Appeared in: 10.1073/pnas.2007246117

Markov State Modeling Reveals Alternative Unbinding Pathways for Peptide-MHC Complexes

Jayvee R. Abella*[a], Dinler Antunes*[a], Kyle Jackson[b], Gregory Lizée[b], Cecilia Clementi[c,d], and Lydia E. Kavraki[a]

[a]Department of Computer Science, Rice University, Houston, TX, USA; [b]Department of Melanoma Medical Oncology - Research, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA; [c]Center for Theoretical Biological Physics, Rice University, Houston, TX, USA; [d]Department of Chemistry, Rice University, Houston, TX, USA

This manuscript was compiled on October 12, 2020

Peptide-MHC complexes are central components of the immune system, and understanding the mechanism behind stable peptide-MHC binding will aid the development of immunotherapies. While MHC binding is mostly influenced by the identity of the so-called anchor positions of the peptide, secondary interactions from non-anchor positions are known to play a role in complex stability. However, current HLA-binding prediction methods lack an atomistic analysis of the major conformational states of the system, and might underestimate the impact of secondary interactions. In this work, we present an atomically-detailed analysis of peptide-MHC binding that can reveal the contributions of any interaction towards stability. We propose a simulation framework that uses both umbrella sampling and adaptive sampling to generate a Markov state model (MSM) for a peptide from SARS-CoV (QFKDNVILL), bound to one of the most prevalent MHC receptors in humans (HLA-A*24:02). While our model reaffirms the importance of the anchor positions of the peptide in establishing stable interactions for binding, our model also reveals the underestimated importance of position 4 (p4), a non-anchor position. We confirmed our results by simulating the impact of specific peptide mutations, and validated these predictions through competitive binding assays. Remarkably, by comparing the MSM of the wild-type system with those of the D4A and D4P mutations, our modeling reveals stark differences in unbinding pathways. The analysis presented here can be applied to any peptide-MHC complex of interest with a 3D model as input, representing an important step towards comprehensive and accurate modeling of the MHC class I pathway.

peptide-MHC binding stability | Markov state modeling | adaptive sampling | competitive binding assay

1. Introduction

Class I major histocompatibility complexes (MHCs), also known as HLAs in humans, are proteins that bind to intracellular peptides and present them at the cellular surface (1). In the endoplasmic reticulum, MHCs are loaded with peptides of length 8–11 amino acids derived from cleaved intracellular proteins. Then the combined peptide-MHC complex is transported to the cell surface to be inspected by surveilling T-cells. T-cell activation normally occurs when a cell presents peptides not found in healthy cells, triggering an immune response. Current efforts in immunotherapy aim to amplify this mechanism to target diseased cells (i.e., infected or tumoral). Since every patient has a different set of MHCs, this problem must be addressed in a personalized manner, i.e., by identifying disease-specific peptides that can bind to the MHCs of a particular patient or to MHCs that will provide broad population coverage.

Therefore, a prerequisite for T-cell activation, or immunogenicity, is stable binding to occur between a given peptide and MHC (2). Peptides bound to MHCs on the cell surface can be identified directly using mass spectrometry, and experiments have been curated into databases such as SysteMHC Atlas (3). Additionally, the binding affinities of peptides can be measured with competitive binding assays, for example, which can provide IC50 values. In turn, results from binding assay experiments have been curated into databases such as the Immune Epitope Database (IEDB) (4). This accumulation of experimental data has led to the popularity of sequence-based methods for peptide-MHC binding prediction. These methods are based on machine learning, typically with neural networks, trained on sequences of known peptide-MHC pairs and can rapidly predict binding affinity (5–8).

Moving beyond a simple measurement or prediction of binding, uncovering the molecular mechanisms for strong binding usually starts with an analysis of a structure of the bound complex. Structures can be from one of the few hundred crystal structures available at PDB, or modeled with a docking-based approach (9–14). However, an analysis of a single conformation may be misleading due to the flexibility of the structure (15), and the dynamics of peptide-MHC binding must be probed. Along this direction, experimental methods such as NMR (16, 17), hydrogen/deuterium exchange (18), and fluorescence anisotropy (19), have been used to gain insight into the flexibility of peptide-MHC complexes. However, these experimental methods have particular limitations regarding the cost, the size of the system, and the resolution of the results.

As an alternative, molecular simulations can be used to analyze the stability and dynamics of peptide-MHC binding.

Significance Statement

Peptide binding to MHC receptors is part of a central biological process that enables our immune system to attack diseased cells. Here we use molecular simulations to illuminate the mechanisms driving stable peptide-MHC binding. Our simulation framework produces an atomistic model of the unbinding dynamics for a given peptide-MHC, which quantifies transitions between the major states of the system (bound, intermediate, and unbound). We applied this framework to study the binding of a SARS-CoV peptide to the HLA-A*24:02 receptor. This work revealed the unexpected importance of peptide's position 4 in driving the stability of the complex, a finding with broader biomedical implications. Our methods can be applied to other peptide-MHC complexes, only requiring a 3D model as input.

*These authors contributed equally to this work

Corresponding Author: Lydia Kavraki (kavraki@rice.edu)

Such analysis can cover the major conformational states of the process, while providing atomistic details that cannot be currently achieved with experimental methods. In this context, many simulation studies have focused on bound peptide-MHC complexes (20). Going even further, Ayres *et al.* built a simplified model for peptide flexibility in the binding site of a particular MHC (21), and Wan *et al.* used the MMPBSA method to compute binding free energy estimates from molecular dynamics (MD) (22). For that, they simulated both bound peptide-MHC conformations and fully unbound conformations (22). While simulating bound/unbound states may be enough for accurate binding affinity prediction, information on the intermediate states and the *transition* between states is lacking. In another study, a coarse-grained Monte Carlo based framework was developed for generating detachment pathways of peptides exiting the MHC binding site (23). These detachment pathways allow some analysis of the transition between bound and unbound states. However, the use of coarse-graining prevents atomic-level predictions of peptide-MHC interactions that could characterize the major states along the binding/unbinding pathways.

Here we propose an analysis that goes beyond previous simulation studies, capable of revealing all the molecular interactions that are driving the stability of a peptide-MHC complex. In other words, we provide a model that can capture all the major conformational states along the binding/unbinding pathway, as well as the transitions between those states, using atomistic MDs. Such models are known as Markov state models (MSMs) (24), and allow for the quantification of both binding affinity and stability for a given peptide-MHC complex (25–27). However, building MSMs of the whole binding process for peptide-MHCs, in atomic-level detail, is computationally challenging. MHCs are large systems comprised of about 380 residues, which contribute to the high computational cost of MD. More importantly, the typical timescales involved in the binding process are significantly longer than current MD simulations are capable of reaching within a reasonable timeframe. For instance, while the timesteps of typical full-atom MD simulations are on the order of femtoseconds, the half-life of the more stable peptide-MHC complexes reaches tens of hours (2).

To address the computational challenges, we propose a simulation framework for peptide-MHCs that splits the problem into two stages: an exploration stage and a connection stage. The exploration stage makes use of umbrella sampling (28), which is a well known technique that can accelerate the sampling along an appropriate reaction coordinate. The connection stage makes extensive use of the relatively newer class of methods called adaptive sampling (27, 29–33). Adaptive sampling works by iteratively performing short MD simulations in parallel. At each iteration, the next round of MD simulations are initialized using conformations that aim to optimize exploration using a restart strategy. The restart strategy selects the conformations using all the simulation data already performed up to the given iteration. Adaptive sampling methods are typically performed in conjunction with MSMs (30, 32). MSMs are built by defining states and counting transitions between states, producing a transition matrix that contains the transition probabilities. Thus, MSMs do not require each individual simulation to be long for construction, only long enough to be able to count transitions. Adaptive sampling methods combined with MSMs are becoming increasingly popular as a way to accelerate the sampling of MD, and recent studies have been investigating how to optimize its use (32–35).

As an example case, we focus this work in studying the binding of the viral peptide QFKDNVILL with the human MHC receptor HLA-A*24:02. The choice of this system is interesting in multiple regards. First, a crystal structure is available for this system (36), which we use to begin our modeling. Second, HLA-A*24:02 is one of the most prevalent HLA allotypes in the human population (4), being therefore highly relevant for several biomedical applications. Third, the displayed peptide is derived from the nucleocapsid protein of SARS-CoV, and this protein has over 90% sequence similarity with that of SARS-CoV-2 (37). Therefore, insights from this system may be relevant for the current and/or future coronavirus epidemics. Finally, the popular sequence-based predictor NetMHC4.0 (5) fails to correctly predict the binding affinity of this peptide, potentially neglecting the role of key secondary interactions.

