Macromolecular Rapid Communications

Supporting Information

for *Macromol. Rapid Commun.*, DOI 10.1002/marc.202400225

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**

Supporting Information

A computational strategy for the rapid identification and ranking of patient-specific T cell receptors bound to neoantigens

Zachary A. Rollins[1], Matthew B. Curtis[2], Roland Faller[1], and Steven C. George[2 *]

Department of Chemical Engineering[1], Department of Biomedical Engineering[2], University of California, Davis, Davis, California

This PDF includes:

Supplementary Figures

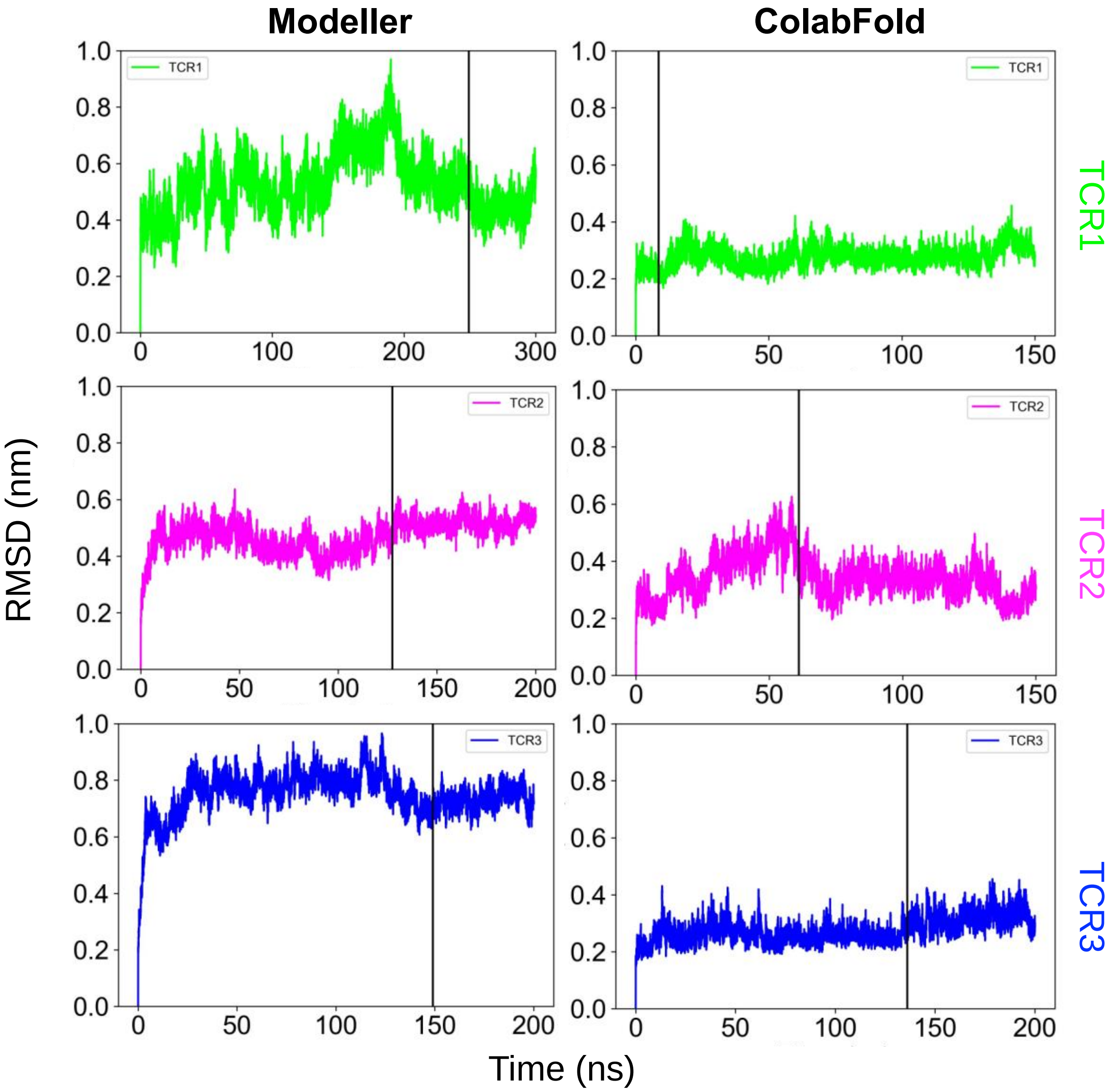


Figure S1: Equilibration of the TCR-pMHC systems. The root mean square deviation (RMSD) is calculated from the initial configuration and plotted versus simulation time (x-axis). RMSD is an all-atom calculation on the entire TCR-pMHC structure. The initial configurations generated by protein-protein structure predictors are identified by columns: Modeller (left) and ColabFold (right). The TCRs are identified by rows: TCR1 (top, green), TCR2 (middle, magenta), and TCR3 (bottom, blue). The equilibrated time-point determined by the variance-bias trade-off algorithm is displayed by the black vertical line in each panel.

