The Atomic-Level Physiochemical Determinants of T Cell Receptor Dissociation Kinetics

Zachary A. Rollins[2], Jun Huang[4], Ilias Tagkopoulos[3], Roland Faller[2], and Steven C. George[1*]

Department of Biomedical Engineering [1], Department of Chemical Engineering [2], Department of Computer Science [3], University of California, Davis, Davis, California; Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL[4]

*Corresponding Author:

Steven C. George, M.D., Ph.D.

Professor and Chair

Department of Biomedical Engineering

451 E. Health Sciences Drive, room 2315

University of California, Davis

Davis, CA 95616

Email: scgeorge@ucdavis.edu

ABSTRACT

The rational design of T Cell Receptors (TCRs) for immunotherapy has stagnated due to a limited understanding of the dynamic physiochemical features of the TCR that elicit an immunogenic response. The physiochemical features of the TCR-peptide major histocompatibility complex (pMHC) bond dictate bond lifetime which, in turn, correlates with immunogenicity. Here, we: i) characterize the force-dependent dissociation kinetics of the bond between a TCR and a set of pMHC ligands using Steered Molecular Dynamics (SMD); and ii) implement a machine learning algorithm to identify which physiochemical features of the TCR govern dissociation kinetics. Our results demonstrate that the total number of hydrogen bonds between the CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide are critical features that determine bond lifetime. We propose that amino acid substitutions to these hypervariable regions of the TCR can efficiently manipulate immunogenicity and thus be used in the rational design of TCRs for immunotherapy.

Keywords: T cell receptor; peptide major histocompatibility complex; immunogenicity; steered molecular dynamics; machine learning

INTRODUCTION

T cell-based immunotherapies (e.g., chimeric antigen receptor-T, or CAR-T; and TCR-engineered-T, or TCR-T) have provided transformative therapeutic responses in a small subset of cancers and patients[1-5]; however, progress in solid tumors has been agonizingly slow. For example, CAR-T cells require an antigen on the tumor cell surface, but the majority (~85%) of identified neoantigens are intracellular[6] and thus are immunogenic only when a representative fragment is presented on the cell surface in a peptide-major histocompatibility complex (i.e., pMHC). Although TCR-T therapy is MHC-restricted, this approach can target intracellular antigens, and the remarkable sensitivity of a TCR to recognize a single pMHC molecule[7] provides an additional strategic advantage. Nonetheless, identifying neoepitopes, matching these with immunogenic TCRs, and minimizing off-target effects remain significant challenges to implementation of these therapies[8].

Recent reports demonstrate that single-cell sequencing and machine learning technologies can identify patient- and tumor-specific neoepitopes[9, 10]. However, identification of partner TCRs remains challenging, despite the fact that tumor-specific T cells can be found in the peripheral blood[11, 12]. The human immune system generates tumor-specific T cells in a process that begins with random V(D)J recombination to create the hypervariable regions of the TCR α and β chains. While this process generates a stunningly large number of *possible* TCRs (>10^{20}-10^{61})[13, 14], including 10^{6}-10^{8} in the peripheral blood, it is inherently inefficient and does not necessarily produce a TCR with appropriate immunogenicity for a given tumor[15]. Alternate strategies of TCR identification have also fallen short; for example, TCR affinity enhancement can lead to a loss of TCR specificity[16, 17] and does not always determine immunogenicity[18].

Computational techniques such as steered molecular dynamics (SMD) and machine learning may enable the creation of highly immunogenic, tumor-specific TCRs through rapid and

efficient screening of the vast number of possible TCRs. The success of these techniques depends on accurate *in vitro* predictions of T cell immunogenicity, a goal that remains elusive. Quantitative descriptors of the TCR-pMHC bond identified in previous studies do not consistently correlate with immunogenicity[18-21]. The majority of these studies measured equilibrium parameters of the TCR-pMHC bond, which do not account for the mechanical forces on the TCR-pMHC bond present *in vivo.* Recent studies using DNA-based tension probes have estimated this force at ~10-20 pN[22, 23], and subsequent studies demonstrate that dissociation kinetics (i.e., bond lifetime) of the TCR-pMHC bond at this physiologic force can predict immunogenicity[24-31]. These correlations are consistent across species, TCR-pMHC pairs, and experimental systems[24-31].

Here, we seek to discern the atomic-level physiochemical features that determine the TCR-pMHC bond lifetime under force (i.e., characterize the TCR-pMHC's force-dependent dissociation kinetics). As a first attempt to manipulate the bond lifetime of the TCR-pMHC over a wide range, we characterized the force-dependent dissociation kinetics of a single TCR (with a known crystal structure) to 17 possible pMHCs using steered molecular dynamics (SMD). Then, we used machine learning to identify the physiochemical features and the specific regions of the TCR regulating bond lifetime. Our results demonstrate that the total number of hydrogen bonds (H-bonds) between the CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide are critical features that determine bond lifetime. This finding may inform the rational design of TCRs for TCR-T cell therapy, and provide a path forward to create more advanced and predictive maching learning algorithms.

METHODS

Molecular Dynamics Setup. The crystal structure of the human DMF5 TCR complexed with agonist pMHC MART1-HLA-A2 (PDB code: 3QDJ)[32] was the initial structure for all simulations (**Figure 1A**). To generate the 17 TCR-pMHC pairs, amino acid substitutions were made to the MART1 peptide (AAGIGILTV) using the Mutagenesis plugin on Pymol

Molecular Graphics System (Schrödinger, New York, New York). Interfacial substructures (**Figure 1B**) were defined by sequential residues from the corresponding chains: TCRα (CDR1α: 24-32, CDR2α: 50-55, CDR3α: 89-99), TCRβ (CDR1β: 25-31, CDR2β: 51-58, CDR3β: 92-103), MHCα (MHCα(β): 50-85, MHCα(α): 138-179), and peptide (1-9). To determine protonation states, pKa values were calculated using propka3.1[33, 34] and residues were considered deprotonated in Gromacs[35] if pKa values were below the physiological pH 7.4. The resulting systems were solvated in rectangular water boxes using the TIP3P water model[36] large enough to satisfy the minimum image convention. Na^+ and Cl^- ions were added to neutralize protein charge and reach physiologic salt concentration of ~150 mM. All

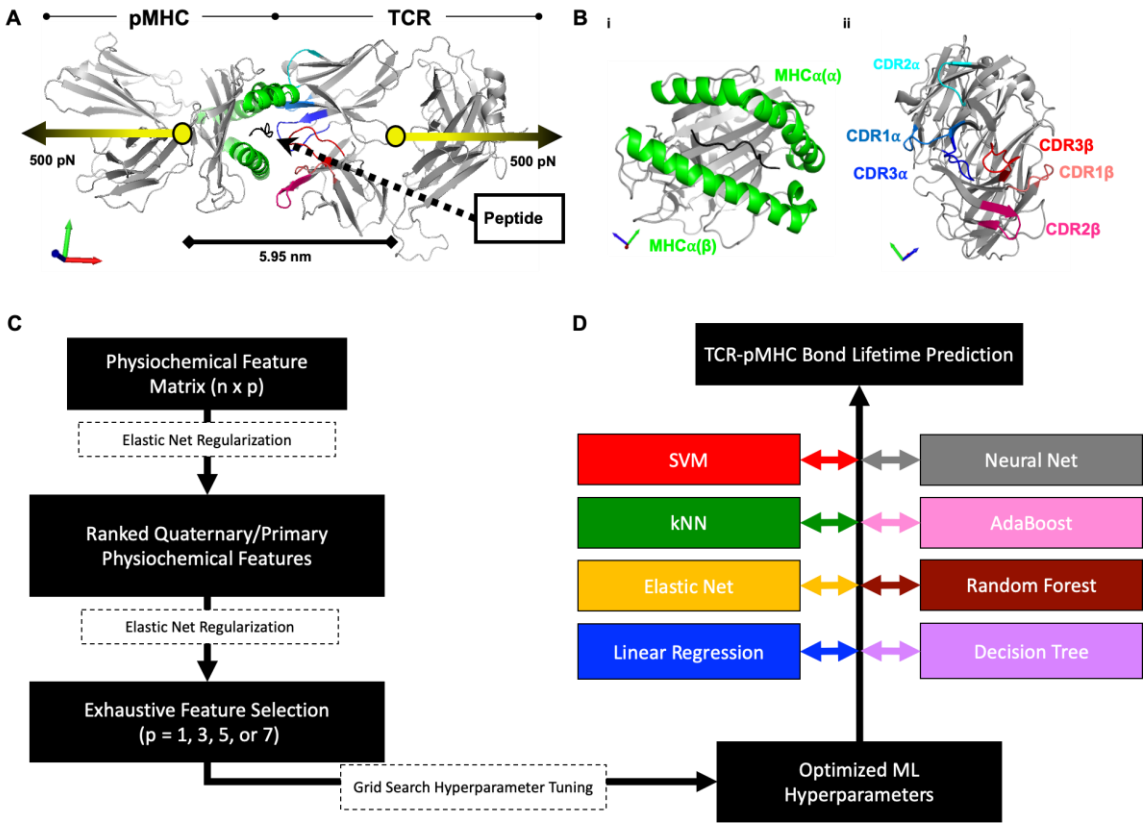

Figure 1: Steered Molecular Dynamics (SMD) simulations and machine learning algorithms were used to identify the physiochemical features that predict TCR-pMHC bond lifetime. **(A)** Starting structure for SMD of TCR and pMHC (shown at the top with black lines and circle arrowheads). The location/direction of pulling are depicted with yellow circles/arrows, respectively; the black scale bar with diamond arrowheads denotes the locality of distance between center of masses. The non-interacting bodies of the TCR and pMHC are colored in gray. Axis directions are indicated in left corner (red: +x-direction, blue: +y-direction, and green: +z-direction). **(B)** The primary interfacial substructures: **(i)** MHCα(α) & MHCα(β) = green, Epitope=black; and **(ii)** TCR CDR1α = light blue, TCR CDR2α = cyan, TCR CDR2α = dark blue, TCR CDR1β = salmon, TCR CDR2β = light red, and TCR CDR3β = red. **(C)** A two-layer Elastic Net-Exhaustive Feature Selection algorithm (dashed boxes) was used to obtain ranked and reduced feature sets. **(D)** Selected features were used to tune hyperparameters (dashed box) for each machine learning model (Linear Regression = blue, Elastic Net = orange, k-Nearest Neighbors = green, Support Vector Machines = red, Decision Tree = purple, Random Forest = brown, AdaBoost = pink, Neural Net = gray).

simulations were performed with Gromacs 2019.1[35] using the CHARM 22 plus CMAP force field for proteins[37] and orthorhombic periodic boundary conditions. All simulations were in full atomistic detail.