Class I MHCs usually bind peptides through dominant inter-molecular interactions that typically involve the residues at both ends of the peptide (so called *anchor residues*). The chemical properties of deeper pockets in the MHC binding cleft determine the "identity" of the preferred anchor residues. As a consequence, we can usually summarize the binding profile of a particular MHC allotype by specifying the types of residues found in the anchor positions. For instance, IEDB data indicates that the anchor residues for peptides binding to HLA-A*24:02 are position 2 (p2 anchor) and the last residue (C-term anchor); with a preference for hydrophobic residues in both positions (4). In particular, the p2 anchor is preferentially a tryptophan (W) or tyrosine (Y), but the corresponding pocket can tolerate a phenylalanine (F). The C-term anchor is preferentially a phenylalanine (F), isoleucine (I), or tryptophan (W), but the corresponding pocket can also tolerate a leucine (L) or methionine (M). Note that the amino acid binding chart at IEDB does not indicate any relevant preferences for peptide positions p3-p6. Although anchor residues vary depending on the MHC allotype, middle positions are usually considered to be more exposed to T-cell interaction, and less relevant for peptide-MHC binding (38). Interestingly, the viral peptide QFKDNVILL, called *WT* in this work, has both anchor positions as "tolerated" residues. The lack of any *preferred* anchors might explain the very low binding affinity predicted by NetMHC4.0 for this complex (7,769.11 nM). While the strongest contacts in the *WT* system are likely to still be formed by the anchor residues, we are interested in the role of secondary interactions involving the other non-anchor peptide positions, which may play a larger role in the absence of strong primary anchors.

Thus, the objective of this work is to investigate the role of secondary interactions in the binding of QFKDNVILL to HLA-A*24:02. Using our proposed simulation framework (Fig. 1), we generate over 150 microseconds of MD data to build a MSM of the entire binding/unbinding process. Our model predicts that QFKDNVILL is capable of binding to HLA-A*24:02, and mutational analysis based on reweighting of this *WT* system reveals the importance of the non-anchor residue in position 4. Additional MSMs of two mutated peptide-variants (*D4A* and *D4P*), generated using around 500 microseconds

of total MD data, were used to predict the relative ranking of these 3 systems, and this ranking was confirmed using competitive binding assays. Detailed analysis of the MSMs for the three different systems has revealed both alternative peptide-unbinding pathways, as well as alternative ways in which p4 can affect peptide-MHC stability. Structural analysis of MHC-binders that lack canonical primary anchors, as the one described here, may provide the key to identify valuable peptide-targets that are being currently missed in vaccine development and T-cell-based immunotherapy efforts.

2. Results

A. New simulation framework enables building MSM for peptide-MHC binding/unbinding. A new simulation framework (Fig. 1) is used to generate MD data to build an MSM of the *WT* system. Characteristics of the exploration and connection stages for the *WT* system can be found in the SI Appendix (SI Appendix, Fig. S1). A total of 160 microseconds of aggregate simulation data was generated, where each simulation takes approximately 15 hours on 1 Tesla V100 GPU, taking about 2,600 GPU-hours total. Time-lagged independent components analysis (TICA) was performed to reduce the dimensionality of the conformations (39, 40). We keep the top two independent components, which adequately capture two different detachment pathways that the peptide takes to go from the native state to the unbound state (SI Appendix, Fig. S2 and S3). One component roughly represents the detachment of the N-term while the second represents the detachment of the C-term. After discretization of the TICA space into microstates, the discrete Transition-based Reweighting Analysis Method (dTRAM) (41) was used to combine the biased and unbiased trajectories from the two stages of the simulation framework into a final MSM (see Materials and Methods, and SI Appendix, Fig. S2–S4).

We partition the microstates into 5 states, which were defined to distinguish between the major metastable states along the binding pathway based on a previous study of detachment pathways (23). Detachment pathways are mainly distinguished by the order in which the anchor residues detach from the corresponding MHC pocket (23), which we captured in the MSM through TICA. The two endpoints of binding are the native state (State 0) and the unbound or dissociated state (State 4). The native state (State 0) is defined as the set of all microstates with an average all-atom RMSD of below 0.2 nm from the crystal structure. The unbound/dissociated state (State 4) is defined as the set of microstates where the minimum distance between the peptide and MHC is greater than 0.5 nm. The next two states define partially bound states where only a single anchor of the peptide is in the corresponding MHC pocket. N-term bound state (State 1) is defined as the set of non-native microstates where the center of mass of position 2 in the peptide is below 0.2 nm from the center of mass of the native position 2 location. C-term bound state (State 2) is defined as the set of non-native microstates where the center of mass of position 9 in the peptide is below 0.2 nm from the center of mass native position 9 location. State 3 defines all the other associated microstates which have the peptide in contact with the MHC. Typical conformations found within each of the 5 states can be found in Fig. 4.

The MSM for *WT* predicts that the native state is the most probable state ($P(\text{native state}) = \pi_0 = 0.906$), despite the lack of strong primary anchors. Therefore, our model predicts the stable binding of QFKDNVILL to HLA-A*24:02, which is in line with crystallographic evidence (36). The predicted free energy of binding was $\Delta G_{WT} = -7.19 \pm 1.02$ kJ/mol.

B. Mutational analysis of the *WT* MSM reveals the importance of peptide's position 4 towards binding. We used the MSM of the *WT* system to perform mutational analysis based on reweighting the state probabilities computed from the MSM, and predict the change to the binding affinity upon alanine mutation (Fig. 2). Unsurprisingly, the F2A and L9A mutations were predicted to be most disruptive to binding, as positions 2 and 9 are the primary anchor residues for this peptide. However, the D4A mutation was also predicted to be remarkably disruptive to peptide binding (Fig. 2). This implies that secondary interactions involving p4 must be particularly relevant for the binding of *WT*.

Table 1. Destabilization of the metastable states upon alanine mutation. The table contains the values $RT[\ln(Z_{wt}^{S_i}/Z_{wt}^{dissociated}) - \ln(Z_{mut}^{S_i}/Z_{mut}^{dissociated})]$ in kJ/mol (see Materials and Methods) for all associated states S_i. Computed values are all in reference to the dissociated state, so the values for State 4 would all be zero.

Mut.\State	0	1	2	3
F2A	38.7	37.7	7.3	6.7
D4A	14.9	17.5	3.5	4.6
L9A	19.8	1.1	15.9	8.7

We can decompose the effect of the alanine-exchanges across the different associated states (i.e., States 0, 1, 2, and 3) (Table 1). Mutating the anchor residues (i.e., p2 and p9) has the expected effect of destabilizing the states associated with the presence of these respective positions in the corresponding MHC pockets. In other words, for the F2A mutation, the native state (State 0) and the N-term bound state (State 1) are most destabilized, while for the L9A mutation, the native state and the C-term bound state (State 2) are most destabilized. The native state (State 0) and the N-term bound state (State 1) are also most destabilized for the *D4A* mutation. Given that this peptide is a 9-mer, position 4 is closer to the N-term side, and is likely playing a role in stabilizing the interactions from that end.

We can use the *WT* MSM to analyze the relevant intermolecular contacts by computing the probability that a given contact exists while the system is within a particular State (SI Appendix, Fig. S5–S8). In the native state (State 0), the aspartic acid in position 4 of the peptide (D4) was more likely to interact with MHC residues K66, Q155, Y159 and T163 (SI Appendix, Fig. S5). Given the 3D arrangement of the binding cleft (Fig. 5), the D4-K66 and D4-T163 interactions are not surprising. On the other hand, the contributions of Q155 and Y159 are less obvious, despite being predicted to be even more important for the N-term bound state (SI Appendix, Fig. S5).

