Molecular Dynamic Simulation of Protein Devices and the Parameterization of Azides and Alkynes for Use in Unnatural Amino Acid Models

Addison Kyle Smith  
*Brigham Young University*

Follow this and additional works at: [https://scholarsarchive.byu.edu/etd](https://scholarsarchive.byu.edu/etd)

*Part of the Engineering Commons*

**BYU ScholarsArchive Citation**

[https://scholarsarchive.byu.edu/etd/8783](https://scholarsarchive.byu.edu/etd/8783)

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Molecular Dynamic Simulation of Protein Devices and the Parameterization of Azides and Alkynes for Use in Unnatural Amino Acid Models

Addison Kyle Smith

A dissertation submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thomas A. Knotts IV, Chair
Bradley C. Bundy
Randy S. Lewis
Dean Wheeler
Matthew J. Memmott

Department of Chemical Engineering
Brigham Young University

Copyright © 2021 Addison Kyle Smith All Rights Reserved
ABSTRACT

Molecular Dynamic Simulation of Protein Devices and the Parameterization of Azides and Alkynes for Use in Unnatural Amino Acid Models

Addison Kyle Smith
Department of Chemical Engineering, BYU
Doctor of Philosophy

Proteins that have been modified by attaching them to a surface or to a polyethylene glycol (PEG) molecule can see many uses in therapeutics and diagnostics – these unique proteins are called protein devices. Current techniques can perform these functionalizations at a specific residue on the protein, but what remains is identifying what happens to protein structure when mutated, and where to perform the attachment. Both of these issues can be examined using molecular dynamic (MD) simulations. Currently, simulations of the unnatural amino acid (uAA) mutations necessary for protein device functionalization cannot be executed, and full-protein screens of all possible protein device models have never been attempted.

Results from this dissertation first employs a new model for simulating PEGylated protein devices building off of previous studies that explore where to attach functional groups. Next, many current assumptions in the community regarding ideal attachment sites are examined. Some of these factors include primary chain location, amino acid type, solvent accessibility, and secondary structure. The focus then turns to novel tertiary structure factors that could influence how well attachment locations affect overall protein device stability. The usefulness of each factor is analyzed to show what factors provided the best predictive power for a site’s performance in the screen. A general heuristic is given that could aid in future screens of other protein devices to reduce compute time and quickly identify sites for experimental examination.

To explore uAA mutation effects on protein structure, parameters are developed for linear moiety R-groups present in these novel amino acids. The CHARMM and CGenFF force fields currently lack parameters for most linear-angle molecular moieties. This work proposes a method that (1) develops CHARMM parameters for four small molecules that contain terminal azido and alkynyl groups using ffTK, (2) addresses linearity issues, and (3) validates ffTK results via in silico MD simulation. Dihedral analysis examines the linear-angle-containing dihedrals and compares methods for the moiety parameterization. Next, the small molecule parameters are combined with CGenFF to generate parameters for unnatural amino acid MD simulation in a protein. Finally, validation confirms the parameters derived in this work to appropriately simulate unnatural amino acids and small molecules with azido and alkynyl groups.

Keywords: Protein device, screen, CHARMM, unnatural amino acid, protein, simulation, molecular dynamics, parameterization, azide, alkyne
TABLE OF CONTENTS

LIST OF TABLES ................................................................. vi

LIST OF FIGURES .............................................................. vii

NOMENCLATURE .............................................................. ix

Chapter 1 Introduction ....................................................... 1

Chapter 2 Objectives ....................................................... 5
  2.1 Objective ............................................................... 5
  2.2 Tasks ................................................................. 5

Chapter 3 Literature Review ................................................ 7
  3.1 Protein Structure .................................................... 7
  3.2 Types of Protein Devices ........................................... 9
    3.2.1 PEGs .............................................................. 9
    3.2.2 Tethering to a Surface ......................................... 10
  3.3 Creation of Protein Devices ........................................ 11
  3.4 Advances in Computer Simulations .................................. 14
  3.5 All-Atom Models ..................................................... 15
    3.5.1 Azide and Alkyne Models ....................................... 16
    3.5.2 All-Atom Protein Simulation .................................... 17
  3.6 Coarse-Grain Simulation ............................................ 20
    3.6.1 Simulations Reveal Device Effects on Protein Stability .... 21
  3.7 Summary of Literature ............................................... 24

Chapter 4 Methods ............................................................ 25
  4.1 All-Atom Model Equations ........................................... 25
  4.2 Coarse-Grain Model Equations ....................................... 27
    4.2.1 Go-Model ....................................................... 27
    4.2.2 Surface Model .................................................. 28
  4.3 Replica Exchange ..................................................... 29
  4.4 WHAM and MBAR ..................................................... 31
  4.5 Heat Capacity .......................................................... 32

Chapter 5 Development of PEG Coarse-grain Model ....................... 34
  5.1 Introduction .......................................................... 34
  5.2 Methods ............................................................... 35
  5.3 PEG Go-Model ......................................................... 36
  5.4 Results ............................................................... 37
  5.5 Conclusions ........................................................... 39
**Chapter 6**  TEM-1 Screen ......................................................... 41  
  6.1 Introduction ................................................................. 41  
  6.2 Models ................................................................. 42  
    6.2.1 Protein ............................................................... 42  
  6.3 Method ............................................................... 43  
    6.3.1 Simulation Protocols ............................................. 43  
    6.3.2 Quantitative Folding Behavior .................................. 44  
  6.4 Results ............................................................... 45  
    6.4.1 Quantitative Site Performance .................................... 45  
  6.5 Discussion ............................................................. 53  
    6.5.1 Domain-specific Native Contact Analysis ....................... 53  
    6.5.2 Domain Stability on Screen Performance ....................... 54  
    6.5.3 Functionalization Type Effect on Protein Stability ............ 56  
  6.6 Procedure for Identifying Poor Sites ............................... 59  
  6.7 Considerations .......................................................... 60  
  6.8 Conclusions ............................................................. 61  

**Chapter 7**  Parameterization of uAA with Azides and Alkynes ........... 63  
  7.1 Introduction ............................................................. 63  
  7.2 Model Parameters .......................................................... 63  
  7.3 Method for Parameterization ........................................... 66  
    7.3.1 Parameter Determination: ffTK .................................... 66  
    7.3.2 Validation of Model Parameters for Azide and Alkyne Moieties ... 71  
  7.4 Results ............................................................... 72  
    7.4.1 Optimized Parameters ............................................. 73  
    7.4.2 LACD Parameterization ........................................... 78  
    7.4.3 Validation .......................................................... 84  
  7.5 Conclusion ............................................................. 87  

**Chapter 8**  Challenges when Simulating LACD Moieties ..................... 88  
  8.1 Introduction ............................................................. 88  
  8.2 Background: Modern Molecular Force Fields ....................... 88  
  8.3 Methods: Simulation .................................................... 89  
    8.3.1 Model ............................................................... 89  
    8.3.2 Simulation Protocols ............................................. 91  
  8.4 Results ............................................................... 93  
    8.4.1 Ramping in NVE Simulations ..................................... 93  
    8.4.2 NVT Simulations with LACD Parameters ......................... 94  
    8.4.3 Examining a Coupling Event ..................................... 96  
    8.4.4 NVE Simulations without LACD Parameters ....................... 107  
  8.5 Conclusion ............................................................. 109  

**Chapter 9**  Future Work ..................................................... 110  
  9.1 Future Work for Protein Devices .................................... 110
9.2 Future Work for LACD Parameterization ............................................. 110
  9.2.1 LACD Parameterization Using Üery-Bradley Potentials ...................... 111
  9.2.2 LACD Parameterization Using Unique Angle Potential ........................ 112

REFERENCES ......................................................................................... 113

Chapter 10 Appendix A: Supplementary Material to Chapter 6 ......................... 126
  10.1 Site Rankings from Tethering Screen .................................................. 126
  10.2 Accessibility Scatter Plots ................................................................. 134
  10.3 Surface Tethering Orientations ........................................................... 135

Chapter 11 Appendix B: Supplementary Material to Chapter 8 ......................... 138
  11.1 Methods: Transformations for Short-Time Simulations .......................... 138
  11.2 Rejection Vector Definition .............................................................. 138
  11.3 Direction of Angle Space .................................................................... 139
  11.4 Transform Cartesian Movement to Parameter Space .............................. 140
      11.4.1 Transform to Angle Space ......................................................... 141
      11.4.2 Transform to Dihedral Space ..................................................... 142
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Best and worst performing sites from the tethering screen</td>
<td>48</td>
</tr>
<tr>
<td>6.2</td>
<td>Best and worst performing sites from the PEGylation screen</td>
<td>49</td>
</tr>
<tr>
<td>6.3</td>
<td>Domain performance based on site rankings</td>
<td>52</td>
</tr>
<tr>
<td>7.1</td>
<td>Optimized CHARMM parameters using ffTK</td>
<td>73</td>
</tr>
<tr>
<td>7.2</td>
<td>Best Fit LACD CHARMM parameters for all relevant atoms in the linear moiety</td>
<td>81</td>
</tr>
<tr>
<td>8.1</td>
<td>Step-by-step analysis of coupling and discontinuities using a 1.0 fs timestep</td>
<td>99</td>
</tr>
<tr>
<td>8.2</td>
<td>Step-by-step analysis of coupling using a 0.01 fs timestep</td>
<td>100</td>
</tr>
<tr>
<td>8.3</td>
<td>Step-by-step analysis of coupling and discontinuities using a 1.0 fs timestep and excluding the LACD parameter</td>
<td>106</td>
</tr>
<tr>
<td>10.1</td>
<td>Complete list of site performances from tethering screen</td>
<td>127</td>
</tr>
<tr>
<td>10.2</td>
<td>Complete list of site performances from PEGylation screen</td>
<td>131</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

1.1 Representation of Tethered and PEGylated protein devices .......................... 2
3.1 Protein Structures ................................................................................. 8
3.2 PRECISE Method Chemistry ................................................................. 13
3.3 All-atom and Coarse-grain Modeling ...................................................... 19
3.4 Ideal and Real Surfaces ........................................................................ 21
4.1 Energy Landscape for Replica Exchange .................................................. 29
5.1 Sites on Lysozyme used for experiments and PEG coarse-grain model simulations. 35
5.2 Simulated $\Delta T_m$ values plotted with experimental $\Delta T_m$ values .................. 37
5.3 Compare simulated and experimental $\Delta T_m$ values in PEG model .................. 38
6.1 TEM-1 (PDB ID: 1XPB) structure and active site ........................................ 43
6.2 Resulting heat capacity curves from the TEM-1 screen when tethering to a surface . 46
6.3 Resulting heat capacity curves from the TEM-1 screen when PEGylating .......... 47
6.4 Site performance observed on the TEM-1 surface structure .......................... 51
6.5 Native contact plot broken-down into domain contribution to stability .............. 53
6.6 Sample structures observed when functionalizing the Beta domain ................. 54
6.7 Sample structures observed when functionalizing the Alpha II domain .......... 55
6.8 Sample structures observed when functionalizing the Loop domains .............. 57
6.9 Sample structures observed when functionalizing the Alpha I domain ............ 57
7.1 uAA used in PRECISE ............................................................................ 64
7.2 Small molecules parameterized in this work ............................................. 67
7.3 Trastuzumab Fab fragments (PDB ID: 5XHF) ............................................ 72
7.4 Torsion profiles for all normal dihedrals .................................................. 78
7.5 Torsion profiles for LACDs produced by ffTK method ............................... 79
7.6 Torsion profiles for LACDs produced by Best Fit method ............................ 80
7.7 NVE Simulations of PAZ ....................................................................... 82
7.8 RMSD plots of azide and alkyne small molecules ....................................... 85
7.9 RMSD plots of Tra-Fab fragment ............................................................ 86
8.1 Four-atom Model .................................................................................. 90
8.2 NVE in Simple Model ........................................................................... 94
8.3 Molecular Structures with LACD Parameters ......................................... 95
8.4 Trajectory Using Long and Short Timesteps ............................................. 97
8.5 Trajectory without LACD Parameters .................................................... 105
8.6 Molecular Structures without LACD Parameters ...................................... 107
10.1 Accessibility and site performance scatter plot for tethering screen .......... 135
10.2 Accessibility and site performance scatter plot for PEGylation screen .......... 135
10.3 Protein-surface orientations during simulation ........................................ 136
11.1 Rejection Vector .................................................................................. 138
11.2 Define Angle Space .......................................................... 140
11.3 Angle Space Transformation ............................................ 141
11.4 Dihedral Space Transformation ........................................ 142
NOMENCLATURE

Constants

\( b \) Bond distance between two adjacent atoms
\( \theta \) Angle value between three consecutive atoms
\( \phi \) Dihedral angle value between four consecutive atoms
\( \omega \) Improper dihedral value between four atoms
\( q \) Charge of an atom
\( S \) Urey-Bradley value between three consecutive atoms
\( K \) Force constant for Parameterization
\( V \) Potential Energy
\( \varepsilon \) Lennard-Jones energy
\( R_{\text{min}} \) Lennard-Jones distance
\( k_b \) Boltzmann constant
\( T \) Temperature
\( \vec{r} \) Radial movement in circular coordinates
\( \Theta \) Angular movement in circular coordinates
\( \varepsilon_0 \) Dielectric constant

Subscripts, superscripts, and other indicators

- \( [\cdot]_0 \) indicates \([\cdot]\) is the equilibrium value for some parameter
- \( [\cdot]_b \) indicates \([\cdot]\) is a bond parameter
- \( [\cdot]_{\theta} \) indicates \([\cdot]\) is an angle parameter
- \( [\cdot]_{\phi} \) indicates \([\cdot]\) is a dihedral parameter
- \( [\cdot]_{\omega} \) indicates \([\cdot]\) is an improper parameter
- \( [\cdot]_S \) indicates \([\cdot]\) is a Urey-Bradley parameter
- \( [\cdot]_{NC} \) indicates \([\cdot]\) is a native contact parameter
- \( [\cdot]_i \) an atom in a nonbonded pair
- \( [\cdot]_j \) an atom in a nonbonded pair
- \( [\cdot] \) indicates atoms in \([\cdot]\) form a bond
- \( [] \) indicates atoms in \([\cdot]\) form an angle
- \( [\cdot] \) indicates \([\cdot]\) is a vector value
CHAPTER 1. INTRODUCTION

Ever since the human genome was sequenced in 2003, research exploring protein sequence and structure has thrived. This field of study, called proteomics, has many branches that all attempt to explain the many different protein functions that give us life. One branch of proteomics seeks to harness a protein’s unique function by attaching man-made chemical features onto protein structure. This can lead to improved function and adds additional human control. The resulting bioengineered creations are called protein devices.

Protein devices are proteins that have undergone biotechnological modification to harness the function of protein molecules for a specific application. 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 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 a protein’s primary sequence [8, 9]. The uAAs commonly used are p-azido phenylalanine (pAz) and p-propargloxy-phenylalanine (pPa) as they 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 two types of protein devices used in this research are proteins tethered to a hydrophilic surface and proteins with polyethylene glycol bonded to its tertiary structure. The former are referred to in this paper as tethered proteins and the latter are referred to as PEGylated proteins. (See Figure 1.1)
Tethering proteins to surfaces (Figure 1.1A) is usually done by attaching one end of a linker molecule to the substrate and the other end to the protein. Many different tethering strategies exist, but the purpose is to create lab-on-a-chip technologies. Applications include biosensors, biocatalysis, therapeutics, microreactors, and complex artificial systems [14–18]. In practice, the usefulness of this approach is in part limited by its unpredictable performance and structural changes that can occur when proteins are deposited onto the surface [19, 20].

PEGylated proteins (Figure 1.1B) have been used for decades as therapeutics. The primary purposes of the polymer is to slow filtration in the body and stabilize the target proteins [6, 21–24]. At present, a combinatorial approach is used to select the size of the PEG and the location of the attachment. Using a more deliberate and rational design approach is challenging because so little is known about the underlying biophysics involved. Obtaining fundamental understanding about the system is difficult because standard methods for determining protein structures, such x-ray diffraction and cryo EM, cannot be used because they require a high degree of crystallinity.

These protein-based devices lack consistent models that identify all protein interactions involved. As this technology progresses these engineered devices become more stable, easily manipulated, highly active, and perform well in extracellular environments. To achieve this, scientists need information and understanding. Computer-aided molecular modeling provides needed detail and scope to more fully grasp structural protein chemistry.

Current simulation research on protein devices primarily focus on how the protein device as a whole affects protein structure. Little attention is paid to the effects of the uAA mutation.
on the protein or the effects on local structures where attachment occurs. The current assumption claims that an uAA mutation does not appreciably affect any protein parameter. The primary chain mutation is therefore ignored and a wild-type sequence is assumed. This work tests this unnatural amino acid assumption and hypothesizes that uAA mutations affects the structure, stability and activity of proteins and protein devices. Investigation will focus on: (1) Capturing the effects uAA exert on bulk-phase protein stability by developing new all-atom computational models. (2) Resolving issues associated with linear-structures in MD simulation. (3) Beginning development of PEGylated protein parameters and (4) refining coarse-grain a priori site selection heuristics.

The remainder of this dissertation is as follows. Chapter 2 identifies the objectives of this work. Chapter 3 contains a literature review of relevant concepts and Chapter 4 contains general information of relevant theory, methods and software for the dissertation. The following chapters (Chapters 5-8) describe the studies completed to further understand protein device chemistry and linear molecule simulation. Chapter 5 shows the first-ever coarse-grain simulations of PEGylated protein systems and includes the parameterization process and experimental work done to validate the model. This work was published in the article The Locational Impact of Site-Specific PEGylation: Streamlined Screening with Cell-Free Protein Expression and Coarse-Grain Simulation where Lysozyme was PEGylated at multiple attachment sites and tested for thermal stability. Newly-developed coarse-grain simulation was then compared to experimental findings to validate the model’s validity. Chapter 6 builds off the findings in the previous chapter by screening a new protein for stable structures. The bulk of these findings were published in the article Coarse-grain Simulation of PEGylated and Tethered Protein Devices at all Experimentally-Accessible Surface Residues on β-lactamase for Stability Analysis and Comparison. This case-study runs an in silico investigation of all experimentally accessible sites on the protein to screen for optimal sites using both tethering and PEGylation functionalization methods. These results provide candidates for experimental study and improves heuristics for future screens. The parameter development process in Chapters 7 and 8 were published in the article Parameterization of Unnatural Amino Acids with Azido and Alkynyl R-Groups for Use in Molecular Simulations. Chapter 7 challenges the previous assumption that the uAA mutation has no affect on protein structure by creating all-atom parameters for these amino acids. During this process, parameters are created for azides and alkynes and represent their first use in the CHARMM force field. Chapter 8 explores the challenges as-
associated with simulating linear molecules in molecular dynamic simulation. Different approaches for parameterization explore numeric stability and the compromises made in the previous chapter. Each of these chapters identify a problem or hypothesis, relevant methods specific to the chapter, and a discussion about the impact of the work. The final chapter (Chapter 9) summarizes all work completed and poses work for future research.
CHAPTER 2. OBJECTIVES

2.1 Objective

The primary purpose of the works in this dissertation is to enable azido- and alkynyl-containing unnatural amino acids in all-atom simulation, and to screen PEGylated and tethered protein devices for functionalization attachment location effectiveness. These results from this work should provide researchers with new information to aid in all-atom and coarse-grain molecular dynamic simulations. All-atom contributions aim to better understand how azido and alkynyl R-groups affect protein structure at atomic level resolution. Coarse-grain tasks improve upon previous protein device heuristics by introducing PEG into protein device simulation and recognizing how functionalization effects protein tertiary structure. Contributions from this work can be used in disciplines that seek to use simulation to better utilize protein chemistry and are as diverse as microsensors, national defence, diagnostics, biocatalytics and therapeutics.

2.2 Tasks

These objectives were fulfilled by accomplishing the following tasks:

1. Create a new method and model that introduces PEGylation effects on protein structure
   (a) Develop a coarse-grain PEG-polymer model that can be attached to a Go-model protein
   (b) Determine in silico PEGylated protein device behavior by functionalizing on multiple solvent-accessible residues and calculating $\Delta T_m$
   (c) Compare simulation results with experiments to evaluate model accuracy

2. Improve upon the current in silico screening methods that lack breadth and depth to better inform the experimental design of protein devices
   (a) Port and benchmark Go-model onto LAMMPS simulation software
(b) Simulate functionalization (Tethering and PEGylation) at all surface-accessible sites on the protein

(c) Compare the effectiveness of each functionalization method

(d) Identify what aspects of protein structure affect protein device stability

(e) Develop heuristic to identify, a priori, good potential attachment location

3. Enable the simulation, and predict the effects, of PRECISE uAA mutation on protein tertiary structure by parameterizing azido and alkynyl groups

(a) Determine method and type of molecules to parameterize

(b) Resolve parameterization issues when simulating sp\(^1\) hybridized orbitals in QM force fields

(c) Validate parameters using small molecules

(d) Validate parameters for use in protein simulation

4. Identify why discontinuities exist when simulating linear angle containing dihedrals and weigh the consequences of current parameterization methods

(a) Create a model that isolates discontinuous motion

(b) Identify source of numeric instability

(c) Determine potential solutions

(d) Identify validity, gains, and losses for each parameterization method

These tasks are accomplished in the following chapters of this dissertation. Chapter 5 (Task 1) shows the method and process used to determine coarse-grain parameters for PEGylated proteins. The model is then compared to experimental data. Chapter 6 (Task 2) reports the case-study of an in silico protein screen using two functionalization methods. Chapter 7 (Task 3) shows the process implemented to determine azido and alkynyl molecular group parameters. Chapter 8 (Task 4) examines why linear angle containing dihedrals are susceptible to numeric overload and the validity of current approaches for their simulation.
CHAPTER 3. LITERATURE REVIEW

3.1 Protein Structure

Protein structure is essential to protein function. At its most basic, proteins consist of a single chain of amino acids (AA) of varying lengths. This is often referred to as the primary structure. The types of AA commonly seen in nature are limited to a list of 20 naturally-occurring amino acids. Each of these 20 AA have a so-called side-chain or R-group, each of which has a unique chemical characteristic and have unique hydrophobicity, charge, aromaticity, reactivity and flexibility. These properties define the functions of an AA.

The interaction of amino acid side-chains with other primary chain AA and the solvent, defines the protein’s primary structure properties (Figure 3.1A). These structures self-aggregate in a process called folding. Aggregation can form into two major structural motifs (Figure 3.1B). α-helical structures occur when the backbone of the amino acids coil due to hydrogen bonding between amino acids separated 3-4 locations away from each other. β-sheet structures form when the chain folds back on itself and hydrogen bonds form between amino acids on chains adjacent to each other. Loop and unstructured regions are found connecting consecutive secondary structures together. These latter are important for higher order structures to form.

Inter-atomic, hydrogen bonding, and hydrophobic forces further “fold” the rigid secondary substructures into complex configurations called tertiary and quaternary structures (Figure 3.1C). The resulting complex structures produce proteins with highly specific functions. For example many proteins useful in diagnostics have an active site in the tertiary structure that selectively bonds to other proteins, catalysts, or compounds. The important concept that defines protein chemistry is structure equals function. A protein structured differently, or modified in any way, that deviates from the naturally occurring wild-type (WT) structure could lose activity or become inert.

Because structure is so important to protein function, it is necessary to understand the thermodynamics behind protein unfolding. This process is called denaturing or unfolding. Though
Figure 3.1: Panel A) Depiction of protein primary structure as a chain of amino acids connected by a peptide bond. Panel B) Secondary structure primarily consists of amino acids folded into α-helices, β-sheets and loop regions Panel C) Tertiary and quaternary structure take secondary structure and form protein domains. Depicted is a protein antibody.

many theories exist on the driving force for folding and unfolding, the most prominent suggests folding is entropically controlled. Changes in Gibbs free energy upon folding are composed of both enthalpic (energetic) and entropic contributions according to:

$$\Delta G = \Delta H - T \Delta S$$

where $\Delta G$ is the change in Gibbs free energy upon folding, $\Delta H$ is the change in enthalpy upon folding, $T$ is the system temperature and $\Delta S$ is the change in entropy between a completely unfolded to completely folded protein in water. In an aqueous environment and at moderate temperatures, unfolded and unstructured hydrophobic AA chains form neighboring water molecules into a dense
ordered shell, called a “water cage,” around the residues [25]. Because most AA are moderately to highly hydrophobic, this water cage can be large and will result in a very low entropic system. To maximize entropy, the system thermodynamically prefers a folded protein state at room temperature that lets water move freely and allows hydrophobic AA residues to cluster in the interior of the molecule. By decreasing the entropy of the protein molecule, the freed water molecules become mobile and increase the entropy of the system.

As the system temperature increases, proteins unfold. At high temperatures the energetic forces driving the system to fold are overcome by thermal fluctuations. Moreover, at high temperatures, maximizing protein entropy become more important than enthalpic contributions to the free energy. The result is that the system is most stable if unfolded. By unfolding its tertiary and secondary structure, the functionally denatured protein looses all activity. This temperature at which the protein becomes inert is called its melting temperature (Tm).

3.2 Types of Protein Devices

Protein devices are proteins that have been optimized for pharmaceutical, industrial, and academic applications to introduce greater control over protein-based system and optimize bioreactions. Protein devices consist of (1) a binding site on the protein, (2) a linker to attach the two components together, and (3) the device that will be attached to the protein. A simple example of this is their use in over-the-counter pregnancy tests, but current applications include cancer therapeutics, biocatalysis, microreactors and other diagnostic tools [6, 14–18, 21–24]. Once created, these devices have potential to revolutionize proteomics because they allow direct human involvement in protein chemistry. However, much is not known about how protein structure is affected when protein devices are created, and because structure determines function, there is much room for better and more robust protein devices. The two types of protein devices that are used in this work are surface-tethered protein devices and PEGylated protein devices.

3.2.1 PEGs

Many current biotechnologies harness the power of proteins and augment these through conjugation to other chemicals. One of the earliest and most common examples of this involves at-
taching polyethylene glycol (PEG) to a protein [26, 27]. Currently, PEGylated proteins are used in pharmaceutical applications to stabilize or purify target proteins [6, 21–24, 28]. PEGylated biocatalysts exhibit superior performance with benefits including superior selectivity, reduced environmental impact, slower filtration in the body, and greater activity in extreme environments [9, 29].

Determining PEG structure for drug design and stability optimization is difficult. Traditional methods for viewing and testing protein structures, like x-ray diffraction and cryo EM, do not work because PEG chains are very unstructured. This makes confirmation of PEG theory and general structure impossible. Our understanding of this science is also limited by attachment methods which typically use non-specific attachment methods [24, 30–33].

Despite their extensive use in research and commercial applications, limitations in structure observation prevents device optimization. Until recently, non-specific functionalization methods have been sufficient, but some therapeutically-relevant proteins are quickly metabolized in the body and cannot be PEGylated because they interfere with active site function [34, 35]. Site-specific functionalization is beneficial in these cases because PEGylation can be strategically placed at a location that is not detrimental to protein activity, but would also slow biologic filtration.

3.2.2 Tethering to a Surface

Multiple proteins covalently tethered to a surface form what is called a protein microarray. Microarrays are used in detection and diagnostics to impart ELISA-like detection in a compact, portable, rapid, and easy-to-read form (Figure 3.4). These labs-on-a-chip seek to accomplish sensitive detection by increasing active detector sites per area to create a high-dense, highly-active protein assay. This advance in uniformity facilitates the scaling of bench-top experiments down to chip-sized tests [2, 36, 37].

The theory behind tethering proteins to non-interacting surfaces should, theoretically, always stabilize the protein [38–40]. The primary basis of this theory focuses on the entropy reduction seen by the protein (Eqn. 3.1). As a folded protein is more compact, the surface affects this compact state significantly less than an unfolded or a transition state configuration because the more unstructured transition state peptide is much more likely to encounter the surface when denaturing. Thus, the surface sterically hinders protein unfolding configurations and, consequently, these bound proteins will always be more stable than their bulk phase counterparts.
In practice, surface tethering does not reap all the benefits the theory would suggest. Enthalpic interactions introduced to the system by the surface proves to destabilize the protein [8, 41]. Stability becomes a balance between entropic stabilization that results from blocked unfolding pathways and enthalpic destabilizations brought on by new surface-induced unfolding and bonding pathways [42, 43].