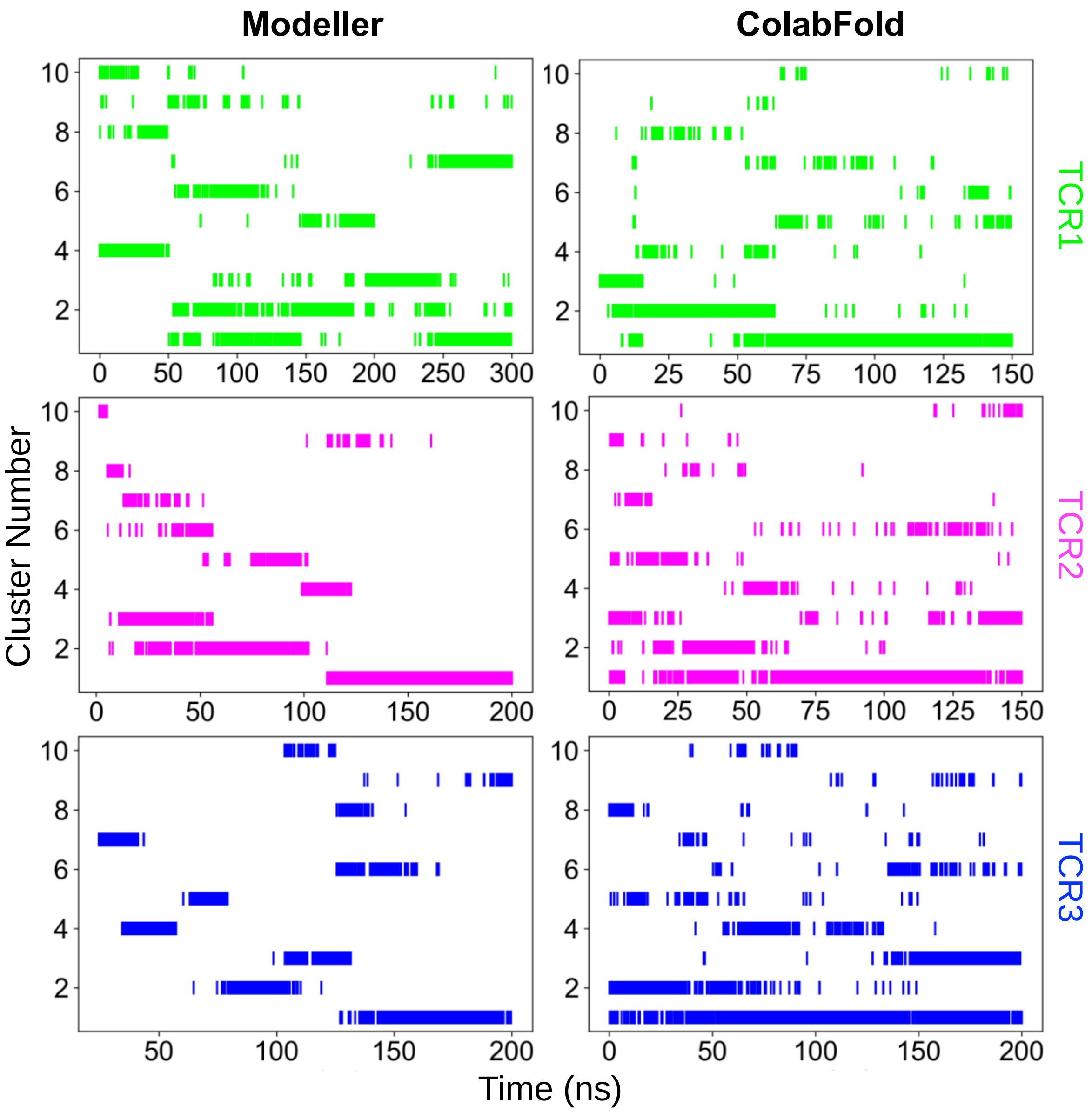
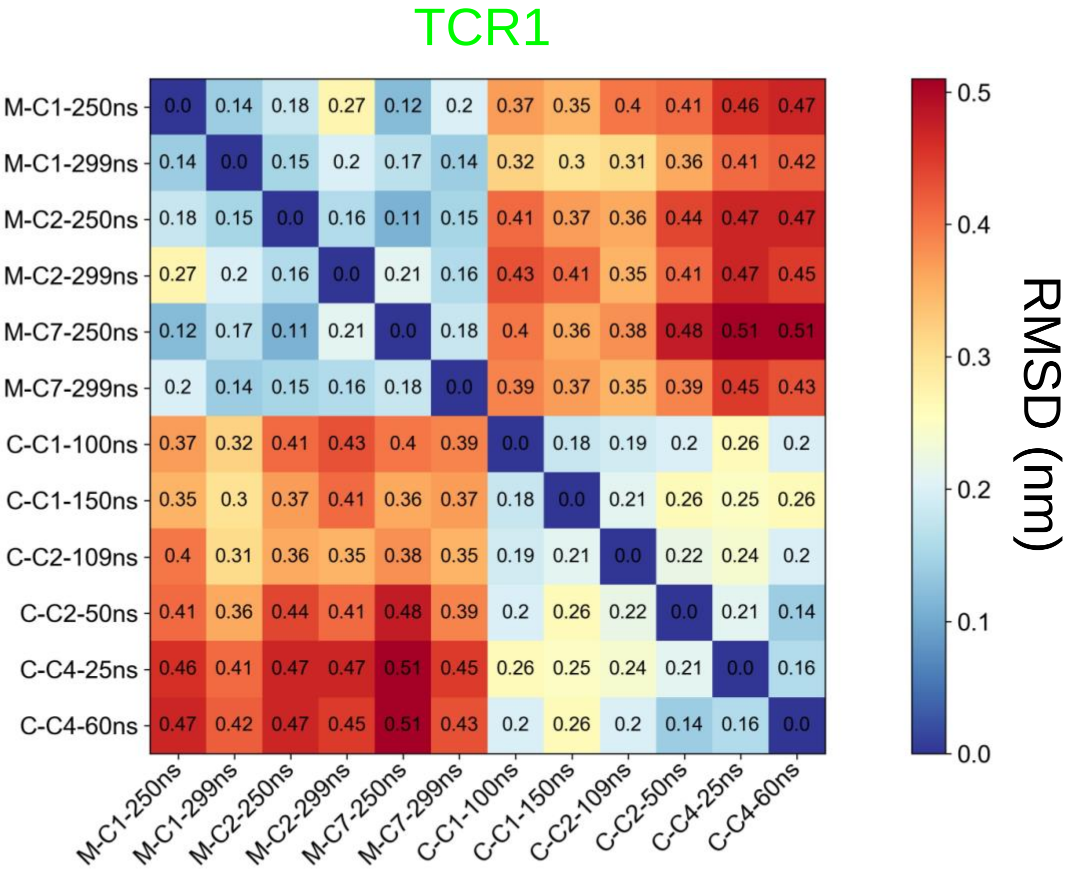
