Supplementary Information: Effect of Thymic Selection on T-Cell Repertoire Recognition of Foreign and Neoantigenic Peptides

J. T. George, D. A. Kessler, H. Levine

August 9, 2017

S1 Overview

We provide here more details and analysis of the three models considered in the main text; the sequential MJ model, the PIRA approach and the RICE construction. We begin in section S2 with a expanded discussion of the general framework of this class of models, followed with a detailed description of the Chakraborty MJ model in Section S3. In Section S4 we specify our modification of this model, the sequential MJ (S-MJ) model and present the details of our calculations of the selection curve and potency distribution in this model. We then define in Section S5 the PIRA model and present the parallel analysis of this model. In Section S6 we detail the RICE model, which forms the main focus of this work, again presenting the analysis of the selection curve and potency distribution. In Section S7, we present the details of our analysis of the recognition of neoantigens and foreign peptides in the RICE model, and Section S8 treats the problem of allogeneic response in this context. In these sections, we use both Gaussian distributed interactions as well as those drawn from a uniform distribution, just to show the generality of the findings. In the last section S9 we investigate changes that would ensue had we used a smaller amino acid alphabet, so as to take into account the chemical similarity between different residues.

S2 General TCR-MHC Interaction and Thymic Selection

We discuss the overall model-building framework which has been widely adopted [1–4]. Our focus is on MHC-I, in the context of CD8+ T-cell recognition. MHC-I comes in 3 varieties, HLA A-C. Each individual has 2 subvarieties of each molecule. The total number of known subvarieties of HLA A, B, and C molecules are 3657, 4459, and 3290, respectively [5].

For a given individual, let $\mathcal{M} = \{M_r : r = 1, ..., 6\}$ denote the collection of MHC-I molecules, which are classified into three groups A, B, and C, and are distributed in the following manner: $M_1, M_2 \sim \mathcal{D}_A$, $M_3, M_4 \sim \mathcal{D}_B$, and $M_5, M_6 \sim \mathcal{D}_C$. Let $\mathscr{m} = \left\{m_M^{(j)} : j = 1, ..., N, M \in \mathcal{M}\right\}$ be a collection of MHC-loaded peptides with distributions $m_M^{(j)} \sim \mathcal{D}_M$, $\mathcal{T} = \left\{T^{(j)} : j = 1, ..., N_t\right\}$ be a collection of variable TCR CDR$_{1,2}$ (MHC-contacting) regions of various thymocyte receptors, and $\tau = \left\{t^{(j)} : j = 1, ..., N_t\right\}$ be a collection of variable TCR CDR$_3$ (peptide-contacting) regions of various thymocyte receptors.

Elements of $\mathcal{T}$ represent variety in the TCR CDR$_{1,2}$ regions, while elements of τ represent variety in the TCR CDR$_3$ regions. These segments are generated from separate mutational events in VDJ recombination, and are thus considered statistically independent here. We identify each TCR by the ordered pair $(T^{(j)}, t^{(j)})$. A given MHC-I molecule may be restricted in the variety of peptides it can bind. Peptide-bound MHC therefore depends on the particular molecule, M_r. We identify each MHC-peptide complex by

$(M_r, m^{(j)}_{M_r})$. Interactions between the TCR and p-MHC are quantified below, and may represent a relevant quantity, such as binding energy or MHC-receptor lifetime. Here, we do not focus on any precise molecular mechanism of activation.

Let f_1 denote the interaction contribution from TCR CDR$_{1,2}$ regions contacting MHC directly, and f_2 denote the interaction contribution from the TCR CDR$_3$ region interfacing with an an MHC-bound peptide, which are assumed to add linearly. It is customary [1–4] to make the simplification that TCR-CDR$_3$ interactions are independent of the TCR-CDR$_{1,2}$ regions. That is, $f_2 = f_2\big(m^{(j)}_{M_r}, t^{(l)}\big)$ (dependencies of f_1 on $m^{(j)}_{M_r}, t^{(l)}$ allow for the possibility that peptide-CDR$_3$ combinations may alter the interaction strength of MHC-CDR$_{1,2}$). Then,

$$E\big(T^{(l)}, M_i, t^{(l)}, m^{(j)}_{M_r},\big) = f_1\big(T^{(l)}, M_r, t^{(l)}, m^{(j)}_{M_r}\big) + f_2\big(t^{(l)}, m^{(j)}_{M_r}\big). \tag{S1}$$

This then represents the general way in which a TCR (labeled by l) interacts with a given MHC-peptide complex, of HLS type r and sequence j.

S3 Chakraborty MJ Model

Working within the approach presented in Section S2, Chakraborty et al. introduced a more restricted framework, which we adopt throughout this work. They focused only on a single MHC molecule per individual and hence dropped the index r ($\mathcal{M} = \{M\}$). For definiteness, they assumed that each MHC-loaded peptide is represented as a decamer peptide so that $\mathscr{m} = \big\{m^{(j)} = \big(m^{(j)}_i\big)^k_{i=1}, j = 1, ..., N\big\}$ with $k = 10$. Moreover, there is only one type of MHC-contacting CDR$_{1,2}$ region ($\mathcal{T} = \{T\}$) considered. Each peptide-contacting TCR CDR$_3$ region is also modeled as a decamer peptide ($\tau = \{(t^{(j)}_i)^k_{i=1}, j = 1, ..., N_t\}$). Both $m^{(j)}_i$ and $t_i(k)$ are distributed according to $P_h(a)$, $a \in A$, where A represents the set of 20 naturally-occurring amino acids, ($|A| = 20$) and P_h represents the probability mass function (pmf) for the amino acids found in the human proteome (Table S2).

In this model, MHC-TCR interactions, E, are interpreted as total binding energy. The MHC-CDH$_{1,2}$ binding energy is taken to be a constant value independent of the specific peptide and peptide-binding CDR$_3$ so that

$$f_1\big(T^{(l)}, M_i, t^{(l)}, m^{(j)}_{M_i}\big) = E_c. \tag{S2}$$

The peptide-CDH$_3$ binding energy is taken to equal a sum of k pairwise amino acid interaction energies, giving

$$f_2\big(t^{(l)}, m^{(j)}\big) = \sum_{i=1}^{k} f^i_2\big(t^{(l)}_i, m^{(j)}_i\big). \tag{S3}$$

Chakraborty further assumed that f^i_2 is independent of the site i and takes the values of pairwise amino acid binding energies captured in the MJ matrix (see Table S1). Thus, for a given TCR $t \in \tau$ and p-MHC $m^{(j)}$ Eq. S1 becomes:

$$E\big(T, M, t, m^{(j)}\big) = E\big(t, m^{(j)}\big) = E_c + \sum_{i=1}^{k} MJ(t_i, m^{(j)}_i), \tag{S4}$$

where $MJ(t_i, m^{(j)}_i)$ represents the pairwise interaction between amino acids of TCR t and peptide $m^{(j)}$ at position i. A T-cell must survive both positive and negative selection to emerge unscathed from the thymus. In the Chakraborty model, positive and negative selection both take place on the same collection of self-peptides $\mathscr{m}$. Positive selection occurs if $E\big(t, m^{(j)}\big) \geq E_p$, for at least one $m^{(j)}$ in $\mathscr{m}$, whereas negative selection is avoided if $E\big(t, m^{(j)}\big) < E_n$, for every $m^{(j)}$ in $\mathscr{m}$. Using this framework, Chakraborty and colleagues were able to formulate selection as an extreme value problem and provided quantitative estimates on the amino acid compositions of those TCRs surviving selection, as well as the number of residues important for T-cell recognition [1–4].

S4 Sequential MJ (S-MJ) Model

S4.1 Model Specification

In this section we present our modification of the Chakraborty MJ model. The main difference is that in our model selection occurs sequentially with positive selection followed by negative selection, to better agree with the biology. Developing thymocytes must pass two separate selection processes in order to become a functional T-cell [6,7]. These thymocytes first pass through the outer thymic cortex where they must receive a sufficient survival signal from cortical self-p-MHC molecules (positive selection). Surviving thymocytes then migrate through the thymic medulla. Binding any medullary self-p-MHC molecule with too high an affinity results in negative selection. The event that a particular TCR survives selection against self-p-MHC is independent from that of other TCRs.

To this end, we partition $\mathscr{m}$ into (disjoint) subsets, $\mathscr{P}$ and $\mathscr{N}$, that represent the collection of positively selecting thymic cortical and negatively selecting medullary peptides, respectively. Formally, $\mathscr{P} = \left\{ \big(p_i^{(j)}\big)_{i=1}^k, j = 1, ..., N_p \right\}$ and $\mathscr{N} = \left\{ \big(q_i^{(j)}\big)_{i=1}^k, j = 1, ..., N_n \right\}$ such that $p_i^{(j)}, q_i^{(j)}$ are IID according to pmf P_h, $\mathscr{P} \dot{\cup} \mathscr{N} = \mathscr{m}$, and typically we will take $N_p = N_n = 10^4$. In this way, positive and negative selection due to each $\mathscr{m}$ are independent of one another. In a similar manner as before, positive selection occurs if $E\big(t, p^{(j)}\big) > E_p$, for at least one $p^{(j)}$ in $\mathscr{P}$, while negative selection is avoided if $E\big(t, q^{(j)}\big) \leq E_n$, for every $q^{(j)}$ in $\mathscr{N}$. In order to maintain contact with prior work, for the S-MJ model, we construct self-peptides and TCRs with amino acids selected according to P_h, the pmf of amino acids in the human proteome (Table S2).

S4.2 Selection Curve

Here we derive the probability that a given TCR t survives both positive and negative selection, referred to subsequently as the selection curve. Let $\mathcal{P}_t$ denote the event that thymocyte $t = \{t_1, t_2, ..., t_k\}$ survives positive selection, and similarly let $\mathcal{N}_t$ denote the event that t survives negative selection. In set-theoretic notation, these events are given by

$$\begin{aligned} \mathcal{N}_t &= \bigcap_{j=1}^{N_n} \left[E\big(t, q^{(j)}\big) \leq E_n \;\middle|\; t = \{t_i\}_{i=1}^k \right], & \mathcal{P}_t &= \bigcup_{j=1}^{N_p} \left[E\big(t, p^{(j)}\big) > E_p \;\middle|\; t = \{t_i\}_{i=1}^k \right], \\ &= \bigcap_{j=1}^{N_n} \left[\sum_{i=1}^k MJ\big(t_i, q_i^{(j)}\big) \leq E_n - E_c \;\middle|\; \{t_i\}_{i=1}^k \right]; & &= \bigcup_{j=1}^{N_p} \left[\sum_{i=1}^k MJ\big(t_i, p_i^{(j)}\big) > E_p - E_c \;\middle|\; \{t_i\}_{i=1}^k \right]. \end{aligned}$$

The key to the analysis is the approximation of the set of interaction energies by an appropriate normal distribution. We will thus have much need of the standard normal cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{S5}$$

Given the heterogeneity of the interaction strengths of the $|A| = 20$ different amino acids, we introduce here the sample mean and variance of the interactions of each, given by

$$\mu_\alpha = \sum_{\beta=1}^{|A|} P_h(\beta) MJ(\alpha, \beta); \qquad \sigma_\alpha^2 = \sum_{\beta=1}^{|A|} P_h(\beta) (MJ(\alpha, \beta))^2 - \mu_\alpha^2, \tag{S6}$$

We first consider the distribution of the interaction energy for a fixed t. Since each t is composed of a particular set of amino acids, $\{t_1, ..., t_k\}$, for a random choice of q_i the $MJ\big(t_i, q_i^{(j)}\big)$ are independent (but not identically distributed) random variables. It can be shown via the Lindeberg-Feller Central Limit Theorem by use of a triangular array with cutoff $Z = 2||f||_\infty^2/\epsilon^2\sigma_{\min}^2$ that sums of the form $\sum_{i=1}^k MJ\big(t_i, q_i^{(j)}\big)$, shifted and scaled by their mean and variance, given by

$$\mu_t \equiv \sum_{j=1}^{k} \mu_{t_i}; \qquad \sigma_t^2 \equiv \sum_{i=1}^{k} \sigma_{t_i}^2. \tag{S7}$$

satisfy Lindeberg's condition. They therefore converge in distribution in the limit of large k to a standard Normal. Here $\sigma_{\min}^2 \equiv \min_{i,j\in A} \operatorname{Var}\big(f(i,j)\big) > 0$ and $||f||_\infty \equiv \max_{i,j\in A} f(i,j)$.

Even for finite k, we observe good agreement between normal approximations and simulations of energies in the specific case of interest ($k = 10$). An example of this convergence is given by Fig. S1A where we have varied k and satisfactory convergence to the normal is already achieved for $k \geq 5$. In each $k = 1, \ldots, 5$ case, normalized cumulative distribution functions are given with respect to the probability space of all k peptide sequences: $(\Omega, \mathcal{F}, \mathbb{P}) = (A^k, 2^{A^k}, P_h^k)$. Fig. S1A demonstrates good agreement for a representative choice of t, and we also observe close agreement over many choices of t. Specifically, $||F^*(x) - \Phi(x)||_\infty < 0.05$ for over 10^3 selections of randomly selected t where F^* is the empirical CDF of the mean-shifted, variance-scaled data. Thus, for our value of $k = 10$, we see that indeed $\sum_{i=1}^k MJ\big(t_i, q_i^{(j)}\big)$ may be approximated by a continuous random variable X_t with distribution $\mathcal{N}(\mu_t, \sigma_t^2)$. From this we can derive an approximate formula for the probability of a given T-cell surviving negative selection,

$$\mathbb{P}(\mathcal{N}_t) = \mathbb{P}\left(\bigcap_{j=1}^{N_n} \left[\sum_{i=1}^{k} MJ\big(t_i, q_i^{(j)}\big) \leq E_n - E_c \;\middle|\; \{t_i\}_{i=1}^k \right] \right),$$

which by virtue of the fact that each $q_i^{(j)}$ is independent and identically distributed equals

$$\begin{aligned} &= \mathbb{P}\left(\sum_{i=1}^{k} MJ(t_i, q_i) \leq E_n - E_c \;\middle|\; \{t_i\}_{i=1}^k \right)^{N_n}, \\ &\approx \mathbb{P}\big(X_t \leq E_n - E_c\big)^{N_n}, \\ &= \Phi\left(\frac{E_n - E_c - \mu_t}{\sigma_t} \right)^{N_n}. \end{aligned} \tag{S8}$$

where the last line employs the Gaussian approximation. We may similarly consider positive selection,

$$\mathbb{P}(\mathcal{P}_t) = \mathbb{P}\left(\bigcup_{k=1}^{N_p} \left[\sum_{i=1}^{k} MJ\big(t_i, p_i^{(j)}\big) > E_p - E_c \;\middle|\; \{t_i\}_{i=1}^k \right] \right),$$

recognizing that the complement of surviving positive selection is an analogous event to negative selection,

$$\begin{aligned} &= 1 - \mathbb{P}\left(\sum_{i=1}^{k} MJ(t_i, p_i) \leq E_p - E_c \;\middle|\; \{t_i\}_{i=1}^k \right)^{N_p}, \\ &\approx 1 - \mathbb{P}\big(X_t \leq E_p - E_c\big)^{N_p}, \\ &\approx 1 - \Phi\left(\frac{E_p - E_c - \mu_t}{\sigma_t} \right)^{N_p}. \end{aligned} \tag{S9}$$

