Parameterization of Unnatural Amino Acids with Azido and Alkynyl R-groups for Use in Molecular Simulations

Addison K. Smith* and Thomas A. Knotts IV*

Department of Chemical Engineering at Brigham Young University

E-mail: addison.smith@byu.edu; thomas.knotts@byu.edu

Supporting Information

The literature is divided on the proper parameterization of linear-angle containing dihedrals (LACD). Some works include LACD parameters while others, including the GAFF and CGenFF databases, exclude LACD parameters. To more fully understand both methods, A 4-bodied model, depicted in Figure S1, is used to explain the cause of ramping in potential energy for LACDs. The purpose of this model is to focus attention on the angle and dihedral by eliminating translational movement of the system and only allowing significant movements for one atom (A) in the 4-bodied system. This was done by increasing the relative mass of sites B, C and D to be magnitudes greater than A and by fixing B, C and D to their initial positions using a strong harmonic spring potential. As a result, sites B-D only experience movements ± 0.001 Å in any direction during simulation. The entire model still uses the CHARMM FF and these changes are only used to help visualize the relevant motion. Specifically, using this set-up allowed the $\overline{BC}$ bond to be fixed along the z-axis and the BCD plane to be restricted to the yz-plane.

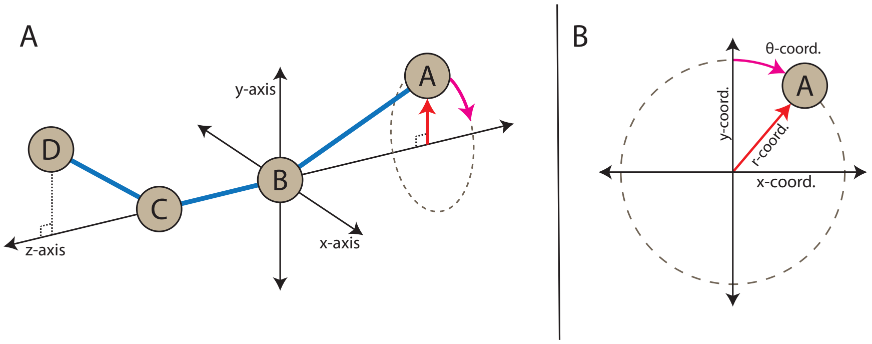

Figure S1: Panel **A** Four-body model used in this work. Sites B, C and D are fixed to their initial positions using a harmonic spring potential. Sites B and C are placed on the z-axis and the BCD plane is on the yz plane. Panel **B** Projection of site A's movement onto the xy plane. Movement in dihedral space is manifested as changes in Θ and movement in angle space results in changes in r.

With this approach, A experiences very limited movement in the z-direction and any changes in z are negligible from a visualization standpoint. Consequently all of A's relevant movement occurs in the xy-plane and traces out a circular path. Said another way, movements in z are allowed, but they are very small compared to changes in x and y. Because only site A remains unconstrained in this model, and the majority of its movement happens solely in angle and dihedral space, this motion can thus be translated to the circular coordinate system shown in Figure S1B. Notice that movements in angle space are manifest as changes in r and movements in dihedral space are manifested as changes in Θ.

Because the motion of B, C, and D is limited by the harmonic spring potential, parameters for the system are only defined for forces that act on A. Specifically, parameters are set for the $\overline{AB}$ bond, $\widehat{ABC}$ angle, and $ABCD$ dihedral. K_b and b_0 parameters were set equal to 200 kcal mol^{-1} Å^{-2} and 1.0 Å, respectively. K_θ was set equal to 120 kcal mol^{-1} rad^{-2}. The equilibrium angle parameter (θ_0) was set to various values between 130° and 180° to demonstrate the difference between linear and nonlinear moieties.

To ensure consistency between all treatments, dihedral parameters are defined to be the same for all simulations where n = 1 and d = 180°. These values give a potential that is easy to analyze as it creates a preference for cis configurations and makes the trans structure unfavorable. K_ϕ is set to a constant value of 0.9 kcal mol^{-1}. This was chosen because it is weak enough to allow full motion around the dihedral but also provides sufficient cis/trans bias for easy comparison.

Simulation Protocols

Multiple molecular dynamics (MD) simulations were performed using the LAMMPS software. Unless the simulation failed due to numeric overload, each system is minimized using 1000 steps of steepest decent and then given sufficient time to equilibrate (typically 10000 timesteps). To observe relevant behavior, the simulation is then run for at least 50000 timesteps.