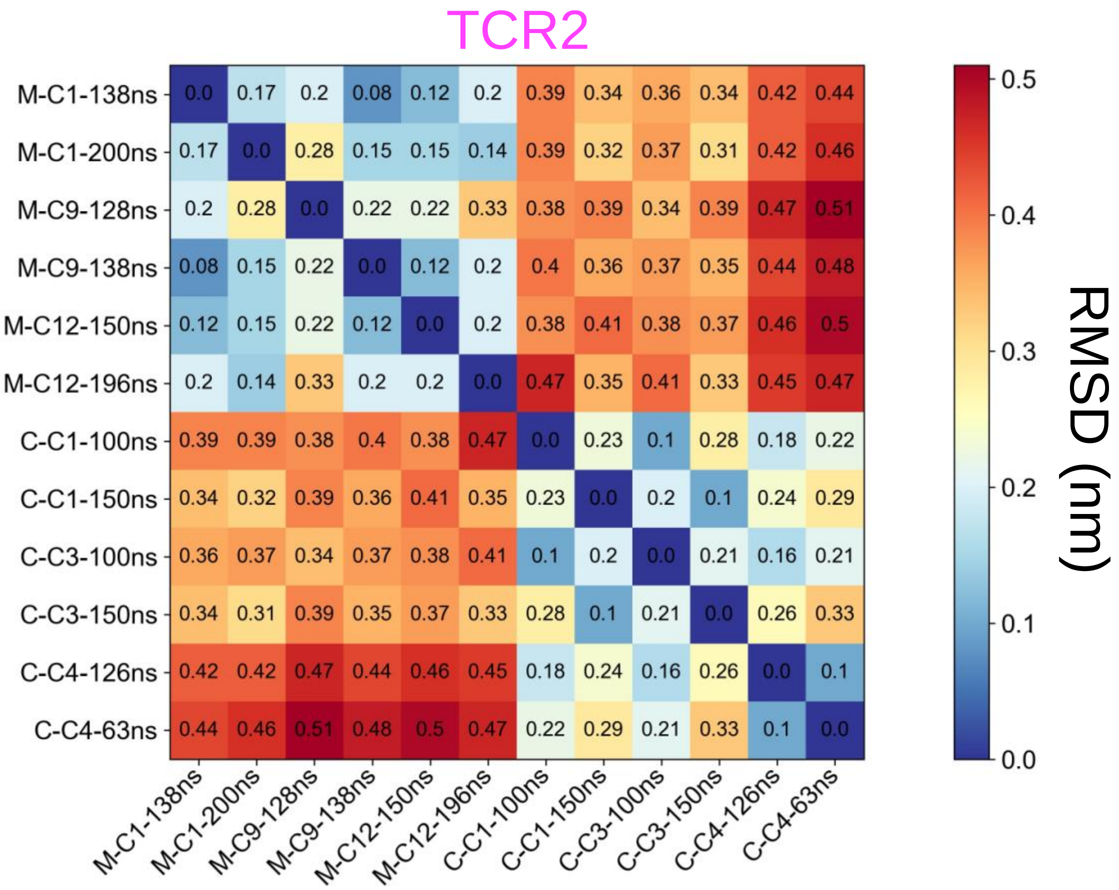


