Predicting TCR-peptide recognition based on residue-level pairwise statistical potential

Vadim K. Karnaukhov[1,2,*], Dmitriy S. Shcherbinin[2,3,*], Anton O. Chugunov[2,4,5], Dmitriy M. Chudakov[1,2,3], Roman G. Efremov[2,4,5], Ivan V. Zvyagin[2,3], Mikhail Shugay[2,3,#]

[1] *Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia*
[2] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia*
[3] *Institute of Translational Medicine, Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Moscow, Russia*
[4] *Higher School of Economics, Moscow, Russia*
[5] *Research Center for Molecular Mechanisms of Aging and Age-related Diseases, Moscow Institute of Physics and Technology, Dolgoprudny, Russia*

* Equal contribution
# Correspondence: mikhail.shugay@gmail.com

Abstract

T cell receptor (TCR) recognition of foreign peptides presented by major histocompatibility complex (MHC) proteins is a crucial step in triggering the adaptive immune response. Prediction of TCR:peptide recognition is important for many clinically relevant problems: prediction of cross-reactivity of TCRs used in adoptive T-cell-based therapies, identification of targets for antigen-specific therapies of autoimmune disorders and cancer, vaccine design. In this work, we describe a novel computational method that ranks TCR-peptide pairs based on statistical estimation of binding energy to select the most likely cognate epitopes. Our method starts from a structure of a peptide-MHC complex with the TCR of interest (either experimentally derived or based on a homology model) and a set of candidate epitopes and generates a ranked list of these peptides with predicted recognition probabilities. In our method interaction energy between TCR and peptide residues is estimated using statistical potential TCRen which we derived via statistical analysis of amino acid contact preferences in TCR-peptide interfaces from crystal structures of TCR-peptide-MHC complexes. We demonstrate excellent performance of our method for two tasks related to TCR-peptide recognition: 1) discrimination between cognate epitopes and unrelated peptides; 2) discrimination between real and mocked TCR-peptide-MHC complexes. Comparison of TCRen with potentials reflecting the rules of general protein-protein interaction and protein folding reveals the distinctive features of TCR-peptide interactions, such as intrinsic asymmetry of the interface, a complex interplay between different physico-chemical properties of contacting residues and lower impact of hydrophobic interactions. Finally, we demonstrate the use of our method for identification of cancer neo-epitopes recognized by tumor-infiltrating lymphocytes.

Introduction

T cells are the key actors of cell-mediated adaptive immune response able to distinguish between self and non-self antigens presented on the cell surface. Specificity of a T cell is determined by its T-cell receptor (TCR), which recognises peptide antigens presented by major histocompatibility complex (MHC) proteins. MHC class I proteins are expressed in almost all cell types and specialized on presentation of peptides from degraded cytoplasmic proteins for inspection of intracellular status by CD8+ T cells; MHC class II proteins are expressed mainly on specialized antigen-presenting cells and present peptides from endocytosed or phagocytosed extracellular proteins for CD4+ T cells. The majority of T cells have αβTCR, the heterodimer composed of α and β chains, both contacting antigen (peptide-MHC complex) mainly by three loops called complementarity-determining regions (CDRs). CDR1 and CDR2 are germline-encoded, whereas CDR3 is formed during T cell maturation via V(D)J-recombination process by random selection of variable (V), diverse (D), and joining (J) gene segments and random nucleotide deletions and insertions on the segment junctions. As a result of the V(D)J-recombination, repertoire of TCRs in each individual is very diverse (more than 10^8 different T cell clones with unique TCRs[1,2]) which enables recognition of a huge number of potential antigenic peptides and provides defence from different pathogens and cancer.

To date, more than 200 crystal structures of TCR-peptide-MHC complexes are deposited in Protein Data Bank (PDB)[3]. Most of these structures share some common features: diagonal binding mode with α-chain contacting with N-terminal part of peptide and β-chain — with C-terminal part of peptide; CDR3s mainly interact with peptide and CDR1s and CDR2s — with MHC.

Prediction of TCR specificity (that is which peptides are recognised by the given TCR) is critical for many clinically relevant problems: identification of targets for antigen-specific therapies of autoimmune disorders and cancer, prediction of cross-reactivity in TCR-based adoptive T cell therapies, vaccine design, TCR repertoire-based diagnostics. In recent years several sequence-based approaches for prediction of TCR specificity was proposed[4,5], utilizing the idea that TCRs which recognise the same antigen have similar sequences. Evidently, these approaches are dependent on experimental data about antigen-specific TCRs and thus can hardly be applied to TCRs with low sequence similarity to TCRs with characterised antigen specificity, as well as for epitopes without identified specific TCRs. Structure-based methods may help to solve these limitations. Approaches for prediction of structure of TCR-peptide-MHC complex have been proposed on the basis of homology modeling[6] or flexible docking[7]. Also in some studies the possibility of structure-based prediction of TCR-peptide interactions was investigated[8,9]. However, limited success was achieved in these works, that may result from using general-class energy functions for protein modeling (such as Rosetta[10] and ZDOCK[11]) rather than energy potentials specialized for TCR-peptide-MHC complex modeling. The TCR-peptide interface significantly differs from general protein-protein interfaces, that comes from biological functions of TCRs which require that each TCR should be specific and cross-reactive at the same time[12]. Typical values of TCR affinity to their cognate peptide-MHC

ligands lie within the range of ~1-100 μM that is several orders of magnitude higher than typical values of affinity in protein-protein complexes[13]. Reaching of specificity in such a low affinity range requires interaction mediated by a number of medium-strength contacts rather than by salt bridges, hydrogen bonds and hydrophobic interactions[14–16]. These features make description of TCR-peptide recognition using conventional approaches (such as MMPBSA or universal scoring functions such as Rosetta) problematic and inspire development of methods specifically focused on prediction of TCR-peptide interactions[17].

In this work, we describe a novel computational method that ranks TCR-peptide pairs based on statistical estimation of binding energy to select the most likely cognate epitopes. Our approach employs calculation of a statistical potential (called TCRen, from “TCR energy”) that describes the interaction energy between TCR and peptide residues. We derived TCRen potential via statistical analysis of amino acid contact preferences in TCR-peptide interfaces from a non-redundant set of crystal structures of TCR-peptide-MHC complexes (**Figure 1**, **upper panel**). Our method starts from a structure of the peptide-MHC complex with the TCR of interest—either experimentally derived or based on a homology model—then extracts a TCR-peptide contact map, estimates the TCRen score for all candidate peptides by convolving the contact map with TCRen pairwise energies, and finally, generates a ranked list of peptides based on this score (**Figure 1**, **lower panel**). Such residue-level scoring is optimal for practical applications in which hundreds or thousands of TCR-peptide pairs are being screened, as it makes the method more robust against inaccuracies in structures and does not require the computationally expensive refinement procedures that are necessary when all-atom scoring functions are used.

Importantly, in contrast to the conventional protein-protein interfaces, the interface of TCR-peptide interaction is essentially asymmetric: TCR repertoire is shaped by V(D)J-recombination and selection processes for minimization of risk of potential self-reactivity, while immunopeptidome is shaped by MHC presentation machinery[19]. Consequently, TCRen that is derived from the natural TCR-peptide-MHC complexes inherits the asymmetric form.

In the first part of the work we describe the derivation of TCRen. Further we assess TCRen ability to discriminate between cognate TCR epitopes and unrelated peptides and between real and mock TCR-peptide-MHC complexes, and compare its performance with other residue-level pairwise potentials, atom-level Rosetta scoring function and MMPBSA approach. Then we perform an analysis of TCRen correlation with physicochemical properties of contacting residues that reveals specific features of TCR-peptide recognition compared to other protein-protein interactions. Finally, we propose a pipeline utilizing TCRen for prediction of cancer neo-epitopes recognised by tumor-infiltrating lymphocytes.

Results

Calculation of TCRen potential

Derivation of TCRen potential is schematically represented in **Figure 1, upper panel**. For calculation of TCRen we used a non-redundant set of 144 TCR-peptide-MHC complexes (**Supplementary Table 1;** see **Supplementary Methods** for details) selected from all available TCR-peptide-MHC crystal structures from PDB (**Supplementary Table 2**). The used non-redundant database includes human and mouse TCR-peptide-MHC complexes with both MHC-I and MHC-II. Although the majority of TCR contacts with peptide are formed by CDR3α/β loops, a significant number of contacts are formed by CDR1α/β and CDR2α/β loops (**Supplementary Figure 1**), so for derivation of TCRen we used all the contacts between TCRs and peptides. For all structures in the non-redundant set we extracted pairwise amino acid contacts between TCR and peptide and summarised the number of occurences of contacts between each of 400 possible amino acid pairs (with explicit assignment which of the residues corresponds to peptide or TCR) across all the structures. Based on these summary contact counts we calculated amino acid contact preferences in TCR:peptide interface, that reflect which residue pairs are found in contact more or less frequently than it is expected from random contact formation. TCRen values for energy of interaction between TCR and peptide residues were calculated from these contact preferences in the assumption that the more energetically favorable residue pair is, the more frequently (with normalization to baseline frequency of contacting amino acids) it is found in real structures.

We considered 3 reference states, which represent different ways to estimate the baseline number of contacts between the given amino acids in the absence of any pairing preferences: “contact-fraction”, which considers only residues which actually form contacts[20], “mole-fraction”, which considers all residues in peptides and central parts of CDR3 regions, and “normalized mole-fraction”, which accounts for different propensity for forming contacts for residues in different positions of peptides and CDR3s (see details in the **Methods** section). As was discussed in the **Introduction**, TCR:peptide interface is asymmetric that results in asymmetric form of TCRen. To investigate the importance of introducing asymmetry we also considered the symmetric form of TCRen, derived from the matrix of TCR-peptide contacts computed regardless of which of the contacting residues come from TCR or peptide. So, in this work we compared the performance of 6 versions of TCRen: symmetric and asymmetric ones with 3 reference states for each. These potentials will further be referred to in the form TCRen.X.Y, when X is either “s” for symmetric or “a” for asymmetric forms, Y is either “cf” for contact-fraction or “mf” for mole-fraction or “nmf” for normalized mole-fraction reference states (notations are summarised in **Supplementary Figure 2A**).

For comparison we considered pairwise potentials from AAindex database[21]. Since most of these potentials are highly correlated with each other (**Supplementary Figure 3A**), we selected a representative set of 8 potentials derived from amino acid contact preferences either in general protein-protein interfaces ("MOOG990101", "KESO980101") or in protein folding

("MIYS990106", "BRYS930101", "MICC010101", "THOP960101", "TOBD000102", "VENM980101").

Pairwise correlation coefficients (r) between entries of different versions of TCRen and representative potentials from AAindex are presented as a correlation matrix in **Supplementary Figure 3B**. Note that $r \approx 0.6$ for symmetric and asymmetric potentials that use the same reference state, $r \approx 0.9$ for potentials of the same type (symmetric or asymmetric) with different reference states, and no statistically significant correlation exists between any of AAindex potentials and any of TCRen versions.

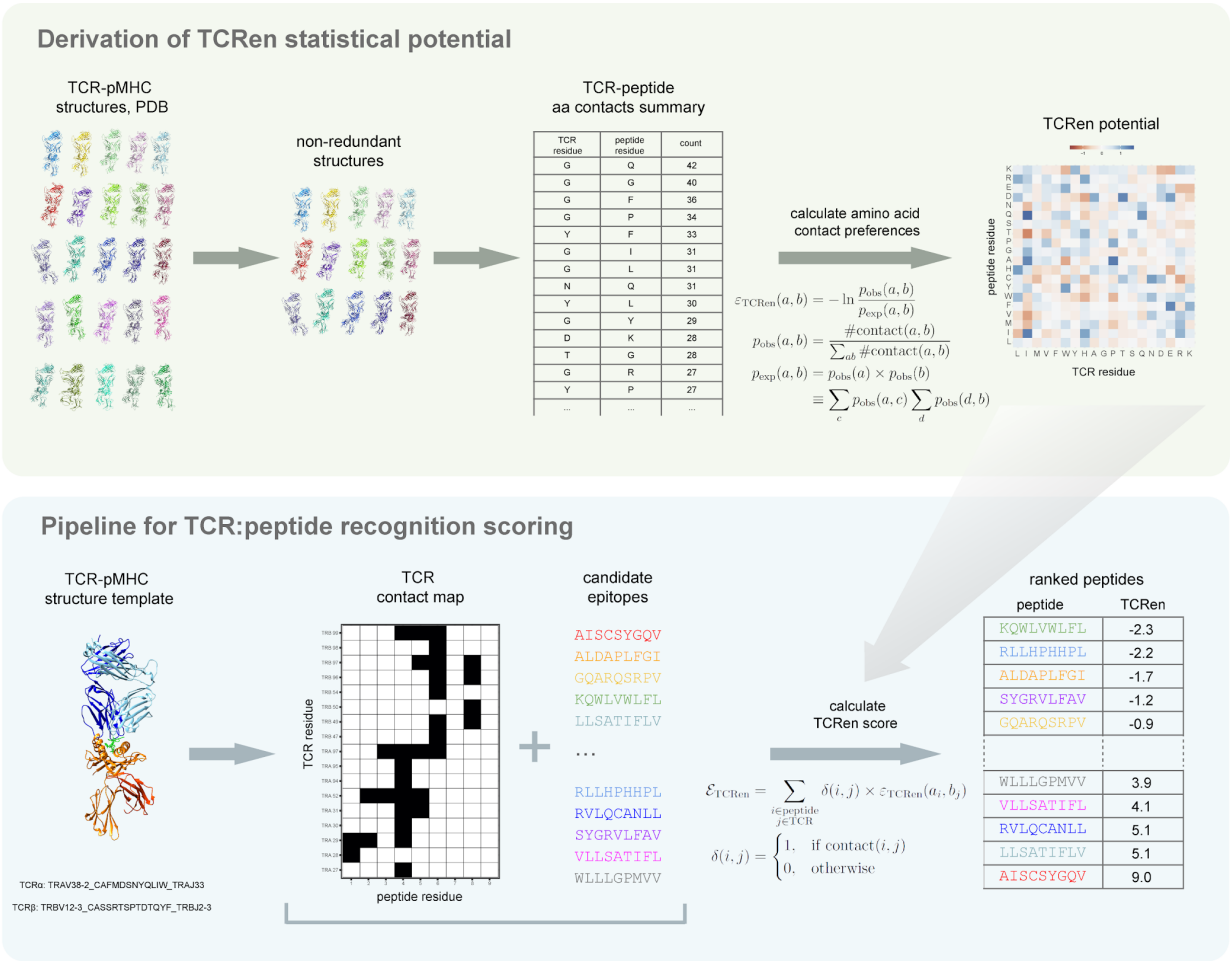


Figure 1. TCRen-based method for prediction of TCR-peptide recognition. The upper panel describes statistical analysis of contact preferences in TCR-peptide interactions. A set of non-redundant (in terms of CDR3α, CDR3β, and peptide sequences) TCR-peptide-MHC complexes (n = 144, which included structures with MHC class I and II from mouse and human) is selected from all available TCR-peptide-MHC crystal structures (n = 239). Contacts between TCR and peptide residues are identified and summarized (total number of contacts = 3,106). A 20 × 20 matrix of contact odds is then derived by normalizing contact frequencies. Rows and columns in this matrix respectively correspond to amino acids from the peptide and TCR. Cells are colored according to TCRen values. The lower panel outlines the bioinformatic pipeline, starting with a structure of the TCR-peptide-MHC complex—either experimentally derived or from homology modeling—followed by TCR-peptide contact map extraction and convolution with TCRen potential, leading to the final TCRen score. Our method outputs a ranked list of candidate peptides according to their estimated TCR binding energy.

TCRen for identification of cognate TCR epitope

First we assessed the performance of TCRen in the task of finding the cognate epitope of the given TCR (that is the peptide contained in the crystal structure of the corresponding TCR-peptide-MHC complex) among a set of unrelated peptides. For each structure in the non-redundant set of TCR-peptide-MHC complexes we extracted a TCR-peptide contact map and calculated TCRen score for the cognate peptide. Then we generated 1000 unrelated peptides (mismatched from the cognate epitope in all non-anchor positions) and calculated TCRen score for each of them (see **Methods** and **Figure 1, lower panel** for details). As a metric of performance we used “cognate epitope rank” - the fraction of unrelated peptides which have lower values of energy of interaction with TCR than the cognate epitope. The lower the cognate epitope rank is, the more accurate distinction between the cognate epitope and unrelated peptides is achived. For a random classifier percentage cognate epitope rank is close to 50%.

Performance of different versions of TCRen is presented in **Supplementary Figure 2B**, where distribution of cognate epitope rank for all TCR-peptide-MHC structures is plotted for each potential. The best results are obtained for TCRen.a.cf: median value of cognate epitope rankis 3%. Notably, for all reference states asymmetric potentials had higher performance than corresponding symmetric versions, likely reflecting the asymmetric nature of TCR-pMHC interaction.

On the same benchmark we compared the performance of TCRen and other pairwise potentials (**Figure 2A**). For potentials from AAindex median value of cognate epitope rank is in the range 32-55% which is significantly worse than for TCRen. TCRen.a.cf potential also has the best performance when comparing all complexes with cognate epitopes against all complexes with unrelated epitopes without segregation based on PDB template with $ROC\ AUC = 0.93$, both among different TCRen versions (**Supplementary Figure 2C**) and in comparison with other pairwise potentials (**Figure 2B**). Results are similar for structures including MHC-I and MHC-II as well as for human and mouse complexes (**Supplementary Figure 4**) that indicate common biophysical features of TCR-peptide recognition among MHC classes and biological species.

Since TCRen.a.cf has the best performance in this benchmark, in further sections we mainly considered this version and refer to it as TCRen.

