COMPUTATIONAL PROTEIN DESIGN FOR PEPTIDE-DIRECTED BIOCONJUGATION

by

Joseph G. Plaks

B.S., University of Pittsburgh, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Chemical and Biological Engineering

2018

This thesis entitled:
Computational Protein Design for Peptide-Directed Bioconjugation
written by Joseph G. Plaks
has been approved for the Department of Chemical and Biological Engineering

_____
(Dr. Joel L. Kaar)

_____
(Dr. Theodore W. Randolph)

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Plaks, Joseph G. (Ph.D., Chemical and Biological Engineering)

Computational Protein Design for Peptide-Directed Bioconjugation

Thesis directed by Dr. Joel L. Kaar

Proteins enable living organisms to perform many of their critical functions, having been applied over evolutionary time to solve problems of overwhelming diversity and complexity. Protein engineering seeks to deploy these versatile molecules in addressing problems of human concern and would benefit from innovations that improve protein utilization in unnatural environments as well as from increased predictive capability in protein design.

Bioconjugation facilitates the use of proteins in unnatural environments by permitting the attachment of molecules, such as polymers, that can modulate protein stability, solubility, and activity and by mediating protein immobilization. We initially explored this propensity of bioconjugation by designing an enzymatic polyurethane-based material in which proteins were covalently immobilized via non-specific, isocyanate reaction chemistries. The resulting material resisted bacterial biofilm formation through the activity of the embedded enzymes, which hydrolyzed signaling molecules involved in quorum sensing.

In order to gain better regional and temporal reaction control, we transitioned to working with an enzyme-mediated conjugation system in which lipoic acid ligase functionalizes a specific 13-residue peptide sequence (the LAP sequence) with an azide-bearing lipoic acid derivative. Using GFP as a model protein, we demonstrated that the LAP sequence could be inserted at internal positions within a protein's structure and successfully ligated and subsequently modified via azide-alkyne click chemistry within that context.

Given the ability of the LAP sequence to site-specifically direct conjugation at internal sites within protein structures, we developed a predictive computational approach to facilitate the design of internal LAP insertion sites within diverse protein targets. The kinematic loop modeling application within the Rosetta framework was adapted for rapid scanning and characterization of insertion sites, using Rosetta's coarse-grained centroid score function for site differentiation. Soluble protein expression of LAP-containing proteins was found to correlate with Rosetta scores, and unintuitive sites were identified relative to *B*-factor and secondary structure considerations. Our results highlighted the role played by residues in the near-loop environment in determining whether a particular site within a protein accommodates an inserted loop, such as the LAP sequence, and suggest that our computational approach, which is highly user accessible, can usefully predict such sites.

DEDICATED

TO

PANTELEJMON PLAKS

ACKNOWLEDGEMENTS

I have had the privilege of encountering many wonderful people over the course of my life who have helped to make this thesis possible and whom I must acknowledge and sincerely thank. Joel has been a tremendous PI, striking a balance in providing me with both support and independence. My committee, likewise, has been very helpful, and I would particularly like to thank Amy Palmer and Ted Randolph for their participation and encouragement. Additionally, I would like to thank Andy VanDemark and Nicolas Sluis-Cremer who were my undergraduate PIs and who taught me many of the foundational skills upon which I have relied so heavily over the years.

The day-to-day business of graduate school is performed alongside fellow graduate students and lab members, all of whom I should also like to thank. My early days in the Kaar Lab were spent with Erik Nordwald, Kelsey MacConaghy, Núria Codina Castillo, and Sean Yu McLoughlin—my latter days with David Faulón Marruecos, James Weltz, Sam Summers, Jared Snell, Garrett Chado, Andres Chaparro Sosa, Becca Falatach, and Louis Sacks. Working with these people was a terrific experience, and I am so grateful for how reliable, pleasant, and engaging they have been. In addition to my CU colleagues, I would like to thank the graduate students and other lab members who mentored me when I was an undergraduate working in the VanDemark and Sluis-Cremer labs, especially Swarna Mohan, Adam Wier, Chris Amrich, Aubrey Lowen, and Kelly Huber. I own so much to the time and effort that these people invested in me.

Nothing that I have ever achieved in science or elsewhere would have been possible without my family, whom I must also thank. My wonderful parents gave me the best of all possible childhoods and made no exception in sacrificing for the sake of my education. Only now that I have my own child do I realize how impossible it will be to ever truly thank them. My brothers, Nick and Pete, were my constant companions growing up and remain my best friends. I am so grateful for all that they have done and continue to do for me. My grandparents, Pantelejmon and Ksenia Plaks, also deserve a great deal of thanks. They were the ones who brought our family to this country and are ultimately responsible for the great opportunities and lives that we have lead here. My thesis is dedicated, with all of the thoughts and experiments and published scientific papers that it contains, to my grandfather, Pantelejmon. He was wonderful, generous, and curious man. I think that he would have been pleased to have seen it.

Finally and most of all, I would like to thank my wife, Amy, who has brought meaning and happiness into my life ever since the day that I met her. Nothing that I do would be possible or even worthwhile without her. Graduate school, with its open-ended timeframe, its insecurities, and, in the case of CU Boulder, its incredible distance from home, is a difficult thing to ask of a partner, and I am so grateful for the sacrifices that she has made in allowing me to pursue it. I am also grateful for the family that she has given me with the birth of our son, Jack. Scientists have an eye for the future, and at one year old, Jack is the future's representative in my life. He is my inspiration and my joy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Protein molecules, commonly referred to as the "workhorses of cells," are absolutely remarkable in their versatility. They have evolved to address many of the fundamental challenges encountered by life. These include challenges associated with energy acquisition and utilization, information replication and transduction, sensing, locomotion, and organization.[1] Certain proteins contribute to the material properties of tissues. Others bind to epitopes with extraordinary affinity and specificity. In the case of enzymes, precise binding specificities are combined with catalytic activities that are capable of increasing reaction rates by many orders of magnitude at mild temperatures, pH, and pressures.[2]

Proteins achieve their biological significance in part by generating structural and conformational complexity from relative chemical simplicity. There are only 20 amino acids commonly employed by the proteins in the natural world, and among these, there are redundancies in chemical characteristics. However, when expressed within particular solvent environments, proteins adopt low energy conformations that can exhibit extremely precise structures.[3] Given the combinatorial space afforded by linking essentially any number of 20 building blocks in essentially any order, the potential sequence and structural diversity that can be achieved with proteins is astronomical. Appearing within these possibilities are conformations with highly useful binding, catalytic, and mechanical properties.

The utility of proteins in nature leads easily to the idea that proteins as a general class of molecules may be usefully applied in solving problems of human interest. There are many problems in medicine,[4] industrial catalysis,[5] and sensing[6] that proteins currently address or that

proteins could in theory address. Additionally, there are many naturally occurring proteins that may underperform within human imposed constrains and might therefore be coxed into working more effectively within those given parameters.[7] The goal of protein engineering is to facilitate the application of protein molecules to these types of novel problems. There are many challenges associated with doing so, but two of the key challenges that are partially addressed within this thesis pertain to extending the utility of proteins in unnatural environments and achieving predictive capability in protein design.

Many novel applications of proteins require tolerance to unnatural environments. Even in cases where human proteins are injected into human patients, upstream and downstream processes in production, transport, storage, and administration all entail some level of interaction with unnatural environments outside of the context in which the proteins evolved. In cases of industrial catalysis and sensing these environments can become particularly extreme. Proteins are often immobilized to solid supports in ways that impact their stability and reusability or, as is the case with sensors, in ways that mediate signal transduction.[6] Additionally, proteins are used in non-physiological solvent environments that can vary in range of severity from subtle changes in pH or ionic strength to the use of proteins in non-aqueous solvents such as alkanes or ionic liquids.[8]

Protein bioconjugation techniques are often employed when engineering proteins for use in unnatural environments and play an important role in protein science more generally. Bioconjugation can mediate protein immobilization, in which case the immobilization matrix itself becomes an unnatural environment with which the protein must contend in addition to whatever environment the protein-containing material is subsequently exposed to.

Bioconjugation can also be used to modify proteins with molecules that enhance their function in unnatural environments. Polymer conjugation is a prime example. Conjugating polymers to therapeutic proteins can lead to improved pharmacokinetics and reduce the immunogenicity of the protein being modified.[9] Additionally, polymers have been used to improve protein solubility in non-aqueous solvents and to mediate protein recycling through temperature dependent phase transitions. Within protein science more generally, conjugation of fluorescent probes can be used to permit protein tracking *in vitro* and *in vivo*. Doubly labeling proteins with fluorophores allows such characterizations to include conformational tracking via FRET,[10] which can also be used to track protein interactions via intermolecular double –fluorophore labeling.[11]

One of the benefits of applying bioconjugation reactions to permit protein utilization in unnatural environments is that doing so enables the pursuit of creative solutions to problems that would otherwise be inaccessible. The prevention of bacterial biofilms through the enzymatic disruption of quorum sensing is an excellent example of this. Preventing biofilm formation on a medical device could be achieved by the slow release of antibiotics or other various antimicrobial compounds, but such strategies eventually lead to the depletion of the released biocidal agent.[12] Additionally, directly killing bacteria, which would occur with this approach, leads to a strong selective pressure for the development of bacterial resistance. By immobilizing a quorum quenching enzyme within a material via multiple covalent conjugation points, both of these shortcomings can potentially be addressed. An enzyme, if immobilized through this technique, could be very stable and unlikely to leach into the surrounding solution. Consequently, unlikely a released biocidal agent, the anti-biofilm activity would not be depleted over time. Additionally, by targeting bacterial signaling that leads to biofilm formation, as is

done with a quorum quenching enzyme, biofilms can be prevented without directly killing bacterial cells.[13] This reduces the selective pressure experienced by the bacteria and reduces the likelihood of an evolved resistance measure.

There are, unfortunately, limitations associated with protein bioconjugation. For example, bioconjugation reactions can have a negative impact on the target protein being conjugated. This is especially true for non-specific chemistries that react with primary amines or other moieties typical to a protein surface. Non-specific reactions, additionally, are difficult to perform in complex chemical environments without leading to unintended side reactions. Consequently, bioconjugation reactions that allow for regional and temporal control, where conjugation only occurs at a specific site on a specific target protein with the addition of a specific stimulus, are highly desirable.

A number of strategies can be used to achieve site specificity with bioconjugation reactions. For example, a reactive residue, such as a cysteine, can be introduced at a particular site in a protein via mutagenesis with the simultaneous removal of all non-target cysteine residues in the protein to create a single, unique conjugation site. This approach, however, is not feasible when the protein contains a cysteine residue that has a critical functional role, and it does not limit conjugation to the target protein in environments that contain other biomolecules, which commonly contain cysteine.

Bioorthogonal reaction chemistries are inert with respect to common chemical functional groups in biology and can be used to site-specifically direct conjugation without the need to mutate away particular amino acids or to perform the conjugation reaction in the absence of other contaminating bio-molecules. Bioorthogonal amino acid residues can be introduced into

proteins via unnatural amino acid (UAA) incorporation using engineered tRNA/synthetase pairs in combination with amber stop codon suppression.[14] This approach has a lot of potential for facilitating bioconjugation reactions and is prevalent in the literature. However, UAA incorporation often reduces target protein expression levels, and amber stop codon suppression is not easily extended for the incorporation of more than one unique UAA without extensive genomic engineering or the adoption of a four-letter genetic code. Additionally, although the amber stop codon is not used for translation termination as commonly within genomes as other stop codons, it is still prevalent, and any protein within the target organism that relies on the amber stop codon will receive a C-terminal UAA. By failing to terminate, these proteins will also be expressed with random peptide tails of arbitrary length because translation will not terminate until another in-frame stop codon is happened upon. As a result, even incorporation of a bioorthogonal reactive UAA will not limit bioconjugation to the target protein when the conjugation reaction is performed in the presence of the host cell cytosol.

The final limitation of using bioorthogonal UAAs to mediate bioconjugation is that the reactive handle must be introduced in the target protein during translation. Although translation can be controlled with various operons and induction agents, once expressed, temporal control of conjugation can only be achieved by controlling when the conjugation partner is introduced into the reaction. As a consequence, a UAA-containing protein cannot, for example, be expressed throughout a cell and then conjugated only within a particular organelle as might be desired for certain live cell protein trafficking experiments.

Enzyme-mediated bioconjugation techniques provide both site-specificity and enhanced temporal control, making them particularly attractive.[15] Within these approaches, a peptide

recognition sequence is genetically fused to a target protein and is site-specifically recognized by an enzyme, which either directly conjugates a second molecule to the peptide sequence or introduces a bioorthogonal reactive handle that can be targeted in subsequent reactions. This type of approach is therefore peptide-directed. Conjugation occurs only within the specific peptide recognition sequence, resulting in a high degree of site specificity with respect to the target protein and precluding conjugation to proteins (such as cellular proteins) that might be present in the conjugation reaction but lack the required peptide. Furthermore, requiring an enzymatic step for conjugation introduces an additional temporal control parameter, as the expression of the enzyme as well as the location of the enzyme within different cellular compartments can be independently controlled. As a result, conjugation to a target protein within a cell can be dependent upon whether that protein enters a particular organelle, provided that the requisite enzyme is uniquely targeted to that organelle.

The genetic fusion of a recognition peptide to a target protein is similar in some ways to the fusion of larger protein molecules that facilitate labeling or conjugation. GFP-fusions, for example, are commonly used for protein trafficking experiments in living cells and SNAP-tags can be introduced to mediate conjugation.[16,17] Small peptide tags have been developed in part to reduce the likelihood of unintended consequences that might arise from the fusion of a target protein with a larger protein partner such as GFP or a SNAP-tag. However, even with their smaller size, the addition of a peptide may lead to unintended consequences, especially if the termini of the target protein are essential to activity, as is the case with human phosphatase and tensin homolog (PTEN).

PTEN is a tumor suppressor protein that has physiological roles in the nucleus as well as the cytoplasm of cells.[18] In the Cytosol, PTEN catalyzes the dephosphorylation of membrane lipids, which would, in the phosphorylated form, contribute to a signal cascade involved in cell division. Consequently, recruitment of PTEN to the cell membrane is critical for its function. The localization of PTEN to the membrane is mediated by N-terminal residues, which form a membrane-binding domain. The C-terminus of PTEN also contributes to localization but with respect to other protein complexes. Specifically the final residues at the C-terminus of PTEN constitute a PDZ-binding domain that interacts with the PDZ domain of a membrane-associated protein. Although the trafficking behavior of PTEN has previously been observed using a GFP-fused construct, such experiments would benefit from the conjugation of a fluorophore to an internal site within PTEN rather than through the fusion of a protein or a peptide to one of the termini.[19] Also, because PTEN has a nuclear role, the ability to control the compartment in which labeling occurs, as can potentially be done with an enzyme-mediated conjugation system, is highly desirable. Further, because of the critical role that it plays, PTEN expression is heavily regulated, and overexpressed PTEN constructs are targeted for degradation. Therefore, a highly bioorthogonal labeling scheme that reduces background labeling within the cell would also be desirable for trafficking experiments with this protein.

Inserting peptide tags at internal positions within proteins would allow the high degree of regional and temporal control that enzyme-mediated, peptide-directed conjugation systems exhibit while avoiding the shortcomings associated with limiting protein conjugation to target protein termini. Inserting a novel peptide at an internal position within a protein without impacting structure, stability, or other target protein properties, however, is a significant

challenge. Larger peptide recognition sequences, such as the 13-residue Lipoic acid Acceptor Peptide (LAP) sequence have high specificity and good reaction kinetics in part due to the number of residues that they contain.[20] But due to their size, they can be especially detrimental when inserted at non-terminal positions within a protein. Larger loop regions within proteins have a destabilizing impact in general,[3] and when they disrupt regions of structure, as would likely occur when inserting a LAP sequence within a protein, the potential for a negative outcome increases.

The difficulty associated with engineering proteins with peptide insertions is particularly frustrating given that peptide loop insertions have broader applications beyond enzyme-mediated bioconjugation. Antibody-binding loops, for example, are often engineered into non-immunoglobulin proteins to create chimeras that combine the binding properties associated with antibody loops with novel properties and functionality associated with the protein scaffold.[21,22] But the disruptive potential of loop insertions limits the types of constructs that can be designed through this approach and, consequently, the types of applications that can be targeted. Together with the benefits of peptide-directed conjugation, the relevance of loop insertions in the design of chimeric protein binders suggests that the ability to predict sites within target proteins that can accommodate loop insertions is of broad general interest.

Predictive capability is an important component of any engineering discipline and is implied in discussions of design. Achieving predictive capability in protein engineering, however, still represents a significant challenge. The same characteristics that make proteins so versatile and useful for living organisms—the complex structural potential of 20 amino acid building blocks that each possess relatively simple and in some ways redundant functional

groups—make proteins particularly difficult to engineer. With the elucidation of the genetic code, perfect predictive capability has been achieved in designing DNA sequences that will produce a particular protein primary sequence. However, predicting how a particular protein sequence will behave in solution relative to folding or the creation or maintenance of any higher-order property is extremely difficult. In other words, predicting the impact of a genetic change on the structure and behavior of a protein persists as a significant challenge and limitation within protein engineering.

Computational approaches have been employed in pursuit of predictive capability in protein design and have been enjoying increasing success, sophistication, and relevance. The Rosetta protein modeling software suite is one such computational platform that has been developed to predict the folded state of a protein based on its primary sequence. Given structure/function relationships that exist in proteins, structural prediction capabilities extend naturally to protein design, and the Rosetta framework has been applied productively in this regard, allowing for the de novo design of enzymes, peptide binders, and multi-protein complexes.[21,23–27] The success of Rosetta stems in part from its use of the massive amounts of information available in the Protein Data Base (PDB) for limiting the conformational space that needs to be searched in a given protein sequence and for scoring different potential protein conformations.

As a potential platform for predicting loop insertion sites and therefore designing peptide-directed conjugation sites, Rosetta has a number of advantages. It is fairly accessible, which is an important quality if it is to be adapted by protein engineers working in diverse areas ranging from live cell protein imaging to protein immobilization. It also has a loop modeling application that, although not previously used for characterizing loop insertion sites, can potentially be

extended to doing so with minimal operational changes.[28] Adapting Rosetta for predicting loop insertion sites in proteins, in addition to generating new predictive capabilities in protein design, would advance the use of peptide-directed, enzyme-mediated bioconjugation strategies. This would allow for broader dissemination of the benefits that these approaches provide in terms of regional and temporal control over conjugation reactions. In turn, these conjugation benefits would extend more generally to the use of proteins in unnatural environments by facilitating protein modification and immobilization.

BACKGROUND

Adapted in part from Enzyme-mediated Ligation Technologies, Springer Protocol (submitted)

2.1 Protein Bioconjugation

2.1.1 Non-Specific Reaction Chemistries

The term, bioconjugation, can generally refer to any persistent interaction between a bio-macromolecule and another chemical entity whether biological or not. Of interest here are bioconjugation reactions that result in a covalent linkage between a protein and another molecule and that are formed with the intention of fulfilling a functional role in a protein engineering or protein science application. The simplest way of achieving a bioconjugation linkage that is consistent with this more particular definition is via non-specific chemical reactions that target function groups, such as particular amino acid side chains, commonly found on the surface of a protein.

A number of protein residues can act as nucleophiles in conjugation reactions, including lysine residues, which contain a commonly targeted primary amino group. Epoxides, isocyanates, and N-hydroxysuccinimide (NHS) groups all react with primary amines and will therefore target lysine residues as well as the primary amine found at the N-terminus of proteins.[29] Other residues, such as cysteines, can also act as nucleophiles, in this case via malemide conjugation reactions. Cysteines, additionally, can form linkages with other thiols under oxidative conditions in order to facilitate conjugation.

Non-specific reaction chemistries are often easy to perform under aqueous conditions, making them appropriate for protein bioconjugation. Although many of the electrophile pairs

mentioned above will also react with water under aqueous conditions (and therefore cannot be stored under those conditions), they tend to react much more quickly with the appropriate protein functional group and can therefore lead to high levels of conversion with respect to the protein when applied in excess within the reaction. The main disadvantage of these types of approaches is the lack of control over the precise location of conjugation. Aside from the inability to limit these reactions to a single protein target in a complex biological environment, such as within the interior of a cell, such approaches can lead to enzyme inactivation when a highly nucleophilic amino acid resides in or near the active site.

The nucleophilicities of protein surface residues can be highly dependent on their environment. Because many enzymatic mechanisms proceed through a step in which one of the amino acid residues in the active site acts as a nucleophile to form a temporary covalent intermediate with the substrate, many active sites are configured in such a way as to increase the nucleophilicity of the pertinent residue.[30,31] This has a substantial consequence when performing bioconjugation reactions with non-specific chemistries. A highly nucleophilic residue in the active site that has a direct mechanistic role in enzyme catalysis cannot be mutated away to prevent bioconjugation at that site as doing so would abolish activity. Further, leaving the residue intact in the hopes that the majority of conjugation events will occur with the same residue type elsewhere on the protein surface is often misplaced, as conjugation may preferentially occur at these residues with increased nucleophilicity.

Adding a competitive enzyme inhibitor in the reaction can in some cases prevent enzyme deactivation associated with non-specific chemical conjugation with active site residues.[32] The inhibitor will bind to the active site and block conjugation. Once the reaction is complete and all

of the reactive groups are quenched, the inhibitor can be dialyzed away to yield active, conjugated enzyme. Naturally, such approaches depend on the existence of an appropriate inhibitor. Additionally, a corollary approach for different types of conjugation reaction-mediated deactivation events is not always available. For example, if a conjugation event were to result in the disruption of an obligate dimer, preventing it with some type of blocking agent, such as the inhibitor used to block the active site in the above discussion, would be difficult.

2.1.2 Unnatural Amino Acids as Bioorthogonal Reactive Handles

Introducing an unnatural amino acid (UAA) into a protein with a bioorthogonal R-group can facilitate highly specific conjugation to that residue and consequently can overcome some of the limitations associated with non-specific conjugation chemistries.[33] This process requires a designed tRNA/synthetase pair that must also be bioorthogonal with respect to other tRNA and amino acid molecules common to the cells in which it is employed. This bioorthogonality is achieved in two ways. Firstly, the tRNA/synthetase pair that will direct UAA incorporation is transplanted into the expression cells from another distantly related organism. This helps to ensure that none of the implanted tRNA will be bound by endogenous synthetases and functionalized with canonical amino acids and that the endogenous tRNAs will not be functionalized with the UAA that is being added to the cells for incorporation. The second way in which UAA incorporation is achieved is through engineering of the synthetase to alter its amino acid substrate specificity. This is commonly done via directed evolution in which positive selection is used to identify mutant constructs with high affinities for the desired UAA, and

negative selections are used to eliminate constructs with low selectivities, that is, constructs that promiscuously bind to and functionalize tRNA with undesired canonical amino acids.

Although many UAAs have been successfully incorporated into proteins with engineered tRNA/synthetase pairs, many of the engineered synthetases used in this process have been derived from one of only two staring points: tyrosine synthetase or pyrrolysine synthetase.[34] As a result, there is some overlap in the chemical characteristics of the UAAs that have been incorporated in proteins. For example, many UAAs that utilize an engineered tyrosine synthetase contain a phenyl ring. Similarly, many of the UAAs that are incorporated by an engineered pyrrolysine synthetase contain an amide bond appended to a 4-carbon chain. In spite of these practical limitations in chemical diversity, however, UAAs with diverse moieties including reactive functional groups that enable click chemistry reactions have successfully been incorporate into proteins with engineered synthetases. Of note are UAAs with azide functionalities that can undergo copper catalyzed cycloaddition reactions with alkynes or strain-promoted cycloaddition reactions with strained alkynes.[35,36] These reactions are bioorthogonal in the sense that neither azides nor alkynes appear commonly in biological molecules as well as in the sense that the reaction kinetics for an azide/alkyne reaction are faster than they are for the reaction of other common biological functional groups with either of these two moieties. As a consequence of these characteristics, azide/alkyne click chemistry reactions can be performed with relatively low concentrations of both reagents in a complicated chemical environment in which the concentration of other, biological functional groups is far greater.

In spite of its advantages over some non-specific conjugation approaches, there are a number of shortcomings associated with UAA incorporation for site-specific bioconjugation. As

discussed in the introduction, UAA incorporation when performed in a cell via amber stop codon suppression results in UAA addition to all endogenous proteins with TAG stop codons in addition to the protein being specifically targeted for UAA incorporation. Additionally, because the UAA is introduced during translation, the reactive handle cannot be added specifically within a particular organelle, as can be done with enzyme-mediated conjugation systems. This dependence on translation for reactive handle incorporation can lead to limitations in other processes besides *in vivo* protein labeling. For example, azide groups are sensitive to UV-induced deactivation. If a purified protein with an azide moiety is required, the azide must be introduced during translation with a UAA approach and care must be taken during the purification process to avoid exposure to light. The azide group cannot be added after the purification is complete.

Limitations associated with the engineered synthetase themselves may also lead to issues with UAA directed approaches to bioconjugation. Using a stop codon to code for UAA incorporation can result in truncation. Although various *E. coli* strains have been engineered to reduce this problem,[37] it persists in other relevant organisms in which *in vivo* protein conjugation may be desirable. Additionally, even though the synthetase has been engineered for its target amino acid, miss-incorporation of a canonical amino acid is still possible, especially for synthetases that have been evolved using tyrosine synthetase as a starting point.

The final, subtler limitation of UAA-mediated conjugation is that it must proceed through a synthetase. Enzyme-mediated conjugation approaches (discussed in the next section) can utilize vastly diverse enzymes to facilitate conjugation. UAA incorporation requires a synthetase and therefore must operate within active site and other constraints associated with these proteins. For example, it is important that mutations that alter amino acid binding specificity do not alter

tRNA binding, which must be preserved. Also, additional constraints are imposed by the amino acid binding cleft, which has a practical upper limit in terms of the size of the UAA that it can be engineered to accommodate. Consequently, conjugation techniques that allow for a greater diversity both in terms of controlling the time at which a reactive handle is incorporated into a protein and in terms of the enzymes that facilitate this incorporation are desirable.

2.1.3 Enzyme-Mediated Conjugation

Protein bioconjugation strategies that proceed through an enzymatic step have several advantages over conventional non-specific chemical conjugation approaches. These advantages arise from characteristics that are often inherent to enzymes, including high selectivity and reaction rates at moderate substrate concentrations, temperatures, pressures, and pH. The ability to express enzymes in cells and to specifically target them to certain organelles allows such approaches to be used *in vivo* and creates additional control parameters that can be exploited when designing an experiment. For example, conjugation of a fluorophore to a protein can be restricted to only occur within a particular organelle.[38] Furthermore, the high substrate specificity of many enzymes allows for bioorthogonal conjugation reactions, resulting in homogeneous products even when performed amid a complex and crowded chemical environment.[15]

Enzyme-mediated bioconjugation typically involves an enzyme/peptide pair derived from a naturally occurring post-translational modification system.[39,40] The peptide consists of a particular recognition sequence that can be fused to a protein of interest (POI) in the form of a tag that is site-specifically modified by the enzyme through various mechanisms. These mechanisms include acyl transfers between glutamine residues and primary amines (*i.e.* using

transglutaminase),[41] peptide transfers (in which a peptide bond is cleaved and re-ligated to a different peptide target molecule, as catalyzed by, for example, sortase),[42] and ATP-dependent ligation reactions of carboxyl groups to the ε-amine of lysine residues (catalyzed by ligases such as biotin ligase).[43] Importantly, enzymatic modification of the peptide does not need to directly lead to conjugation. Formylglycine generating enzyme (FGE), for example, creates an aldehyde moiety within its peptide recognition sequence through the oxidation of a cysteine residue. This aldehyde, once formed, can be targeted for site-specific conjugation reactions given its unique chemical properties relative to the other reactive moieties present in the protein.[44]

The differences in reaction mechanisms and peptide recognition sequences of these systems present various strengths and weaknesses depending on the application. For example, the recognition sequence of transglutaminase, XXQXX, where X can be any proteinogenic amino acid, may not be sufficiently unique when a high degree of site-specificity is required. Sortase reactions result in cleavage of the peptide backbone and therefore cannot be used to perform conjugation reactions at internal sites within a protein in the absence of a specifically designed disulfide staple.[45] Lastly, FGE reactions, which do not suffer from low site-specificity or peptide cleavage, must proceed through aldehyde chemistry for subsequent conjugation steps, which precludes the utilization of more rapid and bioorthogonal click reactions.

Of the various enzyme-mediated bioconjugation approaches, the use of lipoic acid ligase (LplA) offers a number of advantages in terms of recognition peptide specificity and modularity as well as in terms of chemical versatility and user accessibility.[46] *Escherichia coli* (*E. coli*) LplA recognizes and site-specifically functionalizes the 13-residue (GFEIDKVWYDLDA) Lipoic acid Acceptor Peptide (LAP) through an ATP-dependent reaction using magnesium as a cofactor.[47]

The reaction forms a new amide bond between the ε-amino group of the central lysine residue within the LAP sequence and the carboxyl group of lipoic acid, effectively creating a covalent, residue-specific link between the two. The LAP sequence as it appears above was engineered *via* yeast display directed evolution using sequences from several naturally occurring LplA protein targets as a starting point.[20] As a result, the LAP sequence, unlike the natural target sequences of LplA, is highly modular and can be efficiently ligated as a free peptide in solution as well as when fused to the *N*- or *C*- terminus of a target protein. Further, as a consequence of directed evolution through yeast display, the ligation reaction is rapid and has high specificity for the LAP sequence, making it tremendously bioorthogonal.

Another advantageous property of the LplA/LAP system is the amenability of the LplA lipoic acid binding cleft to re-engineering through rational and computational approaches. For example, the W37V mutant of LplA identified by Uttamapinant et al.[38] accommodates a 7-hydroxycoumarin substrate in its binding cleft. This facilitates the direct ligation of a fluorophore to the LAP sequence and is therefore extremely useful for the *in vivo* labeling of LAP-fused target proteins. This type of fluorophore incorporation was taken a step further by Liu et al. who utilized Rosetta-based computational protein design to identify a triple residue mutant (E20A, F147A, H149A) of LplA that binds and directly ligates resorufin to the LAP sequence.[48] Additionally, LplA mutants have been identified that recognize and ligate azide, aldehyde, and hydrazine-functionalized lipoic acid analogues.[49,50] Although wild-type LplA has some activity towards azide and alkyne functionalized fatty acids,[51] the W37V mutant was found to be especially efficient at ligating 10-azidodecanoic acid. The ability to ligate this molecule creates a path for the robust, site-specific addition of an azide reactive group to LAP-fused proteins.

Introducing an azide moiety into a target protein opens the door to numerous and diverse conjugation reactions for protein engineering applications. As is the case with azide-bearing UAAs, azide groups introduced in a protein via an enzyme mediated technique may be targeted for conjugation with alkyne functionalized molecules.[36,35] They can also react with phosphines in Staudinger ligation reactions.[52] Such reactions can be rapid and bioorthogonal, and because many diverse molecules can be functionalized with alkynes or even purchased with such functionality already in place, they can facilitate many types of chemical modifications. Moreover, due to the designable nature of the LplA binding cleft, lipoic acid derivatives with other clickable functionalities (*e.g.*, tetrazines and trans-cyclooctenes) may potentially be usable in the future. These additional chemistries would further expand the utility of LplA-mediated protein modification through improvements in reaction rates, enhanced conjugation selectivity in complex reaction mixtures, and increased stability of the clickable moiety.[53,54]

2.2 Bioconjugation for Facilitating Protein Immobilization

2.2.1 Immobilization Approaches

Proteins can be immobilized through a number of strategies that vary in complexity and provide varying levels of utility. Adsorption, in which the protein noncovalently adheres to a surface based on properties such as charge and hydrophobicity, is one relatively simple and common approach.[55] Entrapment of a protein within a material matrix is another technique employed for immobilization that does not require a covalent linkage between the protein and the immobilization material (or substrate as the material to which a protein is immobilized is commonly called).[56] By avoiding conjugation, adsorption and entrapment can reduce the

likelihood of conjugation-dependent deactivation of the protein. However, leaching of the immobilized protein from the surface or material is more likely when covalent conjugation is not employed. Additionally, controlling specific aspects of the immobilized protein, such as its orientation relative to the surface, is difficult.

Bioconjugation can be used in protein immobilization to address problems associated with protein leaching or the lack of control over orientation that might be seen with the above approaches.[57] Perhaps the simplest application of conjugation chemistry in protein immobilization is seen with the crosslinking of protein aggregates and crystals.[58] In these processes, there is no secondary material. Rather, the proteins are immobilized to each other to form their own macroscopic material. The conjugation reactions typically involve protein surface residues reacting with a chemical crosslinker such as glutaraldehyde. Protein crosslinking can be taken a step further by using a polymer as the crosslinker. Specifically, single and multicomponent polymer formulations that crosslink with themselves as well as with protein surface residues can be used to generate protein-containing materials. This approach benefits from the fact that the polymers, which can be chemically tuned, can dominate the material properties of the system. This allows the materials to be used in more diverse applications while still maintaining the biological properties that the immobilized protein bestows on the material, such as catalysis in cases where enzymes are immobilized. Additionally, the diversity of polymers allows for better control over properties such as enzyme substrate penetration.

Protein immobilization in polymer matrices also often uses non-specific chemistries that target protein surface residues. As a consequence of this type of conjugation chemistry, proteins are usually immobilized at multiple points on their structure, resulting in what is commonly

referred to as multi-point covalent immobilization. Multi-point covalent immobilization commonly arises when non-specific conjugation chemistries are used even with different material architectures. For example, immobilization to a two dimensional self-assembled monolayer can result in multi-point covalent immobilization depending on the spacing of the reactive groups on the surface and the target residues on the protein.[59]

Although site specific chemistries could be employed for multipoint covalent immobilization, doing so would require the incorporation of numerous bioorthogonal reactive handles either via UAAs or enzyme-mediated approaches. Because of this, site-specific bioconjugation reactions typically appear in protein immobilization when proteins are tethered to a surface with a single conjugation point. By tethering a protein to a surface through a single, unique conjugation point, aspects of the immobilized protein, such as its orientation, can be controlled.[60] Controlling the orientation of an immobilized enzyme can increase activity retention of the enzyme by ensuring substrate access and preventing the disruption of the active site that might result depending on where the active site is with respect to the point of conjugation.

2.2.2 Impact of Immobilization on Protein Stability

Immobilization can impact the stability of a protein, notably through confinement. As discussed in section 2.3.1, the difference in conformational entropy between the folded and unfolded states of a protein is a driver for unfolding. When a protein is confined within an area that is similar in size to its folded state, fewer conformations of the unfolded state are available

than would be otherwise. The conformational entropy of the unfolded state is therefore reduced relative to the folded state, as is the driving force for unfolding.[61]

Confinement is easy to envision when a protein is entrapped in a three dimensional material or within a polymer matrix. However, the effect is also relevant for a protein that is tethered to a two-dimensional surface since the existence of the surface limits some of the available unfolded state conformations.[62,63] As a result, immobilization almost always has the potential to have a stabilizing effect on a protein. Whether it does or not depends in large part on the types of interactions that are made between the protein and the surface. Surfaces with different chemical properties and different binding preferences for a protein's folded and unfolded states can impact stability in a negative way in spite of a stabilizing confinement effect.

In addition to the potential of confinement for having a stabilizing effect an immobilized protein, the linkages between an immobilized protein and the material that it is immobilized to can serve to physically hold the protein in place and prevent it from unfolding.[60] This is especially relevant with conjugation approaches that lead to multi-point covalent immobilization. Conjugation to a surface can also allow a protein to be associated with a surface that is stabilizing but that the protein would otherwise have trouble adhering to. For example, nitroreductase is stabilized when immobilized to lipid bilayers under conditions that result in the protein and the bilayer having like charges.[64] In this case, protein association with the bilayer would be difficult to maintain in the absence of conjugation.