Figure S2: Cluster analysis of the TCR-pMHC systems. The GROMOS clustering algorithm with a C$^{\alpha}$ RMSD cutoff of 0.20 nm was used to identify dominant structural configurations during equilibration. The cluster numbers (y-axis) are ordered by the number of configurations in each cluster (i.e., one is the most dominant) and plotted versus simulation time (x-axis). From Modeller, 70, 45, and 102 clusters were identified with 86.8% (26030/30000), 92.7% (18540/20000), and 71.0% (14206/20000) in the top ten clusters and 26.6% (7980/30000), 38.8% (7763/20000), and 23.1% (4621/20000) in the top cluster for TCR1, TCR2, and TCR3, respectively (**left**). From ColabFold, 25, 28, and 31 clusters were identified with 98.3% (14749/15000), 97.0% (14550/15000), and 97.7% (19530/20000) in the top ten clusters and 54.4% (8158/15000), 55.1% (8263/15000), and 50.7% (10133/20000) in the top cluster for TCR1, TCR2, and TCR3, respectively (**right**). The TCRs are identified by rows: TCR1 (top, green), TCR2 (middle, magenta), and TCR3 (bottom, blue).



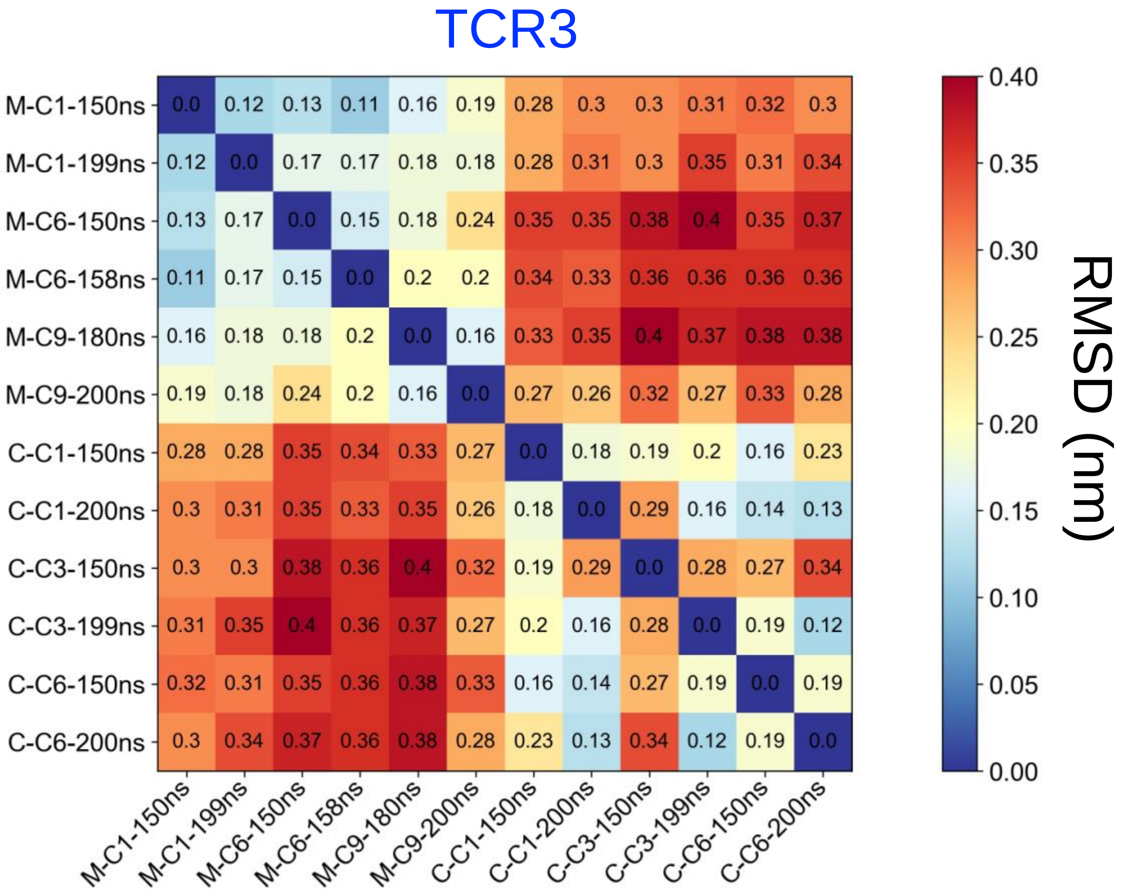
	M-C1-250ns	M-C1-299ns	M-C2-250ns	M-C2-299ns	M-C7-250ns	M-C7-299ns	C-C1-100ns	C-C1-150ns	C-C2-109ns	C-C2-50ns	C-C4-25ns	C-C4-60ns
M-C1-250ns	0.0	0.14	0.18	0.27	0.12	0.2	0.37	0.35	0.4	0.41	0.46	0.47
M-C1-299ns	0.14	0.0	0.15	0.2	0.17	0.14	0.32	0.3	0.31	0.36	0.41	0.42
M-C2-250ns	0.18	0.15	0.0	0.16	0.11	0.15	0.41	0.37	0.36	0.44	0.47	0.47
M-C2-299ns	0.27	0.2	0.16	0.0	0.21	0.16	0.43	0.41	0.35	0.41	0.47	0.45
M-C7-250ns	0.12	0.17	0.11	0.21	0.0	0.18	0.4	0.36	0.38	0.48	0.51	0.51
M-C7-299ns	0.2	0.14	0.15	0.16	0.18	0.0	0.39	0.37	0.35	0.39	0.45	0.43
C-C1-100ns	0.37	0.32	0.41	0.43	0.4	0.39	0.0	0.18	0.19	0.2	0.26	0.2
C-C1-150ns	0.35	0.3	0.37	0.41	0.36	0.37	0.18	0.0	0.21	0.26	0.25	0.26
C-C2-109ns	0.4	0.31	0.36	0.35	0.38	0.35	0.19	0.21	0.0	0.22	0.24	0.2
C-C2-50ns	0.41	0.36	0.44	0.41	0.48	0.39	0.2	0.26	0.22	0.0	0.21	0.14
C-C4-25ns	0.46	0.41	0.47	0.47	0.51	0.45	0.26	0.25	0.24	0.21	0.0	0.16
C-C4-60ns	0.47	0.42	0.47	0.45	0.51	0.43	0.2	0.26	0.2	0.14	0.16	0.0

