Macromolecular Rapid Communications
www.mrc-journal.de

A Computational Strategy for the Rapid Identification and Ranking of Patient-Specific T Cell Receptors Bound to Neoantigens

*Zachary A. Rollins, Matthew B. Curtis, Steven C. George, and Roland Faller**

T cell receptor (TCR) recognition of a peptide–major histocompatibility complex (pMHC) is crucial for adaptive immune response. The identification of therapeutically relevant TCR-pMHC protein pairs is a bottleneck in the implementation of TCR-based immunotherapies. The ability to computationally design TCRs to target a specific pMHC requires automated integration of next-generation sequencing, protein–protein structure prediction, molecular dynamics, and TCR ranking. A pipeline to evaluate patient-specific, sequence-based TCRs to a target pMHC is presented. Using the three most frequently expressed TCRs from 16 colorectal cancer patients, the protein–protein structure of the TCRs to the target CEA peptide–MHC is predicted using Modeller and ColabFold. TCR-pMHC structures are compared using automated equilibration and successive analysis. ColabFold generated configurations require an ≈2.5× reduction in equilibration time of TCR-pMHC structures compared to Modeller. The structural differences between Modeller and ColabFold are demonstrated by root mean square deviation (≈0.20 nm) between clusters of equilibrated configurations, which impact the number of hydrogen bonds and Lennard-Jones contacts between the TCR and pMHC. TCR ranking criteria that may prioritize TCRs for evaluation of in vitro immunogenicity are identified, and this ranking is validated by comparing to state-of-the-art machine learning-based methods trained to predict the probability of TCR-pMHC binding.

1. Introduction

Cytotoxic (CD8+) T cells are part of the adaptive immune system and eradicate potentially harmful cells—including cancer cells—by recognition of the peptide-major histocompatibility complex (pMHC) on target cells. Tumor-specific pMHCs are comprised of a peptide derived from a mutated and/or aberrantly expressed intracellular protein[1] presented to the cell membrane in a pocket formed by the MHC α and β chains.[2] The wide diversity of peptide-MHCs ($\approx 10^{6-12}$)[3] is matched by the even wider diversity of TCRs ($>10^{20-61}$)[4,5] through random V(D)J recombination of the hypervariable complementarity determining regions (CDRs). The function of the adaptive immune response ultimately depends on the ability to produce appropriate immunogenic TCRs (on-target) while minimizing response to self pMHCs (off-target effects).

Despite breakthrough clinical potential for TCR-T cell therapies in solid tumors,[6–10] the implementation is hindered by three central challenges: 1) identifying tumor-specific pMHC ligands; 2) matching immunogenic TCRs with identified pMHCs, and 3) minimizing off-target (side) effects.[11] Combining next generation sequencing and machine learning, significant advancements have been made to identify and rank tumor-specific pMHC ligands,[12–14] thus addressing the first challenge.

Addressing the second challenge has been difficult as the identification of patient-specific TCR repertoires has involved

Z. A. Rollins, R. Faller
Department of Chemical Engineering
University of California
Davis, 1 Shields Ave, Bainer Hall, Davis, CA 95616, USA
E-mail: roland.faller@ttu.edu

M. B. Curtis, S. C. George
Department of Biomedical Engineering
University of California
Davis, 451 E. Health Sciences Dr., GBSF 2303, Davis, CA 95616, USA

R. Faller
Department of Chemical Engineering
Texas Tech University
Lubbock, TX 79409, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/marc.202400225

DOI: 10.1002/marc.202400225

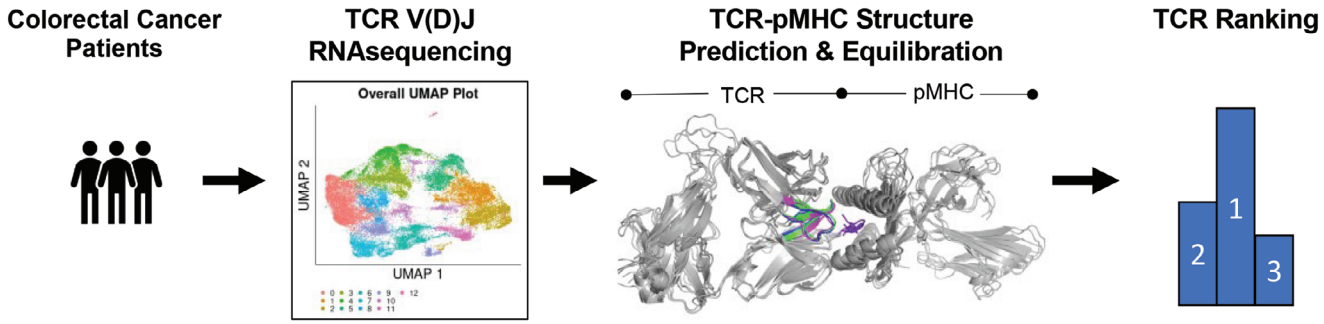


Figure 1. Process flow diagram for the protein–protein structure prediction of TCRs to a target pMHC. The process begins with single cell V(D)J RNA sequencing of the T cells from resected tumors of 16 colorectal cancer patients (left). Then, protein–protein structure prediction of TCRs sequenced from patients bound to a pMHC (HLA-A2) with a restricted target peptide $CEA_{571–579}$ (middle right) is performed. Finally, we run molecular dynamics simulations to equilibrate the structure and rank TCRs based on the number of interactions at equilibrium (right).

methods that are low-throughput or limited to a single chain.[15–17] However, recent breakthroughs in single-cell sequencing allow determination of the CDR3 regions of the α and β chain of the TCR in a high-throughput manner.[18–21] This technological breakthrough facilitates an unprecedented exploration of the vast TCR information space and allows the scientific community to refocus attention on fundamental questions related to recombination, maturation, and intersecting diversity of patient-specific TCR repertoires. This advance also provides an opportunity to leverage machine learning to predict TCR antigen binding specificity from primary amino acid sequence[22–25] or from structural features of TCR-pMHC homology models.[26] However, the training sets to characterize and rank TCRs by their immunogenicity are restricted by either insufficient data on the relevant TCR-pMHC binding parameters[27–32] or a limited number of known TCR-pMHC structures (≥645 on STCRDab).[33] Moreover, machine learning (ML) based methods to predict TCR-pMHC binding probability are strongly biased to the sequence training distributions and fail to generalize to unseen pMHC and TCR sequences.[34,35]

Despite significant advances in protein–protein structure prediction,[36–43] the prediction of TCRs bound to a target pMHC from patient-specific sequences is fundamentally biased to the features of known protein structures. Moreover, ranking TCRs is not possible without detailed information on the relationship between bond strength and immunogenic response.[27–32] Previously, we have identified several physiochemical parameters of the TCR-pMHC interaction that correspond with immunogenicity.[30,44] Herein, we present an automated pipeline to assess TCRs to a target pMHC (**Figure 1**). This pipeline begins with single-cell sequencing to identify the amino acids in the CDR3$\alpha\beta$ loops from T cells resected from the tumors of 16 colorectal cancer (CRC) patients.[20,21] Next, we restrict the carcinoembryonic (CEA) peptide ($CEA_{571–579}$:YLSGANLNL) to the MHC (HLA-0201), known to be expressed in CRC patients[11,45–47] and predict several TCR-pMHC complexes using TCRs sequenced from patients.[20,21] The predicted protein structures are equilibrated at physiological conditions by molecular dynamics simulations and an automated equilibration[48] is implemented to assess the starting structures from either Modeller[39,40] or the recently developed ColabFold[36–38] (Figure 1). Our results demonstrate that ColabFold creates structures that are ≈2.5X faster to equilibrate, and thus reduce overall computational cost compared to Modeller. However, the clusters of structures generated by Modeller and ColabFold are consistently divergent despite structural equilibration. Moreover, we provide potential criteria for ranking the TCRs after structural equilibration including the number of hydrogen bonds and Lennard-Jones contacts. This methodology is generally applicable to identify TCRs with relevant and quantifiable binding parameters to a target pMHC.

