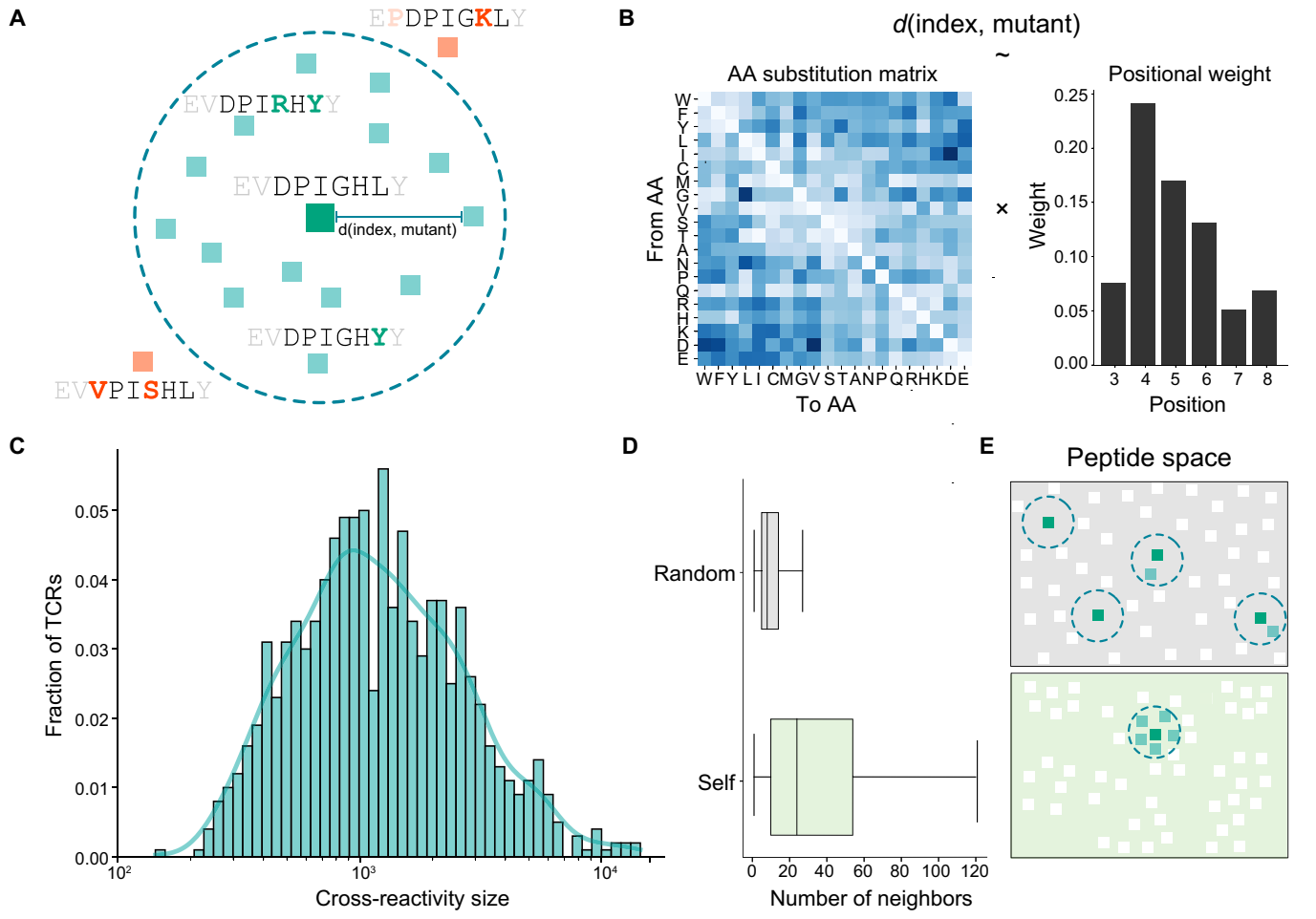

Fig. 2. Cross-reactivity provides a mechanism for T cell generalization. (**A**) Each TCR has a cross-reactivity ball surrounding its index peptide (i.e., the peptide that most strongly binds its receptor), indicated by the large green square. Small mutations to the index peptide could preserve binding (peptides inside the circle) or could eliminate binding (outside the circle). Example peptides shown for the a3a TCR (*77*). (**B**) The BATMAN cross-reactivity model determines which peptides lie in the cross-reactivity ball of a given TCR's index peptide. BATMAN computes the distance between the TCR's index peptide and a given mutant peptide based on an amino acid (AA) substitution matrix and a weight on the position of the mutation in the sequence. (**C**) BATMAN generates a broad spectrum of cross-reactivity sizes (i.e., the number of peptides that lie within the ball centered around a TCR's index peptide). Histogram shown for 1000 random TCRs drawn from the space of 64 million TCRs. (**D**) If self-peptides were randomly distributed in peptide space, then there would be, on average, only 11.9 neighbors in the cross-reactivity ball around a given self-peptide. In contrast, for the actual set of human self-peptides, there are 44.2 (i.e., 3.7 times more) self-neighbors in the cross-reactivity ball around a given human self-peptide. (**E**) The set of human self-peptides are more tightly packed in peptide space compared to random peptides.

peptide, BATMAN outputs a binding distance of the mutant based on two factors (Fig. 2B): a learned amino acid substitution matrix and a learned weight on each position in the peptide sequence, indicating its importance (*38*). These factors allow BATMAN to identify peptides that are similar despite having many mutations separating them. The size of each T cell's cross-reactivity ball is controlled by a radius cutoff parameter. In accord with prior experimental and theoretical work estimating cross-reactivity sizes (see the "Prior work on T cell cross-reactivity" section in Materials and Methods), we set this parameter such that the distribution of sizes has a median of roughly 1000 peptides (from the space of $20^6 = 64$ million possible 6-mer peptides). Despite using a fixed radius, BATMAN can generate a preselection TCR repertoire with cross-reactivity sizes spanning two orders of magnitude, from 100 to 10,000 peptides (Fig. 2C), mimicking the broad spectrum of cross-reactivity sizes found experimentally (*23*, *46*).

The densities of the cross-reactivity balls around each of the N TCRs with an index peptide that lies in self reveals the extent to which binding information can be learned from similar self-peptides. There were a mean of 44 (median, 24) other self-peptides in the cross-reactivity ball around a given T cell's index peptide compared to a mean of 12 (median, 8) if the set of self-peptides was instead randomly distributed in the 20^6 space (Fig. 2D). This implies that self-peptides are three to four times more tightly packed in peptide space than random peptides (Fig. 2D) and, consequently, that cross-reactivity provides a T cell many more "outs" by which it can learn if it is self-reactive. From a machine learning perspective, the relative compactness of the self-distribution (Fig. 2E) means that there is structure in the data that makes generalization more likely to succeed.

In summary, the similarity in the training and testing sets (i.e., the abundance levels of peptides in the thymus and the periphery), along with the ability to generalize binding information across a few dozen self-peptides via cross-reactivity, together provide two key ingredients for immune generalization.

A theoretical model of immune generalization

How much potential peripheral damage (self-reactivity) would T cells cause if negative selection did not exist? How much of this damage is reduced by negative selection as a function of the number of peptides an individual T cell sees during training? Here, we develop a theoretical framework to analytically quantify this ratio that accounts for peptide weights (abundances) and T cell cross-reactivity.

Let $\mathcal{U}$ be the universe of all 20^6 peptides of length 6, and let $\mathcal{S} \subset \mathcal{U}$ be the subset of human MHC-binding self-peptides; in our case, $|\mathcal{S}| = 426,316$. During development, each T cell generates a random receptor (*47*) whose index peptide lies in $\mathcal{U}$. The goal of negative selection is to eliminate any T cell that binds strongly to a peptide in $\mathcal{S}$.

Each peptide $s \in \mathcal{S}$ has two weights, $T(s)$ and $P(s)$, corresponding to the abundance of s in the thymus and periphery, respectively. These abundances are normalized to form a probability distribution, i.e., $\sum_{s\in\mathcal{S}} T(s) = 1$, and similarly for $P(s)$. Consequently, the probability that a T cell sees peptide s in a given random sampling step is $T(s)$.

Each TCR, which is addressed by its 6-mer index peptide t, has a cross-reactivity ball containing all similar peptides that can bind the TCR

$$\mathcal{B}(t) = \{ s \in \mathcal{S} : d(s,t) < r \}$$

where d is the BATMAN distance function with radius r.

During negative selection, each T cell t samples a subset of self-peptides, $\mathcal{S}' \subset \mathcal{S}$. If the T cell samples any self-peptide that lies in its cross-reactivity ball [i.e, if $\mathcal{S}'$ intersects $\mathcal{B}(t)$], then the T cell will be deleted. Deletion can only occur for the subset $\mathcal{X} \subset \mathcal{U}$ of T cells that are self-reactive, that is, for TCRs t for which $\mathcal{B}(t)$ contains at least one self-peptide. For our data, 52.8 million of the $20^6 = 64$ million total preselection TCRs are self-reactive.

With these definitions, we can compute the probability that a random T cell will cause damage without versus with negative selection. Without negative selection, each T cell t survives trivially with probability 1: $\Pr(t \text{ survives}) = 1$. The damage caused by a random T cell is related to the total peripheral weight of all self-peptides that the random T cell can bind

$$\begin{aligned} &= \frac{1}{\sum_{t\in\mathcal{X}} \Pr(t \text{ survives})} \sum\nolimits_{t\in\mathcal{X}} \Pr(t \text{ survives}) \times P[\mathcal{B}(t)] \\ &= \frac{1}{|\mathcal{X}|} \sum\nolimits_{t\in\mathcal{X}} P[\mathcal{B}(t)] \end{aligned} \tag{1}$$

where $P[\mathcal{B}(t)] = \sum_{s\in\mathcal{B}(t)} P(s)$ and similarly for $T[\mathcal{B}(t)]$.

With negative selection, each T cell survives only if it does not interact with any self-peptide in its cross-reactivity ball after k random samples in the thymus with replacement (*17*, *48*)

$$\Pr_{\text{NS}}(t \text{ survives}) = (1 - T[\mathcal{B}(t)])^k \tag{2}$$

Critically, the probability of survival depends on the thymus weights (T), not the peripheral weights (P). However, damage incurred is calculated using the peripheral weights, since damage occurs in the periphery, hence the importance that the two weights are correlated.

Putting it together, the probability that a random T cell that survives negative selection will cause damage is

$$\begin{aligned} &= \frac{1}{\sum_{t\in\mathcal{X}} \Pr_{\text{NS}}(t \text{ survives})} \sum\nolimits_{t\in\mathcal{X}} \Pr_{\text{NS}}(t \text{ survives}) \times P[\mathcal{B}(t)] \\ &= \frac{1}{\sum_{t\in\mathcal{X}} (1 - T[\mathcal{B}(t)])^k} \sum\nolimits_{t\in\mathcal{X}} (1 - T[\mathcal{B}(t)])^k \times P[\mathcal{B}(t)] \end{aligned} \tag{3}$$

The second factor includes two terms: the probability that a random T cell survives negative selection and the amount of damage that T cell would cause if it survived. Ideally, these two values are anticorrelated (i.e., high survival → low damage). The first factor acts as a normalizer of the survival probabilities over all T cells, generating a proper probability distribution. The ratio of Eq. 3 to Eq. 1 is the total amount of damage incurred with versus without negative selection, which is hopefully close to 0. We report 1 − Eq. 3/Eq. 1, i.e., the fraction by which negative selection improves protection, which is ideally close to 1.

