JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

The Effects of *p*-Azidophenylalanine Incorporation on Protein Structure and Stability

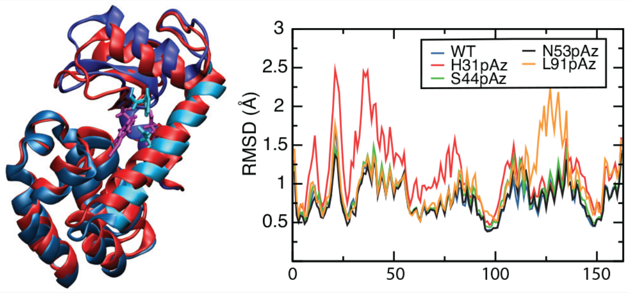
Joshua W. Wilkerson, Addison K. Smith, Kristen M. Wilding, Bradley C. Bundy, and Thomas A. Knotts, IV*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 5117−5125

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Functionalization is often needed to harness the power of proteins for beneficial use but can cause losses to stability and/or activity. State of the art methods to limit these deleterious effects accomplish this by substituting an amino acid in the wild-type molecule into an unnatural amino acid, such as *p*-azidophenylalanine (pAz), but selecting the residue for substitution a priori remains an elusive goal of protein engineering. The results of this work indicate that all-atom molecular dynamics simulation can be used to determine whether substituting pAz for a natural amino acid will be detrimental to experimentally determined protein stability. These results offer significant hope that local deviations from wild-type structure caused by pAz incorporation observed in simulations can be a predictive metric used to reduce the number of costly experiments that must be done to find active proteins upon substitution with pAz and subsequent functionalization.



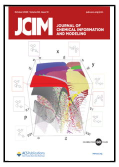
■ INTRODUCTION

Harnessing the functional capabilities of proteins is prevalent in industries ranging from biofuels[1] to pharmaceuticals.[2] Using proteins in new applications is a source of continual research because proteins can have many practical purposes. However, challenges with stability, activity, and bioavailability must be overcome in each case. Many functionalization strategies have been designed to this end. Examples include the following: attachment of glycan, lipid, and other molecules used as chemical reporters for protein research and cancer treatment,[3−5] conjugation of dyes for medical imaging,[6−8] tethering to surfaces or beads to increase stability and for use in protein microarrays,[9−12] and conjugation to the polymer poly(ethylene glycol) (PEG) to decrease immunogenic response and reduce degradation and aggregation.[13−15]

While these strategies make it possible to use the protein outside of its natural environment, each also involves modifying the native structure which can detrimentally affect protein activity and stability.[13,16] For example, peginterferon alpha-2a, a PEGylated cancer and hepatitis drug, shows a 30-fold reduction in activity as compared to its non-PEGylated counterpart.[17] Despite the activity loss, the PEGylated variant is used because it is retained longer in the body, leading to a significant overall increase in disease protection.[18] However, the optimal strategy would both maintain activity and increase retention time.

Very often, a major factor leading to decreases in the activity and stability of functionalized proteins is that the modifications happen nonspecifically.[13,18] For example, it is desirable to attach PEG to the protein at a residue far from the active site so as to reduce interference with function. However, no therapeutic PEGylated protein that is commercially available has been functionalized in a fully site-specific (meaning that a single, unique site is functionalized) and site-selective (meaning that any residue of the protein can be chosen as a site for functionalization) manner.[19] The usual practice is nonspecific PEGylation, such as by targeting amine groups.[18,20] These methods thus produce many different molecules—some with PEG at one location, some with PEG at another location, and others with multiple PEG molecules—resulting in a large population of suboptimal protein−PEG conjugates. Other proteins have been PEGylated by site-specific methods that are not site-selective, including targeting naturally occurring protein moieties, such as the N- or C-terminus.[19] Substitution and targeting of a less-prevalent natural amino acid, such as cysteine, can produce site-specific functionalization, but this method would not be site-specific in a protein with free cysteines.[18,21] What is needed is the ability to optimize activity and stability by targeting a specific position in the amino-acid sequence of any protein so that only one protein−PEG molecule is produced.

Received: June 25, 2020
Published: September 23, 2020



https://dx.doi.org/10.1021/acs.jcim.0c00725

Site-specific and site-selective protein functionalization is possible through unnatural amino acid (uAA) incorporation. With this method, a single amino acid can be replaced by an uAA not found in the standard genetic code, creating a unique site to be targeted for functionalization.[10,13,22−28] For example, the uAA *p*-azidophenylalanine (pAz) can be substituted in place of an amino acid in a protein to create a location for a "click chemistry" reaction to occur at the surface of a protein.[13,25,29] This chemistry is inert to most biological reactions and thus creates a unique site that can be targeted for functionalization at biological conditions.

Despite the existence of a method for functionalization that is both site-specific and site-selective, relatively few sites produce optimal conjugates. At present, this means generating many uAA-substitued proteins/functionalizations and assaying the results to find the best location. This is expensive, in both time and money, so it is highly desirable to shrink the search space. Two questions must be answered in this regard. The first is to determine whether a site can withstand a substitution of an unnatural amino acid and still maintain function. The second is to determine if the subsequent functionalization deleteriously affects the function of the protein. Concerning the latter, it has been suggested that secondary structure and solvent accessibility can be used to guess the efficacy of a functionalization site;[13,25,30] however, studies have shown that no current heuristic is adequate for selecting optimal functionalization sites a priori.[13,30]

Molecular simulation with coarse-grain models has been a popular tool in recent years when examining protein structure, stability, and folding mechanisms.[31−38] Regarding the purposes of this work, they have been able to predict the effects of site-specific functionalization on protein stability and activity, thus providing at least a partial answer to the second question;[10,13,39] however, these simulations were based on the assumption that the effects of the uAA substitution itself on the protein's activity and stability were negligible compared to the effects of the functionalization group. Specifically, the uAA substitution is not parametrized into the coarse-grain model—thus Question 1 is totally ignored. This is problematic because experimental data have suggested that the effects of pAz substitution on activity and stability are often greater than the effects of tether or polymer conjugation.[13]