2. Experimental Section

2.1. Single-Cell RNA V(D)J Sequencing of CRC Patient T Cells

T lymphocyte single cell RNA-Seq data were made available to us from the Han group and has been previously published.[20,21] Raw data were first put through a quality control process to exclude cells with less than 200 unique genes, more than 7500 unique genes, and/or more than 10% mitochondrial gene expression. In addition, any genes that were present in fewer than three total cells were excluded from downstream analysis. All single-cell analysis was performed using the Seurat pipeline.[44] T lymphocytes were clustered using 0.3 as the value for the "resolution" parameter. Cytotoxic T lymphocyte clusters were identified by expression of *Cd3d* and *Cd8a*, and the absence of *Cd4* expression. TCR CDR3α and CDR3β sequences from the 10X Genomics 5′ VDJ analysis pipeline were matched to their corresponding cells for downstream analysis. After the segregation of CD3D+CD4-CD8A+ T cells, the top 3 most frequent TCR clonotypes were identified (Figure S8, Supporting Information).

2.2. TCR-pMHC Protein–Protein Structure Prediction

To demonstrate the feasibility of the proposed pipeline, the starting structures for the three most common TCRs were generated independently using Modeller V10.1[39,40] and ColabFold V 1.2.0[36–38] denoted TCR1, TCR2, and TCR3, respectively. Importantly, this methodology might benefit from recent and future models that fine-tune structure prediction methods on TCR-pMHC structure databases.[42,43] This benefit

Table 1. Rank of TCRs binding to $CEA_{571–579}$ pMHC.

TCR rank method	TCR1	TCR2	TCR3
ERGO-II-AE-VDJdb[22]	2	1	3
ERGO-II-LSTM-McPAS[22]	3	1	2
NetTCR-2.2[23]	2	3	1
pMTNet[24]	3	1	2
pMTnet-Omni[25]	3	1	2
ML – average rank	***2.6***	***1.4***	***2.0***
MD-H-bonds (Modeller)	2	1	3
MD-LJ-contacts (Modeller)	3	1	2
MD-H-bonds (ColabFold)	2	3	1
MD-LJ-contacts (ColabFold	2	1	3
MD – average rank	***2.25***	***1.5***	***2.25***

Note: This includes the rank determined by the probability of binding from numerous machine learning (ML) based methods (top). The rank based on molecular dynamics (MD) interactions is also provided (bottom). The average rank is *italicized* under the respective ranking methodology.

would likely reduce the required molecular dynamics simulation time to equilibrate approximated structures. The primary amino sequence used for multiple sequence alignment was derived from the DMF5 TCR bound to the HLA-A2 (MHC) restricted MART1 (PDB:3QDJ).[49] For sequence alignment, the CDR3α (CAVNFGGGKLIF), CDR3β (CASSLSFGTEAFF), and MART1 peptide (AAGIGILTV) were substituted with the respective CDR3 loops found from patient TCR clonotypes (**Table 1**) and the $CEA_{571–579}$ peptide (YLSGANLNL) known to be restricted to the HLA-A2 (MHC). The TCR CDR3s were substituted as follows: TCR1 (CDR3α: CAVNGDDYKLSF, CDR3β: CASRKRDDSEQYF), TCR2 (CDR3α: CAVSDNARLMF, CDR3β: CASSPFGGGNEQFF), and TCR3 (CDR3α: CAYRISAYDKVIF, CDR3β: CASSQTGGADTDTQYF). For Modeller, the MART1 (PDB:3QDJ) crystal structure was used as the template, ten model structures were generated from the alignment of the respective TCR, and the structure with lowest DOPE score[50] was selected for MD equilibration. ColabFold[38] is, in part, a server that performs rapid MSA/homology search combined with the trained network architecture of AlphaFold2[36,37] for prediction of the 3D atomic coordinates of folded protein structures. For ColabFold, five model structures were generated from the alignment of the respective TCR, and the structure with the highest pTMscore[36,37] was selected for MD equilibration. Multiple structures were generated from both Modeller and ColabFold to maintain best practice at producing the most accurate starting structure as described in their methods.[38,50]

2.3. Molecular Dynamics: Setup, Energy Minimization, and Equilibration

The predicted Modeller or ColabFold structures were used as starting configurations for a seven-step molecular dynamics pipeline to determine their equilibrated structures at physiological conditions. All MD Simulations were performed in full atomistic detail with Gromacs 2019.1[51,52] using the CHARMM22 with CMAP force field[53] in orthorhombic periodic boundary conditions. The force field was chosen to be consistent with earlier studies.[30,44] For the particular question under study here the exact choice of force field is not very relevant. 1) The residue protonation states were determined by calculating pKa values using propka3.1[54,55] and deprotonated if pKa values are below pH 7.4. 2) The properly protonated or deprotonated protein structures were solvated in orthorhombic water boxes large enough to satisfy minimum image convention using the TIP3P water model.[56] 3) Na^+ and Cl^- ions were added to reach salt concentration $\approx 150 \times 10^{-3}$ M and neutral charge. Box sizes were 10.627 × 7.973 × 10.685 nm with ≈48 000 water molecules, ≈300 ions, and ≈157 000 total atoms. Full specifications can be found in the Dryad repository:[57] 4) To avoid steric clashes, steepest descent energy minimization (emtol = 1000 kJ mol^{-1} nm^{-1}) was performed. 5) To relax solute–solvent contacts, a 100 ps simulation was run in the constant volume ensemble (*NVT*) with 0.2 fs timestep ($T = 310$ K). Temperature was maintained by coupling protein and nonprotein atoms to separate baths using a velocity rescale thermostat[58] with a 0.1 ps time constant. 6) To maintain pressure at 1.0 bar, a 100 ps simulation was run in the constant pressure (*NPT*) ensemble using isotropic Berendsen pressure coupling,[59] a 2.0 ps time constant, and 2 fs timestep. Steps (5) and (6) used position restraints (harmonic force constant = 1000 kJ mol^{-1} nm^{-2}) on all protein atoms. 7) Equilibration MD simulations were conducted for 100–300 ns with no restraints. Equilibration runs were extended in 50 ns increments until the root mean square deviation (RMSD) of the TCR-pMHC complex was in equilibrium for a minimum of 50 ns determined by the variance-bias trade-off algorithm.[48] Such equilibration lengths were sufficient for this problem, but additional longer term conformational changes in atomistic simulations were cannot be excluded. To maintain temperature and pressure during the production runs, the Nosé–Hoover thermostat[60] and Parrinello–Rahman barostat[61] were used with time constants 2.0 and 1.0 ps, respectively. The isothermal compressibility of water was used as 4.5×10^{-5} bar^{-1}. Simulations used the particle Ewald mesh algorithm[62,63] for long-range electrostatic calculations with cubic interpolation and 0.12 nm grid spacing. Short-range nonbonded interactions were cut off at 1.2 nm. All water bond lengths were constrained with SETTLE,[64] and all other bond lengths were constrained using the LINCS algorithm.[65] The leap-frog algorithm was used for integrating equations of motion with a 2 fs time step.