2.2.3 Applications

Protein immobilization is tremendously useful for various applications in part due to the increase in stability that can result upon immobilization, which can promote the use of proteins in unnatural and potentially destabilizing environments. Although the immobilization surface or material is itself an unnatural environment, once immobilized, proteins can be subjected to conditions such as elevated temperatures and dry storage that are more extreme than the conditions under which they have evolved. These conditions may result within the context of a bioreactor or when a protein is immobilized to an electrode for sensing applications.[6]

The utility of protein immobilization with regard to diverse applications results from the combination of properties that is possible when proteins are interfaced with macroscopic materials. Proteins have highly specific binding and catalytic properties that are difficult to reproduce with bulk materials. But by immobilizing a protein on or within a material, those properties can be added to the material properties of the immobilization matrix, which can lead to creative solutions to otherwise intractable problems. For example, enzyme containing polymer foams can be used to specifically sense and/or eliminate neurotoxins in a highly selective manner.[65] Enzymes immobilized to metal particles can easily be separated from reaction solutions based on their magnetic properties.[66] Biofuel cells can selectively oxidize target fuels due to the enzymes that are immobilized, and the selectivity of these systems can be extended for sensing applications such as the analysis of blood glucose levels.[6]

2.3 Computational Design of Loop Insertion Sites

2.3.1 Impact of Loop Insertions on Protein Stability

As briefly discussed above, the stability of a protein within a particular solvent environment depends on the energetics associated with the folded and unfolded states. The unfolded state has a high degree of conformational entropy relative to the folded state. Different residues within the protein can have a variable impact on the change in conformation entropy associated with unfolding due to the varying accessibility of torsion angles associated with different amino acids. For example, glycine residues accommodate a greater breadth of phi psi angles, all of which will likely be explored in the unfolded state. Because of this, a structured glycine in the folded state contributes to the change in conformational entropy upon unfolding to a larger degree than a structured β-branched amino acid such as valine that has more restricted torsion angles in the unfolded state.[67]

Various intramolecular interactions within the folded state of a protein counteract the tendency to unfold that results from the disparity of conformational entropy between the folded and unfolded states. These include hydrogen bonds, salt bridges, disulfide bonds, and van der Waals forces.[67,68] The hydrophobic effect also contributes substantially to stabilizing the folded state. Interestingly and perhaps counter intuitively, many of these interactions contribute to increasing the entropy of the total system in spite of the fact that they stabilize the folded state with its low conformational entropy.[3] For example, intramolecular hydrogen bond formation in proteins results in the liberation of a water molecule that was formerly bound to the protein and therefore increases entropy.

Flexible, unstructured loop regions in proteins tend to have a destabilizing impact on the protein's folded state for reasons that pertain to conformational entropy and the presence or

absence of intermolecular interactions. Flexible loop regions in proteins have a greater amount of conformational entropy than more rigidly structured portions of a folded protein. However, flexible loops are still constrained at their termini, where they connect to the rest of the folded, structured protein. These constraints mean that the loop region itself experiences a loss in conformational entropy when the rest of the protein folds that is associated with the confinement of the termini of the loop in three-dimensional space. Additionally, because the loop is flexible and unstructured, there are relatively few intramolecular interactions that can counteract the loss in conformational entropy that results when the protein folds and the two ends of the loop are constrained. Consequently, inserting a flexible loop into a protein that would not otherwise be there has a destabilizing effect. Further this destabilizing effect has a dependence on the size of the loop, with longer loops being more destabilizing than shorter loops. Consistent with this dependence, the loop regions within proteins belonging to extremophiles that have evolved to survive at high temperatures tend to be shorter than those in corresponding mesophile proteins because the shorter loops contribute to higher stability of the folded state at high temperatures.[69–71]

The insertion of novel peptide loops at internal positions within a target protein could also have an impact on stability by disrupting protein structure locally at the site of insertion or by interfering with protein folding. For example, inserting a loop into a protein in such a way so as to disrupt an alpha helix or a beta sheet would likely interfere with the intramolecular hydrogen bonding patterns associated with these secondary structural elements, which would destabilize the folded state of the protein. Interfering with the packing of hydrophobic residues at the loop insertion site might also be expected to be destabilizing to the folded state as

interactions that counteract the decrease in conformational entropy associated with folding would be compromised.

In addition to destabilizing the folded state of a protein, a loop insertion may make the folded state less accessible by interfering with folding either through the disruption of a transition state associated with folding or through the promotion of a miss-folded intermediate. Mutations have been identified in proteins that increase folding rates, by stabilizing a transition state that occurs during folding, but that nevertheless decrease the stability of the final folded state.[3,72] Loop insertions could in theory similarly impact the folding rate of a protein by interfering with a transition state, though unlike the point mutation example described above, a destabilizing interaction with a transition state might be more readily expected. By potentially impacting the folding pathway of a protein, an inserted peptide loop might not only interfere with transition states but might also result in new low-energy intermediates. Such intermediates might deplete the proportion of the protein in the proper folded state at any particular point in time or might promote aggregation with other miss-folded proteins before the proper folded state is accessed. Aggregation might also occur if a loop disrupts the structure of the protein at the site of insertion, exposing hydrophobic residues that are aggregation prone.

2.3.2 Protein Modeling with Rosetta

Because proteins fold spontaneously in particular solvent environments, protein folding could potentially be predicted from first principles. Computational approaches such as MD have made progress toward this goal by attributing force fields to various interactions, explicitly modeling solvent molecules, and tracking the development of chemical systems with time.

However, given the size of proteins and the timescales over which they fold, modeling protein folding with a first-principles based approach is tremendously challenging and requires an extraordinary amount of computational power.[73] This burden is increased when the goal of modeling is extended from protein structural prediction to computational protein design because multiple amino acid substitutions may be considered for many different residues in the design process.

The advantage of the Rosetta protein modeling software suite is that it utilizes heuristics to reduce the computational burden associated with first principle predictions.[74,75] With such a process, instead of modeling changes in the protein and solvent environment in time as the protein folds, different potential conformations of the protein are explored and ranked using an energy function. Challenges associated with this approach include effectively sampling the huge amount of conformational space accessible to a given protein sequence and appropriately scoring different conformations at different stages of the modeling process. Rosetta uses heuristics in addressing both of these challenges.

The Protein Data Base (PDB) contains thousands of solved protein structures, and it is the source of information that Rosetta utilizes in addressing problems of conformational space and structure scoring that are inherent to its modeling approach. The PDB is used by Rosetta to restrict the amount of conformational space that needs to be searched when trying to generate a structure from the primary sequence of a protein. This is done through fragments. Rosetta will match portions of amino acids (referred to as fragments) of only a few residues in length that occur in the primary sequence being modeled with identical sequences that occur within proteins in the PDB. Because of their short length, these fragments will occur many different times within

many different proteins in the PDB and, consequently, they will have many different conformations. In the first phase of modeling, Rosetta will substitute these fragment structures into the sequence of the protein being modeled. Doing so greatly reduces conformational space while providing realistic potential fragment conformations. It is therefore a useful heuristic for reducing the computational burden of modeling.

Rosetta modeling proceeds through two stages: an initial coarse-grained centroid stage, which utilizes fragment files from the PDB, and a subsequent full-atom modeling stage.[76,77] Each has its own energy functions. The first stage does not model each atom explicitly but rather treats each amino acid as a centroid, which is a sphere that has predefined attributes given the amino acid that it represents. These attributes, including torsion angles, charge, hydrophobicity, and minimum sterics requirements will be known to the centroid energy function when a particular conformation is scored. The second, full-atom modeling stage is more precise. This is the point at which conformational space associated with amino acid side chains is explored. In particular, each residue has a rotamer library containing the many different side chain conformations possible when rotation about covalent bonds is permitted. The energy function of this stage must capture more subtle atomic interactions relative to the centroid energy function given this greater level of detail.

In addition to playing a role in the generation and application of fragment files in the first, coarse grained stage of Rosetta modeling, the PDB is a major source of information used to rank centroid models within the context of the centroid energy function. Components of the centroid score function are based on the probability of a particular configuration within a model being observed within the PDB.[77] The env term in the centroid score function is a good example

of this. A particular centroid model will have a total score that results from the summation of various terms, one of which is env. The terms will capture different aspects of the model such as residue solvation as it relates to the hydrophobic effect, electrostatic pair interactions between residues, radius of gyration (which serves as a proxy for protein packing and therefore van der Waals interactions), phi psi angles, and steric repulsion. The env term relates to the first of these. For every residue, it counts the number of neighboring residues within a 10 Å cutoff and determines the probability of that residue having the amino acid type that it does given the observed number of neighboring residues. These probabilities are based on positional data from the accumulated structures in the PDB. If a particular model produces a solvent exposed hydrophobic residue, the env score term for that residue will be less negative (suggesting a less stable conformation) than it would be if that residue were buried because the probability of finding a hydrophobic residue within the PDB with a small number of neighboring residues (as would be the case when a residue is highly solvent exposed) is low.

2.3.3 Loop Modeling with Kinematic Closure

Loop modeling with kinematic closure is a particular application within Rosetta that allows short sequences to be modeled within the context of an existing protein structure. It proceeds through the same modeling stages with the same energy functions as described above, but because the sequences being modeled are kept relatively short, it does not need to utilize fragment files to decrease the amount of conformational space that it explores. Rather, different conformations are generated with a Monte Carlo algorithm.[28,78]

This application is useful for filling in gaps within protein structures. It is not always possible to resolve loop regions with X-ray crystallography even when these regions are relatively structured. For example, a particular loop may interfere with protein crystallization and need to be removed either genetically or via limit proteolysis in order to accommodate crystal formation. Alternatively, the loop may exist in several conformations or be too dynamic within the crystal and may not therefore be easily resolved. In such cases, Rosetta's loop modeling protocol can be used to model a loop when the structure cannot be solved from crystallography data.

This protocol may also be used to fill in gaps in protein structures that arise from homology modeling.[79] Although there are many thousands of protein structures in the PDB, there are still many proteins for which there is no structural data at all. This limitation can partially be overcome with homology modeling. In nature, there is a substantial amount of structural homology between proteins, which can exist even in situations where these is no clear sequence homology. When structural homology exists, the structure of one protein can be used to guide the modeling of another. With this approach, the sequence of one protein is threaded onto the structure of another, and that structure is used as a starting point for further modeling. Threading is facilitated with a sequence alignment, which matches similar residue types between sequences. However, whenever the sequences are not of identical length, which is frequently the case, gaps arise. This can lead to regions of sequence with no corresponding regions of structure after threading. These regions must therefore be modeled from scratch and the loop modeling application with kinematic closure can be used for this.

During loop modeling, many different conformations of the loop are sampled during the centroid modeling stage. At this point only residues within the loop, which are predefined as an input, are modeled. The single conformation with the best centroid energy score from this stage then proceeds to the full atom stage where the side chains are modeled. At this point, side chains of residues outside of the loop in the near-loop environment (within 10 Å) are also remodeled. The lowest energy, full-atom conformation from this second stage then becomes the final model output. In order to accurately predict the structure of a loop, this process must be repeated thousands of times, with the generation of thousands of output models. If a sufficient number of models are created, those with the lowest total energy scores will converge on a single, low energy conformation, which is assumed to be the actual structure of the loop within the protein in solution. This approach has been benchmarked on loops for which structural data exists and has been successful in reproducing those structures.[28]

Given the way in which this loop modeling protocol is designed, it can be used to insert any arbitrary peptide loop into any arbitrary protein target provided that the loop sequence is of a reasonable length. Doing so requires a sequence file that contains the loop sequence inserted at the desired position within the primary sequence of the target protein. When this augmented sequence is aligned to the unadulterated sequence of the target protein, the new sequence can be threaded onto the target protein, at which point the residues in the loop, which have no corresponding structure, can be modeled. Modeling will produce an energy score that could in theory be compared to scores obtained for loop insertions at different sites within the same protein target.

A number of assumptions inherent to the kinematic loop modeling application cease to be relevant when it is used to model a novel loop insertion within a given protein target. The loop modeling protocol is meant to model a loop *within the context* of an existing structure. This means that the structure is assumed to have a larger impact on the conformation of the loop than the loop is expected to have on the conformation of the protein. As such, most of the conformational space explored in this application pertains to residues within the loop sequence itself. It is only during the full-atom modeling stage that conformational space associated with residues in the near loop environment is explored. When these residues are considered, side chain rotamers are focused on as opposed to dramatic backbone rearrangements such as those that occur within the loop during the centroid modeling stage.[28,78]

As a result of focusing on residues that appear only within the loop, there are potential structural aspects of peptide loop insertion that Rosetta will not be able to capture. Any dramatic structural perturbations that occur at the site of an inserted loop will not be modeled since modeling is predominantly restricted to residues within the loop. Additionally, any impact that the loop has on folding rates or the formation of semi-stable miss-folded intermediates will not be captured since Rosetta does produce models that correspond to different conformations along a folding pathway.

In spite of the limitations described above, total scores obtained for models of loop insertions at particular sites within a protein and individual terms within those scores may provide useful information that can be used to infer the impact of a loop on the target protein in which it is inserted. For example, a particular insertion site may result in loop residues or target protein residues flanking the loop insertion site having unfavorable torsion angles. In such a

situation, an impact on structure at the loop insertion site might be expected for a protein in solution. However, a structural perturbation in the Rosetta model would not likely be observed because the Rosetta loop modeling application does not substantially remodel residues that do not appear in the loop. Instead, the rama term, which describes the probability (obtained from the PDB) of a residue having its observed set of torsion angles might be negatively impacted. A change in this score term could therefore be used to infer a negative impact on the protein that might result from loop insertion at this site.[77]

The expectation that components within the Rosetta score function will reflect the impact of a loop insertion at a particular site on the properties of the target protein in combination with the computational cost-cutting measures built into Rosetta via PDB based heuristics makes it an attractive approach for potentially designing loop insertion sites. These loops could potentially be of any particular sequence in the service of numerous applications. The ability to design proteins with internal loop insertions that are well accommodated would be particularly useful for designing peptide-directed conjugation sites for enzyme-mediated bioconjugation.

OBJECTIVE AND SPECIFIC AIMS

3.1 Objective

Protein engineering would benefit from methods that facilitate the use of proteins in unnatural environments as well as from increased predictive capability in protein design. The objective of this thesis is to contribute to both of these challenges by developing a predictive computational approach for identifying accommodating peptide-loop insertion sites in proteins. These peptide loops, once inserted, can serve to direct enzyme-mediated bioconjugation reactions, which can in turn be used to enable the application of proteins in unnatural environments via immobilization, polymer conjugation, and other techniques.

Three specific aims were addressed in pursuing this objective. In the first, we *1) explored the utility of non-specific, multi-point, bioconjugation for the design of enzymatic materials that were resistant to bacterial biofilm formation*. Our system highlighted the potential of using bioconjugation to enable enzyme function in an unnatural environment, namely within a polymer matrix. It also allowed us to pursue an unconventional solution to the problem of biofilm formation, as biofilm disruption via quorum quenching would have been difficult to achieve with a protein-free strategy. Although ultimately effective, this conjugation approach required the brute-force sampling of numerous enzymes in numerous pre-polymer combinations before a suitable formulation was identified, likely due in part to the non-specific nature of the conjugation chemistry employed.

In the second specific aim, we *2) extended the applicability of an enzyme-mediated, peptide-directed conjugation system to facilitate conjugation at internal positions within the*

structure of a protein. The lipoic acid ligase/LAP system used in this aim is capable of conjugating an azide-bearing lipoic acid derivative to the LAP peptide sequence and can therefore mediate protein immobilization and diverse modification reactions via azide/alkyne click chemistry. By demonstrating successful ligation and subsequent modification of the LAP sequence when inserted at internal positions within GFP, we showed that this enzyme-mediated conjugation approach, which allows for precise regional and temporal control over conjugation, could be used not just as a terminal tag but also for site-specific conjugation events at diverse locations within a protein.

Finally, having explored the utility of bioconjugation by specifically applying it to address biofilm formation and having demonstrated the feasibility of using a highly controllable enzyme-mediated conjugation system for modifying proteins at internal sites within their structures, we *3) developed a computational approach for predicting accommodating peptide loop insertion sites within proteins using the LAP sequence as a model loop*. This approach allowed sites with a higher probability of accommodating loop insertions to be identified, reducing the number of actual proteins that needed to be made and experimentally characterized when searching for a successful construct with an internal peptide-directed conjugation site. It also demonstrated that loop accommodation at a particular site could be modulated via mutagenesis of residues in the near loop environment. These factors in combination with the highly accessible nature of the computational approach, which can be performed on a laptop, increase the general applicability of enzyme-mediated bioconjugation.

These three specific aims, introduced in more detail below, are addressed in chapters 4-6 of this manuscript.

3.2 Specific Aim 1: Explore the utility of non-specific, multi-point bioconjugation for designing enzymatic materials that are resistant to bacterial biofilm formation

Due to the prevalence of biofilm-related infections, which are mediated by bacterial quorum sensing, there is a critical need for materials and coatings that resist biofilm formation. In this aim, we developed novel anti-biofilm coatings that disrupt quorum sensing in surface-associated bacteria via the immobilization of acylase in polyurethane films. Specifically, acylase from *Aspergillus melleus* was covalently immobilized in biomedical grade polyurethane coatings via multipoint covalent immobilization. Coatings containing acylase were enzymatically active and catalyzed the hydrolysis of the quorum sensing (QS) molecules *N*-butyryl-L-homoserine lactone (C4-LHL), *N*-hexanoyl-L-homoserine lactone (C6-LHL) and *N*-(3-oxododecanoyl)-L-homoserine (3-oxo-C12-LHL) lactone. In biofilm inhibition assays, immobilization of acylase led to an approximately 60% reduction in biofilm formation by *Pseudomonas aeruginosa* ATCC 10145 and PAO1. Inhibition of biofilm formation was consistent with a reduction in the secretion of pyocyanin, indicating the disruption of quorum sensing as the mechanism of coating activity. Scanning electron microscopy (SEM) further showed that acylase-containing coatings contained far fewer bacterial cells than control coatings that lacked acylase. Moreover, acylase-containing coatings retained 90% activity when stored dry at 37˚C for 7 days and were more stable than the free enzyme in physiological conditions, including artificial urine. Ultimately, such coatings hold considerable promise for the clinical management of catheter-related infections as well as the prevention of infections in orthopedic applications (i.e., on hip and knee prostheses) and on contact lenses.

3.3 Specific Aim 2: Extend the applicability of an enzyme-mediated, peptide-directed conjugation system to facilitate conjugation at internal positions within the structure of a protein

Approaches that allow bioorthogonal and, in turn, site-specific, chemical modification of proteins present considerable opportunities for modulating protein activity and stability. However, the development of such approaches that enable site-selective modification of proteins at multiple positions, including internal sites within a protein, and that exhibit high levels of temporal control over conjugation has remained elusive. To overcome this void, in this aim, we developed an enzymatic approach for multi-site clickable modification based on the incorporation of azide moieties in proteins using lipoic acid ligase (LplA). The ligation of azide moieties to the model protein green fluorescent protein (GFP) at the N-terminus and two internal sites using lipoic acid ligase was shown to proceed efficiently with near complete conversion. Modification of the ligated azide groups with poly(ethylene glycol) (PEG), α-D-mannopyranoside, and palmitic acid resulted in highly homogeneous populations of protein-polymer, protein-sugar, and protein-fatty acid conjugates. The homogeneity of the conjugates was confirmed by mass spectrometry (MALDI-TOF) and SDS-PAGE electrophoresis. In the case of PEG attachment, which involved the use of strain-promoted azide-alkyne click chemistry, the conjugation reaction resulted in highly homogeneous PEG-GFP conjugates in less than 30 mins. As further demonstration of the utility of this approach, ligated GFP was also covalently immobilized on alkyne-terminated self-assembled monolayers. These results underscore the

potential of this approach for, among other applications, site-specific multipoint protein PEGylation, glycosylation, fatty acid modification, and protein immobilization.

3.4 Specific Aim 3: Develop a computational approach for predicting accommodating peptide loop insertion sites within protein structures using the LAP sequence as a model loop

Inserting novel peptide loops at internal positions within target proteins enables the design of chimeric molecules that combine the properties of the target (i.e. catalytic activity) with those of the loop (i.e. recognition by post-translational modification-performing enzymes, antibody-like binding properties, etc.). The design and utilization of these chimeras is impaired by the negative impact that large peptide loop insertions commonly have on aspects, such as stability, of the target protein. In this aim, we addressed these shortcomings by developing a Rosetta-based computational approach for scanning protein structures to identify permissive loop insertion sites. This approach utilizes the centroid modeling stage of the kinematic loop modeling application, and ranks insertion sites by averaging the total centroid energy scores of 10 models produced per site. It is highly user accessible, allowing for every possible insertion site within a protein to be scored with a single laptop computer in a number of days. With this scoring approach, we observed a correlation between soluble protein expression and Rosetta scores for two protein libraries (using lipase A from *Bacillus subtilis* and β-glucosidase from *Trichoderma Reesei*) that contained constructs with 13-residue loop insertions at diverse surface exposed sites. The specific loop sequence, known as the LAP sequence, is recognized and targeted by lipoic acid ligase and can therefore potentially be used to site-specifically direct fluorophore addition as

well as the addition of bioorthogonal, click-reactive handles to facilitate diverse bioconjugation reactions. Within these libraries, Rosetta scores successfully identified permissive sites that were unintuitive with respect to local secondary structure and alpha carbon *B*-factors. By considering the impact of each score term on the correlation between Rosetta total scores and soluble protein expression, the local residue environment was identified as the major structural determinant of loop accommodation that is captured by this approach. *In silico* mutagenesis of residues in the near loop environment demonstrated that Rosetta scores could be modulated for a particular loop insertion site, and experimental characterization of these mutants revealed that changes in score upon mutagenesis correlated with the soluble protein expression of the mutants. This suggests that the permissiveness of a particular loop insertion site can be computationally engineered by designing stabilizing mutations in the near loop environment and highlights the under appreciated role that residues in the near loop environment play in determining loop accommodation. Finally, we use this approach to guide the identification of internal LAP insertions sites within the protein, PTEN, in order to reduce the number of constructs that needed to be built and experimentally characterized when searching for a permissive LAP insertion site. Five constructs with promising Rosetta scores that contained LAP insertions in diverse regions within the protein were expressed in *E. coli* and found to behave similarly to WT PTEN in terms of soluble expression, suggesting that this protein can in theory be labeled at internal sites via conjugation to the LAP sequence without compromising its physiologically active termini.

ACYLASE-CONTAINING POLYURETHANE COATINGS WITH ANTI-BIOFILM ACTIVITY

Adapted from Biotechnology and Bioengineering, 113, 12, 2016

4.1 Introduction

The propensity of bacteria to colonize on the surface of implantable medical devices and materials, including urological catheters, represents a significant clinical challenge. Bacterial colonization, which leads to the formation of stable biofilms, is the result of the accumulation and subsequent coalescence of bacterial cells on the device or material surface. In the case of urological catheters, biofilm formation may occur when a leakage or break in the catheter drainage system occurs through which bacteria may enter the lumen of the catheter.[80] Alternatively, bacteria may also be introduced during catheterization and via transfer from surrounding tissue and skin.[81] Upon assembly into biofilms at sufficiently high cell densities, the microbial consortia often lead to persistent infections, which are difficult to treat. Biofilm-associated infections, which increase in likelihood with catheterization time, are particularly difficult to eradicate due to the high antibiotic resistance of cells in biofilms, which can withstand antibiotic concentrations that are 1000-times greater than free-floating cells.[12,82,83] Despite considerable efforts to develop anti-biofilm materials and coatings, little progress has been made in inhibiting biofilm-associated infections in catheters. Efforts to develop such materials and coatings are ultimately spurred by the staggering prevalence of biofilm-associated infections in catheters and the accompanying burden on healthcare systems worldwide.[84,85]

Conventional approaches to mediate biofilm formation on catheter surfaces rely on the antimicrobial activity of known biocidal materials or molecules. Such materials or molecules may be coated onto catheter surfaces or incorporated into matrices from which they are released in a continuous manner into the near-surface environment. Specifically, the release of silver[86,87] and various antibiotics, such as triclosan,[88] nitrofurazone,[86,87,89] and the combination of minocycline and rifampicin,[90] from catheter surfaces has shown some promise. However, such approaches generally have limited utility due to restrictions associated with the amount and duration over which the biocidal agent can be released. The utility of these approaches is also limited due to the apparent lack of correlation between short-term biocidal activity and long-term biofilm inhibition. Notably, materials with high antimicrobial activity, in most cases, have little or no inhibitory effect on the formation of biofilms over long time periods.[85] Additionally, although also of limited success, the inhibition of bacterial adhesion via modification of material properties (*i.e.,* surface charge and hydrophilicity) has also been investigated as a means of inhibiting biofilm growth. Given these limitations, more effective strategies to mediate biofilm formation are crucial to reduce patient suffering and healthcare costs associated with catheter-related infections.

A novel approach to inhibit biofilm formation on surfaces, which has recently received attention, entails the disruption of signaling pathways involved in microbial community. This approach is specifically based on the quenching of signaling molecules that are secreted by bacteria through which bacteria coordinate the production of extracellular biofilm components. In nature, these signals are degraded by various hydrolytic enzymes, including lactonases and acylases, as a means of controlling quorum sensing (QS) in bacteria.[91–98] In support of this

approach, recent studies have demonstrated that small molecule inhibitors of quorum sensing pathways as well as the addition or overexpression of quorum quenching enzymes negatively impact biofilm formation.[13,99–102] Additionally, Ivanova and co-workers also immobilized acylase on silicone urinary catheters via layer-by-layer assembly, which resulted in the release of the enzyme over time.[103] Release of the enzyme was shown to degrade quorum signals while reducing biofilm formation *in vitro* and *in vivo*, which was enhanced by the co-release of amylase.[104] While these results support the feasibility of the proposed approach, the release of acylase into the bloodstream may interfere with the activity of endogenous human acylase (*e.g.,* aminoacylase-1), thereby disrupting the balance between the catabolism and anabolism of acylated amino acids.[105] Moreover, upon release, the enzyme is susceptible to proteolytic degradation, which may diminish its activity and thus impact on biofilm formation.[106]

In this work, the multipoint covalent immobilization of acylase into polyurethane films to create non-leaching biocatalytic coatings that resist biofilm formation was investigated. Specifically, acylase from *Aspergillus melleus* was reacted with the medical grade polyurethane prepolymers BAYMEDIX FD103 and BAYMEDIX FP 520. The activity and stability of the immobilized enzyme was characterized and, furthermore, the inhibition of biofilm formation by *Pseudomonas aeruginosa*, which is a leading cause of hospital infections and present in virtually all biofilms that form in catheters in clinical settings,[81,82] was characterized. For biofilm inhibition assays, static biofilm formation was measured using a conventional colony forming assay with two strains of *P. aeruginosa* (ATCC 10145 and PAO1). Biofilm inhibition was also characterized via imaging of the surface of acylase-containing and control coatings without acylase by scanning electron microscopy. To confirm that the biofilm inhibition by the coatings

was due to disruption of the quorum sensing system, levels of the quorum sensing-dependent secondary metabolite pyocyanin, were measured. We hypothesized that *N*-acylhomoserine lactone signals secreted by *P. aeruginosa* in the early biofilm would be degraded on contact with the coating, thus disrupting further surface-associated biofilm assembly (**Figure 4.1A**). These results further the development and understanding of this approach for biofilm prevention by demonstrating the utility of such coatings for the prevention of catheter-related infections (*i.e.,* in urological and vascular catheters) as well as infections on other implantable materials and medical devices (*i.e.,* orthopedic prostheses, contact lenses).

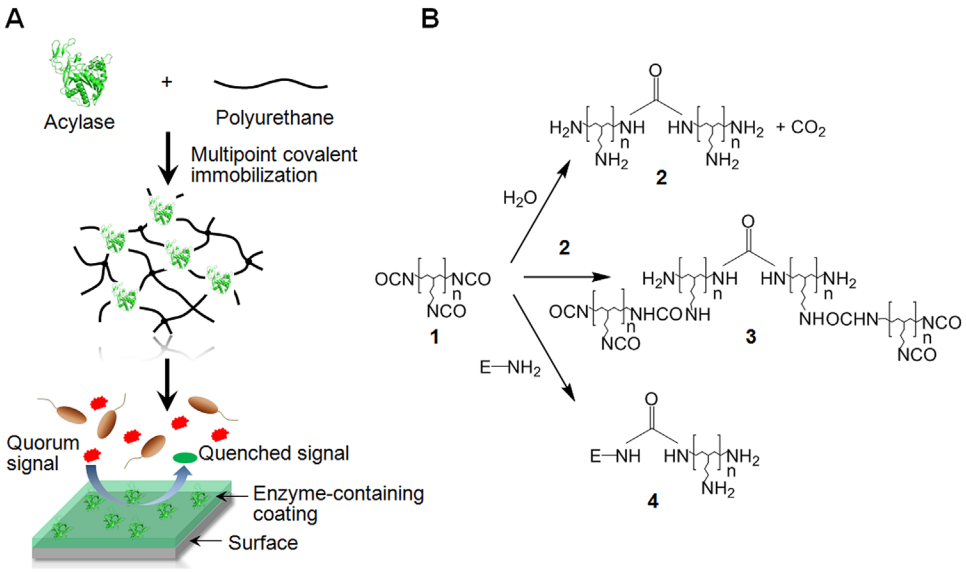

Figure 4.1. Multipoint covalent immobilization of acylase in polyurethane coatings.
(A) Schematic of the preparation of acylase-containing coatings, which disrupt quorum sensing of surface-associated bacteria. Quorum signals secreted in the early biofilm are degraded on contact with the coating surface, thereby quenching the quorum signals responsible for the progression of biofilm formation. (B) Reaction scheme for the irreversible incorporation of acylase into two-component waterborne polyurethane matrices. Isocyanate groups may, in addition to reacting with hydroxyls, react with the primary amines in the enzyme (E), resulting in the covalent crosslinking of the enzyme in the polymer network.

4.2 Materials and Methods

4.2.1 Materials

Acylase I from *A. melleus* (Sigma-Aldrich, MO, USA) with a protein content of 4.5% (w/w) and specific activity of 0.05 U/mg was used as a quorum-quenching enzyme. The non-QS substrate, *N*-acetyl-L-methionine, was purchased from Alfa Aesar (MA, USA). The QS substrates (acyl-homoserine lactones) were purchased from Cayman Chemical Company (MI, USA) and used to model QS signaling compounds produced by Gram-negative bacteria. *P. aeruginosa* ATCC 10145 and PAO1, bacterial strains proficient in biofilm formation, were obtained from ATCC (VA, USA). The medical grade polyurethane monomers (BAYMEDIX FD103 and BAYMEDIX FP 520) were provided by Bayer Material Science (PA, USA).

4.2.2 Preparation of Acylase-Containing Polymeric Films

The polyurethane prepolymers were mixed in 9:1 ratio (BAYMEDIX FD103:BAYMEDIX FP 520) by weight and subsequently 24 mg of acylase (8 mg/mL in sodium phosphate buffer, 50 mM and pH 7.5) was added to the mixture. Prior to use, the acylase, which was supplied as a crude mixture, was purified by size exclusion chromatography (ENrich SEC 70, 10 x 300 column, Bio-Rad) and fractions with enzymatic activity were pooled and used for the preparation of coatings. The resultant mixture was vigorously blended using a custom designed mixing head attached to a 2500 rpm hand held drill for 1 min and put under vacuum for 5 min to remove released carbon dioxide.[107] The final solution was poured on a plexiglass (polymethyl methacrylate) sheet, spread (using JR ROD 12” OA 3/8 DIA #60 wire, Paul N

Gardener Company, Inc., FL, USA) to a thin film, and cured for 5 days. After curing, the film was washed in sodium phosphate buffer (50 mM, pH 7.5) overnight to quench the excess crosslinking. The control film without enzyme was also prepared.

4.2.3 Acylase Activity Assay

The activity of free acylase and acylase-containing coatings was determined via hydrolysis of *N*-acetyl-L-methionine and *N*-butyryl-L-homoserine lactone (C4-LHL). The amount of primary amines released during hydrolysis was measured using a tri-nitrobenzene sulfonic acid (TNBS) assay.[108] Briefly, 20 mM of substrate was incubated with either the free enzyme or enzyme-containing film in 1 mL of sodium phosphate buffer (50 mM, pH 7.5) at 37˚C with constant shaking (100 rpm). Aliquots (50 μL) were withdrawn periodically and transferred to 450 μL of quenching solution (100 mM sodium bicarbonate in 1:5 ethanol:water). The resulting mixture was kept at 90˚C for 10 min to quench the enzymatic reaction. Subsequently, 250 μL of TNBS reagent (0.01% in sodium bicarbonate solution) was added after which the mixture was further incubated at 37˚C and 100 rpm for 2 h. The primary amines released by enzymatic hydrolysis of the substrate react with TNBS in 1:1 ratio. After 2 h of incubation, the TNBS reaction was quenched by adding 100 μL of 1 M hydrochloric acid and 250 μL of 10% sodium dodecyl sulfate. The absorbance of the resultant solution was monitored at 335 nm to determine the enzymatic activity. The kinetic parameters of free and immobilized enzyme were also determined against *N*-acetyl-L-methionine and *N*-butyryl-L-homoserine lactone using the same assay.

To compare the activity of free and immobilized acylase against the QS molecules C4-LHL, *N*-hexanoyl-L-homoserine lactone (C6-LHL) and *N*-(3-oxododecanoyl)-L-homoserine (3-oxo-C12-LHL) lactone, enzyme activity was measured using a fluorometric assay, which, like the TNBS assay, quantifies the release of primary amines. Notably, due to the low solubility of C6-LHL and 3-oxo-C12-LHL, the TNBS assay was not sufficiently sensitive to use for these substrates. Briefly, for the fluorometric assay, 30 μg/mL of each substrate compound was incubated with 0.5 μg/mL of the free enzyme or enzyme-containing coating in 1 mL of sodium phosphate buffer (50 mM, pH 7.5) for 1 h at 37˚C and 100 rpm. The resulting primary amines released by enzymatic hydrolysis were analyzed by a fluorescamine fluorimetric assay. In the fluorimetric assay, 150 μL of the reaction mixture was added to 50 μL of a fluorescamine (Sigma-Aldrich, MO, USA) stock solution (3 mg/mL in dimethyl sulfoxide (Sigma-Aldrich, MO, USA)) followed by incubation at room temperature for 3 min. After incubation, the fluorescence of the assay solution was measured using an excitation and emission wavelength of 390 nm and 470 nm, respectively. The amount of homoserine lactone released in the enzymatic reaction was calculated from a calibration curve using α-amino-γ-butyrolactone hydrobromide (Sigma-Aldrich, MO, USA) as a standard.[103] Moreover, to determine activity retention, the activity of enzyme in the coating was compared to an equivalent amount of free enzyme. An implicit assumption in determining activity retention in this way was that the immobilized enzyme was homogeneously distributed throughout the coating.

4.2.4 Biofilm Inhibition Assay

Antibiofilm activity of the enzyme-containing films in static conditions was assessed against *P. aeruginosa* ATCC 10145 and PAO1 strains using the cell count method. *P. aeruginosa* ATCC 10145 and PAO1 inoculums were prepared from overnight cultures in LB (Bacto, Dickinson and Company, NJ, USA). The control and enzyme-containing films were cut into small pieces (0.5 x 0.5 inches) and placed in a 24-well cell culture plate. Thereafter, 1 mL of bacteria, which was grown overnight and diluted to 10^4 CFU/mL, was inoculated in each well, and the plate was incubated for 24 h at 37˚C. The liquid medium was removed and the biofilms were washed very gently with PBS three times to eliminate non-adhered bacteria. The bacteria embedded in biofilms were suspended in 1 mL of PBS using periodic sonication for 30 s and further diluted in PBS. The resultant bacterial suspensions were plated on LB-agar plates and incubated at 37˚C for 24 h. The corresponding colonies were counted and compared with the colony count on the control coating, which lacked the enzyme. For biofilm inhibition assays, bacterial cells were aliquoted from a single culture for all replicates for both strains of *Pseudomonas*.