Molecular simulation using classical potential models cannot directly probe *activity*, so this work uses *stability* as a surrogate metric. Although activity and stability are not perfectly correlated, changes to protein stability or structure have been found to be the primary causes of decreased activity in most protein mutants.[40] With specific regard to the protein used in the present study, the stability and activity of mutants of T4 lysozyme have been shown to move in the same direction compared to the wild-type protein.[13] Moreover, other work has shown that the optimal pH values for protein stability and activity are correlated.[41] These examples indicate that stability may be used as a surrogate for activity in molecular simulation with high fidelity—a practice that has been done previously.[10,13] Because of the importance of protein thermal stability prediction, various methods involving simulation and machine learning have been proposed to predict the effects of point mutations on thermal stability.[42−44]

The purpose of this work is to show how all-atom molecular simulation can be used to answer the question of whether a pAz substitution will deleteriously affect protein function and stability. The aim is to create an in silico screen to select the best sites for pAz incorporation to reduce the number of expensive experiments that must be performed to identify such a site. The simulation results, which include a detailed analysis of the atomic-level structural changes induced by the substitutions, are validated against experimental thermal-shift stability experiments. The results show, for the first time, that all-atom simulations can correctly predict how protein stability is affected by pAz substitutions—an outcome that offers significant hope that effective protein functionalization can be done in the future with reduced cost.

■ METHODS

Proteins. Five different proteins were investigated in this work. The control was cysteine-free T4 lysozyme (Protein Data Bank ID 1L63), shown in Figure 1. It is a globular monomer 164 residues long[45] and was selected for this work because it has been studied extensively both experimentally and with simulation. Its crystal structure,[46−48] protein folding mechanisms,[46,49−53] thermal stability,[46,50,54] and activity[48,49,55] are well-known. The other four proteins, the treatments, are the following variants of T4 lysozyme: S44pAz, N53pAz, L91pAz, and H31pAz. These sites were chosen to study the effects of substitutions in various types of secondary structure (α helix, β sheet, and unstructured regions).

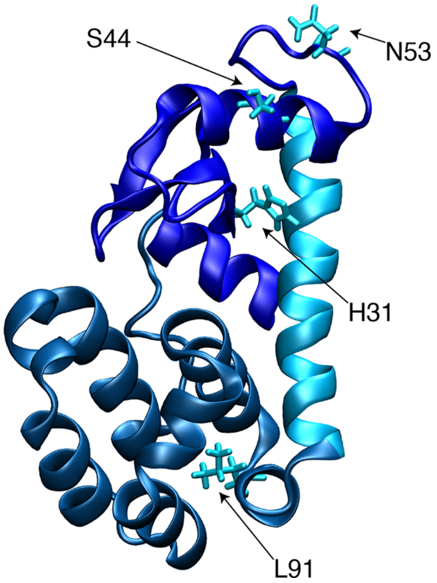


Figure 1. Cysteine-free T4 lysozyme (PDB ID 1L63). The N-terminal domain (top), the α helix connecting the two domains of the protein (middle), and the C-terminal domain (bottom) are colored differently. Residues substituted with pAz in this work are depicted explicitly and labeled.

Experimental Techniques. Proteins for testing thermal stability were produced using *Escherichia coli*-based cell-free protein synthesis.[22,56] Cysteine-free T4 lysozyme from Addgene (Cambridge, MA) was cloned into the pY71 plasmid with a C-terminal strep-tag for purification. Variant proteins were produced using the same plasmid but following the Quikchange II mutagenesis protocol (Agilent Technologies, Santa Clara, CA) to incorporate an amber stop codon. Plasmids were purified using a Qiagen Plasmid Maxi Kit (Valencia, CA) and then used in cell-free protein synthesis. Enzyme purification was performed using Strep-Tactin spin columns (IBA Life Sciences, Gottingen, Germany). Refer to

Wilding et al. for a more detailed description of protein production and purification procedures.[13]

The stability of T4 lysozyme was measured experimentally using the Protein Thermal Shift assay from Thermo Fisher Scientific (Carlsbad, CA). This assay determines the melting temperature of a protein by utilizing a hydrophobic dye that fluoresces when it binds to the internal hydrophobic residues of a protein as it denatures. The protein is exposed to increasing temperatures in a real-time PCR machine, and the amount of fluorescence at each temperature is recorded and analyzed to obtain a fluorescence vs temperature curve. The melting temperature is identified as the temperature at which half of the proteins in solution are found in their native conformation. It is determined by finding the inflection point of the fluorescence vs temperature curve. This melting temperature is thus a quantitative measure that is used to determine how substitutions affect the thermal stability of variants compared to the wild-type protein. If a variant has a higher melting temperature than the wild type, then the substitution stabilized the protein. Conversely, if the melting temperature of the variant is lower than that of the wild type, then the substitution destabilized the protein. A more detailed explanation of the method has been provided previously.[13] In this work, values reported by Wilding et al. were used for S44pAz, N53pAz, and L91pAz.[13] The value for H31pAz is novel to this work and was obtained following a procedure identical to that previously used.[13]

Model. The proteins were modeled using the CHARMM36 force field with CMAP correction.[57,58] pAz was simulated using recently developed parameters by Smith et al.[59] SHAKE[60] was used for bonds to hydrogen atoms. The NTER and CTER patches were used, and histidine was modeled as charge-neutral HSD. A force-switching algorithm[61] with an inner cutoff of 10 Å and an outer cutoff of 12 Å was used for Lennard-Jones interactions. Coulombic interactions were modeled using the particle−particle, particle−mesh (PPPM) Ewald summation technique with a real-space cutoff of 14 Å and the alpha and grid spacing set so that the error is less than 10^{-4}.[62]

CHARMM-compatible input files were obtained using CHARMM-GUI.[63,64] The last two residues were not sufficiently refined in the crystallographic data, so CHARMM-GUI was used to add these to the protein model. In-house software was used to create PDB files for the four pAz-substituted variants of T4 lysozyme by replacing individual amino acids of the wild-type PDB created by CHARMM-GUI with pAz.

The wild-type and pAz-substituted proteins obtained from CHARMM-GUI were then minimized in vacuum using the CHARMM simulation package[57] for 100 steepest descent steps, followed by 100 steps of the Adopted Basis Newton−Raphson method. The charmm2lammps script included with LAMMPS[65] was then used to create LAMMPS input files. The script also created a water box with a 12 Å border around the protein and added Na^+ and Cl^- ions to create a system at typical biological conditions (neutral pH with a concentration of approximately 0.15 M of both ions). The entire hydrated system was minimized in LAMMPS using 5000 steps of the steepest descent method followed by 5000 steps of the conjugate gradient method. Details of the molecular dynamics (MD) simulations are found in the next section.

