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

Addison K. Smith,* Joshua W. Wilkerson,* and Thomas A. Knotts IV*

Department of Chemical Engineering at Brigham Young University

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

Abstract

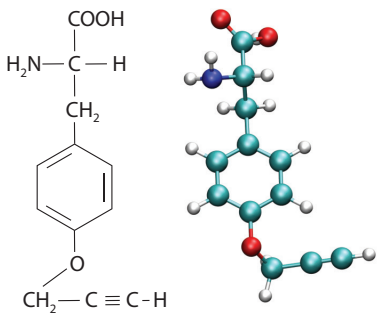
Recent new methods to functionalize proteins at specific amino acid locations use unnatural amino acids that contain azido and alkynyl groups. This capability is unprecedented and enables the creation of site-specific protein devices. Due to the high specificity of these devices, many protein configurations are possible and in silico screens have shown promise in predicting optimal attachment site locations. Therefore, there is significant interest in improving current molecular dynamic models to include the unique chemistries of these linear moieties. This work uses the force field tool kit (ffTK) to obtain the bonded and nonbonded CHARMM parameters for small molecules that contain azido and alkynyl groups. Next, the reliability of these parameters is tested by running simulated MD analysis to prove modeled structures match those found in literature and quantum theory. Finally, protein MD simulation compares this parameter set with crystallographic data to give a greater understanding of unnatural amino acid influence on protein structure.

Introduction

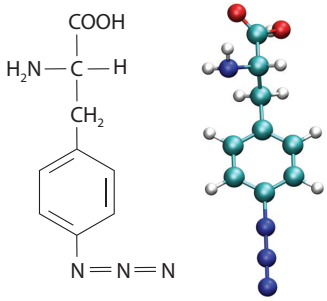
Protein devices are proteins that have undergone biotechnological modification to harness the function of protein molecules for specific applications. These modified proteins introduce control into protein-based systems and have great potential to change how we interact with biology.[1–3] There is significant interest in making protein devices in a site-specific way to reduce cost, optimize function, and maintain device uniformity.[4–7] The state-of-the-art method for site-specific protein device creation is the Protein Residue-Explicit Covalent Immobilization for Stability Enhancement (PRECISE) technique. This method utilizes genetic recoding to mutate unnatural amino acids (uAA) into the primary sequence of a protein.[8,9] Figure 1 shows uAAs commonly used. p-azido phenylalanine (pAz) and p-propargloxy-phenylalanine (pPa) contain terminal azido or alkynyl functional group essential for the 1,3 dipolar cycloaddition "click" reaction.[10,11] Because the chemical moieties needed for this reaction do not occur naturally in proteins, this mutation provides a biologically unique location to functionalize the protein. "Click" chemistry provides high reaction specificity, reaction efficiency, and biologically inert products.[12,13]

The PRECISE method enables site-specific protein device creation, but currently no heuristic exists for determining, a priori, optimal uAA mutation sites. This is problematic because many mutation sites can lower both protein stability and function, which negates the entire purpose of functionalization. Moreover, experimental trial-and-error is tedious and expensive in time and money.[14,15] Previous studies have shown that molecular simulation of surface-accessible sites can be a good *qualitative* predictor of protein device behavior in unnatural environments.[15–18] However, the effects of the uAA (e.g. pAz and pPa) mutation are not considered in these previous efforts due to lack of model parameters for the uAA residues.

Azides and alkynes, in some form, have been simulated before in the COMPASS, CHARMM and AMBER force fields. The general AMBER force field (GAFF) has parameters for azido and alkynyl small molecules attached to aliphatic groups.[19] The CHARMM general force



p-propargyloxy phenylalanine



p-azido-L-phenylalanine

Figure 1: Unnatural amino acids commonly used in the PRECISE technique for protein device creation.

field (CGenFF) has force field parameters for alkynyl groups, but it too is only parameterized when attached to an aliphatic group.[20] Additional parameterization of azides and alkynes have been done in the COMPASS and AMBER FF. These studies indicate that the structures linear moieties are attached to affect the parameterization process and questions the transferability of the general forms provided GAFF and CGenFF.[21–23]

Both uAAs used in the PRECISE method involve azido and alkynyl groups attached to an aromatic ring and therefore requires additional parameterization so that aromatic affects are properly simulated. The purpose of this paper is to report on the the creation of CHARMM-compatible model parameters for the uAA residues used in this protein mutation and functionalization technique. These new parameters expand the realm of protein simulation beyond its current borders and offer researchers a new tool to investigate new and cutting edge technologies involving azides, alkynes, and uAAs in a manner not previously

possible.

Method

General Approach

The general procedure used to develop the new force field parameters for azides and alkynes utilizes a quantum mechanical (QM) basis at all steps of the process. Specifically, the steps in the procedures are:: (1) developing CHARMM parameters for four small molecules that contain terminal azido and alkynyl groups using the *force field Tool Kit* (ffTK),[24] (2) addressing linear structure issues seen during ffTK analysis by modifying the procedure so that all CHARMM parameters can be resolved, (3) validating the resulting force field via in silico molecular dynamic (MD) simulation of the four small molecules, and (4) testing the transferability of the model by simulating a molecule not used in the paramerization but for which experimental data are available. The last step uses CGenFF along with the new parameters to obtain a model for pAz, and then using this uAA residue with the standard CHARMM force field to simulate the Trastuzumab Fab (Tra-Fab) fragment with pAz mutation (PDB ID: 5XHF).[25] 5XHF is the only molecule deposited in the Protein Data Bank which contains one of the uAAs in questions, but because it is not used in the training of the model, it serves as a rigorous test of the new parameters.