Despite sufficient method and theory, insufficient models and unknown interactions prevent efficient microarray designs because protein-surface structures remain unexplored [4, 44]. Structural analysis of these systems are impossible using traditional spectroscopy because the surface interferes with the fundamental physics and optics required for these methods. Of the experimental observations that have been made, the main theme is that activity of these devices vary widely between tethered proteins, with some losing activity altogether [45, 46]. Another finding is that loss in efficiency primarily stems from manufacturing imperfections and cross-reactivity [5].

### 3.3 Creation of Protein Devices

How protein devices are created is essential in understanding what structures will result. Because structure determines function, it is critical to create a protein device that is structured in a way that optimizes the protein’s function. For the duration of this dissertation, the residue on the primary amino acid chain where the linker attaches to the protein is referred to as the “attachment site” or “functionalization site.” Functionalization, in this case, is defined to be some process or method that bonds protein and device.

When creating a tethered protein, the most basic functionalization method is simple adsorption [37, 47, 48]. Simple adsorption (and its cousin physical entrapment) utilizes adsorption and proximity forces to link the protein to a surface. Many current surface-protein products, like the enzyme-linked immunosorbent assays (ELISA) and previously mentioned pregnancy test, utilize an adsorptive technique because it is easy to implement and sufficiently maintains protein stability to accomplish basic medical diagnostic tests [49, 50]. However, simple adsorption significantly decreases overall tethered protein stability compared to the original bulk phase and can result in surface activities as low as 10% [41]. Loss in activity results from (1) unfavorable adsorption orientations where the proteins adsorb such that the antigen binding domains are hidden from the
antigen solution, and (2) the protein partially or fully denatures onto the surface due to dominating surface effects [51].

Covalent attachment methods are used to functionalize both tethered protein devices and PEGylated protein devices. These methods use a covalent bond as the linker between AA and functionalization group [12]. Early covalent bonding techniques add short reactive amino acid chains to the N- or C- termini of the protein [23,52]. These additions allow for selective binding to a surface or PEG. Uniform structures result, but such an approach is limited by a lack of functionalization locations. Moreover, unstructured protein terminal ends tend to easily denature causing the entire protein to unfold [52]. While exceptions exist, the protein termini are not locations that maximize protein activity and often lead to premature denaturing [23].

Lysine (Lys) functionalization creates an amide bond between the Lys nitro group and N-Hydroxysuccinimide (NHS) esters [23, 53]. The NHS ester can be placed on a surface or PEG molecule and will react with any solvent-accessible Lys residue on the protein. Because Lys groups are frequently found on protein surfaces, this method strongly suffers from non-specific binding orientations. In an attempt to produce more uniform attachment, Ladivère et al. looked into a method for more controlled tethering [54]. This was accomplished by inserting a chain of 6 Lys AA into the primary sequence of the HIV-1 protein RH24 at a location where the tethered protein would stand vertical and did not interfere with secondary or tertiary structure. The modified proteins exhibited greater stability after 7 days when compared to the unmodified chain. However, this selective attachment method would not work in all situations as multiple consecutive Lys residues will strongly affect protein structure if placed at a location not tolerant to mutations. Also, because solvent-accessible Lys is still present in other parts of the protein, this method still suffers from non-specificity issues.

Further development in protein device creation utilizes less common solvent-accessible cysteine (Cys) residues on the exterior surface of the protein. This allows for greater selectively when functionalizing the protein [55]. Smith et al. looked at a glucose binding protein (GBP) with an engineered single Cys on the protein’s surface. After coating the surface with bis-maleimidoethane, they then use a zinc-catalyzed disulfide reaction to create a covalent tether to a glass slide. The protein remained stable and active on the surface [55]. Cys attachment provides more uniform structures and some site-specificity is possible if Cys replaces the residue
where functionalization is desired. However, true site-specificity is not possible as proteins with naturally-occurring surface Cys would still suffer from non-uniformity. Many moderate to large proteins contain these surface-accessible Cys residues, and thus only partial site-specificity is possible in most practical applications.

All non-specific functionalization methods inherently contain limitations dictating where a protein device link can be established [56]. This means a protein’s optimal, and most stable configuration, may result from a binding site established at any one of many AA. With only 20 naturally-occurring AA, it is very hard to select a specific solvent-accessible AA to functionalize. Moreover, the previous functionalization methods discussed do not have the ability to explore all possible attachment sites on the tertiary structure and might overlook an optimal configuration.

![Figure 3.2: Panel A) pAz unnatural amino acid. Panel B) pPa unnatural amino acid. Panel C) triazole linking group created between either above uAA and surface bead after the “click” reaction.](image)

The protein residue-explicit covalent immobilization for stability enhancement (PRECISE) method overcomes the limitations of previous controlled-attachment methods by allowing covalent tethering at any single AA location or multiple specific AA locations [8, 57]. Residue-specific selectivity is achieved by mutating the naturally-occurring amino acid to an unnatural amino acid, namely p-azido-L-phenylalanine (pAz) or p-propargyloxyphenylalanine (pPa) (Figure 3.2A & B). Unnatural amino acid (uAA) incorporation is accomplished by modifying the in vivo insertion method first proposed by Wang et al. for use in a cell-free protein synthesis reaction [58]. The insertion reaction recodes the amber stop codon UAG on mRNA to code for a modified orthogonal
tRNA protein that carries the desired uAA [59]. The engineered tRNA and tRNA synthetase then bonds the uAA onto the primary sequence [60].

pPa and pAz were selected as the target uAA because they contains an alkyne or azide group, respectively, at the terminal end of its R-group which allows for a 1,3 dipolar cycloaddition “click” reaction [10, 11]. Upon reaction with uAA, the protein becomes covalently bound to the device via the triazole linking group (Figure 3.2C). Because this “click” chemical reaction is only possible with these uAA residue, the only possible protein device is the one where the protein is bonded to the device at the mutated site. Thus, uniform and site-selective protein devices are possible in a constantly-repeatable manner.

Because PRECISE technology enables the controlled creation of protein devices, attention turns to what protein device structures optimize protein function. Experimental evidence shows that PRECISE attachment at different residues do not produce protein devices with equal activity and stability [8, 9]. Moreover, experimental trial and error to determine optimal attachment locations is expensive in time and money [61,62]. Instead of only using an in vivo (or CFPS) approach for determining optimal functionalization locations, recent advances in computer simulations have enabled in silico screening to inform experimental design.

3.4 Advances in Computer Simulations

Due to experimental limitations, several recent advances in protein technologies have come from computational simulations. Computer simulations overcome problems facing protein-based devices in two ways: (1) provide insight into the molecular-level, structural behavior of PEGylated proteins and tethered proteins, and (2) rapidly assess the success of different tethering techniques and locations. Simulations are limited by either high computational costs or unknown interaction parameters. These limitations echo the two schools of thought when approaching a molecular simulation: atomic vs. coarse-grain models. The primary difference resides in the simulation resolution, i.e. computation “site” size, where sites are defined to be the smallest interacting unit. Both approaches are used in this dissertation to analyze protein device systems. First, all-atom models, findings, and limitation are addressed followed by an analysis and discussion of coarse-grain models.
3.5 All-Atom Models

Atomic models use individual atoms to define the interacting site, and parameters are determined for every unique atom. Organic atomic models predominantly use atoms with few atomic orbitals, and all interactions are sufficiently distant to assume predominantly Newtonian physics. Bonded interactions more than four atoms removed from a given atomic site is assumed insignificant, and all nonbonded interactions are either coulombic or follow a Lennard-Jones potential to replicate van der Waals energies. These “all-atom models” provide the most accurate and most detailed computer simulations of organic molecules without using quantum mechanical (QM) equations.

All-atom simulation has been used extensively in computational biology because in silico analysis can provide atomic-resolution insight into systems that cannot be observed using experimental methods. There are many bio-based all-atom force fields (FF) used computational biology, and one of the oldest and most popular is the CHARMM FF. This FF was specifically created for use in protein and macromolecule systems [63, 64]. The CHARMM set of parameters and equations is the predominant force field used in this dissertation. The AMBER FF is also widely-used in protein and DNA simulation, and differences between CHARMM and AMBER stem from how parameters are derived [65, 66]. The GROMOS FF is used in the GROMACS simulation program, and these parameters optimize protein structure in aqueous solution [67, 68]. Other FFs exist for organic molecule and protein simulation, but their use is less general. The COMPASS FF, for example, is optimized for organic small molecules and takes a purely ab initio approach to parameterization [69].

Many of these bio-based force fields include additions that expand the scope of parameterized atoms beyond the initial published set. The CHARMM general force field (CGenFF) parameterizes many organic structures not seen in protein or DNA models [70]. Parameters were determined by parameterizing molecules with unique structures. The resulting parameters can then be applied to molecules outside the original cases [71]. The general AMBER force field (GAFF) parameterizes unique structures more generally by averaging parameters from a set of ab initio structures that can then be extrapolated to novel organic molecular structures [72]. Both CGenFF and GAFF are constantly receiving additional contributions and revisions as unique chemical structures are studied and parameterized.
3.5.1 Azide and Alkyne Models

As was mentioned in Section 3.3, “click” chemistry is essential for site-specific device creation by reacting azides and alkynes to form a covalent link between two organic molecules. Recent work has shown that uAA mutations that contain these moieties in their R-groups can adversely affect protein stability and activity [73]. In order to use MD to examine these mutation effects, azides and alkynes must be parameterized in the CHARMM FF. Unfortunately, CHARMM FF and CGenFF do not contain these parameters. However, azides and alkynes have been simulated before in the COMPASS and AMBER all-atom force fields.

The general AMBER force field paper reports parameters for many small molecule types. Included in this work are azides, alkynes and other sp$^1$ hybridized structures [72]. Due to the number of molecules parameterized in this paper, angle parameters were determined with minimal QM basis and use an empirical formula to parameterize angle moieties. This method reduces the number of QM simulations and generates good parameters that can be used in most general applications. However, this method sacrifices molecular mechanic (MM) feedback, and parameters are assumed to adhere to Hooke’s law. These omissions in methodology will be shown in this work to be essential in angle parameterization for linear-angle-containing moieties.

Dihedral parameters are most commonly determined using a QM potential energy scan (PES). A PES incrementally rotates the molecular structure around the dihedral angle. At each increment the structure relaxes into its optimized positions and the energy of the system is recorded. GAFF does this for all torsional profiles, but excluded from this list are any sp$^1$ hybridized molecules. The reasoning for this omission is not given and the reader is supposed to assume dihedral contribution is null [72].

Other papers using the COMPASS and AMBER FF parameterize azido moieties with dihedral parameters [74–76]. These papers use a QM basis and MM feedback in all stages of parameterization except dihedral parameterization. Instead, when dihedrals are parameterized, only a best fit approach is used. The QM PES is set as the objective function and the dihedral cosine expansion is then fit to match. No alterations to the QM PES are identified and other MM algorithms to reduce dihedral coupling modes are not included in their analysis.

CGenFF has force field parameters for linear cyano and alkyne groups. However, cyano groups exclude dihedral parameters, and the research specifically recommends against using the
reported alkynyl parameters (which include dihedral parameters) [70]. Again, the literature is silent on the reasoning for these decisions.

Notably, a detailed energy analysis on angle and dihedral potentials is absent from all of these studies. It is likely that energy analysis has been unnoticed in linear moieties due to the fact that ensembles with thermostats or minimization algorithms can mask energetic inconsistencies [77, 78]. Or, inconsistencies have been identified to have resulted from the model parameters themselves, but the cause is unknown [79].

Proper dihedral angle modeling is key to creating accurate force fields for biomolecular simulation [80–84]. Unfortunately, many recent techniques in biotechnology use chemical moieties that are not commonly found in nature and whose molecular motions are not adequately described by current formalisms. The structures found in these azido- and alkynyl-containing uAA exemplify some of these challenges. For conciseness, the term linear angle containing dihedral (LACD) is used to describe the linear structure under investigation. This is a 4-bodied intermolecular structure where three consecutive atoms within the moiety form a linear or near-linear angle. Such structures most commonly form when model parameters attempt to recreate native linear structure (e.g. alkynyl, azido, and cyano groups), but can also form when external forces induce linearity.

All popular biological force fields in current use are inconsistent in LACD parameterization methods and a proper energy analysis is missing. Moreover, numeric instability is prominent when parameterizing LACDs. The source of this instability, the potential solutions, and compromises for current LACD parameterization processes is identified in this work. The goal is to identify the best solution that uses current force field equations and to elucidate all sources of error when simulating linear chemical moieties.

3.5.2 All-Atom Protein Simulation

All-atom simulations are used to compute molecular behavior as close as possible to actual behavior without using QM models. Thus, actual molecular movement can be determined in systems larger than is possible with QM [85]. However, for protein-sized molecules, the cost of calculating the interatomic forces in an all-atom model is still computationally-heavy, and only a relatively few iterations can thus be calculated. Thus, when using all-atom models in molecular
dynamic (MD), only very short timescales can be achieved on most computers and supercomputers.

The results of such a costly effort is very high detail. For small systems and short timescales, this approach works well because ensemble interactions are physically close together and computation times remain manageable. For these small systems, the atomistic model is not computationally demanding and energetic state minima can be found quickly [86]. Computations become very large when simulating proteins larger than 40 amino acid units long. As scale increases, computation scales by $N^2$, but can be scaled to log($N$) if the most state-of-the-art coulombic methods are used [87, 88]. Systems that model long timescales or large molecules dramatically increase CPU computation time because every interaction for every atom in the system must be calculated on every timestep.

Predicting protein denaturing and energy landscapes requires the repeated folding and unfolding of target proteins. Given that a single protein folding event occurs on the order of 50-100 $\mu$s, an all-atom model requires both too many atoms and too long timescales to reasonably simulate denaturing thermodynamics [89]. Only a handful of studies have simulated protein folding events.

In one all-atom study, Freddolino et al. simulated a partial folding pathway for a protein 34 amino acids long. This protein isolates a WW domain and looked at how this tertiary structure partially re-folds by starting with a partially unfolded structure and then running MD to observe how the structures came together. They found that resulting structures did not exactly match WT structure and likely established a meta-stable intermediary. However, this early all-atom research revealed the importance of energetic parallel tempering to allow a full exploration of conformation space. The study simulated only $\sim 10 \, \mu$s, but required millions of iteration steps and thousands of hours of CPU calculations [90].

One of the most significant all-atom protein studies came from the DESRES research group where they also simulated a fast-folding WW domain in the protein FiP35. However, this study goes beyond the previous study by exploring the entire folding pathway using replica exchange. Such an approach requires more computational resources, but is more accurate and can identify all structured states the protein will experience. In total, this study found five structured states for this very basic protein - one of which was the WT crystal structure. Moreover, they identified the
pathway FiP35 takes to go from completely denatured to fully structured protein. To recreate 1 ms of folding behavior, the computations demanded by this study required three months of simulation time using a supercomputer built for the express purpose of running MD simulation [91].

Presently, little simulation research explores unnatural amino acid effects on protein folding mechanics. Simulation work that has been done on uAA use an all-atom model with a quantum mechanical force field (QMFF). This type of force field requires complex QMFF computations for every atom on a timescale short enough to resolve hydrogen bonds, which is on the orders of 0.1 fs per iteration. These models are very computationally intensive, but necessary because unnatural amino acid crystal structures are only occasionally studied [92–95] and simulated interactions are unpredictable [96, 97]. Huang and Roux developed the General automated atomic model parameterization (GAAMP) method that uses QMFF to resolve complex protein charges and structures. They applied this method to uAA and found that the simulations closely follow experimental bond, angle and dihedral structural properties [98, 99]. These high accuracy, all-atom simulations are great for short time-scale analysis, but for large protein analysis, temperature-dependent thermodynamic calculations, and denaturing pathway analysis, these simulations require significantly longer time-scales. Because computation times are so long, research exploring uAA effects on protein structure and folding pathways have not been directly studied.
3.6 Coarse-Grain Simulation

Coarse-grain models utilize monofunctional molecular structures to define larger resolution computation sites. Examples include individual molecules, repeating polymeric units, or amino acid residues [100]. These larger sites decrease overall site count, decrease number of pair calculations and, consequently, result in fewer computations per timestep. Such simulations can model large systems over longer time periods, generally reaching between $10^3$ and $10^7$ CPU-time speed-up when compared to all-atom models [101]. However, by using a coarse-grain computation site, the model sacrifices resolution and a quantum mechanical basis. Consequently, accuracy comes from model design. A good model captures enough system detail that all relevant physics and interactions are fully captured without loss physical relevance. The model must then be validated against experimental data [101].

Coarse-grain work in this dissertation centers on the coarse-grain model proposed by Karanicolas and Brooks [102]. In this model, each amino acid residue is replaced by a single interacting site centered on the C$_\alpha$ carbon, and all site interactions follow a CHARMM-like intermolecular FF. Amino acid bonded parameters are determined from a suite of protein crystal structures from the protein data bank (PDB) [102]. This is done by setting the target’s primary sequence bonded interactions (bond, angle, dihedral) to average values from the suite of proteins. Secondary and tertiary structure are maintained using native contact (NC) interactions between sites [103]. These are interactions between sites deemed necessary for native structure as determined by previous crystallization experiments found in the PDB. NC interactions follow a Lennard-Jones-like interaction profile in implicit solvent, and parameters are based on work by Miyazawa and Jernigan [104].

Studies using this model have explored many varied topics in proteomics as varied as protein-nanoparticle interactions, solvent-accessibility surface area calculations, and enzyme activity [105–109]. Important studies using this model have shown unique folding pathways for proteins. These are used to identify key structures for a protein to maintain structure [110, 111]. Studies using this model in protein device research has primarily focused on proteins tethered to a surface, and will now be explained in more depth.
3.6.1 Simulations Reveal Device Effects on Protein Stability

Molecular simulations have helped explain site-dependent stabilization motifs when functionalizing using the tethering method. These simulations reveal which sites will potentially provide more stability and more active configuration than others [40, 112]. Wei et al. performed a case study on a suite of alpha-bundled protein to look for possible tethering configurations that eliminated the destabilizing intermediates and the possible causes. Sites that most significantly stabilize the protein hold the protein in a configuration where the surface physically prevents the protein from forming intermediate structures. Thus, the surface increase the entropic cost of unfolding and biases the folded protein state (Eqn. 3.1) [43].

Figure 3.4: The surface introduces additional atomic forces that interact with the protein and diminish device performance.

Another surface factor that strongly affects stability is surface hydrophobicity. Because proteins are primarily hydrophobic, they readily collapse onto hydrophobic surfaces. Wei and Knotts developed an experimentally-validated coarse-grain model that incorporates the hydrophobicity between the surface and each individual AA in the protein [113]. This hydrophobically-interacting surface with amino acids (HISAA) introduces enthalpic protein-surface interactions.
Their inclusion creates new ways for the protein to denature, and tethering to certain sites can significantly destabilize the protein [114].

As the surface becomes more hydrophobic, the hydrophobic residues within a protein will interact more with the surface. Conversely, as the surface becomes hydrophilic, interaction behavior should more closely resemble bulk phase chemistry. Simulations predict appropriate responses to surfaces across a full range of hydrophobicities. These simulations reveal a deep adsorptive energy well as hydrophobic AAs encounter the surface. And, as the protein approached both types of surfaces, a solvent exclusion barrier is present. The presence of these interactions suggest this model appropriately models surface effects in a coarse-grain simulation, and thus explores the protein’s previously-unknown enthalpic denaturing pathways [114].

Another factor found through simulation that affects activity is the tether length. Studies exploring this factor show protein stability increases with increasing tether length. The longer the tether, the less the protein interacts with the surface. Thus, protein structure approaches bulk solution configurations and stability [115]. Such systems decrease the number of new intermediates and maintain a uniform microarray. Large proteins also provide the same benefit by limiting surface interactions on the active site because the active site is far away from the surface and thus not subject to surface interactions [115, 116].

A study by Bush and Knotts, using this model, observes the energy landscape generated by hydrophobic/hydrophilic surfaces. They concluded antibodies readily denature onto all types of surfaces, but hydrophobic surfaces destabilized proteins the most. It also provided evidence that, for large proteins, stability is more a function of surface behavior than tether site location [116].

The literature described thus far has all been done to determine the factors affecting the stability of a protein being tethered to a surface. In many protein-based devices, this tethered protein must then binds to another molecule. Recent research has focused on understanding how surface tethering affects protein-protein binding. Liu et al. used Monte Carlo with molecular dynamic simulation and an all-atom model to observe adsorption orientation of an antibody fragment (Fab) onto immunosensors (the type of protein device used in ELISA assays). The small 56-residue protein was placed (“head-on” or “lying-on”) onto a charged surface and ran the simulation until an energetically optimal configuration was found. The study extrapolated their results to concluded that protein activity is maximized when the protein forms uniform monolayers, with the greatest
activity arising from protein in a more upright configuration as seen in Figure 3.4. This trend suggests activity correlates with the protein’s ability keep the active site available to binding agents when adsorbed [117]. While this study did not pose a solution on how to uniformly adsorb proteins or introduce time-dependent denaturing, it does suggest that if an antibody’s active site is free to interact with the bulk solution, it optimizes activity. However, when lying down, surface interference results in a deformed active site.

Bush and Knotts further explore surface-protein-protein interactions by specifically looking at lysozyme binding to a surface-bound Fab. Variables include upright or flat antibody Fab orientations attached to a neutral, hydrophobic or hydrophilic surface. Upright configurations on hydrophilic surfaces produce the best free energy landscape when only varying lysozyme-Fab and lysozyme-surface distances [118]. This optimal configuration most resembles an antibody in bulk solution and places the critical active site furthest away from the surface. Such a conclusion suggests surface entropic benefits do not contribute to protein stability as much as enthalpic destabilization effects for large, antibody-sized, proteins.

Much of the work described thus far in this section has focused on coarse-grain work when simulating tethered protein devices. Currently, PEGylated protein devices have not been simulated before using a Go-model. Generating a Go-model that includes PEG has to overcome parameterization hurdles - these follow a similar pattern established for tethered devices. First, a non-reactive model will reveal the purely-entropic contribution of PEG and establish how such a system can be implemented in simulation. Once this purely-entropic model has been established and compared to experiments, a coarse-grained PEG interaction model can then be parameterized. However, this latter step is outside the scope of this dissertation and will not be addressed further.

Two important assumptions from the protein device literature is examined in this dissertation. These are: (1) effects of uAA mutation are not included in any device analysis, and (2) only sites located on loop regions of a protein have been tested for device stability. The former assumption is made due to a lack of coarse-grain parameters and the belief that mutation effects are negligible when compared to the effect of the additional functional group [43]. The latter assumption was done to reduce the number of replicates, and only loop regions were chose because it was believed these are the most tolerant to the functionalization process [114]. As this technology becomes more established, integrating in silico screening into experimental design requires these
assumptions to be challenged. One of the main purposes of this dissertation is to examine the validity of these assumptions and revise current methods for more accurate computer modeling and screening practices.

3.7 Summary of Literature

The two types of protein-based technologies explored in this literature review focus on PEGylated proteins and proteins tethered to surfaces. Current methods used for building these devices include non-specific adsorption, N- and C-terminal attachment, non-specific AA attachment, and PRECISE attachment. This latter method has the greatest potential to produce highly-specific and uniform protein devices because it utilizes uAA mutation and the highly specific “click” reaction. These milestones in experimental control have enabled protein device creation linked to any part of a protein’s solvent-accessible surface. Because functionalizing at any residue on a protein’s surface is possible, determining where to attach becomes more challenging to determine. Traditional experimental spectroscopy methods do not allow viewing of these devices and experimental trial-and-error is costly, so focus has shifted to analyzing these protein devices with computer simulation.

The two predominant methods for in silico observation of protein folding behavior use all-atom and coarse-grain simulation. All-atom models are the more accurate of the two, but because site density is high and timestep is short, simulations using these models are computationally demanding and only feasible using specialized hardware. Coarse-grain models are better suited for viewing protein folding behavior, but need a strong basis to be used in experimental design.

Studies that utilize computer models have proven to: (1) properly predict stability and activity, (2) optimize physical authenticity verses computational run-time, (3) produce reasonable denaturing pathways and (4) predict likely protein-surface system behavior. Literature also identifies factors that influence binding activity and what protein orientations typically promote activity. These findings suggest that better and more stable protein devices optimize the benefits of functionalization theory and minimize the new structure’s interactions with WT structure. More work is needed to include PEG into the coarse-grain model, and then identify good sites for experimental design independent of functionalization method. To accomplish the latter, uAA mutation effects need to be included, and screening needs to sample more functionalization sites on the protein.
CHAPTER 4. METHODS

4.1 All-Atom Model Equations

The majority of biochemical all-atom force fields use similar governing equations to determine behavior. For the all-atom simulations in this work, the CHARMM FF is used because it is the most widely-used FF for protein structure simulation. This FF breaks-down interatomic interactions into separate and independent contributions. The equations that define the CHARMM FF [119] assume that the potential energy of the system ($V$) can be split into bonded and nonbonded interactions according to:

$$V_{molecule} = V_{bonded} + V_{nonbonded}$$  \hspace{1cm} (4.1)

where nonbonded terms ($V_{nonbonded}$) are composed of contributions for coulombic ($V_{charge}$) and external ($V_{Lennard–Jones}$) interactions:

$$V_{charge} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$ \hspace{1cm} (4.2)

$$V_{Lennard–Jones} = \varepsilon_{ij} \left[ \left( \frac{R_{min,ij}}{r_{ij}} \right)^{12} - \left( \frac{R_{min,ij}}{r_{ij}} \right)^6 \right]$$ \hspace{1cm} (4.3)

the charge terms $q_i$ and $q_j$ are the partial atomic charges for each atom in the charge pair, separated by $r_{ij}$ distance. The $4\pi\varepsilon_0$ term is the system dielectric constant. The Lennard-Jones nonbonded interactions combine parameters from each participating atom using the Lorentz-Berthelodt combination rules. The value $\varepsilon_{ij}$ is the geometric mean of atomic parameters $\varepsilon_i$ and $\varepsilon_j$, and $R_{min,ij}$ is the arithmetic mean of atomic parameters $R_{min,i}$ and $R_{min,j}$.

Bonded terms ($V_{bonded}$) include contributions from intramolecular potentials:
\[ V_{\text{bonded}} = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{dihedrals}} + V_{\text{improper}} + V_{\text{Urey-Bradley}} \] (4.4)

and

\[ V_{\text{bond}}(b) = K_b(b - b_0)^2 \] (4.5)

\[ V_{\text{angle}}(\theta) = K_\theta(\theta - \theta_0)^2 \] (4.6)

\[ V_{\text{dihedral}}(\phi) = K_\phi (1 + \cos(n\phi - d)) \] (4.7)

\[ V_{\text{improper}}(\omega) = K_\omega(\omega - \omega_0)^2 \] (4.8)

\[ V_{\text{Urey-Bradley}}(S) = K_S(S - S_0)^2 \] (4.9)

where parameters \( K_b, K_\theta, K_\phi, K_\omega, \) and \( K_S \) are force constants and all parameters denoted with a 0 subscript are the Hooke’s law equilibrium parameters for their respective potential. The dihedral potential \( V_{\text{dihedral}} \) uses a cosine expansion with a multiplicity \( n \) and phase shift \( d \) to model the periodic nature of the dihedral interaction.

An important assumption about the molecular motion described by \( V_{\text{bonded}} \) is that each contribution is independent from all others (e.g. movement in angle space does not affect dihedral potential) [63]. This assumption was examined in detail in the atomic FF review article by Hagler et al. [80]. This work concluded that the CHARMM FF does well simulating bond and angle potentials, but torsion calculations do not approach the accuracy of other FFs, nor does the model account for anharmonicity and coupling between internal deformations. Coupling is defined to mean changes in one parameter (e.g. angle potential) results in a change in energy of another bonded contribution (e.g. dihedral potential) – which is directly opposite the assumption of independence. For simulations of molecules composed of canonical biological moieties, such as those
in the standard CHARMM force field, the lack of explicit treatment of anharmonicity and coupling produces reasonably accurate simulations while simplifying the computation. [81].