For large $N \equiv N_n = N_p$, these are approximately Gumbel distributions. It is useful to utilize a fairly accurate formula for these asymptotic distributions. We can accomplish this as follows. In general, for large x, N:

$$\begin{aligned}[\Phi(x)]^N &\approx \left(1 - \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}\right)^N \\ &\approx \exp\left(-N \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}\right)\end{aligned} \tag{S10}$$

Introducing $y \equiv x - x_M$, where

$$x_M \equiv \sqrt{2\ln(N/N_0)}, \tag{S11}$$

and N_0 is an N-dependent constant to be specified later, and assuming $y \ll 1$, we have

$$\begin{aligned}[\Phi(x)]^N &\approx \exp\left(-N \frac{1}{\sqrt{4\pi\ln(N/N_0)}} e^{-\left[(2\ln(N/N_0)+2y\sqrt{2\ln(N/N_0)}\right]/2}\right) \\ &\approx \exp\left(-N_0 \frac{1}{\sqrt{4\pi\ln(N/N_0)}} e^{-y\sqrt{2\ln(N/N_0))}}\right)\end{aligned} \tag{S12}$$

If we choose $N_0(N)$ to satisfy the implicit equation

$$N_0^2 = 4\pi\ln(N/N_0) \tag{S13}$$

we then have

$$[\Phi(x)]^N \approx \exp\left(-e^{-(x-x_M)\sqrt{2\ln(N/N_0))}}\right) \tag{S14}$$

which is precisely the CDF of the Gumbel distribution with mode x_M and scale parameter $1/\sqrt{2\ln(N/N_0)}$. Thus, we have

$$\begin{aligned}\mathbb{P}(\mathcal{N}_t) &\approx e^{-e^{-(E_n - E^t_{N_n})/W_{N_n}}} \\ \mathbb{P}(\mathcal{P}_t) &\approx 1 - e^{-e^{-(E_p - E^t_{N_p})/W_{N_p}}}\end{aligned} \tag{S15}$$

where

$$E^t_N = E_c + \mu_t + \sigma_t\sqrt{2\ln(N/N_0(N))}; \qquad W_N = \sigma_t/\sqrt{2\ln(N/N_0(N))}. \tag{S16}$$

Roughly speaking, for $N_n, N_p \gg 1$, the probability of surviving negative selection is a step function $\theta(E_n - E^t_N)$, while the probability of surviving positive selection is approximately $\theta(E^t_N - E_p)$. By virtue of the independence of $\mathcal{P}_t$ and $\mathcal{N}_t$, the probability of thymocyte t surviving both positive and negative thymic selection as a function of E_c is approximated by

$$\begin{aligned}p_s(t) &\equiv \mathbb{P}\big(\mathcal{N}_t \cap \mathcal{P}_t\big) \\ &\approx \Phi\bigg(\frac{E_n - E_c - \mu_t}{\sigma_t}\bigg)^{N_n}\left[1 - \Phi\bigg(\frac{E_p - E_c - \mu_t}{\sigma_t}\bigg)^{N_p}\right] &\text{(S17a)} \\ &\approx \theta(E_p < E^t_N < E_n) = \theta\left(E_p - \mu_t - \frac{\sigma_t^2}{W_N} < E_c < E_n - \mu_t - \frac{\sigma_t^2}{W_N}\right). &\text{(S17b)}\end{aligned}$$

The overall survival probability, then, as a function of E_c is basically a hat function of width $E_n - E_p$, independent of N, whose center moves to lower values as N increases. We have good agreement between Eq. S17a for relevant choices of N (Fig. S1B).

The previous calculation was conditioned on a specific T-cell, t, and we saw that the probability of surviving selection depending on its values of μ_t and σ_t derived from that T-cell's amino acid content. We now wish to calculate the survival probability for a *randomly chosen* developing thymocyte, where we do not know μ_t and σ_t. In theory, this is straightforward given the joint probability distribution of μ_t and σ_t.

We first consider just the marginal distribution of μ_t itself, which as defined in Eq. S7 corresponds to a sum of $k = 10$ IID random variables μ_{t_i}, since each t_i is IID. We can then estimate the distribution of μ_t by the classical Central Limit Theorem,

$$\frac{\sum_{i=1}^{k} \mu_{t_i} - k\mathbb{E}(\mu_{t_i})}{\sqrt{k\mathrm{Var}(\mu_{t_i})}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \quad \text{as} \quad k \to \infty, \tag{S18}$$

From this, we may well approximate the μ_t as a random variable using the following distributions (Fig. S2A):

$$\mu_t = \sum_{i=1}^{k} \mu_{t_i} \sim \mathcal{N}\Big(\bar{\mu}, \overline{\sigma^2}\Big); \qquad \bar{\mu} = k\mathbb{E}(\mu_{t_i}) \approx 29.57, \quad \overline{\sigma^2} = k\mathrm{Var}(\mu_{t_i}) \approx 10.99.$$

The accuracy of this formula is shown in Figs S2A. Turning now to σ_t, it turns out, as can be seen in Fig. S2B, that its primary variation is due to its correlation with μ_t through their joint dependence on t. We may approximate σ_t by its mean value. For P_h, the dependence of the mean value of σ_t on μ_t is found to be approximately linear:

$$\sigma_t \approx a_0 + a_1\mu_t; \qquad a_0 = 1.64, \quad a_1 = 0.0608 \tag{S19}$$

as shown in the figure.

Let us define E^t_{max} as the maximum value of E expected in the full peptide ensemble for a fixed TCR. Eq. S19 implies that E^t_{max} to good approximation depends only on μ_t, and in a linear fashion. This prediction is tested in Fig. S3, which shows a scatterplot of E^t_{max} versus μ_t for 1000 TCRs. The linear regression to the data is essentially identical to the predicted linear relation

$$E^t_{max} \approx E_c + a_0\sqrt{2\ln(N/N_0)} + \mu_t(1 + a_1\sqrt{2\ln(N/N_0)}) \tag{S20}$$

Using the linear relation between σ_t and μ_t, we have $p_s(t) = p_s(\mu_t)$, and

$$\mathbb{P}(\mathcal{P} \cap \mathcal{N}) \approx \int_{-a_0/a_1+\epsilon}^{\infty} p_s(\mu_t) f_\mu(\mu_t) d\mu_t, \tag{S21}$$

where $f_\mu(\mu_t)$ is the pdf of μ_t, and the integral is truncated so that the variance is strictly positive. Thus the unconditional survival of an average TCR may be approximated by:

$$p_s \approx \int_{-a_0/a_1+\epsilon}^{\infty} \Phi\left(\frac{E_n - E_c - \mu_t}{\sqrt{a_0 + a_1\mu_t}}\right)^{N_n} \left[1 - \Phi\left(\frac{E_p - E_c - \mu_t}{\sqrt{a_0 + a_1\mu_t}}\right)^{N_p}\right] \frac{e^{(\mu_t - \bar{\mu})^2/2\overline{\sigma^2}}}{\sqrt{2\pi\overline{\sigma^2}}} d\mu_t. \tag{S22}$$

As can be seen in the figure, this integral evaluated numerically gives us a very good approximation to simulations involving various choices of $N_n = N_p$ (Fig. S2C).

One can obtain an approximately equivalent result using our step function approximation (Eq. S17b). Thus the unconditional survival of an average TCR, with the Gaussian approximation for the μ_t distribution is given by:

$$p_s \approx \int \theta\left(E_p - \mu_t - \frac{\sigma_t^2}{W_N} < E_c < E_n - \mu_t - \frac{\sigma_t^2}{W_N}\right) f_\mu(\mu_t) d\mu_t$$
$$\approx \Phi\left(\frac{\mu_n - \bar{\mu}}{\sqrt{\overline{\sigma^2}}}\right) - \Phi\left(\frac{\mu_p - \bar{\mu}}{\sqrt{\overline{\sigma^2}}}\right), \tag{S23}$$

where

$$\mu_{p,n} = \frac{E_{p,n} - E_c - a_0\sqrt{2\ln(N_{p,n}/N_0(N_{p,n})}}{1 + a_1\sqrt{2\ln(N_{p,n}/N_0(N_{p,n})}}. \tag{S24}$$

This agrees with Eq. S22 when the number of selecting peptide is large, showing that is reasonable for this model to approximate Gumbel distributions as essentially step-like when $N_p, N_n \gg 1$.

S4.3 Potency

Here, we inquire as to the potency of the various self-peptides in the negative selection process. At present, we have considered survival probabilities under the S-MJ model. It is also important to characterize the relative contributions of individual self-peptides to the selection of mature T-cells. For example, some peptides may influence negative selection behavior to a greater degree than others. We say that these self-peptides are more potent. This analysis is independent of positive selection effects. Just as μ_t governs the percentage of peptides that t recognizes, so μ_q, the sample mean of μ_{q_i}, governs the percentage of TCRs that recognize a given peptide. Arguing similarly to the above, the fraction of TCRs that recognize q is

$$r_q \approx 1 - \Phi\left(\frac{E_n - E_c - \mu_q}{\sigma_q}\right). \tag{S25}$$

The most potent peptide clearly is the one with the largest μ_q. Plugging in the actual value of $\mu_q^{max} = 42.77$ and its $\sigma_q = 4.09$, we get, using the actual value of $E_n - E_c = 43.51$ for 50% selection, that $r_q^{max} = 0.428$, extremely close to its actual value in our simulation (with $N_t = 10^5$, $N_n = 10^4$) of 0.431. Thus, we see that the most potent peptide is responsible for roughly 86% of all TCRs eliminated by the negatively selection.

Since μ_q has the same statistics as μ_t, within our Gaussian approximation it is distributed as a Gaussian with mean $\bar{\mu}$, and variance $\overline{\sigma^2}$, and so we can estimate the maximum potency analytically. The maximum of μ_q over the N_n peptides has expected value

$$\mu_q^{max} \approx \bar{\mu} + \sqrt{\overline{\sigma}^2} L \tag{S26}$$

where $L = \sqrt{2\ln(N_n/N_0(N_n))}$. For our value of N_n this gives the result $\mu_q^{max} = 42.0$, comparing well to the simulation result of 42.77 noted above. Plugging this into Eq. (S25), we have

$$r_q^{max} \approx 1 - \Phi\left(\frac{E_n - E_c - \bar{\mu} - \sqrt{\overline{\sigma^2}}L}{a_0 + a_1\mu_q^{max}}\right) \tag{S27}$$

If E_n is chosen so as to give 50% negative selection, then by Eq. (S23) we have

$$E_n - E_c = \bar{\mu}(1 + a_1 L) + a_0 L \tag{S28}$$

and so putting this together

$$r_q^{max} \approx 1 - \Phi\left(L\frac{a_0 + a_1\bar{\mu} - \sqrt{\overline{\sigma^2}}}{a_0 + a_1(\bar{\mu} + \sqrt{\overline{\sigma^2}}L)}\right). \tag{S29}$$

Note that $\sigma = a_0 + a_1\bar{\mu}$ corresponds to the average-strength peptide, so it is very close to $\sqrt{\sigma^2}$. Thus the argument of $\Phi(\cdot)$ is essentially zero, giving a value of r_q^{max} close to 1/2 for all N_n. For example, for $N_n = 10^4$, the formula gives a value of $r_q^{max} = 0.456$. This is fairly close what is found by simulation, 0.431, as reported above, and indicates that the most potent peptide by itself accounts for essentially all the negative selection in the S-MJ model. This also proves that this anomalous behavior is N_n-independent, as claimed in the main text.

S5 Position-Independent Random Affinity (PIRA)

S5.1 Model Definition

In this section, we define and analyze an alternate model, motivated by the unrealistic potency spectrum displayed by the S-MJ model. This model is built on a more randomized T-cell peptide interaction. The PIRA model is specified as follows: The representation of pMHC remains unchanged from the S-MJ model with peptides of length k. In this case, however, amino acids are chosen randomly based on an IID probability distribution, since in any case they are statistically equivalent. We focus only on negative selection by N_n medullary thymic self-peptides and so $\mathscr{m} = \mathscr{N} = \left\{q^{(j)} = \left(q_i^{(j)}\right)_{i=1}^{j}, j = 1, ..., N_n\right\}$. We can set $E_c = 0$ (this is true for the following model as well) as it is only relative interaction strengths that are important.

The T-cell is no longer thought of as a specific peptide sequence, but as a set of interaction grooves that bind specific peptide amino acid with varying affinities. The interaction between each TCR binding groove and amino acid are taken to be IID standard Gaussian random variables. In this model, TCRs interact with amino acids in a position-independent manner. For example, an alanine at position 7 in p-MHC contributes to the total interaction in an identical way as an alanine at position 2. In other words, each TCR is represented by a $|A| = 20$-dimensional vector of IID interaction values, describing interactions with each amino acid type. Thus, $\tau = \left\{t^{(j)}\right\}, j = 1, 2, \ldots, N_t$ and a t is characterized by $X^t = \{X_a^t\}, a = 1, 2, \ldots, |A|$, where X_i^t is the randomly generated interaction between TCR t and amino acid a.

For this model, Eq. S1 thus becomes:

$$E(M, m, T, t) = E(X, q) = f_2(q, X^t) = \sum_{i=1}^{k} X_{q_i}^t. \tag{S30}$$

E_n as before represents the negative interaction threshold, so that negative selection of t is avoided if $E\left(X^t, q^{(k)}\right) \leq E_n$, for every $q^{(k)}$ in $\mathscr{N}$.

S5.2 Selection Curve

As with the S-MJ model, we wish to characterize the TCR recognition percentage as a function of E_n as well as the self-peptide potency. The key to analyzing the PIRA model is to account for the different multiplicities of the amino acids comprising a self-peptide interaction region. For example, if a self-peptide had alanine in all slots, the energy of interaction would be $10E_{Ala}$. Since E_{Ala} has unit variance, the variance of the interaction strength between this self-peptide over the set of TCRs would be 100, which is vastly greater than the variance of 10 for the interaction of a self-peptide with no repeated amino acids with the set of TCRs. Therefore, we first partition the set of self-peptides into subgroups, each with a different pattern of repeats. One class is self-peptides with no repeats, another one with a single amino acid repeated 10 times, another with a two different amino acids each repeated once and the six others unique, and so on. The number of possible classes is the number of partitions of 20, the number of different ways that 20 can be constructed as a sum of natural numbers. An explicit listing of these yields a total of 115975 partitions.

