Title

Bioconjugation and Protein Engineering for the Development of a Peptide-Protein Conjugate Vaccine and Characterization of an N-terminal Modification Reaction

Author

Wucherer, Kristin N

Publication Date

2019

Peer reviewed|Thesis/dissertation

Bioconjugation and Protein Engineering for the Development of a Peptide-Protein Conjugate Vaccine and Characterization of an N-terminal Modification Reaction

by

Kristin N. Wucherer

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Matthew B. Francis, Chair
Professor Jamie H. D. Cate
Professor Sarah A. Stanley

Fall 2019

Bioconjugation and Protein Engineering for the Development of a Peptide-Protein Conjugate Vaccine and Characterization of an N-terminal Modification Reaction

Kristin N. Wucherer

Abstract

Bioconjugation and Protein Engineering for the Development of a Peptide-Protein Conjugate Vaccine and Characterization of an N-terminal Modification Reaction

by

Kristin N. Wucherer

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Matthew B. Francis, Chair

Post-translational protein modification accounts for a significant amount of biodiversity and is essential for many cellular processes. The development of techniques to mimic native biomolecule modification have evolved into the field of modern bioconjugation. The complementary use of both genetic and chemical methods has provided a large toolbox for an endless possibility of potential bioconjugate constructs, using a wide-variety of synthetic and biologically-derived materials. To this end, reproducing these natural modifications of biomolecules provides researchers a way to interrogate and elucidate the intricate functions within biological systems.

Within this bioconjugation toolbox, there exist a large number of different chemical reactions for protein modification. The site-specific covalent link between a protein and synthetic moiety, such as a drug or fluorophore, enables the creation of hybrid material that capitalizes on the properties of both individual components. Thus, bioconjugate materials have a wide variety of applications, such as the study of proteins in a biological context, the elucidation of a multi-protein quartenary structure, creating unique protein-based materials, the development of improved therapeutics, and many more.

One application of site-specific chemical modification of proteins is the development of conjugate vaccines, as described herein. Synthetic vaccines offer great promise as useful therapeutics; however, often individual moieties suffer from poor delivery or weak immunogenicity. Conjugation to a carrier protein can circumvent this issue. Work presented describes the use of cross-reactive material 197 (CRM_{197}) as a carrier protein for the presentation of therapeutic peptide cargo. Heterobifunctional linkers, comprised of orthogonal lysine-reactive and cysteine-reactive handles, were used to modify lysine residues of CRM_{197} to attach cysteine-containing peptide therapeutics. Ultimately, the bioconjugation strategies explored led to a structurally heterogeneous conjugate material. As structure-immunogenicity relationships exist, we turned to protein engineering of CRM_{197} to facilitate the creation of structurally homogeneous conjugate material. Work is ongoing to prepare and characterize the resulting peptide-protein conjugates.

In creating synthetic vaccines, multiple sites of protein modification are often necessary to enable proper immune response. However, often it is desirable to have a site-specific, single modification on a protein of interest. There remains a need for the development of more chemoselective chemical modification of proteins that are mild, efficient, and robust. As a result, novel protein modification techniques target uniquely reactive sites, such as the N-terminal amine, due to its unique environment and pKa. Our group recently reported a single-step N-terminal modification with 2-pyridinecarboxaldehyde (2PCA), which proceeds under physiological conditions. However, certain N-terminal residues were found to have different reactivity and stability toward 2PCA modification. Thus we set out to characterize the reaction mechanism in order to understand this relationship, combining computational analysis, NMR of 2PCA modified peptides, and mass spectrometry on 2PCA modified proteins. With this multifaceted approach, several N-terminal residues were found to strongly promote stable 2PCA modification. Further, we are exploring the key attributes promoting product formation in order to create second generation 2PCA derivatives that will control reaction outcome.

Dedicated to my greatest supporter and motivator, my mother, Yoshiko Otonari.
1961-2015

Contents

List of Figures

Acknowledgments

First and foremost, thank you to Prof. Matt Francis for being an excellent graduate school advisor. Your support and guidance has shaped me into a creative researcher, critical thinker, scientific mixologist, and overall, a more confident person. I'm not sure how you do all that you do, mainly subsisting on KIND bars and coffee, but I am very appreciative and grateful to have you as a mentor!

Thank you to all members of the Francis group for subgroup brainstorming, experimental advice, fun group trips, and a plethora of baked goods. Special thanks to Jim MacDonald and Jake Jaffe for mentorship during the beginning of my graduate school career. Also thanks to Emily Hartman, Jing Dai, Sarah Klass, Amanda Bischoff, and Ariel Furst for being especially supportive labmates. To my 733 roommates (Ioana Aanei, Kanwal Palla, Joel Finbloom, Tyler Hurlburt, Dan Brauer, Celine Santiago, and Wendy Cao, and undergrads Janie Honda and Zoe Merz): whether it be troubleshooting or simply getting coffee, lunch, coffee, cupcakes, coffee, tea, or maybe more coffee (...Dan...), y'all made lab very fun and worthwhile. Special thanks to project collaborators Celine Santiago, Nick Dolan, Byungjin Koo, Ben Horst (Marletta Lab), Madeleine Jensen (Marqusee Lab), and Charlotte Nixon (Marqusee Lab). Also, big thank you to Christine Baolong for maintaining the lab!

Thank you my friends in the 2014 (and 2013) graduate student cohort. I appreciate the help with Bergman's phys org beast, offering instruments/reagents, short elevator conversations, Monday night soccer (go Soctopus!), and many camping/skiing trips. Special heartfelt thanks to classmates Julia Lazzari-Dean and Alice Kunin, who I am fortunate to have fallen upon as housemates. From squid-hat cake-baking kitchen gatherings, to being supportive scientists, caregivers, and listeners, you both kept me afloat (sometimes literally) through the darkest of times and I truly cherish our friendship. Another special thank you goes to Steven Boggess, mon nounours, Super Smash Bros. teacher, and porch gardening and 'bu-brewing partner. Your endless love, patience, and support is so appreciated; I'm excited for our future scientific and non-scientific endeavors.

Final and most heartfelt thank you goes to my family. To my Ji-chan, Uncle Tom, and Uncle Gary, thank you for always looking out for me. To my sisters, Kelly and Karoline, thank you for the sisterly love and gooberness, reminding me of mom's legacy, and ultimately motivating me to be a better person. To my parents, Yoshiko Otonari and E.J. Wucherer, thank you for challenging me throughout my life. I have been blessed with excellent parental role models; you have taught me to work and play hard, from school to soccer, and to truly cherish all the fun adventures in between. Special thanks to pops for your help (and patience) on all those chemistry and physics question phone calls, and for being a superb scientific mentor. I am where I am today because of both of your support and love.

Chapter 1

Development and Characterization of Peptide-CRM197 Conjugate Vaccines

The following is adapted from:
Jaffe, J., Wucherer, K., Sperry, J., Zou, Q., Chang, Q., Massa, M. A., Bhattacharya, K., Kumar, S., Caparon, M., Stead, D., Wright, P., Dirksen, A., Francis, M. B. (2019) *Bioconjugate Chemistry*, **30**, 1, 47-53.

ABSTRACT: Conjugate vaccines prepared with cross-reactive material 197 (CRM_{197}) carrier protein have been successful in the clinic and are of great interest in the field of immunotherapy. One route to preparing peptide-CRM_{197} conjugate vaccines involves an activation-conjugation strategy, effectively coupling lysine residues on the protein to cysteine thiolate groups on the peptide of interest using a heterobifunctional linker as an activation agent. This method has been found to result in two distinct populations of conjugates, believed to be the result of a conformational change of CRM_{197} during preparation. This report explores the factors that lead to this conformational change, pointing to a model in which the unintentional alkylation of histidine-21 by the activating agent promotes the "opening" of the monomeric protein. This exposes a new set of lysine residues that are modified by additional activation agents. Subsequent peptide ligation to these sites results the two conformers. This is the first time that a specific chemical modification is demonstrated to induce a defined conformational change for this carrier protein. Importantly, alternative conditions and reagents have been found to minimize this effect, improving the conformational homogeneity of peptide-CRM_{197} conjugates.

1.1 Introduction

Conventionally, vaccines have been prepared from live attenuated pathogens or inactivated viruses [1]. While these pathogen-derived products have had a profound effect on global health [2], they can present challenges in terms of production and safety [3]. To elicit more specific and predictable immune responses, new technologies have allowed for the rational design of novel and safer vaccines based on synthetic constructs prepared using bioconjugation techniques [4]. These "conjugate vaccines" are composed of hapten molecules (generally B-cell epitopes) that are covalently attached to a protein carrier, which provides a source of T-cell epitopes [5]. Many safety issues can be avoided because these materials are not derived from live pathogens, and hapten-carrier conjugates can be prepared for more diverse targets than existing vaccines. Beyond bacterial and viral targets [4], conjugate vaccines can also be developed against non-infectious diseases, including cancer [6], Alzheimer's disease [7], hypertension [8], and addiction [9–11].