Energy Minimization and Equilibration. Generating equilibrated starting structures for the Steered Molecular Dynamics simulations required four steps: (1) Steepest descent energy minimization to ensure correct geometry and the absence of steric clashes; (2) 100 ps simulation in the constant volume (NVT) ensemble to bring atoms to correct kinetic energies, while maintaining temperature at 310 K by coupling all protein and non-protein atoms to separate baths using the velocity rescale thermostat with a 0.1 ps time constant[38]; (3) 100 ps simulation in the constant pressure (NPT) ensemble using Berendsen pressure coupling[38] and a 2.0 ps time constant to maintain isotropic pressure at 1.0 bar; and (4) Production MD simulations conducted for 50-150 ns with no restraints. The protein structures were evaluated every 50 ns to determine if all protein chains were equilibrated by root mean square deviation. To ensure true NPT ensemble sampling during 100 ns production runs, the Nose-Hoover thermostat[39] and Parrinello-Rahman barostat[40] were used to maintain temperature and pressure, respectively. Time constants were 2.0 and 1.0 ps for pressure and temperature coupling, respectively, utilizing the isothermal compressibility of water, 4.5^{-5} bar^{-1}. Box size for equilibration was 10.627 x 7.973 x.10.685 nm^3 with ~ 48,000 water molecules, ~300 ions, and ~157,000 total atoms. All simulation steps used the Particle Ewald Mesh algorithm[41, 42] for long-range electrostatic calculations with cubic interpolation and 0.12 nm maximum grid spacing. Short-range non-bonded interactions were cut off at 1.2 nm using the Verlet cutoff-scheme and all bond lengths were constrained using LINCS algorithm[43]. The leap-frog algorithm was used for integrating equations of motion with 2 fs time steps. After the preparation runs, three independent MD configurations for each peptide mutant were extracted and used as the three starting points for steered molecular dynamics simulations.

Steered Molecular Dynamics (SMD). The full TCR-pMHC complex structure was extracted from the preparation run for each peptide mutant to generate three SMD starting configurations. The x-axis of these protein complexes was aligned along the *x*-axis and solvated in rectangular water boxes with dimensions 30 x 9.972 x 12.685 nm^3. Solvent was again represented by the TIP3P water model and $Na^{+}$and Cl^{-} ions were added to neutralize protein charge and reach physiologic salt concentration of ~150 mM. This resulted in ~120,000 water molecules, ~700 ions, and ~370,000 total atoms. All Gromacs structure files are uploaded to the Dryad repository for the exact atomic specifications. Before pulling, all systems underwent (1) energy minimization; (2) 100 ps NVT; and (3) and 100 ps NPT to remove high energy contacts without disturbing the configurations. During pull, the Nose-Hoover thermostat and Parrinello-Rahman barostat were used to maintain temperature and pressure. 500 pN linear potential was applied to the center of mass (COM) of the TCR and pMHC in the *x*-direction and simulations continued until distance between COMs reached 0.49 times the box size in *x*-direction (**Figure 1A**). The COM was chosen as the site of applied force because pulling from the TCR and MHC termini resulted in artificial unfolding (not shown). All simulation trajectories and selected frames were visualized using the Pymol Molecular Graphics System (Schrödinger, New York, New York).

Physiochemical Descriptors and Data Analysis. Physiochemical descriptors were evaluated by defining Gromacs index groups (gmx make_ndx) and using Gromacs-suite analysis tools (i.e., gmx hbond, gmx rms, gmx rmsf, gmx sasa, gmx gyrate, gmx distance). Data analyses were performed by standard python packages for data handling and visualization (i.e., numpy[44], pandas[45], seaborn[46], matplotlib[47], statistics[48], and GromacsWrapper[49]), and custom python scripts. Random mutants were generated with a custom python script compatible with Pymol using the random python package and selecting a random location and amino acid to mutate the peptide. The machine learning algorithms were developed using the sklearn package[50, 51] and exhaustive feature selection was performed using mlxtend package[52]. The geometry of a Lennard-Jones contact (LJ-contact) is defined as a distance

less than 0.35 nm between atoms. The L1 peptide bond lifetime was an outlier (z-score = 3.65 > 3). To reduce the effects of the outlier on the dataset, median absolute error was selected as the scoring criterion and L1 was excluded from correlation coefficient calculations. The mean absolute error represents the arithmetic average of median absolute error from repeated three-fold cross validation. The Pearson correlation coefficient (r_p) and Spearman rank correlation coefficients (r_s) were calculated using the correlation method in the pandas python package. Akaike and Bayesian Information Criterion (AIC and BIC) were calculated from the standard deviation of repeated three-fold cross validation of the best machine learning algorithm selected from the hyperparameter grid search. Statistical significance was determined by performing a one-tailed student's t-test ($p<0.05$) for each machine learning algorithm across feature sets. Custom scripts relevant to mutant generation, feature selection, machine learning, and the production of figures have been made available on a GitHub repository: https://github.com/zrollins/TCR.ai.git.

Feature Selection and Machine Learning Algorithms. Features were ranked and reduced utilizing a two-layer Elastic Net – Exhaustive Search algorithm (**Figure 1C**). First, Elastic Net Regularization[53] was used with all physiochemical features and a grid search was performed to optimize hyperparameters. The optimized hyperparameters were implemented into the Exhaustive Feature Selector[52] and the best individual features were ranked by repeated (n_repeats=3) threefold cross-validation. The top ten features were ranked by mean absolute error and feature combinations were exhaustively searched, utilizing Elastic Net Regularization, to determine the best combinations of 3, 5, and 7 features (**Figure 1C**). The best feature combinations were selected by mean absolute error arithmetically averaged over the cross-validation. These feature combinations were then implemented into several machine learning algorithms to determine the most predictive model of bond lifetime (**Figure 1D**)[50, 51]. The machine learning algorithm hyperparameter optimization was performed on a high performance compute cluster at the University of California, Davis College of Engineering and the best model for each feature set was scored on absolute error and

ranked by the arithmetic average of repeated threefold cross-validation (i.e., n_splits=3, n_repeats=3, random_state=1). Detailed documentation regarding the cross validation and hyperparameter optimization of two-layer Elastic Net – Exhaustive Search feature selection and machine learning predictions are provided in the supporting information. In addition, this dataset has been made freely available on the GitHub repository.

RESULTS

Bond lifetime. As the starting point to simulate the force-dependent dissociation kinetics of 17 TCR-pMHC pairs using SMD, we used the previously reported crystal structure (PDB ID: 3QDJ)[32] of the DMF5 TCR (from a melanoma patient) bound to the MART1 peptide (AAGIGILTV)-MHC complex (**Figure 1A**). We then replaced the MART1 peptide with 16 different peptides (**Figure 1—supplement 1**) for a total of 17 TCR-pMHC pairs. Ten peptides were chosen from a set of known pMHCs[54, 55] and 7 were generated through random point mutation of the MART1 peptide. For these 17 TCR-pMHC pairs, the mean bond lifetime in the SMD simulations was 5400 ± 1700 picoseconds (**Figure 2**).

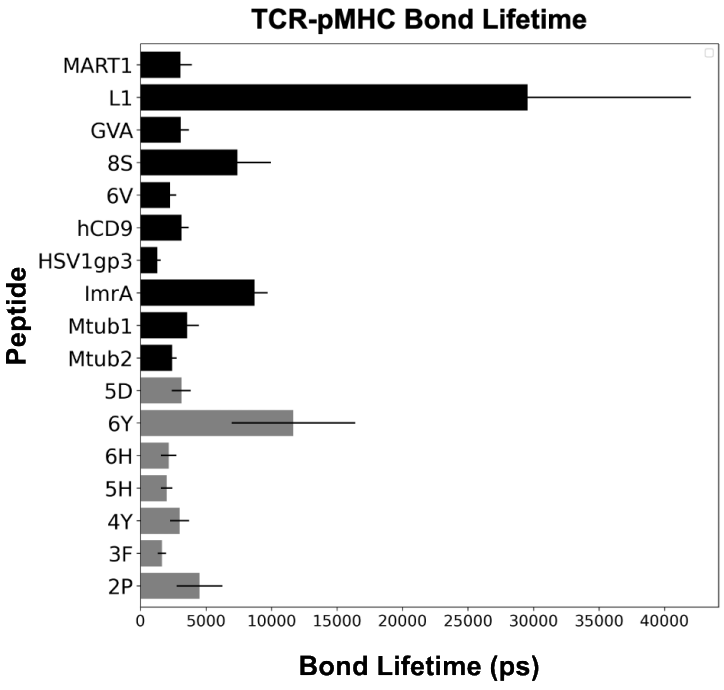

Figure 2. Mean TCR-pMHC bond lifetime for 17 different peptides. Using Steered Molecular Dynamics (SMD), we applied a constant force of 500 pN at the center of mass for the TCR and pMHC and estimated the mean bond lifetime for 17 different peptides. Known peptides and those with random point mutations are denoted with black and gray bars, respectively. Each TCR-pMHC was pulled apart 3 times using different equilibrated structures.

Physiochemical features of the TCR-pMHC. Next, we identified two sets of physiochemical features which, at distinct resolution levels, describe the TCR-pMHC bond during the SMD simulation. The first set characterizes physiochemical

features of the entire TCR-pMHC interaction (e.g., total H-bonds between the TCR and pMHC). This characterization provides an overall assessment of the physiochemical features that might impact bond lifetime and is consistent with the quaternary structure of globular proteins. We considered features likely to impact dissociation kinetics and thus included H-bonds[56], LJ-contacts[57], distance between the TCR and pMHC[58, 59], solvent accessible surface area (SASA)[60], root mean square fluctuations (RMSF)[61], and the gyration tensor of the TCR and pMHC. This approach resulted in 18 features for the first set, and we dubbed these quaternary features (**Figure 1—supplement 2**).

An understanding of the physiochemical features that regulate dissociation kinetics of the global TCR-pMHC bond provides an overall assessment of which physiochemical features regulate bond lifetime. However, this approach does not identify the sub-regions of the TCR-pMHC bond that regulate bond lifetime and thus are suitable targets for rational design of TCRs. The hypervariable regions of the TCR can be divided into 3 complementarity determining regions (CDRs) on the α and β chain, respectively. Within the MHC, the peptide is surrounded by α-helices which also interact with the nearby chains of the TCR (**Figure 1B**). These MHC α-helices are located on the MHCα chain and these substructures are defined by their interaction with the TCR α and β chain, respectively (i.e., MHC$\alpha(\alpha)$ and MHC$\alpha(\beta)$). These TCR CDRs and MHC α-helices form an interface with the peptide antigen – the variable in this study – and based on their physical location are likely to influence TCR-pMHC bond lifetime. Hence, we also identified a second set of features focused on the interface between the TCR and the pMHC (e.g., CDR3α loop of the TCR and the MHC$\alpha(\beta)$ chain, **Figure 1B**). This higher level of resolution is consistent with the secondary structures (e.g., α-helices) of a protein. Again, we considered features that are likely to affect dissociation kinetics and thus included H-bonds, LJ-contacts, distance between the sub-regions, SASA, RMSF, and the gyration tensor of the sub-regions. From these considerations, we identified 79 secondary features (**Figure 1—supplement 3**) that

could potentially impact dissociation kinetics. The quaternary and secondary features were further categorized into chemical – such as H-bonds and LJ-Contacts – and physical – including RMSF, SASA, and the gyration tensor – interaction parameters.

TCR-pMHC Bond Lifetime Prediction using Quaternary Physiochemical Features.

