Editors' Suggestion

Contact map dependence of a T-cell receptor binding repertoire

Kevin 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

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, CDR3α and CDR3β. 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,

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

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

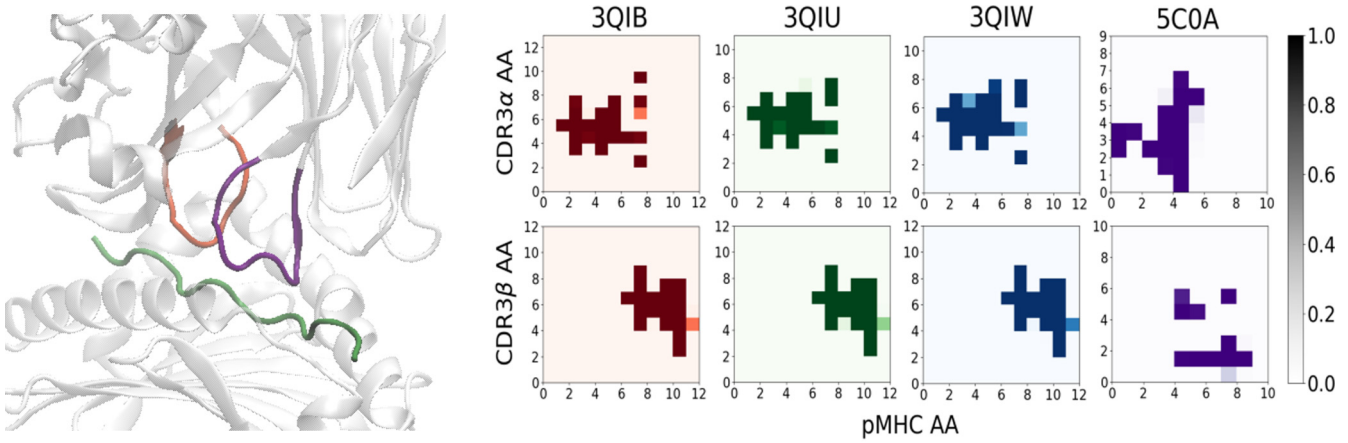

FIG. 1. The TCR-pMHC interface and contact maps. (a) The CDR3-pMHC interface in the crystal structure of the 2B4 TCR binding to the MCC/I-E^k complex (PDB ID 3QIB); with the antigen MCC highlighted in green, the CDR3α loop in purple, and the CDR3β loop in orange. (b) Eight contact maps estimated from four crystal structures, contact maps of the CDR3α-pMHC (CDR3β-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,

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

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}, $i = 1, \ldots, k_t$, $j = 1, \ldots, k_q$. We then compute the corresponding weights W_{ij} from (2) and construct the contact map $\mathbb{W} = (W_{ij})$. Given that both CDR3α and CDR3β 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 $r_{\max} = 9.5$ Å and $\eta = 1$ Å^{-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 MVWGPDPLYV peptide of the Bacteroides fragilis/thetaiotaomicron 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 CDR3α-pMHC interface is plotted at four different $r_{\max}$ values, from 6.5 to 9.5 Å in 1 Å increments, while keeping $\eta = 1$ Å^{-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 $r_{\max} = 6.5$ Å, to about 22 AA pairs in contact at $r_{\max} = 9.5$ Å. For the remainder of this paper, all contact maps are calculated with $r_{\max} = 9.5$ Å and $\eta = 1$ Å^{-1}.

The contact maps in Fig. 1(b) correspond to CDR3α-pMHC interfaces (top row) and CDR3β-pMHC interfaces (bottom row) from crystal structures 3QIB, 3QIU, 3QIW, and 5C0A. The contact profiles of CDR3α-pMHC are different from the CDR3β-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 CDR3α loops primarily contact AAs 2–8 and the CDR3β 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, $E_{mn} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. 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_{mn} 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,

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

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 $C_p = (l_1^{r_1}, l_2^{r_2}, \ldots, l_N^{r_N})$, 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

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

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,

$$\sigma^2 N_C \leqslant \mathrm{Var}U \leqslant \sigma^2 N_C^2. \tag{5}$$

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

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

See the SM [31] (Secs. S5 and S6) for a detailed derivation. In Fig. 2(a), the variances computed by simulation for the CDR3α-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,

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

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 CDR3α-pMHC contact map $\mathbb{W}^{\alpha}_{3\mathrm{QIB}}$ in Fig. 1(b) (top left) and assume that the TCR is a constant sequence of a single repeated AA $t = (t_1, t_1, t_1 \ldots)$. Note that