Model

The CHARMM force field[26] defines the potential energy of a molecular system ($V_{molecule}$) as a summation of bonded (V_{bonded}) and non-bonded ($V_{nonbonded}$) terms:

$$V_{molecule} = V_{bonded} + V_{nonbonded} \tag{1}$$

Bonded terms include contributions from bonds, angles, dihedrals, improper dihedrals, and

Urey-Bradley (UB) interactions according to:

$$V_{bonded} = \sum_{l=Bonds} K_{b,l}(b_l - b_{0,l})^2 + \sum_{m=Angles} K_{\theta,m}(\theta_m - \theta_{0,m})^2 + \sum_{n=Dihedrals} K_{\phi,n}(1 + cos(\mathrm{n}_n\phi_n - \mathrm{d}_n)) + \sum_{u=Impropers} K_{\omega,u}(\omega_u - \omega_{0,u})^2 + \sum_{v=UB} K_{S,v}(S_v - S_{0,v})^2 \quad (2)$$

where all bond pairs (l), angle pairs (m), improper dihedrals (u) and UB 1,3-interactions (v) use a form of Hooke's law. The variables b_l, θ_m, ω_u, and S_v are bond lengths, bond angles, improper torsion angles and UB 1,3-distances, respectively. Parameters for each term include equilibrium values: $b_{0,l}$, $\theta_{0,m}$, $\omega_{0,u}$, and $S_{0,v}$; and their respective force constants: $K_{b,l}$, $K_{\theta,m}$, $K_{\omega,u}$ and $K_{S,v}$.[27,28] For all dihedral pairs (n), the dihedral potential is defined by the dihedral angles (ϕ_n) in relation to their sinusoidal multiplicity (n) and phase shift (d) scaled to their equilibrium potential using a force constant ($K_{\phi,n}$).[27] All intermolecular pair parameters (l, m, n, u, v) must be defined for all combinations within a molecule or protein and are considered unique unless otherwise justified.

Nonbonded terms ($V_{nonbonded}$) are composed of contributions for coulombic and Lennard-Jones interactions according to:

$$V_{nonbonded} = \sum_{charge} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{Lennard-Jones} \epsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \quad (3)$$

Here, q_i and q_j are the partial atomic charges for each atom in the charge pair, and r_{ij} is the

distance between sites i and j, and ϵ_0 is the permitivity of free space. Also, the energy (ϵ_{ij}) and length ($R_{min,ij}$) parameters in the Lennard Jones interaction are obtained through the Lorentz Berthelot combining rules for the atoms involved in each i,j pair.

Small Molecules for Parameterization

The ffTK approach to parameterizations uses small molecules containing the chemical groups in question, rather than larger molecules like entire protein residues, to focus on the relevant biophysics and reduce the effects of atoms that already have parameters. The small molecules used in this study are: propyne (PY), 3-phenoxy-1-propyne (POPY), methyl azide (MAZ) and phenyl azide (PAZ). Figure 2 contains a structural representation of each small molecule and also defines the atom naming structure used in this work. PY and MAZ molecules were chosen because they best represent a terminal linear angle moiety connected to non-aromatic carbons and have not previously been parameterized for CHARMM/CGenFF. The structures of the other two molecules introduce aromaticity and most closely match the uAA chemistry needed to model pAz and pPa. Initial structures for these molecules were built in PyMOL[29] and used the MMFF94 structural optimization algorithm for initial structure approximation.[30]

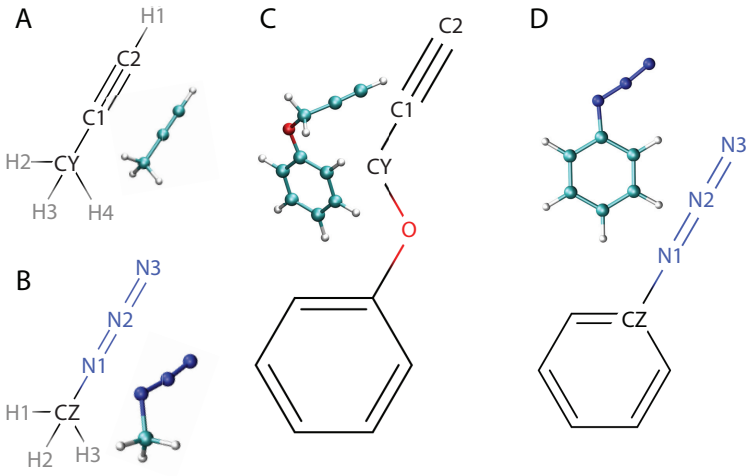


Figure 2: The molecules parameterized in this work: **A** propyne (PY); **B** methyl azide (MAZ); **C** 3-Phenoxy-1-propyne (POPY); **D** phenyl azide (PAZ)

Because no Urey-Bradly nor improper angles exist in linear azide and alkyne structure, new parameters are not needed for these contributions. Atoms two or more sites removed from the

linear angle moiety are assumed to be unaffected by the linear moiety and parameters for these atoms are assumed unchanged from those currently found CGenFF. The rest are considered new atom types and parameters are needed for any $V_{molecule}$ term that includes one of these new atoms. Specifically, all bond, angle, dihedral, charge, and nonbonded term parameters are determined for all atoms in PY and MAZ. For POPY, parameters are determined for all cases that include non-aromatic atoms. Lastly, PAZ parameters are determined for all linear moiety atoms. This includes the phenylic C, but the remaining aromatic atoms are unchanged from CGenFF.