Clearly we will not realize all of these classes in a sample of $N_n = 10^4$ self-peptides. One can explicitly calculate the probability of realizing any particular class, and therefore the expected number of representatives of each class in our set of N_n self-peptides. It turns out that only 20 classes have an expected number of representatives greater than 1/2. We assume that these are in fact the only classes represented in our sample. The most highly represented class is that with six singletons and two doubletons, with 3125 expected representatives. The pure singleton class, in contrast, is expected to have 655 representatives, and is the fifth most highly represented class.

The problem of calculating the distribution of the maximal interaction strength between any given TCR, t, and a particular class, C, of self-peptides proceeds along the line of our calculation of the distribution of maximal interaction energy between a TCR and the entire set of peptides in the S-MJ model. The total interaction energy is the sum of a random subset (with repeat draws) of the X^t_i's. This distribution, a sum of random variables, is in general approximately Gaussian. Employing a Gaussian approximation, the mean is just $10\mu_t$, where μ_t is the mean of the X^t_a. The variance is a little more complicated. If the numbers of singletons, doubletons, tripletons, n-tons is denoted d_n, then the variance is

$$\mathrm{Var}_{t,C} \equiv \mathrm{Var}_{q_C}\big(E(X^t, q_C)\big) = \sigma_t^2 \left(\sum_{n=1}^{10} d_n n^2 - 5 \right) \equiv \sigma_t^2 \sigma_C^2 \tag{S31}$$

Here

$$\sigma_t^2 = \frac{|A|}{|A|-1} \left(\frac{1}{|A|} \sum_{a=1}^{|A|} (X_a^t)^2 - \mu_t^2 \right) \tag{S32}$$

is the sample variance of the X^t_a. The last, -5 term in the above is due to the fact that the energies between different peptides are correlated, as they all depend on the same set of $|A| = 20$ X^t_a. In the general case of a length k interaction region and a set of $|A|$ amino acids, the last term would read $k^2/|A|$, instead of 5.

To get some feeling as to where this last formula comes from, we can work out a specific example. Imagine we are dealing with the class of peptides continuing one amino acid repeated five times, one repeated three times, are two singletons and furthermore let is assume an amino acid alphabet of size 5. Let us assume our TCR has values $\{X_a\}, a = 1, 5$. If we focus on the term in $< E^2 >$ that depends on X_1^2, we get a total of 60 = 5x4x3 peptides and a contribution

$$\frac{1}{60} \times 12 \times (25 + 9 + 1 + 1) = 36/5$$

Likewise the contribution from $< E >^2$ is

$$\frac{1}{60}^2 (12^2 \times 10^2) = 4$$

Note that here the second term is proportional to an extra factor of k^2/A and the first term the factor $\sum_n d_n n^2$ Subtracting gives 16/5. Comparing this to the equations above, the variance is predicted to be 16 multiplied by σ_t^2. The term in this last factor that depends on X_1^2 equals

$$5/4 \times (1/5 - 1/25) = 1/5$$

Given all this, the distribution of the maximum interaction strength for the various amino acids in a given class is Gumbel, and at least for the well-represented classes, can be taken as a delta-function at

$$E^{t,C}_{max} \approx 10\mu_t + \sqrt{2\sigma_t^2 \sigma_C^2 \ln N/N_0} \tag{S33}$$

As a test of this, we present in Fig. S4 a graph of the value of $E^{t,C}_{max}$ averaged over 10^4 TCRs for the twenty classes present in our sample of 10^4 self-peptides. We see that Eq. (S33) works quite well in following the variation from class to class.

Given the maximum interaction of a given TCR t with a given class, the maximum binding energy over all classes is just the maximum of Eq. (S33) taken over all classes. We see from Fig. S4 that this is the E_{max} of the fourth most populous class, consisting of one tripleton, one doubleton and five singletons, with an expected number of 961 representatives. For this class, by Eq. (S31), $\sigma_C^2 = 13$. This gives us a prediction of E^t_{max} of

$$E^t_{max} = 10\mu_t + \sqrt{2 \cdot 13\sigma_t^2 \ln 961/N_0(961)} = 10\mu_t + 11.2\sigma_t \tag{S34}$$

In Fig. S5, we present a scatterplot of E^t_{max} over our sample of 10^4 self-peptides, for a set of 1000 randomly generated TCRs, against the prediction Eq. (S34). The line $y = x$, indicating a perfect prediction, is also plotted, to ease comparison.

For the PIRA model, the width of the Gumbel distribution governing E^t_{max} for a specific TCR is narrower than the width of sample parameters μ_t and σ_t; hence the former can be treated as precise prediction. Using this idea, from $E^t_{max}(\mu_t, \sigma_t)$, we can obtain a rather rough estimate of the selection curve. The main idea as described above is clear from examining Fig. S5. Since the scatter parallel to the theory line is so much greater than that in the perpendicular direction, we may ignore the latter. The former scatter is due to the variation in μ_t and σ_t from TCR to TCR. Since for Gaussian random variables, the sample variance is uncorrelated with the sample mean, we can examine the two sources of variation separately. The mean, μ_t, is Gaussian distributed with mean 0 and standard deviation $1/\sqrt{|A|}$. The variance, σ_t^2 is distributed according to a chi-squared distribution with mean 1 and variance $2/(|A| - 1)$. Thus σ_t has mean 1 and standard deviation $\sqrt{2/(|A| - 1)} = 0.31$. If we approximate the chi-squared distribution by a Gaussian, the E^t_{max} is Gaussian distributed with mean 11.2 and variance $100 * (1/20) + 11.2^2 * (0.31^2) = 17.05$, in which case

$$p_s \approx \Phi\left(\frac{E_n - 11.2}{4.1}\right) \tag{S35}$$

This formula is graphed in Fig. S6, together with the result of a simulation with 10^4 TCRs and 10^4 self-peptides. We see that the agreement is not that quantitatively accurate. There are a few reasons for this. For classes such as class #4, the entire behavior of the energy is determined by just six numbers; hence the Gaussian approximation for the sample mean distribution may not be that accurate. Perhaps more crucially, there is a spread in values for E^t_{max} arising from the class dependence of the expression S33 and simulations show there is a fairly wide distribution of classes for the peptide with maximal binding energy.

S5.3 Potency

We now turn to the problem of the selective potency of the various self-peptides. For this, we have to turn the calculation around, and consider the energy of a random TCR with a given peptide. This of course depends on the class of the peptide, and is a Gaussian random variable with variance

$$\tilde{\sigma}_C^2 = \sum_{n=1}^{k} n^2 d_n = \sigma_C^2 + 5 \tag{S36}$$

The expected number of TCRs that recognize a peptide of class C is then

$$N_C = N_{TCR}\big(1 - \Phi(E_n/\tilde{\sigma}_C)\big) \tag{S37}$$

This is a decreasing function of $\tilde{\sigma}_C$, so the most potent peptide in our sample is clearly that with the largest $\tilde{\sigma}_C^2$. For our $N_n = 10^4$ sample size, this is most likely to be a peptide with three singletons, one doubleton and one 5-ton, giving $\tilde{\sigma}_C^2 = 32$. For $E_n = 11.195$, which as we saw gives 50% selection, this amounts to 2% of the TCRs. This is a vast improvement over the S-MJ model, but it still implies that of order 100 peptides are responsible for the vast majority of the selection. In more detail, since the most potent peptide only removes 2.4% of the TCRs, it does not significantly impact the statistics of the TCR population. Thus, we can calculate the additional impact of the next most potent peptide, which is most likely to also be of the same class. Taking account of the reduction of N_{TCR} due to the most potent peptide, this peptide removes 2.3% of the remaining TCRs. Continuing in this fashion, we can estimate the potency of the other (most likely 3) members of this class, and proceed to the next most potent class, with $\tilde{\sigma}_C^2 = 30$. The result of this process is presented in Fig. S7, together with the result of simulation. We see that the theory works reasonably well as long as the cumulative percentage of negatively selected TCRs is sufficiently small such that the statistical properties of the remaining TCRs is unchanged. In the PIRA model, the 10 most potent peptides are responsible for rejecting 40% of those ultimately selected out, and the top 125 peptides for rejecting 80%. These numbers are still unreasonably small, and motivate the choice of our third, RICE, model.

S6 Random Interaction between Cell receptor and Epitope (RICE)

S6.1 Model Definition

In this section, we define and analyze our third model. Underlying this model is the observation that it is unrealistic to expect the TCR interaction strength to be dependent only on the numbers of different amino acids in the peptide and not at least to some extent on the location. We therefore introduce the RICE model, fashioned to some extent after the random energy model in molecular biophysics. This last model is specified as follows: The pMHC representation remains unchanged from the S-MJ and PIRA models with peptides of length $k = 10$. As in PIRA, we focus only on negative selection by N_n medullary thymic self-peptides and so $\mathscr{m} = \mathcal{N} = \{q^{(j)} = (q_i^{(j)})_{i=1}^{j}, j = 1, ..., N_n\}$. Again, the set $\mathcal{T}$ is conceptualized as a set of binding grooves and the interaction strength between each TCR binding groove and amino acid are comprised of IID random variables. In this model, TCRs interact with amino acids in a position-dependent manner. For example, an alanine at position 7 in p-MHC does not necessarily have the same contribution to total interaction as an alanine at position 2. $\tau = \{t^{(j)}\}, j = 1, \ldots, N_t$. Hence, each t is characterized by $k|A|$ random variables $X_{i,j}^t, i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, |A|$. Here, $X_{i,j}^t$ represents the interaction between TCR t and an amino acid j located at position i.

For fixed peptide q and TCR X, our model will assume that

$$E(T, M, t, m) = E(X, q) = f_2(X, q) = \sum_{i=1}^{k} X_{i,q_i}, \tag{S38}$$

As before, negative selection is avoided if $E\big(X, q^{(j)}\big) \leq E_n$ for every $q^{(j)}$ in $\mathcal{N}$. We typically assume as in PIRA that each of the X's are distributed as a standard mean zero and unit variance Gaussian. We will also consider the alternate assumption that X is distributed uniformly between zero and one. It will turn out that none of the important findings are sensitive to this change.

S6.2 Selection Curve

Let us start with the uniform distribution assumption. We approach this formulation as before by considering selection survival and probabilities taking $X_{i,j}$ as Uniform$[0, 1]$ random variables, representing individual amino acid interaction contributions falling between minimal (0) and maximal (1) values. Gaussian random variables will be used as a convenient approximation for parallel analysis. To this end, we let $X = X_{i,j}$, $i = 1, 2, ..., k,\ j = 1, 2, ..., |A|$ represent a (random) TCR, $\{q^{(j)}\}_{j=1}^{N_n}$ a random collection of negative self-peptides each having length k, and $E(X^t, q)$ the total binding interaction between TCR t and peptide q. We note that $E(X,q) = \sum_{i=1}^{k} X_{i,q_i} \sim$ Irwin-Hall(k), since X_{i,q_i} are IID Uniform$[0,1]$ random variables [8, 9]. The pdf, f_k, and cdf, F_k, of X_{i,q_i}, and hence $E(X,q)$, are given by:

$$f_k(x) = \frac{1}{(j-1)!}\sum_{j=0}^{\lfloor x \rfloor}(-1)^j \binom{k}{j}(x-j)^{k-1}, \qquad F_k(x) = \frac{1}{k!}\sum_{j=0}^{\lfloor x \rfloor}(-1)^j \binom{k}{j}(x-j)^{k}. \tag{S39}$$

Plots of F_k and f_k are provided for $k = 1, 2, \ldots, 10$ (Fig. S8). Thus, survival can be approximated from below by assuming affects from each self-peptide are approximately independent by:

$$\begin{aligned} p_s &= \mathbb{P}\Big(\bigcap_{j=1}^{N_n}\Big[E\big(X, q^{(j)}\big) \leq E_n\Big]\Big) \\ &\approx \prod_{j=1}^{N_n}\mathbb{P}\Big(E\big(X, q^{(j)}\big) \leq E_n\Big) \\ &= F_k\big(E_n\big)^{N_n} \\ &= \left[\frac{1}{k!}\sum_{j=0}^{\lfloor E_n \rfloor}(-1)^j \binom{k}{j}(E_n - j)^k\right]^{N_n}. \end{aligned} \tag{S40}$$

Empirical simulations (Fig. S11A) have been compared to the estimate given by Eq. S40 . We use these results to inform a relevant choice of E_n in subsequent analysis. It is generally agreed upon that survival rates fall between 20% and 70% [10–15]. For 50% survival, E_n is equal to 8.32 in this case, close to the simulated result of $E_n = 8.18$.

The above results rely on the precise distribution for the sum of uniformly distributed random variables. But, because the RICE model constructs energies by summing over $k = 10$ IID random variables, there should be no real difference between our uniform distribution assumption and the baseline Gaussian one. In fact, in the uniform case, F_k above may be approximated by a Gaussian distribution with mean $\mu = k/2$ and variance $\sigma^2 = k/12$; there is a trivial changes in the mean and variance if we instead use a sum of Gaussians centered at zero. Thus, we can proceed to an approximate calculation of the selection curve that would be valid for either formulation.

For the previous two models, we have argued that the Gumbel distribution governing the maximal energy for a given TCR is narrower than the variation in the TCR statistical parameters μ_t and σ_t; hence the former can be ignored and the relationship between these quantities and the maximum energy taken as deterministic. This leads to a selection curve which is given by a Gaussian CDF function. For RICE, the distribution of interaction strengths of different TCRs are narrowly distributed. We can again analyze the dependence on the specific TCR through the dependence on μ_t and σ_t. Here μ_t is the sample mean of all $k|A|$ random variables that characterize t. Similarly, σ_t^2 is the sum over sites i of the sample variances for the $|A|$ different X_{i,q_i}. With these definitions, for the Gaussian model, the energies $E(X_t, q)$ for a given t are Gaussian with mean $k\mu_t$ and variance σ_t^2. μ_t is distributed as a zero-centered Gaussian with width $\sqrt{(|A|-1)/k|A|^2}$, roughly a factor of $1\sqrt{|A|}$ smaller than in PIRA. The quantity σ_t^2 is distributed according to a Chi-squared distribution, with mean $k(|A|-1)/|A|$ and variance $2k(|A|-1)/|A|^2$. Thus, for $N_n = 10^4$, the width of the Gumbel is roughly 1.1. The contribution of the variance of the mean to the variance in E_n is roughly $\sqrt{10/20} = 0.7$, so here it does not swamp the Gumbel width. Similarly, the contribution of the variance in σ_t is roughly $\sqrt{2k/|A|}/\sqrt{k}\sqrt{\ln(10^4/N_0(10^4)} \approx 0.1$.