A variety of carrier proteins have been used for the production of conjugate vaccines, including diphtheria toxoid, diphtheria toxin (DT) cross-reactive material 197 (CRM_{197}; DT G52E), tetanus toxoid (TT), keyhole limpet hemocyanin (KLH), and virus-like particles (VLPs) [5, 12]. Of these, CRM_{197} is the most common carrier used in research studies and in clinically available conjugate vaccines. Advantages of CRM_{197} include its inherent lack of toxicity and, as a result, no need for chemical crosslinking for detoxification [13]. Further, CRM_{197} has apparent lower susceptibility than DT or TT to pre- or co-commitment immunization with other vaccines [14], and the lack of lysine residues in its mapped T-cell epitopes is also advantageous for bioconjugation strategies [15]. Significant efforts have been directed toward the preparation CRM_{197} conjugates by targeting the 39 lysine residues and the N terminus of the carrier through reductive amination or acylation. Carboxylate containing residues, tyrosines, and cysteines have also been targeted for selective modification [16]. In addition to preparing conjugate materials, these studies have demonstrated structure-immunogenicity relationships that are dependent on hapten loading, conjugation chemistry, and the site(s) of modification [16]. However, similarly rigorous studies are rare in the scientific literature for peptide-CRM_{197} conjugates, even though some of these materials are promising vaccine candidates [17–19].

One method for accessing peptide-CRM_{197} conjugate vaccines involves an activation-conjugation procedure in which solvent-accessible lysines in CRM_{197} are first activated with bromoacetic acid *N*-hydroxysuccinimide ester (BAANS, a bromoacetylating agent, **Figure 1.1a**) [20]. Cysteine-containing peptide epitopes of interest are then conjugated to the introduced bromoacetamides through S_N2 chemistry. Finally, unreacted bromoacetamide groups are capped with *N*-acetylcysteamine (NAC) to provide the final conjugate material. Typically, peptide densities for these conjugates range between an average of 5 and 20 peptides per CRM_{197} carrier protein.

Interestingly, characterization of such conjugates by size-exclusion chromatography (SEC) consistently and unexpectedly reveal these conjugates to be structurally non-homogeneous, with up to three distinct populations being observed (**Figure 1.1b**). One might expect these distinct species to have differing stabilities, solubilities, and efficacies as vaccine components, and thus it is important to determine their structures and understand the reasons for their formation. Using a combination of bioconjugation chemistry, product characterization, and structural analysis, we have developed a new structural model that explains the formation of these distinct species. Most importantly, this insight has provided methods for preparing peptide-CRM_{197} vaccine conjugates with greater conformational homogeneity while maintaining sufficient peptide loading to elicit an immune response.

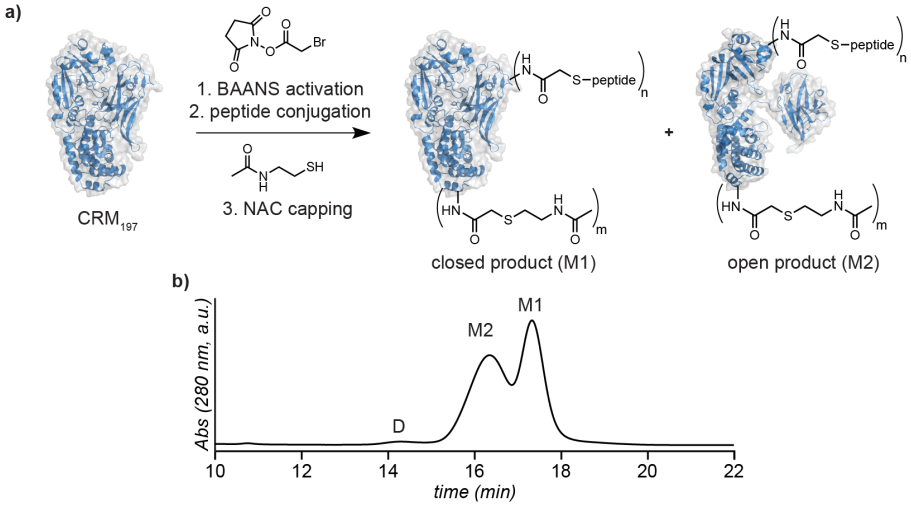


Figure 1.1: Preparation and SEC characterization of peptide-CRM_{197} conjugates. (a) Conjugates are prepared by first activating lysine residues with BAANS (1), and then conjugating a cysteine-containing peptide. Unreacted bromoacetamides are then capped with NAC (2). (b) Characterization of the resulting conjugate material by SEC (monitored at 280 nm) reveals the presence of two populations of monomeric conjugates (M1 and M2) and one dimeric form (D) of the conjugate. M1 and M2 are believed to be peptide-CRM_{197} conjugates in closed and open forms, respectively.

1.2 Results and Discussion

1.2.1 Conformational analysis of peptide-CRM197 material

These studies began with the preparation of CRM_{197} conjugates using a model peptide with the sequence Cys-Thr-Asn-Gln-His-Phe-Arg-Gly (CTNEHFRG) following the protocol outlined in **Figure 1.1a**. The peptide-CRM_{197} conjugates are purified via ultrafiltration. Subsequent analysis by electrospray time-of-flight mass spectrometry (ESI-TOF-MS) indicated a broad distribution of conjugates with an average 14 peptides per CRM_{197} carrier protein (**Figure 1.2a**, top spectrum). Analysis by SEC indicated the formation of two major species with similar abundance (M1 and M2, **Figure 1.1b**), in addition to a much smaller amount of a higher molecular weight species (D).

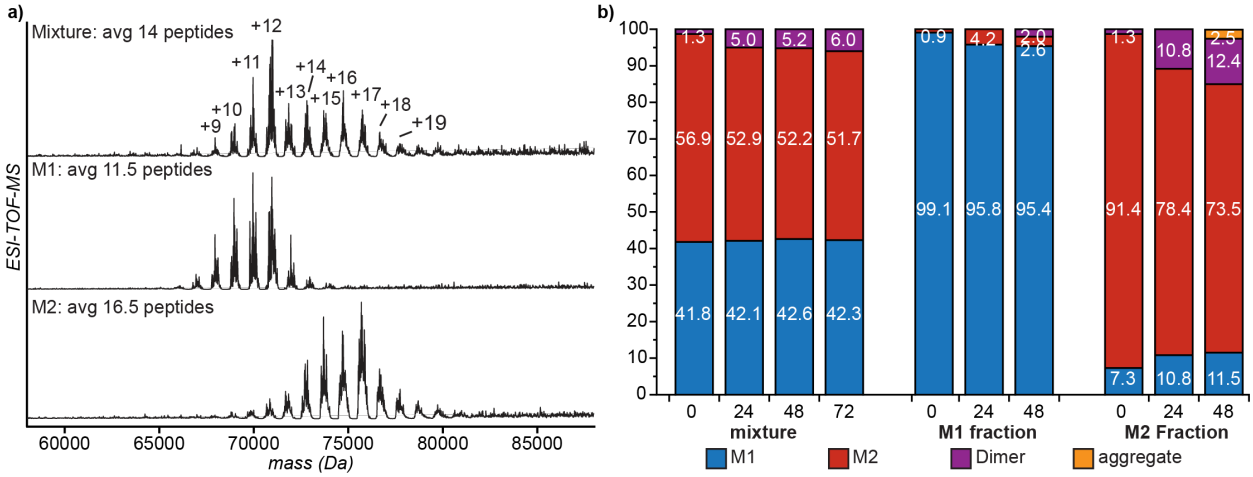
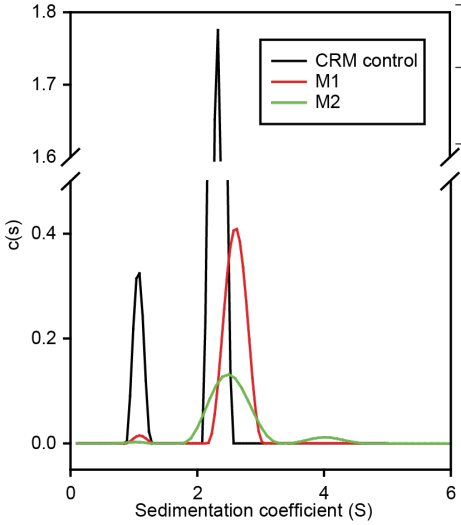


Figure 1.2: Stability and characterization of peptide-CRM_{197} conjugates. (a) A batch of peptide-CRM_{197} was prepared and characterized by ESI-TOF-MS as a crude product mixture (top) and as the M1 and M2 fractions after SEC separation. (b) Unfractionated conjugates, isolated M1 fractions, and isolated M2 fractions were stored at room temperature. The composition of each sample was monitored over time by SEC. Times (x-axis) are reported in hours.