To examine how quaternary physiochemical features influence TCR-pMHC bond dissociation kinetics, we ranked the top ten quaternary features after an Elastic Net grid search for each individual feature (**Figure 3A**). The scoring criterion was mean absolute

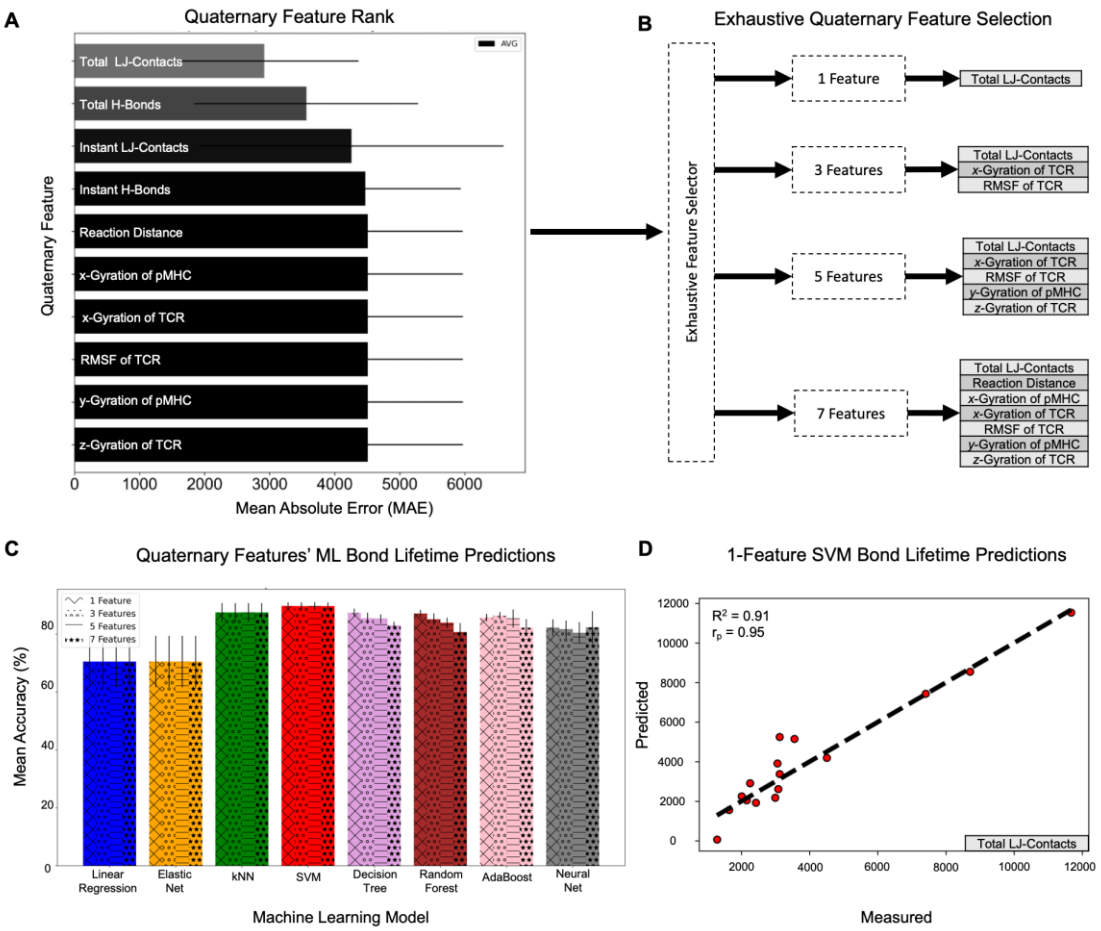

Figure 3: Quaternary Feature Selection and Bond Lifetime Predictions. **(A)** Mean absolute test error from elastic net regularization was used to select the top ten quaternary features. Errors represent the best test set standard deviation from repeated threefold cross-validation. **(B)** According to an exhaustive search, the best feature sets (i.e., p = 1, 3, 5, and 7) to predict bond lifetime. **(C)** The mean accuracies of bond lifetime prediction for all feature sets in (B) and machine learning models after hyperparameter tuning (Linear Regression = blue, Elastic Net = orange, k-Nearest Neighbors = green, Support Vector Machines = red, Decision Tree = purple, Random Forest = brown, AdaBoost = pink, Neural Net = gray). Errors represent the best test set standard error from repeated threefold cross-validation. The machine learning model standard error from cross-validation (n=9) was statistically compared for increasing feature sets by a one-tailed student's t-test: #p<0.10, *p<0.05, **p<0.01. **(D)** The scatter plot of predicted and measured bond lifetimes from the selected one-feature Support Vector Machines algorithm with the coefficient of determination (top left), the Pearson correlation coefficient (top left), and the feature set (bottom right).

error of bond lifetime in picoseconds. After Elastic Net grid search, chemical interaction features, in particular Total LJ-contacts and Total H-bonds, were the most predictive (**Figure 3A**); in particular, the total number of unique LJ-Contacts between TCR and pMHC had the smallest mean absolute error. In addition, the total LJ-Contacts had the highest Pearson and Spearman correlation coefficients (**Figure 3—supplement 1**, **Figure 3—supplement 3**).

We next explored whether a combination of quaternary physiochemical features would improve predictions of bond lifetime. To accomplish this, we applied a regularized regression method (Elastic Net; see **Methods**) as a filter to identify predictive feature sets. To avoid overfitting[62-64], feature sets were reduced utilizing an Elastic Net[53] – Exhaustive Search[52] algorithm (**Figure 1C**) to determine the best combinations of 3, 5, and 7 features. Using these combinations, we then trained and tested 8 different machine learning algorithms to estimate TCR-pMHC bond lifetime **(Figure 1D)**[50, 51]. Although physical quaternary features were selected in this exhaustive search (**Figure 3B**), these did not significantly improve the predictive power of the machine learning models (**Figure 3C**). This finding holds for all machine learning algorithms, as determined by the lack of statistically significant increase in mean accuracy or decrease in information criteria scores (Akaike and Bayesian Information Criteria) with increasing model complexity (**Figure 3—supplement 2**, **Figure 3—supplement 4**).

The best feature combination and machine learning model was chosen based on the lowest error and standard deviation from repeated three-fold cross-validation. Our results demonstrated that a feature set of only LJ-Contacts combined with a Support Vector Machines is best at predicting bond lifetime (**Figure 3D**). The mean absolute error using Support Vector Machines was 560 ± 200 picoseconds producing an accuracy of 90.0 ± 3.7% (i.e., 1-560/5400).

TCR-pMHC Bond Lifetime Prediction Using Secondary Physiochemical Features.

Analogous to our strategy to assess quaternary features of the TCR-pMHC, we examined secondary features. We ranked the top ten secondary features after an Elastic Net grid search for each individual feature (**Figure 4A**). The total number of unique H-bonds between CDR2β -MHCα(β) generated the smallest mean absolute error (**Figure 4A**). In addition, the top three features had the highest Pearson and Spearman correlation coefficients (**Figure 4—supplement 1**, **Figure 3—supplement 3**).

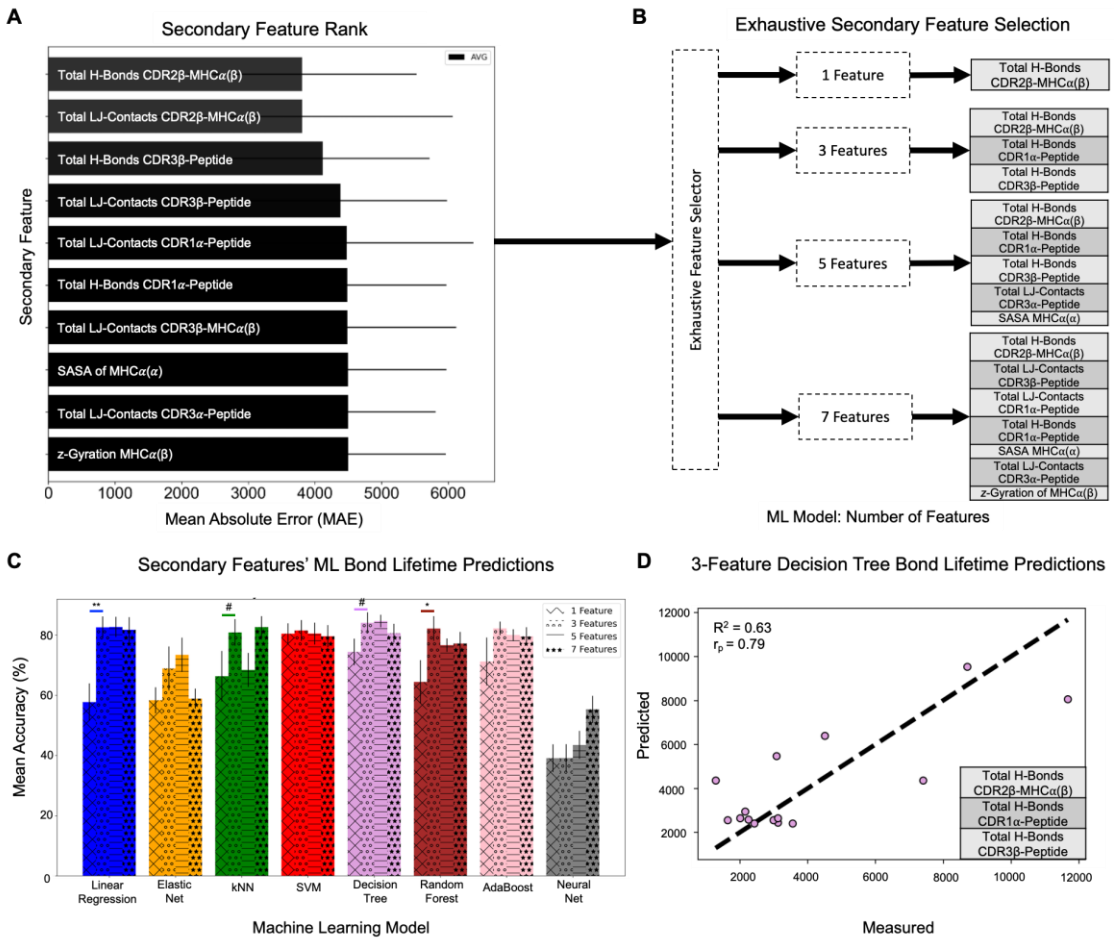

Figure 4. Secondary Feature Selection and Bond Lifetime Predictions. (**A**) Mean absolute test error from elastic net regularization was used to select the top ten secondary features. Errors represent the best test set standard deviation from repeated threefold cross-validation. (**B**) According to an exhaustive search, the best feature sets (i.e., p = 1, 3, 5, and 7) to predict bond lifetime. **(C)** The mean accuracies of bond lifetime prediction for all feature sets in (B) and machine learning models after hyperparameter tuning (Linear Regression = blue, Elastic Net = orange, k-Nearest Neighbors = green, Support Vector Machines = red, Decision Tree = purple, Random Forest = brown, AdaBoost = pink, Neural Net = gray). Errors represent the best test set standard error from repeated threefold cross-validation. The machine learning model standard error from cross-validation (n=9) was statistically compared for increasing feature sets by a one-tailed student's t-test: #p<0.10, *p<0.05, **p<0.01. **(D)** The scatter plot of predicted and measured bond lifetimes from the selected 3-feature Decision Tree algorithm with the coefficient of determination (top left), the Pearson correlation coefficient (top left), and the feature set (bottom right).