Thus,to zeroth approximation we can consider μ_t and σ_t as fixed at their mean values, giving us a pure Gumbel extreme value distribution, since this sets the non-trivial width of the survival curve. For 50% survival, E_n is equal to the median of this distribution. Since for the Gumbel distribution the cumulative probability function is $\exp(-(x-\mu)/\beta)$, it is easy to see that the median is located at $x = \mu - \beta\ln(\ln 2)$ and so

$$E_n = \mu + \sigma\sqrt{2\ln(N_n/N_0(N_n))} - \frac{\sigma\ln(\ln 2)}{\sqrt{2\ln(N_n/N_0(N_n))}} \tag{S41}$$

This leads to a prediction of $E_n \simeq 12$ for the Gaussian formulation, close to the simulation result of $E_n = 11.5$. The predicted shape of the survival profile is significantly narrower than for the other two models, since here there is practically no variation in μ_t as there was for the earlier models. The shape is less symmetric, since it is a Gumbel distribution and not a Gaussian. A graph of the Gumbel distribution versus the empirical curve derived by simulation is presented in (Fig. S9A). Also shown is a simulation curve for the unphysical case of $|A| = 80$, showing that the major source of disagreement is the finite widths of μ_t and σ_t arising from the finite size of the amino acid alphabet.

To investigate this further, we can consider two different issues with finite A. One is the aforementioned finite width of the sample mean and variance; when this is taken into account (Fig. S9B) we obtain a curve which is much closer in shape but still slightly off in parameters from the predicted Gumbel distribution. The small residual deviation(about 3%) is possibly due to correlations in the peptides reducing the effective size of N_n but we leave demonstrating this in detail for future work.

S6.3 Potency

The most notable difference between the RICE formulation of TCR binding and the other two is that in RICE almost all self-peptides take part in selection (Fig. 3). Our hypothesis is that this feature reflects more accurately the underlying biology as it seems unlikely that a large majority of generated self-peptides would be extraneous to the selection process. Motivated by this empirical observation, we provide an analysis of the potency rates under this model, using the Gaussian approximation.

For a collection of N_n peptides and under complete independence between TCR interactions with self-peptides, the probability of detection of a single self-peptide, p_r, by a single TCR is given by

$$p_r = 1 - \Phi\left(\frac{E_n - \mu}{\sigma}\right) \tag{S42}$$

Since in order to survive, a given TCR has to go undetected by all N_n peptides, for N_n large, unless the TCR survival probability is to be tiny, p_r must a very small number. Calling the TCR selection survival probability p_s, we have

$$p_s \approx (1-p_r)^{N_n} \approx e^{-N_n p_r} \tag{S43}$$

Due to independence, the number of TCRs, S, recognized by a given peptide is Poisson distributed with mean $N_t p_r$. If the total number of TCRs, $N_t \gg N_n$, then since p_s is a number of order unity, so is $N_n p_r$, and so $N_t p_r \gg 1$. Thus, we may consider the distribution of S to be Gaussian, with mean and variance $N_t p_r$. The most potent peptide is then given by the extreme value of this distribution over the N_n peptides, so that

$$S_{max} \approx N_t p_r + \sqrt{N_t p_r}\sqrt{2\ln(N_n/N_0(N_n))} \tag{S44}$$

For 50% selection, $N_t = 10^5$ and $N_n = 10^4$, this works out to be $S_{max} \approx 17$, a tiny fraction of the total number of peptides, in contrast to the much higher maximum potency in the MJ and PIRA models. This results in vastly different fluctuations in overall selection rates across individuals for the various interaction formulations (Fig. S14). Again, as with PIRA, the small number of TCRs which recognize the most potent peptide means that the statistical properties of the remaining TCRs is unchanged, and we can calculate the marginal potency of the next most potent peptide by rehashing the above calculation with the reduced value of N_t. Formally, for the j^{th} most potent peptide,

$$S_j \approx {N_t}^{j-1} p_r + \sqrt{{N_t}^{j-1} p_r}\sqrt{2\ln((N_n - j)/N_0(N_n - j))}; \qquad {N_t}^j = {N_t}^{j-1} - S_j; \qquad {N_t}^0 = N_t \tag{S45}$$

This approximation works well as long as the cumulative percentage of recognizing TCRs remains small, and is plotted in Fig. S10.

S7 Neoantigen and Foreign Peptide Recognition in RICE

We now proceed to characterize the recognition rates of (point-mutated) neoantigens and foreign peptides. We will first perform this calculation using the exact CDF for the uniform distribution and then afterwards utilize a Gaussian approximation. Let $\tilde{q} = \tilde{q}^{(1)}$ be a point-mutated version of self-peptide $q = q^{(1)}$, mutated at (random) position $i^* \in 1, 2, ..., k$. All mutation positions are equally likely. The probability in question takes the form:

$$\tilde{p} \equiv \mathbb{P}\Big(\big[E(X,\tilde{q}) > E_n\big] \;\Big|\; \bigcap_{k=1}^{N_n} \Big[E\big(X, q^{(k)}\big) \le E_n\Big]\Big).$$

Dependencies involving the interaction of a given TCR t with many peptides that may share common amino acid positions makes estimating the probability of this event nontrivial. We calculate the probability of the simpler event that t recognizes a single mutant self-peptide $\tilde{q}$ conditioned on t passing negative selection with q. That is,

$$\tilde{p}_1 \equiv \mathbb{P}\Big(E(X^t, \tilde{q}) > E_n \;\Big|\; E(X^t, q) \le E_n \Big).$$

Conditioning on the survival of t by selection with self-peptide q is motivated by the fact that q is most closely related to $\tilde{q}$, and therefore explains a significant amount of the dependency of TCR t recognition ability on t's survival in thymic selection. We observe that $\sum_{i=1}^{k} X^t_{i,q_i} = \sum_{i \ne i^*} X^t_{i,q_i} + X^t_{i^*,q_{i^*}}$ may be viewed as a sum of two independent random variables, with $\sum_{i \ne i^*} X^t_{i,q_i}$ having support $[0, k-1]$. It will prove useful to note also that for any k, integrals of $F_k(x-y)$ taken with respect to y over a unit interval are of the form:

$$\begin{aligned}
\int_0^1 F_k(x-y)dy &= \frac{1}{k!}\Big\{ \int_0^a \sum_{j=0}^{\lfloor x-y \rfloor} (-1)^j \binom{k}{j} (x-y-j)^k dy \\
&+ \int_a^1 \sum_{j=0}^{\lfloor x-y \rfloor} (-1)^j \binom{k}{j} (x-y-j)^k dy \Big\}, \quad \text{for } \; a = x - \lfloor x \rfloor, \\
&= \frac{1}{(k+1)!}\Big\{ -\sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{k}{j} \big[(x-a-j)^{k+1} - (x-j)^{k+1}\big] \\
&- \sum_{j=0}^{\lfloor x-1 \rfloor} (-1)^j \binom{k}{j} (x-1-j)^{k+1} + \sum_{j=0}^{\lfloor x \rfloor} (-1)^j \binom{k}{j} (x-a-j)^{k+1} \Big\}, \\
&= G_k(x) - G_k(x-1),
\end{aligned} \tag{S46}$$

where

$$G_k(x) \equiv \int_0^x F_k(y)dy. \tag{S47}$$

A similar argument can be made for higher order integrals of F_k. Set $Y_i \equiv X^t_{i,q_i}$, $\tilde{Y}_i \equiv X^t_{i,\tilde{q}_i}$, $G_k(x) \equiv$

$\int_0^x F_k(y)dy$, $H_k(x) \equiv \int_0^x G_k(y)dy$, and $x \equiv E_n$. Then

$$\begin{aligned}
\tilde{p}_1 &= \mathbb{P}\Big(\sum_{i=1}^{k} X^t_{i,\tilde{q}_i} > E_n \,\Big|\, \sum_{i=1}^{k} X^t_{i,q_i} \le E_n\Big),\\
&= \mathbb{P}\Big(\sum_{i\neq i^*} Y_i + \tilde{Y}_{i^*} > x \,\Big|\, \sum_{i\neq i^*} Y_i + Y_{i^*} \le x\Big),\\
&= \int_0^1\int_0^1 \mathbb{P}\Big(x - \tilde{y}_{i^*} < \sum_{i\neq i^*} Y_i \le x - y_{i^*}\Big) f_{\tilde{Y}_{i^*}}(\tilde{y}_{i^*}) f_{Y_{i^*}}(y_{i^*}) d\tilde{y}_{i^*} dy_{i^*} \Big/ \int_0^1 \mathbb{P}\Big(\sum_{i\neq i^*} Y_i \le x - y_{i^*}\Big) f_{Y_{i^*}}(y_{i^*}) dy_{i^*},\\
&= \Big\{\int_0^1 F_{k-1}\big(x - y_{i^*}\big)\int_{y_{i^*}}^1 d\tilde{y}_{i^*} dy_{i^*} - \int_0^1\int_{y_{i^*}}^1 F_{k-1}\big(x - \tilde{y}_{i^*}\big) d\tilde{y}_{i^*} dy_{i^*}\Big\} \Big/ \int_0^1 F_{k-1}\big(x - y_{i^*}\big) dy_{i^*}.
\end{aligned}$$

By integrating, we obtain

$$\begin{aligned}
\tilde{p}_1 = \Big\{&G_{k-1}\big(x\big) - G_{k-1}\big(x-1\big) + G_{k-1}\big(x-1\big) - \int_0^1 G_{k-1}\big(x - y_{i^*}\big) dy_{i^*}\\
&+ G_{k-1}\big(x-1\big) - \int_0^1 G_{k-1}\big(x - y_{i^*}\big) dy_{i^*}\Big\} \Big/ \Big(G_{k-1}\big(x\big) - G_{k-1}\big(x-1\big)\Big).
\end{aligned}$$

Additional rearranging yields

$$\tilde{p}_1 = 2\Big(\frac{G_{k-1}\big(E_n\big) - H_{k-1}\big(E_n\big) + H_{k-1}\big(E_n - 1\big)}{G_{k-1}\big(E_n\big) - G_{k-1}\big(E_n - 1\big)}\Big) - 1. \qquad \text{(S48)}$$

Eq. S48 is compared with simulations involving various values of N_n (Fig. S11C). We would like to compare this estimate with the probability of X recognizing a randomly generated peptide $\hat{q}$ conditioned on X having no thymic selection pressure. This serves as an analogous simpler event for the case of random peptide interaction since, in contrast to a point-mutated self-peptide, a self-peptide closely related to foreign peptide is quite rare. This leads to

$$\begin{aligned}
\hat{p}_0 &\equiv \mathbb{P}\Big(E(X,\hat{q}) > E_n\Big) \,\big|\, \Omega\Big)\\
&= \mathbb{P}\Big(\sum_{i=1}^{n} X_{i,\hat{q}_i} > E_n\Big)\\
&= 1 - F_k\big(E_n\big) = 1 - \frac{1}{k!}\sum_{j=0}^{\lfloor E_n\rfloor} (-1)^j \binom{k}{j} (E_n - j)^k. \qquad \text{(S49)}
\end{aligned}$$

Eq. S49 is compared with simulations in Fig. S11B. The relative ratio $\tilde{p}_1/\hat{p}_0$ is presented in Fig. S11D. From this we observe a minimum on the relevant domain.

We can prove the existence of a minimum analytically. To this end, we first calculate the *joint* probability, p_J that q is not recognized by t but $\tilde{q}$ is, again expressing the total interaction strength as the sum over the total contribution of the nonmutated sites plus that of the mutated site. We denote the total energy of the nonmutated sites by E_0, the energy of the mutated site, in its original and mutated form, by ϵ and ϵ' respectively. Thus we require that $E_0 + \epsilon < E_n$ and $E_0 + \epsilon' > E_n$, and since $0 \le \epsilon, \epsilon' \le 1$, we must have

$E_n - 1 \leq E_0 \leq E_n$. We thus have

$$p_J = \int_{E_n-1}^{E_n} P_{k-1}(E_0) dE_0 \int_0^{E_n-E_0} P_1(\epsilon) d\epsilon \int_{E_n-E_0}^{1} P_1(\epsilon') d\epsilon'$$
$$= \int_{E_n-1}^{E_n} P_{k-1}(E_0)(E_n - E_0)(1 - E_n + E_0) dE_0 \tag{S50}$$

where P_1 is the pdf for a single site (i.e., Uniform[0,1]) and P_{k-1} is the distribution for the sum of $k-1$ energies. This expression is exact. From this, we can show that $p_J(E_n)$ is symmetric about $k/2$. To accomplish this, we first note that $P_{k-1}(E_0)$ is symmetric about its mean, $\mu_{k-1} = (k-1)/2$. We write $E_n = k/2 + z$, and define $\Delta_0 \equiv E_0 - \mu_{k-1}$. We then have

$$p_J = \int_{z-1/2}^{z+1/2} P_{k-1}(\Delta_0 + \mu_{k-1})(z + 1/2 - \Delta_0)(\Delta_0 - z + 1/2) d\Delta_0$$
$$= \int_{z-1/2}^{z+1/2} P_{k-1}(\Delta_0 + \mu_{k-1})(1/4 - z^2 + 2z\Delta_0 - \Delta_0^2) d\Delta_0 \tag{S51}$$

It is straightforward to show that $\int_{z-1/2}^{z+1/2} f(x)dx$ is even (resp. odd) in z if f is even(resp. odd) in x. It then follows immediately from the fact that $P_{k-1}(\Delta_0 + \mu_{k-1})$ is even in Δ_0 that p_J is even in z, as we wished to show. The conditional probability, p_C, of t recognizing $\tilde{q}$ given that it does not recognize q, is related to p_J by

$$p_C = \frac{p_J}{\int_{E_n}^{k} P_k(E) dE} \tag{S52}$$

$$p_C/\hat{p}_0 = \frac{p_J}{\left(\int_0^{E_n} P_k(E) dE\right)\left(\int_{E_n}^{k} P_k(E) dE\right)} \tag{S53}$$

Now, as we have seen p_J is even about $E_n = k/2$. It is also clear that the denominator has the same property. Thus, $p_C/\hat{p}_0$ is even as well, and so $k/2$ is an extremum.