Parameter Determination: ffTK

The ffTK method was chosen for parameterization because it has been successfully used to obtain CHARMM-compatible parameters for small molecules of similar size and complexity.[24,31,32] Each ffTK step is briefly outlined in this section, and the reader is referred to the literature[24] for the details of the method. Due to the challenge of modeling linear chemical moieties, the standard ffTK procedure required slight alteration, and these modifications are explained in Steps 3 and 4.

ffTK Step 1: Geometry Optimization

The first step in the ffTK method is to optimize the geometry of the small molecules using quantum mechanics. This was done using the Gaussian09 software package[33] at the ffTK recommended MP2/6-31G* level of theory and basis set. Due to the relative simplicity of the small molecules studied, single-reference correlated wave functions like those specified in the MP2 theory are sufficient to capture relevant behavior, so higher levels of theory were not needed. The structures shown in Figure 2 are those obtained from this geometry optimization approach.

ffTK Step 2: Nonbonded Optimization

Charge and nonbonded Lenard-Jones (LJ) parameterization follow the water-interaction method as proposed by CHARMM.[34] For all aliphatic and aromatic hydrogens, the standard CHARMM charges of +0.09 and +0.015, respectively, were used to maintain consistency with CGenFF.[20] For every other atom in the molecule, donor, acceptor, or non-interacting status was applied and then a TIP3P water molecule was appropriately oriented to optimize hydrogen bonding with the

target atom and minimize steric interactions with the surrounding atoms. Two QM optimization steps were performed using the HF/6-31G(d) level of theory to determine the two remaining free parameters: (1) the distance between the interaction site and the water molecule and (2) the rotation angle of the water molecule to the incident target atom. NAMD, using the CHARMM force field, was then used to simulate the system to determine the molecular mechanic (MM) nonbonded interaction parameters. Additional parameterization iterations are executed until there is sufficient agreement between MM and QM simulation.[24]

ffTK Step 3: Bond and Angle Optimization

To determine bond and angle parameters, ffTK compares the Gaussian09 QM Hessian matrix to a bond and angle potential energy distribution (PED) surface generated from the NAMD Hessian MM calculations. The standard ffTK procedure computes the Hessian entirely in redundant internal coordinates (IC) as opposed to Cartesian or normal mode coordinates because ICs isolate PED distortions in the CHARMM set of parameters. The Hessian in some other coordinate system may contain molecular geometries that have normal modes with multiple contributing force-field coordinates. This runs the risk of parameter coupling where a single distortion may affect a combination of bonds and angles and make convergence challenging.[24]

Unfortunately, for the molecules in this study, Gaussian09 would not run Hessian calculations in IC because of the large gradients involved in systems that contain angles close to 0°or 180°. This numerical instability was overcome by first obtaining the QM Hessian in Cartesian coordinates and then transforming the results into ICs for comparison to the MM scan. Particular care was taken during optimization within the ffTK program to ensure convergence was achieved and effects of parameter coupling were avoided.

ffTK Step 4: Dihedral Optimization

Dihedral parameterization requires a QM dihedral potential energy scan (PES) for each dihedral of interest. For this work, the QM PES used the MP2/6-31G* level of theory and basis set. As none of the dihedrals parameterized were in ring structures, each QM PES was scanned bidirectionally +/-180° in 10° increments. ffTK improves upon previous best fit parameterization methods by

including coupling influences from each dihedral. Instead of parameterizing dihedrals one by one, all QM PESs are compared to MM PESs simultaneously. An additional annealing protocol further improves the model.[24]

As previously described, the parameters of aromatic C, CH, and atoms more than two bonds away from new chemical moieties, were set equal to their values in CGenFF. This leaves 18 dihedrals that require parameterization. Of these 18, 7 are unique to linear molecular moieties. Within the structure of these 7 unique dihedrals, three of the four atoms form a linear angle with $\theta_0 \approx 180°$. For convenience, such dihedrals are given the abbreviation LACD (linear-angle-containing dihedral).

There is some confusion in the literature with regard to the parameterization of LACD parameters. Works that parameterize azido and alkynyl dihedrals when attached to structures outside those found in GAFF and CGenFF *include* LACD parameters,[21–23] but the GAFF and CGenFF databases themselves *exclude* all LACD parameters.[19,20] The reason for the LACD exclusion is because these databases avoid computations in dihedral space where large fluctuations could occur,[35,36] but the compromise is a loss of dihedral contribution that has been shown to cause unrealistic simulation results.[37] Additional simulation and analysis regarding the inclusion or exclusion of LACD parameters was performed to identify what method should be employed for this work and is presented in the Supplemental Material. This supplemental work shows LACD parameters should not be parameterized as their inclusion causes unstable NVE simulations. The reader is referred to the supplementary information for a more detailed discussion on this topic.

Validation of Model Parameters for Azide and Alkyne Moieties

To validate that the resulting MM model reproduces the QM results and available experimental data, multiple MD simulations are performed on two different solvated systems: the small molecules shown in Figure 2 and the Tra-Fab fragment shown in Figure 3 (PDB ID: 5XHF).

The Tra-Fab protein is a tetramer that has 434 residues per Fab dimer and contains an uAA mutation at site 155. In Figure 3 the pAz uAA mutations can be observed using an all-atom ball and stick representation on the secondary structure (circled in red). Protein structure was obtained by X-ray crystallography.[25] The Class, Architecture, Topology Homologous (CATH) structural classification for TEM-1 is a Immunoglobulin-like, Mainly Beta Sandwich for all chains in the

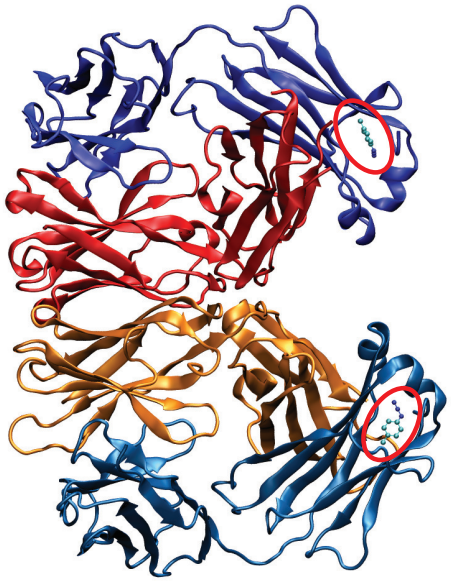


Figure 3: Dimer of Trastuzumab Fab fragments with pAz mutations at residue 155. The pAz uAAs can be observed in the blue regions of the Fab fragments and are circled in red.