Inclusion of LACD parameters on System Energy

This set of simulations highlight the pitfalls of simulating LACD parameters. All of these simulations use the NVE ensemble with a 1 fs timestep. Angle values of $\theta_0 = 130°$, $178°$, and $180°$ were investigated with the latter two being considered linear angles.

Figure S2 contains the energetic results of these NVE simulations. Panel A shows the data for $\theta_0 = 130°$, Panel B for $\theta_0 = 178°$, and Panel C for $\theta_0 = 180°$. The potential energy as a function of simulation time is plotted in each case.

Ramping is present for both linear variants in this model ($\theta_0 = 178°$ and $180°$). The simulation experiences multiple dihedral potential spiking events that translate to increases in PE of the system.

Notice that for a “normal” angle of 130° (Panel A), the potential energy fluctuates around an average value and does not experience long-term drift. This is the expected behavior for such systems. However, for equilibrium angles of 178° and 180°, the potential energy spikes at various times and increases dramatically. These spikes occur in the dihedral potential. Changing the magnitude of K_ϕ in an LACD does nothing to eliminate ramping in the system. Only as θ_0 became more non-linear does the number of PE spike events decrease. In other

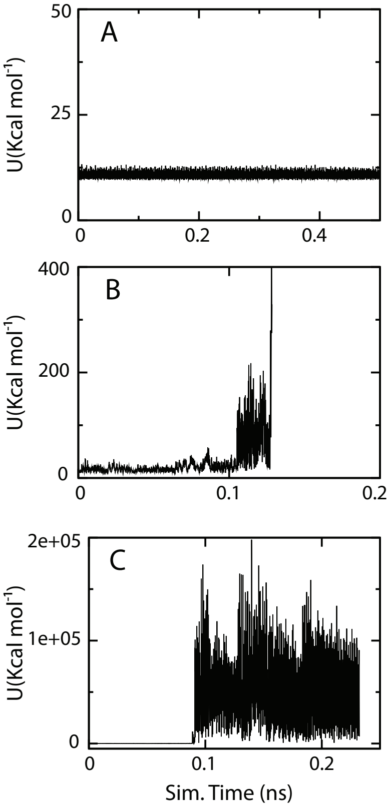

Figure S2: Panel **A** NVE simulation using a θ_0 of 130°. Panel **B** NVE simulation using a θ_0 of 178°. Panel **C** NVE simulation using a θ_0 of 180°.

words, the more linear the θ_0 value, the more frequent destabilizing events occur.

In all literature references mentioned in the main body that include LACD parameters, none of them included simulations using the NVE ensemble. Instead, all use the NVT ensemble with multiple thermostats. We believe the energy analysis that was just demonstrated has been unnoticed in these studies because thermostats or minimization algorithms can mask energetic inconsistencies. Because inconsistencies also indicate discontinuous movements in simulation, LACD parameterization is not considered a viable option for this work.

We also note that the these unstable spiking events could theoretically affect inherently non-linear systems. If some molecule experiences *induced* linear structure by some non-bonded interaction, a spiking event would occur that would skew the conserved energy of the system.

Inclusion vs. Exclusion of LACD on Structure

By excluding the LACD parameter, the dihedral influence is eliminated from linear moieties. The consequences of this assumption is analyzed in this section. Also analyzed in this section is establishing some metric for identifying linearity because there currently does not exist a definition for when some structure is considered “linear.”

To accomplish this, two simulation sets are compared. The first includes LACD parameters in the NVT ensemble with 10 thermostats set to 300 K. The excessive number of thermostats is to ensure ramping effects are as minimal as possible and represents studies that include LACD parameters. The second simulation set excludes LACD parameters and are run in the NVE ensemble. This set represents all studies that exclude LACD parameters. By contrasting the results of both methods, the compromises each approach makes is made apparent.

The metrics used to compare these approaches are angle-space linearity and dihedral-space influence. Linearity was tracked during simulation to produce a histogram of angles. Dihedral influecne was tracked using the xy positions of site A. The position of A was periodically saved during simulation to create a 2D histogram of its position. This histogram thus becomes a heatmap plot identifying the most favorable regions of phase space for site A. Positions on the histogram that are more frequently visited provide molecular structure analysis. Due to placing Site D in a positive y position and the $\overline{BC}$ bond on the z-axis (see Figure S1), positive y values for the position of A indicate cis dihedral preference while negative y values correspond to trans configurations.