MD Simulations. All MD simulations were done using LAMMPS. Simulations consisted of equilibration and production steps. Equilibration involved multiples stages in three different ensembles—NVE, NPT, and NVT. The production phase was done in the NVT ensemble. The temperature was maintained using the Nosé−Hoover method implemented in LAMMPS with 10 thermostats with a damping time of 100 ps for simulations done in the NVT and NPT ensembles.[66−69] Ten barostats with a damping time of 1000 ps were used to control pressure for the NPT ensemble. All equilibration steps used a conservative time step of 0.7 fs to allow the system to equilibrate in the event that the substitution of pAz for the natural amino acid caused atoms to overlap. As explained above, minimization was done prior to equilibration, but this additional precaution ensured that any high-energy states could relax. Each step is now described in more detail.

In the first step of equilibration, the protein was frozen, and the solvent was allowed to equilibrate in a series of NVE stages at different temperatures. The initial stage, done to remove voids produced during the solvation of the protein, consisted of 7 ps of NVE simulation with velocities initially set to produce a temperature of 300 K. Then, the solvent was heated to 350 K over 4.9 ps, held at 350 K for 3.5 ps, cooled to 300 K over 4.9 ps, and held at 300 K for 7 ps. The total NVE simulation time, during which the solvent was free to move around the fixed protein molecule, was thus 27.3 ps.

The next step of equilibration was done in the NPT ensemble. The velocities were initialized with a random uniform velocity distribution to produce a temperature of 275 K. The system was allowed to equilibrate to 300 K and 1 atm for 1.20 ns using NPT simulation. After this, to ensure the correct box size was obtained, multiple 70 ps, NVT simulations were performed with box sizes at or near the average box size produced from the NPT simulation, to obtain a box size that would produce a system pressure of ∼1 atm in the NVT ensemble. Once the proper box size was determined, 420 ps of additional equilibration in the NVT ensemble was used, after which the potential energy of the system leveled off.

As mentioned previously, all production steps were done in the NVT ensemble. The first 1.26 ns of production used the time step of 0.7 fs. This was increased to 2 fs after 1.26 ns for a final time of 10 ns. The combined equilibration and production simulation time of each protein (one wild type and four variants) was thus 11.3 ns per simulation. At least six independent simulations, using the entire equilibration and production scheme just described, were performed for each protein for statstical significance. See the Supporting Information for a discussion of the validity and reasoning of using a relatively short simulation time.

To facilitate visualization and root-mean-square fluctuation (RMSF) calculations, a light spring force ($k = 0.5$ kcal·mol^{-1}·Å^{-1}) was used to keep the center of mass of the protein near the center of the simulation box. Extra care was taken to ensure this weak tether did not affect the simulation results. (See the Supporting Information for the relevant discussion.) Linear momentum was zeroed every 1000 steps for the first 1.26 ns and every 10 000 steps for the remainder of the production.

Trajectory Analysis. Analysis of the simulation results was done by RMSF calculation and visual inspection of the simulation trajectories. The RMSD visualizer tool in VMD[70] was used to align proteins for visual inspection.[71] The MDAnalysis[72−75] Python package was used for RMSF calculations. To analyze the effects of pAz incorporation on a per-residue basis, an RMSF calculation was made for each α

carbon of the protein. Instead of using the time-averaged atom positions of the trajectory as the reference atom positions, the atom positions of the wild-type T4 lysozyme PDB file were used. RMSF calculations typically use the time-averaged positions as the reference position to measure the variation in atom positions over a trajectory. However, using the PDB position values as a reference allowed for comparison between the structure of the protein variants and the wild-type protein structure.

This approach to analysis provided a visual *and* quantitative method of illustrating which regions of the protein were most affected by uAA incorporation. As proteins have natural variations in structure compared to the PDB version due to thermal fluctuations, the RMSF of the wild-type structure was used to normalize the RMSF plots of the variants to highlight differences induced solely by the substitution. For these plots, the RMSF of the control (wild type) is subtracted from the RMSF of the treatment (variant). The term "normalized RMSF" is used for such plots herein.

■ RESULTS

Experimental Data. Figure 2 shows the experimental difference in melting point between each variant and the wild-type protein ($T_{m,treatment} - T_{m,control}$). Plotted in this manner, negative values mean the treatment is less stable than the control. The error bars are the standard error of at least three independent replicates ($N \geq 3$). Notice that all four substitutions destabilize the protein relative to the wild type; however, H31pAz and L91pAz are affected much more than S44pAz or N53pAz.

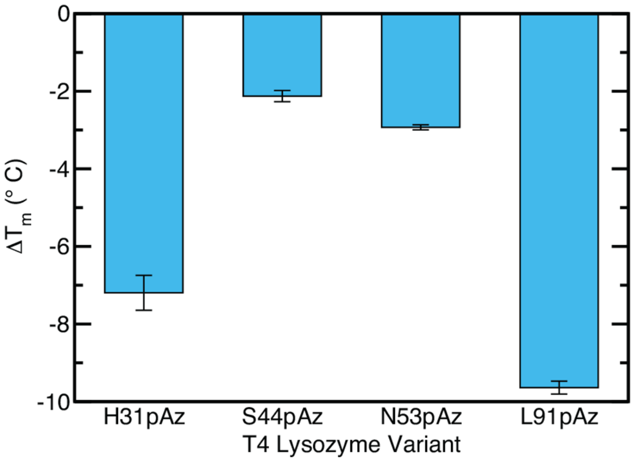


Figure 2. Difference in melting temperature between each treatment (variant) and the control (wild type). Error bars are the standard error of at least three replicates ($N \geq 3$). The data from L91pAz, S44pAz, and N53pAz were previously reported.[13] The data for H31pAz were generated for this work. Adapted in part from from ref 13. Copyright 2018 American Chemical Society.

