





Contact map dependence of a T-cell receptor binding repertoireKevin Ng Chau *Physics Department, Northeastern University, Boston, Massachusetts 02115, USA*Jason T. George **Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA*José N. Onuchic *Center for Theoretical Biological Physics and Departments of Physics and Astronomy, Chemistry and Biosciences, Rice University, Houston, Texas 77005, USA*Xingcheng Lin *Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*Herbert Levine *Center for Theoretical Biological Physics and Departments of Physics and Bioengineering, Northeastern University, Boston, Massachusetts 02115, USA*

(Received 5 January 2022; accepted 10 June 2022; published 27 July 2022)

The T-cell arm of the adaptive immune system provides the host protection against unknown pathogens by discriminating between host and foreign material. This discriminatory capability is achieved by the creation of a repertoire of cells each carrying a T-cell receptor (TCR) specific to non-self-antigens displayed as peptides bound to the major histocompatibility complex (pMHC). The understanding of the dynamics of the adaptive immune system at a repertoire level is complex, due to both the nuanced interaction of a TCR-pMHC pair and to the number of different possible TCR-pMHC pairings, making computationally exact solutions currently unfeasible. To gain some insight into this problem, we study an affinity-based model for TCR-pMHC binding in which a crystal structure is used to generate a distance-based contact map that weights the pairwise amino acid interactions. We find that the TCR-pMHC binding energy distribution strongly depends both on the number of contacts and the repeat structure allowed by the topology of the contact map of choice; this in turn influences T-cell recognition probability during negative selection, with higher variances leading to higher survival probabilities. In addition, we quantify the degree to which neoantigens with mutations in sites with higher contacts are recognized at a higher rate.

DOI: [10.1103/PhysRevE.106.014406](https://doi.org/10.1103/PhysRevE.106.014406)**I. INTRODUCTION**

One of the major components of the human immune system consists of a large repertoire of T lymphocytes (or T cells). Each T cell carries a particular T-cell receptor (TCR) capable of binding to a specific antigen in the form of a peptide (p) displayed by major histocompatibility complex (MHC) molecules (pMHC) on the surface of host cells [1–4]. The activation of the T-cell response depends on the strength [5], and possibly kinetics [6], of this TCR-pMHC binding [7,8]. A typical repertoire of a healthy individual consists of $\sim 10^7$ distinct clonotypes, each with a unique TCR [9]. A growing body of research has been focused on understanding the systems-level interactions between the T-cell repertoire and its recognition of peptide landscapes indicating foreign or cancer threats.

A critical feature of a properly functioning immune system is its ability to discriminate healthy cells of the host from those

infected by pathogens, reacting to the latter ones while tolerating the former ones. In order to achieve the aforementioned discrimination, T cells must survive a rigorous selection process in the thymus before being released into the bloodstream. The first step in this process, called positive selection, ensures that TCRs in thymocytes (developing T cells) can adequately interface with pMHCs. Positive selection occurs in the thymic cortex, where cortical epithelial cells present self-peptides to thymocytes. As long as a thymocyte is able to interface with some presented pMHC, it receives a survival signal and migrates inward to the thymic medulla. This step ensures that the thymocyte has a properly functioning TCR, a rare event as only about 7–35% [10] of thymocytes survive this step. In the inner medulla, they encounter thymic medullary epithelial cells. Here, surviving immature T cells are again presented with a diverse collection of $\sim 10^4$ self-peptides [11,12] representing a variety of organ types. T cells binding too strongly to any self-peptide die off in a process known as negative selection [13,14].

As already pointed out, a key ingredient in the aforementioned process as well as in any subsequent recognition of

*Present address: Department of Biomedical Engineering, Texas A&M University, College Station, Texas 77843, USA.

a foreign antigen by a T cell is the molecular interaction of the TCR and the pMHC molecules. Crystal structures of TCR bound to pMHC show that the interface of the TCR-pMHC interaction is complex, with TCR complementarity determining regions 1 and 2 (CDR1 and CDR2, respectively) primarily binding to the MHC molecule, whereas the CDR3 complex mainly contacts the peptide in the MHC's cleft [15,16]. The CDR3 complex is comprised of two loops, CDR3 α and CDR3 β . Baker *et al.* showed that these loops can exhibit spatial and molecular flexibility during the TCR-pMHC binding process [17]; moreover, the same TCR can bind to different pMHCs [18], for example to a pMHC with a point-mutated peptide [16]. This can involve subtle changes in the CDR3 complexes' spatial conformation. It is clear then that the intricacies of the TCR binding to the pMHC as a dynamic process remain as yet to be fully understood.

In lieu of a complete first-principles understanding, several groups have pioneered the idea of employing relatively simple models so as to get a sense of how negative selection affects the T-cell repertoire. In the original set of models, TCRs and peptides were represented as strings of amino acids (AAs) which interacted in a manner that did not incorporate any structural information. In one such set of models, each AA in the pMHC binding pocket interacted with, and only with, the complementary AA in the TCR CDR3 complex. This interaction was described by either one or a set of 20×20 matrices [19–23]. These works indeed have provided a framework for describing how selection shapes the discrimination ability of the T-cell repertoire, and have been applied to understanding HIV control [24] and for assessing the detectability of cancer neoantigens [22]. In a more recent study, Chen *et al.* [25] introduced nonuniform interaction profiles that translated into some AAs in the TCRs having a more pronounced effect in pMHC recognition, but did not consider how these nonuniformities could vary between TCRs, as shown by existing crystal structures.

In this paper, we introduce the idea of a crystal-structure-dependent contact map that weights the binding energies based on the distance separating the residues on the AAs. A contact map can be thought of as a specific template for a class of TCR interface with the pMHC (TCR-pMHC) interactions, which then will yield an actual binding energy once we specify the specific AA strings on the two molecules. To focus attention on the role of the contact map, we use a simple random energy model which assigns a fixed random energy to each of the possible AA pairs. Our model, described in detail below, can be thought of a more realistic version of the the random interaction between cell receptor and epitope (RICE) model [22], in which contact map effects were simply assumed to decorrelate pair energies at different sites along a uniform binding surface.

The paper is structured as follows. In Sec. II, we present the model description along with how crystal-structure-dependent contact maps are created and also discuss the choice of energy matrix in the model. In Sec. III, we analyze how the variance of the TCR-pMHC binding energy PDF is impacted by the choice of contact map, including the roles of the total number of contacts and the topology of the contact map. We then present two applications of the model that are affected by the choice of contact map: in Sec. IV, we focus on the negative-selection recognition probability, and in Sec. V, we discuss

the point-mutant recognition probability by T cells that have survived negative selection. We present our closing remarks in Sec. VI.

II. CONTACT MAP BASED RANDOM ENERGY MODEL

Our goal is to analyze a model of negative selection in which the TCR-pMHC interaction exhibits antigen specificity of T cells dependent both on the AA occurrence and on the spatial conformation of TCR and pMHC, while retaining enough simplicity so that it can be studied analytically and with feasible computations. We represent a TCR t via its CDR3 loops in the form of a sequence of k_t AAs, $t = \{t(i)\}_{i=1}^{k_t}$, and a pMHC q as a sequence of k_q AAs, $q = \{q(j)\}_{j=1}^{k_q}$. A symmetric energy coefficient matrix of size 20×20 , $\mathbb{E} = (E_{nm})$, has entries E_{nm} that represent the pairwise binding coefficients between AAs n and m . The binding energy contributions are then assumed to be the product of a contact map $\mathbb{W} = (W_{ij})$, containing the weights W_{ij} for the interaction between t and q in a given structure, and the coefficient corresponding to the amino acid interaction. In detail,

$$U(t, q) = U_c + \sum_{i,j} W_{ij} \cdot E_{t(i)q(j)}, \quad (1)$$

where U_c represents the contribution of the TCR's CDR1 and CDR2 complexes interacting with the MHC molecule, as discussed in [19–21,24].

This form of the binding energy in (1) explicitly separates the effects on the CDR3-pMHC interaction due to spatial configuration from the effects due to the rest of the pair-dependent factors, assigning the former ones to \mathbb{W} and coarsely accounting for the latter ones in \mathbb{E} . The particular choices for the contact map \mathbb{W} will depend on the specific TCR-pMHC being used as a template. Also, this formulation does not presuppose any specific choice for \mathbb{E} . We discuss in detail specific choices of \mathbb{E} and \mathbb{W} in the sections below.

We highlight that in Eq. (1), the crystal-structure specific values W_{ij} dictate which AAs are effectively in contact. In [25], a similar equation for TCR-pMHC binding affinity weights energy coefficients with factors $f(c_i)$. However, this formulation limits TCR AA in position i to only interact with its corresponding pMHC AA, and can weight energy coefficients using different interpretations of $f(c_i)$ to accommodate the average number of contacts of position i found on an ensemble of crystal structures, but this then abrogates any capability to account for different interaction pairs for these different contacts.

A. Contact maps

Crystal structures of TCRs bound to pMHCs show a variety of spatial configurations. Each one of these can be thought of as defining a binding template which can be used to determine the energy of a set of possible pairs. In general, we expect there to be a small number of possible templates, as a specific template would presumably be valid for a subset of all pairs; even then, we must necessarily ignore the small structural changes seen between the same TCR-pMHC systems that differ, e.g., by a single AA mutation [16,26–28]. We expect, based on a recent computational study [29], that this approach

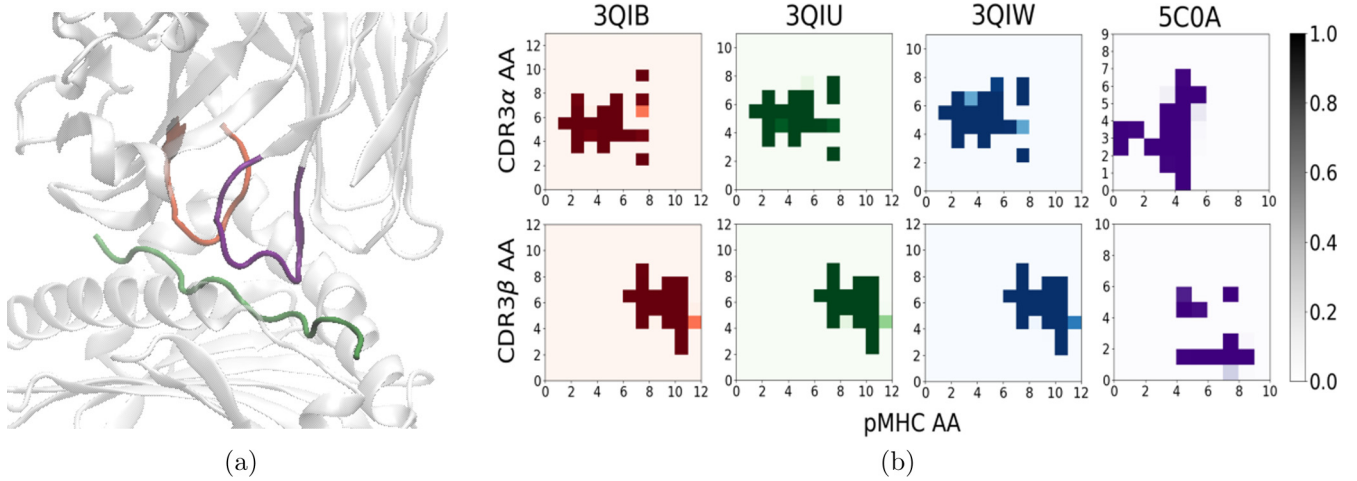


FIG. 1. The TCR-pMHC interface and contact maps. (a) The CDR3-pMHC interface in the crystal structure of the 2B4 TCR binding to the MCC/I-E^k complex (PDB ID 3QIB); with the antigen MCC highlighted in green, the CDR3 α loop in purple, and the CDR3 β loop in orange. (b) Eight contact maps estimated from four crystal structures, contact maps of the CDR3 α -pMHC (CDR3 β -pMHC) interfaces in the top (bottom) row; 3QIB, 3QIU, and 3QIW are MHC class-II restricted, whereas 5C0A is MHC class-I restricted.

will be reasonable if we stick to a fixed MHC allele, as structures with different alleles can look very different. We will see this directly in Fig. 1 below. In the calculations reported in this paper, we typically restrict ourselves to one template.

To derive a contact map from a crystal structure, we utilize the associative memory, water mediated, structure, and energy model (AWSEM) [30], developed in the context of protein folding. We use the position of C β (C α in the case of glycine) atoms to characterize the position of the residues of the AAs in both the TCRs and pMHCs, and to use AWSEM's negative-sigmoid switching function as the screening weight W_{ij} in computing the interaction energy,

$$W_{ij}(r_{ij}) = \frac{1}{2} \{1 - \tanh[\eta \cdot (r_{ij} - r_{\max})]\}. \quad (2)$$

Here, r_{ij} is the distance separating the residues at positions i and j , r_{\max} acts like a cutoff and is the inflection point of W_{ij} after which the function vanishes rapidly for $r_{ij} > r_{\max}$, and η controls how rapidly this vanishing occurs. We use crystal structures [see Fig. 1(a)] of TCR bound to pMHC deposited in the Protein Data Bank (PDB) to determine a list of AAs in the TCR t and in the pMHC q , and to calculate each distance r_{ij} , $i = 1, \dots, k_t$, $j = 1, \dots, k_q$. We then compute the corresponding weights W_{ij} from (2) and construct the contact map $\mathbb{W} = (W_{ij})$. Given that both CDR3 α and CDR3 β loops of the TCR interface with the peptide, we construct a separate contact map for each of these CDR3-loop-pMHC interactions.

To show how the proposed screening weight given by (2) derives from different TCR-pMHC crystal structures, we choose $r_{\max} = 9.5 \text{ \AA}$ and $\eta = 1 \text{ \AA}^{-1}$ and focus on four test cases. For the first three test cases, we use data from Newell *et al.* [16] who present three TCR-pMHC crystal structures: first, of the 2B4 TCR bound to the moth cytochrome *c* peptide presented by MHC molecule I-E^k (MCC/I-E^k) complex (PDB ID 3QIB); second, of the 226 TCR bound to MCC/I-E^k complex (PDB ID 3QIU); and third, of the 226 TCR bound to the MCC peptide with a glutamate in the p5 position (MCC-

p5E/I-E^k) complex (PDB ID 3QIW). For the fourth case, we follow Cole *et al.* [26] who studied the 1E6 TCR bound to human leukocyte antigen (HLA)-A02 carrying a MVWG-PDPLYV peptide of the *Bacteroides fragilis*/thetaitaomicron human pathogen (MVW peptide) (PDB ID 5C0A). For simplicity, we will refer to specific crystal structures by their PDB ID's, unless further details need to be more precisely mentioned about the TCR or the pMHC. Note that 3QIB and 3QIU represent different TCRs bound to the same pMHC complex, whereas 3QIU and 3QIW represent the same TCR bound to two pMHCs that differ by a single AA mutation in the peptide sequence. In addition, 3QIB, 3QIU, and 3QIW share the same mouse MHC class-II restriction and indeed the same I-E^k MHC-II allele, whereas the 5C0A TCR-pMHC system is presented on the human HLA A*02 MHC class-I allele.

As defined here, contact maps are sensitive to the choice of distance cutoff. Clearly, the number of contacts in a contact map for a given crystal structure increases with increasing r_{\max} values. The contact map of the 3QIB's CDR3 α -pMHC interface is plotted at four different r_{\max} values, from 6.5 to 9.5 \AA in 1 \AA increments, while keeping $\eta = 1 \text{ \AA}^{-1}$ fixed (see Fig. S1 in the Supplemental Material (SM) [31]). The contact profile gradually forms with an ever-increasing number of contacts from about 5 AA pairs in contact at $r_{\max} = 6.5 \text{ \AA}$, to about 22 AA pairs in contact at $r_{\max} = 9.5 \text{ \AA}$. For the remainder of this paper, all contact maps are calculated with $r_{\max} = 9.5 \text{ \AA}$ and $\eta = 1 \text{ \AA}^{-1}$.

The contact maps in Fig. 1(b) correspond to CDR3 α -pMHC interfaces (top row) and CDR3 β -pMHC interfaces (bottom row) from crystal structures 3QIB, 3QIU, 3QIW, and 5C0A. The contact profiles of CDR3 α -pMHC are different from the CDR3 β -pMHC contact profiles, as these parts of the TCR contact different residues on the displayed peptide. The contact maps consistently represent the physical proximity of a particular CDR3 loop to a specific portion of the pMHC, as can be seen in 3QIB's crystal structure shown in Fig. 1(a),

wherein the CDR3 α loops primarily contact AAs 2–8 and the CDR3 β loops primarily contact AAs 7–12. The detailed differences among the first three contact maps do capture slight changes in position-dependent interfacing, even when comparing contact maps for the same TCR bound to two pMHCs diverging by peptide single-AA mutation. Different weights of, for example, position pairs $(i, j) = (4, 4), (4, 8), (6, 4)$, and $(7, 6)$ are observed when comparing contact maps of 3QIU and 3QIW in Fig. 1(b) [coordinates in AA pairs are labeled as (i, j) for $t(i)$ and $q(j)$]. But, clearly, from a more coarse-grained perspective, these three can be considered to fall within one template. Conversely, the fourth map is very different, as should be expected because it is based on a different MHC molecule. Our conclusion is that we can use a single map for a class of possible pairings and thereby learn about a significant set of contributors to the T-cell repertoire. We include more contact maps from other crystal structures in the SM [31] to further support our findings (Figs. S2–S4). In general, the TCR-MHC pairing (i.e., independent of the specific peptide) has the most influence on contact map topology, with mutations or even completely altered antigens giving rise to rather small changes to the contact map topology as long as the TCR-MHC pairing remained the same (SM [31], Figs. S2 and S4). A slightly more significant change in topology is observed when different TCRs bind to the same MHC-restricted molecule even when presenting the same antigen (SM [31], Fig. S3).

As mentioned in Sec. I, the CDR3 complexes have a nuanced interaction with the pMHC. One factor that may impact this interaction is the size of AA residues, where larger-sized aromatic AAs can protrude further from the peptide chain into the other complexes in the TCR-pMHC interface and hence have a higher proclivity to contacting smaller AAs. Contact maps can be used to investigate this issue; however, in analyzing the small sample of crystal structures discussed in this manuscript, we found no conclusive evidence as to a unique role for AA size. A more extensive analysis incorporating more TCR-pMHC crystal structures is needed to make a definitive claim; this analysis is beyond the scope of this paper and will be reported upon in future work.