4.2 Coarse-Grain Model Equations

Generating a coarse-grain model for protein devices requires unique parameters for the protein and the attached functional group. Because these models were created separately, they are presented in separate sections. A coarse-grain PEGylation model is not presented here because such a model did not exist prior to work done in this dissertation. A full analysis of the PEG model, parameterization process, and predictive capability is outlined in subsequent chapters.

4.2.1 Go-Model

The Go model is based on already-defined proteins structures submitted to the protein data bank. These PDB structures are used as a basis for all bonded interactions (bond, angle and dihedral). These interactions are based on the CHARMM set of equations (Eqn. 4.4), but applied to AA sites rather then atomic sites. PDB structures also inform nonbonded interactions, but only when AAs participate in a native contact pair [102, 103]. In recent years, NC pair potentials have improved upon the implicit water model to include disassociation/association penalty between interacting sites. This is done to accurately model water displacement in solution. The disassociation/association penalty takes the form of a Lennard-Jones potential with an additional energy barrier just past σ and more accurately simulates protein tertiary structure [42]. \( V_{NC} \) is calculated as:

\[
V_{NC} = \varepsilon_{ij} \left[ 13 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 18 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{10} + 4 \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]
\] (4.10)

where \( r_{ij} \) is the distance between residues \( i \) and \( j \) in the NC pair. The interaction strength (\( \varepsilon_{ij} \)) between the NC residues is parameterized using the interaction energies defined by Miyazawa and Jernigan [104], but scaled based on experimental benchmarks [102]. The \( \sigma_{ij} \) parameter is the length parameter for the NC pair [102]. The modified final term simulates the water displacement barrier for native contact interactions. The addition of the water barrier term has proven to more closely simulate protein folding and unfolding behavior in bulk solution. The value of each
parameter was derived using a top-down approach by comparing simulation thermodynamics to experimental data [42]. All other nonbonded interactions follow a purely-repulsive model.

The coarse-grain protein model input files, including the parameters and coordinates, were obtained using the MMTSB Go Model builder (https://mmtsb.org) [102, 103]. All coarse-grains simulations in this work makes the assumption that uAA mutation effects are negligible compared to protein device effects. As such, the WT protein structure is used when creating protein device models. To functionalize the protein in simulation, a harmonic restraint links the protein to the device at the functionalization residue [115].

4.2.2 Surface Model

All surface-tethered protein devices use the Go-like surface model proposed by Wei and Knotts [113]. This HISAA surface is compatible with Go-model systems and introduces a hydrophobic surface potential that can interact with AA residues. The model is based on Lenard-Jones potentials used in the previously described Go-model, where the surface potential \( V_{\text{surface}} \) describes amino acid interactions with the surface using a modified form of the 12-10-6 Lennard-Jones potential:

\[
V_{\text{surface}} = \sum_i \left\{ \pi \rho \sigma_i^3 \varepsilon_i \left[ \theta_1 \left( \frac{\sigma_i}{r_{is}} \right)^9 - \theta_2 \left( \frac{\sigma_i}{r_{is}} \right)^7 + \theta_3 \left( \frac{\sigma_i}{r_{is}} \right)^3 - (\theta_s \chi_s + \theta_p \chi_p,i) \right] \right\}
\]

where the summation is over all sites in the protein. The \( \rho \) term is 1.0 Å\(^{-3} \) and \( \sigma_i \) and \( \varepsilon_i \) are the amino acid’s nonbonded Lennard-Jones parameters from the Go-model and \( r_{is} \) is the distance of the particle from the surface. The first three term in Equation 4.11 are obtained by integrating over the radial and angular dependencies and result in a potential that is only dependent on the distance of the particle from the surface \( (r_{is}) \). The last two terms define the hydrophobic interaction of the surface and residue \( i \) in the protein. The \( \chi_s \) and \( \theta_i \) parameters account for the hydrophobic contributions of the surface and the \( \chi_p \) and \( \theta_p \) parameters account for the hydrophobic contributions of the amino acid.
4.3 Replica Exchange

A protein’s conformational potential energy and molecular kinetic energy add together to become its structural total energy. Over the course of a simulation, the protein will explore many different configurations. Some configurations are more stable than others, with the most stable resulting in the lowest total energies. This relationship is plotted as an energy landscape. As the protein explores the landscape, it experiences multiple stable (low energy) and transitory (high energy) configurations before reaching a fully denatured state (Figure 4.1). Identifying this denaturing pathway allows for development of denaturing heuristics. For example, if identifying optimal tether site locations on a protein devices, a good tether location orients the protein so that the surface sterically blocks denaturing transition states making a certain transition state require more energy to form. By adding a tether, mutating a site, or changing functionalization location, the protein’s energy landscape changes. The goal is to create a landscape with very low stable states and very high transition states. Thus the protein remains stable and active for longer and at higher temperatures [86].

Figure 4.1: Example of a theoretical energy landscape. The line indicates the energy in a system at some temperature. N, I and D are the theoretical: Naturally occurring (folded), Intermediate and Denatured energetic states, respectively. A protein in one of the energy minima configuration cannot overcome the activation energy needed to enter the other configurations. Larger proteins may have many intermediate folding states.

Simple MD simulations are ill-equipped examine protein folding because these simulations only explore a portion of a protein’s energy landscape. Figure 4.1 shows a hypothetical energy
landscape and can be used as an example to explain the limitations of simple MD simulations. The upper-limit line indicates the max thermal energy available to the system at some temperature. If only using a purely MD simulation, sampled protein configurations only exist on this line and only in the well where the simulation started. Notice the peaks for the protein transition states exceed the energy of the system (the black line). The system cannot overcome this transition and thus will not cross into the other energy wells. This results in unsampled configurations and an inability to accurately quantify protein stability.

The inability of the simulation to sample does not mean they are not plausible. The problem is that the event is not observed over the course of the simulation because of its low probability of occurring. Specifically, constant-temperature MD methods will not capture these configurations with sufficient frequency to obtain reliable thermodynamic data. Proper stability analysis requires sampling from all configurations – all the folded and all the unfolded states. Due to the rough energy landscape of protein folding and its associated barriers, MD simulations will easily become trapped in local energy minima.

The simulation technique used to assess the stability of proteins in different environments is called Replica Exchange [120–122]. This technique fully explores rough energy landscapes, from the fully folded to the completely unfolded states. This then allows calculations of any thermodynamic property of interest as a function of temperature by virtue of improved simulation sampling. Quantifiable protein stability analysis is then possible and can reveal which protein configurations are more stable than others [86].

To perform replica exchange, multiple MD simulations are established at unique temperatures called a simulation box. Normal MD simulations proceed in each box. The key feature of replica exchange is that periodically two randomly chosen boxes swap molecular coordinates. Before the simulations continues, the swap is tested against the Metropolis acceptance criteria [123, 124], where the greater the temperature difference between the two swapped boxes the less likely the swap will be accepted. Whether a box swaps or not, the protein’s configuration is then allowed to adopt a configuration minima. These minima are recorded and swapping is repeated. Over the simulation’s lifetime, all plausible configurations are explored and energetic local and global minima can be determined.
4.4 WHAM and MBAR

Replica exchange post-processing takes all the individual structural energies and puts them in a format where we can directly compare experimentation with simulation. Temperature-dependent heat capacity is commonly used because it captures information regarding stable-state and transition-state protein structures. This value can be directly measured in-lab and can be easily calculated from simulation. In practice simulation and experimentation will not produce the exact, quantitatively similar, heat capacity curves, but simulations will qualitatively predict structural perturbations.

The first method for determining computed heat capacity utilizes the weighted histogram analysis method (WHAM). This method takes all the configurations seen in each simulation box and constructs a structurally-weighted series of histograms that define the protein’s potential of mean force (PMF) across all temperature boxes. The differences between bins $i$ and $j$ are defined as:

$$
\Delta A_{ij} = -k_B T \ln \frac{Q_j}{Q_i}
$$

where $Q$ is the probability densities for $i$ and $j$, and is a function of the density of states $\Omega_i$. If we assume a protein state as a discrete bin ($B_i$), the binned density of states ($\Omega_i$) is then a function $B$ and the free energy ($A$).

$$
\Omega_i(q, \lambda_i) = B_i(q, \lambda_i) \exp \left[ \left( \sum_j \beta_i \lambda_{ji} U(q) \right) - \beta_i A_i \right]
$$

where $\lambda$ is the sets of available states for $q$ positions. The best estimate of the total density of states ($\Omega$) then becomes:

$$
\Omega = \sum_i \omega_i \Omega_i(q, \lambda_i)
$$

where $\omega$ values best minimize statistical noise. The best estimation of a bin value is then:

$$
\langle B_i \rangle = N_i \Omega \exp \left( \beta_i A_i - \beta \sum_j \lambda_{ji} U(q) \right)
$$
Thus, by iterating between free energy \( (A \text{ from Eqn. 4.12}) \) and density of states \( (\Omega \text{ from Eqn. 4.13}) \) allows for the determination of \( B_i \). With this information, the the differences in free energy between states can be calculated and thus determine the changes in heat capacity [125, 126].

If the bin width is reduced to zero, the basis for the MBAR algorithm is established [127]. The estimation of the free energy then becomes:

\[
A_i = -\beta^{-1} \ln \sum_j \sum_n \frac{\exp[-\beta U_i]}{\sum_k N_k \exp[\beta A_k - \beta U_k]},
\]

(4.16)

where \( K \times K \) is the set of weighting functions \( k \). This free energy calculation is the most accurate predictor of free energy from replica exchange simulation [128]. The MBAR algorithm is the method used for all replica exchange post-processing in this dissertation to predict structural energy.

### 4.5 Heat Capacity

As explained in the Chapter 1, achieving the goals of this work requires understanding how functionalization affects the protein compared to the wild type. One useful metric in this regard is the heat capacity. It can be calculated in simulation [113] and can also be obtained using experimental methods such as circular dichroism or thermal shift assays. Protein heat capacity curves reveal folding states and identifies “transition” temperatures in protein.

The heat capacity \( (C) \) is related to potential energy of the simulation \( (U) \) according to:

\[
C(T) = \frac{\langle U^2 \rangle_T - \langle U \rangle_T^2}{RT^2}
\]

(4.17)

where \( R \) is the gas constant, \( T \) is system temperature, and the angle brackets indicate ensemble average values for the given property. For an arbitrary property, \( X \), the ensemble average is given by:

\[
\langle X \rangle_T = \frac{\sum U X(U) \Omega(U) e^{-\beta U}}{\sum U \Omega(U) e^{-\beta U}}, \text{ where } \beta = \frac{1}{k_B T}
\]

(4.18)
where \( k_B \) is the Boltzmann’s constant and \( \Omega(U) \) is the density of states which is determined from the replica exchange simulation data using the MBAR approach as implemented in pyMBAR [125–128] and was just described.

Peaks indicate transition states and the final peak is termed the melting temperature. Multi-state proteins contain at least three folded states (folded, intermediate(s), and unfolded). When some treatment or change to protein structure, like protein device creation, reduces the number of peaks in a heat capacity curve, the protein stays in the more-folded state for longer. A more-folded protein retains a higher activity and is considered more stable. For proteins which fold through two-state mechanisms, only the Tm peak is present. Thus, analysis on these proteins focuses on shifts in the melting temperature.

Determining the effectiveness of a particular functionalization site uses a protein’s melting temperatures as its metric. This information is contained in protein heat capacity curves as a peak in the heat capacity. Specifically, if the melting temperature of the treatment is higher than the control (WT) then the functionalization of the treatment stabilized the protein. If the Tm of the treatment is less than that of the WT, then functionalization decreased stability. For ease in communication, the results presented below are in terms of \( \Delta T_m = T_{m_{\text{treatment}}} - T_{m_{\text{control}}} \). Defined in this way, positive \( \Delta T_m \) values mean the treatment stabilized the protein and negative values means the treatment adversely affected the stability relative to the WT.

Because the coarse-grain model removes many degrees of freedom, the quantitative changes in Tm (or any other thermodynamic property) mean little. However, the model adequately captures the relative improvements. Thus, to determine the best candidates for functionalization, each treatment is ranked according to its \( \Delta T_m \) value. In other words, a simulation result that indicates \( \Delta T_m = +9 \) K will not necessarily correlate to a +9 K shift in melting temperature in experiments. Rather, a \( \Delta T_m = +9 \) K would be expected to be a better candidate for functionalization than a treatment which produced \( \Delta T_m = +2 \) K and much better than a treatment with \( \Delta T_m = -5 \) K. The discrepancy between simulation and experiment is of little consequence for the purposes of this work. The goal is to create a screen that provides a handful of reasonable candidates for experimental functionalization that would then be tested for efficacy.
CHAPTER 5. DEVELOPMENT OF PEG COARSE-GRAIN MODEL

5.1 Introduction

The majority of simulated protein device research centers on surface chemistry and coarse-grain models that incorporate surface effects into protein folding pathways [113, 114]. However, PEGylated proteins pre-date microarrays and are currently being used in therapeutic treatments. Despite this, PEG has yet to be incorporated into the Go-model for coarse-grained protein simulation. Similar to research done in microarrays, PEGylation affects protein folding pathways [8, 9]. Thus, attachment locations on a protein may be in sub-optimal locations such that conjugation would hinder protein stability or hinder important protein-protein interactions [30].

Such an endeavor to incorporate PEG into the previously-introduced Go-model would require coarse-graining the PEG polymer and introducing parameters to explain its interaction with the protein. The predominant theory behind PEGylation suggests PEG increases structured protein stability by changing the unfolded state entropy. Work by Lawrence et al. shows all stabilizing PEG chains exhibit an overall entropic benefit to $\Delta S$ (Eqn. 3.1), but chain interactions with the local structure minimally enthalpically hinder stability [129, 130]. In contrast to this, the competing theory clams long PEG chains enthalpically interact across the protein’s surface promoting greater stability. While this theory has been loosing strength, Pandey et al. shows that PEG can produce localized stability for certain secondary structures [131]. The final hypothesis states that the PEG strand wraps around the protein to sterically hold the protein together, though this theory lacks adequate experimental support [3, 23].

The purpose of this section is to create a coarse-grain model for PEGylated protein simulation and then compare to experiment. PEGylated protein design will examine the following factors for protein stability: (1) PEG attachment location, (2) PEG size, and (3) transferability of well-performing surface tethering sites into PEGylation site efficiency.
5.2 Methods

Experimental method and analysis was performed by Kristen Wilding using the PRECISE method in cell-free protein synthesis (CFPS) [8, 9, 57]. Specificity was achieved with DNA re-coding to mutate-in the pAz uAA. A full detail of the processes and procedures used are reported in the article where this research was published [73] and is not further explained in this dissertation. Only the results from this experimental work is shown and is used as a comparison to the simulation findings.

Lysozyme (Lyz) was chosen as the model protein because it is easily created, mutated, and purified in CFPS experiments. It has also been shown to be a good in silico predictor of protein device behavior when used in protein tethering studies [114]. The three-dimensional structure of T4 Lyz, needed for the Go-model, was obtained from PDB ID 2LZM [132]. All simulations were run in the Knotts Simulation Software (KSS) [42, 112] using the Go-model for proteins and replica exchange algorithm. The replica exchange set-up contained 66 boxes over a range of temperatures with a 1.5 K separation (∆K) between boxes. These ranges surrounded the expected melting temperatures for each of the PEGylated variants (345-366 K for 5 kDa conjugates, 327-348 K for WT and 20 kDa conjugates).

Figure 5.1: SASA for each of the residues mutated for PRECISE experimentation. These residues are also plotted on the PDB structure and colored to according to their secondary structure. Residues depicted in green are part of unstructured loops, while structured sites are represented in pink or orange (pink for beta sheet, orange for alpha helix). The residues surrounding the insertions site are shown in yellow, and are also part of an unstructured loop.
Sites selected for simulation and experimentation were: K16, S44, N53, L91, and K135 (See Figure 5.1). These locations were selected for their relative solvent-accessible surface area value (SASA) [133], location on loop regions [43], and success in previous protein device screens using tethered proteins [114]. The heuristic that dictated these a priori selections was published in a study by Wei et al. and states that stability is found when a functionalization residue is (a) on a long loop and not adjacent to secondary structures or (b) on the vertex of a U-shaped or W-shaped loop [43, 113].

PEGylation site performance is determined using heat-capacity comparisons that was post-processed using the MBAR algorithm as described in the Methods chapter. Because T4 lyz is a multi-state protein (ex. lysozyme contains a folded, intermediate and denatured tertiary structure), stability is achieved by either removing the intermediate folded state or producing a positive $\Delta$Tm value. No PEGylated treatment decreased the number of folded states, and thus only an analysis on $\Delta$Tm is explored.

5.3 PEG Go-Model

The purpose of developing a Go-like PEG model is to enable an in silico tool to narrow candidate PEGylation treatments. This would be useful in further reducing the costs of PEGylated protein design by identifying poor functionalization locations and thus reduce the number of sites for experimental testing. The model in this work simplifies the PEG polymer (Figure 1.1) to be a series of identical PEG monomers. Each interacting site centers on the monomeric oxygen and is 1.48 Å in radius, or the length of the C-O bond. Other relevant parameters, including all Bonded and Lennard-Jones parameters, were chosen to resemble those of the peptide bonds within the protein.

The primary motivation for the PEG model in this work is to isolate the entropic benefit of PEG functionalization. The entropy-only approach has been successfully applied to immobilized proteins [38–40] and is the easiest first-step for parameterizing PEG for coarse-grain simulation. To do this, the PEG interactions were parameterized to be purely repulsive within the governing Go-model nonbonded interaction equation (Eqn 4.10). Thus the nonbonded $\varepsilon$ value is the same value given to all non-NC amino acids interactions in the Go-model, and has been shown to produce the purely repulsive potential [102]. Specifically, $\varepsilon_i = 0.00132$ kcal mol$^{-1}$ and $\sigma_i = 3.831$ Å for
all PEG-PEG and PEG-protein interactions. Therefore, simulated changes in melting temperature correspond to the entropic contribution of the PEG and exclude enthalpic PEG-PEG and PEG-protein interactions (see Eqn. 3.1).

To attach PEG to the protein, a simple harmonic restraint is made between the PEG and WT residue. Notably, pAz is not mutated into the primary sequence to replace the WT AA at the functionalization site (which is done in experiment). Currently pAz cannot be included due to insufficient data to parametrize pAz for inclusion in this coarse-grain model. It is assumed here that the mutation effects are insignificant compared to the effect the PEG has on the protein structure.

5.4 Results

Experimental ΔTm values were determined using thermal shift assays, and PEGylated melting temperatures were determined for the five lysozyme variants. Figure 5.2 shows the experimental change in melting temperature that results from functionalizing on sites 53, 91, and 125. Also shown is the simulated ΔTm values. For the 5 kDa variants, simulated ΔTm values are noticeably higher than their experimental counterparts. This is likely due to the purely-repulsive assumption which means enthalpic destabilization interactions are not accounted for in simulation. A similar phenomenon was observed in protein-surface simulations, where enthalpic destabilization was the contributing difference between purely-repulsive and interactive surfaces. [114]. More research is needed to confirm these observations.

Figure 5.2: Lysozyme stability after PEGylation measured by how functionalization affects melting temperature (ΔTm). Panel A ΔTm values from 5 kDa PEG variants. Panel B ΔTm values from 20 kDa PEG variants.
While the quantitative results from 5.2 indicate simulation can over-predict PEG influence on protein structure, there is evidence that simulation does a good job as predicting qualitative performance. From Figure 5.2, notice that simulation ranks site 91 to significantly under-perform when compared to the other two variants. This result is shared with the experimental findings, and suggests that simulation can properly rank relative site performance. Figure 5.3 compares the simulation-predicted $\Delta T_{m}$ to the experimentally determined values. Plotted this way, the ability of simulation to rank site performance is tested against the experimental values. From this figure, it is apparent that both simulation and experiment show site K16 performing the best and site L91 performing the worst. The remaining functionalized sites fall on a spectrum between these extrema.

![Figure 5.3: Comparison of the coarse-grain simulation predicted impact of PEGylation on thermal stability vs the experimentally determined stability change.](image)

Figure 5.3: Comparison of the coarse-grain simulation predicted impact of PEGylation on thermal stability vs the experimentally determined stability change. The change in thermal stability following PEGylation ($\Delta T_{m}$) is reported for 20 kDa PEG conjugates (gray) and 5 kDa PEG conjugates (orange). The average and standard error of $\Delta T_{m}$ is calculated from simulation ($n \geq 10$) and protein thermal shift assays ($n \geq 3$) are shown. Dashed lines represent a linear regression, with the orange line fit to the 5 kDa PEG data and the gray line fit to the 20 kDa PEG data. The corresponding $R^2$ values are also reported.

To better quantify the ability of the simulation to predict experimental performance, a trend-line on Figure 5.3 correlates the simulation prediction to the experimental results. A positive correlation in the data is evidence that the simulation was able to correctly predict the trend and rank each site for PEGylation stability. This performance can be quantified with the linear regression
$R^2$ value. For both the 5 kDa and 20 kDa variants, the $R^2$ value exceeds 0.94. This high correlation between simulation and experiment shows the entropy-based PEGylated model accurately screened for well and poor performing sites on Lys.

While the simulation did a poor job at predicting the quantitative change in melting temperature, it accurately predicted the relative change in stability after PEGylation. As such, the entropy-only simulation could be a powerful tool for informing design of a screening pool. However, adapting the simulation to include enthalpic effects of PEG-protein interactions may further improve the predictive capacity of the model. Previous experimental results have suggested that the enthalpic interactions of PEGylation can be destabilizing and could account for the over-prediction of simulation $\Delta T_m$ by the entropy-only model.

Interestingly, the simulation seems to provide better quantitative estimates of $\Delta T_m$ for the 20 kDa conjugates. These long-PEG simulations show that the PEG forms its own body separate from the protein and doesn’t surround the protein structure in any way. We hypothesize that this is due to a more dominant impact of entropy with the larger polymer. The two bodies form two centers of mass that are connected by the uAA covalent tether. This system is more entropy-controlled due to the separation between the protein-PEG complexes. Protein instability, then, focuses on the two centers of mass pulling on each other and thus unraveling the structured protein at the functionalization site. By incorporating PEG-PEG and PEG-protein interaction parameters, the model will be better able to predict the impact of PEGylation when different lengths of PEG are used.

5.5 Conclusions

Overall, our use of in silico coarse grain simulations indicate that valuable insight can be gained by examining the effects of PEGylated protein devices by using a simple, entropy-based model. This model could be incorporated as part of the screening process to create a hybrid simulation- and experimental-based screening method in which all of the sites on a target protein that are available to conjugation are first screened with the coarse grain model.

The simulations in this study ran to completion in under 11 hr, potentially enabling simultaneous evaluation of all potential conjugation sites on a protein in less than 24 hr. Sites that are predicted by the simulation to be least stabilizing could then be eliminated from the subsequent
experimental screen in order to conserve resources by reducing the number of candidate conjugation sites. Further improvements in the model include parametrization of pAz and PEG-protein enthalpic interactions. Doing so may enhance the capacity of the model for quantitative $\Delta T_m$ prediction.
CHAPTER 6. TEM-1 SCREEN

6.1 Introduction

With a functional PEGylation model that was proposed in the previous chapter, focuses now switches to both protein devices introduced in this dissertation. Specifically, analysis examines how these significantly different functionalization methods affect protein structure if attachment occurs on the same site. Figure 1.1 shows there is little in common between these functionalization methods except attachment location. Thus, such a case study will reveal many factors that affect protein folding and protein device chemistry.

Previous work that screens functionalization sites for experimental viability, though promising, only tested a limited number of sites that were found in loop regions in the protein and neglect other options in secondary structure elements. Such an approach was taken based on a common thought among the protein engineering community that loops are flexible and can therefore accommodate functionalization without disrupting the structure of the molecule (and thus function) compared to functionalizing at residues within secondary structure regions of the molecule [43, 113]. However, these heuristics have never been subject to rigorous testing, and recent work has shown that good sites for functionalization can be found in regions involved with secondary structure [73].

The working hypothesis of the current research effort is that the effects of functionalization on protein stability and function can be determined from molecular simulation. The aims of this specific work are to 1) examine the affects of tethering at all possible residues in the molecule regardless of its position in secondary structural elements, and 2) move beyond the quintessential model protein lysozyme with the goal of creating better design heuristics that can be used across multiple classes of proteins. The ultimate hope is that better design heuristics be used to reduce the time and expense required to develop a protein device for a specific application, and this work is the next step towards this end. The protein studied here expands on previous work because it comes from a different class of proteins than those previously studied and is industrially and medically
relevant. The latter reason is particularly important in terms of elevating the practical importance of the work beyond model proteins.

6.2 Models

6.2.1 Protein

The protein used in this work is \( \beta \)-lactimase (TEM-1) [Protein Data Bank ID: 1XPB] which has 263 residues [134]. TEM-1 is used in industry to degrade residual beta-lactam antibiotics in dairy products and therapeutically to stop allergic reactions of some people to beta-lactam antibiotics [135–137]. The CATH structural classification for TEM-1 is \( \text{class} = \alpha \beta \); \( \text{architecture} = 3\text{-layer(aba)} \) sandwich. TEM-1 contains 12 alpha helices and a 5-stranded beta sheet [138]. Figure 6.1A shows a cartoon representation of TEM-1 colored to identify different domains. Five domains are found in the protein: the Alpha I domain is composed of Helices 1, 11, and 12 (cyan), the Beta sheet domain is made up of the five beta strands (magenta), the Alpha II domain consists of Helices 2, 5–7, 9, and 10 (steel blue), the Loop I domain connects Helix 2 to Helix 5 (pink), and the Loop II domain connects Helix 7 to Helix 9 (dark purple). The active site is located in the pocket between the Beta and Alpha II domains and contains residues 70, 130, 216 and 234-237 [139]. Figure 6.1B depicts residues comprising the active site in red.

TEM-1 is modeled using the coarse-grain representation proposed by Karanicolous and Brooks [102, 103] and introduced previously. This model has been used to successfully examine protein stability, activity and folding pathways and has been shown to be both qualitatively and quantitatively accurate [73, 105–108, 115, 116, 118, 140].

The model used for the surface is the same hydrophobically interacting surface introduced previously [114] and is compatible with the protein model of Karanicolous and Brooks. This surface model, combined with the protein model of Karanicolous and Brooks, has been used to study protein surface interactions of lysozyme and antibody microarrays [113–116, 118, 140], but has never been used in conjunction with a PEGylation screen.

The PEG model is the same as was introduced in the previous chapter. By pairing this PEG model with the surface model, we can directly compare the effectiveness of sites, compare transferability of site effectiveness, and identify how each functionalization method uniquely affects
Figure 6.1: Panel A) Schematic representation of TEM-1. Domains are as follow: Alpha I domain in cyan, Beta domain in magenta, Alpha II domain in steel blue, Loop I domain in pink, and Loop II domain in dark purple. Panel B) Active site is identified in red.

protein structure. This study will provide a deeper understanding of protein folding and identify key factors for protein folding and stability.

6.3 Method

6.3.1 Simulation Protocols

As mentioned above, the goal of this screen is to sample every protein-surface accessible attachment site on TEM-1. The only criteria for exclusion from the screen was close proximity to the active site (see Figure 6.1B) or an experimental accessibility rating below 10% [133,141]. This left 89 experimentally-accessible sites to be screened for stability analysis upon functionalization.

A simulation treatment uses one of the functionlization methods (PEGylation or tethering) and attaches it to the protein at one of the 89 sites. For each treatment (89 sites x 2 functionalization methods), three independent replicates were performed using the canonical ensemble with 3 Nosé-Hoover thermostats. As a control, 5 replicates of wild-type (WT) TEM-1 was simulated.

