



Rapid assessment of T-cell receptor specificity of the immune repertoire

Xingcheng Lin^{1,2,3,12}, Jason T. George^{1,4,12}, Nicholas P. Schafer^{1,5}, Kevin Ng Chau⁶, Michael E. Birnbaum^{7,8,9}, Cecilia Clementi^{1,2,5,10}, José N. Onuchic^{1,2,5,11} ✉ and Herbert Levine^{1,6} ✉

Accurate assessment of T-cell-receptor (TCR)-antigen specificity across the whole immune repertoire lies at the heart of improved cancer immunotherapy, but predictive models capable of high-throughput assessment of TCR-peptide pairs are lacking. Recent advances in deep sequencing and crystallography have enriched the data available for studying TCR-peptide systems. Here, we introduce RACER, a pairwise energy model capable of rapid assessment of TCR-peptide affinity for entire immune repertoires. RACER applies supervised machine learning to efficiently and accurately resolve strong TCR-peptide binding pairs from weak ones. The trained parameters further enable a physical interpretation of interacting patterns encoded in each TCR-peptide system. When applied to simulate thymic selection of a major-histocompatibility-complex (MHC)-restricted T-cell repertoire, RACER accurately estimates recognition rates for tumor-associated neoantigens and foreign peptides, thus demonstrating its utility in helping address the computational challenge of reliably identifying properties of tumor antigen-specific T-cells at the level of an individual patient's immune repertoire.

The advent of strategies that unleash the host immune system to battle malignant cells represents one of the largest paradigm shifts in treating cancer and has ushered in a new frontier of cancer immunotherapy¹. Various treatments have emerged, including checkpoint blockade therapy², tumor antigen vaccine development³, and the infusion of donor-derived admixtures of immune cells⁴. Successful treatments to date mostly rely on the anti-tumor potential of the CD8⁺ T-cell repertoire, a collection of immune cells capable of differentiating between malignant cells and normal tissue by recognizing tumor-associated neoantigens (TANs) on the cell surface⁵. Therefore, accurately assessing a T-cell repertoire's ability to identify cancer cells by recognizing their tumor antigens lies at the heart of optimizing cancer immunotherapy.

A complete understanding of adaptive immune recognition and the tumor-immune interaction has remained a formidable task, owing in part to the daunting complexity of the system. For example, antigens and self-peptides contained in a space of 20⁹ epitopes (recognizable peptide sequences) are presented to ~10⁷ unique T-cell clones in each individual⁶, a small fraction of the upper limit of TCR diversity (~10²⁰)⁷. Moreover, their behavior is tempered via an elaborate thymic negative selection process in order to avoid auto-recognition⁸. Here, T-cell clones, each with uniquely generated TCRs, interface with numerous (~10⁴) self-peptides presented on the major histocompatibility complex (p-MHC) of thymic medullary epithelial cells via TCR complementary-determining region 3 (CDR3) α and β chains, and survive only if they do not bind too strongly⁹. This process, together with system-level peripheral tolerance¹⁰, imparts T cells with durable tolerance to major self-peptides and influences many of the recognition properties of the resultant repertoire. The complexity of the adaptive immune system has

attracted numerous mathematical modeling efforts quantifying the mechanisms underlying T-cell immune response. Collectively, the field has made significant progress in understanding the population-level effects of tolerance on T-cell recognition and self versus non-self discrimination^{9,11}. This includes the T-cell repertoire's effectiveness at discerning tumor from self-antigens¹², its ability to impart immunity against current and future threats^{13,14}, and the extent of selection pressure that it exerts on an evolving cancer population^{15,16}.

Any attempt at better understanding these system-scale properties must start with a reliable method to evaluate the interaction between specific TCR-p-MHC pairs. Despite this, a comprehensive, biophysical model capable of learning the energy contributions of each contact pair in a TCR-p-MHC system and applying them to new predictions remains elusive. To date, experimental research has integrated solved crystal structures^{17,18} with peptide sequencing^{19,20} to probe the physiochemical hallmarks of epitope-specific TCRs. Publicly available crystal structures have enabled researchers to identify detailed structural features that influence the binding specificity of TCR-p-MHC pairs, and machine learning algorithms have made progress on the complementary task of accurately predicting peptide-MHC binding^{21–24} as well as TCR-peptide binding^{25,26}. However, the limited number of available structures relative to the diversity in MHC alleles and TCR-peptide combinations complicates extrapolation to unsolved systems. Alternate template-based structural modeling²⁷ and docking²⁸ approaches are limited by calculation speeds (at best one structure per minute), thus it is unlikely in the foreseeable future that such strategies will be able to investigate the number of TCR-peptide interactions necessary to study the problem at the immune-repertoire level, as this task

¹Center for Theoretical Biological Physics, Rice University, Houston, TX, USA. ²Department of Physics and Astronomy, Rice University, Houston, TX, USA.

³Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Medical Scientist Training Program, Baylor College of Medicine, Houston, TX, USA. ⁵Department of Chemistry, Rice University, Houston, TX, USA. ⁶Department of Physics, Northeastern University, Boston, MA, USA.

⁷Koch Institute for Integrative Cancer Research, Cambridge, MA, USA. ⁸Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁹Ragon Institute of MIT, MGH, and Harvard, Cambridge, MA, USA. ¹⁰Department of Physics, Freie Universität, Berlin, Germany. ¹¹Department of Biosciences, Rice University, Houston, TX, USA. ¹²These authors contributed equally: Xingcheng Lin, Jason T. George.

✉e-mail: jonuchic@rice.edu; h.levine@northeastern.edu

easily requires the simultaneous assessment of more than 10^9 pairs²⁹. Prior attempts have approximated binding affinity by implementing statistical scores calculated from docking algorithms²⁸. These scores are trained using examples of generic protein binding and thus lose the specific physicochemical features of the TCR–peptide interface.

To deal with this challenge, we have developed a systematic TCR–p-MHC prediction strategy, which we refer to as the Rapid Coarse-grained Epitope TCR (RACER) model, for rapid and accurate assessment of TCR specificity capable of differentiating self from foreign antigens. This approach can evaluate 10^9 similarly MHC-restricted TCR–peptide pairs. This method employs supervised machine learning on known TCR–peptide structures and experimental data to derive a coarse-grained, chemically accurate energy model governing the TCR–p-MHC interaction. This strategy was adapted from earlier efforts that predict protein folding^{30–35} and screen the binding of small molecules^{36,37}. Confining our predictions to TCRs with a given MHC restriction enables the transferability of the method to TCRs that are not included in the training set, but our approach could possibly be generalized with the use of additional training data. This strategy provides a tractable means for affinity predictions based on similarly restricted TCR–peptide primary sequences. We show that RACER accurately distinguishes binding peptides across various TCRs and validation tests. Lastly, we simulate thymic selection and show agreement with previously established estimates of T-cell binding distributions and peptide recognition rates^{38,39}. Our *in silico* results share several features observed in experimental data including the degree to which post-selection TCRs recognize foreign antigens and TANs, in addition to the sequence diversity of epitope-specific TCRs^{20,40}. Taken together, our results demonstrate RACER's utility in learning the interactions relevant for high-throughput TCR–epitope binding predictions.

Results

Distinguishing peptides based on binding affinity. The RACER optimization protocol (Fig. 1a) utilizes high-throughput deep sequencing data on TCR–peptide interactions across a large peptide library¹⁹, together with known physical contacts between TCRs and peptides obtained from deposited crystal structures⁴¹. The training data come from cases where all the peptides are displayed by the same allele of the mouse MHC-II molecule. Binding energies between TCRs and peptides, calculated based on a solvent-averaged coarse-grained pairwise model³⁵, were used as the metric for TCR–peptide binding affinity. The interaction parameters for this solvent-averaged energy model were re-optimized here for recognizing strong TCR–peptide interactions. Adapting an approach previously implemented for studying protein folding^{34,42}, the RACER optimization strategy trains a pairwise energy model that maximizes TCR–peptide binding specificity. This energy model was optimized by maximizing the Z-score defined to separate the affinities of experimentally determined strong-binding peptides, called ‘strong binders’ hereafter, from computationally generated, randomized decoys. (The Z-score is defined as the difference between the average binding energies of strong binders versus decoys, divided by the standard deviation of the decoy energies. Throughout this manuscript, we report the absolute value of the calculated Z-score, unless indicated in the corresponding figure caption.) The optimized residue type-dependent energy model was then used to evaluate the binding energies of an ensemble of new TCR–peptide systems. As will be shown below, we performed three different cases (Fig. 1b), and found that predicted binding energies can differentiate strong binders from weak ones, provided they are displayed by the same MHC allele as that of the training set. Crucially, accurate predictions can be made even without knowledge of the actual crystal structure, although the predictions are improved when this additional information is available.

Figure 2 summarizes RACER's predictive performance for a specific TCR (case I in Fig. 1b). For this fixed TCR, pre-identified

strong-binding peptides and decoy peptides with randomized sequences were used to train the energy model (Methods). Another set of peptides independently verified experimentally as weak binders constitutes the testing set. The resulting energy model was then applied to calculate binding energies for the strong binders in the training set as well as the peptides in the testing set. This approach was repeated on three independent TCRs that are associated with the I-E^k MHC-II allele: 2B4, 5cc7 and 226 (TCR details in Supplementary Table 1). Although the experimentally identified weak binders were omitted from the training set, RACER effectively resolves binding-energy differences between experimentally determined strong and weak binders having Z-scores, calculated in an analogous way as above by replacing decoys with experimentally determined poor binders, larger than 3.5 in all cases (Fig. 2a), thus highlighting the predictive power of this approach.

Despite their relative sparsity in antigen space, strong binders play a central role in T-cell epitope recognition. It is more difficult to predict strong binders than weak binders. To test RACER's ability to identify strong binders, we performed a leave-one-out cross-validation (LOOCV) test, using data from TCR 2B4 as an example. For each test iteration, 1 known strong binder was withheld from the training set of 44 strong binders. Our optimization protocol was applied to train the energy model by using the remaining 43 peptides and then predicting the binding energy of the withheld peptide. This prediction was then compared to predicted binding energies of known weak binders, and the procedure was repeated for each of the 44 peptides. Our model is able to accurately distinguish the withheld strong binder in 43 cases (Fig. 2b). This is in contrast to a cluster-based attempt at strong-binder identification based on peptide sequences alone, which at best correctly identifies 19 out of 44 strong binders (Supplementary Note 1). The same LOOCV test was performed for TCRs 5cc7 and 226, which correctly identified 120 out of 126 strong binders of 5cc7, and 267 out of 274 strong binders of 226. To further test the limits of RACER in detecting strong binders that have a more diverse sequence coverage, we performed a more demanding set of hold-out tests on an extended data set from ref.¹⁹. RACER can recognize peptides sharing little sequence identity (~0.3) with the native peptide (Supplementary Figs. 1 and 2), and is still able to recognize strong binders when a substantial portion of the training data is withheld (Supplementary Notes 2 and 3 and Supplementary Figs. 3 and 4).