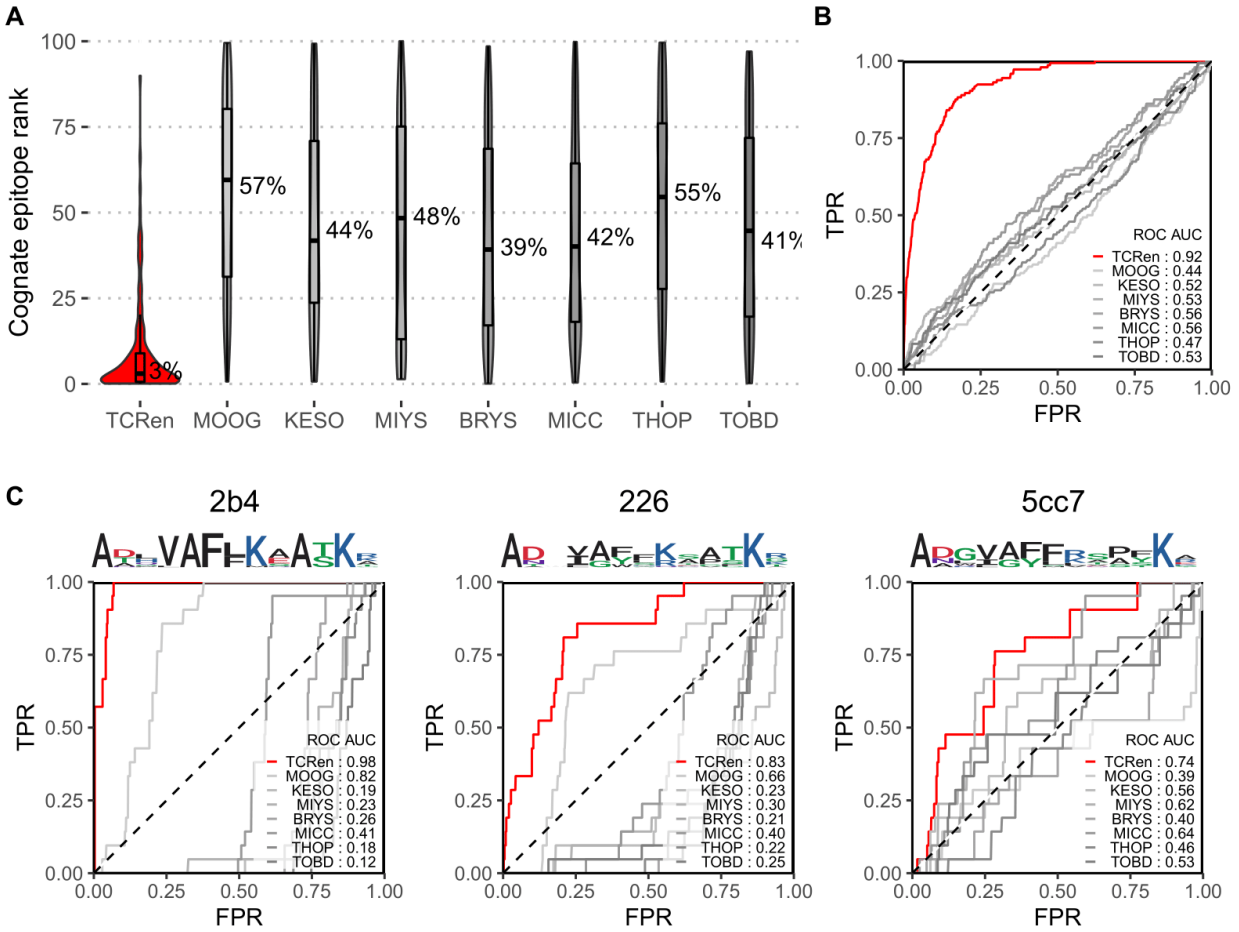


Figure 2. TCRen performance in distinguishing cognate TCR epitopes from unrelated peptides. A. Distributions of percentile ranks of peptide-TCR interaction energies for cognate epitopes among a set of unrelated peptides, as calculated for all structures in a non-redundant TCR-peptide-MHC dataset using TCRen (TCRen.a.cf; red) or pairwise potentials from the AAindex database (grey). The latter were derived from amino acid contact preferences in general protein-protein interfaces (MOOG990101, KESO980101) or in protein cores (MIYS990106, BRYS930101, MICC010101, THOP960101, TOBD000102). See **Methods** for details of the benchmarking experiment. **B.** ROC curves for distinguishing all complexes with cognate epitopes versus all complexes with unrelated peptides using TCRen (red) or potentials from AAindex (grey). **C.** Performance of TCRen (red) and pairwise potentials from AAindex database (grey) in discriminating a set of peptides recognized by TCRs 2B4, 226, and 5cc7, which were previously identified from yeast display screening, and unrelated peptides. Top shows sequence logos for the yeast display hits, bottom shows ROC curves for 2B4, 226, and 5cc7. TPR = true positive rate, FPR = false positive rate.

Assessing TCRen performance on an independent validation set of recently deposited PDB structures

Further we performed validation of TCRen performance on a holdout set of structures recently deposited in PDB. For this benchmark we derived TCRen based on 117 TCR-peptide-MHC structures released before December 31, 2018 and applied it for identification of cognate epitopes among unrelated peptides for the set of 27 non-redundant structures released after

January 1, 2019 and not homologous to the ones in the derivation set (**Supplementary Figure 5AB**). For majority of these cases TCRen had high performance, e.g. for structures containing cancer neoepitopes (derived from KRAS (6ULR), HHAT (6P64) and p53 (6VQO)), tumor-associated antigens (from cancer testis antigen NY-ESO-1 (6RP9), breast cancer antigen NY-BR-1 (6R2L), melanoma antigen gp100 (6VM7)), autoantigen from fibrinogen β (6V13 and 6V0Y), viral peptide from EBV (6VMX) and others (**Supplementary Figure 5A**).

Notably, among these newly deposited structures there were three structures with TCRs recognizing peptides from SARS-CoV2 (7N1E, 7N1E, 7N6E). For two of them, 7N1E (with "RLQ" epitope) and 7N1F (with "YLQ" epitope) TCRen can accurately distinguish cognate epitopes from unrelated peptides (**Supplementary Figure 5B**). For the 7N6E case (with "YLQ" epitope) TCRen performance is worse that could be explained by a high negative contribution to TCR-peptide interaction energy from contact between glutamine (Q) in TCR CDR1α and Q in the epitope (**Supplementary Table 3**). Q-Q is considered by TCRen as highly unfavorable, as it is a very rare contact that occurred only once in the structures used for TCRen derivation, while Q occurs relatively frequently both in peptides and in peptide-contacting positions of TCR in these structures (see equations in *"Derivation of TCRen"* section in **Methods**). Inspection of other cases for which TCRen fail to identify cognate epitope also revealed that all these structures contain contacts that are extremely rare and are not found or occurred only once in the TCRen derivation set and these contacts have the highest unfavorable impact to the interaction energy (**Supplementary Table 3**). However, if in 7N6E case we consider only contacts between peptide and CDR3α/β (which are regarded to be the most important for TCR recognition and comprise the majority of TCR-peptide contacts[16]), TCRen performance is significantly improved (**Supplementary Figure 5B**). This case raises the question about the interplay between CDR3-mediated and germline-encoded specificity (e.g. governed by TRAV12-1[22]), which would be peculiar to investigate in future studies.

Peptide screening for TCRs with multiple known ligands

We further tested the ability of TCRen.a.cf to distinguish all epitopes for TCRs, for which dozens of different recognized peptides were identified using yeast display screening. In this benchmark we considered 3 different TCRs (2B4, 226, 5cc7) from the work of Birnbaum *et al.*[23] and used the TCRen version, derived without inclusion of structures containing 2B4, 226, 5cc7 and TCRs with similar sequences. ROC-curves for distinction between recognized and unrelated peptides are shown in **Figure 2C**. For all the 3 cases TCRen demonstrated excellent performance with ROC AUC 0.98, 0.84 and 0.74 for 2B4, 226, 5cc7, respectively.

TCRen for discrimination between real and mock TCR-peptide-MHC complexes

Next we assessed performance of potentials in distinguishing real TCR-peptide-MHC complexes from the ones obtained by shuffling of TCRs and pMHCs in existing crystal

structures (see **Methods** section for details). Among different versions of TCRen in this benchmark the best results were demonstrated by TCRen.a.cf (**Figure 3A**, **Supplementary Figure 2D**), which also had the best performance in distinction between cognate epitope and unrelated peptides as described above (**Supplementary Figure 2C**). Pairwise potentials from AAindex database had performance in this benchmark close to that of a random classifier (**Figure 3A**). Accuracy of discrimination between real and shuffled complexes increases when using total energy of peptide interaction with TCR (both with α- and β-chains) rather than energy of interaction with separate chains (**Figure 3B**). There is no correlation between calculated peptide-TCRα and peptide-TCRβ energies of interaction (**Figure 3C**) so impact of TCR α- and β-chains to interaction may be considered as complementary.

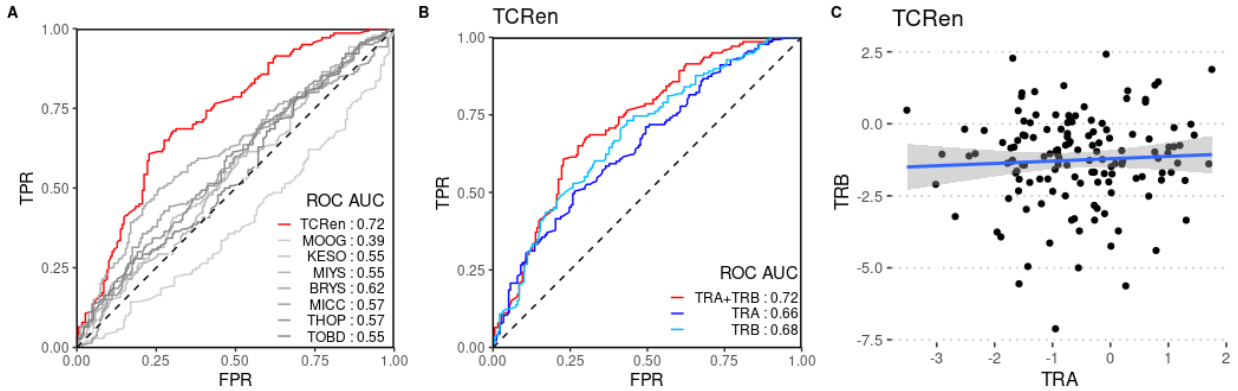


Figure 3. TCRen performance in TCR shuffling benchmark. A. ROC-curves for discrimination between real and shuffled TCR-peptide-MHC complexes using TCRen.a.cf (TCRen; red) or pairwise potentials from the AAindex database (grey). **B**. ROC-curves for discrimination between real and shuffled TCR-peptide-MHC complexes using TCRen when considering only TCRα (TRA) or TCRβ (TRB) or both TCRα and TCRβ (TRA+TRB) chains. **C.** Absence of correlation between energies of TCRα-peptide (X axis) and TCRβ-peptide (Y axis) interactions calculated using TCRen potential.

Comparison of TCRen performance with all-atom Rosetta scoring function and MMPBSA method

In the previous sections we compared TCRen performance with other residue-level pairwise contact potentials. However, more fine-grained atom-level scoring functions (such as Rosetta[10]) and approaches based on physical calculations (such as MMPBSA[24]) exist and are widely used for assessment of protein-protein interactions.

First we compared the robustness of TCRen and Rosetta calculations. For this purpose we examined correlation betweenof TCR-peptide interaction scores calculated for one static structure (obtained using fixbb protocol) and averaged across 1000 frames of 100-ns molecular dynamics (MD) simulations of corresponding TCR-peptide-MHC complexes (**Supplementary Figure 6**). For TCRen excellent correlation between these values is observed that highlights robustness of residue-level TCRen potential to inaccuracies in modeled structures that allows its use without computationally expensive MD simulations. For Rosetta dG values calculated from

MD trajectories were significantly lower than the ones of static structures in all the cases and no correlation between them was observed, that is explained by fitting of TCR and peptide interfaces during MD simulation. Thus, calculation of Rosetta dG for structures obtained using fixbb protocol is unreliable, and using Rosetta potential for scoring of peptide-TCR interactions requires generation of an ensemble of TCR-peptide-MHC complexes, e.g. using MD simulations.

In view of the lack of robustness for Rosetta, for comparison of TCRen and Rosetta performance we carried out MD simulations and averaged TCR-peptide energy of interaction across the trajectory. We also used the obtained trajectories for MMPBSA calculations. We assessed performance of TCRen, Rosetta and MMPBSA in identification of peptides recognized by DMF5 TCR. For this TCR a number of recognised epitopes was identified by yeast display[18] and these peptides are clustered to two groups with different motifs: one with a hydrophobic core (GIGI in positions 4-7) and the second with a charged core (DRG in positions 4-6). We performed MD simulations of DMF5 TCR-peptide-MHC complexes with different peptides from 3 groups: experimentally confirmed epitopes with either “GIGI” or “DRG” motif and unrelated peptides that share neither of these motifs. For each peptide we calculated trajectory-averaged energy of interaction with TCR using either TCRen or Rosetta or MMPBSA. Only TCRen reported lower values of energy for both “GIGI” and “DRG” groups of peptides compared to unrelated peptides. Rosetta scoring function assigned lower energy values for “DRG” but not for “GIGI” peptides, while MMPBSA approach identified no statistically significant differences between unrelated and recognised peptides from both groups (**Figure 4**).

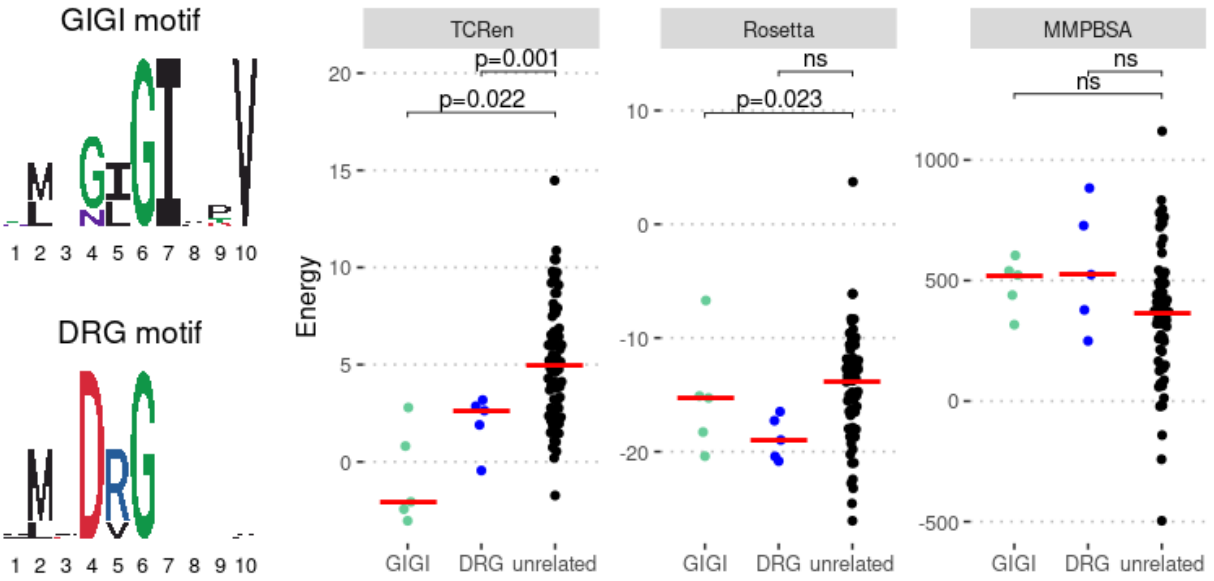


Figure 4. Performance of TCRen, Rosetta, and the MMPBSA approach for identifying peptides recognized by the DMF5 TCR. At left are the sequence logos for peptides with GIGI and DRG motifs recognized by DMF5 TCR. Plots show comparison of energies of interaction of DMF5 TCR with GIGI-containing (n = 5), DRG-containing (n = 5), and unrelated (n = 64) peptides as calculated using TCRen (TCRen.a.cf), Rosetta, and MMPBSA. Energy calculations are averaged across 1,000 frames of MD trajectories of the corresponding TCR-peptide-MHC complexes.

TCRen correlation with physicochemical properties of contacting amino acids

As TCRen was derived based on amino acid contact preferences in TCR:peptide interface, it should reflect the principles of TCR-peptide recognition. We performed the analysis of correlation between TCRen entries (**Figure 5A**) and physicochemical properties of contacting amino acids to shed light on the specific nature of TCR-peptide interactions. For comparison we used the potential derived by Miyazawa and Jernigan[25] (further referred to as MJ; **Figure 5B**), representing the rules governing protein folding, and the potential from the work of Keskin *et al.*[26] (further referred to as Keskin; **Figure 5C**), reflecting preferences in general protein-protein interactions.

To investigate regularity patterns in TCRen, MJ and Keskin potentials we first grouped non-polar, polar and charged amino acids and compared values for contacts between two non-polar residues, one non-polar and one polar residue, one non-polar and one charged residue, etc. (**Figure 5D**). MJ demonstrates clear clustering of non-polar, polar and charged residues (**Figure 5B**) with the best values for contacts of two non-polar residues and the worst values for contacts of two charged residues (**Figure 5D**). Keskin potential follows in general the same patterns as MJ but with more favorable energy values for contacts of oppositely charged residues (**Figure 5E**) and contacts of glutamine and arginine with other amino acids (**Figure 5C**). For TCRen some clustering of amino acids with similar properties (e.g. positively charged R and K, polar T, N and S, hydrophobic W, V, M, F) is observed (**Supplementary Figure 7**), however, generally it doesn't have a simple regular pattern analogous to that of MJ and Keskin.

Further we examined correlation between TCRen, MJ and Keskin values and physicochemical properties of contacting residues represented as 5 Atchley factors (AF) [27], which were derived from a multivariate analysis with dimension reduction of 494 properties of amino acids (**Figure 5G**). MJ and Keskin entries to a greater extent correlate with polarity/hydrophobicity AF1 of individual amino acids. On the opposite, TCRen correlates mostly with the products of AFs of contacting residues that implies that TCRen accounts for interaction of physicochemical properties of contacting amino acids.

We also performed eigenvalue decomposition of TCRen, MJ and Keskin matrices (see **Supplementary Note 1, Supplementary Figure 8**) that also demonstrated mainly hydrophobic nature of driving forces of protein folding[28] and general protein-protein interactions and more complex rules of TCR-peptide interactions.