The LAMMPS simulation software was used for all simulations in this work [142]. The replica exchange algorithm was used to simulate the systems and sufficiently sample the folded and unfolded states of the system, as has been done previously for similar studies. [120–122].
replica exchange temperature range, and the WT melting temperature of TEM-1 was in the approximate center of this range. This range spanned temperatures 150-398 K with a 4 K temperature difference between each box. Each simulation used a 3 fs timestep with 18 ns of equilibration and 90 ns of production.

For tethered treatments, the protein was attached to a hydrophilic surface which is the type typically found in protein devices. This was done by using a harmonic restraint between the functionalization site of the protein and a point on the surface. The length of the restraint was 15 Å to approximate the experimental attachment length. Prior to screening, an additional simulation rotates the protein onto the surface so the tethering attachment potential is minimized. This is done by freezing the protein native structure in its crystallized form and running MD to orient the protein.

The PEG was attached to the protein using a random walk algorithm that started at the attachment site and added additional sites until the desired PEG length was achieved. A PEG length of 20 kDa was used to be consistent with previous works [73]. This equates to 452 PEG monomers. Prior to screening, an additional simulation step is done to ensure the random walk algorithm does not create a high-energy PEG structure. This additional simulation fixes the protein native structure in its crystallized form while the PEG polymer is allowed to equilibrate freely.

6.3.2 Quantitative Folding Behavior

Improving protein activity is the primary goal for creating protein devices. Unfortunately, molecular dynamics simulation cannot directly investigate protein activity if classic models are used. It has been experimentally shown that changes in protein structure and thermal stability is the cause of decreased activity in mutant proteins [143], and various in silico methods predict the effects of these mutations on thermal stability [144–146]. Thus, it is assumed in this work that a protein’s thermal stability performance correlates with its protein activity. This assumption has been used previously to accurately predict protein device activity from protein device thermal stability data. [9, 73, 114].

Determining how functionalization of a particular site affects stability is done using the melting temperatures obtained from the heat capacity curves as described in the Methods section. To review, if the melting temperature of the treatment is higher than the control (WT) then the
functionalization of the treatment stabilized the protein. If the Tm of the treatment is less than that of the WT, then functionalization decreased stability. The results presented below are in terms of $\Delta Tm = Tm_{\text{treatment}} - Tm_{\text{control}}$. Defined in this way, positive values for $\Delta Tm$ mean the treatment stabilized the protein and negative values mean the treatment adversely affected the stability relative to the WT.

6.4 Results

6.4.1 Quantitative Site Performance

As explained previously, 89 total sites were tested for functionalization. In each case, three replicates were simulated, and the average heat capacity curve of these three replicates was determined. Figure 6.2 shows the results of 15 of the 89 tethered treatments. The red lines found at lower temperatures are the heat capacity curves for tether treatments with the five lowest melting temperatures. The blue curves at higher temperatures are for the 10 tether treatments with the highest melting temperature. The dashed black line indicates the Tm for the control (WT). Notice that all the treatment curves peak at temperatures below that of the WT. This indicates that tethering of TEM-1 to a hydrophilic surface will always destabilize the protein compared to the WT. However, there is significant spread (10's of Kelvins) from the highest to the lowest curves. This spread is useful for screening because it allows easy filtering of the potentials sites. Specifically, the 10 sites producing the blue curves are recommended as the best candidates for tethering and subsequent experimental validation.

Notice in Figure 6.2 that some of the poorly-performing sites exhibited multiple peaks in the heat capacity curve indicating that the folding mechanism has change and an intermediate has been introduced. Tethering has been shown in previous studies to change folding mechanisms to decrease the number of transitions, and this results in good candidates for functionalization [113]. Conversely, we see here that increasing the number of transitions only occurred in the least stable cases, indicating such tether points are poor candidates. From these and the previous results where the mechanism changes, it appears that increasing or decreasing transitions compared to the WT is a reliable predictor of whether a candidate is bad or good, respectively.
Figure 6.2: Averaged heat capacity curves for tethering TEM-1 to a hydrophilic surface. The dashed black line indicates the in silico WT Tm. Blue curves are the treatments with the top 10 highest Tm. The red curves are the treatments with the 5 lowest Tm.

Figure 6.3 is the heat capacity curves for 15 of the 89 PEGylation treatments. As with the curves for tethering (Figure 6.2), those plotted in red are treatments with the five lowest PEGylation melting temperatures, and those in blue the highest ten PEGylation melting temperatures. The WT Tm again appears in black. All PEGylation treatments in the screen peak above the WT Tm, and the spread is significantly less than the results from the tethering screen. Previous efforts in simulating PEGylated protein devices have observed spreads of similar degree can sufficiently separate poorly-performing sites from better candidates similar with good correlation to experiment [73], so the degree of spread compared to tethering (Figure 6.2) is unimportant for the aims of this work.

Tables 6.1 and 6.2 rank the best and worst performing sites based on their $\Delta Tm_{tether}$ and $\Delta Tm_{PEGylation}$ values averaged across the three replicates. These $\Delta Tm$ values correlate with the heat capacity peaks in Figures 6.2 and 6.3. Appendix A contains a comprehensive ranking of all treatments and their associated $\Delta Tm$ values with the performance of all 89 sites recorded in Tables 10.1 and 10.2. In addition to a rank, these tables also track primary and secondary structure attributes to identify possible patterns in the performance. These factors are: site location, amino acid type, experimental accessibility, and secondary structure.
Figure 6.3: Averaged heat capacity vs. temperature curves for the PEGylation treatments. The dashed black line indicates the in silico WT Tm. Blue curves are the treatments with the top 10 highest Tms. The red curves are the treatments with the 5 lowest Tms.

These results for accessibility and structure are important to creating accurate heuristics for rational design of protein devices. At present, these two metrics are commonly used to predict good locations for functionalization. Specifically, the current approach is to select sites with high accessibility and that are not in alpha or beta secondary structures. The results presented in Tables 6.1, 10.1, 6.2, and 10.2 indicate that such an approach will not be accurate and will produce many false positives. Of the ten best performing sites for both tethering and PEGylation, six are on non-loop regions and would have been ruled out based on current common practice. This means that a screen that only selected sites on loop structures produce sub-optimal recommendations. For example, Site 198 is on a loop structure and is the 10th best performing loop site for tethering and the 5th best site for PEGylation. Following current practice, such a site would have been recommended for further experiments due to its good performance in both screens. However, if all secondary structures are considered (like is done in this work) the ranking of Site 198 is significantly diminished. In the tethering screen it is ranked 28th, and for PEGylation it ranked 9th. A better site on TEM-1 that performed well in both screens would be Site 153. This site is located on alpha structure and performed 7th in the tethering screen and 1st in the PEGylation screen.

It would be advantageous, from a prediction perspective, if certain secondary structural elements produced more favorable functionalization sites than others, but the data in Tables 6.1 and 6.2 do not support this approach. The overall performance of secondary structures in this work did
not yield conclusive evidence suggesting sites on one secondary structure type performs differently from the others. All such structure analysis resulted in an equal number of good and poor performing sites, or, as is the case of beta secondary structures, not enough sites were functionalized to draw reasonable conclusions.

Table 6.1: Numeric ranking for tethering treatments. The top 10 and bottom 5 sites are presented. See the SI for the data on all 89 surface accessible sites.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>ΔTm (°)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>-8.000 ± 2.494</td>
<td>SER</td>
<td>53.6</td>
<td>Loop</td>
</tr>
<tr>
<td>2</td>
<td>88</td>
<td>-13.667 ± 1.247</td>
<td>GLN</td>
<td>71.1</td>
<td>Loop</td>
</tr>
<tr>
<td>3</td>
<td>192</td>
<td>-14.333 ± 1.414</td>
<td>LYS</td>
<td>35.5</td>
<td>alpha</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>-14.667 ± 2.055</td>
<td>VAL</td>
<td>56.6</td>
<td>alpha</td>
</tr>
<tr>
<td>5</td>
<td>227</td>
<td>-15.333 ± 1.414</td>
<td>ALA</td>
<td>94.4</td>
<td>Loop</td>
</tr>
<tr>
<td>6</td>
<td>92</td>
<td>-15.667 ± 2.055</td>
<td>GLY</td>
<td>66.1</td>
<td>Loop</td>
</tr>
<tr>
<td>7</td>
<td>153</td>
<td>-17.000 ± 2.625</td>
<td>HIS</td>
<td>54.4</td>
<td>alpha</td>
</tr>
<tr>
<td>8</td>
<td>154</td>
<td>-17.333 ± 0.000</td>
<td>ASN</td>
<td>94.9</td>
<td>alpha</td>
</tr>
<tr>
<td>9</td>
<td>281</td>
<td>-17.667 ± 1.886</td>
<td>GLU</td>
<td>58.4</td>
<td>alpha</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>-18.000 ± 1.633</td>
<td>LYS</td>
<td>51.5</td>
<td>alpha</td>
</tr>
<tr>
<td>85</td>
<td>230</td>
<td>-54.667 ± 3.859</td>
<td>PHE</td>
<td>12.5</td>
<td>beta</td>
</tr>
<tr>
<td>86</td>
<td>178</td>
<td>-55.000 ± 1.247</td>
<td>ARG</td>
<td>38.9</td>
<td>Loop</td>
</tr>
<tr>
<td>87</td>
<td>255</td>
<td>-58.333 ± 1.700</td>
<td>GLY</td>
<td>57.2</td>
<td>Loop</td>
</tr>
<tr>
<td>88</td>
<td>275</td>
<td>-60.333 ± 1.886</td>
<td>ARG</td>
<td>8.3</td>
<td>alpha</td>
</tr>
<tr>
<td>89</td>
<td>256</td>
<td>-94.00 ± 0.89</td>
<td>LYS</td>
<td>73.9</td>
<td>Loop</td>
</tr>
</tbody>
</table>
Table 6.2: Numeric ranking for PEGylation treatments. The top 10 and bottom 5 sites are presented. See the SI for the data on all 89 surface accessible sites.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Δ Tm (°)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>153</td>
<td>12.333 ± 1.700</td>
<td>HIS</td>
<td>54.4</td>
<td>alpha</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>12.333 ± 2.828</td>
<td>GLY</td>
<td>86.4</td>
<td>alpha</td>
</tr>
<tr>
<td>3</td>
<td>115</td>
<td>11.667 ± 1.247</td>
<td>ASP</td>
<td>76.9</td>
<td>Loop</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>11.333 ± 2.055</td>
<td>GLY</td>
<td>78.8</td>
<td>alpha</td>
</tr>
<tr>
<td>5</td>
<td>154</td>
<td>11.333 ± 1.633</td>
<td>ASN</td>
<td>94.9</td>
<td>alpha</td>
</tr>
<tr>
<td>6</td>
<td>177</td>
<td>11.000 ± 2.494</td>
<td>GLU</td>
<td>59</td>
<td>Loop</td>
</tr>
<tr>
<td>7</td>
<td>198</td>
<td>11.000 ± 2.625</td>
<td>LEU</td>
<td>37.7</td>
<td>Loop</td>
</tr>
<tr>
<td>8</td>
<td>158</td>
<td>11.000 ± 1.247</td>
<td>HIS</td>
<td>68.7</td>
<td>Loop</td>
</tr>
<tr>
<td>9</td>
<td>209</td>
<td>11.000 ± 3.091</td>
<td>ASP</td>
<td>47.1</td>
<td>alpha</td>
</tr>
<tr>
<td>10</td>
<td>182</td>
<td>11.000 ± 1.414</td>
<td>MET</td>
<td>19.5</td>
<td>beta</td>
</tr>
<tr>
<td>85</td>
<td>51A</td>
<td>5.000 ± 0.816</td>
<td>ALA</td>
<td>14.5</td>
<td>Loop</td>
</tr>
<tr>
<td>86</td>
<td>140</td>
<td>4.667 ± 1.247</td>
<td>THR</td>
<td>66.3</td>
<td>alpha</td>
</tr>
<tr>
<td>87</td>
<td>230</td>
<td>4.667 ± 2.449</td>
<td>PHE</td>
<td>12.5</td>
<td>beta</td>
</tr>
<tr>
<td>88</td>
<td>51</td>
<td>3.333 ± 2.625</td>
<td>LEU</td>
<td>8.7</td>
<td>Loop</td>
</tr>
<tr>
<td>89</td>
<td>62</td>
<td>1.67 ± 0.82</td>
<td>LYS</td>
<td>25.4</td>
<td>Loop</td>
</tr>
</tbody>
</table>

Analysis of the data in Tables 10.1 and 10.2 also show there is no correlation between experimental accessibility and site performance. For example, residues 205 and 209 are located in the Alpha II domain, have near-identical accessibility, and are both located on comparable locations on the same alpha helix. In the PEGylation results, the performance of site 209 is significantly better than site 205. The former is in the 90th percentile of best performing sites, but and the latter
is in the bottom 15%. A similar phenomenon is seen in the tethering screen. Residues 112-115 are located in the Loop I domain. Residues 112 and 114 are two of the ten worst tethering sites with accessibility scores of 31.9% and 93%, respectively. However, the other two sites in this group are in the top 20 and also have vastly different accessibility scores. Specifically, residue 113 has an accessibility of 25.7% and residue 115 has an accessibility of 76.9%. The data contained in Tables 6.1, 10.1, 6.2, and 10.2 indicate the secondary structure factors are poor indicators of site performance. Consequently, focus now show shifts to tertiary structures as an indicator of site performance. Tertiary structure, in this case, is defined by the five domains as outlined in Figure 6.1.

To further examine domain effects on functionalization site effectiveness, Figure 6.4 illustrates the stability performance of each site. The data used to make Figure 6.4 is the same as that contained in Tables 10.1 and 10.2, but presenting the information in pictorial form allows the correlation between domain (tertiary structure) and stability to be ascertained more easily than the tabular data. The stability of each site is sorted into one of five colors. The most stabilizing sites are in green followed by shades of lime and proceeding to orange hues. Red-colored sites indicate sites where functionalization either stabilized the protein the least (PEGylation) or destabilized the protein to the largest degree (tethering). The top row of the figure shows an overview of the performance of the protein at all tested sites. The subsequent rows show the results of specific domains to highlight patterns. For convenience, the Loop I and Loop II domains are combined in this analysis due to similar structure, behavior, location, and performance in the screens.

In Figure 6.4, notice the large concentration of poorly-performing sites in the Beta domain and the high number of well-performing sites in the Alpha II domain for both tethering and PEGylation. Also, the Alpha I domain has a high concentration of well-performing sites in the tethering screen, and the Loop domains perform better in the PEGylation screen. These results are quantified in Table 6.3 which lists the number of well- and poorly-performing sites according to domain. In this table, a site is considered well-performing if it is ranked in the top 25 from the screen (colored green or lime) and a site is considered poorly-performing if it is in the bottom 25 (colored red or dark orange). Using these criteria, the order of best performing domain to worst for surface tethering is: Alpha I, Alpha II, Loops, and Beta. For PEGylation this order is: Loops, Alpha II, Alpha I, and Beta. The Beta domain is unique from the other three because it contains a
Figure 6.4: Relative success of each attachment treatment is depicted as a color on the surface of the protein. Green locations are most stabilizing. Shades of lime to yellow to orange indicate decreasing stability, and dark orange to red sites are least stabilizing. To highlight the domain dependence for protein stability, each of the domains have been isolated on each row of the figure. Domains with more green sites are considered better domains for functionalization, and domains with more red sites are considered poor domains for functionalization.
disproportionate number of poorly performing residues relative to the number of functionalizable sites in the domain. This is further emphasized by the fact that Sites 52, 92, 153, 154 in the Alpha II domain appear in the top performing sites for both PEGylation and surface tethering, and Sites 60, 230, 256 in the Beta domain perform poorly for both functionalization methods.

Table 6.3: The number of best and worst performing sites in each domain for the protein device screens. Performance is identified in Figure 6.4 with well-performing sites depicted in green and lime colors, and poorly-performing sites in red and dark orange colors.

<table>
<thead>
<tr>
<th>Domain(s)</th>
<th>Surface Tethering</th>
<th>PEGylation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of well performing sites</td>
<td>Number of poor performing sites</td>
</tr>
<tr>
<td>Alpha I</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Beta</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Alpha II</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Loop I and II</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Alpha I</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Beta</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Alpha II</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Loop I and II</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

These results suggest a possible new heuristic for designing protein devices. Specifically, functionalization to beta domains should be avoided, but alpha and loop domains likely contain good sites. Future work is needed to fully test this hypothesis.
6.5 Discussion

6.5.1 Domain-specific Native Contact Analysis

This section gives a more detailed analysis of the folding mechanisms of the protein and the changes to the mechanism induced by functionalization. Figure 6.5 plots the fraction of each domain that is folded (percent of the native contacts in each domain that are within 0.2 Å of their equilibrium value) as a function of temperature for the WT simulations [147]. As temperature rises, the protein unfolds (native contacts break), and the protein tertiary structure denatures. The temperature at which the slope is greatest is the melting temperature for the domain. Figure 6.5 shows that different domains within the protein melt at different temperatures than the full protein, indicating some domains are inherently weaker than others.

![Figure 6.5: WT native contact plots of TEM-1 as a function of temperature. NCs are plotted as a percentage of all made NC from their specific domain.](image)

The data in Figure 6.5 indicate that the Beta and Loop domains denature at lower temperatures than the full-protein control (black), or, stated another way, the Beta and Loop domains denature before the other domains. The behavior of the Beta domain, because of its involvement in TEM-1 activity, is of particular importance. This domain melts at a lower temperature than any of the other structured domains and is therefore the weakest domain in the protein. The data in Figure 6.4 indicate that functionalizing at sites in this domain also results in subsequent poor performance. This correlation is important because it suggests that only simulations of the WT protein are needed to screen for potential functionalization sites. Such would greatly reduce the
computer time and resources needed to run a screen. Specifically, only one simulation, rather than dozens, would be needed to identify these weak/strong structures, and the inputs for the needed simulation are much easier to generate than those for PEGylation or surface tethering.

6.5.2 Domain Stability on Screen Performance

![Figure 6.6: Sample structures observed when functionalizing the Beta domain (magenta). Structures on the left are representative of the protein at \( T < T_m \), and structures on the right when the Beta domain begins to thermally denature. Panel A) Tethering at Site 51. Panel B) PEGylating at Site 51A.](image)

The connection between weak domains in the WT (as just described) and poor performance in protein device simulation is now investigated in more detail. Representative snapshots of the functionalized proteins are presented to aid the discussion. Figure 6.6 shows structural conformations observed when Residue 51 is tethered to the surface (Panel A) and PEGylated on Residue 51A (Panel B). For reference, the stable structure at a temperature below the melting temperature of the protein is depicted on the left and the denatured form on the right is at a box temperature where melting of the domain in question first occurs. Notice that at the higher temperature, the Alpha II and Loop domains are folded similar to the lower-temperature stable structure. However, when the Beta domain begins to denature at the elevated temperature, the terminal ends in the
Alpha I domain are heavily influenced by the additional forces from the protein device and both domains become unstable. This can be seen in Figure 6.6 where both the Alpha I (cyan) and the Beta (magenta) domains have lost all tertiary structure. This means that functionalizing on Beta domains not only results in loss of structure on the domain where functionalization occurs, but also promotes loss of structure to additional parts of the protein. In short, inherently weak domains seem to be adversely affected when functionalized, and these domains can be identified from a WT model.

![Figure 6.6](image)

Figure 6.7: Sample structures observed when functionalizing the Alpha II domain (steel blue). Structures on the left are representative of the protein at $T < T_m$, and structures on the right when the Alpha II domain begins to thermally denature. Panel A) Tethering at Site 154. Panel B) PEGylating at Site 154.

Conversely, stable domains in the WT protein seem to remain stable once functionalized. Notice in Figure 6.5 that the Alpha II domain denatures at a higher temperature than all other domains in the protein. When functionalized, this domain also contains many well-performing sites (Table 6.3). Figure 6.7 shows structures formed when TEM-1 is tethered (Panel A) or PEGylated (Panel B) to Site 154 which is located in Alpha II. Notice that all other domains at this elevated temperature have lost most of their structure, but the Alpha II domain, even when functionalized,
still maintains its native structure. Functionalizing on this domain did not significantly alter the tertiary structure resulting in a stable device.

The analysis just described helps clarify important ideas surrounding protein functionalization. Specifically this analysis addresses questions in the community as to whether the tether or PEG should be attached in highly structured and rigid regions of the protein or in flexible regions. One thought is that the functionalization should be in rigid regions because they can withstand the additional forces and interactions that the PEG or surface introduce into the system. The competing opinion is that the functionalization should be placed in flexible regions because they can accommodate the new forces with ease. The results presented above suggest that the former is correct. It seems that more rigid regions remain stable upon functionalization, but that less stable regions are adversely affected. Moreover, the degree of impact is larger for the weaker regions meaning that weaker regions become even weaker upon functionalization whereas the stability of stronger regions shifts by a lesser degree. Thus, a WT-only analysis to a priori site selection will best serve to eliminate poor candidate residues then to identify the best.

6.5.3 Functionalization Type Effect on Protein Stability

Despite the Alpha II region being most stable in the WT simulation, it did not contain the highest concentration of well-performing sites for either tethering or PEGylation. So while simulating just the WT and identifying stable domains answers a large part of the question concerning where to place the functionalization, a screen is significantly improved by simulating tethered and PEGylated models. The main reason for including the following examples is to demonstrate that the computational screening efforts proposed in this paper captures many complexities of functionalization that an analysis based only on the WT protein would miss.

The data in Figure 6.4 and Table 6.3 indicated that the Loops and Alpha I domains did not exhibit transitive performance between the tethering and PEGylation functionalization methods. The Alpha I domain performed well in the tethering simulations, but did not perform well in the PEGylation screen (see Figure 6.4). The converse was observed with the Loop domains. These did not do well in the tethering screen but did well in the PEGylation simulations. The main difference in performance between these domains is caused by how the functionalized domain positioned the protein relative to the hydrophilic surface or PEG. Well-performing domains allowed for stable
Figure 6.8: Sample structures observed when functionalizing the Loop domains (pink and dark purple). Structures on the left are representative of the protein at T<Tm, and structures on the right when the Loop domains begin to thermally denature. Panel A) Tethering at Site 178. Panel B) PEGylating at Site 96.

Figure 6.9: Sample structures observed when functionalizing the Alpha I domain (cyan). Structures on the left are representative of the protein at T<Tm, and structures on the right when the Alpha I domain begins to thermally denature. Panel A) Tethering at Site 28. Panel B) PEGylating at Site 26.
structures to be maintained, and poor performing domains placed the surface or PEG at a location that was destructive to local and adjacent tertiary structures.

**Tethering on Domain Structure**

Previous works have shown that surface interactions can significantly affect protein stability and activity [43, 114, 115, 118]. The work done here for TEM-1 agrees well with these previous results. From Figure 6.1, notice that both termini of both Loop domains attach to the Alpha II domain (steel blue), and recall the Alpha II domain is the most stable domain from the WT domain analysis. Figure 6.8 Panel A shows how tethering in these loop regions affects the structure of the protein – particularly the Alpha II domain. At elevated temperatures (right), notice that the Alpha II domain denatures onto the surface even though the tethering was done outside of this domain. Moreover, simulation shows that the Alpha II domain is denaturing before the Beta or Alpha II domains. In this case, the molecular-level reason for the poorly-performing functionalization site is the denaturing of domains adjacent to the Loop domain where functionalization occurs – an effect that is not captured by examining only the PDB structure of the molecule.

Another example where the simulation identifies effects that are not apparent from examining the WT protein alone is provided by cases where tethering was done in the Alpha I domain. Recall attaching the protein to the surface in the Alpha I domain resulted in many well-performing sites, and that the Beta domain, in general, is a weak part of the molecule. In the WT protein, the Beta domain (pictured in magenta) denatures before the Alpha I domain (see Fig. 6.5). Figure 6.9 Panel A shows representative structures for tethering in the Alpha I (pictured in cyan) domain. The right-side structure of Figure 6.9 Panel A shows that even when the Alpha I domain begins to denature, that the Beta domain remains in tact. In this case, tethering in one region caused an adjacent and WT-weaker region to be stabilized. The result is a well-performing site. Again, the point of this example, is that examining just the WT behavior would not identify the complex stabilization seen by tethering in the Alpha I domain. It would be ideal if heuristics could be developed based only on the WT structure, but this seems infeasible. Rather, simulation is necessary to capture the complex interactions between the protein and the surface to create a proper screen.
**PEGylation on Domain Structure**

With no surface for the protein to interact with, and no evidence of the PEG polymer wrapping around the protein, good candidate sites for PEGylation are dependant on a domain’s ability to maintain local and adjacent domain structure. When PEGylating the Loop domains, Figure 6.8 Panel B show representative of structures where $T < T_m$ on the left, and on the right structures when local domain structure begins to break. Notice how the loop domain to which the PEG is attached (pink) separates from the rest of the protein at the higher temperature. This movement shows how the PEG disrupts the local protein structure surrounding the functionalization site. However, the loss of local protein structure only occurs at high temperatures, is only contained to the domain where it is attached, and does not propagate to adjacent domains. The domain tolerates the additional forces from the PEG and does not fully denature until the other domains have lost structure. Results from the PEGylation screen revealed a large number of well-performing functionalization sites on these loop domains (see Figure 6.4) with the majority of sites ranking in the upper 50th percentile (Table 10.2).

PEGylating on the Alpha I domain resulted in very few well-performing sites. Poor site performance stems from early loss of structure in the adjacent Beta domain. This can be seen in Figure 6.9 where the elevated-temperature structure on Panel B shows both the Alpha I and Beta domains have lost most of their tertiary structure. The additional forces from the PEG overwhelm the intraprotein interactions maintaining the structure of the molecule at temperatures lower then the other domains. Denaturing resembles that observed during simulations when PEGylating the Beta domain. Thus the tertiary structure in this domain cannot maintain native structure.

### 6.6 Procedure for Identifying Poor Sites

Recall that the goal of this work is to create an in silico method to identify sites on a protein that could be functionalized while maintaining protein function. Simulating all relevant, surface-accessible, residues on a protein can produce many candidate sites for screening, but this can be time-consuming and resource-heavy if a protein is large or more replicates are wanted. It is desirous, then, to focus on the regions of the molecule with the highest likelihood of good-performing sites and eliminate poor candidates early in the process. Domain-directed analysis can
help with this process. A procedure to detect poor candidates quickly and then focus resources on better regions of the molecule is now outlined.

1. Run replica exchange simulations on the WT model and analyze the results to identify the stability of each domain (see Figure 6.5). Eliminate the weakest domains from further consideration.

2. Identify the residues comprising the active sites and eliminate these from further consideration.

3. Determine the experimental accessibility rating for the remaining residues in the remaining domains. Eliminate residues with <10% accessibility from further consideration.

4. Perform replica exchange simulations on the functionalized protein at the remaining residues to determine the melting temperature for each case. Optionally, to further reduce the number of candidate sites, a series of screens with increased sampling based on domain performance can bias more promising domains and eliminate domains with unfavorable protein device configurations.

5. Rank sites based on the Tm results to identify the top 5-15 residues that should be further explored experimentally.

Following the targeted screening procedure just described reduces the amount of time required for a screen proportional to the number of weak domains in the protein with negligible loss of effectiveness. For a given replicate simulation, this work used one core per replica exchange box and required 6-8 hrs of computation time. Each of the 89 sites had 3 replicates. By simulating both PEGylated and tethered protein devices, a total of 534 replicates were simulated. Using the above procedure on TEM-1 for this work would have reduced the number of replicates simulated in the screen, and computation time would have been reduced by at least a third. With a moderately-sized computer cluster, the screen would take about a week of real time to run all necessary simulations.