We explored whether a combination of secondary physiochemical features would improve the prediction of bond lifetime. Following the same algorithm as for quaternary features, we applied an Elastic Net[53] – Exhaustive Search[52] algorithm (**Figure 1D**) to identify the best combinations of 3, 5, and 7 secondary features; cross-validated 8 machine learning models with these feature combinations; and selected the best feature combination and machine learning model based on error, standard deviation, and information criteria. Interestingly, the best 3 feature combination (CDR2β -MHCα(β), CDR1α-Peptide, and CDR3β-Peptide) selected by exhaustive search (**Figure 4B**) did not correspond to the top three individual features selected by Elastic Net rank (**Figure 4A)** or correlation coefficients (**Figure 3—supplement 3**). Compared to the single best feature, the best 3-feature combination statistically improved bond lifetime predictions for Linear Regression, k-Nearest Neighbors, Decision Tree, and Random Forest machine learning algorithms (**Figure 4C**). Increases in mean accuracy were not statistically significant beyond 3 features (**Figure 4C, Figure 4—supplement 3**). Moreover, these algorithms reduced information criteria scores (Akaike and Bayesian Information Criteria) when increasing from 1 to 3 features, whereas the Elastic Net, Support Vector Machines, and Neural Net algorithms increased both AIC & BIC (**Figure 4—supplement 2**). These results indicate that, among the secondary features and machine learning algorithms tested, a 3-feature combination utilizing a Decision Tree provides the most accurate prediction of bond lifetime (**Figure 4D**). The absolute error using the Decision Tree was 870 ± 570 picoseconds (**Figure 4—supplement 3**), or an accuracy of 84 ± 10%. In addition, this Decision Tree prediction by the best 3 feature combination exceeded the Pearson correlation coefficient of the individual features (**Figure 4D, Figure 4—supplement 2**).

DISCUSSION

T cell-based immunotherapies, such as TCR-engineered-T cells, provide exciting potential to treat a wide range of cancers, including solid tumors. However, this potential has not been reached, due, in part, to the inability to rapidly and efficiently explore the vast TCR space to

identify optimal tumor-specific TCRs. Experimental methods to design and test potential TCRs are expensive and slow, thus hindering throughput. In contrast, computational algorithms that utilize machine learning have enormous potential to rapidly interrogate the TCR space and identify a small number of candidates for more efficient experimental testing. We tested this premise using SMD to create a small database of TCR-pMHC bond lifetimes, then created machine learning algorithms to predict bond lifetime based on quaternary and secondary features of the TCR-pMHC bond. Using the quaternary features, we found that total LJ-contacts could predict bond lifetime with 90% accuracy. More importantly, we also found that we could predict bond lifetime with an accuracy of 84% using only the total H-bonds between three subregions of the TCR-pMHC: CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide. This result identifies new and unanticipated regions of the TCR to target in the rational design of TCRs for immunotherapy.

Quaternary Features of the TCR-pMHC. Upon quaternary feature investigation, the LJ-Contacts between the TCR and pMHC dominated bond lifetime prediction. In fact, for all machine learning algorithms investigated, there was no statistically significant i) increase in mean accuracy when expanding to larger feature sets (**Figure 3C**) or ii) decrease in information criteria scores (**Figure 4—supplement 2**). Moreover, although physical features (e.g., *x*-Gyration of TCR) were selected in the exhaustive feature selection process (**Figure 3B**), these did not significantly increase mean accuracy. This demonstrates that no selected physical features improve predictive performance and thus the atomic motion of the TCR or pMHC is unlikely to regulate dissociation kinetics.

Secondary Features of the TCR-pMHC. To identify the specific subregions of the TCR that determine the TCR-pMHC bond lifetime, we investigated the TCR-pMHC interface and included substructures, or secondary protein features, that defined the interaction (**Figure 1B**). Physiochemical features within each substructure and between adjacent substructures

(**Figure 1—supplement 3**) were then evaluated to determine the best predictors of bond lifetime. Among the features and machine learning algorithms selected, a 3-feature combination of secondary features (CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide) was selected as the most accurate predictor of TCR-pMHC bond lifetime. This was based on: i) a decrease in information criteria score for 5 of 8 machine learning algorithms; and ii) a statistically significant increase in mean accuracy for 4 of 8 machine learning algorithms when increasing the feature set size from 1 to 3. We found that the combination of total H-bonds between these subregions could predict bond lifetime with the highest accuracy.

The finding that both the total number of unique H-bonds between CDR2β-MHCα(β) and CDR1α-Peptide predict TCR-pMHC bond lifetime is unanticipated. Of particular note, the total number of H-bonds between CDR2β-MHCα(β) remained in all exhaustive search feature sets (**Figure 3B**). Most attention has focused on the heralded CDR3 domains[65] given the proximity to the peptide (**Figure 1A, B**). In contrast, CDR2 flanks the MHCα(α) and MHCα(β) chains. It is perhaps not surprising, given the significantly larger number of residues (MHCα(β) = 42 residues) compared to the peptide (peptide = 9 residues), that interactions between the CDR2β and the MHCα(β) could potentially be the most significant physiochemical features to impact bond lifetime.

The inclusion of CDR1α-Peptide H-bonds draws new attention to the CDR1α region. Similar to the CDR3β region, CDR1α is in proximity to the peptide (**Figure 1A, B**) and thus hydrogen bonding between these substructures may be expected. However, surprisingly, CDR1α-Peptide H-bonds was exhaustively selected despite interactions between CDR3α-Peptide in the exhaustive feature set (**Figure 4A**). Overall, these results suggest that mutagenesis strategies to increase hydrogen bonding between CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide may enhance TCR-pMHC force-dependent bond lifetime. It is

important to acknowledge that the interactions between these interfacial substructures may be specific to the DMF5 TCR and will require further investigation to generalize. Nonetheless, these results bring new attention to the CDR1 and CDR2 regions in the future of TCR design. Finally, in contrast to previous reports[28, 57], peptide radius of gyration and CDR3α-CDR3β distance were not selected in the top ten predictive features. This is likely due to the artificial pMHC unfolding by pulling from TCR-pMHC termini [27] and the lack of diversity in TCR-pMHC pairs evaluated [28, 57], respectively.

Computational Methods. One of the limiting factors of this study is the computational constraint of generating a SMD dataset; here, we examined 17 TCR-pMHC pairs. Larger datasets would likely provide more useful insight into feature combinations that predict TCR-pMHC bond lifetime, but come at a significant additional computational cost. Similarly, although the two-layer Elastic Net – Exhaustive Search feature selection methodology provided a rapid filtering of physiochemical features, this biases the machine learning predictor towards features selected by Elastic Net. At the cost of computation, exhaustive or recursive feature selection for each machine learning predictor may improve predictive performance. However, the focus of this work is to provide an architecture for identifying physiochemical features that dictate TCR-pMHC dissociation kinetics.

Bond Lifetime. The force dependent bond lifetime (at ~10-20 pN) has been reported to correlate with TCR-pMHC immunogenicity. These findings highlight the importance of TCR-pMHC bond lifetime and suggest that the TCR needs to sustain and form transient bonds under load for sufficient time to initiate biochemical signaling. Thus, we utilized force-dependent bond lifetime as an objective function to uncover the physiochemical determinants of this biomolecular design feature. It is important to note that this biomolecular design feature does not necessarily conflict with catch-slip bond behavior [24], and we recognize that our approach may be expanded in the future to include other physiochemical characteristics of the TCR-pMHC bond.

Conclusions. We have demonstrated the utility of combining two computational methods – steered molecular dynamics and machine learning – to create a methodology that can potentially be used to rapidly and efficiently examine the vast TCR space to predict the bond lifetime, and thus the immunogenic response, of a given TCR-pMHC pair. Our initial results suggest that the physiochemical features of three subregions of the TCR-pMHC are of particular importance in determining bond lifetime (CDR2β-MHCα(β), CDR1α-Peptide, and CDR3β-Peptide) and provide new and unanticipated regions of the TCR to manipulate in the rational design of TCR-engineered T cells.

SUPPORTING INFORMATION

Figure 3—supplement 1 Quaternary Features vs Bond Lifetime. **Figure 3—supplement 2** Akaike and Bayesian Information Criterion for Quaternary Features. **Figure 4—supplement 1** Secondary Features vs Bond Lifetime. **Figure 4—supplement 2** Akaike and Bayesian Information Criterion for Secondary Features. **Figure 1—supplement 1** Peptides used in SMD simulations, including their amino acid sequences. **Figure 1—supplement 2** Quaternary Features. **Figure 1—supplement 3** Secondary Features. **Figure 3—supplement 3** Pearson Correlation and Spearman Rank Correlation Coefficients. **Figure 3—supplement 4** Best Machine Learning Models after Hyperparameter Optimization for Quaternary Features. **Figure 4—supplement 3** Best Machine Learning Models after Hyperparameter Optimization for Secondary Features.

ACKNOWLEDGEMENTS

Simulations were performed on the hpc1/hpc2 clusters in the UC Davis, College of Engineering. This work was supported in part by startup funding to SCG from the Department of Biomedical Engineering**.** We thank Nuala Del Piccolo for editing the manuscript.

AUTHOR CONTRIBUTIONS:

ZAR: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review & editing, visualization

JH: writing—review & editing

IT: methodology, software, validation, formal analysis, writing—review & editing, supervision

RF: conceptualization, methodology, formal analysis, resources, writing—review & editing, supervision, project administration

SCG: conceptualization, methodology, formal analysis, resources, writing—original draft, writing—review & editing, supervision, project administration, funding acquisition

COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The datasets of physiochemical features and bond lifetime for each TCR-pMHC pair have been made available on a Dryad repository (https://doi.org/10.25338/B8R33G). Moreover, the root mean square deviation equilibration plots of all TCR-pMHC pairs, starting configurations for all Steered Molecular Dynamics runs, and results from hyperparameter search/cross-validation of the machine learning models have been uploaded to the Dryad repository. The trajectory data generated from Steered Molecular Dynamics simulations are not publicly available due to storage capacity limitations. However, the trajectory data can be reproduced from the initial configurations located in the Dryad repository and are available upon reasonable request, including provision of external storage capacity, to the corresponding author. All scripts including the two-layer Elastic Net – Exhaustive Search algorithm, hyperparameter search of machine learning algorithms with cross validation, and those used to generate figures have been uploaded to a GitHub repository (https://github.com/zrollins/TCR.ai.git). There are no restrictions on data accessibility.

REFERENCES

1. Johnson LA, Morgan RA, Dudley ME, Cassard L, Yang JC, Hughes MS, Kammula US, Royal RE, Sherry RM, Wunderlich JR, Lee CC, Restifo NP, Schwarz SL, Cogdill AP, Bishop RJ, Kim H, Brewer CC, Rudy SF, VanWaes C, Davis JL, Mathur A, Ripley RT, Nathan DA, Laurencot CM, Rosenberg SA. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. Blood. 2009;114(3):535-46. Epub 2009/05/20. doi: 10.1182/blood-2009-03-211714. PubMed PMID: 19451549; PMCID: PMC2929689.

2. Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, Litzky L, Bagg A, Carreno BM, Cimino PJ, Binder-Scholl GK, Smethurst DP, Gerry AB, Pumphrey NJ, Bennett AD, Brewer JE, Dukes J, Harper J, Tayton-Martin HK, Jakobsen BK, Hassan NJ, Kalos M, June CH. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. Blood. 2013;122(6):863-71. Epub 2013/06/19. doi: 10.1182/blood-2013-03-490565. PubMed PMID: 23770775; PMCID: PMC3743463.