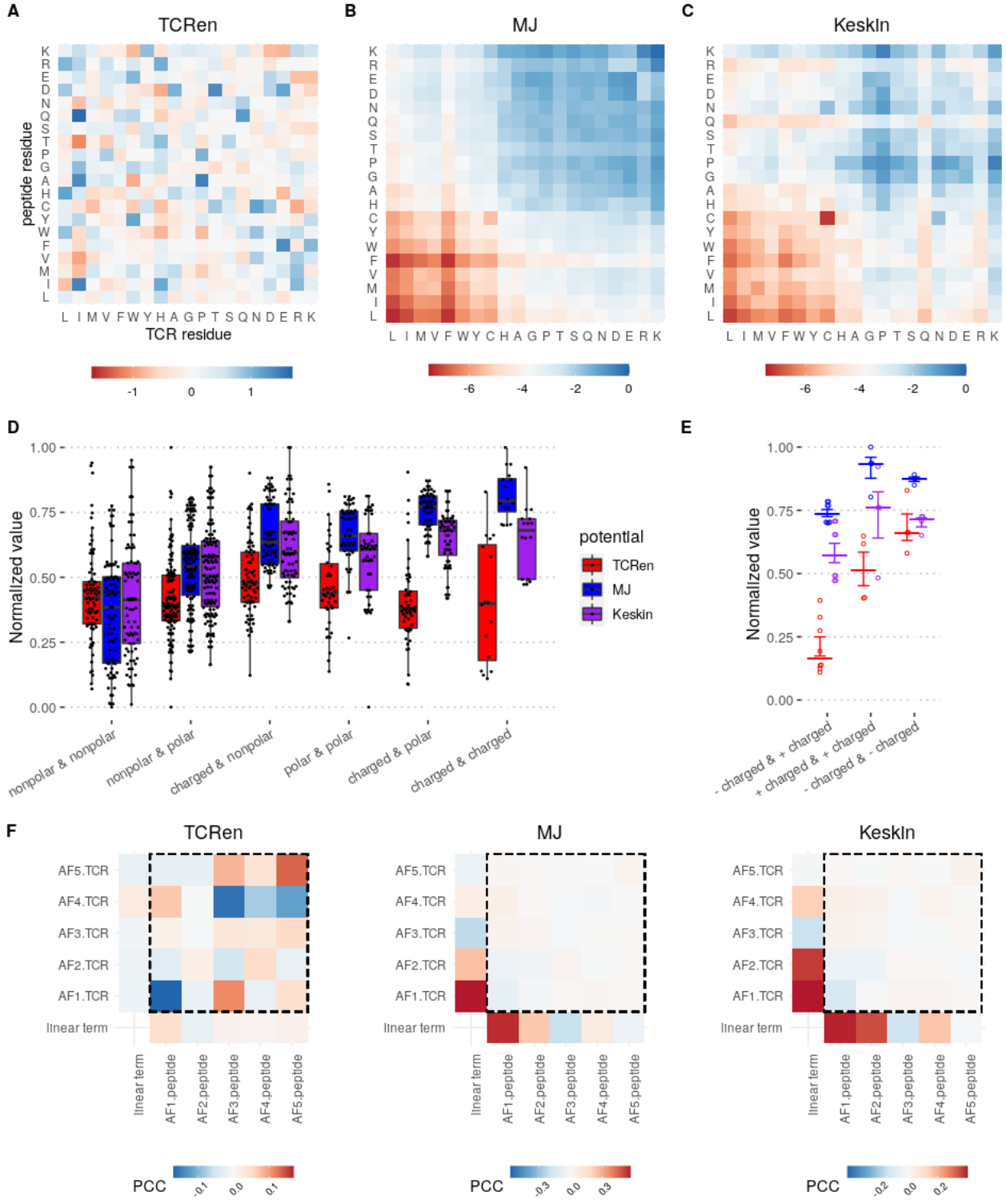

A
TCRen
peptide residue
TCR residue
B
MJ
C
KeskIn
D
Normalized value
potential
TCRen
MJ
Keskin
nonpolar & nonpolar
nonpolar & polar
charged & nonpolar
polar & polar
charged & polar
charged & charged
E
Normalized value
- charged & + charged
+ charged & + charged
- charged & - charged
F
TCRen
MJ
KeskIn
AF5.TCR
AF4.TCR
AF3.TCR
AF2.TCR
AF1.TCR
linear term
AF1.peptide
AF2.peptide
AF3.peptide
AF4.peptide
AF5.peptide
PCC

Figure 5. Comparison of TCRen, MJ and Keskin potentials. A. TCRen.a.cf (TCRen) potential. **B.** MJ potential. **C.** Keskin potential. **D-E.** Comparison of TCRen and MJ entries grouped by properties of the contacting amino acids. Groups in **D.**: “nonpolar” — leucine, isoleucine, valine, methionine, phenylalanine, tryptophan, alanine, glycine, proline; “polar” — serine, threonine, asparagine, glutamine, histidine, cysteine, tyrosine; "charged" — arginine, lysine, aspartic acid, glutamic acid). In **E.** the group of “charged & charged” is decomposed to "+ charged" (arginine, lysine) and "- charged" (aspartic acid, glutamic acid). **F.** Correlation of TCRen, MJ and Keskin entries with AFs of contacting amino acids. Suffixes “.TCR” and “.peptide” denote residues from TCR and peptide, respectively. The first column corresponds to correlation coefficients of matrix entries with AFs of TCR residues, the last row — to correlation coefficients of matrix entries with AFs of peptide residues, and all other cells (marked with dashed rectangles) — to correlation coefficients between matrix entries and products of AFs, the first of which referred to residue from TCR and the second — to the residue from peptide. MJ and Keskin entries better correlate with properties of individual amino acids, while TCRen — with products of AFs of contacting amino acids. AF1: polarity/hydrophobicity, AF2: secondary structure preferences, AF3: molecular volume, AF4: codon diversity, AF5: electrostatic charge (see also **Supplementary Table 4**).

Application of TCRen for identification of neo-epitopes recognized by tumor-infiltrating lymphocytes

Tumor-infiltrating lymphocytes (TILs) recognition of neo-epitopes (peptides from mutated proteins) presented by cancer cells is an important part of anti-tumor immunity. Prediction of neo-epitopes recognised by host TILs is critical for development of efficacious neo-epitope-based therapeutic vaccines [29,30]. Approaches based solely on neo-epitope sequences currently have very limited success [31] and accounting for information about TCR sequences of TILs, which recognise immunogenic neo-epitopes, may improve prediction accuracy.

In previous sections we demonstrated that TCRen can accurately distinguish between cognate TCR epitopes and unrelated peptides. So TCRen may be useful for identification of neo-epitopes recognized by TILs. The overall scheme of the proposed pipeline is presented in **Figure 6A**. The key stage of this approach is assessment of TCRen score for each of candidate neoepitopes based on a model of TCR-peptide-MHC complex (e.g., obtained using TCRpMHCmodels server [6]).

We performed a proof-of-concept demonstration of this approach using recently published data on several TCRs of TILs from patients with recurrent advanced epithelial ovarian cancer [32]. We focused on TCR 302TIL recognising neoepitope $HHAT_{p8F}$ (KQWLVWLFL) in HLA-A*02:06 context as only for this TCR the crystal structure of TCR-peptide-MHC complex with cognate peptide-MHC is available [33]. In general case models rather than crystal structures of TCR-peptide-MHC complexes of interest are used for prediction of TCR-peptide recognition (**Figure 6A**), thus availability of crystal structure is important for benchmarking purposes as it allows to separate between inaccuracy coming either from modeling errors or from TCRen performance.

For the same patient from which TILs with TCR 302TIL were isolated, 20 candidate neo-epitopes were predicted based on exome sequencing of tumor and normal tissues and NetMHC [34] predictions of binding to HLA-A*02:06. We performed homology modeling of the

TCR-peptide-MHC complex for TCR 302TIL using TCRpMHCmodels server[6] (corresponding crystal structure was not used as a template; see **Methods** for details). Using the resulting model we assessed TCRen score of TCR-peptide interaction for each of 20 candidate neoepitopes and ranked them according to this value. Cognate neoepitope $HHAT_{p8F}$, had the best predicted TCRen score (**Figure 6B**). This result perfectly demonstrates the perspective of TCRen for identification of neoepitopes recognized by TILs. We also performed ranking of the candidate neoepitopes using the crystal structure of 302TIL TCR-peptide-MHC complex. In this case TCRen also had an excellent performance, ranking the cognate neoepitope as Top1 (**Figure 6C**). Interestingly, the set of TCR-peptide residue contacts was not identical for the crystal structure and the homology model with ~60% of contacts shared. The fact that cognate peptide got high rank using both for the crystal structure and the homology model, may support the idea that TCR-peptide recognition is robust to small inaccuracies in TCR-peptide-MHC structure that is important in early stages of TCR-pMHC interaction.

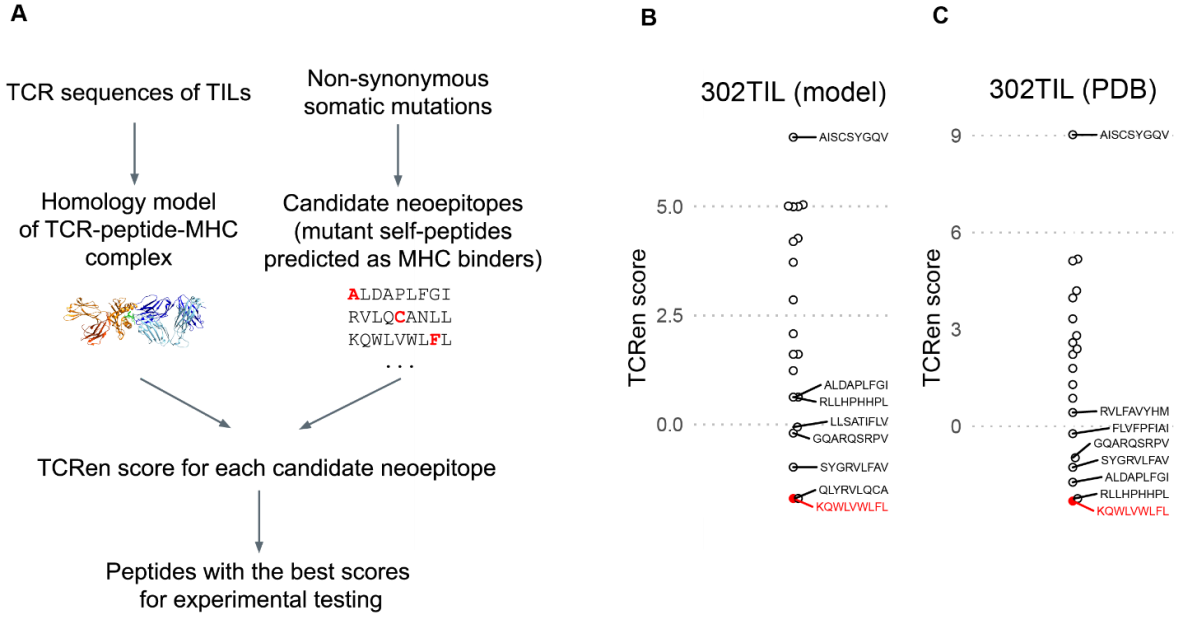


Figure 6. Application of TCRen for identification of neoepitopes recognized by tumor-infiltrating lymphocytes. A. Scheme of the TCRen-based bioinformatic pipeline for identifying neo-epitopes recognized by tumor-infiltrating lymphocytes. **B, C.** TCRen values for energy of interaction of TCR 302TIL with candidate neo-epitopes. Contact map for the 302TIL TCR was taken from either a homology model (**B**) or the actual crystal structure (PDB ID: 6UK4) (**C**). The cognate peptide recognized by TCR 302TIL is highlighted in red. Note that five out of the top seven peptides with the best TCRen score as well as the peptide with the worst score are matched when ranking based on both the homology model and the PDB structure.

Discussion

TCR recognition of peptides presented by MHC molecules on the surface of cells is a critical event in triggering and realization of adaptive immune response. Recent advances in high-throughput sequencing allowed deep and accurate profiling of TCR repertoires[35], and emulsion-based techniques allow for the massive pairing of TCR alpha and beta chain repertoires[36,37]. However, a major challenge still exists in linking sequences of TCRs and their specificities.

In this study we propose a method for prediction of TCR:peptide recognition. We provide the pipeline which requires a structure of TCR-peptide-MHC complex (e.g. obtained by homology modeling in TCRpMHCmodels server[6]) and a list of candidate peptides and outputs a ranked list of these peptides according to their binding probabilities. Given the structure of TCR-peptide-MHC complex, the pipeline first extracts all pairwise contacts between peptide and MHC and then estimates the energy of TCR-peptide interaction for all candidate peptides using a statistical potential named TCRen.

TCRen was derived based on amino acid contact preferences in TCR:peptide interface of TCR-peptide-MHC crystal structures obtained to date. These preferences may be calculated in the assumption of either asymmetry (residues of the same amino acid in TCR and in peptide side are regarded as different residue types) or symmetry of peptide:TCR interface. Asymmetric form of the potential can in part account for multi-body interactions and differences in the environment of residues in CDRs and peptides. We compared both approaches and found that asymmetric potential performs significantly better. This result is in line with the intrinsic asymmetry of TCR-peptide interface. Since TCRs with highly hydrophobic CDR3s may be strongly cross-reactive due to non-specific hydrophobic interactions, amino acid composition of TCR:peptide interfaces is characterized by depletion in hydrophobic and enrichment in polar amino acids in CDR3 regions of TCRs compared to immunopeptidome (**Supplementary Figure 9**). Better performance of asymmetric potential in our study is in concordance with the recent work of Brenke *et al.*[38] which demonstrated superior performance of asymmetric potential compared to symmetric one in the task of antibody-antigen docking and linked these results to enrichment of aromatic residues (tyrosine, phenylalanine and tryptophan) in CDR regions of antibodies.

We showed that TCRen can accurately distinguish between the cognate epitope and unrelated peptides, both for a wide set of TCRs from PDB for which either one or a set of recognized peptides is known, as well as between real and mock TCR-peptide-MHC complexes. On the same benchmarks we demonstrated that other approaches developed for more general tasks related to protein-protein interactions (including residue-level pairwise potentials from AAIndex database[21], all-atom Rosetta scoring function and MMPBSA approach) fail to predict TCR:peptide recognition. The superior performance of TCRen is explained by accounting for the special nature of TCR:peptide recognition, characterized by lower impact of hydrophobic interactions and complex interplay between different properties of contacting amino acids. Due to residue-level contact definition, TCRen is not sensitive to fine details of TCR-peptide-MHC

structure. This is highlighted by high correlation of TCRen energies calculated from one static structure and averaged across 100-ns MD simulations (for Rosetta scoring function dG for a single structure and MD-averaged didn't correlate at all).

Finally, we demonstrate that TCRen can be used for identification of neo-epitopes recognised by tumor-infiltrating lymphocytes.

The major limitation of the TCRen is the small number of available crystal structures of TCR-peptide-MHC complexes accumulated to date. That resulted in low counts for some rare pairs of amino acids (e.g. the ones including cysteine or histidine residues) and consequently high uncertainty for TCRen values for these pairs. So an investigator should carefully examine results obtained for TCR-peptide interfaces which include rare pairs of contacting residues. Increase in the number of available TCR-peptide-MHC crystal structures will provide more statistics for rare amino acid pairs. However, for majority of residue pairs the used non-redundant database has enough statistics for calculation of corresponding e_{ij} values with high precision. Validation in a holdout set of newly deposed TCR-peptide-MHC structures demonstrates that TCRen can identificate epitopes even for TCRs that share no homology to TCRs used for TCRen derivation.

The accuracy of TCRen is dependent to a large extent on correctness of structure of TCR-peptide-MHC complex used for extraction of peptide-TCR contacts. So in practical applications of TCRen for identification of epitopes of TCRs for which no crystal structures are available, the accuracy of TCR-peptide-MHC models is critical. If templates with a close sequence of CDR regions are available, homology modeling may be used, in other cases flexible docking approaches should be used[7]. We expect that TCRen potential may also facilitate improvement of TCR-peptide-MHC modeling algorithms, being incorporated in the stage of rescoring of decoys obtained by homology modeling or docking and/or in sampling procedure.

We suppose that TCRen may be used in development of novel approaches for prediction of TCR specificity and cross-reactivity with applications in antigen-specific therapy in cancer and autoimmunity and vaccine design.

Methods

Derivation of TCRen potential and calculating TCRen score

All available crystal structures of TCR-peptide-MHC complexes (n = 239, **Supplementary Table 2**) were downloaded from PDB website and a non-redundant set of them (n = 144, **Supplementary Table 1**) was composed based on similarity of CDR3α, CDR3β and peptide sequences (see **Supplementary Methods**).

For all structures in the non-redundant set of TCR-peptide-MHC complexes, we extracted contacts between TCR and peptide residues. Residues were considered to be in contact if any two atoms of these residues were separated by a distance ≤ 5 Å. The number of contacts that CDR and framework regions of TCR form with peptides in different crystal structures is shown in **Supplementary Figure 1**, the number of contacts that residues in specific positions in CDRs of different lengths form with peptides is shown in **Supplementary Figure 10**.

Derivation of TCRen potential is based on the assumption that the distribution of frequencies of pairwise amino acid contacts can be described using the inverse Boltzmann relationship:

$$\varepsilon_{TCRen}(a,b) = -\ln\left(\frac{p_{obs}(a,b)}{p_{exp}(a,b)}\right), \qquad (1)$$

where a is a residue from TCR, b is a residue from the peptide, $p_{obs}(a,b)$ is the observed probability of contacts between these residues in the database, and $p_{exp}(a,b)$ is the expected probability of contacts between these residues when assuming random pairing of amino acids.

$p_{obs}(a,b)$ and $p_{exp}(a,b)$ are calculated using the following equations:

$$p_{obs}(a,b) = \frac{\#contact(a,b)}{\sum_{a,b}\#contact(a,b)}, \qquad (2)$$

where $\#contact(a,b)$ is the number of contacts between residues a and b observed in the database plus pseudocount 1, and

$$p_{exp}(a,b) = p_{obs}(a) \times p_{obs}(b), \qquad (3)$$

where $p_{obs}(a)$ and $p_{obs}(b)$ are the frequencies of residues a and b, respectively, in the reference state.

For the asymmetric version of TCRen with the contact-fraction reference state (TCRen.a.cf), which demonstrated the best performance in benchmarking experiments, $p_{obs}(a)$ and $p_{obs}(b)$ were calculated as:

$$p_{obs}(a) = \sum_{c} p_{obs}(a,c);\ p_{obs}(b) = \sum_{d} p_{obs}(d,b),$$

where summation is carried out through all 20 amino acids denoted with indices c and d.

In this work we also considered two other reference states ("mole-fraction" and "normalized mole-fraction") and corresponding symmetric versions of TCRen. Equations used for their derivation are described in detail in **Supplementary Methods**.