Figure S3: TCR1 pairwise RMSDs for clustered Modeller and ColabFold structures. The selection of 12 TCR-pMHC structures (6 Modeller + 6 ColabFold) were chosen to be from distinct clusters after equilibration. The pairwise RMSD between selected structures is displayed with the following format: protein-protein structure predictor (i.e., M=Modeller or C=ColabFold) — cluster number from Figure S2 (e.g., C1= cluster 1) — simulation time (e.g.,100 ns). The RMSD is calculated on the C^{α} atoms and scale bar is displayed in nanometers (**right**). The clusters were selected because they were after the simulation time required for equilibrium (e.g., for Modeller TCR1, cluster 1,2, & 7 were chosen because clusters 3-6 are only dominant before reaching equilibrium at 249 ns simulation time).



	M-C1-138ns	M-C1-200ns	M-C9-128ns	M-C9-138ns	M-C12-150ns	M-C12-196ns	C-C1-100ns	C-C1-150ns	C-C3-100ns	C-C3-150ns	C-C4-126ns	C-C4-63ns
M-C1-138ns	0.0	0.17	0.2	0.08	0.12	0.2	0.39	0.34	0.36	0.34	0.42	0.44
M-C1-200ns	0.17	0.0	0.28	0.15	0.15	0.14	0.39	0.32	0.37	0.31	0.42	0.46
M-C9-128ns	0.2	0.28	0.0	0.22	0.22	0.33	0.38	0.39	0.34	0.39	0.47	0.51
M-C9-138ns	0.08	0.15	0.22	0.0	0.12	0.2	0.4	0.36	0.37	0.35	0.44	0.48
M-C12-150ns	0.12	0.15	0.22	0.12	0.0	0.2	0.38	0.41	0.38	0.37	0.46	0.5
M-C12-196ns	0.2	0.14	0.33	0.2	0.2	0.0	0.47	0.35	0.41	0.33	0.45	0.47
C-C1-100ns	0.39	0.39	0.38	0.4	0.38	0.47	0.0	0.23	0.1	0.28	0.18	0.22
C-C1-150ns	0.34	0.32	0.39	0.36	0.41	0.35	0.23	0.0	0.2	0.1	0.24	0.29
C-C3-100ns	0.36	0.37	0.34	0.37	0.38	0.41	0.1	0.2	0.0	0.21	0.16	0.21
C-C3-150ns	0.34	0.31	0.39	0.35	0.37	0.33	0.28	0.1	0.21	0.0	0.26	0.33
C-C4-126ns	0.42	0.42	0.47	0.44	0.46	0.45	0.18	0.24	0.16	0.26	0.0	0.1
C-C4-63ns	0.44	0.46	0.51	0.48	0.5	0.47	0.22	0.29	0.21	0.33	0.1	0.0

Figure S4: TCR2 pairwise RMSDs for clustered Modeller and ColabFold structures. The selection of 12 TCR-pMHC structures (6 Modeller + 6 ColabFold) were chosen to be from distinct clusters after equilibration. The pairwise RMSD between selected structures is displayed with the following format: protein-protein structure predictor (i.e., M=Modeller or C=ColabFold) — cluster number (e.g., C1= cluster 1) — simulation time (e.g.,100 ns). The RMSD is calculated on the C^{α} atoms and scale bar is displayed in nanometers (**right**). The clusters were selected because they were after the simulation time required for equilibrium (e.g., for Modeller TCR2, cluster 1,9, & 12 were chosen because clusters 2-8 &10-11 are only dominant before reaching equilibrium at 127 ns simulation time).



