Efficient Estimation of Binding Free Energies between Peptides and an MHC Class II Molecule Using Coarse-Grained Molecular Dynamics Simulations with a Weighted Histogram Analysis Method

Ming Huang, Wenjun Huang, Fei Wen, and Ronald G. Larson*

Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136

ABSTRACT: We estimate the binding free energy between peptides and an MHC class II molecule using molecular dynamics (MD) simulations with the Weighted Histogram Analysis Method (WHAM). We show that, owing to its more thorough sampling in the available computational time, the binding free energy obtained by pulling the whole peptide using a coarse-grained (CG) force field (MARTINI) is less prone to significant error induced by inadequate-sampling than using an atomistic force field (AMBER). We further demonstrate that using CG MD to pull 3-4 residue peptide segments while leaving the remaining peptide segments in the binding groove and adding up the binding free energies of all peptide segments gives robust binding free energy estimations, which are in good agreement with the experimentally measured binding affinities for the peptide sequences studied. Our approach thus provides a promising and computationally efficient way to rapidly and reliably estimate the binding free energy between an arbitrary peptide and an MHC class II molecule.

Keywords: MHC, Peptide, peptide-MHC binding affinity, Coarse-Grained, Binding Free Energy, Umbrella Sampling Method

1 Introduction

The Major Histocompatibility Complex (MHC) plays an important role in human adaptive immune responses by binding to antigen-derived peptide fragments and presenting them for T-cell recognition [1]. Due to the variations in the mechanisms mediating antigen processing and presentation [2]–[4], stable binding to the restricting MHC protein is the most selective requisite for T-cell recognition of a peptide. Therefore, identifying MHC-binding peptides is a crucial step in studying T-cell immune responses as well as designing effective vaccines and therapeutics [5], [6]. There are two classes of MHC molecules involved in T-cell antigen presentation - class I and class II. An MHC class I molecule contains a peptide binding groove with closed ends that can only accommodate peptides with 8-10 residues in length [7]. An MHC class II molecule, on the other hand, contains a peptide binding groove that is open on both ends (Fig 1a), allowing the binding of a peptide with 12 to 25 residues [8]. The peptide binding groove of an MHC molecule contains multiple binding pockets that allow strong interactions with certain residues, which are termed anchor residues. The flexibility in the length and registry of binding peptides to MHC class II molecules makes their computational identification more challenging than that for MHC class I molecules. This study focuses on the estimation of binding free energy between peptides and an MHC class II molecule using molecular dynamics simulation technique.

Molecular dynamics (MD) simulation has been successfully applied to model the interactions between peptides and protein molecules. Atomistic MD simulation models all atoms on the amino acid residues explicitly, and offers molecular level insights into peptide-protein binding that are not accessible through conventional experimental efforts. Atomistic MD simulations have been adopted to identify the binding pockets and conformational changes of the binding groove upon binding of the peptide to both MHC class I and II molecules [9]–[11]. However, due to the size of the MHC molecule (~8 nm in diameter), atomistic MD simulations only access short time scales (<100 ns) within reasonable computational time. Coarse-grained (CG) models, on the other hand, combine a few atoms into one single CG bead, and therefore reduce the degrees of freedom of the

John Wiley & Sons, Inc.

system significantly, allowing affordable access to microsecond timescales. Several systematically parameterized CG models are available in the literature. In particular, Marrink and coworkers developed the MARTINI force field for lipids and surfactants [12], and Monticelli *et al.* extended this force field to peptides and proteins [13]. To date, the MARTINI force field has been applied to characterize the properties of lipid membranes and lipid polymorphism, and the interplay between proteins and lipids, such as the interaction between ATP synthase and inner mitochondrial membrane cristae, as well as the self-assembly of soluble peptides and proteins [14].

The Weighted Histogram Analysis Method (WHAM) [15] is often used to estimate the binding free energy between two molecules. WHAM is a statistical analysis technique that combines conformation distributions gathered from multiple biased umbrella sampling simulations along a chosen reaction coordinate between two states of interest and computes the potential of mean force (PMF) as a function of distance along this reaction coordinate based on the distribution of conformations [16]. Several studies have used WHAM to compute the binding free energy between peptides and an MHC class I molecule. For example, Olaposi *et al.* applied WHAM with atomistic force field to estimate the binding free energy between a 9-residue peptide (sequence: SLYNTVATL) and HLA-A2 (PDB ID: 2V2W). They obtained a binding free energy of 12 kJ/mol between the peptide and the MHC class I molecule, which deviates considerably from the 32.2kJ/mol binding free energy obtained from experiments [17]. On the other hand, May *et al.* applied WHAM with MARTINI coarse-grained force field to estimate the binding free energy between the peptide-MHC-Class-I molecule and the T cell receptor (TCR) (PDB ID: 1AO7) and obtained a binding free energy of 80kJ/mol, which compares well to the value of 78.6kJ/mol obtained from experiments [18]. To our knowledge, however, there is no reported study that has accurately estimated the binding free energy between a peptide and a MHC class II molecule. Zhang *et al*. have adopted molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) along with atomistic MD simulations to estimate the binding free energy contribution from each individual residue on the bound peptide, and summed up these contributions to get an estimated total binding free energy between each peptide and an MHC class II molecule. However, the estimated binding free energies did not agree well with experimental results [19]. Therefore, we wish to develop a reliable protocol for using WHAM and MD simulation to predict accurately the binding free energy between peptides and an MHC class II molecule.

The binding groove of the MHC class II molecule DR1 contains five binding pockets for the influenza virus hemagglutinin 306-318 fragment PKYVKQNTLKLAT (termed HA peptide). Each of the pockets accommodates one "anchor residue" on the HA peptide numbered 1 to 5 (Fig 1b). Atoms on the anchor residue interact with the MHC class II binding groove through non-covalent interactions [20], and the free energy difference considered in this study is between the state where the peptide is submerged in the binding groove and the state where the peptide is far eough outside the binding groove that all peptide-binding groove interactions are negligible. Residue TYR308 of the HA peptide is buried in the largest and most hydrophobic pocket, namely binding pocket 1, resulting in a strong binding interaction. Four other HA residues (GLN311, THR313, LEU314, LEU316) are set in the other four smaller shallower and less hydrophobic binding pockets, which result in relatively weaker binding interactions compared to that between TYR308 and binding pocket 1. It is worth noting that the interaction between THR313 and its binding pocket has been shown to be energetically unfavorable [20], [21]. The binding affinity (K_d) of the HA peptide is determined experimentally to be 14nM [22]. We can convert the binding affinity to a binding free energy of 46.6 kJ/mol via the equation (1) below, where the standard reference concentration c^θ, ideal gas constant R, and temperature T are 1M, 8.31J/(mol K), and 310K, respectively.

$$\Delta G = RTln(K_d/c^\theta) \quad (1)$$

To date, no studies have used WHAM to estimate the binding free energy between a peptide and an MHC class II molecule, possibly due to the large number of microstates that a long peptide (>9 residues) needs to explore within each window along the reaction coordinate. The goal of this work is to establish a reliable protocol for using WHAM with MD simulations to obtain the binding free energy between relatively long (13-residue) peptides and an MHC class II molecule. We apply WHAM with both atomistic and CG force fields on a number of peptide sequences and examine the limitations of each approach. We then show that using WHAM with a CG force field on segmented peptides and obtaining the total binding free energy by adding up the free energy contributions from all segments is a robust approach to obtain peptide-MHC class II molecule binding free energies that are in good agreement with the experimentally determined binding affinities. For simplicity, we will hereafter refer to the MHC class II molecule DR1 as "MHCII", and the binding free energy between the peptide and the MHCII as the "binding free energy", unless specified otherwise.

2 Methods

2.1 Peptide Models

The crystal structure of HA-MHCII complex used in this study (HA: PKYVKQNTLKLAT, PDB ID: 1DLH) is obtained from the Protein Data Bank (www.rcsb.org). We generate five additional peptide sequences via the residue mutation utility implemented in PyMOL ver. 1.8 [23] (Table I). Among these six peptide sequences, the binding affinities of HA, CLIP, YAK, and HA_{Y308A} (see the sequence name convention in Table I caption), have been determined experimentally [22], [24]. HA and CLIP are fairly strong binding peptides with binding affinities of 14nM and 25nM respectively; YAK has a weak binding affinity of 118nM; while HA_{Y308A} has a binding affinity of only 23000nM, which is extremely weak. Using Eq 1, these binding affinities can be converted to binding free energies of 46.6 kJ/mol, 45.1kJ/mol, 41.1 kJ/mol, and 27.5kJ/mol respectively.

In addition to these four whole peptide models, segmented peptide models are set up by truncating the peptide based on the location of the anchor residues. For example, we truncate the HA peptide into four segments (PKY | VKQ | NTL | KLAT). We truncate the atomistic peptide by disconnecting the bond between carbon and nitrogen terminus of the two neighboring segments and adding hydrogens to fill the valencies. Only four segments are chosen, instead of five, because both anchor residues THR313 and LEU314 have weak interactions with their corresponding binding pockets, as shown in an experimental study [20], and are therefore combined into one segment. The other segmented peptides are set up in a similar way.

2.2 Atomistic Simulations

All simulations are conducted using the GROMACS simulation package, version 4.6.5 [25]. VMD version 1.9.1 [26] is used for visualization of the structure. Two types of WHAM simulations are set up in this study – (1) pulling the entire peptide, and (2) pulling segments of the peptide from the binding pocket. We use the HA-MHCII binding pair to illustrate the pulling process. Pulling of the whole peptide is simulated with the pulling group chosen as: (1) the whole peptide, (2) the THR318 residue on one end, or (3) the PRO306 residue on the other end. These simulations are referred as (1) center-of-mass (COM) and (2,3) peeling (PEEL), respectively (Fig 2). For the segmented peptide model (e.g. PKY|VKQ|NTL|KLAT), COM pulling simulations are conducted for each of the four segments, leaving the rest of the peptide in the binding groove to prevent the binding groove deformation that would occur if no peptide is present [24], [27]. In all simulations, we define the reaction coordinate to be the distance between the COM of the respective pulling group and that of the MHCII. A spring constant of 600 kJ mol^{-1} nm^{-2} is applied on the pulling group to generate a trajectory with a pulling rate of 0.002nm/ps from the bounded state to unbound state along the reaction coordinate in each simulation. We then select "windows" along the reaction coordinate with a 0.1-0.2nm spacing between two consecutive windows, each with a harmonic potential centered at a position ξ_i^e. Umbrella sampling simulations [28] are conducted for each window, where the harmonic bias potential ω_i (Eq 2) is used with a spring constant value (K) between 800 and 8000 kJ mol^{-1} nm^{-2} to allow adequate sampling of microstates around the equilibrium position of each window.

$$\omega_i(\xi) = \frac{1}{2}K(\xi - \xi_i^e)^2 \quad (2)$$

We apply the WHAM [16] to generate the potential of mean force (PMF) curve, which yields the binding free energy. This is done by using the g_wham [15], [16] utility from the GROMACS package.

The AMBER03 force field is used in the atomistic simulations [29]. Starting structures are placed in a 10 × 18 × 10 nm simulation box solvated with TIP3P water [30] and sodium ions to neutralize the charge. The temperature is maintained at 298K using the Nosé-Hoover weak-coupling method [31], [32] and the pressure is maintained at 1 atm using a Parrinello-Rahman barostat [33], [34]. The cut-off distance for the short-range interaction is 1.4nm. Long-range electrostatic interactions are calculated using the Particle Mesh Ewald (PME) algorithm [35]. Periodic boundary conditions are applied in all three directions. Position restraints are applied on the heavy atoms of the MHCII, namely carbon, oxygen, and nitrogen atoms. The solvated system then undergoes a 10000-step steepest descent minimization followed by a 100ps constant-pressure (NPT) equilibration. A 3ns pulling simulation is used to generate the pulling trajectory, which is divided into windows. Each window is then sampled for 8ns with a 2fs time step. We use data from the last 7ns of each window in the final PMF calculation. Our window selection criteria are shown in Supplemental Information (section S.1).

2.3 Coarse Grain (CG) Simulations

The MARTINI protein force field is used in the CG simulations [13], [36]. Due to the coarse-grained nature of MARTINI, we set position restraints on MHCII backbone beads during the window simulations to stabilize the structure. We note that an alternative method, not used here, to maintain the stability of the CG protein structure during the simulation is to adopt the MARTINI-inherent elastic network approach [37]. We believe that the deformation of the peptide binding groove [27], [38], [39], which occurs during the initial pulling simulation, is more accurately accounted for at the atomistic level, and we therefore convert the configurations generated in the atomistic pulling trajectory to the CG representation, using *martinize.py* (http://cgmartini.nl/index.php/tools2/). We then solvate each starting structure in a simulation box of 10 × 18 × 10 nm with MARTINI water beads and add sodium ions to neutralize the system. The temperature is maintained at 325K using the Berendsen [40] thermostat, and the pressure is maintained at 1atm using the Berendsen [41] barostat. A temperature of 325K is chosen instead of 310K to keep the MARTINI water beads from freezing. We note that the calculated experimental binding free energies using 310K and 325K differ by less than 2kJ/mol, smaller than the typical error bar obtained in the simulations. Energy minimization and a 100ps constant volume (NVT) simulation with the entire peptide-MHCII complex frozen are carried out to allow adequate equilibration of the water beads. A 10ns constant pressure (NPT) simulation with 10fs time step and position restraint on the backbone beads of the MHCII is then conducted, followed by a 500ns umbrella sampling simulation, with a 25fs time step. We use the last 480ns simulation from each window in the final PMF calculation.

2.4 Simulation Systems

A list of all simulation systems studied in this work is given in Table II, and the simulation parameters for window simulations are available from the University of Michigan Library Deep Blue Data Depository (DOI pending). We study the effects of reaction coordinates, mutations, and segmentations using various peptide models. Specifically, the effect of COM and PEEL pulling coordinates on the binding free energy is studied using whole peptide models of both HA and CLIP peptides (System 1 and 2 in Table II). The effects of both single-residue and multiple-residue mutations are studied using System 3-4 in Table II. We test whether the free energy is sensitive to the choice of the segmentation location and/or segment length by using segmented peptides PKYV|KQN|TLK|LAT and PK|YV|KQ|NTL|KLAT (Systems 5-7 in Table II). We also repeat the single-residue mutation studies with the segmented peptides to demonstrate that the use of WHAM with a CG force field on a segmented peptide model is a robust approach to quantify the binding free energy difference resulting from a single-residue mutation. HA pulling simulations and HA segmentation simulation are carried out with both atomistic and CG force fields (Systems 1 and 8 in Table II), while only the CG force field is used in the remaining simulations of Table II. We carry out seven additional single-residue mutation studies on designed peptide sequences (HA_{Q311A}, GGY-$G_{309\text{-}318}$, GGA-$G_{309\text{-}318}$, GGG-$G_{309\text{-}318}$, AAF-$A_{309\text{-}318}$, AAW-$A_{309\text{-}318}$, and AAA-$A_{309\text{-}318}$), but because there are no experimental results available for these, the results are shown in the Supplemental Information in Figure S3, Figure S4, and Table S1.