3. Moore T, Wagner CR, Scurti GM, Hutchens KA, Godellas C, Clark AL, Kolawole EM, Hellman LM, Singh NK, Huyke FA, Wang SY, Calabrese KM, Embree HD, Orentas R, Shirai K, Dellacecca E, Garrett-Mayer E, Li M, Eby JM, Stiff PJ, Evavold BD, Baker BM, Le Poole IC, Dropulic B, Clark JI, Nishimura MI. Clinical and immunologic evaluation of three metastatic melanoma patients treated with autologous melanoma-reactive TCR-transduced T cells. Cancer Immunol Immunother. 2018;67(2):311-25. Epub 2017/10/21. doi: 10.1007/s00262-017-2073-0. PubMed PMID: 29052782; PMCID: PMC5935006.

4. Morgan RA, Dudley ME, Wunderlich JR, Hughes MS, Yang JC, Sherry RM, Royal RE, Topalian SL, Kammula US, Restifo NP, Zheng Z, Nahvi A, de Vries CR, Rogers-Freezer LJ, Mavroukakis SA, Rosenberg SA. Cancer regression in patients after transfer of genetically engineered lymphocytes. Science. 2006;314(5796):126-9. Epub 2006/09/02. doi: 10.1126/science.1129003. PubMed PMID: 16946036; PMCID: PMC2267026.

5. Robbins PF, Morgan RA, Feldman SA, Yang JC, Sherry RM, Dudley ME, Wunderlich JR, Nahvi AV, Helman LJ, Mackall CL, Kammula US, Hughes MS, Restifo NP, Raffeld M, Lee CC, Levy CL, Li YF, El-Gamil M, Schwarz SL, Laurencot C, Rosenberg SA. Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. J Clin Oncol. 2011;29(7):917-24. Epub 2011/02/02. doi: 10.1200/JCO.2010.32.2537. PubMed PMID: 21282551; PMCID: PMC3068063.

6. Weekes MP, Antrobus R, Lill JR, Duncan LM, Hör S, Lehner PJ. Comparative analysis of techniques to purify plasma membrane proteins. Journal of biomolecular techniques : JBT. 2010;21(3):108-15. PubMed PMID: 20808639.

7. Sykulev Y, Joo M, Vturina I, Tsomides TJ, Eisen HN. Evidence that a Single Peptide–MHC Complex on a Target Cell Can Elicit a Cytolytic T Cell Response. Immunity. 1996;4(6):565-71. doi: https://doi.org/10.1016/S1074-7613(00)80483-5.

8. He Q, Jiang X, Zhou X, Weng J. Targeting cancers through TCR-peptide/MHC interactions. J Hematol Oncol. 2019;12(1):139. Epub 2019/12/20. doi: 10.1186/s13045-019-0812-8. PubMed PMID: 31852498; PMCID: PMC6921533.

9. Gartner JJ, Parkhurst MR, Gros A, Tran E, Jafferji MS, Copeland A, Hanada K-I, Zacharakis N, Lalani A, Krishna S, Sachs A, Prickett TD, Li YF, Florentin M, Kivitz S, Chatmon SC, Rosenberg SA, Robbins PF. A machine learning model for ranking candidate HLA class I neoantigens based on known neoepitopes from multiple human tumor types. Nature Cancer. 2021;2(5):563-74. doi: 10.1038/s43018-021-00197-6.

10. Kosaloglu-Yalcin Z, Lanka M, Frentzen A, Logandha Ramamoorthy Premlal A, Sidney J, Vaughan K, Greenbaum J, Robbins P, Gartner J, Sette A, Peters B. Predicting T cell recognition of MHC class I restricted neoepitopes. Oncoimmunology. 2018;7(11):e1492508. Epub 2018/11/01. doi: 10.1080/2162402X.2018.1492508. PubMed PMID: 30377561; PMCID: PMC6204999.
11. Dijkstra KK, Cattaneo CM, Weeber F, Chalabi M, van de Haar J, Fanchi LF, Slagter M, van der Velden DL, Kaing S, Kelderman S, van Rooij N, van Leerdam ME, Depla A, Smit EF, Hartemink KJ, de Groot R, Wolkers MC, Sachs N, Snaebjornsson P, Monkhorst K, Haanen J, Clevers H, Schumacher TN, Voest EE. Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. Cell. 2018;174(6):1586-98 e12. Epub 2018/08/14. doi: 10.1016/j.cell.2018.07.009. PubMed PMID: 30100188; PMCID: PMC6558289.
12. Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, Prickett TD, Gartner JJ, Crystal JS, Roberts IM, Trebska-McGowan K, Wunderlich JR, Yang JC, Rosenberg SA. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. Nat Med. 2016;22(4):433-8. Epub 2016/02/24. doi: 10.1038/nm.4051. PubMed PMID: 26901407; PMCID: PMC7446107.
13. de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermsen R, Chain B, de Boer RJ. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. Elife. 2020;9. Epub 2020/03/19. doi: 10.7554/eLife.49900. PubMed PMID: 32187010; PMCID: PMC7080410.
14. Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. Front Immunol. 2013;4:485. Epub 2014/01/15. doi: 10.3389/fimmu.2013.00485. PubMed PMID: 24421780; PMCID: PMC3872652.
15. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. 2011;21(5):790-7. Epub 2011/02/26. doi: 10.1101/gr.115428.110. PubMed PMID: 21349924; PMCID: PMC3083096.
16. Kunert A, Obenaus M, Lamers CHJ, Blankenstein T, Debets R. T-cell Receptors for Clinical Therapy: In Vitro Assessment of Toxicity Risk. Clin Cancer Res. 2017;23(20):6012-20. Epub 2017/06/25. doi: 10.1158/1078-0432.CCR-17-1012. PubMed PMID: 28645940.
17. Mensali N, Myhre MR, Dillard P, Pollmann S, Gaudernack G, Kvalheim G, Walchli S, Inderberg EM. Preclinical assessment of transiently TCR redirected T cells for solid tumour immunotherapy. Cancer Immunol Immunother. 2019;68(8):1235-43. Epub 2019/06/20. doi: 10.1007/s00262-019-02356-2. PubMed PMID: 31214732; PMCID: PMC6682583.
18. Kersh GJ, Kersh EN, Fremont DH, Allen PM. High- and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling. Immunity. 1998;9(6):817-26. Epub 1999/01/09. doi: 10.1016/s1074-7613(00)80647-0. PubMed PMID: 9881972.
19. van der Merwe PA, Davis SJ. Molecular interactions mediating T cell antigen recognition. Annu Rev Immunol. 2003;21:659-84. Epub 2003/03/05. doi: 10.1146/annurev.immunol.21.120601.141036. PubMed PMID: 12615890.
20. Rudolph MG, Wilson IA. The specificity of TCR/pMHC interaction. Curr Opin Immunol. 2002;14(1):52-65. Epub 2002/01/16. doi: 10.1016/s0952-7915(01)00298-9. PubMed PMID: 11790533.

21. Zhu C, Jiang N, Huang J, Zarnitsyna VI, Evavold BD. Insights from in situ analysis of TCR-pMHC recognition: response of an interaction network. Immunol Rev. 2013;251(1):49-64. Epub 2013/01/03. doi: 10.1111/imr.12016. PubMed PMID: 23278740; PMCID: PMC3539230.
22. Liu Y, Blanchfield L, Ma VP, Andargachew R, Galior K, Liu Z, Evavold B, Salaita K. DNA-based nanoparticle tension sensors reveal that T-cell receptors transmit defined pN forces to their antigens for enhanced fidelity. Proc Natl Acad Sci U S A. 2016;113(20):5610-5. Epub 2016/05/04. doi: 10.1073/pnas.1600163113. PubMed PMID: 27140637; PMCID: PMC4878516.
23. Ma R, Kellner AV, Ma VP, Su H, Deal BR, Brockman JM, Salaita K. DNA probes that store mechanical information reveal transient piconewton forces applied by T cells. Proc Natl Acad Sci U S A. 2019;116(34):16949-54. Epub 2019/08/09. doi: 10.1073/pnas.1904034116. PubMed PMID: 31391300; PMCID: PMC6708336.
24. Liu B, Chen W, Evavold BD, Zhu C. Accumulation of dynamic catch bonds between TCR and agonist peptide-MHC triggers T cell signaling. Cell. 2014;157(2):357-68. Epub 2014/04/15. doi: 10.1016/j.cell.2014.02.053. PubMed PMID: 24725404; PMCID: PMC4123688.
25. Liu B, Chen W, Natarajan K, Li Z, Margulies DH, Zhu C. The cellular environment regulates in situ kinetics of T-cell receptor interaction with peptide major histocompatibility complex. Eur J Immunol. 2015;45(7):2099-110. Epub 2015/05/07. doi: 10.1002/eji.201445358. PubMed PMID: 25944482; PMCID: PMC5642113.
26. Kolawole EM, Andargachew R, Liu B, Jacobs JR, Evavold BD. 2D Kinetic Analysis of TCR and CD8 Coreceptor for LCMV GP33 Epitopes. Front Immunol. 2018;9:2348. Epub 2018/10/31. doi: 10.3389/fimmu.2018.02348. PubMed PMID: 30374353; PMCID: PMC6197077.
27. Sibener LV, Fernandes RA, Kolawole EM, Carbone CB, Liu F, McAffee D, Birnbaum ME, Yang X, Su LF, Yu W, Dong S, Gee MH, Jude KM, Davis MM, Groves JT, Goddard WA, 3rd, Heath JR, Evavold BD, Vale RD, Garcia KC. Isolation of a Structural Mechanism for Uncoupling T Cell Receptor Signaling from Peptide-MHC Binding. Cell. 2018;174(3):672-87 e27. Epub 2018/07/28. doi: 10.1016/j.cell.2018.06.017. PubMed PMID: 30053426; PMCID: PMC6140336.
28. Wu P, Zhang T, Liu B, Fei P, Cui L, Qin R, Zhu H, Yao D, Martinez RJ, Hu W, An C, Zhang Y, Liu J, Shi J, Fan J, Yin W, Sun J, Zhou C, Zeng X, Xu C, Wang J, Evavold BD, Zhu C, Chen W, Lou J. Mechano-regulation of Peptide-MHC Class I Conformations Determines TCR Antigen Recognition. Mol Cell. 2019;73(5):1015-27 e7. Epub 2019/02/04. doi: 10.1016/j.molcel.2018.12.018. PubMed PMID: 30711376; PMCID: PMC6408234.
29. Das DK, Feng Y, Mallis RJ, Li X, Keskin DB, Hussey RE, Brady SK, Wang JH, Wagner G, Reinherz EL, Lang MJ. Force-dependent transition in the T-cell receptor beta-subunit allosterically regulates peptide discrimination and pMHC bond lifetime. Proc Natl Acad Sci U S A. 2015;112(5):1517-22. Epub 2015/01/22. doi: 10.1073/pnas.1424829112. PubMed PMID: 25605925; PMCID: PMC4321250.
30. Robert P, Aleksic M, Dushek O, Cerundolo V, Bongrand P, van der Merwe PA. Kinetics and mechanics of two-dimensional interactions between T cell receptors and different activating ligands. Biophys J. 2012;102(2):248-57. Epub 2012/02/22. doi: 10.1016/j.bpj.2011.11.4018. PubMed PMID: 22339861; PMCID: PMC3260781.
31. Limozin L, Bridge M, Bongrand P, Dushek O, van der Merwe PA, Robert P. TCR-pMHC kinetics under force in a cell-free system show no intrinsic catch bond, but a minimal

encounter duration before binding. Proc Natl Acad Sci U S A. 2019;116(34):16943-8. Epub 2019/07/19. doi: 10.1073/pnas.1902141116. PubMed PMID: 31315981; PMCID: PMC6708305.