4.2.5 Pyocyanin Quantification

The control and enzyme-containing films (0.5 x 0.5 inches) were incubated with 1 mL of 10^4 CFU/mL of *P. aeruginosa* ATCC 10145 and PAO1 strains in LB for 24 h at 37˚C. Pyocyanin was extracted from 1 ml of culture supernatant with 0.5 mL of chloroform. The samples were centrifuged at 4500 rpm for 5 min to separate the organic phase from the aqueous phase. The aqueous phase was discarded and the absorbance of organic phase was measured at 695 nm (ε =

5816 $M^{-1}cm^{-1}$) to quantify pyocyanin levels.[109–112] The *E. coli* BL21 DE3 strain, which does not secrete pyocyanin, was used as a negative control.

4.2.6 Storage Stability

The enzyme-containing coatings (0.5 x 0.5 in) were stored for 7 days at three different temperatures viz. 4˚C, room temperature, and 37˚C. The activity of these coatings was measured periodically against *N*-acetyl-L-methionine using the TNBS assay. The storage stability of free enzyme and the coatings was also determined in physiological conditions viz. in PBS and artificial urine (Wards Science, NY, USA) at 37˚C. The coatings were washed with sodium phosphate buffer (50 mM, pH 7.5) three times before measuring the activity.

4.2.7 Scanning Electron Microscopy (SEM)

Enzyme-containing and control coatings lacking enzyme were incubated with 1 mL of 10^4 CFU/mL of *P. aeruginosa* ATCC 10145 for 24 h at 37˚C to permit biofilm formation under static conditions. After biofilm formation, the coatings were washed with sterilized PBS to remove non-adhering bacteria. The resultant biofilm specimens were fixed with the mixture of 2.5% glutaraldehyde and 2.5% formaldehyde in PBS overnight at 4˚C. The fixed biofilms on polymeric films were then washed 5 times with 1 mL of sterilized water to remove residual glutaraldehyde and formaldehyde as well as buffer salts. Prior to analysis, the specimens were dehydrated using 10, 30, 50, 70 90 and 100% (v/v) of graded ethanol and dried at 37˚C for 6 h. The dried samples were coated with platinum and analyzed using SEM (Hitachi SU3500,

Germany) using the secondary electron emission mode at 1kV. Images were collected at 1000x and 5000x magnifications.

4.3 Results and Discussion

The development of coatings that inhibit biofilm formation via the disruption of quorum sensing represents a new paradigm in the pursuit of materials that prevent catheter-related infections. As evidence of this approach, prior studies have highlighted the link between quorum sensing and biofilm formation and the potential of targeting quorum sensing for biofilm prevention.[13,100–104] Of direct relevance to this work, these studies have included the non-covalent incorporation of acylase in coatings, which inhibited biofilm formation on the coating surface.[103,104] While these studies support the feasibility of this approach, alternative strategies to immobilize quorum quenching enzymes in coatings for medical devices (*i.e.,* that do not result in enzyme leaching) are needed.

4.3.1 Preparation and Activity of Acylase-Containing Coatings

Biocatalytic coatings were prepared via dispersion of acylase in two-component waterborne polyurethane coatings (**Figure 4.1B**). The polymerization of two-component waterborne polyurethane coatings, which have previously been used as matrices for enzymes,[107] is achieved by the reaction of water-dispersible polyisocyanate and polyol prepolymers. Dispersion of the enzyme in the aqueous polymerization reaction facilitates covalent coupling of the enzyme to the polymer network via functional groups on the enzyme surface. Specifically, primary amines that are on the enzyme's surface react with free isocyanate groups, resulting in

multipoint covalent immobilization, which ensures enzyme retention in the coating. In addition to ensuring retention, the formation of linkages between enzyme and polymer restrict the mobility of the enzyme, thereby suppressing it's unfolding due to environmental pressures. Accordingly, multipoint immobilization can lead to significant enhancements in enzyme stability, thus potentially increasing the lifetime of acylase relative to tethering acylase to the coating or catheter surface. Polyurethanes, in particular, are attractive supports due to their rapid and simple preparation and tunable properties that, when used as scaffolds, yield stable bioplastics.[107,113–118] Additionally, as polymers and coatings for implantable materials and medical devices, including catheters and metal stents, polyurethanes are widely used in tissue engineering and regenerative medicine.[119–123] Bakker et al., 2000 have previously shown that acylase could be immobilized in polyurethane foams while retaining activity, although such foams are impractical for use as coatings on catheter surfaces.

For the preparation of acylase-containing coatings, the medical grade polyurethane prepolymers (BAYMEDIX FD103 and BAYMEDIX FP 520) were used. Acylase, which was purified by size exclusion chromatography, was added to the monomers at a 9:1 BAYMEDIX FD103-to- BAYMEDIX FP 520 ratio by wet weight. The extent of irreversible immobilization of acylase upon curing was determined by measuring protein leaching upon extensive washing of the resulting coating. To measure protein leaching, the rinsate was analyzed for enzymatic activity against *N*-acetyl-L-methionine as the substrate at 37˚C. The activity in the wash solution was negligible, indicating near 100% irreversible immobilization where all of the enzyme was retained in the films.

Following washing, the enzymatic activity of the acylase-containing coatings towards the model substrates *N*-acetyl-L-methionine and C4-LHL was assayed. The activity of the coating was monitored spectrophotometrically via titrating the liberation of primary amines with TNBS. While *N*-acetyl-L-methionine is not a quorum sensing substrate, *N*-butyryl-L-homoserine lactone is used by bacteria for QS.[124] Although the activity retention of the immobilized enzyme was low (5% as measured by the more sensitive fluorescence assay), the retained activity was comparable to that for other enzymes that were similarly immobilized via multipoint covalent attachment in waterborne polyurethane coatings.[107] Additionally, the Michaelis-Menten parameters (K_m and V_{max}) for the immobilized and free acylase are shown in **Table 4.1**. Interestingly, the apparent K_m for *N*-acetyl-L-methionine was nearly an order of magnitude lower for immobilized acylase relative to free acylase. The improvement in affinity may be due to partitioning of the substrate into the coating as a result of the substrate and coating being hydrophilic. Such an increase in substrate partitioning into the coating would, in turn, increase the local concentration of substrate in the vicinity of the enzyme, thereby decreasing the apparent K_m.

	K_m (mM)		V_{max} (ΔA/mg-min)	
Substrate	Free acylase	Coatings	Free acylase	Coatings
N-acetyl-L-methionine	29 ± 12	4.9 ± 1.3	8.13 ± 0.16	0.14 ± 0.0
N-butyryl-L-homoserine lactone	4.1 ± 0.0	4.4 ± 0.1	2.96 ± 0.0	0.04 ± 0.0

Table 4.1. Michaelis-Menten parameters for the reaction of free acylase in solution and acylase-containing coatings with the substrates *N*-acetyl-L-methionine and *N*-butyryl-L-homoserine lactone.

4.3.2 Degradation of QS Molecules

To further investigate the activity of the coatings, the hydrolysis of the native QS molecules C4-LHL, C6-LHL, and 3-oxo-C12-LHL by the coatings was compared. For comparison of activity towards the QS molecules, a fluorescent assay based on the formation of a fluorescent complex between the homoserine lactone hydrolysis product and fluorescamine was used. Results of the activity assay using C4-LHL, C6-LHL, and 3-oxo-C12-LHL as substrates by free acylase and acylase-containing coatings are shown in **Figure 4.2**. While active against all of the QS molecules, the activity of both the free enzyme and coatings increased with the length of the carbon chain of the substrate molecule. Specifically, the activity of the free enzyme and coatings was greatest with 3-oxo-C12-LHL, suggesting the preference of the enzyme for more hydrophobic substrates. Similar results were reported by Sio and co-workers,[125] albeit with a different acylase, who found that acylase from *P.aeruginosa* PAO1 was inactive against C4-LHL and C6-LHL, but highly active against 3-oxo-C12-LHL. Notably, C4-LHL and 3-oxo-C12-LHL are the primary QS molecules responsible for the formation of biofilms by *P. aeruginosa*,[124–126] making the activity of the coatings towards these substrates significant.

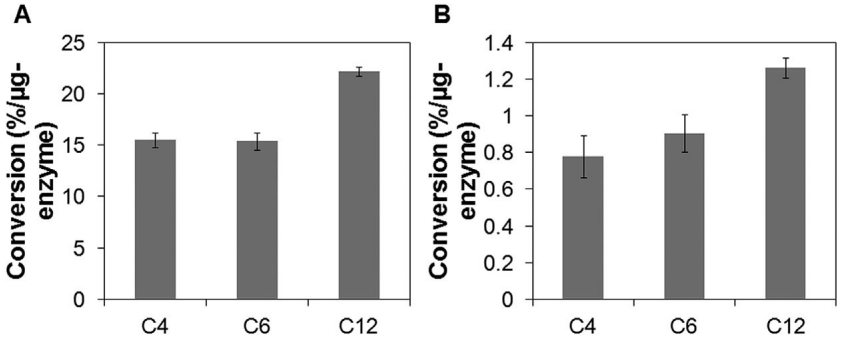

Figure 4.2. Hydrolysis of the QS molecules C4-LHL (C4), C6-LHL (C6), and 3-oxo-C12-LHL (C12) by (A) free acylase and (B) acylase-containing coatings.
Acylase activity was measured in 50 mM sodium phosphate buffer (pH 7.5) at 37 °C in the presence of 0.5 mg/mL of the free enzyme. Error bars represent the standard deviation from the mean for three independent measurements with separate samples.

4.3.3 Biofilm Inhibition

The anti-biofilm activity of acylase-containing coatings was determined under static conditions using *P. aeruginosa* ATCC 10145 and PAO1. To determine biofilm inhibition, the colony forming units per volume (*i.e.,* CFU/mL) of cells embedded within the resulting biofilm after incubation for 24 h at 37˚C was measured. For the biofilm inhibition assays, acylase-containing coatings and control coatings (without acylase) were challenged with 10^4 CFU/mL, which was diluted from an overnight culture. Prior to measuring CFU/mL, the residual biofilm on the coatings was washed gently with sterilized PBS to remove any non-adhered bacteria, which were not part of the biofilm. The coatings used for measuring anti-biofilm activity consisted of the same composition, including enzyme concentration, used to measure acylase activity in enzyme activity assays.

The results of the biofilm inhibition assay found that, for both strains, an approximate 60% reduction in CFU/mL was observed for the acylase-containing coatings relative to control coatings, which did not contain acylase (**Figure 4.3**). Specifically, immobilization of acylase led to a 60% and 58% reduction in biofilm formation by *P. aeruginosa* strains ATCC 10145 and PAO1, respectively. These results clearly illustrate the anti-biofilm properties of the acylase-containing coatings, which presumably is the result of degrading QS molecules in the near-surface environment. Moreover, our results suggest that the enzyme does not need to be released from the coating surface to inhibit biofilm formation. While not investigated here, the anti-biofilm activity of the coatings may be improved by altering the distribution of acylase in the coating as well as increasing coating porosity. For example, directing enzyme localization to the coating surface during curing of the coating may increase the accessibility of the enzyme,

thereby reducing potential diffusional limitations of QS molecules in the coating. One way to direct the localization of acylase to the coating surface during curing is via the attachment of hydrophobic modifiers as demonstrated previously.[127] A similar effect would presumably result from increasing the coating porosity, which may be enhanced by extracting water-soluble poragens (*i.e.,* polyethylene glycol dinaphthylacetate) from the coating. In this case, the coatings may be prepared with poragens that induce the formation of pores, thereby effectively exposing enzyme beneath the coating surface. Additionally, in *Pseudomonas*, there are four quorum signaling systems, although only two (*i.e., las* and *rhl*) are responsive to *N*-acylhomoserine lactones and thus targeted by acylase. The anti-biofilm activity of the coatings may thus potentially be further improved by the addition of inhibitors of the other two systems (*i.e.,* PQS and IQS).

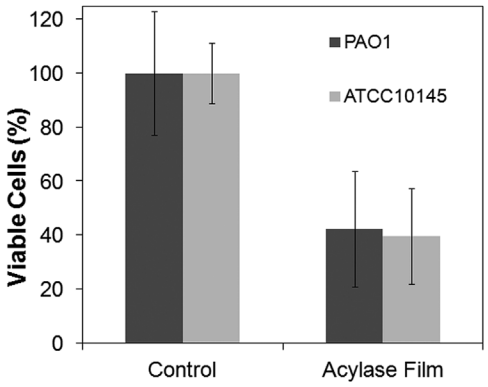

Figure 4.3. Anti-biofilm activity of acylase-containing coatings under static culture conditions using *P. aeruginosa* PAO1 and ATCC 10145.

Coatings were incubated with 10^4 CFU/mL bacteria for 24 h at 37 °C in LB media. The relative amount of residual viable cells, which was used as a quantitative measure of biofilm formation, was determined by measuring colony forming units per mL (CFU/mL). Error bars represent the standard deviation from the mean for 6 independent measurements with separate samples.

The inhibition of biofilm formation on acyalse-containing coating against *P. aeruginosa* strain ATCC 10145 was also characterized by SEM (**Figure 4.4**). On control coatings that lacked acylase, regions consisting of a dense network of bacteria representative of a biofilm were clearly visible. These areas are indicated by the arrows in **Figure 4.4A**. Conversely, in the case of acylase-containing coatings, no such regions were observed, confirming the inhibition of biofilm formation when the enzyme is immobilized (**Figure 4.4B**). Additionally, the surface of coatings containing acylase had qualitatively significantly fewer adhered bacterial cells relative to the control coatings. Furthermore, the morphology of the adhered cells in **Figure 4.4C** and **D** were similar to that observed previously in biofilms formed by *P. aeruginosa*.[128] Although not shown, live/dead staining was used to determine if the residual cells on the coating surface were viable. As expected, given the lack of anti-bacterial activity of acylase, the residual cells on acylase-containing surfaces were, similar to on the control coatings, predominately alive. As such, the decrease in cell number on the coatings with acylase was the result of a decrease in the adherence of cells, which is reflective of biofilm inhibition, rather than cell death. By inhibiting biofilm formation, the residual cells may be eradicated by other mechanisms, including antibiotics or attack by the immune system.

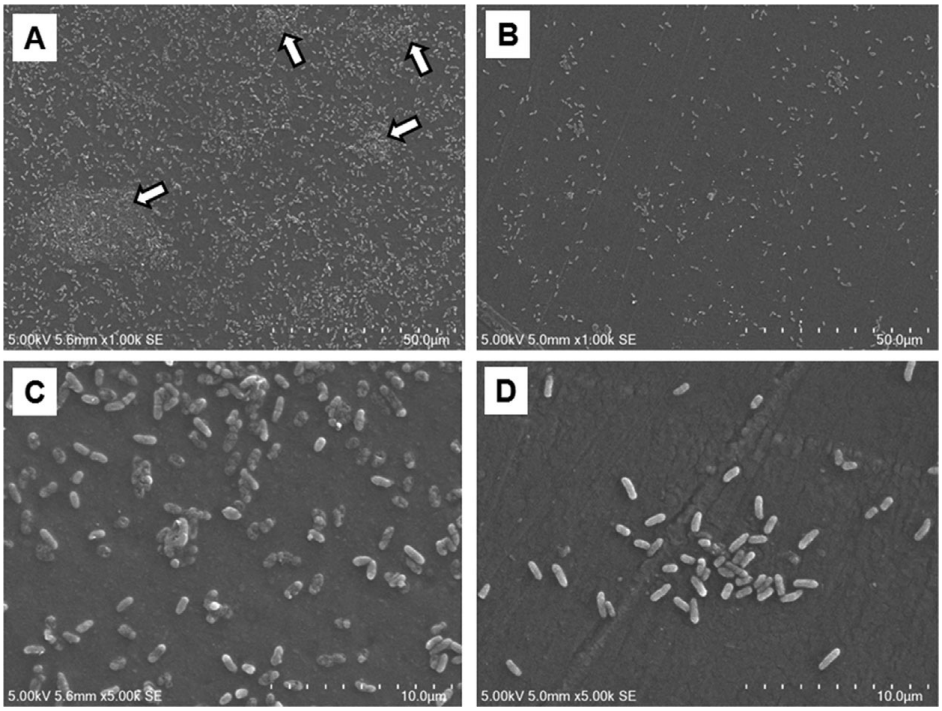

Figure 4.4. SEM images of biofilm formation by P. aeruginosa ATCC 10145 on (A and C) control coatings without acylase and (B and D) acylase-containing coatings.
Images were collected at 1000 (A and B) and 5000 (C and D).

4.3.4 Characterization of QS

The production of pyocyanin was quantified by UV/vis in order to demonstrate that the anti-biofilm activity of acylase-containing coatings was consistent with a reduction in quorum sensing activity. It has previously been demonstrated that the secondary metabolite pyocyanin, a blue pigment, is secreted by *P. aeruginosa* upon induction of quorum sensing.[129–131] In nature, pyocyanin secretion is regulated by C4-LHL,[103,104] which was previously shown to be degraded by the acylase-containing coatings. Given the hydrolytic activity of the acylase-containing coatings in degrading C4-LHL, we expected that pyocyanin secretion by *P. aeruginosa* ATCC 10145 and PAO1 would decrease in the presence of acylase-containing coatings relative to

control coatings without acylase. As expected, a significant decrease in pyocyanin secretion by *P. aeruginosa* ATCC 10145 and PAO1 upon incubation with acylase-containing coatings was observed (**Figure 4.5**).

Specifically, the secretion of pyocyanin by both strains of *P. aeruginosa* was reduced by approximately 60%, which is consistent with the extent of biofilm inhibition by the coatings. A similar reduction in pyocyanin production by *P. aeruginosa* PAO1 was observed on nanoalumina-functionalized membranes containing the lactonase SsoPox from *Sulfolobus solfataricus*, which was immobilized non-covalently via electrostatic interaction on the membrane surface.[111] Our results ultimately confirm the mechanism of the anti-biofilm activity of acylase-containing coatings and help to understand this approach to biofilm inhibition.

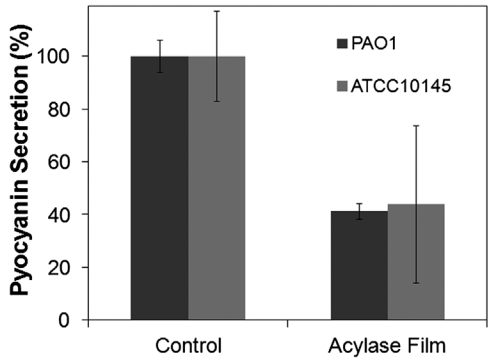

Figure 4.5. Pyocyanin secretion by *P. aeruginosa* PAO1 and ATCC 10145 in the presence of control (acylase-free) and acylase-containing coatings.

Coatings were incubated with 10^4 CFU/mL bacteria for 24 h at 37 °C in LB media. Pyocyanin production was measured spectrophotometrically at 695 nm upon extraction from culture media in chloroform. Error bars represent the standard deviation from the mean for three independent measurements with separate samples.

4.3.5 Stability of Acylase-Containing Coatings

An important question related to the potential clinical utility of acylase-containing coatings for combating biofilm formation is: *how stable are the coatings upon storage and at physiological conditions?* To address this question, the thermal stability of the enzyme-containing coatings was initially measured when stored dry at 4˚C, room temperature, and 37˚C. After 7 days, the enzyme-containing coating retained almost 90% of its initial activity at all three temperatures (**Figure 4.6A**), indicating a significant enhancement in stability upon immobilization of acylase. The initial rapid drop in activity retention at short times may be due to heterogeneity in the enzyme preparation with respect to glycosylation patterns.[108] Specifically, the enzyme used for immobilization in the coatings presumably contains different populations with varying extents of glycosylation, which, in turn, likely have different stabilities. Accordingly, the initial drop in activity retention is likely the result of the denaturation of the least stable isoforms, while the stability at longer times is due to the thermostablity of the more stable isoforms.

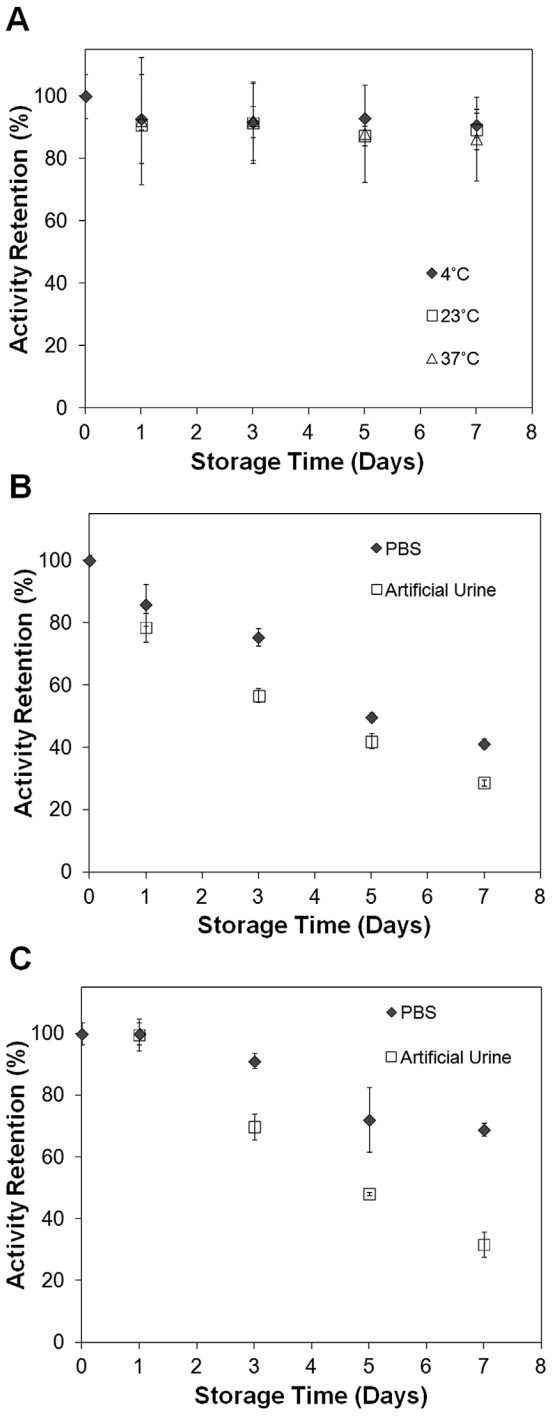

Figure 4.6. Impact of immobilization on acylase stability.
(A) Stability of acylase- containing coatings at 4 °C, room temperature, and 37 °C upon incubation in the dry state. (B) Stability of free acylase in PBS and artificial urine at 37 °C. (C) Stability of acylase-containing coatings in PBS and artificial urine at 37 °C. Residual acylase activity was assayed in 50 mM sodium phosphate buffer (pH 7.5) at 37 °C. Error bars represent the standard deviation from the mean for three independent measurements with separate samples.

In addition to measuring the stability of the coatings in the dry state, the stability of the coatings in PBS solution as well as artificial urine at 37˚C was also measured. The use of artificial urine, in particular, was intended to mimic physiological conditions within a urological catheter. Results of stability studies indicated that the enzyme-containing coatings were significantly more stable than the free enzyme in PBS (**Figure 4.6 B** and **C**). Specifically, in PBS, free acylase retained only 40% of its initial activity after 7 days, whereas the coatings retained 69% of their initial activity over the same period. In artificial urine, both fee acylase and acylase-containing coatings were significantly less stable than in PBS. However, up until 3 days, the immobilized enzyme retained a higher fraction of activity than the free enzyme, which may be critical for inhibiting biofilm formation *in vivo*. As evidence of the enhancement in stability upon immobilized, the free enzyme lost approximately 22% and 41% at 1 and 3 days, respectively, whereas the immobilized enzyme retained virtually 100% activity at 1 day and lost only 30% activity after 3 days (**Figure 4.6 B** and **C**). Notably, the artificial urine as a significantly higher ionic strength and lower pH relative to PBS (4.3 for artificial urine versus 7 for PBS), which is presumably the cause of the decrease in stability of both forms of the enzyme.

4.4 Conclusions

In conclusion, we have immobilized acylase (from *A. melleus*) covalently within polyurethane based polymers to prepare coatings with permanent (*i.e.,* non-leaching) anti-biofilm activity. Acylase activity assays confirmed that the resulting coatings hydrolyzed *N*-acetyl-L-methionine as well as the QS molecules C4-LHL, C6-LHL, and 3-oxo-C12-LHL and were stable upon storage dry and in solution relative to free acylase. Of particular interest, the activity

retention of the immobilized enzyme was greater than free acylase in artificial urine, which has important implications for the use of these coatings for the clinical management of infections in urological catheters. Most importantly, acylase-containing coatings were shown to reduce biofilm formation by approximately 60% in biofilm inhibition studies in static culture conditions using *P. aeruginosa* ATCC 10145 and PAO1. To confirm biofilm inhibition was related to the disruption in quorum sensing, pyocyanin secretion was quantified. Furthermore, SEM characterization of the residual biofilm on enzyme-containing coatings also showed the inhibition of biofilm as compared to control coatings. Such results warrant the further development of these coatings for preclinical testing for potential use in *in vivo* settings.

MULTI-SITE CLICKABLE MODIFICATION OF PROTEINS USING LIPOIC ACID LIGASE

Adapted from Bioconjugate Chemistry, 26, 2015

5.1 Introduction

Chemical modification has widespread utility in modulating the activity and stability of proteins as well as monitoring biological processes (i.e., *in vivo* protein trafficking, protein folding) and as such is an important tool in many fields.[132–135] For example, the modification of proteins with natural and synthetic polymers and sugars may significantly enhance the circulatory lifetime and biological properties of protein drugs.[136–138] The covalent conjugation of responsive polymers (i.e., poly(N-isopropylacrylamide), azobenzene) has also been used to create novel protein switches that turn protein and enzyme activity "on" and "off".[139,140] To date, in addition to various polymers and sugars, proteins have been modified with a broad spectrum of molecular species, including labeling reagents (i.e., fluorophores, radiolabels), affinity tags, lipids, and derivatizing agents for attachment to supports.[57,141,142] The use of such modifying agents has ultimately found application in, among other areas, pharmaceutical development, drug delivery, tissue engineering, enzyme immobilization, and protein-based polymer engineering.[57,143–145]

A major challenge in the modification or derivatization of proteins with various agents entails preserving the protein's three-dimensional state such that the protein remains active and stable. Conventional approaches for chemical modification entail the formation of linkages via reaction with functional groups (i.e., amines or thiols) randomly located on the protein surface.[57]

The covalent modification of proteins via reaction of random residues inherently leads to large heterogeneity in the resulting conjugates. Because the site of modification cannot be controlled, the sites that are critical for protein function may be partially or completely blocked upon modification, impacting protein activity.[146] Additionally, the activity of the protein may be further altered by the modification of residues that are involved in protein dynamics (i.e., hinge motions) as well as by the disruption of protein structure. Although natural residues, including cysteines or lysines, may be introduced at specific sites for modification, non-specific conjugation may still occur if such residues at other sites are not mutated or removed. Furthermore, while techniques to modify the termini of a protein exist, these methods may not have the same desired effect as modification of internal sites and may be of little utility in cases where the termini are critical for function.

As a means to create highly uniform and active modified protein conjugates, the use of bioorthogonal conjugation chemistries presents considerable opportunities.[15,147] Bioorthogonal conjugation chemistries may specifically be enabled through the incorporation of non-canonical amino acids with non-native reactive handles that are unique in nature.[14,148,149] However, such approaches are hampered by low expression yields and incorporation efficiencies of non-canonical amino acids, which are amplified for more than one non-canonical amino acid.[150]

Herein, we present a novel bioorthogonal approach for the site-selective chemical modification of proteins at multiple positions via the enzymatic attachment of click reactive groups. This approach, which is both general and facile, entails the use of the enzyme lipoic acid ligase (LplA) to attach an azide-containing molecule (10-azidodecanoic acid) to a short peptide tag, which is a substrate for the ligase (**Figure 5.1**). The peptide tag, which is a 13 amino acid

sequence known as the LAP sequence (GFEIDKVWYDLDA), may be inserted at the N- or C-terminus of the protein or in internal loop regions, provided its insertion does not perturb protein folding. Following the ligase-mediated reaction, the azide group may subsequently be modified via click chemistry with complementary alkyne-functionalized modifying agents or surfaces (i.e., for protein immobilization). The LAP/LplA system was initially developed by the Ting lab, which demonstrated ligase-mediated attachment of unique azide groups for site-specific labeling of the N- and C-terminus of proteins with fluorophores both *in vitro* and *in vivo*.[20,38,46,50,51] Until now, however, the system has yet to be adopted for making general modifications to proteins at multiple internal sites for protein engineering purposes.

Similar enzyme-mediated approaches to conjugate proteins with modifying agents have been reported using biotin ligase, sortase, transglutamase, farnesyltransferase, phosphopantetheinyl transferase, and formylglycine generating enzyme.[43,151,152] While these approaches allow for single site modifications, such modifications are generally limited to the N- or C-terminus, non-specific, and, in some cases, result in cleavage of the polypeptide backbone of the target protein.[45,153] Additionally, the substrates for these enzymes may not be readily available or easily synthesized, limiting widespread use, and the chemical properties of subsequent reactions, such as slow reaction kinetics, may not be appropriate for particular applications.

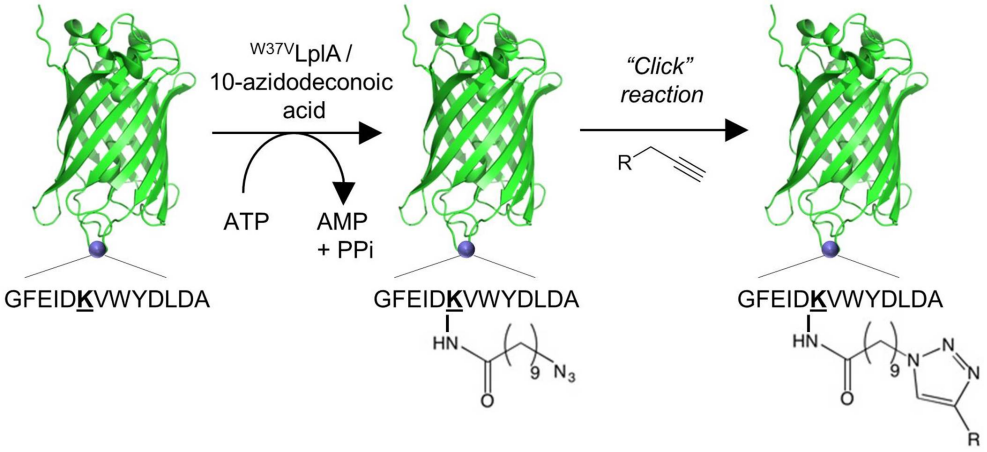

Figure 5.1. Strategy schematic for site-specific ligation of click reactive azide groups using lipoic acid ligase and subsequent chemical modification of GFP by click chemistry.

The utility of the lipoic acid ligase-mediated approach for multi-site selective chemical modification was demonstrated using the model protein green fluorescent protein (GFP). This approach was demonstrated through the design and modification of GFP constructs that contain the lipoic acid acceptor tag at single and multiple sites with polyethylene glycol (PEG). Modification sites included the N-terminal position as well as multiple internal sites, showing the flexibility of this approach with respect to modification position. Ligated GFP constructs were also modified with sugar molecules as well as palmitic acid as a means of demonstrating site-specific glycosylation and modification with fatty acids. Finally, we also used this approach to site-specifically immobilize GFP on a self-assembled monolayer (SAM) with controlled orientation.

5.2 Materials and Methods

5.2.1 Materials

The lipoic acid ligase (^{W3V}LplA) containing plasmid, pYFJ16-LplA(W37V)[50], was obtained from addgene (addgene plasmid 34838). The superfold GFP construct, pET-stGFP,[154] was kindly provided by David Liu of Harvard University and the tRNA/synthetase pair for 4-azido-L-phenylalanine incorporation was provided in the form of the pDule2 pCNF RS plasmid[155] courtesy of Ryan Mehl of Oregon State University. Mutagenic primers were purchased from Integrated DNA Technologies (Coralville, IA). DBCO-PEG (M_w 5kDa) was purchase from Jena Bioscience GmbH (Jena, Germany), propargyl α-D-mannopyranoside from LC Scientific Inc. (Concord, Ontario) and 15-hexadecynoic acid (palmitic acid alkyne) from Cayman Chemical (Ann Arbor, MI). The ligand tris(3-hydroxypropyltriazolylmethyl)amine (THPTA) for copper-catalyzed click reactions was obtained from Click Chemistry Tools (Scottsdale, AZ), 4-azido-L-phenylalanine from Chem-Impex International, Inc. (Wood Dale, IL), and 5,6-epoxyhexyltriethoxysilane from Gelest (Morrisville, PA). All other reagents were purchased from Sigma Aldrich (St. Louis, MO).

5.2.2 Cloning, Expression, and Purification of LAP-Containing GFP Constructs

The LAP sequence GFEIDKVWYDLDA was introduced into stGFP within pET-stGFP using the QuickChange Lightning site-directed mutagenesis kit (Agilent Technologies, Santa Clara, CA). Primers for mutagenesis were designed with 25 to 30 bp of DNA complementary to

pET-stGFP flanking both sides of the LAP sequence. Successful mutagenesis for all LAP-containing stGFP constructs was confirmed via sequencing (Eurofins Genomics, Huntsville, AL).

Wild type GFP and LAP-containing constructs were expressed in BL21 DE3 *E. coli* cells using ampicillin as a selective marker. After initial inoculation, transformed cells were grown in LB to an OD of ~0.6 at 37° C with shaking at 200 rpm and subsequently induced with 1 mM IPTG. Cells were then harvested after an overnight incubation at 25° C and at 200 rpm via centrifugation and lysed by homogenization in 50 mM Tris (pH 8.0), 250 mM NaCl, 5 mM imidazole, 0.01 % β-mercaptoethanol, and 2 % glycerol. The resulting lysate was clarified by centrifugation and loaded onto Ni^{2+} charged Bio-Scale Mini immobilized metal affinity chromatography cartridges (Bio-Rad, Hercules, CA). Protein containing an N-terminal polyhistidine tag was eluted in lysis buffer containing 150 mM imidazole and subsequently dialyzed into 20 mM sodium phosphate buffer (pH 7.0). Purified protein was flash frozen in liquid nitrogen and stored at -80° C.

Constructs for 4-azido-L-phenylalanine incorporation were similarly prepared via mutagenesis except that the amber stop codon TAG was inserted in place of the LAP sequence. The resulting constructs were co-transformed with pDule2 pCNF RS in BL21 DE3 cells and expressed using the same procedure as the other GFP constructs with the exception of the addition of 2 mM 4-azido-L-phenylalanine and the second selective marker, spectinomycin, to the induction media.

5.2.3 GFP Fluorescence Measurements

Excitation spectra (300-550 nm) were obtained for the purified GFP constructs using a Fluoromax-4 spectrofluorometer (Horiba Scientific, Edison, NJ). The excitation maxima for each was found to be 467 nm, which was subsequently used to excite each construct to obtain emission spectra (450-600 nm). Relative stGFP expression levels were obtained by measuring the fluorescence of clarified cell lysate from 250 mL induction volumes. The fluorescent intensities of these samples were converted to protein concentrations with extinction coefficients obtained from purified variants. Total protein expression of each variant was normalized to WT expression levels. Expression data was obtained in duplicate.