Simulation Results. Figure 3 shows the RMSF results from the simulations. The RMSF is plotted for each residue in the protein for all five systems (one control, four treatments). The top graph is the unnormalized data. The bottom graph is the normalized values where the RMSF of each residue in the control is subtracted from that of the respective residue in the treatment. Plotted in this manner, the wild-type results are the horizontal line at RMSF = 0, positive RMSF values indicate regions of a treatment that vary more from the PDB structure compared to simulations of the wild-type protein, and negative values indicate regions of a treatment that vary less. Each treatment is considered in more detail below.

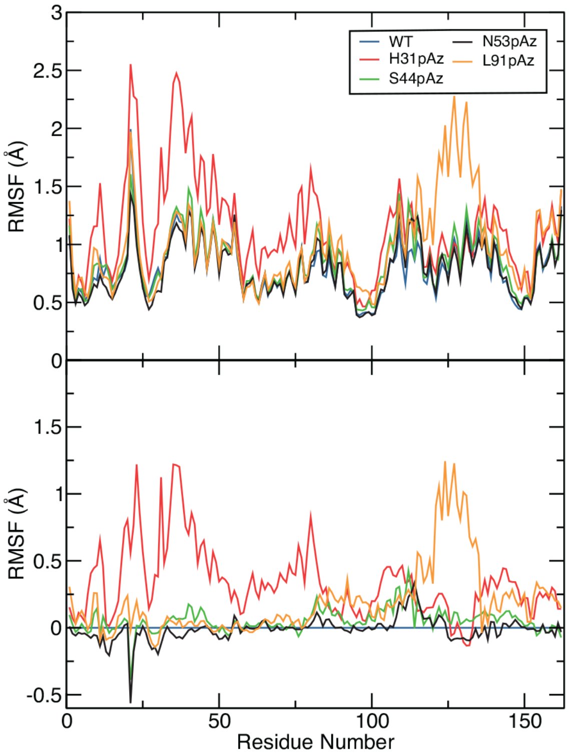


Figure 3. α Carbon RMSF of all variants of T4 lysozyme simulated. The top is the standard, unnormalized RMSF, and the bottom is a normalized RMSF where the value of the wild-type protein RMSF is subtracted from each variant's RMSF. Note the difference in the y-axis scale between the two panels. S44pAz and N53pAz are very similar to the wild type. Certain domains of H31pAz and L91pAz display observable changes in structure from the wild type. See the RMSF Error section and Figure S1 in the Supporting Information for plots that include the standard error. The gaps seen between the wild type and H31pAz and L91pAz are statistically significant.

■ DISCUSSION

S44pAz. Previous experimental research has suggested that every natural amino acid except for proline can be substituted for S44 without a significant change in the stability or structure of the protein despite S44 being located in the middle of an α helix.[76] Testing S44pAz with the thermal shift assay showed that this substitution led to only a minimal decrease in the stability of lysozyme (see Figure 2). The simulation results provide a potential explanation for this observation. Notice in Figure 3 and Figure 4 that although S44 is located in a highly structured region, substituting pAz at this site leads to little change in the backbone structure of the protein. Specifically, RMSF values show very little difference between the structure of the protein in the vicinity of the substitution compared to the natural variation observed in the wild-type protein. Notice that in Figure 3 the RMSF plots of the wild-type protein and S44pAz are nearly indistinguishable. No residue of S44pAz has a normalized RMSF value greater than 0.5 Å.

N53pAz. As with S44pAz, thermal-shift data for N53pAz suggest that the protein is only marginally less stable than the wild type (Figure 2). RMSF analysis of the simulation

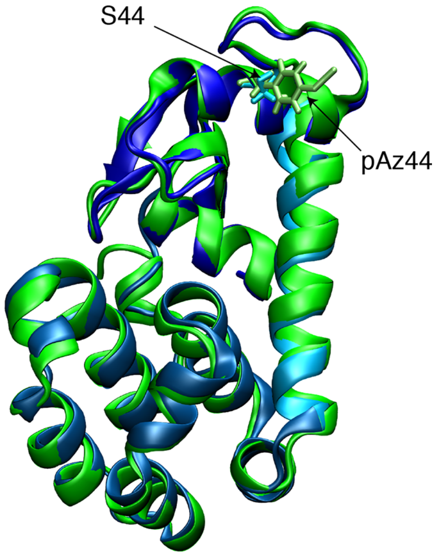


Figure 4. S44pAz (green) overlaid with wild type (blue). The structures vary from each other very little.

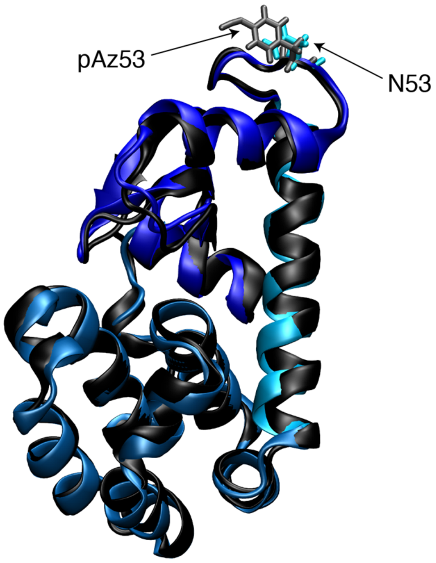


Figure 5. N53pAz (black) overlaid with wild type (blue). The structures vary from each other very little.

trajectory corroborated the experimental data. Notice in Figures 3 and 5 that, similar to S44pAz, no residue is more than 0.5 Å further from its PDB position than the corresponding residue of the wild-type protein. N53 is located in a flexible region linking two α helices. This location specifically showed an imperceptible difference in its RMSF from that of the wild-type protein.

H31pAz. H31 is part of the last strand of a β sheet located near the active site and at the boundary between the two domains of the protein.[77] H31 is the only histidine in T4 lysozyme and has been found to have a highly stabilizing effect on the protein because of a strong salt bridge it forms with D70.[78] The experiments in this work showed that H31pAz was very destabilized as compared to the wild-type protein (Figure 2). The data in Figure 3 reveal that this substitution causes significant deviations from the wild-type structure throughout the protein. The largest deviations primarily occur in the N-terminal domain (residues 1−59).