6.7 Considerations

A notable observation from this work is that PEGylating or tethering the Alpha I domain is not necessarily as destructive as functionalizing on the Beta domain. This is unique because all
secondary structures in the Alpha I only attach to structures in the Beta domain and, additionally, this Alpha I contains both the terminal ends. Attaching to the weak Beta domain structure and containing the terminal ends would intuitively suggest poor functionalization performance, but the simulation results indicate otherwise. No clear or obvious mechanism for why the Alpha I domain performs better than the Beta domain can be concluded from the data. However, the authors hypothesize that when functionalizing the Beta domain there is a much greater energy barrier to reforming both the Beta and Alpha I domains – compared to only reforming one domain when functionalizing the Alpha I domain. Thus once unfolded, reforming native structure is less favorable.

6.8 Conclusions

In this paper TEM-1 was functionalized by either tethering to a surface or PEGylation and then the protein device was simulated to determine its thermal stability. All experimentally-accessible sites for functionalization on TEM-1 were analyzed and ranked according to the degree the device affects protein melting temperature. Analysis of the relative performance shows little significant correlation with primary and secondary structure factors like accessibility, amino acid type, and secondary structures type. Tertiary structure factors do seem to correlate with site performance and analysis compared domain performances as a predictor of functionalization effectiveness. Using domain-specific unfolding analysis, strong and weak domains can be identified prior to functionalization simulations and could be used in the future to reduce the number of sites tested during a screen.

When comparing performances of tethered and PEGylated protein devices, primary and secondary structure factors do not seem to have much correlation. Often, individual residue performance from one protein device did not predict performance in the other. The strongest and weakest tertiary structure domains for functionalization were shown to be transferable between the two protein devices tested. This was evident when results from both screens showed many poor performing sites located in the Beta domain, and many well-performing site located in the Alpha II domain. However, domain-specific analysis cannot replace an exhaustive screen due to unforeseeable influences of functionalization on stability. Sites identified as good candidates still require
additional experimental examination, but work here shows the significance of tertiary structure on the effectiveness of a protein device.
CHAPTER 7. PARAMETERIZATION OF UAA WITH AZIDES AND ALKYNES

7.1 Introduction

The PRECISE method enables site-specific protein device creation. However, currently no heuristic exists for determining, a priori, optimal uAA mutation sites and instead the mutation is considered insignificant. When creating protein devices, this is problematic because many mutation sites can lower both protein stability and function, which negates the entire purpose of functionalization. Experimental trial-and-error is tedious and expensive in time and money \[42, 113\]. Previous studies show that an in silico screen using molecular simulation can be a good qualitative predictor of protein device behavior in unnatural environments such as those present in protein devices \[73, 113, 118\]. 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 azide and alkyne moieties.

The purpose of this section is to report on the creation of CHARMM-compatible model parameters for molecules that contain azido or alkynyl groups. All structures in this chapter involve moieties that contain a linear angle between three atoms, a feature not found in most biological molecules, and whose unique structure has not been properly addressed in current literature (see Chapter 3). This report explains the methods used to create parameters, the subsequent results, and the pitfalls encountered when modeling linear angles and their associated dihedrals. These new parameters expand the realm of 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.

7.2 Model Parameters

The general procedure used to develop the new force field parameters needed for azides and alkynes consists of: (1) developing CHARMM parameters for four small molecules that con-
p-propargyloxy phenylalanine

p-azido-L-phenylalanine

Figure 7.1: Unnatural amino acids commonly used in the PRECISE technique for protein device creation.

To obtain terminal azido and alkynyl groups using the force field Tool Kit (ffTK) [148], (2) addressing linear structure issues seen during ffTK analysis by modifying the procedure so that all CHARMM parameters can be resolved, and (3) validating the resulting force field via in silico molecular dynamic (MD) simulation of the four small molecules. Once the parameters for the linear moieties were obtained, their transferability to model uAAs in proteins is tested. This consists of using 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) [95].

Because an all-atom model does not exist for the four small molecules used in this study, the CHARMM FF is again presented to identify the exact parameters that need to determined. CHARMM defines the potential energy of a molecular system ($V_{molecule} = V_{bonded} + V_{nonbonded}$) to include contributions from bonds, angles, dihedrals, improper dihedrals, Urey-Bradley (UB), charge and external interactions. All bonded interactions are calculated according to:
\[ V_{\text{bonded}} = \sum_{l=\text{Bonds}} K_{b,l}(b_l - b_{0,l})^2 + \]
\[ \sum_{m=\text{Angles}} K_{\theta,m}(\theta_m - \theta_{0,m})^2 + \]
\[ \sum_{n=\text{Dihedrals}} K_{\phi,n}(1 + \cos(n_n\phi_n - d_n)) + \]
\[ \sum_{u=\text{Impropers}} K_{\omega,u}(\omega_u - \omega_{0,u})^2 + \]
\[ \sum_{v=\text{UB}} K_{S,v}(S_v - S_{0,v})^2 \]

where all bond pairs \((l)\), angle pairs \((m)\), improper dihedrals \((u)\) and UB 1,3-interactions \((v)\) use a form of Hook’s law. The values \(b_l, \theta_m, \omega_u, \) 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 \(K_{S,v}\) [64]. For all dihedral pairs \((n)\), the dihedral potential is defined by the dihedral angles \((\phi_n)\) in relation to their sinusoidal multiplicity \((n)\) and phase shift \((d)\) scaled to their equilibrium potential using a force constant \((K_{\phi,n})\) [64]. Most importantly, all bonded pair parameters \((l, m, n, u, v)\) must be defined for all combinations within the molecule and are considered unique unless otherwise justified.

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

\[ V_{\text{nonbonded}} = \sum_{\text{charge}} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{\text{Lenard-Jones}} \varepsilon_{ij} \left[ \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^6 \right] \]

where charge terms \(q_i\) and \(q_j\) are the partial atomic charges for unique atom-charge pair. The \(r_{ij}\) parameter is the distance between the two atoms in the pair and is obtained from simulation. The \(4\pi\varepsilon_0\) term defines the system dielectric constant. The Lennard-Jones nonbonded interactions combine parameters from each participating atom using the Lorentz-Berthelodt combination rules. The \(\varepsilon_{ij}\) value is the geometric mean of atomic parameters \(\varepsilon_i\) and \(\varepsilon_j\), and \(R_{\text{min},ij}\) is the arithmetic
mean of atomic parameters $R_{\text{min},i}$ and $R_{\text{min},j}$. All nonbonded atomic parameters must be defined for all new atom types introduced into the the CHARMM FF.

7.3 Method for Parameterization

To determine unique parameters, the Force Field tool kit (ffTK) is used [148]. The ffTK approach to parameterization 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 7.2 contains a cartoon representation of each small molecule and also defines the atom naming structure used in this work. PY and PAZ molecules were chosen because they best isolate the linear angle moiety and connect it to non-aromatic carbons. 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 [149] and used the MMFF94 structural optimization algorithm for initial structure approximation [150]. Multiple iterations of the ffTK method are used if a molecule presents evidence during validation that parameter refinement is needed [151].

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 atoms and parameters are needed for any $V_{\text{molecule}}$ 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. PAZ is similarly parameterized, but include in our list of new atoms the phenylic C.

7.3.1 Parameter Determination: ffTK

The ffTK method was chosen for parameterization because it has been successfully used to obtain CHARMM-compatible parameters for small molecules [148,151,152]. Each ffTK step
is briefly outlined in this section. Due to the challenge of modeling linear chemical moieties, the standard ffTK procedure is altered slightly and these modifications are explained at the appropriate ffTK step.

**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 [153] 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. Thus higher levels of theory were not needed. The structures shown in Figure 7.2 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 [154]. For all aliphatic and aromatic hydrogens, the standard CHARMM charges of +0.09 and +0.015, respectively, were used to maintain consistency with CGenFF [70]. For every other atom in the molecule, donor, acceptor, or non-interacting status was applied and then a TIP3P water molecule is appropriately oriented to optimize hydrogen bonding with the target atom and minimize steric interactions with the surrounding atoms. Two QM optimization steps are performed using the HF/6-31G(d) level of theory to determine the two remaining free parameters: (1) distance between interaction site and water molecule and (2) rotation angle of the water molecule to the incident target atom. NAMD, using the CHARMM force field, then simulates the same system to determine the MM nonbonded interaction parameters. Additional parameterization iterations are executed until there is sufficient agreement between MM and QM simulation [148].

**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 [148].

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 PED. 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. The QM PES uses 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 [148].

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 12 dihedrals that require parameterization. Of these 12, 7 required modification of the ffTK protocol. Within the structure of these 7 unique dihedrals, three of the four atoms form a linear angle whose $\theta_0 \approx 180°$. For convenience, such dihedrals are given the abbreviation LACD (linear-angle-containing dihedral).

As is demonstrated later, LACDs are difficult to parameterize due to numerical discontinuities associated with the calculation of the dihedral angle and the sensitivity of the dihedral potential to small perturbations of the linear angle. Said another way, the movements of the linear angle and the LACD are coupled in a very stiff region of angle/dihedral phase space. The ffTK method, which seeks to obtain decoupled parameters, failed because it is not intended to be used for inherently-coupled systems. Thus, a different approach is needed to parameterize LACDs. Three different methods were used towards this end and are now described.

ffTK Step 4.1: Modified ffTK

As previously mentioned, normal ffTK procedure for dihedral parameterization requires a QM PES, but LACDs would not complete a PES within Gaussian09 using a completely unrestrained system. To alleviate the problem, the linear angle within the LACD was fixed to its optimal angle for the duration of the dihedral scan. This means that all portions of the molecule except the linear moiety were allowed to relax during structure optimization steps.

Fixing the linear angle to the equilibrium value is considered a viable candidate approach to solving the difficulties surrounding the parameterization of the LACD because the strength of
linear moiety structure does not allow large deviations from linearity to occur. The LACD PES scan was then parameterized and optimized in ffTK, but only after all non-LACD parameters had already been determined. By doing so, any potential numeric instability of the LACD would not affect the validity of the non-LACD parameters.

**ffTK Step 4.2: Linear Fit**

This second approach undertaken to parameterize LACD parameters uses best fit regression. In this procedure the LACD was not included with the other dihedrals when ffTK perform simultaneous dihedral optimization. The QM LACD PES was still obtained from Gaussian09 using the same procedures as was outlined in the previous section, but the generation of model parameters for the mathematical description of the LACD was created by manual least-squares fitting. This was done by using the CHARMM dihedral cosine expansion (see Equation 7.1) as the objective function, setting \( n = 1 \) and varying \( K_\phi \) and \( d \) to reproduce the PES. This approach for fitting the PES to the dihedral equation in isolation is considered a dated but viable candidate as it was the first dihedral parameterization approach proposed by CHARMM [119].

**ffTK Step 4.3: Rational Design**

This last approach for parameterizing LACDs completely eliminates LACD contribution from the force field. Specifically, \( K_\phi \) is set equal to zero for each LACD. The reasoning behind this approach came after the results of the previous two methods were obtained which showed that energetic contributions from LACDs are much less than regular, non-LACD, dihedrals and approach \( k_BT \). This suggests that LACD energetic contribution is on the same order as the thermal background energy in the system and that 4-bodied dihedrals are not needed to reproduce the correct dynamics in linear moieties.

**ffTK Step 4.4: Compare LACD Parameterization Methods**

To compare the validity of these three LACD parameterization methods, MD simulations of each new molecule are compared to QM structure. All MD simulations are done using the LAMMPS simulation software [142]. Each molecule is solvated in a bath of TIP3P waters with at least 17\( \text{Å} \) of water between the molecule and the sides of the periodic simulation box. LAMMPS-compatible input files are generated from CHARMM files using the charmm2lammps tool within
the LAMMPS software and SHAKE is used to fix the bond lengths between hydrogens and heavy atoms. Next, these models are simulated using an NVE ensemble for at least 10 ns. The LACD parameterization method that does not present numeric instability and most accurately reproduces linear angle structure is then selected. For brevity, only the simulation results for the AZB small molecule are shown as it is representative of the other 3 small molecules.

7.3.2 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 7.2, and the Tra-Fab fragment shown in Figure 7.3 (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. Protein structure was obtained by X-ray crystallography [95]. To simulate this system, parameters are needed for the pAz uAA. The pAz model was determined using the same procedure as defined in ParamChem [155, 156] but with its scope of reference molecules expanded to include the azide small molecules parameterized in this work. This protein system is completely independent from the parameterization model. Thus successfully modeling this protein will provide suitable validative test case for LACD parameters and pAz uAA simulation.

Protocol for all MD simulations used during validation are as follows: First, the 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 ensure equilibrium and then cooled 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. 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 for small molecule analysis or 18.78 ns of production for large protein analysis. These production-phase simulations are then compared to the QM ab initio structures or crystal structure using RMSD analysis, respectively.
Figure 7.3: 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.

7.4 Results

The results are presented in the following order. First, all optimized parameters (except the LACD parameters) from the ffTK method are given for all four small molecules. These parameters describe molecular structure, charge, bond, angle and normal dihedral parameters. The next section shows the results of the three approaches for parameterizing LCADs.
7.4.1 Optimized Parameters

All parameters, with the exception of LACDs, are determined in accordance with ffTK procedure as defined in the Methods section and without issue. Table 7.1 contains the list of CHARMM-compatible atom types along with partial charges and non-bonded parameters according to atom reference shown in Figure 7.2. The names for each atom type are chosen to be distinct from other types found in the CHARMM36 protein force field and CGenFF.

Alkyne charges correlate with previous ab initio studies that show the CH group is more positively charged than the CH$_3$ group [157]. Azide charges are more complicated as they can have up to 4 possible charge configurations [158]. The 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 [159, 160]. All charge and nonbonded parameters are within the optimization standards set in ffTK.

Table 7.1: Optimized CHARMM parameters for all relevant atoms in the linear moiety. Dihedral CHARMM parameters in this table exclude LACDs, thus only select dihedrals from POPY, MAZ, and PAZ are reported.

<table>
<thead>
<tr>
<th>Atoms</th>
<th>Atom Type</th>
<th>Charge</th>
<th>Nonbonded $\varepsilon$</th>
<th>Nonbonded $\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>CY1</td>
<td>0.040</td>
<td>-0.1670</td>
<td>1.840</td>
</tr>
<tr>
<td>C2</td>
<td>CY2</td>
<td>-0.469</td>
<td>-0.1032</td>
<td>1.9925</td>
</tr>
<tr>
<td>H1</td>
<td>HY</td>
<td>0.298</td>
<td>-0.0090</td>
<td>0.875</td>
</tr>
<tr>
<td>CY</td>
<td>CYA3</td>
<td>-0.139</td>
<td>-0.0320</td>
<td>2.000</td>
</tr>
<tr>
<td>H2-4</td>
<td>HYA3*</td>
<td>0.090</td>
<td>-0.0240</td>
<td>1.340</td>
</tr>
<tr>
<td><strong>POPY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>CY1O</td>
<td>0.236</td>
<td>-0.1670</td>
<td>1.840</td>
</tr>
<tr>
<td>C2</td>
<td>CY2O</td>
<td>-0.544</td>
<td>-0.1032</td>
<td>1.9925</td>
</tr>
</tbody>
</table>

Continued on next page
Table 7.1 – *Continued from previous page*

<table>
<thead>
<tr>
<th>Bond</th>
<th>$K_b$ ($kcal$ mol$^{-1}$ Å$^{-2}$)</th>
<th>$b_0$ (Å)</th>
<th>Ref. [74] (Å)</th>
<th>Exp. [161] (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYA3 CY1</td>
<td>370.73</td>
<td>1.466</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CY1 CY2</td>
<td>999.30</td>
<td>1.221</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CY2 HY</td>
<td>423.09</td>
<td>1.065</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYA3 HYA3</td>
<td>353.90</td>
<td>1.094</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>POPY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYA2 CY1O</td>
<td>373.31</td>
<td>1.460</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Angle Parameters</th>
<th>( K_\theta ) (kcal mol(^{-1}) rad(^{-2}))</th>
<th>( \theta_0 ) (°)</th>
<th>Ref. [74] (°)</th>
<th>Exp. [161] (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY1O CY2O HY</td>
<td>70.052</td>
<td>179.936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYA2 CY1O CY2O</td>
<td>76.604</td>
<td>179.934</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table 7.1 – Continued from previous page

<table>
<thead>
<tr>
<th></th>
<th>Kφ (kcal mol⁻¹)</th>
<th>n</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POPY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1 CY O CG2R61</td>
<td>0.730</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>C1 CY O CG2R61</td>
<td>0.826</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>C1 CY O CG2R61</td>
<td>1.337</td>
<td>3</td>
<td>180</td>
</tr>
<tr>
<td>H2/3 CY O CG2R61</td>
<td>0.058</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>H2/3 CY O CG2R61</td>
<td>0.651</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>H2/3 CY O CG2R61</td>
<td>0.981</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>CY O CG2R61 CG2R61</td>
<td>1.549</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CY O CG2R61 CG2R61</td>
<td>1.139</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>CY O CG2R61 CG2R61</td>
<td>0.213</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Continued on next page
**Table 7.1 – Continued from previous page**

<table>
<thead>
<tr>
<th>Bond Type</th>
<th>Atom 1</th>
<th>Atom 2</th>
<th>Atom 3</th>
<th>Atom 4</th>
<th>Bond Length (Å)</th>
<th>Bond Order</th>
<th>Bond Angle (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAZ</td>
<td>N2 N1</td>
<td>CZ</td>
<td>H</td>
<td></td>
<td>0.364</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>N2 N1</td>
<td>CZ</td>
<td>H</td>
<td></td>
<td>0.180</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>PAZ</td>
<td>N2 N1</td>
<td>CZ</td>
<td>CG261</td>
<td></td>
<td>3.537</td>
<td>2</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>N2 N1</td>
<td>CZ</td>
<td>CG261</td>
<td></td>
<td>0.249</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>N1 CZ</td>
<td>CG2R61</td>
<td>HGR61*</td>
<td></td>
<td>2.000</td>
<td>2</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>N1 CZ</td>
<td>CG2R61</td>
<td>CG2R61**</td>
<td></td>
<td>1.000</td>
<td>2</td>
<td>180</td>
</tr>
</tbody>
</table>

*Aliphatic and aromatic hydrogen charges fixed to CHARMM standard. **These parameters were assumed similar to CGenFF and were not explicitly parameterized 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 errors are close to the 0.02 Å and 3° standards set in CGenFF [70]. Linear moiety angle parameters had errors ≥ 5°, but these errors are the consequence of simulation error and not parameterization error. Further analysis is discussed in the next chapter. Normal dihedral (non-LACD) parameters derived using ffTK are summarized in Table 7.1. These parameters were optimized with good RMSE agreement for both high and low energy states and were within the 5 kcal/mol standard set by CGenFF (Figure 7.4) [70].
Figure 7.4: Torsion profiles for all normal dihedrals. Panel A The N2 N1 CZ H dihedral in MAZ Panel B The N2 N1 CZ CG2R61 dihedral in PAZ Panel C The C1 CY CG2R61, H2/3 CY O CG2R61, AND CY O CG2R61 CG2R61 dihedrals in POPY.

7.4.2 LACD Parameterization

As explained in Section 7.3.1, parameterization of LACDs requires a novel approach because the standard ffTK method fails to accurately capture the relevant physics. This section reports the results from the three approaches – Modified ffTK, Direct Fit, and Rational Design.

Modified ffTK Method

During a normal quantum-mechanical PES scan, the entire molecule is allowed to relax into the lowest-energy conformation for each step of the PES. However, this normal procedure failed for LACDs. For the Modified ffTK procedure, instead of allowing the entire molecule to relax during the QM PES scan, all linear angles were fixed to their QM optimized angle. After
Figure 7.5: ffTK dihedral optimization of the LACD in the MAZ (A) and PAZ (B) molecule, respectively. PY did not produce an LACD QM PES and the LACD QM PES in POPY had zero agreement with the MM PES and thus these plots are not presented.

all non-LACD dihedrals were parameterized, the LACD dihedrals were parameterized in a second implementation of ffTK dihedral parameterization.

Figure 7.5 shows how the linear-angle-constrained QM model (black) maintains a smooth PES whereas the MD simulation (blue) tries to mimic this trend, but points of discontinuity prevent a good fit. Under normal ffTK optimization, MD simulations are performed to obtain an internal MM PES using a dihedral model that is optimized to match the QM PES. This internal MM PES should produce a smooth curve where structural changes in angle do not appreciably affect dihedral structure [81]. However, in the case of linear moieties, linear angle parameters couple with dihedral parameters (see Chapter 8). This means that during MM PES structural relaxation steps, the forces from the angle potential induce changes in dihedral space resulting in discontinuous movements in dihedral space. Due to these discontinuities, it is impossible to obtain a smooth MM PES for any LACD (Figure 7.5). The ffTK procedure fails to produce reliable LACD parameters because it tries to match an unconstrained and discontinuous MM PES to the linear-angle-constrained QM PES. Moreover, in the case of propyne and because this molecule is so symmetric, PES energetic perturbations were inconsequential and Gaussian09 could not resolve a PES for this molecule.
Best Fit

This method used the same set-up as the method in the previous section, except after all non-LACD dihedrals were parameterized, the LACD dihedrals parameters were determined by using a linear regression on the dihedral cosine expansion equation. These LACD parameters determined using this method closely model the QM PES and in all cases the modeled torsion profiles (Figure 7.6) do not exhibit significant deviations from the QM.

Figure 7.6: Torsional profiles using the Best Fit method for all LACD dihedrals in POPY, MAZ, and PAZ, as labeled at the top of each plot. The QM PES (black) and the Best Fit PES (red) are shown for each dihedral. The PY profile is not shown because a QM PES could not be computed.

The sinusoidal multiplicity is the same for all LACDs (n = 1) with differences only being found in phase shift (d) and force constant ($K_\phi$) values. $K_\phi$ is scaled to appropriately match the magnitude of torsion influence on structure with 4 of the 5 PES revealing torsional contributions a magnitude or more below $k_b T$, and the remaining PES on the same order. Phase shifts for most of the LACDs are either in-phase (d = 0°) or out-of-phase (d = 180°). The only exception is the H CY C1 C2 dihedral in POPY that is best represented by a 300° shift. As many simulation programs do
not allow for atypical phases, analysis using this parameter assumes it is close enough to $d = 0^\circ$ to not significantly alter the results, and in practice a $d = 0^\circ$ phase is used.

Table 7.2: Best Fit LACD CHARMM parameters for all relevant atoms in the linear moiety.

<table>
<thead>
<tr>
<th>Dihedral Parameters</th>
<th>$K_\phi$ (kcal mol$^{-1}$)</th>
<th>n</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CY C1 C2 H1</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>H CY C1 C2</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>POPY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CY C1 C2 H1</td>
<td>0.0095</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>H CY C1 C2</td>
<td>0.0017</td>
<td>1</td>
<td>300</td>
</tr>
<tr>
<td>O CY C1 C2</td>
<td>0.0017</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td><strong>MAZ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CZ N1 N2 N3</td>
<td>0.562</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>PAZ</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CZ N1 N2 N3</td>
<td>0.055</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

* QM PES could not be resolved for these LACD and, consequently, the Best Fit method could not generate parameters for these moieties.

Table 7.2 presents all LACD parameters using the Best Fit approach. The values obtained can be used in simulation and produce reasonable structures, but a unique problem arose that requires a very small timestep to obtain good energy conservation. This peculiar difficulty is inherent to all LACDs and is now described. Figure 7.7 shows the results of several NVE simulations of
PAZ that were started from an equilibrated state. The potential energy as a function of iteration number is plotted in each case. Panel A shows the PE vs. time plot when using a timestep of 1.0 fs. Under such equilibrated conditions, the potential energy of the system should fluctuate around an average value and should not drift. In this case, however, the potential energy increases over the entire simulation time and experiences large discontinuous jumps. Because the system never establishes equilibrium and potential energy increases throughout the simulation, this phenomenon is termed “ramping.”

Figure 7.7: NVE Simulations of PAZ with varying timestep or LACD dihedral parameter. Panel A Optimized parameters from Best Fit approach for dihedral parameterization. Panels B & C Decreasing the timestep to 0.5 and 0.1fs, respectively. Panel D Same timestep as Panel A, but excludes the LACD parameter.

Ramping is not novel to our simulations. Rather, it is observed whenever discontinuities in the potential energy landscape occurs. Many factors can produce discontinuities in system PE such as improper radial cutoffs, exclusion of SHAKE on hydrogen bonds, or too large of timesteps. However, the system established in Figure 7.7A was devoid of these potential pitfalls and ran at
conditions appropriate for all-atom simulation. Thus, there is a discontinuity introduced by the LACD that is unknown in current literature.

The most common method to reduce the effects of discontinuous motion is to reduce timestep. Such an approach was taken here. Simulated were timesteps at: 1.0, 0.5, 0.25 and 0.1 fs. As already mentioned, Figure 7.7A shows the results of using a 1.0 fs timestep. Panels B & C on Figure 7.7 show the same system, except with timesteps of 0.5 and 0.1 fs, respectively. Notice that the degree of numeric ramping decreases as timestep decreases. Numeric ramping is absent for a timestep of 0.1 fs for PAZ (Figure 7.7C) and for all other LACDs (not shown). These results demonstrate that the $K_\phi$ Best Fit parameter can yield a stable system without noticeable simulation ramping, but running all-atom simulations with such a low timestep is computationally infeasible for modern applications.

Extensive work was undertaken to understand the origin of this ramping, and these numeric inconsistencies are of such complexity and import that they are the subject of the next chapter in this dissertation. It shows that ramping occurs due to coupling between the angle and dihedral potentials in a stiff region of dihedral space and is only present in LACDs. Specifically, coupling causes molecular movement within a single timestep that appears physically impossible with the assumption that dihedral space is independent from all other intermolecular forces during simulation. For the present purposes, it is sufficient to know that numerically-stable simulations that contain LACD parameters do not present evidence of ramping if timestep is sufficiently small.

Rational Design

Resolving the issue of simulation ramping at reasonable timesteps begins with three observations from the previous results. (1) Energetic contributions of the LACD QM PES dihedral scans are on the same order as $k_b T$. (2) $K_\phi$ for LACDs using the Best Fit method are very small compared to other dihedral parameters. For example, the average magnitude of the $K_\phi$’s in the CHARMM 36 Protein Force File is $\sim$2 kcal/mol compared to the Best Fit $K_\phi$ values for LACDs which are at least an order of magnitude smaller (Table 7.2). Finally, (3) Numeric inconsistencies during MD simulation cause dihedral/angle parameter coupling. Due to these factors, the force field preferentially biases linear angle contributions over dihedral forces. By considering these three factors, it was decided to remove the LACD potential altogether. This was justified because: a) the energetic
barriers that must be traversed to move through dihedral space are very small compared to both the thermal energy and the other intramolecular forces in present in the system and b) doing so would remove the source of ramping in the simulation. Said another way, it was hypothesized that LACDs do not behave in the same manner as normal dihedrals in restricting movement through dihedral space, so they should not be included in the mathematical model. This third approach to parameterizing, termed the Rational Design approach, sets $K_\phi = 0$ for all LACDs.

Figure 7.7D shows the outcome of using this Rational Design approach for LCADs. As was done before, the potential energy of the system is plotted as a function of simulation time. Ramping is absent in this case (PAZ) and in all other cases (not shown). For practical applications of linear moiety simulation, it appears that removing LACD parameters generates stable simulations without introducing a new FF. The validity of this approach is now tested by comparing the molecular structures that result from simulation using this model to QM and crystallographic structures.

### 7.4.3 Validation

Of the three methods for determining LACD parameters, two methods produce energetically stable simulations of linear molecular structures: Best Fit and Rational Design. However, the former produces ramping for any timestep greater than 0.1 fs and is thus of little practical importance. Fortunately, linear molecular moieties do not require dihedral FF parameters to maintain molecular structure at or near 180° because the other intermolecular forces are sufficient to simulate proper behavior. This fact is demonstrated in this section.