The mutational analysis can be performed on the MHC side as well, and we used the MSM of the *WT* system to evaluate the impact of mutations Q155A and Y159A. Interestingly, the MSM predicts Y159A to have a similar detrimental impact on binding ($\Delta G_{Y159} = 4.86 \pm 0.77$ kJ/mol) as that observed for the D4A mutation. The same impact was not predicted for Q155A ($\Delta G_{Q155A} = -7.52 \pm 0.37$ kJ/mol). Visual inspection of conformations obtained from State 0 and State 1 indicate a network of hydrogen bonds involving D4 and MHC residues

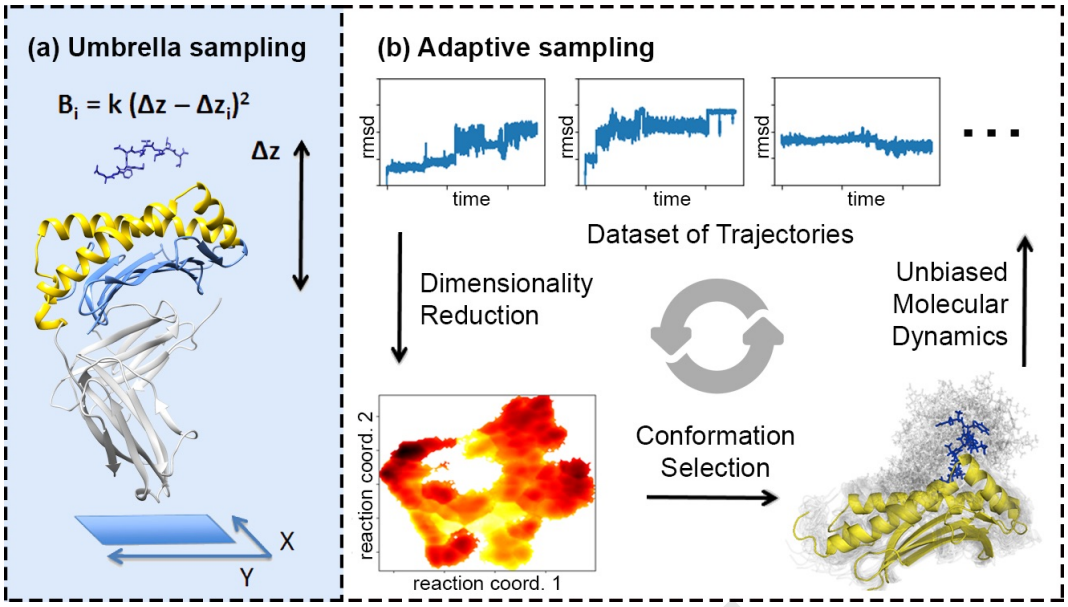


Fig. 1. Overview of the simulation framework. a) The exploration stage involves running umbrella sampling simulations along the *z-dist* reaction coordinate, which approximates the unbinding direction. B_i is the energy bias, while k is the force constant. The β-sheet floor of the MHC (light blue) is aligned to the XY-plane, then the Z-coordinate is used to define *z-dist*. The truncated portion of the MHC (light gray) is not included in any of the simulations. b) The connection stage involves running unbiased simulations in an adaptive sampling fashion until most of the states are connected. Restarting conformations are chosen by analyzing the trajectories in a dimensionality-reduced space using TICA that adequately capture the binding/unbinding pathway. Then the selection of conformations is biased towards the less densely sampled regions of the TICA space.

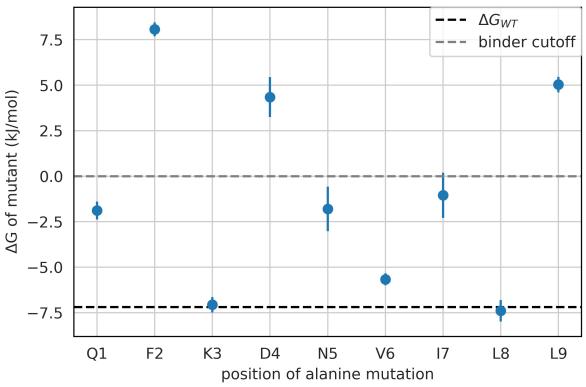


Fig. 2. $\Delta\Delta G$ predictions from the mutational analysis. The black dotted line represents the predicted ΔG_{WT} of -7.19 kJ/mol. The gray dotted line represents the separation between predicted binders and nonbinders. Alanine mutations in positions 2, 4, and 9 are all predicted to significantly impair binding, while alanine mutations in positions 1, 5, and 7 are predicted to reduce the binding affinity.

K66 and T163. Due to the side chain flexibility of D4, direct hydrogen bonds between D4-Q155 and D4-Y159 can also be observed in some conformations.

C. MSMs of *D4A* and *D4P* indicate alternative roles for p4. To confirm the dominant role of hydrogen bonds on the beneficial role of p4 for peptide binding, we created MSMs with two peptide variants: *D4A* and *D4P*. Characteristics of the exploration and connection stages for the *D4A* system can be found in the SI Appendix (SI Appendix, Fig. S9). A total of 213 microseconds of aggregate simulation data was used to build the MSM (see Materials and Methods, and SI Appendix, Fig. S10–S12), taking approximately 3,000 GPU-hours to complete. Our model for *D4A* predicts that the unbound state is the most probable state (P(unbound state) $= \pi_4 = 0.601$). We predict $\Delta G_{D4A} = 1.02 \pm 1.01$ kJ/mol, thus corroborating the mutational analysis prediction based on the *WT* network (Fig. 2), and predicting QFKANVILL to be a much weaker binder to HLA-A*24:02.

Characteristics of the exploration and connection stages for the *D4P* system can be found in the SI Appendix (SI Appendix, Fig. S13). A total of 293 microseconds of aggregate simulation data was used to build the MSM (see Materials and Methods, and SI Appendix, Fig. S14–S16), taking approximately 4,300 GPU-hours to complete. By replacing the flexible polar D4 with a rigid nonpolar P4, we expected to observe similar results to that of *D4A*. Surprisingly, the resulting MSM predicted *D4P* to be a stronger binder ($\Delta G_{D4P} = -8.01 \pm 0.18$ kJ/mol) than *WT*. We also evaluated the impact of the MHC mutations Q155A and Y159A using the MSM of *D4P*, but these mutations were not predicted to affect the binding of the peptide. Taken together, these results indicate that P4 benefits peptide-MHC binding through a mechanism that is different from that observed for D4 (i.e., does not rely on hydrogen bonds with the aforementioned MHC residues).

D. Competitive binding assays confirm predicted ranking of relative binding affinities. To validate our MSM-derived predictions we performed competitive binding assays with *WT*, *D4A* and *D4P* (Fig. 3). First, QFKDNVILL (*WT*) shows partial inhibition across a variety of concentrations ($IC50_{WT}$ = 1,600 nM), but does not reach the level of the positive control. This confirms the MSM prediction of weak yet stable binding of *WT* towards HLA-A*24:02. Note that NetMHC4.0 not only predicts this peptide to be a much weaker binder (7,769

nM), but also predicts *D4A* to be a stronger binder (4,154 nM). However, our binding assay with *D4A* shows little to no inhibition across concentrations ($IC50_{D4A} > 6{,}000$ nM), thus confirming the MSM prediction that this mutation significantly impairs binding to HLA-A*24:02. Finally, the binding assay of *D4P* confirmed the MSM prediction that this mutation in fact enhances binding to HLA-A*24:02 ($IC50_{D4P} = 600$ nM).

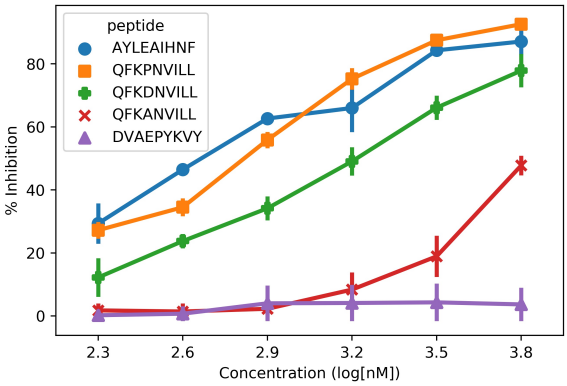


Fig. 3. Competitive binding assays to determine the ranking of *WT*, *D4A* and *D4P*. Based on the relative position of the *WT* curve (green plus) versus the positive control (blue circle), we see that QFKDNVILL is indeed a weak binder to HLA-A*24:02 ($IC50_{WT} = 1{,}600$ nM). Upon mutation of D4 to an alanine, inhibition is significantly reduced ($IC50_{D4A} > 6{,}000$ nM) as the *D4A* curve (red cross) is most similar to the negative control (purple triangle). Upon mutation of D4 to a proline, inhibition is increased ($IC50_{D4P} = 600$ nM) as the *D4P* curve (orange square) is most similar to the positive control.