The M1 and M2 species were isolated by SEC fractionation and independently analyzed by ESI-TOF-MS. Each species comprised a distribution of peptide conjugates, but with distinct average levels of modification (**Figure 1.2a**, middle and bottom traces). M1 was conjugated to 11.5 peptides on average, while M2 was conjugated to an average of 16.5 peptides. The two fractions were also found to have different solution stabilities. Purified samples of M1 and M2, in addition to a non-fractionated sample, were kept at room temperature and reanalyzed by SEC at 24 h intervals (**Figure 1.2b**). While M1 was found to be quite stable under these conditions, M2 was found to convert slowly to dimer (D). However, little-to-no interconversion between M1 and M2 was observed, as the appearance of small amounts of M2 in isolated M1 and small amounts of M1 in isolated M2 were considered to be within the error of the peak fitting procedure used to determine composition. These findings suggested that M1 and M2 are not in equilibrium, but rather species trapped in two distinct states that correspond to differing levels of modification.

To determine the approximate sizes of the M1 and M2 populations observed in the SEC trace, M1 and M2 fractions were analyzed by analytical ultracentrifugation (AUC, **Figure 1.3**). The AUC data indicated that M1 and M2 were still monomeric (i.e. smaller than 116 kDa expected for a dimeric species), but they differed in both size and shape. M2 was determined to be significantly larger and more elongated than M1, suggesting that a large conformational change had occurred, which cannot simply be explained by the higher average peptide density of M2 compared to M1. It is important to note that no properties of the M2 population were observed in samples of unmodified CRM_{197}, activated CRM_{197}, or activated CRM_{197} fully capped with NAC. The M2 population therefore results upon installation of the peptide moieties.



Sample	Sedimentation coefficient (S)	Monomer MW (Da)	Monomer frictional ratio (shape factor)	Stokes radius (nm)	Oblate (a/b)	Prolate (a/b)
CRM_{197}	3.52	57716	1.51	3.29	7.1	6.5
M1	3.95	75035	1.61	4.49	8.8	7.9
M2	3.84	79947	1.74	4.96	11.3	10.1

NOTE ON AUC PROFILE: The M1 fraction contains a monomeric species with 98% purity. The M2 fraction contains some aggregate and about 90% monomer. The majority of the aggregate in the M2 fraction is dimer (6.8% of total sample population). The frictional coefficient (f/f0) provides measurement of the maximum shape asymmetry from a sphere; a value of 1 would correlate to a compact sphere. For most globular proteins, the value will be around 1.2; thus the larger the value, the more asymmetric or elongated the molecule. Such, while M1 and M2 have similar molar mass, M2 has a f/f0 value of 1.74, compared to 1.61 for M1. This difference suggests that M2 is in a different conformation than M1, and is likely more elongated in shape. This is consistent with the observation in the SEC results that M2 elutes earlier than M1, suggesting that M2 has a much larger hydrodynamic volume. The unconjugated CRM_{197} has a smaller f/f0 value (1.51) than the conjugated M1 and M2, which suggests that the conjugation of peptides increases the asymmetry of the molecule. Assuming that the protein hydration is at 0.3 g of water per gram of protein, the calculated hydrodynamic shape equivalent to a prolate ellipsoid has an axial ratio of 10.06 for M2 monomer, which is larger than the calcuated axial ratio of 7.87 of M1 monomer, and 6.49 of unconjugated CRM_{197} monomer.

Figure 1.3: Analytical ultracentrifugation (AUC) data for CRM_{197} and the M1 and M2 species after modification. Data analyzed from AUC profile of the M1 (red) and M2 (green) fractions, along with the CRM_{197} control (black) sample is tabulated. Experiment and analysis by Qin Zou.

DT is known to undergo significant conformational changes during cellular entry [21]. Additionally, both DT and CRM_{197} are known to exhibit three-dimensional domain-swapping behavior [13, 22]. In domain-swapping, the "closed" interface between the receptor-binding domain (R) and the translocation (T) and catalytic (C) domains of the DT or CRM_{197} monomer is first disrupted to form an "open" monomer, wherein R is connected to T and C through a hinge-loop (**Figure 1.4**). Two open monomers can then interact to form a non-covalent dimer, wherein the R domain of each monomer forms a closed interface with the T and C domains of the other. The characteristic ability of CRM_{197} to "open" suggested that the M1 and M2 populations observed in peptide-CRM_{197} conjugates represent some form of closed and open monomers, respectively.

Earlier studies on conjugate vaccines have reported observations regarding conformational shifts of CRM_{197} upon conjugation. CRM_{197} glycoconjugates have been reported to take on more open conformation than unconjugated material [22–25]. However, distinct populations of open and closed monomers have not previously been observed through SEC characterization [14, 24]. In a study of an anti-nicotine vaccine, where a small molecule hapten was conjugated to CRM_{197} using *N*-hydroxysulfosuccinimide / 1-ethyl-3-(3-dimethylamino)propyl carbodiimide (sNHS/EDC) chemistry, two distinct populations were reported [9]. In that particular case, the conformational change was suggested to correlate with increased levels of crosslinking.

In the case of the peptide-CRM_{197} conjugate presented here, the structural hypothesis could suggest that after increasing the number of peptides per CRM_{197} to a certain point, the protein simply undergoes a conformational change. However, three pieces of evidence suggest

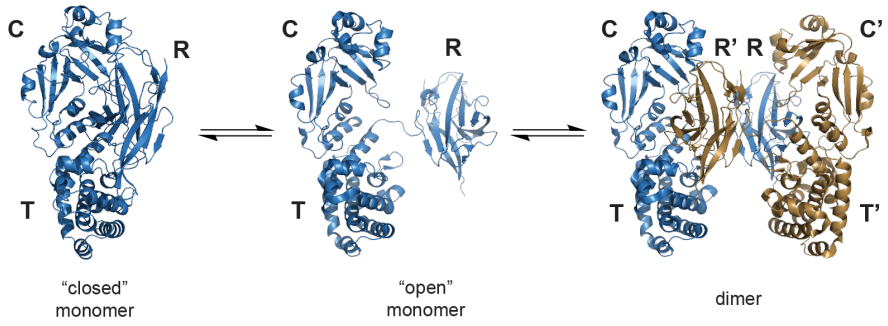


Figure 1.4: Conformational changes and dimerization of CRM_{197}. Disruption of the interface between the catalytic (C) and receptor-binding (R) domains of monomeric CRM_{197} results in "open" form monomer. The open monomer can revert to the closed form or interact with another open monomer to form a noncovalent dimer, where the R domain of one monomer forms an interface with the catalytic domain (C') of the other. The transmembrane domain is labeled as "T". Structures modeled from DT monomer structure (PDB ID: 1MDT) and CRM_{197} dimer structure (PDB ID: 4AE0).

that a more complex mechanism is occurring. First, the mass spectra of isolated M1 and M2 partially overlap (**Figure 1.2a**). For example, the 14-peptide conjugate can be found in both the M1 and M2 population (although it is possible that some cross-contamination between M1 and M2 could in part explain the overlap). Second, decreasing the peptide loading does not eliminate M2. For CRM_{197} activated to a given level with BAANS, increasing peptide loading does lead to increased M2 formation (**Figure 1.5**, red trace). However, significant amounts of M2 (45%) are observed in batches with only a few peptides. Third, the increasing peptide loading by increasing the level of BAANS-activation prior to conjugation has a greater effect on M2 formation than does increasing the degree of peptide loading on CRM_{197} activated to a consistent degree (**Figure 1.5**, black trace).

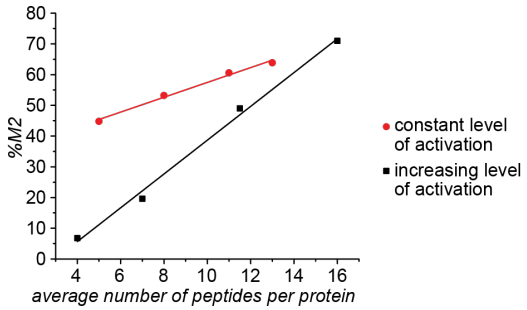


Figure 1.5: Effect of peptide loading on M2 formation in peptide-CRM_{197} conjugates. Variable levels of peptide loading were achieved in two ways: 1) CRM_{197} was activated with BAANS to an average of 16 activating agents per protein, then treated with increasing amounts of peptide (0.1875-1.5 g peptide per g protein, red curve), and 2) CRM_{197} was activated with increasing amounts of BAANS (10-100 equiv.), then treated with a constant amounts of peptide (1.5 g peptide per g protein, black curve). Composition of M2 was determined by SEC. Experiment and analysis by Jake Jaffe.