The validity of all ffTK model parameters with the Rational Design LACD parameterization method is tested following the standard practice for force field generation. The new model is used in MD simulation and the resulting structures are then compared to standard structures. This validation MD is executed on two systems sets. The first set are simulations of the small molecules as depicted in Figure 7.2 at 300 K and 1 atm and the second set are simulations of the Tra-Fab protein with pAz mutation also at 300 K and 1 atm. The data from each NVT simulation is analyzed to determined the root mean squared deviation (RMSD) between structures produced by the simulation and the appropriate standard structure. The standard used for the small molecules is the QM ab initio structures, and the standard for the Tra-Fab protein is the crystallographic data.
Figure 7.8: Panel A PY RMSD with an average value of 0.13 Å. Panel B MAZ RMSD with an average value of 0.09 Å. Panel C PAZ RMSD with an average value of 0.17 Å. Panel 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).

Figure 7.8 shows RMSD results for the Rational Design LACD parameterization method. RMSD calculations are averaged across all production-phase simulation frames to produce an average RMSD value. Figure 7.8A shows the RMSD result for PY and generates an average RMSD value of 0.13 Å. Figures 7.8B-C show the MAZ and PAZ molecules generate an average RMSD value of 0.09 and 0.17 Å respectively. For POPY, two 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 7.8D shows the POPY RMSD using this criteria and generates an average RMSD value of 0.08 Å. In all cases, RMSD never exceeds 0.2 Å which proves good agreement between model parameters and QM predicted structures.
Protein Simulation

To test the transferability of the newly-developed parameters for linear moieties, simulations are done on a dimer of Tra-Fab domains where both domains have mirrored pAz mutations within their primary structure. Figure 7.3 depicts the uAA mutations located on chains A and C (colored blue). 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. As before, the temperature was set to 300 K and the box size equal to the average lengths found during the 1 ns NPT simulation.

![Graph](image)

**Figure 7.9:** Panel A Full protein RMSD with an average RMSD of 2.2 Å. Panel B Local structure containing the pAz uAA on chain A showed an average RMSD of 1.3 Å. Panel 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).

Previous research on MD uAA incorporation shows that analysis of an entire protein (150-200 residues) allows for 2 Å RMSD from crystal structure to remain consistent during simulation [99]. Figure 7.9 shows the protein backbone RMSD for the Tra-Fab dimer and that MD uAA incorporation of pAz successfully recreates crystal structures. Analysis of the entire protein’s
backbone produces an average RMSD of 2.2 Å. Additionally, local structures that include and surround the mutant (sites 151-159) perform better than the protein average to generate average RMSD values of 1.3 or 1.4 Å (Figure 7.9B and C, respectively).

These findings establish a strong case supporting the validity of our linear moiety parameters. Prior to simulation none of the LACD parameters used this crystal structure for parameterization input. Simulation not only agrees with crystal structure for the entire protein, but in the regions where the new parameters were implemented there is better-than-standard agreement. Thus in a full-protein system, simulation can accurately predict how an azido group affects protein structure.

7.5 Conclusion

Modern biotechnology techniques often use unnatural amino acids to add chemistries not typically seen in biology for protein modification, such as the linear structures present in azides and alkynes. Molecular simulation of these linear-angle-containing molecules has not been previously possible because model parameters did not exist. This work reports the first CHARMM-compatible parameters for azide and alkyne chemical moieties. The parameterization largely follows the ffTK methodology, but a novel approach is needed to address dihedral parameters that contain linear angles. Three methods for parameterizing these linear-angle-containing dihedrals are tested: modified ffTK, Best Fit and Rational Design. Modified ffTK did not produce any reasonable parameters. Best Fit and Rational Design methods produce viable parameters, but MD simulation of the Best Fit parameters produces energetic ramping in simulations that do not use a very small time-step. The Rational Design, or elimination of all LACD parameters, is currently the best method for parameterization of LACDs because it produces reasonable structures for linear moieties and does not require any additional FF modification. 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. In short, the parameters derived in this work sufficiently capture the structural geometry of linear moieties and can be used in the CHARMM force field to model molecules that contain azido or alkylnyl groups.
CHAPTER 8. CHALLENGES WHEN SIMULATING LACD MOIETIES

8.1 Introduction

Proper dihedral angle modeling is key to creating accurate force fields for biomolecular simulation [82–84]. Unfortunately, many recent techniques in biotechnology use chemical moieties not commonly found in nature whose molecular motions are not adequately described by current formalisms. One of these types of moieties is linear molecular groups which are found in azide and alkyne-containing molecules needed in so-called “click” reactions [10, 11, 162]. This reaction is used in organic chemistry and biology to covalently link molecules without the risk of side-reactions. “Click” chemistry is useful in several applications including drug delivery, medicinal chemistry, site-specific protein devices, and drug docking [8, 10, 163–166].

The previous chapter (Chapter 7) describes efforts to create CHARMM-compatible parameters for unnatural amino acids composed of azide and alkyne groups, and how problems quickly arose with the linear angle parameters. Specifically, when standard procedures were followed, the resulting parameters produced simulations where the potential energy of the molecule increased throughout the entire simulation, a phenomenon termed “ramping” and experienced by all linear moieties. With particular emphasis on LACDs, this chapter explains why these structures are vulnerable to computational issues, reveals the cause of ramping, and explains why linear angle parameters display large errors during parameterization.

8.2 Background: Modern Molecular Force Fields

Modern molecular force fields, such as CHARMM, AMBER, OPLS, COMPASS, etc. are each based on similar assumptions, one of which breaks down for LACDs. To illustrate, we describe key assumptions in the CHARMM FF, but the underlying arguments apply to the others. The equations that define the CHARMM FF (Eqn. 4.1) assume that the potential energy of the
system can be split into bonded and nonbonded interactions. In this model, nonbonded interaction parameters (Eqn. 4.2 and 4.3) for LACDs may be found without problems using standard procedures and are not referenced further. Bonded terms (Eqn. 4.4) include contributions from bond, angle, dihedral, improper and Urey-Bradley intramolecular potentials.

No Urey-Bradley nor improper angles exist in linear moiety structures so these contributions to the energy are not discussed further. Additionally, bond potentials did not present difficulties during parameterization, so they too are not examined in this work. Rather, the focus is on the angle and dihedral contributions to the potential energy function. To review:

\[
V_{\text{angles}}(\theta) = \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2
\]

\[
V_{\text{angles}}(\phi) = \sum_{\text{dihedrals}} K_\phi (1 + \cos(n\phi - d))
\]

where \( K_\theta \) and \( K_\phi \) are constants and \( \theta_0 \) is the equilibrium value in angle space. The dihedral potential uses a cosine expansion with a multiplicity (n) and phase shift (d) to model the periodic nature of the dihedral interaction. These potential are assumed to be independent from each other and thus changes angle space do not affect the dihedral potential.

Molecules that contain LACDs present two significant problems for simulations using the CHARMM FF: (1) structural kinetic energy at or near angles of 180° produces an unrealistic angle distribution and (2) dihedral and angle modes couple in a very stiff region of dihedral space which causes discontinuous dihedral movements that artificially add energy to the system (“ramping”). Coupling is defined to mean changes in one parameter (e.g. angle) results in a change in energy of another bonded contribution (e.g. dihedral), which is directly opposite the assumption of independence.

8.3 Methods: Simulation

8.3.1 Model

A 4-bodied model, depicted in Figure 8.1, 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 $\pm 0.001 \text{ Å}$ in any direction during simulation. The entire model still uses the CHARMM FF as defined in Equation 2, and these changes are only used to help visualize the relevant motion. Specifically, using this set-up allowed the $BC$ bond to be fixed along the z-axis and the $BCD$ plane to be restricted to the yz-plane.

Figure 8.1: 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 $\Theta$ 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 8.1. Notice that movements in angle space manifest as changes in $r$ and movements in dihedral space manifest as changes in $\Theta$. 

90
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 \(AB\) bond, \(ABC\) angle, and \(ABCD\) dihedral. \(K_b\) and \(b_0\) parameters were set equal to 200 kcal mol\(^{-1}\) \(\text{Å}^{-2}\) and 1.0 Å, respectively. \(K_\theta\) was set equal to 120 kcal mol\(^{-1}\) rad\(^{-2}\). \(\theta_0\) values were 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. \(K_\phi\) 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.

### 8.3.2 Simulation Protocols

Multiple molecular dynamics simulations were performed in this study using the LAMMPS software [142]. Unless otherwise specified, 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 40000 timesteps.

#### Recreation of Simulation Ramping

The first set of simulations demonstrate that the ramping phenomenon seen in Chapter 7 for azide and alkyne moieties is found in any model involving linear moieties – including the one used in this work. All 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. Simulations were then run for 0.1 ns of equilibration and 0.5 ns of production to ensure either equilibrium was reached or the program crashed due to numerical overload caused by ramping of the energy.

#### LACD Influence on Structure

The second set of simulations were done to observe the effects of LACDs on molecular structure over a longer period of time. As mentioned above, ramping produces numerical problems, but the effects can be eliminated using an NVT ensemble because the thermostat effectively removes
the energy added to the system to maintain the temperature at the set point value. These NVT simulations were done at 300K with 10 Nosé-Hoover thermostats. Variation between simulations are defined in the \( \theta_0 \) parameter whose values vary between 130° and 180°. 

Linearity was tracked during simulation to produce a histogram of angles. Additionally, the x,y positions of Site A were periodically saved 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 the \( \overline{BC} \) bond on the z-axis (see Figure 8.1), positive y values for the position of A indicate cis dihedral preference while negative y values correspond to trans configurations.

**Short-Time Simulation**

A third set of simulations were performed to show how angle and dihedral motions couple in linear structures by inducing a coupling event and comparing paths taken by simulations using a 1.0 fs and 0.01 fs timestep. The former is a common timestep frequently used in MD simulation, and the latter utilizes a very small timestep to reduce numerical instabilities and is used as a comparison. Because the coupling events occur quickly, these simulations are only run for 20 fs and use the NVE ensemble. Such a short time is sufficient because the purpose of these simulations is to explain and demonstrate that coupling between different bonded modes occurs and that it produces numerical instabilities.

Since the purpose of these simulations was to focus on how coupling occurs, it was desirable to confine movement in the system to only Site A to ensure post-simulation analysis was not affected by movements of sites B, C, or D. This was done by fixing these latter sites to their initial positions by setting the forces on these molecules to 0 at every timestep. The initial configuration was a cis structure that was at dihedral equilibrium and angular non-equilibrium. Specifically, \( \theta = 174° \), \( \theta_0 = 180° \), and \( \phi = \phi_0 = 0° \) so that the initial force is positive in angle space and directed toward \( \theta_0 \). This setup is identical to what is depicted in Figure 8.1. A is given an initial velocity in angle space, 0.02 Å s\(^{-1}\), which is observed as \(-y\) velocity in this initial set-up. The system also starts with a small amount of dihedral space velocity, 0.0002 Å s\(^{-1}\) manifested as \(+x\) velocity, and is needed to initiate coupling. This initial configuration, along with confining the movement of \( B-D \), confines phase space to the x,y plane to easily follow the movement of the system. Additional
post-simulation analysis performed on the data is used to compare theoretical movements in angle and dihedral space. The theory and equations used for this analysis is outlined in Appendix B.

**Removing LACDs in Simulation**

The final set of simulations demonstrate how to eliminate ramping and generate stable simulations for systems containing linear structures. These NVE simulations set $K_\phi$ equal to 0 to eliminate the LACD contribution to potential. Variation between treatment simulations is defined in the $\theta_0$ parameter whose values vary between $130^\circ$ and $180^\circ$. Linearity and structure is again identified using histograms and heatmap plots to determine the effect removing LACD parameters has on MD movement.

8.4 **Results**

8.4.1 **Ramping in NVE Simulations**

As explained previously, the first goal of this paper was to recreate the numeric instability when simulating LACD, but with a model simple enough to isolate movement yet detailed enough to produce ramping. When simulating azides and alkynes in the previous chapter, there were repeated spikes in the PE (and thus total energy) that would continue until the simulation crashed from numerical overflow. Figure 8.2 contains the results of NVE simulations using the simplified model from this work. Panel A shows the data for $\theta_0 = 130^\circ$, Panel B for $\theta_0 = 178^\circ$, and Panel C for $\theta_0 = 180^\circ$. The potential energy as a function of simulation time is plotted in each case.

Using the same simulation set-up as the azido and alkynyl small molecules, we were able to achieve ramping for both linear variants of our model ($\theta_0 = 178^\circ$ and $180^\circ$). The simulation experiences multiple dihedral potential spiking events that translate to increases in PE of the system.

Notice that for a “normal” angle of $130^\circ$ (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 $178^\circ$ and $180^\circ$, the potential energy spikes at various times and increases dramatically. As will be shown later, these spikes occur in the dihedral potential. Changing the magnitude of $K_\phi$ in an LACD does nothing to eliminate ramping in the system.
Only as $\theta_0$ became more non-linear does the number of PE spike events decrease. In other words, the more linear the $\theta_0$ value, the more frequent destabilizing events occur.

### 8.4.2 NVT Simulations with LACD Parameters

We next observe the structures formed from temperature-controlled simulations with LACDs and, consequently, observe how discontinuous coupling events affects molecular structure. To maintain consistent structures and avoid run-away structures, the NVT ensemble is used with multiple thermostats to ensure only structures at 300 K are sampled.

Figure 8.3 contains the results from multiple simulations for different values of $\theta_0$ between $130^\circ$ and $180^\circ$. Two plots are shown for each $\theta_0$ value tested: a histogram of angles sampled (top graph for each angle) and an x,y heatmap plot of the positions of Site A from the model (bottom plot). Because model parameters establish a cis dihedral structure, the heatmap plot observes this
Figure 8.3: NVT plots of the angle distribution and 2D histogram of A’s x,y position in space using LACD parameters.

as more sampling of positive y-values. The top set of simulations shows results for cases where no ramping events occur and for which NVE simulations produced stable trajectories. The bottom set shows results from simulations that experience ramping and were numerically unstable in an NVE ensemble.

We first analyze the top row of simulations that show no signs of ramping. Notice that for all simulations where $\theta_0 < 170^\circ$, angle histogram plots produce a Gaussian distribution of values centered on the $\theta_0$ 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 $\theta$ values centered on $\theta_0$ 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 would exhibit numeric instability if the NVE ensemble were used, (bottom row from Figure 8.3) 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 $\theta_0 \approx 170^\circ$. 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 conformations. Thus, in the transition region, ramping causes trans structures to be disproportionately sampled.

Beyond the transition region ($\theta_0 \approx 180^\circ$), 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 $\theta_0$ values, angle sampling no longer centers on $\theta_0$, 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. Moreover, as will be described in the next section, the dihedral and angle potentials become coupled in a very stiff region of dihedral space – a condition also contrary to the assumptions found in modern force fields.

8.4.3 Examining a Coupling Event

In order to show how angle and dihedral potentials are coupled when $\theta_0$ approaches $180^\circ$, multiple short simulations in which coupling events occur are now described step by step. The initial positions are such that the system is at dihedral equilibrium ($\phi = \phi_0$) while the angle is in nonequilibrium ($\theta \neq \theta_0$). The initial velocities are set to force a coupling event to occur. The simulation duration and select simulation values are reported so that the effects of coupling can be viewed by the reader. These results reveal that the origins of energetic ramping is caused by coupling between the angle and dihedral modes.
Three types of NVE simulations were done, and coupling occurs in each. The first was run with a 1.0 fs timestep which was chosen because it is a common timestep used in all-atom simulation. The second type of simulation uses a very small timestep to show how smooth trajectories can be generated. This second simulation uses a 0.01 fs timestep. The final simulation uses a 1.0 fs timestep and excludes the LACD contribution to the potential energy in an effort to eliminate ramping. The first two will be analyzed together, followed by a discussion of the third.

Figure 8.4: Simulations using two different timesteps. The green dot indicates the starting position and the black triangle indicates the origin and represents the equilibrium $\theta_0=180$ set point. Panel A has data for simulations using a 1.0 fs and shows spacial (top) and dihedral space (bottom) movements. Panel B contains plots using a 0.01 fs timestep and similarly plots spacial (top) and dihedrals space (bottom) movements.

**Trajectories When Including LACD Parameters**

Figure 8.4 shows the x,y position of Site A for the first two cases (Panel A is for the 1.0 fs timestep and Panel B is for the 0.01 fs timestep). Each case was simulated for 20 fs. The top plots in each panel are the position of Site A in space and the bottom plots show how this
positional movement correlates to movement in dihedral space. In Panel A, x,y position data is plotted over two different scales. The left plot better identifies the discontinuous movements while the larger scale on the right emphasizes the small movements involved in the trajectory (fractions of an angstrom). The red points located along the trajectories are at 1 fs intervals. The green dot indicates the starting position and the black triangle is the origin. Parameters are set such that the equilibrium point the system is trying to achieve is the origin \(\theta_0 = 180^\circ\).

The first thing to notice in Figure 8.4 is that the systems follow the same trajectories for both timesteps (Panels A and B) until 4 fs have passed and Site A is closest to the origin. At this time the 0.01 fs simulation experiences a relatively small impulse in its trajectory but movements remain smooth. Conversely, the 1.0 fs simulation experiences a sharp change in trajectory and shoots off in an unexpected direction for the duration of the simulation. It is this break in smooth movements that causes numeric instability within simulation.

To help understand how coupling of different bonded modes affects the simulation and why this instability occurs, it is useful to calculate where a system should move based on the individual forces in the system compared to where it eventually ends up. Ideally, these should be the same, but coupling induces forces cause the system to ramp in energy and move away from rather than towards the equilibrium angle and dihedral values. The result is that the system does not end up where it is expected to be.

Tables 8.1 and 8.2 report angle \(\theta_{ABC}\) and dihedral \(\phi_{ABCD}\) numerical data for the 1.0 fs and 0.01 fs timestep simulations. The purpose of these tables is to show where angles and dihedrals are expected to be based on their current position and where they actually end up after a simulation iteration. If the motions of the angles and dihedrals are independent (as is often assumed) then the calculated and simulated values should be the same (to within numerical error) because the angle is only affected by the angle potential and the dihedral only by its own potential. Table 8.1 contains data from the 1.0 fs timestep simulation for the first 10 iteration or a total of 10 fs in time. Table 8.2 contains data from the 0.01 fs timestep simulation. However, only select values from this data set are reported that are relevant to the concepts presented.

The columns in each table are as follows. The first column reports the current iteration’s total simulation time. The second column is the current value of the angle or dihedral (Cur. \(\theta\) and Cur. \(\phi\)). The third column is the Simulation \(\Delta\theta\) or Sim. \(\Delta\phi\) data. These values are the changes in
that occur in simulation as a result of all the Cartesian forces and velocities in the system. Thus adding a given iteration’s Cur. and Sim. Δ values together will generate the next iteration’s Cur. value. The values in these columns are directly taken from simulation and are susceptible to coupling.

The fourth column calculates changes in angle and dihedral space that should occur if the system were independent. These values transform the current iteration’s Cartesian velocity and force vectors into angle or dihedral space, and then performs a single Eulerian integration to calculate the theoretical Δθ or Δφ (see Appendix B). Thus, columns 3 and 4 will be the same if the motions of angles and dihedrals are independent; if different, they are coupled.

Table 8.1: Step-by-step analysis of coupling and discontinuities using a 1.0 fs timestep.

<table>
<thead>
<tr>
<th>Time (fs)</th>
<th>Cur. θ (°)</th>
<th>Sim. Δθ (°)</th>
<th>Calc. Δθ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>174.106</td>
<td>1.326</td>
<td>1.355</td>
</tr>
<tr>
<td>1</td>
<td>175.432</td>
<td>1.431</td>
<td>1.428</td>
</tr>
<tr>
<td>2</td>
<td>176.863</td>
<td>1.487</td>
<td>1.474</td>
</tr>
<tr>
<td>3</td>
<td>178.350</td>
<td>1.418</td>
<td>1.487</td>
</tr>
<tr>
<td>4</td>
<td>179.768</td>
<td>-1.020</td>
<td>0.873</td>
</tr>
<tr>
<td>5</td>
<td>178.748</td>
<td>-1.308</td>
<td>-1.350</td>
</tr>
<tr>
<td>6</td>
<td>177.440</td>
<td>-1.213</td>
<td>-1.297</td>
</tr>
<tr>
<td>7</td>
<td>176.227</td>
<td>-1.091</td>
<td>-1.209</td>
</tr>
<tr>
<td>8</td>
<td>175.136</td>
<td>-0.954</td>
<td>-1.096</td>
</tr>
<tr>
<td>9</td>
<td>174.182</td>
<td>-0.808</td>
<td>-0.963</td>
</tr>
<tr>
<td>10</td>
<td>173.374</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table 8.1 – *Continued from previous page*

<table>
<thead>
<tr>
<th>Time (fs)</th>
<th>Sim. φ (°)</th>
<th>Sim. Δφ (°)</th>
<th>Calc. Δφ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.633</td>
<td>0.051</td>
</tr>
<tr>
<td>1</td>
<td>0.633</td>
<td>1.183</td>
<td>0.037</td>
</tr>
<tr>
<td>2</td>
<td>1.816</td>
<td>3.225</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>5.041</td>
<td>47.544</td>
<td>0.094</td>
</tr>
<tr>
<td>4</td>
<td>52.585</td>
<td>120.364</td>
<td>1.166</td>
</tr>
<tr>
<td>5</td>
<td>172.949</td>
<td>4.293</td>
<td>-0.137</td>
</tr>
<tr>
<td>6</td>
<td>177.242</td>
<td>1.367</td>
<td>-0.030</td>
</tr>
<tr>
<td>7</td>
<td>178.609</td>
<td>0.692</td>
<td>-0.003</td>
</tr>
<tr>
<td>8</td>
<td>179.301</td>
<td>0.432</td>
<td>-0.015</td>
</tr>
<tr>
<td>9</td>
<td>179.733</td>
<td>0.305</td>
<td>-0.026</td>
</tr>
<tr>
<td>10</td>
<td>180.038</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.2: Step-by-step analysis of coupling using a 0.01 fs timestep.

<table>
<thead>
<tr>
<th>Time (fs)</th>
<th>Cur. θ (°)</th>
<th>Sim. Δθ (°)</th>
<th>Calc. Δθ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>174.106</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>1.00</td>
<td>175.428</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>2.00</td>
<td>176.857</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>3.00</td>
<td>178.342</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>4.09</td>
<td>179.815</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*Continued on next page*
Table 8.2 – *Continued from previous page*

<table>
<thead>
<tr>
<th>Time (fs)</th>
<th>Cur. $\theta$ (°)</th>
<th>Sim. $\Delta\theta$ (°)</th>
<th>Calc. $\Delta\theta$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.10</td>
<td>179.816</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4.11</td>
<td>179.816</td>
<td>-0.002</td>
<td>-0.001</td>
</tr>
<tr>
<td>4.12</td>
<td>179.814</td>
<td>-0.003</td>
<td>-0.002</td>
</tr>
<tr>
<td>7.00</td>
<td>176.107</td>
<td>-0.011</td>
<td>-0.013</td>
</tr>
<tr>
<td>9.00</td>
<td>174.002</td>
<td>-0.009</td>
<td>-0.010</td>
</tr>
<tr>
<td>10.00</td>
<td>173.172</td>
<td>-0.008</td>
<td>-0.009</td>
</tr>
</tbody>
</table>

**Dihedral Data**

<table>
<thead>
<tr>
<th>Time (fs)</th>
<th>Cur. $\phi$ (°)</th>
<th>Sim. $\Delta\phi$ (°)</th>
<th>Calc. $\Delta\phi$ (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.000</td>
<td>0.005</td>
<td>0.001</td>
</tr>
<tr>
<td>1.00</td>
<td>0.630</td>
<td>0.009</td>
<td>0.000</td>
</tr>
<tr>
<td>2.00</td>
<td>1.805</td>
<td>0.017</td>
<td>0.000</td>
</tr>
<tr>
<td>3.00</td>
<td>4.992</td>
<td>0.061</td>
<td>0.001</td>
</tr>
<tr>
<td>4.09</td>
<td>84.212</td>
<td>4.561</td>
<td>0.016</td>
</tr>
<tr>
<td>4.10</td>
<td>88.773</td>
<td>4.560</td>
<td>0.017</td>
</tr>
<tr>
<td>4.11</td>
<td>93.333</td>
<td>4.503</td>
<td>-0.013</td>
</tr>
<tr>
<td>4.12</td>
<td>97.836</td>
<td>4.393</td>
<td>-0.013</td>
</tr>
<tr>
<td>7.00</td>
<td>176.046</td>
<td>0.009</td>
<td>0.001</td>
</tr>
<tr>
<td>9.00</td>
<td>177.113</td>
<td>0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>10.00</td>
<td>177.400</td>
<td>0.003</td>
<td>-0.000</td>
</tr>
</tbody>
</table>
Details of a Coupling Event

The first thing to notice about the data reported in Tables 8.1 and 8.2 is that differences exist between the expected and actual changes in $\theta$ and $\phi$. The effect is more pronounced in Table 8.1, so the discussion will focus here, but the same phenomenon is observable in the other simulation as well.

From the very start of the simulation a difference exists between the actual change in angle space and the calculated change in angle space. The first step predicts a change in angle space to be $+1.383^\circ$, but the simulation produces a change of $+1.326^\circ$. The opposite happens in dihedral space because angle space adds energy to dihedrals.

Minor breaks in independence is expected because these types of FFs have modes that are not completely independent for such short simulations [80]. Angle motion is lost to dihedral space, and dihedral space gains an equal amount of energy and thus total energy is conserved. From a practical standpoint, when the difference between the expected and actual changes in $\theta$ and $\phi$ are smaller than the error associated with the overall simulation errors (FF + numeric), the ensemble averages can considered accurate [81]. However, as the system approaches angular equilibrium (identified as the triangle at the origin in Figure 8.4) the amount of coupling increases. Multiple magnitudes of disagreement in dihedral space is concerning because it represents a significant deviation from actual behavior beyond normal error.

What is important to notice is how insignificant the dihedral potential affects dihedral space. We use the first 4 fs of this 1.0 fs timestep simulation as an example. In these first 4 fs, compare Simulation $\Delta \phi$ to the Calculated $\Delta \phi$ values. In the time between 3 and 5 fs only two iterations occur in simulation, yet the dihedral experiences over $167^\circ$ of movement. Calculated changes in dihedral space were much smaller and in the opposite direction during this time. These numbers suggest dihedral space is, in addition to being coupled, very stiff when sampling molecular configurations near $180^\circ$. Said another way, a very small change in angle space results in a very large change in dihedral space when structures are very close to $180^\circ$. At all steps, the Calculated $\Delta \phi$ values are a magnitude or more smaller than the Simulation $\Delta \phi$ values. This suggests significant movement in dihedral space that was not caused by the dihedral potential.

Moreover, the changes in Sim. $\Delta \phi$ values continually increase. This causes unphysical acceleration in dihedral space during this time. In other words, dihedral space momentum *increases*
in a direction opposite the direction of the dihedral force in a system without any nonbonded forces. In a truly independent FF there is no situation where a change in dihedral space velocity increases in a direction opposite its potential. For such a situation to occur there must be some other force influencing the potential. In this model and with the initial conditions specified, the only parameter that could influence the dihedral is the angle potential. Thus significant coupling between the two modes has occurred. Because this effect is observed in both long and short timestep simulations, decreasing timestep size does not eliminate significant coupling within linear moiety systems. Coupling in both cases cause path deviations inconsistent with real movement and show that effects of unstable coupling cannot be avoided even with very small timesteps.

**Discontinuous Trajectories**

Discontinuities in position only occur in LACD-containing systems with long timesteps and result from the significant coupled movements in this regime. Using a longer timestep, and as \( \theta \) approaches 180°, the system experiences what is called a “node flip” in angle space. A node flip occurs because the computer/compiler defines angle space to be between 0-180°. At these extremes, any force that pushes angle space beyond the extrema will be perceived incorrectly by the computer and the forces will push the system farther from equilibrium rather than closer. For example, in Table 1 at the 4 fs time marker, the change in theta space should be +0.873° (Calc. \( \Delta \theta \)). Assuming complete independence, this would push the new \( \theta \) value to be 180.641°. This value is beyond 180° and would be interpreted by the computer to be 179.359°. Thus a positive change in angle space is instead interpreted by the computer as negative movement and a discontinuous traversal of angle potential where \( A \) is now located on the opposite side. We also note this node is inherent to the FF and causes the skewed angle potentials shown in Figures 8.3 and 8.6 as repeated incorrect sampling events skew the distribution. To summarize, what should have been a positive movement in angle space is instead interpreted as negative motion in simulation, and the computer perceives the position of \( A \) to be on the opposite side of the angle potential. Because this system is heavily coupled, this discontinuous behavior in angle space also affects dihedral space.

