Supporting Information for Rapid Assessment of T-Cell Receptor Specificity of the Immune Repertoire

Xingcheng Lin[1,2,3,*], Jason T. George[1,4,*], Nicholas P. Schafer[1,5], Kevin Ng Chau[6], Michael E. Birnbaum[7,8,9], Cecilia Clementi[1,5,7], José N. Onuchic[1,2,5,11,†], and Herbert Levine[1,6,†]

[1]Center for Theoretical Biological Physics, Rice University, Houston, TX
[2]Department of Physics and Astronomy, Rice University, Houston, TX
[3]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA
[4]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX
[5]Departments of Chemistry, Rice University, Houston, TX
[6]Department of Physics, Northeastern University, Boston, MA
[7]Koch Institute for Integrative Cancer Research, Cambridge, MA
[8]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA
[9]Ragon Institute of MIT, MGH, and Harvard, Cambridge, MA
[10]Department of Physics, Freie Universität, Berlin, Germany
[11]Department of Biosciences, Rice University, Houston, TX
[*]Equal contribution
[†]To whom correspondence should be addressed: jonuchic@rice.edu, h.levine@northeastern.edu

Contents

1 Sequence-based peptide clustering

To assess the ability to separate strong binders from weak binders based only on the peptide sequences, CD-Hit [1, 2], a greedy incremental clustering algorithm, was used for classifying the peptide sequences of TCR 2B4. The input sequences contained a total of 44 experimentally determined strong binders (including the native peptide present in the crystal structure), as well as 231 experimentally determined weak binders as described in the main text. A strong binder was identified if it is classified in the same category of the native sequence. The best performance of CD-Hit correctly identifies the cluster including 19 strong binders (native peptide included) and no weak binders, using a sequence identity threshold of 0.5.

2 Additional hold-out tests on an extended dataset

To test the limit of RACER's transferability over a more diversified coverage of peptide sequence, we included more strong binders from the [3], where all peptides of TCR 2B4 that ends up with more than one copy from the deep-sequencing experiments were included, constituting an extended dataset. RACER was applied on a more demanding set of hold-out tests on the original dataset used in the main text, as well as this more extended dataset.

In the leave-50%-out-test, these strong binders were randomly shuffled before partitioned into two sets. One set was used in the training set, the other set, together with the experimentally determined weak binders, were used as the testing set. We switched two sets of strong binders for an equivalent testing, therefore constituting two testing cases for each data set. As shown in Fig. 4, the leave-50%-out-test demonstrates that RACER can fully separate strong binders from the weak ones, with an average recognition Z-score equal to 5.26 for the original dataset, and 4.63 for the extended dataset.

In the leave-90%-out-test, the strong binders were randomly shuffled into 10 sets, and we only used one of them for training, the other 9 sets, together with the experimentally determined weak binders, were used as the testing set. Therefore, we have 10 testing cases for each dataset. As shown in Fig. 4, the leave-90%-out-test again demonstrate RACER's success in distinguishing strong binders from weak ones in both the original dataset (average Z-score of 4.74) and the extended dataset (average Z-score of 4.50).

To push the limit of RACER's predictive power, we added an additional leave-99%-out test for the extended dataset, where ~ 4 strong binders were included in the training set, and a leave-one-in test, where only one strong binder was included in the training set. We summarize the percentage of strong peptides that failed to be detected (failure is defined to occur whenever the binding energies of the withheld binders are larger than the median of the weak binders). As shown in Fig. 3, the amount of peptides that can be recognized by RACER gradually decreases as fewer peptides were included in training, and if we only include one peptide in our training, the performance of RACER is worse than an alternative test that uses the identity of sequences based on the native peptide of the crystal structure (PDB ID: 3QIB).

3 Sequence diversity in the leave-one-out test

To see the coverage of sequence diversity of peptides that succeeded or failed to be recognized by RACER, we calculated the sequence identity of the peptide sequences that were used in our leave-one-out test of the TCR 2B4, based on the native peptide presented in the crystal structure (PDB ID: 3QIB). As shown in Fig. 1, RACER capably recognizes strong binders with small sequence identity in both the original dataset and the extended dataset, with some cases having little to no sequence identity.

4 The standard protein force field cannot fully resolve strong binders from weak binders

Two commonly used force fields were utilized to test the performance of standard protein force fields for distinguishing strong TCR binders from weak ones. This analysis was applied to the three TCRs investigated in the main text. The default AWSEM force field [4] was previously optimized for folding protein structures [5]. The Miyazawa-Jernigan (MJ) potential [6] is one of the most widely used knowledge-based force fields,

derived from a statistical analysis of large protein repositories. Both force fields have been demonstrated to perform well in describing the structural dynamics of generic proteins. However, when replacing our optimized TCR parameters with these two force fields, the resulting model is unable to clearly separate the strong binders from the weak ones. (Fig. 5).

5 Improving predictive accuracy by including target TCR structures

For blind assessment of TCR transferability, we ask whether we can improve prediction accuracy if there are available strong binders presented in the structures of the target TCRs. To test this, we added the set of interactions calculated from the crystal structures of the other two TCRs as two additional strong binders in the training set. For example, in the case of TCR 2B4, the set of interactions from the crystal structures of TCR 5CC7 and 226 were added into the training set of TCR 2B4, constituting a total of 46 strong binders. Same procedure was repeated for TCR 5CC7 and TCR 226. We find that the new energy model demonstrates significant improvement in Z-scores (Fig. 7, compared with Fig. 5a). These results suggest that further incorporation of additional crystal structures of target TCRs may lead to improved resolution of strong and weak binders via refinement of the optimized energy model.

6 Extended test of RACER's transferability across different TCRs restricted to the same MHC-II allele

To test the transferability of RACER beyond the coverage of the TCR-p-MHCs used in the main text, we further tested the performance of RACER on the data provided in [7]. The data includes all the TCRs associated with different MHC-II alleles until 2017. One strongly binding peptide, and four weakly binding peptides were provided for each TCR. We used RACER to perform a leave-one-out prediction for all the cases where more than one TCRs are shared among the same allele, by excluding one TCR from training and using the optimized energy model to predict the binding affinity of the withheld TCR. The first 50 eigenvectors of the B matrix in Eq. (5) (see Method for details) were found to be well-determined. The influence of the remaining eigenvectors of the B matrix on the optimized interaction parameters in γ was reduced according to a filtering scheme (see Method for details). As shown in Figure 6, RACER was able to recognize the strong peptide (Z-score > 1) for 21 out of the 26 tests. It is worth noting that there are many cases where TCRs associated with the same MHC allele share different Vα and Vβ genes. RACER works less well for cases where only 2 TCRs are available, and better when there are $3 \sim 5$ cases, regardless of whether they share the same or different Vα and Vβ genes. This additional test further supports the predictive power of RACER trained with a small set (around 3 to 5 copies) of available TCR-peptide structures/sequences. To challenge RACER's predictive capacity when statistical learning is performed on a TCR-pMHC pair distinct from either the target predicted TCR or peptide, we intentionally selected a TCR-pMHC structure with the same MHC allele having different Vα and Vβ genes, where available, from the target as the template. We replaced the CDR3 loops with target sequences using trivial alignment (case III of Fig. 1), and repeat the same test as above. As shown by Fig. 9, RACER can still recognize 19 out of the 26 examples (Z-score > 1). The success of this test highlight RACER's predictive power across different TCRs associated with the same MHC allele.