The interaction energy between TCR and peptide in the given complex is estimated as:

$$E_{TCRen} = \sum_{i \in peptide,\, j \in TCR} \delta(i,j) \times \varepsilon_{TCRen}(a_i, b_j), \qquad (4)$$

where summation is carried out through all residues i in the peptide and all residues j in the TCR; $\delta(i,j) = 1$ if residues i and j are in contact and $= 0$ otherwise.

Cognate epitope identification among unrelated peptides

For each structure in the non-redundant set of TCR-peptide-MHC complexes, we extracted a contact map: a list of contacts between particular residues in the TCR and peptide with residue positions indicated (see **Figure 1** for an example).

For each entry in our non-redundant database of TCR-peptide-MHC structures, we composed a list of 1,000 unrelated peptides with the same length as the cognate epitope, but which were mismatched from the cognate epitope in all non-anchor positions. The second and last residues were considered anchors, as they are responsible for MHC binding. For each unrelated peptide, we generated contact maps from the corresponding crystal structure and substituted residues of the cognate epitope with residues located at the same positions in the unrelated peptide.

For each peptide (both cognate epitope and unrelated ones) we calculated TCRen score using equation 4. Then for each structure in our non-redundant set we calculated coganate epitope rank - the fraction of unrelated peptides which have TCRen score lower (better) than the score of the cognate epitope.

Construction of shuffled TCR-peptide-MHC complexes

In the database of TCR-peptide-MHC complexes we found pairs of structures which have the same lengths of CDR3α, CDR3β and peptide, but mismatched in at least 3 positions in each of these 3 regions. Shuffled complexes were constructed by combining TCRs and pMHC from different complexes in these pairs. Contact map for a shuffled complex was taken from the original crystal structure with TCR included in the shuffled complex. For this benchmark we used only complexes with MHC-I as peptides in complexes with MHC-II may accommodate different binding registers that makes comparison of contact maps in different complexes ambiguous.

Molecular dynamics simulations of DMF5 TCR-peptide-HLA-A02 complexes

MD simulations were performed for DMF5-peptide-HLA-A02 complexes with 42 different peptides from 3 groups (“GIG”, “DRG” and “random”) using two different starting structures. This yields 84 systems, for each 100 ns MD simulations were carried out in at least 3 replicas for “GIG” and “DRG” peptides and 2 replicas for “random” peptides (a total simulation time was 17,8 μs). Sequences for “GIG” and “DRG” peptides were taken from the work by Riley et al. [39], both groups consist of 5 peptides for each of them recognition by DMF5 TCR in HLA-A02 context was confirmed using SPR and INFγ release assay. 32 “Random” peptides were taken from

IEDB database [40] with the following conditions: they are experimentally determined to be HLA-A*02:01 binders, derived from human proteins and share neither “GIG” nor “DRG” motif. Starting structures for MD simulations were constructed based on PDB crystal structure templates (6AM5 for “standard” register, 6AMU for “shifted” register) with corresponding peptide sequence introduced by “mutate_residue” PyRosetta protocol.

Preparation of the system and MD simulations were carried out using GROMACS v.5.1.4 software with Amber99sb-ildn force field (see **Supplementary Methods** for details).

Rosetta Interface analyser module was used for calculation of the binding energy (dG_separated) according to Rosetta scoring function [10] between TCR and peptide. Molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) was applied for estimation of binding energy between TCR and peptide-MHC using g_mmpbsa tool [24]. 100-ns MD trajectories were splitted to 1000 frames, for each frame calculated ΔG using Rosetta scoring function or MMPBSA and then ΔG was averaged over all the frames.

Identification of cancer neo-epitopes recognized by tumor-infiltrating lymphocytes

Sequences of 302TIL α- and β-chains were taken from the crystal structure PDB ID: 6UK4 (from Devlin *et al.*[13]). The model of TCR-peptide-MHC complex for 302TIL with the random peptide VVVVVVVVV (chosen because valine has an intermediate value of side-chain volume amongst the 20 canonical amino acids) and HLA-A*02:06 was obtained using the TCRpMHCmodels server[14], which performs homology modeling based on an internal database of TCR-peptide-MHC templates. Templates selected by TCRpMHCmodels for TCR-peptide-MHC modeling are shown in **Supplementary Figure 11A**. Superimposition of the resulting homology model and crystal structure 6UK4 is shown in **Supplementary Figure 11B**.

Sequences of candidate neo-epitopes were from Bobisse *et al.*[12]. In this study, 20 peptides were identified that were predicted to bind to HLA-A*02:06 and include non-synonymous somatic mutations. For both the homology model and the crystal structure, we derived contact maps, and for each of the 20 candidate neo-epitopes, we estimated the energy of TCR-peptide interaction based on this contact map using TCRen as described above. Sequences of candidate neo-epitopes and the corresponding TCRen scores of TCR-peptide interaction are shown in **Supplementary Table 6.**

Acknowledgements

Supported by grant from the Ministry of Science and Higher Education of Russian Federation (075-15-2019-1789). MD simulations were supported by the HSE University Basic Research Program and carried out with the use of computational facilities of the Supercomputer Center “Polytechnical” at the St. Petersburg Polytechnic University.

Supplementary Materials

(Supplementary Methods, Supplementary Figures, Supplementary Tables, Supplementary Note 1)

Supplementary Methods

Selection of a nonredundant set of TCR-peptide-MHC structures

For each pair of TCR-peptide-MHC crystal structures from PDB we calculated Levenstein distance between their CDR3α, CDR3β and peptide sequences ($d_{CDR3\alpha}$, $d_{CDR3\beta}$ *and* $d_{peptide}$, respectively) and the overall sequence distance between these two structures were calculated as:

$$d = d_{CDR3\alpha} + d_{CDR3\beta} + d_{peptide}.$$

Then we performed hierarchical clustering of PDB structures based on this measure with cutting the resulting tree at the height 6. This procedure yielded in 135 clusters.

If selection of a non-redundant set is based only on sequence information, this may result in filttering out structures with similar TCRs recognising similar epitopes using different mechanisms (see, for example, structures with PDB ID 1NAM and 2OL3; 5M02 and 6G9Q). To include such cases in our database, we performed additional clustering based on contact information. For each structure we extracted all TCR-peptide contacts and represented them as a 400-mer vector containing information about the number of contacts of residues of the given type (i.e. if we label amino acids from A to W as 0 to 19, the contact between *i*-th and *j*-th amino acid has an index of $i * 20 + j$). For each pair of structures we calculated the Manhattan (rectilinear) distance between these 400-mer vectors normalized on the sum of lengths of these vectors. Then we performed hierarchical clustering of PDB structures based on this measure with cutting the resulting tree at the height 0.7. This procedure yielded in 91 clusters.

Finally, we divided all the PDB structures to groups containing structures assigned to the same clusters according to both sequence-based and contact-baseded clustering. We select a single representative from each group to include in the nonredundant set of TCR-peptide-MHC structures (n = 144).

Calculation of asymmetrical and symmetrical versions of TCRen for different reference states

We also explored other approaches for calculating TCRen. We considered three different reference states—contact-fraction, mole-fraction, and normalized mole-fraction—and derived symmetric and asymmetric versions for each of these. In the symmetric version, we considered

contacts without assignment of which molecule type (peptide or TCR) the contacting residues come from:

$$\#contact(a,b) = \#contact(a^{peptide}, b^{TCR}) + \#contact(a^{TCR}, b^{peptide}).$$

That resulted in

$\#contact(a,b) = \#contact(b,a)$ and consequently $\varepsilon_{TCRen}(a,b) = \varepsilon_{TCRen}(b,a)$.

In the asymmetric version, we explicitly considered chain types corresponding to both interacting residues:

$$\#contact(a,b) = \#contact(a^{peptide}, b^{TCR}).$$

In general cases, this resulted in:

$$\#contact(a,b) \neq \#contact(b,a)$$

and consequently

$$\varepsilon_{TCRen}(a,b) \neq \varepsilon_{TCRen}(b,a).$$

$p_{obs}(a)$ and $p_{obs}(b)$ are also calculated separately for peptide and TCR, and equation (3) for asymmetric potential may be rewritten as:

$$p_{exp}(a,b) = p_{obs}^{pepide}(a) \times p_{obs}^{TCR}(b)$$

Definitions of the reference states, which represent different approaches for calculation of $p_{exp}(a,b)$, were the following:

1) for contact-fraction, $p_{obs}(a)$ and $p_{obs}(b)$ were calculated as the aggregated fraction of contacts that residue a or b forms with all other residues:

$$p_{obs}(a) = \sum_{c} p_{obs}(a,c),\ p_{obs}(b) = \sum_{d} p_{obs}(d,b);$$

2) for mole-fraction, $p_{obs}(a)$ was the mole fraction of residue a in all peptides in the database:

$$p_{obs}(a) = \frac{\sum_{peptide \in S} n_a}{\sum_{peptide \in S} \sum_{c} n_c};$$

and $p_{obs}(b)$ was the mole fraction of residue b in all CDR3s in the database:

$$p_{obs}(b) = \frac{\sum_{TCR \in S} n_b}{\sum_{TCR \in S} \sum_d n_d},$$

where n_i was the number of occurrences of residue i in the given peptide/CDR3 and summation was carried out through all peptides/CDR3s in database S;

3) for normalized mole-fraction:

$$p_{obs}(a) = \frac{\sum_{peptide \in S} \sum_{position=1}^{length(s)} w(position, length(s)) \times \delta(a, position)}{\sum_c \sum_{peptide \in S} \sum_{position=1}^{length(s)} w(position, length(s)) \times \delta(c, position)}$$

And

$$p_{obs}(b) = \frac{\sum_{TCR \in S} \sum_{position=1}^{length(s)} w(position, length(s)) \times \delta(b, position)}{\sum_d \sum_{TCR \in S} \sum_{position=1}^{length(s)} w(position, length(s)) \times \delta(d, position)},$$

where $w(position,\ length)$ was the average number of contacts that a residue in a given position of peptide/CDR3 of the given length forms with its counterpart. $\delta(i,\ position) = 1$ if there was residue i in the given position of peptide/CDR3, and $= 0$ if there was another residue in this position.

Note, that the mole-fraction reference state is a special case of the normalized mole-fraction reference state when stating $w = 1$ for all positions of peptide/CDR3 of all lengths.

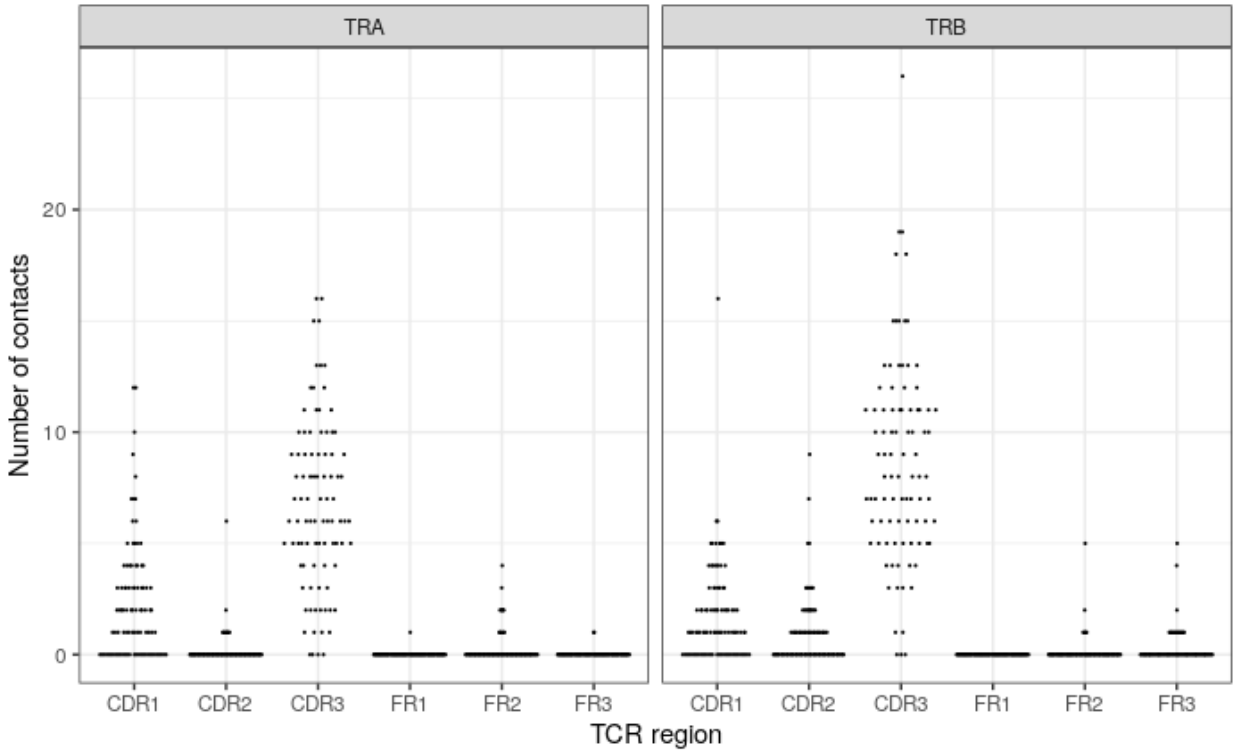
The reason for introducing a normalized mole-fraction reference state is the different propensities of forming contacts for residues at different positions in peptides and CDRs of different lengths. As can be seen from **Supplementary Figure 10**, residues that are close to the center of peptide or CDR3 tend to form more contacts. However, substantial variation in the number of contacts for given peptide and CDR3 positions exists between structures, so the observation that a residue in the center of the peptide or CDR3 does not form contacts with its counterpart in the given structure may indicate that all available contacts are energetically disfavored and that it is more favorable to avoid any contact.

In calculating the mole fractions of residues in TCRs, we considered only residues in CDR3, with trimming of the first three and the last four residues. This is because these residues in CDR3 almost never contact the peptide in available TCR-peptide-MHC crystal structures (**Supplementary Figure 10**). Although CDR1α/β and CDR2α/β often form a number of contacts with the peptide (**Supplementary Figure 1**), position-specific interaction patterns are not regular for these regions, and the median number of contacts is 0 for almost all residue positions in CDR1α/β and CDR2α/β (**Supplementary Figure 10**).

Molecular dynamics simulations

The starting structure was placed in a rectangular box filled with explicit TIP3P water with size allowing the distance between the complex and the box boundaries to be no less than 1.2 nm. Then the system was neutralized by addition of a necessary number of Na^+ ions, minimized by a steepest descent method and heated to 298 K with positional restraints for all heavy atoms. MD replicas were performed from the structures obtained after heating using different seeds for initial velocities of atoms. In order to prevent rotation of the complex leading to its interaction with periodic images we used positional restraints on Cα-atoms of MHC residues 182 and 197 along the two smaller axes of the box with force constants 10-times less than the ones used in the heating stage. These residues were chosen as they are located in MHC α3-domain so these restraints should have no impact on TCR-pMHC interface behavior. The obtained trajectories were visually checked on absence of interactions of the complex with its periodic images. Use of rectangular box with imposed positional restraints resulted in ~5-time reduction of computational time of MD compared to simulation of the same systems in dodecahedronic box.

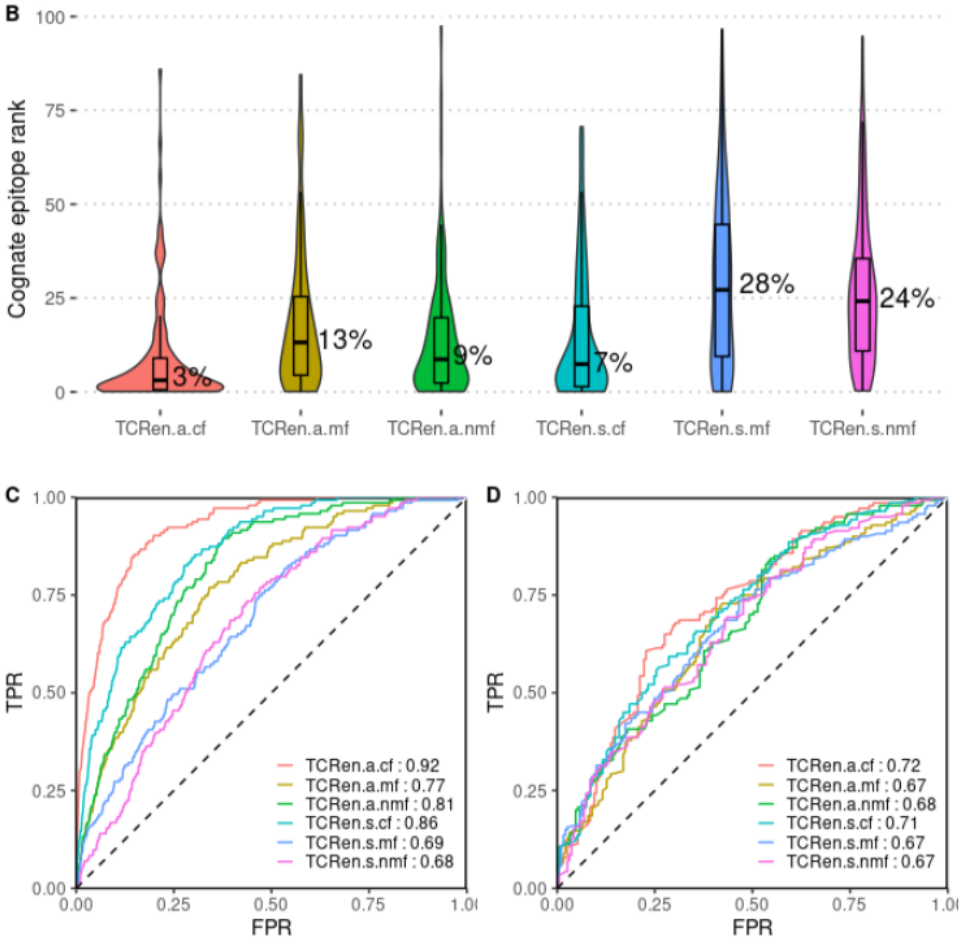
Supplementary Figures



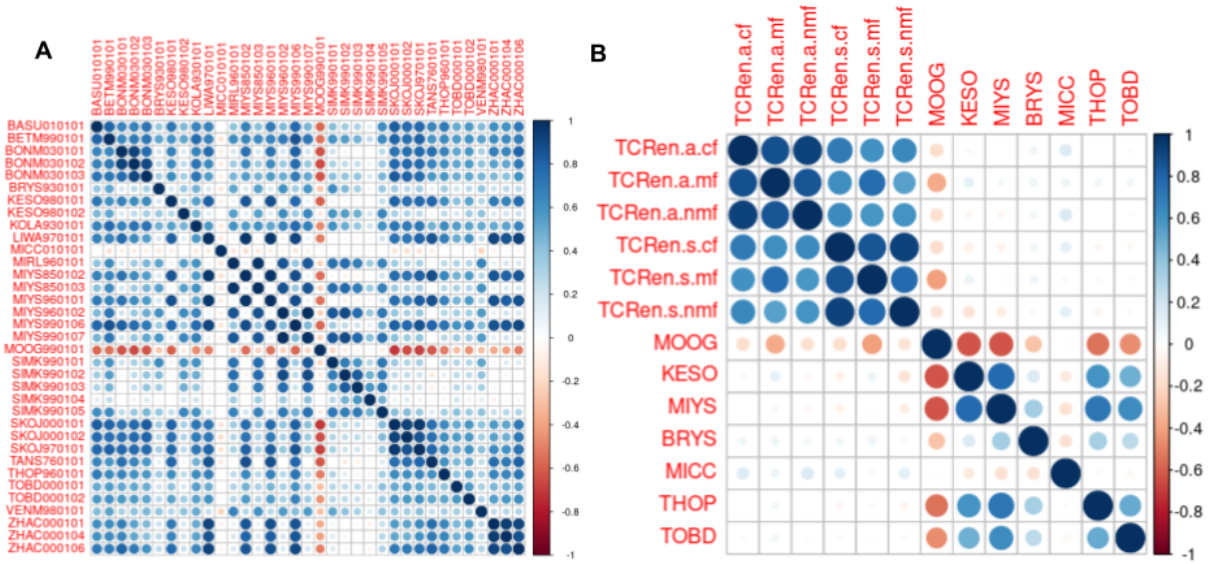
Supplementary Figure 1. Distribution of contacts between complementarity determining (CDR) and framework (FR) regions of TCR with peptide in TCR-pMHC crystal structures of the non-redundant set used for TCRen derivation.