2.4. Data and Statistical Analysis

Selected TCR-pMHC structures from MD trajectories were visualized using the Pymol v2.4.0 Molecular Graphics System (Schrodinger, LLC; New York, NY). The selected frames for visualization were chosen to be from the top three clusters (two from each cluster) after TCR-pMHC structure equilibration (Figures S1 and S2, Supporting Information). This resulted in a total of 12 structures for each TCR: 6 from Modeller (TCR1: Cluster 1, 2, & 7, TCR2: Cluster 1, 9, & 12, and TCR3: Cluster 1, 6, & 9) and 6 from ColabFold (TCR1: Cluster 1, 2, & 4, TCR2: Cluster 1, 3, & 4 and TCR3: Cluster 1, 3, & 6) (Figures S3–S5, Supporting Information). The clusters were selected because they were after the simulation time required for equilibration (e.g., for Modeller TCR1, Clusters 1, 2, & 7 were chosen because Clus-

ters 3–6 were only dominant during the 249 ns equilibration time). The all-to-all alignment of TCR-pMHC structures was performed in Pymol using the align command on the C^{α} atoms to compute RMSD between pairs of structures. Data analysis from MD simulations was performed with tools from the Gromacs suite:[51,52] gmx make_ndx, gmx hbond, gmx rms, gmx rmsf, and gmx cluster. These results were complemented with a secondary analysis utilizing python packages for data handling and visualization including: numpy,[66] pandas,[67] matplotlib,[68] GromacsWrapper,[69] scipy,[70] and pingouin,[71] and pymbar.[48] Custom bash shell python scripts relevant to the production of figures were deposited in a GitHub repository.[72] The geometry of a Lennard-Jones contact is defined as a distance of less than 0.35 nm between atoms. Results were presented as mean ± SEM. As indicated in figures, statistics were performed in python using scipy for one-way analysis of variance (ANOVA), and pingouin for pairwise Tukey-HSD post hoc tests. Detailed outputs of statistical analysis were written to excel and are provided in a Dryad repository.[57]

2.5. Machine Learning Based TCR-pMHC Binding Predictions

The TCR ranking method based on physical interactions was compared with the pMHC to numerous state-of-the-art machine learning methods. Recent machine learning methods were typically trained on sequence representations of the TCR-pMHC and their binary binding label. The positive binding label data were derived from 10X Genomics sequencing datasets after redundancy reductions.[73] The dataset sizes vary depending on the methodology used for redundancy reductions as well as the resolution capability of the sequencing dataset: CDR3β-peptide, CDR3α-CDR3β-peptide, or CDR1α-CDR2α-CDR3α-CDR1β-CDR2β-CDR3β-peptide. The negative binding label data were usually generated assuming that known TCR-peptide binders do not cross react with additional peptides. For more details, the authors referred to the methods of the machine learning based method used to predict the TCR-pMHC binding probability of TCR1, TCR2, and TCR3 to the $CEA_{571–579}$ peptide: ERGO-II-AE-VDJdb,[22] ERGO-II-LSTM-McPAS,[22] NetTCR-2.2,[23] pMTNet,[24] and pMTNet-Omni.[25] For each method, the instructions on the respective GitHub repository were followed, and only the sequence information used to train the model was provided. The TCR-pMHC pairs were then ranked by the predicted binding probability (Table 1). Results from the machine learning based model predictions are made available on the GitHub repository:[72] https://github.com/zrollins/TCR_homology.git.

3. Results

To design TCRs to target the $CEA_{571–579}$ peptide restricted to the HLA-A2 (MHC), TCR clonotypes were identified utilizing the single cell RNA V(D)J sequenced T cells resected from colorectal tumors of 16 CRC patients.[20,21] This technique identifies TCR clonotypes by matching the amino acids from the CDR3 regions of the α and β chain of the TCR.[18–21] First, T cells were identified by the expression of *Cd3d* and *Cd8a* (and the absence of *Cd4* expression), as only CD8+ T cells can bind to the HLA-A2 (MHC). Of the 37931 T cells analyzed (**Figure 2**Ai), there were 9709 *Cd3d+/Cd4−/Cd8a+* T cells (corresponding to clusters 2, 6, 7, 9, and 11; Figure 2Aii–iv), and 3931 identified *Cd3d+/Cd4−/Cd8a+* TCR$\alpha\beta$ clonotypes (Figure 2B and Table 1). The three most frequently identified TCRs (Table 1) from patient tumors (denoted clonotype TCR1, TCR2, and TCR3) were then used to predict TCR-pMHC structures.

To determine the best method for generating TCR-pMHC starting configurations, we generated the configurations by either Modeller[39,40] or ColabFold.[36–38] The structures generated in Modeller utilized the DMF5 TCR bound to the HLA-A2 (MHC) restricted MART1 (PDB:3QDJ)[49] as template structure, and the ColabFold structures were predicted from trained neural networks.[36–38] The resulting starting configurations were then solvated in all-atom molecular dynamics simulations at physiological conditions (see the Experimental Section) for 150–300 ns to equilibrate the protein structures. Equilibration is indicated by the flattening of the RMSD from the initial configuration with fluctuations less than 0.2 nm for the entire TCR-pMHC structure (Figure S1, Supporting Information). A bias-variance tradeoff algorithm[48] was used to automate the detection of the equilibrated TCR-pMHC structure (Figure S1, Supporting Information). The equilibration time required for TCR1 (249 & 9 ns), TCR2 (127 & 61 ns), and TCR3 (149 & 136 ns) from Modeller and ColabFold, respectively, demonstrates ≈61% reduction in computational cost—on average—using ColabFold.

We next evaluated the structural similarity of the TCR-pMHC structures throughout the equilibration using the GROMOS clustering algorithm (Figure S2, Supporting Information).[74] Using a C^{α} RMSD cutoff of 0.2 nm, the top ten equilibrated clusters contain most of the configurations for TCR1 (86.8% and 98.3%), TCR2 (92.7% and 97.0%), and TCR3 (71.0% and 97.7%) from Modeller and ColabFold, respectively. Interestingly, the top cluster contains a plurality of structures and occurs after the estimated equilibration time indicating not a single, but a set of converged TCR-pMHC structures (Figure S2, Supporting Information).