7 Comparison with ERGO, a sequence-based predictor trained by neural network

ERGO [8] is a sequence-based TCR-peptide prediction tool trained by neural networks. ERGO implemented two types of models: Long short-term memory (LSTM) and Autoencoder, together with two training datasets: McPAS-TCR (2×10^4 TCR by 300 peptides) and VDJdb (4×10^4 TCR by 200 peptides). We applied ERGO to calculate the binding scores of the strong and weak binders of the three TCRs in [3]. We found ERGO performs best when the Autoencoder was applied based on the VDJdb database. As shown in Fig. 10, ERGO can only recognize the strong binders of TCR 5CC7 with a Z-score of 2.85.

8 Additional statistical evaluation of RACER-derived thymic selection

We have argued in our previous paper [9] that a well-functioning immune system should utilize a majority of thymic self-peptides in the deletion of self-reactive T-cells. This desideratum can be used to determine if a high-throughput model is behaving in a statistically sensible manner; specifically, a reasonable model of thymic selection would feature most self-peptides contributing to T-cell selection. A rank order of self-peptides based on their ability to recognize T-cells, defined as the peptide potency, characterizes the extent to which each self-peptide is utilized in thymic selection. The RACER-derived potency using 2B4-optimized data generates reasonable behavior with respect to this criterion (Fig. 11a). In addition, our prior theoretical work posited that optimal thymic selection occurs at survival cutoffs near $1/e$ [10, 9]. Calculating the product of survival and recognition probabilities yields a broad curve maximized at intermediate selection cutoffs, including $1/e$ (Fig. 11c).

We also compared RACER-derived repertoire-level CDR3 sequence similarity to experimentally determined antigen-specific T-cell repertoires [11]. Our post-selection simulated TCRs recognizing the top 10 foreign antigens were collected and clustered using a discrete Hamming metric with CDR3 sequence weights as in [11]. Dendrograms obtained from hierarchical clustering identified a diverse set of TCRs (Fig. 12a). Because our model sampled a sparse (10^5) subset of CDR3 sequence space, we then augmented our repertoire by *in silico* site-directed mutagenesis to include 100 closely related TCRs for each foreign antigen. This augmented antigen-specific repertoire recapitulates features of experimentally determined antigen-specific repertoires comprised of diverse and homologous clusters of TCR sequences (Fig. 12b), and demonstrates RACER's ability to identify diverse TCRs with shared antigen specificity.

9 Detailed future development of RACER

While RACER effectively resolved strong and weak binders of [7] in all cases where the training and test peptide were identical, approximately half of the cases examined here contained training and test peptides that are dissimilar (Fig. 9). For these cases where training and test peptides are different, RACER correctly predicts 67% of the examples (Z-score > 1.0). The resulting predictive accuracy demonstrates that our structurally informed pairwise model can resolve TCR-p-MHC specificity in a majority of available test cases. Further experimental validation will be required to definitively assess RACER's ability to resolve TCR-p-MHC specificity across all possible TCR-peptide pairs within a given MHC allele. This challenge remains a top priority for future investigations on repertoire-level TCR-peptide assessment.

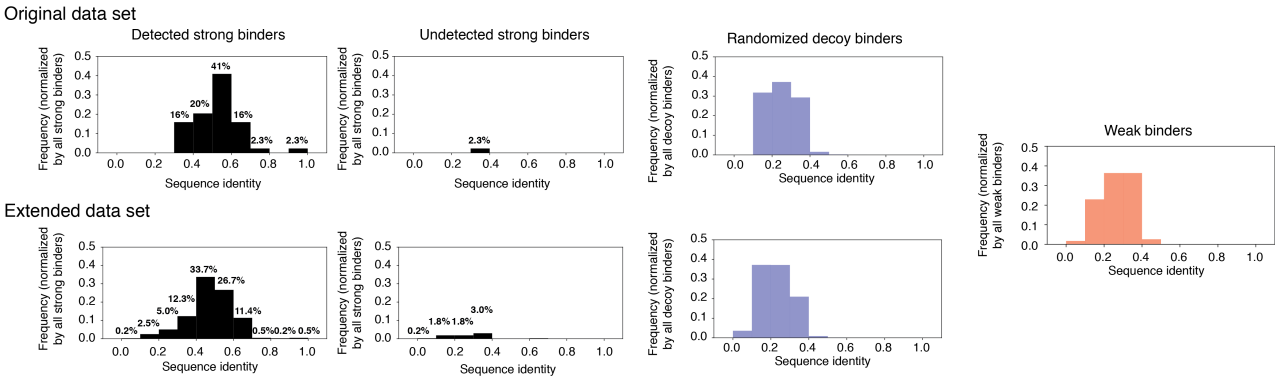
In designing RACER to achieve rapid and accurate predictions, our calculation only includes pairwise energetic interactions, while omitting other contributions to peptide binding affinity, such as conformational entropy. This is in line with analogous assumptions when computing mutational free energies utilizing potentials derived from protein co-evolutionary information (discussed in [12]) where there is a careful discussion on how the leading order binding energy differences are mostly due to direct residue-residue interactions. While RACER maintains reasonably high predictive accuracy, more accurate assessments of the TCR-p-MHC binding free energy will likely lead to improvements and will be a focus of subsequent work.

10 Contact maps and their influence of T-cell survival and recognition

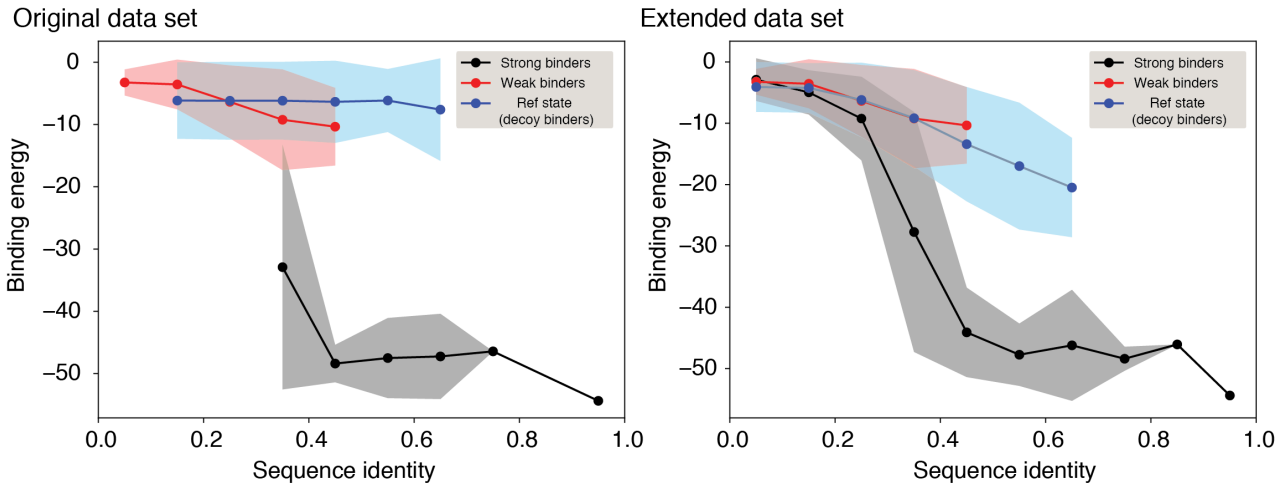
In cases with available crystal structures, contact map analysis revealed a largely conserved interaction pattern for TCR-peptide pairs associated with the IE^k MHC-II allele (Fig. 4), providing an explanation for the transferability of RACER-derived interactions when trained on a particular crystal structure. Moreover, these results demonstrate how differences in the contact maps may manifest as shifts in the mean binding energy between T-cells and thymic self-peptides, thereby affecting a TCR's post-thymic selection inclusion probability (Fig. 6). Previous investigations have characterized the probability distribution for generating particular TCR sequences in VDJ recombination, and have even suggested that the *a posteriori* observed post-selection TCRs had greater generation probabilities [13, 14], with so-called "public" TCR sequences being observed in multiple

individuals. Incorporation of contact maps into our generative model contributes to variations in T-cell survival probability, and may offer a physical interpretation of why public repertoires survive thymic selection at higher rates [15], in addition to providing an explicit means of estimating post-selection T-cell prevalence within a given MHC-class restriction.