A key observation was the fact that activated CRM_{197} that was subsequently treated with a thiol-containing discrete polyethylene glycol (m-$dPEG_{12}$-SH) resulted in formation of both M1 and M2, much like the peptide conjugates. However, direct conjugation of equivalent

levels of a discrete PEG-NHS ester (m-dPEG$_{12}$-NHS) to the lysine residues in CRM$_{197}$ did not result in M2 formation (*vida infra*, **Figure 1.8**). Taken together, this suggests that the peptide loading, and the peptide conjugation step in general, is not the root cause of the irreversible conformational change. Rather, it is the bromoacetylation activation step that causes or promotes it.

1.2.2 Modification of Histidine-21 and crosslinking

While the bromoacetyl functionality is generally selective towards cysteines (or other thiol-containing molecules), cross-reactivity with other nucleophilic residues is not unprecedented [26]. In the case of bromoacetylation of CRM$_{197}$, some off-target reactivity of the bromoacetyl group is observed in the form of crosslinking formation with other nearby nucleophiles on the protein surface. As CRM$_{197}$ does not contain any free cysteine residues, the crosslinking must occur between lysines and other non-cysteine side chains.

In previous work towards the preparation of homogeneous CRM$_{197}$ conjugate material, Chang *et al.* discovered that histidine-21 of CRM$_{197}$ displayed pronounced reactivity towards iodoacetamide-containing reagents [27]. Furthermore, this group demonstrated that specific crosslinking between His21 and Lys24 (**Figure 1.6a**) was possible using the bis(iodoacetamide) reagent HPBIA. We confirmed that this reagent modified CRM$_{197}$ selectively by mass spectrometry (**Figure 1.6bc**). Based on this finding, crosslinking between His21 and Lys24 might also be expected in the case of bromoacetylation.

To examine whether bromoacetylation results in a specific crosslink with His21 (versus other nucleophilic residues), His21 was first modified to roughly 50% yield with HPBIA. The HPBIA-CRM$_{197}$ was then capped with m-dPEG$_{12}$-SH and activated with varying amounts of BAANS. No subsequent peptides or capping agents were coupled in this experiment. The resulting mixture contained two populations of activated CRM$_{197}$, which were distinct by ESI-TOF-MS (**Figure 1.6d**). Comparison of the two populations revealed significantly greater crosslinking in the portion of the sample where His21 was not capped, relative to that in which it was capped with PEG-HPBIA. This experiment therefore suggests that the majority of the crosslinking observed in samples of activated CRM$_{197}$ occurs at His21, presumably with Lys24.

While this reactivity is interesting in its own right, we hypothesized that this specific crosslink plays a role in the formation of the open form conjugate (M2). His21 is located on the catalytic domain, close to the C-R interface [28]. Perhaps modification of His21 through crosslinking could disrupt the C-R interface, partially opening the protein through a "bumping" effect. To explore this possibility, Lys24-His21 crosslinking was mimicked through complete modification of His21 with HPBIA, followed by NAC capping. Peptide mapping of (NAC-HPBIA)-CRM$_{197}$ confirmed the high specificity of this reagent for His21 (**Figure 1.6c**). (NAC-HPBIA)-CRM$_{197}$ was then activated with varying levels of BAANS,

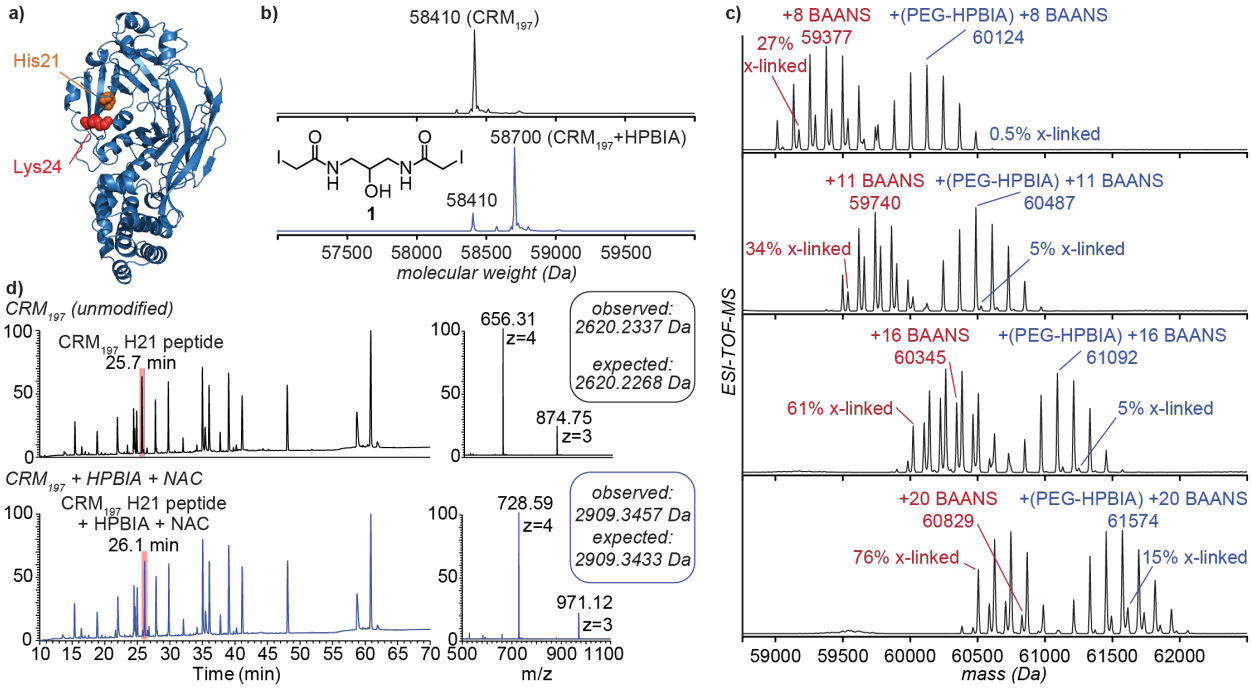


Figure 1.6: BAANS-activated CRM_{197} has intraprotein crosslinks. (a) The locations of His21 and Lys24 are shown on the CRM197 structure (PDB ID: 1MDT). (b) HPBIA (**1**) selectively modifies CRM_{197} at His21. (c) MS analysis of a tryptic digest of unmodified and modified (NAC-HPBIA)-CRM_{197}. The results indicated the high selectivity of HPBIA for His21. Experiment and analysis by Justin Sperry. (d) Using **1**, His21 was selectively modified to about 50% conversion. The remaining iodide was capped with m-$dPEG_{12}$-SH. The resulting batch of 1:1 CRM_{197}:(PEG-HPBIA)-CRM_{197} was then treated with BAANS in increasing amounts (from top to bottom). Mass spectra of the activated material clearly showed significantly lower amounts of crosslinking in material that had been treated with HPBIA (higher molecular weight series), suggesting that His21 was the primary target of BAANS crosslinking.

and subsequently conjugated with the peptide of interest (CTNEHFRG). Comparison by SEC of this series of conjugates prepared from (NAC-HPBIA)-CRM_{197} to an analogous series prepared from standard CRM_{197} revealed that modification of His21 has a distinct effect on M2 formation (**Figure 1.7**). Selective capping of His21 with HPBIA-NAC prior to activation resulted in an increase in M2 formation over standard material with equivalent levels of peptide loading. However, as the level of modification of His21 through cross-linking is significantly lower in standard activated CRM_{197}, the actual contribution of M2 formation due to His21 crosslinking is likely smaller than observed for the (NAC-HPBIA)-CRM_{197} system.