	M-C1-150ns	M-C1-199ns	M-C6-150ns	M-C6-158ns	M-C9-180ns	M-C9-200ns	C-C1-150ns	C-C1-200ns	C-C3-150ns	C-C3-199ns	C-C6-150ns	C-C6-200ns
M-C1-150ns	0.0	0.12	0.13	0.11	0.16	0.19	0.28	0.3	0.3	0.31	0.32	0.3
M-C1-199ns	0.12	0.0	0.17	0.17	0.18	0.18	0.28	0.31	0.3	0.35	0.31	0.34
M-C6-150ns	0.13	0.17	0.0	0.15	0.18	0.24	0.35	0.35	0.38	0.4	0.35	0.37
M-C6-158ns	0.11	0.17	0.15	0.0	0.2	0.2	0.34	0.33	0.36	0.36	0.36	0.36
M-C9-180ns	0.16	0.18	0.18	0.2	0.0	0.16	0.33	0.35	0.4	0.37	0.38	0.38
M-C9-200ns	0.19	0.18	0.24	0.2	0.16	0.0	0.27	0.26	0.32	0.27	0.33	0.28
C-C1-150ns	0.28	0.28	0.35	0.34	0.33	0.27	0.0	0.18	0.19	0.2	0.16	0.23
C-C1-200ns	0.3	0.31	0.35	0.33	0.35	0.26	0.18	0.0	0.29	0.16	0.14	0.13
C-C3-150ns	0.3	0.3	0.38	0.36	0.4	0.32	0.19	0.29	0.0	0.28	0.27	0.34
C-C3-199ns	0.31	0.35	0.4	0.36	0.37	0.27	0.2	0.16	0.28	0.0	0.19	0.12
C-C6-150ns	0.32	0.31	0.35	0.36	0.38	0.33	0.16	0.14	0.27	0.19	0.0	0.19
C-C6-200ns	0.3	0.34	0.37	0.36	0.38	0.28	0.23	0.13	0.34	0.12	0.19	0.0

Figure S5: TCR3 pairwise RMSDs for clustered Modeller and ColabFold structures. The selection of 12 TCR-pMHC structures (6 Modeller + 6 ColabFold) were chosen to be from distinct clusters after equilibration. The pairwise RMSD between selected structures is displayed with the following format: protein-protein structure predictor (i.e., M=Modeller or C=ColabFold) — cluster number (e.g., C1= cluster 1) — simulation time (e.g.,100 ns). The RMSD is calculated on the C^α atoms and scale bar is displayed in nanometers (**right**). The clusters were selected because they were after the simulation time required for equilibrium (e.g., for Modeller TCR3, cluster 1,6, & 9 were chosen because clusters 2-5 & 7-8 are only dominant before reaching equilibrium at 149 ns simulation time).