3 Results and Discussion

As mentioned earlier, the previous attempt to apply WHAM using atomistic MD simulations to estimate the binding free energy between a 9 residue peptide and an MHC class I molecule did not agree well with experimental value [17]. We suspect that this may be due to the overwhelmingly large number of microstates that a long peptide needs to explore during each window along the reaction coordinate. The HA peptide contains 13 residues, which has considerably more degrees of freedom than a 9-residue peptide does. Therefore, we expect that the CG simulation might yield more accurate estimation of the binding free energy than atomistic pulling simulation by reducing the degrees of freemdom. In this study, we performed and compared both methods.

3.1 Whole Peptide Simulation

We first show a PMF diagram for pulling the entire peptide out of the binding groove using an atomistic simulation (Figure 3). The estimated binding free energy is 53.8 ± 2.9kJ/mol, which is in reasonably good agreement with the experimentally determined value of 46.6 kJ/mol. The standard deviation is calculated by the bootstrap analysis method [15] with 100 bootstraps. We label four regions on the PMF diagram, based on the location along the reaction coordinate where each anchor residue leaves the corresponding binding pocket. Note that we combine anchor residues THR313 and LEU314 into one region, because these two residues leave their binding pockets at roughly the same time in our pulling trajectory. For each region, we show a snapshot from the simulation near the point in the simulation at which the anchor residue(s) jumps out of the binding pocket. We observe two free energy barriers in the PMF diagram. The main barrier occurs in regions III and IV, which

corresponds to the free energy cost of pulling residues TYR308 and GLN311 out of binding pockets 1 and 2. The secondary energy barrier occurs in region II. After examining the trajectories carefully, we find this barrier is associated with the free energy cost of pulling anchor residue LEU316 out of binding pocket 5. The PMF decreases along the reaction coordinate after this first energy barrier between 2.2 and 2.5nm. Based on the COM distance between residues on the peptide and MHCII, this decrease in PMF is the result of the departure of segment NTL (anchor residue THR313, LEU314) from the MHCII binding groove. Although this is generally in line with the experimental result that THR313 interacts favorably with its binding pocket [20], [21], we cannot finely resolve the contribution to a single anchor residue, due to the limitations of the reaction coordinate we chose.

It is surprising to see that an atomistic full peptide pulling simulation gives a binding free energy that is in reasonable agreement with that reported experimentally. We therefore decided to test the limitation of using WHAM with an atomistic force field for whole-peptide pulling simulations. A true free energy does not depend on the path taken between the two states of interest. Therefore, pulling a peptide out of the binding groove along a different reaction coordinate should result in the same binding free energy, assuming that each reaction coordinate permits adequate sampling of configurations within each window and that the windows overlap sufficiently. So far we have discussed pulling the peptide out of the binding groove by using the entire peptide as the pulling group. Here we discuss the effect of choosing the two additional reaction coordinates in which the pulling groups are residues closer to the ends of the peptide, namely THR318 and PRO306 respectively, which we refer to as "PEEL" simulations.

We peel the entire peptide from the THR318 residue end using the atomistic force field, giving the results shown in Figure 4a which differs drastically from that of COM pulling. Based on the COM distances between residues on the peptide and MHCII, we can identify the portion of each PMF that corresponds to the pulling of NTLKLAT residues from the pocket. In the COM pulling, this corresponds to the distance along the reaction coordinate between 1.6 and 2.5nm (reproduced as Fig 4b), and in the peeling PMF, to the distance between 0.7 and 4.0nm (Fig 4a). The estimated binding free energies of NTLKLAT residues from both PMFs are around 9 kJ/mol. The remainder of the PMFs corresponds to the pulling of PKYVKQ residues. For this portion of the pulling the COM pulling simulation gives a free energy of 44 kJ/mol, while the peeling one gives a much higher value, 217 kJ/mol.

This significant difference in binding free energy suggests that it is extremely costly energetically to pull either one of the two anchor residues (GLN311 or TYR308) out of its corresponding binding pocket, which has not been reported in the experimental studies. We have inspected the relevant frames and found the following sources of this significant energy barrier. 1) An accurate WHAM free energy requires adequate sampling of all possible microstates in each umbrella-sampling window. However, in multiple cases, we find that the peptide becomes stuck in a localized set of configurations caused by strong interactions between a single residue and the binding pocket, preventing adequate sampling in that window. This biases the WHAM calculation, giving an unrealistically high energy barrier. We demonstrate one such incident in the following briefly, and in much more detail in the Supplemental Information (section S.3). Figure 5a shows the initial state of the window in which residue GLN311 is at the edge of leaving its binding pocket 2 (highlighted by the blue surface). Ideally, LYS310 and GLN311 should explore all microstates near this initial configuration in this umbrella simulation window, including configurations in which GLN311 resides both in and out of the binding pocket and LYS310 is both in and out of the surrounding binding groove. However, in our umbrella sampling simulation, LYS310 slips into the binding pocket 2, replacing GLN311, which has been pulled out, after which LYS310 remains inside the binding pocket throughout the 8ns simulation period of the window (Fig 5b). A thorough sampling of configurations between these two microstates therefore does not occur. A close examination of the distance between the COM of LYS310 on the peptide and of the residues GLN9 on the α chain of the MHCII molecule (termed GLN α9), and of TYR78 on the β chain of the MHCII molecule (termed TYR β 78), reveals that LYS310 remains in close proximity to these two residues throughout this window (Figure S5). In the following window along the reaction coordinate, both LYS310 and GLN311 explore microstates outside the binding groove throughout the simulation, but do not sample properly the transitional states through which LYS310 moves in and out of the binding groove. A positively charged residue such as LYS310 should not bind strongly within pocket 2, due to the nearby presence of residue ARG β 71 on MHCII β chain [20]. Therefore, the apparently strong interaction between LYS310 and binding pocket 2 imputed by WHAM is an artifact created by the choice of reaction coordinate, along which not all microstates between the bounded peptide and non-bounded peptide states have been sampled. The localized set of configurations where LYS310 is in close proximity to GLN α9 and TYR β 78 produced large

positive "pulling" forces, which resulted in positive PMF values when taking the weighted average during the WHAM calculation [28]. To resolve this issue, we need to identify an additional reaction coordinate that allows a thorough equilibration of the process by which LYS310 slips into the binding pocket and then leaves. If equilibration of these microstates are thorough, the PMF should show first a downward trend in free energy, corresponding to the favorable interaction between the LYS310 residue and the pocket, followed by a steep upward trend, corresponding to the large free energy required to pull LYS310 out of the binding pocket. Our chosen reaction coordinate, namely the COM of the whole peptide and that of the end residue, does not specifically track the COM distance between LYS310 and the MHCII pocket, and therefore does not force a sampling of the microstates involving this interaction. Note that we do not find large sampling errors in this window when using COM of the whole peptide as the pulling group, because the LYS310 leaves the binding groove before GLN311 does, unlike the trajectory we describe above in which the COM of the end residue is used as the pulling group. A 2-D reaction coordinate described by Huston *et al*. [42] might resolve the improper sampling described here, but requires much greater computational resources than are available to us. 2) Due to the size of the system and the number of simulation windows needed for constructing the PMF, our individual atomistic umbrella sampling simulation time is limited to 8ns per window, which we will soon demonstrate is likely far short of the time actually needed. We attempted the following two approaches to test if the sampling could be improved. First, we extended the simulation time of the two windows in question to 20 ns each. Second, we used two slightly different starting structures selected from the pulling trajectory whose equilibrium positions are very close (~0.01 nm) to those in the two original windows in question, and ran simulations of 20 ns for each of these. Based on the histogram, we find that although new microstates are sampled by using a new starting structure (Figure S5b), neither approach changed the histograms obtained along the defined reaction coordinate significantly (Figure S5c). This suggests that the simulation time is still insufficient to allow adequate sampling of all degrees of freedom. It would be computationally too expensive to extend the simulation time beyond 20 ns for all windows, however. later in the manuscript, we will estimate the simulation time required to achieve thorough sampling of all microstates using the coarse-grained simulation results.

It is evident from this work and Olaposi's work [17] that estimating the binding free energy by pulling the full peptide using an atomistic force field is prone to large error due to the inadequate sampling caused by small time step and limited simulation time. We note that there are other advanced sampling methods at the atomistic level that provide estimations of the binding free energies. One such method is thermodynamic integration, wherein the non-bonded interactions between the peptide and the solvents are slowly turned off along the reaction coordinate [43]. However, such methods do not reveal the valuable free energy landscape provided by a PMF plot. Another method is metadynamics or well-tempered metadynamics, which allows one to explore the two- or multi-dimensional free energy landscape. This method has been used widely to extract the binding free energy surface of the constituents of many organic small-molecule systems [44–48]. Compared to the umbrella sampling method used in this study, this approach is as computationally intensive. The small time step and limited simulation time can be also addressed with CG simulations, which employ larger time steps and achieve much longer simulation times. We therefore used the CG force field in whole-peptide pulling simulations and assessed the limitations of this approach. The PMF from the CG simulation (Fig 6) gives a binding free energy of 47.3 ± 1.4 kJ/mol, which is in excellent agreement with the binding affinity measured in the experiments. We divide the CG PMF diagram into three regions, each corresponding to at least one anchor residue leaving the binding pocket. In the CG simulations, all three anchor residues LEU316, LEU314, and THR313 leave the binding groove at nearly the same position on the reaction coordinate, and therefore are combined into a single region on the PMF diagram. Otherwise, the CG PMF is similar to the atomistic one. We do not observe a large secondary energy barrier in the CG PMF, while we do in the atomistic one. Carefully examining the individual window simulation trajectories, we find that NTL residues (anchor residue THR313) leave the binding pocket 3 (at reaction coordinate around 2.4 nm) soon after KLAT residues (anchor residue LEU316) leave the binding pocket 4 (at reaction coordinate around 2.0 nm).

We show the PMF curves generated from CG PEEL simulations in Figure 7(a). From these two PMF curves, we obtain binding free energies of 43.1±1.6 kJ/mol and 48.1±1.7 kJ/mol by pulling from the PRO306 end and the

THR318 end, respectively. Both results are in good agreement with the experimentally determined value, as well as with the CG result obtained when the entire peptide is the pulling group. In addition, we are able to achieve much longer time scales (500ns per window) using CG simulations.

Although the sequence of the CLIP peptide is different from that of the HA peptide, the former has a binding free energy of 45.1 kJ/mol [38], which is comparable to that of the HA peptide. We mutate the HA peptide to CLIP and carry out the pulling simulation at the CG level with three reaction coordinates, namely COM and two PEEL reaction coordinates and obtain a binding free energy of 44.5 ± 2.2 kJ/mol using the COM reaction coordinate and 42.0 ± 1.5 kJ/mol and 53.8 ± 1.1 kJ/mol using the two PEEL reaction coordinates (Fig 7b). These results are in reasonable (although not excellent) agreement with each other and again demonstrate that, compared to atomistic simulations, the binding free energies from the CG simulations are less sensitive to the choice of the reaction coordinate.

Painter *et al.* has deposited a crystal structure of CLIP-MHCII in the protein data bank (PDB ID: 3QXA) [39]. However, comparing the MHC crystal structures in 1DLH (MHCII-HA peptide) and 3QXA (MHCII-CLIP peptide), we find the protein structures around the binding pocket 1(P1) are different, even though the sequences of both MHCII molecules were identical. This is likely due to the different pH conditions where these two crystal structures were determined (pH=5.5 and 7.0 respectively). Painter *et al*. has shown that minor structural difference in the binding pocket, particularly at the P1 position, leads to difference in the binding free energy [39]. Indeed, when we conduct a whole-peptide pulling simulation of CLIP-MHCII using this crystal structure (3QXA), we obtain a modestly different binding free energy of 35.7 ± 2.8 kJ/mol (PMF shown in Figure S6) from the one that we obtain using the CLIP-MHCII structure where the CLIP peptide is mutated from the HA peptide (44.5 ± 2.2 kJ/mol, shown in Figure 7b). Note that due to the position restraints set on the MHC backbone during the CG simulations, the differences in the crystal structures would be inherited throughout the simulation. Because removing the position restraints in the CG simulation causes the CG protein structure to destabilize, we are not able to investigate further the effect of MHC II conformational change on the binding free energy.

To estimate the simulation time required to achieve the adequate sampling of microstates needed to obtain a converged PMF, we plot in Figure 8 the PMFs for HA and CLIP peptides obtained using the data from the first 20 ns, 50 ns, 200 ns, 500 ns, and 1000 ns of each window simulation. The PMFs depend on the window simulation time up to around 200 ns runs, with good convergence for runs longer than this for both HA and CLIP. The thoroughness of sampling in each window is also demonstrated by the symmetry of the histogram. To quantify this symmetry, for each PMF, we fit Gaussian distributions to all histograms used to construct it, and plot the total fitting error in the insets of Figure 8. We find the fitting errors decrease greatly with increasing window duration, up to around 200-500 ns of sampling, and plateau thereafter, supporting the conclusion that the minimum simulation time to achieve thorough sampling of microstates at the coarse-grained level is around 200 ns for each window. Since the MARTINI force field, with its softer interactions, speeds up molecular mobility by a factor of around four relative to atomistic simulations, the 200ns minimum simulation time per window at coarse-grained level corresponds to 800 ns at atomistic level [49]. Such a time scale is very challenging to achieve at atomistic scale, where a total of 16 μs simulation time would be required for a single PMF with 20 windows. Moreover, there are significantly more degrees of freedom at the atomistic level, therefore the total number of microstates need to be sampled at atomistic level is likely to be significant greater than at coarse-grained level and the minimum simulation time required for each atomistic simulation window may be well beyond 800 ns.

3.2 Estimating the Binding Free Energy between Mutated Peptides and MHCII

We have obtained an accurate estimation of the binding free energy of the HA peptide and of the CLIP peptide using WHAM with CG simulations. Since our goal is to develop a robust method to estimate the binding free energy between an arbitrary peptide and MHCII, we further apply WHAM with this CG force field to estimate the binding free energies of mutated peptide sequences (Systems 3 and 4 in Table II).