In order to further characterize RACER's predictive power, an independent set of dissociation constant (K_D) values measured by surface plasmon resonance (SPR)¹⁹ was compared with predicted affinities. The SPR experiments were performed on nine independent peptides for each of the aforementioned three TCRs. RACER was used to predict the binding energies of each of those TCR–peptide pairs, each modeled with the structure of the corresponding TCR as the template. The free energies, $k_B T \log(K_D)$, were compared to calculated binding energies from RACER as a quantitative test of binding affinity prediction accuracy. Lower binding energies indicate stronger binding affinity so that a positive correlation between the $k_B T \log(K_D)$ values and calculated binding energies implies a successful prediction. As shown in Fig. 2c, RACER's prediction of binding affinities for these nine peptides correlates well with the experimental measurements, with an average Pearson correlation coefficient of 0.74. The predicted order of binding affinities is also consistent with those from the experiment, with an average Spearman's rank correlation coefficient of 0.65.

Optimized specific interactions for TCR–peptide recognition.

The data utilized by RACER includes strong binders and an input crystal structure, as well as TCR and peptide primary sequences, which determine an interaction pattern that was then used to construct a system-specific force field. To illustrate this, we focus on

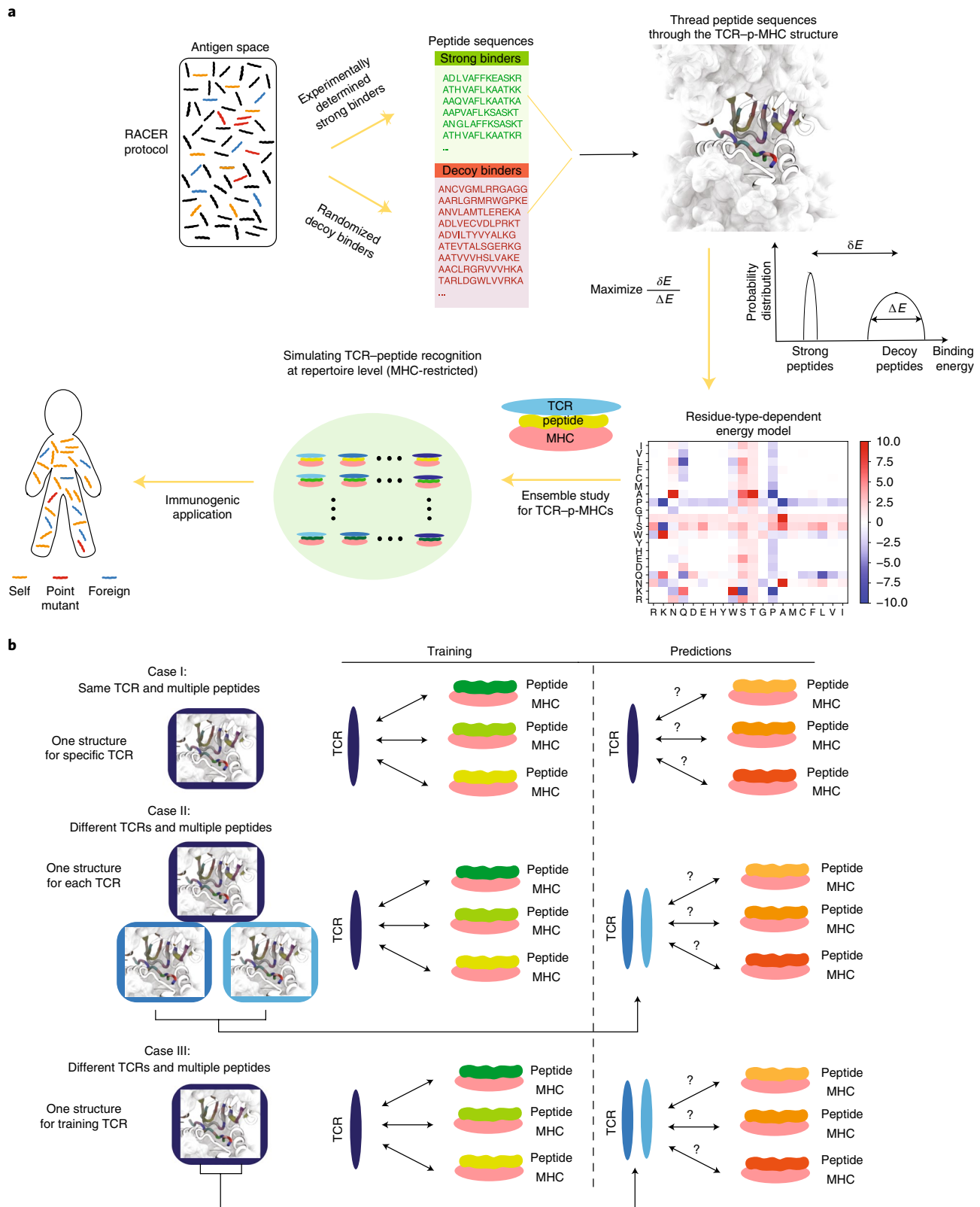


Fig. 1 | Summary of the modeling approach employed in this study. a, The protocol of RACER optimization (Methods). The color scale is presented with a reduced unit as described in the Methods. **b**, Three tests were conducted to evaluate RACER's performance. Case I: the training set includes one TCR-p-MHC structure and multiple peptide sequences. The test set includes the same TCR structure and a separate set of peptide sequences. Case II: the training set includes one TCR-p-MHC structure and multiple peptide sequences. The test set includes two different TCR structures (restricted to the same MHC allele) and two separate sets of peptide sequences. Structures for the two additional test TCRs are included in predictions. Case III: The training set includes one TCR-p-MHC structure and multiple peptide sequences. The test set includes only the sequences of two different TCRs (restricted on the same MHC allele) and two separate sets of peptides. Only the structure from the original training TCR was used in prediction. (The interactions of interest are indicated by double-sided arrows between TCR and p-MHC.)

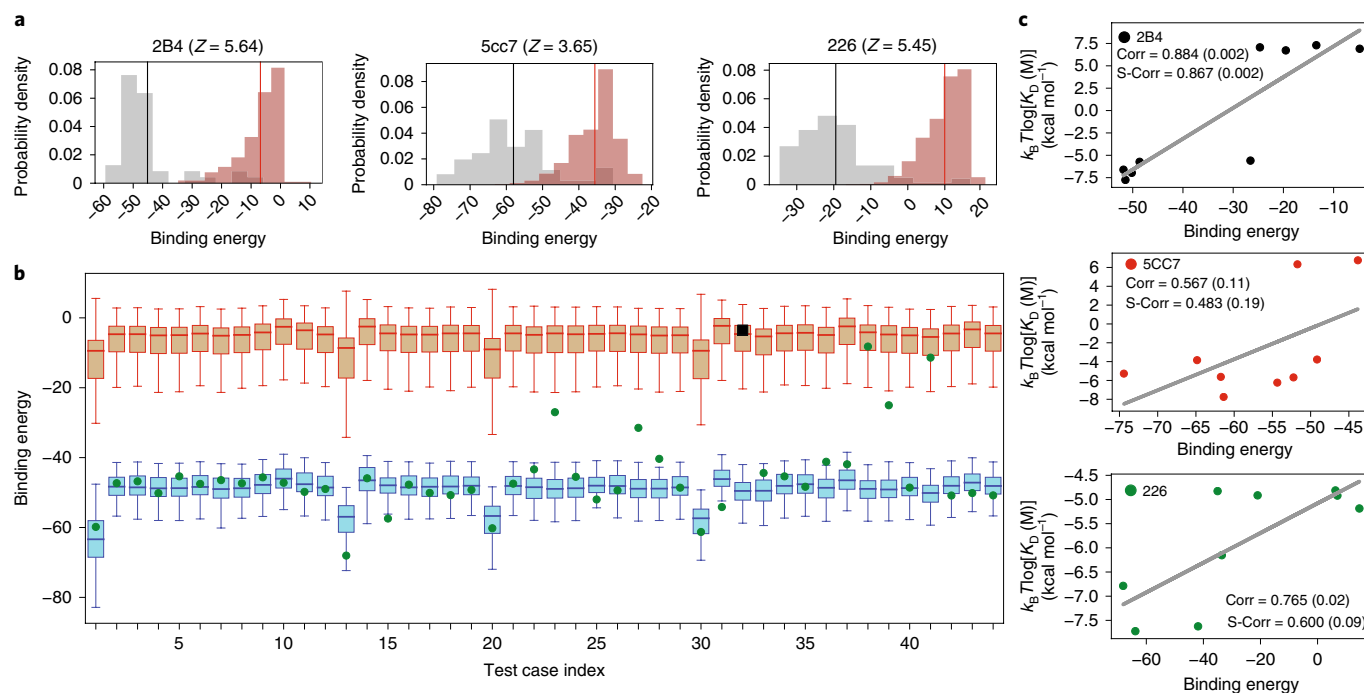


Fig. 2 | RACER identification of TCR-specific strong and weak binders. **a**, Probability density distributions of the predicted binding energies of experimentally determined strong (brown, with mean depicted in red) and weak (gray, with mean depicted in black) binders of three TCRs (2B4, 5cc7 and 226). **b**, Summary of the predicted binding energies from the leave-one-out cross-validation tests using TCR 2B4. Each test case represents one example using 1 of the 44 strong binders (green or black), as well as the experimentally determined weak binders (brown) as the test set and the other 43 strong binders as the training set (blue). Each box plot represents the lower (Q1) to upper (Q3) quartiles of the predicted binding energies, with a horizontal line at the median. Withheld strong binders are depicted as green circles when successfully recognized (binding energy lower than the median of the experimentally determined weak binders), and as black squares otherwise. The whiskers are placed at the first and last data points that fall within (m , M), where $m = Q1 - 1.5IQR$ and $M = Q3 + 1.5IQR$, IQR is the interquartile range. **c**, In a completely independent testing data set measured by SPR (ref.¹⁹), the calculated binding energies of test peptides were compared with the binding affinity converted from their experimentally determined dissociation constant K_D . Best-fit linear regression is depicted for each case. Corr: Pearson correlation coefficient. S-Corr: Spearman's rank correlation coefficient. The p-value of each correlation coefficient is reported in the parentheses. The predicted binding energies are presented with reduced units as described in Methods.