How many self-peptides does a developing T cell "see" in the thymus?

A critical missing parameter is k, the number of self-peptides sampled by each T cell during negative selection. As k increases, the growth rate of the number of unique peptides seen decreases since the same peptide can be seen multiple times (i.e., samples are taken with replacement). This growth rate is further reduced if samples are taken with probability proportional to peptide abundances (Fig. 3A).

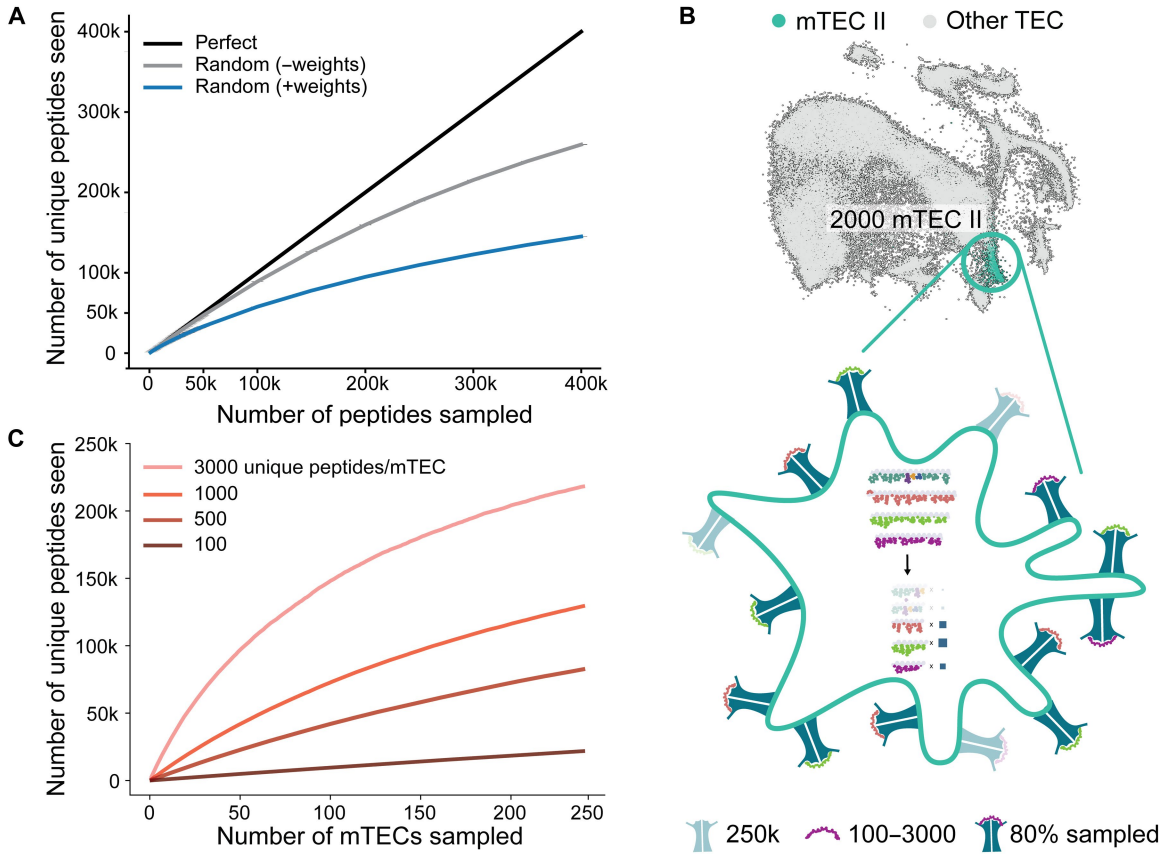

Fig. 3. Estimating the number of peptides an individual T cell sees in the thymus. (**A**) Relationship between the number of peptides sampled by a T cell and the total number of unique peptides seen. "Perfect" sampling indicates that each sample is unique. "Random" sampling with replacement with uniform weights ("–weights") or with peptide abundances ("+weights") both show diminishing returns. For example, after 100,000 samples of peptides (with probability proportional to abundance levels), only 58,000 unique peptides are seen. (**B**) Pipeline to generate MHC binding weighted 6-mer peptides for mTECs with the largest gene expression diversity (mTECII). Peptide abundances were determined starting from single-cell gene expression data of mTECII (low-dimensional embedding of single-cell data at top of panel) and processed as in Fig. 1B (steps summarized inside the highlighted cell in the bottom panel). Peptides for MHC loading were selected on the basis of expression weights (blue boxes) and the number of unique pMHCs, which we varied from 100 to 3000. During simulated scanning of an mTEC, 80% of pMHC molecules are sampled by a T cell (indicated by dark teal). (**C**) Relationship between the number of mTECs sampled by a T cell and the total number of unique peptides seen. For example, after sampling 240 mTECs with 1000 unique peptides per mTEC, the T cell sees 130,000 unique peptides.

To provide bounds on k, we estimated two quantities: the set of self-peptides presented across different antigen-presenting cells (mTECs) and the number of mTEC interactions by each T cell. For the first quantity, it is difficult to experimentally determine the identities of the self-peptides present across the pMHC molecules on the surface of each mTEC. Instead, we estimated this distribution using cell line–derived estimates for MHC abundance (*49*) and peptide presentation (*50*–*52*), as well as single-cell gene expression from 2000 mTECs of the human thymus (Fig. 3B and the "Datasets" section in Materials and Methods). On the basis of these, we derived that each mTEC displays between 100 to 3000 unique peptides distributed over roughly 250,000 pMHC slots on its surface. For each mTEC, we mapped gene expression to peptide abundances; then, we sampled 100 to 3000 unique peptides with probability proportional to its abundance and distributed these peptides proportionally across the 250,000 MHC slots. When interacting with an mTEC, we estimate that each T cell scans up to 80% of its slots (*53*), and the union of the peptides across these slots are counted as seen. For the second quantity, we multiplied the average number of mTEC interactions per T cell [5/hour (*8*)] with the average dwell time in the medulla, focusing on the time with highest susceptibility to apoptosis by negative selection (2 days) (*54*, *55*). Consequently, each T cell interacts with roughly 240 mTECs. Full details of these derivations are in Materials and Methods (see the "Peptide sampling statistics in the thymus" section).

Simulating these 240 mTEC interactions with a range of 100 to 3000 unique peptides per mTEC, we estimate that each T cell encounters between 20,000 and 200,000 unique self-peptides, of the $N \approx 426,000$ total unique self-peptides, during its 2-day dwell in the thymus (Fig. 3C). Thus, T cells are challenged to generalize tolerance to a testing set in the periphery that is at least as large [i.e., $(426{,}000-200{,}000)/200{,}000 \approx 1$] or up to 20 times larger [i.e., $(426{,}000-20{,}000)/20{,}000 \approx 20$] than its training set in the thymus.

Sparse thymic sampling is sufficient to protect most of peripheral self

To summarize our model, for each self-reactive TCR $t \in \mathcal{X}$, we compute the probability that it samples any self-peptide in its cross-reactivity ball $\mathcal{B}(t)$ based on the total weight of the those peptides

$T[\mathcal{B}(t)]$ in the thymus (Fig. 4A). We then use the total weight of those same peptides $P[\mathcal{B}(t)]$ in the periphery to compute the damage-causing probability (self-reactivity) of the TCR with negative selection (Eq. 3) versus without negative selection (Eq. 1); one minus their ratio tells us how much negative selection reduces damage. Following our estimates of T cell sampling rates, we varied k such that the number of unique peptides seen ranged from 20,000 (i.e., 5% of the total) to 100,000 (i.e., 23%) (Fig. 4B, x axis).

Notably, random, sparse sampling by T cells of 5% of the total number of unique self-peptides in the thymus was sufficient to reduce self-reactivity in the periphery by >80% (Fig. 4B). Increasing sampling to 23% of unique self-peptides reduced peripheral self-reactivity by over 95%. This range, derived under IID sampling assumptions, closely matches experimental estimates of the burden peripheral tolerance faces due to imperfect negative selection (5 to 20%) (*10*, *13*).

This level of protection required both generalization conditions outlined above. Without correlated peptide weights in the thymus and periphery (i.e., by setting all thymus weights to be the same), the benefit of negative selection reduces by 5 to 20% (Fig. 4B, "No weights"), and this benefit is further reduced to 10 to 30% if thymus peptide weights are randomly shuffled. Without cross-reactivity, a self-reactive T cell is only deleted if it samples its index peptide, which occurs with near-zero probability (Fig. 4B, "No xreact"). To quantify the benefit of cross-reactivity toward generalization when self-peptides lie compactly versus randomly distributed in peptide space (Fig. 2E), we applied our model to a random set of N peptides (drawn from the 20^6 space); this offered 15 to 30% less protection than when applied to the actual set of human self-peptides.

Cross-reactivity was also critical for protecting self-peptides with low weight (abundance). The average expression of peptides in the cross-reactivity ball of a high-weight versus a low-weight self-peptide was very similar; i.e., there was no correlation between the weight of a self-peptide and the average weight of other peptides in its cross-reactivity ball (Pearson $r = 0.09$; fig. S2). This means that cross-reactivity enables even TCRs with a low-weight self-index peptide to be deleted because of having high-weight neighbors.

Overall, these results suggest that negative selection best generalizes to unseen peptides using a combination of peptide weights and cross-reactivity. Generalization is further enabled by the relatively compact structure of self-peptides in space. Moreover, generalization is not unique to the selection on HLA-A02:01–restricted peptides but is robust across three common human HLA molecules and their haplotype (see the "MHC and haplotype selection" section in Materials and Methods, fig. S3, and table S2).