E. MSM flux analysis reveal alternative unbinding pathways. By comparing the *WT* MSM with the MSM of the mutants (*D4A* and *D4P*), we can identify differences in unbinding pathways. This analysis was done by computing the percentage of flux that goes from the native state (State 0) to the unbound state (State 4). Fig. 4a shows that the majority of *WT* unbinding pathways first detach from the C-term end. However, upon D4A mutation, the majority of unbinding pathways detach first from the N-term end (Fig. 4b). Note that both pathways are accessible for the *D4A* system, but the lack of stabilizing interactions involving position 4 allows for the alternate unbinding route. In addition, *D4A* prefers to stay in the unbound state (State 4), as opposed to *WT*'s preference of staying in the bond state (State 0). The stabilizing effect of D4 on *WT* seems primarily related to the interaction with MHC positions K66, T163, Y159 and Q155, respectively. Interestingly, these positions are mostly conserved across HLA allotypes (SI Appendix, Fig. S17). In particular, D4 interactions with K66 and T163 can be easily observed both in State 0 and State 1 (Fig. 5), which is consistent with the role of stabilizing the N-term portion of the peptide.

The *D4P* mutation revealed a different picture. Like *D4A*, the *D4P* system has a preference to unbind from the N-term first. In fact, all sampled unbinding trajectories for the *D4P* system showed the N-term detaching first, and there were zero trajectories sampled where the C-term detaches first (i.e., although the MSM included transitions from State 0 to State 1, and from State 1 back to State 0, none of the trajectories included transitions from State 1 to States 3 and 4). However, unlike *D4A*, *D4P* is a more stable binder, and the various bound states (States 0, 1 and 2) have higher equilibrium probabilities (Fig. 6a). Therefore, the inability of *D4P* to detach first from the C-term side represents a decrease in unbinding options of the system, even offsetting any destabilizing effect from the lack of a salt-bridge with p4.

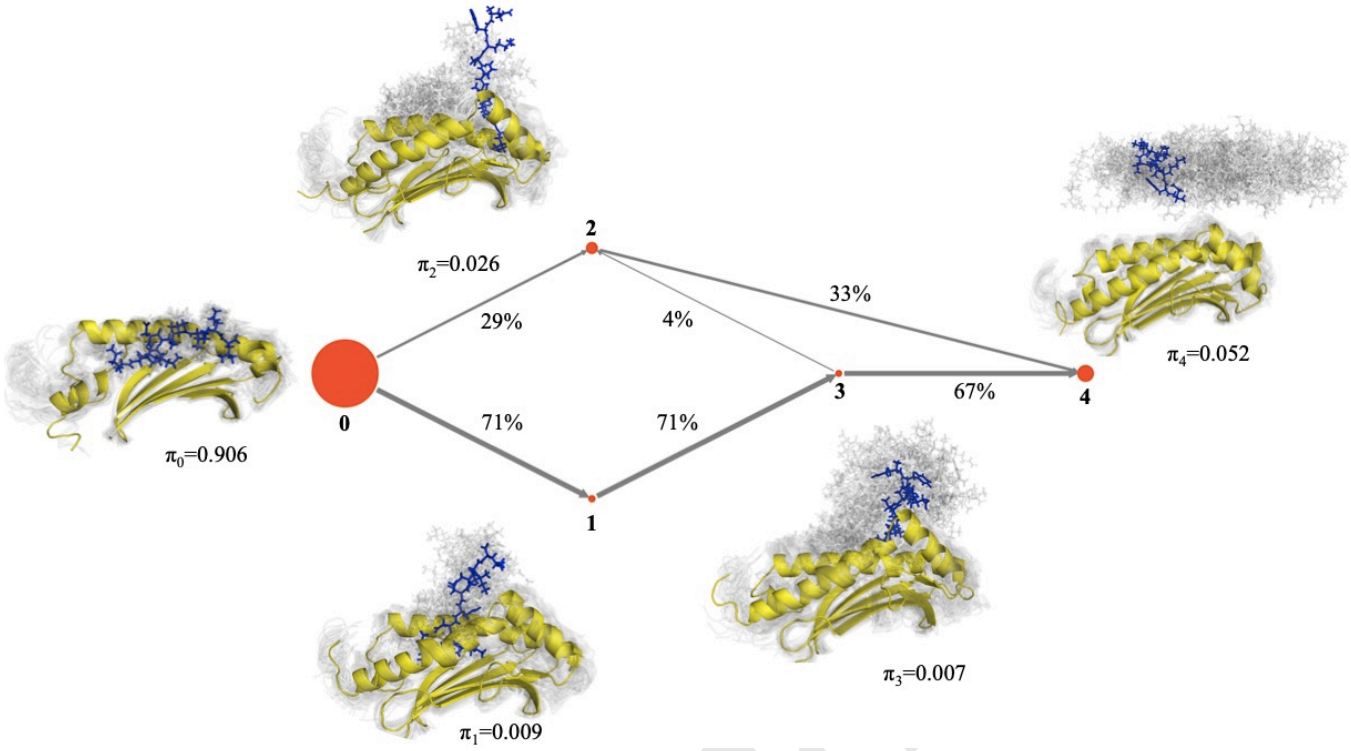
Finally, Fig. 6a shows that the native state for the *D4P* system appears to be relatively less stable than other intermediate states as compared to the *WT* system, despite being a stronger binder. Currently, it is not known whether QFKPNVILL is immunogenic. In addition to the lack of a charged residue in the TCR binding interface, T-cell recognition of this complex may be impaired by a less stable peptide-MHC native state. However, further experiments are needed to investigate the immunigenicity of the *D4P* system.

F. Proline's rigid backbone prevents torsions that would facilitate unbinding. The *D4P* system has a strong preference to unbind from the N-term side first. While it is possible for the *D4P* system to be in a state with the C-term unbound (State 1, Fig. 6a), our sampling suggests that it is difficult for conformations to then progress to a state in which the N-term is subsequently unbound (State 3). To investigate why, the backbone torsions of position 4 were extracted from the unbinding trajectories of *WT* and *D4A* where the C-term unbinds first and compared with the Ramachandran plot of prolines (42). In Fig. 6b, we see that trajectories starting in the native state (State 0) lie in regions overlapping with the possible phi/psi angles for prolines. However, as the *WT*/*D4A* transitions to having the C-term unbind first (State 1), p4 adopts a backbone conformation that is inaccessible for prolines. Unbinding trajectories continue to be outside the accessible region of prolines as *WT*/*D4A* transition from State 1 to State 3 (anchors unbound, but peptide in contact with MHC). Therefore, the rigidity of the proline backbone in *D4P* prevents transitions from State 1 to State 3, and subsequently from becoming fully unbound.