5.2.4 Synthesis of 10-Azidododecanoic Acid

The substrate for lipoic acid ligase, 10-azidodecanoic acid, was synthesized following the method of Yao and co-workers.[50] Briefly, 1 g (3.98 mmol) of bromodecanoic acid was added to 10 ml of N,N-dimethylformamide (DMF). To this, 0.5 g (7.69 mmol) of sodium azide was added and allowed to stir at room temperature overnight. The reaction was monitored by thin layer chromotagraphy using 1:2 hexanes:ethyl acetate after which the solvent was removed under vacuum. Following solvent removal, 15 ml of 1M HCl was added to the resulting product, which was then washed 3 times with 15 ml of ethyl acetate. The organic layers were combined and dried over sodium sulfate and subsequently placed under vacuum. The product was then separated by silica gel chromatography using a solvent gradient of hexane and ethyl acetate. The identity of the acid was confirmed by infrared spectroscopy and nuclear magnetic resonance. ^{1}H

NMR ($CDCl_3$, 300 MHz): 3.23ppm (t, 2H), 2.32ppm (t, 2H), 1.57ppm (m, 5H), 1.31ppm (m, 9H). IR (15%EtOAc/Hexanes): 1706, 2093 cm^{-1}.

5.2.5 Ligation Reaction

Ligation of 10-azidodecanoic acid to LAP-containing GFP constructs was performed as described by Yao and co-workers.[50] For this, LplA was expressed in BL21 DE3 cells from the pYFJ16-LplA(W37V) plasmid. Purified LplA was dialyzed into 20 mM Tris (pH 7.5), 0.01 % β-mercaptoethanol, and 10 % glycerol, flash frozen in liquid nitrogen, and stored at -80° C. Ligation reactions containing 0.1 μM LplA, 10 μM GFP, 600 μM 10-azidodecanoic acid, 2 mM ATP, 2 mM $MgCl_2$, and 25 mM phosphate buffer (pH 7.2) were incubated at 30° C. Reactions were quenched at various time points via chelation by the addition of EDTA to a final concentration of 300 mM. Quenched ligation products were characterized via ESI mass spectrometry (Synapt G2), which was preceded by exchanging the buffer with 40 % acetonitrile and 0.1% formic acid. Spectra were deconvoluted with MaxEnt, and conversions were obtained by dividing the peak height of ligated GFP by the sum of ligated and un-ligated GFP peak heights. Ligated protein was thoroughly dialyzed to remove all components of the ligation reaction before being used in subsequent click reactions. Protein solutions were loaded into seamless cellulose dialysis tubing from Fisher Scientific (Waltham, MA) with a molecular weight cutoff of 12-16 kDa and were deposited into 4 L of target buffer. No more than 20 mL of protein solution were included per 4 L of target buffer, and the target buffer was exchanged a minimum of three times over the course of 12 hours, resulting in a theoretical dilution factor of 8 million for contaminating components of the ligation reaction.

5.2.6 PEGylation of GFP Constructs

Non-specific PEGylation of wild type GFP was performed in 50 mM sodium phosphate buffer (pH 7.5) and at room temperature for 3 h via reaction with NHS-PEG. A 5:1 molar ratio of PEG-to-primary amine was used for the reaction. Site-specific PEG attachment using DBCO-PEG was similarly performed in 50 mM sodium phosphate buffer (pH 7.0) using a 25 molar excess of DBCO-PEG for each ligation site to obtain samples for MALDI and a 50 molar excess of DBCO-PEG to obtain samples for SDS-PAGE. The click reaction was incubated at room temperature for 0.5-6 h. Reaction samples were removed at varying time points and quenched with 100 mM tris(2-carboxyethyl)phosphine. Following the reaction, the reaction products were characterized by SDS-PAGE and MALDI-TOF mass spectrometry. For iodine stained SDS-PAGE gels, the gels were stained for 10-30 min and imaged after destaining with water. The iodine stain consisted of 1.3 % (w/v) iodine, 1.0 % (w/v) potassium iodide, and 2.5 % (w/v) barium chloride in a 0.6 M HCl solution. After imaging, gels were re-stained with coomassie and reimaged. Images of coomassie stained gels were analyzed quantitatively by densitometry using ImageJ.[156]

5.2.7 Copper-Catalyzed Click Modification of Ligated GFP Constructs

Propargyl α-D-mannopyranoside conjugation to ligated GFP by copper-catalyzed azide-alkyne cycloaddition was performed using the optimized conditions described by Hong et al.[157] Initially, 865 µL of GFP (10 µM in 50 mM sodium phosphate, pH 7.0) was added to 20 µL of propargyl α-D-mannopyranoside (20 mM in deionized water) in a microcentrifuge tube. To the GFP mixture, 15 µL of a solution of $CuSO_4$ and THPTA, which was prepared by initially mixing

50 μL of 20 mM $CuSO_4 \cdot 5H_2O$ (in deionized water) and 100 μL of 50 mM THPTA (in deionized water) for 20 mins, and 50 μL of aminoguanidine hydrochloride (100 mM in deionized water) was added. The click reaction was initiated by the addition of 50 μL of sodium ascorbate (100 mM in deionized water), which was allowed to proceed for 4 h at 37 °C.

Palmitic acid alkyne conjugation to GFP was performed using conditions identical for the α-D-mannopyranoside attachment except the palmitic acid-alkyne stock was prepared in DMSO. Additionally, the click reaction was performed in the presence of 2% sodium deoxycholate to solubilize the palmitic acid.

Following the attachment of the α-D-mannopyranoside or palmitic acid, the resulting reaction mixtures were dialyzed with deionized water. The GFP conjugates were then analyzed by MALDI-TOF MS.

5.2.8 Click Immobilization of Ligated GFP Constructs

Glass cover slides were initially washed with detergent and thoroughly rinsed with ultrapure water. The slides were subsequently immersed in warm piranha solution for 1 h followed by thorough rinsing with purified water, drying with ultrapure nitrogen, and exposure to UV-ozone for 15 mins. The surface was then functionalized with epoxide groups by forming a SAM of 5,6-epoxyhexyltriethoxysilane. The SAM was formed by exposing the cleaned glass to vapors of a mixture of the silane (10% v/v), n-butylamine (5% v/v), and toluene (85% v/v) for 20 h. Attachment of the silane was confirmed by using a custom built goniometer to measure the static water contact angle. The static water contact angle after silane modification was 36.3° ± 0.8° where the error represents the standard deviation from 12 total measurements over 4 slides.

Following deposition of the SAM, the terminal epoxide groups were reacted with dibenzocyclooctyne-amine (DBCO-NH_2) to introduce alkyne groups to the surface. For this reaction, DBCO-NH_2 was dissolved in DMSO at a concentration of 5 mg/mL and added to a borate buffer (100 mM, pH 9.5), resulting in a final concentration of 1 mM. The DBCO-NH_2 solution was subsequently added to a small petri dish containing the epoxide-modified surfaces to start the coupling reaction. The reaction mixture was placed in an orbital shaker at 37°C and allowed to react for 30 h. After the reaction, the surfaces were thoroughly washed with deionized water and then immersed 1 % w/v ammonium chloride (in 100 mM borate buffer, pH 9.5) overnight at 37 °C to quench the remaining epoxides. To attach GFP, 10 nM of the ligated GFP (or WT-GFP for control) in sodium phosphate buffer (100 mM, pH 7.0) was allowed to react with the DBCO functionalized surface for 4 h under gentle agitation. Finally, non-covalently bound GFP was washed from the surface by gently shaking with sodium phosphate buffer (100 mM, pH 7.0) for 4 hours at 37 °C. The rinse buffer was replaced approximately every 30 mins during washing.

5.2.9 Imaging of GFP Attachment Using Epifluorescence Microscopy

Glass slides were imaged in epifluorescence mode using a Nikon TE-2000 microscope, 60x objective, and an Andor iXon 3 EMCCD camera (model DU 888). Light was provided by a metal halide arc lamp (X-Cite 120, Lumen Dynamics) and a filter cube to select light from 450-490 nm for excitation and capture light from 505-550 nm for fluorescence imaging. A spot on the glass slide was photobleached for 10 mins at a power density of 20 W/cm^2 using an iris to

block light from the surrounding area. Following bleaching, a video was recorded at ~19 fps (52 ms per frame) while the iris was opened. The 10 frames immediately after the iris was fully open were averaged to image the contrast between bleached and unbleached areas of the surface. All surfaces were imaged using identical camera settings and illumination power density.

5.3 Results and Discussion

5.3.1 Design and Expression of LAP-Containing GFP Constructs

Site-specific chemical modification of protein using the LAP/LplA system was investigated by designing GFP constructs containing the LAP sequence at terminal and internal positions. For demonstrating the utility of this approach, GFP is an ideal model protein due to its secondary structure, which contains a large number of flexible loop regions (**Figure 5.2a**).[158] These loop regions, which can accommodate the insertion of large polypeptides as well as tolerate altered connectivity,[159–161] provide numerous potential LAP insertion sites on either end of the protein's beta barrel structure. Furthermore, because the fluorescence of GFP is sensitive to its folding state and specifically the structure of its chromophore,[162] the structural impact of mutations and subsequent chemical modifications may easily be monitored. Accordingly, the fluorescence spectra of LAP-containing GFP constructs may be compared to that of wild-type GFP to determine if GFP folding and chromophore formation is perturbed. The intrinsic fluorescence of GFP can also be used to estimate expression levels of folded protein within cell lysate[163] and to observe surface immobilization, which can be imaged by fluorescence microscopy.

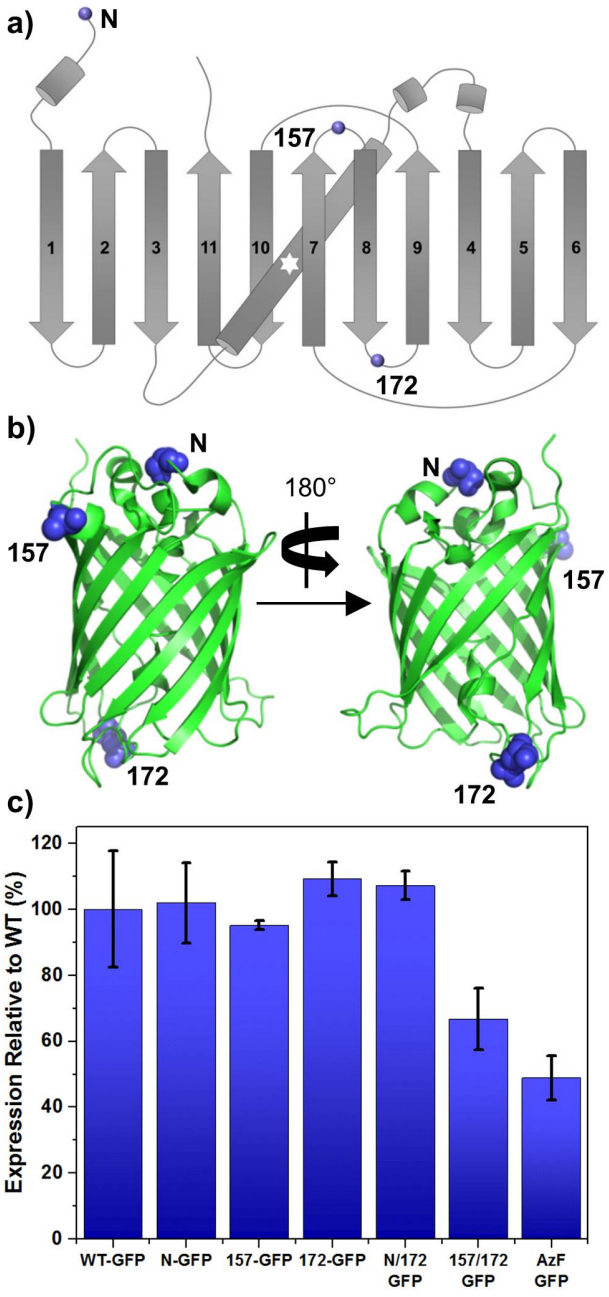

Figure 5.2. Location of LAP insertion sites in GFP and expression of LAP-GFP constructs relative to WT

(a) The location of LAP insertion sites within GFP relative to primary sequence and secondary structure. Insertion sites are represented with blue spheres. (b) The position of LAP insertion sites in the context of folded GFP. Residues preceding insertion sites are depicted as spherical models in blue. (c) Expression data for all LAP-containing GFP constructs relative to WT-GFP. For AzF-GFP, the non-canonical amino acid AzF was incorporated at the N and 172 sites.

Five different LAP-containing GFP constructs were prepared by site-directed mutagenesis. Constructs with single LAP insertions at the N-terminus (N-GFP) as well as between Q157 and K158 (157-GFP) and between E172 and D173 (172-GFP) were designed. Additionally, constructs containing two simultaneous LAP insertions were prepared, the first with LAP sequences at the N-terminus and between E172 and D173 (N/172-GFP) and the second with LAP sequences between Q157 and K158 as well as between E172 and D173 (157/172-GFP). Notably, the sites at positions 157 and 172 are in solvent-exposed loops on either end of the beta barrel structure (**Figure 5.2b**). As such, the 157/172-GFP construct contains LAP insertion sites at opposite ends of the protein, permitting modification of both ends of the protein simultaneously. Moreover, the site at position 172 is on the opposite side of the N and C-termini and thus enables modification of a region of GFP that is not accessible by modifying the termini alone. Although the LAP sequence has previously been introduced at internal sites within fusion constructs[38] (i.e. between a protein and a polyhistidine tag), the use of the LAP/LplA system for labeling internal, constrained sites within a single protein in a manner that disrupts the protein's primary sequence as done here has not previously been reported to the best of our knowledge.

All of the LAP-containing constructs, including double insertion constructs, were found to express well relative to WT-GFP as determined by measuring intrinsic GFP fluorescence (**Figure 5.2c**). To determine expression levels, the intrinsic fluorescence of GFP in crude *Escherichia coli* lysate for each construct was measured and normalized per liter of culture. The expression level of the double insertion construct 157/172-GFP was slightly lower than the other constructs, but still greater than that of GFP containing the azide-functionalized non-canonical amino acid 4-azido-L-phenylalanine. In this case, a double non-canonical amino acid construct

(AzF-GFP) was expressed with 4-azido-L-phenylalanine at the N-terminus and at position 172 for comparison. The improved expression with the LAP sequence, which does not require altered transcriptional machinery, relative to AzF-GFP represents a major advantage of the LAP/LplA system. Additionally, use of the LAP tag evades translational inefficiencies of non-canonical amino acid incorporation such as release factor binding and mis-incorporation of native amino acids.

To confirm that fluorescence can be used to quantify expression in lysate, we confirmed that the fluorescence properties of the purified GFP constructs were similar. For all of the purified constructs, the excitation and emission maxima was 467 nm and between 509-511 nm, respectively. Similar fluorescence intensities, within 20 % of WT-GFP, were also observed per μmol of purified protein (normalized through measurement of OD 280 nm) for each construct and used to calculate extinction coefficients. In addition to allowing the use of fluorescence measurements to compare expression levels, these results ultimately indicate that LAP insertion within GFP had little impact on protein structure, suggesting that the LAP sequence, in theory, may be inserted into flexible loop sites on other proteins as well without disrupting function.

5.3.2 Characterization of Ligase Reaction with GFP Constructs

Having expressed and purified LAP-containing GFP constructs, a critical next step was to investigate the efficiency and kinetics of the post-translational LplA-catalyzed ligation reaction. In performing this reaction, GFP constructs (10 μM) were ligated with the azide-containing substrate, 10-azidodecanoic acid (600 μM), using the previously described enzyme mutant, ^{W37V}LplA.[50] To determine the ligation efficiency as well as monitor the ligation rate, the addition

of 10-azidodecanoic acid was monitored by electrospray ionization (ESI) mass spectrometry. As shown in the ESI spectra for N-GFP in **Figure 5.3a**, ligation of 10-azidodecanoic acid resulted in a shift of 195 Da, which is consistent with the expected mass increase. The peak for the expected ligation product increased over time while the peak for the native (i.e., unmodified protein) diminished until near complete conversion was reached at approximately 3 h. Similar rates and extent of final conversion were observed for both single internal LAP-containing constructs relative to the N-terminal LAP construct (**Figure 5.3b**). No mass shift was observed either for the control ligation reaction with WT-GFP under the same conditions or for LAP-containing constructs in reactions without ^{W37V}LplA, confirming that ligation is enzyme dependent and is specific to the LAP sequence.

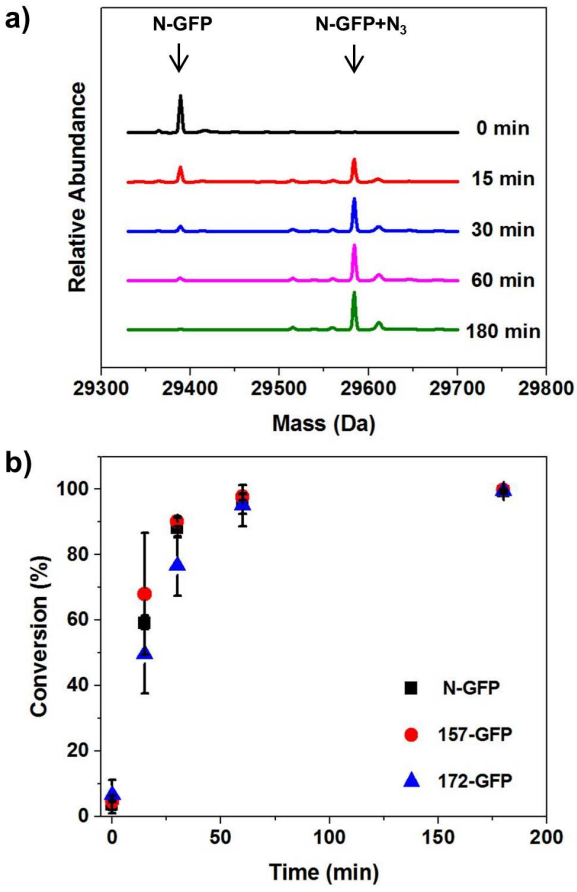

Figure 5.3. Ligation of 10-azidodecanoic acid to LAP-GFP constructs.
(a) ESI results for 10-azidodecanoic acid ligation to N-GFP. Relative abundance measurements are offset to illustrate change in spectra as a function of reaction time. (b) Kinetics of ligation reaction for GFP constructs with single LAP insertions at positions N, 157, and 172.

In addition to the ligation reaction being highly efficient, ^{W37V}LplA expresses well in *E. coli* and its substrate, 10-azidodecanoic acid, is easy to synthesize. This represents an advantage of the LAP/LplA system that is critical to its potential widespread utility for protein modification. Another important advantage of the LAP/LplA system for modification is the ability to temporally control the introduction of azide groups at desired sites. Specifically, using this approach, a LAP-containing protein can be expressed and purified without azide functionalization, which can avoid reduction of the azide group. The reduction of azides when introduced via non-canonical amino acid incorporation is particularly problematic as many proteins must be stored under reducing conditions for stability. Disruption of azides can also occur within the reducing environment in cells[164] or through UV-induced photolysis.[165] In all cases, azide reduction ultimately leads to greater heterogeneity within a protein population.

Apart from decreasing heterogeneity by limiting opportunities for reduction, temporal control of the LAP/LplA system also presents intriguing possibilities for modification strategies that simultaneously use non-canonical amino acid incorporation and LAP modification for sequentially labeling different sites. For example, a protein containing 4-azido-L-phenylalanine and a LAP site could potentially be modified by reaction of the non-canonical amino acid prior to ligation of the LAP site with an azide group. The combination of the LAP/LplA system with non-canonical amino acid incorporation could also be, in theory, exploited to introduce more than one bioorthogonal functionality within a single protein. These possibilities were not

explored in the present work, but the chemistry employed during modification with the LAP/LplA system and during non-canonical amino acid incorporation is, in theory, compatible. A combined modification approach, therefore, may represent an exciting subject for future inquiry.

5.3.3 PEGylation of Ligated GFP Constructs

An interesting area in which the LAP/LplA system may be particularly useful is PEGylation of therapeutic proteins. The conjugation of polyethylene glycol (PEG) is a widely used strategy for improving the efficacy of therapeutic proteins by increasing circulatory lifetime *in vivo*.[5] Such an increase in circulatory lifetime can result from reduced proteolytic degradation, immunogenicity to the therapeutic protein, and renal clearance. However, non-specific approaches to PEGylation lead to highly heterogeneous populations with individual protein molecules varying in the number and location of PEG attachment sites. Such heterogeneity requires often challenging purification of the desired PEG-protein conjugate, which may be necessary and critical to ensure safety and therapeutic efficacy.

To demonstrate the utility of the LAP/LplA system for site-specific protein PEGylation, the ligated GFP constructs were reacted with dibenzocyclooctyne-polyethylene glycol (DBCO-PEG; M_w 5kDa). The reaction of DBCO-PEG with the ligated azide groups occurs via strain-promoted azide-alkyne cycloaddition and yields a stable covalent triazole linkage between the PEG and protein. Reaction products were characterized by matrix-assisted laser desorption ionization (MALDI) time-of-flight mass spectrometry (**Figure 5.4a**) as well as SDS-PAGE (**Figure 5.4b**). In the case of SDS-PAGE, the presence of PEG in a particular protein band was confirmed by staining with iodine in addition to coomassie staining. The reaction of ligated,

LAP-containing constructs with DBCO-PEG resulted in large proportions of protein migrating as higher molecular weight bands that stained with iodine, suggesting PEG attachment. MALDI data confirmed that these bands corresponded to PEGylated protein as the predominant peaks within the spectra matched expected molecular weights for the addition of one PEG molecule to single LAP-containing constructs and the addition of two PEG molecules to double LAP-containing constructs. Of note, when characterized by SDS-PAGE, the band for the PEGylated 157/172-GFP construct was unexpectedly higher than that for the PEGylated N/172-GFP construct. This apparent mass difference, which was not observed by MALDI (data not shown), was determined to be an artifact related to differences in SDS-PAGE mobility upon PEGylation at different sites.

For all of the GFP constructs, over 70 % of the total protein was converted to the desired product (i.e., one attached PEG molecule per LAP insertion) within 30 minutes as determined by densitometry measurements of the gel images, and generally, only a single protein band was observed that corresponded to undesired products, where an undesired product is any protein molecule that does not contain one attached PEG for every inserted LAP sequence. Moreover, no conjugation was observed for control reactions in which GFP constructs, which had been subjected to LplA ligation in the absence of 10-azidodecanoic acid and therefore did not contain and azide moiety, were treated with DBCO-PEG, indicating that PEGylation was restricted to ligated LAP sites (data not shown). This combination of high site specificity and conversion enables the production of highly homogeneous PEGylated protein populations while reducing or eliminating the need to separate undesirable conjugates and potentially reducing the number of conjugates that require characterization.

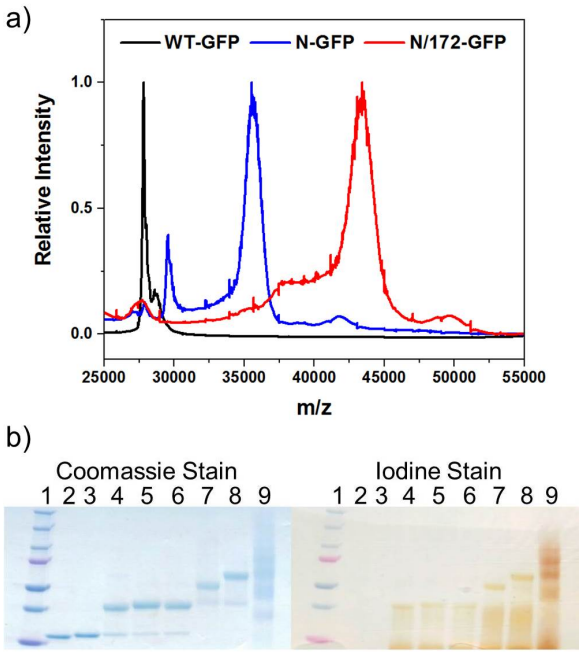

Figure 5.4. Single PEGylation of N-GFP and double PEGylation of N/172-GFP via strain promoted cycloaddition with ligated 10-azidodecanoic acid.
(a) Representative MALDI-TOF spectra of WT-GFP, ligated N-GFP+DBCO-PEG (labeled as N-GFP), and ligated N/172-GFP+DBCO-PEG (labeled as N/172-GFP). (b) SDS-PAGE of PEGylated GFP samples after 30 min reaction time. Coomassie staining indicates the presence of protein while iodine staining indicates the presence of PEG. Lane 1: Precision Plus Dual Color Ladder. Lane 2: WT-GFP. Lane 3: N-GFP. Lane 4: ligated N-GFP+DBCO-PEG. Lane 5: ligated 157-GFP+DBCO-PEG. Lane 6: ligated 172-GFP+DBCO-PEG. Lane 7: ligated N/172-GFP+DBCO-PEG. Lane 8: ligated 157/172-GFP+DBCO-PEG. Lane 9: WT-GFP+NHS-PEG.

To illustrate the advantage of the present approach for PEGylation over non-specific techniques, WT-GFP was randomly reacted with methoxypolyethylene glycol propionic acid N-succinimidyl ester (NHS-PEG; 5kDa M_w). Primary amines, such as a protein's N-terminus and lysine residues, serve as reaction sites for NHS chemistry, and as such, there are 22 potential reaction sites in WT-GFP. Unlike with the LAP-mediated approach, conventional PEGylation with this chemistry yielded a heterogeneous population of PEG-protein conjugates with four major products as observed by SDS-PAGE (**Figure 5.4b**). These products correspond to proteins

that vary in either the number or the location of attached PEG molecules, and thus the potential heterogeneity of this molecular population is substantial.

Although the limitations of non-specific reaction chemistries for therapeutic protein PEGylation may alternatively be overcome with non-canonical amino acids,[163] one possible downside to this approach is that the incorporation of non-canonical amino acids may elicit unwanted immunogenicity. For example, *p*-nitrophenylalanine has been shown to trigger an immune response to otherwise self-tolerant proteins, including TNF-α and RBP4, when administered in mice.[166] Restricting residues to native amino acids may reduce the risk of such adverse responses, although additional studies would be required to determine if the LAP insertion is itself immunogenic, which may be protein and site dependent. Additionally, the immunogenicity of other components of the reaction chemistry including DBCO and triazole groups would need to be determined.

5.3.4 Glycosylation and Fatty Acid Modification of LAP-Containing GFP Constructs

In addition to PEGylation, site-specific glycosylation and fatty acid modification can modulate protein structure and function. However, site-specific glycosylation, in particular, represents a major challenge since many bacterial expression strains (e.g., *E. coli*) lack the necessary glycosylation machinery and pathways.[167,168] Additionally, eukaryotic expression generally results in diversity in position, length, and branching pattern of attached sugars, depending on organism and growth conditions.[169] Because many glycosylated proteins possess therapeutic potential, the development of approaches to produce well-defined glycosylated proteins has received considerable attention. Fatty acid modifications, though less prevalent in

nature than glycosylation, can enhance transport across lipid barriers (e.g., the blood brain barrier) and stability *in vivo* (e.g., via binding albumin), and thus well characterized attachment of fatty acids also has important implications in improving the properties of therapeutic proteins.[142]

To further illustrate the breadth of feasible modifications, the LAP/LplA system was used to site-specifically glycosylate as well as to modify GFP with fatty acids. This was demonstrated by attaching alkyne-functionalized α-D-mannopyranoside and palmitic acid to both ligated LAP sites on the 157/172-GFP construct using copper-catalyzed click chemistry (**Figure 5.5**). Unmodified 157/172-GFP has a molecular weight of 30,967 Da and increases in mass by 824 Da (theoretical) with ligation and subsequent mannose attachment at both sites and by 892 Da (theoretical) with ligation and palmitic acid attachment at both sites. The apparent shifts in mass when measured by MALDI-TOF were found to be within 10 percent of these theoretical mass increases, and no prominent peaks or shoulders within the observed peaks were found to correspond with unmodified protein. Taken together, these results qualitatively indicate successful protein modification with both mannose and palmitic acid.

Protein glycosylation at internal positions has been carried out in a similar manner previously through the use of formylglycine generating enzyme (FGE), which converts a cysteine residue within the CXPXR recognition sequence to a formylglycine residue that can then be used for glycan attachment via aldehyde dependent bioorthogonal reactions.[170,171] Our results indicate that the LAP/LplA system may serve as an alternative to the FGE approach, increasing the available chemical diversity that may be used to attach a glycan to a recognition sequence, namely through the azide-alkyne click reaction that the LAP system employs.

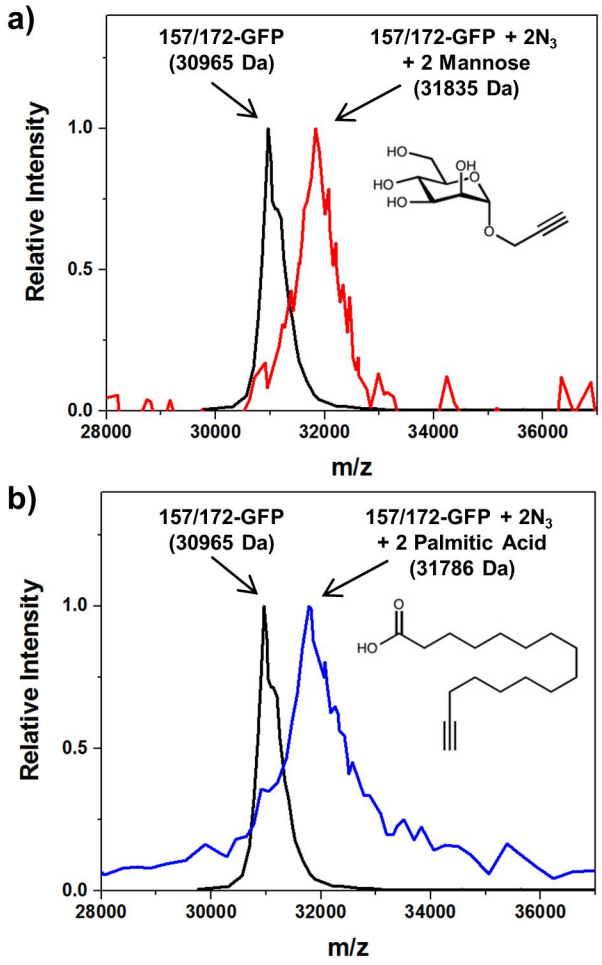

Figure 5.5. Double glycosylation and fatty acid modification of 157/172-GFP via copper catalyzed click chemistry with ligated 10-azidodecanoic acid
MALDI mass spectra for single charge states of 157/172-GFP with mannose (a) and palmitic acid (b) attachment at both internal LAP sites. The peak for 157/172-GFP is in black while the peaks for the attachment of two α-D-mannopyranoside and two palmitic acid molecules are in red and blue, respectively. Structures of mannose and palmitic acid are shown in each panel. Observed mass values for each major peak are provided in parentheses.

5.3.5 Immobilization of Ligated GFP Constructs on Self-Assembled Monolayers

The bioorthogonal nature of the LAP/LplA system combined with the ability to design internal insertion sites suggests that it may also serve as a convenient strategy for orientation-controlled protein immobilization. To explore this possibility, ligated 172-GFP was immobilized

on a DBCO-modified epoxide-terminated SAM, which was used as a model surface. Protein samples were thoroughly dialyzed to remove any contaminating components from the ligation reaction prior to immobilization, and surfaces were extensively washed after immobilization to remove any non-covalently bound protein. Surfaces were imaged using fluorescence microscopy, illuminating the protein molecules bound on the surfaces (**Figure 5.6a**). The spot in the middle of the fluorescent image was photobleached using long exposure times as a control to show the contrast between the fluorescence of the rest of the surface containing GFP and the spot. This contrast was large on the surface treated with ligated 172-GFP and negligible with WT-GFP (i.e., unligated GFP), which was used as a control, indicating a much higher presence of bound 172-GFP. To quantify this contrast, the relative fluorescence intensity across the red line in the fluorescent images of the surfaces was plotted using image analysis (**Figure 5.6b**). These results suggest that 172-GFP is immobilized, and that immobilization requires the azide functionality. This requirement implies that bound GFP molecules possess a single uniform orientation, and, because the 172 LAP insertion site is on the opposite side of GFP's beta barrel than both the N and C-termini, it is a particular orientation that would not be possible without an internal insertion of the LAP sequence. These results demonstrate the utility of the LAP/LplA system for controlling the orientation of immobilized biomolecules and thus its possible applicability to many areas of general importance, including biosensing and the creation of immobilized biocatalysts.

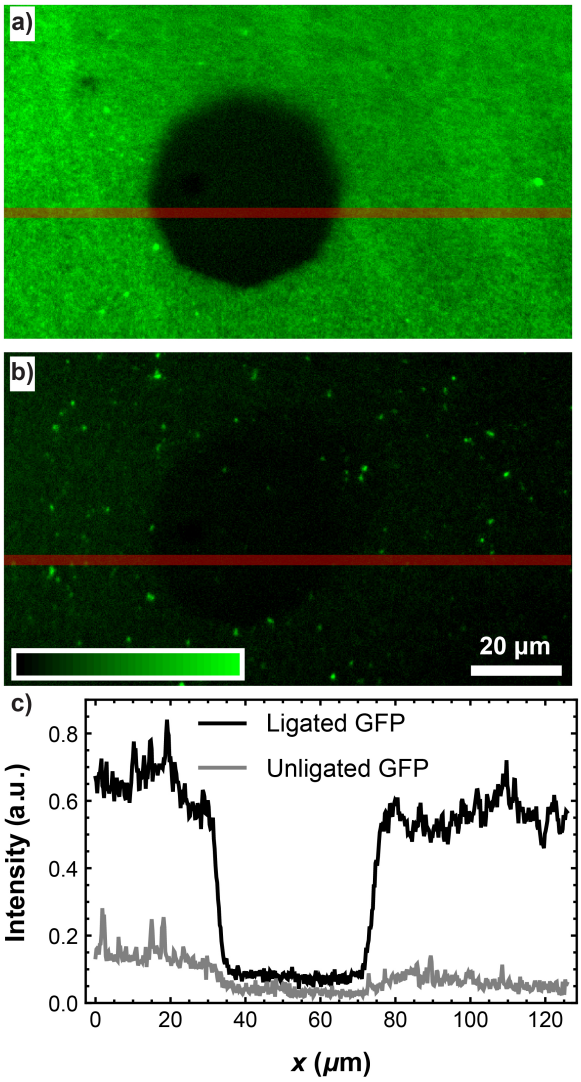

Figure 5.6. Covalent immobilization of 172-GFP to a strained-alkyne functionalized self-assembled monolayer.
(a) Image of surface after reaction of ligated 172-GFP with strained alkyne functionalized SAM. (b) Image of surface treated with WT-GFP control. The dark octagons in (a) and, less visible, in (b) were produced by photobleaching. The fluorescence intensity measured along the red lines in (a) and (b) was plotted in (c).

5.4 Conclusion

In summary, we have developed an enzymatic approach for multi-site clickable modification based on the incorporation of azide moieties in protein using LplA. This approach was applied to the site-specific attachment of PEG, α-D-mannopyranoside, and palmitic acid to

GFP as well as the immobilization of GFP on model SAM surfaces. Importantly, strong expression levels were obtains for the various GFP-tagged constructs, which contained single as well as double, simultaneous LAP insertions. Furthermore, the ligase-mediated modification of GFP with 10-azidodecanoic acid at the N-terminus and two internal positions was rapid and highly efficient. Our findings ultimately demonstrate the considerable potential utility of the LAP/LplA system for protein modification. Of particular advantage using this approach is the flexibility to incorporate multiple sites, including internal sites, at the same time. Additionally, this approach provides unprecedented temporal control of both azide functionalization as well as subsequent modification. Although we have demonstrated that the LAP/LplA system is quite general in the types of protein modifications it allows at internal sites within GFP, an interesting question that remains is how well other, diverse proteins and sites can accommodate LAP insertion and whether the LAP sequence can be engineered to be compatible with sites that are less accommodating than the flexible loops of GFP. Addressing this question is an important and natural next step; however, the advantages of the LAP/LplA system combined with its demonstrated utility within GFP nevertheless make it an important system to consider when designing protein modifications.