Model predictions and validation

1) How many mTECs does a T cell need to sample such that its survival probability is invariant to the exact mTECs sampled? Using the single-cell expression data described above, we grouped mTECs into m random, nonoverlapping sets, each consisting of $2000/m$ mTECs. Each of these sets represents the mTECs sampled by a single T cell during negative selection. For each set, we computed an expression vector corresponding to the sum of expression values for all presented peptides across the $2000/m$ mTECs in the set. Using these peptide expression values, we computed the survival probabilities (Eq. 2) for each of the 52.8 million self-reactive T cells. Last, for each pair of sets, we computed the Spearman correlation between the survival probabilities. This analysis includes two parameters—the number of mTECs sampled in each set (which we varied from 20 to 1000) and the number of unique peptides per mTEC (ranging from 100 to 3000) (see the "Peptide sampling statistics in the thymus" section in Materials and Methods). We found that when each set contains 200 to 250 mTECs, the correlation in the TCR survival probabilities ranged from 84 to 95% for the two largest estimates for unique peptides per mTEC (Fig. 5A). Doubling to 500 mTECs seen per T cell increases the correlation range to 93 to 97%, suggesting that doubling development time provides only marginal benefits. This estimate for the number of

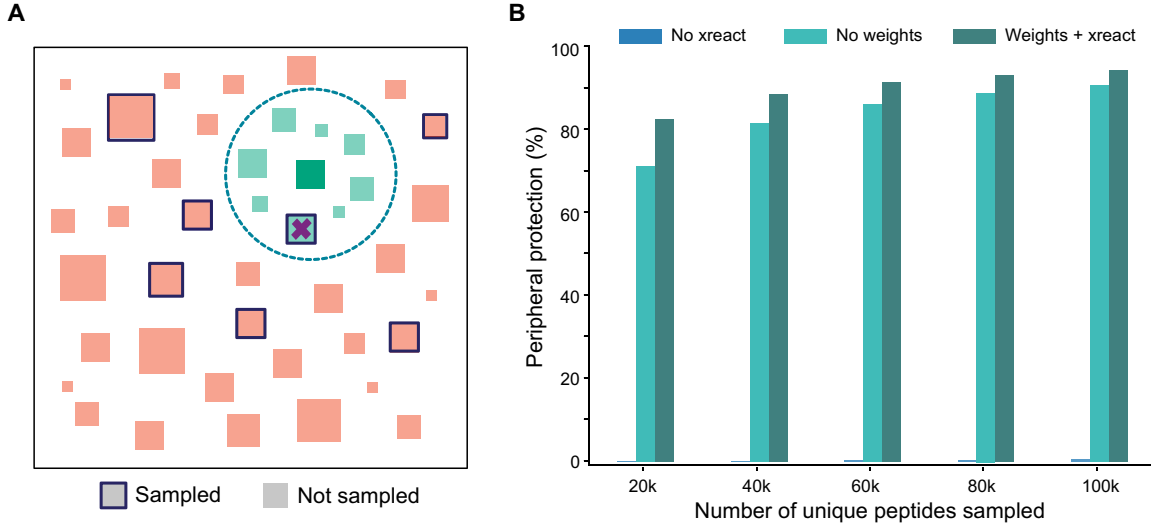

Fig. 4. Sparse, random sampling is sufficient to protect most of peripheral self. (**A**) In our theoretical model, individual T cells sample self-peptides (outlined squares) according to their abundance (size of square) in the thymus. If the T cell encounters any peptide within its cross-reactivity (xreact) ball, shown in green squares, then the T cell is deleted (**X**); a T cell survives if it only samples peptides outside its cross-reactivity ball (orange squares). (**B**) Relationship between the number of unique peptides each T cell sees and peripheral protection as a result of negative selection. For example, with 60,000 unique peptides sampled by each T cell, peripheral protection for the full model ("Weights + xreact") is 91.6%.

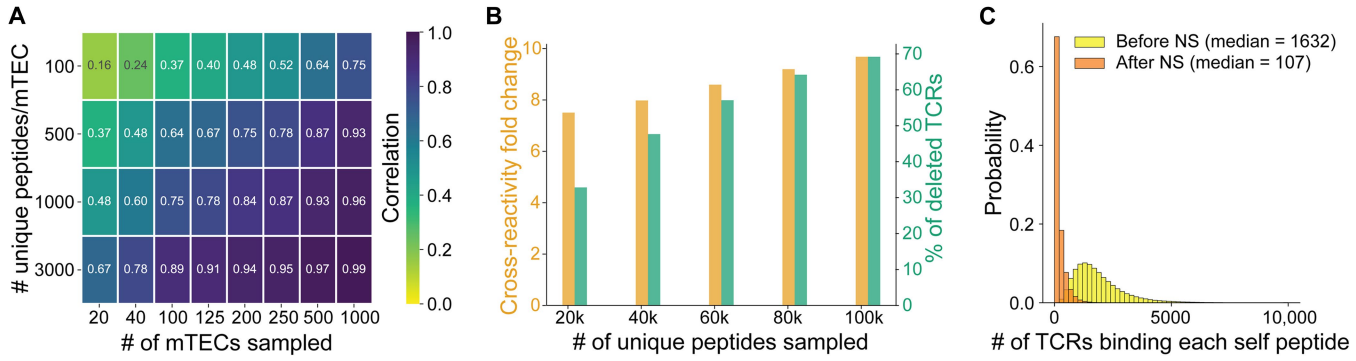

Fig. 5. Validating model predictions. (**A**) Heatmap showing the correlation between the survival probabilities of TCRs as a function of the number of mTECs sampled and the number of unique peptides on the surface of the mTEC. For example, with 250 mTECs sampled and 1000 peptides per mTEC, survival probabilities across different mTEC sets visited have a 0.87 Spearman correlation. (**B**) Relationship between the number of peptides sampled by a T cell and the cross-reactivity fold change of low versus high survival TCRs (left *y* axis) or the percentage of deleted TCRs (right *y* axis). For example, with 60,000 unique peptides sampled, the cross-reactivity size of TCRs that are likely to survive was 8.7× lower than TCRs that are unlikely to survive, and the percentage of deleted TCRs was 58.1%. (**C**) After negative selection (NS) (with 60,000 unique peptides seen), the median number of TCRs that recognize each self-peptide is reduced 15-fold compared to the preselection repertoire.

mTECs sampled by a T cell aligns very closely with experimentally derived estimates described above [5 mTECs/hour (*8*) × 48 hours (*54*, *55*) = 240 mTECs].

2) How well do model estimates on the impact of negative selection on T cell repertoires align with those approximated experimentally? First, we asked what percentage of the set $\mathcal{X}$ of 52.8 million self-reactive T cells are eliminated by negative selection. In our model, the percentage of deleted T cells can be calculated as

$$\frac{\sum_{t \in \mathcal{X}} 1 - \Pr_{NS}(t \text{ survives})}{|\mathcal{X}|}$$

This quantity ranged from 35 to 70% as a function of the number of unique peptides sampled (Fig. 5B, right *y* axis), which is close to the range reported by theoretical and experimental studies (37 to 75%) (*48*, *56*). Second, we asked how negative selection affects the range of cross-reactivity sizes of TCRs (*57*, *58*). We calculated the change in average cross-reactivity toward self-peptides as $|\overline{\mathcal{B}(s)}| / |\overline{\mathcal{B}(t)}|$, where s ranges over all low survival probability TCRs [$\Pr_{NS}(s \text{ survives}) < 0.5$] and t ranges over all high survival probability TCRs [$\Pr_{NS}(t \text{ survives}) \geq 0.5$]. We found that cross-reactivity to self is narrowed by 7- to 10-fold comparing low versus high likelihood of survival TCRs (Fig. 5B, left *y* axis). In contrast, cross-reactivity of these TCRs to all peptides in $\mathcal{U}$ (as opposed to only self-peptides) changes by only twofold (fig. S4A), suggesting that the narrowing of cross-reactivity is more targeted toward self-reactive TCRs than to nonself-reactive TCRs. Although this distinction between self–cross-reactivity and overall cross-reactivity has not to our knowledge been made, these predicted ranges bookend the roughly fivefold reduction in cross-reactivity after negative selection previously estimated (*23*, *57*, *59*).

3) Is our model consistent with alternative mechanisms proposed to overcome the challenge of generalization? One prevailing theory of peripheral tolerance is quorum sensing (*17*). The idea is that a minimum number of TCRs must recognize a self-peptide in the periphery to trigger an immune response. As a result, the inevitable leakage by negative selection can be controlled by thresholding responses in the periphery. We found that, before negative selection, each self-peptide could bind to a median of 1632 TCRs (Fig. 5C). After negative selection, this number was reduced to a median of 107 TCRs—a 15-fold reduction. In contrast, random peptides bind to essentially the same number of TCRs before versus after negative selection (fig. S4B), suggesting that quorum sensing is particularly effective toward tolerating self-peptides. Thus, quorum sensing in the periphery complements our model and can further reduce the error of generalization and the possibility of autoimmunity.

In summary, our model can recapitulate several important statistics of negative selection and is compatible with a prevailing theory of peripheral tolerance.

Identifying vulnerabilities in generalization

We asked whether our generalization model can identify weaknesses in negative selection that leave certain tissues vulnerable and whether these tissues are implicated in common autoimmune diseases. We first considered all high (>98%) survival probability TCRs after negative selection and the associated set of self-peptides these TCRs bind. We computed the sum of these peptide abundances in each of the 29 GTEx tissues. We found that six tissues (liver, pancreas, pituitary, salivary, stomach, and testis) had between 1.9- and 8.9-fold higher abundance compared to other tissues (Benjamini-Hochberg adjusted $P < 0.05$, with empirical $P < 0.01$; table S3), and these tissues are implicated in 15 autoimmune diseases (see the "Immune disease statistics" in Materials and Methods and table S4). This indicates an inherent tolerance vulnerability reflected in tissues implicated in autoimmune diseases.

We next asked whether our model could recapitulate the disease phenotype of APS1, characterized by mucocutaneous candidiasis, hypoparathyroidism, and adrenal insufficiency. In addition, other tissues implicated in this disease include the liver, skin, and pancreas (all APS1 tissues present in GTEx are indicated in Fig. 1A) (*20*). APS1 is caused by genetic mutations in the AIRE gene (*21*, *22*), a crucial transcriptional regulator in mTECs. Upon loss of AIRE, as tested in mouse models (*3*, *60*), mTECs lose expression of more than 3000 genes, and mice suffer from subsequent APS1-like autoimmunity as a result of impaired thymic selection.

To test whether our model could predict autoreactivity to tissues afflicted in APS1 as a result of losing AIRE-related peptides, we estimated

the effect of AIRE deletion on thymic expression using human orthologs of murine genes known to be controlled by Aire (*3*) (see the "Aire deficiency model" in Materials and Methods). Removing these 3361 genes from human thymus gene expression resulted in 363,344 peptides, i.e., a reduction of 15% compared to the fully intact thymus. We then applied our generalization model using estimated peptides abundances from this impaired thymus; peripheral abundances remain unchanged.

Analogous to our baseline vulnerability analyses above, we considered all high (>98%) survival probability TCRs in the AIRE-deleted thymus. We found that their associated self-peptides were strongly enriched in APS1 tissues ($P = 1.35 \times 10^{-21}$, with a 1.6-fold higher abundance; Table 1). The observed level of enrichment was significantly higher compared to when choosing the same number of random TCRs (empirical $P < 0.01$). We also considered all TCRs whose survival probability was 50% higher in the AIRE-deleted thymus compared to the fully intact thymus. The associated peptides of these TCRs were again significantly enriched in APS1 tissues ($P = 3.89 \times 10^{-4}$; Table 1), and this level of enrichment was roughly two orders of magnitude higher compared to when choosing random TCRs (empirical $P < 0.01$).