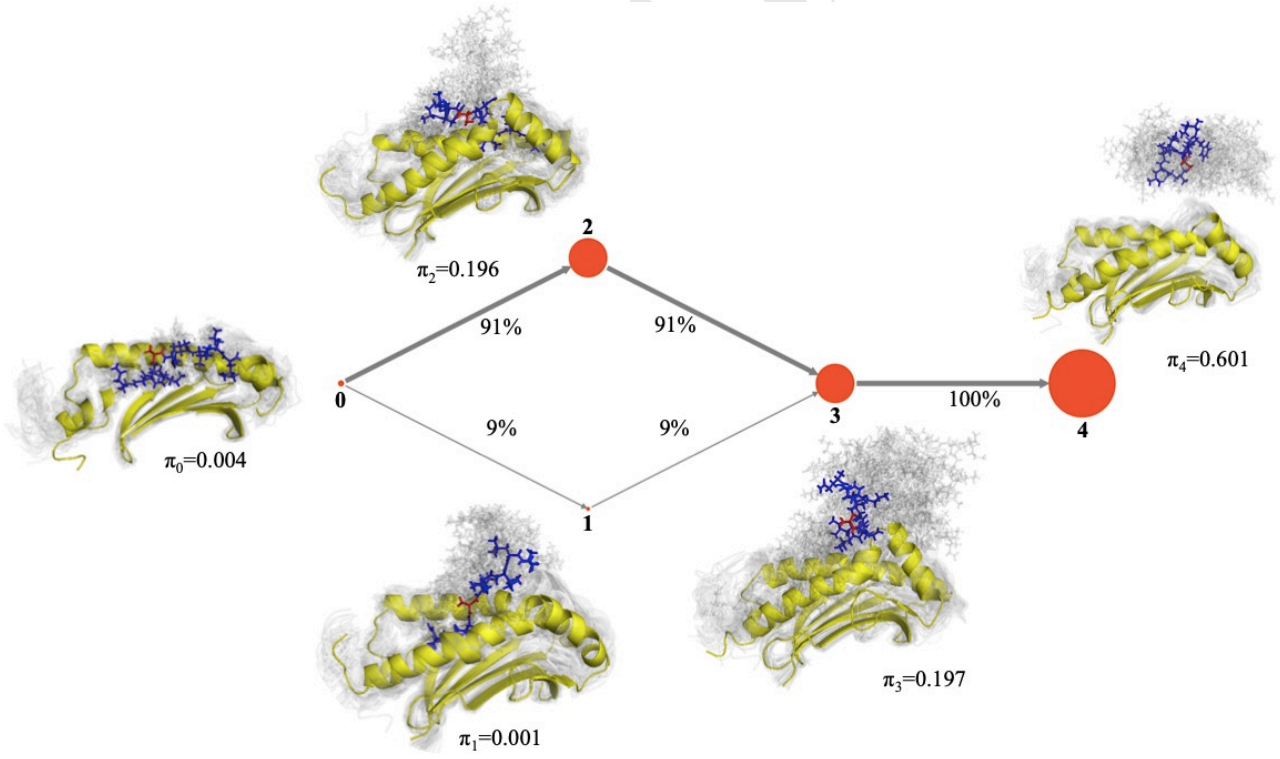
3. Discussion

In this work, we studied the mechanism behind stable binding of QFKDNVILL to HLA-A*24:02. We proposed a simulation framework that makes it feasible to generate MD data to build an MSM of the entire binding/unbinding process. As expected, our model predicted the importance of the anchor residues in positions 2 and 9, as demonstrated by mutational analysis. Interestingly, these analyses also singled out the contribution of the non-anchor position 4 to the stability of the system. To further explore the role of this position on peptide binding, we used our model to estimate the impact of two different mutations over peptide's binding affinity, and later confirmed our prediction with competitive binding assays. While *D4A* significantly impairs peptide binding, *D4P* leads to stronger binding.

In addition, by building the MSMs for each of these systems we were able to observe alternative unbinding pathways. While the *WT* system is more likely to start unbinding from the C-term end, both *D4A* and *D4P* are more likely to unbind the N-term first. This behavior is consistent with the lost of key interactions observed in the *WT* system, particularly between p4 and MHC residues K66, Q155, Y159 and T163. Interaction with K66 is not surprising, since a D4-K66 salt-



(a) *WT* flux network



(b) *D4A* flux network

Fig. 4. Flux network of unbinding trajectories for the *WT* system. States 0, 1, 2, 3 denote the set of associated states that have the peptide in contact to the MHC. State 4 represents the dissociated or unbound state. Size of the nodes (depicted in red) indicate the equilibrium probabilities of each state (π_i). a) The *WT* system prefers to unbind through detaching first on the C-term end (State 0 to State 1 transition) due to the stronger interactions on the N-term end, which include the aspartic acid in position 4. b) With a single mutation, the *D4A* system prefers to unbind through detaching first on the N-term end (State 0 to State 2 transition), and the accessibility of both detachment pathways favors the instability of the *D4A* system. Note that the MSM model includes all transitions between nodes, in all directions. However, this flux network depicts only trajectories starting from State 0 and reaching State 4 (i.e., unbinding pathways).

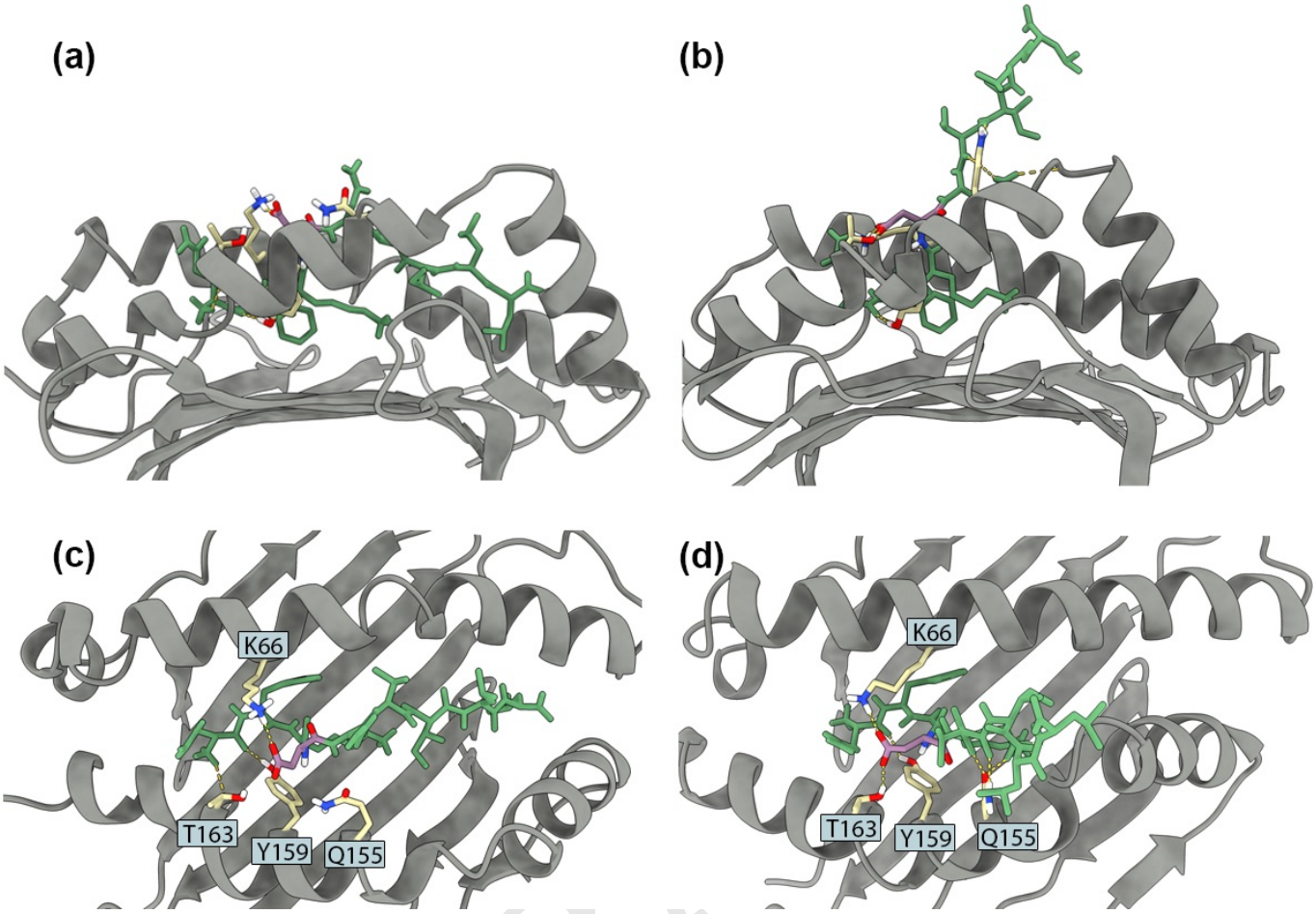
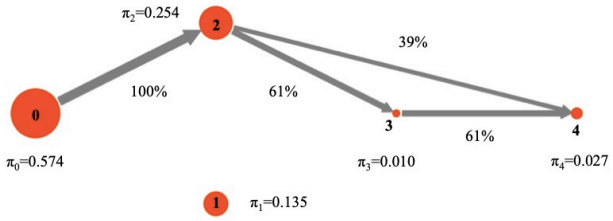


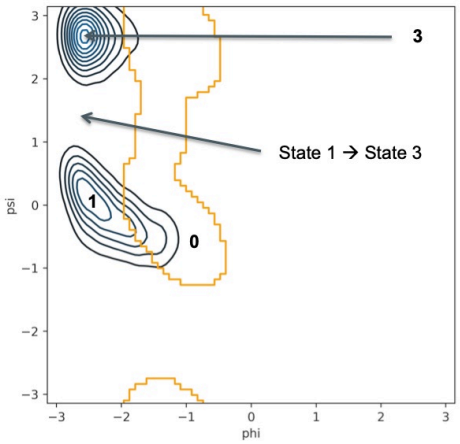
Fig. 5. Representative conformations in the *WT* system from State 0 (native state) and State 1 (N-term bound state). Panels (a) and (b) depict the side views of States 0 and 1, respectively. These states can be distinguished by the location of the C-term of the peptide relative to the MHC binding cleft (i.e., proximity to the F pocket). Panels (c) and (d) depict the top views of States 0 and 1, respectively. Peptide's position 4 (p4) residue (aspartic acid, D) is depicted in magenta (carbon atoms in magenta; oxygen atoms depicted in red). Other peptide positions are depicted in green. Key MHC residues predicted to interact with p4 are depicted in yellow (carbon atoms in yellow; oxygen atoms depicted in red; nitrogen atoms in blue; hydrogen atoms in white), including lysine 66 (K66), threonine 163 (T163), tyrosine 159 (Y159) and glutamine 155 (Q155). Hydrogen bonds involving any of these residues are depicted in yellow dashed lines.