ROSETTA-ENABLED STRUCTURAL PREDICTION OF PERMISSIVE LOOP INSERTION SITES IN PROTEINS

In preparation for publication

6.1 Introduction

Loop regions within proteins often contribute substantially to function. They participate in determining the binding specificity of antibodies,[172] modulating substrate recognition in enzymatic catalysis,[173,174] and by serving as targeted recognition sequences for post-translational modifications (PTMs).[175] As a consequence, altering or introducing peptide loops within a protein may generate novel function. Indeed, loop alterations have been observed in enzyme evolution as a route by which proteins within the same family diversify.[176] In the field of protein engineering, the combination of novel loop regions with diverse protein targets has resulted in chimeric molecules with multifunctional characteristics.[177] A prominent example is the introduction of antibody-like loops into non-immunoglobulin protein scaffolds.[21,178] Another is the implantation of peptide recognition sequences that are functionalized by PTM-performing enzymes within protein conjugation targets.[45,179] These recognition sequences can facilitate the addition of fluorophores for protein trafficking experiments as well as the addition of bioorthogonal reactive handles that can direct the conjugation of diverse molecules and thereby address a wide range of applications.[46,50]

In spite of the potential associated with engineering chimeric proteins to contain novel peptide loops, the ability to identify sites within a target protein that can accommodate a loop insertion persists as a major challenge limiting its practical use. Often times, in what is referred

to as loop grafting, existing loops within the target protein are extended or replaced by the novel loop that is being introduced.[180,181] In identifying grafting sites, areas with high *B*-factors and low amounts of secondary structure are typically considered to be promising candidates. However, these characteristics do not guarantee loop accommodation, and inserting a large peptide into the structure of a target protein commonly interferes with folding. Additionally, not all proteins contain highly flexible loop regions to target, and, in the event that they do, such regions often serve pre-existing functional roles, which make them unsuitable for modification. The limitations in identifying permissive loop insertion sites are especially cumbersome when the exact location of an inserted loop is critical as would likely be the case for a designed PTM site. Given these circumstances, a systematic approach for accurately predicting permissive loop insertion sites, including unintuitive sites, within proteins is highly desirable.

The Rosetta protein modeling software suite represents a potential computational platform for such a systematic approach. The Rosetta framework has been successfully applied to a large variety of problems in structure prediction and protein design, including the design of chimeric proteins.[23–27,182] Additionally, it contains loop modeling protocols that are used for filling in unresolved portions of crystal structures or homology models and that work by modeling short strands of amino acids within the context of a preexisting protein structure.[28,79] Although Rosetta loop modeling has not thus far been used to predict permissive loop insertion sites in proteins via full structural scans, it is highly amenable to the task due to its ability to introduce and model a novel peptide loop at any position within a target protein scaffold without remodeling the entire target protein in the process.

In the present work, we adapt a Rosetta loop modeling application for the rapid identification of loop insertion sites within proteins using lipase A (lipA) from *Bascillus subtilis*, β-glucosidase from *Tichoderma reesei*, and human phosphatase and tensin homolog (PTEN) as model target proteins and the Lipoic acid ligase Acceptor Peptide (LAP) sequence as a model loop. The LAP sequence is a 13-residue peptide (GFEIDKVWYDLDA) that is site specifically ligated with lipoic acid by the enzyme, lipoic acid ligase (LplA). It is highly modular and can be recognized by LplA as a free peptide in solution as well as when genetically fused to a protein of interest.[20] Prior work has shown that the LAP sequence can be efficiently ligated when inserted at internal positions within the structure of GFP, making it a tremendous model for designing internal PTM sites within proteins via peptide loop insertions.[179] Additionally, LplA has been engineered through a number of strategies (including Rosetta computational design)[48] to attach diverse molecules including fluorophores and azide-bearing lipoic acid derivatives to the LAP sequence.[50] The ligation of an azide moiety extends the utility of the LAP sequence to a wide range of protein engineering applications as it allows subsequent modification with diverse alkyne-functionalized molecules through azide-alkyne click chemistry. These modifications can include glycosylation and polymer attachment. Additionally the attached azide can be used to direct protein immobilization.[179]

Using the LAP sequence as a model loop, we show that energy scores obtained from our modeling approach correlate with soluble protein expression and that they accurately predict LAP accommodation at highly unintuitive sites based on *B*-factor and secondary structure considerations. Additionally, we find this approach to be extremely user accessible, allowing an entire protein structure to be scanned with a single laptop computer within a number of days. We

find the predictive capability to be especially effective at differentiating between potential LAP insertion sites locally within particular regions of a protein structure and discover that these predictions are primarily derived from the identity of the amino acid residues in the near loop environment. These results indicate that the residue environment of a loop insertion site is a critical structural determinant in loop accommodation. Based on our observations, we present a computational strategy for decreasing the number of constructs that need to be built and tested in order to identify permissive loop insertion sites within target proteins. We demonstrate the effectiveness of this strategy by rapidly identifying five LAP insertion sites within the structure of PTEN that behave similarly to WT and that avoid modifying the physiologically active termini of the protein.

6.2 Materials and Methods

6.2.1 Materials

All primers and g-block gene fragments were ordered from IDT (San Jose, CA). All reagents including InVision™ His-tag IN-gel stain and restriction enzymes were ordered from Thermo Fisher Scientific (Waltham, MA) unless otherwise stated. Bis-acrylamide, Precision Plus Protein Dual Color Standards (gel ladders) and all gel casting and running materials were purchased from Bio-Rad. Rosetta source code was downloaded from the Rosetta Commons website using an academic license. Plate One 2-mL 96-well plates were purchased from USA Scientific.

6.2.2 Cloning, Expression, and Purification of Protein Constructs

A pet21B-lipA plasmid that contained lipA PCR amplified from *Bacillus subtilis* with a stop codon immediately after the gene and cloned between the NdeI and SacI sites of pET21B+ (previously described)[183] was modified so as to express with a C-terminal his-tag. Specifically, the stop codon and base pairs from the multiple cloning region between the end of the lipA gene and the plasmid's intrinsic C-terminal his-tag coding sequence were deleted via site-directed mutagenesis. The resulting product, which expressed lipA with a his-tag directly fused to its C terminus with no linker in between was used for WT lipA expression and was genetically modified to produce all LAP-containing lipA constructs.

A pet21B-β-glucosidase plasmid was made via restriction cloning of the β-glucosidase gene between NdeI and XhoI restriction sites of pET21B. A stop codon was not included in the gene so that translation resulted in a C-terminal his-tag with a two residue linker corresponding to the base pairs associated with the XhoI restriction site. The β-glucosidase gene was synthesized by IDT as a g-block gene fragment that was codon optimized for *E. coli* expression. The NdeI and XhoI sites were appended to either end along with 8 random base pairs flanking the restriction sites to facilitate restriction enzyme nuclease activity.

The human PTEN gene was also synthesized by IDT as a g-block gene fragment, and was cloned into the pET28A plasmid as described previously.[19]

LAP-protein constructs were made either via site directed mutagenesis using large insertion primers to add the LAP sequence as previously described or by ordering a complete gene that contained the LAP sequence, which was cloned into the appropriate plasmid as

described above. The same DNA sequence, GGC TTC GAG ATC GAC AAG GTG TGG TAC GAC CTG GAC GCC, was used to code for the LAP sequence in all circumstances.

Point mutations in the lipA LAP-43 construct were made using QuikChange Lightning mutagenesis kits.

All DNA constructs were confirmed by sequencing and transformed in BL21 (DE3) *E. coli* cells for expression. 96-well glycerol stock plates were made, with each well containing a unique construct to enable high throughput expression of the libraries (discussed below). These were made by mixing a 50 % glycerol solution with overnight cultures for each of the constructs and stored at -80 °C.

LAP-lipA constructs that were purified for CD and specific activity analysis were expressed at 13 °C overnight in LB-amp after induction with IPTG. Cells were lysed via sonication, cell lysate was cleared via centrifugation and vacuum filtration with a 0.2 μm cutoff filter, and protein was purified on a Ni-NTA column. Purified protein was dialyzed overnight into 20 mM sodium phosphate buffer (pH 7.0), flash frozen in liquid nitrogen, and stored at -80 °C.

6.2.3 Test Expression of LAP-Protein Libraries

Test expressions for all protein libraries were performed in a 96-well plate format with 2 mL capacity wells. Briefly, glycerol stock plates were used to seed overnight culture plates using a multichannel pipette with sterile pipette tips. Each well in the overnight culture plate was filled with 1.8 mL of LB-amp (or LB-kan for the pET28A-PTEN plasmids) before inoculation. Plates were grown overnight at 37 °C with shaking at 200 rpm. The following day, fresh plates

containing 1.8 mL LB-antibiotic per well were inoculated with 30 μL of overnight culture. Six replicate wells were inoculated for each construct. Cultures were grown until an optical density (OD 600) of 0.6 was achieved. Culture plates were cooled to the appropriate expression temperature (37 °C for lipA constructs, 25 °C for β-glucosidase constructs, and 13, 25, and 37 °C for PTEN constructs). Plates were induced by the addition 30 μL of an LB-antibiotic solution that contained IPTG. The final IPTG concentration in each well was 1mM. Plates were then incubated at the appropriate temperature overnight (16 hours) at 200 rpm.

The following day, plates were centrifuged at 1100 x g for 20 minutes with a plate-compatible swinging bucket rotor in order to pellet the cells in each well. LB supernatant was removed. Samples were made by combining pellets from two wells in 300 μL of BugBuster Master Mix, resuspending those pellets, and incubating them in BugBuster at room temperature for 20 minutes to allow for cell lysis. By combining 2 pellets, an initial 6 induction wells per construct was reduced to three replicates. Before lysis, protease inhibitor tablets were dissolved in the BugBuster at a concentration of one tablet per 50 mL, per manufacturer recommendation.

After lysis, total protein samples were taken for SDS-PAGE. 30 μL of crude, unclarified were added to 10 μL of 4x SDS loading dye, incubated at 95 °C for 10 minutes, and frozen at -20 °C pending characterization by SDS-PAGE. Crude cell lysate for each replicate of each construct was transferred to 1.5-mL Eppendorf the cell lysate was cleared by centrifugation in a table top centrifuge for 5 minutes at 13000 x g. 30 μL was taken from the soluble fraction and made into SDS-PAGE samples as described above for the total protein samples. After sampling the soluble fraction, all of the supernatant was removed and the pellets of cell debris and insoluble protein

was dissolved in 240 μL of 10 % SDS. Once dissolved, 30 μL samples of this insoluble fraction were taken and prepared for SDS-PAGE in the same manner as the total protein samples. Green fluorescent protein with an N-terminal polyhistidine-tag was included in all SDS-PAGE samples at a final concentration of 1.8 μM to serve as an internal reference for each sample when run on a gel.

6.2.4 Densitometry Measurements of Protein Library Expression

Samples from protein library test expressions were characterized via SDS-PAGE. 10 μL of each replicate for each sample was loaded on 12 % acrylamide gels and run at 150 V for 55 min. After running, gels were fixed in a 10 % acetic acid, 50 % ethanol fixing solution for 1 hour with gentle shaking. Gels were rinsed with ddH_2O and incubated in ddH_2O for 20 min with the water being replaced every 10 min to thoroughly remove fixing solution. Gels were then incubated in InVision™ stain solution containing 30 mM imidazole overnight at room temperature with gentle shaking. The following day, InVision™ stain was removed and gels were destained in 20 mM sodium phosphate buffer (pH 7.8) over the course of 24 hours with the buffer being refreshed at least three times.

Destained gels were imaged on a GE Typhoon FLA 9500 laser scanner with an excitation wavelength of 532 nm. Because InVision™ stain selectively stains proteins with polyhistidine-tags, only the library proteins and GFP standards were visible in spite of the fact that many cellular proteins were present in the samples. Densitometry measurements for each gel lane were made with ImageJ. Do to potential variation in gel loading and in the staining and destaining

process, all protein bands corresponding to constructs of interest were normalized to the GFP band in the same gel lane. Protein:GFP ratios were averaged for the three expression replicates obtained for each construct, and expression values were reported as fractions of WT protein.

Because point mutations made to LAP43-lipA resulted in changes in the actual codon composition of the gene, these samples were also normalized to total protein samples. Values for these constructs were therefore expressed as "fraction soluble relative to WT." Additionally, because PTEN constructs were expressed at three different temperatures, which led to different cell densities post-induction, PTEN samples were likewise normalized relative to total protein samples.

6.2.5 Library Persistence in Cell Lysate

Constructs from the LAP-lipA library that did not express in the soluble fraction when induced at 37 °C were expressed again at 13 °C. They were lysed with BugBuster as described above. Soluble fraction SDS-PAGE samples were taken for each (time zero) and the remaining soluble fraction was stored at room temperature for 4 hours. After storage, the 1.5-mL Eppendorf tubes containing the soluble fraction were again centrifuged for 5 min at 13000 x g, and the soluble fraction sample. Time zero and 4 hour sample were compared to establish protein persistence in the cell lysate at room temperature. No bands suggestive of protease degradation were observed for any of the constructs over this timeframe.

6.2.6 Characterization of LAP-LipA Constructs

Select LAP-lipA constructs were purified and characterized to determine the deferential impact of LAP insertions at diverse sites within the protein. Specific activity experiments were performed in 20 mM sodium phosphate buffer (pH 7.0) using p-nitrophenylbutyrate as a substrate at a final concentration of 3 mM. Specific activity values are based on activity measurements performed with at least three different concentrations of protein for each construct. Activity increased linearly with protein over the concentration ranges tested. Protein concentrations were determined via absorbance at 280 nm, using extinction coefficients calculated from protein primary sequences.

Circular dichroism spectra were obtained in a 1 mm cuvette at room temperature and consist of an average of at least 5 scans per protein. Composition percentages of alpha helices, beta sheets, and random coils were calculated from CD spectra as previously described. Single measurements at 222 nm were made for the various constructs at different concentrations of guanidinium chloride to generate denaturation curves.

6.2.7 Rosetta Installation and Modeling

All coarse-grained protein scans were performed on a MacBook Pro laptop computer on which Rosetta source code that contained binaries pre-built for Mac had been downloaded. Full-atom models were generated on the Janus supercomputer at the University of Colorado Boulder. Briefly, Rosetta source code (without binaries) was uploaded to the supercomputer and an MPI installation was performed. Input alignment files were generated on a laptop through the command line with downloaded clustal source code.

Protein scans were performed without specifying the -loops:refine refine_kic option, which is needed for full-atom modeling.[28] Additionally, build attempts were limited to 100 during an initial scan to eliminate sites that were too sterically hindered to allow for successful model generation. Mutations were generated by modifying the fasta inputs into the clustal alignments, which were threaded onto target PDBs.

Scores were obtained for isolated loops by creating new PDB files with the loop coordinates generated by the modeling. These isolated loop PDBs were cleaned with Rosetta's clean_pdb script and used as an input for a second round of modeling. However, in the second round of modeling, the .loop file contained two adjacent residues as the termini of the loop. As a consequence, no modeling occurred, but scores were generated for all of the residues in the loop that reflect the loop as an isolated entity. A similar approach was used to obtain background scores for loop-free versions of the target protein used. Namely, the entire protein was included in the modeling run, but the loop was defined as existing between two adjacent residues. As a consequence, no modeling took place, but all of the residues in the protein (and the protein as a whole) were assigned scores.

6.2.8 Rosetta Data Analysis

Total scores from centroid models produced per insertion site were averaged to generate a single characteristic score per site. 95 % confidence intervals were calculated from standard deviations. Delta Rosetta scores were calculated by subtracting total scores calculated from proteins with no loops from total scores generated from LAP-protein models. Individual terms from the score function are based on averages for those terms obtained from the 10 models

produced per site. Individual residue scores were obtained by average the scores for that particular residue over the 10 models produced.

6.2 Results and Discussion

6.2.1 Impact of LAP Insertion on Protein Target Properties

Inserting the LAP sequence at internal sites within the structure of a target protein typically has a severely negative impact on the properties of that protein. Constructing a library of lipA constructs with LAP sequences genetically inserted at 36 unique, solvent exposed sites at internal positions within the protein's primary sequence, we found that most constructs did not appear in the soluble fraction when expressed in *E. coli* during an overnight induction at 37 °C. Rather, most of the constructs were found in the insoluble fraction, suggesting that inserting the LAP sequence at most sites within the structure of a protein has a substantial impact on protein folding, stability, or both. Expressing this same library at 13 °C resulted in observed soluble expression for all but six of the constructs. However, those that were not soluble when expressed at 37 °C did not persist in cell lysate at room temperature. For these constructs, the amount of protein in the soluble fraction decreased by over 80% after a four-hour incubation. The presence of protease inhibitors and the absence of degradation products observed via SDS-PAGE suggest that this behavior was due to a loss in protein stability resulting from the addition of the LAP sequence (**Figure 6.1a-c**).

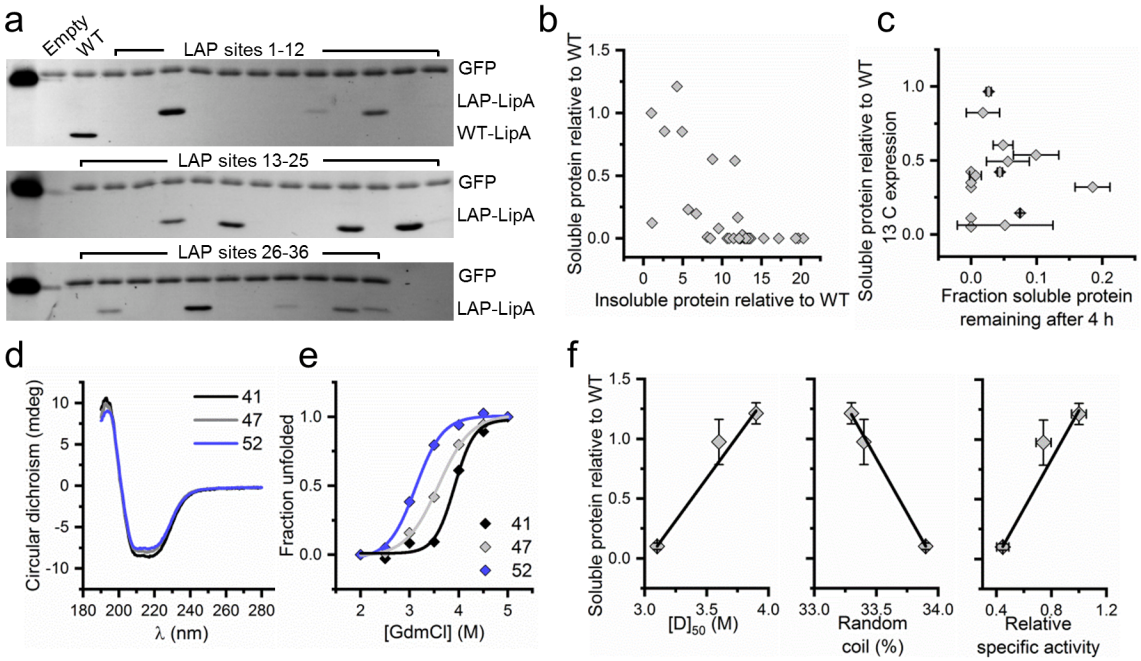

Figure 6.1. Impact of LAP insertions at diverse sites on target protein properties. Representative gels of LAP-lipA expression in the soluble fraction at 37 °C (a). Samples were spiked with a known concentration of GFP as an internal loading and staining control. Correlation between soluble and insoluble protein expression for all LAP-lipA constructs at 37 °C (b). Proteins that do not express in the soluble fraction appear in the insoluble fraction. Soluble expression at 13 °C of constructs that fail to express in the soluble fraction at 37 °C versus persistence in cell lysate after 4 hours at room temperature (c). CD spectra of 3 purified constructs with LAP inserts at sites 41, 47, and 52 (d). Chemical denaturation of purified constructs (e). Correlation of soluble protein expression at 37 °C with stability toward denaturant, secondary structure, and specific activity (f).

Of the constructs that were found in the soluble fraction after induction at 37 °C, the extent of soluble protein was highly variable. Three of these constructs, with LAP insertions at sites 41, 47, and 52 respectively, were purified and characterized (**Figure 6.1d-f**). (Site 41 refers to a construct with a LAP insertion between residue 41 and residue 42. This convention of naming insertion sites after the adjacent protein residue immediately upstream of the inserted LAP sequence is maintained throughout.) Soluble expression of these constructs was found to correlate with stability toward guanidinium chloride denaturation, demonstrating that LAP

insertions can impact the stability of the target protein in which they are inserted to varying degrees depending on the particular insertion site. Further, this data suggests that changes in stability upon LAP insertion contribute to the variable levels of soluble protein expression observed for the LAP-lipA protein library.

Interestingly, in addition to correlating with stability, soluble expression of these constructs inversely correlated with secondary structure. Specifically, the site 52 construct, which expressed more poorly in the soluble fraction than constructs 41 and 47, had the largest percentage of random coil structure based on CD spectra analysis. These data suggest that LAP insertions can perturb the structure of a target protein and that the extent of perturbation may depend upon the specific location of the insertion site. Additionally, they suggest that structural perturbations may also affect the variations in soluble expression seen with the LAP-lipA library. These observations are consistent with specific activity data obtained for the three constructs, which was found to correlate with soluble protein expression and to inversely correlate with random coil percentages.

The failure of so many LAP-containing constructs illustrates the need for predictive capability in identifying accommodating insertion sites. Additionally, the above data demonstrate the utility of treating soluble protein measurements as a metric for characterizing the permissiveness of particular LAP insertion sites. Constructs that were well expressed in the soluble fraction were generally more stable and better maintained their structural integrity as well as their catalytic activity.

6.2.2 Basis for Rosetta Modeling Approach

To establish a highly accessible computational platform for predicting permissive LAP insertion sites, the Rosetta kinematic loop modeling application was adapted for rapid structural screening of proteins.[28] With this approach, the LAP sequence is modeled at every possible site (between every two adjacent residues) within a protein target. The full-atom stage of the protocol is omitted, and as a consequence, Rosetta's coarse energy function is used to produce energy scores for each model. Ten models are generated at every insertion site for which modeling is possible and the total energy scores of these 10 models are averaged to achieve one unique and reproducible score per site. Lower, more negative scores indicate that a site is permissive toward LAP insertion whereas higher, more positive scores indicate that the site is unaccommodating.

The kinematic loop modeling application is typically used to model short structured loop regions within proteins. It does so by proceeding through a coarse-grained centroid modeling stage followed by a full-atom stage. Using this protocol, we were able to remodel a 13-residue fragment of lipA (residues 7-19), observing convergence of low energy models upon the correct structure with the generation of 3000 full-atom models (**Figure 6.2a-b**). Adapting kinematic loop modeling for the identification of permissive LAP insertion sites, however, required a number of operational changes relative to this standard protocol. These were made in part due to the possibility that the LAP sequence may not be structured in solution. As opposed to the results obtained when modeling residues 7-19, producing 3000 full-atom models of the LAP sequence inserted at site 13 in lipA did not result in convergence on any one conformation at low energies (**Figure 6.2c-d**). Rather, the distribution of models more closely resembled that observed for a 13-residue glycine loop inserted at the same site (**Figure 6.2e-f**), which would not be expected to

be structured within an actual protein in solution or to converge upon a single structure when modeled.

The observation that LAP insertions at different sites perturb protein secondary structure and affect specific activity further motivated the pursuit of operational changes to the kinematic loop modeling protocol. If structural changes occurring within the protein target but outside of the LAP sequence lead to these effects, they would not likely be captured by Rosetta's loop modeling protocols, which primarily models residues within the loop itself. The final motivation behind making operational changes was the amount of computing power that would be required to produce thousands of full-atom models at every possible insertion site within the structure of a target protein. Although possible, such a structural scan would require supercomputer access and would therefore limit its practical application by diverse users.

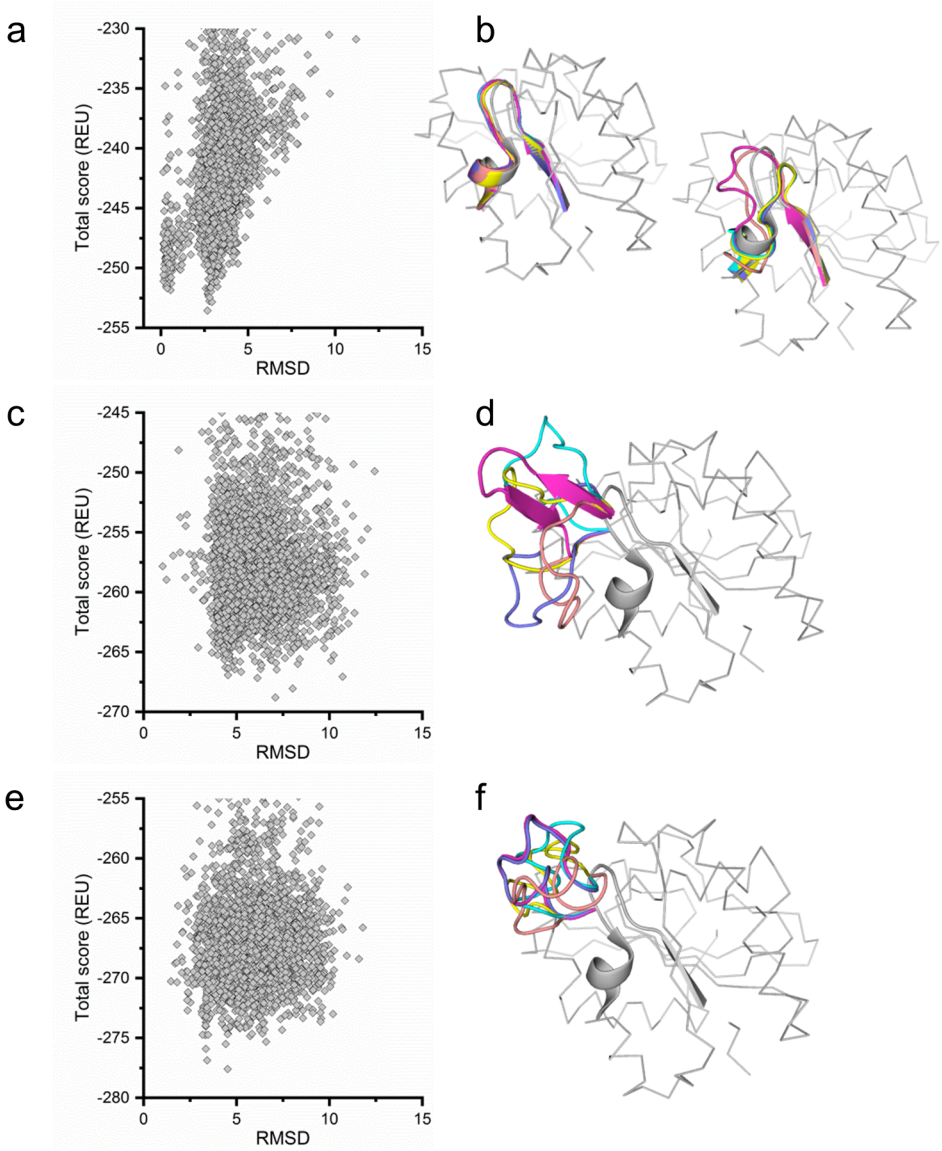

Figure 6.2. Full-atom modeling of a structured loop (residues 7-19) in lipA (a), a LAP insertion at site 13 (c), and a 13-residue glycine loop insertion at site 13 (e).
Producing 3000 full-atom models of residues 7-19 results in convergence upon two conformations at low energy. An overlay of the 5 lowest energy structures from each convergence point (b) reveals a strong degree of overlap with the crystal structure for models at low RMSD. No convergence is observed for an inserted LAP sequence at site 13 and the five lowest energy structures take on very diverse conformations (d). The behavior of the LAP sequence when modeled is similar to that of a 13-residue glycine loop at site 13 for which the 5 lowest energy structures also display diverse conformations (f).

The first change that was made to address the above concerns was the elimination of the full-atom modeling step in the kinematic loop modeling protocol. The second was the characterization of each site through the averaging of scores from a small sample of structures rather than through the score achieved by the single lowest energy structure obtained from producing thousands of models. By eliminating full-atom modeling, the protocol is restricted to the centroid stage, which does not model each atom explicitly. This reduces computational time and removes elements from the total score that may not be relevant to a highly dynamic loop. Taking an average of coarse energy scores for a small sample of models per site also dramatically reduces computational time and is appropriate relative to finding a minimum given that LAP insertions are potentially unstructured. We found for a number of LAP insertion sites in lipA that the average of a sample of 10 centroid models reproduces the average obtained from a much larger population (2500 centroid models) within a calculated 95% confidence interval. We also found that the standard deviation of the sample, which is used to calculate confidence interval, is a good approximation of the standard deviation of the population (**Table 6.1**). These results indicate that a small number of centroid models can be used to characterize a single insertion site.

Eliminating the full-atom stage of the kinematic loop modeling protocol and producing only 10 models per site does not lend itself to the generation of a single model that accurately predicts the structure of a LAP-target protein. Rather, in taking this approach, the coarse energy function is used to infer the impact of a LAP insertion on the structure and/or stability of a protein without explicitly modeling it. A more positive score for a particular site suggest that LAP insertion at that site may have more of a detrimental impact on the target protein than

insertion at a site with a lower score. This type of approach makes scanning an entire protein structure for insertion sites extremely rapid. Producing 10 centroid models takes approximately 30 minutes, which means that 50-200 unique sites can be characterized on a conventional laptop within a single day depending on the number of cores that are used, and the entire structure of a reasonably sized protein can be scanned on the order of days.

Number of models	3000 Full atom	2500 Centroid		10 Centroid		
	RMSD spread	Avg	Stdev	Avg	Stdev	95 % CI
Residues 7-19	0.261	--	--	--	--	--
Glycines site 13	3.235	--	--	--	--	--
LAP site 13	6.173	-66	2	-66	2	1
LAP site 53	4.179	-50	2	-51	2	1
LAP site 108	1.167	-61	3	-62	3	2

Table 6.1. Characterization of full-atom and centroid model populations.
3000 full-atom models were generated to remodel residues 7-19 in lipA, a 13-residue glycine loop inserted at position 13, and a LAP sequence inserted as a loop at sites 13, 53, and 108. The RMSD spread is recorded for the 5 lowest energy models at each site and suggests convergence for residues 7-19 but not for any of the glycine loop or LAP insertions. Average energy scores obtained from 2500 centroid models are accurately reproduced by averaging the scores from a sample of only 10 models within the calculated 95% confidence interval (CI), with the standard deviation of the sample closely approximating the standard deviation of the larger population.

We initially used our approach to scan the entire structure of lipA (**Figure 6.3**). While doing so, we found that modeling failed at sites that were totally buried within the protein. This was due to the fact that every possible conformation of the LAP sequence at these sites resulted in the overlap of loop residues with structural residues. To prevent Rosetta from searching through loop conformations at these sites, the number of build attempts per insertion site was limited to 100 during an initial scan of the structure in which only a single model was generated

per site. Only sites that produced a successful model within 100 build attempts were revisited to complete the additional 9 models needed to achieve a characteristic site score.

Modeling was successful and scores were obtained for LAP insertions at 86 solvent exposed sites in lipA. For these sites, a wide distribution of scores was achieved relative to the average 95% confidence interval, demonstrating that this approach is competent at effectively and reproducibly differentiating between large numbers of potential insertion sites. Mapping the scores for each site onto the structure of lipA revealed unintuitive predictions. Specifically, some sites within loop regions of the protein were predicted to be highly unaccommodating whilst some sites that interrupted helices were predicted to be relatively permissive. Additionally, some nearly adjacent sites within lipA were observed to have radically different scores, suggesting that our approach might have high resolution in distinguishing between neighboring insertion sites. At the same time, a crude trend was observed between sites near the active site (at the top of the image in **Figure 6.3c**) and sites that are further from it (at the bottom of image).

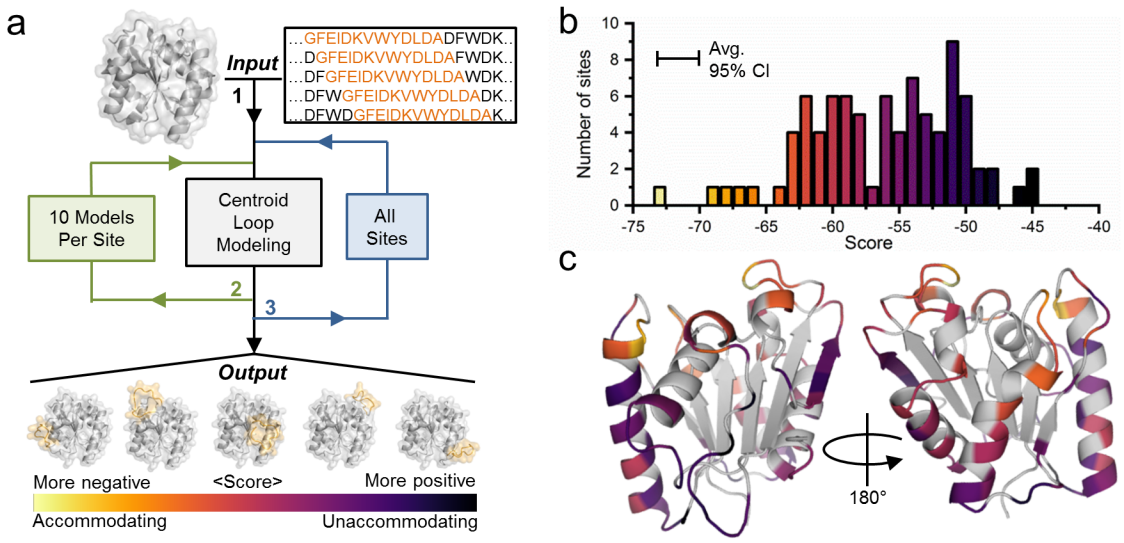

Figure 6.3. Structural scan of all possible LAP insertion sites in lipA.
Modeling produces a score for every solvent exposed LAP insertion site in the structure of lipA by averaging scores from 10 models per site (a). A color scale is used to represent relative scores for each site in the protein. Yellow indicates a more negative score and a more accommodating site. Black indicates a more positive score and a less accommodating site. Grey indicates buried sites for which modeling could not be performed. Score distribution for all of the successful modeling sites in lipA (b). The distribution of scores for the various sites is much larger than the average 95% CI, indicating that the approach can be used to distinguish between different sites. Heat map of LAP insertion site scores within the structure of lipA (c).

6.2.3 Model Validation with LAP-Protein Libraries

Plotting soluble protein expression for all of the lipA constructs in our library versus scores obtained for each construct's particular LAP insertion site revealed a striking correlation (**Figure 6.4a**). Constructs with lower scores were more likely to express in the soluble fraction and to do so to a higher degree. Specifically, 7 of the 9 constructs with the most negative scores exhibited a measurable amount of soluble protein expression with 5 of them exceeding 50% of WT lipA expression levels. In contrast, 20 of the 26 constructs with the most positive scores

produced no measurable protein in the soluble fraction and none of them expressed at levels above 50% of WT.

To place the correlation between soluble expression and Rosetta scores in context with other structural metrics that might be used to predict the permissiveness of a loop insertion site, soluble protein expression was plotted as a function of crystallographic *B*-factor. *B*-factors for each site were determined by averaging Cα atom *B*-factors for the two protein residues between which the LAP sequence was inserted. The expectation was that sites with higher *B*-factors would be more permissive to loop insertions since correlations have previously been observed between the *B*-factor of a residue and its mutational plasticity. Interesting, Rosetta scores were found to be more predictive than local *B*-factor measurements, and soluble protein was not found to increase with increasing B-factor. Rather, those constructs with the greatest levels of soluble expression had a slight tendency toward lower *B*-factors.

In order to identify any regional or structural dependencies of the LAP insertion sites on accommodation, the extent of soluble expression and the Rosetta scores for each construct were plotted as a function of insertion site location within the primary sequence of lipA. Doing so revealed scores to be especially predictive locally within different regions of the protein. For example, Rosetta scores successfully differentiate between sites 41 and 43, which have very different soluble expression levels in spite of the fact that they are nearly adjacent.