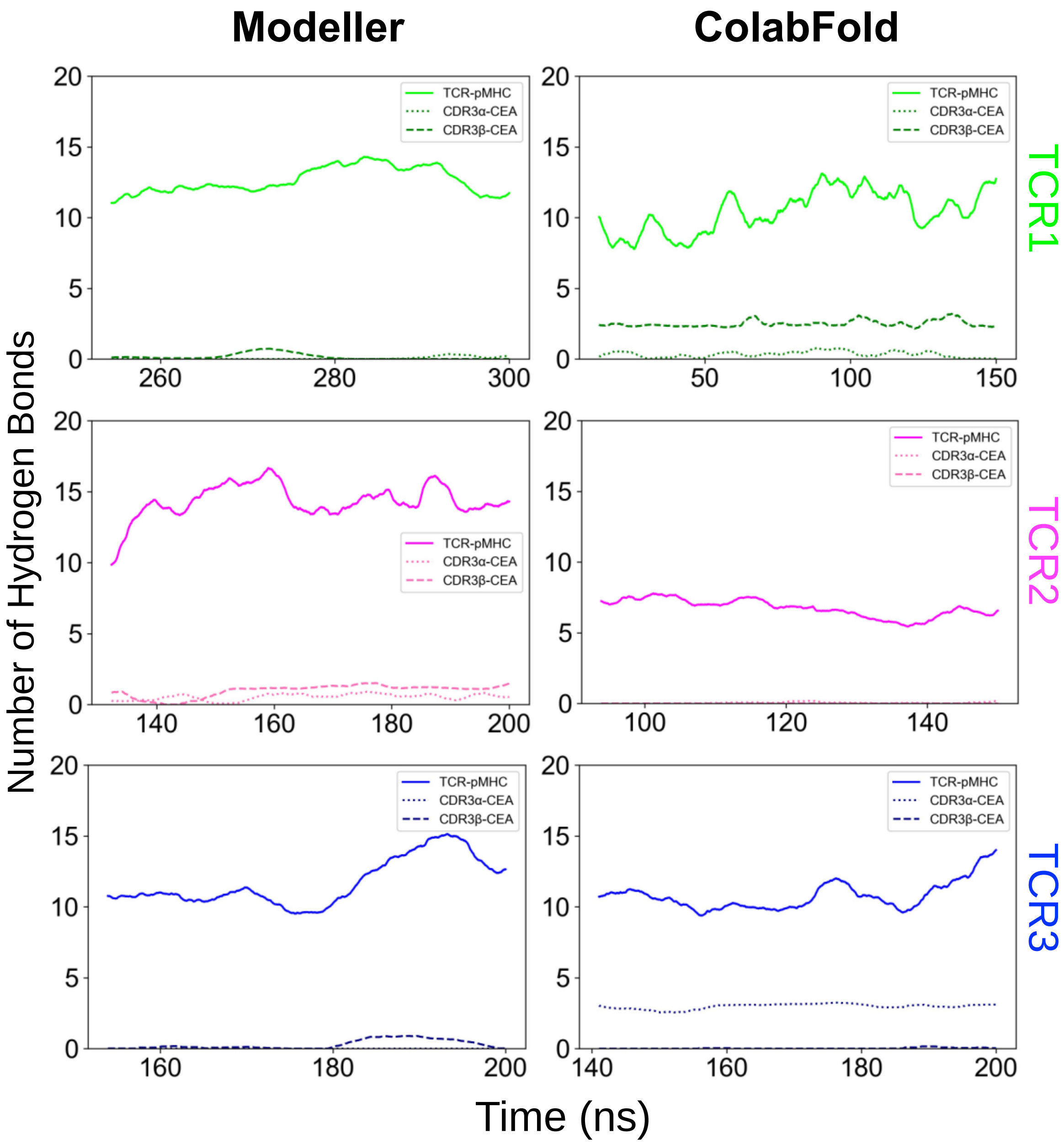


Figure S6: Hydrogen bonds for the TCR-pMHC systems. The number of hydrogen bonds (y-axis) is plotted versus simulation time (x-axis) after the determined equilibrated time-point. The initial configurations generated by protein-protein structure predictors are identified by columns: Modeller (left) and ColabFold (right). The TCRs are identified by rows: TCR1 (top, green), TCR2 (middle, magenta), and TCR3 (bottom, blue). The number of hydrogen bonds between the TCR-pMHC (solid line), CDR3α-CEA (dotted line), and CDR3β-CEA (dashed line) are plotted versus simulation time.