protein.[38] To simulate this system, parameters are needed for the pAz uAA. The pAz model was determined using the same procedure as defined in ParamChem[39,40] but with its scope of reference molecules expanded to include the azide small molecules parameterized in this paper. This protein system was not used in the parameterization of the uAA or small molecules in any way, so provides a rigorous validative test case for the ability of the new model to capture the correct biophysics of pAz/uAA simulations.

All validation simulations were performed in the LAMMPS simulation software.[41] Protocol for all MD simulations used during validation are as follows: First, the solvated systems are equilibrated using the NVE ensemble to remove voids in the model. Next NVT simulations using 10 Nose-Hoover thermostats ramp the temperature to 350 K. It is held at this temperature for 0.1 ns to remove any unphysical molecular configurations that may be present at lower temperatures before cooling to 300K. The small molecule or protein is then allowed to come to temperature and the system is allowed to fully equilibrate at 300 K for 0.5 ns. NPT simulations are then done at 1 atm and 300 K for 0.5 ns to obtain the correct box length/density for the system under normal conditions. This is followed by NVT simulations at 300 K with the box size set equal to the average size found in the previous NPT simulation step. This NVT simulation step consists of at least 3 ns of equilibration followed by 30 ns of production time for small molecule analysis or 18.78 ns of production time for

large protein analysis. These production-phase simulations are then compared to the QM ab initio structures, in the case of the small molecules, or crystal structure, in the case of Tra-Fab, using RMSD analysis. RMSD analysis is the standard metric used to validate that new model parameter produce correct structures[22,32,42]

Results and Discussion

Optimized Parameters

All parameters, with the exception of LACDs, were determined in accordance with ffTK procedure as defined in the Methods section and without issue. As mentioned above, the LACD's were studied extensively (see Supplementary Information), and are discussed in brief below. Table 1 contains the list of CHARMM-compatible atom types along with partial charges and non-bonded parameters according to the atom nomenclature shown in Figure 2. The names for each atom type are chosen to be distinct from other types found in the CHARMM36 protein force field and CGenFF. The table is arranged in four sections with the non-bonded parameters and partial charges in the first section, the bonds in the second section, the angles in the third section, and the dihedrals in the fourth section. Each section is divided into atoms according to the the small molecule (i.e. PY, POPY, MAZ, or PAZ) in which it is found.

Table 1: Optimized CHARMM parameters for all relevant atoms in the linear moiety. Dihedral CHARMM parameters in this table exclude LACDs.

Atoms	Atom Type	Charge	ϵ	R_{min}
PY				
C1	CY1	0.040	-0.1670	1.840
C2	CY2	-0.469	-0.1032	1.9925
H1	HY	0.298	-0.0090	0.875
CY	CYA3	-0.139	-0.0320	2.000

Continued on next page

Table 1 – *Continued from previous page*

H2-4	HYA3*	0.090	-0.0240	1.340
POPY				
C1	CY1O	0.236	-0.1670	1.840
C2	CY2O	-0.544	-0.1032	1.9925
H1	HY	0.328	-0.0090	0.875
O	OG301	-0.334	-0.0320	2.000
CY	CYA2	-0.014	-0.0560	2.010
H2-3	HYA2*	0.090	-0.0350	1.340
MAZ				
CZ	CZA	0.076	-0.078	2.050
N1	NZ1	-0.327	-0.200	1.850
N2	NZ2	0.000	-0.200	1.850
N3	NZ3	-0.019	-0.180	1.790
H1-3	HZA*	0.090	-0.024	1.340
PAZ				
CZ	CZRA	0.314	-0.070	1.9924
N1	NZ1C6	-0.301	-0.200	1.850
N2	NZ2C6	0.000	-0.200	1.850
N3	NZ3C6	-0.013	-0.180	1.850

Bond Parameters	K_b ($kcal\ mol^{-1}$ Å^{-2})	b_0(Å)
PY		
CYA3 CY1	370.73	1.466

Continued on next page

Table 1 – *Continued from previous page*

CY1 CY2	999.30	1.221
CY2 HY	423.09	1.065
CYA3 HYA3	353.90	1.094
POPY		
CYA2 CY1O	356.33	1.466
CY1O CY2O	983.93	1.219
CY2O HY	413.00	1.069
CYA2 HYA2	346.44	1.095
OG301 CYA2	339.08	1.425
MAZ		
NZ1 CZA	321.76	1.475
NZ1 NZ2	717.05	1.245
NZ2 NZ3	999.99	1.163
HZA CZA	360.02	1.093
PAZ		
NZ1C6 CZRA	305.85	1.430
NZ1C6 NZ2C6	700.40	1.254
NZ2C6 NZ3C6	953.00	1.167
CG2R61 CZRA	303.30	1.383
Angle Parameters	K_θ $(kcal\ mol^{-1}\ rad^{-2})$	$\theta_0(^\circ)$
PY		
CY1 CY2 HY	5.088	179.994