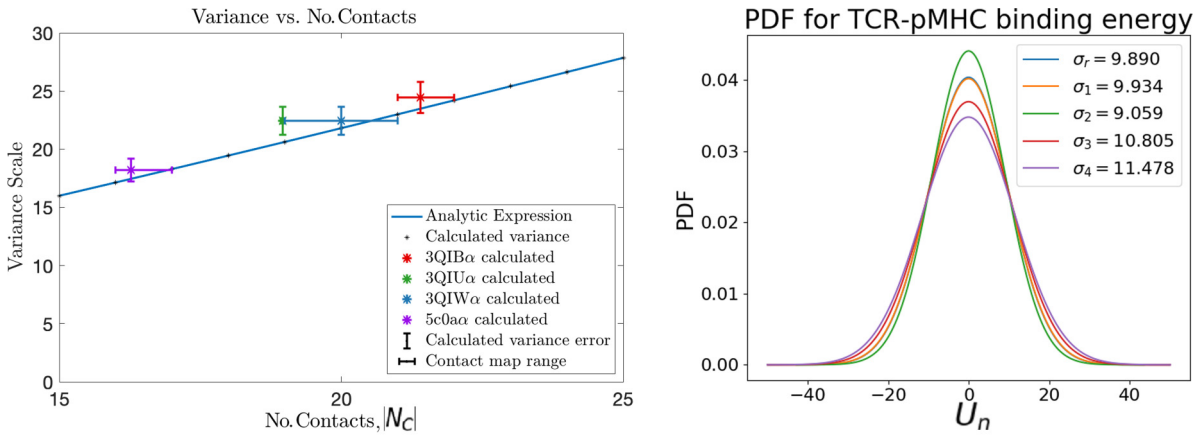

FIG. 2. The variance of the TCR-pMHC binding energy distribution depends on the total number of contacts and on the repeat structure allowed by the topology in the contact map. (a) Binding energy $U(t, q)$ variance scaling with the number of contacts, $|N_C|$; calculated variances with their variance (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 $\mathbb{W}^{\alpha}_{3\text{QIB}}$, 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 $C_{q,n}$ $(n = 1, \ldots, N)$ 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 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

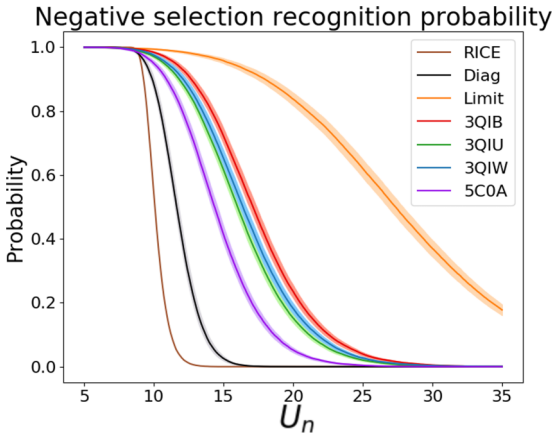

FIG. 3. Negative-selection recognition probability as a function of the survival energy threshold for T cells auditioning for negative selection. All curves involving the use of contact maps are generated from simulations sharing the same parameters apart from the contact maps. The prediction of the 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 5C0A; 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

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

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 CDR3α-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.)

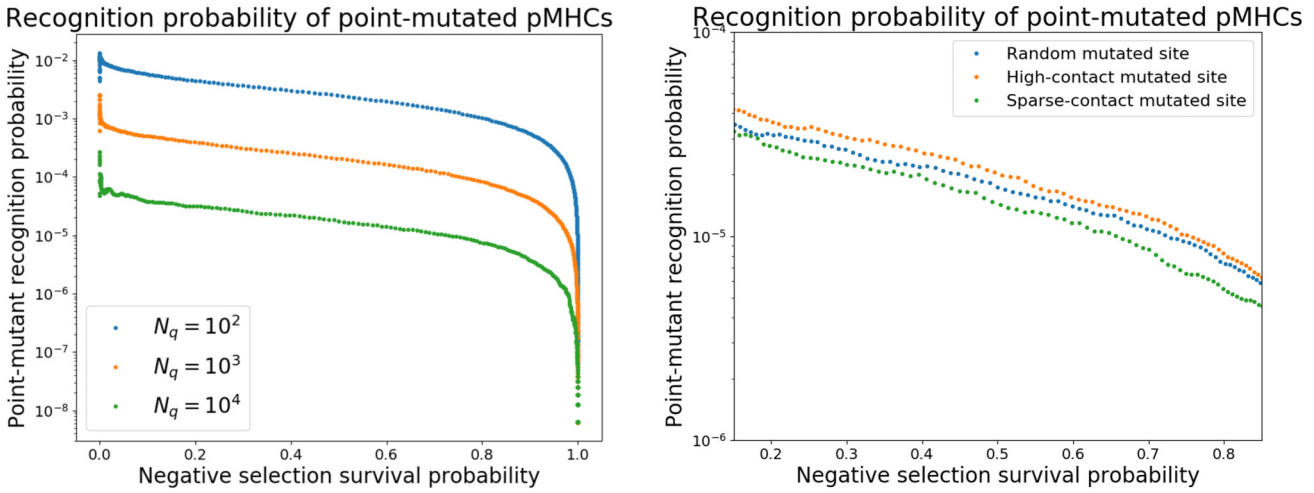

FIG. 4. Recognition probability of point-mutated peptides by T cells that have undergone negative selection. (a) The point-mutant recognition probability from simulations plotted for T cells that have received negative selection against self-peptide repertoires of three different sizes, $N_q = \{10^2, 10^3, 10^4\}$. (b) The point-mutant recognition probability from simulations that changed the site of the mutated AA; for the CDR3α-pMHC interface of 3QIB [top left panel of Fig. 1(b)] in use, pMHC-AAs in high-contact sites are in contact with 5 TCR-AAs, whereas pMHC-AAs in sparse-contact sites are in contact with only 1 TCR-AA; and when picking random sites to mutate, the number of peptide-AAs 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 CDR3α-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 probability 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 nonvanishing 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 variances 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. Marcatili, 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–i-ek, 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 interresidue contact energies from protein crystal structures: Quasichemical 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).