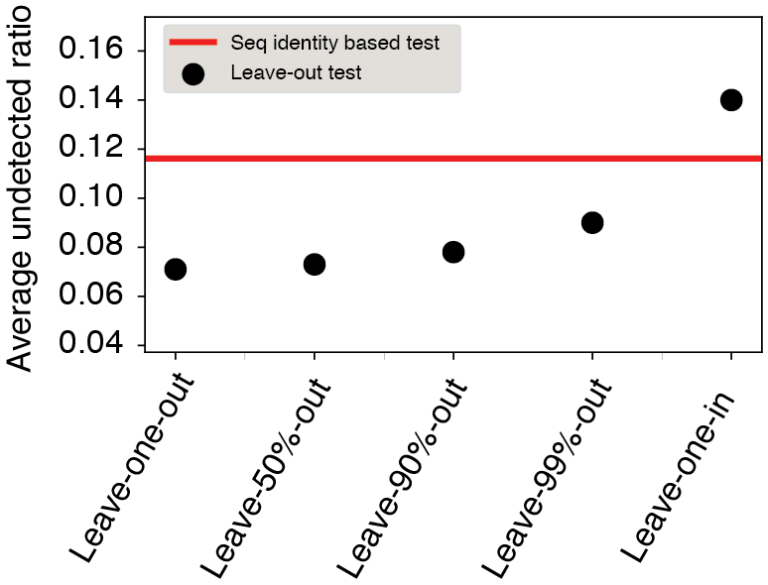
11 Supplementary Figures 1-13 and Table 1



Supplementary Figure 1: The normalized count frequency of the sequence identity, calculated based on the native peptide of TCR 2B4, between the experimentally determined strong binders (black), randomized decoy binders (blue), and experimentally determined weak binders (orange). For the strong binders, the probability distributions are further organized based on whether or not they are successfully detected in the leave-one-out test. For the strong binders, the normalization was carried out with the total number of strong binders being the normalization factor, to emphasize the contrast between the number of detected/undetected cases. Detailed percentage of histograms within each bin are noted. We show the distribution of sequence identities for the original dataset (including all strong binders where the final copies of peptides from the experiment were amplified to contain at least 50 copies following the affinity-based selection), and an extended dataset (including additional peptides where the final copy numbers were larger than 1).



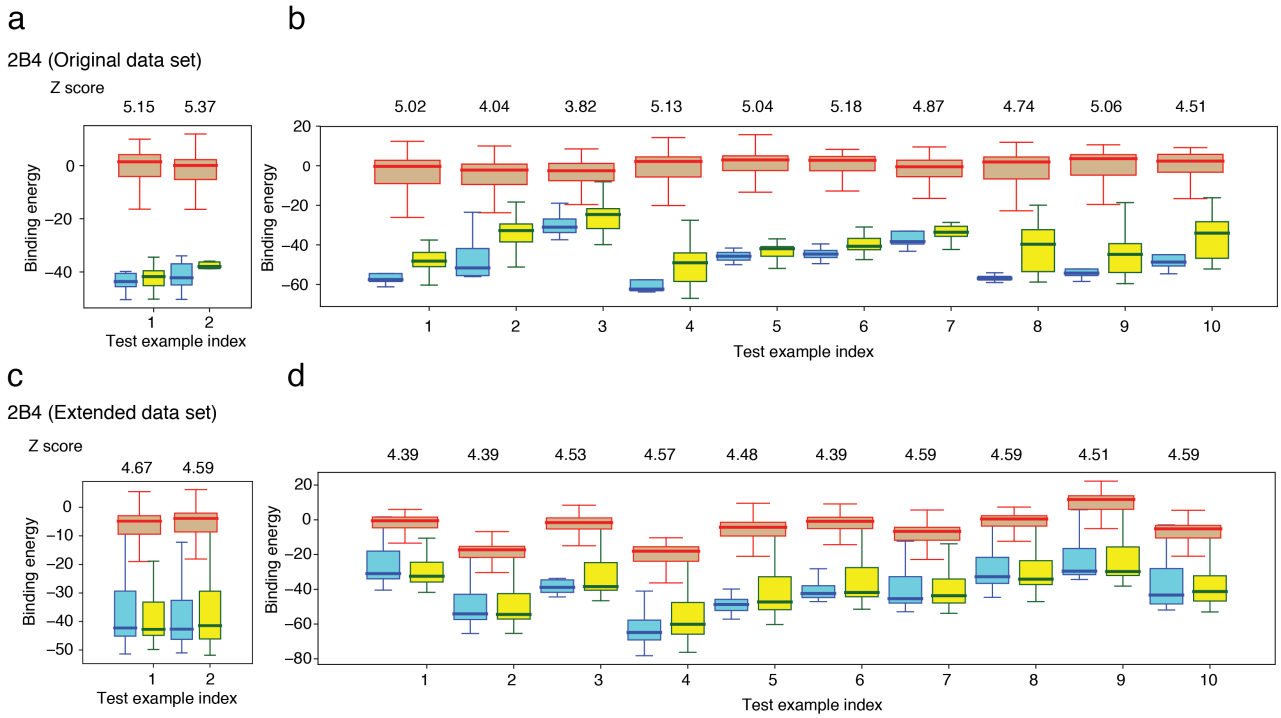
Supplementary Figure 2: RACER-derived binding energies of peptides of TCR 2B4 (Fig. 2a) as a function of sequence identity, calculated based on the native peptide of the crystal structure (PDB ID: 3QIB). Solid lines indicate average binding energy, and their corresponding shaded regions depict standard deviations. Binding energies were calculated based on strong binders (black), experimentally determined weak binders (red) and the randomized decoys of strong binders (blue). The energies for the original data set (**left**) and extended data set (**right**) were both presented.