Continued on next page

Table 1 – *Continued from previous page*

CYA3 CY1 CY2	13.082	179.873
HYA3 CYA3 CY1	49.740	111.335
HYA3 CYA3 HYA3	42.389	108.118
POPY		
CY1O CY2O HY	52.805	179.378
CYA2 CY1O CY2O	24.883	178.329
CY1O CYA2 HYA2	60.984	110.853
CY10 CYA2 OG301	85.136	112.648
OG301 CYA2 HYA2	46.580	107.263
CG2R61 OG301 CYA2	116.593	117.125
HYA2 CYA2 HYA2	76.304	108.307
MAZ		
NZ1 NZ2 NZ3	46.144	177.340
CZA NZ1 NZ2	92.630	128.104
HZA CZA NZ1	3.336	108.045
HZA CZA HZA	52.543	109.385
PAZ		
NZ1C6 NZ2C6 NZ3C6	107.69	179.577
CZRA NZ1C6 NZ2C6	132.17	134.897
CG2R61 CZRA NZ1C6	94.00	120.000
CG2R61 CG2R61 CZRA**	40.00	120.000
CG2R61 CZRA CG2R61**	40.00	120.000
HGR61 CG2R61 CZRA**	30.00	120.000

Continued on next page

Table 1 – *Continued from previous page*

Dihedral Parameters	K_ϕ ($kcal\ mol^{-1}$)	n	d
PY			
CYA CY1 CY2 HY	0.00	1	0
HYA CYA CY1 CY2	0.00	1	0
POPY			
CYA2 CY1O CY2O HY	0.00	1	0
HYA2 CYA2 CY1O CY2O	0.00	1	0
OG301 CYA2 CY1O CY2O	0.00	1	0
CY1O CYA2 OG301 CG2R61	0.730	1	180
CY1O CYA2 OG301 CG2R61	0.826	2	0
CY1O CYA2 OG301 CG2R61	1.337	3	180
HYA2 CYA2 OG301 CG2R61	0.058	1	180
HYA2 CYA2 OG301 CG2R61	0.651	2	0
HYA2 CYA2 OG301 CG2R61	0.981	3	0
CYA2 OG301 CG2R61 CG2R61	1.549	1	180
CYA2 OG301 CG2R61 CG2R61	1.139	2	180
CYA2 OG301 CG2R61 CG2R61	0.213	3	180
MAZ			
NZ3 NZ2 NZ1 CZA	0.00	1	0
NZ2 NZ1 CZA HZA	0.364	1	180
NZ2 NZ1 CZA HZA	0.180	3	0
PAZ			
NZ3C6 NZ2C6 NZ1C6 CZRA	0.00	1	0
NZ2C6 NZ1C6 CZRA CG261	3.537	2	180

Continued on next page

Table 1 – *Continued from previous page*

NZ2C6 NZ1C6 CZRA CG261	0.249	3	0
NZ1C6 CZRA CG2R61 HGR61	2.997	2	180
NZ1C6 CZRA CG2R61 CG2R61	2.885	2	180
CG2R61 CG2R61 CG2R61 CZRA**	3.100	2	180
CG2R61 CG2R61 CZRA CG2R61**	3.100	2	180
HGR61 CG2R61 CG2R61 CZRA**	4.200	2	180
HGR61 CG2R61 CZRA CG2R61**	4.200	2	180

*Aliphatic and aromatic hydrogen charges fixed to CHARMM standard. **These parameters were assumed the same as those in CGenFF and were not explicitly parameterized in ffTK.

Alkyne charges agree well with previous ab initio studies that show the CH group is more negatively charged than the CH_3 group.[43] Selecting the correct partial charges for the atoms in the azide is more complicated as azides can have up to four possible charge configurations.[44,45] A 1,3 dipole configuration for MAZ and PAZ is chosen as this is the precursor to click reaction initiation and is most stable according to frontier molecular orbital models.[46,47] All charge and nonbonded parameters are within the optimization standards set in ffTK.

Bond length alignment in all cases did not exceed 0.03 Å for any bonded pair and the angle degree alignment did not exceed 5° for any non-linear angle. These values are congruent with the standards set in CGenFF.[20] All bonds that result from sp^1 hybridization produced very strong K_b values. An upper limit of 999.99 $kcal\ mol^{-1}$ was set to ensure convergence was achieved. Linear moiety angle parameters had errors $\geq 5°$, but these are the consequence of simulation error and not parameterization error (see Supplementary Material).

Dihedral parameters were optimized with good root-mean-square error (RMSE) and were within the tolerances of ffTK. Figure 4 shows the QM and MM PES comparison for all non-LACD dihedrals. Because it was assumed LACD contribution is zero, LACDs were not included during this step. Multiple low-energy configurations were observed in POPY and presented as coupled dihedral

parameters for dihedrals CY O CG2R61 CG2R61 and C1 CY O CG2R61. Using a similar process proposed by Yu et al., multiple iterations of the ffTK method was used to identify proper structure and dihedral potential.[31] All dihedral errors for both the high and low energy states are within the 0.5 kcal/mol standard set by CGenFF.[20]

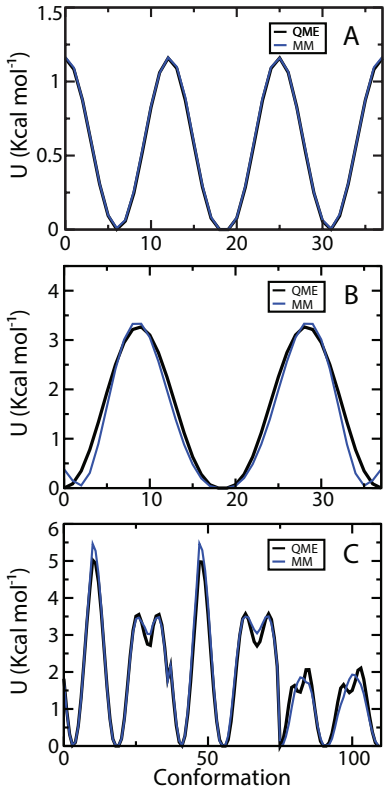


Figure 4: Torsion profiles for all normal dihedrals. **A** The N2 N1 CZ H dihedral in MAZ **B** The N2 N1 CZ CG2R61 dihedral in PAZ **C** The C1 CY CG2R61, H2/3 CY O CG2R61, AND CY O CG2R61 CG2R61 dihedrals in POPY.