In the remainder of this paper, we will explore the segment of the repertoire that depends on one template and its corresponding contact map, and determine how the features of that map affect repertoire properties.

B. Energy matrix

As discussed above, we propose, for the recognition of an antigen by a T cell, an affinity-based criterion in which the TCR-pMHC binding energy $U(t, q)$ given in (1) equates to recognition (evasion) if $U(t, q)$ is above (below) a particular energy threshold U_n . Thus, we need to specify a symmetric energy coefficient matrix $\mathbb{E} = (E_{nm})$. The first example of matrix choice was one based primarily on hydrophobicity, as developed by Miyazawa and Jerningan (MJ) [32] and used in studies of thymic selection [19,25]. More recent efforts have focused on developing immune-specific energy matrices [33]. A recent study [29] used machine learning to derive the optimal matrix separating strong from weak binders within a single contact map template; this optimization approach

would lead to a different such matrix for each assumed template. Here, our interest is in the role of the contact map and so we have opted for the expedient choice of a random model where all matrix elements are chosen to be independent, mean-zero, unit-variance normally distributed random variables, $E_{nm} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Note the assumption that the n - m interaction coefficient has the same value independently of the AAs' location in the TCR or the pMHC sequences. Thus, our model is distinct from the RICE approach [22], which assumed that the spatial location of the amino acid directly affected the energy coefficient.

The position independence of E_{nm} ignores structural information such as the specific AA orientation, or to some extent the size of the residue. That this will be sufficient is at the moment uncertain, but we note that such approaches have proven useful in protein folding and related molecular biophysics computations (see [32]).

III. DISTRIBUTION OF TCR-pMHC BINDING ENERGY

The TCR-pMHC binding energy $U(t, q)$ is the indicator of the affinity between a T cell and an antigen. When assuming the pairwise AAs' interaction energies to be independent Gaussian random variables, $U(t, q)$ in (1) becomes a weighted sum of these variables with weights given by the contact map \mathbb{W} . Hence, $U(t, q)$ is also a normally distributed random variable, and since its mean is automatically zero, knowledge of the variance σ_{tq}^2 of its PDF allows us to fully characterize how $U(t, q)$ varies as we vary the particular realization of \mathbb{E} . The contact map dependence of $U(t, q)$ has a twofold impact on the variance of its PDF when compared to the case of the addition of equal variance random variables (as in the RICE approach from [22]). On one hand, the total number of nonvanishing contacts W_{ij} given by the contact map directly determines the number of random energies E_{ij} contributing to $U(t, q)$, thus increasing σ_{tq}^2 as the number of nonvanishing W_{ij} 's increases. On the other hand, the particular repeat structure of AAs in the TCR sequence and in the pMHC sequence also influences σ_{tq}^2 , as a particular pair of AAs that appears multiple times in the energy summation gives rise to a variance increase. In this section, we explore how the variance of the PDF of $U(t, q)$ depends on the two aforementioned factors.

Before proceeding, we must discuss various statistical ensembles of interest here. So far, we have focused on varying the coefficient matrix, thus generating ensemble values for each specific t, q . However, we imagine that the biophysical problem is defined by a fixed \mathbb{E} , which may be chosen (as done here) in a random fashion but, as mentioned above, may be learned from the data as done in other work [29]. Thus, we are actually interested in the distribution of binding energies as we vary either the peptide (fixing the TCR), the TCR (fixing the peptide), or both, as these are what is necessary to determine the effects of negative selection. To see how to determine these distributions, we return to the basic equation,

$$U(t, q) = \sum_i^{k_t} \sum_j^{k_q} W_{ij} \cdot E_{t(i)q(j)}, \quad (3)$$

where we have limited ourselves to one class of MHC molecule and hence U_c becomes an irrelevant constant. Also, we will assume for the purpose of our analysis that W_{ij} is either 0 or 1; this is true for all but a very small number of possible pairs. Finally, we will assume to take the distribution over AA to be uniform, although it might be useful in future work to use the known AA distribution in the human proteome. With these number of assumptions, the mean value of $U(t, q)$ sampled over the peptide sequence and/or TCR sequence constrained to have no repeats is just the sample mean of drawing a number of values from a mean-zero, variance σ^2 Gaussian distribution. This number is very much peaked around zero. Similarly, the mean value of U^2 will be strongly peaked around the variance times the contact number N_c . Perhaps not surprisingly, these are the same answers we get when averaging over \mathbb{E} ; in other words, as long as we average over sufficient numbers of sequence choices, the results for all choices of coefficient matrices are the same; see the SM [31] (Sec. S8) for a more complete discussion.

Let us now extend this analysis to the more general case. We introduce the following notation: A pair repeat structure is denoted as $C_p = (l_1^{r_1}, l_2^{r_2}, \dots, l_N^{r_N})$, with $\sum r_i \cdot l_i = N_c$, where l_i denotes the number of times an amino acid pair is repeated in different contacts and r_i denotes how many such l_i repetitions there are. For example, for a total of 20 contacts, if there are three contacts with the same AA pair and two sets of two contacts with the same AA pair, this would be denoted as $C_p = (3, 2^2, 1^{13})$. An extension of the previous argument allows us to determine the most likely value of the mean energy and its variance, averaged over all possible peptide and TCR sequences that do not change the class. The mean is still zero and the variance now becomes

$$\text{Var}(C_p) = \sigma^2 \sum r_i l_i^2. \quad (4)$$

Again, this is exactly the same as the result obtained when averaging over energy coefficient matrices. A more precise version of this correspondence is presented in the SM [31] (Secs. S5 and S6). If one wants to find the total variance, we have to average over different choices of C weighted by their respective probabilities of occurrence given the assumed uniform distribution of residue choice.

We note that while a string model may also contain pair repeats, the structural topology of the contact map matters significantly and influences the likelihood of repeated amino acids. In a string model, the likelihood of repeated AA pairs is determined by the length of the TCR and pMHC sequences and by the underlying AA distribution. In the contact map dependent model, repeated AA pairs are much more likely. First, there are in general more contacts than can be accommodated by a string model. But also, for a given peptide AA contacting many TCR AAs, there is an increased likelihood that a repeated AA pair will occur once choices are made for the interacting TCR AAs. Therefore, the overall probability of obtaining certain repeat structures is directly dependent on the contact map topology. This is most evident when comparing extreme cases, say comparing a diagonal contact map and a contact map with one row of nonvanishing contacts. The latter has much higher proclivity to show repeated AA pairs.

All these amount to the repeat structures emerging from the number of contacts and topology of the contact map of choice.

A. Variance scales with the number of contacts

It is clear from the previous analysis that the variance in the binding energy distribution increases with N_c , the total number of contacts. It is easy to see from the above that there are bounds on the total variance,

$$\sigma^2 N_c \leq \text{Var}U \leq \sigma^2 N_c^2. \quad (5)$$

The lower bound comes from the case where all pairs are distinct, whereas the upper bound arises from assuming that all contacts are the same AA pair, i.e., $C = (N_c)$. From the size of the AA alphabet $|\mathcal{A}|$, the total number of AA pairs (irrespective of ordering) is $M = \binom{|\mathcal{A}|+1}{2}$. Now, we have just seen that the precise value of the variance depends on the exact repeat structure of the peptide (q) and TCR (t) AA sequences, together with the contact map. In the case where we wish to obtain the variance of the PDF obtained by varying both t and q , we can obtain a useful approximation of this variance by ignoring the exact configuration of \mathbb{W} and instead simply counting the number of times each of the M AA pairs is selected with equal probability, where there are N_c total opportunities. In this case, the number of times each AA pair is realized follows a multinomial distribution, and the variance can be calculated from the second moment of this distribution as

$$\text{Var}[U(t, q)|\mathbb{W}] \approx \frac{1}{M} N_c^2 + \left(1 - \frac{1}{M}\right) N_c. \quad (6)$$

See the SM [31] (Secs. S5 and S6) for a detailed derivation. In Fig. 2(a), the variances computed by simulation for the CDR3 α -pMHC interfaces of 3QIB, 3QIU, 3QIW, and 5C0A [top row of Fig. 1(b)] are presented along with the predicted variance from (6). As we can see, this approximation captures the basic dependence on the total number of contacts. In the SM [31] (Fig. S5), we provide further evidence for this result by considering the effects of varying the cutoff used in the definition of the contact matrix.

B. Variance depends on the repeat structures of the TCR and pMHC AA sequences

If we are looking for the distribution of energies for a fixed TCR sequence, there is no simple formula that can encompass the dependence of the variance on the exact TCR sequence and on the exact contact map. As already mentioned, we have to find the variance for different possible repeat structures and then weight them appropriately by their occurrence probability. Specifically,

$$\sigma_t^2 = \sum_{n=1}^{N_R} p_n \sigma_n^2, \quad (7)$$

where N_R is the total number of different possible structures.

We would like to work out a specific and relatively simple example to illustrate how this works. To simplify the analysis, we focus on the 3QIB CDR3 α -pMHC contact map \mathbb{W}_{3QIB}^α in Fig. 1(b) (top left) and assume that the TCR is a constant sequence of a single repeated AA $t = (t_1, t_1, t_1 \dots)$. Note that

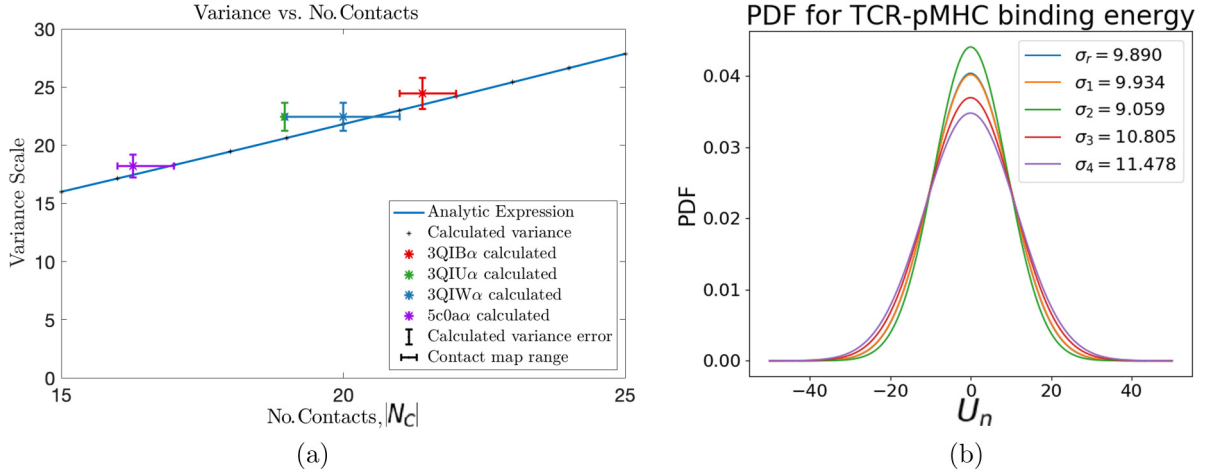


FIG. 2. The variance of the TCR-pMHC binding energy distribution depends on the total number of contacts and on the repeat structure allowed by the topology in the contact map. (a) Binding energy $U(t, q)$ variance scaling with the number of contacts, $|N_C|$; calculated variances with their variance (vertical error bars) were plotted as a function of total contacts $|N_C|$ in the contact map. Horizontal error bars represent the range of threshold used for determining each contact map (lower estimates corresponding to counting contacts >0.9 , and upper estimates corresponding to including contacts >0.1). (b) The binding energy PDFs and corresponding simulated standard deviations (σ_r, σ_1 , etc.) for pMHC repertoires of randomly chosen AA sequences (blue) and with all TCRs constrained to the same repeated AAs motif; repertoires constrained to each of the four most likely pMHC repeat motifs are shown with different colors and are labeled in decreasing order of likelihood. In the simulations, $\sigma^2 = 1$.

this makes labeling of repeat motifs dependent on the pMHC's primary sequence only. In \mathbb{W}_{3QIB}^α , only 7 AAs in t and 7 AAs in q make significant contacts, so the effective lengths are $k_t = k_q = 7$.

We will break down the problem of computing the terms in this sum as follows: We will first focus on the probable configurations of the peptide by itself and consider how the different sites are chosen. Drawn from a $|\mathcal{A}| = 20$ AA alphabet, there are $N = 15$ different repeat configurations of length 7; when randomly generating AA sequences, the four most likely repeat configurations $C_{q,1} = (2, 1^5)$, $C_{q,2} = (1^7)$, $C_{q,3} = (2^2, 1^3)$, and $C_{q,4} = (3, 1^4)$ [in the section above, C is the repeat structure of the TCR-pMHC pairing, whereas $C_{q,n}$ ($n = 1, \dots, N$) here indicate the repeat structure only of the pMHC] cover about $p_c = 96.66\%$ of the AA sequence space. A complete breakdown of these probabilities can be found in the SM [31], Table S2. We thus truncate the sum in (7) to the pairings that can be obtained from these leading order structures.

Now, each peptide configuration can give rise to a set of different possible pairing structures, depending on the specific nonvanishing elements of the contact matrix. These then need to be averaged together (with proper weighting). This somewhat complicated calculation is presented in the SM [31] (Sec. S6) and is carried out by using the self-averaging property to allow for computing the average over different realizations of the energy coefficient matrix; no rounding to 0 or 1 for the values W_{ij} is made in this calculation and the results to follow. Finally, we obtain $\sigma_t(p_c) = 9.7833\sigma$ and, extrapolating this value to approximate the full analytical value in (7), we get

$$\sigma_t \approx \sqrt{\frac{1}{p_c}} \cdot \sigma_t(p_c) = 9.95\sigma.$$

This estimation has relative error of 0.6% as compared to the simulated value of the standard deviation; see the blue plot in Fig. 2(b). The simulated PDFs related to the four most likely repeat structures are also shown in Fig. 2(b).

It is worth noting that in (7), the contributions of higher values of variances are dominated by the even faster vanishing of the corresponding probabilities. For reference, the standard deviation for this contact map ranges from $\sigma_2 = 9.0761$ for $C_{q,2} = (1^7)$ to $\sigma_{15} = 21.4090$ for $C_{q,15} = (7)$; whereas the probabilities are $p_2 = 30.52\%$ and $p_{15} = 1.56 \times 10^{-6}\%$, respectively.

IV. NEGATIVE-SELECTION RECOGNITION PROBABILITY

Negative selection trains the naïve T-cell repertoire to avoid host cells by eliminating T cells that bind too strongly to any of the self-peptides. We now wish to consider the effects on the postselection repertoire due to incorporating crystal-structure motivated contact maps into the negative-selection process.

We focus on determining the negative selection recognition probability as a function of the energy survival threshold U_n . For a T cell to survive negative selection, it must not bind too strongly, i.e., $U < U_n$, to any of the self-selecting pMHCs it encounters during selection. This is described by the probability that the maximum of the TCR-pMHC binding energies, $\max\{U(t, q_i)\}_{i=1}^{N_q}$, resulting from a T cell t undergoing negative selection against a repertoire, $\mathcal{Q} = \{q_i\}_{i=1}^{N_q}$ of N_q self-pMHCs, is below the threshold U_n [22]. This recognition probability is thus a monotonically decreasing function that gradually transitions from 1 to 0 with ever increasing values of U_n . For a fixed TCR, the scale of the transition correlates with a typical value of σ_t^2 . Averaging this over different TCRs will give rise to a width that strongly correlates with the number of

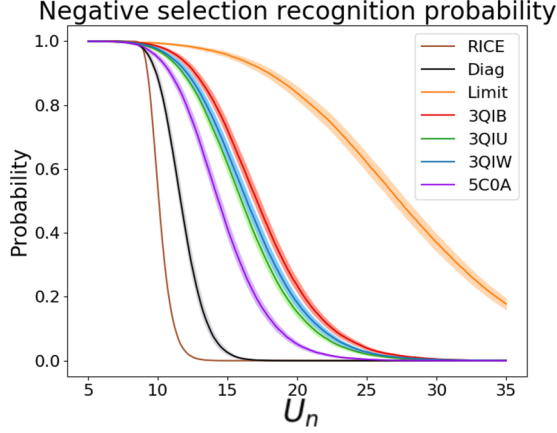


FIG. 3. Negative-selection recognition probability as a function of the survival energy threshold for T cells auditioning for negative selection. All curves involving the use of contact maps are generated from simulations sharing the same parameters apart from the contact maps. The prediction of the RICE model (brown), the identity matrix giving a diagonal contact map case (black), and the limiting case where all AAs in the CDR3 loop interact with all AAs in the pMHC (yellow) are included for comparison. Plots are averaged over the different random energy matrices in use, and shaded areas indicate the corresponding standard error of the mean.

contacts, as suggested by the phenomenological relationship given above and verified in the SM [31].

We simulate negative selection for various CDR3-pMHC interfaces (contact maps), using fixed randomly generated TCR and pMHC repertoires and 16 zero-mean, unit-variance randomly generated energy matrices \mathbb{E} . In Fig. 3, we show the recognition probability averaged over energy matrices \mathbb{E} for seven different simulations, four of them using contact maps 3QIB, 3QIU, 3QIW, and 5COA; along with a 7×7 identity-matrix contact map case, as well as the original RICE model, and a 7×7 contact map with all unit entries case simulating the scenario where all AAs in t are interacting with all AAs in q . At a given U_n , the recognition probability is higher for those contact maps with higher σ^2 [see, also, Fig. 2(a)], giving a higher probability for a pair of t and q to bind strongly enough and thus for t to face deletion. Here, the independence of RICE energy terms eliminates any possibility of the effects due to repeated AA pairs, which therefore yields a minimal variance estimate for a given number of contacts. The comparatively greater variance of the diagonal contact model is the result of possible repeated interaction terms. This leads to higher negative selection recognition probability for the diagonal contact map case and makes it closer to an actual contact map dependent calculation. Interestingly, the data in the figure show directly that similar to what we argued earlier, the recognition probability curve for a single realization is quite accurately given by the average over energy matrices.