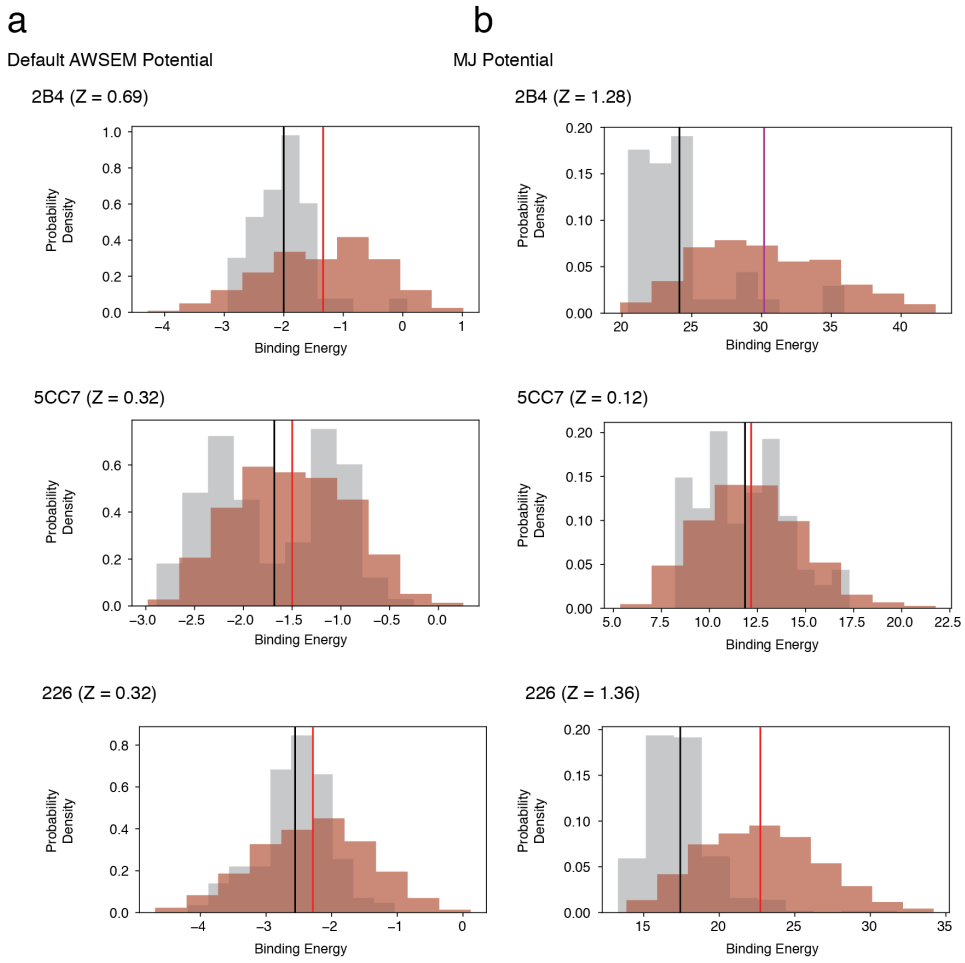
Supplementary Figure 3: Cross-validation test of TCR 2B4 with RACER for the extended dataset, where more diversified strong binders were included for a more comprehensive test. RACER performs well, reliably detecting $> 99\%$ of the strong binders, until leaving 99% of the peptides out. At this point RACER's performance deteriorates due to a lack of training data. When only one peptide is included in the training set, RACER performs worse than a selection based only on the sequence identity of peptides calculated based on the native peptide in the crystal structure (PDB ID: 3QIB) of the TCR 2B4.

TCR	Gene usage (alpha)	Gene usage (beta)	CDR sequences (alpha)	CDR sequences (beta)	Native peptide sequence
2B4	V:TRAV4N-4*01 J:TRAJ56*01	V:TRBV26*01 J:TRBJ2-5*01	CDR1:TTMRA CDR2:LASGT CDR3:AALRAT GGNNKLT	CDR1:KGHPV CDR2:FQNQEV CDR3:ASSLNW SQDTQY	ADLIAYLKQA TKG
5CC7	V:TRAV4N-4*01 J:TRAJ34*01	V:TRBV26*01 J:TRBJ1-2*01	CDR1:TTMRA CDR2:LASGT CDR3:AAEASN TNKVV	CDR1:KGHPV CDR2:FQNQEV CDR3:ASSLNN ANSDYT	ANGVAFFLTP FKA
226	V:TRAV4N-4*01 J:TRAJ16*01	V:TRBV26*01 J:TRBJ1-2*01	CDR1:TTMRA CDR2:LASGT CDR3:AAEPSS GQKLV	CDR1:KGHPV CDR2:FQNQEV CDR3:ASSLNN ANSDYT	ADLIAYLKQA TKG

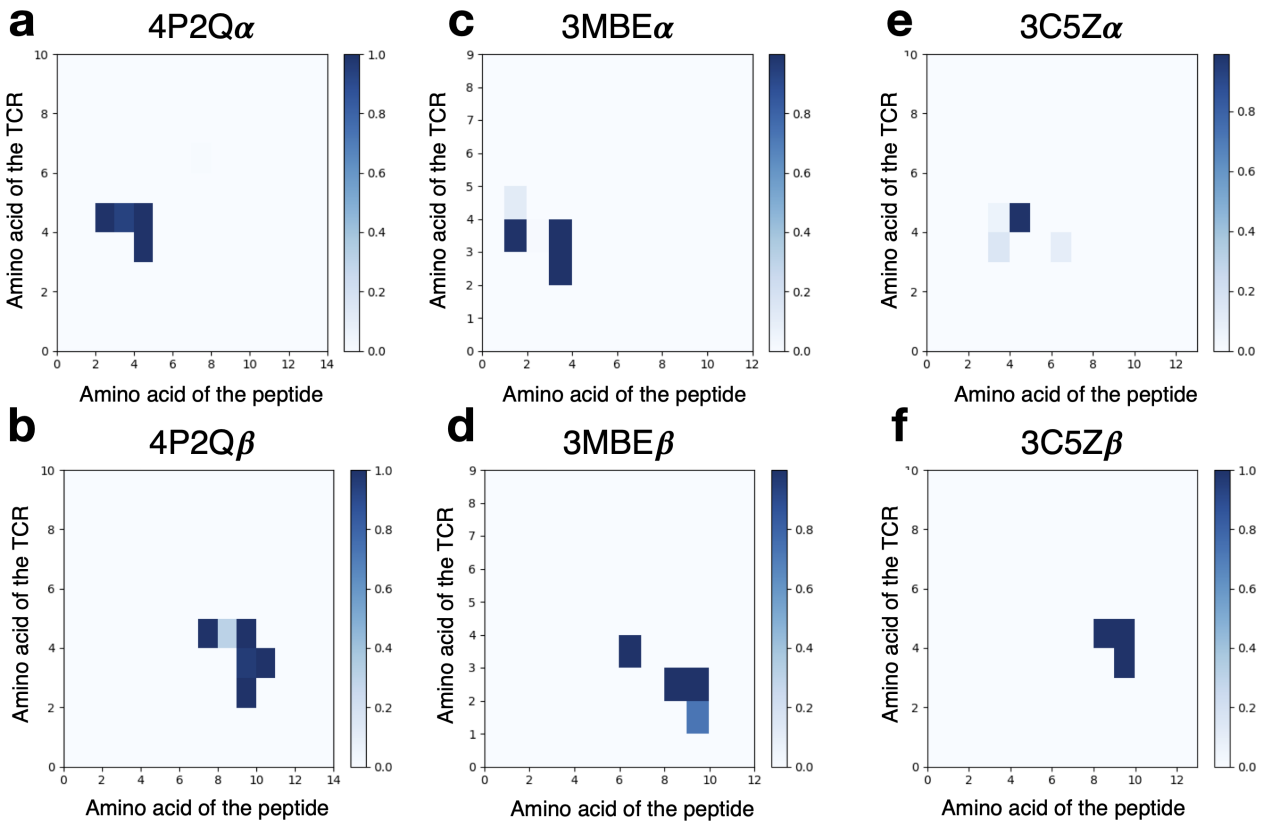
Supplementary Table 1: Detailed information[16] of TCR 2B4, 5CC7 and 226 used in the main text. The three TCRs used in our test shared the same V gene, their J genes are different from each other, resulting in different CDR3 sequences.