The correlation between score and soluble expression within local regions of primary sequence became more striking when the secondary structure of these regions was considered. Rosetta scores predicted LAP accommodation at sites that were unintuitive given their secondary structure. For example, sites 17, 47-49, and 108 were predicted and found to be relatively

permissive in spite of the fact that they interrupt alpha helices, and sites 67, 116, 119, and 121 were predicted and found to be unaccommodating though they occur in highly exposed loop regions. Lastly, ordering expression and modeling data for each construct by Lap insertion location revealed that many of the false positives (sites with good scores but poor expression) were confined to a single region of the protein (between residues 133 and 139). This suggests that model scores may be more predictive when applied to different sites locally within a particular region of a protein than when used to identify the single best site relative to all possible insertion sites. Using Rosetta scores in this way is beneficial when regions of the protein are found to be intolerant to insertions due to factors that are not encompassed by this modeling approach, as seems to be the case with this particular region of lipA.

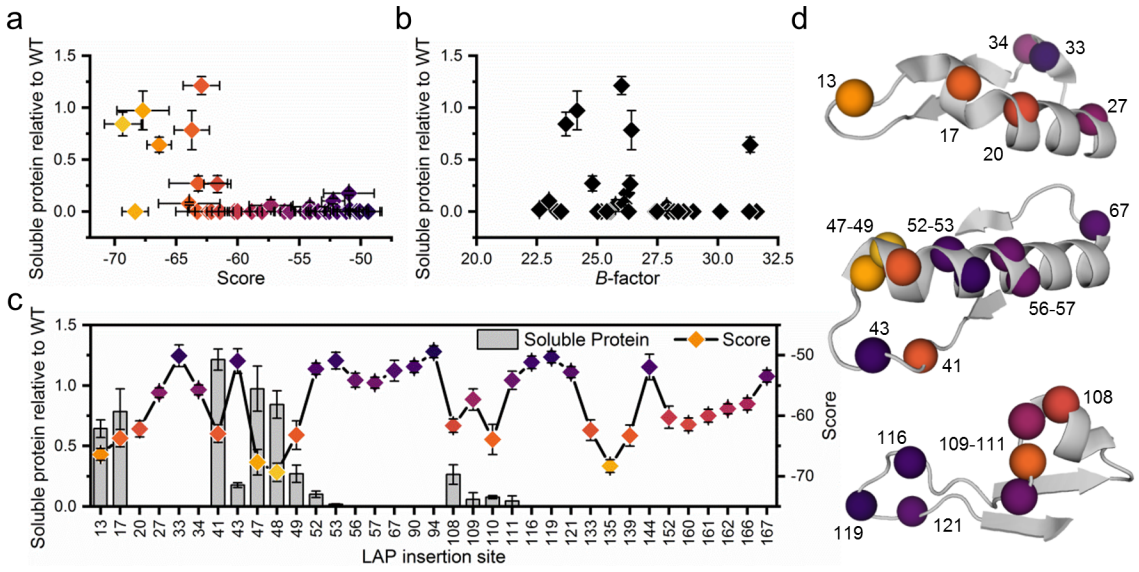

Figure 6.4. Correlation of soluble protein expression for the LAP-lipA construct library with Rosetta scores and B-factor and LAP insertion site with respect to primary sequence and secondary structure.
Soluble protein expression vs. Rosetta score (a). Soluble protein expression vs. *B*-factor of LAP insertion sites (b). Soluble protein expression and score for each site ordered by lipA primary sequence (c). Location of select LAP insertion sites within lipA (d).

In order to determine if the observed trends between soluble protein expression and Rosetta scores were unique to lipA or if a similar predictive capability could reasonably be expected for other proteins, a second model protein, β-glucosidase, was computationally scanned for accommodating LAP insertion sites, and select constructs were cloned and experimentally characterized (**Figure 6.5**). β-glucosidase is larger and less stable than lipA and has difficulty folding when expressed in *E. coli*. Consequently, only a single LAP-β-glucosidase construct was found to express in the soluble fraction at levels greater than 25% of WT β-glucosidase expression. Nevertheless, in spite of these lower overall expression levels, the correlation observed between soluble protein and Rosetta scores for LAP-β-glucosidase constructs was qualitatively similar to that observed for the LAP-lipA protein library. Specifically, the five best expressing constructs, which appeared in the soluble fraction at between 9 and 40 % of WT levels, had LAP insertions at sites with scores within the lowest quintile of all scored sites. Of the constructs with LAP sites scoring in one of the more positive scoring (and therefore expected to be less permissive) quintiles, 16 out of 20 experimentally characterized constructs had no measurable expression in the soluble fraction. Of the four constructs for which soluble expression was observed, none expressed at levels greater than 5 % of WT expression.

As was the case with the LAP-lipA protein library, the correlation between soluble expression and Rosetta scores for LAP-β-glucosidase constructs compared favorable to B-factor correlations. Additionally, although the library coverage of insertion sites was less complete with β-glucosidase than it was for lipA, regions of primary sequence with greater sampling density again demonstrated that Rosetta scores are especially predictive locally within different regions of primary sequence. Further, in spite of the fact that most of the relatively permissive

LAP insertion sites in β-glucosidase were found in loop regions, site 175, which partially interrupts a helix, was found to be relatively accommodating, again suggesting that permissive sites within this protein may be unintuitive with respect to secondary structure, as was found to be the case with lipA.

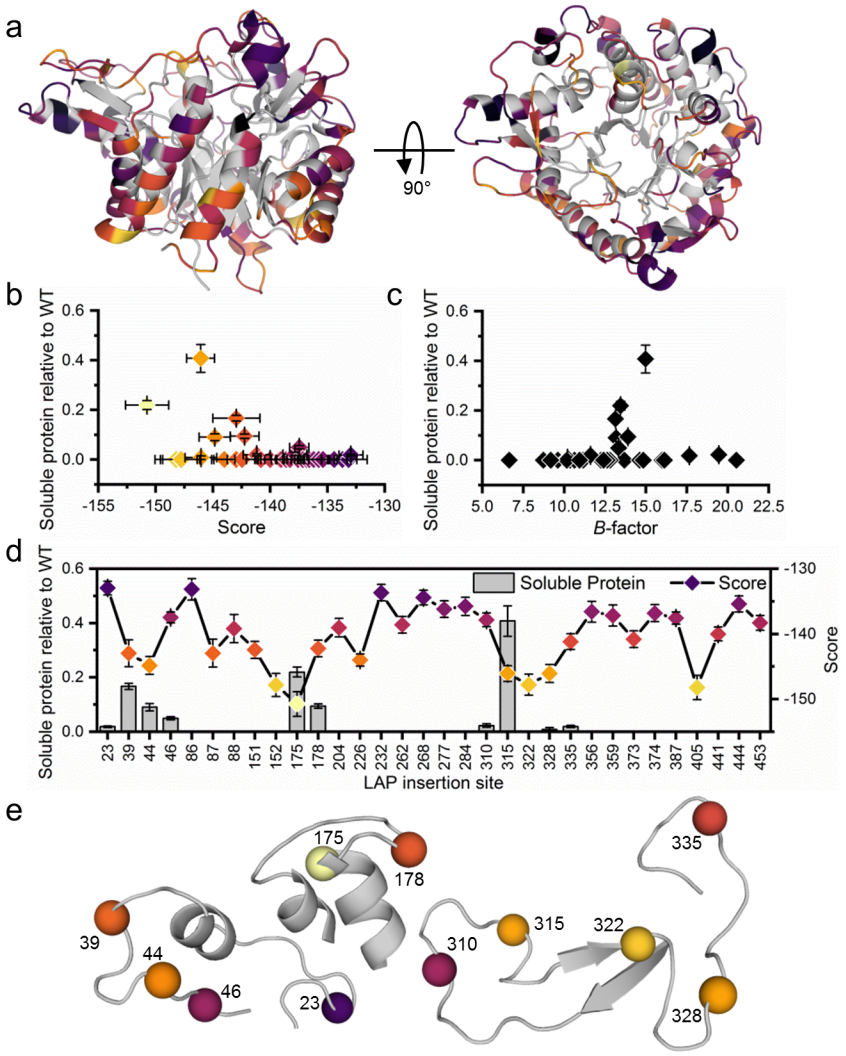

Figure 6.5. β-glucosidase full structural scan of potential LAP insertion sites and correlation between soluble protein expression and score for select constructs.
Heat map of scores for every possible solvent exposed LAP insertion site (a). Soluble expression vs. score for select insertion sites (b). Soluble expression vs. *B*-factor of insertion site (c). Soluble protein expression and score for each site in order of location in the primary sequence of β-glucosidase (d). Location of successful LAP insertion sites in β-glucosidase (e).

6.2.4 Structural Determinants of Model Predictions

Individual score terms within the Rosetta centroid energy function were compared with the total energy score output to identify score components that contribute most strongly to differentiating between the different LAP insertion sites in lipA. Doing so revealed that the ability to differentiate between sites in LipA is dominated by two score terms, env and pair (**Figure 6.6**). The env term is the only one that correlates strongly with total scores for all of the LAP sites in lipA. Because all of these terms are summed to generate the total score, we looked at which other terms, when added to the env term, to see if any would substantially improve this correlation (**Table 6.2**). The pair term was the only one that, when added to the env term, substantially improved the correlation between terms and total score (R^2 increased from 0.78 to 0.97). Adding a third term does not substantially improve the correlation, suggesting that these two terms dominate the predictive capability of the total score in successfully differentiating between insertion sites in lipA.

The env and pair terms, which are generated for every residue and summed over the entire structure, capture the hydrophobic and electrostatic environment of the residue for which they are generated. This environment is dependent on the number and identify of neighboring resides within 10-12 Å as defined by the score terms themselves. Therefore, the fact that these two terms dominate the total score and, consequently, the predictive capability of this approach suggests that residues outside of the loop are playing a role in determining whether a particular site is or is not permissive to loop insertion. To test the extent to which residues outside of the inserted loop contribute to env and pair terms, loops from all of the models were removed from the rest of the protein and, env/pair terms were recalculated for these loops, which retained their

conformations, in isolation from the rest of the protein (**Figure 6.7a**). No correlation was observed between env/pair terms for isolated loops and the total scores from the intact models, indicating that the residue environment of the loop insertion, as opposed to the allowed conformations of the loop at that site, dominate env/pair scores and therefore are the dominant contributing factor to the predictive capability of this approach.

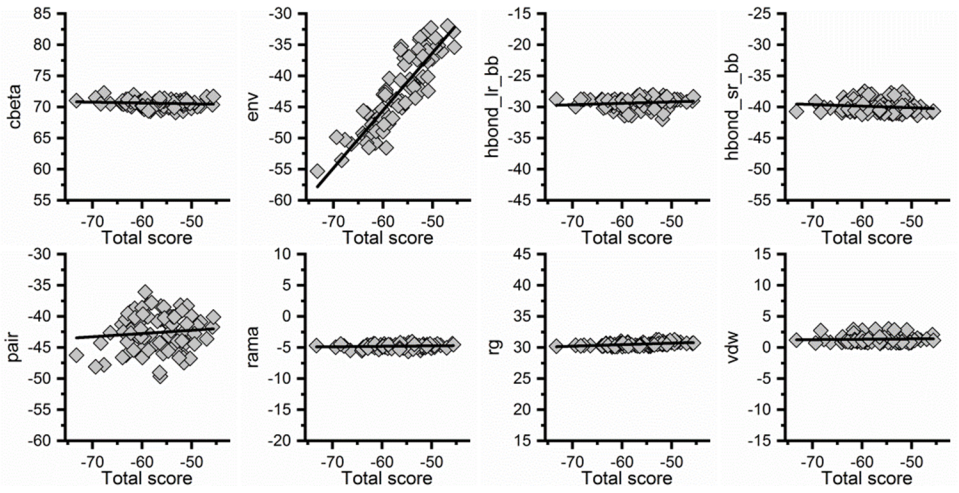

Figure 6.6. Individual terms within the centroid score function vs. total score for all modeled LAP insertion sites in lipA.

	R^2 from linear fit							
Term	cbeta	env	hbond_lr_bb	hbond_sr_bb	pair	rama	rg	vdw
Term vs. total score	0.02	**0.78**	0.02	0.02	0.01	0.02	0.28	0.00
Term+env vs. total score	0.76	--	0.77	0.74	**0.97**	0.78	0.78	0.79
Term+env+pair vs. total score	0.97	--	0.96	0.94	--	0.97	0.96	**0.98**

Table 6.2. R^2 values for linear correlations for each term and combined terms in the centroid score function vs. total score.

To further explore the impact that residues in the near-loop environment have on determining total scores and, consequently, on predicting LAP accommodation, we examined the residue environment of LAP site 43 in lipA. Rosetta correctly predicted this site to be poorly accommodating. It falls between sites 41 and 47, which were themselves predicted to be and found to be permissive. The close proximity of these sites and the fact that energy scores are dominated by residues in the near loop environment suggested that only a small number of residues played a role in the differing levels of accommodation observed for these sites. Plotting the contribution that each local residue makes to the total energy score illustrates that LAP insertion at site 43 creates a particularly unfavorable environment for a single amino acid, K44, which is directly adjacent to the inserted loop in terms of primary sequence (**Figure 6.7b-c**). Other residues in the near loop environment are also destabilized by the loop but to a lesser degree. These include residue D43, which directly flanks the loop on the opposite side. Interestingly, it also includes residue T46, which is not directly adjacent to the loop in terms of primary sequence but is brought into close proximity by a turn in the structure.

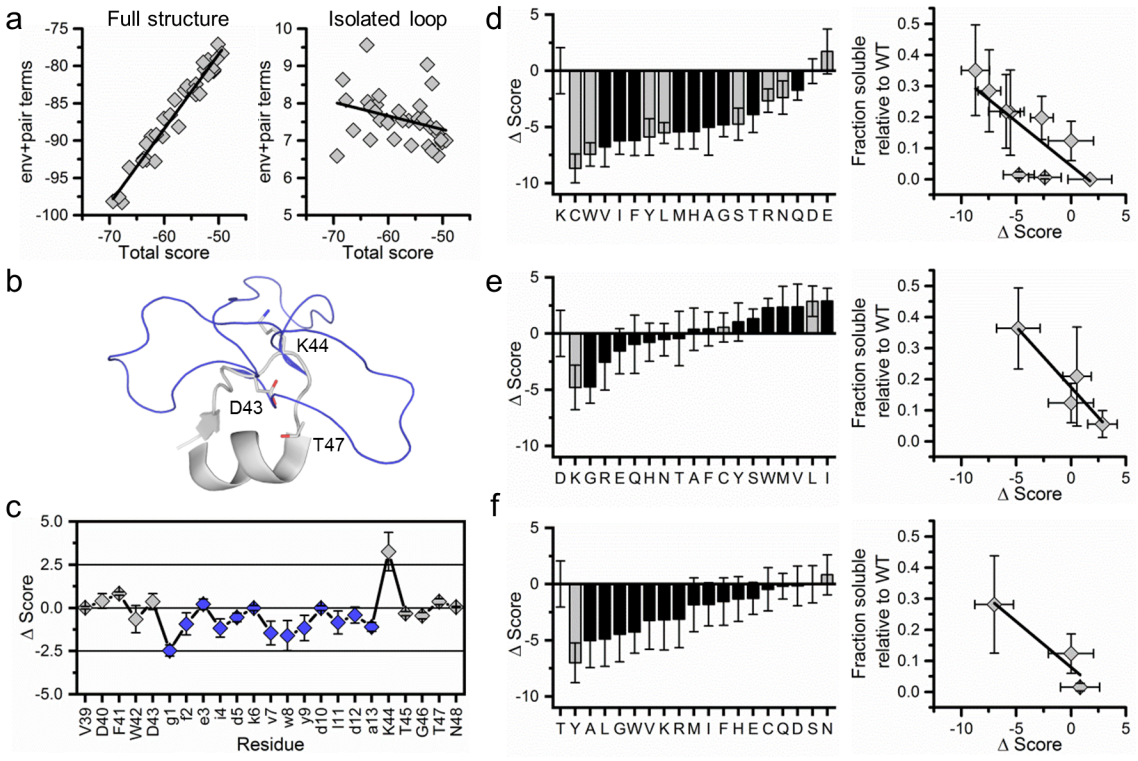

Figure 6.7. Impact of the local residue environment on LAP accommodation.
Correlation between the sum of the env and pair terms with total scores for all experimentally produced constructs (a). Results are show for LAP sequences within the context of the protein (left) as well as in isolation (right). Overlay of three individual models of the LAP sequence inserted at site 43 showing notable residues (D43, K44, and T47) in the near-loop environment (b). Score profile for residues in the near-loop environment as well as in the LAP sequence based on averages of 10 models (c). *In silico* saturation mutagenesis of K44 with the LAP sequence at site 43 (left) and correlation of soluble protein expression with scores (right) for mutants highlighted in grey (d). Saturation in silico mutagenesis (left) of residue D43 (e) and T47 (f) with the LAP sequence at site 43 and correlation of soluble expression with select constructs (right).

A strong dependence on the residue environment suggests that LAP accommodation can be modulated by site directed mutagenesis of local residues. To determine the responsiveness of site 43 energy scores to changes in the near residue environment, *in silico* saturation mutagenesis was performed for residues K44, D43, and T46 (**Figure 6.7d-f**). The coarse energy function is not meant for evaluating the impact of single point mutations on a protein, and therefore highly penalizes the mutation of single solvent exposed surface residues from polar to non-polar

functional groups. To prevent this penalization and to focus on the impact of a mutation on loop accommodation, Rosetta scores for each mutation were normalized to scores calculated for structures that contain the mutation but lacked the LAP insertion. Delta scores obtained in this way show the impact of each mutation on LAP accommodation rather than on the protein itself in the absence of the loop. Several mutations for each residue were chosen, made, and experimentally characterized. Each construct contained the LAP sequence inserted at site 43 with a single point mutation that was either predicted to improve or reduce accommodation (and therefore soluble expression of the construct) for the LAP sequence at that site. Soluble expression relative to total lipase expression was plotted as a function of normalized Rosetta score, and a correlation was observed for between soluble expression and score for the various mutants for each of the targeted residues. These observations further demonstrate the importance of the near loop environment in determining LAP accommodation and the important contributing role that residues in the near loop environment play with regard to the observed predictive capability of this approach. Interestingly as can be seen with T46 mutations, this near loop environment extends beyond residues that are directly adjacent to the loop.

General trends were observed in the chemical characteristics of mutations that were favored or disfavored for each residue in terms of Rosetta score. *In silico* saturation mutagenesis for residue K44 revealed a preference for hydrophobic residues at this position, which tended to have a stabilizing impact on the loop. Polar residues tended to be less stabilizing. The WT residue, lysine, was predicted to be the second worst possible residue to occupy this position, and glutamic acid was predicted to be the worst. With a positive delta score, it was predicted to destabilize the protein. Indeed, the K44E construct was found to be worse than the WT LAP-43

construct; it did not express at all in the soluble fraction. Of the hydrophobic mutations that were experimentally characterized (K44W, K44Y, K44L), all were found to have a slight if statistically insignificant improvement over WT LAP-43 expression and to contribute to the general trend observed between soluble expression and delta Rosetta scores. The mutation that was predicted to be the best at this residue position and was found to result in the highest levels of soluble expression was K44C. Though not typically considered as hydrophobic, cysteine is more hydrophobic than serine, which is similar in side chain architecture but more polar due to the electronegativity of oxygen and which as a K44 mutant did not improve soluble expression of LAP-43. Because there are no other cysteine residues in lipA or in the LAP sequence, it is unlikely that the favorable Rosetta score for this mutation arises from an artifact in the energy function derived from encouraging disulfide formation in models but rather that the mutation improved the residue environment of the loop in a constructive way.

Interestingly, trends were also observed between the chemical characteristics of mutations and delta Rosetta scores for *in silico* saturation mutagenesis at positions D43 and T47. The trend at position T47 was similar to that at K44, with hydrophobic residues tending to be preferred. However, the trend at D43 was the opposite. Rosetta scored polar residues more favorably with the D43K mutation receiving the best score and hydrophobic residues such as leucine and isoleucine receiving positive scores relative to WT LAP-43, suggesting that these mutations would actually reduce soluble expression. Again, delta Rosetta scores were found to be predictive. D43K had higher expression in the soluble fraction than WT LAP-43 and D43L had less soluble expression.

6.2.5 Identification of Permissive LAP Sites in PTEN

The ability of Rosetta to predict accommodating sites especially locally within different regions of a protein suggests a strategy for substantially reducing the number of constructs that need to be built and tested to identify permissive LAP insertion sites. Specifically, by scanning a protein structure and choosing sites with the most negative local scores within a number of diverse regions of the protein, the probability of identifying an accommodating LAP insertion site within a small sample of experimentally produced and characterized constructs is maximized.

Such an approach was used to identify accommodating LAP insertion sites in PTEN. PTEN is a phosphatase and a tumor suppressor protein that is active in both the cytoplasm and the nucleus of cells.[18] Trafficking experiments for this protein in live cells are of interest due to the fact that proper localization of PTEN is critical for its function. Specifically, it must bind to the cell membrane where it interacts with a membrane associated protein complex and dephosphorylates membrane lipids. Although trafficking experiments have been performed with PTEN-GFP fusion constructs,[19] a labeling approach that avoids modification of the protein's termini would represent an important advance due to the physiological importance of the N- and C- terminus of PTEN for localization. Specifically, the membrane-binding domain on the N-terminus of PTEN is required for association with the membrane and the PDZ-binding domain on the C-terminus is required for forming a complex with other membrane-associated proteins. For this reason, a LAP insertion site within the structure of the protein that did not negatively impact protein properties would be useful as it could serve as a handle for site-specific fluorophore conjugation.

In order to identify accommodating LAP insertion sites within the structure of PTEN, the structure was scanned with our Rosetta-based approach and five potential insertion sites within different regions of the protein with Rosetta scores representing local minima were selected (**Figure 6.8a**). These sites avoid residues within the protein that are post-translationally modified within the cell including the unstructured flexible loop between residues 281 and 313 that is ubiquitinated in order to facilitate transport into the nucleus. Interestingly, when the scores of these sites are compared to all other sites within PTEN, lipA, and β-glucosidase, they are found to be particularly accommodating. For example, 4 of the 5 PTEN sites chosen for expression have lower scores than all of the modeled insertion sites in β-glucosidase and score better than the vast majority of sites in lipA (**Figure 6.8b**). These results suggest that LAP mediated modification may be a particularly appropriate bioconjugation approach for PTEN given the identity and configuration of its surface residues.

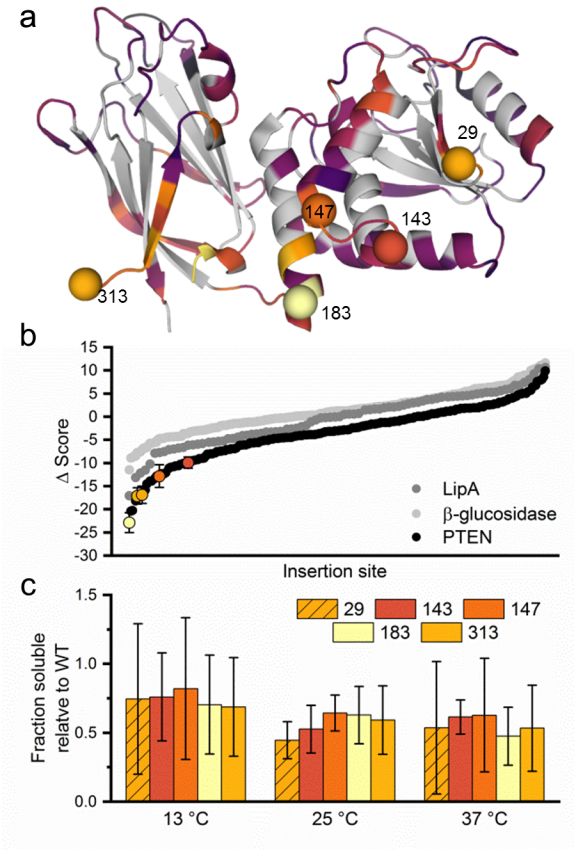

Figure 6.8. Prediction and expression of PTEN constructs that accommodate LAP insertions at internal (non-terminal) positions.
Heat map of scores for every possible solvent exposed LAP insertion site in PTEN with sites chosen for expression highlighted with spheres (a). Scores for all LAP insertion sites modeled in LipA, B-glucosidase, and PTEN evenly distributed on the x-axis (b). Fraction of protein soluble relative to WT for all of the LAP-containing constructs when expressed at 13, 25, and 37 °C (c).

LAP-PTEN constructs were expressed at 13, 25, and 37 °C. All five of the constructs were well expressed in the soluble fraction relative to WT at each of these temperatures. Even at the highest expression temperature, none of the constructs are completely insoluble and 4 of the 5 express at levels greater than 50% of WT soluble expression levels. A slight increase in relative soluble expression is observed at lower temperatures, suggesting that the LAP sequence is having an impact on the stability of the constructs, which is less apparent at lower temperatures.

The success of these constructs relative to most of the LAP-containing constructs within the lipA and β-glucosidase libraries, which mostly produced insoluble protein constructs, suggests that a Rosetta guided approach can successfully reduce the experimental burden associated with identifying accommodating LAP insertion sites within proteins and that doing so can help identify potential internal labeling sites for proteins with physiologically active termini.

6.3 Conclusion

Leveraging the coarse energy function within the kinematic loop modeling application of Rosetta to rank loop insertion sites by taking the average scores from 10 models produced per site represents a highly user accessible and predictive approach for identifying accommodating LAP insertion sites in proteins. The user accessibility stems from the relatively small amount of computational power that a total protein structural scan requires, which makes this approach feasible for users who only have access to a laptop computer. The predictive capability of this approach stems from the apparent impact of the residues in the near-loop environment on loop accommodation. Rosetta scores will be predictive to the extent that the residue environment determines accommodation, and in the case of an inserted LAP sequence, this dependence seems to be significant.

The observation that residues in the near loop environment dominate the predictive capability of this approach was somewhat of a surprise and suggests that the residue environment of a loop insertion site plays an underappreciated role in determining loop accommodation relative to factors such as *B*-factor and secondary structure. Notably, constraints imposed by the architecture of the insertion site did not impact the torsion angles or sterics of residues in the

loop in a way that led to differentiation between sites by Rosetta scores, though modeling failed outright at sites that were buried. Also, the residue environment created by the particular conformation of the loop at a particular site did not have a large impact on site differentiation, as observed by scoring loop conformations in isolation from the protein.

The dependence of loop accommodation on residues in the near-loop environment suggests that mutating those residues can modulate the permissiveness of a particular site. This was indeed found to be the case. The types of residue substitutions that led to enhanced or reduced accommodation were found to be site-dependent and difficult to predict by intuition alone. In some cases, polar residue substitutions were preferred whereas in other cases hydrophobic residues were. In all cases, however, Rosetta scores from *in silico* saturation mutagenesis of residues in the near-loop environment correlated with changes in soluble protein expression for those mutants, suggesting that our approach can be used to redesign insertion sites in order to alter loop accommodation. Substantially redesigning an insertion site with this approach by making multiple mutations to the near-loop environment, however, is not practical because the centroid scoring system used only reflects the impact of a point mutation on loop accommodation and not its impact on the folding or stability of the protein more generally.

As was demonstrated with PTEN, the predictive capability achieved with this approach is well suited for reducing the number of protein constructs that need to be made and experimentally characterized in order to identify sites within a protein that will accommodate a LAP insertion. Although Rosetta does produce false positives, choosing sites with local score minima within different regions of the protein can increase the likelihood of finding an accommodating site. Additionally, Rosetta seems to be particularly useful at ruling out sites, as

false negatives were rare especially when considered with respect to other local sites with better scores. Additionally, like the impact of individual point mutations on accommodation, the permissiveness of any particular site is unintuitive and difficult to predict in the absence of this approach. Together, these observations suggest that a Rosetta-enabled approach for predicting loop accommodation can serve as a useful tool in facilitating LAP-mediated bioconjugation and loop insertions in proteins more generally.

BIBLIOGRAPHY

1. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2008).

2. Nelson, D. L. & Cox, M. M. *Lehninger Principles of BIochemistry*. (W. H. Freeman and Company, 2008).

3. Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. (W. H. Freeman and Company, 1999).

4. Leader, B., Baca, Q. J. & Golan, D. E. Protein therapeutics: a summary and pharmacological classification. *Nat. Rev. Drug Discov.* **7,** 21 (2008).

5. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485,** 185 (2012).

6. Gorton, L., Bremle, G., Csöregi, E., Jönsson-Pettersson, G. & Persson, B. Amperometric glucose sensors based on immobilized glucose-oxidizing enzymes and chemically modified electrodes. *Anal. Chim. Acta* **249,** 43–54 (1991).

7. Jensen, R. G. Activation of Rubisco regulates photosynthesis at high temperature and CO₂ *Proc. Natl. Acad. Sci.* **97,** 12937 LP-12938 (2000).

8. Kaar, J. L., Jesionowski, A. M., Berberich, J. a, Moulton, R. & Russell, A. J. Impact of ionic liquid physical properties on lipase activity and stability. *J. Am. Chem. Soc.* **125,** 4125–31 (2003).

9. Harris, J. M., Martin, N. E. & Modi, M. Pegylation. *Clin. Pharmacokinet.* **40,** 539–551 (2001).

10. Weltz, J. S., Schwartz, D. K. & Kaar, J. L. Surface-Mediated Protein Unfolding as a Search Process for Denaturing Sites. *ACS Nano* **10,** 730–738 (2016).

11. Kastantin, M. *et al.* Connecting Protein Conformation and Dynamics with Ligand–Receptor Binding Using Three-Color Förster Resonance Energy Transfer Tracking. *J. Am. Chem. Soc.* **139,** 9937–9948 (2017).

12. Hetrick, E. M. & Schoenfisch, M. H. Reducing implant-related infections: active release strategies. *Chem. Soc. Rev.* **35,** 780–789 (2006).

13. Fetzner, S. Quorum quenching enzymes. *J. Biotechnol.* **201,** 2–14 (2015).

14. Lang, K. & Chin, J. W. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chem. Rev.* **114,** 4764–4806 (2014).

15. Stephanopoulos, N. & Francis, M. B. Choosing an effective protein bioconjugation strategy. *Nat. Chem. Biol.* **7,** 876–884 (2011).

16. Zimmer, M. Green Fluorescent Protein (GFP): Applications, Structure, and Related Photophysical Behavior. *Chem. Rev.* **102,** 759–782 (2002).

17. Keppler, A. *et al.* A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nat. Biotechnol.* **21,** 86 (2002).

18. Song, M. S., Salmena, L. & Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* **13,** 283 (2012).

19. Das, S., Dixon, J. E. & Cho, W. Membrane-binding and activation mechanism of PTEN. *Proc. Natl. Acad. Sci.* **100,** 7491 LP-7496 (2003).

20. Puthenveetil, S., Liu, D. S., White, K. A., Thompson, S. & Ting, A. Y. Yeast display evolution of a kinetically efficient 13-amino acid substrate for lipoic acid ligase. *J. Am. Chem. Soc.* **131,** 16430–16438 (2009).

21. Binz, H. K., Amstutz, P. & Pluckthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat Biotech* **23,** 1257–1268 (2005).

22. Kiss, C. *et al.* Antibody binding loop insertions as diversity elements. *Nucleic Acids Res.* **34,** e132 (2006).

23. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190 (2008).

24. Siegel, J. B. *et al.* Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science (80-.).* **329,** 309 LP-313 (2010).

25. Chevalier, A. *et al.* Massively parallel de novo protein design for targeted therapeutics. *Nature* **550,** 74 (2017).

26. Bhardwaj, G. *et al.* Accurate de novo design of hyperstable constrained peptides. *Nature* **538,** 329 (2016).

27. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510,** 103 (2014).

28. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods* **6,** 551–552 (2009).

29. Hermanson, G. T. *Bioconjugate Techniques*. (Academic Press, 2013).

30. Harris, T. K. & Turner, G. J. Structural Basis of Perturbed pKa Values of Catalytic Groups in Enzyme Active Sites. *IUBMB Life* **53,** 85–98 (2008).

31. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. Analysis of Catalytic Residues in Enzyme Active Sites. *J. Mol. Biol.* **324,** 105–121 (2002).

32. Berberich, J. A., Yang, L. W., Madura, J., Bahar, I. & Russell, A. J. A stable three-enzyme creatinine biosensor. 1. Impact of structure, function and environment on PEGylated and immobilized sarcosine oxidase. *Acta Biomater.* **1,** 173–181 (2005).

33. Young, T. S. & Schultz, P. G. Beyond the canonical 20 amino acids: expanding the genetic lexicon. *J. Biol. Chem.* **285,** 11039–11044 (2010).

34. Liu, C. C. & Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **79,** 413–444 (2010).

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

36. Meldal, M. & Tornøe, C. W. Cu-Catalyzed Azide−Alkyne Cycloaddition. *Chem. Rev.* **108,** 2952–3015 (2008).

37. Johnson, D. B. F. *et al.* RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat. Chem. Biol.* **7,** 779 (2011).

38. Uttamapinant, C. *et al.* A fluorophore ligase for site-specific protein labeling inside living cells. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10914–10919 (2010).

39. Dierks, T. *et al.* Multiple sulfatase deficiency is caused by mutations in the gene encoding the human C(alpha)-formylglycine generating enzyme. *Cell* **113,** 435–444 (2003).

40. Mazmanian, S. K., Liu, G., Ton-That, H. & Schneewind, O. Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall. *Science* **285,** 760–763 (1999).

41. Strop, P. Versatility of Microbial Transglutaminase. *Bioconjug. Chem.* **25,** 855–862

(2014).

42. Clancy, K. W., Melvin, J. A. & McCafferty, D. G. Sortase transpeptidases: insights into mechanism, substrate specificity, and inhibition. *Biopolymers* **94,** 385–396 (2010).

43. Rashidian, M., Dozier, J. K. & Distefano, M. D. Enzymatic labeling of proteins: techniques and approaches. *Bioconjug. Chem.* **24,** 1277–1294 (2013).

44. Rush, J. S. & Bertozzi, C. R. New aldehyde tag sequences identified by screening formylglycine generating enzymes in vitro and in vivo. *J. Am. Chem. Soc.* **130,** 12240–12241 (2008).

45. Guimaraes, C. P. *et al.* Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. *Nat. Protoc.* **8,** 1787–1799 (2013).

46. Uttamapinant, C., Sanchez, M. I., Liu, D. S., Yao, J. Z. & Ting, A. Y. Site-specific protein labeling using PRIME and chelation-assisted click chemistry. *Nat. Protoc.* **8,** 1620–1634 (2013).

47. Fujiwara, K. *et al.* Global conformational change associated with the two-step reaction catalyzed by Escherichia coli lipoate-protein ligase A. *J. Biol. Chem.* **285,** 9971–9980 (2010).

48. Liu, D. S. *et al.* Computational design of a red fluorophore ligase for site-specific protein labeling in living cells. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E4551-9 (2014).

49. Cohen, J. D., Zou, P. & Ting, A. Y. Site-specific protein modification using lipoic acid ligase and bis-aryl hydrazone formation. *Chembiochem* **13,** 888–894 (2012).

50. Yao, J. Z. *et al.* Fluorophore targeting to cellular proteins via enzyme-mediated azide ligation and strain-promoted cycloaddition. *J. Am. Chem. Soc.* **134,** 3720–3728 (2012).

51. Fernandez-Suarez, M. *et al.* Redirecting lipoic acid ligase for cell surface protein labeling with small-molecule probes. *Nat. Biotechnol.* **25,** 1483–1487 (2007).

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

53. Blizzard, R. J. *et al.* Ideal Bioorthogonal Reactions Using A Site-Specifically Encoded Tetrazine Amino Acid. *J. Am. Chem. Soc.* **137,** 10044–10047 (2015).