We first carry out the pulling simulation using the peptide HA_{Y308A}, which is a single-residue mutation of the HA peptide at position 308. This mutation changes the tyrosine at binding pocket 1 to alanine, which results in a

much weaker binding interaction. The simulated binding free energy of the HA_{Y308A} is 33.5 ± 1.2 kJ/mol (Fig 9a), which is about 6 kJ/mol higher than the experimental result of 27.5 kJ/mol [50].

Next, we carry out a pulling simulation with multiple residues mutated from the HA peptide. We consider the YAK peptide, which shares only two residues, namely TYR308 and LYS315, with the HA peptide, with all remaining residues mutated to alanine. Because TYR308 and LYS315 are two major anchor residues, the experimental binding free energy of the YAK peptide is 41.1 kJ/mol [42], only 5 kJ smaller than that of the HA peptide. The simulated binding free energy for the YAK peptide is 44.6 ± 2.6 kJ/mol (Fig 9b), which is about 3 kJ higher than the experimental result.

We also carry out four additional multiple residues mutation studies on designed peptide sequences (HA_{Q311A}, GGY-$G_{309\text{-}318}$, GGA-$G_{309\text{-}318}$, GGG-$G_{309\text{-}318}$), but because there are no experimental results available for these, the results are shown in the Supplemental Information in Figure S4. We found GGA-$G_{309\text{-}318}$ peptide has a fairly strong binding free energy of 28.3 ± 1.9 kJ/mol, which is not expected because the peptide only contains residues that have weak interactions with binding pockets, namely glycine and alanine. We inspect the window simulations for GGA-$G_{309\text{-}318}$ peptide and find inadequate samples in a number of windows. In Figure 10 we illustrate one such instance. Figure 10a shows the starting structure of a window simulation at a distance 2.7nm along the reaction coordinate, where ALA308 is leaving the binding pocket. If the reaction coordinate is defined properly, the peptide should explore all possible configurations where ALA308 is near the binding pocket 1 (P1, shown in orange). However, we find that most of samples collected in the window simulation correspond to the GGA peptide interacting with various regions on the MHCII surface (Figure 10b). These inadequate samples, as we discussed in section 3.1, result in errors in the PMF curve. We note that we have not found similar instances in the GGY-$G_{309\text{-}318}$ peptide simulation. This is likely due to the strong interaction between TYR308 and the binding pocket 1 (P1), preventing the peptide from slipping out of the binding groove and interacting with the MHCII surface. Nevertheless, even with the improved sampling allowed by CG MD, a 2-D reaction coordinate is still recommended to ensure proper sampling of the whole peptide configuration space, especially for peptides that have weak interactions with the binding groove.

3.4 Segmented Peptide Simulation

From the above, we find that umbrella sampling simulation with the CG force field can provide a better estimation of the binding free energy than is generally obtained with atomistic force field, but there remain some limitations. As we discussed, the chosen 1-D reaction coordinate in this work, namely the COM distance between the peptide and MHCII, does not guarantee adequate sampling of the entire configuration space of a 13-residue peptide, which may result in errors in the PMF curves (e.g. Fig 5). In the hope of addressing this issue, we propose segmenting the peptide, wherein the 13-residue long peptide is truncated into four small segments, based on the locations of their anchor residues. Then, we estimate the binding free energy of each segment and sum up these up to obtain the total binding free energy for the whole peptide. Although we are still employing a 1-D reaction coordinate defined as the COM distance between the peptide segment and MHCII, the configuration space is far smaller for a 3-mer peptide segment than for a 13-mer peptide. Therefore, combining the small configuration space and the long simulation time, we expect that CG MD with a segmented peptide using the 1-D reaction coordinate might be sufficient to allow adequate sampling in all window simulations.

We have tabulated the simulated binding free energies between each peptide segment and its corresponding binding pocket from CG simulations in Table III. We combine the two weak anchor residues THR315 and LEU316 into one segment, and no segment has fewer than three residues. The estimated total binding free energy for the entire peptide, obtained by summing up the four individual contributions from the segments, is 45.2±3.8 kJ/mol, which is in excellent agreement with the experimentally determined binding free energy of 46.6 kJ/mol. Segment PKY has the strongest binding interaction with the MHCII binding pocket 1, contributing over half the total binding free energy. Segments VKQ and KLAT have weaker binding interactions with their corresponding binding pockets 2 and 5. Segment NTL shows a negative binding free energy, suggesting that this segment has an energetically unfavorable interaction with the binding pocket. We compare the rank order of the binding affinity with the known experimental observation of the HA-MHCII interaction. From the experiments, TYR308 has the largest binding affinity with MHCII, followed by GLN311, LEU314 and LEU316 [20]. Specifically, TYR308 stays in the deepest and largest binding pocket and has the strongest binding interaction. GLN311 and LEU316 are located in two separate shallow binding pockets and have weaker binding interactions with MHCII. Jardetzky

et al. have shown that THR313 does not bind to MHCII favorably [21], which is consistent with the negative binding free energy obtained for segment NTL from our CG simulations. We also conduct atomistic simulations with the segmented peptide model, but the results only agreed with the experimental trend qualitatively (Figure S8).

3.6 Effect of the Segmentation Location and Peptide Segment Length

Although it is very promising that we have obtained the correct total binding free energy using the segmented peptide, we wish to investigate whether the resulting binding free energy is sensitive to the segmentation location and/or segment length. This will not only demonstrate the robustness of this approach, but will also be beneficial for predicting the change of peptide binding free energy caused by a few mutated residues by conducting pulling simulations of short segments containing each single-point mutation.

We test the effect of the segmentation locations along the HA peptide by using a different segmentation pattern (PKYV|KQN|TLK|LAT). The simulated binding free energies are shown in Table IV. We again observe that the segment containing TYR308 residue has the largest binding free energy, and the segment with THR313 contributes a negative binding free energy. These results agree with the ones shown in Table III; the summed binding free energy is 43.3±3.4 kJ/mol, again in good agreement with the experimentally determined binding free energy of 46.6 kJ/mol.

We also test the effect of segment length using the segmentation pattern PK|YV|KQ, with the rest of the segmentation (NTL|KLAT) remaining in the binding pocket. From Table V, the total binding free energy of the PKYVKQ residues obtained by combining the contributions from three 2-residue segments is 35.1±2.6 kJ/mol, which deviates significantly from a binding free energy of 47.1±2.1 kJ/mol for the same sequence obtained by combining the contributions from two 3-residue segments (adding the first two rows in Table III). In addition, experimental results [20] showed that the YV segment has a stronger interaction with MHCII does the KQ segment, but the simulation results for 2-residue segments show otherwise. We suspect a short segment can slide within the binding groove causing errors in the PMF curves. Thus, the 3-residue segmentation appears to be an appropriate segment length for obtaining robust binding free energy.

3.7 Estimating the Binding Free Energy of Mutated Peptides using Segmented Peptide Model

To consider the possibility that a mutation on one segment might affect the binding free energy of the neighboring segment, we repeat the pulling simulations of the mutant HA_{Y308A} peptide using the segmented peptide. Specifically, we pull the segment PKA (which contains the mutation) and segment VKQ (which neighbors segment PKA) from the binding groove and compare the simulated binding free energies of these two segments to those of the corresponding segments in the HA peptide, leaving the other segments of the peptide in the binding pocket.

From Table VI, the binding free energy of the segment PKA clearly decreases due to the mutation of Y to A. The binding free energy for the segment VKQ, however, appears to be almost the same, which suggests the mutation in the PKA segment has little effect on the neighboring 3-residue VKQ segment. In addition, we observe no conformational change in segments NTLKLAT and their corresponding binding pockets on MHCII based on visual inspection before and after the mutation, suggesting the mutation has no effect on the binding free energy of these two segments either. Therefore, we can add the binding free energy of PKA|VKQ (Table VI) and NTL|KLAT (Table III) to obtain the total binding free energy, which is close to the experimental result.

4 Conclusions

We performed both atomistic and course-grained (CG) Weighted Histogram "pulling" simulations of both whole peptides and segmented peptides from the MHCII molecule along center of mass (COM) and PEEL reaction coordinates. All simulation results are summarized in Table VII. We find the binding free energy estimations obtained using CG MD are less prone to large error caused by inadequate sampling than when using atomistic MD (Tests 1, 2 and 3 in Table VII). This is likely because we can achieve a much longer simulation time (almost 500-fold) with CG MD than with atomistic MD, allowing sampling of enough microstates in the former. Nevertheless, estimating the binding free energy by pulling the whole peptide using CG MD still has limitations when sampling weak peptide-binding groove interactions. This is due to the fact that the chosen 1-D reaction coordinate in our work does not guarantee appropriate sampling of the voluminous configuration space that a 13-

resdiue peptide can adopt. We therefore proposed segmenting the peptide to demonstrate that a 1-D reaction coordinate is sufficient to extract accurate binding free energies of 3-mer segments (Tests 5, 6, 7, and 9 in Table VII) with the remainder of the peptide left in the pocket, and that with a single anchor residue in each segment the total free energy can be obtained by adding the free energy contributions of these short segments. In addition, the segmented peptide model allows us to rapidly predict the effect of single-residue mutations on the binding free energy. Using much larger-scale computations, it should be possible to apply our CG umbrella sampling method with this segmented peptide strategy to predict binding free energies of hundreds or thousands of peptide variants to MHC class II molecules, with implications for the rational design of peptides for immunotherapy.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rlarson@umich.edu

NOTES

The authors declare no competing financial interest.

ACKNOWLEDGEMENTS

We thank Dr. Shihu Wang and Kyle Huston for helping with the simulation setup. We also thank Mason Smith and Eshita Khera for many helpful and stimulating discussions. This work was supported by a University of Michigan MCube grant.

SUPPORTING INFORMATION

Additional discussions regarding the details of the pulling simulation, binding free energies of designed peptide sequences, and additional PMFs are supplied in the Supporting Information.

REFERENCES

[1] J. S. Blum, P. A. Wearsch, and P. Cresswell, "Pathways of Antigen Processing," *Annu. Rev. Immunol.*, vol. 31, pp. 443–473, 2013.

[2] J. M. Blander and P. Nair-Gupta, "An updated view of the intracellular mechanisms regulating cross-presentation," *Mol. Innate Immun.*, vol. 4, p. 401, 2013.

[3] "Alternative generation of MHC class II-restricted epitopes: Not so exceptional?" [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0161589012004348. [Accessed: 31-Jul-2016].

[4] C. Watts, "The endosome–lysosome pathway and information generation in the immune system," *Biochim. Biophys. Acta BBA - Proteins Proteomics*, vol. 1824, no. 1, pp. 14–21, Jan. 2012.

[5] P. E. Jensen, "Recent advances in antigen processing and presentation," *Nat. Immunol.*, vol. 8, no. 10, pp. 1041–1048, Oct. 2007.

[6] G. L. Pira, F. Ivaldi, P. Moretti, and F. Manca, "High Throughput T Epitope Mapping and Vaccine Development," *J. Biomed. Biotechnol.*, p. 325720, 2010.

[7] M. Nielsen, O. Lund, S. Buus, and C. Lundegaard, "MHC Class II epitope predictive algorithms," *Immunology*, vol. 130, no. 3, pp. 319–328, Jul. 2010.

[8] H.-G. Rammensee, "Chemistry of peptides associated with MHC class I and class II molecules," *Curr. Opin. Immunol.*, vol. 7, no. 1, pp. 85–96, Feb. 1995.

[9] D. Rognan, L. Scapozza, G. Folkers, and A. Daser, "Molecular Dynamics Simulation of MHC-Peptide Complexes as a Tool for Predicting Potential T Cell Epitopes," *Biochemistry (Mosc.)*, vol. 33, no. 38, pp. 11476–11485, Sep. 1994.

[10] R. Yaneva, S. Springer, and M. Zacharias, "Flexibility of the MHC class II peptide binding cleft in the bound, partially filled, and empty states: A molecular dynamics simulation study," *Biopolymers*, vol. 91, no. 1, pp. 14–27, Jan. 2009.

[11] U. Omasits *et al.*, "Analysis of key parameters for molecular dynamics of pMHC molecules," *Mol. Simul.*, vol. 34, no. 8, pp. 781–793, Jul. 2008.

[12] S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse Grained Model for Semiquantitative Lipid Simulations," *J. Phys. Chem. B*, vol. 108, no. 2, pp. 750–760, Jan. 2004.
[13] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The MARTINI Coarse-Grained Force Field: Extension to Proteins," *J. Chem. Theory Comput.*, vol. 4, no. 5, pp. 819–834, May 2008.
[14] S. J. Marrink and D. P. Tieleman, "Perspective on the Martini model," *Chem. Soc. Rev.*, vol. 42, no. 16, pp. 6801–6822, Jul. 2013.
[15] J. S. Hub, B. L. de Groot, and D. van der Spoel, "g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates," *J. Chem. Theory Comput.*, vol. 6, no. 12, pp. 3713–3720, Dec. 2010.
[16] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comput. Chem.*, vol. 13, no. 8, pp. 1011–1021, Oct. 1992.
[17] Omotuyi I Olaposi and Hamada Tsuyoshi, "Theoretical Dynamics and Energetics of HLA-A2/SLYNTVATL Interaction," *Am. J. Bioinforma. Res.*, vol. 5, no. 1, pp. 1–8, 2015.
[18] A. May *et al.*, "Coarse-grained versus atomistic simulations: realistic interaction free energies for real proteins," *Bioinformatics*, pp. 1–9, Nov. 2013.
[19] H. Zhang *et al.*, "Limitations of Ab Initio Predictions of Peptide Binding to MHC Class II Molecules," *PLOS ONE*, vol. 5, no. 2, p. e9272, Feb. 2010.
[20] L. J. Stern *et al.*, "Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide," *Nature*, vol. 368, no. 6468, pp. 215–221, Mar. 1994.
[21] T. S. Jardetzky, J. C. Gorga, R. Busch, J. Rothbard, J. L. Strominger, and D. C. Wiley, "Peptide binding to HLA-DR1: a peptide with most residues substituted to alanine retains MHC binding.," *EMBO J.*, vol. 9, no. 6, pp. 1797–1803, Jun. 1990.
[22] A. K. Sato *et al.*, "Determinants of the Peptide-induced Conformational Change in the Human Class II Major Histocompatibility Complex Protein HLA-DR1," *J. Biol. Chem.*, vol. 275, no. 3, pp. 2165–2173, Jan. 2000.
[23] Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version 1.8," Nov-2015.
[24] C. A. Painter, A. Cruz, G. E. López, L. J. Stern, and Z. Zavala-Ruiz, "Model for the Peptide-Free Conformation of Class II MHC Proteins," *PLOS ONE*, vol. 3, no. 6, p. e2403, Jun. 2008.
[25] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A message-passing parallel molecular dynamics implementation," *Comput. Phys. Commun.*, vol. 91, no. 1–3, pp. 43–56, Sep. 1995.
[26] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, Feb. 1996.
[27] M. Zacharias and S. Springer, "Conformational Flexibility of the MHC Class I α1-α2 Domain in Peptide Bound and Free States: A Molecular Dynamics Simulation Study," *Biophys. J.*, vol. 87, no. 4, pp. 2203–2214, Oct. 2004.
[28] J. Kästner, "Umbrella sampling," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 1, no. 6, pp. 932–942, Nov. 2011.
[29] Y. Duan *et al.*, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *J. Comput. Chem.*, vol. 24, no. 16, pp. 1999–2012, Dec. 2003.
[30] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Jul. 1983.
[31] S. Nosé, "A molecular dynamics method for simulations in the canonical ensemble," *Mol. Phys.*, vol. 52, no. 2, pp. 255–268, Jun. 1984.
[32] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, no. 3, pp. 1695–1697, Mar. 1985.