In Table 1 all LACDs are assumed to have zero contribution to molecular potential. In the Supplementary Material this assumption was challenged by looking at LACD influence on structure and energy conservation. Previous efforts have obtained LACD parameters by first running a QM PES and then fitting the the dihedral cosine expansion.[21–23] However, when LACD parameters are simulated in the NVE ensemble, energy is not conserved and the simulation will crash from numeric overload. This instability has likely gone unnoticed because thermostats in the NVT ensemble prevent temper system energy. This results in simulations with LACD influence on structure.

Observable in all methods of LACD simulation (including the null contribution) is that the

Hooke's Law assumption does not hold for the angle space potential. Any form of LACD simulation skews the angle distribution so that θ_0 is not the mode angle sampled. All figures and results can be found in Supplementary Material. Despite the challenges parameterizing LACDs, to ensure functionality in any ensemble this work assumed LACD contribution is negligible and LACD parameterization set $K_{\phi,LACD} = 0\ kcal\ mol^{-1}$.[19,20] As a note, LACD parameterization in ffTK was impossible because the MM PES required structural relaxation steps that were energetically unstable in the linear domain and caused by the instabilities just described.

Validation

As explained in the methods section, the validity and transferability of all ffTK model parameters was tested following the standard practice for force field generation. Validation MD simualtions were done on two systems sets. The first set were simulations of the small molecules depicted in Figure 2 at 300 K and 1 atm. The second set was simulations of the Tra-Fab protein with the pAz mutation also at 300 K and 1 atm. The metric for structural analyzes was the root mean squared deviation (RMSD) between structures produced by the simulation and the appropriate standard structure. The standard for each small molecule was its QM ab initio structures, and the standard for the Tra-Fab protein was the crystallographic data.

Small Molecules

Figure 5 shows RMSD results for the small molecules parameterized using ffTK. RMSD calculations are averaged across all production-phase simulation frames to produce an average RMSD value. Figure 5A shows the RMSD result for PY and generates an average RMSD value of 0.13 Å. Figures 5B-C show the MAZ and PAZ molecules generate an average RMSD value of 0.09 and 0.17 Å respectively.

For POPY, two non-LACD dihedral configurations are structurally preferred. So, instead of comparing the entire structure of the molecule to the ab initio structure – which only reveals one of the conformations – only the LACD moiety is compared to POPY's ab initio LACD structure. This ensures deviations in the non-LACD regions within POPY do not skew analysis on the linear moiety structures being examined. Figure 5D shows the RMSD using this criterion and generates

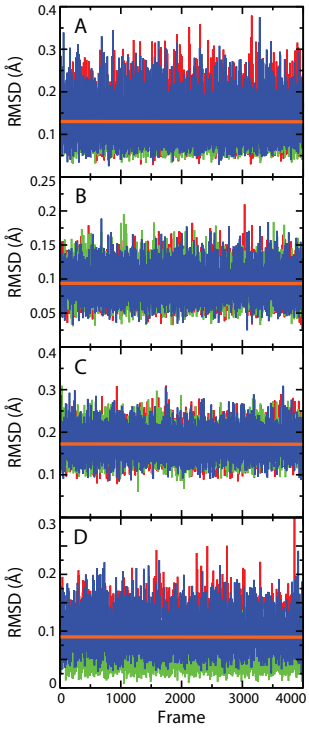


Figure 5: **A** PY RMSD with an average value of 0.13 Å. **B** MAZ RMSD with and average value of 0.09 Å. **C** PAZ RMSD with and average value of 0.17 Å. **D** POPY RMSD of the molecule's LACD moieties with an average value of 0.08 Å. Three replicates are represented as different colors (red, green, blue) with the average RMSD line (orange).

an average RMSD value of 0.08 Å. In all cases, RMSD never exceeds 0.2 Å which indicates that the MM model is in excellent agreement with the QM-predicted structures.

Protein Simulation

To test the transferability of the newly-developed parameters for linear moieties on relevant biological molecules, simulations are done on a dimer of Tra-Fab domains where both domains have mirrored pAz mutations within their primary structure. Figure 3 depicts the uAA mutations (circled in red) located on chains A and C (colored in blue hues). The simulation process is similar to that used for the small molecule validation except the NVT simulations ran for 18.78 ns of production rather than 30 ns.

Previous research on MD uAA incorporation indicate that the parameters are valid if the average backbone RMSD is on the order of 2 Å..[42] Figure 6 shows the backbone RMSD from three, independent validation the simulations. Panel A shows the RMSD calculated for the full protein.

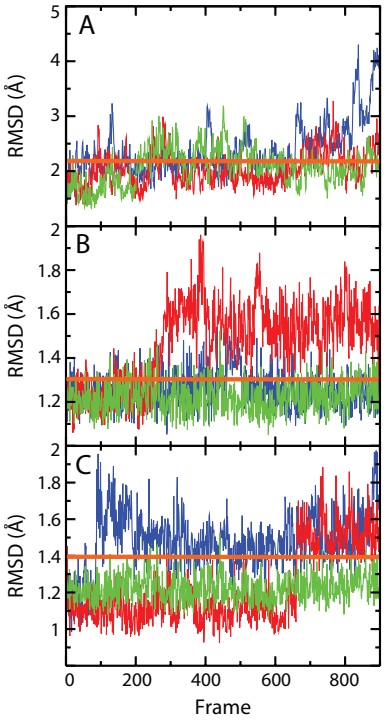


Figure 6: **A** Full protein RMSD with an average RMSD of 2.2 Å. **B** Local structure containing the pAz uAA on chain A showed an average RMSD of 1.3 Å. **C** Local structure containing the pAz uAA on chain C showed an average RMSD of 1.4 Å. Three replicates are represented as different colors (red, green, blue) with the average RMSD line (orange).