A

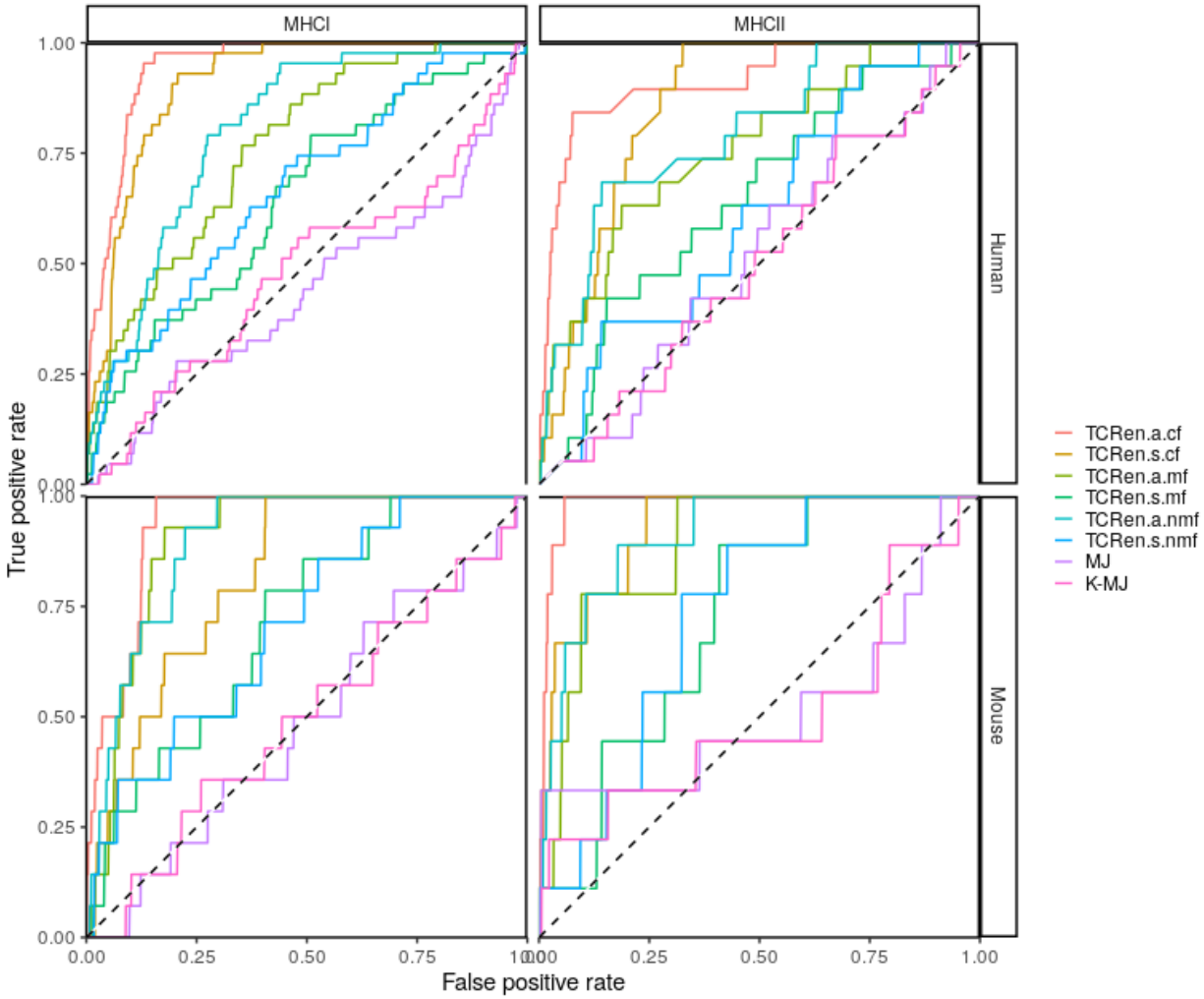
	Reference state		
	contact-fraction	**mole-fraction**	**normalized mole-fraction**
asymmetric	TCRen.a.cf	TCRen.a.mf	TCRen.a.nmf
symmetric	TCRen.s.cf	TCRen.s.mf	TCRen.s.nmf



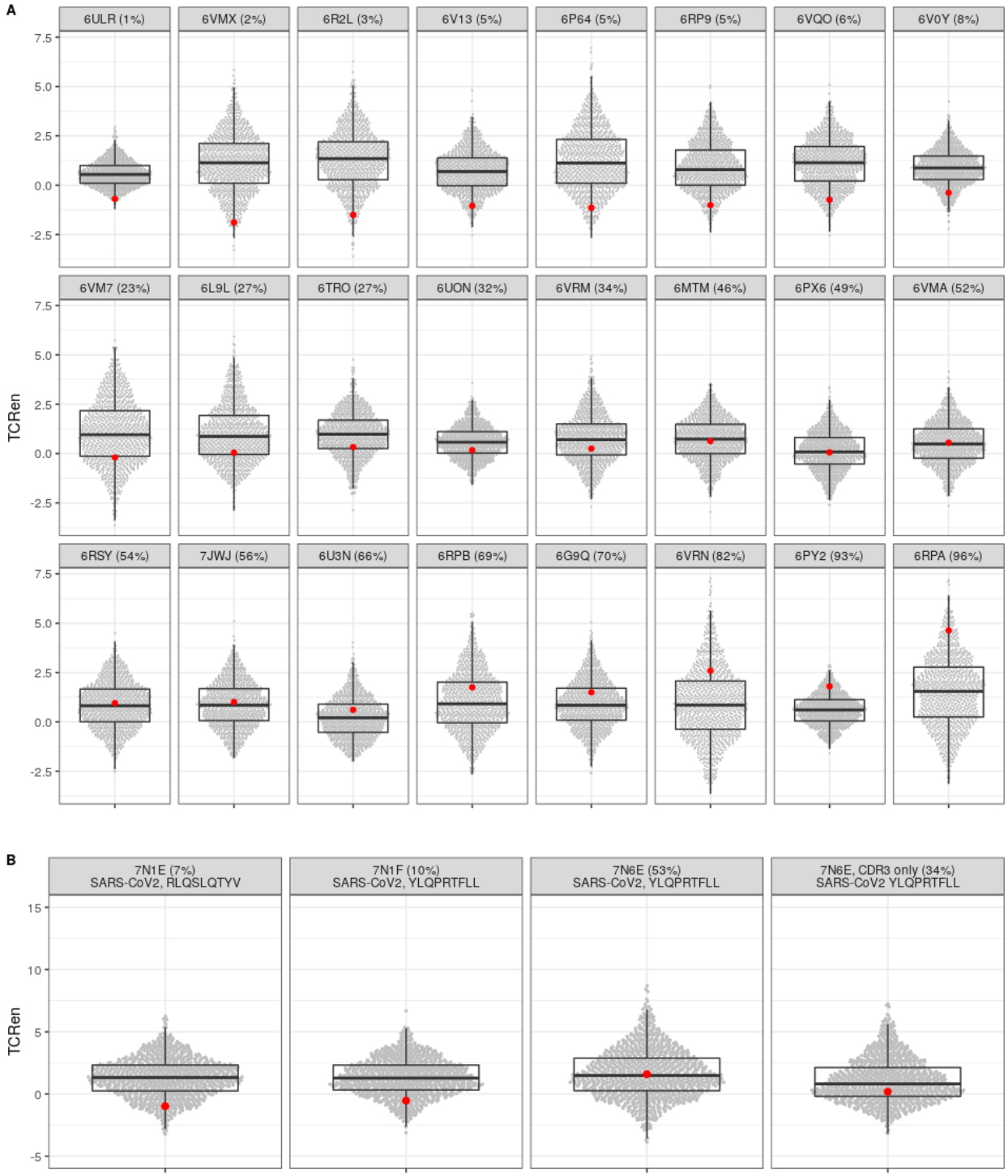
Supplementary Figure 2. Comparison of performance of different versions of TCRen. A. Summary of notations for different versions of TCRen potential. All versions are called as TCRen.X.Y, when X is either "s" for symmetric or "a" for asymmetric forms, Y is either "cf" for contact-fraction or "mf" for mole-fraction or "nmf" for normalized mole-fraction reference states. **B.** Distributions of cognate epitope ranks for different versions of TCRen in cognate/unrelated peptides benchmark (benchmark setup is the same as in **Figure 2A,** see **Methods** for the details). Median value for each potential is marked. **C.** ROC-curves for distinction between all complexes with cognate epitopes and all complexes with unrelated peptides using different versions of TCRen. **D.** ROC-curves for discrimination between real and shuffled TCR-peptide-MHC complexes using different versions of TCRen (benchmark setup is the same as in **Figure 3A,** see **Methods** for the details).



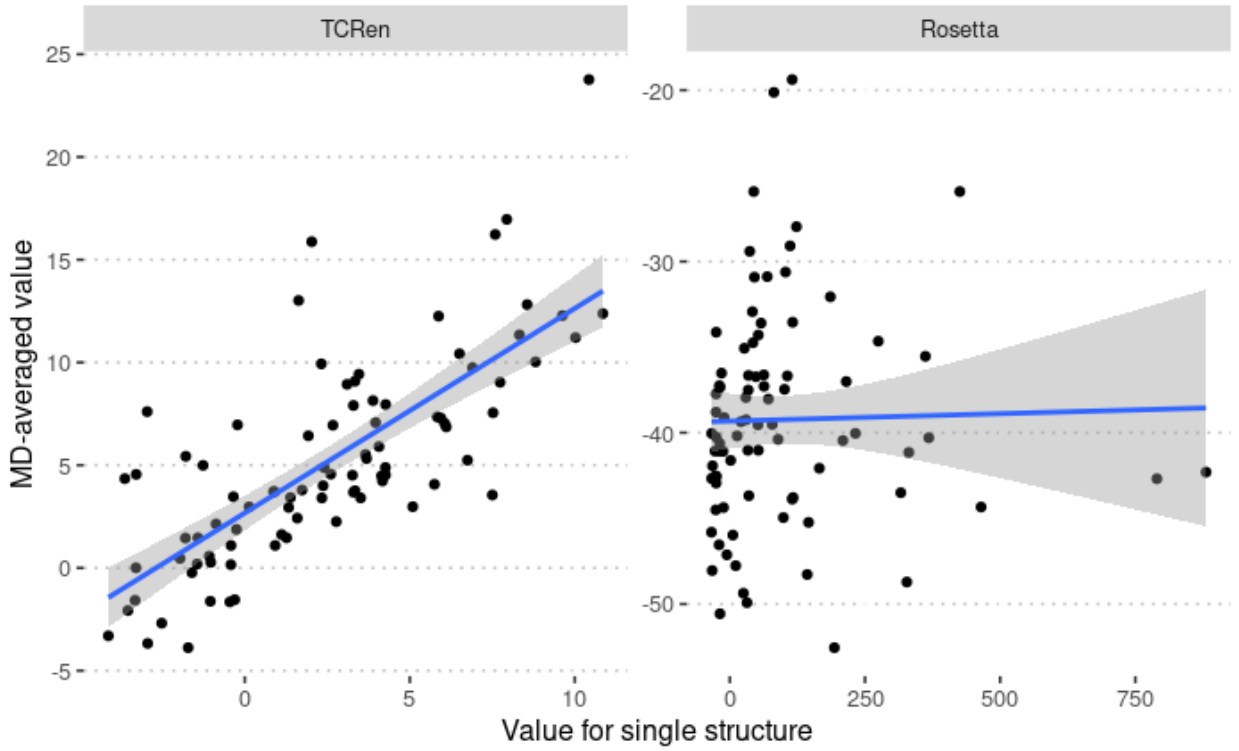
Supplementary Figure 3. Pairwise correlations between the potentials from AAindex and different versions of TCRen. A. Corrplot for pairwise correlation coefficients between all potentials from AAindex database. **B.** Corrplot for pairwise correlation coefficients between selected representative potentials from AAindex database and different versions of TCRen potential. Note that none of TCRen versions correlate with AAindex potentials, symmetric and asymmetric versions TCRen with the same reference state only modestly correlate with each other while correlation of potentials with the same (a)symmetry and different reference states is high.



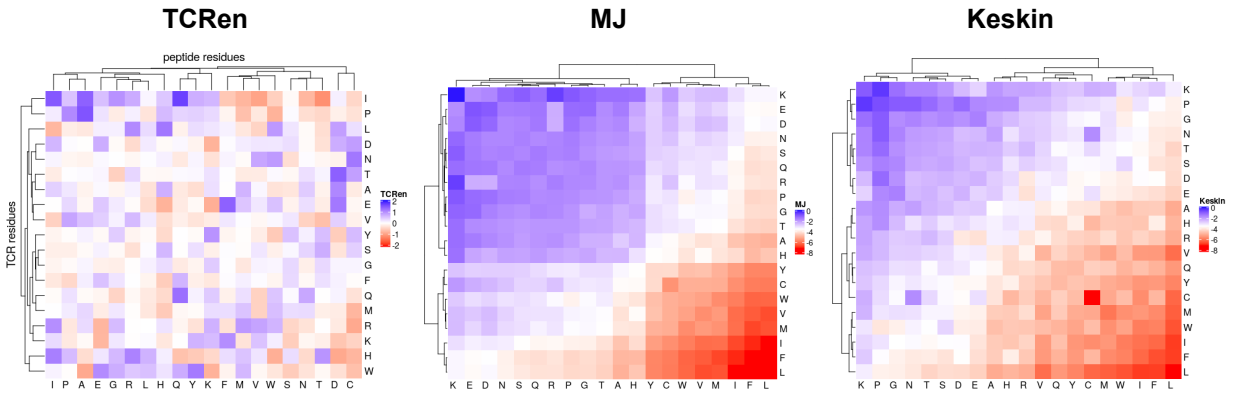
Supplementary Figure 4. Performance of different potentials in peptide screening benchmark splitted by MHC class and species. Results presented in **Figure 2B** (MIYS (MJ) and KESO (K-MJ) were selected as representative potentials) and **Supplementary Figure 2C** are splitted according to MHC class in the corresponding complex (MHCI or MHCII) and species of the host organism (human or mouse).



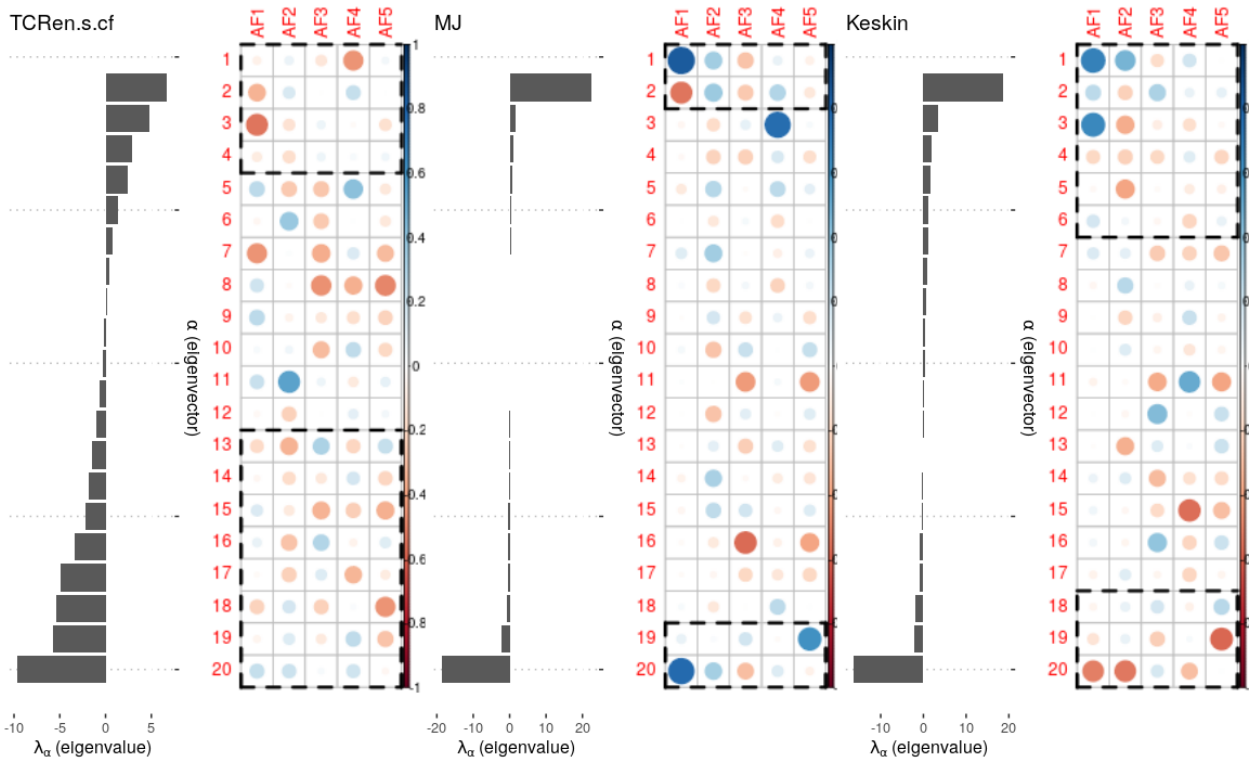
Supplementary Figure 5. Validation of TCRen performance using an independent test set of recently released PDB structures. **A.** Results for test structures, shown in the ascending order of cognate epitope rank. **B.** Results for structures with TCRs recognising peptides from SARS-CoV2. Red points show values of energy of interaction with TCR for cognate epitopes, grey points - values for unrelated peptides. Cognate epitope ranks are shown in brackets in panel titles. TCRen was derived based on structures released before December 31, 2018, test structures are those released after January 1, 2019.