To better understand how this substitution affects the structure of the protein, three additional RMSF calculations were done for H31pAz. For these calculations, the treatment was aligned to three different parts of the control to ascertain which domains were most affected. (Recall that the data in Figure 3 were generated by aligning the entire backbone of the molecule.)

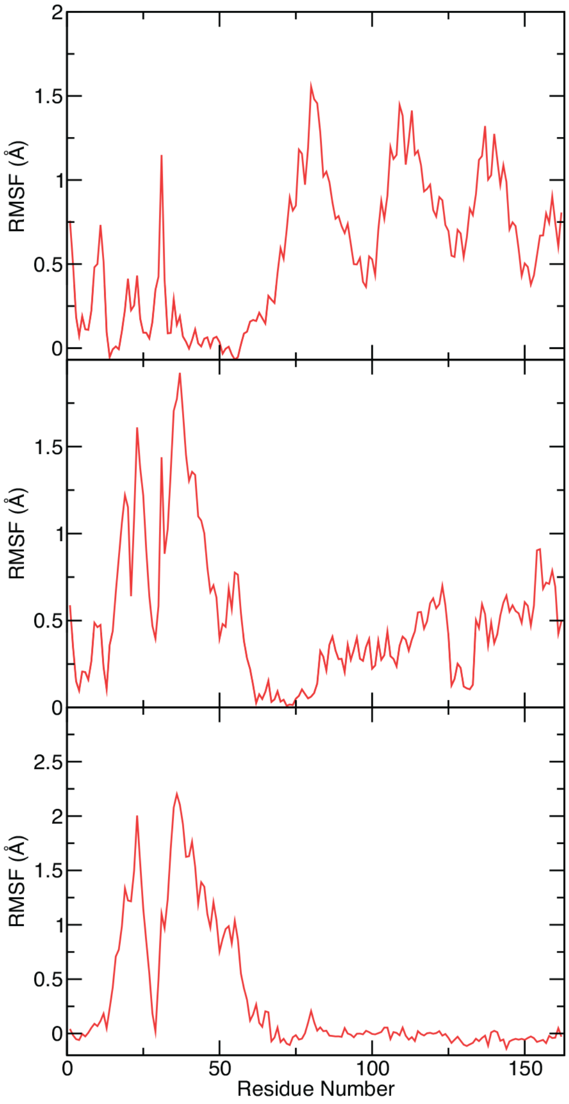


Figure 6. H31pAz hinge-bending, showing the normalized RMSF of H31pAz when aligned along residues 1−59 (top), 60−80 (middle), and 81−162 (bottom).

Figure 6 shows the results of this detailed RMSF analysis. The top plot is for alignment of just the N-terminal domain (residues 1−59). The middle plot is for alignment of residues 60−80 which form the α helix connecting the two domains of the protein. The bottom plot is for alignment done in the C-terminal domain (residues 81−162). Notice that alignment of the N-terminal domain results in significant deviations from the wild-type structure in both the N-terminal and C-terminal domains. When the α helix connecting the two domains (residues 60−80) is aligned, the N-terminal domain exhibits large deviations from the wild-type protein, but the deviations of the C-terminal domain are much smaller. Finally, when alignment is only done on the C-terminal domain, its deviations from the control are small while those of the N-terminal domain are very large. These results indicate that the deviations induced by the H31pAz substitution are largely concentrated in the N-terminal domain.

Previous works have shown that hinge-bending, where the angle between the N- and C-terminal domains of the protein can change by over 50°, occurs in both wild-type T4 lysozyme and variants.[46,79,80] When hinge-bending occurs, the N- and C-terminal domains typically shift as nearly rigid bodies with only minor disturbances to the structure of the domains.[46,79] However, in this work, the deviations in the N-terminal domain of H31pAz even when the N-terminal domain is aligned for RMSF calculations cannot be explained by hinge-bending because the normalized deviations of some residues are near 1 Å (see Figure 6, top).

One potential explanation of this deviation is the disruption of the salt bridge between residues H31 and D70 that has been shown to significantly stabilize the protein.[78] Figure 7 shows that the relative positions and orientations of residues 31−70 in the wild type (blue-highlighted residues) and H31pAz (pink-highlighted residues) are very different. The substitution of a much bulkier amino acid (pAz) for H31 appears to eliminate this salt bridge as pAz31 moves away from D70 in H31pAz. The result is significant structural deviations throughout the protein, which lead to protein destabilization and a decreased melting temperature.

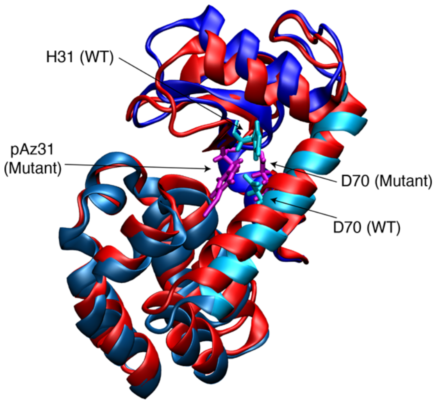


Figure 7. H31pAz (red) overlaid with wild type (blue). The salt bridge between residues 31 and 70 appears to be broken by substitution of pAz.

L91pAz. L91 is located far from the active site of the protein in a short linker region between two alpha helices. Previous simulation and experimental research has determined that L91 is an optimal site for protein immobilization through tethering but is a poor site for PEGylation.[13,25,81] Using Replica Exchange MD simulation, tethering has been shown to eliminate a stable intermediate in the protein-folding pathway. This makes unfolding the protein less kinetically favorable and leads to the protein remaining in a folded configuration a greater fraction of the time.[81] Conversely, PEGylating the protein does not eliminate this stable intermediate; rather, the PEG usually interferes with correct folding instead of preventing unfolding.[13]

The experimental results shown in Figure 2 indicate that without any functionalization L91pAz is very thermally unstable compared to the wild type. Interestingly, simulation results shown in Figure 3 show very little difference in the RMSF of the protein in the vicinity of site 91. However, large structural changes are observed in the region comprising residues 118−138. Although this region is not sequentially close to L91, it is spatially close to this site in the folded state. Figure 8 suggests that steric effects, due to the large size of pAz, push this region away from what would be its equilibrium position in the wild-type protein. Replacing a hydrophobic residue with a charged residue may also contribute to the displacement seen. In particular, residue L121 in the wild-type protein is located in a space that is filled by the bulky, charged pAz side chain in L91pAz. The previously observed low stability of PEGylated L91pAz could be a result of steric effects as well: attaching a large polymer to an amino acid that already produces steric effects in the protein is likely to exacerbate these effects. When L91pAz is tethered, the effects of the elimination of the stable intermediate on the folding pathway appear to be stronger than the steric effects observed in this work.