[33] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, Dec. 1981.
[34] S. Nosé and M. L. Klein, "Constant pressure molecular dynamics for molecular systems," *Mol. Phys.*, vol. 50, no. 5, pp. 1055–1076, Dec. 1983.
[35] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, Jun. 1993.
[36] D. H. de Jong *et al.*, "Improved Parameters for the Martini Coarse-Grained Protein Force Field," *J. Chem. Theory Comput.*, vol. 9, no. 1, pp. 687–697, Jan. 2013.
[37] X. Periole, M. Cavalli, S.-J. Marrink, and M. A. Ceruso, "Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition," *J. Chem. Theory Comput.*, vol. 5, no. 9, pp. 2531–2543, 2009.
[38] J. A. Zarutskie *et al.*, "A conformational change in the human major histocompatibility complex protein HLA-DR1 induced by peptide binding," *Biochemistry (Mosc.)*, vol. 38, no. 18, pp. 5878–5887, May 1999.
[39] C. A. Painter, M. P. Negroni, K. A. Kellersberger, Z. Zavala-Ruiz, J. E. Evans, and L. J. Stern, "Conformational lability in the class II MHC 310 helix and adjacent extended strand dictate HLA-DM susceptibility and peptide exchange," *Proc. Natl. Acad. Sci.*, vol. 108, no. 48, pp. 19329–19334, Nov. 2011.
[40] H. J. C. Berendsen, "Transport Properties Computed by Linear Response through Weak Coupling to a Bath," in *Computer Simulation in Materials Science*, M. Meyer and V. Pontikis, Eds. Springer Netherlands, 1991, pp. 139–155.
[41] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984.
[42] K. J. Huston and R. G. Larson, "Reversible and Irreversible Adsorption Energetics of Poly(ethylene glycol) and Sorbitan Poly(ethoxylate) at a Water/Alkane Interface," *Langmuir*, vol. 31, no. 27, pp. 7503–7511, Jul. 2015.
[43] S. Genheden and U. Ryde, "How to obtain statistically converged MM/GBSA results," *J. Comput. Chem.*, vol. 31, no. 4, pp. 837–846, Mar. 2010.
[44] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc. Natl. Acad. Sci.*, vol. 99, no. 20, pp. 12562–12566, Oct. 2002.
[45] Alessandro Laio and Francesco L Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Rep. Prog. Phys.*, vol. 71, no. 12, p. 126601, 2008.
[46] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, "Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics," *J. Am. Chem. Soc.*, vol. 128, no. 41, pp. 13435–13441, Oct. 2006.
[47] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein, "Metadynamics as a Tool for Exploring Free Energy Landscapes of Chemical Reactions," *Acc. Chem. Res.*, vol. 39, no. 2, pp. 73–81, Feb. 2006.
[48] T. Mandal, R. L. Marson, and R. G. Larson, "Coarse-grained modeling of crystal growth and polymorphism of a model pharmaceutical molecule," *Soft Matter*, vol. 12, no. 39, pp. 8246–8255, Oct. 2016.
[49] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations," *J. Phys. Chem. B*, vol. 111, no. 27, pp. 7812–7824, Jul. 2007.
[50] F. Wen, O. Esteban, and H. Zhao, "Rapid identification of CD4+ T-cell epitopes using yeast displaying pathogen-derived peptide library," *J. Immunol. Methods*, vol. 336, no. 1, pp. 37–44, Jul. 2008.
[51] F. Wen, "Engineering of Yeast Displaying Singe-chain MHC-peptide Complexes for Biomedical Applications," 2006.

Table I: List of peptide sequences used in this work. The HA peptide is the wild type peptide sequence in PDB entry 1DLH [20]. The YAK and HA_{Y308A} peptides are mutated sequences studied in Aaron et al. [22]. For the HA_{Y308A} peptide sequence, a single point mutation from tyrosine to alanine is introduced. The peptide is named using the convention that all residue numbering follows that of the HA peptide. The peptide name HA_{Y308A} therefore refers to HA peptide with tyrosine at 308 position mutated to alanine.

Peptide name	Peptide Sequence
HA	PKYVKQNTLKLAT
CLIP	VSKMRMATPLLMQA
YAK	AAYAAAAAAKAAA
HA_{Y308A}	PKAVKQNTLKLAT

Table II: Summary of the simulations. The vertical separators (|) indicate the segmentation locations

No.	Peptide Model Type	System	Peptide Sequence	No.	Peptide Model Type	System	Peptide Sequence
1	Whole	HA	PKYVKQNTLKLAT	5	Segmented	HA	PKY|VKQ|NTL|KLAT
2		CLIP	SKMRMATPLLMQAV	6		HA	PKYV|KQN|TLK|LAT
3		YAK	AAYAAAAAAKAAA	7		HA	PK|YV|KQ|NTL|KLAT
4		HA_{Y308A}	PKAVKQNTLKLAT	8		HA_{Y308A}	PKA|VKQ|NTL|KLAT

Table III: Simulated CG binding free energies for all four segments in the segmented peptide model. PMF curves for each segment are provided in the Supplementary Information (Figure S7)

Segment Sequence	Binding Free Energy (kJ/mol)
PKY	26.2±1.1
VKQ	20.9±2.5
NTL	-20.2±1.4
KLAT	18.4±2.2
Sum of the four segments	45.3±3.8

Table IV: Same as Table III, except with a shifted segmentation pattern

Segment Sequence	Binding free energy (kJ/mol)
PKYV	26.5±1.4
KQN	11.3±0.9
TLK	-12.4±2.6
LAT	17.9±1.4
Sum of the four segments	43.3±3.4

Table V: CG pulling simulation results for 2-residue segments

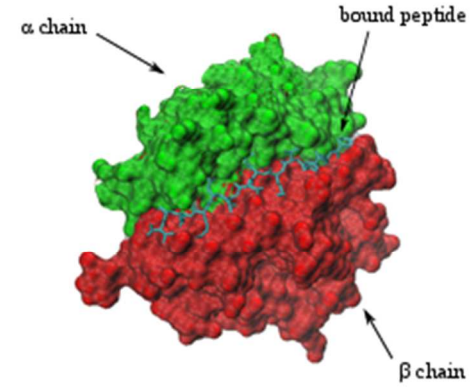
Segment Sequence	Binding free energy(kJ/mol)
PK	3.2±0.6
YV	14.5±1.5
KQ	17.4±2.1
Sum of the three segments	35.1±2.6

Table VI: CG simulation results for HA_{Y308A} and HA

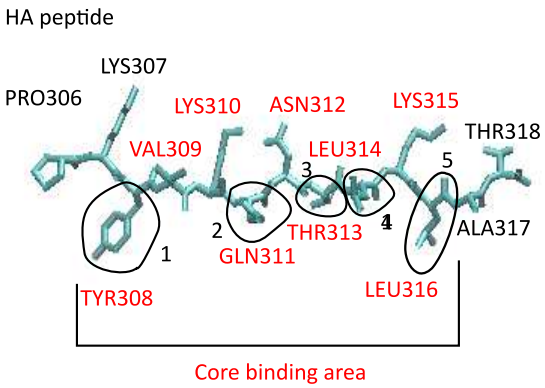
Segment Sequence	Binding free energy(kJ/mol)	Segment Sequence	Binding free energy (kJ/mol)
PKA	9.1±1.8	PKY	26.2±1.1
VKQ (in HA_{Y308A})	22.8±1.4	VKQ (in HA)	20.9±2.5

Table VII: Summary of main simulation results compared with experimental results

	Effect Studied	Peptide Model	Pulling Coordinate	Peptide	Sim. ΔG (kJ/mol)	Exp. ΔG (kJ/mol)	Sim. $\Delta\Delta G_{Y308A}$ (kJ/mol)	Exp. $\Delta\Delta G_{Y308A}$ (kJ/mol)
1	Reaction Coordinate (Atomistic)	Whole	COM	HA	53.8 ± 2.9			
		Whole	PEEL-THR318	HA	>217			
2	Reaction Coordinate (CG)	Whole	COM	HA	47.3 ± 1.4	46.6		
		Whole	PEEL-PRO306	HA	43.1 ± 1.6			
		Whole	PEEL-THR318	HA	48.1 ± 1.7			
3	Reaction Coordinate (CG)	Whole	COM	CLIP	44.6 ± 2.6			
		Whole	PEEL-VAL305	CLIP	42.0 ± 1.5	45.1		
		Whole	PEEL-ALA318	CLIP	53.8 ± 1.1			
4	Multiple Residues Mutation	Whole	COM	HA	47.3 ± 1.4	46.6		
		Whole	COM	YAK	44.6 ± 2.6	41.1		
5	Peptide Segmentation	Whole	COM	HA	47.3 ± 1.4			
		Segmented	COM	HA	45.3 ± 3.8			
6	Segmentation Location	Segmented	COM	HA	45.3 ± 3.8	46.6		
		Segmented	COM	HA	43.3 ± 3.4			
7	Segment Length	Segmented	COM	HA	45.3 ± 3.8			
		Segmented	COM	HA	33.3 ± 4.4			
8	Single Residue Mutation	Whole	COM	HA	47.3 ± 1.4		13.8 ± 1.8	19.1
		Whole	COM	HA_{Y308A}	33.5 ± 1.2			
9	Single Residue Mutation	Segmented	COM	HA	45.3 ± 3.8		15.2 ± 5.1	
		Segmented	COM	HA_{Y308A}	30.1 ± 3.4			



(a)



(b)

Figure 1: a) Crystal structure of the MHC class II molecule with influenza virus haemagglutinin 306-318 (HA) peptide (PDB ID: 1DLH). The α and β chains of the MHC II molecule are shown in green and red, respectively. The HA peptide inside the MHC class II binding groove is shown in cyan. b) Detailed view of the HA peptide (sequence: PKYVKQNTLKLAT). Throughout, we identify residues by their 3-letter name followed by location ID (e.g. TYR308 for the 308th residue tyrosine of the HA peptide). The nine residues between TYR308 and LUE316 are buried within the core binding area of the MHCII. Specifically, residues highlighted in circles (i.e. TYR308, GLN311, THR313, LEU314, and LEU316) are the anchor residues, and are buried in the five binding pockets in the MHCII binding groove, which we number sequentially from left to right (e.g. TYR308 is in binding pocket 1, and LEU316 is in binding pocket 5).

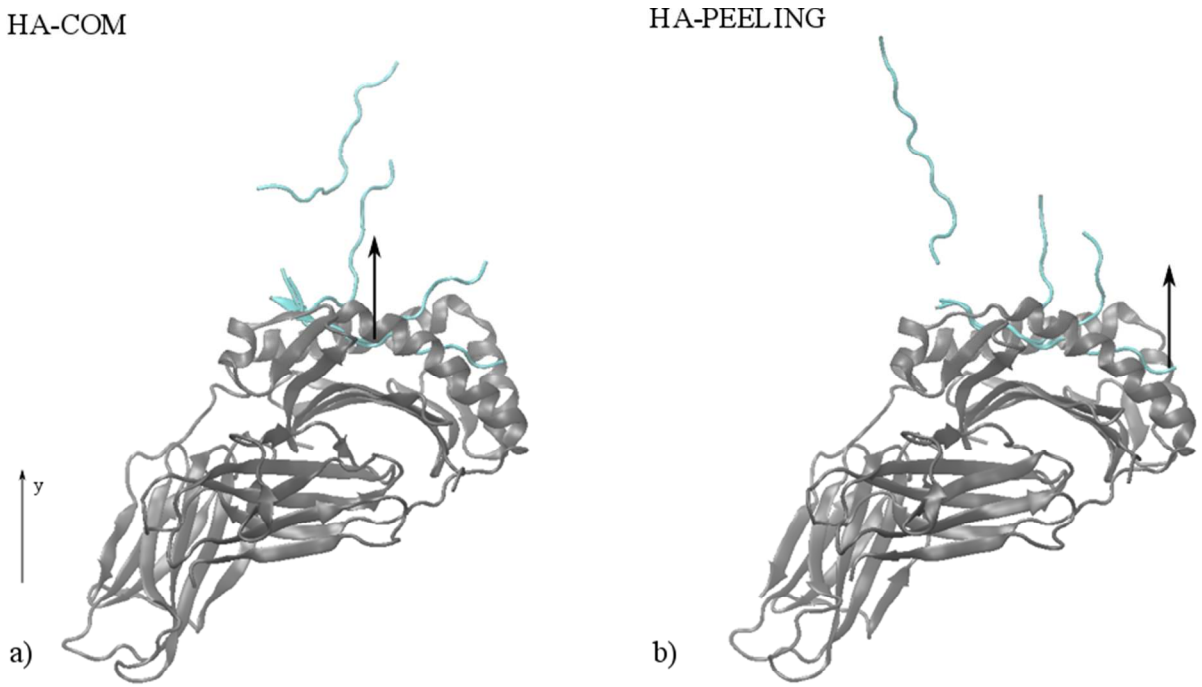


Figure 2: Reaction coordinates are generated from pulling simulations with different pulling groups. The structure in grey is the MHCII molecule. The structure in cyan is the peptide being pulled. The three cyan structures in both a) and b) are the same peptide, captured at different times during the pulling simulation. a) The pulling group is the entire peptide, and the pulling direction is along the y axis shown. This reaction coordinate is referred to as the center-of-mass (COM) reaction coordinate; b) the pulling group is residue THR318, and the pulling direction is also along the y axis; this reaction coordinate is referred as the peeling (PEEL) reaction coordinate. A similar PEEL reaction coordinate is obtained by defining residue PRO306 at the opposite end of the peptide as the pulling group.