Using our Gaussian approximation for the $k-1$ energies, we can produce a formula for this ratio. We approximate P_{k-1} in Eq. S51 by a Gaussian with mean μ_{k-1} and width $\sigma_{k-1} = \sqrt{(k-1)/12}$. Then

$$p_J \approx p_J^G = \frac{1}{\sqrt{2\pi\sigma_{k-1}^2}} \int_{E_n-1}^{E_n} e^{-\frac{(E_0-\mu_{k-1})^2}{2\sigma_{k-1}^2}} (E_n - E_0)(1 - E_n + E_0) dE_0$$
$$= \frac{\sigma_{k-1}}{\sqrt{2\pi}} \left[(E_n - \mu_{k-1}) e^{-\frac{(E_n-\mu_{k-1}-1)^2}{2\sigma_{k-1}^2}} - (E_n - \mu_{k-1} - 1) e^{-\frac{(E_n-\mu_{k-1})^2}{2\sigma_{k-1}^2}} \right]$$
$$+ \left((E_n - \mu_{k-1})(E_n - \mu_{k-1} - 1) + \sigma_{k-1}^2\right) \left[\Phi\left(\frac{E_n - \mu_{k-1}}{\sigma_{k-1}}\right) - \Phi\left(\frac{E_n - \mu_{k-1} - 1}{\sigma_{k-1}}\right) \right] \tag{S54}$$

Then

$$p_C/\hat{p}_0 \approx \frac{p_J^G}{\Phi\left(\frac{E_n-\mu_k}{\sigma_k}\right)\left(1 - \Phi\left(\frac{E_n-\mu_k}{\sigma_k}\right)\right)} \tag{S55}$$

Our Gaussian approximation of $p_C/\hat{p}_0$ captures the minimum at $E_n = \mu_k$, where its value is 0.29, which agrees well with our prior estimates of $\tilde{p}_1/\hat{p}_0$ using exact values for $\tilde{p}_1$ (Fig. 3D).

Simulations of $\tilde{p}_1$ and $\hat{p}_0$ are compared alongside these analytical expressions as well as simulations of $\tilde{p}$ and $\tilde{p}$ for larger N_n (Fig. 3B-C). We find that estimates for the $N_n = 1$ case are close to empirical estimates of the $N_n = 10^4$ (see below; simulations for $N_n = 10^4$ averaged over 10 iterations of self-peptide recognized with $N_t = 10^5$, each averaged with 10^4 non-self peptides).

Value	Minimal ($\tilde{p}_1$) or Absent ($\hat{p}_0$) Selection	Physiologic Selection ($N_n = 10^4$)
Point-mutated peptide	$\tilde{p}_1 = 6.6 * 10^{-5}$	$\tilde{p} = 3.4 * 10^{-5}$
Random peptide	$\hat{p}_0 = 9.8 * 10^{-5}$	$\hat{p} = 4.8 * 10^{-5}$

In the sequential MJ model we found that TCR selection behavior depended almost exclusively on a few 'extreme' self-peptides. For the most extreme such peptide, the $N_n = 1$ case selected the same TCRs as the $N_n = 10^4$ case, which is inconsistent with the underlying biology. In the RICE model, nearly all self-peptides participate in selection. If we restrict our attention to selection with E_n chosen to approximate empirical negative selection survival estimates, we find that TCR recognition probabilities is relatively insensitive to the number of self-peptides (i.e. small changes in the number of self-peptides loaded on various MHC molecules would become by itself not influence the degree of TCR recognition ability). Moreover, we find that the recognition rates of point-mutated self-peptides (representing tumor-associated neoantigens) and random peptides (representing foreign antigens or mislocalized/aberrantly-displayed self-peptides) are comparable to one another and on the same order of magnitude for $N_n \in [1, 10^4]$ (Fig. 3D).

S7.1 Optimal Selection

We remark that our choice of E_n was selected based on targeting experimentally-observed negative selection survival rates. From an optimization standpoint, the thymus functions to produce mature T-cells with the ability to effectively recognize foreign threats. The VDJ recombination relies on random generation of TCR sequences to cover epitope space. This, along with cell division and finite resources, places a limit on thymocyte output. However, selection thresholds (E_n), in theory, could be controlled by the organism. An effective adaptive immune system would therefore be able to quickly distinguish foreign threats from self. Prior to thymic migration, this would require the most efficient production of effective thymocytes (those able to survive thymic selection and recognize a random, unknown foreign antigen). This is approximated in the language above by,

$$\begin{aligned}
\mathbb{P}\Big(\big[E(X,\hat{q}) > E_n\big] \cap \bigcap_{j=1}^{N_n} \big[E\big(X, q^{(j)}\big) \le E_n\big]\Big) &= \mathbb{P}\Big(E(X,\hat{q}) > E_n \mid \bigcap_{j=1}^{N_n} E\big(X, q^{(k)}\big) \le E_n\Big) \\
&\quad \cdot \mathbb{P}\Big(\bigcap_{j=1}^{N_n} \big[E\big(X, q^{(j)}\big) \le E_n\big]\Big) \\
&\approx \hat{p} p_s \\
&\approx \hat{p}_0 p_s \\
&= \big[1 - F_k(E_n)\big] F_k(E_n)^{N_n}. \qquad \text{(S56)}
\end{aligned}$$

Let $R(x) \equiv \big[1 - F_k(x)\big] F_k(x)^{N_n}$. This is maximized whenever

$$\begin{aligned}
\frac{\partial R}{\partial x} &= \frac{\partial \hat{p}_0 p_s}{\partial x} \\
&= N_n F_k(x)^{N_n - 1} F_k'(x) - (N_n + 1) F_k(x)^{N_n} F_k'(x) \\
&= f_k(x) F_k(x)^{N_n - 1} \big[N_n - (N_n + 1) F_k(x)\big] = 0.
\end{aligned}$$

$f_k, F_k > 0$ implies that,

$$F_k(x) = \frac{N_n}{N_n + 1},$$

which occurs uniquely in the Uniform[0,1] formulation at

$$E_n^* = F_k^{-1}\left(\frac{N_n}{N_n+1}\right) = 8.197, \quad \text{for } \; N_n = 10^4.$$

In both cases, however, optimal selection rates are predicted to occur at

$$p_s = [F_k(x)]^{N_n} = \left(\frac{N_n}{N_n+1}\right)^{N_n} \approx 1/e. \tag{S57}$$

Thus, selection is optimal in this sense when roughly 1/3 of the TCRs survive selection. At this level of selection, by Eqs. S40 and S43, the recognition probability of a peptide is just $1/N_n$, and the mean number of TCRs recognizing a peptide is precisely N_t/N_n, which is the optimal sharing of the selection burden.

This agrees with and reinforces empirical observations on negative selection rates. We would like to make a statement about the maximizer x_2^* of $\hat{p}p_s$ relative to the maximizer x_1^* of $\hat{p}_0 p_s$, despite not having the explicit expression for $\hat{p}$. All hypotheses in the following claim are derived by considering the results of Fig. S11B for a generous region of the relevant parameter range of interest ($E_n = x \in [7.5, 10]$).

Proposition S1. *Let $\hat{p}_0$, $\hat{p}_s \in [0,1]$ be continuous, nonincreasing functions of x on $[7.5, 10]$, with $\hat{p}_0, \hat{p}_s \to 0$ as $x \to 10^-$ such that there exists some nondecreasing $\alpha(x)$ for which $0 \leq \alpha \leq 1$, $\alpha \uparrow 1$ as $x \to 10^-$, and $\hat{p} = \alpha \hat{p}_0$ on $[7.5, 10]$. Let x_1^* be a global maximizer for $\hat{p}_0 p_s$. Then, there exists a maximizer x_2^* of $\hat{p} p_s$ such that $x_2^* \geq x_1^*$.*

Proof.
Put $g_1 \equiv \partial(\hat{p}_0 p_s)/\partial x$, $g \equiv \partial(\hat{p} p_s)/\partial x$. Then,

$$\begin{aligned} g &= \frac{\partial p_s}{\partial x}\hat{p} + p_s \frac{\partial \hat{p}}{\partial x} \\ &= \alpha \hat{p}_0 \frac{\partial p}{\partial x} + p_s \frac{\partial \alpha \hat{p}_0}{\partial x} \\ &= \alpha \Big(\hat{p}_0 \frac{\partial p_s}{\partial x} + p_s \frac{\partial \hat{p}_0}{\partial x} \Big) + p_s \hat{p}_0 \frac{\partial \alpha}{\partial x} \\ &= \alpha g_1 + \frac{\partial \alpha}{\partial x} p_s \hat{p}_0 \\ &\geq \alpha g_1, \end{aligned}$$

since α is nondecreasing and both $p, \hat{p}_0$ are nonnegative. Therefore,

$$g(x_1^*) \geq \alpha g_1(x_1^*) = 0. \tag{S58}$$

This demonstrates that $\hat{p} p_s$ is increasing at $x = x_1^*$. Moreover, $\hat{p} p_s$ is a nonnegative function of x such that $\hat{p}(x_1^*) p_s(x_1^*) > 0$ and $\hat{p} p \to 0$ as $x \to 10^-$. The existence of $x_2^* \geq x_1^*$ follows from continuity of $\hat{p} p_s$.

□

From this we conclude that if thymic selection is indeed optimized with regard to the generation of 'useful' TCRs, then we may have confidence that the effective cutoff regime (E_n) under physiological conditions is no less than our estimate, in which case we expect differences between $\tilde{p}_1$ (resp. $\hat{p}_0$) and $\tilde{p}$ (resp. $\hat{p}$) are no more than the above estimate (Fig. S11B-C). We remark that the argument above holds for the general features of the Gaussian formulation as well (Fig. 4A).

S8 Allogeneic Response in RICE

We finish with a brief application to the allogeneic response present in the setting of MHC-matched HSCT. Let Y be the number of point-mutated peptides present in an average host cell, where the mutation is described with respect to Donor T-cells. We consider $P_A(Y)$, the (nondecreasing) fraction of alloreactive TCRs for increasing values of Y. TCRs surviving the same selection criteria are identically distributed. If we make an additional approximation that the number of new TCRs that react to peptide y but not to any prior peptides $\{1, 2, \ldots, y-1\}$, is independent of prior peptides, then we may approximate $P_A(Y)$ by

$$P_A(Y) \approx 1 - \big(1 - \tilde{p}\big)^Y, \tag{S59}$$

where we recall from Section S7 that $\tilde{p}$ (resp. $\hat{p}$) represents the probability that a TCR surviving selection recognizes point-mutated self-peptide (resp. random peptide). A similar argument can be made with $\hat{p}$ instead for calculating the fraction of CTLs responding to foreign (e.g. pathogenic) antigens. Analytical predictions of allorecognition percentages from Eq. S59 can be compared against simulations that record the percentage of responding T-cells with increasing differences in presented antigens, either foreign or point-mutant self-peptides (Fig. 4).

Our final goal is to characterize the distribution of Y in order to estimate the fraction of allogeneic TCRs in the setting of MHC-matched (through $\tilde{p}$) and unmatched (through $\hat{p}$) transplant. Recognition of point-mutated self-peptide is relevant to minor histocompatibility differences in the form of SNPs between

MHC-matched host and donor. Assuming that SNPs occur in the genome with frequency 300^{-1} bp^{-1}, we can view the number of SNP differences per 10-mer peptide between two individuals (host and donor) as, $X{\sim}\text{Poisson}(\lambda = N_n/10)$, given that each amino acid is constructed from a trinucleotide codon. Each SNP position provides an (independent) opportunity for the donor to differ from host, and the likelihood of this depends on the underlying frequency of DNA base pairs. Given the relative frequencies of base pairs A,C,G,T in the exome (approximated in this estimate as 1/4) the probability of missense mutations is $p_d \approx 0.6$, calculated directly by considering all possible DNA codons. Therefore, the number of self-peptides that differ conditioned on x SNPs is binomially distributed:

$$[Y|X = x] \sim \text{Binomial}(x, p_d). \tag{S60}$$

Let X be the number of (random) SNP differences between host and donor. Let $G_Y(z)$ (resp. $G_X(z)$) be the probability generating function of Y (resp. X). Then, by definition,

$$G_Y(z) = \mathbb{E}\big[z^Y\big] = \big(1 - p_d + p_d z\big)^X; \qquad G_X(z) = \mathbb{E}\big[z^X\big] = e^{\lambda(z-1)}.$$

Therefore,

$$\begin{aligned} G_Y(z) &= \mathbb{E}\big[\mathbb{E}\big(z^Y|X\big)\big] \\ &= \mathbb{E}\Big[\big(1 - p_d + p_d z\big)^X\Big] \\ &= G_X\big(1 - p_d + p_d z\big) \\ &= e^{\lambda[(1-p_d+p_d z)-1]} \\ &= e^{\lambda p_d (z-1)}. \end{aligned}$$

Therefore $Y{\sim}\text{Poisson}(\lambda p_d)$. From this, we wish to characterize the mean and variance of the fraction of alloreactive T-cells using Eq. S59. Let m_Y be the probability mass function of Y. Then,

$$\begin{aligned} \mathbb{E}\big[P_A(Y)\big] &= \sum_{y=0}^{\infty} P(y) m_Y(y) \\ &= \sum_{y=0}^{\infty} \big(1 - r^y\big)\frac{e^{-\nu}\nu^y}{y!}, \quad \text{for} \quad r \equiv 1 - \tilde{p}, \quad \nu \equiv \lambda p_d, \\ &= \sum_{y=0}^{\infty} \frac{e^{-\nu}\nu^y}{y!} - \frac{e^{-\nu}(r\nu)^y}{y!} \\ &= 1 - \sum_{y=0}^{\infty} e^{-\nu\hat{p}} \frac{e^{-r\nu}(r\nu)^y}{y!} \\ &= 1 - e^{-\lambda p_d \hat{p}}. \end{aligned} \tag{S61}$$

Additionally,

$$\begin{aligned}
\mathbb{E}\big[P_A(Y)^2\big] &= \sum_{y=0}^{\infty} P(y)^2 m_Y(y) \\
&= \sum_{y=0}^{\infty} \big(1 - 2ry + r^{2y}\big) \frac{e^{-\nu}\nu^y}{y!} \\
&= 1 - 2e^{-\nu\tilde{p}} + \sum_{y=0}^{\infty} e^{-\nu\left(2r\tilde{p}+\tilde{p}^2\right)} \frac{e^{-r^2\nu}(r^2\nu)^y}{y!} \\
&= 1 - 2e^{-\nu\tilde{p}} + e^{-\nu\tilde{p}(2r+\tilde{p})} \\
&= 1 - 2e^{-\lambda p_d\tilde{p}} + e^{-\lambda p_d\tilde{p}(2-\tilde{p})}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(P_A) &= \mathbb{E}\big[P_A^2\big] - \mathbb{E}\big[P_A\big]^2 \\
&= 1 - 2e^{-\nu\tilde{p}} + e^{-\nu\tilde{p}(2r+\tilde{p})} - 1 + 2e^{-\nu\hat{p}} - e^{-2\nu\tilde{p}} \\
&= e^{-\nu(2r\tilde{p}+\tilde{p}^2)} - e^{-2\nu\tilde{p}} \\
&= e^{-2\nu\tilde{p}}\big(e^{\nu\tilde{p}^2} - 1\big) \\
&= e^{-2\lambda p_d\tilde{p}}\big(e^{\lambda p_d\tilde{p}^2} - 1\big). \qquad \text{(S62)}
\end{aligned}$$

Using the parameters above, we estimate 2.02%±0.08% (mean ± s.d.) allogeneic TCRs for MHC-matched individuals with roughly 600 foreign p-MHC. We generally expect that 1-24% of cells are reactive in a typical allogeneic response [16]. This simple estimate assumes that a sufficient amount of relevant peptides be successfully displayed to TCRs. Variability in this process may be responsible for increased variance between individuals.