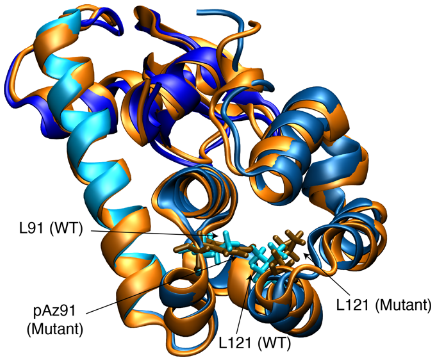


Figure 8. L91pAz (orange) overlaid with wild type (blue). Substitution of pAz causes steric hindrance with L121, leading to deviations from the wild-type structure in the vicinity of L121.

Using Simulation as a Screen. As mentioned in the Introduction, one of the goals of this work is to show that simulation can be used to reduce the number of pAz-substituted proteins that need to be experimentally generated before an active and stable variant is found. Exact prediction of the stability of pAz-substituted proteins compared to wild-type proteins is desirable but not feasible with current methods. However, this work shows promise that simulation can be used to give reasonable candidates for substitution to reduce the search space needed. Even an accuracy rate of 20% would greatly improve the current situation which involves a costly guess and check approach.

The data presented in Figure 3, particularly the bottom panel, illustrate how such an in silico screen could be done. Specifically, simulations, like those presented here, could be done on several candidate substitution sites. The normalized RMSF for each variant could then be generated from which the best candidates for experimental testing could be specified. If Figure 3 were a result of such an approach, S44 and N53, which have only small RMSF deviations from the wild type, would be selected for experiment while L91 and H31 would be avoided because the simulations indicate that these substitutions exhibit a large deviation in some regions compared to the wild-type protein.

Although the time required to perform such a screen would vary depending on the size of protein used, this approach saves time compared to experimentally testing all sites. Running the proteins for this study took approximately 304 h per protein from minimization through the end of production using 16 cores per simulation with Intel Haswell and Intel Sandy Bridge processors. Thus, with typical high-performance computing resources, running an entire simulation screen to test a suite of reasonable sites, similar to the one presented here, could feasibly be done in about 2 weeks. Analysis would vary depending on the protein and number of modifications studied, but a streamlined process focused on eliminating sites unlikely to be optimal for optimization could be done in 1−2 days, leading to a total time of approximately 2.5−3.0 weeks. Of course, the exact number of modifications that could be tested depends greatly on the computer resources available. With sufficient resources, all desired protein simulations could be run simultaneously, adding very little total time to simulation and analysis.

■ CONCLUSION

This work demonstrates that all-atom molecular simulation of T4 lysozyme can be used to predict the locational impact of unnatural amino acid incorporation on thermal stability compared to wild-type protein. Specifically, the experimental results on the thermal stability of pAz-substituted T4 lysozyme variants correlated very well with the deviation in RMSF from the MD simulations. This work complements other studies that have examined the effects of protein functionalization on protein stability and activity. Though they are preliminary in nature, and more work is needed to fully validate and understand the predictive capability of the simulations, these results offer a clear path to utilizing a computational screen to assist in determining optimal sites for site-specific and site-selective functionalization through pAz incorporation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00725.

Further discussion of simulation data including simulation length, the effects of a light spring force, and RMSF error (PDF)

Accession Codes

T4 Lysozyme PDB code: 1L63

■ AUTHOR INFORMATION

Corresponding Author

Thomas A. Knotts, IV − *Department of Chemical Engineering, Brigham Young University, Provo, Utah 84602, United States;* orcid.org/0000-0001-6248-4459; Email: thomas.knotts@byu.edu

Authors

Joshua W. Wilkerson − *Department of Chemical Engineering, Brigham Young University, Provo, Utah 84602, United States;* orcid.org/0000-0002-8247-3826

Addison K. Smith − *Department of Chemical Engineering, Brigham Young University, Provo, Utah 84602, United States;* orcid.org/0000-0003-3153-418X

Kristen M. Wilding − *Department of Chemical Engineering, Brigham Young University, Provo, Utah 84602, United States;* orcid.org/0000-0001-8476-8946

Bradley C. Bundy − *Department of Chemical Engineering, Brigham Young University, Provo, Utah 84602, United States;* orcid.org/0000-0003-4438-183X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00725

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

(1) Ebaid, R.; Wang, H.; Sha, C.; Abomohra, A. E.-F.; Shao, W. Recent trends in hyperthermophilic enzymes production and future perspectives for biofuel industry: A critical review. *J. Cleaner Prod.* **2019**, *238*, 117925.

(2) Dimitrov, D. S. *Therapeutic Proteins*; Springer: New York, 2012; pp 1−26.

(3) Grammel, M.; Hang, H. C. Chemical reporters for biological discovery. *Nat. Chem. Biol.* **2013**, *9*, 475.

(4) Wu, A. M.; Senter, P. D. Arming antibodies: prospects and challenges for immunoconjugates. *Nat. Biotechnol.* **2005**, *23*, 1137.

(5) Patterson, D. M.; Nazarova, L. A.; Prescher, J. A. Finding the right (bioorthogonal) chemistry. *ACS Chem. Biol.* **2014**, *9*, 592−605.

(6) Zhu, S.; et al. Molecular imaging of biological systems with a clickable dye in the broad 800-to 1,700-nm near-infrared window. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 962−967.

(7) Heyer, E.; Lory, P.; Leprince, J.; Moreau, M.; Romieu, A.; Guardigli, M.; Roda, A.; Ziessel, R. Highly Fluorescent and Water-Soluble Diketopyrrolopyrrole Dyes for Bioconjugation. *Angew. Chem., Int. Ed.* **2015**, *54*, 2995−2999.

(8) Krall, N.; Da Cruz, F. P.; Boutureira, O.; Bernardes, G. J. Site-selective protein-modification chemistry for basic biology and drug development. *Nat. Chem.* **2016**, *8*, 103.