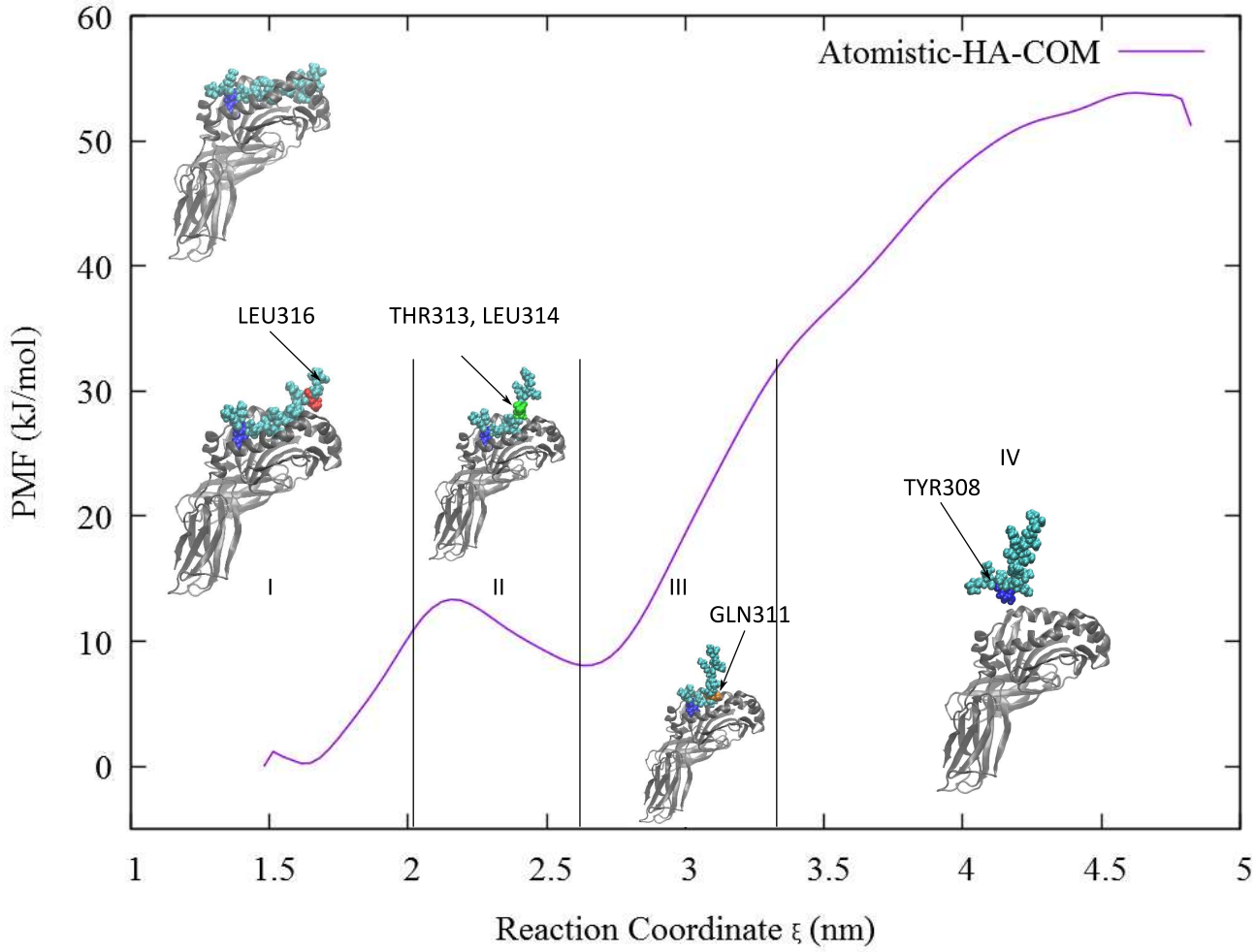


Figure 3: Atomistic PMF curve for pulling the HA peptide from MHCII along a COM reaction coordinate. Four regions are defined, along with one representative snapshot from each. MHCII is colored grey and the peptide is colored cyan, with red LEU316, green THR313 and LEU314, orange GLN311, and blue TYR308. We include TYR308 colored in blue in all snapshots to mark the orientation of the peptide. The snapshot at the upper left corresponds to the early stage of the pulling process where the entire peptide is buried in the binding groove. In Region I, the residue LEU316 leaves the binding pocket 5, while in Region II anchor residues THR313 and LEU314 are leaving binding pockets 3 and 4. In Region III anchor residue GLN311 is pulled out of binding pocket 2. Region IV corresponds to the final stage where TYR308 and whole peptide are pulled out of the binding groove.

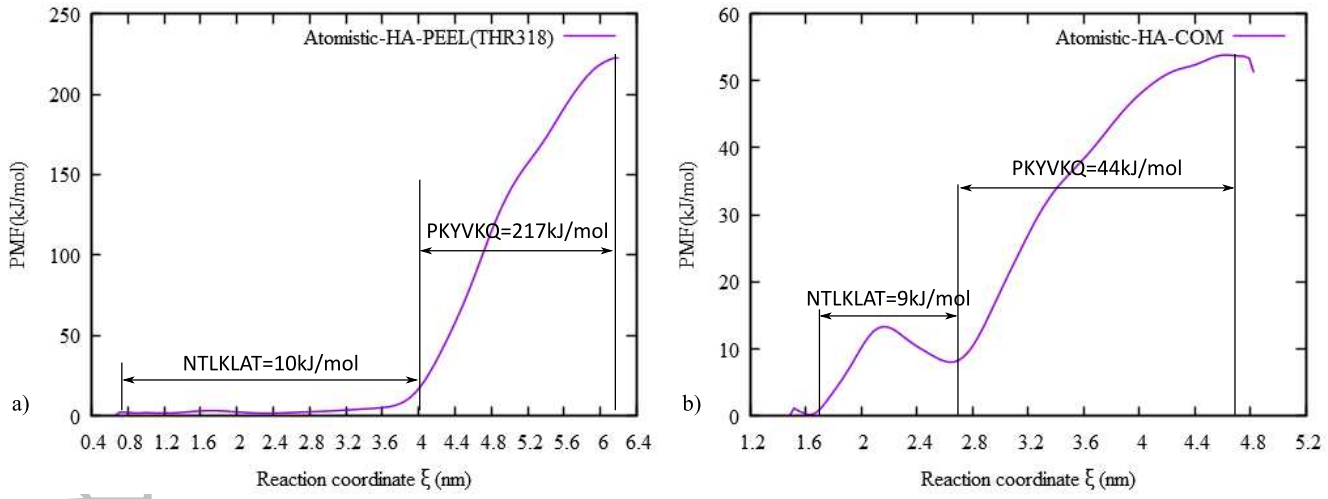


Figure 4: Atomistic PMF curves. a) PMF from peeling the HA peptide from the THR318 end; b) Same as Figure 3 where the entire peptide is pulled.

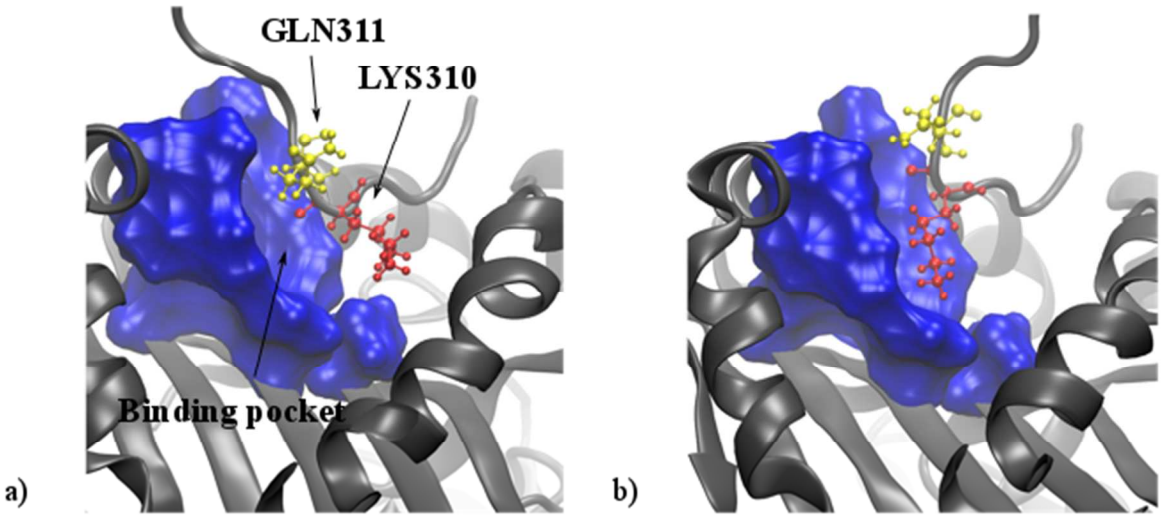


Figure 5: Figures show 2 snapshots in the window simulation at 4.9nm along the peeling reaction coordinate. Figure a) shows the initial configuration. GLN311 is colored yellow, LYS310 red, and the blue surface is the binding pocket for GLN311 (P3). Figure b) is a snapshot from later in this same window simulation.

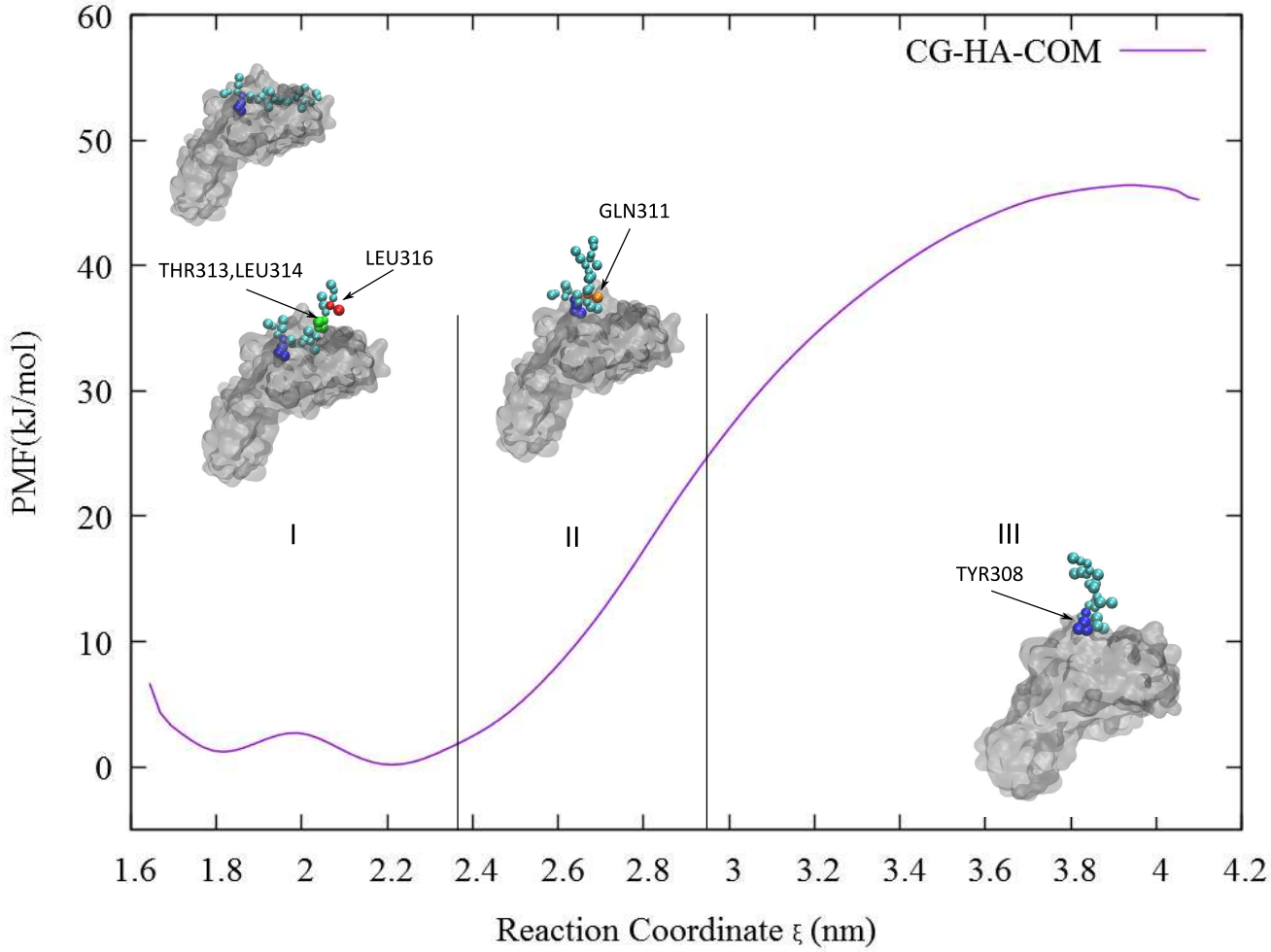


Figure 6: CG PMF curve for pulling the HA peptide from MHCII along a COM reaction coordinate. Notation is the same as in Figure 3

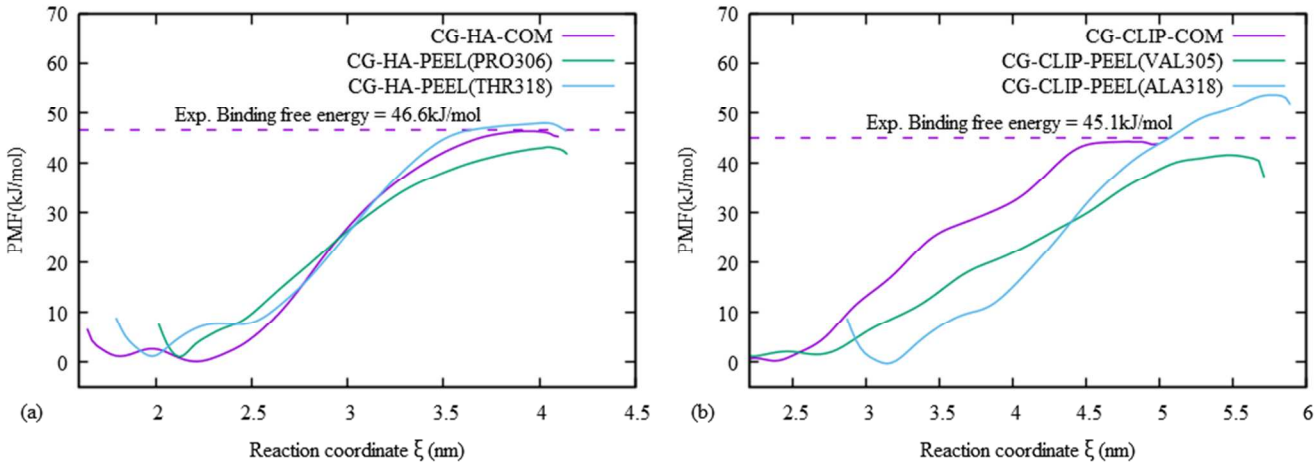


Figure 7: Figure (a) shows the PMF curves from CG COM and PEEL simulations of the HA peptide from the MHCII molecule using the COM, PRO306 end and THR318 end as the pulling group. Figure (b) shows the PMF curves from CG simulations of the CLIP peptide from the COM and both peptide ends. The dashed lines in the both figures stand for the experimental results of binding free energy for HA and CLIP peptides.