bridge can be observed on the original crystal structure (PDB code 3I6L), as well as in other conformations corresponding to the bound state (Fig. 5c). In particular, K66 and T163 seem to be able to keep D4 in place, even when the peptide is already partially unbound from the C-term end (Fig. 5d). Visual inspection also suggest other roles for these MHC residues, notably interactions between p1-Y159 and p5/p6-Q155 (Fig. 5)

Interestingly, our model also predicts direct interactions between D4 and both Q155 and Y159 (Fig. S5-S6). In fact, the Y159A exchange had a negative impact on the binding of the *WT*, similar to that observed for *D4A*. The same impact was not detected when introducing Y159A on the *D4P* system. Taken together these results suggest two different mechanisms through which p4 can contribute to peptide-MHC stability. Polar residues, particularly negatively charged residues, such as aspartic acid, can benefit from a network of conserved interactions that help stabilize the N-term end of the peptides. On the other hand, having a proline at p4 makes it harder for the peptide backbone to bend in ways that would favor peptide detachment (Fig. 6). Although our analysis was limited to a few peptide-MHCs of interest, we believe the two binding mechanisms involving p4 might be of broader relevance to peptide-MHC binding in general. Two interesting observations provide additional support to this hypothesis. First, all the aforementioned MHC residues, that are potential p4 contacts, are present in the consensus sequence produced by aligning over 10,000 protein sequences including HLA-As, HLA-Bs and HLA-Cs (SI Appendix, Fig. S17). The prevalence of K66 is not very high, about 40% across all types, being often replaced with N in HLA-As and I in HLA-Bs. T163 is particularly high among HLA-A sequences (74%). Most notably, Q155 and Y159 are present in over 99.9% of the sequences for all HLA types, and the peptide-binding contribution of these specific MHC positions has been observed in previous studies (43, 44). Second, across sequences of HLA-binders, the observed frequencies of aspartic acid and proline were shown to be 2.2 times more frequent than expected relative to the proteome (7). Another negatively charged residue, glutamic acid, was also found to be 1.6 times more frequent than expected (7). Further experimental studies will be needed to investigate the differential contribution of these interactions on the binding of different peptides, and across different HLA allotypes.



(a) *D4P* flux network



(b) Ramachandran plot of p4 for unbinding trajectories in *WT*/*D4A*

Fig. 6. (a) Flux network of unbinding trajectories for the *D4P* system. The introduction of a proline forces the unbinding starting from the N-term side (State 2). (b) (Blue contour) Phi/Psi angles (in radians) of position 4 from *WT*/*D4A* unbinding trajectories where the C-term side unbinds first. The bottom region cover States 0 and 1, while the top region covers State 3. (Orange border) Ramachandran plot of accessible phi/psi angles of proline. Unbinding trajectories during the transition from State 1 to State 3 lie in regions that do not overlap with the accessible phi/psi angle of proline. Thus, the unbinding trajectories adopt backbone conformations of p4 that are incompatible with the rigidity of proline. Note that the MSM of *D4P* (a) includes transitions from State 0 to State 1, and from State 1 back to State 0. However, these transitions are not depicted in the flux network, since none of the paths passing by State 1 were able to progress to State 4.

This is the first work to apply MSMs to describe the preferred unbinding pathways for peptide-MHC complexes. In addition, to the best of our knowledge, this is also the largest computational exploration of peptide-MHC dynamics to date (over 650 microseconds). This unique combination of methods provided a wealth of information on the studied systems, including the contributions of particular interactions to peptide binding and complex stability. Such analysis can also be done for any other peptide-MHC of interest, provided an initial 3D structure of the complex. In the absence of a crystal structure, an appropriate 3D model could be used, and our group has also contributed tools for this particular task (13, 14). The computational cost to build the MSMs was manageable and was done using local GPU computing clusters (about 10000 GPU-hours compared to 115,000 GPU-hours in (26)).

While this work demonstrates the feasibility of using MD and MSMs to study peptide-MHC dynamics, it is important to note that the approximations performed here could have an impact on obtained results. The use of implicit solvent, for instance, can have an effect on the dynamics of the system and artificially accelerate the time for events to occur. In addition, hydrophobic interactions are typically the major contributions of peptide-MHC binding, particularly for the anchor residues, and the finite size of water molecules may need to be accounted for. Finally, there is evidence of allostery where peptide binding affected the dynamics of remote regions in HLA-A2, including the α_3 and β-2 microglobulin domains (45). While we used positional restraints on the β-sheet floor to minimize the potential impact, the full effect of the MHC truncation in our simulations is unknown.

Future work can focus on ways to improve the accuracy of the final MSM. This is likely in the form of including more atoms into the system, such as the β-2 microglobulin portion of the MHC, explicit water molecules, or even the other proteins involved in keeping MHCs in the peptide-receptive state (46). However, the simulation output similarly needs to be kept high in order for enough statistics to be generated. Other enhanced sampling approaches (47) could conceivably be done as long as there is a way to produce an unbiased MSM in the end. The use of coarse graining is also promising, however it is highly nontrivial to perform in such a way that does not negatively influence the computation of kinetic quantities (48, 49).

Finally, it is worth nothing that the peptide studied here (QFKDNVILL) was derived from the nucleocapsid protein of SARS-CoV, and a highly similar peptide exists in the nucleocapsid protein of SARS-CoV-2 (NFKDQVILL). The differences between the two peptides do not appear to be significant, as asparagine and glutamine are both polar, uncharged residues. More importantly, both peptides share the same residues in positions 2, 4, and 9, which means that the analysis we have performed here likely apply to both systems. Finally, given that D4 and K66 are exposed for the recognition by T-cells, this conserved interaction could be the focus of cross-reactive T-cell responses (i.e., T-cells primed with QFKDNVILL may also recognize NFKDQVILL). In fact, cross-reactivities involving D4 in other viral peptides have already been predicted (50) and confirmed experimentally (51). Regardless of its role in T-cell recognition, the alternative roles of p4 in peptide-MHC binding and stability highlight the importance of structure-based methods in the analysis of peptide-MHC binding, and the discovery of peptide-targets for several immunotherapy

applications.

Materials and Methods

Molecular dynamics protocol. In this work, we simulate only the binding site of the MHC in order to make the whole framework more computationally tractable. While the entire peptide-MHC complex is a large system of around 380 residues total, we exclude the β-2 microglobulin and portions of the α chain (α-3) of the MHC, leaving two α-helices (α-1 and α-2 in yellow, Fig 1a) and the β-sheet floor (in light blue, Fig 1a) that enclose the bound peptide. This roughly results in a system half the size of the original (around 190 residues total). The MHC portion that was truncated is likely important for overall stability of the MHC, so in all simulations we include a positional restraint on the C_α atoms of the β-sheet floor (force constant: 100 kJ/mol/nm^2), which include the main contacts formed between the simulated binding site and the truncated portion.

In all simulations, the AMBER99sbildn (52) force field was used with implicit solvent (GBSA OBC) (53). Simulations were performed at 300 K with the Langevin integrator (friction coefficient: 0.1 ps^{-1}). The hydrogen masses were artificially increased to 4 amu to allow a 4 fs timestep. Starting conformations were equlibrated for 500 ns with the positional restraints on the C_α atoms of the whole system.

Exploration stage: umbrella sampling. Umbrella sampling is used to accelerate the exploration of the relevant states of the binding process. Biased sampling is needed here since the half-life of peptide-MHC binding can be on the order of seconds or greater (2). Starting with the crystal structure of *WT* (PDB: 3I6L), we generate detachment/unbinding pathways of the peptide.

The geometry of the MHC allows us to define a convenient reaction coordinate for the umbrella sampling. Bound peptides are enclosed between two α-helices atop a β-sheet floor. In order to detach, peptides must essentially unbind in a direction that is approximately normal to the β-sheet floor (23), which is roughly planar (50). We can see from Fig. 1a that the principal axis of the (non-truncated) system happens to roughly align with this direction. Thus, if the principal axis is aligned to the Z direction in Euclidean space, the β-sheet floor becomes approximately aligned to the XY plane, and a bias along the Z direction can be used to accelerate sampling along the binding/unbinding pathway. The biases for the umbrella sampling simulations are based on the distance between the center of masses of the peptide and the MHC along the Z-coordinate. We will call this distance the *z-dist*. We use the C_α atoms of the β-sheet floor as a stable set of atoms to compute the center of mass for the MHC; these are the same atoms from which we add positional restraints.