V. RECOGNITION PROBABILITY OF POINT-MUTATED ANTIGENS BY NEGATIVELY SELECTED T CELLS

One of the motivations to model negative selection is to understand how the rejection of T cells that detect self-peptides negatively impacts the chances that T cells can detect tumor

neo-antigens; after all, these neo-antigens are typically just one mutated amino acid away from a self-peptide sequence. We therefore turn to the probability that a T cell (t) that has survived negative selection is able to recognize an antigen (\tilde{q}) whose primary sequence differs by only one AA from a self-peptide (q) included in the negative-selecting repertoire (\mathcal{Q}). We call such antigen a point mutant. In general, this probability for a fixed T cell is defined via

$$\tilde{D}_t(N_q) = \mathbb{P}[U(t, \tilde{q}) \geq U_n | \max\{U(t, \mathcal{Q})\} < U_n], \quad (8)$$

where we have averaged over all possible point mutants with nontrivial contacts. Here, \mathcal{Q} denotes the selecting repertoire of N_q peptides, one of which is q . Prior modeling (cf. [22]) has demonstrated the utility of considering two analytic approximations for the selection and recognition process. Since \tilde{q} is closely related to q , we approximate the recognition of \tilde{q} based on selection trained to *only* avoid q , \tilde{q} 's most closely related peptide, corresponding to the $N_q = 1$ case. Similarly, since a randomly generated peptide not participating in selection shares little overlap with any self-peptides, we approximate the postselection recognition of a random peptide by the *unconditional* recognition probability, corresponding to the $N_q = 0$ case. In the limiting case where t has not undergone negative selection ($N_q = 0$), Eq. (8) reduces to the recognition probability of a randomly generated antigen. The case corresponding to t negatively trained only on q ($N_q = 1$), where the point-mutant position has k contacts, results in the expression

$$D_t(1) = 1 - F_R(U_n)^{-1} \left[\int_{\mathbb{R}} F_{R-k}(U_n - x) F_k(x) f_k(x) dx + \int_{\mathbb{R}} \int_{[x, \infty)} F_{R-k}(U_n - \tilde{x}) f_k(\tilde{x}) f_k(x) d\tilde{x} dx \right], \quad (9)$$

where $F_k(x)$ and $f_k(x)$ denote the distribution function and density function of mean-zero normal random variables with variance $\sigma^2 k$ (see the SM [31], Sec. S7, for a full derivation). We expect that for relatively small N_q , it is unlikely that any of the peptides in the training set will be close enough to q or \tilde{q} to help distinguish the two binding energies; hence, \tilde{p}_1 should be a reasonable approximation to D_t . This agreement should decrease as N_q increases. The accuracy of this approximation is explored in the SM [31], Fig. S9.

More generally, we ran a set of simulations with varying sizes $N_q = \{10^2, 10^3, 10^4\}$ to assess the detection of \tilde{q} by a T cell trained to evade q . We used the CDR3 α -pMHC interface of 3QIB [top left of Fig. 1(b)] as the contact map for the simulations, for simplicity. Figure 4(a) shows the simulated point-mutant recognition probabilities as a function of T-cell negative-selection survival probability at three different sizes of the selecting repertoire. At lower (higher) values of negative-selection survival probability, i.e., when the negative selection is more (less) stringent during T-cell maturation, a mature T cell's sense of an antigen resembling self-antigens is relatively more strict (lenient); this means that the mature T cell is less (more) tolerant to changes in the peptide sequence. Therefore, recognition of the point mutant is more (less) easily triggered by deviations caused by single AA mutations; this results in higher (lower) point-mutant recognition probability at lower (higher) T-cell negative-selection survival probability. (See the SM [31], Sec. S7, for a more detailed explanation.)

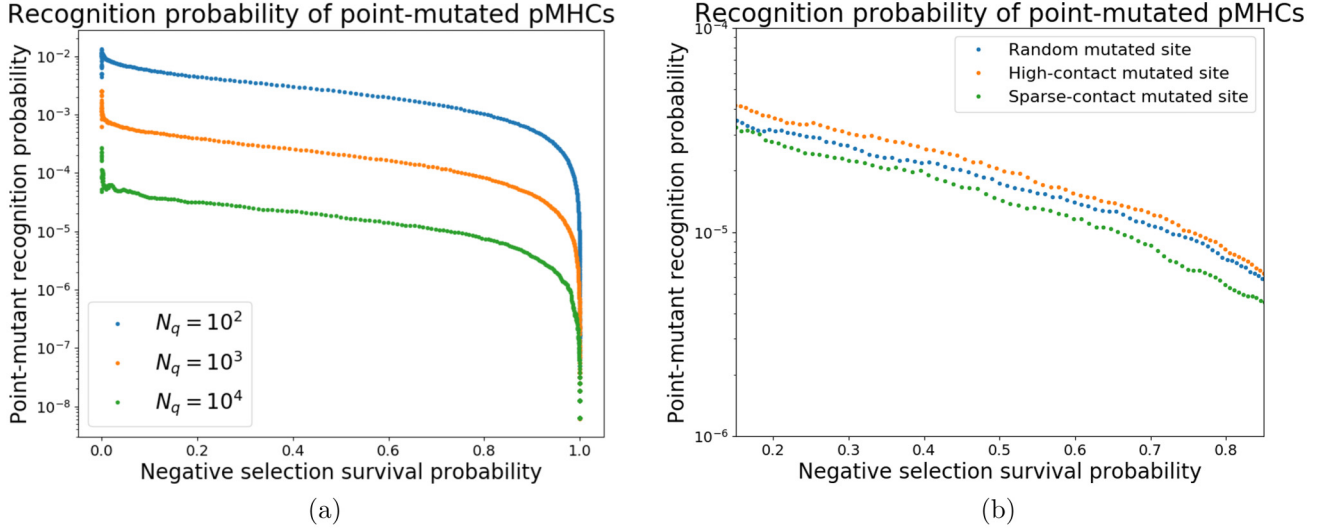


FIG. 4. Recognition probability of point-mutated peptides by T cells that have undergone negative selection. (a) The point-mutant recognition probability from simulations plotted for T cells that have received negative selection against self-peptide repertoires of three different sizes, $N_q = \{10^2, 10^3, 10^4\}$. (b) The point-mutant recognition probability from simulations that changed the site of the mutated AA; for the CDR3 α -pMHC interface of 3QIB [top left panel of Fig. 1(b)] in use, pMHC-AA in high-contact sites are in contact with 5 TCR-AA, whereas pMHC-AA in sparse-contact sites are in contact with only 1 TCR-AA; and when picking random sites to mutate, the number of peptide-AA that a given TCR-AA can contact ranges from 1 to 5.

Next, we compare the results at different N_q . This is a bit tricky because fixing the negative-selection probability leads to different thresholds U_n at different training set sizes. This accounts for a large part, but not all, of the difference in the curves seen in Fig. 4(a); see the SM [31], Fig S9. By increasing the size of the negative-selecting repertoire N_q , a mature T cell’s sense for self-antigen resemblance broadens; thus leading to higher tolerance (less detectability) for point mutants at higher N_q values.

Another feature impacting point-mutant recognition probability that stems from incorporating contact maps into the model pertains to the site in the pMHC sequence of the mutated AA. As can be seen in the contact maps in Fig. 1(b), some pMHC AAs make more significant contacts with TCR AAs than other pMHC AAs. In the case of the 3QIB’s CDR3 α -pMHC contact map [top left of Fig. 1(b)], the number of nonvanishing contacts for a particular pMHC AA ranges from 1 (sparse-contact site) to 5 (high-contact site), with an averaged 3.06 TCR AAs in contact by the 7 pMHC AAs with nonvanishing contacts. Accordingly, a point mutant \tilde{q} with its mutation occurring in a sparse-contact site (high-contact site) bears higher (lower) resemblance with the nonmutant q for a T cell. This effect clearly should impact the point-mutant recognition probability, with high-contact site point mutants having higher recognition probability than their sparse-contact counterparts, and point mutants with randomly chosen mutation sites having recognition probability somewhere in between the aforementioned two. We investigated this idea by running three simulations as explained in the paragraph above, but with the additional constraint that in each round of simulations, the mutated site was as follows: one, always a high-contact site; two, always a sparse-contact site; and three, randomly chosen. The negative-selection repertoire was fixed at $N_q = 10^4$. The point-mutant recognition proba-

bility of these simulations is shown in Fig. 4(b) and exhibits agreement with the expected behavior.

The aforementioned RICE framework cannot adequately distinguish high-contact sites from sparse ones on either the TCR or pMHC amino acid sequences. RICE’s prediction for neo-epitope recognition probability therefore represents fixed estimates for a typical “one-contact” mutation. On the other hand, the approach in this paper enables a quantitative estimate of this obvious dependence. This aligns with previous strategies calling for mutations to target TCR-facing peptide amino acids; see, for example, [34,35].

In [36], Karapetyan *et al.* showed that amino acids in the peptide that face the TCR are less tolerant to substitution, resulting in a drastic decrease in T-cell binding, activation, and killing when the TCR-facing amino acids are swapped; other amino acids in the peptide were more tolerant to substitutions. Also, Wilson *et al.* [37] found that for the *Plasmodium berghei* peptide (SYIPSAEKI), four peptide amino acid positions (S1, I3, S5, and E7) outside of the known TCR-contacting position (K8) moderately decreased T-cell re-stimulation *in vitro* when swapped with alanine. In addition, the T-cell re-stimulation response was modest for alanine substitution in all positions but K8 when testing with three different adjuvant or delivery systems, suggesting that only K8 hinders cross reactivity when replaced by alanine. Taken together, these two papers highlight a more influential role of TCR-facing (potentially high-contacting) peptide amino acids over other peptide amino acids.

VI. CONCLUSIONS

In this manuscript, we considered the role of a nontrivial contact map acting as a template for the explicit interactions between the TCR and pMHC AA sequences. This

approach is a compromise between making an arbitrary rule as to how these sequences interact (for example, assuming only diagonal coupling as done in previous models) or using a measured crystal structure for each considered pair, an obvious impossibility for anything resembling a large repertoire undergoing negative selection. The formulation isolates contributions from spatial conformation of CDR3 loops and pMHC complexes into these contact maps, while the remaining features are encapsulated in energy coefficient matrices. The above model takes into account the spatial proximity of TCR-peptide amino acid pairs through the contact map and implicitly contains information regarding amino acid sizes. It does not encode other AA pair-specific structural information, for example, orientation. The RICE model makes the alternate assumption, namely, that additional structural details make each pair energy completely independent of each other, even for the exact same AAs. This makes a very big difference in the variance calculations, as has been seen in the selection curves. Also, if every contacting pair has a different energy, we could not possibly learn useful energy matrix models from existing datasets of strong binders. We therefore have chosen to proceed with the simpler assumption, recognizing that this may need to be modified in the future.

Although all the analysis here was done using randomly generated energy matrices, serving as a baseline “toy” model, the methodology is not restricted to such a choice and other energy matrices, such as the hydrophobicity-driven MJ matrix [32,38], or data-driven matrices [29] can be used instead. Herein, we compared negative-selection recognition probabilities of the contact map dependent model with that of the RICE model; in [22], there is a more in-depth comparison of the RICE model with an approach that uses MJ energy coefficients. Since our focus here is on the role of structural information, we restricted our analysis to models with the simplest approach to the energy matrix, namely, assuming it is composed of Gaussian random variables. Future efforts will combine our analysis here with more realistic energy matrices, as determined, e.g., by the machine learning methods in our recent paper [29].

We observed that the inclusion of contact maps gave rise to several features impacting the variance of the TCR-pMHC binding energy: a density-related one, as the number of non-vanishing contacts correlates with increased variance, and a topology-related one, in which the repeat structure of the AAs in CDR3-loops’ and in pMHC-complexes’ sequences also skews the variance, with additional repeats correlating with increased variance. These changes in variance also affect negative-selection recognition probabilities, with larger vari-

ances driving higher recognition probabilities. The proposed generalization is therefore useful for characterizing the distributional behavior of TCR systems with a relatively fixed contact structure. Given that even at fixed MHC allele, there are likely to be several distinct spatial conformations that can give rise to effective binding, a full treatment of the repertoire should include finding the set of templates that give rise to the largest possible binding for the sequences under consideration. This extension will be reported elsewhere.

Another influence of the topology of the contact map manifest in the recognition probability of point-mutated antigens by T cells that have been negatively selected. Here, some pMHC-AAs have a higher number of nonvanishing contacts with TCR-AAs, that upon mutation make the antigen to be perceived more like foreign by the T cells than when mutating pMHC-AAs with fewer nonvanishing contacts. This results in higher recognition probability of high-contact site point mutants. Conversely, this notion can provide at least some information about which mutations in a previously detected peptide could prevent the detection of an evolved virus by memory T cells generated in an earlier infection. Data to this effect are now becoming available in the context of COVID-19-specific T cells in never infected individuals resulting from prior responses to other endemic coronaviruses [39].

As seen here, the problem of dissecting the generation and functioning of the postselection T-cell repertoire is incredibly complex, even utilizing a number of vastly simplifying assumptions. The full problem requires attention to biases in the generation of the naïve repertoire [40], inclusion of a set of different MHC alleles for different individuals, a better handle on the statistical properties of the negative-selection training set, and, of course, the full range of molecular biophysics effects that contribute to binding energy and on-off kinetics. These cannot all be included in any useful theoretical model. By isolating and improving our understanding of the effects of specific contact geometries, we hope to build intuition for how different aspects of this complex system contribute to different functional aspects of the full T-cell arm of adaptive immunity.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Michael E. Birnbaum for fruitful discussion on systems-level TCR-antigen specificity. This work was supported by the National Science Foundation (NSF) Grant No. NSF PHY-2019745 (Center for Theoretical Biological Physics).

- [1] L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan *et al.*, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing, *Nature (London)* **481**, 506 (2012).
- [2] J. Robinson, A. R. Soormally, J. D. Hayhurst, and S. G. E. Marsh, The ipd-imgt/hla database - New developments in reporting hla variation, *Hum. Immunol.* **77**, 233 (2016).
- [3] T. N. Schumacher and R. D. Schreiber, Neoantigens in cancer immunotherapy, *Science* **348**, 69 (2015).

- [4] E. M. E. Verdegaal, N. F. C. C. de Miranda, M. Visser, T. Harryvan, M. M. van Buuren, R. S. Andersen, S. R. Hadrup, C. E. van der Minne, R. Schotte, H. Spits *et al.*, Neoantigen landscape dynamics during human melanoma-T cell interactions, *Nature (London)* **536**, 91 (2016).
- [5] D. K. Das, Y. Feng, R. J. Mallis, X. Li, D. B. Keskin, R. E. Hussey, S. K. Brady, J.-H. Wang, G. Wagner, E. L. Reinherz *et al.*, Force-dependent transition in the T-cell receptor β -subunit allosterically regulates peptide discrimination and pmhc bond lifetime, *Proc. Natl. Acad. Sci.* **112**, 1517 (2015).

- [6] P. François and G. Altan-Bonnet, The case for absolute ligand discrimination: Modeling information processing and decision by immune T cells, *J. Stat. Phys.* **162**, 1130 (2016).
- [7] S. M. Alam, P. J. Travers, J. L. Wung, W. Nasholds, S. Redpath, S. C. Jameson, and N. R. J. Gascoigne, T-cell-receptor affinity and thymocyte positive selection, *Nature (London)* **381**, 616 (1996).
- [8] M. Krogsgaard and M. M. Davis, How T cells “see” antigen, *Nat. Immunol.* **6**, 239 (2005).
- [9] T. P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky, A direct estimate of the human $\alpha\beta$ T cell receptor diversity, *Science* **286**, 958 (1999).
- [10] C. Sinclair, I. Bains, A. J. Yates, and B. Seddon, Asymmetric thymocyte death underlies the cd4:cd8 T-cell ratio in the adaptive immune system, *Proc. Natl. Acad. Sci.* **110**, E2905 (2013).
- [11] R. J. De Boer and A. S. Perelson, How diverse should the immune system be? *Proc. R. Soc. London B* **252**, 171 (1993).
- [12] A. Yates, Theories and quantification of thymic selection, *Front. Immunol.* **5**, 13 (2014).
- [13] V. Detours and A. S. Perelson, Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection, *Proc. Natl. Acad. Sci.* **96**, 5153 (1999).
- [14] L. Klein, B. Kyewski, P. M. Allen, and K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don’t see), *Nat. Rev. Immunol.* **14**, 377 (2014).
- [15] E. Lanzarotti, P. Marcantili, and M. Nielsen, Identification of the cognate peptide-mhc target of T cell receptors using molecular modeling and force field scoring, *Mol. Immunol.* **94**, 91 (2018).
- [16] E. W. Newell, L. K. Ely, A. C. Kruse, P. A. Reay, S. N. Rodriguez, A. E. Lin, M. S. Kuhns, K. C. Garcia, and M. M. Davis, Structural basis of specificity and cross-reactivity in T cell receptors specific for cytochrome c- β , *J. Immunol.* **186**, 5823 (2011).
- [17] B. M. Baker, D. R. Scott, S. J. Blevins, and W. F. Hawse, Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism, *Immunol. Rev.* **250**, 10 (2012).
- [18] L. A. Colf, A. J. Bankovich, N. A. Hanick, N. A. Bowerman, L. L. Jones, D. Kranz, and K. C. Garcia, How a single T cell receptor recognizes both self and foreign mhc, *Cell* **129**, 135 (2007).
- [19] A. Košmrlj, A. K. Jha, E. S. Huseby, M. Kardar, and A. K. Chakraborty, How the thymus designs antigen-specific and self-tolerant T cell receptor sequences, *Proc. Natl. Acad. Sci.* **105**, 16671 (2008).
- [20] A. Košmrlj, A. K. Chakraborty, M. Kardar, and E. I. Shakhnovich, Thymic Selection of T-cell Receptors as an Extreme Value Problem, *Phys. Rev. Lett.* **103**, 068103 (2009).
- [21] A. K. Chakraborty and A. Košmrlj, Statistical mechanical concepts in immunology, *Annu. Rev. Phys. Chem.* **61**, 283 (2010).
- [22] J. T. George, D. A. Kessler, and H. Levine, Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides, *Proc. Natl. Acad. Sci.* **114**, E7875 (2017).
- [23] I. Wortel, C. Keşmir, R. J. de Boer, J. N. Mandl, and J. Textor, Is T cell negative selection a learning algorithm?, *Cells* **9**, 690 (2020).
- [24] A. Košmrlj, E. L. Read, Y. Qi, T. M. Allen, M. Altfeld, S. G. Deeks, F. Pereyra, M. Carrington, B. D. Walker, and A. K. Chakraborty, Effects of thymic selection of the T-cell repertoire on hla class I-associated control of HIV infection, *Nature (London)* **465**, 350 (2010).
- [25] H. Chen, A. K. Chakraborty, and M. Kardar, How nonuniform contact profiles of T cell receptors modulate thymic selection outcomes, *Phys. Rev. E* **97**, 032413 (2018).
- [26] D. K. Cole, A. M. Bulek, G. Dolton, A. J. Schauenberg, B. Szomolay, W. Rittase, A. Trimby, P. Jothikumar, A. Fuller, A. Skowera *et al.*, Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity, *J. Clin. Invest.* **126**, 2191 (2016).
- [27] D. K. Sethi, S. Gordo, D. A. Schubert, and K. W. Wucherpfennig, Crossreactivity of a human autoimmune TCR is dominated by a single TCR loop, *Nat. Commun.* **4**, 2623 (2013).
- [28] Y. T. Ting, S. Dahal-Koirala, H. S. K. Kim, S.-W. Qiao, R. S. Neumann, K. E. A. Lundin, J. Petersen, H. H. Reid, L. M. Sollid, and J. Rossjohn, A molecular basis for the T cell response in hla-dq2.2 mediated celiac disease, *Proc. Natl. Acad. Sci.* **117**, 3063 (2020).
- [29] X. Lin, J. T. George, N. P. Schafer, K. Ng Chau, M. E. Birnbaum, C. Clementi, J. N. Onuchic, and H. Levine, Rapid assessment of T-cell receptor specificity of the immune repertoire, *Nat. Comput. Sci.* **1**, 362 (2021).
- [30] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, Awsem-md: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing, *J. Phys. Chem. B* **116**, 8494 (2012).
- [31] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.106.014406> for additional figures, detailed derivation of equations, and further supporting arguments and analysis.
- [32] S. Miyazawa and R. L. Jernigan, Estimation of effective inter-residue contact energies from protein crystal structures: Quasi-chemical approximation, *Macromolecules* **18**, 534 (1985).
- [33] A. L. Woelke, J. von Eichborn, M. S. Murgueitio, C. L. Worth, F. Castiglione, and R. Preissner, Development of immune-specific interaction potentials and their application in the multi-agent-system vaccimm, *PLoS One* **6**, e23257 (2011).
- [34] D. Chowell, S. Krishna, P. D. Becker, C. Cocita, J. Shu, X. Tan, P. D. Greenberg, L. S. Klavinskis, J. N. Blattman, and K. S. Anderson, TCR contact residue hydrophobicity is a hallmark of immunogenic cd8+ T cell epitopes, *Proc. Natl. Acad. Sci.* **112**, E1754 (2015).
- [35] X. Shang, L. Wang, W. Niu, G. Meng, X. Fu, B. Ni, Z. Lin, Z. Yang, X. Chen, and Y. Wu, Rational optimization of tumor epitopes using in silico analysis-assisted substitution of TCR contact residues, *Eur. J. Immunol.* **39**, 2248 (2009).
- [36] A. R. Karapetyan, C. Chaipan, K. Winkelbach, S. Wimberger, J. S. Jeong, B. Joshi, R. B. Stein, D. Underwood, J. C. Castle, M. van Dijk *et al.*, TCR fingerprinting and off-target peptide identification, *Front. Immunol.* **10**, 2501 (2019).
- [37] K. L. Wilson, S. D. Xiang, and M. Plebanski, Functional recognition by cd8+ T cells of epitopes with amino acid variations outside known mhc anchor or T cell receptor recognition residues, *Intl. J. Mol. Sci.* **21**, 4700 (2020).