(9) Gupta, S.; Manubhai, K.; Kulkarni, V.; Srivastava, S. An overview of innovations and industrial solutions in Protein Microarray Technology. *Proteomics* **2016**, *16*, 1297−1308.

(10) Grawe, R. W.; Knotts, T. A., IV The effects of tether placement on antibody stability on surfaces. *J. Chem. Phys.* **2017**, *146*, 215102.

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

(12) Bush, D. B.; Knotts, T. A., IV Communication: Antibody stability and behavior on surfaces. *J. Chem. Phys.* **2015**, *143*, No. 061101.

(13) Wilding, K. M.; Smith, A. K.; Wilkerson, J. W.; Bush, D. B.; Knotts, T. A., IV; Bundy, B. C. The locational impact of site-specific PEGylation: streamlined screening with cell-free protein expression and coarse-grain simulation. *ACS Synth. Biol.* **2018**, *7*, 510−521.

(14) Webber, M. J.; Appel, E. A.; Vinciguerra, B.; Cortinas, A. B.; Thapa, L. S.; Jhunjhunwala, S.; Isaacs, L.; Langer, R.; Anderson, D. G. Supramolecular PEGylation of biopharmaceuticals. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 14189−14194.

(15) Mishra, P.; Nayak, B.; Dey, R. PEGylation in anti-cancer therapy: An overview. *Asian J. Pharm. Sci.* **2016**, *11*, 337−348.

(16) Zaghmi, A.; Mendez-Villuendas, E.; Greschner, A.; Liu, J.; de Haan, H.; Gauthier, M. Mechanisms of activity loss for a multi-PEGylated protein by experiment and simulation. *Mater. Today Chem.* **2019**, *12*, 121−131.

(17) Dhalluin, C.; Ross, A.; Leuthold, L.-A.; Foser, S.; Gsell, B.; Müller, F.; Senn, H. Structural and biophysical characterization of the 40 kDa PEG- interferon-α2a and its individual positional isomers. *Bioconjugate Chem.* **2005**, *16*, 504−517.

(18) Bell, S. J.; Fam, C. M.; Chlipala, E. A.; Carlson, S. J.; Lee, J. I.; Rosendahl, M. S.; Doherty, D. H.; Cox, G. N. Enhanced circulating half-life and antitumor activity of a site-specific pegylated interferon-α protein therapeutic. *Bioconjugate Chem.* **2008**, *19*, 299−305.

(19) Dozier, J.; Distefano, M. Site-specific PEGylation of therapeutic proteins. *Int. J. Mol. Sci.* **2015**, *16*, 25831−25864.

(20) Schumacher, D.; Hackenberger, C. P. More than add-on: chemoselective reactions for the synthesis of functional peptides and proteins. *Curr. Opin. Chem. Biol.* **2014**, *22*, 62−69.

(21) Goel, N.; Stephens, S. Certolizumab pegol. *mAbs* **2010**, *2*, 137−147.

(22) Bundy, B. C.; Swartz, J. R. Site-specific incorporation of p-propargyloxyphenylalanine in a cell-free environment for direct protein- protein click conjugation. *Bioconjugate Chem.* **2010**, *21*, 255−263.

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

(24) Shrestha, P.; Smith, M. T.; Bundy, B. C. Cell-free unnatural amino acid incorporation with alternative energy systems and linear expression templates. *New Biotechnol.* **2014**, *31*, 28−34.

(25) Wu, J. C. Y.; Hutchings, C. H.; Lindsay, M. J.; Werner, C. J.; Bundy, B. C. Enhanced enzyme stability through site-directed covalent immobilization. *J. Biotechnol.* **2015**, *193*, 83−90.

(26) Schinn, S.-M.; Bradley, W.; Groesbeck, A.; Wu, J. C.; Broadbent, A.; Bundy, B. C. Rapid in vitro screening for the location-dependent effects of unnatural amino acids on protein expression and activity. *Biotechnol. Bioeng.* **2017**, *114*, 2412−2417.

(27) Saleh, A. M.; Wilding, K. M.; Calve, S.; Bundy, B. C.; Kinzer-Ursem, T. L. Non-canonical amino acid labeling in proteomics and biotechnology. *J. Biol. Eng.* **2019**, *13*, 43.

(28) Yuet, K. P.; Doma, M. K.; Ngo, J. T.; Sweredoski, M. J.; Graham, R. L.; Moradian, A.; Hess, S.; Schuman, E. M.; Sternberg, P. W.; Tirrell, D. A. Cell-specific proteomic analysis in Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 2705−2710.

(29) Thirumurugan, P.; Matosiuk, D.; Jozwiak, K. Click chemistry for drug development and diverse chemical−biology applications. *Chem. Rev.* **2013**, *113*, 4905−4979.

(30) Reddington, S. C.; Tippmann, E. M.; Jones, D. D. Residue choice defines efficiency and influence of bioorthogonal protein modification via genetically encoded strain promoted Click chemistry. *Chem. Commun.* **2012**, *48*, 8419−8421.

(31) Sanyal, T.; Mittal, J.; Shell, M. S. A hybrid, bottom-up, structurally accurate, Gō-like coarse-grained protein model. *J. Chem. Phys.* **2019**, *151*, No. 044111.

(32) Zhao, Y.; Cieplak, M. Proteins at curved fluid-fluid interfaces in a coarse-grained model. *J. Phys.: Condens. Matter* **2020**, *32*, 404003.

(33) Lee, H. Molecular Simulations of PEGylated Biomolecules, Liposomes, and Nanoparticles for Drug Delivery Applications. *Pharmaceutics* **2020**, *12*, 533.

(34) Dandekar, B. R.; Mondal, J. Capturing Protein−Ligand Recognition Pathways in Coarse-Grained Simulation. *J. Phys. Chem. Lett.* **2020**, *11*, 5302−5311.

(35) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-grained protein models and their applications. *Chem. Rev.* **2016**, *116*, 7898−7936.

(36) Wu, C.; Shea, J.-E. Coarse-grained models for protein aggregation. *Curr. Opin. Struct. Biol.* **2011**, *21*, 209−220.

(37) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.