Panels B and C the RMSD of the residues close to the uAA (residues 151-159) on each chain the uAA is located. The average of the RMSDs depicted in Panel A is 2.2 Å, that of Panel B is 1.3 Å, and that of Panel C is 1.4 Å.

All of the average RMSD values are consistent with that expected of a typical protein simulation, findings which establish a strong case supporting the validity of our linear moiety parameters. As stated previously, the parameterization done in this work did not use the Tra-Fab crystal structure as an input. Simulation results not only agree with crystal structure for the entire protein, but the regions very near the uAA have better-than-standard agreement. These parameters can thus be trusted to accurately predict how an azido group affects protein structure in molecular simulation..

Conclusion

Recent biotechnological techniques offer unprecedented control over where functionalization of a protein can occur, but these methods use unnatural amino acids to add chemistries not typically seen in biology, such as the linear structures present in azides and alkynes. Molecular simulation of these linear-angle-containing molecules would aid in optimization and utilization of these methods as a means to identify whether a mutation will deleteriously affect the structure of the protein, but model parameters for the relevant unnatural amino acids were not previously available. This paper reported the first CHARMM-compatible parameters, obtained using the ffTK methodology, for azide and alkyne chemical moieties. Validation of all optimized parameters was done by testing the ability of the model to reproduce quantum mechanical structures of small molecules *and* experimental crystallographic data of a Tra-Fab fragment with an unnatural amino acid mutation that contains an azido linear moiety. The results of these validation simulations show that the parameters derived in this work accurately capture the biophysics of the system and reproduce the structural geometry of the linear moieties. Thus, these new chemistries can now be used with the existing CHARMM force field to model molecules that contain azido or alkynyl groups.

Supporting Information

- Analysis examining the energetics and structures of linear molecules when including or excluding LACD parameters in the CHARMM FF

Acknowledgements

The authors are grateful for funding from the National Science Foundation (DMR: 1710574) and for computer resources from the Office of Research Computing at Brigham Young University.

References

(1) Sadik, O.; Land, W.; Wang, J. Targeting chemical and biological warfare agents at the molecular level. *Electroanal.* **2003**, *15*, 1149–1159.

(2) Ellington, A. A.; Kullo, I. J.; Bailey, K. R.; Klee, G. G. Antibody-Based Protein Multiplex Platforms: Technical and Operational Challenges. *Clin. Chem.* **2010**, *56*, 186–193.

(3) Veronese, F.; Pasut, G. PEGylation, successful approach to drug delivery. *Drug Discov. Today* **2005**, *10*, 1451–1458.

(4) Balboni, I.; Chan, S. M.; Kattah, M.; Tenenbaum, J. D.; Butte, A. J.; Utz, P. J. Multiplexed protein array platforms for analysis of autoimmune diseases. *Annu. Rev. Immunol.* **2006**, *24*, 391–418.

(5) Angenendt, P. Progress in protein and antibody microarray technology. *Drug Discov. Today* **2005**, *10*, 503–511.

(6) Pelegri-O'Day, E. M.; Lin, E.-W.; Maynard, H. D. Therapeutic Protein-Polymer Conjugates: Advancing Beyond PEGylation. *J. Am. Chem. Soc.* **2014**, *136*, 14323–14332.

(7) Illanes, A.; Cauerhff, A.; Wilson, L.; Castro, G. R. Recent trends in biocatalysis engineering. *Bioresource Technol.* **2012**, *115*, 48–57.

(8) Smith, M. T.; Wu, J. C.; Varner, C. T.; Bundy, B. C. Enhanced protein stability through minimally invasive, direct, covalent, and site-specific immobilization. *Biotechnology Progress* **2013**, *29*, 247–254.

(9) Wu, J. C. Y.; Hutchings, C. H.; Lindsay, M. J.; Werner, C. J.; Bundy, B. C. Enhanced Enzyme Stability Through Site-Directed Covalent Immobilization. *J. Biotechnol.* **2015**, *193*, 83–90.

(10) Kolb, H.; Finn, M.; Sharpless, K. Click chemistry: Diverse chemical function from a few good reactions. *Angew. Chem. Int. Edit.* **2001**, *40*, 2004–2021.

(11) Jewett, J. C.; Bertozzi, C. R. Cu-free click cycloaddition reactions in chemical biology. *Chem. Soc. Rev.* **2010**, *39*, 1272–1279.

(12) Rusmini, F.; Zhong, Z.; Feijen, J. Protein immobilization strategies for protein biochips. *Biomacromolecules* **2007**, *8*, 1775–1789.

(13) Seo, M.-H.; Han, J.; Jin, Z.; Lee, D.-W.; Park, H.-S.; Kim, H.-S. Controlled and Oriented Immobilization of Protein by Site-Specific Incorporation of Unnatural Amino Acid. *Anal. Chem.* **2011**, *83*, 2841–2845.

(14) Knotts, T. A.; Rathore, N.; de Pablo, J. J. An entropic perspective of protein stability on surfaces. *Biophysical Journal* **2008**, *94*, 4473–4483.

(15) Wei, S.; Knotts, T. A. Effects of tethering a multistate folding protein to a surface. *J. Chem. Phys.* **2011**, *134*, 185101.