- [38] S. Miyazawa and R. L. Jernigan, Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading, *J. Mol. Biol.* **256**, 623 (1996).
- [39] J. Braun, L. Loyal, M. Frentsch, D. Wendisch, P. Georg, F. Kurth, S. Hippenstiel, M. Dingeldey, B. Kruse, F. Fauchere *et al.*, Sars-cov-2-reactive T cells in healthy donors and patients with COVID-19, *Nature (London)* **587**, 270 (2020).
- [40] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan, Statistical inference of the generation probability of T-cell receptors from sequence repertoires, *Proc. Natl. Acad. Sci.* **109**, 16161 (2012).

SI - Contact map dependence of a T cell receptor binding repertoire

Kevin Ng Chau, Jason T. George, José N. Onuchic, Xingcheng Lin, Herbert Levine

May 10, 2022

S1 Introduction

We provide complementary information for the paper that includes additional plots and derivations for the data shown and ideas discussed in the main text of the paper. We start by presenting our generalization to the RICE model, referred to as a contact-map-based random energy model, that uses crystal-structure informed contact maps to help determine the TCR-pMHC interaction energy. We show additional contact maps that were not included in the main text, and then show how the number of contacts in a contact map changes with the choice of distance cut-off r_{\max} . We provide a detailed derivation of how to estimate how the variance of the TCR-pMHC binding energy distribution scales with the number of contacts. We follow up with a discussion on how the topology of the contact map makes the AA repeat structure in a TCR's or pMHC's sequence influence the aforementioned variance. We work out an explicit TCR-pMHC repeat-structure pairing case that illustrates how to proceed with other repeat-structure pairings, and include a comparison of our predictions with direct simulations. We show an explicit derivation of the point-mutant recognition probability where the negative selection training is performed only by the non-mutated antigen, and evaluate the extent to which this provides an accurate estimate of the full recognition probability. We finish by presenting an argument for the confidence of estimating mean binding energy and its variance by self-averaging pairwise energies.

S2 Contact map based random energy model

Previously, we introduced the RICE model as a mathematical framework for T cell selection that focuses on the TCR-pMHC interface. This interface is described as sequences of AAs interacting in a site-to-site basis with no further inclusion of information regarding the spatial conformation of the AA chains in the interface. Our generalization in this paper incorporates, using this type of information obtained from crystal structures of TCRs bound to pMHCs, in the form of distance-dependent weights for the pairwise interacting energies in the TCR-pMHC interface.

The amino acid (AA) alphabet \mathcal{A} is comprised of $|\mathcal{A}| = 20$ different AAs. We define an energy matrix $\mathbb{E} = (E_{nm})$ as a $|\mathcal{A}| \times |\mathcal{A}|$ symmetric matrix containing all the AA pairwise interaction scores. Specifically, the AAs interact as pairs with the interaction between AAs a_n and a_m ($a_n, a_m \in \mathcal{A}$) contributing an energy $E_{nm} = E_{mn}$ to the overall interaction strength. Here, to allow for a better understanding on how the contact maps impact the model of the TCR-pMHC interaction, we select the entries of \mathbb{E} from the standard normal distribution so that $E_{nm} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

The TCR has two CDR3 loops (CDR3 α and CDR3 β). Because we will proceed by analyzing the energy contribution of each loop independently, we perform our analysis using a single contact map with the understanding that the complete CDR3-pMHC interaction is the addition of the contributions of each of the two CDR3 loops calculated separately. We depict the TCR as a single sequence of AAs $t = \{t(i)\}_{i=1}^{k_t}$ and the pMHC is an AA sequence $q = \{q(j)\}_{j=1}^{k_q}$, with k_t and k_q the total number of

TCR and pMHC AAs, respectively, subject to the choice of contact map. The TCR-pMHC binding energy is given by

$$U(t, q) = U_c + \sum_{i,j} W_{ij} \cdot E_{t(i)q(j)}, \quad (\text{S1})$$

where the sum of the pairwise interacting AA scores $E_{t(i)q(j)}$ are weighted by the contacts W_{ij} from the contact map. The additional element U_c accounts for the contribution of the CDR1 and CDR2 complexes interacting with the rest of the MHC molecule, which are largely conserved across specific MHC systems.

We utilize the definition of a contact weighting function between sites i and j , on respectively the TCR and the displayed peptide, utilized previously in studies of protein folding [Davtyan2012] as a negative-sigmoid that depends on three parameters: the distance separating the C_β (C_α for glycine) atoms in the crystal structure (r_{ij}), a cut-off distance r_{\max} , and a transition variable η controlling how rapidly the negative-sigmoid transitions in the vicinity of r_{\max} . The weight is calculated as

$$W_{ij}(r_{ij}) = \frac{1}{2} (1 - \tanh[\eta \cdot (r_{ij} - r_{\max})]). \quad (\text{S2})$$

The contact map-based random energy model determines the recognition of a pMHC q by a T cell t by affinity strength. An energy threshold U_n determines a binary outcome, with $U(t, q) \geq U_n$ (resp. $U(t, q) < U_n$) that represents T cell recognition (resp. no recognition). Consequently, the PDF for $U(t, q)$ determines recognition probability at a given threshold. As we assumed $E_{nm} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$, $U(t, q)$ is also normally distributed around $\mu_{tq} = 0$, leaving the variance to be determined. Note that the variance is defined by averaging over the energy matrix realization and does not directly connect to the value of $U(t, q)$ for fixed t and q is a specific realization. This point will be discussed later.

We devote later parts of this SI to a discussion of the relationship of the variance of $U(t, q)$ to the number of non-vanishing contacts in \mathbb{W} and the repeat structure of t and q . Note that we will usually define the variance with respect to averaging over the choice of the energy matrix from its normally distributed ensemble. In addition, we may investigate quantities that are defined as averages over choices of peptide sequence and/or TCR sequence. We will discuss below the extent to which these sequence-averaged quantities for a single energy matrix realization is well-approximated by its average value over the energy ensemble.

S3 Contact maps

Here we display some additional contact maps. In figure S1 we focus on the CDR3 α -pMHC interface of the crystal structure PDB ID 3QIB, with $\eta = 1 \text{ \AA}^{-1}$ fixed and various values of $r_{\max} = \{6.5, 7.5, 8.5, 9.5\} \text{ [\AA]}$ in eq. (S2), and we illustrate how the number of contacts in a contact map changes noticeably in that range of r_{\max} .

In figures S2-S4 the contact maps for different crystal structures are calculated with fixed $\eta = 1 \text{ \AA}^{-1}$ and $r_{\max} = 9.5 \text{ \AA}$. We observe that the TCR-MHC pairing is more influential than the particular peptide to the overall topology of the contact maps. When comparing the contact maps in Fig. S2, the same TCR-MHC pairing (human T cell hy.1B11 bound to MHC type II HLA-DQ1 molecule) shows a similar contact-map topology in spite of two different peptides being displayed by the MHC molecule in the two crystal structures. This feature is also present in Fig. S4, where the contact maps for four crystal structures with the same TCR-MHC pairing (human 1E6 TCR bound to MHC type I HLA-A02 molecule) bear peptides that only share the same AAs (GPD) in the core 4-6 sites. However, the same degree of homology conservation is not observed across TCRs that can bind to the same pMHC as in Fig. S3. One should therefore expect that calculations that use a fixed contact map with nonetheless varying peptides but with a fixed TCR will likely be more quantitatively accurate than calculations that also vary the TCR sequences.

To better illustrate some degree to which point-mutations in antigens are reflected in TCR-pMHC contact maps, we used PDB ID 3QIU and PDB ID 3QIW as they satisfy the point-mutation condition. The matrix subtraction the contact maps of 3QIU minus 3QIW (Fig. 1b in main) is shown in Fig. S5. In each interface, we observe up to four rather mild differing contacts out of 140 total contacts. To provide a numerical estimate, we treat each contact map as a vector (each matrix element taken as a coordinate) and defined a similarity measure as

$$M(\vec{u}, \vec{v}) = \frac{|\vec{u} - \vec{v}|}{\min(|\vec{u}|, |\vec{v}|)}. \quad (\text{S3})$$

Note that two contact maps are more similar if M is closer to zero. We found $M(3\text{QIU}, 3\text{QIW})_\alpha = 0.2508$ and $M(3\text{QIU}, 3\text{QIW})_\beta = 0.0729$; this indicates less variation in the CDR3 β -pMHC interface than in the CDR3 α -pMHC counterpart, consistent with what Fig. S5 reflects. These results are an indication of the fact that point-mutants have rather small effects on CDR3 complex-pMHC binding profile as represented through contact maps. Of course, Many more test cases should be considered to test the robustness of this claim, and this will be reported elsewhere.

S4 Repeated amino acids

To develop intuition, we consider the case where TCR t encounters a constant peptide $q = \{a, a, \dots, a\}$, $a \in \mathcal{A}$. This represents a simple, upper-bound on added variance.

S4.1 Random Energy Model Without Nonadjacent Spatial Interactions

In this case $n_t = n_q \equiv n$, and \mathbb{W} is the identity matrix, and so Eq. S1 becomes

$$U(t, q) = \sum_{k=1}^n X_{t(k), q(k)} \quad (\text{S4})$$

giving, for the fixed t and \mathbb{W} being considered,

$$\text{Var}(U(t, q) | t, \mathbb{W}) = \left\langle \sum_{j=1}^M \left(\sum_{k=1}^n X_{t_k, q_k} \right)^2 \cdot \mathbb{P}(q_k = j) \right\rangle = \frac{1}{M} \sum_{j=1}^M \left\langle \left(\sum_{k=1}^n X_{t_k, j} \right)^2 \right\rangle, \quad (\text{S5})$$

assuming that each letter in the alphabet is uniformly likely. Because we are averaging over the energy matrix ensemble, each choice of a is equivalent and the average over peptide AA choice can be dropped.

S4.1.1 Case I. Distinct TCR sequence

If t has no repeats, then the $X_{t_k, j}$ are IID random variables distributed $\mathcal{N}(0, \sigma^2)$. We refer to this as a *distinct TCR*, and denote its sum for any specific choice of j comprising the constant peptide as

$$Y \equiv \sum_{k=1}^n X_{t_k, j}. \quad (\text{S6})$$

Here, $Y \sim \mathcal{N}(0, n\sigma^2)$ so that

$$\text{Var}(U(t, q) | t, \mathbb{W}) = n\sigma^2. \quad (\text{S7})$$

S4.1.2 Case II. Constant TCR sequence

Repeated TCR amino acid entries introduces additional variance. In the extreme case, and without loss of generality, $t_k = 1$ is assumed to be constant at all positions, which refer to as a *constant TCR*. In this case,

$$Y = \sum_{k=1}^n X_{1,1} = nX_{1,1} \quad (\text{S8})$$

so that $Y \sim \mathcal{N}(0, n^2\sigma^2)$ and

$$\text{Var}(U(t, q) | q) = n^2\sigma^2. \quad (\text{S9})$$

S4.1.3 Case III. General TCR sequence

We now consider t having arbitrary repeats. We will define a sequence r of length M that counts the number of each amino acid present in the TCR sequence as follows:

$$r_i = \sum_{k=1}^n \mathbb{1}_{[r_k=i]}, \quad (\text{S10})$$

where $\mathbb{1}$ is the usual indicator function

$$\mathbb{1}_{[r_k=i]} = \begin{cases} 1, & r_k = i; \\ 0, & r_k \neq i. \end{cases} \quad (\text{S11})$$

We shall refer to this alternative representation of t as the *amino acid contact sequence*. It follows immediately by definition that $\sum_i r_i = n$. An example TCR and r pair is given below ($n = 10$, $M = 20$):

$$t = \{1, 2, 16, 5, 1, 1, 16, 8, 18, 20\} \quad r = \{3, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 1\}. \quad (\text{S12})$$

In case I (S4.1.1) above, r has n 1-entries and $M - n$ 0-entries. In case II (S4.1.2), r has a single non-zero entry equal to n . This representation allows TCRs to be written as a sum of independent random variables. Y in this case can be represented as

$$Y = \sum_{i=1}^M r_i X_{i,j}. \quad (\text{S13})$$

The r_i weight each random variable $X_{i,j}$ according to the relative abundance of that particular amino acid in the TCR sequence. This yields

$$\langle Y^2 \rangle = \sigma^2 \sum_{i=1}^M r_i^2, \quad (\text{S14})$$

so that the variance may be represented as

$$\text{Var}(U(t, q) | t, \mathbb{W}) = \alpha n \sigma^2. \quad (\text{S15})$$

where,

$$\alpha \equiv \frac{1}{n} \sum_{i=1}^M r_i^2. \quad (\text{S16})$$

The effect of repeated entries becomes apparent, as we note that Eq. S16 is minimized and equal to 1 when the peptide entries are distinct (Case I above), while $\alpha = n$ when the peptide is a single repeat (Case II).

S4.2 Random Energy Model With Nonadjacent Spatial Interactions

We now consider the addition of a contact map $\mathbb{W} = w_{k,\ell}$. The previous section details one extreme, when $w_{k,\ell} = \delta_{k,\ell}$.

S4.2.1 Case I. Adjacent interactions only

This reduces to Sec. [S4.1](#)

S4.2.2 Case II. All Interactions

If all interactions are allowed then $w_{k,\ell} = 1$. Eq. [S1](#) reduces to

$$U(t, q) = \sum_{\ell=1}^{n_q} \sum_{k=1}^{n_t} X_{t_k, q_\ell}. \quad (\text{S17})$$

so that the sum across each TCR conditioned on the peptide amino acid and rearranged according to repeats becomes

$$Y = \sum_{\ell=1}^{n_q} \sum_{i=1}^M r_i X_{i,j} = n_q \sum_{i=1}^M r_i X_{i,j}. \quad (\text{S18})$$

where again the choice of j is irrelevant. In this case, we have

$$\langle Y^2 \rangle = \sigma^2 n_q \sum_{i=1}^M r_i^2, \quad (\text{S19})$$

and

$$\text{Var}(U(t, q) \mid t, \mathbb{W}) = \alpha n_t n_q \sigma^2, \quad (\text{S20})$$

with

$$\alpha \equiv \frac{1}{n_t} \sum_{i=1}^M r_i^2. \quad (\text{S21})$$

We now can understand the range of allowable noise in these models. To compare with the behavior in Sec. [S4.1](#), we summarize the range of the conditional variance assuming that $n_t = n_q = n$.

Conditional Variance	Distinct TCR	Constant TCR
Adjacent Interactions	$n\sigma^2$	$n^2\sigma^2$
All Interactions	$n^2\sigma^2$	$n^3\sigma^2$

Table S1: Range of conditional variances in for the random energy model with spatial interactions when encountering a constant peptide.