Node-flipping is concerning for coupled motion because the motion (and thus energy and momentum) imparted into dihedral space (from angle space) is of different magnitude and in the opposite direction. This causes a significant dihedral response opposite the direction of motion and
non-smooth movements in space result. Thus, incorrect and highly coupled potential calculations artificially inflate system energy. The effect of one impulse can lead to additional compounding non-smooth events and a break-down in the conservation of energy expected in NVE simulations. As multiple non-smooth events build upon each other, and as more energy artificially enters the system through these unphysical movements, the system energy will increase until the program crashes from numeric overload.

Using a small timestep does not exempt LACD simulations from experiencing dihedral space discontinuities. Close examination of the path plot in Figure 8.4 Panel B shows a mild impulse in movement when the simulation is closest to the origin (the minimum of the system). Table 8.2 indicates that as $\theta$ approaches $180^\circ$, coupling between angle and dihedral spaces predominates and significant differences occur between the calculated and simulated changes in dihedral space. These movements are non-smooth in dihedral space as evidenced by the large changes in the dihedral relative to the timestep as seen in the bottom plot of Panel B of Figure 8.4. These dramatic changes in dihedral space affect the path of Site A, represent significant deviation from real behavior, but do not cause the simulation to crash because node-flips in angle space do not occur. They do, however, cause path deviations inconsistent with real movement and shows that effects of coupling cannot be avoided even with very small timesteps.

**Removing the LACD Parameter**

Because angle and dihedral space are coupled for LACDs, new approaches are needed to simulate these moieties in molecular simulation. One approach is to parameterize the coupled motion into a combined intramolecular motion that is dependent on $\theta$ and $\phi$ simultaneously. This approach has two problems. First, the data needed for such a parameterization is not available in the literature as such moieties have not been a focus of study previously and would be time-consuming to generate from first-principles. Second, most simulation programs would need new coding to accomplish the task. Both of these problems could be overcome with time and effort, but an alternate option is now proposed which can largely supply the simulation community with an immediate and adequate solution without such effort. This approach is to remove the dihedral potential for LACDs completely by simply setting $K_\phi = 0$ for any LACD, and is similar to the Rational Design approach used in Chapter 7.
Figure 8.5 Panel A) NVE simulation with $K_\phi h\dot{\theta} = 0$. Panel B) All three simulations from this section.

Figure 8.5 Panel A shows the x,y position of Site A for a simulation with such an approach ($K_\phi = 0$). This path is void of discontinuous movement and doesn’t show any signs of coupled behavior. Figure 8.5 Panel B plots all 3 simulations described in this section on the same plot for comparison. The paths for all three are identical up to 4 fs, but once a near-linear configuration is sampled in either of the LACD-containing simulations, the system experiences unphysical movements as the discontinuities in dihedral space affect position. Only the $K_\phi = 0$ LACD assumption prevents discontinuous movements in position or dihedral.

Table 8.3 shows the angle space computations for the $K_\phi = 0$ simulation. Notice how the simulated and calculated changes are still not identical (similar to the data in Table 8.1). However, because the LACD contribution to molecular potential is zero, coupled motion in dihedral space and thus ramping in dihedral potential is prevented. What is not avoided is coupled motion in angle space. This can be seen as movements in angle space being lost to dihedral space movements. This results in a portion of angle potential being lost to a null dihedral space and thus is eliminated from the system. This loss of movement to null dihedral space causes minor energy loss to the system.
and energy conservation is not ensured. To be clear, eliminating the LACD parameter does not eliminate coupling in the system, but unstable numerical issues are avoided.

Table 8.3: Step-by-step analysis of coupling and discontinuities using a 1.0 fs timestep and excluding the LACD parameter.

<table>
<thead>
<tr>
<th>Simulation Data Using a Timestep of 1.0 fs and $K_\phi = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle Data</td>
</tr>
<tr>
<td>Sim. Time (fs)</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

The result of losing dihedral influence on structure will cause minor differences between the null dihedral space assumption and the theoretical “real” movement that would occur in the presence of a dihedral potential. In practice however, the differences are negligible when simulating large molecules using CHARMM or CHARMM-like force fields. In other words, the inaccuracies introduced by setting $K_\phi = 0$ for the practical applications of this model, like uAA simulation,
are acceptable and this approach produces a better approximation of molecular response than any alternative.

8.4.4 NVE Simulations without LACD Parameters

The previous discussion used the NVE ensemble to highlight the problems encountered when simulating an LACD coupling event. The compromises of setting $K_\phi = 0$ in the NVE ensemble are now explored in more depth. Results are in the form of histograms of angle values and heat maps of the x,y positions (dihedral values) similar to our previous results when running NVT simulations that include LACD parameters.

Figure 8.6: Plots of the angle distribution and 2D histogram of A’s x,y position in space without LACD parameters.
Figure 8.6 shows results from the simulations at various values of $\theta_0$ 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 8.6, the angle space histograms appear relatively unchanged from those shown in Figure 8.3 when LACDs were included. This is important because it reveals that the GAFF and CGenFF assumption about LACDs contains inherent flaws. These skewed histograms indicate that the system is not adhering to the Hooke’s law assumption. Because of this, these widely-used databases are producing angle structure that do not replicate actual behavior in the linear regime. This will result in linear angle parameters with larger errors, and these errors have not been noticed or addressed due to the parameterization design of these general force fields [70, 72].

In the Non-Linear region the dihedral space heatmap plots in Figure 8.6 show that both cis and trans structures are equally probable. This is undesirable because dihedral influence is important for molecular structures at these angles. However, this result is expected because the lack of a dihedral parameter will result in no bias for cis or trans configurations. For any real system these dihedral parameters would be included.

As $\theta_0$ 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°, but does not produce a skewed angle distribution. When $\theta_0 = 170°$, the probability band dissolves into the Transition region to produce an angle space prediction well. The onset of this Transition region is important because it identifies the limit for stable simulations that include dihedral parameters, the $\theta_0$ parameter where significant coupling events first occur, and when $K_\phi$ should be set equal to 0.

All simulations in the Transition and Linear Angle region produce reasonable structures expected for a linear system. These 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 $K_\theta$ values are appropriate for the system. Thus, dihedral influence on structure is compromised for a stable LACD-containing system. As explained previously, the validity of this approach for practical application was explored in the previous chapter.
8.5 Conclusion

This research provides analysis on the modeling and simulation of linear or near-linear angle molecular moieties. Simulations of these structures reveal two significant inconsistencies within typical force field theory. First, sampling of linear angles do not center on the \( \theta_0 \) value due to the numerical limitations that restrict angles to be between 0° and 180°. Second, the dihedral force field significantly couples with angle parameters in a very stiff region of dihedral space. As a result, coupling introduces both steady and discontinuous contributions to total energy that numerically destabilize the simulation. Effects of this coupling are unavoidable when using a FF that assumes angle and dihedral independence.

Additional short-time simulations juxtaposed molecular models with and without LACD parameters and found that LACD discontinuities significantly disrupt MD simulations. It was shown that the independence assumption is not valid for linear moieties and angle space significantly influences dihedral movements. Decreasing timestep does not eliminate the problem. Simulations without LACD parameters are the best approximation of actual behavior at the present time. However, potential negative ramping could result as angle space couples and looses energy to a null dihedral potential.

In the absence of better force fields, there must be a choice between NVE stability and dihedral influence. The most concerning observation from this work 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. It is our recommendation that if linear structure appears in simulation, that the dihedral potential be excluded from energetic calculations to eliminate ramping from the system.
9.1 Future Work for Protein Devices

The analysis from the TEM-1 screens in Chapter 6 proposed a heuristic to explore all experimentally accessible sites for functionalization. The basis for this heuristic was using domain analysis to identify locations that could be good for functionalization and then screening these sites for experimental testing. The next step in exploring the concept of domain-directed functionalization site selection would be to experimentally functionalize TEM-1 using the recommended sites in this screen. Such a study would allow for a direct comparison between the in silico theoretical work and the experimental protein device effectiveness.

This work is the first time PEGylated or tethered beta sheet regions have been screened in a coarse grain model for the purpose of identifying optimal functionalization sites. More work is needed that tests functionalization on beta sheets (or structures in close proximity to beta sheets) to identify if domains containing beta sheets correlate with unstable protein devices, or if the poor performances seen by the beta-containing domain is isolated to TEM-1. More research is needed that expands the scope of proteins that have been screened for functionalization effectiveness. Specifically, testing a diverse range of tertiary structures classes would improve our understanding of functionalization on protein tertiary structure.

9.2 Future Work for LACD Parameterization

This next section of future work centers on the compromises made during LACD parameterization in Chapters 7 & 8. To review, in order to simulate an energy-stable linear moiety we have to remove the LACD parameter. The consequence of this is three-fold: (1) a loss of dihedral influence to the model and (2) a potential for negative ramping as angle space loses energy to a null dihedral space. Moreover, (3) the solutions given in this work do not resolve the fact that the angle
potential does not adhere to the Hooke’s law and a skewed angle distribution still persist. Thus the focus of this section is in addressing these issues.

9.2.1 LACD Parameterization Using Urey-Bradley Potentials

One assumption made during the parameterization process was that Urey-Bradley contributions did not exist for these moieties. This still remains true as they are not present or needed in non-aromatic moieties, but lets examine the potential anyways. Urey-Bradley (UB) interactions are a 1-3 harmonic potential and can be interpreted as a spring between atoms A and C in the ABC angle:

\[ V_{Urey-Bradley}(S) = \sum_{UB} K_S (S - S_0)^2 \]  (9.1)

where \( S \) is the distance between an atom and another atom two positions away. The harmonic value \( S_0 \) is the equilibrium distance and \( K_S \) is the spring constant. The primary use for UB parameters is to maintain a rigid aromatic structure for molecules like benzene, but it could have a use in an LACD setting.

As has been shown in this work, the angle potential does not conform to Hooke’s law, and with no way of overcoming the “node-flip” problem, there is very little theoretical use for the angle potential. It does create structures that resemble the ab initio structures, but the additional negative ramping risk makes it a questionable choice. Similar to what was done with the LACD potential, assume the angle potential is also zero because it inherently-contains numeric inconsistencies. Instead, structure can be maintained with a new use of the UB potential. The new use holds atoms 1-3 at a bond distance equal to what would be seen at an angle of 180°. Because this is linear, the UB potential essentially creates an additional bond that holds the linear structure by setting-up a bonded matrix between the \( \overline{AB} \), \( \overline{BC} \) and \( \overline{AC} \) bonds. Because \( S_{0,AC} \approx b_{0,AB} + b_{0,BC} \) and if \( K_{S,AC} \) is strong enough to limit significant angular motion, then the theory is that linear structures results without any energetic consequences and adheres to Hooke’s law.

The parameter \( S_{0,AC} \) is easily determined using ab inito calculations, but the main hurdle is in developing a method to parameterize the \( K_{S,AC} \) parameter. Because this is a new implementation
of the UB potential, such a method does not exist and would require a new approach. More work needs to be done to examine the feasibility of using UB-potentials for LACD systems.

9.2.2 LACD Parameterization Using Unique Angle Potential

The most interesting concept from the UB potential is the proposal that skipping atoms in sequence can maintain proper structure. In addition to harmonically bonding the 1-3 atoms, a completely new use of an atom-skipping potential would be using a 1-3-4 angle potential. This has never been done before, but could theoretically introduce dihedral influence to the a linear system that currently has no dihedral bias. The new 1-3-4 angle creates an angle that will hold to Hooke’s law and maintain dihedral influence on linear structure. However, removing the 1-2-3 angle potential could produce unexpected structures or unreasonable structure bias. More work needs to be done to examine the feasibility of this option. Another consideration if this method is implemented, must look at two consecutive sp\(^1\) hybridized orbitals (as is present in alkynes). This system would contain a 1-3-4 angle that would be linear, thus defeating the purpose of the new potential.

These atypical implementations of atom-skipping would require significant research on making sure the parameters maintained theoretical significance and structural consistence. However, if done properly, it could be a new method that resolves all issues with LACD parameterization and does not require additional coding or FF potentials.
REFERENCES


[34] C Dhalluin, A Ross, LA Leuthold, S Foser, B Gsell, F Muller, and H Senn. Structural and biophysical characterization of the 40 kDa PEG-interferon-alpha(2a) and its individual positional isomers. *Bioconj. Chem.*, 16(3):504–517, MAY-JUN 2005. 10


115


[80] Pnina Dauber-Osguthorpe and A. T. Hagler. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? J. Comput. Aid. Mol. Des., 33(2), 2019. 17, 26, 102


[133] Mizuguchi Laboratory. 2018. 36, 43


123


The pymol molecular graphics system, version 1.2r3pre. Schrodinger, LLC. 66


Benjamin Schulze and Ulrich S. Schubert. Beyond click chemistry - supramolecular interactions of 1,2,3-triazoles. *Chem. Soc. Rev.*, 43(8), 2014. 73

Fang-Fang Chen and Feng Wang. Electronic structure of the azide group in 3 ‘-azido-3 ‘-deoxythymidine (azt) compared to small azide compounds. *Molecules*, 14(7), 2009. 73

K FUKUI. Role of frontier orbitals in chemical-reactions. *Science*, 218(4574), 1982. 73


R Huisgen. 1.3-dipolare cycloadditionen - ruckschau und ausblick. *Angew. Chem. Int. Edit.*, 75(13), 1963. 88


A Brik, J Alexandratos, YC Lin, JH Elder, AJ Olson, A Wlodawer, DS Goodsell, and CH Wong. 1,2,3-triazole as a peptide surrogate in the rapid synthesis of hiv-1 protease inhibitors. *Chembiochem*, 6(7), 2005. 88
The purpose the work in Chapter 6 was to develop a rapid computer screen to identify sites to functionalize on TEM-1 to improve stability and activity. Many metrics and properties were calculated to attempt to correlate structure to stability, but not all of these resulted in an observable correlation. Moreover, for brevity, the body of the paper did not present results for all the residues where functionalization could occur even though each was analyzed. In this appendix all sites and rankings are shown. These tables track primary and secondary structure influences that were examined as a possible heuristic for a priori site selection.

The last part of this appendix examines how protein orientation is affected by tether placement in more detail. In terms of the purposes of this paper (screening), the main point of this orientation analysis is to show that WT domain analysis alone cannot replace an exhaustive screen because of the intricacies introduced by surface tethering that the PDB structure for the unfunctionalized protein doesn’t capture. This point was made in the main body of the paper, and the additional information presented here is the numerical analysis quantifying the orientation effects.

10.1 Site Rankings from Tethering Screen

Numeric ranking of all sites that were functionalized by either tethering to a hydrophillic surface or PEGylation. Rankings are based on the $\Delta$Tm values as identified from simulations screens and their subsequent heat capacity curves. Three replicate simulations were performed for every site treatment. $\Delta$Tm is averaged across all replicates, and the errors shown are the standard error for the value.
Table 10.1: Surface Tethering quantitative analysis. Numeric ranking of all 89 surface accessible sites.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>$\Delta Tm$ (°C)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>-8.000 ± 2.494</td>
<td>SER</td>
<td>53.6</td>
<td>Loop</td>
</tr>
<tr>
<td>2</td>
<td>88</td>
<td>-13.667 ± 1.247</td>
<td>GLN</td>
<td>71.1</td>
<td>Loop</td>
</tr>
<tr>
<td>3</td>
<td>192</td>
<td>-14.333 ± 1.414</td>
<td>LYS</td>
<td>35.5</td>
<td>alpha</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>-14.667 ± 2.055</td>
<td>VAL</td>
<td>56.6</td>
<td>alpha</td>
</tr>
<tr>
<td>5</td>
<td>227</td>
<td>-15.333 ± 1.414</td>
<td>ALA</td>
<td>94.4</td>
<td>Loop</td>
</tr>
<tr>
<td>6</td>
<td>92</td>
<td>-15.667 ± 2.055</td>
<td>GLY</td>
<td>66.1</td>
<td>Loop</td>
</tr>
<tr>
<td>7</td>
<td>153</td>
<td>-17.000 ± 2.625</td>
<td>HIS</td>
<td>54.4</td>
<td>alpha</td>
</tr>
<tr>
<td>8</td>
<td>154</td>
<td>-17.333 ± 0.000</td>
<td>ASN</td>
<td>94.9</td>
<td>alpha</td>
</tr>
<tr>
<td>9</td>
<td>281</td>
<td>-17.667 ± 1.886</td>
<td>GLU</td>
<td>58.4</td>
<td>alpha</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>-18.000 ± 1.633</td>
<td>LYS</td>
<td>51.5</td>
<td>alpha</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>-18.667 ± 7.587</td>
<td>LEU</td>
<td>20.2</td>
<td>alpha</td>
</tr>
<tr>
<td>12</td>
<td>273</td>
<td>-19.000 ± 2.055</td>
<td>ASP</td>
<td>73.4</td>
<td>alpha</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
<td>-19.667 ± 2.357</td>
<td>PRO</td>
<td>25.4</td>
<td>Loop</td>
</tr>
<tr>
<td>14</td>
<td>156</td>
<td>-19.667 ± 6.342</td>
<td>GLY</td>
<td>59.7</td>
<td>alpha</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
<td>-20.333 ± 2.494</td>
<td>GLU</td>
<td>56.1</td>
<td>alpha</td>
</tr>
<tr>
<td>16</td>
<td>113</td>
<td>-20.667 ± 4.989</td>
<td>LEU</td>
<td>25.7</td>
<td>Loop</td>
</tr>
<tr>
<td>17</td>
<td>115</td>
<td>-21.000 ± 0.816</td>
<td>ASP</td>
<td>76.9</td>
<td>Loop</td>
</tr>
<tr>
<td>18</td>
<td>200</td>
<td>-21.000 ± 1.633</td>
<td>THR</td>
<td>49</td>
<td>Loop</td>
</tr>
<tr>
<td>19</td>
<td>175</td>
<td>-21.000 ± 3.742</td>
<td>ASN</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>20</td>
<td>146</td>
<td>-21.333 ± 3.300</td>
<td>SER</td>
<td>52</td>
<td>alpha</td>
</tr>
<tr>
<td>21</td>
<td>26</td>
<td>-21.667 ± 1.633</td>
<td>HIS</td>
<td>51.1</td>
<td>alpha</td>
</tr>
<tr>
<td>22</td>
<td>38</td>
<td>-21.667 ± 3.300</td>
<td>ASP</td>
<td>63</td>
<td>alpha</td>
</tr>
<tr>
<td>23</td>
<td>41</td>
<td>-22.000 ± 1.700</td>
<td>GLY</td>
<td>78.8</td>
<td>alpha</td>
</tr>
<tr>
<td>24</td>
<td>120</td>
<td>-22.333 ± 4.320</td>
<td>ARG</td>
<td>46.7</td>
<td>alpha</td>
</tr>
</tbody>
</table>

Continued on next page
Table 10.1 – Continued from previous page

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>$\Delta$Tm ($^\circ$)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>39</td>
<td>-22.667 ± 2.055</td>
<td>GLN</td>
<td>61.6</td>
<td>alpha</td>
</tr>
<tr>
<td>26</td>
<td>61</td>
<td>-23.333 ± 2.055</td>
<td>ARG</td>
<td>32.7</td>
<td>beta</td>
</tr>
<tr>
<td>27</td>
<td>188</td>
<td>-23.333 ± 5.437</td>
<td>THR</td>
<td>38.9</td>
<td>alpha</td>
</tr>
<tr>
<td>28</td>
<td>198</td>
<td>-23.333 ± 2.494</td>
<td>LEU</td>
<td>37.7</td>
<td>Loop</td>
</tr>
<tr>
<td>29</td>
<td>96</td>
<td>-23.333 ± 2.944</td>
<td>HIS</td>
<td>71.5</td>
<td>beta</td>
</tr>
<tr>
<td>30</td>
<td>151</td>
<td>-23.333 ± 0.943</td>
<td>PHE</td>
<td>17.9</td>
<td>alpha</td>
</tr>
<tr>
<td>31</td>
<td>258</td>
<td>-23.333 ± 1.700</td>
<td>SER</td>
<td>35</td>
<td>Loop</td>
</tr>
<tr>
<td>32</td>
<td>176</td>
<td>-23.667 ± 6.549</td>
<td>ASP</td>
<td>27.7</td>
<td>Loop</td>
</tr>
<tr>
<td>33</td>
<td>83</td>
<td>-24.000 ± 4.989</td>
<td>ARG</td>
<td>23.1</td>
<td>alpha</td>
</tr>
<tr>
<td>34</td>
<td>267</td>
<td>-24.667 ± 2.160</td>
<td>GLY</td>
<td>58.4</td>
<td>beta</td>
</tr>
<tr>
<td>35</td>
<td>163</td>
<td>-24.667 ± 0.471</td>
<td>ASP</td>
<td>38.8</td>
<td>Loop</td>
</tr>
<tr>
<td>36</td>
<td>86</td>
<td>-25.000 ± 2.944</td>
<td>ALA</td>
<td>59</td>
<td>alpha</td>
</tr>
<tr>
<td>37</td>
<td>191</td>
<td>-25.667 ± 2.625</td>
<td>ARG</td>
<td>29.3</td>
<td>alpha</td>
</tr>
<tr>
<td>38</td>
<td>269</td>
<td>-26.000 ± 1.247</td>
<td>GLN</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>39</td>
<td>51</td>
<td>-26.000 ± 9.274</td>
<td>LEU</td>
<td>8.7</td>
<td>Loop</td>
</tr>
<tr>
<td>40</td>
<td>290</td>
<td>-26.333 ± 7.760</td>
<td>TRP</td>
<td>40.3</td>
<td>alpha</td>
</tr>
<tr>
<td>41</td>
<td>201</td>
<td>-26.333 ± 2.944</td>
<td>LEU</td>
<td>62.8</td>
<td>alpha</td>
</tr>
<tr>
<td>42</td>
<td>165</td>
<td>-26.667 ± 2.055</td>
<td>TRP</td>
<td>37</td>
<td>Loop</td>
</tr>
<tr>
<td>43</td>
<td>202</td>
<td>-26.667 ± 1.700</td>
<td>ALA</td>
<td>69.9</td>
<td>alpha</td>
</tr>
<tr>
<td>44</td>
<td>213</td>
<td>-27.000 ± 0.943</td>
<td>ALA</td>
<td>35.4</td>
<td>alpha</td>
</tr>
<tr>
<td>45</td>
<td>51A</td>
<td>-27.000 ± 3.682</td>
<td>ALA</td>
<td>14.5</td>
<td>Loop</td>
</tr>
<tr>
<td>46</td>
<td>94</td>
<td>-27.333 ± 2.449</td>
<td>ARG</td>
<td>48.9</td>
<td>beta</td>
</tr>
<tr>
<td>47</td>
<td>43</td>
<td>-27.333 ± 3.091</td>
<td>ARG</td>
<td>34.5</td>
<td>Loop</td>
</tr>
<tr>
<td>48</td>
<td>87</td>
<td>-27.667 ± 2.625</td>
<td>GLY</td>
<td>86.4</td>
<td>alpha</td>
</tr>
<tr>
<td>49</td>
<td>177</td>
<td>-27.667 ± 3.859</td>
<td>GLU</td>
<td>59</td>
<td>Loop</td>
</tr>
<tr>
<td>50</td>
<td>288</td>
<td>-28.000 ± 2.494</td>
<td>LYS</td>
<td>81.2</td>
<td>alpha</td>
</tr>
<tr>
<td>51</td>
<td>84</td>
<td>-28.667 ± 5.312</td>
<td>VAL</td>
<td>20.8</td>
<td>alpha</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>$\Delta \text{Tm} , (^\circ)$</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>93</td>
<td>$-29.000 \pm 2.449$</td>
<td>ARG</td>
<td>47.2</td>
<td>Loop</td>
</tr>
<tr>
<td>53</td>
<td>174</td>
<td>$-29.333 \pm 2.055$</td>
<td>PRO</td>
<td>59.9</td>
<td>Loop</td>
</tr>
<tr>
<td>54</td>
<td>100</td>
<td>$-29.667 \pm 8.485$</td>
<td>ASN</td>
<td>100</td>
<td>alpha</td>
</tr>
<tr>
<td>55</td>
<td>182</td>
<td>$-30.000 \pm 4.784$</td>
<td>MET</td>
<td>19.5</td>
<td>beta</td>
</tr>
<tr>
<td>56</td>
<td>161</td>
<td>$-30.667 \pm 3.399$</td>
<td>ARG</td>
<td>31.4</td>
<td>Loop</td>
</tr>
<tr>
<td>57</td>
<td>105</td>
<td>$-30.667 \pm 1.247$</td>
<td>TYR</td>
<td>52.4</td>
<td>Loop</td>
</tr>
<tr>
<td>58</td>
<td>209</td>
<td>$-31.333 \pm 3.682$</td>
<td>ASP</td>
<td>47.1</td>
<td>alpha</td>
</tr>
<tr>
<td>59</td>
<td>143</td>
<td>$-31.333 \pm 2.160$</td>
<td>GLY</td>
<td>64.8</td>
<td>alpha</td>
</tr>
<tr>
<td>60</td>
<td>277</td>
<td>$-31.667 \pm 7.789$</td>
<td>ARG</td>
<td>57.6</td>
<td>alpha</td>
</tr>
<tr>
<td>61</td>
<td>27</td>
<td>$-32.333 \pm 3.091$</td>
<td>PRO</td>
<td>59.2</td>
<td>alpha</td>
</tr>
<tr>
<td>62</td>
<td>141</td>
<td>$-33.333 \pm 2.867$</td>
<td>THR</td>
<td>20.9</td>
<td>alpha</td>
</tr>
<tr>
<td>63</td>
<td>195</td>
<td>$-33.333 \pm 2.160$</td>
<td>THR</td>
<td>46.1</td>
<td>alpha</td>
</tr>
<tr>
<td>64</td>
<td>159</td>
<td>$-33.667 \pm 2.625$</td>
<td>VAL</td>
<td>24.7</td>
<td>Loop</td>
</tr>
<tr>
<td>65</td>
<td>102</td>
<td>$-33.667 \pm 2.055$</td>
<td>LEU</td>
<td>26.2</td>
<td>Loop</td>
</tr>
<tr>
<td>66</td>
<td>158</td>
<td>$-33.667 \pm 3.300$</td>
<td>HIS</td>
<td>68.7</td>
<td>Loop</td>
</tr>
<tr>
<td>67</td>
<td>222</td>
<td>$-33.667 \pm 1.700$</td>
<td>ARG</td>
<td>28.4</td>
<td>Loop</td>
</tr>
<tr>
<td>68</td>
<td>90</td>
<td>$-34.000 \pm 5.558$</td>
<td>GLN</td>
<td>81.2</td>
<td>Loop</td>
</tr>
<tr>
<td>69</td>
<td>110</td>
<td>$-34.667 \pm 10.530$</td>
<td>GLU</td>
<td>62.4</td>
<td>alpha</td>
</tr>
<tr>
<td>70</td>
<td>228</td>
<td>$-36.333 \pm 2.055$</td>
<td>GLY</td>
<td>76.2</td>
<td>Loop</td>
</tr>
<tr>
<td>71</td>
<td>63</td>
<td>$-37.333 \pm 2.494$</td>
<td>GLU</td>
<td>83.6</td>
<td>Loop</td>
</tr>
<tr>
<td>72</td>
<td>205</td>
<td>$-38.667 \pm 1.633$</td>
<td>GLN</td>
<td>47</td>
<td>alpha</td>
</tr>
<tr>
<td>73</td>
<td>54</td>
<td>$-38.667 \pm 1.886$</td>
<td>LYS</td>
<td>54.9</td>
<td>Loop</td>
</tr>
<tr>
<td>74</td>
<td>197</td>
<td>$-39.000 \pm 8.602$</td>
<td>GLU</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>75</td>
<td>56</td>
<td>$-39.667 \pm 2.449$</td>
<td>LEU</td>
<td>28.4</td>
<td>beta</td>
</tr>
<tr>
<td>76</td>
<td>140</td>
<td>$-40.000 \pm 3.300$</td>
<td>THR</td>
<td>66.3</td>
<td>alpha</td>
</tr>
<tr>
<td>77</td>
<td>149</td>
<td>$-40.000 \pm 2.055$</td>
<td>THR</td>
<td>27.4</td>
<td>alpha</td>
</tr>
<tr>
<td>78</td>
<td>60</td>
<td>$-42.667 \pm 3.399$</td>
<td>PHE</td>
<td>21.9</td>
<td>beta</td>
</tr>
</tbody>
</table>