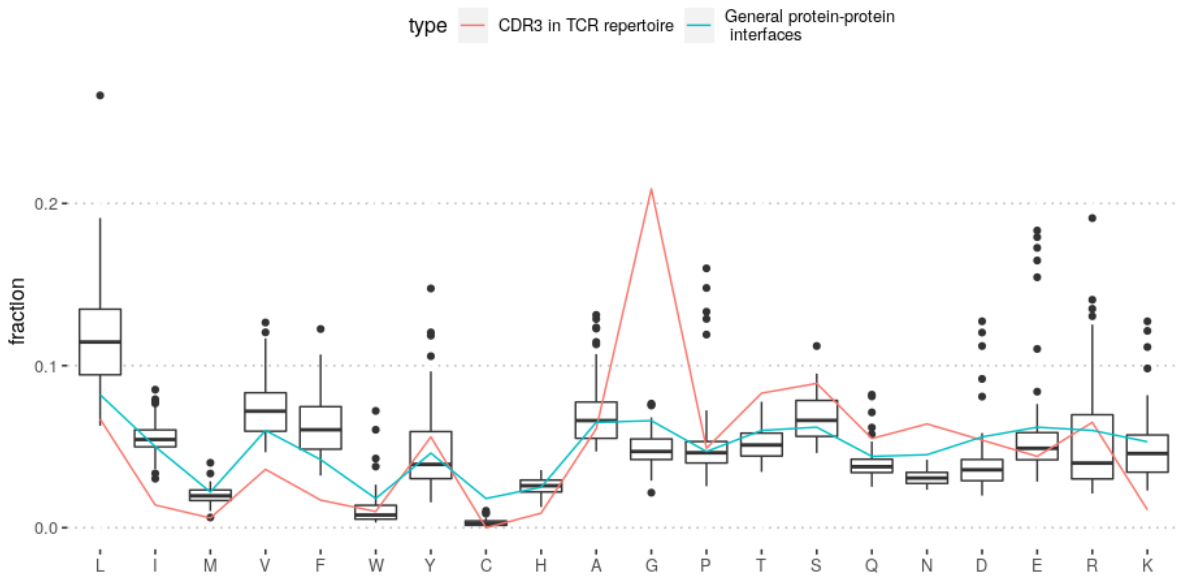
Supplementary Figure 6. Robustness of TCRen.a.cf and Rosetta potentials. X axis correspond to values of energy of TCR-peptide interaction averaged across 1000 frames of 100-ns MD trajectories, Y axis - values calculated from one static structure obtained using “fixbb” protocol.



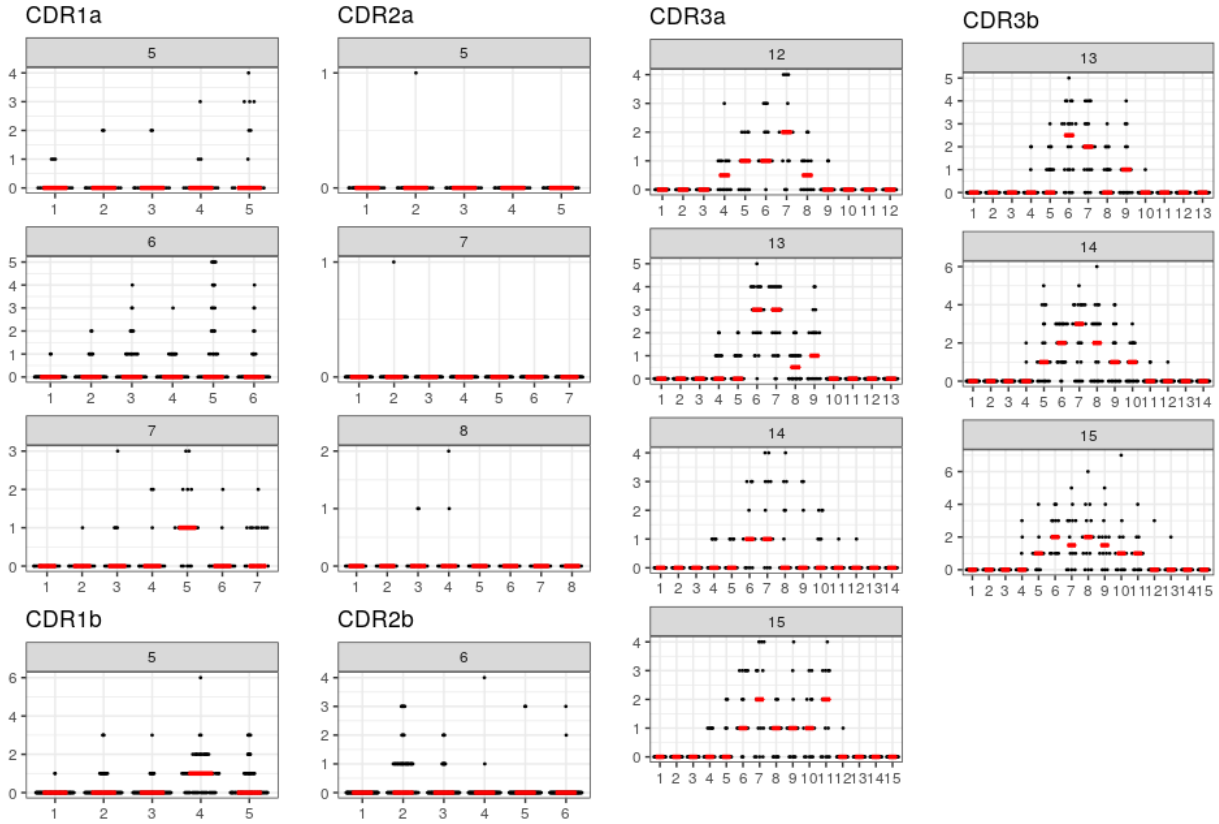
Supplementary Figure 7. Clustering of TCRen, MJ and Keskin matrices. For TCRen matrix axis X corresponds to peptide residues, axis Y corresponds to TCR residues. Hierarchical clustering was performed based on the euclidean distance using stats::hclust R function.



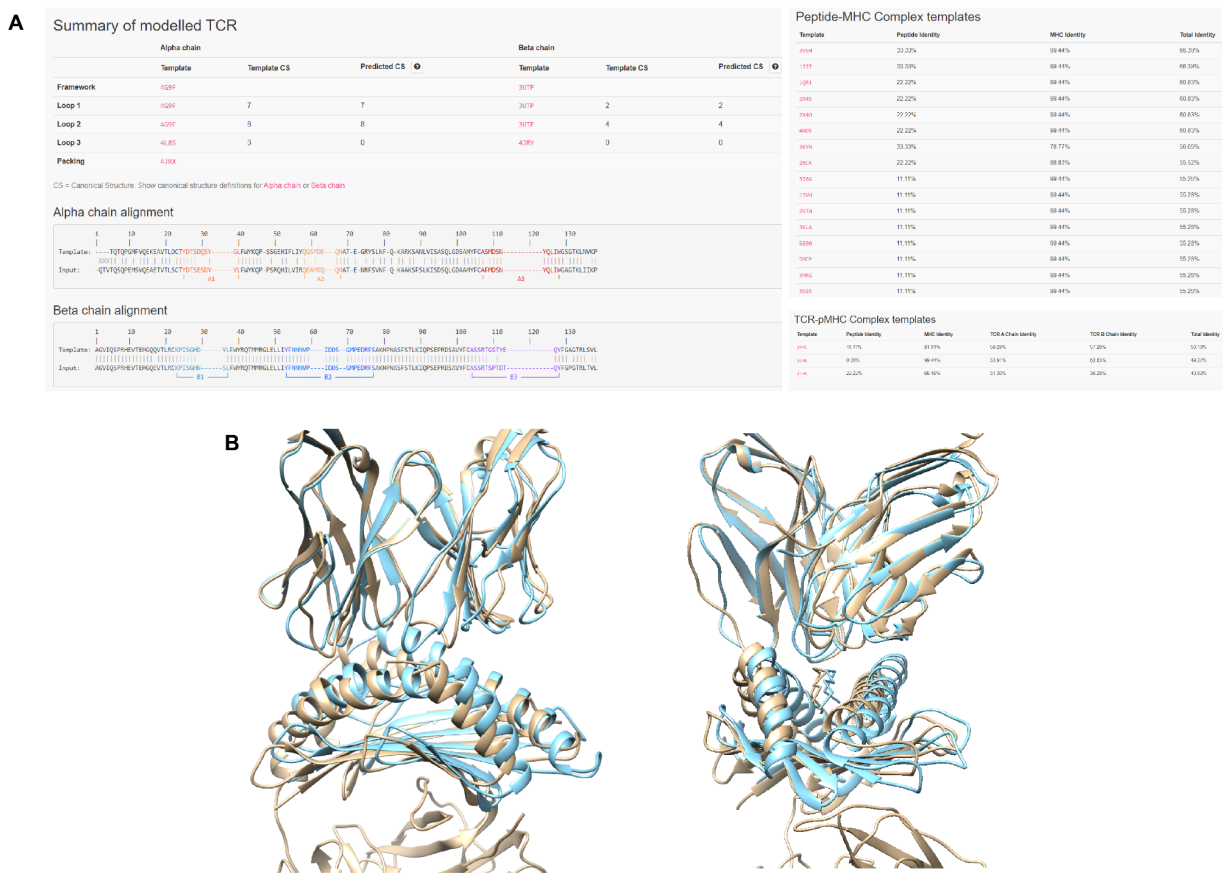
Supplementary Figure 8. Eigenvalue decomposition for matrices of TCRen.s.cf, MJ and Keskin potentials. The left half represents eigenvalues, the right half shows correlation coefficients between corresponding eigenvectors and AFs of residues. For each matrix eigenvectors which cumulatively explain 90% of variance are marked with dashed rectangles.



Supplementary Figure 9. Amino acid composition of ligands for different HLA alleles. Boxplot shows distribution of fractions of amino acids in ligands for different HLA alleles (only HLA-A, HLA-B and HLA-C alleles, for which more than 1000 unique ligands are available in IEDB, are included, n = 54). Red and blue lines correspond to amino acid composition of CDR3 regions in TCR repertoire and general protein-protein interfaces, respectively. Note that median values of amino acid fractions in HLA ligands are closer to amino acid composition of general protein-protein interfaces than that of CDR3 regions.



Supplementary Figure 10. Number of contacts that residues in different positions of CDR1, CDR2 and CDR3 of different lengths form with peptides in structures from the non-redundant set of TCR-peptide-MHC complexes. X axis corresponds to the numbering of residues in CDR, Y axis is the number of contacts of corresponding residue with peptide. Red lines mark median values. Note that median values are 0 for almost all residues in CDR1 and CDR2. Only lengths of CDRs that have more than 10 corresponding structures in the non-redundant database are shown.



Supplementary Figure 11. Homology modeling of 302TIL TCR-pepide-MHC complex. A. Templates used by TCRpMHCmodels server for homology modeling of TCR, peptide-MHC complex and TCR-pMHC relative orientation. **B.** Superposition of homology model (blue) and crystal structure (brown) of TCR-pepide-MHC complex with 302TIL.

Supplementary Tables

Supplementary Table 1. Non-redundant set of crystal structures of TCR-peptide-MHC complexes from PDB used in the work for derivation of TCRen potential.

PDB ID	Species	MHC class	peptide	CDR3a	CDR3b
1AO7	Human	MHCI	LLFGYPVYV	CAVTTDSWGKLQF	CASRPGLAGGRPEQYF
1BD2	Human	MHCI	LLFGYPVYV	CAAMEGAQKLVF	CASSYPGGGFYEQYF
1D9K	Mouse	MHCII	GNSHRGAIEWEGIESG	CAATGSFNKLTF	CASGGQGRAEQFF
1FO0	Mouse	MHCI	INFDFNTI	CAMRGDYGGSGNKLIF	CTCSADRVGNTLYF
1FYT	Human	MHCII	PKYVKQNTLKLAT	CAVSESPFGNEKLTF	CASSSTGLPYGYTF
1G6R	Mouse	MHCI	SIYRYYGL	CAVSGFASALTF	CASGGGGTLYF
1KJ2	Mouse	MHCI	KVITFIDL	CAARYQGGRALIF	CTCSAAPDWGASAETLYF
1LP9	Human	MHCI	ALWGFFPVL	CALFLASSSFSKLVF	CASSDWVSYEQYF
1MI5	Human	MHCI	FLRGRAYGL	CILPLAGGTSYGKLTF	CASSLGQAYEQYF
1NAM	Mouse	MHCI	RGYVYQGL	CAMRGDYGGSGNKLIF	CTCSADRVGNTLYF
1OGA	Human	MHCI	GILGFVFTL	CAGAGSQGNLIF	CASSSRSSYEQYF
1U3H	Mouse	MHCII	SRGGASQYRPSQ	CAASANSGTYQRF	CASGDAGGGYEQYF
1YMM	Human	MHCII	ENPVVHFFKNIVTP	CATDTTSGTYKYIF	CSARDLTSGANNEQFF
1ZGL	Human	MHCII	VHFFKNIVTPRTPG	CALSGGDSSYKLIF	CASSLADRVNTEAFF
2AK4	Human	MHCI	LPEPLPQGQLTAY	CALSGFYNTDKLIF	CASPGLAGEYEQYF
2BNQ	Human	MHCI	SLLMWITQV	CAVRPTSGGSYIPTF	CASSYVGNTGELFF
2E7L	Mouse	MHCI	QLSPFPFDL	CAVSHQGRYLTF	CASGGGGTLYF
2ESV	Human	MHCI	VMAPRTLIL	CIVVRSSNTGKLIF	CASSQDRDTQYF
2IAM	Human	MHCII	GELIGILNAAKVPAD	CAALIQGAQKLVF	CASTYHGTGYF
2NX5	Human	MHCI	EPLPQGQLTAY	CAVQASGGSYIPTF	CATGTGDSNQPQHF
2OL3	Mouse	MHCI	SQYYYNSL	CAMRGDYGGSGNKLIF	CTCSADRVGNTLYF
2PXY	Mouse	MHCII	RGGASQYRPSQ	CALSENYGNEKITF	CASGDASGAETLYF
2YPL	Human	MHCI	KAFSPEVIPMF	CAVSGGYQKVTF	CASTGSYGYTF
3C5Z	Mouse	MHCII	FEAQKAKANKAVDG	CALVISNTNKVVF	CASIDSSGNTLYF
3C60	Mouse	MHCII	FEAQKAKANKAVD	CAANSGTYQRF	CASGDFWGDTLYF
3C6L	Mouse	MHCII	FEAQKAKANKAV	CAASDNRIFF	CASGDAWGYEQYF
3DXA	Human	MHCI	EENLLDFVRF	CIVWGGYQKVTF	CASRYRDDSYNEQFF
3FFC	Human	MHCI	FLRGRAYGL	CAMREDTGNQFYF	CASSFTWTSGGATDTQYF
3GJF	Human	MHCI	SLLMWITQV	CAGELLPYYGMDVW	CWSFAGSYYVF
3GSN	Human	MHCI	NLVPMVATV	CARNTGNQFYF	CASSPVTGGIYGYTF
3HG1	Human	MHCI	ELAGIGILTV	CAVNVAGKSTF	CAWSETGLGTGELFF
3KPR	Human	MHCI	EEYLKAWTF	CILPLAGGTSYGKLTF	CASSLGQAYEQYF
3MBE	Mouse	MHCII	GAMKRHGLDNYRGYSLG	CAAEDGGSGNKLIF	CASSWDRAGNTLYF
3MV7	Human	MHCI	HPVGEADYFEY	CAVQDLGTSGSRLTF	CASSARSGELFF
3NFJ	Human	MHCI	RYPLTFGWCF	CALGELARSGGYQKVTF	CAWSVSVGAGVPTIYF
3O4L	Human	MHCI	GLCTLVAML	CAEDNNARLMF	CSARDGTGNGYTF
3O6F	Human	MHCII	FSWGAEGQRPGF	CTVYGGATNKLIF	CSARGGSYNSPLHF
3PL6	Human	MHCII	NPVVHFFKNIVTPR	CAASSFGNEKLTF	CATSALGDTQYF
3PQY	Mouse	MHCI	SSLENFRAYV	CILSGGSNYKLTF	CASSFGREQYF
3PWP	Human	MHCI	LGYGFVNYI	CAVTTDSWGKLQF	CASRPGLAGGRPEQYF
3QDG	Human	MHCI	ELAGIGILTV	CAVNFGGGKLIF	CASSLSFGTEAFF
3QDM	Human	MHCI	ELAGIGILTV	CAGGTGNQFYF	CAISEVGVGQPQHF
3QIB	Mouse	MHCII	ADLIAYLKQATKG	CAALRATGGNNKLTF	CASSLNWSQDTQYF
3QIU	Mouse	MHCII	ADLIAYLKQATKG	CAAEPSSGQKLVF	CASSLNNANSDYTF
3RDT	Mouse	MHCII	FEAQKAKANKAVD	CAASKGADRLTF	CASGDFWGDTLYF
3RGV	Mouse	MHCI	WIYVYRPMGCGGS	CAANSGTYQRF	CASGDFWGDTLYF
3SJV	Human	MHCI	FLRGRAYGL	CVVRAGKLIF	CASGQGNFDIQYF
3TFK	Mouse	MHCI	QLSDVPMDL	CAVSAKGTGSKLSF	CASSDAPGQLYF
3TJH	Mouse	MHCI	SPLDSLWWI	CAVSAKGTGSKLSF	CASSDAPGQLYF
3TPU	Mouse	MHCI	FLSPFWFDI	CAVSAKGTGSKLSF	CASSDAPGQLYF
3UTS	Human	MHCI	ALWGPDPAAA	CAMRGDSSYKLIF	CASSLWEKLAKNIQYF
3VXM	Human	MHCI	RFPLTFGWCF	CAVGAPSGAGSYQLTF	CASSPTSGIYEQYF