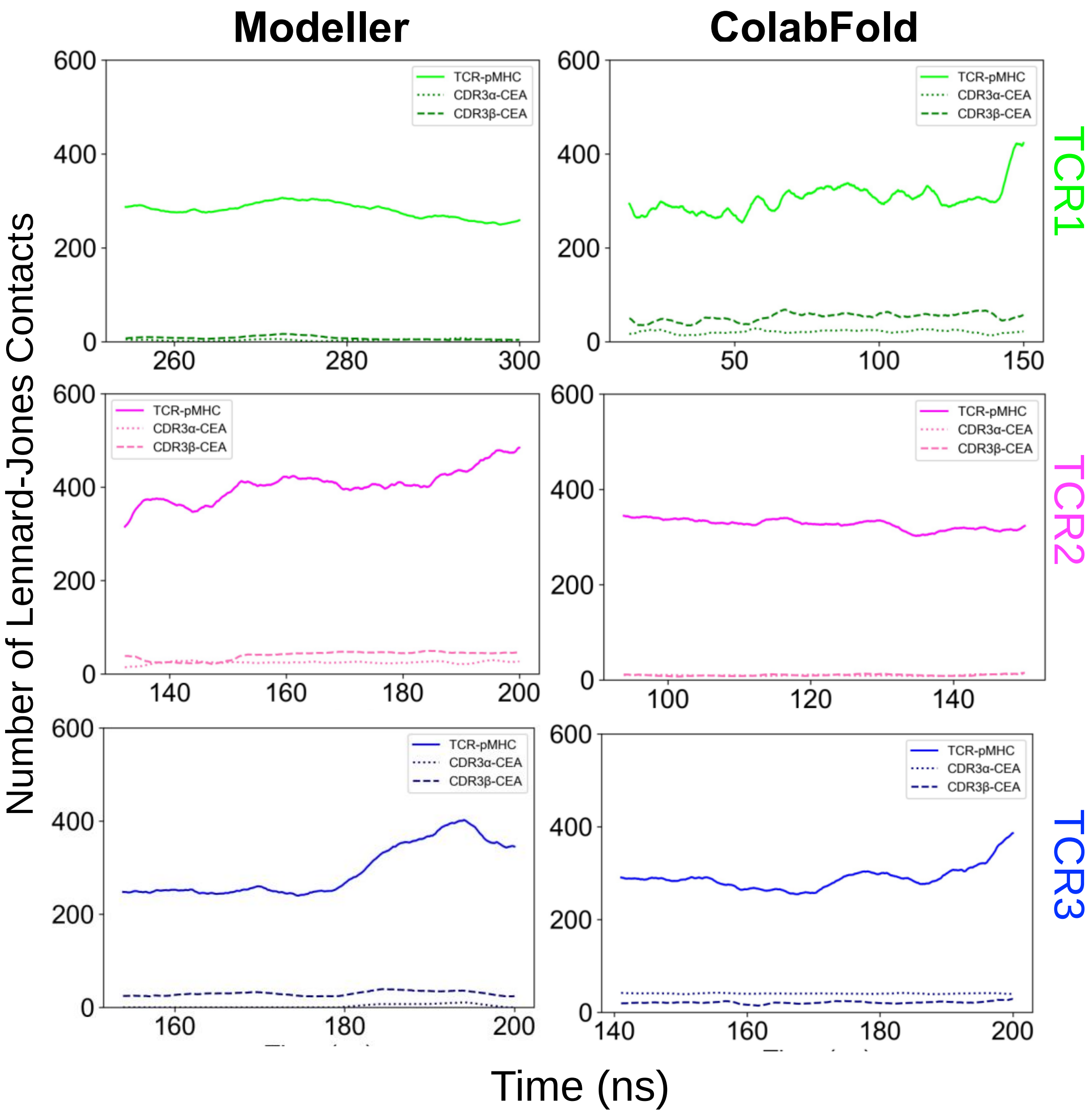


Figure S7: Lennard-Jones contacts for the TCR-pMHC systems. The number of Lennard-Jones contacts (y-axis) is plotted versus simulation time (x-axis) after the determined equilibrated time-point. The initial configurations generated by protein-protein structure predictors are identified by columns: Modeller (left) and ColabFold (right). The TCRs are identified by rows: TCR1 (top, green), TCR2 (middle, magenta), and TCR3 (bottom, blue). The number of Lennard-Jones contacts between the TCR-pMHC (solid line), CDR3α-CEA (dotted line), and CDR3β-CEA (dashed line) are plotted versus simulation time.

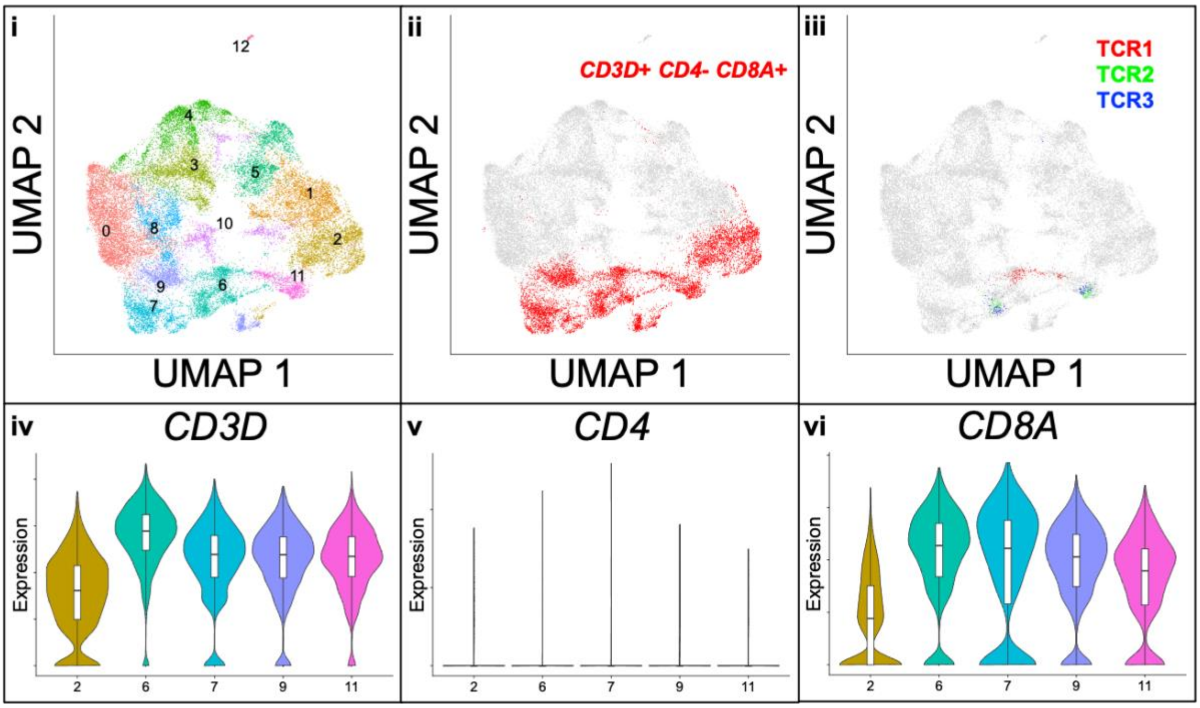


Figure S8: **Identification of the top 3 most frequent TCR clonotypes in the dataset. i)** UMAP projection of T cell gene expression data from Han, *et al*. with unsupervised clusters (same as Figure 2Ai). **ii)** distribution of cells identified as *CD3D+CD4-CD8A+* by gene expression from Figure 2Aii-2Aiv. **iii)** distribution of cells expressing the top 3 most frequent TCR clonotypes in the dataset. **iv)** Violin plot on expression of *CD3D* for all cells identified as *CD3D+CD4-CD8A+*. **v)** Violin plot on expression of *CD4* for all cells identified as *CD3D+CD4-CD8A+*. **vi)** Violin plot on expression of *CD8A* for all cells identified as *CD3D+CD4-CD8A+*.