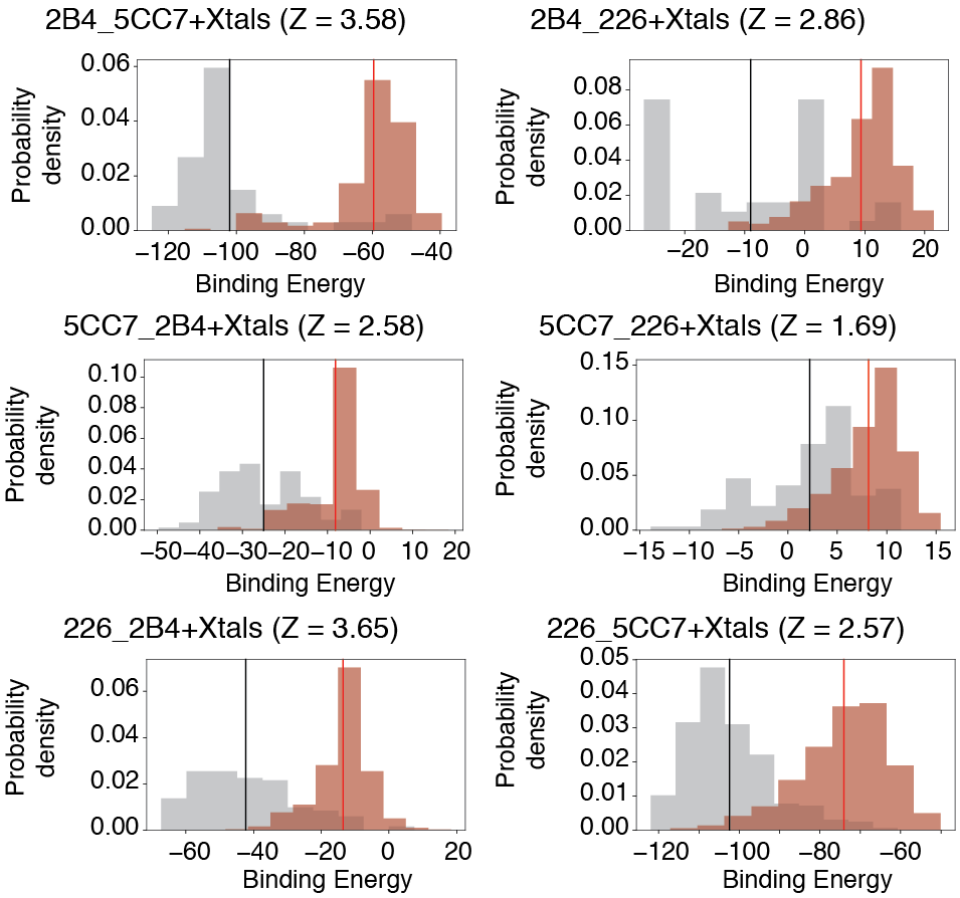
Supplementary Figure 4: Cross-validation test of TCR 2B4 with RACER (main text Fig. 2) where 50% (**a, c,**) and 10% (**b**, **d**) of the strong binders were used as the training set (blue). The predicted binding energies of the 50% of withheld strong binders (yellow) are lower than the binding energies of the experimentally determined weak binders (brown). The median of each set of binders was shown as a bar in the corresponding box plot. The whiskers are placed at the first and last datum points that fall within (m, M), where m = Q1 - 1.5IQR and M = Q3 + 1.5IQR, with IQR = Q3 - Q1 representing the interquartile range. The calculated Z-score of each test was shown at the top. In both the original (**a, b,**) and extended (**c**,**d**) dataset, the leave-50%-out and leave-90%-out test demonstrate RACER's predictive capacity for recognizing strong binders of the TCR 2B4.



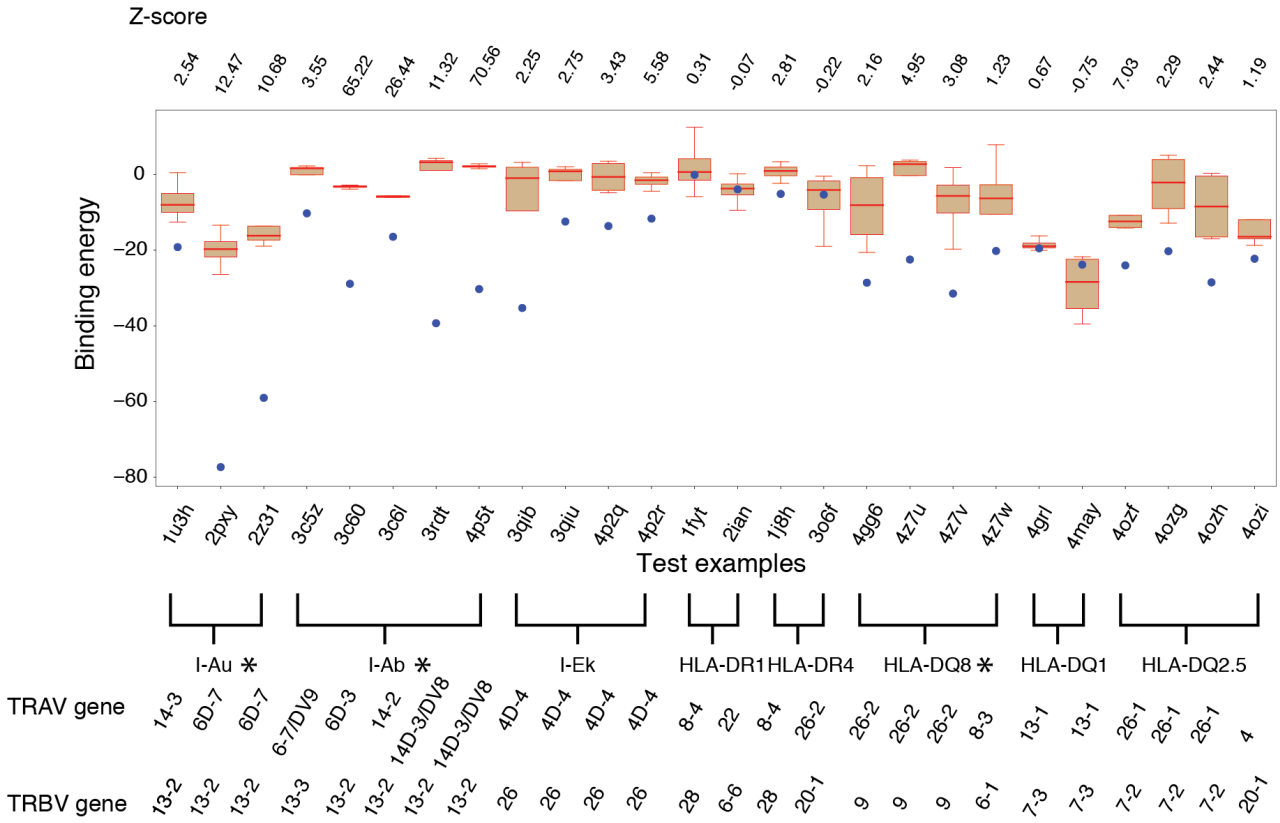
Supplementary Figure 5: The protein interaction parameters from standard force field cannot fully separate strong binders from weak ones. **a,** The performance using the default parameters from the AWSEM force field [4]. **b,** The performance using the parameters from the Miyazawa-Jernigan (MJ) potential [6]. Compare this with the main text Figure 2 shows the advantage of RACER in terms of identifying strong binders from weak binders.