TCR 2B4 as an example (Fig. 3). The crystal structure of TCR 2B4 (Fig. 3a) reveals that there can be many threonine (T) and asparagine (N) residues contained in the CDR loops region of the TCR. In the strong-binder set, these residues tend to interact with specific peptide residues such as alanine (A), as seen for the specific peptide given in Fig. 3. This notion can be formalized by showing the set of observed probabilities of close proximity of specific residue pairs. Thus, we see that certain pairs such as A–T and A–N are significantly enriched in the strong binders, while much less so in the decoy binders (Fig. 3b). This leads to strongest attractions between the A–T, A–N residue pairs in the optimized energy model (Fig. 3c). By contrast, the TCR tryptophan (W) residue frequently interacts with alanine (A) in both strong binders and decoy peptides. As a result, the optimized energy model does not favor the A–W interaction.

This energy model is rather distinct from those typically used for studying protein folding. In order to compare the RACER-derived energy model to well-established force fields described in the protein-folding literature, we substitute for our energy model either the standard AWSEM³⁵ (optimized on deposited folded proteins) force field or the Miyazawa–Jernigan (MJ) statistical potential⁴³ (constructed using the probability distribution of contacting residues from deposited proteins) and calculate the corresponding binding energy predictions for the TCR 2B4 peptides. We find that neither of them effectively resolves these groups, with Z-scores of 0.69 and 1.28, respectively (Supplementary Note 4 and Supplementary Fig. 5). Similar trends were observed utilizing the peptides corresponding to TCRs 5cc7 and 226, demonstrating the necessity of

RACER's de novo identification of pertinent structural information for studying the TCR–peptide system.

Predicting TCR–peptide binding affinity for fixed MHC allele.

Given RACER's accuracy in resolving test peptides presented to the specific TCR used for training, we next explored the feasibility of extending predictions to additional TCR–peptide pairs albeit with the same MHC restriction. To this end, we assessed whether the physical contacts implicitly encoded in RACER's optimized force field were conserved within I-E^k-restricted TCR–peptide pairs. The three I-E^k-restricted TCRs considered in our analysis have all been tested with peptides bound to the I-E^k mouse MHC molecule. The available crystal structures have a significant degree of structural similarity at the TCR CDR3–peptide binding interface (figure 5 in ref.¹⁹). We further quantified the TCR CDR3–peptide contacts for each pair, constructing a contact map based on their crystal structures (Fig. 4, Methods). Our results suggest that, despite differences in TCR and peptide primary sequences, similarly MHC-restricted TCR–peptide pairs share common structural features. By contrast, these contact maps are not preserved across different MHC alleles (Supplementary Fig. 6).

We next examined RACER's ability to predict binding peptides of TCRs restricted to the same MHC alleles. Toward this end, we apply the energy model optimized using binding data for one of the three TCRs to predict the TCR–peptide binding energies of the remaining two holdout TCRs (case II in Fig. 1b). To do this, we initially use a known structure for each of the holdouts, and the energy model learned from the training TCR to predict the binding energies of the

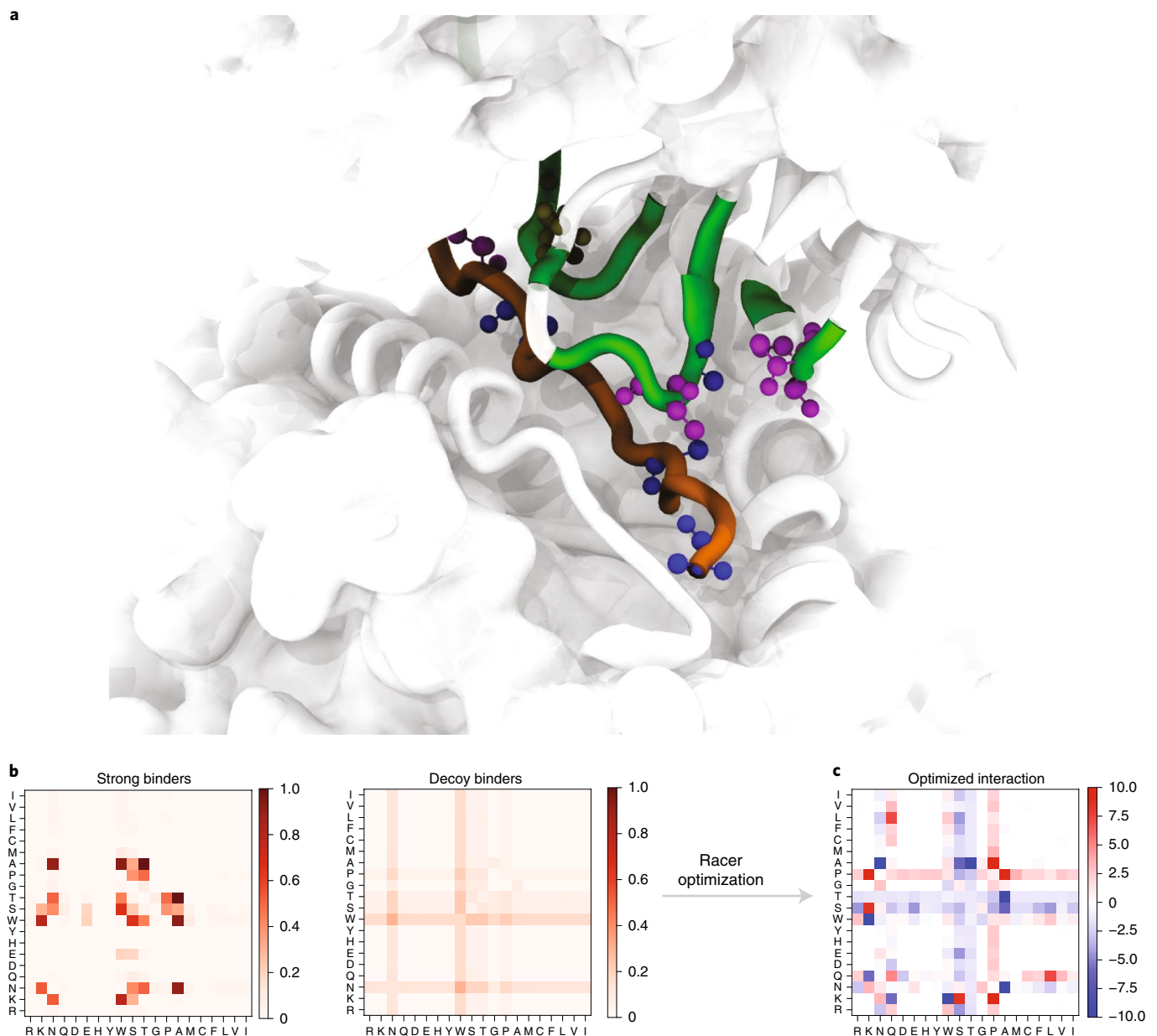


Fig. 3 | The RACER-derived energy model. a, The 3D crystal structure of TCR 2B4 bound to a specific peptide (PDB: 3QIB). The parts of the structure that are in contact between the TCR and peptide are highlighted as green (TCR) and orange (peptide). Also shown are alanine (blue), threonine (magenta) and asparagine (tan) residues, which are discussed in the main text (CPK representation²⁵). **b**, The probability of contact formation (interaction set) between each two of the 20 amino acids in the set of strong binders (left) and the set of randomized decoy binders (right) of TCR 2B4. **c**, The residue-based interaction strength (energy model) determined by RACER for TCR 2B4. A more negative value indicates a stronger attractive interaction between the two corresponding residues. The predicted color scales are presented with a reduced unit as described in Methods.

experimentally determined strong and weak binders of those hold-out TCRs. Although the Z-scores measured for these alternate TCRs are lower than those found previously (section ‘Distinguishing peptides based on binding affinity’), RACER still successfully distinguishes a majority of strong binders from weak binders, with an average Z-score of 1.8 (Fig. 5a). Further incorporation of target TCR structural information in the optimization step improves RACER’s predictive accuracy (Supplementary Note 5; Supplementary Fig. 7).

To provide an additional test and to quantify our discrimination capability, we used an independent dataset from ref.⁴⁴. Four independent TCRs (Protein Data Bank (PDB) IDs: 3QIB, 3QIU, 4P2Q, 4P2R) from their curated benchmark dataset are associated with the I-E* allele; note that three of these overlap with the TCRs in

ref.¹⁹. To test the performance of RACER for different TCR–peptide pairs, we used the energy model trained on 2B4 (3QIB) to predict the binding energies of both strong and weak binders for the three remaining TCRs. This calculation again uses the structure found for the one strong binding peptide for each of the three TCRs. Our calculation re-emphasizes that RACER can successfully distinguish strong binders even when it is trained based on a different TCR (Fig. 5c), with an area under the curve (AUC) of 0.89. As a more comprehensive test of RACER’s transferability, we included other TCR–peptide pairs from ref.⁴⁴. RACER capably recognizes most strong binders across the same MHC allele-restricted TCRs with different V α and V β genes, and does so more effectively when there are multiple copies of TCR–peptide pairs available for