To extend these findings from the tissue level to the protein level, we asked whether the peptides identified in our AIRE-deleted thymus map to target proteins identified in a proteome-wide survey of the autoimmune target repertoire in APS1 (*61*). Of 21 high-confidence targets in the screen, 19 are predicted targets of TCRs with 50% increased survival probability in the AIRE-deleted thymus (empirical $P = 0.002$ and $P = 0.03$ compared to random TCRs and random targets, respectively; see the "Autoimmune target proteins in APS1" section in Materials and Methods).

Overall, these results suggest that (i) TCRs that likely survive negative selection preferentially target peptides in six tissues linked to 15 autoimmune disorders and (ii) TCRs that likely survive after thymic deletion of AIRE preferentially target APS1 tissues and associated proteins, i.e., APS1 tissues are relatively less protected from autoimmunity compared to non-APS1 tissues. The latter result further highlights the importance of correlated peptide abundances in the thymus and periphery toward accurate generalization.

DISCUSSION

In biology, the generalization problem is most commonly studied in the brain, where learning a concept from a few examples (e.g., recognizing someone's face) is a regular challenge. Here, we explored how generalization is also faced during the development of T cells. There are two conditions necessary for any learning system, be it biological or otherwise, to properly generalize. We showed that these two conditions—namely, that the training and testing data are correlated and that similar data points have similar outcomes—are satisfied by the immune system. The former relates to peptide abundance levels in the thymus (training) and periphery (testing), respectively, which, to our knowledge, have not been considered in previous models of negative selection and which we showed are highly correlated. The latter relates to the fact that an individual T cell can react to many similar peptides (called cross-reactivity), which serves as a mechanism by which a T cell can learn if it is self-reactive without having to see every self-peptide. Together, we showed that sparse, random sampling of self-peptides in the thymus is sufficient to avoid reactivity to most of peripheral self and that this observations holds across multiple HLA alleles. Last, we showed that our generalization model can predict vulnerabilities inherent to central tolerance, reflected in enrichment of autoimmune target tissues and proteins under both baseline and AIRE deletion conditions.

While we focused on overcoming the challenge of sparse peptide sampling during thymic selection, there may also be some benefits to this strategy. For example, sparse sampling allows T cells to develop faster than complete sampling of all self-peptides, which takes exponentially longer since the rate of seeing new peptides diminishes over time. Second, complete sampling would produce a perfectly self-tolerant T cell repertoire but with a deterministic set of "holes" outside of self-peptide space, which pathogens could evolve to exploit (*10*, *13*). In contrast, sparse sampling means that self-tolerance is compromised for smaller, stochastic holes in nonself-space. Third, low levels of self-reactive T cells in the periphery may be beneficial for diverse processes, such as wound healing, tissue homeostasis, and T regulatory cell sustenance [as reviewed by Richards *et al.* (*13*)]. Thus, sparse sampling coupled with peripheral tolerance could be an effective strategy to trade-off generalization among speed, discrimination, and other physiological functions.

Our work opens the door to many future questions. First, while our assumption of random, IID samples in the thymus makes our model simple and analytically tractable, T cell encounters may be more structured and encompass a wide spectrum of antigen-presenting cells in the thymus. We focused on the most transcriptionally diverse subset of mTECs; however, T cells also encounter other antigen-presenting cells, including mTEC subsets with different transcriptional activities (*62*, *63*), as well as B cells and dendritic cells, which can present both intrinsic

Table 1. Generalization model under thymic AIRE deletion predicts peptide enrichment in APS1 tissues. Mean peptide abundance in APS1 versus non-APS1 tissues and their associated *P* values (one-sided [>] paired). "N_{TCRs}" and "$N_{peptides}$" specify the number of TCRs and their associated target peptides in the "Model." For comparison, the first row shows all self-peptides stratified into APS1-affected tissues and non-APS1–affected tissues, with no significant difference in mean abundance.

Model	N_{TCRs}	$N_{peptides}$	*P* value	Mean abundance	
				APS1	Non-APS1
All peptides	–	426,316	0.83	49.17	49.42
>98% survival ΔAIRE model	2,803,613	82,565	1.35×10^{-21}	18.23	11.46
>50% increased survival ΔAIRE model	118,464	61,515	3.89×10^{-4}	20.39	16.99

antigens and antigens derived by transfer from mTECs (*64*–*66*). Single-cell expression data from these antigen-presenting cells (*62*, *67*) may enable a more comprehensive model of thymus encounters, although quantitative experimental measurements of antigen transfer remain lacking. Second, while our analysis of single-cell mTEC expression allowed us to estimate how many peptides are seen by individual T cells in the thymus, this approach required some assumptions, such as extrapolating the actual set of peptides present on the surface of mTECs from whole-cell expression and cell line–derived protein and peptide abundance estimates. Although peptide presentation is correlated with both RNA and protein expression levels (*68*–*70*), future work performing immunopeptidomics on antigen-presenting cells would allow for more precise estimates of peptide sampling. Third, while the BATMAN model of cross-reactivity takes into account two factors—biochemical similarity of amino acid substitutions and a weight on each position of the peptide sequence—both of these factors were derived from a database of only 25 index peptides and from peptide mutations that only systematically spanned the one-dimensional (1D) Hamming distance [although, reassuringly, BATMAN could reasonably interpolate outside the 1D space (*28*)]. Understanding the underlying distribution from which these factors are drawn across TCRs, alongside more expansive mutational scan data, would enable more accurate modeling of the diversity of TCR cross-reactivity sizes. Further, while our focus on 6-mer peptides allowed us to explore the $20^6 = 64$ million TCR space without subsampling, exploring TCR-pMHC interactions in the 9-mer space ($20^9 = 512$ billion TCRs), while unlikely to be possible exhaustively, could better account for potential allosteric interactions between MHC-binding residues and TCR-pMHC binding (*71*, *72*). Fourth, our model can be used to make HLA-specific predictions of target autoimmune epitopes to detect individuals at risk for tissue-specific autoimmune diseases. Fifth, generalization to avoid self must be balanced against overgeneralization, which may hinder detection of non-self (*9*), given the similarity between self and non–self-peptides (*24*, *73*). Understanding how this fine line can be achieved is reminiscent of anomaly detection algorithms (*74*).

Furthering this connection with machine learning, the T cell selection problem includes some unique twists on standard generalization problems. For example, in typical machine learning setups, the training and testing sets would not overlap, but in our case, they almost exactly overlap (i.e., the thymus and periphery contain essentially the same set of peptides), although with slightly shifted weights (peptide abundances). This represents a unique type of domain adaptation problem (*75*), where the training and testing distributions are close but not identical (i.e., mild distribution shift). In addition, the training set is typically fully accessible to the model, but in our case, the training set is only partially accessible to each T cell. Versions of this setup are used in instances where each learner (T cell) is trained on a different slice of the training data; for example, in ensemble learning, this idea is used to improve robustness or, more recently, in federated learning (*76*), to preserve privacy. Expansion of these classically studied machine learning paradigms to better model immunological realities could be mutually beneficial to both fields.

MATERIALS AND METHODS

MHC and haplotype selection

We analyzed alleles and haplotype frequency data from http://www.allelefrequencies.net (*78*). Population ancestries from http://www.allelefrequencies.net/datasets.asp#tag_5 were curated into six broadly continental and one mixed ancestry: Amerindian, Asian, European, Hispanic, Oriental, Sub-Saharan and Mixed. The most common allele is HLA-A02:01, which we focused on in our main analyses. To extend our analyses beyond a single HLA-A allele, we chose a haplotype common in 6 out of the 7 ancestries (figure S3): HLA-A01:01, HLA-B08:01, and HLA-C07:01.

Datasets

Our generalization model relies on peptide abundance estimates from peripheral tissues and thymic antigen-presenting cells. We infer the peptide abundances from gene expression data as described in the following sections. For direct comparison of peripheral and thymic expression, we analyzed bulk expression data because consistent bulk technologies were used across a large number of peripheral tissues (*27*) and from our published thymus study (*26*) (sample characteristics and tissues summarized in table S1). For estimating sampling statistics and peptide coverage in the thymus, we needed more fine-grained data, and hence used single-cell expression data from human thymic epithelial cells.

Bulk RNA sequencing data from 29 human tissues in the Genotype Tissue Expression dataset. We downloaded bulk gene expression from 948 donors of the Genotype Tissue Expression (GTEx) portal (version 8, https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression, accessed June 11, 2024, gene expression TPM from RNASeQCv1.1.9). Using sample mapping (GTEx Analysis v8 Annotations SampleAttributesDS.txt) and TPM files (GTEx Analysis 2017-06-05 v8 RNASeQCv1.1.9 gene tpm.gct.gz), we computed the mean expression per gene and per tissue across biological replicates. The final dataset contained 56,200 genes across the following 29 tissues (numbers in parentheses indicate number of donors per tissue): adrenal gland (258), bladder (21), blood (767), blood vessel (786), brain (382), breast (459), cervix uteri (14), colon (555), esophagus (710), fallopian tube (9), heart (561), kidney (88), liver (226), lung (578), muscle (803), nerve (619), ovary (180), pancreas (328), pituitary (283), prostate (245), salivary gland (162), skin (912), small intestine (187), spleen (241), stomach (359), testis (361), thyroid (653), uterus (142) and vagina (156).

Thymus bulk RNA sequencing data. We used the Transcript per Million (TPM) normalized expression data from immature and mature human bulk-sorted medullary thymic epithelial cells (mTECs) generated by Carter *et al.* (2022) (*26*) (GEO accession: GSE201719) to obtain mTEC (thymus) peptide weights. We first obtained the mean TPM expression for each transcript across three biological replicates of immature and mature mTEC samples and then summed these TPMs to obtain the total mTEC transcript expression. We then summed all transcripts mapping to a single gene to obtain TPMs for 40,481 genes.

Thymus single-cell RNA sequencing data. We re-processed five public single-cell RNA sequencing datasets comprising 23 human thymus samples (*62, 63, 67, 79*) and seven new single cell and single-nucleus RNAseq datasets of human pediatric thymus samples through a common pipeline: (i) alignment with STARsolo (*80*) (reference genome: GRCh38 (primary assembly); Gencode annotation (version V44), which includes gene count estimation with expectation maximization and cell filtering to retain singlets; and (ii) quality control with scanpy v 1.9.3 (*81*), where cells with mitochondrial gene percentages greater than 15% and with fewer than 200 expressed genes were removed from downstream analysis. We integrated single-cells across studies using scVI v 1.0.3 (*82*), correcting for sequencing protocol, batch and sample. Then, we identified all epithelial cells in the datasets,

followed by iterative subclustering of these cells to obtain 17 thymic epithelial cell types (data analyses manuscript in preparation). For all downstream analyses, we focused on cells that we annotated as mTEC II, which are characterized by high AIRE and promiscuous gene expression (*26*), thus providing an upper bound for gene diversity within a single cell. The low dimensional representation of each cell's gene expression is computed as latent factors by 'single-cell annotation using variational inference' (*81*) and is shown in Figure 3B, upper panel.