While His21 crosslinking appears to promote M2 formation, it is not independently sufficient to cause M2 population. As mentioned above, PEGylated CRM_{197} prepared by lysine modification with m-$dPEG_{12}$-NHS does not result in observable M2 formation. If His21 crosslinking is the direct cause of M2 formation, (NAC-HPBIA)-CRM_{197} should form M2 after direct PEGylation. However, PEGylation of (NAC-HPBIA)-CRM_{197} resulted in very slight (almost negligible) M2 formation, as determined by SEC (**Figure 1.8**). Moreover,

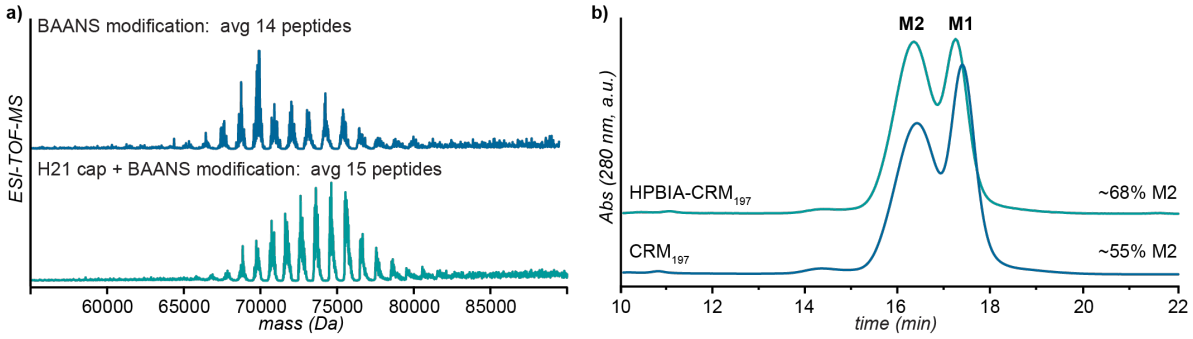


Figure 1.7: Comparison of His21 capping on BAANS-activated peptide-CRM_{197}. Two seperate samples of peptide-CRM_{197} were synthesized and then analyzed by ESI-TOF-MS and SEC. One was prepared from HPBIA-CRM_{197} (His21 capped) prior to BAANS activation (teal trace) while the other was unmodified prior to activation (blue trace). As anticipated, the HPBIA-CRM_{197} peptide conjugate results in a larger M2 population (about 68% by SEC).

HPBIA capping after PEGylation of CRM_{197} does not result in M2 formation. Given that His21 crosslinking promoted, but did not directly result in M2 formation, the "C-R domain interface bumping" hypothesis was clearly missing an important factor.

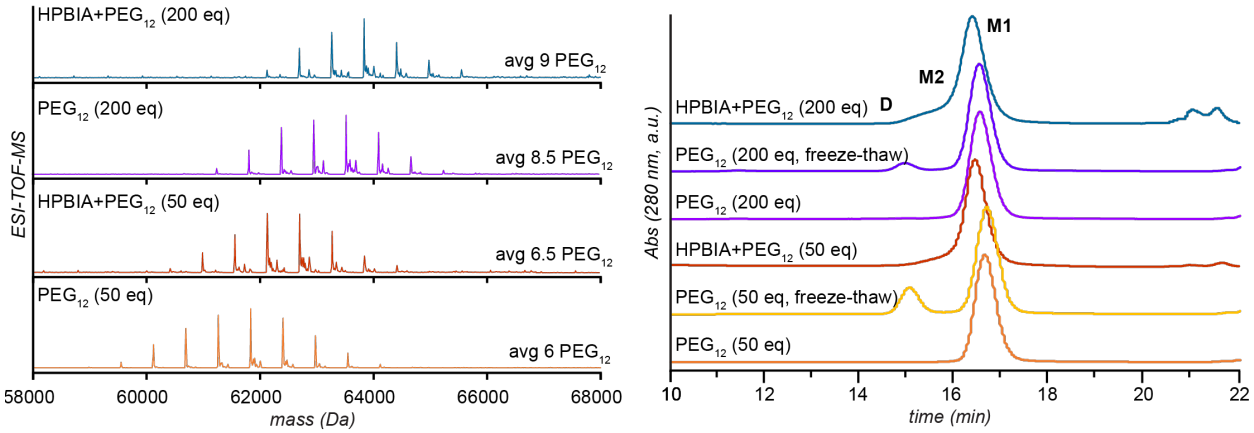


Figure 1.8: ESI-TOF-MS and SEC analysis of PEGylated CRM_{197}. Varying levels of descrete PEG-NHS (m-$dPEG_{12}$-NHS) were directly conjugated to the lysine residues of CRM_{197} and analyzed by ESI-TOF-MS (a) and SEC (b). No observed M2 population was promoted (orange (50 equiv. PEG loading) and magenta (200 equiv. PEG loading) traces). The dimer configuration can be induced after a freeze-thaw cycle (observed in yellow and purple SEC traces), but no significant M2 shoulder peak is observed in the SEC traces. However, if HPBIA is used to cap the His21 prior to PEG conjugation, the M2 shoulder can be detected in the 200 equiv. PEG sample (teal trace).

Examination of the structures for "closed" and "open" forms of CRM_{197} (as shown in **Figure 1.4**), led to a "bump and label" hypothesis; the C-R interface is partially disrupted ("bump") by His21 crosslinking or other factors, allowing for activation ("label") of the interfacial lysine residues. A comparison of the differences between the solvent accessible surface area (SASA) of the lysine residues in the closed-and open-form monomers revealed that most lysine residues remained similarly accessible (**Figure 1.9**). However, residues K419, K445, K447, K456, and K474 become significantly more solvent accessible in the open

conformation. Unsurprisingly, these five residues reside at the C-R interface in the closed monomer. This collection of lysine residues is roughly the same in number as the difference in average peptide loading between M1 and M2 in the isolation experiment described in **Figure 1.2**. BAANS activation of this set of interfacial lysine residues may be the necessary factor for M2 formation after conjugation. If this is the case, M2 formation can be explained by a "bump and label" mechanism. It should be noted that previous studies probing CRM_{197} lysine reactivity through peptide mapping have not shown these interfacial lysine residues to be reactive [23, 25, 29]. However, the conjugates examined in those studies were prepared by direct conjugation methods rather than through an activation-conjugation strategy like that used in the present work (and are therefore analogous to the direct m-$dPEG_{12}$-NHS coupling results discussed above).

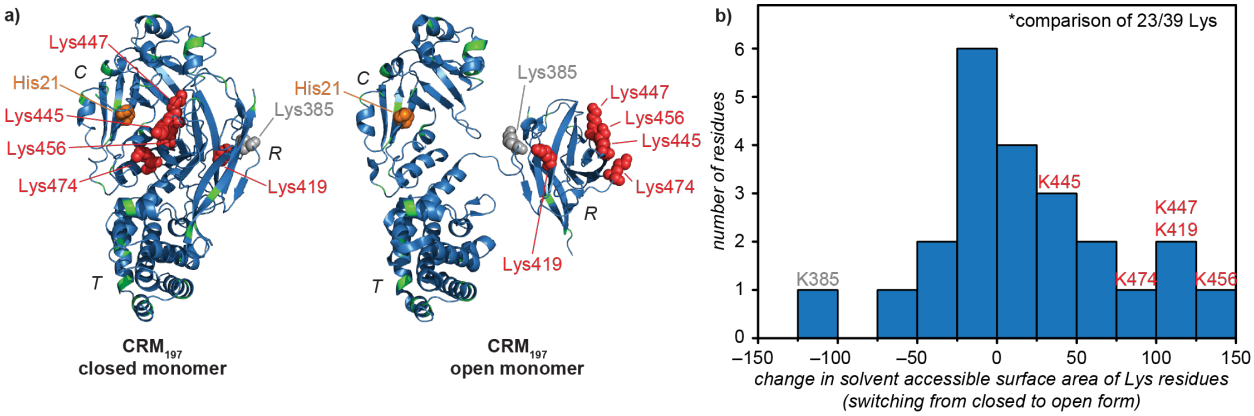


Figure 1.9: Lysines at the interface of the catalytic (C) and receptor binding (R) domains of CRM_{197} are unique. (a) K419, K445, K447, K456, and K474 become more exposed upon transition from the closed form to the open form of CRM_{197}. (b) Differences in solvent accessible surface area (SASA) of the lysine residues between the two conformations of CRM_{197} were quantified with BioLuminate. Lysine residues that were incomplete or missing from either one or both crystal structures were not included in this analysis. Analysis by Jake Jaffe and Sandeep Kumar.

1.2.3 Alternative bifunctional linkers

As intraprotein crosslinking was found to be a contributing factor to the conformational changes in peptide-CRM_{197} conjugates, alternative bifunctional linkers were explored for generating homogeneous products. Initially we hypothesized that simply extending the linker could decrease the crosslinking. Thus, succinimidyl 3-(bromoacetamido)propionate (SBAP), a 6.2 Å linker length compound, was used to synthesize peptide-CRM_{197} conjugate material. Using the same protocol mentioned in **Figure 1.1**, peptide conjugates from SBAP-activated CRM_{197} were prepared and analyzed by SEC (**Figure 1.10**). The resulting material had a decreased M2 population, as compared to BAANS-activated CRM_{197} for conjugates at similar peptide loading. Likewise, with HPBIA-CRM_{197}, SBAP activation followed by peptide conjugation resulted in a higher percentage of M2 (62% M2 for average peptide loading of 14) as compared to the uncapped SBAP-activated CRM_{197} (62% versus 30% M2, respectively, for an average peptide loading of 14, **Figure 1.10**). While the amount

of crosslinking cannot be accurately quantified, we deduce that the increased length of the linker limited the amount of crosslinking.