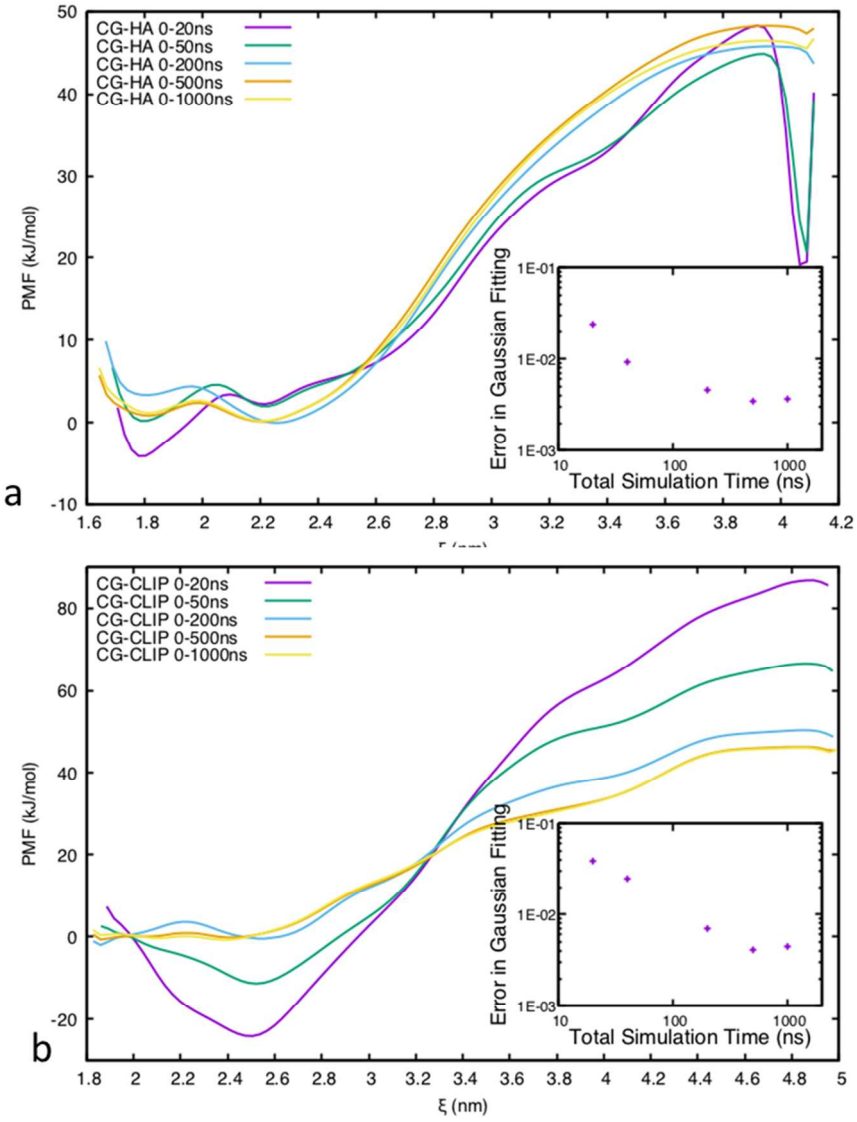


Figure 8: Convergence of the PMF curves obtained from COM CG pulling of (a) HA peptide and (b) CLIP peptide. The PMFs are obtained using data from first 20, 50, 200, 500, and 1000ns of the simulation respectively. The insets show the total error obtained by fitting Gaussian distributions to the histograms used to construct the corresponding PMF.

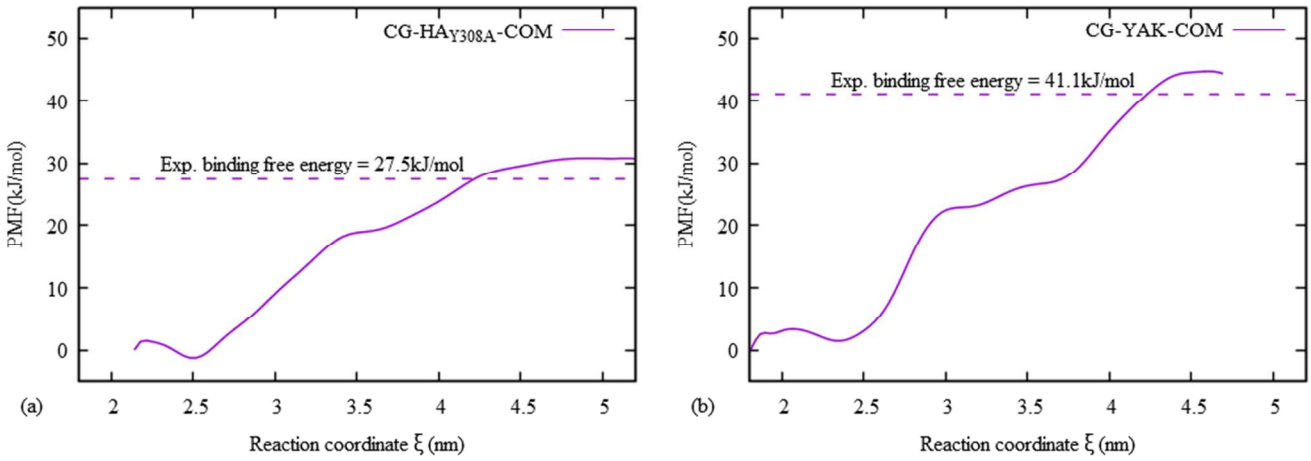


Figure 9: PMF curves for pulling a) HA_{Y308A} and b) YAK along COM reaction coordinate

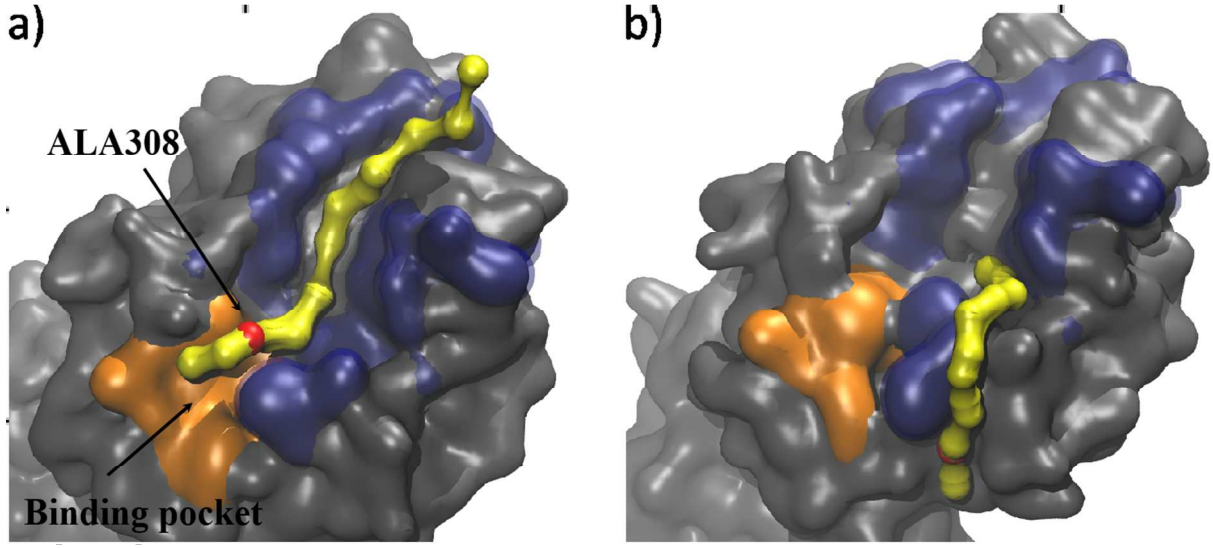
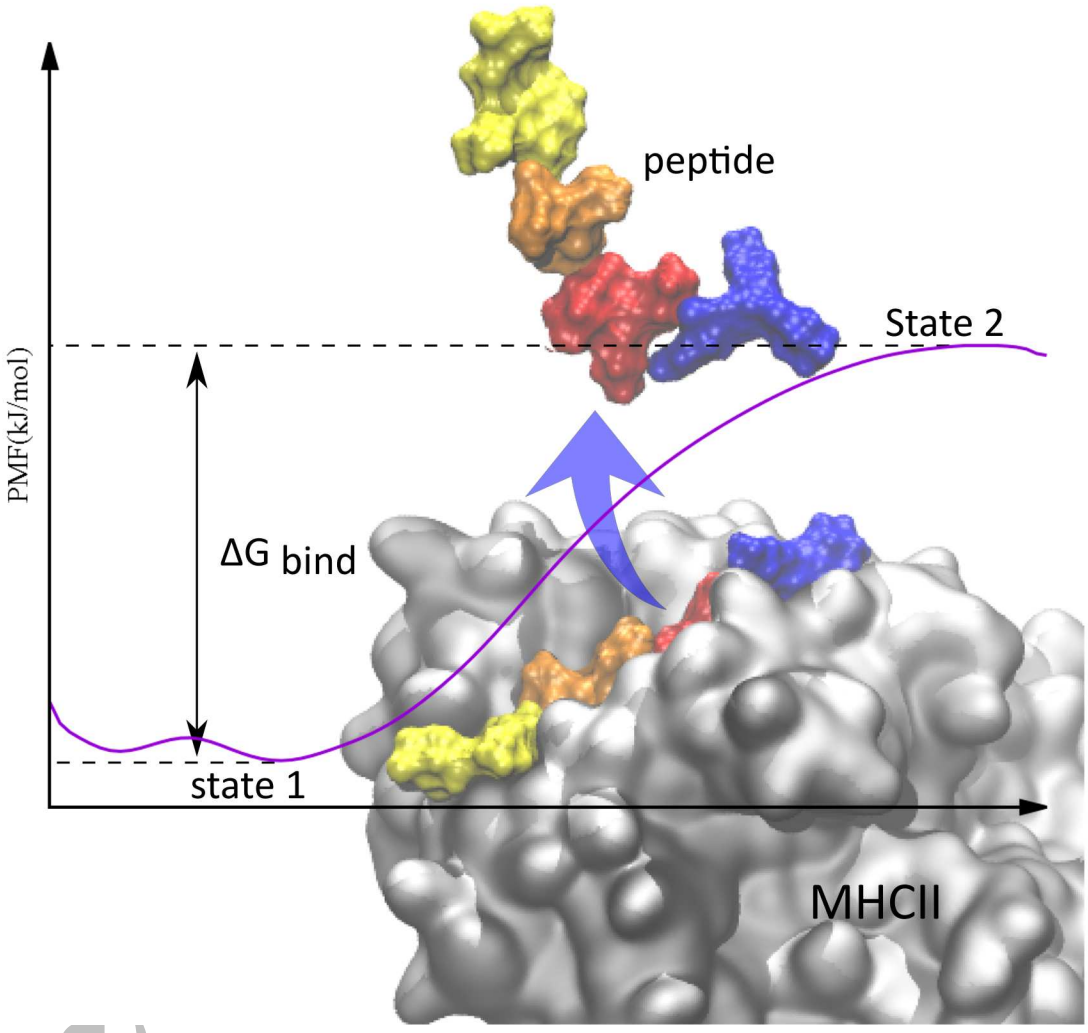


Figure 10: Figures show 2 snapshots in the window simulation at 2.7nm along the COM reaction coordinate of GGA-$G_{309-318}$ pulling simulation. Figure a) is the starting structure of the window simulation. Peptide is colored yellow, ALA308 red, the binding pocket for ALA309 (P1) orange, and the rest of the binding groove dark blue. Figure b) is from a later time in the same window.



Knowing the binding free energy between bound peptide and Major Histocompatibility Complex Class II (MHCII) is important to design effective vaccines and therapeutics. In this paper, we use Molecular Dynamics simulation with umbrella sampling method and coarse-grained force field to estimate the binding free energy. We also propose a segmented peptide model, which can robustly estimate the binding free energy between bound peptide and MHCII efficiently.

Supporting Information

Efficient Estimation of Binding Free Energies between Peptides and an MHC Class II Molecule Using Coarse-Grained Molecular Dynamics Simulations with a Weighted Histogram Analysis Method

Ming Huang, Wenjun Huang, Fei Wen, and Ronald G. Larson*

Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136

S.1 Details of Umbrella Sampling Simulation

We employ two criteria to assess the adequacy of our umbrella sampling simulations. One criterion is that the windows are sufficiently overlapped. The starting structures for window simulations are selected according to the COM distance between the pulling group and MHCII, where the distance between neighboring windows is chosen to be 0.1-0.2nm. We regard the overlap between neighboring windows to be sufficient if the overlap is greater than approximately 1/3 of the area of any of these two windows (as shown in Figure S1a); otherwise (as shown in Figure S1b) an additional window needs to be inserted. The second criterion is that the "set position" of a window at which the applied harmonic potential is zero should be close to that of the peak of the resulting histogram from the window simulation. If window's peak position is within 0.02nm of the set position, we regard the set position of the window to be unbiased. As shown in Figure S2, the set position of the window considered is 3.20nm. With a spring constant K=500 kJ/(mol*nm), the peak position of the simulated window is 3.16nm, which is 0.04nm away from the set position. We deem this window to be biased, and by increasing the spring constant, the simulated peak position moves towards the set position. However, increasing the spring constant decreases the range of the window, which can lead to insufficient window overlapping. If this is the case, a new window should be inserted. A table containing all set positions, simulation positions and spring constants of all our window simulations can be accessed from the University of Michigan Library Deep Blue Data Depository[1].