Peptide abundance

We downloaded the human reference proteome (id: UP000005640, Uniprot release: 2022 _02, downloaded 14 July 2022) and, following prior work (*24, 73*), used a sliding window to generate all possible 9-mer peptide sequences. We obtained the unique set of these peptides (11,136,576) and predicted their binding to HLA-A02:01 using NetMHCpan 4.1 (*83*). We combined weak and strong binders (netMHCpan binding affinity rank < 2) to obtain the set of 434,276 human 9-mer peptides with predicted HLA-A0:201 binding. For each binder, we removed the 1st, 2nd and 9th amino acid to yield 426,316 6-mer peptides with predicted HLA-A02:01 binding. We mapped each peptide to the protein it was derived from and then to the gene encoding the protein, and assigned each peptide a thymus and peripheral abundance (weight) based on the gene expression values in the bulk thymus and peripheral datasets, respectively. Specifically, we used Ensembl Biomart (version: grcg38.p14, genes110) to extract all ensembl gene ids and corresponding uniprot protein ids; any gene id without corresponding uniprot entry was removed from downstream analyses. Gene identifiers in the thymus (both bulk and single-cell) and peripheral datasets were then mapped to these uniprot ids and each peptide mapping to a given gene was assigned this gene's TPM expression level. Finally, we calculated unique peptide abundance levels by summing all TPMs for a given peptide.

Peptide sampling statistics in the thymus

There are four critical parameters needed to estimate the number of peptides sampled per T cell during negative selection: the total number of MHC complexes on the mTEC surface, the number of unique peptides that occupy these complexes, the percentage of pMHC complexes scanned by a T cell per mTEC encounter, and the total number of TCR-mTEC interactions.

1) Total number of MHC complexes on the mTEC surface: The number of MHC molecules on the surface of an mTEC differs by cell type and can change over time based on infection status and environment. However, across both murine and human unstimulated cells, different MHC classes, and different MHC alleles, roughly 10^5 molecules have been observed in measurements by immunoprecipitation using MHC-specific antibodies (*49, 84, 85*). Here, we chose an upper bound estimate of 250K MHC complexes/cell based on immunoprecipitation using the monoclonal antibody B22 recognizing the murine MHC I molecule D^b on EL4 thymoma cells (*49*).

2) Total number of unique peptides per mTEC: A common experimental strategy to estimate the number of unique peptides bound on MHC molecules is to take cells of interest and wash peptides off of MHC alleles by mild acid elution followed by peptide separation by microcapillary high-performance liquid chromatography. After identification of peptides by mass-spectrometry, peptide abundances can be estimated by comparison to a standard abundance curve of synthetic peptides. However, for the above processing pipeline, a starting cell number of 10^8 is often required. Despite recent progress in reducing the cell numbers down to 10^6 (*86, 87*), this assessment remains challenging for rare cell types from limited patient material, such as TECs, and to date there are no comprehensive estimates in these cells. Thus, we extrapolate from data derived from cell lines expressing the HLA-A02:01 allele, which detected a range of 100–3000 unique peptides per cell (*51, 88*). This estimate is also conserved across other MHC alleles and species (*50, 52*), allowing us to apply these estimates to other MHC loci.

3) Percentage of pMHC scanned per T cell encounter: Quantitative measurements of pMHC encounters on antigen-presenting cells (APC) such as mTECs are often focused on estimating the lower bound of specific pMHCs required to stimulate a T cell response (*89–91*). Here, we are interested not only in response-eliciting pMHC encounters, but an overall estimate of sampling of the pMHC space per APC. We approached this by estimating the surface area explored by a developing T cell upon encounter of an APC in the thymus. We relied on imaging tracks recorded by Bousso *et al.* (2002) (*53*, Figure 2), measuring the crawling of fluorescently labeled thymocytes on the APC surface in reaggregate thymic organ cultures. While variable from encounter to encounter, we estimate a lower and upper bound of surface area covered to be 10–80%. In our analyses, we used the upper bound estimate (80%), though this parameter had the smallest effect on the number of unique peptides seen because each peptide reoccurs many times across pMHC complexes on a cell (i.e., at least 250K/3000 ≈ 83 times).

4) Total number of TCR-APC interactions during negative selection: The estimated number of interactions between thymocytes and APCs are also derived from thymocyte imaging in a thymus culture system, here, agarose-embedded sections of thymic tissue explants (*8*). In addition, traceable thymocytes were introduced in this system by hematopoietic chimerism with fluorochrome-expressing bone marrow. Measuring thymocyte motility and interaction in this system, Le Borgne *et al.* (2009) (*8*) estimated on average 5 interactions between thymocytes and medullary APCs/hour. To estimate the total number of interactions, we consulted estimates from GFP-reporter mice where the level of GFP can indicate how much time has passed since being licensed for negative selection (*54*). These experiments conclude that thymocytes spend up to 4 days in the medulla, but are only in an apoptosis susceptible state, i.e., actively undergoing negative selection, for about two days (*55*). Together, we estimate about 5 interactions/hour x 48 hours = 240 interactions with APCs for each T cell.

Prior work on T cell cross-reactivity

Modeling T cell cross-reactivity requires a distance function that can be used to predict all the peptides that activate a given TCR. Extensive progress has been made on the opposite problem — predicting all the TCR sequences that can bind a given peptide (*92*) — but fewer studies have attempted to characterize the cross-reactivity function itself (*93, 94*), largely due to insufficient data detailing how small mutations in a TCR's index peptide affects TCR binding. Consequently, prior work (*25*) has modeled cross-reactivity using general sequence-based distance functions, such as r-contiguous matching (*24, 74*), Hamming distance (*73, 95, 96*), and BLOSUM distance (*38, 97–100*), which takes biochemical similarity of substitutions into account.

Determining the size of the cross-reactivity ball has been of intense interest in the literature (*42*), given its wide-spread importance to

pathogen detection, autoimmunity, and transplant rejection (*46*). Since it remains daunting to quantify this number comprehensively, most estimates in the literature combine sparse experimental probing of antigenic space with theoretical extrapolations. The reported estimates are also sensitive to the length of the peptide tested and whether MHC binding is considered. For example, Mason (1998) (*41*) estimated that a single T cell can cross-react to 1M peptides (length 9) without factoring MHC binding, and only 0.1% (*101*) to 1% (*102*) of peptides are believed to bind to a given MHC. Consequently, the cross-reactivity size for MHC-binding peptides may be closer to 1–10K. Hybrid experimental-theoretical estimates by Woolridge *et al.* (2012) (*102*) and Hiemstra *et al.* (1999) (*103*) support the 1M number and for MHC-binders, but these studies used longer peptides (lengths 10 and 11, respectively), which occupy a 20–400x larger space, and thus again, the average cross-reactivity size for length 9 MHC-binders may be in the low to mid thousands. Wortel *et al.* (2020) (*24*) considered length 6 peptides, noting that positions 2 and 9 are MHC anchor residues (*24, 25, 37, 38*) and position 1 mutations have the lowest effect on TCR binding (*24*, *37, 38*). Within the 6-mer space, Wortel *et al.* (2020) estimate that each TCR can bind to one in every 30K peptides (*104*), which again amounts to a cross-reactivity size of a few thousand.

Modeling T cell cross-reactivity using BATMAN

To model T cell cross-reactivity, we used BATMAN (*28*), a recent method employing a hierarchical Bayesian model to predict TCR activation by a given peptide based on its distance to the TCR's index peptide. This peptide-to-index distance is a product of two factors: (i) a learned amino acid (AA) substitution distance matrix from the index AA to the mutant AA, and (ii) a learned weight on each position in the peptide sequence. In the original work, these factors were learned from a large database, called BATCAVE, covering over 22,000 TCR-pMHC pairs and 151 mouse and human TCRs tested against 25 mutagenized index peptides.

In our main analysis, we focused on one HLA allele (A02:01) and on 6-mer peptides (positions 3–8 of each peptide). Consequently, we re-trained BATMAN using a subset of the BATCAVE data (i.e., only HLA-A02:01-binding peptides, as predicted by NetMHCpan4.1 (*83*), resulting in 1,429 TCR-pMHC pairs, containing 10 TCRs binding to 10 unique index peptides). Since the original 9-mer BATCAVE data had all peptide positions mutagenized, extracting positions 3–8 could result in multiple identical 6-mer peptide entries with potentially different TCR activation levels, even for the same TCR. Nevertheless, in five-fold cross-validation tests for TCR activation prediction, the AUCs were very consistent when using 9-mers (AUC = 0.709) versus 6-mers (AUC = 0.704). Moreover, the AA substitution matrices inferred using 9-mers versus 6-mers were highly correlated (Pearson r > 0.99), and, when using all 9 peptide positions, the inferred positional weights of positions 1, 2, and 9 were the smallest compared to positions 3–8. Together, these results further support our focus on peptide positions 3–8 towards accurate T cell cross-reactivity modeling.

For the other individual alleles (A01:01, B08:01, C07:01), as well as the haplotype analysis on the combined alleles (A01:01-B08:01-C07:01), we re-trained BATMAN on the full BATCAVE database, since there was not sufficient data for an allele-specific model.

Immune disease statistics

Immune disease statistics were based on a comprehensive literature survey by Hayter and Cook (*105*). They identified 81 autoimmune diseases that fulfil at least two of the following five criteria (as defined by the original study): (i) the specific adaptive immune response is directed to the affected organ or tissue; (ii) autoreactive T cells and/or autoantibodies are present in the affected organ or tissue; (iii) autoreactive T cells and/or autoantibodies can transfer the disease to healthy individuals or animals; (iv) immunization with the autoantigen induces the disease in animal models; and (v) elimination or suppression of the autoimmune response prevents disease progression or even ameliorates the clinical manifestation. We referred to Table 2 in the original publication, matching disease name (as indicated) with target tissue (not indicated) based on the provided molecular targets and references cited (table S4).

Aire deficiency model

Human Aire-dependent genes

We followed previous studies (*26*, *106, 107*) to obtain AIRE-dependent genes in human using orthologues of murine Aire-dependent genes (*3*). Aire-dependent genes were defined by Sansom *et al.* (2014) as differentially expressed genes between *Aire* knockout mTECs and mature *Aire* expressing mTECs in mice (Benjamini-Hochberg corrected pvalues ≤ 0.05). Supplemental Table 3, sheet 16 of Sansom *et al.* (2014) (*3*) provides all differentially expressed genes, and applying a fold change threshold of ≥ 2 yields the 3,980 AIRE dependent genes described in the paper. We matched this set of *Aire*-dependent genes to human orthologues, using biomaRT v2.46.3 (*108*), via attribute *hsapiens_homolog_ensembl_gene*, to obtain 3,361 unique human AIRE dependent genes (ensembl gene identifiers).