Continued on next page
Table 10.1 – *Continued from previous page*

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>ΔTm (°)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>99</td>
<td>-42.667 ± 3.266</td>
<td>GLN</td>
<td>83.4</td>
<td>Loop</td>
</tr>
<tr>
<td>80</td>
<td>114</td>
<td>-44.333 ± 5.437</td>
<td>THR</td>
<td>93</td>
<td>Loop</td>
</tr>
<tr>
<td>81</td>
<td>254</td>
<td>-45.000 ± 2.449</td>
<td>ASP</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>82</td>
<td>147</td>
<td>-45.000 ± 2.625</td>
<td>GLU</td>
<td>57.2</td>
<td>alpha</td>
</tr>
<tr>
<td>83</td>
<td>112</td>
<td>-46.333 ± 4.028</td>
<td>HIS</td>
<td>31.9</td>
<td>Loop</td>
</tr>
<tr>
<td>84</td>
<td>212</td>
<td>-53.667 ± 4.190</td>
<td>GLU</td>
<td>49.2</td>
<td>alpha</td>
</tr>
<tr>
<td>85</td>
<td>230</td>
<td>-54.667 ± 3.859</td>
<td>PHE</td>
<td>12.5</td>
<td>beta</td>
</tr>
<tr>
<td>86</td>
<td>178</td>
<td>-55.000 ± 1.247</td>
<td>ARG</td>
<td>38.9</td>
<td>Loop</td>
</tr>
<tr>
<td>87</td>
<td>255</td>
<td>-58.333 ± 1.700</td>
<td>GLY</td>
<td>57.2</td>
<td>Loop</td>
</tr>
<tr>
<td>88</td>
<td>275</td>
<td>-60.333 ± 1.886</td>
<td>ARG</td>
<td>8.3</td>
<td>alpha</td>
</tr>
<tr>
<td>89</td>
<td>256</td>
<td>-94.00 ± 0.89</td>
<td>LYS</td>
<td>73.9</td>
<td>Loop</td>
</tr>
</tbody>
</table>
Table 10.2: PEGylation quantitative analysis. Numeric ranking of all 89 surface accessible sites.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>$\Delta$Tm ($^\circ$)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>153</td>
<td>12.333 ± 1.700</td>
<td>HIS</td>
<td>54.4</td>
<td>alpha</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>12.333 ± 2.828</td>
<td>GLY</td>
<td>86.4</td>
<td>alpha</td>
</tr>
<tr>
<td>3</td>
<td>115</td>
<td>11.667 ± 1.247</td>
<td>ASP</td>
<td>76.9</td>
<td>Loop</td>
</tr>
<tr>
<td>4</td>
<td>154</td>
<td>11.333 ± 2.055</td>
<td>ASN</td>
<td>94.9</td>
<td>alpha</td>
</tr>
<tr>
<td>5</td>
<td>41</td>
<td>11.333 ± 1.633</td>
<td>GLY</td>
<td>78.8</td>
<td>alpha</td>
</tr>
<tr>
<td>6</td>
<td>177</td>
<td>11.000 ± 2.494</td>
<td>GLU</td>
<td>59</td>
<td>Loop</td>
</tr>
<tr>
<td>7</td>
<td>92</td>
<td>11.000 ± 2.625</td>
<td>GLY</td>
<td>66.1</td>
<td>Loop</td>
</tr>
<tr>
<td>8</td>
<td>158</td>
<td>11.000 ± 1.247</td>
<td>HIS</td>
<td>68.7</td>
<td>Loop</td>
</tr>
<tr>
<td>9</td>
<td>198</td>
<td>11.000 ± 3.091</td>
<td>LEU</td>
<td>37.7</td>
<td>Loop</td>
</tr>
<tr>
<td>10</td>
<td>209</td>
<td>11.000 ± 1.414</td>
<td>ASP</td>
<td>47.1</td>
<td>alpha</td>
</tr>
<tr>
<td>11</td>
<td>182</td>
<td>11.000 ± 1.700</td>
<td>MET</td>
<td>19.5</td>
<td>beta</td>
</tr>
<tr>
<td>12</td>
<td>96</td>
<td>11.000 ± 0.471</td>
<td>HIS</td>
<td>71.5</td>
<td>beta</td>
</tr>
<tr>
<td>13</td>
<td>86</td>
<td>10.667 ± 1.700</td>
<td>ALA</td>
<td>59</td>
<td>alpha</td>
</tr>
<tr>
<td>14</td>
<td>159</td>
<td>10.667 ± 1.886</td>
<td>VAL</td>
<td>24.7</td>
<td>Loop</td>
</tr>
<tr>
<td>15</td>
<td>288</td>
<td>10.333 ± 5.312</td>
<td>LYS</td>
<td>81.2</td>
<td>alpha</td>
</tr>
<tr>
<td>16</td>
<td>99</td>
<td>10.333 ± 0.816</td>
<td>GLN</td>
<td>83.4</td>
<td>Loop</td>
</tr>
<tr>
<td>17</td>
<td>52</td>
<td>10.333 ± 2.625</td>
<td>SER</td>
<td>53.6</td>
<td>Loop</td>
</tr>
<tr>
<td>18</td>
<td>197</td>
<td>10.333 ± 3.682</td>
<td>GLU</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>19</td>
<td>202</td>
<td>10.000 ± 4.110</td>
<td>ALA</td>
<td>69.9</td>
<td>alpha</td>
</tr>
<tr>
<td>20</td>
<td>112</td>
<td>10.000 ± 0.816</td>
<td>HIS</td>
<td>31.9</td>
<td>Loop</td>
</tr>
<tr>
<td>21</td>
<td>39</td>
<td>10.000 ± 3.399</td>
<td>GLN</td>
<td>61.6</td>
<td>alpha</td>
</tr>
<tr>
<td>22</td>
<td>254</td>
<td>10.000 ± 1.700</td>
<td>ASP</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>23</td>
<td>120</td>
<td>9.667 ± 2.160</td>
<td>ARG</td>
<td>46.7</td>
<td>alpha</td>
</tr>
<tr>
<td>24</td>
<td>94</td>
<td>9.667 ± 2.944</td>
<td>ARG</td>
<td>48.9</td>
<td>beta</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Δ Tm (°)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>114</td>
<td>9.667 ± 2.055</td>
<td>THR</td>
<td>93</td>
<td>Loop</td>
</tr>
<tr>
<td>26</td>
<td>28</td>
<td>9.667 ± 3.399</td>
<td>GLU</td>
<td>56.1</td>
<td>alpha</td>
</tr>
<tr>
<td>27</td>
<td>61</td>
<td>9.667 ± 2.494</td>
<td>ARG</td>
<td>32.7</td>
<td>beta</td>
</tr>
<tr>
<td>28</td>
<td>273</td>
<td>9.667 ± 2.494</td>
<td>ASP</td>
<td>73.4</td>
<td>alpha</td>
</tr>
<tr>
<td>29</td>
<td>161</td>
<td>9.667 ± 0.471</td>
<td>ARG</td>
<td>31.4</td>
<td>Loop</td>
</tr>
<tr>
<td>30</td>
<td>110</td>
<td>9.667 ± 2.867</td>
<td>GLU</td>
<td>62.4</td>
<td>alpha</td>
</tr>
<tr>
<td>31</td>
<td>258</td>
<td>9.333 ± 0.471</td>
<td>SER</td>
<td>35</td>
<td>Loop</td>
</tr>
<tr>
<td>32</td>
<td>146</td>
<td>9.333 ± 1.700</td>
<td>SER</td>
<td>52</td>
<td>alpha</td>
</tr>
<tr>
<td>33</td>
<td>195</td>
<td>9.000 ± 0.943</td>
<td>THR</td>
<td>46.1</td>
<td>alpha</td>
</tr>
<tr>
<td>34</td>
<td>267</td>
<td>9.000 ± 1.247</td>
<td>GLY</td>
<td>58.4</td>
<td>beta</td>
</tr>
<tr>
<td>35</td>
<td>200</td>
<td>9.000 ± 1.700</td>
<td>THR</td>
<td>49</td>
<td>Loop</td>
</tr>
<tr>
<td>36</td>
<td>84</td>
<td>9.000 ± 2.160</td>
<td>VAL</td>
<td>20.8</td>
<td>alpha</td>
</tr>
<tr>
<td>37</td>
<td>163</td>
<td>9.000 ± 3.682</td>
<td>ASP</td>
<td>38.8</td>
<td>Loop</td>
</tr>
<tr>
<td>38</td>
<td>34</td>
<td>8.667 ± 0.000</td>
<td>LYS</td>
<td>51.5</td>
<td>alpha</td>
</tr>
<tr>
<td>39</td>
<td>143</td>
<td>8.667 ± 2.357</td>
<td>GLY</td>
<td>64.8</td>
<td>alpha</td>
</tr>
<tr>
<td>40</td>
<td>255</td>
<td>8.667 ± 0.943</td>
<td>GLY</td>
<td>57.2</td>
<td>Loop</td>
</tr>
<tr>
<td>41</td>
<td>90</td>
<td>8.667 ± 1.886</td>
<td>GLN</td>
<td>81.2</td>
<td>Loop</td>
</tr>
<tr>
<td>42</td>
<td>174</td>
<td>8.667 ± 0.816</td>
<td>PRO</td>
<td>59.9</td>
<td>Loop</td>
</tr>
<tr>
<td>43</td>
<td>54</td>
<td>8.667 ± 2.357</td>
<td>LYS</td>
<td>54.9</td>
<td>Loop</td>
</tr>
<tr>
<td>44</td>
<td>213</td>
<td>8.667 ± 1.633</td>
<td>ALA</td>
<td>35.4</td>
<td>alpha</td>
</tr>
<tr>
<td>45</td>
<td>192</td>
<td>8.667 ± 2.828</td>
<td>LYS</td>
<td>35.5</td>
<td>alpha</td>
</tr>
<tr>
<td>46</td>
<td>290</td>
<td>8.333 ± 2.055</td>
<td>TRP</td>
<td>40.3</td>
<td>alpha</td>
</tr>
<tr>
<td>47</td>
<td>102</td>
<td>8.333 ± 2.494</td>
<td>LEU</td>
<td>26.2</td>
<td>Loop</td>
</tr>
<tr>
<td>48</td>
<td>93</td>
<td>8.333 ± 1.886</td>
<td>ARG</td>
<td>47.2</td>
<td>Loop</td>
</tr>
<tr>
<td>49</td>
<td>176</td>
<td>8.333 ± 3.559</td>
<td>ASP</td>
<td>27.7</td>
<td>Loop</td>
</tr>
<tr>
<td>50</td>
<td>113</td>
<td>8.333 ± 2.944</td>
<td>LEU</td>
<td>25.7</td>
<td>Loop</td>
</tr>
<tr>
<td>51</td>
<td>178</td>
<td>8.000 ± 1.414</td>
<td>ARG</td>
<td>38.9</td>
<td>Loop</td>
</tr>
</tbody>
</table>

*Continued on next page*
Table 10.2 – Continued from previous page

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>$\Delta$ Tm ($^\circ$)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>175</td>
<td>8.000 ± 2.625</td>
<td>ASN</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>53</td>
<td>165</td>
<td>8.000 ± 2.944</td>
<td>TRP</td>
<td>37</td>
<td>Loop</td>
</tr>
<tr>
<td>54</td>
<td>26</td>
<td>8.000 ± 1.633</td>
<td>HIS</td>
<td>51.1</td>
<td>alpha</td>
</tr>
<tr>
<td>55</td>
<td>56</td>
<td>8.000 ± 2.055</td>
<td>LEU</td>
<td>28.4</td>
<td>beta</td>
</tr>
<tr>
<td>56</td>
<td>147</td>
<td>8.000 ± 2.160</td>
<td>GLU</td>
<td>57.2</td>
<td>alpha</td>
</tr>
<tr>
<td>57</td>
<td>228</td>
<td>8.000 ± 0.471</td>
<td>GLY</td>
<td>76.2</td>
<td>Loop</td>
</tr>
<tr>
<td>58</td>
<td>201</td>
<td>7.667 ± 1.700</td>
<td>LEU</td>
<td>62.8</td>
<td>alpha</td>
</tr>
<tr>
<td>59</td>
<td>212</td>
<td>7.667 ± 2.828</td>
<td>GLU</td>
<td>49.2</td>
<td>alpha</td>
</tr>
<tr>
<td>60</td>
<td>88</td>
<td>7.667 ± 1.414</td>
<td>GLN</td>
<td>71.1</td>
<td>Loop</td>
</tr>
<tr>
<td>61</td>
<td>83</td>
<td>7.667 ± 2.449</td>
<td>ARG</td>
<td>23.1</td>
<td>alpha</td>
</tr>
<tr>
<td>62</td>
<td>275</td>
<td>7.667 ± 1.414</td>
<td>ARG</td>
<td>8.3</td>
<td>alpha</td>
</tr>
<tr>
<td>63</td>
<td>149</td>
<td>7.667 ± 0.471</td>
<td>THR</td>
<td>27.4</td>
<td>alpha</td>
</tr>
<tr>
<td>64</td>
<td>63</td>
<td>7.667 ± 2.160</td>
<td>GLU</td>
<td>83.6</td>
<td>Loop</td>
</tr>
<tr>
<td>65</td>
<td>43</td>
<td>7.667 ± 1.700</td>
<td>ARG</td>
<td>34.5</td>
<td>Loop</td>
</tr>
<tr>
<td>66</td>
<td>30</td>
<td>7.333 ± 1.633</td>
<td>LEU</td>
<td>20.2</td>
<td>alpha</td>
</tr>
<tr>
<td>67</td>
<td>269</td>
<td>7.333 ± 1.633</td>
<td>GLN</td>
<td>100</td>
<td>Loop</td>
</tr>
<tr>
<td>68</td>
<td>191</td>
<td>7.333 ± 1.247</td>
<td>ARG</td>
<td>29.3</td>
<td>alpha</td>
</tr>
<tr>
<td>69</td>
<td>156</td>
<td>7.333 ± 0.471</td>
<td>GLY</td>
<td>59.7</td>
<td>alpha</td>
</tr>
<tr>
<td>70</td>
<td>105</td>
<td>7.000 ± 0.816</td>
<td>TYR</td>
<td>52.4</td>
<td>Loop</td>
</tr>
<tr>
<td>71</td>
<td>38</td>
<td>7.000 ± 3.682</td>
<td>ASP</td>
<td>63</td>
<td>alpha</td>
</tr>
<tr>
<td>72</td>
<td>222</td>
<td>6.667 ± 1.700</td>
<td>ARG</td>
<td>28.4</td>
<td>Loop</td>
</tr>
<tr>
<td>73</td>
<td>256</td>
<td>6.667 ± 0.943</td>
<td>LYS</td>
<td>73.9</td>
<td>Loop</td>
</tr>
<tr>
<td>74</td>
<td>31</td>
<td>6.667 ± 1.247</td>
<td>VAL</td>
<td>56.6</td>
<td>alpha</td>
</tr>
<tr>
<td>75</td>
<td>277</td>
<td>6.667 ± 2.625</td>
<td>ARG</td>
<td>57.6</td>
<td>alpha</td>
</tr>
<tr>
<td>76</td>
<td>281</td>
<td>6.333 ± 2.944</td>
<td>GLU</td>
<td>58.4</td>
<td>alpha</td>
</tr>
<tr>
<td>77</td>
<td>188</td>
<td>6.000 ± 1.700</td>
<td>THR</td>
<td>38.9</td>
<td>alpha</td>
</tr>
<tr>
<td>78</td>
<td>205</td>
<td>6.000 ± 3.300</td>
<td>GLN</td>
<td>47</td>
<td>alpha</td>
</tr>
</tbody>
</table>

Continued on next page
Table 10.2 – Continued from previous page

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Δ Tm (°C)</th>
<th>AA</th>
<th>Accessibility (%)</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>227</td>
<td>6.000 ± 4.899</td>
<td>ALA</td>
<td>94.4</td>
<td>Loop</td>
</tr>
<tr>
<td>80</td>
<td>100</td>
<td>6.000 ± 2.494</td>
<td>ASN</td>
<td>100</td>
<td>alpha</td>
</tr>
<tr>
<td>81</td>
<td>141</td>
<td>6.000 ± 2.357</td>
<td>THR</td>
<td>20.9</td>
<td>alpha</td>
</tr>
<tr>
<td>82</td>
<td>27</td>
<td>5.667 ± 1.700</td>
<td>PRO</td>
<td>59.2</td>
<td>alpha</td>
</tr>
<tr>
<td>83</td>
<td>151</td>
<td>5.333 ± 1.700</td>
<td>PHE</td>
<td>17.9</td>
<td>alpha</td>
</tr>
<tr>
<td>84</td>
<td>60</td>
<td>5.333 ± 1.700</td>
<td>PHE</td>
<td>21.9</td>
<td>beta</td>
</tr>
<tr>
<td>85</td>
<td>51A</td>
<td>5.000 ± 0.816</td>
<td>ALA</td>
<td>14.5</td>
<td>Loop</td>
</tr>
<tr>
<td>86</td>
<td>140</td>
<td>4.667 ± 1.247</td>
<td>THR</td>
<td>66.3</td>
<td>alpha</td>
</tr>
<tr>
<td>87</td>
<td>230</td>
<td>4.667 ± 2.449</td>
<td>PHE</td>
<td>12.5</td>
<td>beta</td>
</tr>
<tr>
<td>88</td>
<td>51</td>
<td>3.333 ± 2.625</td>
<td>LEU</td>
<td>8.7</td>
<td>Loop</td>
</tr>
<tr>
<td>89</td>
<td>62</td>
<td>1.67 ± 0.82</td>
<td>LYS</td>
<td>25.4</td>
<td>Loop</td>
</tr>
</tbody>
</table>

10.2 Accessibility Scatter Plots

Data in Tables 10.1 and 10.2 indicate little correlation exists between accessibility and site performance. The following scatter plots below demonstrates this visually. Each point represents a replicate in the protein device screen (i.e. entries in Tables 10.1 and 10.2), and are color-coded to according to the domain each site is located. The abscissa is the solvent accessibility and the ordinate is the change in melting temperature for the treatment. Figure 10.1 is for surface tethering, and Figure 10.2 for PEGylation. Both plots in Figures 10.1 and 10.2 demonstrate random scatter indicating little to no correlation between accessibility and stabilization for either functionalization case.


10.3 Surface Tethering Orientations

The orientation of the protein on the surface is an important factor for tethered devices. The main body of the paper mentioned how there was little correlation between surface orientation and screen performance. The numerical data for this assertion is presented here.

For purposes of discussion, two types of orientations are discussed: “upright” and “flat.” For proteins with a high aspect ratio, as is the case with TEM-1, an upright configuration can be defined where the longest dimension is perpendicular to the surface. Conversely, a flat configuration will have the longer dimension more parallel with the surface plane. For this study, the longest
dimension is defined between residues 34 and 88, and represents the direction of the vertical height of TEM-1.

An angle is also defined to distinguish between upright and flat configurations in simulation. This angle is defined by the surface normal passing through Site 34 and the line between Sites 34 and 88. The top graphic in Figure 10.3 shows that as $\theta$ approaches 0 or 180° the protein is in an “upright” position, and as $\theta$ approaches 90° the protein is lying “flat” on the surface.

Figure 10.3: The graphic above the plots show the longest dimension in TEM-1 between Sites 34 and 88. Also shown is the line normal to the surface plane that goes through Site 34. “Upright” or “flat” configurations are identified by the angle between these two dimensions. The two figures below this graphic show the value of protein angle over the course of the NVT equilibration simulations.

As outlined in the main body of the paper, only sites in the Alpha I and Loop domains are strongly affected by the surface, so only orientations resulting from functionalizing these two
domains are analyzed here. The plots in Figure 10.3 show the protein-surface orientations as a function of time for representative simulations. The left plot of Figure 10.3 shows orientations resulting from functionalizing the Alpha I domain, and large variations between functionalization site can be seen. Tethering on Sites 34 and 227 result in flat configurations, and tethering on Sites 28 and 281 result in more upright configurations. Because all these sites perform well in the screen (Table ??), conformational orientation does not appear correlate with site performance.

Orientations resulting from functionalizing Loop domains (left plot of Figure 10.3) were flat on the surface with little variation between the sites examined. Notably, these flat configurations expose adjacent domains to the surface and likely contribute to adjacent domain denaturing and subsequent poor performance observed Table ??.

Results from Table III in the main body identify the Alpha I domain as performing the best in the tethering screen, and the Loop domains performing the worst. An upright protein configuration would reduce the destabilizing effect of a hydrophobically-interacting surface and result in better performing sites. However, the orientation variation in the strongly performing residues from the Alpha I domain is more nuanced, and good performing residues produced more flat configurations. Sites in the Loop domains all produce flat configurations (see the left plot of Fig. S3) and also performed poorly in the screen.

More work is needed that compares protein orientation on the surface with thermal stability. These findings corroborate the conclusions of this work that emphasizes the need to perform the computational screen to identify the best-performing sites because the simulations capture the effects of surfaces or PEGylation that an analysis based solely on the PDB structure ignores.
CHAPTER 11. APPENDIX B: SUPPLEMENTARY MATERIAL TO CHAPTER 8

11.1 Methods: Transformations for Short-Time Simulations

One important aspect for the short-time simulations in 8.3.2 is predicting movements in angle and dihedral space. This is done by transforming Cartesian velocity and force vectors into $\theta$ and $\phi$ space and performing a single Eulerian integration step to compute $\Delta \theta$ or $\Delta \phi$. The theory for angle and dihedral space transformations is now explained. All transformation calculations for these simulations assumes $\theta_0 = 180^\circ$ and the dihedral potential prefers cis structures over tans. These parameters assumptions are necessary for the transformations descriptions in Sections 11.4.1 and 11.4.2.

11.2 Rejection Vector Definition

To fully understand the process for transforming Cartesian vectors into angle and dihedral space, a few principals must be defined. Vector projection is the parallel projection of one vector onto the other. This is shown in Figure 11.1 where the projection vector of $\vec{A}$ onto $\vec{B}$ is labeled $\vec{A}_p$.

![Figure 11.1: Calculating the rejection vector from $\vec{A}$ to $\vec{B}$. Component vectors are black, the projection vector in blue and the rejection vector in green.](image)
In contrast, vector rejection is the perpendicular projection of one vector onto the other. Using the same vectors in Figure 11.1, the rejection vector is labeled $\vec{A}_R$ and is computed according to:

$$\vec{A}_R = \vec{A} - \vec{A}_P$$  \hspace{1cm} (11.1)

where $A_R$ is called the rejection vector from $A$ to $B$. If the projection vector is unknown, $\vec{A}_R$ can be determined relative to vectors $\vec{A}$ and $\vec{B}$ alone by substituting $\vec{A}_P$ for the general projection equation:

$$\vec{A}_R = \vec{A} - \frac{\vec{A} \cdot \vec{B}}{\vec{B} \cdot \vec{B}} \vec{B}$$  \hspace{1cm} (11.2)

where this form of the equation does not require any additional information about projection angle. The rejection vector is important for our model system because it defines a vector normal to $\vec{B}$ and in a direction desirable for transformations.

### 11.3 Direction of Angle Space

Angle space is defined as motion that results from angle potential. Figure 11.2A and 11.2B show all possible positions in angle space for some $\theta$ value. This is represented as an blue cone. The cone eventually converges onto a single line when $\theta = 180^\circ$ (Figure 11.2C). This convergence line is in-line with the $\vec{BC}$ bond which is, for our model, the $z$-axis. With the knowledge that angle space always includes this convergence line, the direction of angle space can be determined. Assume a vector on the convergence line called the Linear Convergence Vector (LCV), Figure 11.2D shows that the direction of angle space is determined by calculating the rejection vector from the LCV (orange) to the bond vector. Note that when doing the angle space transformations, angle space is defined to go toward minimized angle space ($\theta = \theta_0$). Thus the direction of this rejection vector is dependant on $\theta$’s value relative to $\theta_0$ (see Figure 11.3 Step I).
Figure 11.2: Panels A-C) Angle space for increasingly linear $\theta$ values. Panel D The convergence line that is used to determine the direction of angle space.

11.4 Transform Cartesian Movement to Parameter Space

The transformation of Cartesian force and velocity into angle/dihedral space movement requires five steps:

I Find the direction of angle/dihedral space (direction vector)

II Determine property vector (velocity and force vectors in Cartesian coordinates) contribution to angle/dihedral space

III Add velocity and force contributions to find the magnitude of movement

IV Determine magnitude of change in curved space

V Perform a single Eulerian iteration to find total change in angle/dihedral space
Figure 11.3: Steps for determining $\Delta \theta'$. For steps I and II we determine the direction of angle space and then project velocity and force vectors (orange) into this space (red). In step III we add these vectors together to determine the magnitude of movement in angle space and in step IV we use this magnitude to determine $\Delta \theta'$.

### 11.4.1 Transform to Angle Space

Step I is done by computing the vector normal to the $\overline{AB}$ bond and pointing toward minimized angle space (the z-axis for this model). Thus, the direction vector is computed by determining the z-axis’ rejection vector from Site A’s position vector, a.k.a. the green vector in Figure 11.3. Step II computes the projection of the Site A’s property vectors onto the direction vector. In Step III, the velocity and force property vectors are added together. The magnitude of this vector is then determined to find the magnitude of movement. For Step IV, because angle space is curved, the arc-length equation is used to determine the magnitude of change ($\Delta \theta'$):

$$\Delta \theta' = \frac{s \cdot 360}{r \cdot 2\pi} \quad \text{(11.3)}$$

where arc-length ($s$) is the magnitude of movement from step 3, and the current bond length is the arc radius ($r$). Finally in Step V, total change in angle space is determined by performing a single Eulerian integration:

$$\Delta \theta = \Delta \theta' \cdot \delta t \quad \text{(11.4)}$$
where simulation timestep is $\delta t$.

### 11.4.2 Transform to Dihedral Space

![Diagram illustrating steps for determining $\Delta \phi'$](image)

Figure 11.4: Steps for determining $\Delta \phi'$. For steps I and II we determine the direction of dihedral space and then project velocity and force vectors (orange) into this space (violet). In step III we add these vectors together to determine the magnitude of movement in angle space and in step IV we use this magnitude to determine $\Delta \phi'$.

For Step I, the direction vector is a vector normal to the $\overline{AB}$ bond and tangent to the circle of dihedral motion. Thus this vector only exists in an xy-plane for this model. The sign of this vector is dependant on the dihedral potential. For this model, when $x \geq 0$ positive dihedral motion is clockwise, otherwise positive dihedral motion is counter-clockwise, a.k.a the green vector in Figure 11.4. Steps II and III mirror the process for angle space. The projection of each property vector onto the dihedral direction vector is computed and add together. This vector’s magnitude is the magnitude of movement in dihedral space. To calculate the magnitude of change for Step IV, the arc-length equation (Eqn. 11.3) is used except the arc-radius ($r$) differs from angle space. Dihedral motion rotates around the $\overline{BC}$ bond, thus $r$ is the magnitude of the vector that results from
computing the rejection vector of position from the z-axis. Finally for Step V, Eqn. 11.4 is again used to determine total change in dihedral space where simulation timestep is $\delta t$. 