[1] DOI: 10.7302/Z2M906KK

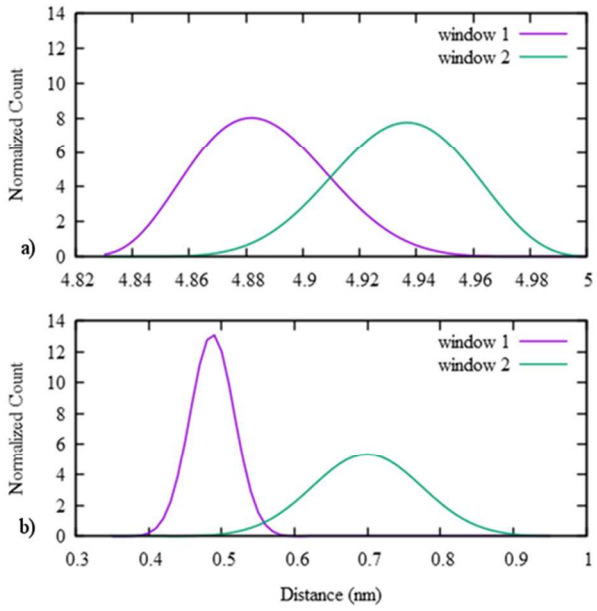


Figure S1. Histograms generated from two neighboring window simulations along two reaction coordinates a) COM distance between THR318 and MHC which is the reaction coordinate used to compute the biasing potential and b) COM distance between LYS310 and the binding pocket. The COM of the binding pocket is taken as the mean COM of residues GLN α9 and TYR β78. Window 1 (purple) has peak positions of 4.89 and 0.55 nm along the two coordinates, and Window 2 (green) has peak positions of 4.94 and 0.70 nm along these two coordinates.

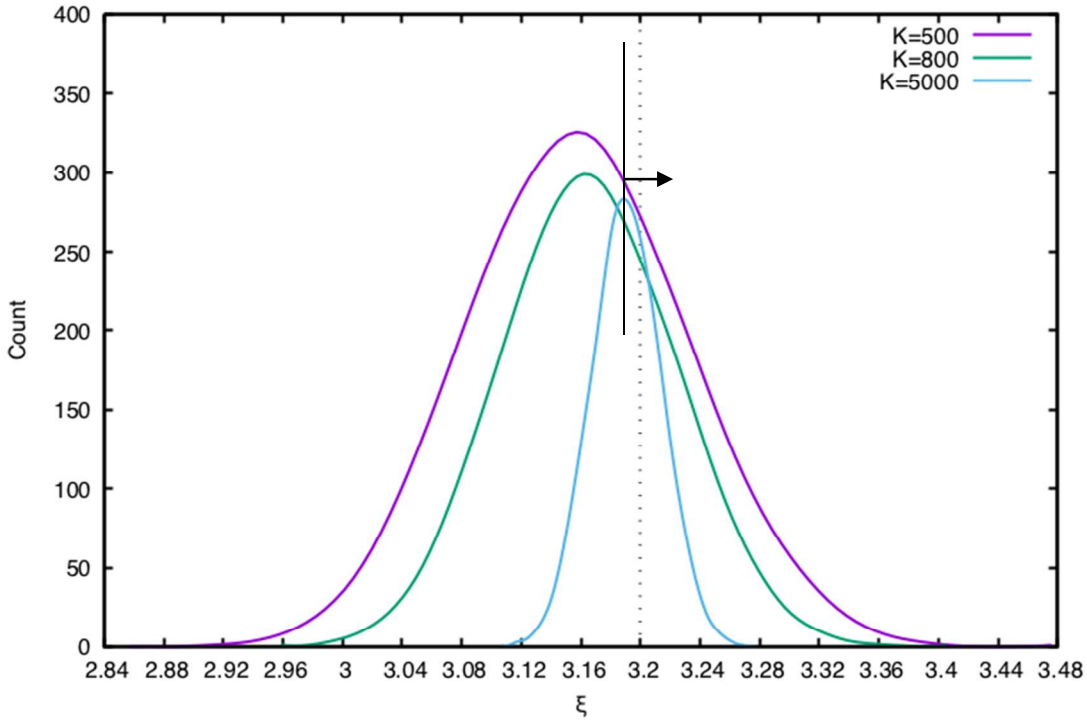


Figure S2. Histograms for the same window but different spring constants (K), from 500 (kJ/mol*nm) to 5000 (kJ/mol*nm). The dotted line is the set position, while the solid line shows the peak position of the simulation histogram with K = 500 (kJ/mol*nm).

S.2 Binding free energy of designed peptide sequence

We introduce seven peptide sequences, HA_{Q311A}, GGY-$G_{309\text{-}318}$, GGA-$G_{309\text{-}318}$, GGG-$G_{309\text{-}318}$, AAF-$A_{309\text{-}318}$, AAW-$A_{309\text{-}318}$, and AAA-$A_{309\text{-}318}$. We conducted whole peptide pulling on the first five of these (Figure S3 and S4). For GGA-$G_{309\text{-}318}$ and the GGY-$G_{309\text{-}318}$, we hypothesize that the binding free energy change due to the single residue mutation is independent of the remaining residues in the sequence, and we expect the difference in the simulated binding free energies between these two peptides (i.e. $\Delta\Delta G_{Y\to A} = \Delta G_{GGY\_sim} - \Delta G_{GGA\_sim}$) to be in good agreement with the binding free energy difference between HA and HA_{Y308A} peptides (i.e. $\Delta\Delta G_{Y\to A} = \Delta G_{HA\_exp} - \Delta G_{HA_{Y308A}\_exp}$). However, the simulated result ($\Delta\Delta G_{Y\to A} = -4.4\mathrm{kJ/mol}$) deviates significantly from the experimental result ($\Delta\Delta G_{Y\to A} = 19.1\mathrm{kJ/mol}$). We find biased samples in a number of the GGA-$G_{309\text{-}318}$ peptide's windows.

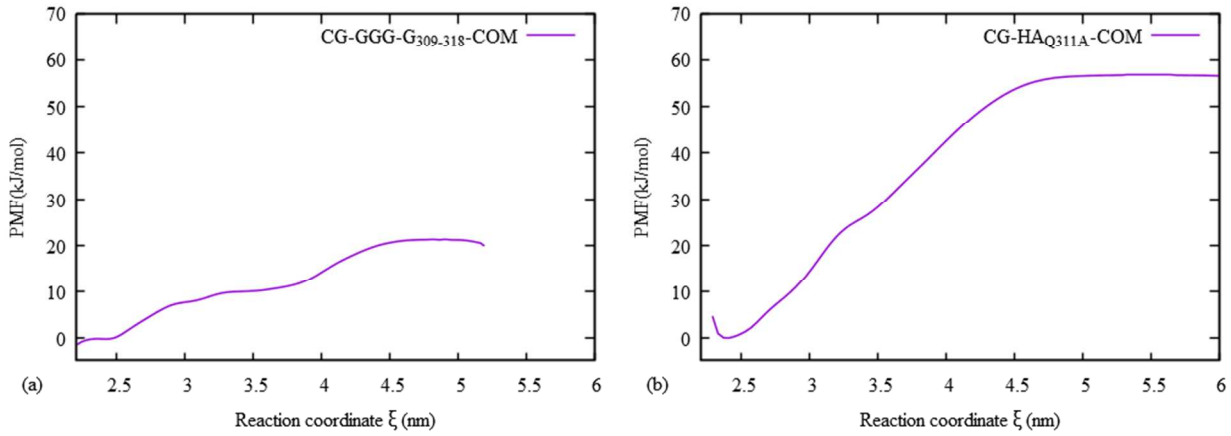


Figure S3: PMF for pulling whole peptides a) GGG-$G_{309\text{-}318}$ and b) HA_{Q311A} along COM reaction coordinate

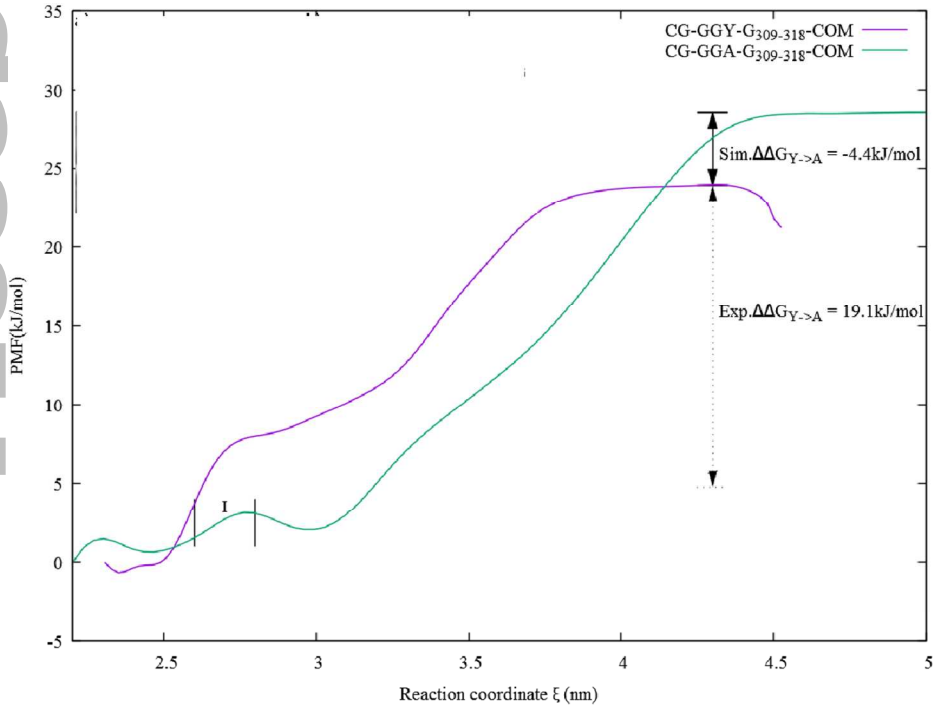


Figure S4: PMFs for pulling GGA-$G_{309-318}$ and GGY-$G_{309-318}$ peptides along COM reaction coordinate.

Table S1 shows the binding free energy of first segment for the peptide sequences GGY-$G_{309-318}$ (GGY|GGGGGGGGGG), GGA-$G_{309-318}$ (GGA|GGGGGGGGGG), GGG-$G_{309-318}$ (GGG|GGGGGGGGGG), AAA-$A_{309-318}$ (AAA|AAAAAAAAAA), AAF-$A_{309-318}$ (AAF|AAAAAAAAAA), AAW-$A_{309-318}$ (AAW|AAAAAAAAAA), and the second segment of peptide sequence HA_{Q311A} (PKY|VKQ|NTLKLAT). As expected, the segment GGA has a weak binding free energy of 2.0 ± 0.2 kJ/mol and the segment GGY has a strong binding free energy of 20.7 ± 0.9 kJ/mol. The mutation from TYR308 to ALA308 thus resulted in a difference of 18.7 ± 1.1 kJ/mol in binding free energy, agreeing well with the experimental result of 19.1 kJ/mol.

Table S1: Simulated CG binding free energy for the first segment of peptide sequences GGY-$G_{309-318}$, GGA-$G_{309-318}$, GGG-$G_{309-318}$, AAA-$A_{309-318}$, AAF-$A_{309-318}$, AAW-$A_{309-318}$, and the second segment (VKA) for peptide HA_{Q311A}

Segment Sequence	Binding Free Energy (kJ/mol)
GGY	20.7 ± 0.9
GGA	2.0 ± 0.2
GGG	2.2 ± 0.8
AAA	5.9 ± 0.6
AAF	12.3 ± 0.9
AAW	19.7 ± 0.8
VKA	14.9 ± 1.4

S.3 Inadequate WHAM window sampling in atomistic simulations

Here we demonstrate a case of inadequate sampling within a WHAM window when using the COM distance between residues on the peptide and residues in the binding pocket as a reaction coordinate (ξ') instead of the COM or PEEL reaction coordinates (ξ). The improper sampling occurs when residue LYS310 remains in close proximity to a few residues, namely GLN α 9 and TYR β 78 in binding pocket 2, throughout one particular window due to the favorable interaction between charged residues. Hereafter we refer to the COM of all beads in GLN α 9 and TYR β 78 as the COM of binding pocket 2. First, note that the COM distance between LYS310 and the COM of the binding pocket remains around 0.65nm until LYS310 is pulled out of the binding groove (Figure S5a). We next select two neighboring simulation windows, "Window 21" and "Window 22", along the reaction coordinate (ξ) that we used to construct the PMF diagram (Peeling from THR318). These two windows should sample the microstates that are just before and just after LYS310 leaving the binding groove, and have equilibrium distances of 0.65nm and 0.70nm between the LYS310 COM and binding pocket COM. We note that a proper reaction coordinate assumes that the molecule is able to sample all relevant microstates contributing to the free energy along the coordinate; only then will WHAM give the correct PMF. We show that this is not the case for these two windows by plotting the same histogram data used to construct the PMF along a different coordinate, (ξ ') (Figure S5a inset). Judging from the histogram, LYS310 is stuck to microstates at a COM distance of around 0.5nm in Window 21, instead of 0.6nm. Moreover, the overlap between these two histograms is insufficient, which accounts for roughly only 10% of the total microstates sampled. This inadequate sampling in the window simulation eventually led to a large error in the PMF curve. As a reference, the histograms obtained from Window 21 and Window 22 along the reaction coordinate (ξ) used to construct the PMF show reasonable overlap (Solid lines in Figure S5c), thus giving no sign of inadequate sampling; using reaction coordinate ξ ', however, shows the poor sampling as shown in the inset to Fig. S5a. These results show the danger in using WHAM to compute binding free energies;

the method requires that all degrees of freedom be adequately sampled in each window, and that all degrees of freedom have overlap with neighboring windows. For large flexible molecules, such as 13-residue peptides, well-overlapping windows of a single reaction coordinate can be misleading; other degrees of freedom may have poor sampling, and this poor sampling may go undetected unless great care is taken. We attempt two approaches to improve the sampling 1) extend the window simulations to 20ns and 2) restart the window simulations with a very similar starting structure for 20ns. In Fig S5b and Fig S5c, we plot the histograms obtained from these two approaches, in addition to the histograms obtained from the original 8ns simulations, along the reaction coordinates ξ' and ξ respectively. Although additional microstates are sampled when using a new starting structure (Fig S5b dotted curve), neither of these two 20ns simulation changes the histograms obtained along the reaction coordinate ξ significantly (Fig S5c).