Generalization under AIRE-deficiency

To model negative selection in an AIRE-deficient thymus, we estimated thymus peptide abundances by removing the 3,361 human orthologs of Aire-dependent genes from the thymus bulk expression data, followed by peptide abundance estimation (Materials and Methods, *Peptide abundance*). We then applied our negative selection model on the AIRE-deficient thymus and analyzed all TCRs with survival probability > 0.98; we also analyzed TCRs that had a > 50% increased chance of survival in the AIRE deletion versus the baseline model (Table 1, column N_{TCRs}). For each set of TCRs, we obtained the union of peptides in their cross-reactivity balls (Table 1, column N_{peptides}). We then used a paired, one-sided t-test to determine if these sets of peptides were enriched in APS1 tissues (adrenal gland, liver, ovary, pancreas, skin, small intestine, thyroid, testis) versus non-APS1 tissues (bladder, blood, blood vessel, brain, breast, cervix uteri, colon, esophagus, fallopian tube, heart, kidney, lung, muscle, nerve, pituitary, prostate, salivary gland, spleen, stomach, uterus, vagina), measured at the sum of peptide weights across the respective tissues. To estimate empirical null distributions on these statistics, we randomly chose N_{TCRs} TCRs for each condition from the set 52.8 million self-reactive TCRs, and repeated the above analyses. For each random test, we drew 100 random sets of TCRs and computed the p-value associated with their t-test statistic. To obtain an empirical p-value, we counted the fraction of how often our observed p-value was smaller than the p-value across random sets, out of the total number of random sets.

Autoimmune target proteins in APS1

Landegren *et al.* (2016) examined the sera from 51 patients with APS1 and 21 healthy subjects for IgG autoantibodies on proteome arrays containing over 9,000 full-length human proteins (*61*). They identified the following 21 high-confidence immune targets in APS1: IFNW1, IFNA21, IFNA4, IFNA17, IFNA1, IFNA13, IFNA5, IFNA2, IFNA6, IFNA8, IFNA14, IL22, TPH1, DDC, GAD1, GAD2, GIF,

TGM4, MAGEB2, MAGEB4, PDILT. We mapped the peptide targets of TCRs with 50% increased survival probability in the AIRE-deleted thymus to their corresponding proteins and counted the overlap between the TCR target proteins and this set of high-confidence immune targets in APS1. All APS1 target proteins (except IFNA1 and GIF) overlapped with the TCR targets. We employed two statistical tests to derive empirical p-values. First, we generated 10,000 randomized protein sets, each of size 21 proteins (i.e., the number of observed true APS1 immune targets) out of the total set of proteins, and we counted the overlap of these random sets with the TCR-target proteins. Second, we generated 10,000 randomized protein sets, each of the same size as the number of observed TCR-target proteins, and we counted the overlap of these random sets with the APS1 immune targets. For both approaches, we obtained the empirical p-value by dividing the number of times the empirical test statistic was greater than or equal to the observed, divided by the total number of random sets (10,000).

Supplementary Materials

The PDF file includes:

Supplementary Text
Figs. S1 to S4
Legends for tables S1 and S4
Tables S2 and S3
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 and S4

REFERENCES

1. J. Sprent, H. Kishimoto, The thymus and central tolerance. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 609–616 (2001).
2. B. Kyewski, L. Klein, A central role for central tolerance. *Annu. Rev. Immunol.* **24**, 571–606 (2006).
3. S. N. Sansom, N. Shikama-Dorn, S. Zhanybekova, G. Nusspaumer, I. C. Macaulay, M. E. Deadman, A. Heger, C. P. Ponting, G. A. Holländer, Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res.* **24**, 1918–1931 (2014).
4. 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).
5. P. Brennecke, A. Reyes, S. Pinto, K. Rattay, M. Nguyen, R. Küchler, W. Huber, B. Kyewski, L. M. Steinmetz, Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat. Immunol.* **16**, 933–941 (2015).
6. M. J. Bevan, In thymic selection, peptide diversity gives and takes away. *Immunity* **7**, 175–178 (1997).
7. V. Detours, R. Mehr, A. S. Perelson, Deriving quantitative constraints on T cell selection from data on the mature T cell repertoire. *J. Immunol.* **164**, 121–128 (2000).
8. M. Le Borgne, E. Ladi, I. Dzhagalov, P. Herzmark, Y. F. Liao, A. K. Chakraborty, E. A. Robey, The impact of negative selection on thymocyte migration in the medulla. *Nat. Immunol.* **10**, 823–830 (2009).
9. T. Mora, A. M. Walczak, Towards a quantitative theory of tolerance. *Trends Immunol.* **44**, 512–518 (2023).
10. W. Yu, N. Jiang, P. J. R. Ebert, B. A. Kidd, S. Müller, P. J. Lund, J. Juang, K. Adachi, T. Tse, M. E. Birnbaum, E. W. Newell, D. M. Wilson, G. M. Grotenbreg, S. Valitutti, S. R. Quake, M. M. Davis, Clonal deletion prunes but does not eliminate self-specific αβ $CD8^+$ T lymphocytes. *Immunity* **42**, 929–941 (2015).
11. J. J. Calis, R. J. de Boer, C. Keşmir, Degenerate T-cell recognition of peptides on MHC molecules creates large holes in the T-cell repertoire. *PLOS Comput. Biol.* **8**, e1002412 (2012).
12. M. M. Davis, Not-so-negative selection. *Immunity* **43**, 833–835 (2015).
13. D. M. Richards, B. Kyewski, M. Feuerer, Re-examining the nature and function of self-reactive T cells. *Trends Immunol.* **37**, 114–125 (2016).
14. F. Camaglia, A. Ryvkin, E. Greenstein, S. Reich-Zeliger, B. Chain, T. Mora, A. M. Walczak, N. Friedman, Quantifying changes in the T cell receptor repertoire during thymic development. *eLife* **12**, e81622 (2023).
15. D. L. Mueller, Mechanisms maintaining peripheral tolerance. *Nat. Immunol.* **11**, 21–27 (2010).
16. R. Marsland, O. Howell, A. Mayer, P. Mehta, Tregs self-organize into a computing ecosystem and implement a sophisticated optimization algorithm for mediating immune response. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2011709118 (2021).
17. T. C. Butler, M. Kardar, A. K. Chakraborty, Quorum sensing allows T cells to discriminate between self and nonself. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11833–11838 (2013).
18. G. Altan-Bonnet, T. Mora, A. M. Walczak, Quantitative immunology for physicists. *Phys. Rep.* **849**, 1–83 (2020).
19. R. H. Schwartz, T cell anergy. *Annu. Rev. Immunol.* **21**, 305–334 (2003).
20. E. M. Orlova, L. S. Sozaeva, M. A. Kareva, B. E. Oftedal, A. S. B. Wolff, L. Breivik, E. Y. Zakharova, O. N. Ivanova, O. Kämpe, I. I. Dedov, P. M. Knappskog, V. A. Peterkova, E. S. Husebye, Expanding the phenotypic and genotypic landscape of autoimmune polyendocrine syndrome type 1. *J. Clin. Endocrinol. Metabol.* **102**, 3546–3556 (2017).
21. K. Nagamine, P. Peterson, H. S. Scott, J. Kudoh, S. Minoshima, M. Heino, K. J. Krohn, M. D. Lalioti, P. E. Mullis, S. E. Antonarakis, K. Kawasaki, S. Asakawa, F. Ito, N. Shimizu, Positional cloning of the APECED gene. *Nat. Genet.* **17**, 393–398 (1997).
22. J. Aaltonen, P. Björses, J. Perheentupa, N. Horelli–Kuitunen, A. Palotie, L. Peltonen, Y. S. Lee, F. Francis, S. Henning, C. Thiel, H. Leharach, M.-L. Yaspo, An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nat. Genet.* **17**, 399–403 (1997).
23. D. L. Chao, M. P. Davenport, S. Forrest, A. S. Perelson, The effects of thymic selection on the range of T cell cross-reactivity. *Eur. J. Immunol.* **35**, 3452–3459 (2005).
24. I. M. N. Wortel, C. Keşmir, R. J. de Boer, J. N. Mandl, J. Textor, Is T cell negative selection a learning algorithm? *Cells* **9**, 690 (2020).
25. B. Koncz, G. M. Balogh, M. Manczinger, A journey to your self: The vague definition of immune self and its practical implications. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2309674121 (2024).
26. J. A. Carter, L. Strömich, M. Peacey, S. R. Chapin, L. Velten, L. M. Steinmetz, B. Brors, S. Pinto, H. V. Meyer, Transcriptomic diversity in human medullary thymic epithelial cells. *Nat. Commun.* **13**, 4296 (2022).
27. The GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
28. A. Banerjee, D. J. Pattinson, C. L. Wincek, P. Bunk, A. Axhemi, S. R. Chapin, S. Navlakha, H. V. Meyer, T cell receptor cross-reactivity prediction improved by a comprehensive mutational scan database. *Cell Syst.* **16**, 101345 (2025).
29. T. Cover, P. Hart, Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
30. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algo-rithms* (Cambridge Univ. Press, 2014).
31. D. Malhotra, J. L. Linehan, T. Dileepan, Y. J. Lee, W. E. Purtha, J. V. Lu, R. W. Nelson, B. T. Fife, H. T. Orr, M. S. Anderson, K. A. Hogquist, M. K. Jenkins, Tolerance is established in polyclonal $CD4^+$ T cells by distinct mecha-nisms, according to self-peptide expression patterns. *Nat. Immunol.* **17**, 187–195 (2016).
32. L. Klein, M. Klugmann, K.-A. Nave, V. K. Tuohy, B. Kyewski, Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nat. Med.* **6**, 56–61 (2000).
33. H. Lv, E. Havari, S. Pinto, R. V. S. R. K. Gottumukkala, L. Cornivelli, K. Raddassi, T. Matsui, A. Rosenzweig, R. T. Bronson, R. Smith, A. L. Fletcher, S. J. Turley, K. Wucherpfennig, B. Kyewski, M. A. Lipes, Impaired thymic tolerance to α-myosin directs autoimmunity to the heart in mice and humans. *J. Clin. Invest.* **121**, 1561–1573 (2011).
34. R. V. S. R. K. Gottumukkala, H. J. Lv, L. Cornivelli, A. J. Wagers, R. Y. Kwong, R. Bronson, G. C. Stewart, P. C. Schulze, W. Chutkow, H. A. Wolpert, R. T. Lee, M. A. Lipes, Myocardial infarction triggers chronic cardiac autoimmunity in type 1 diabetes. *Sci. Transl. Med.* **4**, 138ra80 (2012).
35. S. Pinto, D. Sommermeyer, C. Michel, S. Wilde, D. Schendel, W. Uckert, T. Blankenstein, B. Kyewski, Misinitiation of intrathymic MART-1 transcription and biased TCR usage explain the high frequency of MART-1-specific T cells. *Eur. J. Immunol.* **44**, 2811–2821 (2014).
36. N. J. Burroughs, R. J. de Boer, C. Keşmir, Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics* **56**, 311–320 (2004).
37. J. J. A. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Keşmir, B. Peters, Properties of MHC class I presented peptides that enhance immunogenicity. *PLOS Comput. Biol.* **9**, e1003266 (2013).
38. S. Frankild, R. J. de Boer, O. Lund, M. Nielsen, C. Kesmir, Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLOS ONE* **3**, e1831 (2008).
39. F. Dhalla, J. Baran-Gale, S. Maio, L. Chappell, G. A. Holländer, C. P. Ponting, Biologically indeterminate yet ordered promiscuous gene expression in single medullary thymic epithelial cells. *EMBO J.* **39**, e101828 (2019).
40. M. E. Moses, J. L. Cannon, D. M. Gordon, S. Forrest, Distributed adaptive search in T cells: Lessons from ants. *Front. Immunol.* **10**, 1357 (2019).
41. D. Mason, A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunol. Today* **19**, 395–404 (1998).
42. A. K. Sewell, Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
43. J. A. Gaevert, D. L. Duque, G. Lythe, C. Molina-París, P. G. Thomas, Quantifying T cell cross-reactivity: Influenza and coronaviruses. *Viruses* **13**, 1786 (2021).