Inclusion of LACD parameters on Structure

Figure S3 shows structures formed from temperature-controlled simulations with LACDs and allows for observation of molecular structures despite significant coupling being present. Different treatments focus on changing values of θ_0 to equal values between 130° and 180°.

Two plots are shown for each θ_0 value tested: a histogram of angles sampled (top graph for each angle) and an xy heatmap plot of the positions of site A from the model (bottom plot). The model was created so that movement of site A may be viewed as motion on a circle in the x-y plane. The top set of simulations shows results for cases where no ramping events occur and, if using NVE, produced stable trajectories. The bottom set shows results from simulations that experience ramping and were numerically unstable in an NVE ensemble.

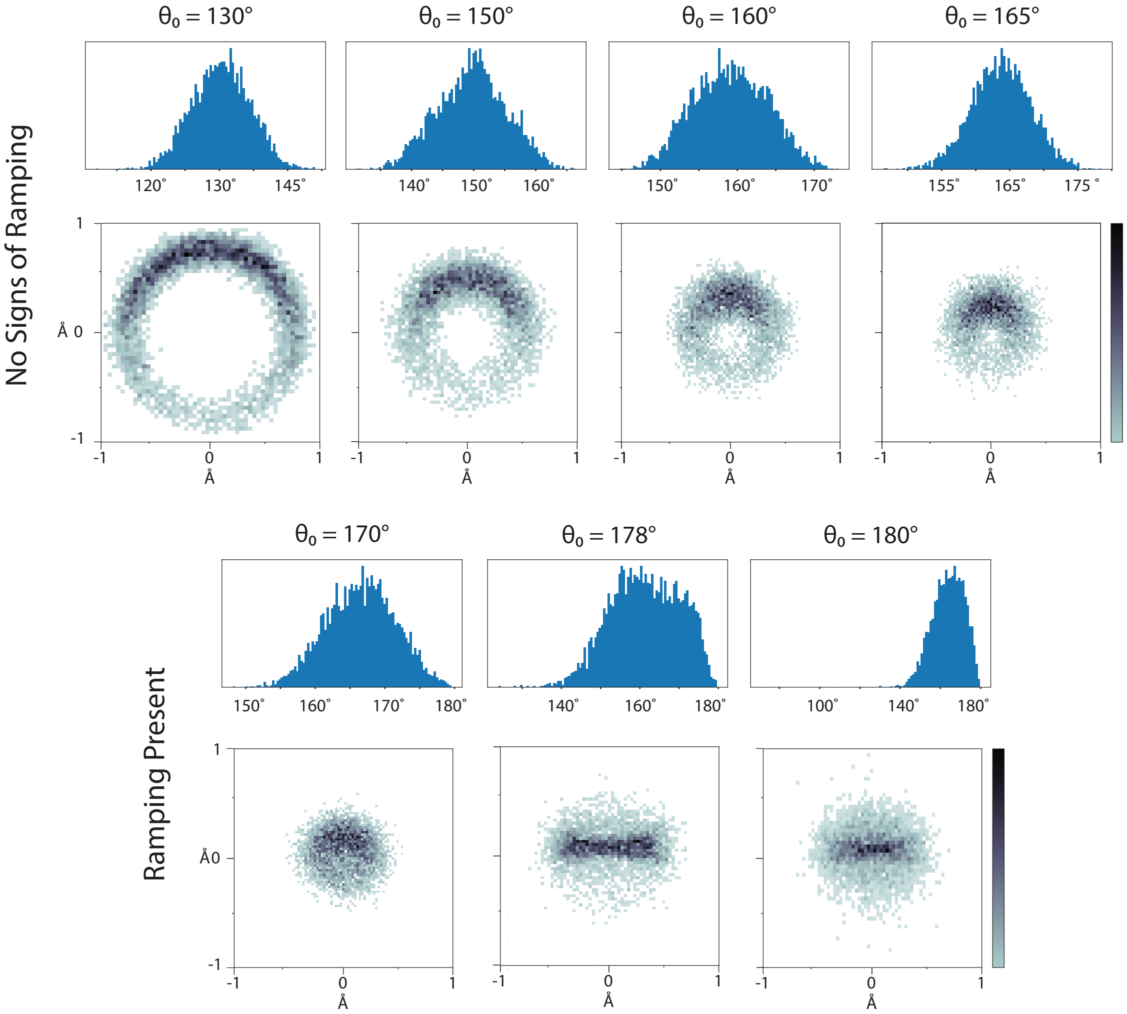

Figure S3: NVT plots of the angle distribution and 2D histogram of A's xy position in space using LACD parameters.