32. Borbulevych OY, Santhanagopolan SM, Hossain M, Baker BM. TCRs used in cancer gene therapy cross-react with MART-1/Melan-A tumor antigens via distinct mechanisms. J Immunol. 2011;187(5):2453-63. Epub 2011/07/29. doi: 10.4049/jimmunol.1101268. PubMed PMID: 21795600; PMCID: PMC3166883.

33. Olsson MH, Sondergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput. 2011;7(2):525-37. Epub 2011/02/08. doi: 10.1021/ct100578z. PubMed PMID: 26596171.

34. Sondergaard CR, Olsson MH, Rostkowski M, Jensen JH. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. J Chem Theory Comput. 2011;7(7):2284-95. Epub 2011/07/12. doi: 10.1021/ct200133y. PubMed PMID: 26606496.

35. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. GROMACS: fast, flexible, and free. J Comput Chem. 2005;26(16):1701-18. Epub 2005/10/08. doi: 10.1002/jcc.20291. PubMed PMID: 16211538.

36. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983;79(2):926-35. doi: 10.1063/1.445869.

37. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. The Journal of Physical Chemistry B. 1998;102(18):3586-616. doi: 10.1021/jp973084f.

38. Berendsen HJC, Postma JPM, Gunsteren WFv, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics. 1984;81(8):3684-90. doi: 10.1063/1.448118.

39. Evans DJ, Holian BL. The Nose–Hoover thermostat. The Journal of Chemical Physics. 1985;83(8):4069-74. doi: 10.1063/1.449071.

40. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics. 1981;52(12):7182-90. doi: 10.1063/1.328693.

41. Di Pierro M, Elber R, Leimkuhler B. A Stochastic Algorithm for the Isobaric–Isothermal Ensemble with Ewald Summations for All Long Range Forces. Journal of Chemical Theory and Computation. 2015;11(12):5624-37. doi: 10.1021/acs.jctc.5b00648.

42. Ewald PP. Die Berechnung optischer und elektrostatischer Gitterpotentiale. Annalen der Physik. 1921;369(3):253-87. doi: https://doi.org/10.1002/andp.19213690304.

43. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry. 1997;18(12):1463-72. doi: https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

44. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. Array programming with NumPy. Nature. 2020;585(7825):357-62. doi: 10.1038/s41586-020-2649-2.

45. McKinney W. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference. 2010;445. doi: 10.25080/Majora-92bf1922-00a.
46. Waskom MB, Olga; O'Kane, Drew; Hobson, Paul; Lukauskas, Saulius; Gemperline, David C; Augspurger, Tom; Halchenko, Yaroslav; Cole, John B; Warmenhoven, Jordi; de Ruiter, Julian; Pye, Cameron; Hoyer, Stephan; Vanderplas, Jake; Villalba, Santi; Kunter, Gero; Quintero, Eric; Bachant, Pete; Martin, Marcel; Qalieh, Adel. mwaskom/seaborn: v0.8.1. 0.8.1 ed. Meyrin, Switzerland: Zenodo; 2017.
47. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007;9(3):90-5. doi: 10.1109/MCSE.2007.55.
48. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli AP, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik DV, Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold G-L, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y, SciPy C. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17(3):261-72. doi: 10.1038/s41592-019-0686-2.
49. Beckstein OD, Jan; Somogyi, Andy. GromacsWrapper: v0.3.3 (release-0.3.3). 0.3.3 ed. Meyrin, Switzerland: Zenodo; 2015.
50. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, Vanderplas J, Joly A, Holt B, Varoquaux G, editors. API design for machine learning software: experiences from the scikit-learn project. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases; 2013 2013-09-23; Prague, Czech Republichttps://hal.inria.fr/hal-00856511/document https://hal.inria.fr/hal-00856511/file/paper.pdf.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.
52. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. The Journal of Open Source Software. 2018;3(24):638. doi: https://doi.org/10.21105/joss.00638.
53. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67(2):301-20. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.
54. Rivoltini L, Squarcina P, Loftus DJ, Castelli C, Tarsini P, Mazzocchi A, Rini F, Viggiano V, Belli F, Parmiani G. A superagonist variant of peptide MART1/Melan A27-35 elicits anti-

melanoma CD8+ T cells with enhanced functional characteristics: implication for more effective immunotherapy. Cancer Res. 1999;59(2):301-6. Epub 1999/02/02. PubMed PMID: 9927036.

55. Hellman LM, Foley KC, Singh NK, Alonso JA, Riley TP, Devlin JR, Ayres CM, Keller GLJ, Zhang Y, Vander Kooi CW, Nishimura MI, Baker BM. Improving T Cell Receptor On-Target Specificity via Structure-Guided Design. Mol Ther. 2019;27(2):300-13. Epub 2019/01/09. doi: 10.1016/j.ymthe.2018.12.010. PubMed PMID: 30617019; PMCID: PMC6369632.

56. Solbach F, Bernardi A, Bansal S, Budamagunta MS, Krep L, Leonhard K, Voss JC, Lam KS, Faller R. Determining structure and action mechanism of LBF14 by molecular simulation. Journal of Biomolecular Structure and Dynamics. 2021:1-12. doi: 10.1080/07391102.2021.1967783.

57. Hwang W, Mallis RJ, Lang MJ, Reinherz EL. The alphabetaTCR mechanosensor exploits dynamic ectodomain allostery to optimize its ligand recognition site. Proc Natl Acad Sci U S A. 2020;117(35):21336-45. Epub 2020/08/17. doi: 10.1073/pnas.2005899117. PubMed PMID: 32796106; PMCID: PMC7474670.

58. Huang Y, Harris BS, Minami SA, Jung S, Shah PS, Nandi S, McDonald KA, Faller R. SARS-Cov-2 Spike binding to ACE2 is stronger and longer ranged with glycans. bioRxiv. 2021:2021.07.15.452507. doi: 10.1101/2021.07.15.452507.

59. Welch DA, Woehl TJ, Park C, Faller R, Evans JE, Browning ND. Understanding the Role of Solvation Forces on the Preferential Attachment of Nanoparticles in Liquid. ACS Nano. 2016;10(1):181-7. doi: 10.1021/acsnano.5b06632.

60. Xiong Y, Karuppanan K, Bernardi A, Li Q, Kommineni V, Dandekar AM, Lebrilla CB, Faller R, McDonald KA, Nandi S. Effects of N-Glycosylation on the Structure, Function, and Stability of a Plant-Made Fc-Fusion Anthrax Decoy Protein. Frontiers in Plant Science. 2019;10(768). doi: 10.3389/fpls.2019.00768.

61. Martínez L. Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis. PLOS ONE. 2015;10(3):e0119264. doi: 10.1371/journal.pone.0119264.

62. Trunk GV. A problem of dimensionality: a simple example. IEEE Trans Pattern Anal Mach Intell. 1979;1(3):306-7. Epub 1979/03/01. doi: 10.1109/tpami.1979.4766926. PubMed PMID: 21868861.

63. McLachlan GJ. Discriminant analysis and statistical pattern recognition. Hoboken, N.J.2004. xv, 526 p. p.

64. Zollanvari A, James AP, Sameni R. A Theoretical Analysis of the Peaking Phenomenon in Classification. Journal of Classification. 2020;37(2):421-34. doi: 10.1007/s00357-019-09327-3.

65. Danska JS, Livingstone AM, Paragas V, Ishihara T, Fathman CG. The presumptive CDR3 regions of both T cell receptor alpha and beta chains determine T cell specificity for myoglobin peptides. J Exp Med. 1990;172(1):27-33. Epub 1990/07/01. doi: 10.1084/jem.172.1.27. PubMed PMID: 1694219; PMCID: PMC2188142.

Supporting Information

The Atomic-Level Physiochemical Determinants of T Cell Receptor Dissociation Kinetics

Zachary A. Rollins[2], Jun Huang[4], Ilias Tagkopoulos[3], Roland Faller[2], and Steven C. George[1*]

Department of Biomedical Engineering [1], Department of Chemical Engineering [2], Department of Computer Science [3], University of California, Davis, Davis, California; Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL[4]

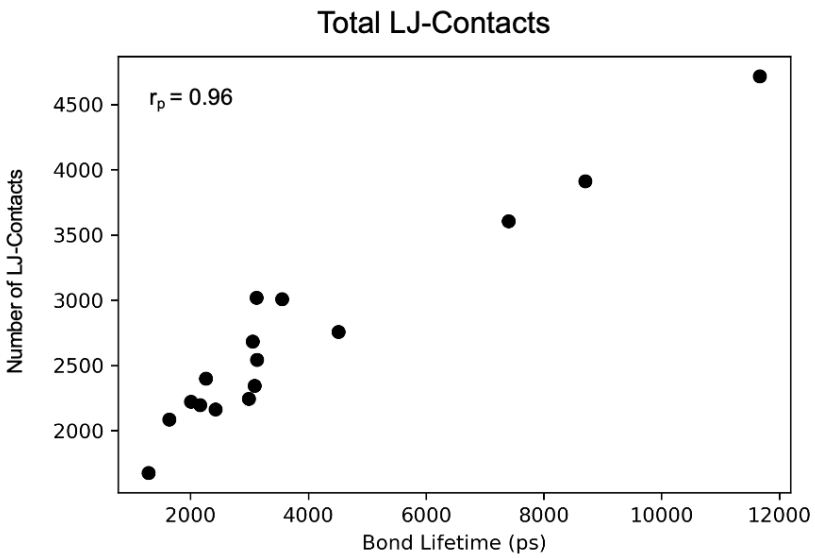

Figure 3—supplement 1: Quaternary Features vs Bond Lifetime. *Scatter plot of the total number of LJ-contacts vs bond lifetime for all TCR-pMHC pairs; the Pearson correlation coefficient is listed in the top left corner.*