S4.2.3 Case III. Arbitrary Spatial interactions

For a general \mathbb{W} , we have

$$U(t, q) = \sum_{\ell=1}^{n_q} \sum_{k=1}^{n_t} w_{k,\ell} X_{t_k, q_\ell} = \frac{1}{M} \sum_{j=1}^M \sum_{\ell=1}^{n_q} \sum_{k=1}^{n_t} w_{k,\ell} X_{t_k, j}. \quad (\text{S22})$$

Again, once we average over the energy matrix choice, the average over M becomes irrelevant. t 's corresponding amino acid contact sequence R is defined as

$$R_i \equiv \sum_{k=1}^{n_t} N_k \mathbb{1}_{[t_k=i]}; \quad N_k \equiv \sum_{\ell=1}^{n_q} w_{k,\ell} \quad (\text{S23})$$

where R_i (resp. N_k) represents the total number of interactions of TCR amino acid i (resp. position k) with the peptide sequence. Of course, $\sum_{i=1}^M R_i = \sum_{k=1}^{n_t} N_k$ and the special case where $\max_k N_k = 1$ reduces to Eq. [S10](#). The inner-double sum can be written as:

$$Y = \sum_{\ell=1}^{n_q} \sum_{k=1}^{n_t} w_{k,\ell} X_{t_k,j} = \sum_{k=1}^{n_t} N_k X_{t_k,j} = \sum_{i=1}^M R_i X_{i,j}, \quad (\text{S24})$$

which is analogous to Eq. [S13](#). From here,

$$\langle Y^2 \rangle = \sigma^2 \sum_{i=1}^M R_i^2, \quad (\text{S25})$$

giving

$$\text{Var}(U(t, q) \mid t, \mathbb{W}) = \alpha n_t n_q \sigma^2, \quad (\text{S26})$$

with

$$\alpha \equiv \frac{1}{n_t} \sum_{i=1}^M R_i^2. \quad (\text{S27})$$

In order to calculate conditional variance in this case, it suffices to record the number of total interactions N_k for each t amino acid position, and then aggregate their values together if the amino acid is repeated, giving R_k . From there, the final variance is related to original variance scaled according to the sum of squared R_k terms.

S4.2.4 Case IV. Non-identical variance

The above procedure can be applied to the more general case where each amino acid pair may interact with different variance: $a_{i,j} \sim \mathcal{N}(0, \sigma_{i,j}^2)$. In this case, the variance can be expressed as

$$\text{Var}(U(t, q) \mid t, \mathbb{W}) = \alpha n_q \sigma_{\text{eff}}^2, \quad (\text{S28})$$

where α is given by Eq. [S27](#), and

$$\sigma_{\text{eff}}^2 = \frac{1}{M} \sum_{j=1}^M \sqrt{\frac{\sum_{i=1}^M (\sigma_{i,j} R_i)^2}{\sum_{i=1}^M R_i^2}}. \quad (\text{S29})$$

S5 Non-repeated amino acids

Here, we consider the general case where an arbitrary TCR-peptide pairs are considered.

S5.1 Unconditional variance

In this section, the contact map \mathbb{W} is fixed, and the goal is to characterize the variance over the distribution of possible q and t sequences. We must therefore keep track of the number of unordered repeated amino acid pairs. Each of the $N_C \equiv |\mathbb{W}|$ contacts is chosen out of a total of $\tilde{M} = \binom{M+1}{2}$ possible unordered amino acid pairs $(i, j) = (j, i)$, where there are at most N_C repeats (i.e. if all contacts are the same), and at most N_C distinct pairings (provided $N_C < \tilde{M}$). We define $Z = \{Z_1, Z_2, \dots, Z_j, \dots, Z_{\tilde{M}}\}$ to be a vector of length \tilde{M} that counts the number of times each amino acid pair repeats. Ordering of the sequence elements Z matter in this representation. Each of the N_C pairs is equally likely and occurs with probability $1/\tilde{M}$. In this way, Z follows a multinomial distribution:

$$\mathbb{P}(Z) = \frac{N_C!}{\tilde{M}^{N_C} \prod_{j=1}^{\tilde{M}} Z_j!}. \quad (\text{S30})$$

We are more interested in characterizing the total number of repeats, not the particular sequence. This is best represented by integer partitions of N_C :

$$\mathcal{F}_{N_C} = \left\{ F = (1^{\eta_1}, 2^{\eta_2}, \dots, N_C^{\eta_{N_C}}) : \sum_{s=1}^{N_C} s\eta_s = N_C \right\}, \quad (\text{S31})$$

where F refers to an explicit partition of the equivalence classes \mathcal{F}_{N_C} . η_s describes the number of amino-acid pairs that are repeated s times. It can be shown that the number of elements in a given partition F , $|F|$, is given by

$$|F| = \prod_{s=1}^{N_C} \binom{\tilde{M} - \sum_{h=0}^{s-1} \eta_h}{\eta_s} = \frac{\tilde{M}!}{\left(\prod_{s=1}^{N_C} \eta_s!\right) (\tilde{M} - \eta)!}, \quad \text{with } \eta = \sum_{j=1}^{N_C} \eta_j. \quad (\text{S32})$$

The advantage of this representation is in the straightforward variance calculation:

$$\mathbb{V}\text{ar}(F) = \sigma^2 \sum_{s=1}^{N_C} (s^2 \eta_s) \quad (\text{S33})$$

which is a generalization of Eq. [S25](#). Putting this together, with $\mathbb{P}(F) = |\{z \in F\}| \mathbb{P}(z)$, and noting that $\prod_{j=1}^{\tilde{M}} z_j! = \prod_{s=1}^{N_C} (s!)^{\eta_s}$ we can calculate the total variance assuming N_C contacts as:

$$\begin{aligned} \mathbb{V}\text{ar}(U(t, q) | \mathbb{W}) &= \sum_{F \in \mathcal{F}_{N_C}} \mathbb{V}\text{ar}(U(t, q) | F) \mathbb{P}(F) \\ &= \sigma^2 \sum_{F \in \mathcal{F}_{N_C}} \frac{\tilde{M}! N_C! \sum_{s=1}^{N_C} (s^2 \eta_s)}{\tilde{M}^{N_C} (\tilde{M} - \eta)! \prod_{s=1}^{N_C} \eta_s! (s!)^{\eta_s}}. \end{aligned} \quad (\text{S34})$$

Eq. [S34](#) provides a convenient way to calculate variance by categorizing interactions based on their repeat structure. For the problem at hand, $n \sim 10$ so that $|\mathbb{W}| \leq 100$. In general, the issue of calculating variance requires an enumeration of the integer partitions of $|\mathbb{W}|$. This may become computationally infeasible for large interactions, (for example, the number of integer partitions of 100 is over $1.9 \cdot 10^8$). This variance is bounded below (resp. above) by the extreme case of having no (resp. all) repeats so that clearly

$$N_C \sigma^2 \leq \mathbb{V}\text{ar}(U(t, q) | W) \leq N_C^2 \sigma^2. \quad (\text{S35})$$

If all we are interested in is the variance scale as a function of the number of contacts, we may provide another convenient approximation by calculating the variance of a random energy model with nonadjacent spatial interactions (i.e. diagonalized contact map), this time augmented so that the total length is N_C . In this case, the number of repeats of each type may be represented by the vector Z with distribution given in Eq. [S30](#). Since Z follows a multinomial distribution with equal probabilities ($p \equiv 1/\tilde{M}$),

$$\langle Z_j^2 \rangle = \text{Var}(Z_j) + \langle Z_j \rangle^2 = N_C p(1-p) + (N_C p)^2 = p [N_C^2 p + N_C(1-p)]. \quad (\text{S36})$$

Thus this variance may be expressed as

$$\begin{aligned} \text{Var}(U(t, q) | \mathbb{W}) &= \sum_{i=1}^{\tilde{M}} \langle Z_i^2 \rangle \\ &= \tilde{M} p [N_C^2 p + N_C(1-p)] \\ &= \frac{1}{\tilde{M}} N_C^2 + \left(1 - \frac{1}{\tilde{M}}\right) N_C. \end{aligned} \quad (\text{S37})$$

This is a straightforward, closed-form estimate detailing how variance scales with the number of contacts, and states that the variance estimate is a convex combination of the two extreme cases (sums of independent energies versus dependent energies) weighted according to the total number of possible contact pairs. We remark that this estimate is a lower bound since the assumption of nonadjacent spatial interactions ignores the increased likelihood of repeated amino acid pairs arising from t and q positions having multiple contacts. Consequently, its accuracy is greater for contact maps that exhibit behavior closer to that of a diagonal map.

S5.2 Conditional variance

As before, \mathbb{W} is taken to be fixed. In this section we assume, without loss of generality for the conditional peptide case, that t is also fixed. To each TCR there corresponds an amino-acid contact sequence R_i that captures dependencies both due to repeated amino acids in the TCR binding sequence as well as repeated peptide contacts for the same TCR amino acid position. As before, sums of energy contributions across the R_i are conditionally independent. Since peptide amino acid sequences are now assumed arbitrary, we must consider the effects of distinct terms and repeats for a given amino-acid contact element. We will do this term-by-term, so that for amino acid i , there are R_i pairings with complementary peptide amino acids, each equally likely. Since TCR amino acid i is fixed, there are a total of M possible configurations determined solely by the peptide amino acids, where we have at most R_i repeats (if all peptide amino acids are the same), and at most R_i distinct pairings (provided $R_i < M$). In an approach similar to before, we define $Z = \{Z_1, Z_2, \dots, Z_j, \dots, Z_M\}$ to count the number of times each amino acid pair repeats $\{(i, 1), (i, 2), \dots, (i, j), \dots, (i, M)\}$ with $(i, j) = (j, i)$. Each pair is equally likely with probability $1/M$.

Here, the i^{th} out of M amino acids on the T-cell CDR3 sequence corresponds to R_i repeated contacts with peptide amino acids. From the above, we have that the variance for a given TCR amino acid may be given by

$$\text{Var}(U(i, q) | t, \mathbb{W}) = R_i^2/M + R_i(1 - 1/M). \quad (\text{S38})$$

Thus,

$$\text{Var}(U(t, q) | t, \mathbb{W}) = \sum_{i=1}^M R_i^2/M + R_i(1 - 1/M), \quad \text{with} \quad \sum_{i=1}^M R_i = |\mathbb{W}|. \quad (\text{S39})$$

In the special case of no repeats, with each of n_t TCR amino acids having a single contact, Eq. [S39](#) yields n_t as the scale factor, in agreement with the simplest case.

S6 Numerical approximation for TCR-pMHC binding energy variance

Here we show how we estimate the TCR-pMHC binding energy variance. The exact calculation of the aforementioned variance is the sum of the variances for the possible repeat structures weighted by the corresponding occurrence probability of the repeat structure (see eq. (S45) below). The number of elements in the sum increases rapidly with the length of the TCR and pMHC sequences, but many of the repeat structures have a very low probability of occurrence; this allows for accounting for an approximated value of the variance by only considering the most likely repeat structures in the sum, and extrapolating this approximation to estimate the expected TCR-pMHC binding energy variance.

In order to isolate the effects of repeat structures on the calculation of the variance, we restrict ourselves to equal unit variances, $E_{ij} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. We also assume that the overall contribution of the TCR binding to the MHC molecule is a constant, that for practical purposes we set to zero, $U_c = 0$. The TCR-pMHC binding energy is reduced to

$$U(t, q) = \sum_{i,j} \mathbb{W}_{ij} \cdot E_{t(i)q(j)}. \quad (\text{S40})$$

where the mean energy $\langle U(t, q) \rangle$ and variance $\text{Var}(U(t, q))$ are the first and second moments of $U(t, q)$, respectively. One way to proceed in calculating the aforementioned mean and variance is to think of $U(t, q)$ as the sum of the binding energies between a TCR AA $t(i)$ and the entire pMHC sequence $q = \{q(j)\}_{j=1}^{k_q}, U_{t(i)}$. Once the variance of $U_{t(i)}$, that is $\text{Var}(U_{t(i)}) = \sigma_{t(i)}^2$, is calculated, we can use these values to compute variance for $U(t, q)$. Therefore,

$$U(t, q) = \sum_i^{k_t} \left[\sum_j^{k_q} \mathbb{W}_{ij} \cdot E_{t(i)q(j)} \right] \equiv \sum_i^{k_t} U_{t(i)}, \quad (\text{S41})$$

and

$$\text{Var}(U_{t(i)}) = \text{Var} \left(\sum_{j=1}^{k_q} \mathbb{W}_{ij} \cdot E_{t(i)q(j)} \right) \quad (\text{S42})$$

$$= \sum_{j=1}^{k_q} \mathbb{W}_{ij}^2 \sigma_{t(i)q(j)}^2 + 2 \sum_{j < k} \mathbb{W}_{ij} \mathbb{W}_{ik} \text{Cov}(E_{t(i)q(j)}, E_{t(i)q(k)}) \equiv \sigma_{t(i)}^2. \quad (\text{S43})$$

We obtain

$$\text{Var}(U(t, q)) \equiv \sigma_{tq}^2 = \sum_{i=1}^{k_t} \sigma_{t(i)}^2 + 2 \sum_{i < k} \text{Cov}(U_{t(i)}, U_{t(k)}). \quad (\text{S44})$$

Given that $E_{t(i)q(j)}$ are IID, the second sum in (S44) involving correlation vanishes.

Equations for the variance $\text{Var}(U(t, q))$ in (S44) is a general formula that applies to any given contact map. The calculation of the variance $\text{Var}(U(t, q))$, as alluded in the paper, is contact-map specific. This contact-map dependence is two-fold: on one side, the contact map defines the values \mathbb{W}_{ij} ; on the other side, the non-vanishing weights define a topology for the CDR3-pMHC interface that influence the possible repeat structures present in the interface at a repertoire level.

At repertoire level for a TCR undergoing negative selection, one randomly generated pMHC sequence q can have as many different repeat structures as the partitions of k_q , the length of q . The variance in (S44) is in general dependent on the particular repeat structure and even the specific location of the repeated AAs of q in the CDR3-pMHC interface. The expected energy variance for a randomly generated TCR undergoing selection against a randomly generated pMHC repertoire is the occurrence probability-weighted sum of the variances for each particular repeat structure. If σ_n^2

is the average variance for the n -th repeat structure, p_n is the probability for the corresponding n -th repeat structure to occur, and N_R the total number of possible repeat structures, the expected energy variance is given by

$$\sigma_{tq}^2 = \sum_{n=1}^{N_R} p_n \sigma_n^2. \quad (\text{S45})$$

In principle, one can compute each p_n and σ_n^2 , making σ_{tq}^2 an exact estimation of the variance. However, N_R increases rapidly with the lengths of the TCR and the pMHC sequences k_t and k_q , respectively, which typically are $k_t \sim k_q \sim 10$; given that there are 42 partitions of 10, $N_R = 1764$ when $k_t = k_q = 10$, making the computation in (S45) unpractical. Nonetheless, some of the repeat structures are very unlikely (up to six orders of magnitude smaller for $k_t = k_q = 7$, this gap increases with k_t and k_q) to be present in the repertoire, while the variances fluctuate in one or two orders of magnitude; the compounding effect in (S45) is that the probabilities of the unlikely cases dominate over their respective variances, rendering the contribution of these cases to be negligible for σ_{tq}^2 . We take this simplification one step further by selecting a cut-off probability p_c of the most likely pMHC present in the repertoire, calculate a truncated variance σ_{approx}^2 , and extrapolating to obtain an approximated value of σ_{tq}^2 as

$$\sigma_{tq}^2 \approx \left(\frac{1}{p_c}\right) \sigma_{approx}^2. \quad (\text{S46})$$

We further discuss on the details and illustrate the approximation method below with an example.

S6.1 Example using 3QIB's CDR3 α -pMHC contact map

We focus on illustrating the method of approximating the variance of the energy distribution on one contact map; this method that can be applied for any contact map. We use for illustration purposes the CDR3 α -pMHC contact map of crystal structure 3QIB (see top left panel in Fig. 1B). Only 7 pMHC AAs and 7 CDR3 α AAs have significant non-vanishing contacts, leaving $k_t = k_q = k = 7$. For further simplicity, assume the TCR to be a sequence of a repeated AA, $t = \{t_r, t_r, \dots, t_r\}$, so that all repeat structure nuance is encoded in the pMHC sequence.

Under the assumptions laid out above, the exact value of the variance σ_{tq}^2 in (S45) has a sum of $N_R = 15$ elements. We illustrate how to obtain an estimation of σ_{tq}^2 by extrapolating the approximated variance σ_{approx}^2 , that is calculated by only considering for the sum the four most likely repeat structures. These four repeat structures cover about $p_c = 96.67\%$ of the randomly generated sequences in the pMHC repertoire (see table S2 for full breakdown of the probabilities of each repeat structure).

We now need the average variance of each repeat structure σ_n^2 . Here, σ_n^2 is an average of the variances of all possible permutations of repeated AAs within repeat structure C_n because the values in the contact map and the number of contacts for each pMHC AA $q(j)$ vary. For example, $C_1 = (2, 1^5)$ has one repeat among the interacting AAs in q , the location of the repeated AAs has $\binom{7}{2} = 21$ permutations, $C_2 = (1^7)$ has no repetitions, $C_3 = (2^2, 1^3)$ has $\frac{1}{2} \binom{7}{2,2} = 105$ permutations, $C_4 = (3, 1^4)$ has $\binom{7}{3} = 35$ permutations, and so on. The average σ_n 's of the repeat structures of interest are

$$\begin{aligned} \sigma_1 &= 9.9923, \\ \sigma_2 &= 9.0761, \\ \sigma_3 &= 10.8350, \\ \sigma_4 &= 11.6772. \end{aligned}$$

We show step by step how to calculate σ_1 in the section below to serve as an example for the other repeat structures. With these four values we get

$$\sigma_{approx}^2 = \sum_{n=1}^4 p_n \cdot \sigma_n^2 \quad \implies \quad \sigma_{approx} \approx 9.7833.$$

Label	Class	Probability (%)
C_2	(1^7)	30.52
C_1	$(2, 1^5)$	45.79
C_3	$(2^2, 1^3)$	15.26
C_6	$(2^3, 1)$	0.95
C_4	$(3, 1^4)$	5.09
C_5	$(3, 2, 1^2)$	1.91
C_8	$(3, 2^2)$	5.61×10^{-2}
C_{10}	$(3^2, 1)$	3.74×10^{-2}
C_7	$(4, 1^3)$	0.32
C_9	$(4, 2, 1)$	5.61×10^{-2}
C_{12}	$(4, 3)$	1.04×10^{-3}
C_{11}	$(5, 1^2)$	1.12×10^{-2}
C_{13}	$(5, 2)$	6.23×10^{-4}
C_{14}	$(6, 1)$	2.01×10^{-4}
C_{15}	(7)	1.56×10^{-6}

Table S2: Classes (repeat structures) classifying AA sequences with length $k = 7$ and their respective probability of occurrence when the AA alphabet is $A = 20$ characters in size. Repeat structures are labelled in descending order of likelihood

Thus,

$$\sigma_{tq} \approx \sqrt{\left(\frac{1}{0.9666}\right)} 9.7833 = 9.9510.$$

The accuracy of this estimation is tested with simulations (see figure S8). We found good agreement with simulations, with 0.61 % relative error from the simulated value.