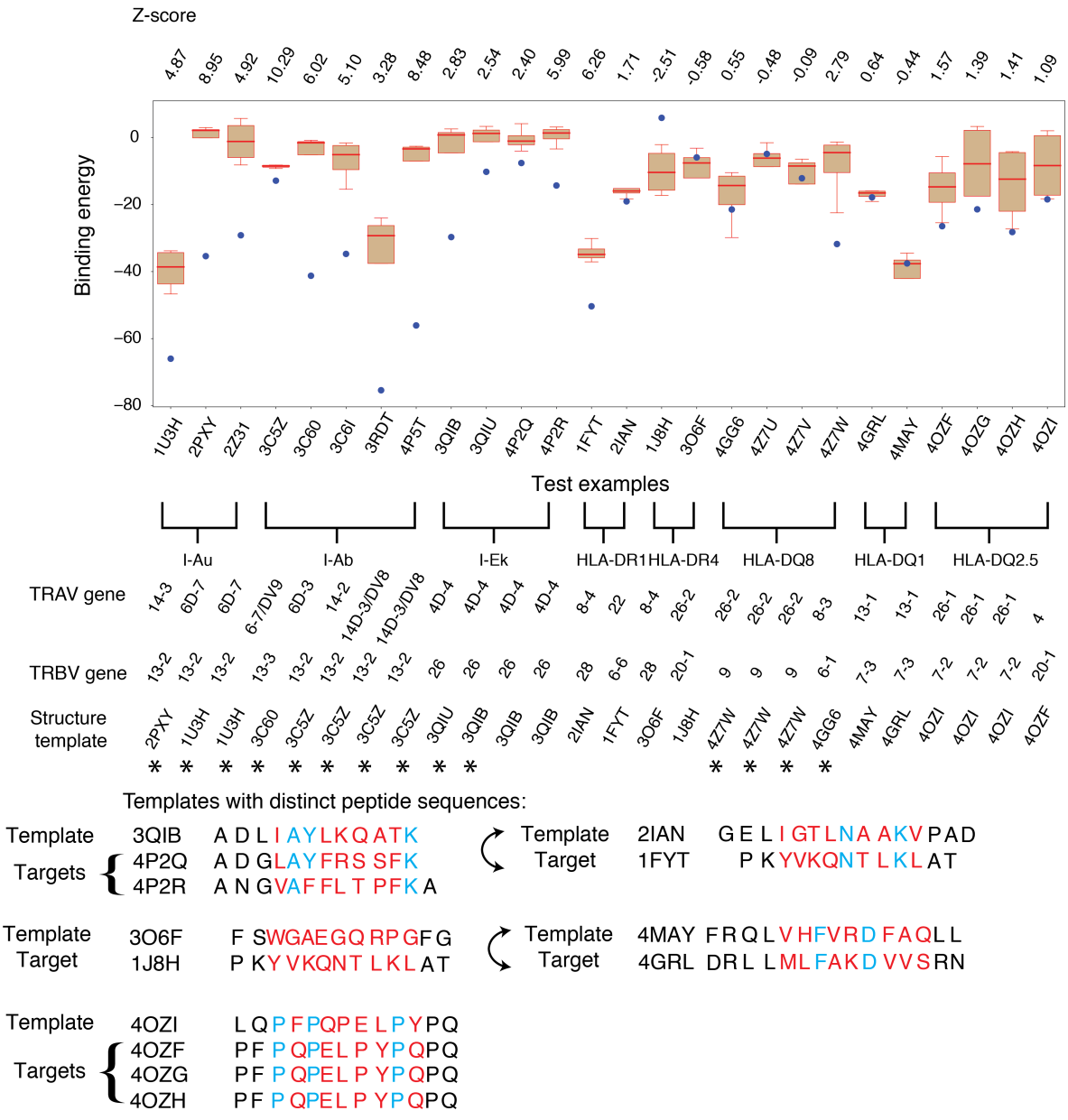
Supplementary Figure 6: The contact maps of TCR-peptide pairs associated with different MHCII alleles share less structural similarity, compared with the main text Figure 4. Contact maps are calculated using distances from each pairwise TCR-peptide amino acid combination using Eq. 6 for the following TCR-p-MHC pairs: 4P2Q - peptide ADGLAYFRSSFK presented by MHC-II IE^k to TCR 5cc7 **a,** CDR3α (AAEASNTNKVV) and **b,** CDR3β (ASSLNNANSDYT); 3MBE - peptide AMKRHGLDNYRG presented by MHC-II IA^g to TCR 21.3 **c,** CDR3α (AAEDGGSGNKLI) and **d,** CDR3β (ASSWDRAGNTLY); 3C5Z - peptide FEAWKAKANKA presented by MHC-II IA^b to TCR B3K506 **e,** CDR3α (ALVISNTNKVV) and **f,** CDR3β (ASIDSSGNTLY).



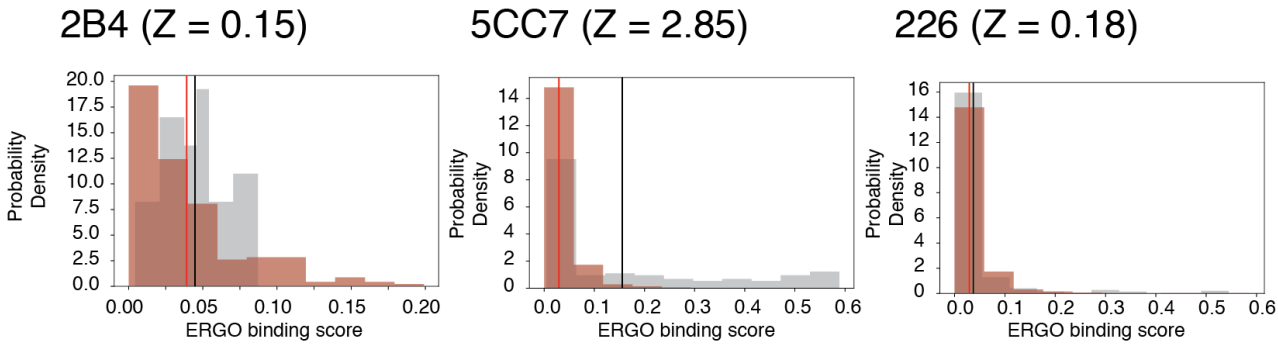
Supplementary Figure 7: Probability density distributions of the predicted binding energies of experimentally determined strong (brown, with mean depicted in red) and weak (grey, with mean depicted in black) binders of each of the three TCRs (2B4, 5CC7 and 226), using another TCR for training. In addition, structures of the other two TCRs are included into the training sets. The title of each figures follows the format of "target_training TCRs+Xtals", with "Xtals" means the crystal structures of the other two TCRs were added into the training set.