To evaluate the structural similarity at the TCR-pMHC interface of TCR1, TCR2, and TCR3, we selected and aligned a subset of the equilibrated structures (12 structures—6 Modeller + 6 ColabFold—see the Experimental Section) (**Figure 3**A–C). An all-to-all structural alignment (72 unique comparisons) was performed to calculate the pairwise RMSD (after equilibration) within and between TCR-pMHC structures generated by Modeller and ColabFold. The average RMSD for TCRs generated within either Modeller or ColabFold is consistent for TCR1-pMHC (0.19 ± 0.04 nm), TCR2-pMHC (0.20 ± 0.07 nm), and TCR3-pMHC (0.19 ± 0.05 nm). However, there is a consistent increase in average RMSD when comparing equilibrated structures between Modeller and ColabFold for TCR1-pMHC (0.41 ± 0.05 nm), TCR2-pMHC (0.40 ± 0.05 nm), and TCR3-pMHC (0.33 ± 0.04 nm) (Figures S3–S5, Supporting Information). The increase in RMSD occurs despite selecting TCR-pMHCs from distinct equilibration clusters (Figure S2, Supporting Information). The increase in configuration dissimilarity between Modeller and ColabFold at the TCR-pMHC interface can be visualized by the aligned and overlaid structures (Figure 3A–C). To investigate the relative fluctuations of the substructures at the TCR-pMHC interface, the root mean square fluctuations were calcu-

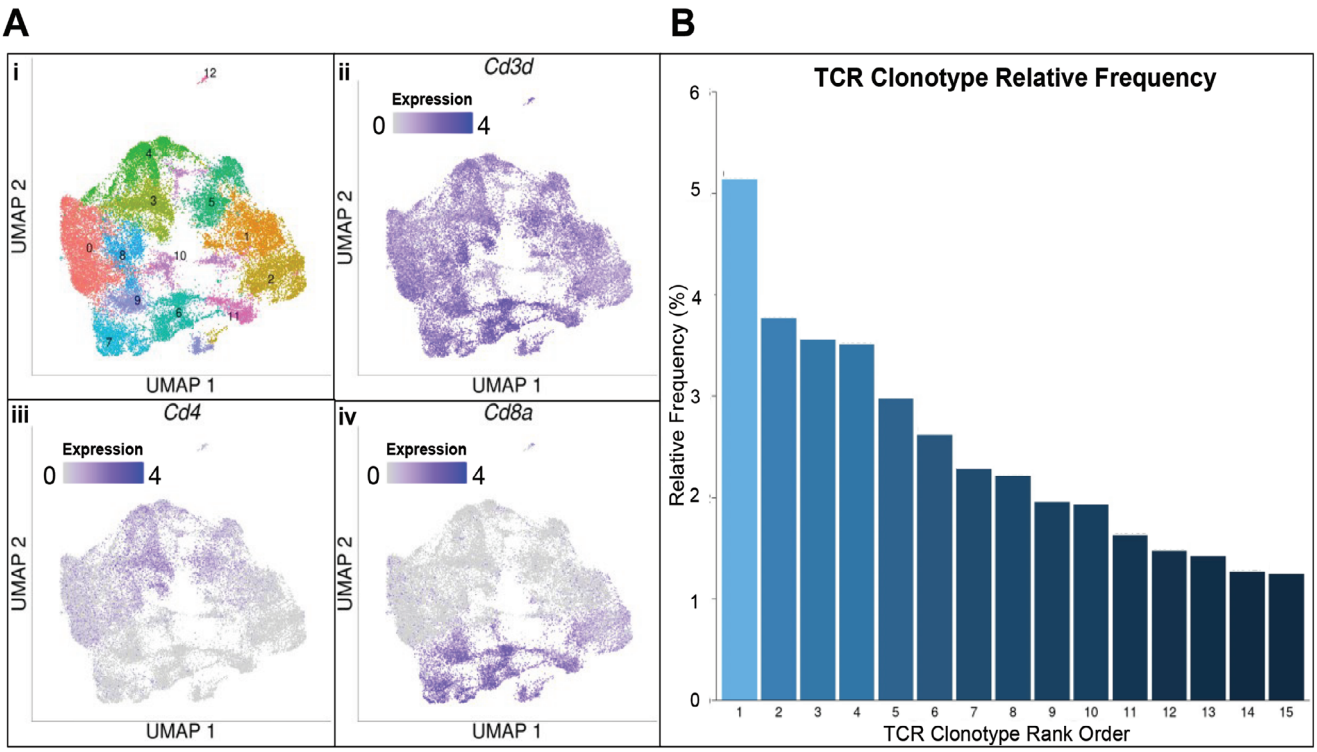


Figure 2. Identified TCR$\alpha\beta$ clonotypes from CRC patient tumors. UMAP projection of T cell gene expression data from Han et al. include the A) i) total number of unsupervised clusters, ii) distribution of *Cd3d* expression, iii) distribution of *Cd4* expression, and iv) distribution of *Cd8a* expression across the dataset. TCR$\alpha\beta$ clonotypes were identified from the subset of T cells with high *Cd8a* expression and the relative frequency B) of those *Cd8a*$^+$ clonotypes. Single cell data from refs. [20, 21].