S6.1.1 Explicitly working out the case when the pMHC has only one repeated amino acid

Let us work out the approximated estimation for the variance when the pMHC q has only one repeated AA, namely, q belongs to the class $C_1 = (2, 1^2)$; and the TCR $t = \{t(i) = t_r\}_{i=1}^{k_t}$ is a sequence of one AA t_r repeated in all its sites. Recall that the contact map for this example only has seven AAs in t and seven AAs in q making significant contributions to the binding energy between t and q , in this sense, $k_t = k_q = 7$.

Assume that q has the repeated AAs in sites r_1 and r_2 , $r_1, r_2 \in \{1, 2, \dots, k_q\}$, $r_1 \neq r_2$, such that $q(r_1) = q(r_2)$. Notice that the assumption that t the same AA repeated in all its sites simplifies the calculation of the binding energy $E(t, q)$ in (S40) to the addition of only k_q elements

$$\begin{aligned}
U(t, q) &= \sum_{i,j} \mathbb{W}_{ij} E_{t(i)q(j)} \\
&= \sum_{j=1}^{k_q} \left(\sum_{i=1}^{k_t} \mathbb{W}_{ij} \right) E_{t_r q(j)} \\
&= \sum_{j=1}^{k_q} \tilde{\mathbb{W}}_j E_{t_r q(j)} = \sum_{j=1}^{k_q} \tilde{\mathbb{W}}_j E_j \\
&= (\tilde{\mathbb{W}}_{r_1} + \tilde{\mathbb{W}}_{r_2}) E_r + \sum_{j \neq r_1, r_2}^{k_q} \tilde{\mathbb{W}}_j E_j,
\end{aligned} \tag{S47}$$

where $\tilde{W}_j = \sum_{i=1}^{k_t} W_{ij}$ becomes the weight of the contribution of the pairwise interacting energy $E_{t,r,q(j)} = E_j$ of AAs t_r and $q(j)$. Note that \tilde{W}_j can be calculated directly from the addition of the elements in the j -th column of the contact map (W_{ij}). We assign the same symbol to the energy of the repeated AAs $E_{t_r,q(r_1)} = E_{t_r,q(r_2)} \equiv E_r$. When E_j has unit variance $\sigma^2 = 1$ normal distribution PDF, the resulting variance (S43) reduces to

$$\text{Var}(U[C_1 = (2, 1^5)]) = \left[(\tilde{W}_{r_1} + \tilde{W}_{r_2})^2 + \sum_{j \neq r_1, r_2}^{k_q} \tilde{W}_j^2 \right] \sigma^2 = (\tilde{W}_{r_1} + \tilde{W}_{r_2})^2 + \sum_{j \neq r_1, r_2}^{k_q} \tilde{W}_j^2 \quad (\text{S48})$$

This variance (S48) depends on the sites of the repeating AAs $\{r_1, r_2\}$. Across the peptides with repeat structure C_1 in a randomly generated repertoire, sites r_1 and r_2 vary among peptides. However, in a randomly generated selecting repertoire all allowed combinations for $\{r_1, r_2\}$ are expected to be equally occurring, therefore, under central limit theorem, the expected value of $\text{Var}(U[C_1])$ converges to the average of (S48) across all possible permutations for $\{r_1, r_2\}$; there are $n_p = \binom{7}{2} = 21$ permutations for our choice of contact map. One can use (S48) n_p times replacing the appropriate values of \tilde{W}_{r_1} and \tilde{W}_{r_2} for each permutation, and average the resulting values.

Provided the $k_q = 7$ values of $\tilde{W}_j \in \{1.0000, 4.9806, 2.0100, 4.0006, 3.9999, 0.9988, 4.4190\}$, the average $\text{Var}(U[C_1]) = 99.8467$. From here, the average standard deviation is $\sigma_1 = 9.9923$; this value differs by 0.58 % from the estimations in simulations (see figure S7 top left panel).

S7 Point Mutated Variants: Constant TCR

To simplify calculations and to match our analysis of variance, we first assume that the TCR is a repeated amino acid. We also assume that a self-peptide is known, so that the repeat structure $(Z_1, Z_2, \dots, Z_M) = (z_1, z_2, \dots, z_M)$ is also known. Let X_1, X_2, \dots, X_M be their corresponding random energies, with X_i IID $\mathcal{N}(0, \sigma^2)$. We condition on a particular point-mutant substitution at amino acid i to j , with $1 \leq i, j \leq M$ (determined by the location in the contact map and entity of the AA at that position) with k total contacts. Letting

$$Y_{i,j} \equiv \sum_{\ell \in Q} z_\ell X_\ell; \quad Q = \{1, 2, \dots, M\} \setminus \{i, j\}, \quad (\text{S49})$$

we can express the event that the TCR survives selection on self-peptide as:

$$\mathcal{S}_Y = [z_1 X_1 + z_2 X_2 + \dots + z_M X_M \leq U_n] = [Y_{i,j} + z_i X_i + z_j X_j \leq U_n]. \quad (\text{S50})$$

Similarly, the event that the TCR recognizes the point-mutated neopeptide may be written as:

$$\mathcal{R}_Y = [Y_{i,j} + (z_i - k)X_i + (z_j + k)X_j > U_n]. \quad (\text{S51})$$

We remark that $Y_{i,j} \sim \mathcal{N}(0, \sigma^2 \sum_{\ell \in Q} z_\ell^2)$ with pdf f_Y . Conditioning on $Y_{i,j}$ gives us the desired probability:

$$\mathbb{P}(\mathcal{R} | \mathcal{S}) = \frac{\int_{\mathbb{R}} \mathbb{P}(\mathcal{R}_y \cap \mathcal{S}_y) f_Y(y) dy}{\int_{\mathbb{R}} \mathbb{P}(\mathcal{S}_y) f_Y(y) dy}. \quad (\text{S52})$$

The region $\mathcal{R}_y \cap \mathcal{S}_y \subset \mathbb{R}^2$ is located below $z_i X_i + z_j X_j = U_n - y$ and above $(z_i - k)X_i + (z_j + k)X_j = U_n - y$. Their intersection is at $X_i = X_j = \frac{U_n - y}{z_i + z_j} \equiv C_y$. Expressing these lines as functions of X_i gives

$$\ell_U(x_i) = \frac{U_n - y - z_i x_i}{z_i + z_j}; \quad \ell_L(x_i) = \frac{U_n - y - (z_i - k)x_i}{z_j + k} \quad (\text{S53})$$

Furthermore, the joint distribution of X_i and X_j is

$$f(x_i, x_j) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_i^2 + x_j^2)}{2\sigma^2}}. \quad (\text{S54})$$

Thus, we may write

$$\mathbb{P}(\mathcal{R}_y \cap \mathcal{S}_y) = \int_{-\infty}^{C_y} \int_{\ell_L(x_i)}^{\ell_U(x_i)} f(x_i, x_j) dx_j dx_i, \quad (\text{S55})$$

and

$$\mathbb{P}(\mathcal{S}_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\ell_U(x_i)} f(x_i, x_j) dx_j dx_i. \quad (\text{S56})$$

Numerical approximation can be implemented for integration over $M \times M \subset \mathbb{R}^2$ choosing M sufficiently large to obtain the desired convergence

$$\int_{-M}^M \frac{\int_{-M}^{C_y} \int_{\ell_L(x_i)}^{\ell_U(x_i)} f(x_i, x_j) dx_j dx_i}{\int_{-M}^M \int_{-M}^{\ell_U(x_i)} f(x_i, x_j) dx_j dx_i} f_Y(y) dy. \quad (\text{S57})$$

S7.1 Point mutants on distinct TCR/peptides

If the amino acids are non-repeated on both the peptide and amino acid sides, then each of the N_C contacts in the contact map contribute an independent and statistically identical $\mathcal{N}(0, \sigma^2)$ energy. Assume that there are k contacts which can change by point-mutating one peptide amino acid. Let X_{N_C-k} denote those energies which do not change with the mutation, and X_k the pre-mutated energies and \tilde{X}_k the post-mutated energies. Their corresponding distributions are:

$$X_{N_C-k} \sim \mathcal{N}(0, (N_C - k)\sigma^2), \quad X_k, \tilde{X}_k \sim \mathcal{N}(0, k\sigma^2). \quad (\text{S58})$$

Let $F_\ell(x)$ and $f_\ell(x)$ denote the CDF and PDF of normal random variables with standard deviation $\sigma\sqrt{\ell}$. The conditional probability that a point-mutated antigen is recognized by a TCR conditioned on that TCR not recognizing the non-mutated antigen may be expressed as:

$$\mathbb{P}(X_{N_C-k} + \tilde{X}_k > U_n \mid X_{N_C-k} + X_k \leq U_n). \quad (\text{S59})$$

Conditioning on the distributions of X_k, \tilde{X}_k , and noting that the joint distributions factor for independent random variables, we have, by definition of conditional probability:

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{P}(U_n - \tilde{x} < X_{N_C-k} < U_n - x) f_k(\tilde{x}) f_k(x) d\tilde{x} dx \Big/ \mathbb{P}(X_{N_C-k} + X_k \leq U_n). \quad (\text{S60})$$

The denominator of Eq. [S60](#) is just $F_{N_C}(U_n)$. The numerator can be further simplified as:

$$D = \int_{\mathbb{R}} \int_{\mathbb{R}} [F_{N_C-k}(U_n - x) - F_{N_C-k}(U_n - \tilde{x})] \cdot \mathbb{1}_{[x < \tilde{x}]} f_k(\tilde{x}) f_k(x) d\tilde{x} dx, \quad (\text{S61})$$

where

$$\mathbb{1}_E = \begin{cases} 1, & \text{on } E; \\ 0, & \text{on } E^c. \end{cases} \quad (\text{S62})$$

Thus,

$$D = \int_{\mathbb{R}} \int_x^{\infty} \left(F_{N_C-k}(U_n - x) - F_{N_C-k}(U_n - \tilde{x}) \right) f_k(\tilde{x}) f_k(x) d\tilde{x} dx \quad (\text{S63})$$

$$= \int_{\mathbb{R}} \left(F_{N_C-k}(U_n - x) [1 - F_k(x)] - \int_x^{\infty} F_{N_C-k}(U_n - \tilde{x}) f_k(\tilde{x}) d\tilde{x} \right) f_k(x) dx. \quad (\text{S64})$$

Observing that

$$\int_{\mathbb{R}} F_{N_C-k}(U_n - x) f_k(x) dx = \int_{\mathbb{R}} \mathbb{P}(X_{N_C-k} \leq U_n - x) f_k(x) dx \quad (\text{S65})$$

$$= \mathbb{P}(X_{N_C-k} \leq U_n + X_k) = F_k(U_n), \quad (\text{S66})$$

so that Eq. [S64](#) reduces to

$$F_{N_C}(U_n) - \int_{\mathbb{R}} \left(F_{N_C-k}(U_n - x) F_k(x) + \int_x^{\infty} F_{N_C-k}(U_n - \tilde{x}) f_k(\tilde{x}) d\tilde{x} \right) f_k(x) dx. \quad (\text{S67})$$

Plots of this conditional recognition probability are plotted as a function of thymic negative selection cutoff U_n using $N_C = 21$ in 3QIB α CDR3 for a variety of contacts assumed in the peptide in Fig. [S9d](#).

In the main manuscript we discussed how the point-mutant recognition probability is influenced by selection stringency. We follow-up that discussion with a more detailed explanation. Selection stringency is driven by many factors including the affinity threshold, size of the selecting repertoire, diversity of the selecting repertoire, etc, each of which bias the post-selection TCRs towards non self-reactivity. At fixed selection repertoire size, say $Nq = 10^4$, lenient selection (high negative selection survival probability) is usually accomplished by choosing a high affinity threshold (U_n). This case gives rise to low point-mutant recognition probability, because of the increased baseline unlikeliness of any TCR binding to a pMHC, Fig. [S9](#). In this regard, recall that we use for defining successful binding the same threshold as used in the negative selection step. On the other end, stringent selection (low negative selection survival probability) means having a low affinity threshold. Now, the baseline binding probability is higher, giving a TCR a better chance to recognize a point mutant ($U(t, \tilde{q}) > U_n$), driving higher point-mutant recognition probability. In other words, the closeness of point mutants to self-peptides reduces the chances of detection by roughly a fixed amount and hence the resulting detection curve roughly tracks the baseline binding probability.

S8 Self-Averaging of pairwise energies

Given a particular contact map \mathbb{W} , symmetric pairwise amino acid interaction matrix \mathbb{E} , and given random TCR $\{t_1, \dots, t_n\}$ and peptide $\{p_1, \dots, p_n\}$, we can represent the total energy of interaction as a sum of (fixed!) energies $e_k \in \mathbb{E}$, with corresponding repeats r_k such that $\sum_k r_k = |\mathbb{W}| = N_C$. We assume there are N distinct repeats so that the energy is given by

$$U(t, q) = \sum_{k=1}^N r_k e_k. \quad (\text{S68})$$

Conditional on a realized repeat pattern and energy choices $\{r_k, e_k\}_{k=1}^N$, we are interested in understanding the mean and variance of first- and second-moments of U over the set of possible amino acid choices. Because each e_k may take values in \mathbb{E} with $|\mathbb{E}| \equiv M = 210$, we will need to sum over all realizations of these pairs. Let \mathcal{F} be the set of all ordered realizations of the $\{e_1, \dots, e_P\}$ energy terms taking values in \mathbb{E} without replacement. Then $|\mathcal{F}| = M!/(M-N)!$ and each ordered realization $F \in \mathcal{M}$ is equally likely so that $\mathbb{P}(F) = 1/|\mathcal{M}|$.

Denote by $\mathbb{A}_{\mathcal{F}}[\cdot]$ the operator which averages over all possible realizations of $F \in \mathcal{F}$. In contrast with the expectation operator $\langle \cdot \rangle$ assigning a real number to the random variable U , $\mathbb{A}_{\mathcal{F}}[U]$ is itself a random variable. Averaging U over the finite realizations in \mathcal{F} gives,

$$\mathbb{A}_{\mathcal{F}}[U] = \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} \sum_{k=1}^N r_k e_k = \sum_{k=1}^N r_k \cdot \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} e_k. \quad (\text{S69})$$

Note that for fixed k ,

$$\sum_{F \in \mathcal{F}} e_k = \frac{(M-1)!}{(M-N)!} \sum_{i=1}^M e_i, \quad (\text{S70})$$

since taking the sum of e_k over all $M!/(M-N)!$ possible sequence realizations gives $(M-1)!/(M-N)!$ repeats of each of M possible choices for e_k . This implies that the right hand side of Eq. [S69](#) becomes

$$\mathbb{A}_{\mathcal{F}}[U] = \sum_{k=1}^N r_k \cdot \frac{1}{|\mathcal{F}|} \frac{(M-1)!}{(M-N)!} S_{k,M} = \sum_{k=1}^N r_k \cdot S_{k,M}/M, \quad (\text{S71})$$

where $S_{k,M}$ is a sum of M IID $\mathcal{N}(0, \sigma^2)$ random variables. Thus, for $\gamma \equiv \sum_{k=1}^N r_k^2/M$, we have that

$$\mathbb{A}_{\mathcal{F}}[U] \sim \mathcal{N}\left(0, \sigma^2 \sum_{k=1}^N \frac{r_k^2}{M}\right) \equiv \mathcal{N}(0, \gamma \sigma^2). \quad (\text{S72})$$

Note that this quantifies the deviation from a much simpler calculation of the expected value assuming the fixed energies e_k in Eq. [S68](#) are idealized by IID random variables given by their common distribution $E_k \sim \mathcal{N}(0, \sigma^2)$, which clearly yields

$$\left\langle \sum_{k=1}^N r_k E_k \right\rangle = 0 = \langle \mathbb{A}_{\mathcal{F}}[U] \rangle. \quad (\text{S73})$$

In the most extreme case, $N = 1$ and $r_k = N_C = |\mathbb{W}|$ so that $\gamma = N_C^2/M$. Using Eq. [S72](#) this equates to $\gamma = 1.904$ for 20 contacts. A majority of cases are however far from here. As examples utilizing the 3QIB α system with $n = 7$, the partitions whose union occurs with probability $> 99\%$ and their associated variance scales are given in Table [S3](#). Because these factors scale the variance so tightly, it is reasonable to represent the distribution $\mathbb{A}_{\mathcal{F}}[U]$ by its expected value, which corresponds to the expected value of the idealized quantity with e_k replaced by E_k as in Eq. [S73](#).

Partition	(1^7)	$(2, 1^5)$	$(2^2, 1^3)$	$(3, 1^4)$	$(3, 2, 1^2)$
γ	0.0324	0.0428	0.0524	0.0620	0.0713

Table S3: Partitions represented by their repeat structure are given with their corresponding variance scales γ as in Eq. [S72](#).

We desire to investigate whether an analogous result holds for the second moment calculation.