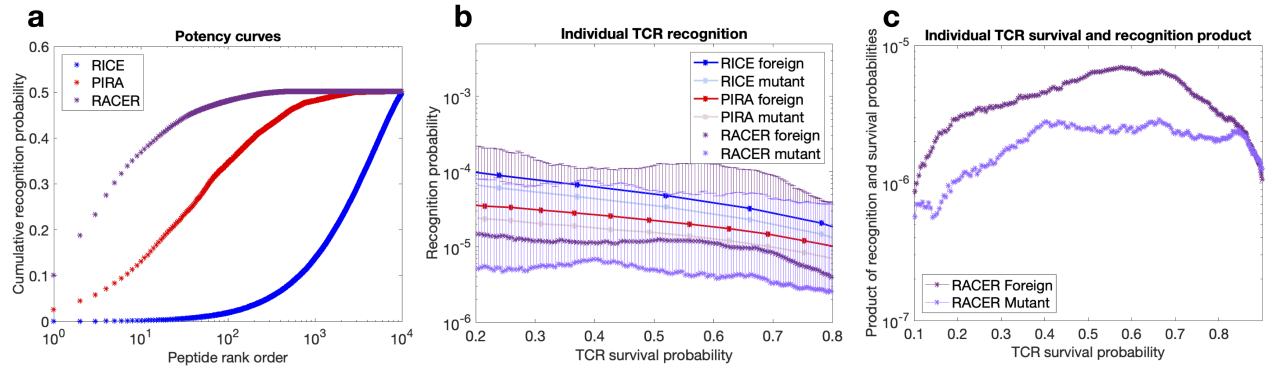
Supplementary Figure 8: Further leave-one-out test of RACER's predictive transferability over TCRs, using data from [7]. The TCRs were grouped by their associated MHC allele, with their Vα and Vβ genes noted at the bottom. Asterisks are marked for TCRs sharing identical peptides within the same allele. RACER successfully predicted lower binding energies for strong binders (blue) relative to weak binders (brown). The prediction Z-score is provided above each case. RACER was able to successfully recognize the strong-binding peptide (Z-score > 1) for 21 out of the 26 tests. RACER's predictive accuracy is reduced in cases where only 2 TCRs are available, and shows improvement when there are 3$\sim$5 cases, regardless of whether they share the same Vα and Vβ genes.



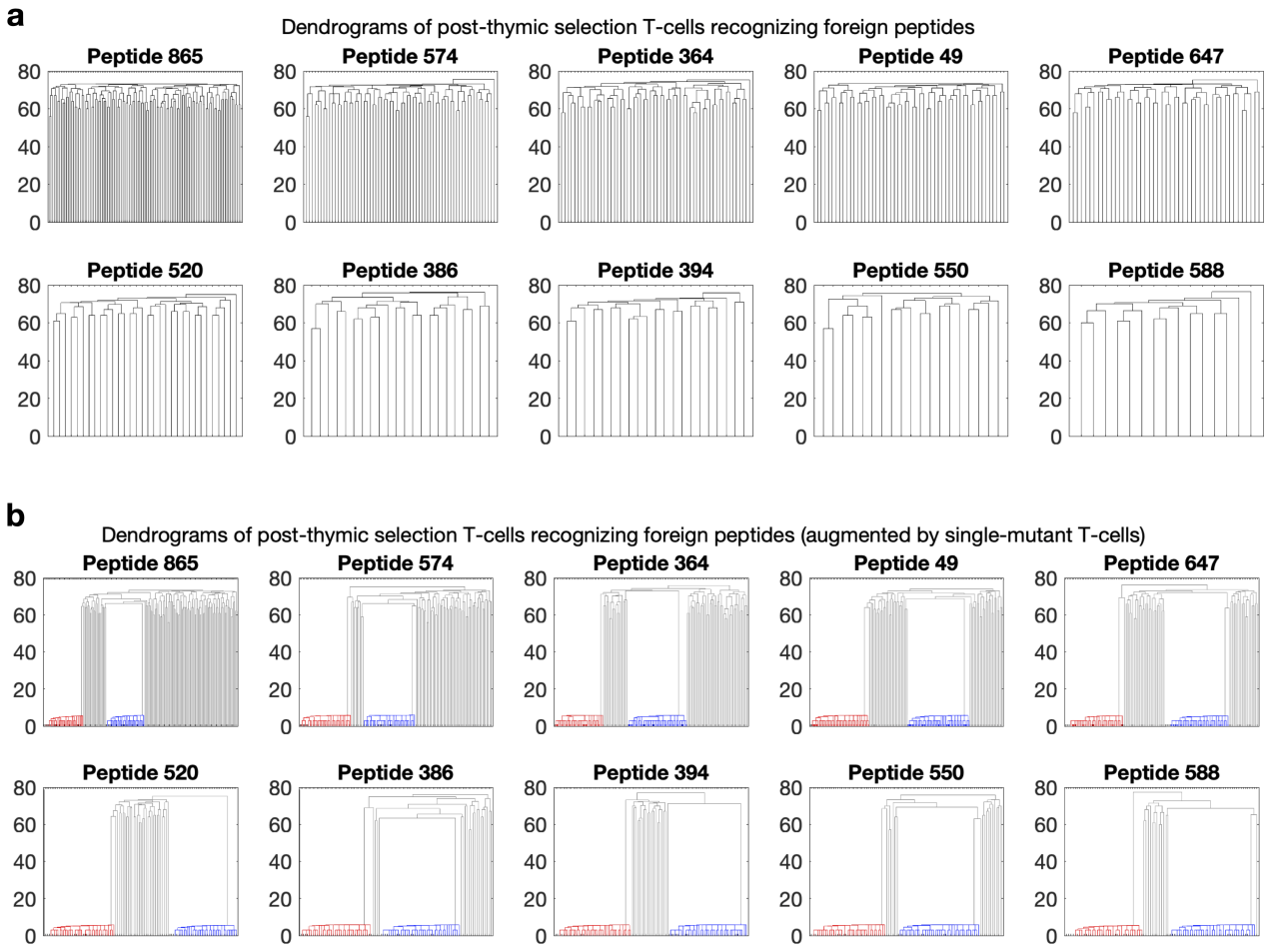
Supplementary Figure 9: Additional leave-one-out test of RACER's predictive transferability over TCRs, using data from [7]. One template structure, intentionally selected to have Vα and Vβ genes that are distinct from the target structure (where available), is used to build the target structure; this is accomplished by replacing the CDR3 loops with target sequences based on trivial alignment (case III of Fig. 1). Asterisks mark examples where the template structure shares the same peptide sequence as the target structure. For all remaining cases target amino acid sequences are explicitly compared to the template sequence, with red (resp. cyan) positions having different (resp. identical) amino acid entries. Double-ended arrows indicate those examples where the template and the target have also been switched for another prediction test. RACER successfully predicted lower binding energies of the strong binders (blue) relative to the weak binders (brown). The TCRs were grouped by their associated MHC allele, with their Vα and Vβ genes, as well as the template structures labeled at the bottom. RACER-predicted Z-scores are listed at the top of each case. RACER was able to recognize the strong peptide (Z-score > 1) for 19 out of the 26 challenging tests, and maintains predictive accuracy when only restricted to test cases which include distinct peptides from the training step (accuracy 67%.