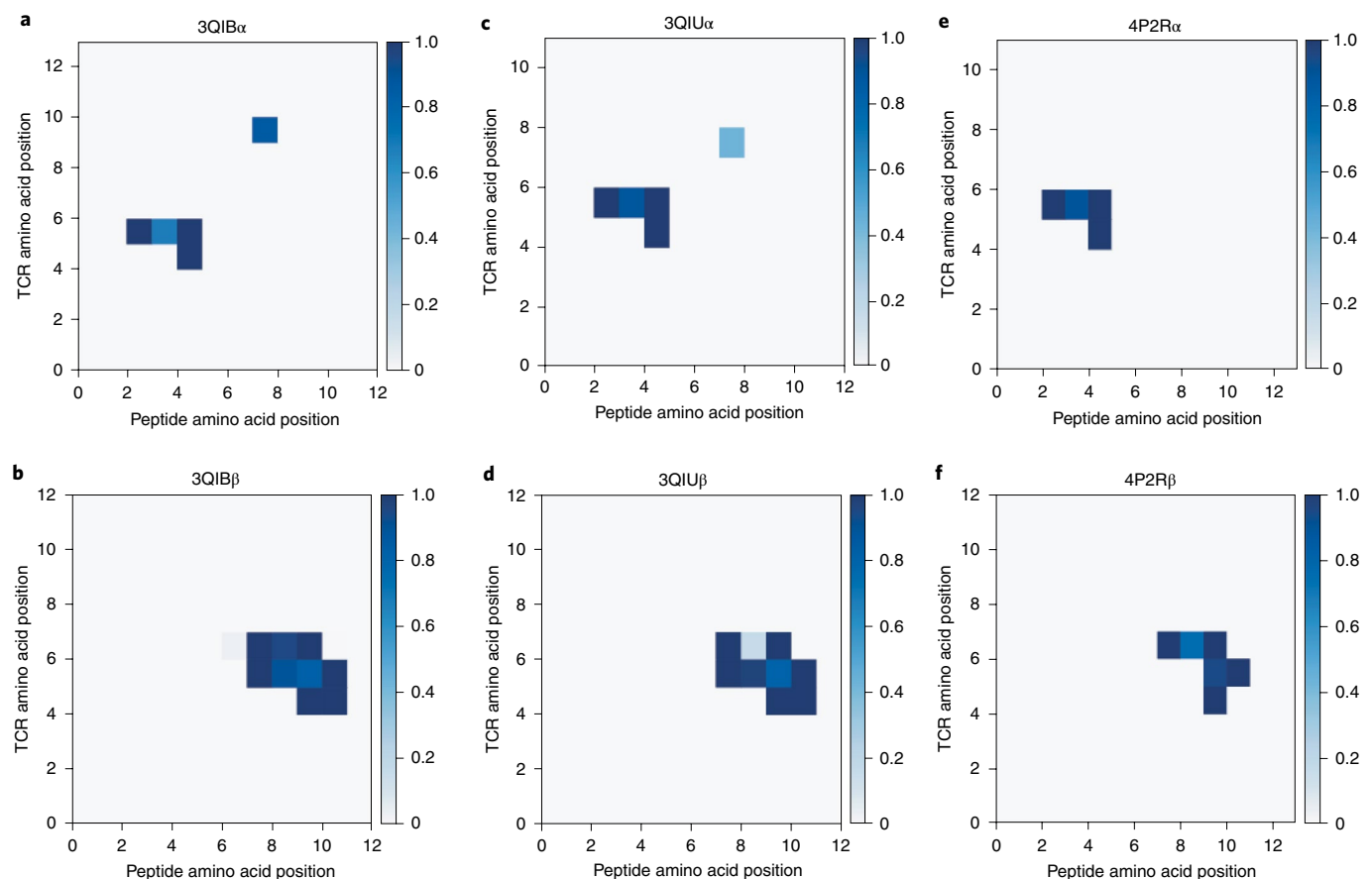


Fig. 4 | Contact maps for MHC-II I-E^k-restricted TCR-peptide pairs. Contact maps are calculated using distances from each pairwise TCR-peptide amino acid combination using equation (6) for the following MHC-II I-E^k-restricted TCR-peptide pairs. **a,b**, 3QIB peptide ADLIAYLKQATK with TCR 2B4 CDR3α (AALRATGGNNKLT, **a**) and CDR3β (ASSLNWSQDTQY, **b**) chains. **c,d**, 3QIU peptide ADLIAYLKQATK with TCR 226 CDR3α (AAEPSSGQKLV, **c**) and CDR3β (ASSLNNSDYSY, **d**) chains. 4P2R peptide ADGVAFLLTPFKA with TCR 5cc7 CDR3α (AAEASNTNKVV, **e**) and CDR3β (ASSLNNSDYSY, **f**) chains. Similarity in interaction topology across TCR-peptide pairs is observed by comparing the contact silhouette of interacting coordinates for the α (top row) and β (bottom row) TCR sequences.

training (Supplementary Note 6; Supplementary Fig. 8). Of note, when we tested data from the same study involving TCR-p-MHCs with different MHC alleles, RACER could not isolate strong binders, presumably due to the substantially different TCR-peptide interacting patterns (Supplementary Fig. 6).

Next we examine the need to have at least one TCR-p-MHC crystal structure in order to use the optimized energy model for identifying other strong binders (case III in Fig. 1b). Of course to evaluate the binding energy we must have a structure; the alternative to having a measured structure for a new sequence is to thread that new CDR3 sequence into the crystal structure used for the training data, which potentially introduces an uncertainty in registration. For the cases at hand, this issue arises only for the CDR3α chain as the β chains for all three TCRs have 12 residues and there is no residual ambiguity. We tested the simplest possible assumption, where all three α chains are aligned to the left⁴⁵. Figure 5b shows that this procedure again leads to successful discrimination between strong and weak binders, with an average Z-score of 2.36. As a comparison, the best performance of a recent sequence-based predictor trained by using artificial neural networks²⁶ can recognize the strong binders of TCR 5cc7, but not TCR 2B4 and 226 (Supplementary Note 7; Supplementary Fig. 10). Similar tests were also performed for the TCR-peptide pairs from ref.⁴⁴. RACER still capably recognizes the strong binders across TCRs with different Vα and Vβ genes (Supplementary Note 6; Supplementary Fig. 9). Thus, we conclude that the MHC-restricted TCR structures

are sufficiently similar so that not only can we use the energy model derived from a single TCR training set for other TCRs but we can also use the same structure. This then allows us to make estimates at the repertoire scale without creating extremely large numbers of TCR-p-MHC structures.

RACER-optimized representation of thymic selection. We then applied RACER's ability to reasonably assess binding strengths using a single crystal structure and associated energy model to study statistical properties of the high-throughput TCR-p-MHC binding observed in thymic negative selection. Using the 2B4 TCR-peptide crystal structure, we simulated 10^5 TCRs and 10^4 self-peptides by uniform randomization of the CDR3 and peptide sequences over amino acid space. To avoid registration issues, simulated TCRs were chosen to have the same number of α and β chain residues as TCR 2B4. This was repeated with 2,000 TCRs and 10^4 self-peptides, this time weighing CDR3-peptide interactions by each of the three contact maps in Fig. 4. The same approach was applied to a model that assumes a strictly diagonal contact map (motivated by previous analytical work¹²) with randomization of the TCR sequence taken over each non-null position in the contact map.

Using this framework, a given TCR survives only if it binds every self-peptide below a fixed activation threshold. The maximum binding interaction over all self-peptides for each TCR defines a selection curve (Fig. 6a), which describes the percentage of negatively selected T cells as a function of the cutoff activation threshold.

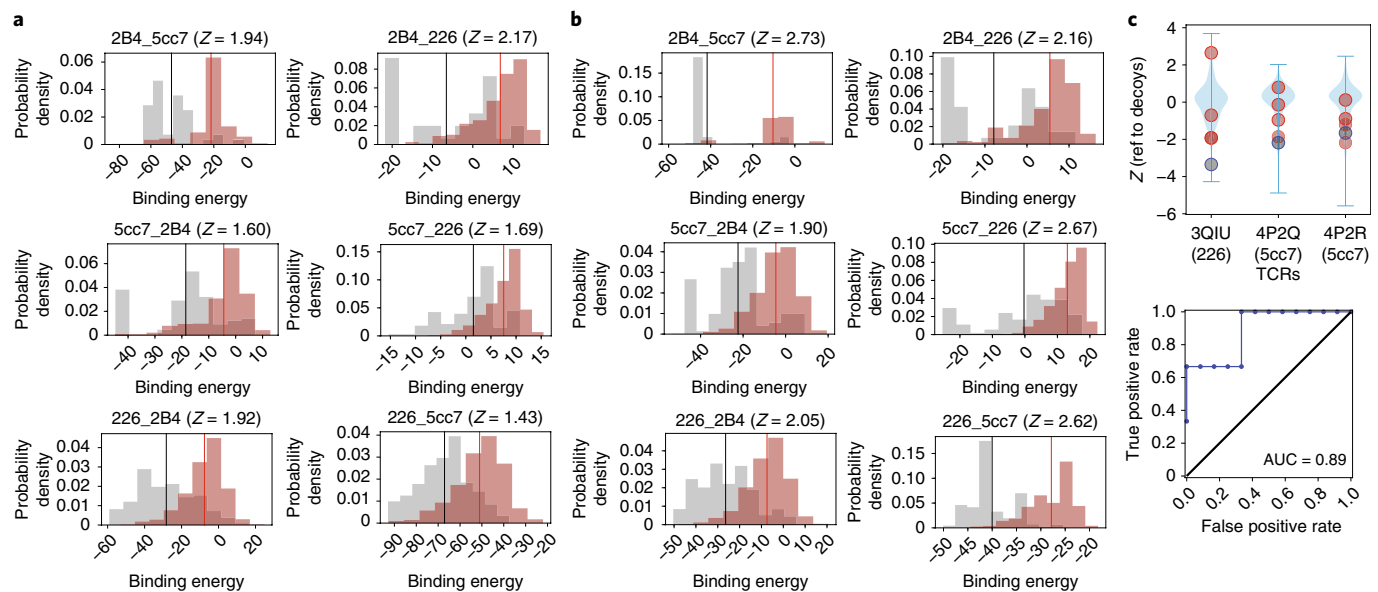


Fig. 5 | RACER predictive transferability across distinct TCRs. a, Probability density distributions of the predicted binding energies of experimentally determined strong (brown, with mean depicted in red) and weak (gray, with mean depicted in black) binders of each of the three TCRs (2B4, 5cc7 and 226), using another TCR for training. The title of each figure follows the format of 'target_training TCRs', for example, '2B4_5cc7' utilizes the energy model trained on TCR 5cc7 for predicting peptide binding affinities of TCR 2B4. **b**, Probability density distributions of the predicted binding energies of the same cases as in **a**, but without utilizing any new structure for the new TCR. The panel is formatted in the same way as **a**. **c**, Top: the energy model trained on TCR 2B4 is used to predict the binding energies of sequences from the other I-E_k-associated TCRs⁴⁴. Z-scores of known strong binders (gray) and weak binders (orange) provided by ref.⁴⁴ were calculated relative to a set of 1,000 decoy peptides with randomized sequences (blue violin plot), with lower (not absolute) Z-scores indicating better predictive performance. Bottom: the calculated Z-scores of each TCR were used to depict the corresponding ROC curve and AUC score (0.89).