Given the description of the reaction coordinate above, we run umbrella sampling simulations across *z-dist* umbrellas centered from 1.0–3.0 nm (in increments of 0.1 nm) with a force constant of 100 kJ/mol/nm^2, where the the *z-dist* of the native state is approximately 1.0 nm. Each simulation was run for approximately 1 microsecond, producing many detachment trajectories across the runs. Additional umbrella sampling simulations were done for *D4A* with a looser force constant (10 kJ/mol/nm^2) given that the peptide is known to be a nonbinder and is less stable. Several replicates were performed, particularly for umbrellas centered in the 2.0–3.0 nm range in order to sample more association/dissociation events.

Connection stage: Generating transition statistics with adaptive sampling. In this stage, we use adaptive sampling to run enough unbiased molecular dynamics to produce a final MSM that connects most of the states generated (Fig. 1b). At each iteration, a new set of about 20 unbiased molecular dynamics simulations are spawned from starting conformations chosen from less densely sampled regions of the conformational space. The conformations are chosen based on the analysis of the set of trajectories that have already been generated. Trajectories are first featurized using residue-residue contacts (defined as the the closest heavy atom distance) between peptide with MHC and peptide with itself. Then the conformations are mapped to the two leading independent components using time-lagged independent components analysis, or TICA (39, 40) (lag 10 ns), and the space is discretized into microstates with K-means (100 clusters). Next, microstates are chosen with probability inversely proportional to the number of conformations mapped to it, and a conformation is uniformly randomly chosen from the microstate as a starting point for the next round of simulations. We repeat the adaptive sampling iterations until a MSM can be built using more than 90% of the microstates (SI Appendix, Fig. S1, S9, and S13). All simulations were run using CUDA and OpenMM (54) and performed on NOTS as part of Rice University's Center of Research Computing.

Building the MSMs. Similar to the adaptive sampling process, the trajectories were featurized using residue-residue contacts between peptide with MHC and peptide with itself, resulting in 1692 contacts. We extract 2 independent components using TICA using a lag time of 10 ns based on the convergence of timescales (SI Appendix, Fig. S2a, S10a, and S14a). The two leading independent components adequately capture the transition to and from the native and unbound states (SI Appendix, Fig. S3, S11, and S15). This space was discretized into microstates using K-means with 100 clusters. From the trajectories on the discretized space, discrete Transition-based Reweighting Analysis Method (dTRAM) was used to build a Markov state model (41), taking into account the biases introduced with the umbrella sampling simulations. A final MSM was constructed using a lag time based on the convergence of timescales (SI Appendix, Fig. S2b, S10b, and S14b). Error bars are computed based on a moving block procedure for bootstrapping (55). The final MSMs are self-consistent based on the Chapman-Kolmogorov test (SI Appendix, Fig. S4, S12, and S16). All analysis was performed using MDTraj (56) and Pyemma (57). The data and scripts for analysis are available upon request.

Mutational analysis. We can estimate the changes in the free energy of binding upon mutation ($\Delta\Delta G$) for all nine residues in the peptide. We do this with free energy perturbation theory (58, 59). The change in binding free energy is computed as

$$\begin{aligned}\Delta\Delta G &= \Delta G_{mut} - \Delta G_{wt} \\ &= (G_{mut}^{\text{associated}} - G_{mut}^{\text{dissociated}}) - (G_{wt}^{\text{associated}} - G_{wt}^{\text{dissociated}}) \\ &= (G_{mut}^{\text{associated}} - G_{wt}^{\text{associated}}) - (G_{mut}^{\text{dissociated}} - G_{wt}^{\text{dissociated}}) \\ &= -RT\ln(\frac{Z_{mut}^{\text{associated}}}{Z_{wt}^{\text{associated}}}) + RT\ln(\frac{Z_{mut}^{\text{dissociated}}}{Z_{wt}^{\text{dissociated}}})\end{aligned} \quad [1]$$

where $RT = 2.479\frac{\text{kJ}}{\text{mol}}$ at temperature $T = 298K$, and Z is the configurational partition function for the corresponding system. The last two terms represent $\Delta G_{wt->mut}^{\text{associated}}$ and $-\Delta G_{wt->mut}^{\text{dissociated}}$, thus completing the free energy cycle. Positive values of $\Delta\Delta G$ indicate that the mutant is a weaker binder, while negative values of $\Delta\Delta G$ indicate that the mutant is a stronger binder.

The ratio of configurational partition functions over a state S, can be manipulated as follows:

$$\begin{aligned}\frac{Z_{mut}^S}{Z_{wt}^S} &= \frac{1}{Z_{wt}^S}\int_S e^{-\beta U_{mut}(x)}dx \\ &= \frac{1}{Z_{wt}^S}\int_S e^{-\beta U_{mut}(x)}e^{\beta U_{wt}(x)}e^{-\beta U_{wt}(x)}dx \\ &= \frac{1}{Z_{wt}^S}\langle e^{-\beta(U_{mut}(x)-U_{wt}(x))}\rangle_{S,wt}\end{aligned} \quad [2]$$

where $U(x)$ is the potential energy. The average is taken using the stationary probabilities, $\mu(x)$, of the *WT* system computed from the MSM/dTRAM analysis. Thus, the following ratios can be finally computed as:

$$\begin{aligned}\frac{Z_{mut}^{\text{dissociated}}}{Z_{wt}^{\text{dissociated}}} &= \frac{\sum_{x\in S_D} e^{-\beta(U_{mut}(x)-U_{wt}(x))}\mu(x)}{\sum_{x\in S_D}\mu(x)} \\ \frac{Z_{mut}^{\text{associated}}}{Z_{wt}^{\text{associated}}} &= \frac{\sum_{x\in S_A} e^{-\beta(U_{mut}(x)-U_{wt}(x))}\mu(x)}{\sum_{x\in S_A}\mu(x)}\end{aligned} \quad [3]$$

where a configuration, x, is in S_D, the dissociated state, if the minimum distance between the peptide and MHC is greater than 0.5 nm. Otherwise, x is in S_A, the associated state.

The original and mutation energies are computed using the same force field from the molecular dynamics simulations (AMBER99sbildn force field (52) with GBSA OBC implicit solvent (53)) but only nonbonded terms were considered. Mutated structures were generated with PyMOL where the original amino acid was cut back to the C_β-atom and hydrogen atoms were added, resulting in an alanine structure. The value of the dihedral angle $C-C_\alpha-C_\beta-H_{\beta 1}$ was taken to be the dihedral angle of the original residue, $C-C_\alpha-C_\beta-C_\gamma$ (or $C-C_\alpha-C_\beta-C_{\gamma 1}$ for the valine in position 6 and isoleucine in position 7).

Competitive binding assays. We run competitive binding assays to find the binding affinities of QFKDNVILL (WT), QFKANVILL ($D4A$), and QFKPNVILL ($D4P$) with HLA-A*24:02. Fluorescent and unlabeled peptides were synthesized by BioSynthesis Inc. EBC-1 cells used for assay were transduced with HLA-A*2402 for increased expression. Competition peptide assay followed protocol established by Kessler *et al* (60). In brief, EBC-1 cells were washed with elution buffer then incubated overnight in the dark with a fixed concentration of a known HLA-A*24:02 binding peptide tagged with GFP and varying concentrations of test peptides. Cells were analyzed on a FACs CANTO II analyzer and median fluorescence intensity was measured. IC50 values were determined using nonlinear regression from GraphPad Prism 8.0.

Multiple sequence alignment. A total of 19,689 protein sequences were downloaded from IMGT/HLA (61), corresponding to the three classical class I HLA genes (HLA-A, HLA-B, HLA-C). Since many sequences did not cover the entire protein length, we removed entries with less than 3/4 of the complete sequence, resulting in a total of 10,435 sequences (HLA-A: 3,160, HLA-B: 3,788, HLA-C: 3,487). A multiple sequence alignment was performed with MUSCLE (62), and the visual inspection was performed with Jaview (63).

Code repository. Code for umbrella sampling, adaptive sampling, and MSM analysis can be found at https://github.com/KavrakiLab/adaptive-sampling-pmhc.

ACKNOWLEDGMENTS. This work was supported by a training fellowship from the Gulf Coast Consortia on the Training Program in Biomedical Informatics, National Library of Medicine T15LM007093. This work has also been supported in part by Cancer Prevention and Research Institute of Texas (CPRIT) through Grant award RP170508, through a Fellowship from the Computational Cancer Biology Training Program (RP170593), Einstein Foundation Berlin (Einstein Visiting Fellowship to C.C.), the National Science Foundation (CHE-1740990, CHE-1900374, and PHY-1427654 to C.C.), and the Welch Foundation (grant C-1570 to C.C.). This work was also supported by the Blue Waters supercomputer, XSEDE resources (Stampede2 and Comet), and the Center of Research Computing at Rice University (NOTS).