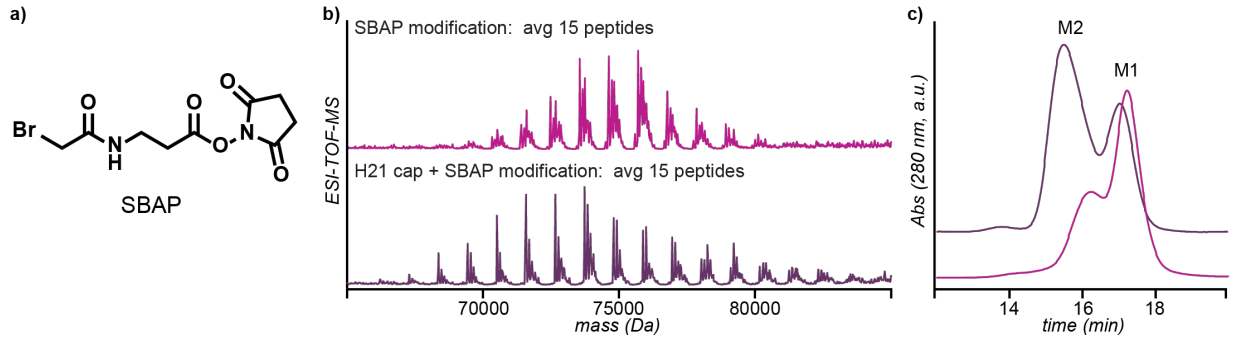


Figure 1.10: Conformational analysis of SBAP-activated CRM_{197} conjugates. (a) SBAP, a 6.2 Å bromoactamide-NHS ester bifunctional linker, was used to activate CRM_{197}. Two peptide-CRM_{197} samples were prepared using SBAP activation: one with uncapped CRM_{197} (magenta trace), and one with His21 capping (purple trace). Peptide loading was assessed via ESI-TOF-MS (b), and M2 population was analyzed by SEC (c). For the His21 capped CRM_{197}, a drastic increase in M2 population was observed.

Additionally, a maleimide-NHS ester bifunctional linker was explored as an alternative for generating homogeneous conjugates. Like haloacetamides, maleimides are capable of labeling residues with nucleophilic side chains other than the primarily targeted cysteine [30]. However, it was unclear whether a maleimide-NHS ester crosslinking reagent, such as *N*-β-maleimidopropyl-oxysuccinimide ester (BMPS), would form the Lys24-His21 M2-promoting crosslink that is associated with BAANS activation. Peptide conjugates from BMPS-activated CRM_{197} were prepared and analyzed by SEC. Intraprotein crosslinks were observed to a similar extent during ESI-TOF-MS analysis after thiol addition, and increasing activation and peptide loading resulted in increased M2 formation (**Figure 1.11a-c**). Arguably, the impact of capping His21 on crosslinking in the case of BMPS-activation is less pronounced than in the case of BAANS-activation. This suggests that crosslinking seen in BMPS-activated conjugates is not completely directed at His21. Further, conjugates prepared from BMPS-activated CRM_{197} did not result in as much M2 formation as those prepared from BAANS-activated CRM_{197} on a per peptide basis (**Figure 1.11d**).

This trend with BMPS further supports the "bump and label" hypothesis discussed prior. While the small BAANS agent, with its zero-length linker, is able to access the interfacial lysine residues fairly readily, we also explored bifunctional moieties with longer linkers: *N*-γ-maleimidobutyryl-oxysuccinimide ester (GMBS, 7.3 Å C4 linker) and maleimide-PEG_4-NHS ester (27.4 Å PEG_4 linker). We observed that while BMPS offers an intermediate case, in which the maleimide moiety minimizes reactivity towards interfacial lysine acylation through steric hindrance, increasing the length of the linker decreases the percentage of M2 (**Figure 1.12**). Also consistent with the "bump and label" model, preincubation of CRM_{197} with HPBIA-NAC increased the level of M2 formation in all cases. Compiled data is summarized in **Supplemental Figure 1.16**.

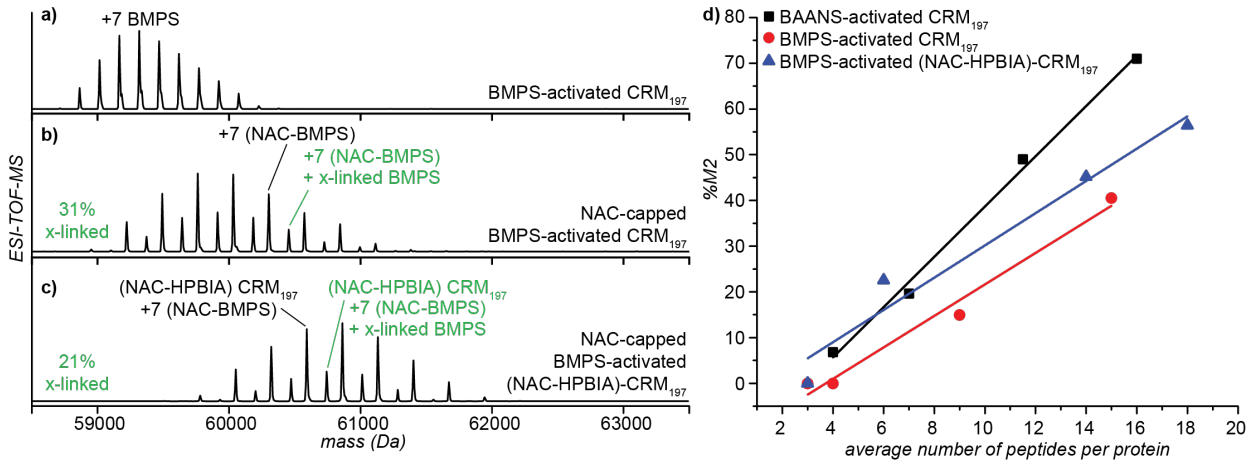


Figure 1.11: Crosslinking in BMPS-activated CRM_{197} and comparison of BAANS-activated and BMPS-activated peptide-CRM_{197} conjugates. (a) Crosslinking in BMPS-activated CRM_{197} cannot be observed directly as the Michael addition of a nucleophilic residue to the maleimide of BMPS results in no change of mass. (b) Crosslinking can be observed indirectly by coupling a small molecule thiol (NAC) to the activated material. Unreacted activating groups are assumed to have formed crosslinks, as labeled in red. (c) BMPS-activation of (NAC-HPBIA)-CRM_{197} results in lower levels of crosslinking. However, crosslinking must occur at sites other than His21 in BMPS-activated CRM_{197}. (d) Conjugates prepared from BMPS-activated CRM_{197} (red curve) resulted in lower levels of M2 formation than did comparable conjugates from BAANS-activated CRM_{197} (black curve). BMPS-activated (NAC-HPBIA)-CRM_{197} (blue curve) was found to contain more M2 than did comparable material without His21-capping.

One intriguing result was that of succinimidyl 4-(*N*-maleimidomethyl)cyclohexane-1-carboxylate (SMCC) activated CRM_{197}. This bifunctional linker contains a 8.3 Å cyclohexane linker. We hypothesized that this linker would reduce the amount of crosslinking, and subsequent M2 population, much like like GMBS and SBAP. ESI-TOF-MS analysis of peptide conjugates of SMCC-activated CRM_{197} depicted significantly less crosslinking, however, SEC analysis showed drastically increased M2 population (**Figure 1.12**). Understanding the number of lysine residues at the interface of the C and R domain (**Figure 1.9**), it can be reasoned that upon modification with SMCC, the cyclohexane linker steric bulk at this interface would promote the "bump" and subsequent opening of the C domain to promote M2 product. On this note, it would be interesting to quantify the reactivity of the lysine residues on the surface of CRM_{197} to understand chemoselective bioconjugation.