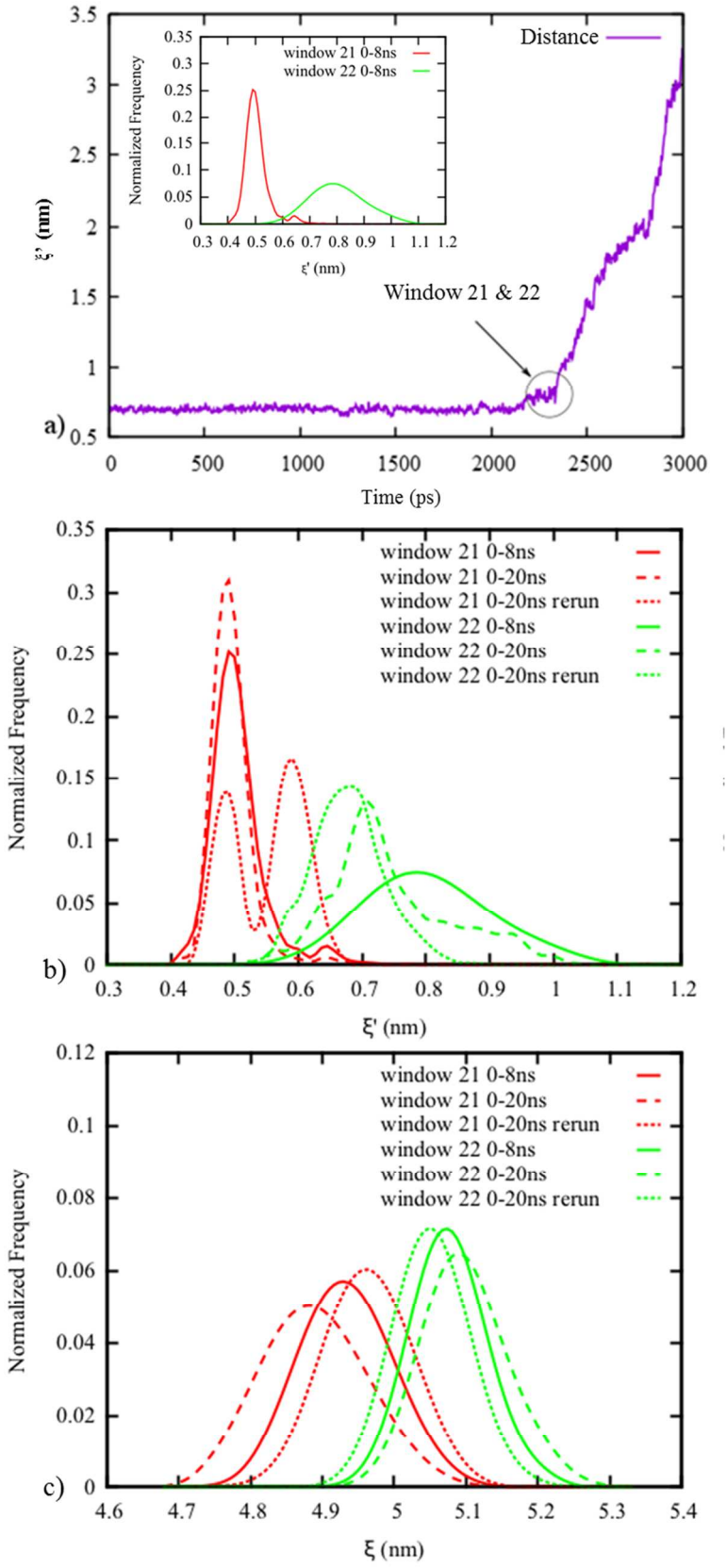


Figure S5. We define a new reaction coordinate ξ' as the distance between LYS310 COM and the binding pocket COM (defined as the COM of the atoms contained in the two binding pocket residues GLN α9 and TYR β78), in addition to the reaction coordinate ξ used to construct the PMF (distance between THR318 COM and MHC COM). a) The time dependent COM distance between LYS310 COM and the binding pocket COM. "Window 21" and "Window 22" are selected from the circled region. These two windows sample the microstates that are just before and just after LYS310 leaving the binding groove. The inset shows the histograms from these two windows, calculated along the reaction coordinate ξ'. Figure b) and c) are three sets of histograms calculated along reaction coordinates ξ' and ξ respectively. The solid lines are calculated from the original 8ns simulations. The dashed lines and the dotted lines represent runs that are extended to 20ns long and runs that are started from slightly different initial structures respectively and run for 20ns, respectively.

S.4 Potential of Mean Force (PMF) curve for CLIP-MHCII using 3QXA crystal structure

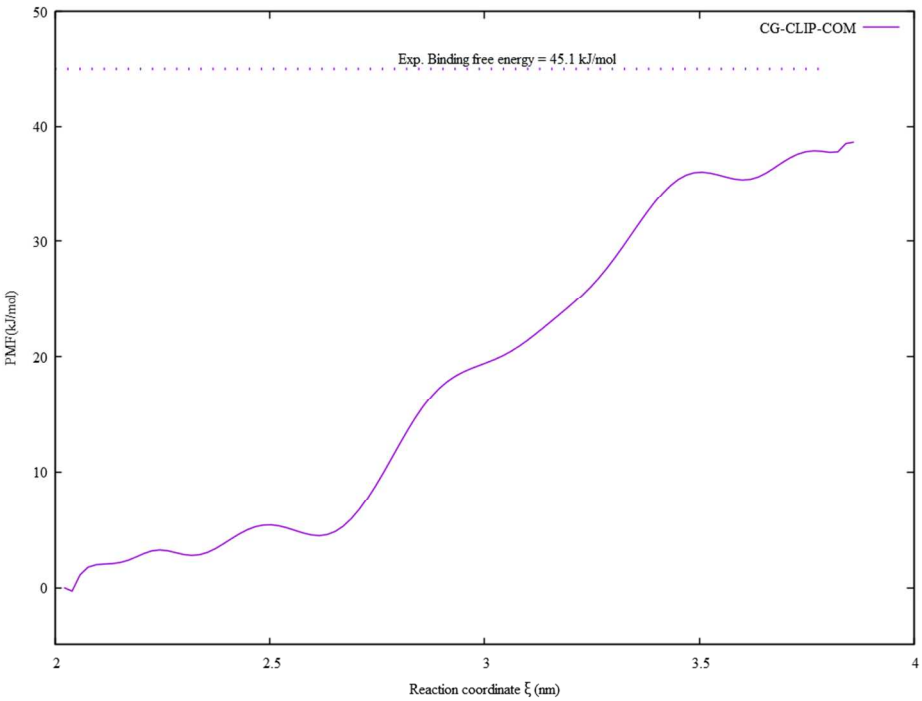


Figure S6: The PMF curve for pulling CLIP peptide along COM reaction coordinate.

S.5 Potential of Mean Force (PMF) curves for segmented peptide

The PMF curves for pulling peptide segments out of their corresponding binding pockets, using coarse-grained (CG) force fields (Figure S7) and atomistic (Figure S8), respectively are shown.

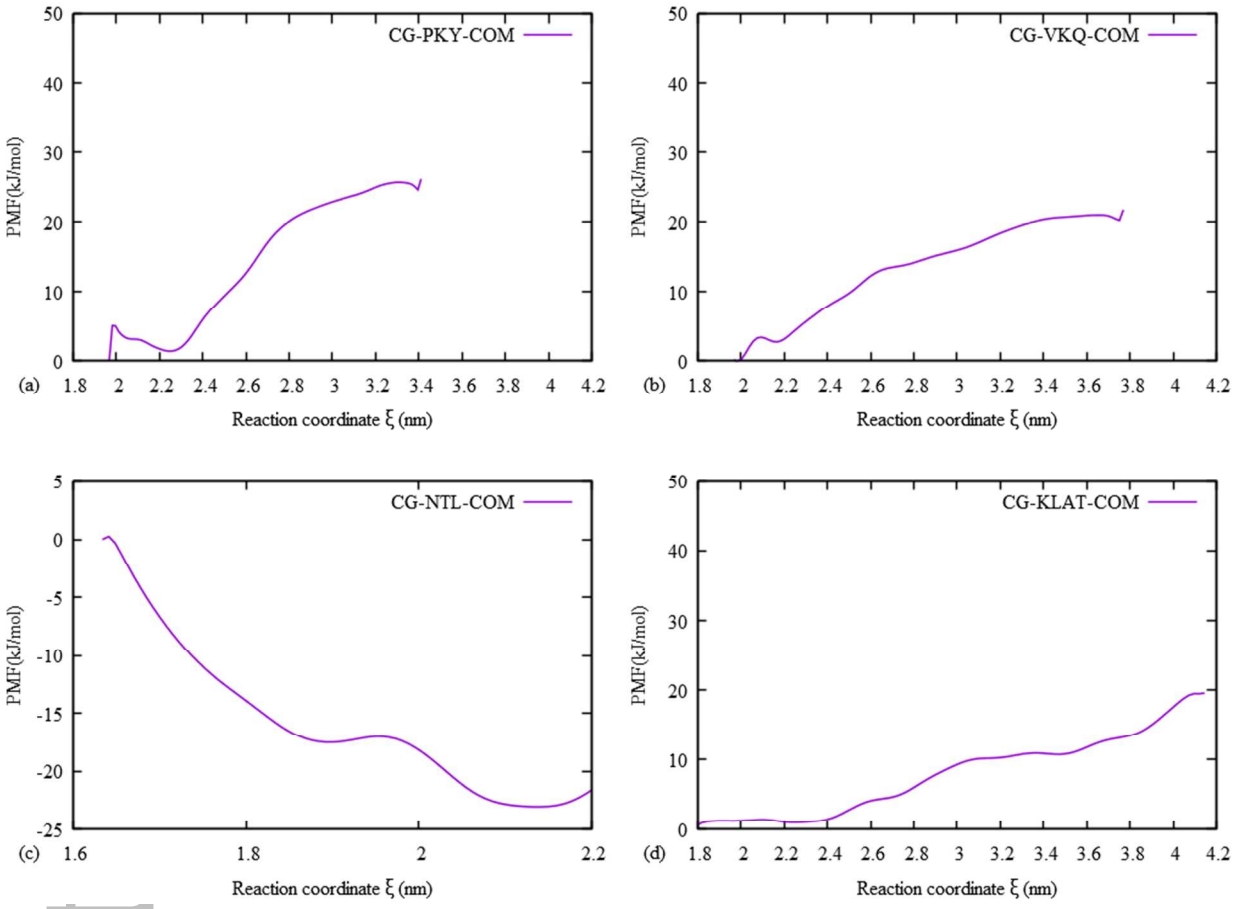


Figure S7. PMF diagram for pulling segments of HA peptide using CG simulations.

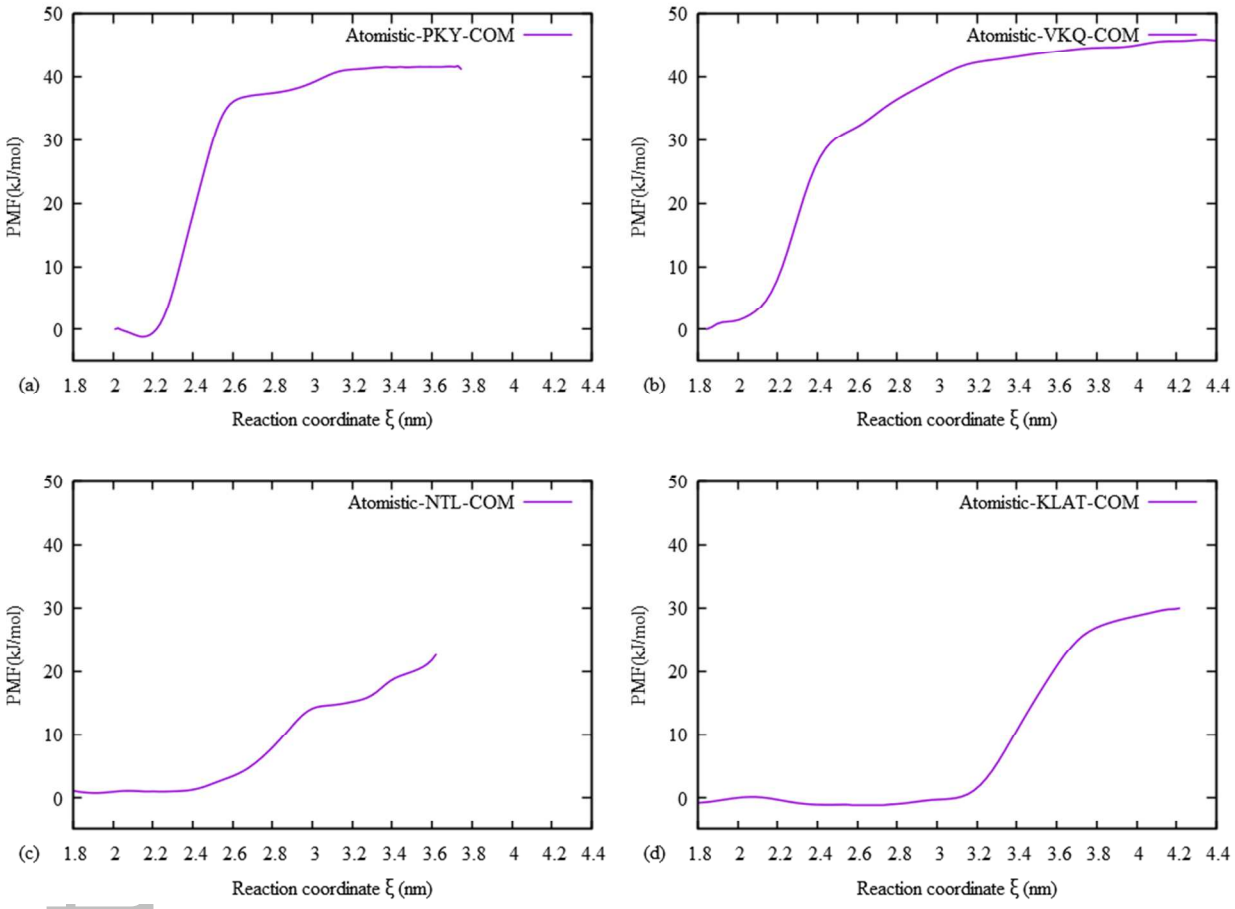


Figure S8. PMF diagram for pulling segments of HA peptide using atomistic simulations.