Selection curves for the three TCR sets using Fig. 4 contact maps and the RACER energy model compare reasonably to the diagonal contact map motivated by previous analytical work (Fig. 6b red curve). Here, variances are similar in each case with mean-shifts correlated directly with the number of peptide-CDR3 contacts (Fig. 4). These findings reinforce the relevance of TCR-p-MHC-specific structural interactions encoded in the RACER-derived energy potential for binding prediction and T-cell repertoire generation. Although empirical estimates of TCR thymic selection survival rates vary (20–50%)^{46,47}, we assess recognition across all survival rates, restricting our analysis to 50%, when applicable. Given these assumptions, we demonstrate that RACER-generated thymic selection makes reasonable use of available self-peptides (Supplementary Note 8; Supplementary Fig. 11a) and generates a sensible regime of optimal selection, consistent with previous analytical estimates¹² (Supplementary Note 8; Supplementary Fig. 11c).

One key issue influencing adaptive immune recognition of tumor-associated neoantigens (TANs) is the recognition of peptides closely related to self (for example, point-mutants) relative to foreign peptide recognition. The fact that T cells can in fact recognize tumors suggests that thymic selection leaves intact the ability to strongly bind TANs. Post-selection individual TCRs exhibit minor recognition differences between foreign peptides and TANs (Fig. 6b) with overlapping variances in line with previous theoretical estimates (Supplementary Fig. 11b). Moreover, the recognition capacity of the MHC-restricted post-selection T-cell repertoire demonstrates that this minimal difference is maintained at the aggregate immune level (Fig. 6c). These findings explain the ability of the immune system to target cancers in a manner dependent on their mutational load. Moreover, comparisons of RACER-derived post-selection T-cell maximal binding energy to the immunogenicity scores for empirically observed thymic self-peptides, foreign peptides, and TANs⁴⁰ demonstrates RACER's ability to capably

classify TANs having immunogenicity intermediary to those of foreign and self-peptides with their distribution closer to the foreign group (Fig. 6d). Additional assessments of RACER-derived TCR repertoire CDR3 sequence similarity recapitulate key features observed in experimentally studied repertoires²⁰ (Supplementary Note 8; Supplementary Fig. 12). Collectively, our results reinforce RACER's utility for performing realistic post-selection T-cell repertoire-level analyses.

Discussion

TCR-p-MHC structures encode a system-specific energy model, whose identification can uncover the rules underlying TCR-antigen specificity. The preserved sequence and structural features of TCR-peptide systems^{18–20} limit the physicochemical space explorable by TCR-peptide interface. When optimized on TCR-peptide pairs, the arrangement of the residue contacts between TCR and its cognate peptide (Fig. 4) leads to an energy model (Fig. 3) distinct from the traditional hydrophobic-hydrophilic interaction patterns⁴⁸ used for studying protein folding, such as the MJ potential⁴³. This system-specific energy model enables RACER to identify strong binders of corresponding TCRs (Fig. 2) while standard protein-folding energy models fall short (Supplementary Fig. 5).

RACER offers an approach for developing models that incorporate available protein structural information. Prior investigations have applied artificial neural networks for predicting strong binders of TCR^{25,26} and MHC⁴⁹ molecules based only on the primary sequences. Although deep learning can implicitly account for higher-order interactions, such approaches may still be limited by available sequences. RACER alleviates the high demands for sequences by including existing crystal structures in a pairwise potential. To comprehensively characterize RACER's predictive power, our training set was limited to cases that had pre-identified TCR-peptide pairs given their known crystal structure^{19,44}. While

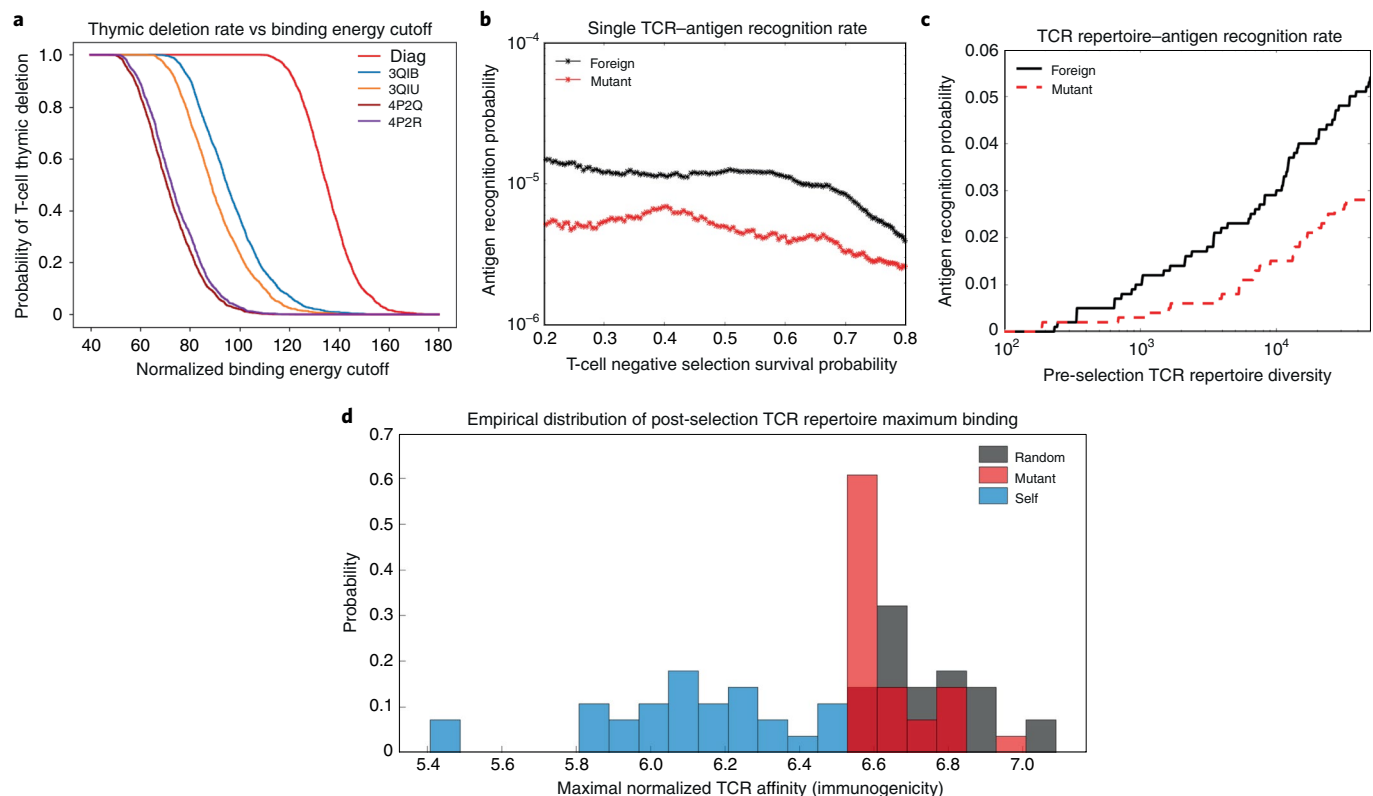


Fig. 6 | RACER-derived T-cell repertoire simulations of thymic selection and antigen recognition. **a**, Simulated thymic selection curves (T-cell recognition probability as a function of negative selection binding energy cutoff) incorporating the effects of non-adjacent contacts (given in Fig. 4) using $N_p = 10^4$ uniformly randomized self-peptides and $N_t = 2,000$ randomized I-E^k-restricted TCRs. 4P2Q and 4P2R (purple) use T-cells generated by randomizing the CDR3 region of TCR 5cc7, while 3QIB (blue) randomizes the CDR3 of TCR 2B4, and 3QIU (yellow) randomizes the CDR3 of TCR 226 (in all cases, randomized CDR3 lengths were unchanged from the original TCR) (the red curve uses RACER energy using a diagonal contact map model whose study here is motivated by previous work¹²). **b**, Utilizing RACER-derived energy assessments from the 2B4 crystal structure, the probability of recognizing foreign and point-mutant antigens for individual post-selection T-cells is plotted as a function of the percentage of TCRs surviving negative selection (ordinate of the graph in **a**, simulations averaged over all post-selection TCRs with pairwise interactions amongst 10^3 random peptides and 10^3 point-mutant peptides). **c**, The recognition probability of foreign (black) and mutant (red) peptides by the entirety of the TCR repertoire is plotted as a function of pre-selection TCR repertoire diversity (the number of unique post-selection TCRs), with negative selection thresholds giving 50% survival. **d**, RACER-derived immunogenicity of foreign, mutant and self antigen. The distribution of maximum binding affinity over all post-selection T cells for immunogenic random (gray) and point-mutated self-peptides (red) is compared to that of thymic self-peptides (blue). (There were 28 point-mutated peptides that had at least one T-cell recognition event. To keep an equal number of peptides in each distribution, these were compared with the top 28 similarly ordered foreign peptides and 28 randomly chosen self-peptide groups.)

limited by the diversity of experimentally determined strong binders, RACER correctly resolves most of the strong binders even in the most challenging training scenario (Fig. 5b, Supplementary Fig. 9). While the pairwise potential of RACER maintains reasonably high predictive accuracy, it might be further improved by including additional contributions to peptide binding affinity (Supplementary Note 9).

RACER's application to CDR3 α , β chains obtained from T-cell sequencing, together with possible TAN lists generated by cancer deep sequencing could provide a rapid and reliable method of generating clinically actionable information for cancer-specific TCRs in the form of putative TCR–TAN pairs, provided those TANs are similarly presented on the original MHC^{38,39}. While we focused our analysis on a single MHC restriction, our approach could also be applied to the crystal structure of another TCR–p-MHC pair, together with several known strong and weak binder candidates. Analysis of the crystal structure associated with the I-E^k MHC-II allele reveal largely conserved interaction patterns (Fig. 4), and their differences may have broader implications for preferential selection and recognition behavior of T cells (Supplementary Note 10). In the future, RACER's predictive accuracy can be further improved by

incorporating additional strong binders and structural data as they become available (Supplementary Fig. 7).