1.2.4 Alternative conjugation strategies

Bifunctional linkers offer a way to site-specifically modify resides on a protein of interest with a modular linker. However, direct conjugation methods avoid the need for a distinct two-step process step. Namely, glycoconjugate vaccines can be prepared by reduction of the lysine residues followed by akylation by an aldehyde-containing oligosaccharide. It is unknown whether CRM_{197} glycoconjugates form the M1/M2 populations observed in peptide-CRM_{197} conjugates as analysis of the glycoconjugates is highly complicated due to the nature of the oligosaccharide epitopes. Direct conjugation might avoid the "bump and label" situation,

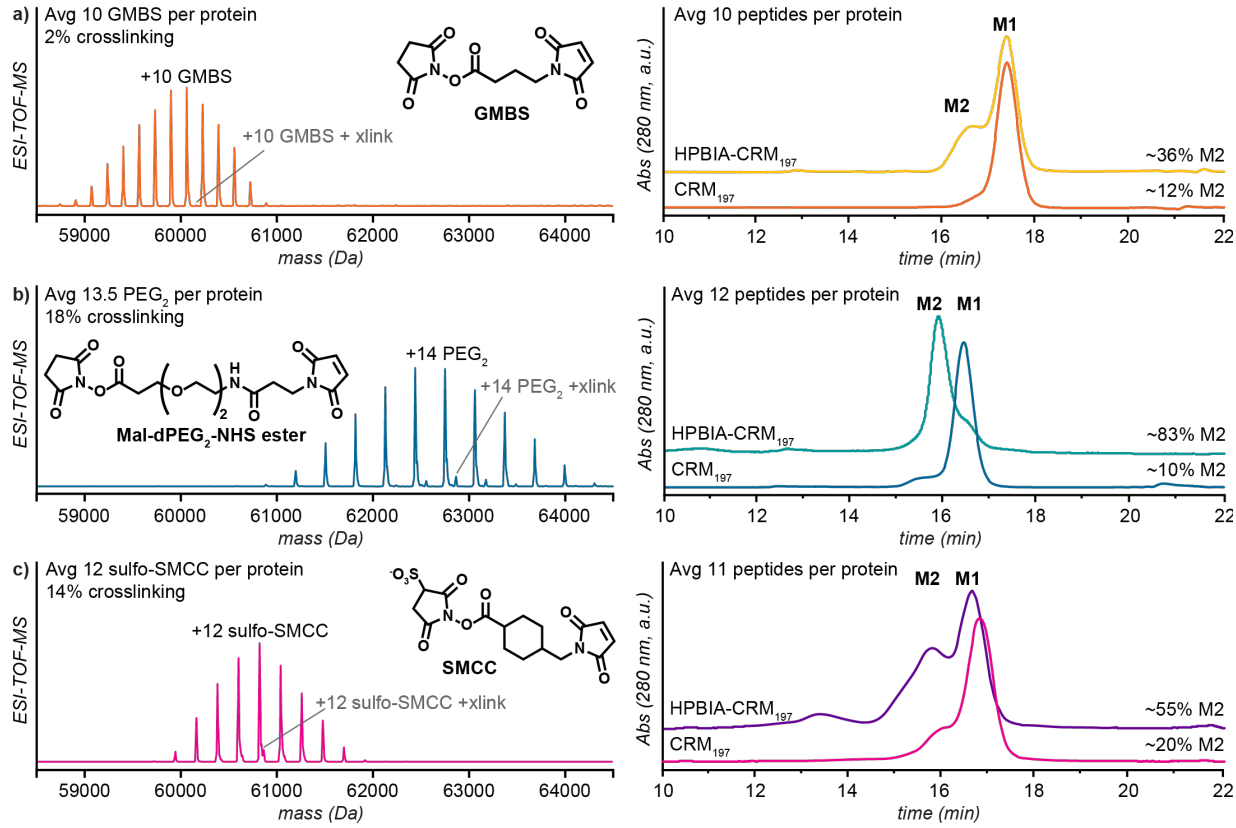


Figure 1.12: Effect of longer bifunctional linker on crosslinking and M2 formation. (a) GMBS-activated CRM_{197} appears to have little-to-no crosslinking in the ESI-TOF-MS analysis, but after peptide conjugation, M2 population was observed by SEC. GMBS-activated peptide-CRM_{197} was observed to be about 12% M2 (orange trace), which is a significant decrease in M2 population as compared to activation with the shorter bifunctional linkers at similar peptide loadings. GMBS-activated CRM_{197} with His21 capping was calculated to be about 36% M2 (yellow trace). Solubility issues were encountered at higher levels of peptide loading, accounting for the lower conjugation levels. (b) PEG_2-activated CRM_{197} showed around 18% crosslinking, but interestingly only slight M2 population after peptide conjugation. Around 10% M2 was observed in the SEC analysis of PEG_2-activated peptide-CRM_{197} (blue trace), which is a significant decrease in M2 population as compared as compared to activation with the shorter bifunctional linkers at similar peptide loading. PEG_2-activated (HPBIA-NAC)-CRM_{197} (teal trace) was found to contain drastically higher M2 population (around 83%). This unprecedented increase could be due to changes in the local protein environment, due to the PEG_2 linker itself. (c) Sulfo-SMCC-activated CRM_{197} appeared to have 12% crosslinking, and 20% M2 was observed after peptide conjugation (magenta trace). Interestingly, sulfo-SMCC-activated (HPBIA-NAC)-CRM_{197} (purple trace) was found to contain higher M2 population (around 55%) than anticipated. The steric bulk of the cyclohexyl linker could contribute structural perturbations that are not crosslinking related.

leading to conformationally homogeneous conjugate material, as prescendented by previous studies [23, 25, 29].

The canonical method for activating the amines on lysine residues is by using a strong reducing agent, such as a hydride source [26]. Sodium cyanoborohydride ($NaCNBH_4$) was used as a protein compatible reagent, with a model substrate, benzaldehyde, to test small molecule loading onto CRM_{197}. An average of four benzaldehyde moieties were successfully

loaded onto CRM_{197} (**Figure 1.13**). While a variety of buffers were examined, the yield of conjugated benzaldehyde molecules was not optimal. To mimic a short peptide cargo on the aldehyde, a PEG_3-benzaldehyde was synthesized and loaded onto CRM_{197}, but did not improve loading yield. Creating a compatible peptide aldehyde substrate would be useful; a synthetic route for the production of aldehyde-functionalized peptide (sequence: ACTNEHFRG) was proposed, using a solid-phase peptide synthesis on a special Weinreb amide resin (available from MiliporeSigma), but the aldehyde could not be recovered. For the reductive amination reaction overall, it is suggested that increased reaction concentrations and times might facilitate higher imine/iminium formation and subsequent loading.

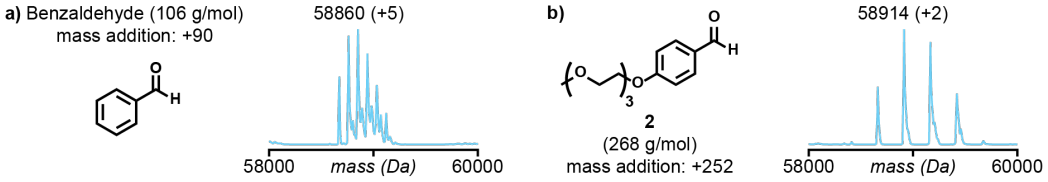


Figure 1.13: Direct conjugation of aldehydes on CRM_{197} via reductive amination. Using $NaCNBH_4$ to activate the surface available amines, (a) benzaldehyde and (b) 4-[tri(ethylene glycol) monomethyl ether] benzaldehyde (**2**) were loaded onto the protein. The resulting material was analyzed by ESI-TOF-MS, with an average of +5 benzaldehyde and +2 **2** modifications per conjugate. Reaction conditions: 10 μM CRM_{197}, 1 mM aldehyde, 20 mM $NaCNBH_4$ in DPBS pH 8.0 buffer, at 22 °C for 20 h.

To alleviate this and the potential downsides of using the harsher $NaCNBH_4$, we used an iridium catalyst with formate as a reductant to accelerate the reaction, as previously reported by McFarland and Francis [31]. Synthesis of the Ir-catalyst proceeded as reported, and subsequent conjugation of benzaldehyde moieties to CRM_{197} was observed by ESI-TOF-MS. Unfortunately, a side reaction between the Ir-catalyst, sodium formate, and CRM_{197} was observed, even with thorough spin filtering and buffer washes, which complicated analysis (**Figure 1.14**). The contamination with Ir-catalyst was unavoidable and thus this route was abandoned.

Another direct conjugation approach examined involved oxidative coupling of chemically manipulated surface avaliable lysine residues. Surface available amines can be modified with isatoic anhydride, converting these to aniline moieties [32]. The newly installed aniline can be further modified through exposure to a phenol derivative, such as *ortho*-methoxyphenol, and an oxidizer, such as potassium ferricyanide [33]. When applied to CRM_{197}, this reaction led to high yield of lysine conversion, but unfortunately less-than-optimal small molecule loading (**Figure 1.15**). Analysis of the converted product (aniline modification) was difficult because the mass addition of aniline (+121) and *o*-methoxyphenol (+121) overlap. Instead, phenol derivatives, functionalized off of the *para*-position with PEG or a peptide, would facilitate analysis. For this reason, a 3-(2-methoxyphenol)-propionic acid compound is proposed as starting material to create a functionalized peptide for oxidative coupling.