(38) Wei, S.; Ahlstrom, L. S.; Brooks, C. L., III Exploring Protein−Nanoparticle Interactions with Coarse-Grained Protein Folding Models. *Small* **2017**, *13*, 1603748.

(39) Wei, S.; Knotts, T. A., IV A coarse grain model for protein-surface interactions. *J. Chem. Phys.* **2013**, *139*, No. 095102.

(40) Pakula, A. A.; Sauer, R. T. Genetic analysis of protein stability and function. *Annu. Rev. Genet.* **1989**, *23*, 289−310.

(41) Talley, K.; Alexov, E. On the pH-optimum of activity and stability of proteins. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 2699−2706.

(42) Fang, J. A critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. *Briefings Bioinf.* **2020**, *21*, 1285−1292.

(43) Pucci, F.; Bourgeas, R.; Rooman, M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Sci. Rep.* **2016**, *6*, 23257.

(44) Wang, B.; Qi, Y.; Gao, Y.; Zhang, J. Z. A method for efficient calculation of thermal stability of proteins upon point mutations. *Phys. Chem. Chem. Phys.* **2020**, *22*, 8461−8466.

(45) Poteete, A. R.; Hardy, L. W. Genetic analysis of bacteriophage T4 lysozyme structure and function. *J. Bacteriol.* **1994**, *176*, 6783.

(46) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the lysozyme of phage T4. *Protein Sci.* **2010**, *19*, 631−641.

(47) Feher, V. A.; Schiffer, J. M.; Mermelstein, D. J.; Mih, N.; Pierce, L. C.; McCammon, J. A.; Amaro, R. E. Mechanisms for Benzene Dissociation through the Excited State of T4 Lysozyme L99A Mutant. *Biophys. J.* **2019**, *116*, 205−214.

(48) Rennell, D.; Bouvier, S. E.; Hardy, L. W.; Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **1991**, *222*, 67−88.

(49) Matthews, B. W. Structural and genetic analysis of the folding and function of T4 lysozyme. *FASEB J.* **1996**, *10*, 35−41.

(50) Hawkes, R.; Grutter, M. G.; Schellman, J. Thermodynamic stability and point mutations of bacteriophage T4 lysozyme. *J. Mol. Biol.* **1984**, *175*, 195−212.

(51) Kato, H.; Feng, H.; Bai, Y. The folding pathway of T4 lysozyme: the high-resolution structure and folding of a hidden intermediate. *J. Mol. Biol.* **2007**, *365*, 870−880.

(52) Kato, H.; Vu, N.; Feng, H.; Zhou, Z.; Bai, Y. The folding pathway of T4 lysozyme: An on-pathway hidden folding intermediate. *J. Mol. Biol.* **2007**, *365*, 881−891.

(53) Lu, J.; Dahlquist, F. W. Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two-dimensional NMR. *Biochemistry* **1992**, *31*, 4749−4756.

(54) Nicholson, H.; Anderson, D.; Dao Pin, S.; Matthews, B. Analysis of the interaction between charged side chains and the alpha-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* **1991**, *30*, 9816−9828.

(55) Brockerman, J. A.; Okon, M.; Withers, S. G.; McIntosh, L. P. The pKa values of the catalytic residues in the retaining glycoside hydrolase T26H mutant of T4 lysozyme. *Protein Sci.* **2018**, *28*, 620−632.

(56) Smith, M. T.; Wilding, K. M.; Hunt, J. M.; Bennett, A. M.; Bundy, B. C. The emerging age of cell-free synthetic biology. *FEBS Lett.* **2014**, *588*, 2755−2761.

(57) Brooks, B. R.; et al. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(58) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.

(59) Smith, A. K.; Wilkerson, J. W.; Knotts, T. A., IV Parameterization of Unnatural Amino Acids with Azido and Alkynyl R-Groups for Use in Molecular Simulations. *J. Phys. Chem. A* **2020**, *124*, 6246−6253.

(60) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(61) Steinbach, P. J.; Brooks, B. R. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* **1994**, *15*, 667−683.

(62) Hockney, R. W.; Eastwood, J. W. *Computer simulation using particles;* CRC Press: Boca Raton, FL, 1988.

(63) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−1865.

(64) Jo, S.; Cheng, X.; Islam, S. M.; Huang, L.; Rui, H.; Zhu, A.; Lee, H. S.; Qi, Y.; Han, W.; Vanommeslaeghe, K.; MacKerell, A. D.; Roux, B.; Im, W. *Advances in Protein Chemistry and Structural Biology*; Elsevier: Amsterdam, Netherlands, 2014; Vol. 96; pp 235−265.

(65) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.

(66) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511−519.

(67) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.

(68) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177−4189.

(69) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal−isobaric ensemble. *J. Phys. A: Math. Gen.* **2006**, *39*, 5629.

(70) Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(71) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.

(72) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J.; Melo, M. N.; Seyler, S. L.; Domański, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. Proceedings of the 15th Python in Science Conference, 2016.

(73) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(74) Theobald, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2005**, *61*, 478−480.

(75) Liu, P.; Agrafiotis, D. K.; Theobald, D. L. Fast determination of the optimal rotational matrix for macromolecular superpositions. *J. Comput. Chem.* **2009**, *31*, 1561−1563.

(76) Blaber, M.; Zhang, X.-j.; Lindstrom, J. D.; Pepiot, S. D.; Baase, W. A.; Matthews, B. W. Determination of α-helix propensity within the context of a folded protein: sites 44 and 131 in bacteriophage T4 lysozyme. *J. Mol. Biol.* **1994**, *235*, 600−624.

(77) Matthews, B.; Remington, S. The three dimensional structure of the lysozyme from bacteriophage T4. *Proc. Natl. Acad. Sci. U. S. A.* **1974**, *71*, 4178−4182.

(78) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. pH-induced denaturation of proteins: a single salt bridge contributes 3−5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry* **1990**, *29*, 2403−2408.

(79) Zhang, X.-j.; Wozniak, J. A.; Matthews, B. W. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. *J. Mol. Biol.* **1995**, *250*, 527−552.

(80) Yirdaw, R. B.; Mchaourab, H. S. Direct observation of T4 lysozyme hinge-bending motion by fluorescence correlation spectroscopy. *Biophys. J.* **2012**, *103*, 1525−1536.

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