1. Rock KL, Reits E, Neefjes J (2016) Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol.* 37(11):724–737.
2. Harndahl M, et al. (2012) Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* 42(6):1405–1416.
3. Shao W, et al. (2018) The SysteMHC Atlas project. *Nucleic Acids Res.* 46(D1):D1237–D1247.
4. Vita R, et al. (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* 47(D1):D339–D343.
5. Nielsen M, et al. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12(5):1007–1017.
6. O'Donnell TJ, et al. (2018) MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* 7(1):129–132.
7. O'Donnell TJ, Rubinsteyn A, Laserson U (2020) MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst* 11(1):42–48.
8. Sarkizova S, et al. (2020) A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.*
9. Alam N, Schueler-Furman O (2017) Modeling Peptide-Protein Structure and Binding Using Monte Carlo Sampling Approaches: Rosetta FlexPepDock and FlexPepBind. *Methods Mol. Biol.* 1561:139–169.
10. Kyeong HH, Choi Y, Kim HS (2018) GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics* 34(3):469–476.
11. Antunes DA, Devaurs D, Moll M, Lizée G, Kavraki LE (2018) General prediction of peptide-MHC binding modes using incremental docking: A proof of concept. *Scientific Reports* 8.
12. Antunes DA, Abella JR, Devaurs D, Rigo MM, Kavraki LE (2018) Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes. *Curr Top Med Chem* 18(26):2239–2255.
13. Abella JR, Antunes DA, Clementi C, Kavraki LE (2019) APE-Gen: A Fast Method for Generating Ensembles of Bound Peptide-MHC Conformations. *Molecules* 24(5).
14. Antunes DA, et al. (2020) HLA-Arena: a customizable environment for the structural modeling and analysis of peptide-HLA complexes for cancer immunotherapy. *JCO Clinical Cancer Informatics* 4:623–636.
15. Fodor J, Riley BT, Borg NA, Buckle AM (2018) Previously Hidden Dynamics at the TCR-Peptide-MHC Interface Revealed. *J. Immunol.* 200(12):4134–4145.
16. Beerbaum M, et al. (2013) NMR spectroscopy reveals unexpected structural variation at the protein-protein interface in MHC class I molecules. *J. Biomol. NMR* 57(2):167–178.
17. Yanaka S, Sugase K (2017) Exploration of the Conformational Dynamics of Major Histocompatibility Complex Molecules. *Front Immunol* 8:632.
18. van Hateren A, et al. (2017) Direct evidence for conformational dynamics in major histocompatibility complex class I molecules. *J. Biol. Chem.* 292(49):20255–20269.
19. Hawse WF, et al. (2013) Peptide modulation of class I major histocompatibility complex protein molecular flexibility and the implications for immune recognition. *J. Biol. Chem.* 288(34):24372–24381.
20. Wieczorek M, et al. (2017) Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front Immunol* 8:292.
21. Ayres CM, Riley TP, Corcelli SA, Baker BM (2017) Modeling Sequence-Dependent Peptide Fluctuations in Immunologic Recognition. *J Chem Inf Model* 57(8):1990–1998.
22. Wan S, Knapp B, Wright DW, Deane CM, Coveney PV (2015) Rapid, Precise, and Reproducible Prediction of Peptide-MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *J Chem Theory Comput* 11(7):3346–3356.
23. Knapp B, Demharter S, Deane CM, Minary P (2016) Exploring peptide/MHC detachment processes using hierarchical natural move Monte Carlo. *Bioinformatics* 32(2):181–186.
24. Husic BE, Pande VS (2018) Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* 140(7):2386–2396.
25. Wieczorek M, et al. (2016) MHC class II complexes sample intermediate states along the peptide exchange pathway. *Nat Commun* 7:13224.
26. Paul F, et al. (2017) Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nat Commun* 8(1):1095.
27. Plattner N, Doerr S, De Fabritiis G, Noe F (2017) Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat Chem* 9(10):1005–1011.
28. You W, Tang Z, Chang CA (2019) Potential Mean Force from Umbrella Sampling Simulations: What Can We Learn and What Is Missed? *J Chem Theory Comput* 15(4):2433–2443.
29. Bowman GR, Ensign DL, Pande VS (2010) Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J Chem Theory Comput* 6(3):787–794.
30. Doerr S, De Fabritiis G (2014) On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J Chem Theory Comput* 10(5):2064–2069.
31. Preto J, Clementi C (2014) Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys Chem Chem Phys* 16(36):19181–19191.
32. Hruska E, Abella JR, Nuske F, Kavraki LE, Clementi C (2018) Quantitative comparison of adaptive sampling methods for protein dynamics. *J Chem Phys* 149(24):244119.
33. Zimmerman MI, Porter JR, Sun X, Silva RR, Bowman GR (2018) Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *J Chem Theory Comput* 14(11):5459–5475.
34. Betz RM, Dror RO (2019) How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *J Chem Theory Comput* 15(3):2053–2063.
35. Wan H, Voelz VA (2020) Adaptive Markov state model estimation using short reseeding trajectories. *J Chem Phys* 152(2):024103.
36. Liu J, et al. (2010) Novel immunodominant peptide presentation strategy: a featured HLA-A*2402-restricted cytotoxic T-lymphocyte epitope stabilized by intrachain hydrogen bonds from severe acute respiratory syndrome coronavirus nucleocapsid protein. *J. Virol.* 84(22):11849–11857.
37. Ahmed SF, Quadeer AA, McKay MR (2020) Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* 12(3).
38. He L, De Groot AS, Bailey-Kellogg C (2015) Hit-and-run, hit-and-stay, and commensal bacteria present different peptide content when viewed from the perspective of the T cell. *Vaccine* 33(48):6922–6929.
39. Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noe F (2013) Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* 139(1):015102.
40. Schwantes CR, Pande VS (2013) Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput* 9(4):2000–2009.
41. Wu H, Mey AS, Rosta E, Noe F (2014) Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J Chem Phys* 141(21):214106.
42. Lovell SC, et al. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* 50(3):437–450.
43. Baker BM, Turner RV, Gagnon SJ, Wiley DC, Biddison WE (2001) Identification of a crucial energetic footprint on the alpha1 helix of human histocompatibility leukocyte antigen (HLA)-A2 that provides functional interactions for recognition by tax peptide/HLA-A2-specific T cell receptors. *J. Exp. Med.* 193(5):551–562.
44. Uchtenhagen H, et al. (2013) Proline substitution independently enhances H-2D(b) complex stabilization and TCR recognition of melanoma-associated peptides. *Eur. J. Immunol.* 43(11):3051–3060.
45. Ayres CM, et al. (2019) Dynamically Driven Allostery in MHC Proteins: Peptide-Dependent Tuning of Class I MHC Global Flexibility. *Front Immunol* 10:966.

46. Mage MG, et al. (2012) The peptide-receptive transition state of MHC class I molecules: insight from structure and molecular dynamics. *J. Immunol.* 189(3):1391–1399.
47. Yang YI, Shao Q, Zhang J, Yang L, Gao YQ (2019) Enhanced sampling in molecular dynamics. *J Chem Phys* 151(7):070902.
48. Wang J, et al. (2019) Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent Sci* 5(5):755–767.
49. Nuske F, Boninsegna L, Clementi C (2019) Coarse-graining molecular systems by spectral matching. *J Chem Phys* 151(4):044116.
50. Antunes DA, et al. (2017) Interpreting T-Cell Cross-reactivity through Structure: Implications for TCR-Based Cancer Immunotherapy. *Front Immunol* 8:1210.
51. Kamga L, et al. (2019) CDR3a drives selection of the immunodominant Epstein Barr virus (EBV) BRLF1-specific CD8 T cell receptor repertoire in primary infection. *PLoS Pathog.* 15(11):e1008122.
52. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8):1950–1958.
53. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55(2):383–394.
54. Eastman P, et al. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13(7):e1005659.
55. Kunsch HR (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* 17(3):1217–1241.
56. McGibbon RT, et al. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 109(8):1528–1532.
57. Scherer MK, et al. (2015) PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* 11(11):5525–5542.
58. Matysiak S, Clementi C (2004) Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J. Mol. Biol.* 343(1):235–248.
59. Matysiak S, Clementi C (2006) Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.* 363(1):297–308.
60. Kessler JH, et al. (2004) Competition-based cellular peptide binding assay for HLA class I. *Curr Protoc Immunol* Chapter 18:Unit 18.12.
61. Robinson J, Halliwell JA, Hayhurst JD, , et al. (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 43(Database issue):D423–431.
62. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.
63. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.