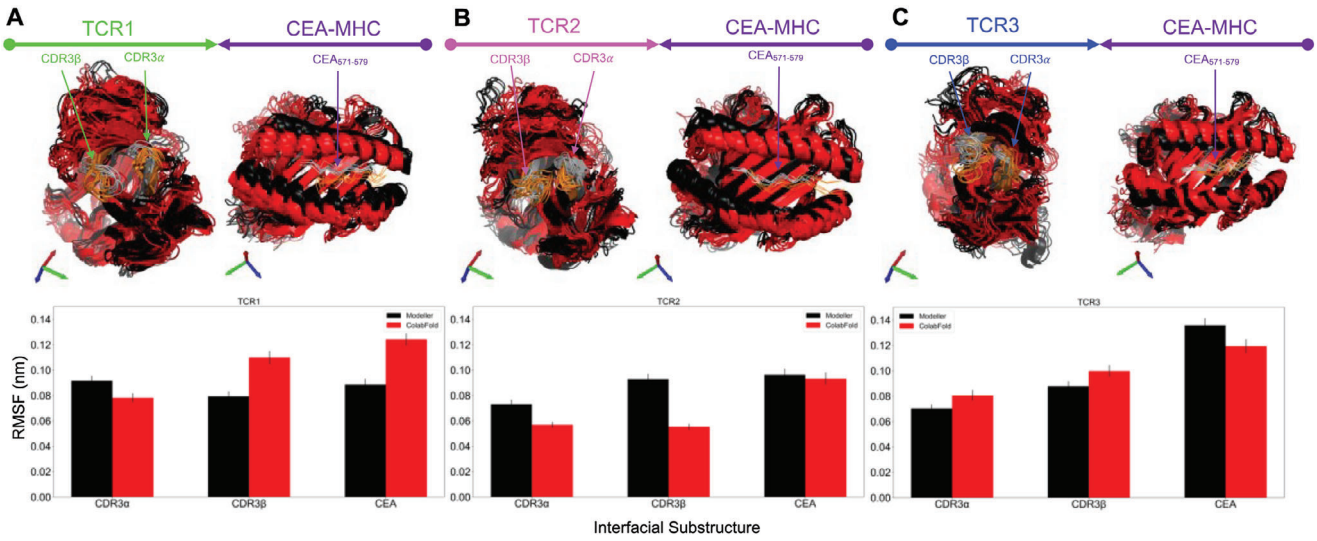
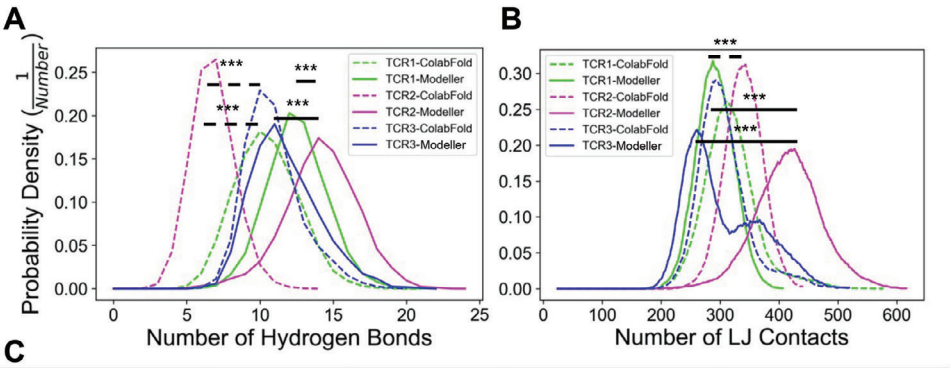


Figure 3. Equilibrated patient-specific TCRs bound to $CEA_{571–579}$ pMHC. The structures of the three most frequently found TCRs from 16 CRC patients were predicted (using Modeller or ColabFold) and equilibrated at physiological conditions using molecular dynamics simulations. A–C) The most frequent TCRs: TCR1, TCR2, and TCR3 are displayed with colors green, magenta, and blue, respectively. The starting structures were created from Modeller (black) and ColabFold (red) and 12 TCR-pMHC structures (6 from Modeller + 6 from ColabFold) are aligned after equilibration with the TCR (on the left, in respective color) and pMHC (on the right, in purple). In addition, the mutated substructures of the TCR and pMHC are indicated by arrows and highlighted in the following colors: CDR3α, CDR3β, and $CEA_{571–579}$ (Modeller: gray and ColabFold: orange). After structural equilibration, the root mean square fluctuation (RMSF) for each TCR is calculated for the regions that were mutated: CDR3α, CDR3β, and $CEA_{571–579}$ (bottom). The RMSF is calculated for both the Modeller (black) and ColabFold (red) generated starting structures.



C

TCR	Freq. Rank	Rel. Freq.	TCR-pMHC H-Bonds	TCR-pMHC LJ-Contacts
TCR1- ColabFold	1	0.051	10.5 ± 2.2	306 ± 45
TCR1 - Modeller	1	0.051	12.5 ± 2.0	279 ± 31
TCR2 - ColabFold	2	0.038	6.77 ± 1.5	328 ± 31
TCR2 - Modeller	2	0.038	14.2 ± 2.5	404 ± 56
TCR3 - ColabFold	3	0.036	10.9 ± 2.0	294 ± 43
TCR3 - Modeller	3	0.036	11.7 ± 2.3	294 ± 64

Figure 4. Interactions of patient specific TCRs with pMHC in equilibrium. A) The probability density for the number of hydrogen bonds between the pMHC and TCR1 (green), TCR2 (magenta), and TCR3 (blue), respectively. B) The probability density for the number of Lennard-Jones Contacts between the pMHC and TCR1 (green), TCR2 (magenta), and TCR3 (blue), respectively. The interaction distributions after equilibration are separated for Modeller (solid line) and ColabFold (dashed line). C) The expected value and standard deviation of H-bonds and LJ-contacts with the TCRs ranked by relative frequency found in CRC patients. The number of interactions throughout the simulations was statistically compared: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$ by one-way ANOVA followed by Tukey-HSD post hoc test. Statistical significance was only displayed for comparisons with Cohen effect size $d > 0.5$. Significance was displayed by solid and dashed lines for Modeller and ColabFold comparisons, respectively.

lated after equilibration for CDR3α, CDR3β, and the CEA peptide (Figure 3A–C). Fluctuations for all TCRs and TCR substructures are approximately 0.10 nm.

After structural equilibration, the number of molecular-level interactions between the TCRs and pMHC were evaluated to assess potential differences between the TCR-pMHCs complexes, and thus provide insight into potential methods to rank the TCRs (**Figure 4**). We selected hydrogen bonds (H-bonds) and Lennard-Jones contacts (LJ-contacts) between the TCRs and pMHC to understand the relative importance of coulombic and hydrophobic interactions. The number of hydrogen bonds (Figure S6, Supporting Information) and the number of Lennard-Jones contacts (Figure S7, Supporting Information) were calculated as a function of simulation time, and these plots were used to calculate the probability densities. The probability density of interactions is an index to describe the relative likelihood of interactions that occur at any timepoint during the equilibration. Our results demonstrate that TCR2 (expected value: 14.2) is significantly more likely to have more hydrogen bonds than TCR1 (expected value: 12.5) and TCR3 (expected value: 11.7) for structures generated by Modeller (Figure 4). In contrast, for structures generated by ColabFold, TCR1 (expected value: 10.5) and TCR3 (expected value: 10.9) are significantly more likely to have more hydrogen bonds than TCR2 (expected value: 6.8) (Figure 4A). In addition, TCR2 (expected value: 404) is more likely to have more Lennard-Jones contacts than TCR1 (expected value: 279) and TCR3 (expected value: 294) for structures generated by Modeller. Consistently, for structures generated by ColabFold, TCR2 (expected value: 328) is more likely to have more Lennard-Jones contacts than TCR3 (expected value: 294) and TCR1 (expected value: 306) (Figure 4B). Despite TCR1 being detected more frequently in patient tumors (Figure 2B), TCR2 may have more binding interactions to $CEA_{571–579}$ pMHC at equilibrium (Figure 4C).

To get a composite score for the MD based TCR rank, we categorized TCR-pMHC pairs as higher probability of binding the $CEA_{571–579}$ pMHC if they had more interactions at equilibrium: more H-bonds or LJ-contacts for both the Modeller and ColabFold based starting structures (Table 1). We found that, on average, TCR2 has more physical interactions with the based $CEA_{571–579}$ pMHC with a MD rank of 1.5. In addition, TCR1 and TCR3 had the same MD rank of 2.25. Interestingly, TCR2 was also predicted to have the highest probability of binding the $CEA_{571–579}$ pMHC in 4/5 machine learning based methods (Table 1) and achieved an average ML rank of 1.4. In addition, TCR2 and TCR3 resulted in a ML rank of 2.6 and 2.0, respectively.