44. C. Gouttefangeas, R. Klein, A. Maia, The good and the bad of T cell cross-reactivity: Challenges and opportunities for novel therapeutics in autoimmunity and cancer. *Front. Immunol.* **14**, 1212546 (2023).
45. D. J. Smith, S. Forrest, R. R. Hightower, A. S. Perelson, Deriving shape space parameters from immunological data. *J. Theor. Biol.* **189**, 141–150 (1997).
46. M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Ozkan, M. M. Davis, K. W. Wucherpfennig, K. C. Garcia, Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).
47. D. G. Schatz, Y. Ji, Recombination centres and the orchestration of V(D) J recombination. *Nat. Rev. Immunol.* **11**, 251–263 (2011).
48. A. J. Yates, Theories and quantification of thymic selection. *Front. Immunol.* **5**, 13 (2014).
49. E. R. Christinck, M. A. Luscher, B. H. Barber, D. B. Williams, Peptide binding to class I MHC on living cells and quantitation of complexes required for CTL lysis. *Nature* **352**, 67–70 (1991).
50. K. Falk, O. Rötzschke, K. Deres, J. Metzger, G. Jung, H. G. Rammensee, Identification of naturally processed viral nonapeptides allows their quan-tification in infected cells and suggests an allele-specific T cell epitope forecast. *J. Exp. Med.* **174**, 425–434 (1991).
51. D. F. Hunt, R. A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. L. Cox, E. Appella, V. H. Engelhard, Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
52. E. L. Huczko, W. M. Bodnar, D. Benjamin, K. Sakaguchi, N. Z. Zhu, J. Shabanowitz, R. A. Henderson, E. Appella, D. F. Hunt, V. H. Engelhard, Characteristics of endogenous peptides eluted from the class I MHC molecule HLA-B7 determined by mass spectrometry and computer modeling. *J. Immunol.* **151**, 2572–2587 (1993).
53. P. Bousso, N. R. Bhakta, R. S. Lewis, E. Robey, Dynamics of thymocyte-stromal cell interac-tions visualized by two-photon microscopy. *Science* **296**, 1876–1880 (2002).
54. T. M. McCaughtry, M. S. Wilken, K. A. Hogquist, Thymic emigration revisited. *J. Exp. Med.* **204**, 2513–2520 (2007).
55. M. A. Weinreich, K. A. Hogquist, Thymic emigration: When and how T cells leave home. *J. Immunol.* **181**, 2265–2270 (2008).
56. J. P. van Meerwijk, S. Marguerat, R. K. Lees, R. N. Germain, B. J. Fowlkes, H. R. MacDonald, Quantitative impact of thymic clonal deletion on the T cell repertoire. *J. Exp. Med.* **185**, 377–384 (1997).
57. E. S. Huseby, J. White, F. Crawford, T. Vass, D. Becker, C. Pinilla, P. Marrack, J. W. Kappler, How the T cell repertoire becomes peptide and MHC specific. *Cell* **122**, 247–260 (2005).
58. A. Košmrlj, E. L. Read, Y. Qi, T. M. Allen, M. Altfeld, S. G. Deeks, F. Pereyra, M. Carrington, B. D. Walker, A. K. Chakraborty, Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* **465**, 350–354 (2010).
59. R. J. De Boer, C. Kesmir, A. S. Perelson, J. A. M. Borghans, Is the exquisite specificity of lymphocytes generated by thymic selection or due to evolution? *Front. Immunol.* **15**, 1266349 (2024).
60. M. S. Anderson, E. S. Venanzi, L. Klein, Z. Chen, S. P. Berzins, S. J. Turley, H. von Boehmer, R. Bronson, A. Dierich, C. Benoist, D. Mathis, Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395–1401 (2002).
61. N. Landegren, D. Sharon, E. Freyhult, Å. Hallgren, D. Eriksson, P. H. Edqvist, S. Bensing, J. Wahlberg, L. M. Nelson, J. Gustafsson, E. S. Husebye, M. S. Anderson, M. Snyder, O. Kämpe, Proteome-wide survey of the autoimmune target repertoire in autoimmune polyendocrine syndrome type 1. *Sci. Rep.* **6**, 20104 (2016).
62. J.-E. Park, R. A. Botting, C. Domínguez Conde, D. M. Popescu, M. Lavaert, D. J. Kunz, I. Goh, E. Stephenson, R. Ragazzini, E. Tuck, A. Wilbrey-Clark, K. Roberts, V. R. Kedlian, J. R. Ferdinand, X. He, S. Webb, D. Maunder, N. Vandamme, K. T. Mahbubani, K. Polanski, L. Mamanova, L. Bolt, D. Crossland, F. de Rita, A. Fuller, A. Filby, G. Reynolds, D. Dixon, K. Saeb-Parsy, S. Lisgo, D. Henderson, R. Vento-Tormo, O. A. Bayraktar, R. A. Barker, K. B. Meyer, Y. Saeys, P. Bonfanti, S. Behjati, M. R. Clatworthy, T. Taghon, M. Haniffa, S. A. Teichmann, A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
63. B. D. Huisman, D. A. Michelson, S. A. Rubin, K. Kohlsaat, W. Gomarga, Y. Fang, J. M. Lee, P. del Nido, M. Nathan, C. Benoist, L. Zon, D. Mathis, Cross-species analyses of thymic mimetic cells reveal evolutionarily ancient origins and both conserved and species-specific elements. *Immunity* **58**, 108–123.e7 (2025).
64. A. M. Afzali, L. Nirschl, C. Sie, M. Pfaller, O. Ulianov, T. Hassler, C. Federle, E. Petrozziello, S. R. Kalluri, H. H. Chen, S. Tyystjärvi, A. Muschaweckh, K. Lammens, C. Delbridge, A. Büttner, K. Steiger, G. Seyhan, O. P. Ottersen, R. Öllinger, R. Rad, S. Jarosch, A. Straub, A. Mühlbauer, S. Grassmann, B. Hemmer, J. P. Böttcher, I. Wagner, M. Kreutzfeldt, D. Merkler, I. B. Pardàs, M. S. Supprian, V. R. Buchholz, S. Heink, D. H. Busch, L. Klein, T. Korn, B cells orchestrate tolerance to the neuromyelitis optica autoantigen AQP4. *Nature* **627**, 407–415 (2024).
65. J. N. Lancaster, H. M. Thyagarajan, J. Srinivasan, Y. Li, Z. Hu, L. I. R. Ehrlich, Live-cell imaging reveals the relative contributions of antigen-presenting cell subsets to thymic central tolerance. *Nat. Commun.* **10**, 2220 (2019).
66. M. Vobo˘ril, J. Březina, T. Brabec, J. Dobeš, O. Ballek, M. Dobešová, J. Manning, R. S. Blumberg, D. Filipp, A model of preferential pairing between epithelial and dendritic cells in thymic antigen transfer. *eLife* **11**, e71578 (2022).
67. J. L. Bautista, N. T. Cramer, C. N. Miller, J. Chavez, D. I. Berrios, L. E. Byrnes, J. Germino, V. Ntranos, J. B. Sneddon, T. D. Burt, J. M. Gardner, C. J. Ye, M. S. Anderson, A. V. Parent, Single-cell transcriptional profiling of human thymic stroma uncovers novel cellular heterogeneity in the thymic medulla. *Nat. Commun.* **12**, 1096 (2021).
68. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14**, 658–673 (2015).
69. A. S. Juncker, M. V. Larsen, N. Weinhold, M. Nielsen, S. Brunak, O. Lund, Systematic characterisation of cellular localisation and expression profiles of proteins containing MHC ligands. *PLOS ONE* **4**, e7448 (2009).
70. J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, K. R. Clauser, N. Hacohen, M. S. Rooney, S. A. Carr, C. J. Wu, Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326 (2017).
71. K. Natarajan, A. C. McShan, J. Jiang, V. K. Kumirov, R. Wang, H. Zhao, P. Schuck, M. E. Tilahun, L. F. Boyd, J. Ying, A. Bax, D. H. Margulies, N. G. Sgourakis, An allosteric site in the T-cell receptor Cβ domain plays a critical signalling role. *Nat. Commun.* **8**, 15260 (2017).
72. S. Rangarajan, Y. He, Y. Chen, M. C. Kerzic, B. Ma, R. Gowthaman, B. G. Pierce, R. Nussinov, R. A. Mariuzza, J. Orban, Peptide-MHC (pMHC) binding to a human antiviral T cell receptor induces long-range allosteric communication between pMHC- and CD3-binding sites. *J. Biol. Chem.* **293**, 15991–16005 (2018).
73. A. Tiffeau-Mayer, J. A. Levine, C. J. Russo, Q. Marcou, W. Bialek, B. D. Greenbaum, How different are self and nonself? *PRX Life* **4**, 013027 (2026).
74. D. Dasgupta, *Artificial immune systems and their applications* (Springer Berlin Heidelberg, 2012).
75. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains. *Mach. Learn.* **79**, 151–175 (2010).
76. J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, W. Zhang, A survey on federated learning: Challenges and applications. *Int. J. Mach. Learn. Cybern.* **14**, 513–535 (2023).
77. A. C. Kohlgruber, M. H. Dezfulian, B. M. Sie, C. I. Wang, T. Kula, U. Laserson, H. B. Larman, S. J. Elledge, High-throughput discovery of MHC class I- and II-restricted T cell epitopes using synthetic cellular circuits. *Nat. Biotechnol.* **43**, 623–634 (2025).
78. F. Gonzalez-Galarza, A. McCabe, E. J. M. D. Santos, J. Jones, L. Takeshita, N. D. Ortega-Rivera, G. M. D. Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, D. Middleton, A. R. Jones, Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788 (2020).
79. S. Campinoti, A. Gjinovci, R. Ragazzini, L. Zanieri, L. Ariza-Mc Naughton, M. Catucci, S. Boeing, J.-E. Park, J. C. Hutchinson, M. Muñoz-Ruiz, P. G. Manti, G. Vozza, C. E. Villa, D.-E. Phylactopoulos, C. Maurer, G. Testa, H. J. Stauss, S. A. Teichmann, N. J. Sebire, A. C. Hayday, D. Bonnet, P. Bonfanti, Reconstitution of a functional human thymus by postnatal stromal progenitor cells and natural whole-organ scaffolds. *Nat. Commun.* **11**, 6372 (2020).
80. B. Kaminow, D. Yunusov, A. Dobin, STARsolo: Accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv 2021.05.05.442755 [Preprint] (2021). https://doi.org/10.1101/2021.05.05.442755.
81. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
82. C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, N. Yosef, Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
83. B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
84. S. Demotz, H. M. Grey, A. Sette, The minimal number of class II MHC-antigen complexes needed for T cell activation. *Science* **249**, 1028–1030 (1990).
85. J. W. Yewdell, J. R. Bennink, Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* **17**, 51–88 (1999).
86. S. Klaeger, A. Apffel, K. R. Clauser, S. Sarkizova, G. Oliveira, S. Rachimi, P. M. le, A. Tarren, V. Chea, J. G. Abelin, D. A. Braun, P. A. Ott, H. Keshishian, N. Hacohen, D. B. Keskin, C. J. Wu, S. A. Carr, Optimized liquid and gas phase fractionation increases HLA-peptidome coverage for primary cell and tissue samples. *Mol. Cell. Proteomics* **20**, 100133 (2021).
87. D. Oliinyk, H. R. Gurung, Z. Zhou, K. Leskoske, C. M. Rose, S. Klaeger, diaPASEF analysis for HLA-I peptides enables quantification of common cancer neoantigens. *Mol. Cell. Proteomics* **24**, 100938 (2025).
88. R. A. Henderson, A. L. Cox, K. Sakaguchi, E. Appella, J. Shabanowitz, D. F. Hunt, V. H. Engelhard, Direct identification of an endogenous peptide recognized by multiple HLA-A2.1-specific cytotoxic T cells. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 10275–10279 (1993).
89. Y. Sykulev, M. Joo, I. Vturina, T. J. Tsomides, H. N. Eisen, Evidence that a single peptide–MHC complex on a target cell can elicit a cytolytic T cell response. *Immunity* **4**, 565–571 (1996).