3VXR	Human	MHCI	RYPLTFGWCF	CAVRMDSSYKLIF	CASSSWDTGELFF
4E41	Human	MHCII	GELIGILNAAKVPAD	CAVDRGSTLGRLYF	CASSQIRETQYF
4EUP	Human	MHCI	ALGIGILTV	CAVSGGGADGLTF	CASSFLGTGVEQYF
4FTV	Human	MHCI	LLFGYPVYV	CAVTTDSWGKLQF	CASRPGLMSAQPEQYF
4G8G	Human	MHCI	KRWIILGLNK	CAMRDLRDNFNKFYF	CASREGLGGTEAFF
4GG6	Human	MHCII	SGEGSFQPSQENP	CILRDGRGGADGLTF	CASSVAVSAGTYEQYF
4GRL	Human	MHCII	RLLMLFAKDVVSRN	CAASSFGNEKLTF	CATSALGDTQYF
4H1L	Human	MHCII	QHIRCNIPKRISA	CAVGASGNTGKLIF	CASSLRDGYTGELFF
4JFD	Human	MHCI	ELAAIGILTV	CAVNDGGRLTF	CAWSETGLGMGGWQF
4JRY	Human	MHCI	LPEPLPQGQLTAY	CAVGGGSNYQLIW	CASSRTGSTYEQYF
4MAY	Human	MHCII	QLVHFVRDFAQL	CAASSFGNEKLTF	CATSALGDTQYF
4MJI	Human	MHCI	TAFTIPSI	CATDDDSARQLTF	CASSLTGGGELFF
4MNQ	Human	MHCI	ILAKFLHWL	CAVDSATALPYGYIF	CASSYQGTEAFF
4MS8	Mouse	MHCI	SPAEAGFFL	CAVSAKGTGSKLSF	CASSDAPGQLYF
4MXQ	Mouse	MHCI	SPAPRPLDL	CAVSAKGTGSKLSF	CASSDAPGQLYF
4N5E	Mouse	MHCI	VPYMAEFGM	CAVSAKGTGSKLSF	CASSDAPGQLYF
4OZF	Human	MHCII	APQPELPYPQPGS	CIAFQGAQKLVF	CASSFRALAADTQYF
4OZG	Human	MHCII	APQPELPYPQPG	CIVLGGADGLTF	CASSFRFTDTQYF
4OZH	Human	MHCII	APQPELPYPQPGS	CIVWGGATNKLIF	CASSVRSTDTQYF
4OZI	Human	MHCII	QPFPQPELPYP	CLVGDGGSFSGGYNKLIF	CSAGVGGQETQYF
4P2O	Mouse	MHCII	PADPLAFFSSAIKGGGGSLV	CAALRATGGNNKLTF	CASSLNWSQDTQYF
4P2Q	Mouse	MHCII	ADGLAYFRSSFKGG	CAAEASNTNKVVF	CASSLNNANSDYTF
4P2R	Mouse	MHCII	ANGVAFFLTPFKA	CAAEASNTNKVVF	CASSLNNANSDYTF
4P4K	Human	MHCII	QAFWIDLFETIG	CALSLYSGAGSYQLTF	CASSLAQGGETQYF
4P5T	Mouse	MHCII	FEAQKAKANKAVD	CAASRDSGQKLVF	CASGDFWGDTLYF
4QRP	Human	MHCI	HSKKKCDEL	CALSDPVNDMRF	CASSLRGRGDQPQHF
4QRR	Human	MHCI	IPSINVHHY	CALGELAGAGGTSYGKLTF	CASSLEGGYYNEQFF
4WUU	Human	MHCI	RMFPNAPYL	CARVQYSGYYDWFDPW	CAAWDDSLNGWVF
4Y19	Human	MHCII	QPLALEGSLQKRG	CAASVYAGGTSYGKLTF	CASRPRRDNEQFF
4Y1A	Human	MHCII	LQPLALEGSLQKRG	CAASSSAGGTSYGKLTF	CASRPRDPVTQYF
4Z7U	Human	MHCII	APSGEGSFQPSQENPQ	CILRDRSNQFYF	CASSTTPGTGTETQYF
4Z7V	Human	MHCII	SGEGSFQPSQENP	CILRDSRAQKLVF	CASSAGTSGEYEQYF
4Z7W	Human	MHCII	PSGEGSFQPSQENPQ	CAVGETGANNLFF	CASSEARRYNEQFF
5BRZ	Human	MHCI	EVDPIGHLY	CAVRPGGAGPFFVVF	CASSFNMATGQYF
5C07	Human	MHCI	YQFGPDFPIA	CAMRGDSSYKLIF	CASSLWEKLAKNIQYF
5C0C	Human	MHCI	RQFGPDWIVA	CAMRGDSSYKLIF	CASSLWEKLAKNIQYF
5D2L	Human	MHCI	NLVPMVATV	CAFITGNQFYF	CASSQTQLWETQYF
5D2N	Human	MHCI	NLVPMVATV	CILDNNNDMRF	CASSLAPGTTNEKLFF
5E6I	Human	MHCI	GILGFVFTL	CAGPGGSSNTGKLIF	CASSLIYPGELFF
5E9D	Human	MHCI	ELAGIGILTV	CAVTKYSWGKLQF	CASRPGWMAGGVELYF
5EU6	Human	MHCI	YLEPGPVTV	CAVLSSGGSNYKLTF	CASSFIGGTDTQYF
5EUO	Human	MHCI	GILGFVFTL	CAGAIGPSNTGKLIF	CASSIRSSYEQYF
5ISZ	Human	MHCI	GILGFVFTL	CAFDTNAGKSTF	CASSIFGQREQYF
5IVX	Mouse	MHCI	RGPGRAFVTI	CAASASFGDNSKLIW	CASSLGHTEVFF
5JHD	Human	MHCI	XGILGFVFTL	CAWGVNAGGTSYGKLTF	CASSIGVYGYTF
5JZI	Human	MHCI	XKLVALGINAV	CAYGEDDKIIF	CASRRGPYEQYF
5KS9	Human	MHCII	APSGEGSFQPSQENPQ	CAVALNNNAGNMLTF	CASSVAPGSDTQYF
5KSA	Human	MHCII	QPQQSFPEQEA	CAVQFMDSNYQLIW	CASSVAGTPSYEQYF
5KSB	Human	MHCII	GPQQSFPEQEA	CAVQASGGSYIPTF	CASSNRGLGTDTQYF
5M00	Mouse	MHCI	KAVANFATM	CAALYGNEKITF	CASSDAGGRNTLYF
5NHT	Human	MHCI	ELAGIGILTV	CAVGGGADGLTF	CASSQGLAGAGELFF
5NME	Human	MHCI	SLYNTVATL	CAVRTNSGYALNF	CASSDTVSYEQYF
5SWS	Mouse	MHCI	ASNENMETM	CAASEGSGSWQLIF	CASSAGLDAEQYF
5SWZ	Mouse	MHCI	ASNENMETM	CAASETSGSWQLIF	CASSRDLGRDTQYF
5TEZ	Human	MHCI	GILGFVFTL	CAASFIIQGAQKLVF	CASSLLGGWSEAFF
5W1V	Human	MHCI	VMAPRTLIL	CAGQPLGGSNYKLTF	CASSANPGDSSNEKLFF
5WKF	Human	MHCI	GTSGSPIVNR	CGLGDAGNMLTF	CASSLGQGLLYGYTF
5WLG	Mouse	MHCI	SQLLNAKYL	CATVYAQGLTF	CASSDWGDTGQLYF
5XOT	Human	MHCI	IPLTEEAEL	CALGEGGAQKLVF	CASRTRGGTLIEQYF
6AMU	Human	MHCI	MMWDRGLGMM	CAVNFGGGKLIF	CASSLSFGTEAFF
6AVF	Human	MHCI	APRGPHGGAASGL	CLVGEILDNFNKFYF	CASSQRQEGDTQYF
6AVG	Human	MHCI	APRGPHGGAASGL	CLVVDQKLVF	CASSGGHTGSNEQFF
6BJ2	Human	MHCI	IPLTEEAEL	CALSHNSGGSNYKLTF	CASSFRGGKTQYF

6CQL	Human	MHCII	RFYKTLRAEQASQ	CAFKAAGNKLTF	CASSRLAGGMDEQFF
6G9Q	Mouse	MHCI	KAPYDYAPI	CAALYGNEKITF	CASSDAGGRNTLYF
6L9L	Mouse	MHCI	SPSYAYHQF	CILQGTGSKLSF	CASSDGDYEQYF
6MTM	Human	MHCI	FEDLRVLSF	CGTERSGGYQKVTF	CASSMSAMGTEAFF
6P64	Human	MHCI	KQWLVWLFL	CAFMDSNYQLIW	CASSRTSPTDTQYF
6PX6	Human	MHCII	APFSEQEQPVLG	CAVHTGARLMF	CASSHGASTDTQYF
6PY2	Human	MHCII	APFSEQEQPVLG	CASPQGGSEKLVF	CASSSGGWGGGTEAFF
6R2L	Human	MHCI	SLSKILDTV	CAVGGNDWNTDKLIF	CASSPLDVSISSYNEQFF
6RP9	Human	MHCI	SLLMWITQV	CALTRGPGNQFYF	CASSSPGGVSTEAFF
6RPA	Human	MHCI	SLLMWITQV	CAVRDINSGAGSYQLTF	CSVGGSGGADTQYF
6RPB	Human	MHCI	SLLMWITQV	CAVKSGGSYIPTF	CASSYLNRDSALDF
6RSY	Human	MHCI	RMFPNAPYL	CIGGGTTSGTYKYIF	CASSLGFGRDVMRF
6TRO	Human	MHCI	GVYDGREHTV	CVVNHSGGSYIPTF	CASSFLMTSGDPYEQYF
6U3N	Human	MHCII	APMPMPELPYP	CAVGAGSNYQLIW	CASSLEGQGASEQFF
6ULR	Human	MHCI	GADGVGKSA	CLVGDMDQAGTALIF	CASSLGEGRVDGYTF
6UON	Human	MHCI	GADGVGKSAL	CAAAMDSSYKLIF	CASSDPGTEAFF
6V0Y	Human	MHCII	GGYAPAKAAAT	CALSDSGSFNKLTF	CASSLDWGGQNTLYF
6V13	Human	MHCII	GGYRAPAKAAAT	CALSPSNTNKVVF	CASSLDWGVNTLYF
6VM7	Human	MHCI	ITDQVPFSV	CAVNARRNTPLVF	CASSITLSSYNEQFF
6VMA	Human	MHCI	ITDQVPFSV	CAVNALLGNQFYF	CASSMGGTYEQYF
6VMX	Human	MHCI	RPPIFIRRL	CAFGSSNTGKLIF	CASSQDLFTGGYTF
6VQO	Human	MHCI	HMTEVVRHC	CAMSGLKEDSSYKLIF	CASSIQQGADTQYF
6VRM	Human	MHCI	HMTEVVRHC	CVVQPGGYQKVTF	CASSEGLWQVGDEQYF
6VRN	Human	MHCI	HMTEVVRHC	CAFMGYSGAGSYQLTF	CAISELVTGDSPLHF
7JWJ	Mouse	MHCI	ASNENMETM	CAAVTGNTGKLIF	CASSRGTIHSNTEVFF
7N1E	Human	MHCI	RLQSLQTYV	CALSGFNNAGNMLTF	CASSLGGAGGADTQYF
7N1F	Human	MHCI	YLQPRTFLL	CAVNRDDKIIF	CASSPDIEQYF
7N6E	Human	MHCI	YLQPRTFLL	CVVNRNNDMRF	CAGQVTNTGELFF

Supplementary Table 2. PDB IDs of all crystal structures of TCR-peptide-MHC complexes available at PDB. Structures included in the non-redundant set (**Supplementary Table 1**) are marked in bold. Non-redundant structures released after January 1, 2019 and used as a holdout set (**Supplementary Figure 4**) are underlined.

1AO7	**2AK4**	2UWE	**3KPR**	3QIW	**4GG6**	**4OZI**	5BS0	**5KS9**	5WKH	6EQB	**6V0Y**
1BD2	**2BNQ**	2VLJ	3KPS	**3RDT**	**4GRL**	4P23	**5C07**	**5KSA**	**5WLG**	**6G9Q**	**6V13**
1D9K	2BNR	2VLK	3KXF	**3RGV**	**4H1L**	**4P2O**	5C08	**5KSB**	**5XOT**	**6L9L**	6V15
1FO0	2CKB	2VLR	**3MBE**	**3SJV**	**4JFD**	**4P2Q**	5C09	**5M00**	5XOV	**6MTM**	6V18
1FYT	**2E7L**	**2YPL**	**3MV7**	3T0E	4JFE	**4P2R**	5C0A	5M01	5YXN	**6P64**	6V19
1G6R	**2ESV**	2Z31	3MV8	**3TFK**	4JFF	**4P4K**	5C0B	5M02	5YXU	**6PX6**	6V1A
1J8H	2F53	**3C5Z**	3MV9	**3TJH**	4JRX	**4P5T**	**5C0C**	5MEN	6AM5	**6PY2**	**6VM7**
1JTR	2F54	**3C60**	**3NFJ**	**3TPU**	**4JRY**	4PRH	**5D2L**	**5NHT**	**6AMU**	6Q3S	6VM8
1KJ2	2GJ6	**3C6L**	**3O4L**	**3UTS**	4L3E	4PRI	**5D2N**	**5NME**	**6AVF**	**6R2L**	**6VMA**
1LP9	**2IAM**	3D39	**3O6F**	3UTT	**4MAY**	4PRP	**5E6I**	5NMF	**6AVG**	**6RP9**	6VMC
1MI5	2IAN	3D3V	**3PL6**	**3VXM**	**4MJI**	4QOK	**5E9D**	5NMG	**6BJ2**	**6RPA**	**6VMX**
1MWA	2J8U	**3DXA**	**3PQY**	**3VXR**	**4MNQ**	**4QRP**	**5EU6**	5NQK	6BJ3	**6RPB**	**6VQO**
1NAM	2JCC	3E2H	**3PWP**	3VXS	**4MS8**	**4QRR**	**5EUO**	**5SWS**	6BJ8	**6RSY**	**6VRM**
1OGA	**2NX5**	3E3Q	**3QDG**	**4C56**	4MVB	**4WUU**	5HHM	**5SWZ**	**6CQL**	6TMO	**6VRN**
1QRN	2OI9	**3FFC**	3QDJ	**4E41**	**4MXQ**	**4Y19**	5HHO	**5TEZ**	6CQN	**6TRQ**	7JWI
1QSE	**2OL3**	**3GJF**	**3QDM**	**4EUP**	4N0C	**4Y1A**	5HYJ	5TIL	6CQQ	**6U3N**	**7JWJ**
1QSF	2P5E	**3GSN**	3QEQ	4EUQ	**4N5E**	**4Z7U**	**5ISZ**	5TJE	6CQR	6U3O	**7N1E**
1U3H	2P5W	3H9S	3QFJ	**4FTV**	**4OZF**	**4Z7V**	**5IVX**	**5W1V**	6D78	6UK4	**7N1F**
1YMM	**2PXY**	3HAE	**3QIB**	**4G8G**	**4OZG**	**4Z7W**	**5JHD**	5W1W	6DKP	**6ULR**	**7N6E**
1ZGL	2PYE	**3HG1**	**3QIU**	4G9F	**4OZH**	**5BRZ**	**5JZI**	**5WKF**	6EQA	**6UON**	

Supplementary Table 3. Examination of failures in the holdout set validation. Contact maps for PDB 7N6E, 6VRN, 6PY2 and 6RPA with corresponding values of TCRen and number of contacts of these amino acids in the database used for TCRen derivation. Note that for all cases the most disadvantageous contribution to total TCRen energy is from contacts that are not found or occur only once in the structures used for TCRen derivation. Number of contacts for all residue pairs: max 39, mean ± standard error 6.6 ± 0.3.