Although less clinically relevant, the analysis of allorecognition in the context of MHC-unmatched pairs requires additional assumptions. In theory, one could compare the number of differences in p-MHC as a result of distinct peptide loading on each MHC. As seen in Fig. 5, these contributions will on average always contribute more to alloreognition percentages. It is also hypothesized that direct recognition of the MHC complex by TCRs contributes to GVHD, which is at present not considered. Analogous results for Fig. 5 using the Uniform[0,1] distribution are depicted in Fig. S12.

S9 Reduced amino acid alphabets

The major recognition results obtained above are compared with an identical analysis, this time with a reduced number of possible amino acids (Fig. S13. This is motivated by the fact that correlations between functionally similar amino acids may be partitioned into functionally related equivalence classes. While the number of classes is depends on the application, approximations from the molecular biophysics community would place realistic numbers at 5-10 amino acids [17, 18].

We find predictably that selection behavior approaches the 20 amino acid case as the number of equivalence classes increases (Figure S13A). Under severe restriction (3 amino acid equivalence classes), the extreme reduction in the variety of allowable TCR binding grooves leads to peptide potency behavior resembling that found in the PIRA model. This issue is significantly mitigated in the case of 5 equivalence classes, and by 10 equivalence classes resembles the full alphabet case (Figure S13B). Recognition of point-mutated self peptide and foreign peptide effectively remains unchanged in the simulations that were compared to empirical estimates $\tilde{p}_1$ and $\hat{p}_0$.

We are therefore confident that the above analysis still holds under reasonable assumptions of reduced amino acid alphabets.

References

[1] Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK (2008) How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proceedings of the National Academy of Sciences of the United States of America* 105(43):16671–16676.

[2] Košmrlj A, Chakraborty AK, Kardar M, Shakhnovich EI (2009) Thymic selection of T-cell receptors as an extreme value problem. *Physical Review Letters* 103(6):3–6.

[3] Kosmrlj A, et al. (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* 465(7296):350–4.

[4] Chakraborty AK, Kosmrlj A (2010) Statistical mechanical concepts in immunology. *Annual review of physical chemistry* 61:283–303.

[5] Marsh SGE, et al. (2010) Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* 75(4):291–455.

[6] Hernandez JB, Newton RH, Walsh CM (2011) Life and death in the thymus - cell death signaling during T cell development. *Curr Opin Biol* 22(6):865–871.

[7] Klein L, Kyewski B, Allen PM, Hogquist K (2014) Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature reviews. Immunology* 14(6):377–91.

[8] Irwin J (1927) On the Frequency Distribution of the Means of Samples from a Population Having any Law of Frequency with Finite Moments , with Special Reference to Pearson ' s Type II. *Biometrika* 19(3):225–239.

[9] Hall P (1927) The Distribution of Means for Samples of Size N Drawn From a Population in which the Variate Takes Values Between 0 and 1, All Such Values Being Equally Probable. *Biometrika* 19(3/4):240–245.

[10] Sinclair C, Bains I, Yates AJ, Seddon B (2013) Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proceedings of the National Academy of Sciences of the United States of America* 110(31):E2905–14.

[11] Itano A, Robey E (2000) Highly efficient selection of CD4 and CD8 lineage thymocytes supports an instructive model of lineage commitment. *Immunity* 12(4):383–389.

[12] Merkenschlager M, et al. (1997) How many thymocytes audition for selection? *The Journal of experimental medicine* 186(7):1149–58.

[13] Ignatowicz L, et al. (1997) T cells can be activated by peptides that are unrelated in sequence to their selecting peptide. *Immunity* 7(2):179–186.

[14] Tourne S, et al. (1997) Selection of a Broad Repertoire of CD4+ T Cells in H-2Ma 0 / 0 Mice. *Immunity* 7:187–195.

[15] Zerrahn J, Held W, Raulet DH (1997) The MHC Reactivity of the T Cell Repertoire Prior to Positive and Negative Selection. *Cell* 88(5):627–636.

[16] Detours V, Perelson AS (1999) Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection. *Proceedings of the National Academy of Sciences of the United States of America* 96(9):5153–8.

[17] Truong HH, Kim BL, Schafer NP, Wolynes PG (2013) Funneling and frustration in the energy landscapes of some designed and simplified proteins. *The Journal of Chemical Physics* 139:1–15.

[18] Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering* 13(3):149–152.

[19] Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18(3):534–552.

[20] Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology* 256(3):623–644.

[21] Flicek P, et al. (2008) Ensembl 2008. *Nucleic Acids Research* 36(SUPPL. 1):707–714.

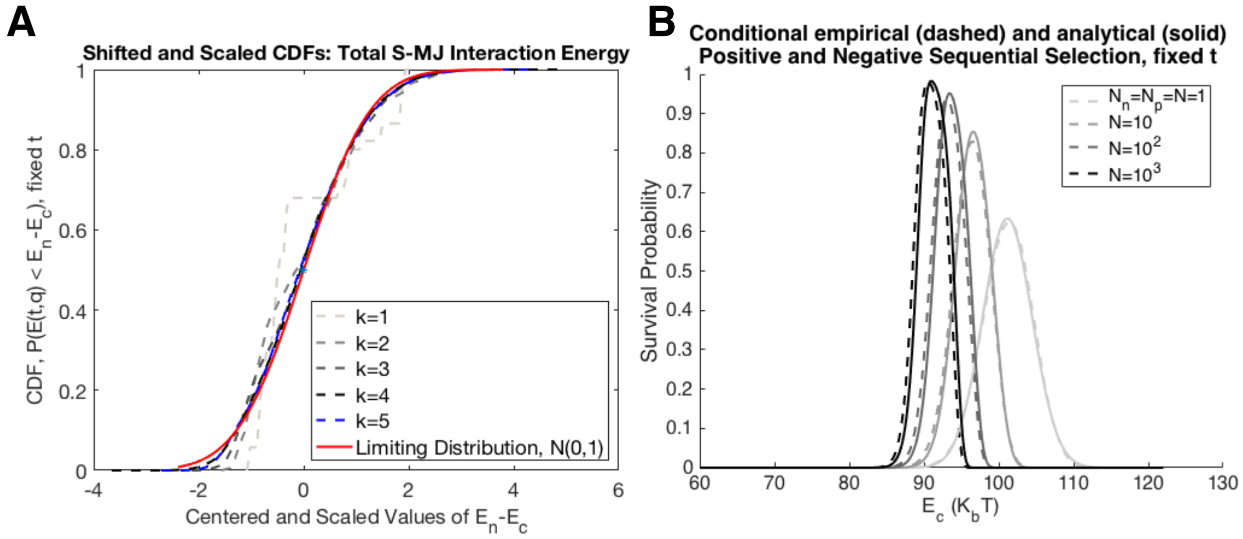

Figure S1: Properties of S-MJ model conditioned on a fixed TCR, $t = \{\text{P, R, S, D, E, D, K, R, R, M}\}$. (A) Lindeberg-Feller Central Limit Theorem convergence of sums of k random MJ interactions in the sequential MJ formulation. (B) Analytical and empirical overall survival rates conditioned on a single TCR versus absolute constant energy interaction for the sequential MJ model for various numbers of thymic self-peptides. Independent peptide populations were used for cortical and medullary selection steps ($E_n = 127, E_p = 122$ so that $E_n - E_p = 5K_\text{b}\text{T}$ as in [1–4]).

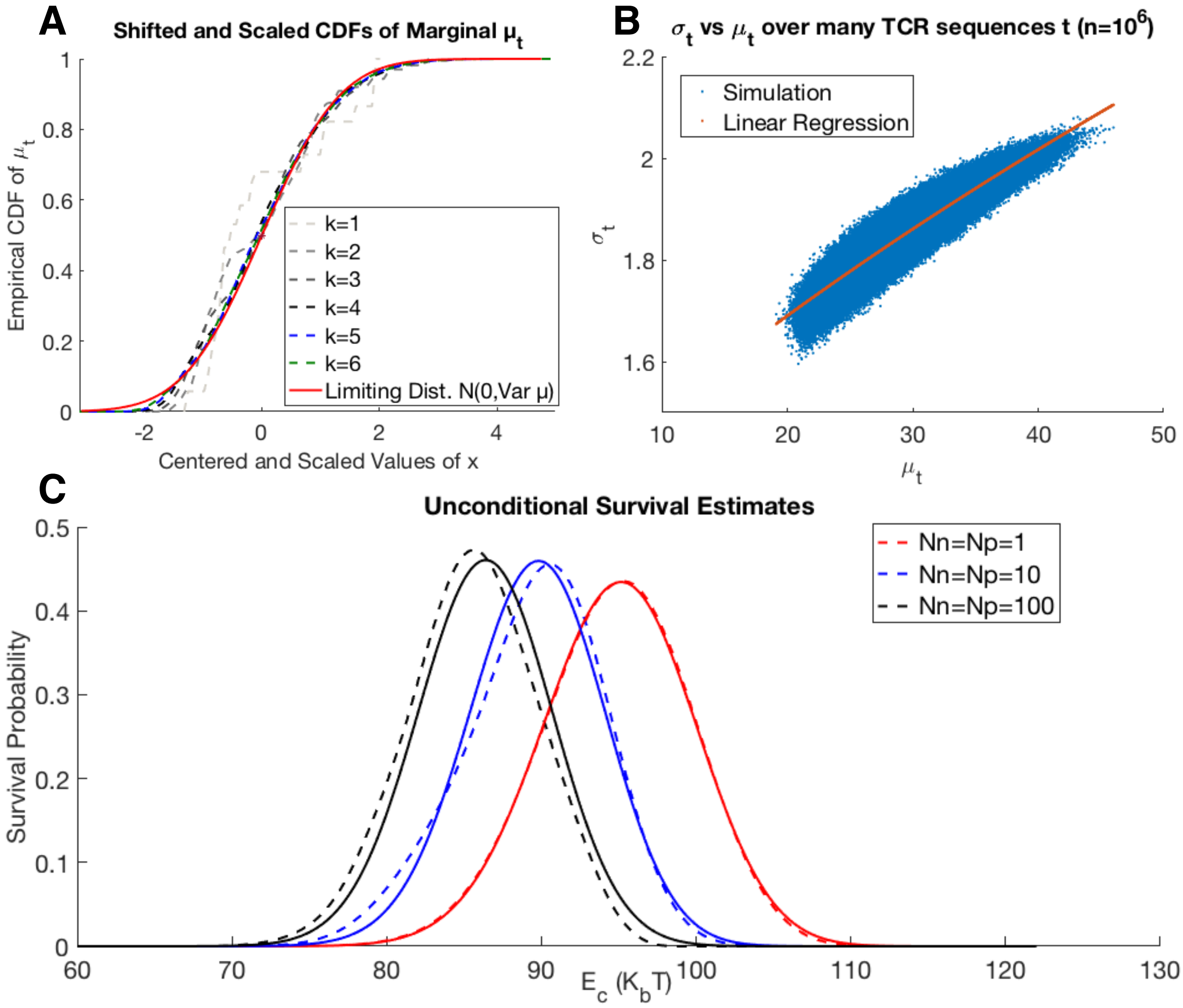

Figure S2: Unconditional TCR survival rates in the S-MJ model. (A) Classical Central Limit Theorem convergence of the distribution of, μ_t, the mean total interaction energy (averaged over all peptides) for a TCR, t. (B) Empirical plots of standard deviation versus mean MJ-interactions (n=10^6). (D) Unconditional analytical and empirical overall survival rates for the S-MJ model ($E_n = 127, E_p = 122$ so that $E_n - E_p = 5K_\mathrm{b}$T as in [1–4]).

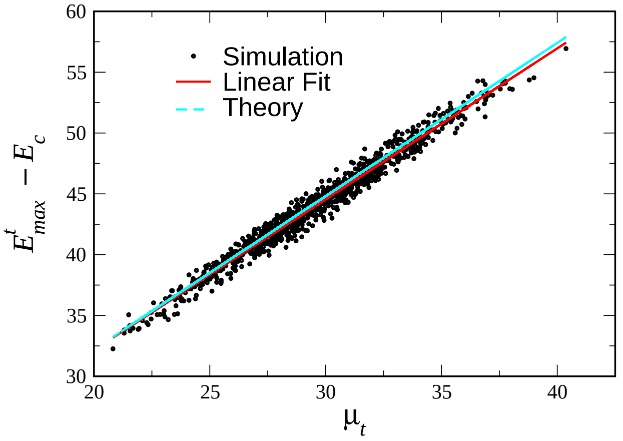

Figure S3: Correlation between μ_t and total interaction energy in the S-MJ model. The maximum interaction energy between a given TCR t and a collection of $N = 10^4$ peptides, E^t_{max} versus μ_t, for 1000 different TCRs. Also show is the linear regression and the straight-line prediction, Eq. (S20).

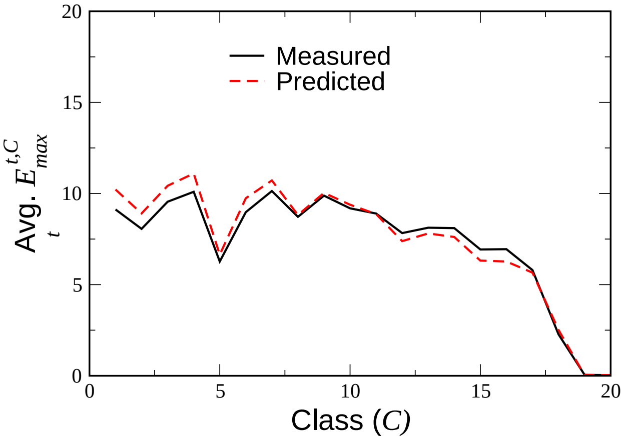

Figure S4: Maximum Binding Energy vs. Repetition Class for the PIRA model. The maximum binding energy of a TCR with elements of a given class present in a randomly generated sample of 10^4 self-peptides, averaged over 10^4 randomly generated TCRs, together with our prediction, Eq. (S33), with μ_t, σ^2_t replaced by the average values, 0 and 1, respectively. The classes are numbered in decreasing order of their number of representative peptides in the sample.