That is, we wish to compare $\mathbb{A}_{\mathcal{F}} [U^2]$ with $\langle \mathbb{A}_{\mathcal{F}} [U^2] \rangle$ and $\text{Var} (\mathbb{A}_{\mathcal{F}} [U^2])$. Note that,

$$\mathbb{A}_{\mathcal{F}} [U^2] = \mathbb{A}_{\mathcal{F}} \left[\sum_{k=1}^N (r_k e_k)^2 + \sum_{i \neq j} r_i r_j e_i e_j \right] \quad (\text{S74})$$

$$= \sum_{k=1}^N r_k^2 \cdot \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} (\sigma \tilde{e}_k)^2 + \sum_{i \neq j} r_i r_j \cdot \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} (\sigma \tilde{e}_i)(\sigma \tilde{e}_j), \quad \text{for } \tilde{e}_k = e_k/\sigma \sim \mathcal{N}(0, 1) \quad (\text{S75})$$

$$= \sigma^2 \sum_{k=1}^N r_k^2 \frac{1}{|\mathcal{F}|} \cdot \frac{(M-1)!}{(M-N)!} \tilde{S}_{k,M} + \sigma^2 \sum_{i \neq j} r_i r_j \cdot \frac{1}{|\mathcal{F}|} \cdot \frac{(M-2)!}{(M-N)!} \tilde{T}_{k,M}, \quad (\text{S76})$$

$$= \sigma^2 \sum_{k=1}^N r_k^2 \tilde{S}_{k,M}/M + \sigma^2 \sum_{i \neq j} r_i r_j \tilde{T}_{k,M}/M(M-1), \quad (\text{S77})$$

for $\tilde{S}_{k,M} \sim \chi^2(M) = \text{Gamma}(M/2, 2)$ the sum of M IID squared standard normals, and with $\tilde{T}_{k,M}$ the sum of $M(M-1)$ IID random variables of the form XY , where X, Y are IID standard normals. Unlike in the first moment case, we cannot represent the second sum as a standard distribution. However, we can still evaluate the mean and variance of $\mathbb{A}_{\mathcal{F}} [U^2]$. From the above, Eq. [S76](#) may be written in the form

$$\mathbb{A}_{\mathcal{F}} [U^2] = \sigma^2 \left(\sum_{k=1}^N \tilde{X}_k + \sum_{i \neq j} \tilde{Y}_{i,j} \right) \quad (\text{S78})$$

with

$$\langle \tilde{X}_k \rangle = r_k^2, \quad \text{Var} (\tilde{X}_k) = 2r_k^4/N; \quad \langle \tilde{Y}_{i,j} \rangle = 0, \quad \text{Var} (\tilde{Y}_{i,j}) = \frac{M(M-1) \cdot 1}{[M(M-1)]^2} = \frac{1}{M(M-1)}. \quad (\text{S79})$$

or alternatively,

$$\left\langle \sum_{k=1}^N \tilde{X}_k \right\rangle = \sum_{k=1}^N r_k^2, \quad \text{Var} \left(\sum_{k=1}^N \tilde{X}_k \right) = \frac{2}{M} \sum_{k=1}^N r_k^4/M; \quad (\text{S80})$$

$$\left\langle \sum_{i \neq j} \tilde{Y}_{i,j} \right\rangle = 0, \quad \text{Var} \left(\sum_{i \neq j} \tilde{Y}_{i,j} \right) = \frac{N(N-1)}{M(M-1)}. \quad (\text{S81})$$

Thus, we may write

$$\langle \mathbb{A}_{\mathcal{F}} [U^2] \rangle = \sigma^2 \|r\|_2^2; \quad \text{Var} (\mathbb{A}_{\mathcal{F}} [U^2]) = \frac{\sigma^4}{M} \left(2\|r\|_4^4 + N \frac{N-1}{M-1} \right), \quad (\text{S82})$$

using the usual l^p -norm

$$\|r\|_p \equiv \left(\sum_{k=1}^N |r_k|^p \right)^{1/p}. \quad (\text{S83})$$

Now, we may calculate the coefficient of variation, c_v , by:

$$c_v = \frac{\sqrt{\text{Var} (\mathbb{A}_{\mathcal{F}} [U^2])}}{\langle \mathbb{A}_{\mathcal{F}} [U^2] \rangle} \quad (\text{S84})$$

$$= \left(\sqrt{\frac{2}{M}} \right) \sqrt{\frac{\|r\|_4^4}{\|r\|_2^4} + \frac{N(N-1)}{(M-1)\|r\|_2^4}}. \quad (\text{S85})$$

We may achieve a convenient upper-bound on c_v by

$$c_v \leq \left(\sqrt{\frac{2}{M}} \right) \sqrt{\max_r \frac{\|r\|_4^4}{\|r\|_2^4} + \max_r \frac{N(N-1)}{(M-1)\|r\|_2^4}} \quad (\text{S86})$$

For the first quantity, we observe that

$$\|r\|_4^4 = \sum_{k=1}^N r_k^4 \leq \sum_{k=1}^N r_k^4 + \sum_{i \neq j} r_i r_j = \left(\sum_{k=1}^N r_k^2 \right)^2 = \|r\|_2^4 \quad (\text{S87})$$

with equality holding in Eq. [S87](#) whenever there is a single $r = N_C$, which implies that the left maximum of Eq. [S86](#) is equal to 1. In contrast, the denominator of the second term is minimized whenever all partitions are unique so that $r_k = 1$, for $k = 1, \dots, N = N_C$, and $\sum_k r_k = N_C = \|r\|_2^2$. In this case, $\|r\|_2^4 = N^2$, which implies that the right maximum of Eq. [S86](#) is $1/(M-1)$, ultimately yielding

$$c_v \leq \sqrt{\frac{2}{M-1}} = 9.78\% \quad (\text{S88})$$

for M determined by the 20-by-20 symmetric amino acid energy matrix. Estimates of c_v for the most commonly occurring partitions are given in Table [S4](#). Eq. [S88](#) represents a conservative bound for most common partitions. As in the case of first moment calculations, we obtain agreement in the expectation taken over the finite realized e_k and the idealized distributions E_k :

$$\left\langle \left(\sum_{k=1}^N r_k E_k \right)^2 \right\rangle = \sigma^2 \sum_{k=1}^N r_k^2 = \langle \mathbb{A}_{\mathcal{F}} [U^2] \rangle \quad (\text{S89})$$

This, together with the small coefficient of variation implies that $\mathbb{A}_{\mathcal{F}} [U^2]$ may also be associated with its mean value by neglecting a small amount of fluctuation relative to its mean.

Partition	(1 ⁷)	(2, 1 ⁵)	(2 ² , 1 ³)	(3, 1 ⁴)	(3, 2, 1 ²)
c_v	3.74%	4.99%	5.26%	6.92%	6.48%

Table S4: Partitions represented by their repeat structure are given with their corresponding coefficient of variation, c_v , calculated from Eq. [S85](#).

S9 Figures

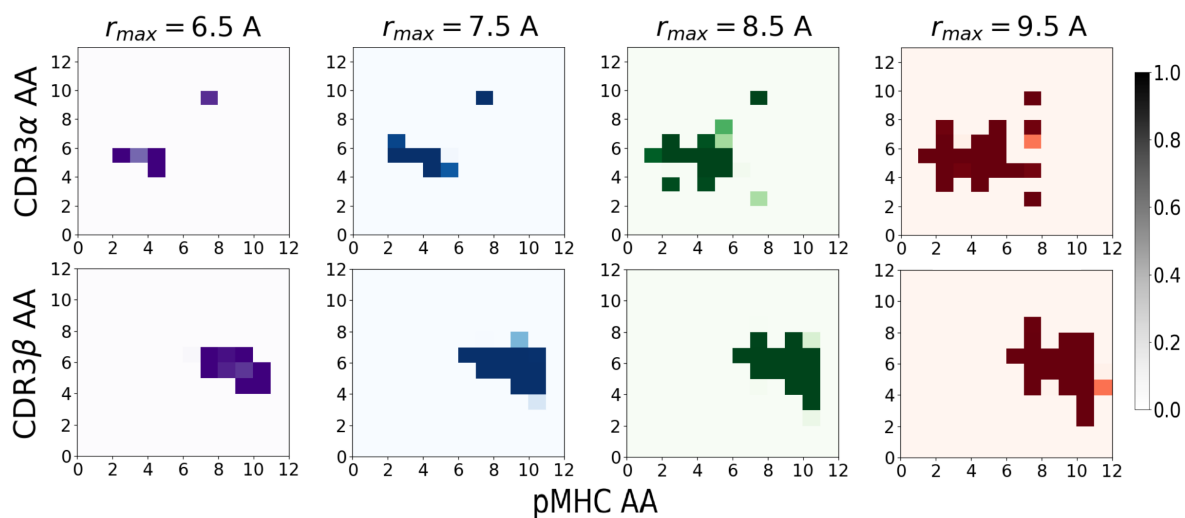


Figure S1: Number of contacts in a contact map increases with increasing values of distance cut-off, r_{max} . Contact maps for CDR3 α -pMHC (top row) and CDR3 β -pMHC (bottom row) interfaces of crystal structure PDB ID 3QIB are plotted at four cut-off distances, $r_{max} = \{6.5, 7.5, 8.5, 9.5\}$ Å. As r_{max} increases, the distance at which two AAs are considered to be in contact also increases (S2), resulting in more contacts for higher values of r_{max} .

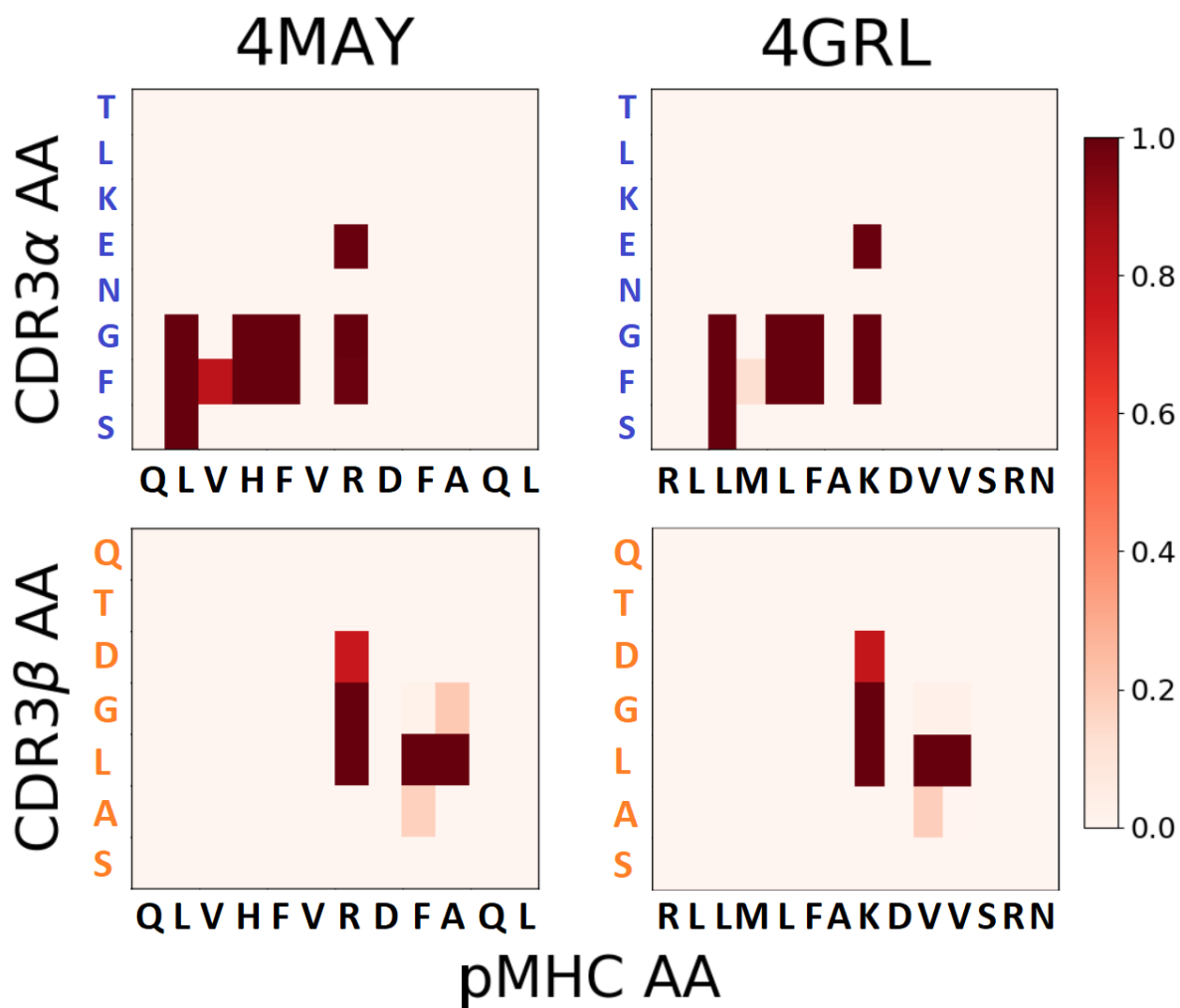


Figure S2: The topology of a contact map is mostly preserved when a TCR binds to MHC molecules with the same allele restriction. In both cases, PDB ID 4MAY (left) and 4GRL (right), the contact maps show a similar topology. In these contact maps, human T cells hy.1B11 are bound to MHC type II HLA-DQ1 molecules and show cross-reactivity to two different antigens: *Herpes simplex virus*, UL15_{154–166} (4MAY); and *Pseudomonas aeruginosa*, PMM_{260–274} (4GRL).

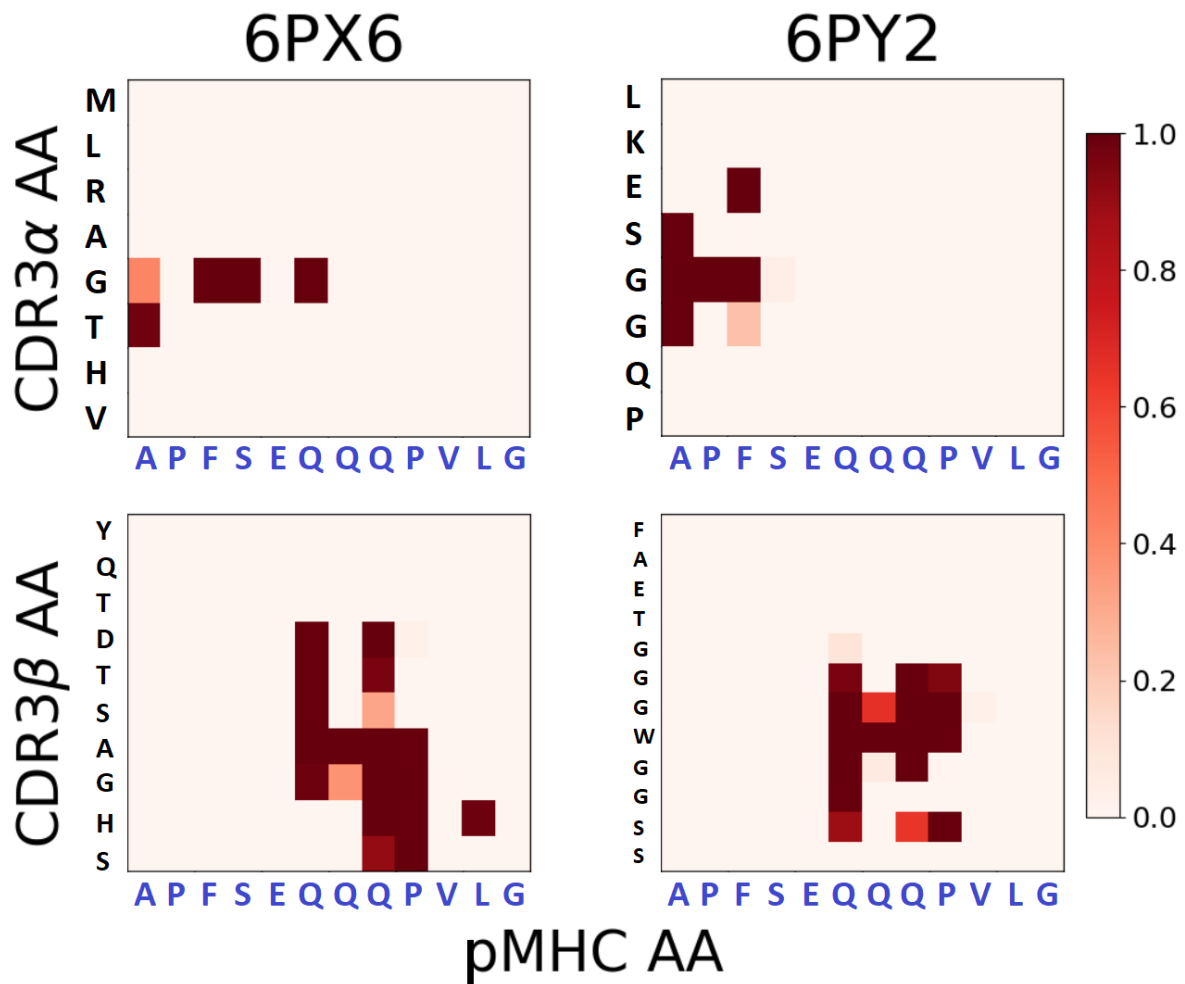


Figure S3: The contact map topology differs for two different TCR-MHC pairings even under the MHC restriction. Crystal structures PDB ID 6PX6 and PDB ID 6PY2 correspond to two different human TCRs, T1005.2.56 and T594, respectively, bound to the same MHC type II HLA-DQ2.2 molecule presenting the same DQ2.2-glut-L1 *Triticum aestivum* peptide.

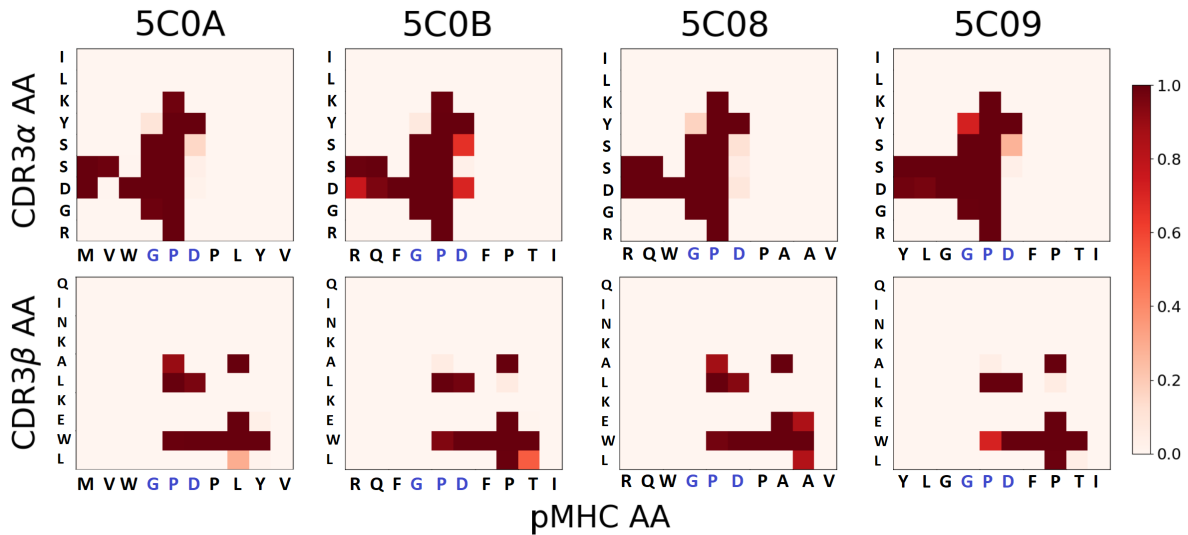


Figure S4: Contact map topology is mostly preserved for the same TCR-MHC pairing. The four contact maps shown belong to crystal structures of human 1E6 TCR bound to MHC type I HLA-A02, the peptides presented by the MHC complex in these four cases have common AAs G, P, and D in sites 4-6 (highlighted in blue).