90. G. Altan-Bonnet, R. N. Germain, Modeling T cell antigen discrimination based on feedback control of digital ERK responses. *PLOS Biol.* **3**, e356 (2005).
91. M. Krogsgaard, Q. J. Li, C. Sumen, J. B. Huppa, M. Huse, M. M. Davis, Agonist/endogenous peptide–MHC heterodimers drive T cell activation and sensitivity. *Nature* **434**, 238–243 (2005).
92. D. Hudson, R. A. Fernandes, M. Basham, G. Ogg, H. Koohy, Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**, 511–521 (2023).
93. V. P. Balachandran, M. Łuksza, J. N. Zhao, V. Makarov, J. A. Moral, R. Remark, B. Herbst, G. Askan, U. Bhanot, Y. Senbabaoglu, D. K. Wells, C. I. O. Cary, O. Grbovic-Huezo, M. Attiyeh, B. Medina, J. Zhang, J. Loo, J. Saglimbeni, M. Abu-Akeel, R. Zappasodi, N. Riaz, M. Smoragiewicz, Z. L. Kelley, O. Basturk, Australian Pancreatic Cancer Genome Initiative, Garvan Institute of Medical Research, Prince of Wales Hospital, Royal North Shore Hospital, University of Glasgow, St Vincent's Hospital, QIMR Berghofer Medical Research Institute, University of Melbourne, Centre for Cancer Research, University of Queensland, Institute for Molecular Bioscience, Bankstown Hospital, Liverpool Hospital, Royal Prince Alfred Hospital, Chris O'Brien Lifehouse, Westmead Hospital, Fremantle Hospital, St John of God Healthcare, Royal Adelaide Hospital, Flinders Medical Centre, Envoi Pathology, Princess Alexandria Hospital, Austin Hospital, Johns Hopkins Medical Institutes, ARC-Net Centre for Applied Research on Cancer, M. Gönen, A. J. Levine, P. J. Allen, D. T. Fearon, M. Merad, S. Gnjatic, C. A. Iacobuzio-Donahue, J. D. Wolchok, R. DeMatteo, T. A. Chan, B. D. Greenbaum, T. Merghoub, S. D. Leach, Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551**, 512–516 (2017).
94. F. Drost, E. Dorigatti, A. Straub, P. Hilgendorf, K. I. Wagner, K. Heyer, M. López Montes, B. Bischl, D. H. Busch, K. Schober, B. Schubert, Predicting T cell receptor functionality against mutant epitopes. *Cell Genom.* **4**, 100634 (2024).
95. D. Santoni, Viral peptides-MHC interaction: Binding probability and distance from human peptides. *J. Immunol. Methods* **459**, 35–43 (2018).
96. D. Vergni, R. Gaudio, D. Santoni, The farther the better: Investigating how distance from human self affects the propensity of a peptide to be presented on cell surface by MHC class I molecules, the case of *Trypanosoma cruzi*. *PLOS ONE* **15**, e0243285 (2020).
97. A. Bresciani, S. Paul, N. Schommer, M. B. Dillon, T. Bancroft, J. Greenbaum, A. Sette, M. Nielsen, B. Peters, T-cell recognition is shaped by epitope sequence conservation in the host proteome and microbiome. *Immunology* **148**, 34–39 (2016).
98. A.-M. Bjerregaard, M. Nielsen, V. Jurtz, C. M. Barra, S. R. Hadrup, Z. Szallasi, A. C. Eklund, An analysis of natural T cell responses to predicted tumor neoepitopes. *Front. Immunol.* **8**, 1566 (2017).
99. L. P. Richman, R. H. Vonderheide, A. J. Rech, Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* **9**, 375–382.e4 (2019).
100. A. Gao, Z. Chen, A. Amitai, J. Doelger, V. Mallajosyula, E. Sundquist, F. P. Segal, M. Carrington, M. M. Davis, H. Streeck, A. K. Chakraborty, B. Julg, Learning from HIV-1 to predict the immunogenicity of T cell epitopes in SARS-CoV-2. *iScience* **24**, 102311 (2021).
101. P. M. Bruno, R. T. Timms, N. S. Abdelfattah, Y. Leng, F. J. N. Lelis, D. R. Wesemann, X. G. Yu, S. J. Elledge, High-throughput, targeted MHC class I immunopeptidomics using a functional genetics screening platform. *Nat. Biotechnol.* **41**, 980–992 (2023).
102. L. Wooldridge, J. Ekeruche-Makinde, H. A. van den Berg, A. Skowera, J. J. Miles, M. P. Tan, G. Dolton, M. Clement, S. Llewellyn-Lacey, D. A. Price, M. Peakman, A. K. Sewell, A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* **287**, 1168–1177 (2012).
103. H. S. Hiemstra, P. A. van Veelen, S. J. M. Willemen, W. E. Benckhuijsen, A. Geluk, R. R. P. de Vries, B. O. Roep, J. W. Drijfhout, Quantitative determination of TCR cross-reactivity using peptide libraries and protein databases. *Eur. J. Immunol.* **29**, 2385–2391 (1999).
104. J. Ishizuka, K. Grebe, E. Shenderov, B. Peters, Q. Chen, Y. C. Peng, L. Wang, T. Dong, V. Pasquetto, C. Oseroff, J. Sidney, H. Hickman, V. Cerundolo, A. Sette, J. R. Bennink, A. McMichael, J. W. Yewdell, Quantitating T cell cross-reactivity for unrelated peptide antigens. *J. Immunol.* **183**, 4337–4345 (2009).
105. S. M. Hayter, M. C. Cook, Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun. Rev.* **11**, 754–765 (2012).
106. S. Pinto, C. Michel, H. Schmidt-Glenewinkel, N. Harder, K. Rohr, S. Wild, B. Brors, B. Kyewski, Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3497–E3505 (2013).
107. S. Cepeda, C. Cantu, S. Orozco, Y. Xiao, Z. Brown, M. K. Semwal, T. Venables, M. S. Anderson, A. V. Griffith, Age-associated decline in thymic B cell expression of aire and aire-dependent self-antigens. *Cell Rep.* **22**, 1276–1287 (2018).
108. S. Durinck, P. T. Spellman, E. Birney, W. Huber, Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

Acknowledgments: We thank V. Kovaleva, D. Pattinson, Y. Xie, and QED science for feedback on the manuscript; V. Kovaleva for help in generating figure schematics; and F. Gonzalez-Galarza for providing frequency files from www.allelefrequencies.net. **Funding:** The research was supported by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory; the Cold Spring Harbor Laboratory and Northwell Health Affiliation; US National Institutes of Health grants S10OD028632-01 and 1R01AI167862 (to H.V.M.); and the Simons Pivot Fellowship (to H.V.M. and S.N.). The funders had no role in the template design or decision to publish. **Author contributions:** Investigation: A.B., R.K.P., S.R.C., Y.L., H.V.M., and S.N. Conceptualization: H.V.M. and S.N. Methodology: S.D., H.V.M., and S.N. Formal analyses: A.B., S.D., H.V.M., and S.N. Data curation: R.K.P., S.R.C., Y.L., H.V.M., and S.N. Software: C.K., A.B., H.V.M., and S.N. Resources: R.K.P., S.R.C., Y.L., H.V.M., and S.N. Visualization: A.B., H.V.M., and S.N. Writing—original draft: H.V.M. and S.N. Writing—review and editing: All authors; Project administration: H.V.M. and S.N. Validation: H.V.M. and S.N. Funding acquisition: H.V.M. and S.N. Supervision: H.V.M. and S.N. **Competing interests:** C.K. is a cofounder of Ocean Genomics Inc. All other authors declare that they have no competing interests. **Data, code, and materials availability:** All data and code needed to evaluate and reproduce the conclusions in the paper are present in the paper and/or the Supplementary Materials. All preprocessed data to regenerate results and manuscript figures are available on figshare (https://doi.org/10.6084/m9.figshare.32952404). Custom analysis code was written in either R (version ≥4.3.3) or Python (version ≥3.6.8) and is freely available on Zenodo (https://doi.org/10.5281/zenodo.21297214) and GitHub (https://github.com/meyer-lab-cshl/central-tolerance-generalization/). No new materials were generated.

Submitted 2 March 2026
Accepted 14 July 2026
Published 19 August 2026
10.1126/sciadv.aeg8240

ScienceAdvances

Central T cell tolerance from sparse peptide sampling

Hannah V. Meyer, Sanjoy Dasgupta, Amitava Banerjee, Yong Lin, Rishvanth K. Prabakar, Sarah R. Chapin, Carl Kingsford, and Saket Navlakha

Sci. Adv. **12** (34), eaeg8240. DOI: 10.1126/sciadv.aeg8240

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.