The relative efficacy of targeting TANs remains an important question with significant clinical implications. Our findings suggest that thymic selection affords little-to-no recognition protection of peptides closely related to self, thus supporting the notion that T cells undergoing central tolerance to thymic self-peptides are essentially memorizing a list of antigens to avoid. Given that a large class of TANs is generated via point-mutations in self-peptides, our results provide a quantitative argument for the efficacy of immunotherapies that target point-mutated neoantigens. We expect that RACER's ability to identify a diverse set of antigen-specific TCRs within high-dimensional CDR3 sequence space will accelerate therapeutic T-cell discovery by providing a quick and inexpensive screening tool that can then inform more costly confirmatory TCR repertoire sequencing and affinity tests. Currently, we have focused on predicting binding affinities of TCR–peptide pairs restricted to a particular MHC allele, offering a proof-of-principle for epitope identification. This procedure can in general be repeated for other MHC alleles and could be applied to a broad set of clinical scenarios by training on a relatively small number of the most

common MHC class-I alleles, each having ample available crystal structure data.

While important, TCR–p-MHC pairwise interactions are only one factor influencing adaptive immune system recognition. Signaling between other adaptive immune elements and intracellular factors influences antigen generation and abundance, and availability also affects recognition rates. We propose our optimized framework as a tool for understanding general questions regarding the interactions between the T-cell repertoire and relevant antigen landscape. Although we calculate static antigen recognition probabilities, the temporal tumor–immune interaction leads to dynamic co-evolution¹⁶ reliant on the quality, abundance and systems-level signaling of antigens⁵⁰. The availability of time-series assessments of immune-cell repertoires, self-peptides and tumor antigens will enable the development of optimized immunotherapeutic treatments by uncovering the T-cell–tumor-antigen specificity map.

Methods

RACER model. The optimization of RACER (Fig. 1a) starts from a series of TCR binders obtained from the deep-sequencing experiments¹⁹, as well as the corresponding TCR–p-MHC crystal structures deposited in the database⁴¹. The sequences of the strong binders, as well as the generated decoy binders from randomizing the non-anchoring sequences of the strong binders, are collected for parameterizing a pairwise energy model, which maximizes the energetic gap between the strong binders and a randomized set of decoys. The resulting energy model can be used to quickly evaluate the binding affinities of an ensemble of TCR–peptide interactions at the population level. The calculated binding affinities can be used for simulating the negative selection process in the thymus, as well as measuring the recognition probability of the post-selection TCRs. Finally, this kind of ensemble study can be used for immunogenic applications that require input from an entire T-cell repertoire.

Energy model. To evaluate the binding energies on the basis of a structurally motivated molecular energy model, the framework of a coarse-grained protein energy model, AWSEM force field³⁵, was utilized for calculating the binding energies between the T-cell receptors (TCRs) and the peptide displayed on top of an MHC molecule. AWSEM is a coarse-grained model with each residue described by the positions of its three atoms—C α , C β and O atoms (except for glycine, which does not have C β atoms)³⁵. We used the C β atom (except for glycine, where we used C α) of each residue to calculate inter-residue interactions. The original AWSEM energy includes both bonded and non-bonded interactions.

$$V_{\text{total}} = V_{\text{bonded}} + V_{\text{nonbonded}} \quad (1)$$

Since those residue pairs that contribute to the TCR–peptide binding energy, specifically those from the CDR loops and peptides, are in separate protein chains, only non-bonded interactions are considered. $V_{\text{nonbonded}}$ is composed of three terms:

$$V_{\text{nonbonded}} = V_{\text{pairwise}} + V_{\text{burial}} + V_{\text{database}} \quad (2)$$

Among them, V_{burial} is a one-body term describing the propensity of residues to be buried in or exposed on the surface of proteins. V_{database} is a protein-sequence-specific term that uses information from the existing protein database, such as secondary and tertiary interactions, to ensure locally accurate chemistry of protein structure. Since the TCR–p-MHC system features pairwise interactions between a TCR and its corresponding peptide, only the term V_{pairwise} is used for this study.

The pairwise energy of AWSEM potential describes the interactions between any two non-bonded residues and can be further separated into two terms:

$$V_{\text{pairwise}} = V_{\text{direct}} + V_{\text{mediated}} \quad (3)$$

V_{direct} captures the direct protein–protein interaction of residues that are in between 4.5 and 6.5 Å. The functional form of V_{direct} is

$$V_{\text{direct}} = \sum_{\substack{i \in \text{TCR} \\ j \in \text{peptide}}} \gamma_{ij}(a_i, a_j) \Theta_{ij}^I \quad (4)$$

in which $\Theta_{ij}^I = \frac{1}{4}(1 + \tanh[5.0 \times (r_{ij} - r'_{\min})])(1 + \tanh[5.0 \times (r'_{\max} - r_{ij})))$ is a switching function capturing the effective range of interactions between two residues (here taken between $r'_{\min} = 4.5$ Å and $r'_{\max} = 6.5$ Å). Thus, two residues are defined to be ‘in contact’ if their distance falls between 4.5 Å and 6.5 Å. $\gamma_{ij}(a_i, a_j)$ describes the residue-type-dependent interaction strength, and is the most important parameter that enters the optimization of RACER. V_{mediated} describes the longer range interactions of two residues and is not used in this study.

Maximizing specificity of TCR–peptide recognition. For each interaction type, the $\gamma_{ij}(a_i, a_j)$ parameters constitute a 20-by-20 matrix of parameters that describes the pairwise interaction between any two residues i, j , each with one of the 20 residue types, a_i, a_j . Guided by the principle of minimum frustration³², $\gamma_{ij}(a_i, a_j)$ was previously optimized self-consistently to best separate the folded states from the misfolded states of proteins. Distilled into mathematical details, the energy model was optimized to maximize the functional $\delta E/\Delta E$, where δE is the energy gap between folded and misfolded proteins, and ΔE measures the standard deviation of the energies of the misfolded states. An energy model was optimized based on a pool of selected protein structures⁵¹, where a series of decoy structures were generated by either threading the sequences along the existing crystal structures, or by biasing the proteins into molten globule structures using MD simulations³⁴. The resultant γ parameter thus determines an energy model that facilitates the folding of proteins with given sequences.

Motivated by this idea, RACER was parameterized to maximize the Z-scores for fully separating TCR strong binders from weak ones. Strong binders were chosen to be those top peptides that survive and were amplified to contain at least 50 copies after 4 rounds of experimental deep-sequencing processes (details of input data are provided in Section Data input)¹⁹, together with the peptides present in the deposited crystal structures⁴¹. In the experiment in ref. ¹⁹, to ensure the correct display of peptides on the MHC, limited diversity was introduced for most distal residues and anchoring residues of peptides. The decoy binder sequences were generated by randomizing the non-anchoring residues of each strong binder thereby generating 1,000 copies, and excludes the strong-binder sequences. The γ parameters were then optimized to maximize the stability gap between strong and randomized set of decoy binders, $\delta E = \mathbf{A}^T \gamma$, and the standard deviation of decoy energies, $\Delta E^2 = \gamma^T B \gamma$, where the vector \mathbf{A} and matrix B are defined as:

$$\begin{aligned} \mathbf{A} &= \langle \langle \phi_{\text{direct}} \rangle^{\text{db}} - \phi_{\text{direct}}^{\text{sb}} \rangle \\ B &= \langle \langle \phi_{\text{direct}} \phi_{\text{direct}} \rangle^{\text{db}} - \langle \phi_{\text{direct}} \rangle^{\text{db}} \langle \phi_{\text{direct}} \rangle^{\text{db}} \rangle \end{aligned} \quad (5)$$

In equation (5), direct stands for the interaction type, V_{direct} . ϕ_{direct} is the functional form for V_{direct} . ϕ_{direct} also summarizes the probability of contacts formation (interaction set) between pairs of amino acids in a specific TCR–peptide system. The superscript db stands for decoy binders and sb for strong binders. The first average is over the 1,000 decoy binders generated from one specific strong binder. The second average is over all the strong binders. The maximization of $\delta E/\Delta E = \mathbf{A}^T \gamma / \sqrt{\gamma^T B \gamma}$ can be performed effectively by maximizing the functional objective $R(\gamma) = \mathbf{A}^T \gamma - \lambda_1 \Delta$, where $\Delta^2 = \gamma^T B \gamma$. The solution of this optimization gives $\gamma \propto B^{-1} \mathbf{A}$. \mathbf{A} is a vector containing the difference in the number of interactions of each type in the strong and decoy binders. B is a covariance matrix, which contains information about which types of interactions tend to co-occur in the decoy binders. Finally, γ is a vector that encodes the optimized strengths of the interactions. The dimension of the vector \mathbf{A} is (1, 210), that of the matrix B is (210, 210) and that of the vector γ is (210, 1). To aid visual presentation, we reshape the γ vector into a symmetric 20-by-20 matrix in Fig. 3c. Finite sampling of decoy binders introduces noise in the optimization process, particularly in B . As such, a filter is applied to reduce the effects of the noise. The filtering scheme was performed by first diagonalizing the B matrix such that $B^{-1} = P \Lambda^{-1} P^{-1}$, where P is composed of the eigenvectors of B and Λ is made up of B 's eigenvalues. The first N modes of B (sorted in descending order by eigenvalue) are retained and the other $(210 - N)$ eigenvalues in Λ are replaced with the N th eigenvalue of B . The final result is robust to the choice of N . In practice, N is chosen so that no eigenvalue is close to zero. The Wolynes group performed this optimization in an iterative way where the optimized parameters were used for generating a new set of decoy protein structures⁵². In this study, since different peptides are structurally degenerate on top of MHC as observed from experiments¹⁹, only one round of optimization was performed. Since the optimization leaves a scaling factor as a free parameter, throughout this manuscript, the binding energies are presented with reduced units, which is kcal mol⁻¹ multiplied by this scaling factor. The value of this scaling factor affects neither the calculation of Z-scores for identifying strong binders of a specific TCR nor the subsequent ensemble study of peptide recognition, where only the order of binding affinities among individual TCR–p-MHC pairs matters for our results. To obtain binding energies that have physical units, the scaling factor can be further calibrated to fit the experimentally determined binding affinities, such as the K_D values measured by SPR experiments (Fig. 2c).