54. Row, R. D., Shih, H.-W., Alexander, A. T., Mehl, R. A. & Prescher, J. A. Cyclopropenones for Metabolic Targeting and Sequential Bioorthogonal Labeling. *J. Am. Chem. Soc.* **139,**

7370–7375 (2017).

55. Fernandez-Lafuente, R., Armisén, P., Sabuquillo, P., Fernández-Lorente, G. & M. Guisán, J. Immobilization of lipases by selective adsorption on hydrophobic supports. *Chem. Phys. Lipids* **93,** 185–197 (1998).

56. Luckarift, H. R., Spain, J. C., Naik, R. R. & Stone, M. O. Enzyme immobilization in a biomimetic silica support. *Nat. Biotechnol.* **22,** 211 (2004).

57. Wong, L. S., Khan, F. & Micklefield, J. Selective covalent protein immobilization: strategies and applications. *Chem. Rev.* **109,** 4025–4053 (2009).

58. Sheldon, R. A. Cross-linked enzyme aggregates (CLEA®s): stable and recyclable biocatalysts. *Biochem. Soc. Trans.* **35,** 1583 LP-1587 (2007).

59. Mateo, C., Palomo, J. M., Fernandez-Lorente, G., Guisan, J. M. & Fernandez-Lafuente, R. Improvement of enzyme activity, stability and selectivity via immobilization techniques. *Enzyme Microb. Technol.* **40,** 1451–1463 (2007).

60. Seo, M.-H. *et al.* Controlled and Oriented Immobilization of Protein by Site-Specific Incorporation of Unnatural Amino Acid. *Anal. Chem.* **83,** 2841–2845 (2011).

61. Lucent, D., Vishal, V. & Pande, V. S. Protein folding under confinement: A role for solvent. *Proc. Natl. Acad. Sci.* **104,** 10430 LP-10434 (2007).

62. Knotts, T. A., Rathore, N. & de Pablo, J. J. An Entropic Perspective of Protein Stability on Surfaces. *Biophys. J.* **94,** 4473–4483 (2008).

63. Friedel, M., Baumketner, A. & Shea, J.-E. Stability of a protein tethered to a surface. *J. Chem. Phys.* **126,** 95101 (2007).

64. Chaparro Sosa, A. F. *et al.* Stabilization of Immobilized Enzymes via the Chaperone-Like Activity of Mixed Lipid Bilayers. *ACS Appl. Mater. Interfaces* **10,** 19504–19513 (2018).

65. LeJeune, K. E. & Russell, A. J. Covalent binding of a nerve agent hydrolyzing enzyme within polyurethane foams. *Biotechnol. Bioeng.* **51,** 450–457 (2018).

66. Liu, X., Guan, Y., Shen, R. & Liu, H. Immobilization of lipase onto micron-size magnetic beads. *J. Chromatogr. B* **822,** 91–97 (2005).

67. Baker, D. A surprising simplicity to protein folding. *Nature* **405,** 39–42 (2000).

68. Dobson, C. M. Protein folding and misfolding. *Nature* **426,** 884 (2003).

69. Wang, L., Rivera, E. V, Benavides-Garcia, M. G. & Nall, B. T. Loop Entropy and Cytochrome c Stability. *J. Mol. Biol.* **353,** 719–729 (2005).

70. Ladurner, A. G. & Fersht, A. R. Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates11Edited by J. Karn. *J. Mol. Biol.* **273,** 330–337 (1997).

71. Nagi, A. D. & Regan, L. An inverse correlation between loop length and stability in a four-helix-bundle protein. *Fold. Des.* **2,** 67–75 (1997).

72. Fersht, A. R. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci.* **97,** 1525 LP-1529 (2000).

73. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **19,** 120–127 (2009).

74. Das, R. & Baker, D. Macromolecular Modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).

75. Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49,** 2987–2998 (2010).

76. Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).

77. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383,** 66–93 (2004).

78. Stein, A. & Kortemme, T. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS One* **8,** e63090 (2013).

79. Combs, S. A. *et al.* Small-molecule ligand docking into comparative models with Rosetta. *Nat. Protoc.* **8,** 1277 (2013).

80. Maki, D. G. & Tambyah, P. A. Engineering out the risk for infection with urinary catheters. *Emerg. Infect. Dis.* **7,** 342–347 (2001).

81. Daifuku, R. & Stamm, W. E. Association of rectal and urethral colonization with urinary tract infection in patients with indwelling catheters. *JAMA* **252,** 2028–2030 (1984).

82. Stewart, P. S. & Costerton, J. W. Antibiotic resistance of bacteria in biofilms. *Lancet (London, England)* **358,** 135–138 (2001).

83. Smith, A. W. Biofilms and antibiotic therapy: is there a role for combating bacterial resistance by the use of novel drug delivery systems? *Adv. Drug Deliv. Rev.* **57,** 1539–1550 (2005).

84. Davies, D. Understanding biofilm resistance to antibacterial agents. *Nat. Rev. Drug Discov.* **2,** 114–122 (2003).

85. Nowatzki, P. J. *et al.* Salicylic acid-releasing polyurethane acrylate polymers as anti-biofilm urological catheter coatings. *Acta Biomater.* **8,** 1869–1880 (2012).

86. Johnson, J. R., Johnston, B. D., Kuskowski, M. A. & Pitout, J. In vitro activity of available antimicrobial coated Foley catheters against Escherichia coli, including strains resistant to extended spectrum cephalosporins. *J. Urol.* **184,** 2572–2577 (2010).

87. Desai, D. G., Liao, K. S., Cevallos, M. E. & Trautner, B. W. Silver or nitrofurazone impregnation of urinary catheters has a minimal effect on uropathogen adherence. *J. Urol.* **184,** 2565–2571 (2010).

88. Bayston, R., Fisher, L. E. & Weber, K. An antimicrobial modified silicone peritoneal catheter with activity against both Gram-positive and Gram-negative bacteria. *Biomaterials* **30,** 3167–3173 (2009).

89. Johnson, J. R., Berggren, T. & Conway, A. J. Activity of a nitrofurazone matrix urinary catheter against catheter-associated uropathogens. *Antimicrob. Agents Chemother.* **37,** 2033–2036 (1993).

90. Raad, I. *et al.* Anti-adherence activity and antimicrobial durability of anti-infective-coated catheters against multidrug-resistant bacteria. *J. Antimicrob. Chemother.* **62,** 746–750 (2008).

91. Ulrich, R. L. Quorum quenching: enzymatic disruption of N-acylhomoserine lactone-mediated bacterial communication in Burkholderia thailandensis. *Appl. Environ. Microbiol.* **70,** 6173–6180 (2004).

92. Wang, Y.-J., Huang, J. J. & Leadbetter, J. R. Acyl-HSL signal decay: intrinsic to bacterial cell-cell communications. *Adv. Appl. Microbiol.* **61,** 27–58 (2007).

93. Roche, D. M. *et al.* Communications blackout? Do N-acylhomoserine-lactone-degrading enzymes have any role in quorum sensing? *Microbiology* **150,** 2023–2028 (2004).

94. Dong, Y.-H. & Zhang, L.-H. Quorum sensing and quorum-quenching enzymes. *J. Microbiol.* **43 Spec No,** 101–109 (2005).

95. Kim, M. H. *et al.* The molecular structure and catalytic mechanism of a quorum-quenching N-acyl-L-homoserine lactone hydrolase. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 17606–17611 (2005).

96. Liu, D. *et al.* Mechanism of the quorum-quenching lactonase (AiiA) from Bacillus thuringiensis. 1. Product-bound structures. *Biochemistry* **47,** 7706–7714 (2008).

97. Momb, J. *et al.* Mechanism of the quorum-quenching lactonase (AiiA) from Bacillus thuringiensis. 2. Substrate modeling and active site mutations. *Biochemistry* **47,** 7715–7725 (2008).

98. Kisch, J. M., Utpatel, C., Hilterhaus, L., Streit, W. R. & Liese, A. Pseudomonas aeruginosa biofilm growth inhibition on medical plastic materials by immobilized esterases and acylase. *Chembiochem* **15,** 1911–1919 (2014).

99. Geske, G. D., Wezeman, R. J., Siegel, A. P. & Blackwell, H. E. Small molecule inhibitors of bacterial quorum sensing and biofilm formation. *J. Am. Chem. Soc.* **127,** 12762–12763 (2005).

100. Parker, B. M., Taylor, I. N., Woodley, J. M., Ward, J. M. & Dalby, P. A. Directed evolution of a thermostable l-aminoacylase biocatalyst. *J. Biotechnol.* **155,** 396–405 (2011).

101. Xu, F., Byun, T., Deussen, H.-J. & Duke, K. R. Degradation of N-acylhomoserine lactones, the bacterial quorum-sensing molecules, by acylase. *J. Biotechnol.* **101,** 89–96 (2003).

102. Ma, F. *et al.* Heterologous expression of human paraoxonases in Pseudomonas aeruginosa inhibits biofilm formation and decreases antibiotic resistance. *Appl. Microbiol. Biotechnol.* **83,** 135–141 (2009).

103. Ivanova, K., Fernandes, M. M., Mendoza, E. & Tzanov, T. Enzyme multilayer coatings inhibit Pseudomonas aeruginosa biofilm formation on urinary catheters. *Appl. Microbiol. Biotechnol.* **99,** 4373–4385 (2015).

104. Ivanova, K. *et al.* Quorum-Quenching and Matrix-Degrading Enzymes in Multilayer Coatings Synergistically Prevent Bacterial Biofilm Formation on Urinary Catheters. *ACS Appl. Mater. Interfaces* **7,** 27066–27077 (2015).

105. Perrier, J., Durand, A., Giardina, T. & Puigserver, A. Catabolism of intracellular N-

terminal acetylated proteins: involvement of acylpeptide hydrolase and acylase. *Biochimie* **87,** 673–685 (2005).

106. Pisal, D. S., Kosloski, M. P. & Balu-Iyer, S. V. Delivery of therapeutic proteins. *J. Pharm. Sci.* **99,** 2557–2575 (2010).

107. Drevon, G. F. & Russell, A. J. Irreversible immobilization of diisopropylfluorophosphatase in polyurethane polymers. *Biomacromolecules* **1,** 571–576 (2000).

108. Bode, M. L., van Rantwijk, F. & Sheldon, R. A. Crude aminoacylase from aspergillus sp. is a mixture of hydrolases. *Biotechnol. Bioeng.* **84,** 710–713 (2003).

109. Essar, D. W., Eberly, L., Hadero, A. & Crawford, I. P. Identification and characterization of genes for a second anthranilate synthase in Pseudomonas aeruginosa: interchangeability of the two anthranilate synthases and evolutionary implications. *J. Bacteriol.* **172,** 884–900 (1990).

110. Sarkisova, S., Patrauchan, M. A., Berglund, D., Nivens, D. E. & Franklin, M. J. Calcium-induced virulence factors associated with the extracellular matrix of mucoid Pseudomonas aeruginosa biofilms. *J. Bacteriol.* **187,** 4327–4337 (2005).

111. Ng, F. S. W., Wright, D. M. & Seah, S. Y. K. Characterization of a phosphotriesterase-like lactonase from Sulfolobus solfataricus and its immobilization for disruption of quorum sensing. *Appl. Environ. Microbiol.* **77,** 1181–1186 (2011).

112. Muller, M. & Merrett, N. D. Pyocyanin production by Pseudomonas aeruginosa confers resistance to ionic silver. *Antimicrob. Agents Chemother.* **58,** 5492–5499 (2014).

113. LeJeune, Swers, Hetro, Donahey & Russell. Increasing the tolerance of organophosphorus hydrolase to bleach. *Biotechnol. Bioeng.* **64,** 250–254 (1999).

114. Gordon, R. K. *et al.* Organophosphate skin decontamination using immobilized enzymes. *Chem. Biol. Interact.* **119–120,** 463–470 (1999).

115. LeJeune, K. E. *et al.* Dramatically stabilized phosphotriesterase-polymers for nerve agent degradation. *Biotechnol. Bioeng.* **54,** 105–114 (1997).

116. Gill, I. & Ballesteros, A. Bioencapsulation within synthetic polymers (Part 2): non-sol-gel protein-polymer biocomposites. *Trends Biotechnol.* **18,** 469–479 (2000).

117. Drevon, G. F., Hartleib, J., Scharff, E., Ruterjans, H. & Russell, A. J. Thermoinactivation of diisopropylfluorophosphatase-containing polyurethane polymers. *Biomacromolecules*

2, 664–671 (2001).

118. Vasudevan, P. T. *et al.* A novel hydrophilic support, CoFoam, for enzyme immobilization. *Biotechnol. Lett.* **26,** 473–477 (2004).

119. Santerre, J. P., Woodhouse, K., Laroche, G. & Labow, R. S. Understanding the biodegradation of polyurethanes: from classical implants to tissue engineering materials. *Biomaterials* **26,** 7457–7470 (2005).

120. Guelcher, S. A. Biodegradable polyurethanes: synthesis and applications in regenerative medicine. *Tissue Eng. Part B. Rev.* **14,** 3–17 (2008).

121. Klement, P., Du, Y. J., Berry, L. R., Tressel, P. & Chan, A. K. C. Chronic performance of polyurethane catheters covalently coated with ATH complex: a rabbit jugular vein model. *Biomaterials* **27,** 5107–5117 (2006).

122. Du, Y. J. *et al.* Protein adsorption on polyurethane catheters modified with a novel antithrombin-heparin covalent complex. *J. Biomed. Mater. Res. A* **80,** 216–225 (2007).

123. Trigwell, S., De, S., Sharma, R., Mazumder, M. K. & Mehta, J. L. Structural evaluation of radially expandable cardiovascular stents encased in a polyurethane film. *J. Biomed. Mater. Res. B. Appl. Biomater.* **76,** 241–250 (2006).

124. Vikstrom, E., Tafazoli, F. & Magnusson, K.-E. Pseudomonas aeruginosa quorum sensing molecule N-(3 oxododecanoyl)-l-homoserine lactone disrupts epithelial barrier integrity of Caco-2 cells. *FEBS Lett.* **580,** 6921–6928 (2006).

125. Sio, C. F. *et al.* Quorum quenching by an N-acyl-homoserine lactone acylase from Pseudomonas aeruginosa PAO1. *Infect. Immun.* **74,** 1673–1682 (2006).

126. Singh, P. K. *et al.* Quorum-sensing signals indicate that cystic fibrosis lungs are infected with bacterial biofilms. *Nature* **407,** 762–764 (2000).

127. Lele, B. S. *et al.* Enhancing bioplastic-substrate interaction via pore induction and directed migration of enzyme location. *Biotechnol. Bioeng.* **86,** 628–636 (2004).

128. Branda, S. S., Vik, S., Friedman, L. & Kolter, R. Biofilms: the matrix revisited. *Trends Microbiol.* **13,** 20–26 (2005).

129. Wagner, V. E. & Iglewski, B. H. P. aeruginosa Biofilms in CF Infection. *Clin. Rev. Allergy Immunol.* **35,** 124–134 (2008).

130. Williams, P. & Camara, M. Quorum sensing and environmental adaptation in

Pseudomonas aeruginosa: a tale of regulatory networks and multifunctional signal molecules. *Curr. Opin. Microbiol.* **12,** 182–191 (2009).

131. Reyes, E. A., Bale, M. J., Cannon, W. H. & Matsen, J. M. Identification of Pseudomonas aeruginosa by pyocyanin production on Tech agar. *J. Clin. Microbiol.* **13,** 456–458 (1981).

132. Giepmans, B. N. G., Adams, S. R., Ellisman, M. H. & Tsien, R. Y. The fluorescent toolbox for assessing protein location and function. *Science* **312,** 217–224 (2006).

133. Cummings, C., Murata, H., Koepsel, R. & Russell, A. J. Tailoring enzyme activity and stability using polymer-based protein engineering. *Biomaterials* **34,** 7437–7443 (2013).

134. Cummings, C., Murata, H., Koepsel, R. & Russell, A. J. Dramatically increased pH and temperature stability of chymotrypsin using dual block polymer-based protein engineering. *Biomacromolecules* **15,** 763–771 (2014).

135. Nordwald, E. M. & Kaar, J. L. Stabilization of enzymes in ionic liquids via modification of enzyme charge. *Biotechnol. Bioeng.* **110,** 2352–2360 (2013).

136. Harris, J. M. & Chess, R. B. Effect of pegylation on pharmaceuticals. *Nat. Rev. Drug Discov.* **2,** 214–221 (2003).

137. Sola, R. J. & Griebenow, K. Effects of glycosylation on the stability of protein pharmaceuticals. *J. Pharm. Sci.* **98,** 1223–1245 (2009).

138. Furuhashi, M. & Hotamisligil, G. S. Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. *Nat. Rev. Drug Discov.* **7,** 489–503 (2008).

139. De, P., Li, M., Gondi, S. R. & Sumerlin, B. S. Temperature-regulated activity of responsive polymer-protein conjugates prepared by grafting-from via RAFT polymerization. *J. Am. Chem. Soc.* **130,** 11288–11289 (2008).

140. Gorostiza, P. & Isacoff, E. Y. Optical switches for remote and noninvasive control of cell signaling. *Science* **322,** 395–399 (2008).

141. Young, C. L., Britton, Z. T. & Robinson, A. S. Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnol. J.* **7,** 620–634 (2012).

142. Yi, X. & Kabanov, A. V. Brain delivery of proteins via their fatty acid and block copolymer modifications. *J. Drug Target.* **21,** 940–955 (2013).

143. Tessmar, J. K. & Gopferich, A. M. Matrices and scaffolds for protein delivery in tissue

engineering. *Adv. Drug Deliv. Rev.* **59,** 274–291 (2007).

144. Murata, H., Cummings, C. S., Koepsel, R. R. & Russell, A. J. Rational tailoring of substrate and inhibitor affinity via ATRP polymer-based protein engineering. *Biomacromolecules* **15,** 2817–2823 (2014).

145. Murata, H., Cummings, C. S., Koepsel, R. R. & Russell, A. J. Polymer-based protein engineering can rationally tune enzyme activity, pH-dependence, and stability. *Biomacromolecules* **14,** 1919–1926 (2013).

146. Berberich, J. A., Yang, L. W., Madura, J., Bahar, I. & Russell, A. J. A stable three-enzyme creatinine biosensor. 1. Impact of structure, function and environment on PEGylated and immobilized sarcosine oxidase. *Acta Biomater.* **1,** 173–181 (2005).

147. Prescher, J. A. & Bertozzi, C. R. Chemistry in living systems. *Nat. Chem. Biol.* **1,** 13–21 (2005).

148. Peeler, J. C. *et al.* Genetically encoded initiator for polymer growth from proteins. *J. Am. Chem. Soc.* **132,** 13575–13577 (2010).

149. Kim, C. H., Axup, J. Y. & Schultz, P. G. Protein conjugation with genetically encoded unnatural amino acids. *Curr. Opin. Chem. Biol.* **17,** 412–419 (2013).

150. Schmied, W. H., Elsasser, S. J., Uttamapinant, C. & Chin, J. W. Efficient multisite unnatural amino acid incorporation in mammalian cells via optimized pyrrolysyl tRNA synthetase/tRNA expression and engineered eRF1. *J. Am. Chem. Soc.* **136,** 15577–15583 (2014).

151. Carrico, I. S., Carlson, B. L. & Bertozzi, C. R. Introducing genetically encoded aldehydes into proteins. *Nat. Chem. Biol.* **3,** 321–322 (2007).

152. Witte, M. D. *et al.* Site-specific protein modification using immobilized sortase in batch and continuous-flow systems. *Nat. Protoc.* **10,** 508–516 (2015).

153. Lin, C.-W. & Ting, A. Y. Transglutaminase-catalyzed site-specific conjugation of small-molecule probes to proteins in vitro and on the surface of living cells. *J. Am. Chem. Soc.* **128,** 4542–4543 (2006).

154. Lawrence, M. S., Phillips, K. J. & Liu, D. R. Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* **129,** 10110–10112 (2007).

155. Hammill, J. T., Miyake-Stoner, S., Hazen, J. L., Jackson, J. C. & Mehl, R. A. Preparation of site-specifically labeled fluorinated proteins for 19F-NMR structural characterization.

Nat. Protoc. **2,** 2601–2607 (2007).

156. Schneider, C. a, Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9,** 671–675 (2012).

157. Hong, V., Presolski, S. I., Ma, C. & Finn, M. G. Analysis and optimization of copper-catalyzed azide-alkyne cycloaddition for bioconjugation. *Angew. Chem. Int. Ed. Engl.* **48,** 9879–9883 (2009).

158. Pedelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24,** 79–88 (2006).

159. Pavoor, T. V, Cho, Y. K. & Shusta, E. V. Development of GFP-based biosensors possessing the binding properties of antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 11895–11900 (2009).

160. Cetinkaya, M., Zeytun, A., Sofo, J. & Demirel, M. C. How do insertions affect green fluorescent protein? *Chem. Phys. Lett.* **419,** 48–54 (2006).

161. Baird, G. S., Zacharias, D. A. & Tsien, R. Y. Circular permutation and receptor insertion within green fluorescent proteins. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11241–11246 (1999).

162. Reid, B. G. & Flynn, G. C. Chromophore formation in green fluorescent protein. *Biochemistry* **36,** 6786–6791 (1997).

163. Deiters, A., Cropp, T. A., Summerer, D., Mukherji, M. & Schultz, P. G. Site-specific PEGylation of proteins containing unnatural amino acids. *Bioorg. Med. Chem. Lett.* **14,** 5743–5745 (2004).

164. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 19–24 (2002).

165. Kyba, E. P. & Abramovitch, R. A. Photolysis of alkyl azides. Evidence for a nonnitrene mechanism. *J. Am. Chem. Soc.* **102,** 735–740 (1980).

166. Grunewald, J. *et al.* Mechanistic studies of the immunochemical termination of self-tolerance with unnatural amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 4337–4342 (2009).

167. Merritt, J. H., Ollis, A. A., Fisher, A. C. & DeLisa, M. P. Glycans-by-design: engineering bacteria for the biosynthesis of complex glycans and glycoconjugates. *Biotechnol. Bioeng.*

110, 1550–1564 (2013).

168. Valderrama-Rincon, J. D. *et al.* An engineered eukaryotic protein glycosylation pathway in Escherichia coli. *Nat. Chem. Biol.* **8,** 434–436 (2012).

169. Durocher, Y. & Butler, M. Expression systems for therapeutic glycoprotein production. *Curr. Opin. Biotechnol.* **20,** 700–707 (2009).

170. Appel, M. J. & Bertozzi, C. R. Formylglycine, a post-translationally generated residue with unique catalytic capabilities and biotechnology applications. *ACS Chem. Biol.* **10,** 72–84 (2015).

171. Smith, E. L. *et al.* Chemoenzymatic Fc glycosylation via engineered aldehyde tags. *Bioconjug. Chem.* **25,** 788–795 (2014).

172. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science (80-.).* **255,** 959 LP-965 (1992).

173. Ramasubbu, N., Ragunath, C. & Mishra, P. J. Probing the Role of a Mobile Loop in Substrate Binding and Enzyme Activity of Human Salivary Amylase. *J. Mol. Biol.* **325,** 1061–1076 (2003).

174. Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop — a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15,** 430–434 (1990).

175. Adams, J. A. Activation Loop Phosphorylation and Catalysis in Protein Kinases: Is There Functional Evidence for the Autoinhibitor Model? *Biochemistry* **42,** 601–607 (2003).

176. Tawfik, D. S. Loop Grafting and the Origins of Enzyme Species. *Science (80-.).* **311,** 475 LP-476 (2006).

177. Park, H.-S. *et al.* Design and Evolution of New Catalytic Activity with an Existing Protein Scaffold. *Science (80-.).* **311,** 535 LP-538 (2006).

178. Kiss, C. *et al.* Antibody binding loop insertions as diversity elements. *Nucleic Acids Res.* **34,** e132 (2006).

179. Plaks, J. G., Falatach, R., Kastantin, M., Berberich, J. A. & Kaar, J. L. Multisite clickable modification of proteins using lipoic acid ligase. *Bioconjug. Chem.* **26,** 1104–1112 (2015).

180. Toma, S. *et al.* Grafting of a calcium-binding loop of thermolysin to Bacillus subtilis neutral protease. *Biochemistry* **30,** 97–106 (1991).

181. Poth, A. G., Chan, L. Y. & Craik, D. J. Cyclotides as grafting frameworks for protein engineering and drug design applications. *Pept. Sci.* **100,** 480–491 (2013).

182. Azoitei, M. L. *et al.* Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold. *Science (80-.).* **334,** 373 LP-376 (2011).

183. Nordwald, E. M., Armstrong, G. S. & Kaar, J. L. NMR-Guided Rational Engineering of an Ionic-Liquid-Tolerant Lipase. *ACS Catal.* **4,** 4057–4064 (2014).

184. Chado, G. R., Holland, E. N., Tice, A. K., Stoykovich, M. P. & Kaar, J. L. Modification of Lipase with Poly(4-acryloylmorpholine) Enhances Solubility and Transesterification Activity in Anhydrous Ionic Liquids. *Biomacromolecules* **19,** 1324–1332 (2018).

APPENDIX A: LIPOIC ACID LIGASE-PROMOTED BIOORTHOGONAL PROTEIN MODIFICATION AND IMMOBILIZATION PROTOCOL

Adapted from Enzyme-mediated Ligation Technologies, Springer Protocol (submitted)

A.1 Introduction

Protein bioconjugation benefits from precise regional and temporal control. One notable way of achieving this control is through the enzymatic attachment of bioorthogonal reactive handles to peptide recognition sequences that are genetically fused to target proteins of interest. The lipoic acid ligase variant, $LplA^{W37V}$, functionalizes proteins through this mechanism, covalently attaching an azide-bearing lipoic acid derivative to a 13-amino acid recognition sequence known as the Lipoic acid ligase Acceptor Peptide (LAP). Once attached, the azide group can be modified with diverse chemical entities through azide-alkyne click chemistry, enabling conjugation of chemical probes such as fluorophores and facilitating polymer attachment, glycosylation, and protein immobilization in addition to many other possible chemical modifications. The versatility of the attached azide group is complemented by the modular nature of the LAP sequence, which can be introduced within a protein at internal and/or terminal sites as well as at multiple sites simultaneously. This chapter describes the *in vitro* $LplA^{W37V}$-mediated ligation of 10-azidodecanoic acid to a LAP-containing target protein (*i.e.* green fluorescent protein (GFP)) and the characterization of the ligation reaction products. Additionally, methods for the modification and immobilization of azide-functionalized LAP-GFP are discussed.

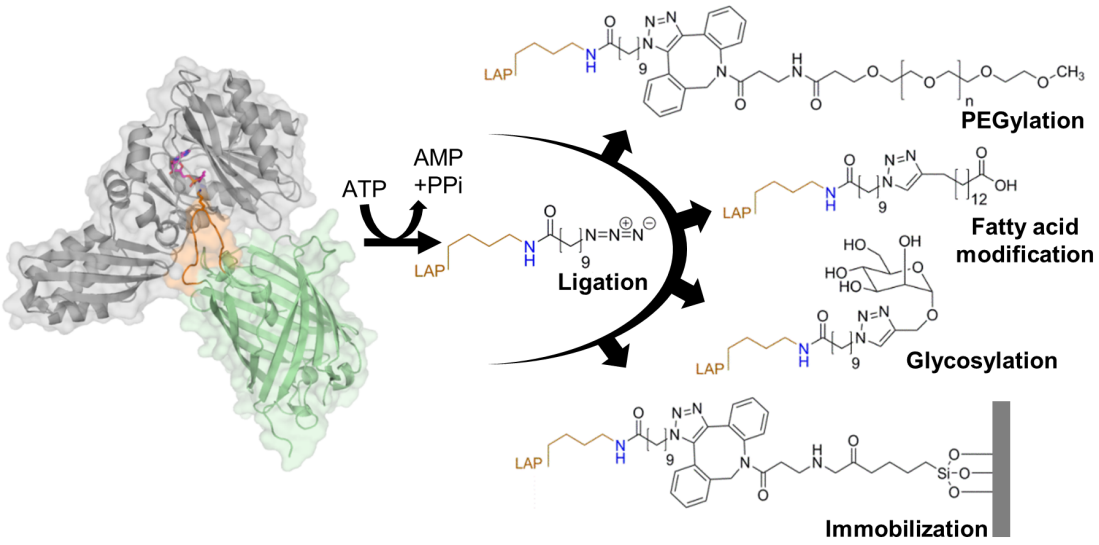

Figure A.1. Chemistry of ligation and subsequent modification reactions described in Appendix A protocols.
Ligation and subsequent modification and immobilization reactions of LAP-stGFP (green with LAP sequence in orange)[158] by LplA^{W37V} (grey).[47] Ligation of 10-azidodecanoic acid by LplA^{W37V} introduces an azide moiety within the LAP-stGFP target protein. Subsequent copper catalyzed azide-alkyne cycloaddition reactions allow for fatty acid modification and glycosylation of the LAP sequence *via* the appended azide moiety and strain-promoted cycloaddition reactions allow for PEGylation and immobilization.

A.2 Materials

Prepare all solutions using ultrapure water (ddH_2O, 18.2 MΩ • cm at 25 °C), analytical grade reagents and solvents (unless stated otherwise).

A.2.1 Reagents

1. LplA^{W37V} plasmid: pYFJ16-LplA(W37V), Addgene no. 34838.[50]
2. stGFP plasmid: pET-stGFP as described by Lawrence et al.[154]
3. Electrocompetent *E. coli* BL21 (DE3) cells.

4. Luria Broth (LB)-Amp medium: 10 g/L tryptone, 10 g/L sodium chloride, 5 g/L yeast extract and 100 μg/mL ampicillin. Sterilize by autoclaving.
5. Lysis buffer: 50 mM Tris, 250 mM NaCl, 5 mM imidazole, 2% glycerol, 0.01% β-mercaptoethanol, pH 8.0.
6. $LplA^{W37V}$ storage buffer: 20 mM Tris, 10% glycerol, 0.01% β-mercaptoethanol, pH 7.5.
7. stGFP storage buffer: 20 mM sodium phosphate, pH 7.0 (*see* **Note 1**).
8. 10-azidodecanoic acid: purchased from commercial supplier or synthesized as a pure liquid as previously described.[179]
9. IPTG stock solution: 1 M isopropyl-β-D-thiogalactoside (IPTG) in ddH_2O. Filter sterilize through a 0.22 μm syringe filter before use.
10. Ligation buffer: 25 mM sodium phosphate, 2 mM $MgCl_2$, 2 mM ATP, pH 7.2.
11. TAMRA-DBCO: TAMRA-PEG4-DBCO (Broadpharm, catalog no. BP-22456).
12. Detergent: 2% MICRO-90® Concentrated Cleaning Solution in ddH_2O.
13. Piranha solution: 70% (v/v) sulfuric acid, 30% (v/v) of an aqueous hydrogen peroxide solution (for a final hydrogen peroxide concentration of 9% (v/v)).
14. Epoxy-silane solution: 10% (v/v) 5,6-epoxyhexyltriethoxysilane (Gelest), 5% (v/v) *n*-butylamine, 85% (v/v) toluene.
15. DBCO-NH_2 solution: 1 mM dibenzocyclooctyne-amine (added from an 18 mM stock in DMSO), 100 mM borate buffer, pH 9.5.
16. Ammonium chloride solution: 1% (w/v) ammonium chloride, 100 mM borate buffer, pH 9.5.
17. DBCO-PEG (5 kDa M_w) (Jena Bioscience).

18. $CuSO_4$-THPTA solution: 6.7 μM $CuSO_4.5\ H_2O$, 33.3 mM tris(3-hydroxypropyltriazolylmethyl)amine (THPTA) in ddH_2O. Store at 4 °C and incubate at room temperature for 20 min before use.
19. Propargyl α-D-mannopyranoside (LC Scientific).
20. Palmitic acid alkyne: 5-hexadecynoic acid (Cayman Chemical).
21. ESI infusion solution: 40% (v/v) acetonitrile, 0.1% (v/v) formic acid in water (Fisher Scientific Optima™ LC/MS grade for all components including water).
22. Succinic acid matrix solution.
23. 10x Tris/Glycine/SDS buffer *e.g.* Bio-Rad product number 1610732.
24. SDS polyacrylamide gel electrophoresis (PAGE) loading dye: 9:1 (v/v) Bio-Rad Laemmli sample buffer (4x):β-mercaptoethanol.
25. Fluorescent protein ladder: Bio-Rad Precision Plus Protein™ Dual Color Standard.
26. Coomassie stain solution: 10% acetic acid, 50% ethanol, and 0.1% Coomassie Blue R-250.

A.2.2 Equipment

1. 1.5-mL microcentrifuge tubes.
2. Electroporator, *e.g.* Eppendorf™ Eporator™ or equivalent.
3. 1 mm Electroporation cuvettes, *e.g* Bulldog Bio 12358345 or equivalent.
4. Incubator shaker, *e.g.* New Brunswick™ I26 or equivalent.
5. LB-agar plates.

6. 250-mL culture flask.
7. 2-L baffled culture flask.
8. Centrifuge, *.e.g.* Thermo Scientific™ Sorvall™ Legend™ XTR or equivalent.
9. Spectrophotometer, *e.g.* Eppendorf BioPhotometer® 6131 or equivalent.
10. GEA PandaPLUS 2000 homogenizer.
11. 250 mL Vacuum Filter with 0.2 μm pore size.
12. Fast Protein Liquid Chromatography (FPLC) system, *e.g.* Bio-Rad BioLogic DuoFlow™ or equivalent.
13. 5 mL BioRad Bio-Scale™ Mini Profinity™ IMAC Cartridge nickel column.
14. Vivaspin® 20 mL centrifugal concentrator with 10000 Da MWCO (molecular weight cut-off).
15. ElectroSpay Ionization Mass Spectrometer (ESI-MS) and quadrupol/ ToF mass analyzer, *e.g.* Waters Synapt G2 High Definition Mass Spectrometer.
16. Trajan SGE 50 μL syringe (part number 004312).
17. Laser scanner, *e.g.* GE Typhoon™ FLA 9500.
18. Glass cover slide, *e.g.* VWR® Micro Cover Glasses (catalog no. 48393-251).
19. Novascan PSD Digital UV-Ozone cleaner.
20. MALDI mass spectrometer, *e.g.* Applied Biosystems Voyager-DE™ STR MALDI mass spectrometer or equivalent.
21. SDS-PAGE gels, *e.g.* Bio-Rad 4-20% Mini-PROTEAN® TGX™ Precast Protein Gels.
22. Gel apparatus, *e.g.* Bio-Rad Mini-PROTEAN Tetra Cell.
23. Electrophoresis power source, e.g. Bio-Rad PowerPac™ Basic power supply.

24. Microwave oven.

A.3 Methods

A.3.1 LplA^{W37V} Transformation, Expression, and Purification

1. Transform the pYFJ16-LplA(W37V) plasmid into BL21 (DE3) *E. coli* cells (*see* **Note 2**). Prepare transformation samples by precooling 1 μL of plasmid (diluted in ddH_2O to a final concentration of ~50 ng/μL) in a 1.5-mL Eppendorf tube on ice. Thaw *E. coli* cells on ice and add 50 μL to the cooled plasmid. Mix thoroughly by flicking the tube and incubate on ice for several minutes.
2. Transfer the above mixture to an electroporation cuvette that has also been pre-cooled on ice and electroporate at 1200-1500 V.
3. Add 1 mL of LB without any ampicillin to the electroporation cuvette to act as a recovery medium and pipette up and down to thoroughly mix in the transformed cells. Transfer the solution to a 1.5-mL Eppendorf tube and incubate at 37 °C with shaking at 200 rpm for 1 hour.
4. Plate 50 μL of recovered cells onto an LB agar plate containing ampicillin at a concentration of 100 μg/mL. Incubate plates at 37 °C overnight.
5. Pick a single colony, inoculate a 250-mL flask containing 50 mL liquid culture of LB-Amp medium and incubate the culture overnight at 37 °C shaking >200 rpm.
6. Pellet cells *via* centrifugation at 4000 x g for 10 min. Resuspend the cells in fresh LB-Amp.