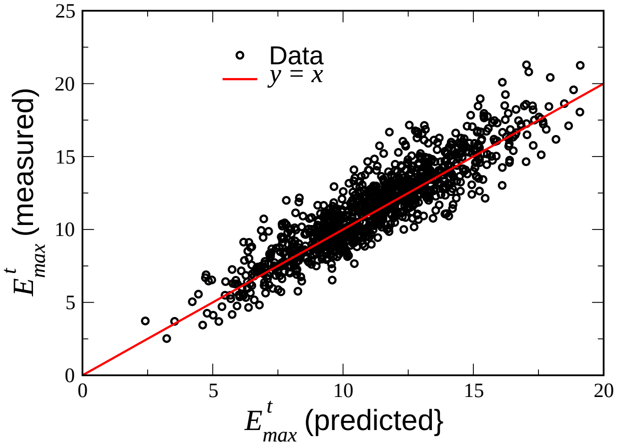

Figure S5: Maximum Binding Energy for Different TCRs in the PIRA model. The maximum binding energy for a set of 1000 TCRs with a randomly generated sample of 10^4 self-peptides, together with our prediction, Eq. (S34), with μ_t, σ_t^2 replaced by the average values, 0 and 1, respectively.

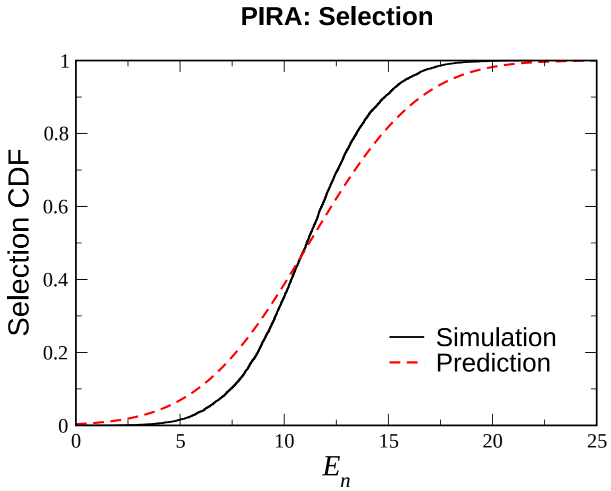

Figure S6: Selection Curve for the PIRA model. The probability of selection as measured for a set of 10^4 TCRs with a randomly generated sample of 10^4 self-peptides, together with our prediction, Eq. (S35).

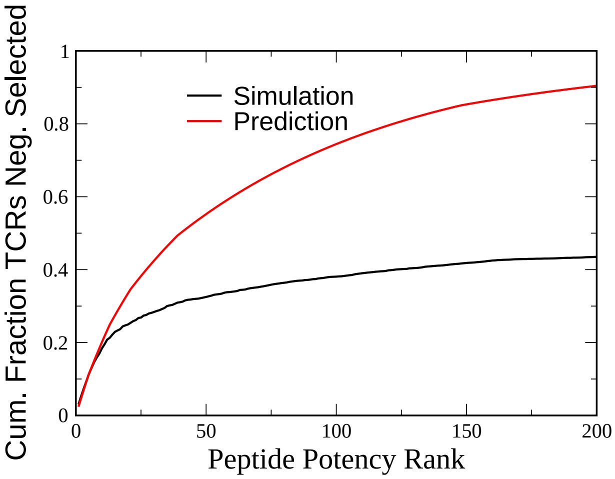

Figure S7: Potency Curve for the PIRA model. The cumulative fraction of negatively selected TCRs due to the n most potent peptides (out of 10^4) as a function of n, as measured in simulation with $E_n = 11.195$, resulting in negative selection of 50% of the TCRs.

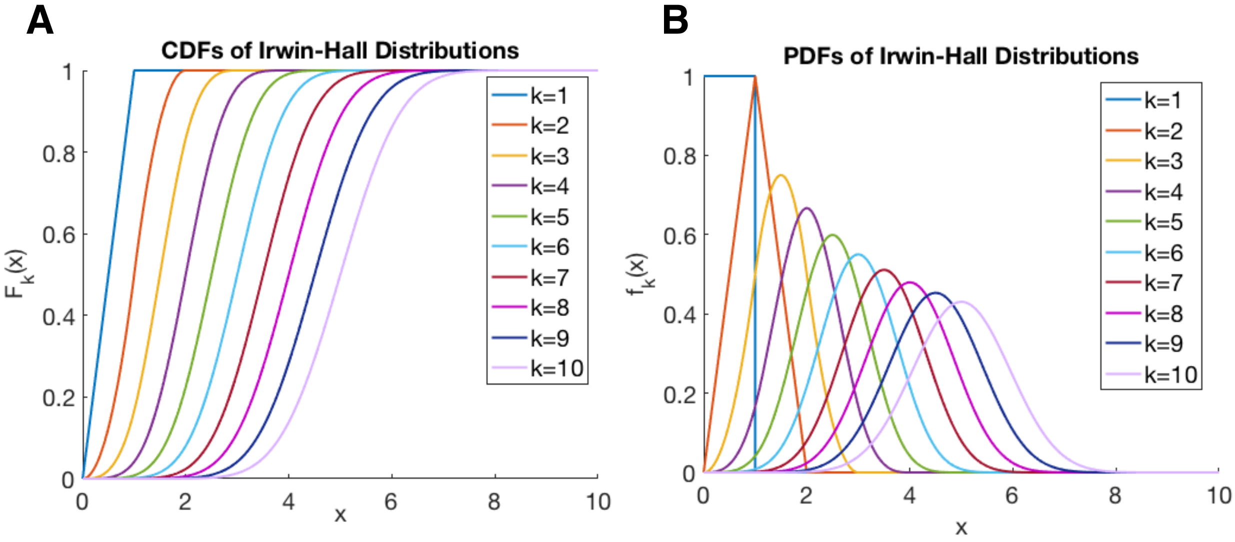

Figure S8: Statistics of the Total Energy in the RICE model. (A) Cumulative distribution functions (CDFs) and (B) probability density functions (PDFs) of the Irwin-Hall distribution across all relevant values of RICE selection thresholds, for various n.

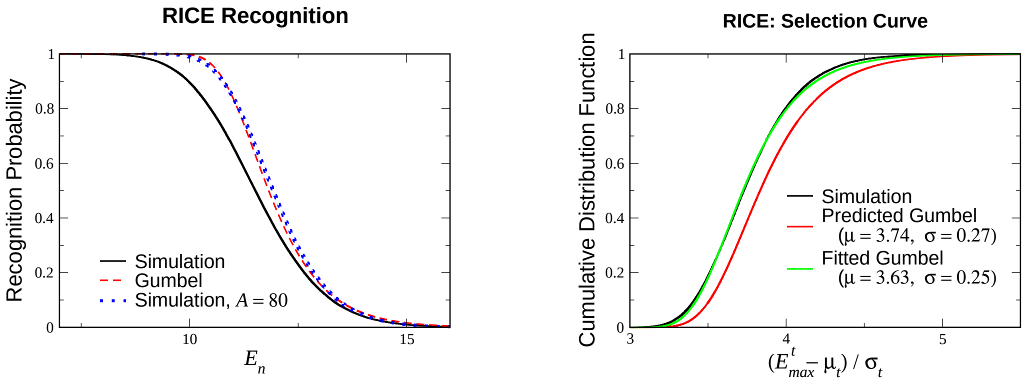

Figure S9: Gaussian Selection Curve for the RICE model. (A) The probability of selection as measured for a set of 10^4 TCRs with a randomly generated sample of 10^4 self-peptides, together with our prediction. Deviations from the simulation may be explained by the finiteness of the amino acid alphabet, with agreement occurring for $|A|$ sufficiently large. (B) Including the effect of the finite width of the sample mean and variance gives a Gumbel distribution for the maximum energy which is slightly shifted from the best fit Gumbel distribution of the data (Green curve).

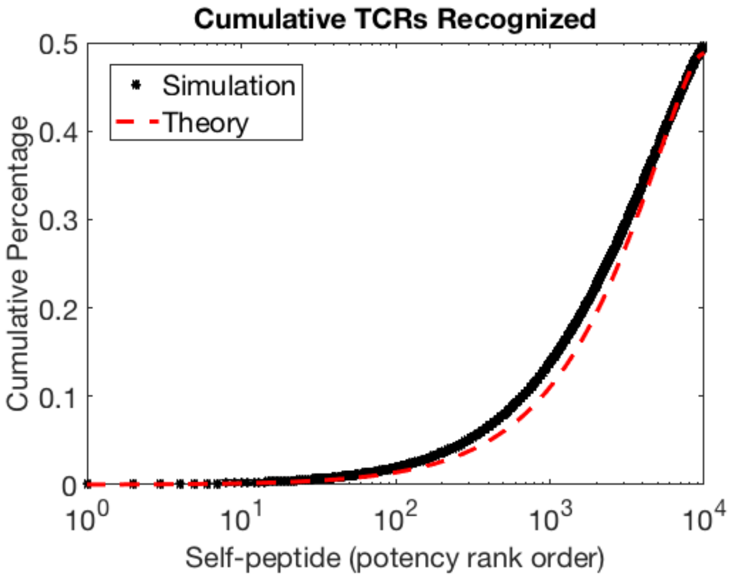

Figure S10: RICE Peptide Potency. $N_n = 10^4$ self-peptides were ordered by 'potency', or the fraction of (10^5) thymocytes recognizing them during selection simulations. 'Potent' self-peptides were those that were recognized most often by the TCRs. The cumulative contributions of each self-peptide to negative selection was plotted in decreasing order of self-peptide potency for the RICE model. Simulations were compared with theoretical estimates given by Eqs. S45.

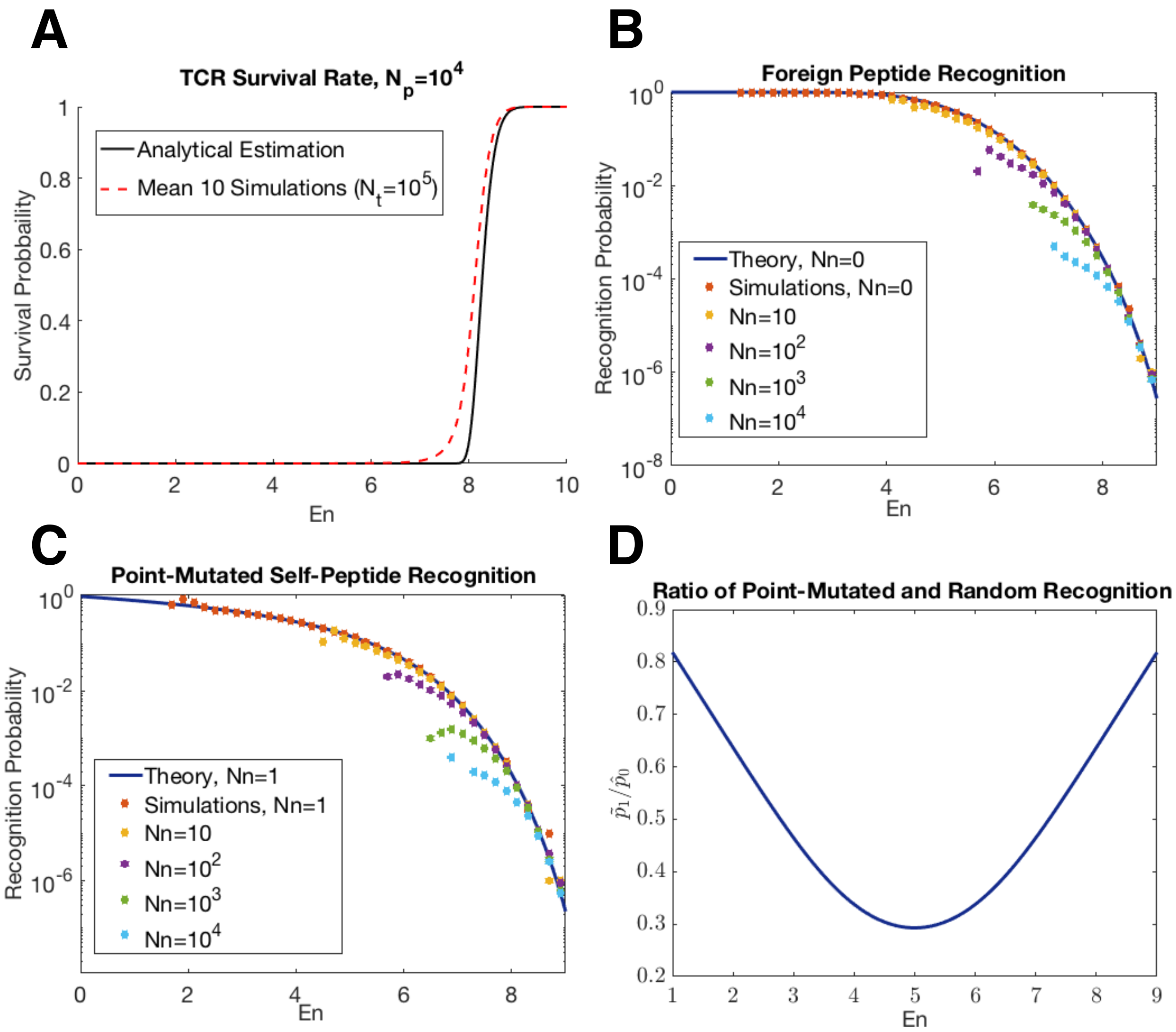

Figure S11: RICE Survival and Recognition Behavior. Analytical estimations of (A) TCR selection, (B) Foreign peptide recognition, and (C) point-mutated self-peptide recognition are compared with simulations for assuming interaction energies selected as IID Uniform[0,1] random variables. (D) The ratio $(\tilde{p}_1/\hat{p}_0)$ between recognition of point-mutated self-peptide and foreign peptide is never less than 30%.