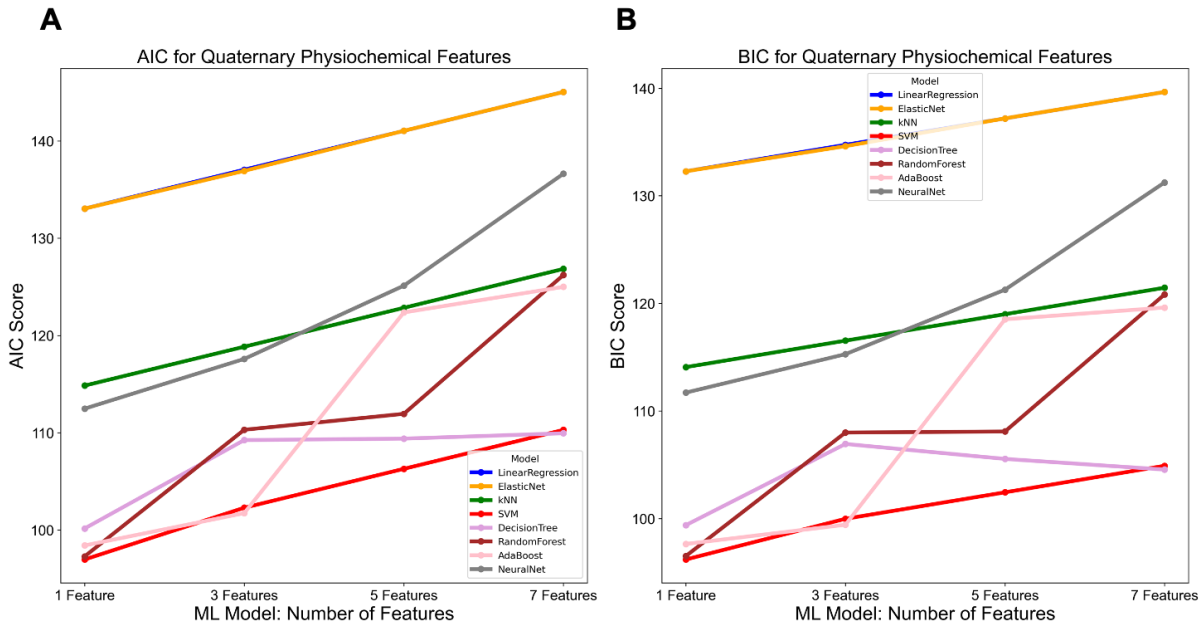

Figure 3—supplement 2: Akaike and Bayesian Information Criterion Scores for Quaternary Features. *The* ***(A)*** *Akaike Information Criterion (AIC) and* ***(B)*** *Bayesian Information Criterion (BIC) for all the quaternary feature sets (i.e., p = 1, 3, 5, and 7) and machine learning models (Linear Regression = blue, Elastic Net = orange, k-Nearest Neighbors = green, Support Vector Machines = red, Decision Tree = purple, Random Forest = brown, AdaBoost = pink, Neural Net = gray).*

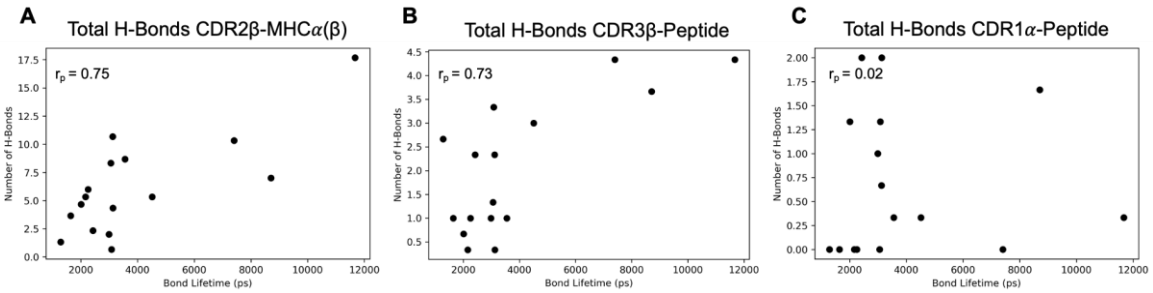

Figure 4—supplement 1: Secondary Features vs Bond Lifetime. *Scatter plots of total H-bonds for* ***(A)*** *CDR2β-MHCα(β),* ***(B)*** *CDR3β-Peptide, and* ***(C)*** *CDR1α-Peptide vs Bond Lifetime for all TCR-pMHC pairs; the Pearson correlation coefficient is listed in the top left corner.*

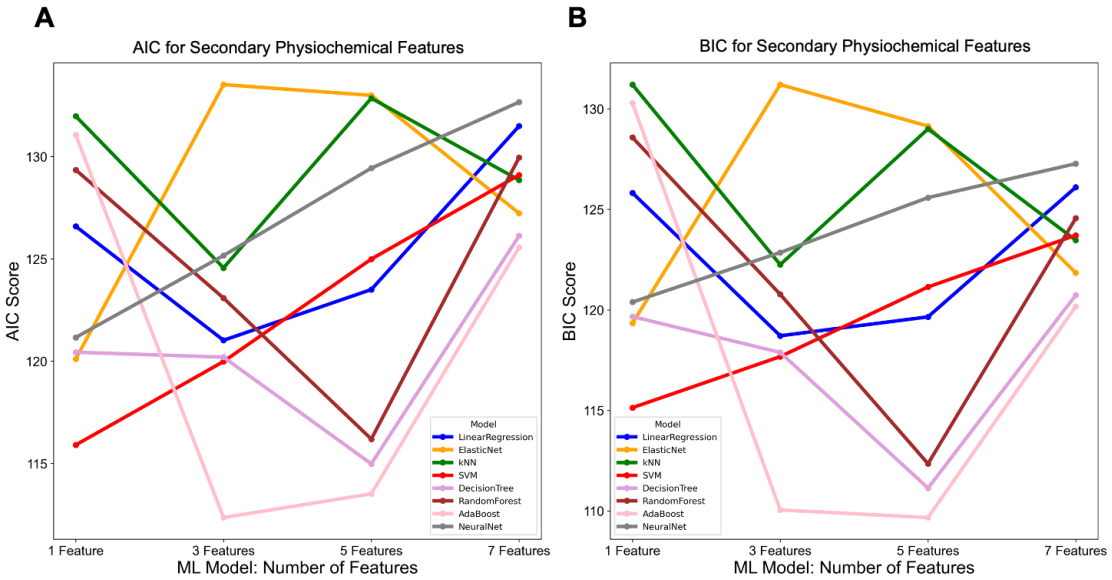

Figure 4—supplement 2: Akaike and Bayesian Information Criterion Scores for Secondary Features. (A) *The Akaike Information Criterion (AIC) and* ***(B)*** *Bayesian Information Criterion (BIC) scores for all secondary feature sets (i.e., p = 1, 3, 5, and 7) and machine learning models (Linear Regression = blue, Elastic Net = orange, k-Nearest Neighbors = green, Support Vector Machines = red, Decision Tree = purple, Random Forest = brown, AdaBoost = pink, Neural Net = gray).*

Peptide (Amino Acid Sequence)
MART1 (AAGIGILTV)
L1 (LAGIGILTV)
GVA (GAGIGVLTA)
8S (AAGIGILSV)
6V (AAGIGVLTV)
hCD9 (AVGIGIAVV)
HSV1gp3 (IAGIGILAI)
ImrA (LAGIGLIAA)
Mtub1 (LGGLGLFFA)
Mtub2 (IAGPGTITL)
5D (AAGIDILTV)
6Y (AAGIGYLTV)
6H (AAGIGHLTV)
5H (AAGIHILTV)
4Y (AAGYGILTV)
3F (AAFIGILTV)
2P (APGIGILTV)

Figure 1—supplement 1: *Peptides used in SMD simulations, including their amino acid sequences.*

Quaternary Features
Total H-Bonds
Total LJ-Contacts
Instantaneous H-Bonds
Instantaneous LJ-Contacts
Reaction Distance
Maximum Frequency Distance
SASA TCR
SASA pMHC
RMSF TCR
RMSF pMHC
Gyration of TCR
***x*-Gyration of TCR**
***y*-Gyration of TCR**
***z*-Gyration of TCR**
Gyration of pMHC
***x*-Gyration of pMHC**
***y*-Gyration of pMHC**
***z*-Gyration of pMHC**

Figure 1— supplement 2: *Quaternary Features.*

Secondary Features	
CDR3 Distance	Total LJ-Contacts CDR2α-MHCα(α)
SASA CDR1α	Total LJ-Contacts CDR3α-Peptide
SASA CDR2α	Total LJ-Contacts CDR3α-MHCα(α)
SASA CDR3α	Total LJ-Contacts CDR1β-Peptide
SASA CDR1β	Total LJ-Contacts CDR1β-MHCα(β)
SASA CDR2β	Total LJ-Contacts CDR2β-Peptide
SASA CDR3β	Total LJ-Contacts CDR2β-MHCα(β)
SASA MHCα(α)	Total LJ-Contacts CDR3β-Peptide
SASA MHCα(β)	Total LJ-Contacts CDR3β-MHCα(β)
SASA Peptide	Gyration CDR1α
RMSF CDR1α	*x*-Gyration CDR1α
RMSF CDR2α	*y*-Gyration CDR1α
RMSF CDR3α	*z*-Gyration CDR1α
RMSF CDR1β	Gyration CDR2β
RMSF CDR2β	*x*-Gyration CDR2β
RMSF CDR3β	*y*-Gyration CDR2β
RMSF MHCα(α)	*z*-Gyration CDR2β
RMSF MHCα(β)	Gyration CDR3β
RMSF Peptide	*x*-Gyration CDR3β
Total H-Bonds CDR1α-Peptide	*y*-Gyration CDR3β
Total H-Bonds CDR1α-MHCα(α)	*z*-Gyration CDR3β
Total H-Bonds CDR2α-Peptide	Gyration MHCα(α)
Total H-Bonds CDR2α-MHCα(α)	*x*-Gyration MHCα(α)
Total H-Bonds CDR3α-Peptide	*y*-Gyration MHCα(α)
Total H-Bonds CDR3α-MHCα(α)	*z*-Gyration MHCα(α)
Total H-Bonds CDR1β-Peptide	Gyration MHCα(β)
Total H-Bonds CDR1β-MHCα(β)	*x*-Gyration MHCα(β)
Total H-Bonds CDR2β-Peptide	*y*-Gyration MHCα(β)
Total H-Bonds CDR2β-MHCα(β)	*z*-Gyration MHCα(β)
Total H-Bonds CDR3β-Peptide	Gyration Peptide
Total H-Bonds CDR3β-MHCα(β)	*x*-Gyration Peptide
Total LJ-Contacts CDR1α-Peptide	*y*-Gyration Peptide
Total LJ-Contacts CDR1α-MHCα(α)	*z*-Gyration Peptide
Total LJ-Contacts CDR2α-Peptide	

Figure 1—supplement 3: *Secondary Features.*

A

Bond Lifetime Correlation	Total Contacts	Total H-Bonds	Instant Contacts	Instant H-Bonds	RXN Distance	GYRx_pMHC	GYRx_TCR	RMSF_TCR	GYRy_pMHC	GYRz_TCR
Pearson (r_p)	0.96	0.82	0.63	0.57	0.31	-0.19	-0.56	0.45	-0.10	0.29
Spearman (r_s)	0.93	0.78	0.64	0.69	0.25	-0.09	-0.41	0.42	0.0045	0.31

B

Bond Lifetime Correlation	Total H-Bonds CDR2β_MHCα(β)	Total Contacts CDR2β_MHCα(β)	Total H-Bonds CDR3β_pep	Total Contacts CDR3β_pep	Total Contacts CDR1α_pep	Total H-Bonds CDR1α_pep	Total Contacts CDR3β_MHCα(β)	SASA_MHCα	Total Contacts CDR3α_pep	GYRz_MHCα(β)
Pearson (r_p)	0.75	0.80	0.73	0.76	0.39	0.02	0.32	0.30	0.62	0.63
Spearman (r_s)	0.62	0.66	0.57	0.66	0.49	0.28	0.41	0.35	0.60	0.24

Figure 3—supplement 3: *Pearson correlation and Spearman rank correlation coefficients. This includes correlation coefficients for the list of top ten Quaternary Features* ***(A)*** *and the list of top ten Secondary Features* ***(B)****.*