Analysis on the top row of simulations show no signs of ramping. Notice that for all simulations where $\theta_0 < 170°$, angle histogram plots produce a Gaussian distribution of values centered on the θ_0 value and the heatmap plots show molecular structures are limited to a circular band or ring. Also notice that the heatmaps indicate the cis state is sampled more often during these simulations as expected from the model parameters. The combination of a Gaussian distribution of θ values centered on θ_0 with the correct cis conformations on the heat plots shows independence of the angle and dihedral contributions to the force field in this region.

In contrast, all simulations that exhibited numerical instability in the NVE ensemble (bottom row from Figure S3) lose the previously described angle space ring structure. Instead, the heatmaps show the most-frequently-visited structures as a well of high probability in the middle of the circle. Ramping events in simulations first occur when $\theta_0 \approx 170°$. For generalization, we define this threshold as a transition region which indicates the onset of numeric ramping, the collapse of the angle-space ring, and the presence of spikes in system energy. Another important feature of this transition region is that the dihedral potential begins to have less impact on the structure. Specifically, the system should sample only a few trans conformations. However, the heatmap plots show frequent visits to trans conformation. Thus, in the transition region, ramping causes trans structures to be disproportionately sampled.

Beyond the transition region ($\theta_0 \approx 180$), ramping dominates the behavior of the system and many thermostats are required to maintain numeric stability. This reduces the influence of the dihedral potential. Moreover, at the most linear θ_0 values, angle sampling no longer centers on θ_0, but skews towards one side. This observation is important because it emphasizes the fact that linear angles do not adhere to the Hooke's law assumption required for angle potential calculation and parameterization in most FFs.

Exclusion of LACD parameters on Structure

This next set of simulation assumes LACD parameters are inconsequential to gain NVE stability. Results are again in the form of histograms of angle values and heatmaps of the xy positions of site A and are again presented for simulations at various values of θ_0.