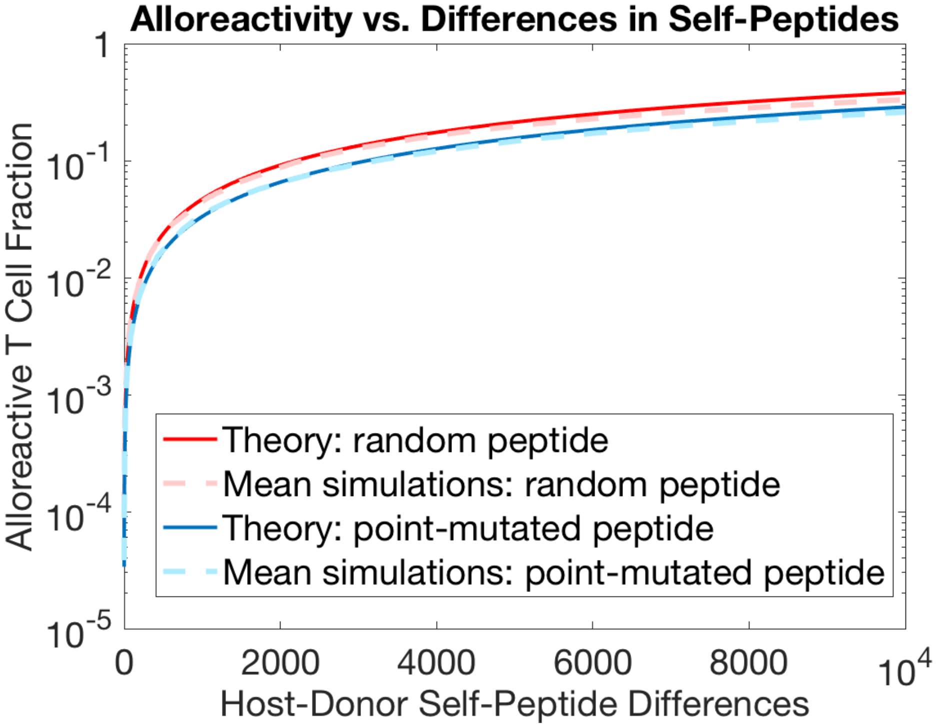

Figure S12: The Effects of Increasing Differences in Host and Donor Thymic Self-Peptides on Alloreactivity percentages. The results of Fig 5 were repeated here for the IID Uniform[0,1] formulation. For the case of maximal single amino acid sequence differences in the uniform distirubtion model model with $N_n = 10^4$ would correspond to an alloreactive rate of 26%, while maximal numbers of random peptides would correspond to rates as high as 38%.

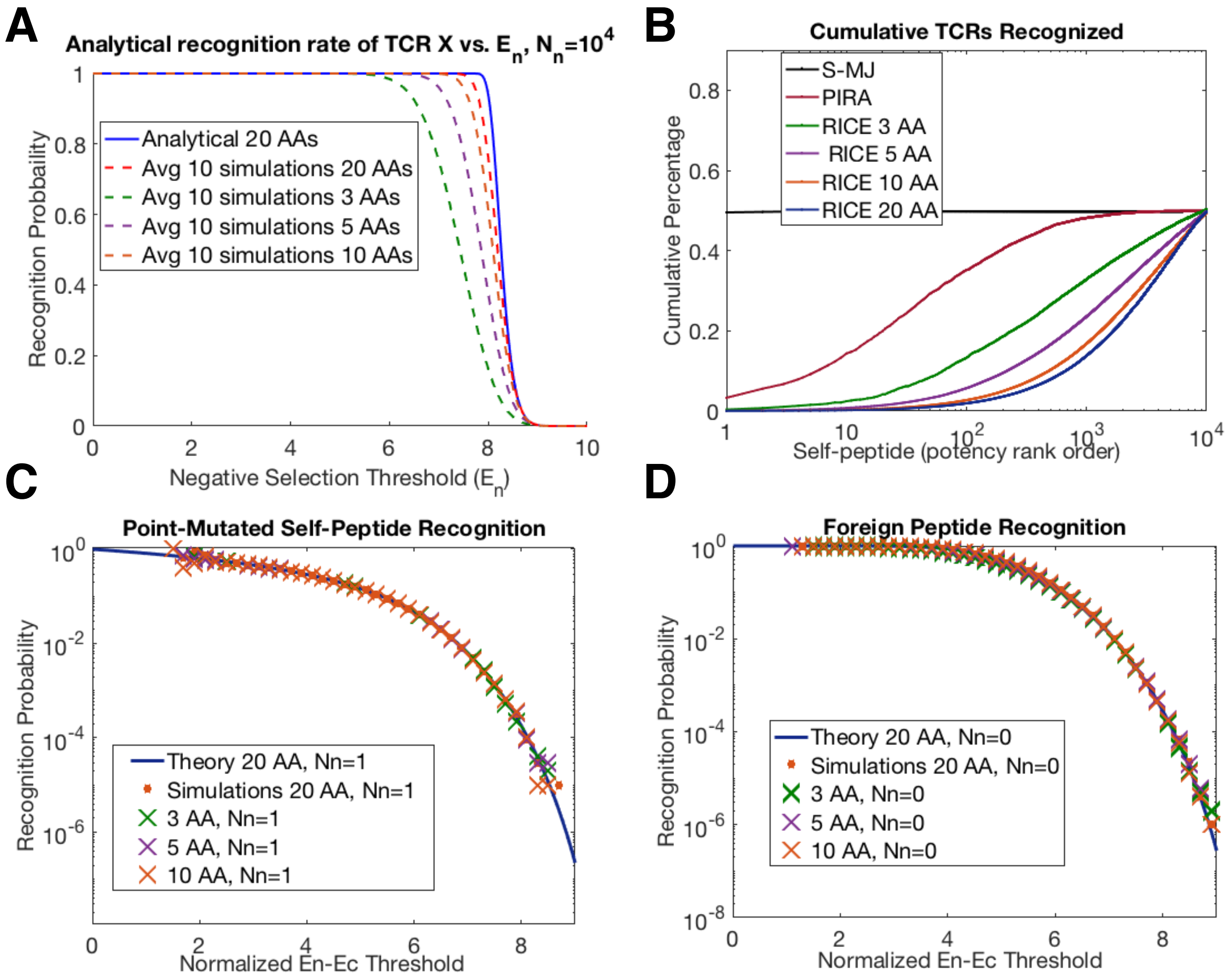

Figure S13: Key results for systems with reduced amino acid alphabets: Simulations and analytical predictions based on the uniform distribution energy RICE model of (A) survival profiles, (B) potency, (C) TAN recognition, and (D) foreign peptide recognition are presented for reduced amino acid alphabets of sizes 3, 5, and 10. In each case, $N_n = 10^4$, $N_t = 10^5$.

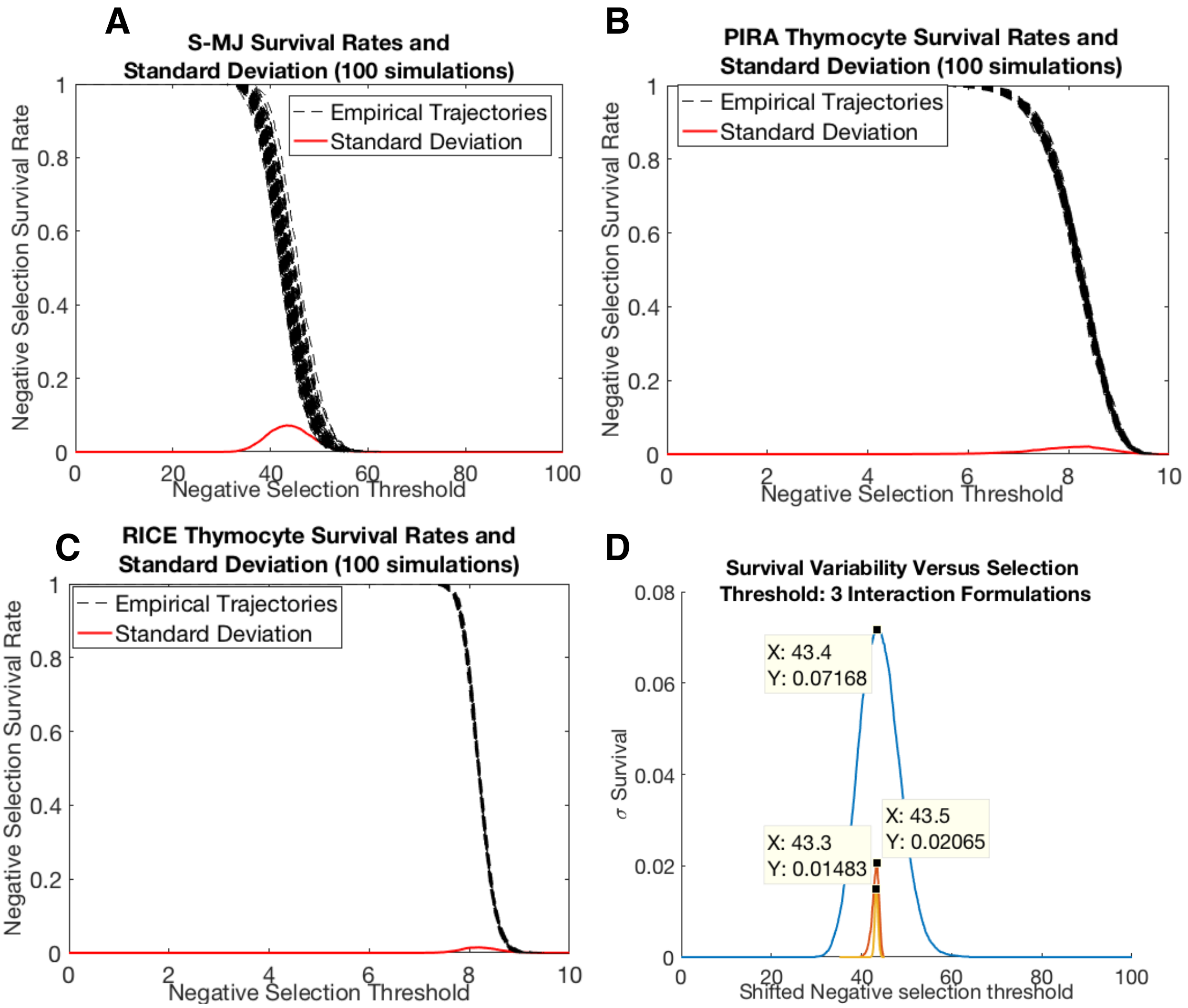

Figure S14: Fluctuations of the Selection Curve: Comparisons between the S-MJ, PIRA and RICE models. Simulated trajectories of negative selection rates and fluctuations (n=100) for (A) sequential MJ, (B) PIRA, and (C) RICE formulations. (D) Comparisons of thymocyte survival variability for each formulation overlaid with shifting selection threshold (variability for relevant selection thresholds indicated). In each case, $N_n = 10^4$, $N_t = 10^5$.

MJ Interaction Matrix																				
	C	M	F	I	L	V	W	Y	A	G	T	S	N	Q	D	E	H	R	K	P
C	5.44	4.99	5.80	5.50	5.83	4.96	4.95	4.16	3.57	3.16	3.11	2.86	2.59	2.85	2.41	2.27	3.60	2.57	1.95	3.07
M	4.99	5.46	6.56	6.02	6.41	5.32	5.55	4.91	3.94	3.39	3.51	3.03	2.95	3.30	2.57	2.89	3.98	3.12	2.48	3.45
F	5.80	6.56	7.26	6.84	7.28	6.29	6.16	5.66	4.81	4.13	4.28	4.02	3.75	4.10	3.48	3.56	4.77	3.98	3.36	4.25
I	5.50	6.02	6.84	6.54	7.04	6.05	5.78	5.25	4.58	3.78	4.03	3.52	3.24	3.67	3.17	3.27	4.14	3.63	3.01	3.76
L	5.83	6.41	7.28	7.04	7.37	6.48	6.14	5.67	4.91	4.16	4.34	3.92	3.74	4.04	3.40	3.59	4.54	4.03	3.37	4.20
V	4.96	5.32	6.29	6.05	6.48	5.52	5.18	4.62	4.04	3.38	3.46	3.05	2.83	3.07	2.48	2.67	3.58	3.07	2.49	3.32
W	4.95	5.55	6.16	5.78	6.14	5.18	5.06	4.66	3.82	3.42	3.22	2.99	3.07	3.11	2.84	2.99	3.98	3.41	2.69	3.73
Y	4.16	4.91	5.66	5.25	5.67	4.62	4.66	4.07	3.36	3.01	3.01	2.78	2.76	2.97	2.76	2.79	3.52	3.16	2.60	3.19
A	3.57	3.94	4.81	4.58	4.91	4.04	3.82	3.36	2.72	2.31	2.32	2.01	1.84	1.89	1.70	1.51	2.41	1.83	1.31	2.03
G	3.16	3.39	4.13	3.78	4.16	3.38	3.42	3.01	2.31	2.24	2.08	1.82	1.74	1.66	1.59	1.22	2.15	1.72	1.15	1.87
T	3.11	3.51	4.28	4.03	4.34	3.46	3.22	3.01	2.32	2.08	2.12	1.96	1.88	1.90	1.80	1.74	2.42	1.90	1.31	1.90
S	2.86	3.03	4.02	3.52	3.92	3.05	2.99	2.78	2.01	1.82	1.96	1.67	1.58	1.49	1.63	1.48	2.11	1.62	1.05	1.57
N	2.59	2.95	3.75	3.24	3.74	2.83	3.07	2.76	1.84	1.74	1.88	1.58	1.68	1.71	1.68	1.51	2.08	1.64	1.21	1.53
Q	2.85	3.30	4.10	3.67	4.04	3.07	3.11	2.97	1.89	1.66	1.90	1.49	1.71	1.54	1.46	1.42	1.98	1.80	1.29	1.73
D	2.41	2.57	3.48	3.17	3.40	2.48	2.84	2.76	1.70	1.59	1.80	1.63	1.68	1.46	1.21	1.02	2.32	2.29	1.68	1.33
E	2.27	2.89	3.56	3.27	3.59	2.67	2.99	2.79	1.51	1.22	1.74	1.48	1.51	1.42	1.02	0.91	2.15	2.27	1.80	1.26
H	3.60	3.98	4.77	4.14	4.54	3.58	3.98	3.52	2.41	2.15	2.42	2.11	2.08	1.98	2.32	2.15	3.05	2.16	1.35	2.25
R	2.57	3.12	3.98	3.63	4.03	3.07	3.41	3.16	1.83	1.72	1.90	1.62	1.64	1.80	2.29	2.27	2.16	1.55	0.59	1.70
K	1.95	2.48	3.36	3.01	3.37	2.49	2.69	2.60	1.31	1.15	1.31	1.05	1.21	1.29	1.68	1.80	1.35	0.59	0.12	0.97
P	3.07	3.45	4.25	3.76	4.20	3.32	3.73	3.19	2.03	1.87	1.90	1.57	1.53	1.73	1.33	1.26	2.25	1.70	0.97	1.75

Table S1: Miyazawa-Jernigan (MJ) pairwise amino acid interaction matrix [19, 20].

A.A. Distribution	
A.A. (j)	mass, $\mathbb{P}_h(j)$
C	0.0225
M	0.0215
F	0.0359
I	0.0434
L	0.0985
V	0.0598
W	0.0123
Y	0.0263
A	0.0692
G	0.0658
T	0.0536
S	0.0836
N	0.0481
Q	0.0360
D	0.0718
E	0.0476
H	0.0261
R	0.0568
K	0.0576
P	0.0636

Table S2: Distribution of amino acids obtained by their estimate from the human proteome obtained from [21].