	Model	Mean	STD	Log-Likelihood	Hyperparameters	AIC	BIC	t	p
1 Q-Feature	LinearRegression	1605.5	1451.4	-65.5	{'alpha': 0.0001, 'power': 0}	133.0	132.3	-	-
	ElasticNet	1605.5	1451.4	-65.5	{'alpha': 0.0, 'l1_ratio': 0.0}	133.0	132.3	-	-
	kNN	683.0	528.0	-56.4	{'leaf_size': 1, 'n_neighbors': 7, 'p': 1}	114.8	114.1	-	-
	SVM	559.7	195.5	-47.5	{'C': 100, 'degree': 3, 'gamma': 10, 'kernel': 'linear'}	97.0	96.2	-	-
	DecisionTree	689.2	233.5	-49.1	{'max_depth': 7, 'max_features': 'log2', 'max_leaf_nodes': 7, 'splitter': 'random'}	100.2	99.4	-	-
	RandomForest	702.3	199.0	-47.6	{'max_features': 'log2', 'n_estimators': 10}	97.3	96.5	-	-
	AdaBoost	779.6	211.8	-48.2	{'loss': 'linear', 'n_estimators': 10}	98.4	97.6	-	-
	NeuralNet	967.7	462.9	-55.2	{'activation': 'identity', 'alpha': 10.0, 'hidden_layer_sizes': 7}	112.5	111.7	-	-
3 Q-Feature	LinearRegression	1605.4	1451.4	-65.5	{'alpha': 10.0, 'power': 0}	137.0	134.7	0.00	0.50
	ElasticNet	1602.7	1441.6	-65.5	{'alpha': 0.01, 'l1_ratio': 0.61}	136.9	134.6	0.00	0.50
	kNN	683.0	528.0	-56.4	{'leaf_size': 1, 'n_neighbors': 7, 'p': 1}	118.8	116.5	0.00	0.50
	SVM	562.0	210.4	-48.1	{'C': 0.1, 'degree': 3, 'gamma': 10, 'kernel': 'linear'}	102.3	100.0	0.02	0.49
	DecisionTree	789.4	309.7	-51.6	{'max_depth': 4, 'max_features': 'auto', 'max_leaf_nodes': 8, 'splitter': 'random'}	109.2	106.9	0.78	0.22
	RandomForest	810.5	328.5	-52.2	{'max_features': 'auto', 'n_estimators': 410}	110.3	108.0	0.85	0.21
	AdaBoost	745.1	204.1	-47.9	{'loss': 'exponential', 'n_estimators': 10}	101.7	99.4	0.35	0.36
	NeuralNet	997.5	492.4	-55.8	{'activation': 'identity', 'alpha': 1e-05, 'hidden_layer_sizes': 4}	117.6	115.3	0.13	0.45
5 Q-Feature	LinearRegression	1605.5	1451.1	-65.5	{'alpha': 100.0, 'power': 0}	141.0	137.2	0.00	0.50
	ElasticNet	1605.5	1451.4	-65.5	{'alpha': 10.0, 'l1_ratio': 0.0}	141.0	137.2	0.00	0.50
	kNN	683.0	528.0	-56.4	{'leaf_size': 1, 'n_neighbors': 7, 'p': 1}	122.8	119.0	0.00	0.50
	SVM	561.9	210.4	-48.1	{'C': 0.1, 'degree': 3, 'gamma': 10, 'kernel': 'linear'}	106.3	102.4	0.00	0.50
	DecisionTree	802.5	250.0	-49.7	{'max_depth': 4, 'max_features': 'auto', 'max_leaf_nodes': 4, 'splitter': 'random'}	109.4	105.5	0.10	0.46
	RandomForest	876.0	288.1	-51.0	{'max_features': 'auto', 'n_estimators': 60}	111.9	108.1	0.45	0.33
	AdaBoost	795.6	514.2	-56.2	{'loss': 'exponential', 'n_estimators': 510}	122.4	118.5	0.27	0.39
	NeuralNet	1063.8	599.2	-57.6	{'activation': 'identity', 'alpha': 0.001, 'hidden_layer_sizes': 6}	125.1	121.3	0.26	0.40
7 Q-Feature	LinearRegression	1605.7	1451.0	-65.5	{'alpha': 100.0, 'power': 0}	145.0	139.7	0.00	0.50
	ElasticNet	1605.7	1451.0	-65.5	{'alpha': 100.0, 'l1_ratio': 0.0}	145.0	139.7	0.00	0.50
	kNN	683.0	528.0	-56.4	{'leaf_size': 1, 'n_neighbors': 7, 'p': 1}	126.8	121.5	0.00	0.50
	SVM	561.9	210.5	-48.1	{'C': 0.1, 'degree': 3, 'gamma': 10, 'kernel': 'linear'}	110.3	104.9	0.00	0.50
	DecisionTree	926.4	206.4	-48.0	{'max_depth': 6, 'max_features': 'auto', 'max_leaf_nodes': 9, 'splitter': 'random'}	109.9	104.6	1.15	0.13
	RandomForest	1052.3	510.0	-56.1	{'max_features': 'auto', 'n_estimators': 260}	126.2	120.8	0.90	0.19
	AdaBoost	969.2	476.6	-55.5	{'loss': 'exponential', 'n_estimators': 860}	125.0	119.6	0.74	0.23
	NeuralNet	962.0	909.4	-61.3	{'activation': 'identity', 'alpha': 0.0001, 'hidden_layer_sizes': 9}	136.6	131.2	0.28	0.39

Figure 3—supplement 4: Best Machine Learning Models after Hyperparameter Optimization for Quaternary Features. *Table includes the best performing model hyperparameters as well as the mean and standard deviation from repeated threefold cross validation. Akaike and Bayesian Information Criterion are calculated for each model and feature set, based on mean absolute error standard deviation from repeated threefold cross validation, to assess the improved accuracy with increasing complexity. The respective algorithms are statistically compared across feature sets using a one-tailed student's t-test.*

	Model	Mean	STD	Log-Likelihood	Hyperparameters	AIC	BIC	t	p
1 S-Feature	LinearRegression	2301.1	1014.1	-62.3	{'alpha': 0.01, 'power': 0}	126.6	125.8	-	-
	ElasticNet	2268.0	707.2	-59.1	{'alpha': 10.0, 'l1_ratio': 0.39}	120.1	119.3	-	-
	kNN	1832.8	1367.6	-65.0	{'leaf_size': 1, 'n_neighbors': 1, 'p': 1}	132.0	131.2	-	-
	SVM	1062.0	560.1	-57.0	{'C': 1000, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf'}	115.9	115.1	-	-
	DecisionTree	1392.5	720.4	-59.2	{'max_depth': 6, 'max_features': 'auto', 'max_leaf_nodes': 6, 'splitter': 'random'}	120.4	119.7	-	-
	RandomForest	1935.5	1181.9	-63.7	{'max_features': 'sqrt', 'n_estimators': 10}	129.3	128.6	-	-
	AdaBoost	1566.0	1299.6	-64.5	{'loss': 'linear', 'n_estimators': 160}	131.1	130.3	-	-
	NeuralNet	3314.9	749.8	-59.6	{'activation': 'identity', 'alpha': 10.0, 'hidden_layer_sizes': 9}	121.2	120.4	-	-
3 S-Feature	LinearRegression	948.8	595.9	-57.5	{'alpha': 1.0, 'power': 2}	121.0	118.7	3.45	0.00
	ElasticNet	1692.9	1192.9	-63.8	{'alpha': 10.0, 'l1_ratio': 0.74}	133.5	131.2	1.24	0.12
	kNN	1040.8	725.4	-59.3	{'leaf_size': 1, 'n_neighbors': 2, 'p': 1}	124.6	122.3	1.53	0.07
	SVM	1005.6	562.5	-57.0	{'C': 1000, 'degree': 3, 'gamma': 0.01, 'kernel': 'rbf'}	120.0	117.7	0.21	0.42
	DecisionTree	865.7	569.1	-57.1	{'max_depth': 6, 'max_features': 'auto', 'max_leaf_nodes': 3, 'splitter': 'random'}	120.2	117.9	1.72	0.05
	RandomForest	969.4	668.2	-58.5	{'max_features': 'auto', 'n_estimators': 310}	123.1	120.8	2.13	0.02
	AdaBoost	966.7	368.2	-53.2	{'loss': 'square', 'n_estimators': 10}	112.4	110.0	1.33	0.10
	NeuralNet	3312.2	750.1	-59.6	{'activation': 'identity', 'alpha': 1.0, 'hidden_layer_sizes': 8}	125.2	122.9	0.01	0.50
5 S-Feature	LinearRegression	940.3	547.6	-56.8	{'alpha': 1.0, 'power': 2}	123.5	119.7	0.03	0.49
	ElasticNet	1444.9	927.8	-61.5	{'alpha': 10.0, 'l1_ratio': 0.84}	133.0	129.1	0.49	0.31
	kNN	1719.5	920.6	-61.4	{'leaf_size': 1, 'n_neighbors': 2, 'p': 1}	132.9	129.0	1.74	0.05
	SVM	1061.3	594.8	-57.5	{'C': 1000, 'degree': 3, 'gamma': 0.1, 'kernel': 'rbf'}	125.0	121.1	0.20	0.42
	DecisionTree	833.7	341.2	-52.5	{'max_depth': 2, 'max_features': 'log2', 'max_leaf_nodes': 3, 'splitter': 'random'}	115.0	111.1	0.14	0.44
	RandomForest	1270.9	364.9	-53.1	{'max_features': 'sqrt', 'n_estimators': 110}	116.2	112.3	1.19	0.13
	AdaBoost	1087.0	314.5	-51.8	{'loss': 'square', 'n_estimators': 910}	113.5	109.7	0.75	0.23
	NeuralNet	3075.4	761.4	-59.7	{'activation': 'identity', 'alpha': 100.0, 'hidden_layer_sizes': 8}	129.4	125.6	0.66	0.26
7 S-Feature	LinearRegression	992.3	683.5	-58.7	{'alpha': 0.001, 'power': 3}	131.5	126.1	0.18	0.43
	ElasticNet	2238.2	539.5	-56.6	{'alpha': 100.0, 'l1_ratio': 0.0}	127.2	121.8	2.22	0.02
	kNN	944.1	590.3	-57.4	{'leaf_size': 1, 'n_neighbors': 2, 'p': 2}	128.9	123.5	2.13	0.02
	SVM	1108.1	598.5	-57.5	{'C': 100, 'degree': 3, 'gamma': 0.0001, 'kernel': 'rbf'}	129.1	123.7	0.17	0.43
	DecisionTree	1053.1	507.3	-56.1	{'max_depth': 4, 'max_features': 'sqrt', 'max_leaf_nodes': 4, 'splitter': 'best'}	126.1	120.7	1.08	0.15
	RandomForest	1241.5	627.4	-58.0	{'max_features': 'log2', 'n_estimators': 10}	129.9	124.6	0.12	0.45
	AdaBoost	1110.5	491.6	-55.8	{'loss': 'linear', 'n_estimators': 60}	125.6	120.2	0.12	0.45
	NeuralNet	2430.6	729.4	-59.3	{'activation': 'identity', 'alpha': 0.1, 'hidden_layer_sizes': 8}	132.7	127.3	1.83	0.04

Figure 4—supplement 3: Best Machine Learning Models after Hyperparameter Optimization for Secondary Features. *Table includes the best performing model hyperparameters as well as the mean and standard deviation from repeated threefold cross validation. Akaike and Bayesian Information Criterion are calculated for each model and feature set, based on mean absolute error standard deviation from repeated threefold cross validation, to assess the improved accuracy with increasing complexity. The respective algorithms are statistically compared across feature sets using a one-tailed student's t-test.*