7. Inoculate four 2-L baffled flasks that each contain 250 mL LB-Amp with the resuspended cells, adding an appropriate volume so as not to exceed a final OD_{600} of 0.1 as measured by UV-VIS spectrophotometry with a 1 cm path length.
8. Incubate flasks at 37 °C at >200 rpm until an OD_{600} ~0.6. Induce LplA expression by adding 250 μL IPTG stock solution to each flask (final IPTG concentration of 1 mM). Allow protein expression to proceed by continuing to incubate the cells at 37 °C and >200 rpm overnight.
9. Pellet cells by centrifugation at 4000 x g for 10 min. Remove the supernatant (*see* **Note 3**).
10. Resuspend cell pellets in approximately 200 mL lysis buffer that has been pre-cooled to 4 °C. Ensure that all cell clumps have been broken up before proceeding.
11. Lyse cells by homogenization. Begin by rinsing the homogenizer with ddH_2O followed by chilled lysis buffer. Add resuspended cells and pass them through the homogenizer 2-3 times at 800-1000 bar. Be careful to keep the cell lysate on ice between the different rounds of homogenization.
12. Clear cell lysate by centrifugation at 15000 x g for 30-40 min and filter the supernatant, which will contain soluble $LplA^{W37V}$, with a 0.2 μm cutoff vacuum filter.
13. Attach a Ni-NTA IMAC column to the FPLC. Rinse the column with 2 column volumes (CV) of deionized water and pre-equilibrate it with 5 CV of lysis buffer using a flow rate of 5-10 mL/min and pressures of less than 45 psi.
14. Load cell lysate onto the column maintaining the flow rates and pressures described above.

15. Wash the column with 40 CV (200 mL) lysis buffer.
16. Elute LplA^{W37V} from the column by gradually ramping the concentration of imidazole in the lysis buffer from 5-250 mM over the course of 20 CV while collecting eluent in 2-5 mL fractions.
17. Run an SDS-PAGE gel (see section A.3.5.2) to determine which elution fractions contain LplA^{W37V}. Pool product-containing fractions and dialyze them against 4 L of LplA^{W37V} storage buffer at 4 °C. Refresh the buffer at least three times over the course of the dialysis.
18. Aliquot 25 μL of LplA^{W37V} into 1.5-mL Eppendorf tubes, flash freeze in liquid nitrogen, and store at -80 °C until further use (*see* **Note 4**).

A.3.2 LAP-stGFP Target Protein Cloning, Expression, and Purification

1. The LAP sequence can be genetically introduced into a POI by site-directed mutagenesis using long primers to achieve the required 39 base pair insertion, or, more simply, by having a new gene synthesized with the LAP sequence included within it. The peptide sequence, GFEIDKVWYDLDA, should be codon optimized for *E. coli* expression. We generally use the following DNA sequence to code for it in all of our work: GGC TTC GAG ATC GAC AAG GTG TGG TAC GAC CTG GAC GCC (*see* **Note 5**).
2. Express and purify LAP-stGFP from a pET-stGFP plasmid that has been mutated to contain an *N*-terminal LAP sequence. Follow the transformation, expression, and

purification procedure described above for LplA^{W37V} but use stGFP storage buffer in place of LplA^{W37V} storage buffer (*see* **Note 6**).

3. Additionally, express and purify wild-type stGFP control from the stGFP plasmid again as described above using stGFP storage buffer (*see* **Note 7**).

A.3.3 Ligation of 10-Azidodecanoic Acid to LAP-stGFP

1. Add LAP-stGFP, LplA^{W37V}, and 10-azidodeanoic acid to ligation buffer to final concentrations of 10 μM, 0.1 μM, and 600 μM respectively (*see* **Note 8**).
2. Perform identical control ligation reactions as described above but lacking LplA^{W37V} (in one reaction) and 10-azidodecanoic acid (in a different reaction) for negative controls. As a third negative control, set up another ligation reaction with both LplA^{W37V} and 10-azidodecanoic acid but with wild-type stGFP in place of LAP-stGFP.
3. Incubate the ligation reactions at 30 °C. Conversions higher than 90% have been observed for LAP-stGFP constructs after 1 h incubation, but appropriate reaction times may vary. Less exposed sites may require longer reaction times to achieve high conversion, which will need to be determined empirically.
4. After the ligation, remove unreacted 10-azidodecanoic acid by dialyzing against 4 L of 50 mM phosphate buffer (pH 7.0) at 4 °C. Refresh the buffer at least 3 times, and perform this same dialysis for all controls. Store the ligated protein and un-ligated controls at 4 °C until ready for use (*see* **Note 9**).

5. Analyze the ligation product according to section A.3.5.1. If an appropriate mass spectrometer is unavailable, label the ligation product by reacting it with a strained alkyne functionalized fluorophore (see section A.3.4.3), and run an SDS-PAGE gel to qualitatively assess azide mediated fluorophore addition as a proxy for azide ligation (section A.3.5.2).

A.3.4 Azide-Mediated Conjugation Reactions

Once 10-azidodecanoic acid has been ligated to the LAP sequence with LplA, the appended azide can serve as a bioorthogonal reactive handle, directing site specific bioconjugation of diverse alkyne functionalized molecules. Reactions of alkynes with the appended azide may proceed through a copper catalyst or may be enabled by strained alkyne molecules. A number of highly relevant azide-directed protein modification reactions enabled by both copper-dependent and copper-free mechanisms are described below including immobilization, PEGylation, fluorophore labeling, glycosylation, and fatty acid modification.

A.3.4.1 Azide Mediated Protein Immobilization

1. Wash at least two glass cover slides in detergent and rinse them in ddH_2O.
2. Incubate cover slides in piranha solution for 1 h (*see* **Note 10**).
3. Remove cover slides from the piranha solution, and rinse them thoroughly with ddH_2O.
4. Dry slides with ultra-pure nitrogen and place them in a UV-ozone machine for 15 min (*see* **Note 11**).

5. Add several milliliters of epoxy-silane solution to a wide-mouthed jar or beaker and set the cleaned cover slides face down across the rim of the jar so that the cover slides act as an imperfect lid.
6. Allow plates to sit in this configuration in a fume hood for 20 h over which time vapor deposition of the epoxy-silane to the downward facing glass surfaces of the cover slides will occur.
7. Place cover slides into a Petri dish face up and add several milliliters of DBCO-NH_2 solution.
8. Incubate cover slides in DBCO-NH_2 solution for 30 h at 37 °C with gentle shaking.
9. Remove cover slides from the DBCO-NH_2 solutions and rinse with ddH_2O.
10. Transfer cover slides into a new Petri dish containing several milliliters of ammonium chloride solution to quench unreacted epoxide groups. Incubate overnight at 37 °C with gentle shaking.
11. Thoroughly rinse cover slides with ddH_2O. Incubate one or more cover slides in a new Petri dish containing several milliliters of ligated protein (10 μM LAP-stGFP in 50 mM phosphate buffer pH 7.0) for 4 h with gentle shaking. Additionally, incubate at least one cover slide in an un-ligated protein solution (10 μM LAP-stGFP that has not been azide functionalized in 50 mM phosphate buffer pH 7.0) to act as a negative control.
12. Remove the protein solution from the Petri dish and add several milliliters of 50 mM phosphate buffer (pH 7.0) to remove unreacted, nonspecifically adsorbed protein. For thorough rinsing, replace this buffer every 30 min over the course of a 4 h incubation at 37 °C with shaking (*see* **Note 12**).

13. If imaging cannot be performed immediately after the above wash step, store protein functionalized cover slides in 50 mM phosphate buffer (pH 7.0) at 4 °C.
14. Assess protein immobilization by fluorescent imaging with a laser scanner using an excitation wavelength of 467 nm (or 473 if using the GE Typhoon FLA 9500). Compare protein modified surfaces to surfaces treated with protein that lacks the azide functionality to qualitatively determine the extent of immobilization.

A.3.4.2 Azide Mediated Chemical Functionalization: PEGylation

1. Mix ligated LAP-stGFP with a 50-fold molar excess of PEG-DBCO by, for example, dissolving 1.25 mg of DBCO-PEG in 500 μL of 50 mM phosphate buffer pH 7.0 and adding it to 500 μL of ligated LAP-stGFP. This may be scaled up as desired.
2. Incubate for at least 30 min at room temperature.
3. Remove excess PEG-DBCO by dialyzing against 4 L of 50 mM phosphate buffer (pH 7.) at 4 °C or by Ni-NTA purification as described for protein purification from crude cell lysate in section 3.1 (*see* **Note 13**).
4. Store PEGylated protein at 4 °C until ready for characterization and use.

A.3.4.3 Azide Mediated Chemical Functionalization: Fluorophore Labeling

1. Make a 10 mM stock solution of TAMRA-DBCO in DMSO by adding 533.3 μL of DMSO directly to the bottle sent by the supplier containing 5 mg of powdered TAMRA-DBCO (*see* **Note 14**).

2. Add 0.9 μL of TAMRA-DBCO stock to 29.1 μL of ligated LAP-stGFP that has been dialyzed to remove excess 10-azidodecanoic acid (Step 3 in section 3.3). Make similar samples for all of the negative controls described in step 2 of section A.3.3.
3. Incubate samples for 5 min at room temperature.
4. Remove excess dye by dialyzing against 4 L of 50 mM phosphate buffer (pH 7.0). This step does not need to be performed before SDS-PAGE characterization as the free dye will migrate away from the protein on the gel. However, it may be necessary depending on the intended final application of the fluorescently labeled product *i.e.* single molecule protein tracking.[64]
5. Characterize fluorophore attachment *via* mass spectrometry (section A.3.5.1) or qualitatively *via* SDS-PAGE (section A.3.5.2).

A.3.4.4 Azide Mediated Chemical Functionalization: Copper Catalyzed Glycosylation and Fatty Acid Modification

1. For a 1 mL total volume glycosylation reaction, mix 856 μL of ligated LAP-stGFP, 20 μL of 20 mM propargyl α-D-mannopyranoside (dissolved in ddH_2O), 15 μL of $CuSO_4$-THPTA solution, 50 μL of 100 mM aminoguanidine hydrochloride (dissolved in ddH_2O), and, at the very end to initiate the reaction, 50 μL of 100 mM sodium ascorbate (dissolved in ddH_2O).
2. Incubate for 4 h at 37 °C.

3. Remove unreacted components by dialyzing against 4 L of 50 mM sodium phosphate buffer (pH 7.0) at 4 °C and store at 4 °C if needed before characterization or use. Dialyze against 4 L of pure ddH_2O at 4 °C for samples that are to be characterized by MALDI mass spectrometry.
4. For a 1 mL fatty acid conjugation reaction, follow steps 1-3 but replace the 20 μL of propargyl α-D-mannopyranoside solution with 20 μL of 20 mM 15-hexadecynoic acid (dissolved in DMSO) and add sodium deoxycholate to a final concentration of 2% (v/v). Remove unreacted fatty acid by dialysis as described above for the glycosylation product, dialyzing in buffer for storage or use and into ddH_2O if in preparation for MALDI MS characterization.

A.3.5 Analysis and Characterization of Ligation and Modification Products

A.3.5.1. Electrospray Ionization (ESI) Mass Spectrometry

1. Characterize ligation reaction products and fluorophore label modification reaction products with ESI MS.
2. Prepare ESI samples by adding 0.5-1 mL of ligation or fluorophore-labeled product to 5-10 mL of infusion solution.
3. Transfer this mixture to a 10000 MWCO ultrafiltration tube and centrifuge at 4000 x g in a swinging bucket rotor until the volume in the concentrator above the membrane has been reduced to ~200 μL.

4. Add 2 mL of pure infusion solution to the 200 μL of sample above the membrane and discard the solution below the membrane that has flown through during centrifugation. Mix well and centrifuge the samples again until the volume above the membrane is again approx. 200 μL.
5. Repeat step 3 five times to reduce the concentration of all buffer components by 6 orders of magnitude (*see* **Note 15**).
6. Inject samples into the ESI mass spectrometer using a 50-μL glass syringe. Proteins will be positively charged due to the formic acid in the infusion solution, and consequently the instrument must be set to detect positively charged molecules. Injections can be performed by hand using flow rates of about 2-10 μL/min. Record spectra during injections over the course of 1 minute using a 1 s scan time and a mass range of 500-3000 m/z.
7. Deconvolute ESI spectra with MaxEnt software using a resolution of 1 Da and setting the maximum number of iterations to 15. Ligation of 10-azidodecanoic acid should result in a mass increase of 194 Da (*see* **Note 16**). Compare peaks for ligated and un-ligated proteins to estimate the percent conversion of the reaction.

A.3.5.2 Matrix-Assisted Laser Desorption/Ionization (MALDI) Mass Spectrometry

1. Characterize PEGylation, glycosylation, and fatty acid modification reactions with MALDI MS (*see* **Note 17**).

2. Dialyze samples of all protein modification reaction products (from PEGylation, glycosylation, and fatty acid modification reactions) and a sample of ligated but unmodified LAP-stGFP as a negative control against 4 L of ddH_2O at 4 °C.
3. Add trifluoroacetic acid (TFA) to these samples for a final TFA concentration of 0.1% (v/v).
4. Add 5 μL of each product-TFA solution to 5 μL of succinic acid matrix solution, and mix thoroughly by pipetting up and down.
5. Drop 1-2 μL of each sample mixed with matrix solution onto a gold MALDI plate and allow these samples to dry thoroughly before analyzing.

A.3.5.3 SDS-PAGE

1. Characterize PEGylated and fluorophore labeled protein constructs by SDS-PAGE.
2. Add 10 μL SDS loading dye to 30 μL of each modified protein sample being characterized.
3. Incubate at 95 °C for 10 min (*see* **Note 18**).
4. Load an SDS-PAGE gel into a gel apparatus. Dilute 100 mL 10x Tris/Glycine/SDS buffer in 900 mL ddH_2O and add this solution to the gel apparatus.
5. Load 10 μL of each sample on the gel along with a fluorescently labeled protein ladder to serve as a molecular weight reference.
6. Run the gel at 150 V for ~55 min (*see* **Note 19**).

7. If characterizing fluorophore labeled protein, image the gel with a Typhoon imager set with a 532 nm excitation wavelength and perform densitometry measurements for the protein bands with ImageJ software (**Figure A.2**) (*see* **Note 20**).[156]
8. If characterizing PEGylation products, place the gel in 100 mL of coomassie stain solution and microwave for 1 min.
9. Incubate the gel in the warm stain for 10 min.
10. Transfer the gel into ddH_2O and microwave for 5 minutes to de-stain. Repeat as necessary for a high quality image.

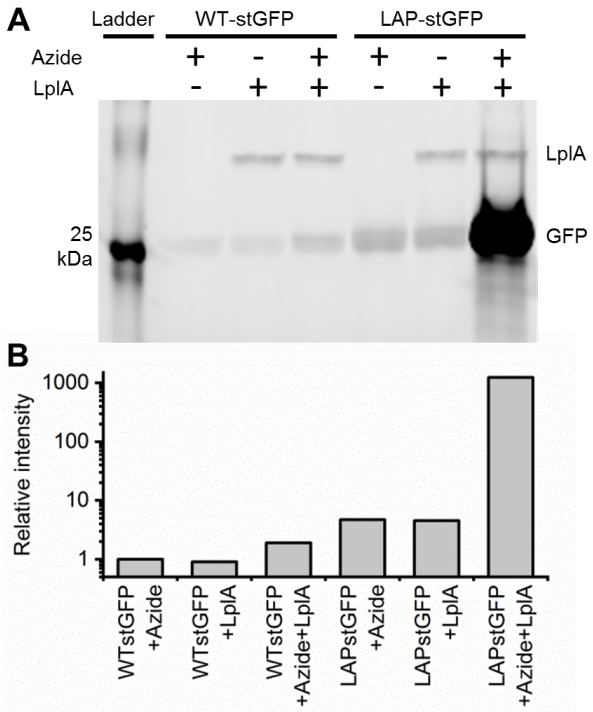

Figure A.2. Labeling with TAMRA-DBCO as a proxy to characterize 10-azidodecanoic acid ligation.
SDS-PAGE gel of TAMRA-DBCO-reacted LAP-stGFP with WT-stGFP and azide and $LplA^{W37V}$-free negative controls (A). Quantification of gel bands for GFP constructs (B). Successfully ligated protein reacts with TAMRA-DBCO, resulting in a fluorescence signal ~2 orders of magnitude greater than that of the negative controls.

A.4. Notes

1. Although storage buffer components can be removed by dialysis before performing the ligation reaction, tailoring the storage buffer of the LAP-target protein to be compatible with LplA^{W37V} ligation conditions is convenient for streamlining the ligation process. In particular, a pH of near 7.2 should be used and the concentration of reducing agents, which negatively impact azides, should be kept to a minimum.
2. There is a T5 promoter for LplA on the pYFJ16-LplA(W37V) plasmid, and as such, DE3 cell lines that contain T7 RNA polymerases are not absolutely required. The plasmid has a high copy number, producing high DNA yields and high protein titers when expressed. The plasmid has a lac operator, making it IPTG inducible, and produces LplA^{W37V} with an *N*-terminal polyhistidine-tag.
3. At this point, cell pellets may be frozen and stored at -80 °C or transferred directly to lysis buffer for purification. However, once resuspended and lysed LplA^{W37V} should be immediately purified to prevent digestion by endogenous *E. coli* proteases and other forms of degradation.
4. Because a very small amount of LplA^{W37V} can be used to label a large amount of target protein, small volume aliquots (<50 μL) are preferred. For best results, aliquots should only be put through one freeze-thaw cycle. However, we have observed residual LplA activity even after three cycles.
5. Different LAP insertion sites within a target protein may have different labeling efficiencies and may differentially impact the stability/behavior of the target protein. This is especially true when the LAP sequence is inserted at internal positions within a protein

rather than at one of the termini. Nonetheless, internal LAP insertion sites that do not substantially interfere with protein folding have previously been identified within superfolder GFP, *Bacillus subtilis* lipase A, *Trichoderma reesei* β-glucosidase, and human PTEN. Accommodating sites within these proteins tend to be unintuitive and difficult to identify *a priori* on the basis of *B*-factor and secondary structure. In order to address this, we have developed a computational approach to predict accommodating internal LAP insertion sites that utilizes the Rosetta Protein Modeling Software Suite (results in preparation for publication). Still, it is recommended that in the event a terminal conjugation site is undesirable, multiple protein constructs with internal LAP insertions should be built to increase the likelihood of identifying a site that accommodates the insertion.

6. Expression and purification of the LAP-modified target protein will necessarily vary depending on the properties of the target protein being modified. Additionally, LAP-target proteins may exhibit different behavior than their wild-type counterparts (the target protein without an inserted LAP sequence). For example, the LAP sequence contains several hydrophobic residues, which may reduce the colloidal stability of proteins that have large hydrophobic patches as well as influence elution times from a hydrophobic column. The LAP sequence also has a net negative charge, which alters the pI of the target protein to which it is fused or inserted and will consequently alter loading and elution conditions on ionic exchange columns.
7. Wild-type stGFP behaves very similarly to LAP-containing versions when expressed and purified on a nickel column. Having a “wild-type” control for the ligation reaction is

useful as it demonstrates that 10-azidodecanoic acid is in fact being ligated to the LAP sequence and not to some other region or residue within the protein.

8. For convenience, 10-azidodecanoic acid can be dissolved in DMSO before adding it to the ligation reaction. Specifically, 100x stock solutions of 10-azidodecanoic acid were generally prepared in DMSO. Although this resulted in a 1% (v/v) concentration of DMSO in the final ligation reaction, a negative impact on ligation efficiency was not observed. When preparing 10-azidodecanoic acid and when handling proteins that have been labeled, care should be taken to avoid excessive exposure to light to prevent UV-induced degradation of the azide moiety.
9. It is important to thoroughly remove excess 10-azidodecanoic acid as contaminating azide will compete with functionalized proteins in subsequent azide/alkyne click reactions, reducing conjugation reaction efficiencies.
10. Piranha solution should be made fresh and treated with extreme care. The solution will be warm due to the heat of dilution associated with mixing sulfuric acid and hydrogen peroxide. The incubation of the cover slides is performed at this elevated temperature.
11. At this point in the procedure, the plates become sided; the side of the plate facing up in the UV-ozone machine will be the side of the plate that is functionalized with DBCO and reacted with 10-azidodecanoic acid-ligated protein.
12. Non-specifically adsorbed protein can be difficult to remove from certain materials including glass. It is important to perform negative control immobilization reactions using protein that has not been functionalized with an azide to confirm that un-reacted protein is indeed being removed. Lowering the concentration of protein in the

immobilization reaction may help to reduce irreversible, non-specific adsorption. Additionally, the architecture of the solid support can make a difference in how easily non-specifically adsorbed protein is removed. For example, we have found that removing excess protein from a two-dimensional glass cover slide is easier than removing it from nanoparticles.

13. Excess PEG-DBCO can interfere with characterization by MALDI MS, and its removal may be required for particular applications of protein-PEG conjugates. In our experience, thorough removal of excess PEG is difficult to achieve with dialysis alone. Interestingly, we have found that some 6x-polyhistidine-tagged proteins maintain Ni-binding affinity after being covalently modified with polymers and that, consequently, a Ni-column may be used to remove excess, unreacted polymer.[184] In addition to MALDI MS, protein PEGylation products can be characterized by observing mobility shifts on SDS-PAGE gels depending on the size of the attached PEG molecules. This type of characterization does not require the removal of excess PEG. However, because PEG does not migrate in the same manner as a protein, it is not possible to determine an absolute mass change from an apparent one. Therefore, this approach cannot be used to determine the number of attached PEG molecules in situations where more than one LAP sequence has been introduced into a single protein.
14. TAMRA-DBCO dissolves easily in DMSO, and we have found that the stock solution, which was stored at -20 °C between use, tolerates freeze thaw cycles well in terms of maintained DBCO reactivity.

15. Contaminating buffer components must be thoroughly removed as salt adducts may form with the protein and interfere with ESI. Generally, protein concentrations between 10-100 μM are detectable by MS. Lower concentrations are preferred as they reduce the amount of time needed to wash the lines between sample injections. GFP loses fluorescence in the infusion solution, presumably due to unfolding, but is, nevertheless, very soluble.
16. One of the reasons to prepare a negative control of un-ligated protein as described in step 2 of section A.3.3 is to have an experimental value to compare with the ligated product. It is easier to measure a change in mass that corresponds to 10-azidodecanoic acid ligation than it is to confirm a calculated mass because the protein might not have the same mass as that which is calculated from its primary sequence. For example, *N*-terminal methionine cleavage of proteins with small residues directly adjacent to the start methionine can occur post-translationally when those protein are expressed in *E. coli*, altering the mass of the target protein. We observed this previously with a GFP construct that had a glycine residue directly following the start methionine. Additionally, in the case of GFP, chromophore formation results in a mass decrease that can be observed by ESI. Having an unligated sample also makes it easier to identify the peak corresponding to unligated protein in the ligated sample spectra should the ligation not be quantitative. This peak must be identified in order to calculate a conversion (percentage) for the ligation reaction.
17. MALDI MS is a good characterization approach for these reactions because it is less sensitive to a sample's polydispersity than ESI. Although proteins are very consistent in terms of mass since translation utilizes an mRNA template, PEG molecules are usually

polydisperse. When polydisperse PEG's are covalently conjugated to a protein, the mass of the protein-PEG conjugate also becomes disperse. When using ESI MS, the unligated protein peak is unaffected after such a reaction, but the product is detected as many different peaks, corresponding to the masses of the different length PEG molecules that have been attached. In the case of MALDI, where generally only one charge state is visible for proteins, this effect results in peak broadening. In the case of the much more sensitive ESI, the spectra becomes very difficult to interpret. Interestingly, we have observed this phenomenon not just with protein PEGylation, but also when comparing ESI and MALDI spectra for glycosylation and fatty acid modification reactions. Although the molecules being attached in these cases were not polydisperse, side reactions between copper and residues in the protein may have occurred, leading to apparent polydispersity.

18. stGFP is amazingly stable. Consequently, reducing the temperature or incubation time for these samples results in poor gel mobility, likely from incomplete protein unfolding and SDS binding.
19. Excess TAMRA-DBCO if not previously dialyzed away will run through the gel along with the dye front. To improve the quality of the fluorescent image of the gel, insure that all excess dye, which appears pink on the gel, runs off of the bottom. If the dye does not run completely off the gel, remove it by cutting off the bottom of the gel containing the dye front with a razor blade.
20. Reacting the protein conjugate with a fluorescent dye and characterizing it *via* SDS-PAGE can serve as a proxy for ESI characterization of the ligation reaction. It is much more qualitative than analysis by MS and cannot be used to determine ligation

efficiencies. However, it is extremely simple to perform, and successfully ligated constructs tend to have fluorescent signals that are at least an order of magnitude above background.

APPENDIX B: ROSETTA SCRIPTS

B.1 Preparing Input Files

Rosetta requires a unix-based operating system. It can be run on a PC but only through a virtual machine that is running a unix-based operating system. The following instructions and scripts are for running the software on a MacBook Pro.

1. Download Rosetta source code that includes binaries for MacBook Pro and unpack the directories in a designated area of your computer.
2. Create a working directory, which will contain all relevant input files as well as any output files (such as score files and output PDBs) created as a result of modeling.
3. In the command line, use shell scripting to navigate to the working directory.

B.1.1 Target Protein Structure Files

1. Download the PDB file for the target protein that is to be scanned in order to identify potential loop insertion sites.
2. Run the Rosetta "clean_pdb" script in order to remove water molecules and other extraneous details from the PDB file by typing the following into the command line:

```
/path-to-rosetta/tools/protein_tools/scripts/clean_pdb.py
target.pdb A
```

The letter "A" designates that chain A within the PDB will be preserved in the cleaned output PDB file. Chain A should therefore be the protein subunit or monomer within the unit cell that will be scanned. If an error is returned from running this script it may be

overcome by removing ".pdb" from the input file argument. Another common error that is encountered can be overcome by moving the "rosettautil" directory, which is found in the rosetta/tool/protein_tools directory, to the protein_tools directory.

3. In addition to outputting the cleaned PDB file (named target_A.pdb), the "clean_pdb" script will also output a fasta sequence file that contains the protein primary sequence derived from the input structure file (named target_A.fasta). The contents of this file for lipA are as follows:

```
>1ISP_A
EHNPVVMVHGIGGASFNFAGIKSYLVSQGWSRDKLYAVDFWDKTGTNYNNGPVLSRFVQKVLDET
GAKKVDIVAHSMGGANTLYYIKNLDGGNKVANVVTLGGANRLTTGKALPGTDPNQKILYTSIYSS
ADMIVMNYLSRLDGARNVQIHGVGHIGLLYSSQVNSLIKEGLNGGGQNT
```

This file will be required for generating an alignment file, which is used to introduce the LAP sequence into the target protein before modeling.

B.1.2 Loop-Construct Sequence Files

1. Create a library of sequence files that contain the inserted peptide loop sequence (in this case the LAP sequence) inserted at every possible position within the primary sequence of the target protein. This can be done by running the following python script:

```
python insert.py arg1 arg2
```

Where arg1 is the entire target protein sequence (which must be pasted into the command line) and arg2 is the sequence of the peptide loop insertions (GFEIDKVWYDLDA in the case of the LAP sequence). The python code in the insert.py file is as follows:

```
import csv
import sys

name=sys.argv[1]
seq = sys.argv[2]
newname = ""
textfile = []

for i in range(len(name)):
    newname = name[:i] + seq + name[i:]
    with open("insertAt"+str(i)+".txt","w") as text_file:
              text_file.write(newname)
```

This code, which was kindly written by Garrett Chado, will output individual sequence files for loop insertions at each possible site (between every possible residue) within the protein. They will be named for the residue number directly preceding the loop insertion. For example “insertAt23” will be the protein sequence with a loop insertion between residues 23 and 24.

2. Using the above library of sequence files, create a library of clustal alignment input files. The following script is a for loop that uses shell scripting to generate a clustal alignment input file for every possible LAP insertion site. It first creates a unique directory for every possible insertion site, which are numbered for the residue directly proceeding LAP insertion. This is the directory where modeling will eventually take place for each LAP insertion site construct. It then writes a clustal input file, which contains the sequence of the protein with the LAP insertion generated from the python script in the above step with a preceding descriptive label “>LAP”, as well as the contents of the fasta file generated from the “clean_pdb” script.

```
for var in {1..#amino acids in protein minus 1}
do
mkdir $var
echo ‘>LAP’ > $var/$var.fasta
cat sites/insertAt$var.txt >> $var/$var.fasta
cat sites/1ISP_A.fasta >> $var/$var.fasta
```

```
done
```

The output files from this script will look like the example bellow for LAP insertion between the first and second residues of lipA. They will be named with a number corresponding to the residue preceding the inserted LAP sequence followed by ".fasta".

```
>LAP
EGFEIDKVWYDLDAHNPVVMVHGIGGASFNFAGIKSYLVSQGWSRDKLYAVDFWDKTGTNYNNGP
VLSRFVQKVLDETGAKKVDIVAHSMGGANTLYYIKNLDGGNKVANVVTLGGANRLTTGKALPGTD
PNQKILYTSIYSSADMIVMNYLSRLDGARNVQIHGVGHIGLLYSSQVNSLIKEGLNGGGQNT
>1ISP_A
EHNPVVMVHGIGGASFNFAGIKSYLVSQGWSRDKLYAVDFWDKTGTNYNNGPVLSRFVQKVLDET
GAKKVDIVAHSMGGANTLYYIKNLDGGNKVANVVTLGGANRLTTGKALPGTDPNQKILYTSIYSS
ADMIVMNYLSRLDGARNVQIHGVGHIGLLYSSQVNSLIKEGLNGGGQNT
```

3. Download the source code for clustal sequence alignments from http://www.clustal.org/omega/
4. Once the clustal source code is downloaded, alignments can be performed through the command line using the above .fasta clustal input file as the argument as shown below.

```
~/path-to-clustal-code/clustalw2 LAP-insert-site-number.fasta
```

5. Write a for loop to generate a clustal alignment file (.aln) for every possible LAP insertion site. An example is shown below. It generates a .aln file in all of the individual directories that were made for the various LAP insertion sites.

```
for var in {1..#amino acids minus 1}
do
cd ~/working-directory/$var
~/pat-to-clustal-code/clustalw2 $var.fasta
cd ~/working-directory
done
```

B.1.3 Loop-Construct Structure Files

The .aln files from above are used to create a Rosetta input PDB file that has the LAP-containing primary sequence threaded onto the structure of the cleaned input PDB. This is done

using the “thread_pdb” python script that is included in the Rosetta source code. The atoms for the inserted residues of the LAP sequence will be assigned coordinate values of zero until modeling is completed, which is when they are given real coordinates. The for loop below will generate a threaded.pdb file for each LAP insertion site in its corresponding directory, again using the lipA scan as an example.

```
for var in {1..#amino acids minus 1}
do
cd ~/working-directory/$var
python /path-to-rosetta/tools/protein_tools/scripts/
thread_pdb_from_alignment.py --template=1ISP_A --target=LAP --
chain=A --align_format=clustal $var.aln ~/Computation/LAP/Lipase/
scan/sites/1ISP_A.pdb $var-threaded.pdb
cd ~/working-directory
done
```

The input files for this script are the .aln files generated by the clustal alignment and the cleaned PDB file of the target protein structure. The output is the threaded PDB file. In the case of the above for loop, each is named with the position of the LAP insertion site followed by “-threaded.pdb”. The threaded PDB file is one of two input files required for loop modeling.

B.1.4 Loop Files

Along with a threaded PDB input file, loop modeling requires a “loop file” (.loop) as an input, which identifies the residues that are to be modeled as part of the loop and differentiates them from the rest of the protein structure. The text from an example .loop file is below:

```
LOOP 1 15 0 0 1
```

This loop file was written for a threaded input PDB that has the LAP sequence inserted between residues 1 and 2 of the target protein. The first two numbers in the loop file, 1 and 15, designate the residues flanking those within the loop that are to be modeled. There are 13 residues between

residues 1 and 15 and they correspond to the LAP sequence. The following for loop was written to generate .loop files for all of the various LAP insertion sites modeled in our structural scan of lipA in the corresponding directory using shell scripting.

```
for var in {1..#amino acids minus 1}
do
cd ~/working-directory/$var
echo LOOP $var $[$var+14] 0 0 1 > $var.loop
cd ~/working-directory
done
```

B.2 Modeling

B.2.1 Initial Structural Scan

At this point, there should be a unique directory for every possible LAP insertion site (LAP insertion between every two residues) of the target protein, which in this case is lipA. These directories each contain a unique threaded PDB file and .loop file specifically made for the LAP insertion site that is being modeled in the directory in which they are found. With these directories and files, the initial structural scan of the target protein can be performed.

The kinematic loop modeling application is used for modeling at each site. The executable file for this application should already to be built in the rosetta/main/source/bin directory if the source code was downloaded with binaries for Mac as described above. Modeling the LAP sequence at each of the sites requires calling this file with the necessary options while in the specific directory that contains the custom threaded PDB and .loop files. The options that are included specify the location of the Rosetta database as well as the requisite input files generated above, indicate which stages of modeling to perform (centroid or full-atom or both), and indicate how many build and modeling attempts should be performed in order to find a successful model.

```
~/path-to-rosetta/main/source/bin/loopmodel.macosclangrelease -
database ~/path-to-rosetta/main/database -loops:remodel
perturb_kic -in:file:s 1-threaded.pdb -in:file:fullatom -
loops:loop_file 1.loop -max_kic_build_attempts 100 -max_retry_job
1 -nstruct 1
```

The above command will generate a single model (-nstruct 1) after having attempted to generate a loop conformation no more than twice (-max_retry_job 1) with 100 build attempts permitted for each attempt (-max_kic_build_attempts 100). Importantly, the -loops:refine refine_kic option that is usually specified for the kinematic loop modeling application to enable the full-atom modeling stage is omitted here so that only the centroid stage is used, which runs in response to the "-loops:remodel perturb kic" option.

By running the above command as part of a for loop, modeling can be attempted at every possible LAP insertion site in a target protein. Modeling will fail at sites that are buried (a successful model will not be generated after 100 build attempts) and the for loop will proceed to the next site. The above command was written for LAP insertion between residues 1 and 2 of lipA. The for loop below can be used to scan through the entire structure of lipA by moving from one individual LAP insertion site directory to another and running the loop modeling command in each.

```
for var in {1..#amino acids minus 1}
do
cd ~/working-directory/$var
~/path-to-rosetta/main/source/bin/loopmodel.macosclangrelease -
database ~/path-to-rosetta/main/database -loops:remodel
perturb_kic -in:file:s $var-threaded.pdb -in:file:fullatom -
loops:loop_file 1.loop -max_kic_build_attempts 100 -max_retry_job
1 -nstruct 1
done
```

B.2.2 Final Scan

After running the above for loop and attempting to build a single model for each LAP insertion site in a protein, there will be many sites for which modeling failed. No output files will be generated in the directories corresponding to these sites. Those sites at which modeling was successful (surface exposed sites) will contain score files and output structures in their directories.

Shell scripting can be used to generate a list of all directories that contain these output files or, more precisely, a list of all LAP insertion sites at which modeling was successful. Specify all of these sites in a new for loop that will only access directories at which modeling was successful but that is otherwise similar to the for loop used in the initial scan. By changing the value after the –nstruct option in this for loop from 1 to 9, nine additional models will be made at each of the sites for which a successful model can be built in 100 build attempts. The scores for these nine models will be appended to the score file that already exists in each of these directories and that contains the score from the first model that was produced.

B.3 Data Analysis

Shell scripting can again be used to concatenate all of the score files produced from the final scan into a single text file that can be transferred to Microsoft Excel for further analysis. The average total scores from the 10 models produced per site should be used to characterize each unique insertion site. Standard deviations should be used to calculate 95 % confidence intervals.