4. Discussion

Using a combination of single cell sequencing of T cells derived from patient tumors, protein structure prediction algorithms,

and MD simulations, we present a pipeline to rank TCRs based on their molecular level interactions with a target pMHC at equilibrium. To commence a pipeline to assess TCRs in silico, we chose only the three most frequently expressed clonotypes in 16 patients with colorectal cancer as a case study. Although the selection of the most frequent TCRs from the clonal pool present in the tumor microenvironment is somewhat arbitrary, the relatively expanded clonal pool in the tumor microenvironment is more likely to be immunogenic to the tumor than a random selection of low-frequency clones. Nonetheless, our pipeline is easily adapted to selecting clonotypes from alternate sources (e.g., peripheral blood), or alternate strategies (e.g., selecting from the entire clonal pool in the tumor microenvironment.

Our study also presented an opportunity to assess two fundamentally different protein predictions tools: ColabFold—a recently released trained deep learning network; and Modeller—a traditional template-based protein-prediction model. We found that the MD equilibration of 3D atomic coordinates of TCR-pMHC structures was ≈2.5× faster using ColabFold generated structures (Figure S1, Supporting Information). This finding is based on a small number of protein–protein structures, and there was significant variation. Nonetheless, our results demonstrate that ColabFold may be a superior computational tool for protein structure prediction, and thus may have implications on the scale-up of assessing a larger set (i.e., thousands) of TCRs generated from sequencing data. Moreover, the compute cost may be further reduced by utilizing recently fine-tuned structure prediction methods on TCR-pMHC databases[43] which may reduce the required simulation time to equilibrate approximated protein-protein structures. A reason why ColabFold is faster in equilibration might be that it uses a multitude of templates and not only one.

To automate and remove human bias from determining the required simulation time to reach an equilibrated TCR-pMHC structure, we used a variance-bias trade-off algorithm[48] and required a minimum equilibration time of 50 ns (Figure S1, Supporting Information). In addition, we performed a cluster analysis to identify the set of converged clusters after equilibration (Figure S2, Supporting Information). Moreover, we found that the root-mean-square fluctuations after equilibration for CDR3α, CDR3β, and the CEA peptide (Figure 3A–C) were ≈0.10 nm. The fluctuations for CDR3α, CDR3β, and peptide are consistent with equilibrated TCRs with a known crystal structure,[30] and thus consistent with equilibration.

After equilibration, we assessed several clusters of configurations and found that within a protein-protein structure predictor (i.e., Modeller or ColabFold) there is a pairwise RMSD of ≈0.20 nm. Interestingly, across structure predictors there is an increase in pairwise RMSD ≈0.40 nm (Figures S3–S5, Supporting Information). This trend is consistent for TCR1, TCR2, and TCR3 indicating that the structure prediction method can influence the set of equilibrated TCR-pMHC configurations, and molecular level interactions. We found that TCR2 had more hydrogen bonds compared to TCR1 and TCR3 when generated by Modeller, but less hydrogen bonds when generated by ColabFold (Figure 4A). These results indicate that the differences in configurations generated by Modeller and ColabFold can also influence the number of molecular level interactions between the TCR and pMHC. We observed consistent results between Modeller and ColabFold for the number Lennard-Jones contacts across the three TCRs (Figure 4B). The probability density of hydrogen bonds and Lennard-Jones contacts may provide a rudimentary criterion to rank TCRs based on their relative strength of interaction[30] at equilibrium. For example, TCR2 may be a more ideal target to the $CEA_{571–579}$ pMHC because of the consistent increase in Lennard-Jones contacts.

To develop a composite TCR ranking index, we summarized the results found from the MD simulations (Table 1). After ranking the TCRs based on the number of interactions with the $CEA_{571–579}$ pMHC, we found that TCR2 had the highest average rank of 1.5. This was consistent with machine learning based methods that predict the probability of TCR-pMHC binding (Table 1). In fact, TCR2 had the highest probability of binding in 4/5 methods and achieved a similar average rank of 1.4 across the methodologies. Although these results are based on a small set of TCRs, this demonstrates that the MD rank of TCRs derived from molecular interactions at equilibrium is consistent with state-of-the-art ML methods.[22–25] Moreover, a major limitation in ML methods is difficulty at predicting TCR-pMHC binding pairs that are not included in the training distribution.[34,35] This methodology may be better suited for generalization to disparate TCR-pMHC pairs because the ranking is derived from physical interactions. Future work will require a comprehensive dataset that will assess the physiochemical properties of TCR-pMHC interactions that determine immunogenicity. Also, extensive experimental validation would be very useful.

5. Conclusions

The identification of tumor-specific TCRs will be augmented by computational methodologies that accurately rank TCRs based on the immunogenic response to a target pMHC. We have integrated next-generation sequencing with protein-protein structure prediction and MD to introduce a potential pipeline to evaluate TCRs. We found that ColabFold outperforms Modeller (≈2.5×) in the required simulation time to generate equilibrated TCR-pMHC structures, and thus may be a superior computational tool to utilize in a computational algorithm built to predict TCR immunogencity. In addition, the protein structure prediction method influences the set of equilibrated configurations and the number of interactions between the TCR and pMHC, and thus may impact the accuracy of predicting TCR-pMHC bond strength or immunogenic response. On average, the MD-based ranking of TCR-pMHC pairs was consistent with state-of-the-art ML based ranking methods and may provide an additional benefit of generalizability to unseen TCR-pMHC pairs.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Single cell V(D)J and single cell gene expression RNA sequencing data was a generous gift from Prof. Arnold Han at Columbia University. Simulations were performed on the hpc1/hpc2 clusters at UC Davis. This work was supported in part by startup funding to SCG from the Department of Biomedical Engineering.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

S.C.G. and R.F. contributed equally to this work. Z.A.R. performed the simulations, analyzed, and interpreted the data, and wrote the manuscript. M.B.C. analyzed and interpreted the scRNA-Seq data and wrote the manuscript. R.F. designed the experiments, analyzed and interpreted the data, wrote the manuscript, and secured computer time. S.C.G. designed the experiments, analyzed, and interpreted the data, wrote the manuscript, and secured the funding.

Data Availability Statement

TCR-pMHC structures generated from protein–protein structure predictors have been made available ArrayExpress #E-MTAB-9455. In addition, the structures, box sizes, and atom counts are all deposited in a Dryad repository:[57] https://doi.org/10.25338/B83S70. All scripts relevant to the production of figures have been made available on GitHub:[72] https://github.com/zrollins/TCR_homology.git.