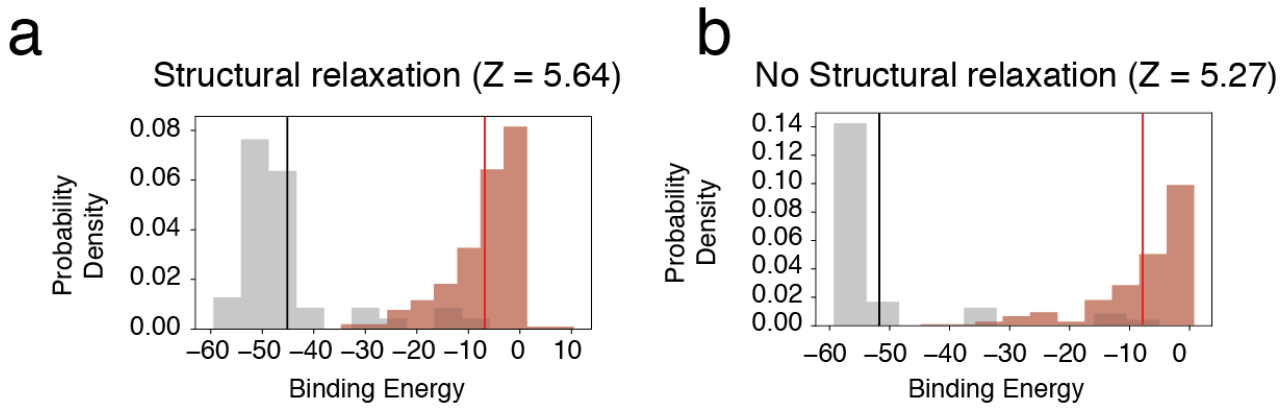
Supplementary Figure 10: The performance of ERGO [8] for differentiation of the strong and weak binders of three TCRs. The best performance of ERGO, using Autoencoder trained on the database VDJdb, can recognize the strong binders of TCR 5CC7, with a recognition Z score of 2.85. However, ERGO cannot fully differentiate the strong binders of TCR 2B4 and 226.



Supplementary Figure 11: Comparison of RACER recognition characteristics to previous models [9]. RACER-derived estimates of post-thymic selection T-cell repertoire recognition rates reveal similarity in the ability to recognize foreign and point-mutated self antigen, in agreement with previous models. **a,** Participation of self-peptides in the deletion of reactive TCRs is quantified by plotting the total number of unique TCRs recognized as a function of each self-peptide rank-ordered based on selection potency. **b,** The probability of post-selection individual TCR recognition of foreign and point-mutated self-peptide as a function of the percentage of surviving TCRs following negative selection; positive standard deviations are given for estimates obtained in the RACER model (in all plots purple represents outputs of the RACER model, red and blue correspond to PIRA and RICE models from [9], respectively) **c,** The product of thymic selection survival probability and recognition probability of random and point-mutated self-peptides as a function of T-cells survival probability in the RACER model.



Supplementary Figure 12: Quantification of the foreign epitope-specific T-cell repertoire. Each of the 10^4 T-cells were filtered based on their ability to recognize foreign epitopes in the RACER simulation, performed for the pre-thymic selection and post-selection T-cell repertoire. Each foreign epitope was then sorted by the number of total post-selection T-cells that recognized that antigen. **a,** Dendrograms of the identified post-selection recognizing T-cells for the top 10 recognized peptides are constructed using hierarchical clustering using an averaged hamming distance on the primary CDR3 sequence of each T-cell in a similar manner as in [11]. **b,** For each peptide, two dissimilar TCRs were selected randomly from the left and right side of the highest dendrogram clade. Each of these TCR underwent mutagenesis by point-mutating a single entry in the CDR3 region 50 times for a total of 100 mutated (closely-related) TCRs (blue and red clusters). These TCRs were then subject to the same thymic selection and foreign peptide recognition steps as previously, and dendrograms of these and the original TCRs were constructed (the following peptide sequences were identified by RACER as the most immunogenic: 865=ADWINQGSDWWKG, 574=ADLIALLLMWWKG, 364=ADAIAEAANCSKG, 49=ADEINKHEKWWKG, 647=ADMIDSKSTSAKG, 520=ADSIAHCGKFSKG, 386=ADWITHNWALWKG, 394=ADCIAYPKRDAKG, 550=ADYINACKSDAKG, 588=ADYINPTWAHAKG).



Supplementary Figure 13: RACER ensures comparable accuracy with or without structural relaxation after changing peptide sequences. The binding energies of experimentally determined strong and weak binders as predicted by RACER with **a,** and without **b,** structural relaxation after switching the peptide sequences. The coarse-grained nature of RACER significantly reduces the chance for steric clashes to occur after changing peptide residues, resulting in comparable modeling performance.

References

[1] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," Bioinformatics, vol. 22, pp. 1658–1659, July 2006.

[2] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," Bioinformatics, vol. 28, pp. 3150–3152, Dec. 2012.

[3] M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Özkan, M. M. Davis, K. W. Wucherpfennig, and K. C. Garcia, "Deconstructing the Peptide-MHC Specificity of T Cell Recognition," Cell, vol. 157, pp. 1073–1087, May 2014.

[4] 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, vol. 116, pp. 8494–8503, July 2012.

[5] N. P. Schafer, B. L. Kim, W. Zheng, and P. G. Wolynes, "Learning To Fold Proteins Using Energy Landscape Theory," Isr. J. Chem., vol. 54, pp. 1311–1337, Aug. 2014.

[6] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation," Macromolecules, vol. 18, pp. 534–552, May 1985.

[7] 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., vol. 94, pp. 91–97, Feb. 2018.

[8] I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, and Y. Louzoun, "Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs," Front. Immunol., vol. 11, Aug. 2020.

[9] 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. U. S. A., vol. 114, no. 38, pp. E7875–E7881, 2017.

[10] R. J. De Boer and A. S. Perelson, "How diverse should the immune system be?," Proc. Biol. Sci., vol. 252, pp. 171–175, June 1993.

[11] P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, and P. G. Thomas, "Quantifiable predictive features define epitope-specific T cell receptor repertoires," Nature, vol. 547, no. 7661, pp. 89–93, 2017.

[12] F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, and P. G. Wolynes, "Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection," Proceedings of the National Academy of Sciences, vol. 111, pp. 12408–12413, Aug. 2014.

[13] Y. Elhanati, A. Murugan, C. G. Callan, T. Mora, and A. M. Walczak, "Quantifying selection in immune receptor repertoires," Proc. Natl. Acad. Sci. U. S. A., vol. 111, pp. 9875–9880, July 2014.

[14] P. G. Thomas and J. C. Crawford, "Selected before selection: A case for inherent antigen Bias in the T-cell receptor repertoire," Curr. Opin. Syst. Biol., vol. 18, pp. 36–43, 2019.

[15] A. Madi, E. Shifrut, S. Reich-Zeliger, H. Gal, K. Best, W. Ndifon, B. Chain, I. R. Cohen, and N. Friedman, "T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity," Genome Res., vol. 24, no. 10, pp. 1603–1612, 2014.

[16] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters, "The immune epitope database (IEDB) 3.0," Nucleic Acids Res., vol. 43, pp. D405–412, Jan. 2015.