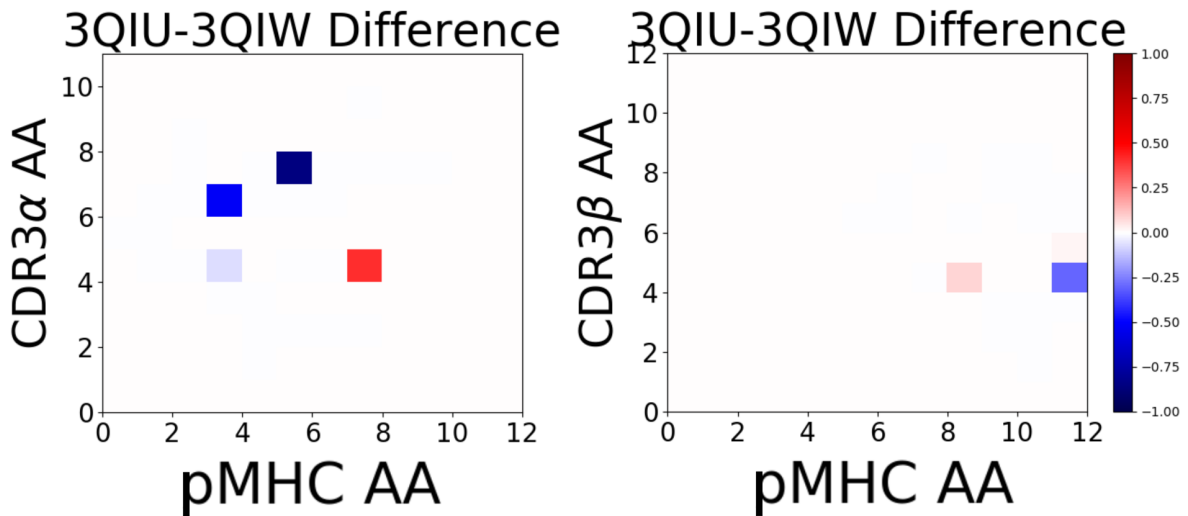


Figure S5: Difference between the contact maps of CDR3 α -pMHC (left) and CDR3 β -pMHC (right) binding interfaces, obtained by directly subtracting one contact map matrix from the other.

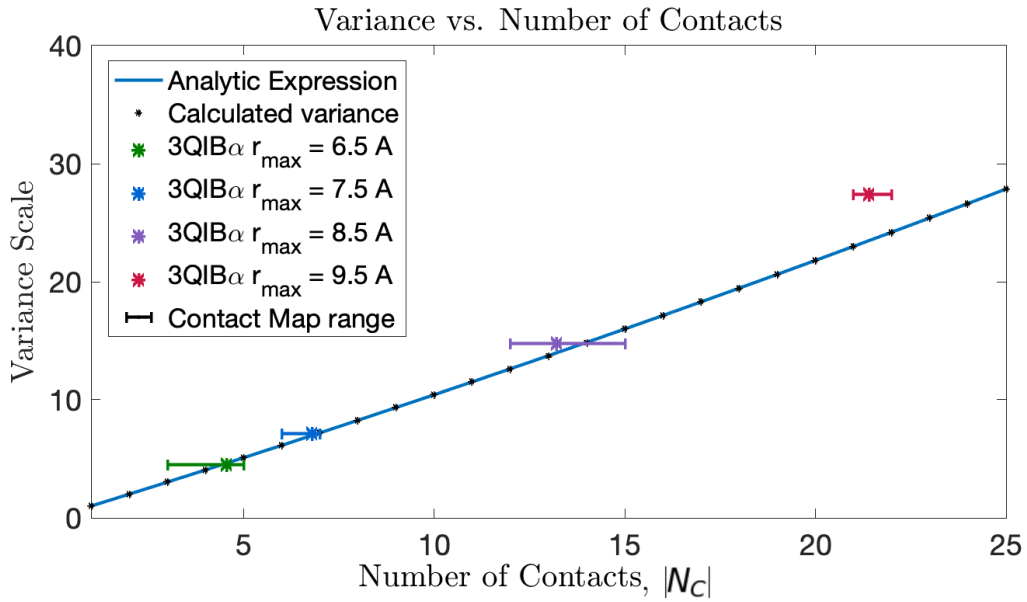


Figure S6: TCR-pMHC energy $U(t, q)$ variance scales with the number of contacts. Analytical vs real variances as a function of total contacts. Variance scale as a function of contact number. This plots the scale factor to σ^2 as a function of total contact number. Agreement is seen for summing over all possible partitions (black dots, calculation given by Eq. S34) and the analytic solution (blue dashed line, calculation given by Eq. S37). These values are compared to exact simulations involving variances in the contact maps of real examples.

PDF for TCR-pMHC binding energy. 3QIB crystal structure, CDR3 α

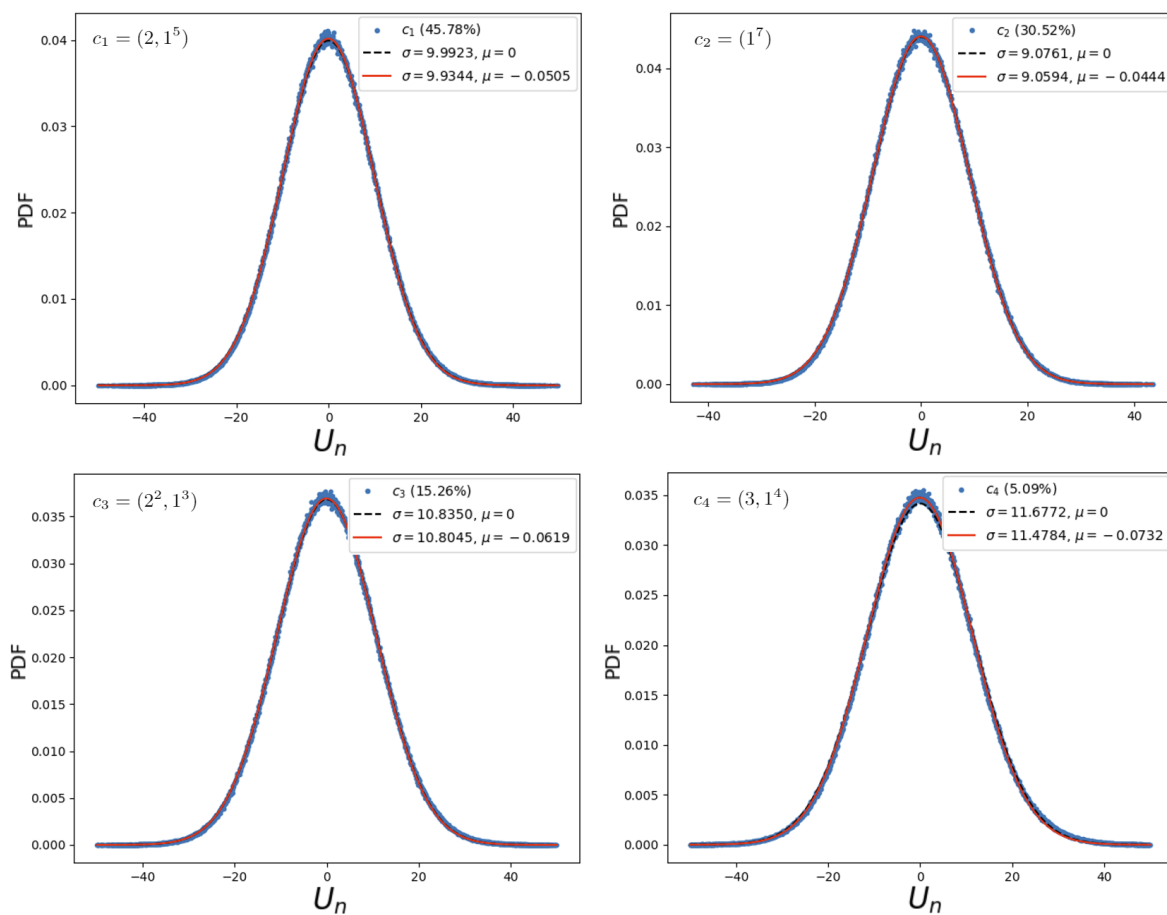


Figure S7: Normal distribution PDF for TCR-pMHC binding energy in the four most likely repeat structures from a randomly generated pMHC repertoire. Each repeat structure c_n , $n = 1, 2, 3, 4$, has a probability shown in parenthesis. In each panel, blue dots are simulated values from TCRs undergoing negative selection against pMHCs with the same repeat structure, dashed black lines show theoretical predictions, and solid red lines are best least-square fit.

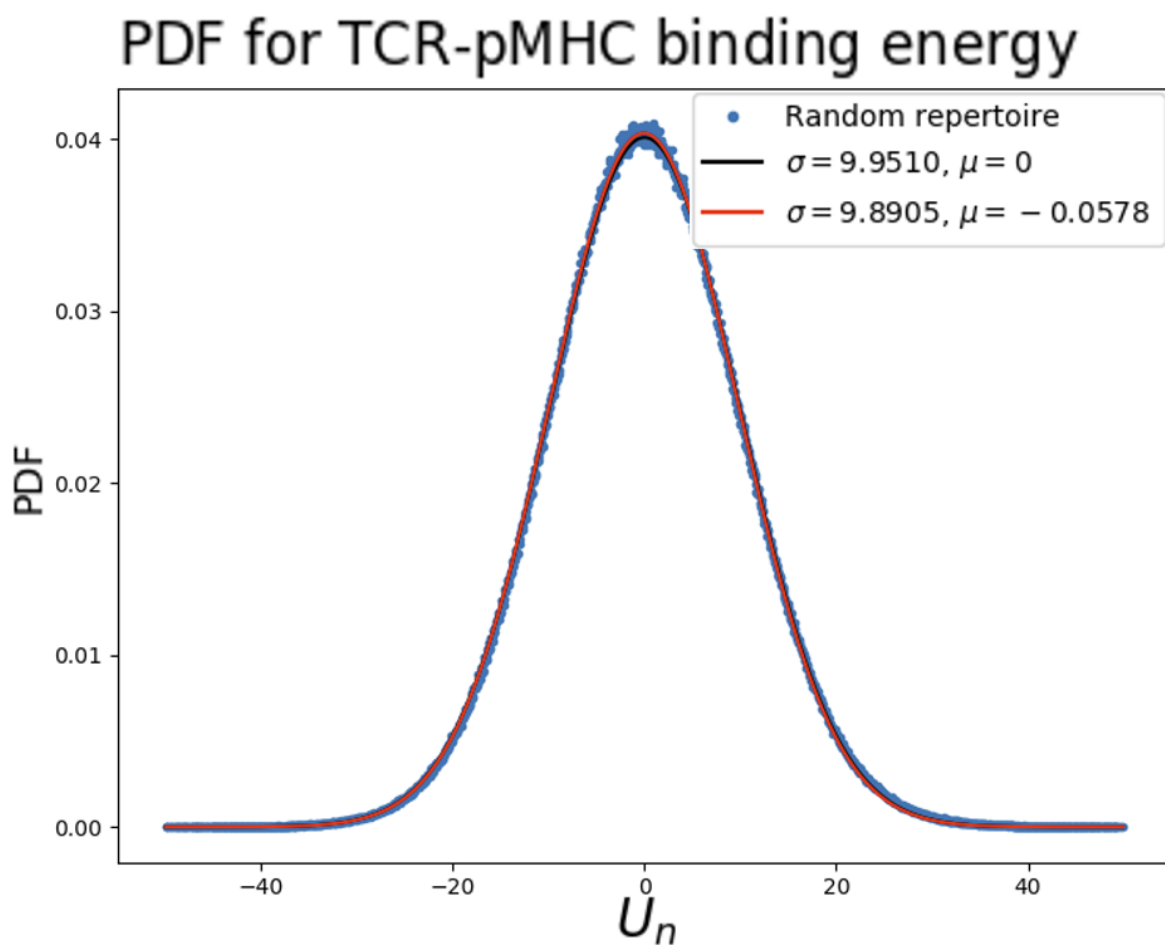


Figure S8: Normal distribution PDF for TCR-pMHC binding energy between TCRs of all-repeated AAs against a randomly generated selecting peptide repertoire. Crystal structure 3QIB's CDR3 α -pMHC interface is used as the contact map. Blue dots are simulated data, solid red line corresponds to least-square fit and dashed black line is the theoretical prediction.

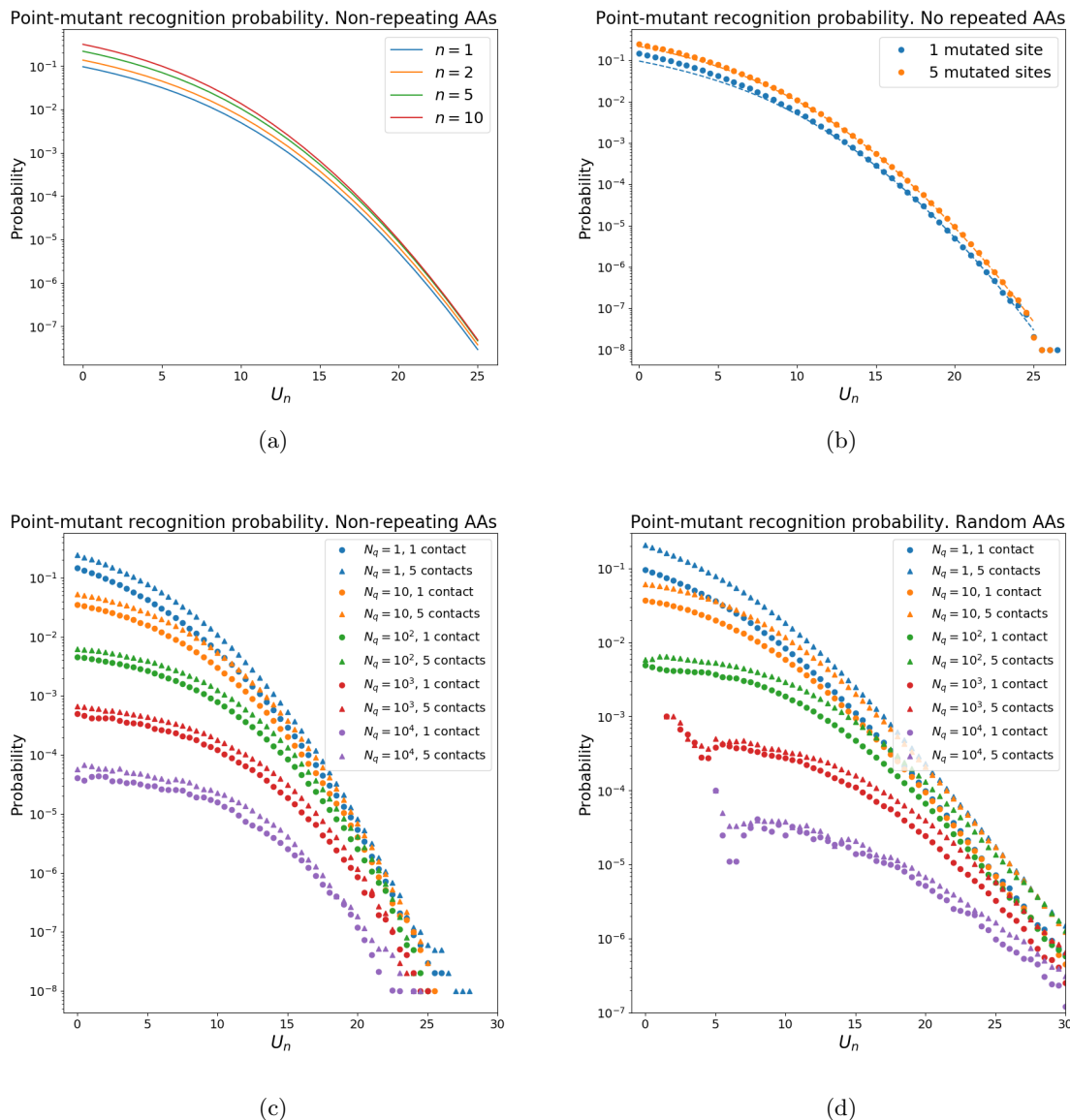


Figure S9: Point-mutant recognition probability changes with number of mutated contacts and with the size of the selecting repertoire. Panel [S9a](#) shows the change in recognition probability when the point mutation occurs in a site with $n = \{1, 2, 5, 10\}$ contacts, with overall probability increasing with n due to more mutated contacts making the mutant peptide appear more like a foreign antigen. Panel [S9b](#) shows good agreement between predicted point-mutant recognition probability (dashed lines) and simulations (dots) for $n = 1$ and $n = 5$ mutated contacts, peptides are generated with no repeated amino acids in their sequences. Panels [S9c](#) and [S9d](#) show simulated point-mutant recognition probabilities at $N_q = 1$ (blue), 10 (orange), 10^2 (green), 10^3 (red), and 10^4 (purple) selection repertoire sizes, with point mutations affecting $n = 1$ (dots) and $n = 5$ (triangles) contacts in each N_q case; in [S9c](#) peptide sequences are generated with no repeated amino acids, whereas in [S9d](#) the peptide sequences are randomly generated.