Keywords

molecular dynamics, protein structure, T cells

Received: April 11, 2024
Revised: June 2, 2024
Published online: June 12, 2024

[1] M. P. Weekes, R. Antrobus, J. R. Lill, L. M. Duncan, S. Hör, P. J. Lehner, *J. Biomol. Tech.* **2010**, *21*, 108.
[2] K. Murphy, C. Weaver, *Janeway's Immunbiology*, Garland Science, New York **2017**.
[3] K. L. Rock, E. Reits, J. Neefjes, *Trends Immunol.* **2016**, *37*, 724.
[4] P. C. de Greef, T. Oakes, B. Gerritsen, M. Ismail, J. M. Heather, R. Hermsen, B. Chain, R. J. de Boer, *Elife* **2020**, *9*, 49900.
[5] V. I. Zarnitsyna, B. D. Evavold, L. N. Schoettle, J. N. Blattman, R. Antia, *Front. Immunol.* **2013**, *4*, 485.
[6] L. A. Johnson, R. A. Morgan, M. E. Dudley, L. Cassard, J. C. Yang, M. S. Hughes, U. S. Kammula, R. E. Royal, R. M. Sherry, J. R. Wunderlich, C.-C. R. Lee, N. P. Restifo, S. L. Schwarz, A. P. Cogdill, R. J. Bishop, H. Kim, C. C. Brewer, S. F. Rudy, C. Van Waes, J. L. Davis, A. Mathur, R. T. Ripley, D. A. Nathan, C. M. Laurencot, S. A. Rosenberg, *Blood* **2009**, *114*, 535.
[7] G. P. Linette, E. A. Stadtmauer, M. V. Maus, A. P. Rapoport, B. L. Levine, L. Emery, L. Litzky, A. Bagg, B. M. Carreno, P. J. Cimino, G. K. Binder-Scholl, D. P. Smethurst, A. B. Gerry, N. J. Pumphrey, A. D. Bennett, J. E. Brewer, J. Dukes, J. Harper, H. K. Tayton-Martin, B. K. Jakobsen, N. J. Hassan, M. Kalos, C. H. June, *Blood* **2013**, *122*, 863.
[8] T. Moore, C. R. Wagner, G. M. Scurti, K. A. Hutchens, C. Godellas, A. L. Clark, E. M. Kolawole, L. M. Hellman, N. K. Singh, F. A. Huyke, S.-Y. Wang, K. M. Calabrese, H. D. Embree, R. Orentas, K. Shirai, E. Dellacecca, E. Garrett-Mayer, M. Li, J. M. Eby, P. J. Stiff, B. D. Evavold, B. M. Baker, I. C. Le Poole, B. Dropulic, J. I. Clark, M. I. Nishimura, *Cancer Immunol. Immunother.* **2018**, *67*, 311.
[9] R. A. Morgan, M. E. Dudley, J. R. Wunderlich, M. S. Hughes, J. C. Yang, R. M. Sherry, R. E. Royal, S. L. Topalian, U. S. Kammula, N. P. Restifo, Z. Zheng, A. Nahvi, C. R. de Vries, L. J. Rogers-Freezer, S. A. Mavroukakis, S. A. Rosenberg, *Science* **2006**, *314*, 126.
[10] P. F. Robbins, R. A. Morgan, S. A. Feldman, J. C. Yang, R. M. Sherry, M. E. Dudley, J. R. Wunderlich, A. V. Nahvi, L. J. Helman, C. L. Mackall, U. S. Kammula, M. S. Hughes, N. P. Restifo, M. Raffeld, C.-C. R. Lee, C. L. Levy, Y. F. Li, M. El-Gamil, S. L. Schwarz, C. Laurencot, S. A. Rosenberg, *J. Clin. Oncol.* **2011**, *29*, 917.
[11] Q. He, X. Jiang, X. Zhou, J. Weng, *J. Hematol. Oncol.* **2019**, *12*, 139.
[12] M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund, S. Buus, *PLoS One* **2007**, *2*, e796.
[13] Z. Koşaloğlu-Yalçın, M. Lanka, A. Frentzen, A. L. R. Premlal, J. Sidney, K. Vaughan, J. Greenbaum, P. Robbins, J. Gartner, A. Sette, B. Peters, *Oncoimmunology* **2018**, *7*, e1492508.
[14] J. J. Gartner, M. R. Parkhurst, A. Gros, E. Tran, M. S. Jafferji, A. Copeland, K.-I. Hanada, N. Zacharakis, A. Lalani, S. Krishna, A. Sachs, T. D. Prickett, Y. F. Li, M. Florentin, S. Kivitz, S. C. Chatmon, S. A. Rosenberg, P. F. Robbins, *Nat. Cancer* **2021**, *2*, 563.
[15] S.-M. Kim, L. Bhonsle, P. Besgen, J. Nickel, A. Backes, K. Held, S. Vollmer, K. Dornmair, J. C. Prinz, *PLoS One* **2012**, *7*, e37338.
[16] M. A. Turchaninova, O. V. Britanova, D. A. Bolotin, M. Shugay, E. V. Putintseva, D. B. Staroverov, G. Sharonov, D. Shcherbo, I. V. Zvyagin, I. Z. Mamedov, C. Linnemann, T. N. Schumacher, D. M. Chudakov, *Eur. J. Immunol.* **2013**, *43*, 2507.
[17] A. Han, J. Glanville, L. Hansmann, M. M. Davis, *Nat. Biotechnol.* **2014**, *32*, 684.
[18] B. Howie, A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, H. S. Robins, *Sci. Transl. Med.* **2015**, *7*, 301ra131.
[19] T. Dupic, Q. Marcou, A. M. Walczak, T. Mora, *PLoS Comput. Biol.* **2019**, *15*, e1006874.
[20] K. Masuda, A. Kornberg, J. Miller, S. Lin, N. Suek, T. Botella, K. A. Secener, A. M. Bacarella, L. Cheng, M. Ingham, V. Rosario, A. M. Al-Mazrou, S. A. Lee-Kong, R. P. Kiran, M. Stoeckius, P. Smibert, A. Del Portillo, P. E. Oberstein, P. A. Sims, K. S. Yan, A. Han, *JCI Insight* **2022**, *7*, e154646.
[21] K. Masuda, A. Kornberg, S. Lin, P. Ho, K. Secener, N. Suek, A. M. Bacarella, M. Ingham, V. Rosario, A. M. Al-Masrou, S. A. Lee-Kong, P. R. Kiran, K. S. Yan, M. Stoeckius, P. Smibert, P. E. Oberstein, P. A. Sims, A. Han, bioRxiv 2020, https://www.biorxiv.org/content/10.1101/2020.09.27.313445v2 (accessed: June 2023).
[22] I. Springer, N. Tickotsky, Y. Louzoun, *Front. Immunol.* **2021**, *12*, 664514.
[23] M. F. Jensen, M. Nielsen, *Elife* **2024**, *12*, RP93934.
[24] T. Lu, Z. Zhang, J. Zhu, Y. Wang, P. Jiang, X. Xiao, C. Bernatchez, J. V. Heymach, D. L. Gibbons, J. Wang, L. Xu, A. Reuben, T. Wang, *Nat. Mach. Intell.* **2021**, *3*, 864.
[25] Y. Han, Y. Yang, Y. Tian, F. J. Fattah, M. S. von Itzstein, M. Zhang, X. Kang, D. M. Yang, J. Liu, Y. Xue, C. Liang, I. Raman, C. Zhu, O. Xiao, Y. Hu, J. E. Dowell, J. Homsi, S. Rashdan, S. Yang, M. E. Gwin, D. Hsiehchen, Y. Gloria-McCutchen, K. Pan, F. Wu, D. Gibbons, X. Wang, C. Yee, J. Huang, A. Reuben, C. Cheng, et al., *bioRxiv* **2023**, https://doi.org/10.1101/2023.12.01.569599.
[26] M. Milighetti, J. Shawe-Taylor, B. Chain, *Front. Physiol.* **2021**, *12*, 1358.
[27] C. Zhu, N. Jiang, J. Huang, V. I. Zarnitsyna, B. D. Evavold, *Immunol. Rev.* **2013**, *251*, 49.
[28] B. Liu, W. Chen, B. D. Evavold, C. Zhu, *Cell* **2014**, *157*, 357.
[29] L. Limozin, M. Bridge, P. Bongrand, O. Dushek, P. A. van der Merwe, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 201902141.
[30] Z. A. Rollins, R. Faller, S. C. George, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2124.
[31] Y. Feng, X. Zhao, A. K. White, K. C. Garcia, P. M. Fordyce, C. Z. Biohub, *bioRxiv* **2021**, https://www.biorxiv.org/content/10.1101/2021.04.24.441194v1. (accessed: June 2023).
[32] S. Kitano, A. Ito, Y. Kim, M. Inoue, M. Fuse, K. Tada, K. Yoshimura, *Cell. Immunol.* **2015**, *6*, 2.