PDB ID 7N6E					
TCR aa	peptide aa	TCR residue	peptide residue	TCRen	count in database
Q	Q	30 (CDR1)	2	1,03	1
A	P	28 (CDR1)	3	0,36	5
T	R	95 (CDR3)	4	0,36	5
T	R	93 (CDR3)	4	0,36	5
N	Y	26 (CDR1)	0	0,35	8
Q	P	30 (CDR1)	3	0,27	8
D	L	27 (CDR1)	7	0,21	10
M	R	95 (CDR3)	4	0,18	0
Q	R	30 (CDR1)	4	0,13	7
Q	R	91 (CDR3)	4	0,13	7
T	F	93 (CDR3)	6	0,09	14
N	F	94 (CDR3)	6	0,07	19
N	F	93 (CDR3)	6	0,07	19
T	T	93 (CDR3)	5	0,04	9
N	R	90 (CDR3)	4	-0,02	13
N	R	94 (CDR3)	4	-0,02	13
N	R	92 (CDR3)	4	-0,02	13
N	R	93 (CDR3)	4	-0,02	13
N	L	94 (CDR3)	7	-0,03	20
N	L	94 (CDR3)	8	-0,03	20
Q	T	47 (CDR2)	5	-0,07	10
N	T	93 (CDR3)	5	-0,11	15
T	L	93 (CDR3)	7	-0,12	17
N	P	93 (CDR3)	3	-0,13	20
N	P	90 (CDR3)	3	-0,13	20
N	P	92 (CDR3)	3	-0,13	20
G	R	96 (CDR3)	4	-0,13	22
Q	F	30 (CDR1)	6	-0,19	19
Q	L	47 (CDR2)	7	-0,21	18
A	Y	28 (CDR1)	0	-0,22	10
R	P	91 (CDR3)	3	-0,23	17
S	R	31 (CDR1)	4	-0,25	17

PDB ID 6VRN					
TCR aa	peptide aa	TCR residue	peptide residue	TCRen	count in database
L	R	94 (CDR3)	6	0,52	1
V	R	95 (CDR3)	6	0,49	0
S	V	90 (CDR3)	4	0,48	5
S	V	90 (CDR3)	5	0,48	5
E	E	27 (CDR1)	3	0,47	0
T	R	96 (CDR3)	6	0,36	5
R	H	28 (CDR1)	7	0,25	2
S	E	26 (CDR1)	3	0,11	10
S	E	90 (CDR3)	3	0,11	10
N	E	29 (CDR1)	3	0,08	11
N	E	28 (CDR1)	3	0,08	11
Y	H	29 (CDR1)	7	0,03	7
Y	H	95 (CDR3)	7	0,03	7
Y	H	48 (CDR2)	7	0,03	7
Y	V	89 (CDR3)	4	0,03	13
Y	V	89 (CDR3)	5	0,03	13
Y	V	95 (CDR3)	5	0,03	13
Y	R	95 (CDR3)	6	-0,12	17
Y	R	31 (CDR1)	6	-0,12	17
Y	R	89 (CDR3)	6	-0,12	17
G	R	88 (CDR3)	6	-0,13	22
Y	E	89 (CDR3)	3	-0,14	17
R	C	28 (CDR1)	8	-0,15	3
S	R	90 (CDR3)	6	-0,25	17

PDB ID 6PY2					
TCR aa	peptide aa	TCR residue	peptide residue	TCRen	count in database
E	F	96 (CDR3)	2	0,87	0
W	Q	96 (CDR3)	5	0,59	1
W	Q	96 (CDR3)	7	0,59	1
W	E	96 (CDR3)	6	0,54	0
W	P	96 (CDR3)	8	0,20	4
G	A	94 (CDR3)	0	0,02	21
G	F	94 (CDR3)	2	-0,07	32
G	F	93 (CDR3)	2	-0,07	32
E	Q	96 (CDR3)	5	-0,09	6
G	Q	95 (CDR3)	5	-0,09	27
G	Q	98 (CDR3)	7	-0,09	27
G	Q	95 (CDR3)	7	-0,09	27
G	Q	94 (CDR3)	5	-0,09	27
G	Q	97 (CDR3)	7	-0,09	27
G	P	97 (CDR3)	8	-0,14	29
G	E	95 (CDR3)	6	-0,19	23

PDB ID 6RPA					
TCR aa	peptide aa	TCR residue	peptide residue	TCRen	count in database
N	W	96 (CDR3)	4	0,82	1
R	W	93 (CDR3)	4	0,59	1
S	T	97 (CDR3)	6	0,55	5
S	M	101 (CDR3)	3	0,46	3
S	M	97 (CDR3)	3	0,46	3
S	M	97 (CDR3)	3	0,46	3
D	I	101 (CDR3)	5	0,43	5
G	T	98 (CDR3)	6	0,41	11
G	T	99 (CDR3)	6	0,41	11
A	M	99 (CDR3)	3	0,35	1
G	S	98 (CDR3)	0	0,33	7
D	W	94 (CDR3)	4	0,14	4
Y	M	102 (CDR3)	3	0,11	8
S	W	97 (CDR3)	4	0,09	7
S	W	97 (CDR3)	4	0,09	7
G	W	98 (CDR3)	4	0,07	12
G	W	99 (CDR3)	4	0,07	12
G	M	100 (CDR3)	3	0,03	12
G	M	98 (CDR3)	3	0,03	12
T	Q	29 (CDR1)	7	-0,03	13
D	Q	101 (CDR3)	7	-0,03	12
S	I	97 (CDR3)	5	-0,08	16
G	Q	96 (CDR3)	7	-0,09	27
I	W	95 (CDR3)	4	-0,10	1
Y	L	102 (CDR3)	2	-0,12	25
G	I	98 (CDR3)	5	-0,14	26
D	T	101 (CDR3)	6	-0,15	11
Y	I	102 (CDR3)	5	-0,16	21
Y	W	102 (CDR3)	4	-0,38	17

Supplementary Table 4. Atchley factors

Atchley factor	Property	Higher values	Lower values
AF1	Polarity/hydrophobicity	Hydrophilicity	Hydrophobicity
AF2	Secondary structure	Bends, coils, turns	alpha-helical
AF3	Molecular volume	Small size	Large size
AF4	Codon diversity	Higher amino acid usage	Heat capacity, refractivity
AF5	Electric charge	Positively charged	Negatively charged

Supplementary Table 5. Contact map for TCR-peptide-MHC complex 6BJ3

PDB ID	TCR chain	TCR region	TCR residue index	TCR residue	peptide residue index	peptide residue
6BJ3	TRA	CDR3	94	G	3	T
6BJ3	TRA	CDR3	95	G	3	T
6BJ3	TRA	CDR3	95	G	4	E
6BJ3	TRA	CDR3	97	Q	4	E
6BJ3	TRB	CDR3	91	R	4	E
6BJ3	TRB	CDR3	92	T	4	E
6BJ3	TRB	CDR3	93	R	4	E
6BJ3	TRB	CDR3	94	G	4	E
6BJ3	TRB	CDR3	98	I	4	E
6BJ3	TRB	CDR3	93	R	5	E
6BJ3	TRB	CDR3	94	G	5	E
6BJ3	TRB	CDR3	93	R	6	A
6BJ3	TRB	CDR3	94	G	6	A
6BJ3	TRB	CDR3	95	G	6	A
6BJ3	TRB	CDR3	93	R	7	E

Supplementary Table 6. Candidate neoepitopes for 302TIL TCR. Column “TCRen” corresponds to the energy of TCR-peptide interaction estimated using TCRen based on the homology model of TCR-peptide-MHC complex. Cognate neoepitope KQWLVWLFL has the best TCRen value and marked by red color.

Gene	Mutation	Peptide	WT Peptide	Mutation position	HLA	TCRen
HHAT	L75F	KQWLVWLFL	KQWLVWLLL	8	A02:06	-1,697
NAB2	R51C	QLYRVLQCA	QLYRVLQRA	8	A02:06	-1,695
OR2L8	Q219R	SYGRVLFAV	SYGQVLFAV	4	A02:06	-0,978
SYN1	G530A	GQARQSRPV	GQGRQSRPV	3	A02:06	-0,197
OR2L8	F199L	LLSATIFLV	FLSATIFLV	1	A02:06	-0,058
PDE4D	H56P	RLLHPHHPL	RLLHPHHHL	8	A02:06	0,627
OR10G7	G20A	ALDAPLFGI	GLDAPLFGI	1	A02:06	0,628
SLC35G5	G46D	ALLDGGLPA	ALLGGGLPA	4	A02:06	1,235
HHAT	H78P	WLLLGPMVV	WLLLGHMVV	6	A02:06	1,609
HHAT	L75F	WLFLGHMVV	WLLLGHMVV	3	A02:06	1,611
NAB2	R51C	RVLQCANLL	RVLQRANLL	5	A02:06	2,08
OR2L8	F199L	WVYEGTVLL	WVYEGTVFL	8	A02:06	2,857
SLC35G5	G46D	LLDGGLPAG	LLGGGLPAG	3	A02:06	3,717
TTLL12	A25G	GQALAEFAA	AQALAEFAA	1	A02:06	4,193
FOXD4L5	Y145S	FISGRFPYS	FISGRFPYY	9	A02:06	4,268
SLAMF8	G96E	SEDSGNFSV	SGDSGNFSV	2	A02:06	4,991
OR2L8	Q219R	RVLFAVYHM	QVLFAVYHM	1	A02:06	4,995
OR2L8	F199L	VLLSATIFL	VFLSATIFL	2	A02:06	5,009
OR2L8	G212A	FLVFPFIAI	FLVFPFIGI	8	A02:06	5,039
OR2L8	G212A	AISCSYGQV	GISCSYGQV	1	A02:06	6,586

Supplementary Note 1

Eigenvalue decomposition of TCRen, MJ and Keskin matrices

To assess intrinsic regularity of TCRen potential, we performed eigenvalue decomposition of the corresponding matrix. As this method is applicable only to symmetric matrices, we performed eigenvalue decomposition of TCRen.s.cf matrix, which also has excellent performance in all benchmarks (**Table 1**) and thus may be considered as appropriate for representation of rules of TCR:peptide interactions. Eigenvalue decomposition of asymmetric TCRen.a.cf matrix can be performed after a simple transformation and its analysis results in qualitatively the same conclusions as reported further in this section (see **Supplementary Note 1**).

Eigenvalue decomposition allows to represent all entries of the matrix as:

$$e_{ij} = \sum_{\alpha=1}^{N} \lambda_{\alpha} V_{\alpha,i} V_{\alpha,j},$$

where λ_{α} are eigenvalues and $V_{\alpha,i}$ are elements of corresponding eigenvectors.

If some eigenvalues prevail and have higher absolute values compared to other eigenvalues, corresponding eigenvectors may be considered as reflecting some properties of residues that govern interaction (in case of TCRen — TCR-peptide recognition).

Eigenvalue spectrum of TCRen.s.cf matrix and correlation coefficients of eigenvectors and AFs are shown in **Supplementary Figure 8, left panel**. For comparison, eigenvalue decomposition was performed also for MJ and Keskin matrices (before analysis they were modified by subtracting mean values $< e_{ij} >$) (**Supplementary Figure 8, central and right panels**). In the case of MJ as little as two eigenvalues prevail, both with a strong negative correlation with polarity, that points to the hydrophobic nature of driving forces of protein folding [28]. The form of the eigenvalue spectrum of Keskin matrix is similar to that of MJ matrix, with two dominating eigenvectors negatively correlated with polarity. The notable difference between Keskin and MJ eigenvalue spectrums is that Keskin prevailing eigenvectors also correlate with AF2 (related to secondary structure preferences) and absolute values of other eigenvectors of Keskin are a bit higher compared to MJ matrix, that explains higher correlation of Keskin entries with AF2 of contacting amino acids (**Figure 6D**) compared to MJ. However, both entries and prevailed eigenvectors of Keskin have the highest correlation with hydrophobicity AF1, so the major driving force of general protein-protein interactions may be also considered as hydrophobic. Eigenvalue spectrum of TCRen.s.cf (**Figure 6E**) is much more complex than that of MJ and Keskin and a higher number of eigenvalues prevail. Top eigenvectors correlate with different properties of contacting residues: AF4 for eigenvector 1, AF1 for eigenvectors 2 and 3, AF5 for eigenvectors 18 and 19. This analysis suggests that different properties of contacting residues contribute to specificity of TCR-peptide recognition. Although hydrophobicity is important for this process (that is evident from correlation of TCRen.a.cf entries with products of AF1s of

contacting residues in **Figure 6B** and correlation of TCRen.s.cf eigenvectors 2 and 3 with AF1 in **Figure 6E**) it is not as dominant as in protein folding and protein-protein interaction and other properties (such as electric charge) are also important for TCR-peptide interaction.

As eigenvalue decomposition may be performed only for symmetric matrices, a simple transformation of 20×20TCRen.a.cf matrix was made: residues of the same type in TCR and peptide were considered as distinct entities that resulted in a symmetric matrix of size 40×40 with zero values in cells for which both dimensions correspond to residues from the same molecule (TCR or peptide). Eigenvalue spectrum of this 40×40 matrix is shown in **Figure SN1A**. To examine the relation between eigenvectors and properties of amino acids, correlation coefficients of each eigenvector with 5 AFs were computed separately for residues in TCR and peptide (**Figure SN1B-C)**. Top eigenvectors correlate with different properties of contacting residues and this pattern slightly differs for residues in TCR and peptide side (AF1, AF3, AF4 and AF5 for TCR residues; AF1, AF2 and AF4 for peptide residues; see **Figure SN1B-C**).

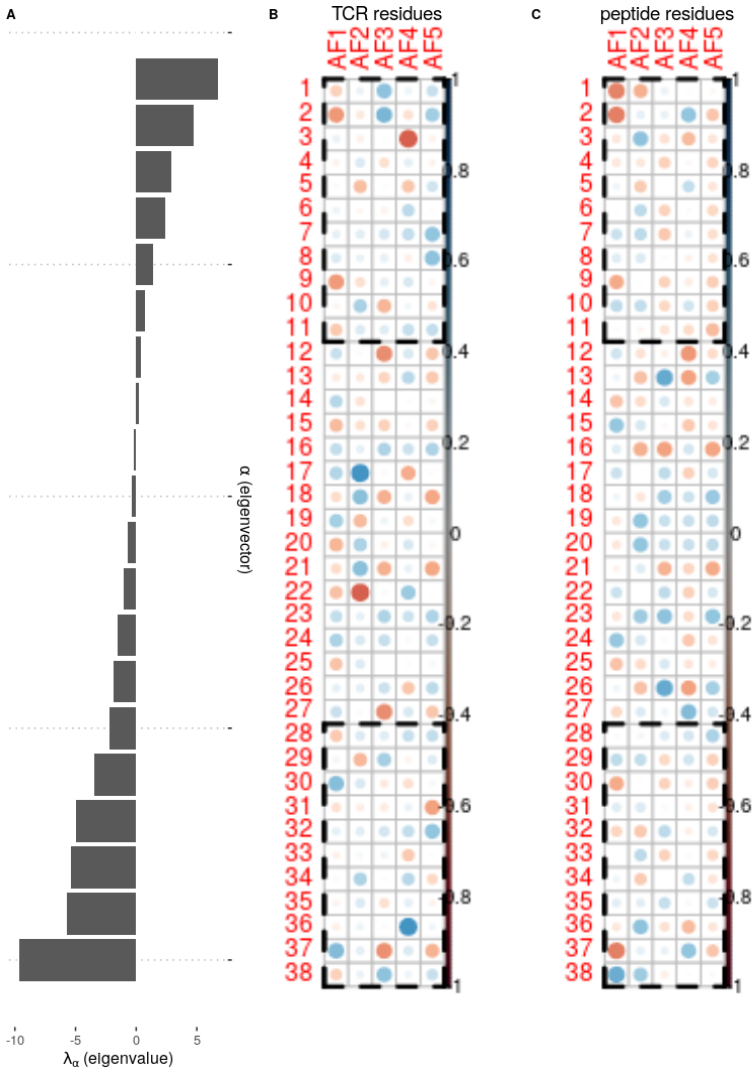


Figure SN1. A. The eigenvalue spectrum for TCRen.a.cf matrix. **B-C.** Correlation of TCRen matrix eigenvectors with AF of TCR (**B**) and peptide (**C**) residues.

References

1. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13139–13144 (2014).

2. Britanova, O. V. *et al.* Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014).

3. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

4. Glanville, J. *et al.* Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).

5. Dash, P. *et al.* Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).

6. Jensen, K. K. *et al.* TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. *Sci. Rep.* **9**, 14530 (2019).

7. Pierce, B. G. & Weng, Z. A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Sci.* **22**, 35–46 (2013).

8. Liu, I.-H., Lo, Y.-S. & Yang, J.-M. Genome-wide structural modelling of TCR-pMHC interactions. *BMC Genomics* **14 Suppl 5**, S5 (2013).

9. 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).

10. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

11. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–87 (2003).

12. Sewell, A. K. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).

13. Kastritis, P. L. *et al.* A structure-based benchmark for protein-protein binding affinity. *Protein Sci.* **20**, 482–491 (2011).

14. 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. U. S. A.* **105**, 16671–16676 (2008).

15. Singh, N. K. *et al.* Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. *J. Immunol.* **199**, 2203–2213 (2017).

16. Rossjohn, J. *et al.* T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).

17. Borrman, T. *et al.* ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins* **85**, 908–916 (2017).

18. Gee, M. H. *et al.* Antigen Identification for Orphan T Cell Receptors Expressed on Tumor-Infiltrating Lymphocytes. *Cell* **172**, 549–563.e16 (2018).

19. Karnaukhov, V. *et al.* HLA binding of self-peptides is biased towards proteins with specific molecular functions. *bioRxiv* 2021.02.16.431395 (2021) doi:10.1101/2021.02.16.431395.

20. Moont, G., Gabb, H. A. & Sternberg, M. J. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**, 364–373 (1999).

21. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**,

D202–5 (2008).

22. Cole, D. K. *et al.* Germ line-governed recognition of a cancer epitope by an immunodominant human T-cell receptor. *J. Biol. Chem.* **284**, 27281–27289 (2009).

23. Birnbaum, M. E. *et al.* Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).

24. Kumari, R., Kumar, R., Open Source Drug Discovery Consortium & Lynn, A. g_mmpbsa--a GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* **54**, 1951–1962 (2014).

25. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

26. Keskin, O., Bahar, I., Badretdinov, A. Y., Ptitsyn, O. B. & Jernigan, R. L. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.* **7**, 2578–2586 (1998).

27. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6395–6400 (2005).

28. Li, H., Tang, C. & Wingreen, N. S. Nature of Driving Force for Protein Folding: A Result From Analyzing the Statistical Potential. *Phys. Rev. Lett.* **79**, 765–768 (1997).

29. Peng, M. *et al.* Neoantigen vaccine: an emerging tumor immunotherapy. *Mol. Cancer* **18**, 128 (2019).

30. Melero, I. *et al.* Therapeutic vaccines for cancer: an overview of clinical trials. *Nat. Rev. Clin. Oncol.* **11**, 509–524 (2014).

31. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818–834.e13 (2020).

32. Bobisse, S. *et al.* Sensitive and frequent identification of high avidity neo-epitope specific CD8 + T cells in immunotherapy-naive ovarian cancer. *Nat. Commun.* **9**, 1092 (2018).

33. Devlin, J. R. *et al.* Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nat. Chem. Biol.* **16**, 1269–1276 (2020).

34. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* **36**, W509–12 (2008).

35. Bradley, P. & Thomas, P. G. Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. *Annu. Rev. Immunol.* **37**, 547–570 (2019).

36. De Simone, M., Rossetti, G. & Pagani, M. Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges. *Front. Immunol.* **9**, 1638 (2018).

37. Turchaninova, M. A. *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43**, 2507–2515 (2013).

38. Brenke, R. *et al.* Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* **28**, 2608–2614 (2012).

39. Riley, T. P. *et al.* T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability. *Nat. Chem. Biol.* **14**, 934–942 (2018).

40. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).