Data input. A deep-sequencing technique was developed to assess the binding affinity of a diverse repertoire of MHC-II-presented peptides towards a certain type of TCR¹⁹. Specifically, three types of TCR: 2B4, 5cc7 and 226, were used for selecting peptides upon four rounds of purification. The peptides that survived and enriched with multiple copies bind strongly with the corresponding TCR. By contrast, the peptides that are present initially but become extinct during purification represent experimentally determined weak binders. For each of the three TCRs, the peptides that end up with more than 50 copies after the purification process, together with the peptides presented in the crystal structures, were selected as strong binders. 1,000 decoy sequences were generated for each of the strong binders by randomizing the non-anchoring residues. Both

strong binders and decoys were included in the training set. In addition, to test the performance of RACER, peptides with at least eight copies initially but disappearing during purification were selected as experimentally determined weak binders and were assigned to the test set for each TCR. To test the transferability of the model, we used weak-binding peptides of two different TCRs (for example, 5cc7 and 226) as additional test sets distinct from the TCR used in training (for example, 2B4).

When structural data for a specific TCR–peptide pair of interest were unavailable, we built the structure by homology modeling⁴⁵, based on a known TCR–peptide crystal structure. Since potential steric clashes after switching peptide sequences may disfavor the strong binders used in our training set, we used Modeller⁴⁵ to refine the structures located at the TCR–peptide interface of strong binders before including them in the training process. Likewise, the binding energies of the experimentally determined weak binders were also evaluated after structural relaxation. The structural relaxation adds several seconds of computational time for each TCR–peptide pair, and thus poses a challenge for large-scale repertoire analysis. However, the coarse-grained nature of the RACER framework may significantly reduce the probability of side-chain clashes after switching peptide sequences. To test the accuracy of our model prediction without structural relaxation, we calculated the binding energies of strong and weak binders of TCR 2B4 by only switching the peptide sequences, omitting any structural adjustment. Our result (Supplementary Fig. 13) shows comparable accuracy in separating strong from weak binders, similar to that reported in Fig. 2a. In the same vein, the transferability of RACER was also maintained without structural relaxation (Fig. 5b). Encouraged by the accuracy of our coarse-grained model without relaxation, we modeled large pairwise collections of TCR–peptide interactions by only altering their corresponding sequences.

For an additional independent test of the transferability of RACER under the same MHC allele, we used the benchmark set reported in ref. ⁴⁴. Four crystal structures (three TCRs) are curated in their benchmark set: 3QIB (2B4), 3QIU (226), 4P2Q (5cc7) and 4P2R (5cc7). Each of them have one strong-binding peptide presented in the crystal structure, and four weakly binding peptides. All the TCR–peptide pairs are associated with MHC-II allele I-E*, and three of them overlap with the main dataset reported in ref. ¹⁹. We therefore used the energy model previously trained from TCR 2B4 to test its transferability for the other three TCR–peptide pairs. The calculated binding energies were converted into a Z score by referencing to a set of 1,000 randomized peptides of corresponding TCRs: $Z = \frac{E_{\text{binding}} - E_{\text{decoys}}}{\sigma(E_{\text{decoys}})}$, with $\sigma(E_{\text{decoys}})$ being the standard deviation of E_{decoys} . The ROC curve and AUC score were calculated by scanning through different thresholds of the Z score. A further test by including more examples from ref. ⁴⁴ is available in Supplementary Note 6 and Supplementary Figs. 8 and 9.

Transferability without target TCR–peptide structure. To test the transferability of RACER without requiring any measured structure for a new TCR, we threaded the sequences of the CDR3 loops of the new TCR on the TCR structure used in our training. The length of the CDR3 β chain is the same among three TCRs (2B4: ASSLNWSQDTQY; 5cc7: ASSLNNSNDYT; 226: ASSLNNSNDYT), but the length of the CDR3 α chain varies (2B4: AALRATGGNNKLT; 5cc7: AAASNTNKVV; 226: AAEPSSGQKLV). To accommodate the difference when threading the CDR3 α sequences, we used a simple approach: aligning them based on the first two AA residues, leaving two gaps for TCR 5cc7 and 226. Modeller⁴⁵ was used to build the new loop structure based on these aligned new sequence, using the single structure in the training set as the template. These homology-modeled structures were then used for calculating the binding energies of the strong and weak binders of the new TCRs, using the trained energy model. We also omitted the step of structural relaxation when replacing a new peptide sequence on the built structure. Such an approach is unlikely to reduce RACER's performance, as demonstrated in Supplementary Fig. 13.

Leave-one-out cross validation. Leave-one-out cross validation (LOOCV) was used to test the predictive power of RACER to identify strong binders. Specifically, 1 of the 44 strong binders of TCR 2B4 was removed from the training set, and its predicted binding energy E_{pred} was compared with the experimentally determined weak binders. If the median of the weak binders' binding energies is larger than E_{pred} (a larger binding energy is associated with a smaller affinity), the testing strong binder is successfully identified. Similar tests were performed for TCR 5cc7 and TCR 226. The performance of RACER is compared with that from the clustering of peptide sequences using the algorithm from CD-Hit⁵³ (Supplementary Note 1).

Comparing results from surface plasmon resonance experiments. Surface plasmon resonance (SPR) was performed to assess the binding affinities of the three TCRs towards nine selected peptides¹⁹. The correlation between the predicted binding energies from RACER and the dissociation constant K_D evaluated from the SPR experiments thus constitutes a separate set of tests for the accuracy of RACER. We first built a relaxed structure with Modeller⁴⁵ for each of those TCR–peptide pairs, using the corresponding TCR structure as the template. We then used the optimized energy model of the corresponding TCR to evaluate the binding energy of each of those TCR–peptide pairs. The K_D values were obtained from fitting the SPR titration curves (supplementary figure 4f of ref. ¹⁹) using equation

$R_{\text{eq}} = \frac{C \times R_{\text{max}}}{C + K_D}$ with C , K_D and R_{max} as free parameters. The Pearson correlation coefficient and the Spearman's rank correlation coefficient between K_D Tlog (K_D) and predicted binding energies were used to quantify this correlation.

Evaluating contact residues of TCR–peptide pairs. The contact map of a given TCR–peptide structure was constructed by measuring the proximity W_{ij} between each residue of peptide (residue i) and CDR loops (residue j) based on their mutual distance, using a smoothed step function:

$$W_{ij} = \frac{1 - \tanh(d - d_{\text{max}})}{2} \quad (6)$$

Where $d_{\text{max}} = 6.5$ Å. Only C β atoms were included in our calculation (except for glycine, where the C α atom was used). The CDR3 loops were utilized as defined in the IEDB database⁵⁴. The constructed contact map represents those residues that are spatially close to each other in the given crystal structure.

Evaluating repertoire-level TCR–p-MHC interactions. To assess the statistical behavior of the inferential model, we calculated the pairwise binding interactions between a simulated T-cell population of size N_t and collection of $N_n = 10^4$ thymic self-peptides. For this proof-of-principle study, we used TCR 2B4 as an example, uniformly varying the 10^4 amino acids of the peptides, as well as those residues from the TCR that are in spatial contact with the peptide. TCR–peptide pairwise energies were calculated for $N_t = 10^5$ randomized TCR sequences using the RACER energy model optimized for TCR 2B4, and $N_t = 2,000$ for each of the TCR–p-I-E* systems given in Fig. 4 using energies weighted according to their contact maps, along with a model using a contact map with diagonal interactions (Fig. 6a). Substitution of TCR–peptide sequences with the newly generated ensemble yielded a total of $N_t \times N_n$ (10^9 in the 2B4 case; 2×10^7 for each of the cases involving the TCR–p-I-E* and diagonal contact maps) TCR–peptide pairs representing interactions occurring during thymic selection. Given our previous results (Supplementary Fig. 13), we avoid the computationally expensive task of structural relaxation, and instead calculate pairwise interactions with the original structure, requiring 5,000 CPU hours on an Intel Xeon CPU E5-2650 v2 for the large-scale 2B4-optimized simulation.

Thymic selection. Each T cell survives if the maximal interaction over all self-peptides does not exceed some upper threshold. Selection thresholds were chosen to achieve 50% (ref. ⁷). In all cases, the RACER-optimized energy model was used for energy assignment. Thymic selection was performed for each of the TCR–p-I-E* examples and their corresponding contact maps given in Fig. 4 (Fig. 6a). For each TCR–p-I-E* example, $N_t = 2,000$ pre-selection TCRs were created by uniformly varying the original TCR CDR3 α and β sequences over the amino acid space, keeping the sequence lengths unchanged. A similar randomization yielded $N_n = 10^4$ randomized peptide sequences representing self-peptides. For each of the 2,000 randomized TCRs, binding energies were calculated against the 10^4 self-peptides by selecting the corresponding entries in the RACER-optimized energy model weighted by the original TCR–p-I-E* contact maps, and the maximum energy was recorded. The fraction of TCRs whose maximal binding energy exceeded the selection threshold E_n traces the survival curves. This procedure, utilizing the RACER-optimized energy model, was repeated for a simplified model that utilizes only adjacent contacts (that is a strictly diagonal contact map with each entry having weight one) in the TCR–peptide interaction. The number of diagonal elements in the diagonal contact model was taken to be 20 (10 for each of the CDR3 α –peptide and CDR3 β –peptide pairs).

Self-peptide potency. Most self-peptides present in thymic selection are expected to participate in the deletion of self-reactive T cells. Thus, a reasonable model of thymic selection would feature a majority of self-peptides contributing to the selection of immature T cells. A rank order of these self-peptides based on their ability to recognize unique T cells, or potency, characterizes the extent to which each self-peptide is utilized in thymic selection. The rank order of potency was created for the RACER model utilizing the crystal structure of the 2B4 TCR (PDB ID: 3QIB) and its corresponding energy model derived from the set of experimentally determined strong binders. The thymic selection process using 10^4 self-peptides and 10^5 TCRs for the 2B4-optimized RACER model described above generates a total of 10^9 pairwise binding energies. The negative selection threshold E_n was selected to yield 50% selection, resulting in $\sim 5 \times 10^4$ deleted TCRs. The number of TCRs deleted by each self-peptide was recorded. The peptide deleting the most TCRs defines the most potent self-peptide. TCRs recognized by this peptide are removed from the list of total TCRs, and this peptide is similarly removed from the list of self-peptides. This process is repeated on the smaller TCR and self-peptide list to determine the second most potent peptide. Additional iteration until no TCRs remain provides the rank order of self-peptides in decreasing order of potency. The cumulative fraction of deleted relative to total TCRs is plotted in decreasing order of peptide potency.