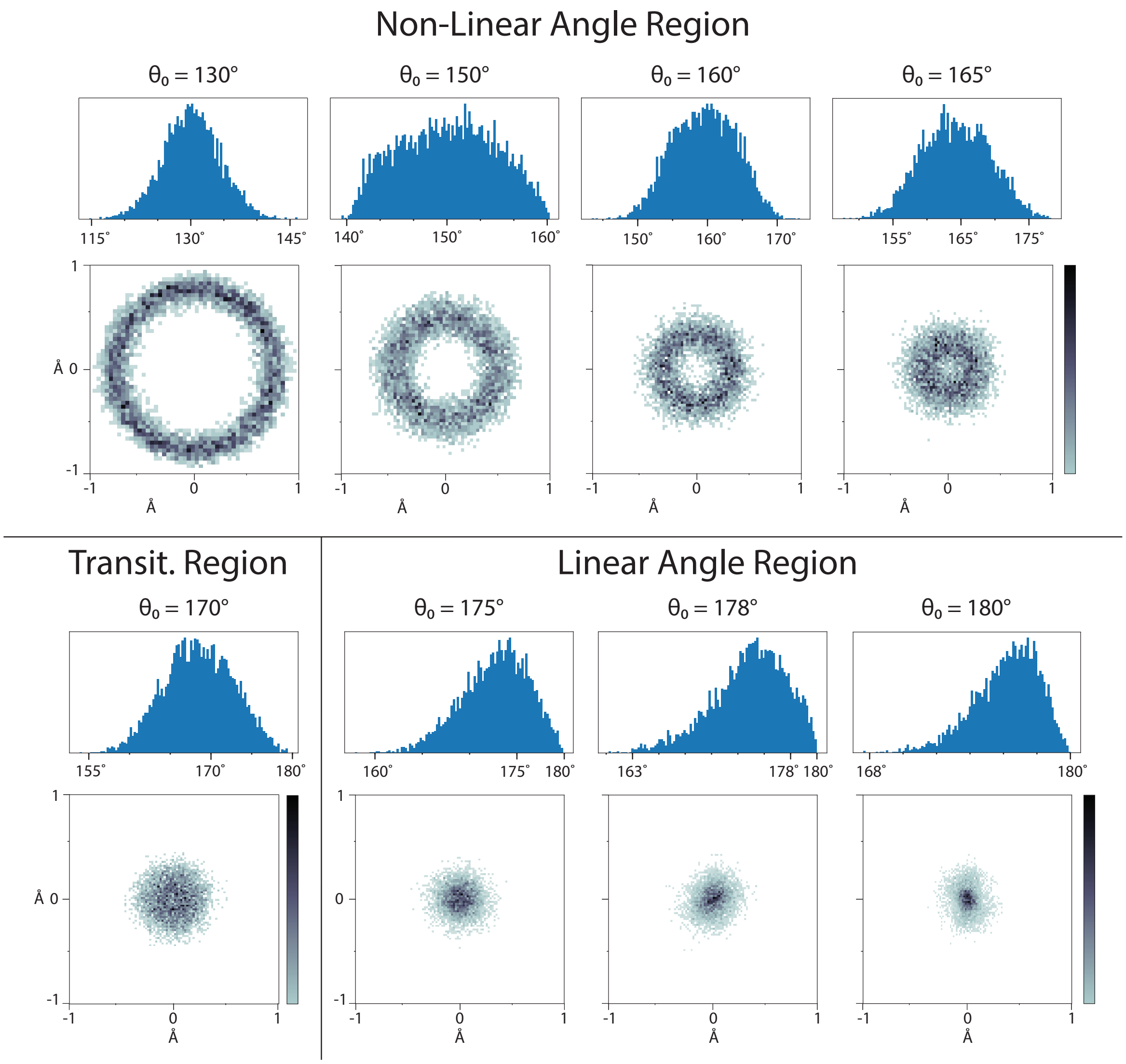

Figure S4: Plots of the angle distribution and 2D histogram of A's xy position in space without LACD parameters.

Figure S4 shows these results and is split into different sections. The top row are the results that are done with a non-linear equilibrium angle. The other two sections would present signs of numeric ramping if the LACD parameters were included.

For all simulations in Figure S4, the angle space histograms appear relatively unchanged from those shown in Figure S3 when LACDs were included. This is important because it reveals that the GAFF and CGenFF assumption about LACDs contains inherent flaws. Because angle space does not adhere to Hooke's law, these widely-used databases are producing angle structure that do not replicate actual behavior and in the linear regime the fundamental equation used to model this space is flawed.

The dihedral space heatmap plots in Figure S4 show that both cis and trans structures are equally probable. This is undesirable because dihedral influence is important for proper molecular structure (non-linear structures in particular). As θ_0 increases in linearity, the angle space prediction band decreases in diameter until the histogram of angles begins to sample values at or very near 180°. When $\theta_0 = 170°$, the probability band dissolves into the Transition region to produce an angle/dihedral space prediction well. The onset of this Transition region is important because it (1) identifies the limit for stable simulations that include dihedral parameters, (2) the θ_0 parameter where significant coupling and unstable NVE events first occur, (3) the point when angle space does not adhere to the Hooke's law assumption and (4) when K_ϕ should be set equal to 0. The unfortunate characteristic about the transistion region is that the θ_0 values that defines its onset appears to be a function of temperature, K_θ and, theoretically, nonbonded interactions. This means there is no way to predict the θ_0 onset values prior to simulation.

The simulations that do occur in the Transition and Linear Angle regions produce structures that are rational for a linear system. The heatmap structures are held to linear-looking angles, but the nuance that presents in the QM structural optimization is lost. Moreover, because angle space is not held to Hooke's law, there is no way of confirming K_θ values are appropriate for the system. Thus, dihedral influence on structure is compromised for a stable LACD-containing system in the NVE ensemble that appears to produce reasonable structures.

Discussion

In the absence of better force fields, there must be a choice between NVE stability and dihedral influence. The most concerning observation from this Supplementary Material is the break in the Hooke's law assumption for *all* simulations using a classical model like the CHARMM FF. This shows a need for a new LACD force field that holds to all assumptions in its construction. For this work, excluding LACD parameters was chosen because it is the desire of the authors that the parameters derived in the main body be used in all ensembles. NVE is important for solvation and void removal and the additional energy added to the system, if included, would skew subsequent simulations or analysis more than the minor structural accuracy gained with LACD inclusion.