(16) Wei, S.; Knotts, T. A. Predicting stability of alpha-helical, orthogonal-bundle proteins on surfaces. *J. Chem. Phys.* **2010**, *133*, 115102.

(17) Bush, D. B.; Knotts, T. A. Probing the effects of surface hydrophobicity and tether orientation on antibody-antigen binding. *J. Chem. Phys.* **2017**, *146*, 155103.

(18) Wilding, K. M.; Smith, A. K.; Wilkerson, J. W.; Bush, D. B.; Knotts, T. A.; Bundy, B. C. The Locational Impact of Site-Specific PEGylation: Streamlined Screening with Cell-Free Protein Expression and Coarse-Grain Simulation. *ACS Synth. Biol.* **2018**, *7*, 510–521.

(19) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(20) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D., Jr. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(21) McQuaid, M.; Sun, H.; Rigby, D. Development and validation of COMPASS force field parameters for molecules with aliphatic azide chains. *J. Comput. Chem.* **2004**, *25*, 61–71.

(22) Pieffet, G.; Petukhov, P. A. Parameterization of aromatic azido groups: application as photoaffinity probes in molecular dynamics studies. *J. Mol. Model.* **2009**, *15*, 1291–1297.

(23) Carvalho, A. T. P.; Fernandes, P. A.; Ramos, M. J. Parameterization of AZT - A widely used nucleoside inhibitor of HIV-1 reverse transcriptase. *Int. J. Quantum Chem.* **2007**, *107*, 292–298, 2nd International Theoretical Biophysics Symposium, Orebro Univ, Orebro, Sweden, Jun 28-Jul 01, 2005.

(24) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, *34*, 2757–2770.

(25) Kato, A.; Kuratani, M.; Yanagisawa, T.; Ohtake, K.; Hayashi, A.; Amano, Y.; Kimura, K.; Yokoyama, S.; Sakamoto, K.; Shiraishi, Y. Extensive Survey of Antibody Invariant Positions for Efficient Chemical Conjugation Using Expanded Genetic Codes. *Bioconjugate Chem.* **2017**, *28*, 2099–2108.

(26) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. CHARMM - A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(27) MacKerell, A. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem B* **1998**, *102*, 3586–3616.

(28) MacKerell, A.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2001**, *56*, 257–265.

(29) The PyMOL Molecular Graphics System, Version 1.2r3pre. Schrödinger, LLC.

(30) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011**, *3*, 33.

(31) Yu, Y.; Fursule, I. A.; Mills, L. C.; Englert, D. L.; Berron, B. J.; Payne, C. M. CHARMM force field parameters for 2 '-hydroxybiphenyl-2-sulfinate, 2-hydroxybiphenyl, and related analogs. *J. Mol. Graph. Model.* **2017**, *72*, 32–42.

(32) Pavlova, A.; Parks, J. M.; Gumbart, J. C. Development of CHARMM-Compatible Force-Field Parameters for Cobalamin and Related Cofactors from Quantum Mechanical Calculations. *J. Chem. Theory Comp.* **2018**, *14*, 784–798.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al., Gaussian 09, Revision A.02. Gaussian, Inc., Wallingford, CT, 2016.

(34) Hopkins, C. W.; Roitberg, A. E. Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem. *J. Chem. Inf. Model.* **2014**, *54*, 1978–1986.

(35) Hagler, A. T. Force field development phase II: Relaxation of physics-based criteria... or inclusion of more rigorous physics into the representation of molecular energetics. *J. Comput. Aid. Mol. Des.* **2019**, *33*, 205–264.

(36) Bulacu, M.; Goga, N.; Zhao, W.; Rossi, G.; Monticelli, L.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved Angle Potentials for Coarse-Grained Molecular Dynamics Simulations. *J. Chem. Theory Comp.* **2013**, *9*, 3282–3292.

(37) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comp.* **2009**, *5*, 350–358.

(38) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. CATH - a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093–1108.

(39) Vanommeslaeghe, K.; MacKerell, A. D., Jr. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

(40) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155–3168.

(41) Plimpton, S. Fast Parallel Algorithms for Short-range Molecular-dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(42) Jo, S.; Cheng, X.; Islam, S. M.; Huang, L.; Rui, H.; Zhu, A.; Lee, H. S.; Qi, Y.; Han, W.; Vanommeslaeghe, K.; MacKerell, A. D., Jr.; Roux, B.; Im, W. In *Biomolecular Modeling and Simulations*; KarabenchevaChristova, T., Ed.; 2014; Vol. 96; pp 235–265.

(43) Saethre, L.; Berrah, N.; Bozek, J.; Borve, K.; Carroll, T.; Kukk, E.; Gard, G.; Winter, R.; Thomas, T. Chemical insights from high-resolution X-ray photoelectron spectroscopy and ab initio theory: Propyne, trifluoropropyne, and ethynylsulfur pentafluoride. *J. Am. Chem. Soc.* **2001**, *123*, 10729–10737.

(44) Schulze, B.; Schubert, U. S. Beyond click chemistry - supramolecular interactions of 1,2,3-triazoles. *Chem. Soc. Rev.* **2014**, *43*, 2522–2571.

(45) Anderson, D.; Rankin, D.; Robertson, A. Electron-diffraction Determination of Molecular Structures of Methylazide, Methylisocyanate and Methylisothiocyanate in Gas-phase. *J. Mol. Struct.* **1972**, *14*, 385–396.

(46) Chen, F.-F.; Wang, F. Electronic Structure of the Azide Group in 3 ‘-Azido-3 ‘-deoxythymidine (AZT) Compared to Small Azide Compounds. *Molecules* **2009**, *14*, 2656–2668.

(47) Fukui, K. Role of Frontier Orbitals in Chemical-reactions. *Science* **1982**, *218*, 747–754.

TOC Graphic