T-cell antigen recognition probability. Utilizing the same post-selection T-cell repertoire from the previous section, post-selection T cells were quantified for their ability to recognize random non-self antigens and tumor neoantigens that differ from the N_n thymic self-peptides by one residue. 50% selection of TCRs result

in approximately 5×10^4 surviving, for which pairwise interactions are generated against 10^3 random and 10^3 point-mutated self-peptides, representing foreign and tumor-associated neoantigens, respectively (randomly generated peptides were checked to ensure non-membership in the set of thymic self-peptides). Estimates of individual TCR recognition probability were calculated by averaging the 5×10^4 -by- 10^3 indicator matrix, having values of 1 (0) corresponding to recognition (no recognition). Estimates of the corresponding recognition probability for the entire post-selection MHC-restricted T-cell repertoire was calculated by assessing the 1-by- 10^3 vector indicating the presence or absence of at least one recognizing TCR. The post-selection individual and repertoire T-cell recognition probabilities of random and point-mutant antigens were then compared with previously derived analytic results for two random energy models⁴².

Data availability

The data comprising the peptides recognized by the three TCRs, used for RACER training and testing, are available in ref.¹⁹. An extended data set of these three TCRs was uploaded at Zenodo⁵⁶. The additional data used for training and testing on different MHC-II TCRs can be found in ref.⁴⁴.

Code availability

The full code, along with a demo for predicting TCR-peptide interaction, as well as being applied to a collection of randomly generated TCRs and peptides, have been deposited to Code Ocean⁵⁷ and can also be found at <https://github.com/XingchengLin/RACER.git>.

Received: 16 September 2020; Accepted: 23 April 2021;

Published online: 24 May 2021

References

- Couzin-Frankel, J. Cancer immunotherapy. *Science* **342**, 1432–1433 (2013).
- Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271**, 1734–1736 (1996).
- Ott, P. A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
- Moldrem, J. J., Komanduri, K. & Wieder, E. Overexpressed differentiation antigens as targets of graft-versus-leukemia reactions. *Curr. Opin. Hematol.* **9**, 503–508 (2002).
- Abbas, A. K., Lichtman, A. H. & Pillai, S. *Cellular and Molecular Immunology* (Elsevier, 2018).
- De Boer, R. J. & Perelson, A. S. How diverse should the immune system be? *Proc. Biol. Sci.* **252**, 171–175 (1993).
- Yates, A. J. Theories and quantification of thymic selection. *Front. Immunol.* **5**, 13 (2014).
- Nossal, G. J. Negative selection of lymphocytes. *Cell* **76**, 229–239 (1994).
- Kosmrlj, A., Jha, A. K., Huseby, E. S., Kardar, M. & Chakraborty, A. K. How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl Acad. Sci. USA* **105**, 16671–16676 (2008).
- Davis, M. M. Not-so-negative selection. *Immunity* **43**, 833–835 (2015).
- Detours, V., Mehr, R. & Perelson, A. S. A quantitative theory of affinity-driven T cell repertoire selection. *J. Theor. Biol.* **200**, 389–403 (1999).
- George, J. T., Kessler, D. A. & Levine, H. Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides. *Proc. Natl Acad. Sci. USA* **114**, E7875–E7881 (2017).
- Mayer, A., Balasubramanian, V., Walczak, A. M. & Mora, T. How a well-adapting immune system remembers. *Proc. Natl Acad. Sci. USA* **116**, 8815–8823 (2019).
- Altan-Bonnet, G., Mora, T. & Walczak, A. M. Quantitative immunology for physicists. *Phys. Rep.* **849**, 1–83 (2020).
- George, J. T. & Levine, H. Stochastic modeling of tumor progression and immune evasion. *J. Theor. Biol.* **458**, 148–155 (2018).
- George, J. T. & Levine, H. Sustained coevolution in a stochastic model of cancer-immune interaction. *Cancer Res.* **80**, 811–819 (2020).
- Riley, T. P. et al. T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability. *Nat. Chem. Biol.* **14**, 934–942 (2018).
- Singh, N. K. et al. Emerging concepts in TCR specificity: rationalizing and (maybe) predicting outcomes. *J. Immunol.* **199**, 2203–2213 (2017).
- Birnbaum, M. E. et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).
- Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
- Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
- Abella, J. R., Antunes, D. A., Clementi, C. & Kavraki, L. E. APE-Gen: a fast method for generating ensembles of bound peptide-mhc conformations. *Molecules* **24**, 881 (2019).
- Abella, J. R., Antunes, D. A., Clementi, C. & Kavraki, L. E. Large-scale structure-based prediction of stable peptide binding to class I HLAs using random forests. *Front. Immunol.* **11**, 1583 (2020).
- Chen, B. et al. Predicting HLA class II antigen presentation through integrated deep learning. *Nat. Biotechnol.* **37**, 1332–1343 (2019).
- Jurtz, V. I. et al. NetTCR: Sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. Preprint at *bioRxiv* <https://doi.org/10.1101/433706> (2018).
- Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* **11**, 1803 (2020).
- Gowthaman, R. & Pierce, B. G. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Res.* **46**, W396–W401 (2018).
- Pierce, B. G. & Weng, Z. A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Sci.* **22**, 35–46 (2013).
- Ishizuka, J. et al. Quantitating T cell cross-reactivity for unrelated peptide antigens. *J. Immunol.* **183**, 4337–4345 (2009).
- Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953 (2000).
- Wang, J. & Verkhivker, G. M. Energy landscape theory, funnels, specificity, and optimal criterion of biomolecular binding. *Phys. Rev. Lett.* **90**, 188101 (2003).
- Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA* **84**, 7524–7528 (1987).
- Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. Improved design of stable and fast-folding model proteins. *Fold. Des.* **1**, 221–230 (1996).
- Schaefer, N. P., Kim, B. L., Zheng, W. & Wolynes, P. G. Learning to fold proteins using energy landscape theory. *Isr. J. Chem.* **54**, 1311–1337 (2014).
- Davtyan, A. et al. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **116**, 8494–8503 (2012).
- Wang, J. et al. Quantifying intrinsic specificity: a potential complement to affinity in drug screening. *Phys. Rev. Lett.* **99**, 198101 (2007).
- Yan, Z., Zheng, X., Wang, E. & Wang, J. Thermodynamic and kinetic specificities of ligand binding. *Chem. Sci.* **4**, 2387–2395 (2013).
- Alspach, E. et al. MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* **574**, 696–701 (2019).
- Castle, J. C. et al. Exploiting the mutanome for tumor vaccination. *Cancer Res.* **72**, 1081–1091 (2012).
- Ogishi, M. & Yotsuyanagi, H. Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* **10**, 827 (2019).
- Newell, E. W. et al. Structural basis of specificity and cross-reactivity in T cell receptors specific for cytochrome c –I-E^k. *J. Immunol.* **186**, 5823–5832 (2011).
- Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl Acad. Sci. USA* **89**, 9029–9033 (1992).
- Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).
- Lanzarotti, E., Marcatili, P. & Nielsen, M. Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. *Mol. Immunol.* **94**, 91–97 (2018).
- Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* **54**, 5.6.1–5.6.37 (2016).
- Sinclair, C., Bains, I., Yates, A. J. & Seddon, B. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proc. Natl Acad. Sci. USA* **110**, E2905–E2914 (2013).
- Zerrahn, J., Held, W. & Raulet, D. H. The MHC reactivity of the T cell repertoire prior to positive and negative selection. *Cell* **88**, 627–636 (1997).
- Kapcha, L. H. & Rossy, P. J. A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J. Mol. Biol.* **426**, 484–498 (2014).
- Nielsen, M. et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017 (2003).
- George, J. T. & Levine, H. Implications of tumor-immune coevolution on cancer evasion and optimized immunotherapy. *Trends Cancer* **7**, 373–383 (2021).
- Papioan, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z. & Wolynes, P. G. From the cover: water in protein structure prediction. *Proc. Natl Acad. Sci. USA* **101**, 3352–3357 (2004).
- Koretke, K. K., Luthey-Schulten, Z. & Wolynes, P. G. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc. Natl Acad. Sci. USA* **95**, 2932–2937 (1998).
- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
- Vita, R. et al. The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–412 (2015).

55. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
56. Lin, X. XingchengLin/RACER v.1.0 *Zenodo* <https://doi.org/10.5281/zenodo.4699821> (2021).
57. Lin, X. et al. Codes for rapid assessment of t-cell receptor specificity of the immune repertoire. *Code Ocean* <https://doi.org/10.24433/CO.3204967.v2> (20 April 2021).

Acknowledgements

Work at the Center for Theoretical Biological Physics was sponsored by the NSF (grant PHY-2019745). H.L. was also supported by the NSF (grant PHY-1935762). J.N.O. was also supported by the NSF (grant CHE-1614101) and the Welch Foundation (grant C-1792). J.T.G. was also supported by the National Cancer Institute of the NIH (F30CA213878). J.N.O. is a CPRIT Scholar in Cancer Research. C.C. was also supported by NSF (grant CHE-1738990 and CHE-1900374), Welch Foundation (grant C-1570), Deutsche Forschungsgemeinschaft (SFB/TRR 186/A12, and SFB 1078/C7) and Einstein Foundation Berlin, Germany.

Author contributions

X.L., J.T.G., J.N.O. and H.L. conceived of and designed the research. X.L., J.T.G., N.P.S. and K.N.C. performed the research. X.L., J.T.G., N.P.S., K.N.C., C.C., J.N.O. and H.L.

analyzed the data. M.E.B. supplied additional experimental data. X.L., J.T.G., J.N.O. and H.L. created new software used in the work. X.L., J.T.G., N.P.S., K.N.C., C.C., J.N.O. and H.L. drafted the work or substantively revised it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43588-021-00076-1>.

Correspondence and requests for materials should be addressed to J.N.O. or H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Peer review Information *Nature Computational Science* thanks Sergei Grudinin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Handling editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021