[33] J. Leem, S. H. P. De Oliveira, K. Krawczyk, C. M. Deane, *Nucleic Acids Res.* **2018**, *46*, D406.
[34] F. Grazioli, A. Mösch, P. Machart, K. Li, I. Alqassem, T. J. O'Donnell, M. R. Min, *Front. Immunol.* **2022**, *13*, 1014256.
[35] L. Deng, C. Ly, S. Abdollahi, Y. Zhao, I. Prinz, S. Bonn, *Front. Immunol.* **2023**, *14*, 1128326.
[36] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Z. dek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, et al., *Nature* **2021**, *596*, 583.
[37] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, *bioRxiv* **2021**, https://www.biorxiv.org/content/10.1101/2021.10.04.463034v1 (accessed: June 2023).
[38] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, *bioRxiv* **2022**, https://www.biorxiv.org/content/10.1101/2021.08.15.456425v3 (accessed: June 2023).
[39] A. S. Fiser, A. Šali, *Methods Enzymol.* **2003**, *374*, 461.
[40] B. Webb, A. Sali, *Curr. Protoc. Bioinf.* **2016**, *54*, 5.6.1.
[41] M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen, P. Marcatili, *Nucleic Acids Res.* **2015**, *43*, W349.
[42] K. K. Jensen, V. Rantos, E. C. Jappe, T. H. Olsen, M. C. Jespersen, V. Jurtz, L. E. Jessen, E. Lanzarotti, S. Mahajan, B. Peters, M. Nielsen, P. Marcatili, *Sci. Rep.* **2019**, *9*, 14530.
[43] R. Yin, H. V. Ribeiro-Filho, V. Lin, R. Gowthaman, M. Cheung, B. G. Pierce, *Nucleic Acids Res.* **2023**, *51*, W569.
[44] Z. A. Rollins, J. Huang, I. Tagkopoulos, R. Faller, S. C. George, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3473.
[45] P. Gold, N. A. Goldberg, *McGill J. Med.* **2020**, *3*, 46.
[46] K. Y. Tsang, S. Zaremba, C. A. Nieroda, M. Z. Zhu, J. M. Hamilton, J. Schlom, *JNCI, J. Natl. Cancer Inst.* **1995**, *87*, 982.
[47] M. R. Parkhurst, J. C. Yang, R. C. Langan, M. E. Dudley, D. A. N. Nathan, S. A. Feldman, J. L. Davis, R. A. Morgan, M. J. Merino, R. M. Sherry, M. S. Hughes, U. S. Kammula, G. Q. Phan, R. M. Lim, S. A. Wank, N. P. Restifo, P. F. Robbins, C. M. Laurencot, S. A. Rosenberg, *Mol. Ther.* **2011**, *19*, 620.
[48] J. D. Chodera, *J. Chem. Theory Comput.* **2016**, *12*, 1799.
[49] O. Y. Borbulevych, K. H. Piepenbrink, B. M. Baker, *J. Immunol.* **2011**, *186*, 2950.
[50] M. Y. Shen, A. Sali, *Protein Sci.* **2006**, *15*, 2507.
[51] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, *J. Comput. Chem.* **2005**, *26*, 1701.
[52] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1*, 19.
[53] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586.
[54] M. H. M. Olsson, C. R. SØndergaard, M. Rostkowski, J. H. Jensen, *J. Chem. Theory Comput.* **2011**, *7*, 525.
[55] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, J. H. Jensen, *J. Chem. Theory Comput.* **2011**, *7*, 2284.
[56] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.
[57] Z. Rollins, R. Faller, S. C. George, TCR-pMHC Starting Configurations & Atomic Motion Supplementary Videos UC Davis, Dryad Dataset 2021, https://doi.org/10.25338/B8FK8D.
[58] G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 014101.
[59] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, J. R. Haak, *J. Chem. Phys.* **1984**, *81*, 3684.
[60] D. J. Evans, B. L. Holian, *J. Chem. Phys.* **1985**, *4069*, 83.
[61] M. Parrinello, A. Rahman, *J. Appl. Phys.* **1981**, *52*, 7182.
[62] P. P. Ewald, *Ann. Phys.* **1921**, *369*, 253.
[63] M. Di Pierro, R. Elber, B. Leimkuhler, *J. Chem. Theory Comput.* **2015**, *11*, 5624.
[64] S. Miyamoto, K. P. A. Settle, *J. Comput. Chem.* **1992**, *13*, 952.
[65] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, *J. Comput. Chem.* **1997**, *18*, 1463.
[66] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, *Nature* **2020**, *585*, 357.
[67] W. Mckinney, Data Structures for Statistical Computing in Python. **2010**.
[68] J. D. Hunter, *Comput. Sci. Eng.* **2007**, *9*, 90.
[69] O. Beckstein, GromacsWrapper [Internet], https://github.com/Becksteinlab/GromacsWrapper (accessed: June 2023).
[70] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, et al., *Nat. Methods* **2020**, *17*, 261.
[71] R. Vallat, *J. Open Source Software* **2018**, *3*, 1026.
[72] Z. A. Rollins, https://github.com/zrollins/TCR_homology.git **2022**.
[73] 10 X Genomics, A New Way of Exploring Immunity – Linking Highly Multiplexed Antigen Recognition to Immune Repertoire and Phenotype | Technology Networks [Internet], https://www.technologynetworks.com/immunology/application-notes/a-new-way-of-exploring-immunity-linking-highly-multiplexed-antigen-recognition-to-immune-repertoire-332554 (accessed: June 2023).
[74] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, A. E. Mark, *Angew. Chem., Int. Ed. Engl.* **1998**, *31*, 1387.