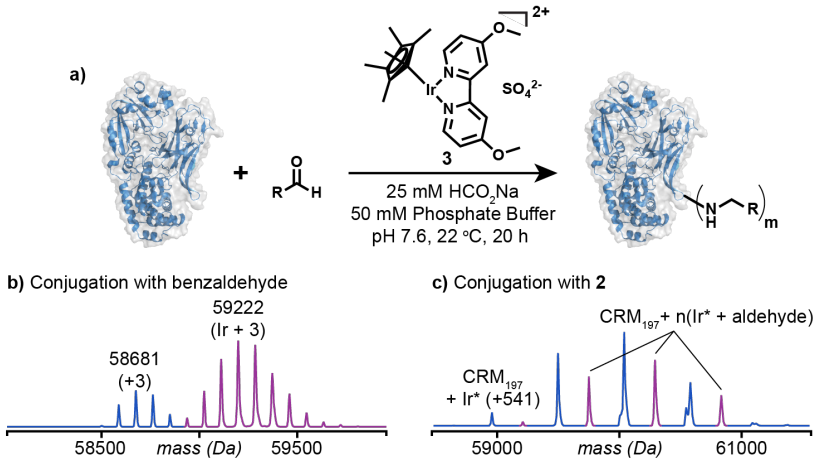


Figure 1.14: Direct conjugation of aldehydes on CRM_{197} via reductive alkylation. (a) A [Cp*Ir(4,4'-dimethoxy-2,2'bipyridine)]SO_4 (**3**) catalyst was reported to reduce imines formed from the condensation of aldehydes with surface available amines, in the presence of formate ions. The reaction proceeds at or above 22 °C, in pH 7.6 buffer. Benzaldehyde (b) and **2** (c) were conjugated to CRM_{197} using this method. Upon analysis by ESI-TOF-MS, **3** was observed to associate with CRM_{197}, even after buffer exchange purification washes. Reaction conditions: 50 μM CRM_{197}, 1 mM aldehyde, 20 μM Ir-catalyst, 50 mM formate in 50 mM phosphate buffer pH 7.6, incubated at 22 °C for 20 h.

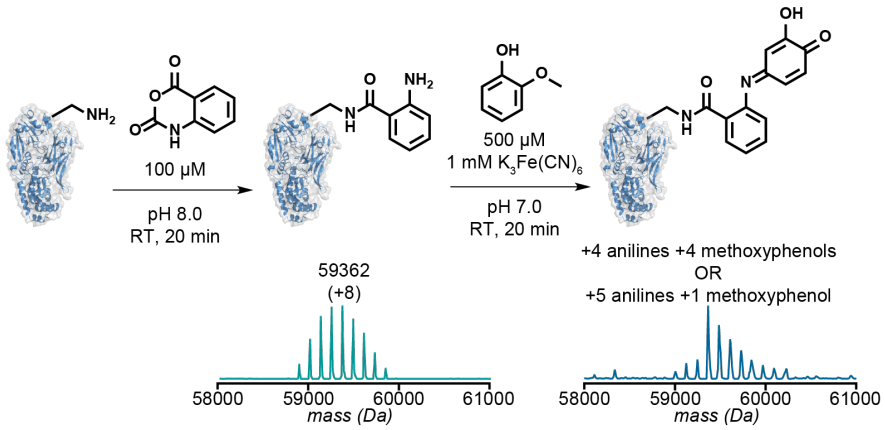


Figure 1.15: Isatoic anhydride mediated oxidative coupling of small molecules to CRM_{197}. Surface available amines are converted to anilines with isatoic anhydride. This intermediate is then subjected to oxidizing conditions (potassium ferricyanide), which subsequently facilitates the coupling of aniline to 4-methoxyphenol, resulting in a stable, 4-iminoquinone product. However, analysis of the converted product was difficult because the mass addition of aniline (+121) and methoxyphenol (+121) overlap.

1.3 Conclusions

These studies have improved our understanding of the conformational changes that occur during the synthesis of peptide-CRM_{197} conjugate materials. It was determined that crosslinking, or direct modification, of His21 in CRM_{197} promotes conformational change when using an activation-conjugation strategy. This exposes a new set of lysine residues that, upon modification, lead to the formation of the M2 species. Ultimately, these well-

defined constructs can be used to gain additional insights into structure-immunogenicity relationships in peptide-CRM_{197} conjugate vaccines.

1.3.1 Acknowledgements

Special acknowledgements to Jake Jaffe, Qin Zou, Justin Sperry, and Jane Honda for experimental contributions and project support.

1.4 Supplemental Figure

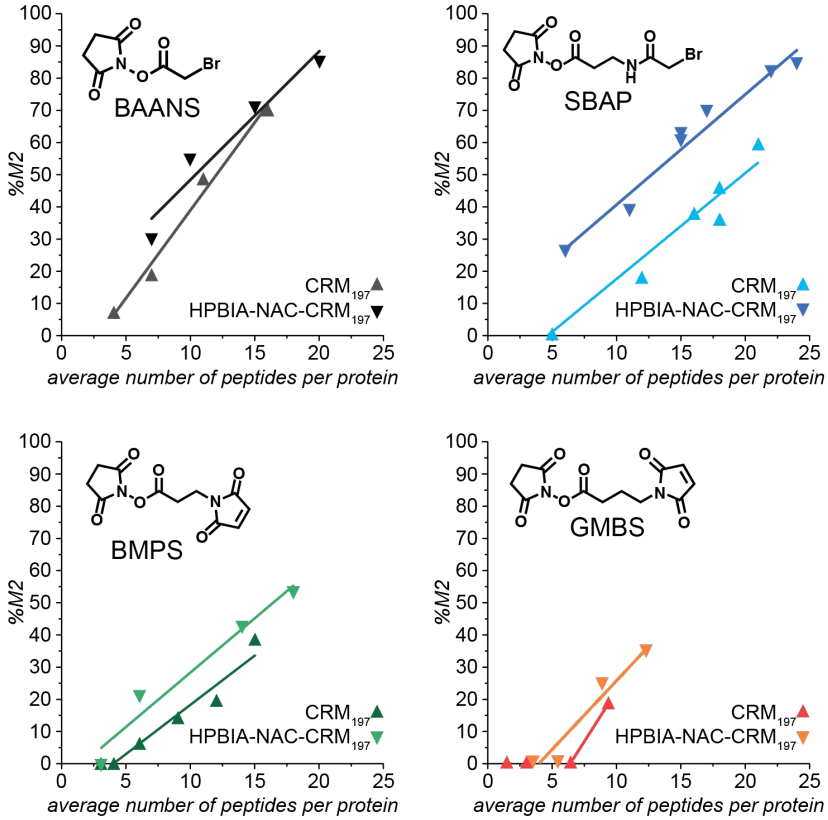


Figure 1.16: Effect of HPBIA capping on M2 formation in peptide-CRM_{197} conjugates. In each case, CRM_{197} was either capped at His21 with HPBIA prior to activation (HPBIA-NAC-CRM_{197}), or directly activated with the bifunctional linker (CRM_{197}). Overall, the peptide-CRM_{197} conjugates capped with HPBIA prior to activation resulted in higher levels of M2 formation than did uncapped peptide-CRM_{197} conjugates, of comparable peptide loading. This trend was observed throughout various activation strategies (bromoacetyl-NHS bifunctional linkers BAANS and SBAP, and NHS-maleimide bifunctional linkers BMPS and GMBS.

1.5 Materials and Methods

1.5.1 General methods and instrumentation

Unless otherwise noted, all reagents were obtained from commercial sources and used without any further purification. Analytical thin layer chromatography (TLC) was performed on EM Reagent 0.25 mm silica gel 60-f254 plates and visualized by ultraviolet (UV) irradiation at 254 nm and/or staining with potassium permanganate. Purifications by flash silica gel chromatography were performed using EM silica gel 60 (230–400 mesh). All organic solvents were removed under reduced pressure using a rotary evaporator. Water (dd-H_2O) used in all procedures was deionized using a NANOpure^{TM} purification system (Barnstead, USA). Centrifugations were performed with an Eppendorf 5424 R at 4 °C (Eppendorf, Hauppauge, NY). CRM_{197} from *Corynebacterium diphtheriae* was obtained from Pfizer, Inc. (St. Louis, MO). All samples of CRM_{197}, activated CRM_{197}, and CRM_{197} conjugates were handled and stored at or below 4 °C. All containers (e.g., Eppendorf tubes, spin columns, spin filters, and LC-MS vials) were chilled on ice prior to addition of samples containing CRM_{197} or its derivatives. Peptides were procured from GenScript (Piscataway, NJ). Discrete PEG reagents were purchased from Quanta Biodesign (Plain City, OH).

Nuclear Magnetic Resonance spectroscopy (NMR). ^1H and ^{13}C spectra were measured with a Bruker AVQ-400 (400 MHz) spectrometer. ^1H NMR chemical shifts are reported as δ in units of parts per million (ppm) relative to residual $CHCl_3$ (δ 7.26, singlet) or DMSO-d_6 (δ 2.50, pentet). Multiplicities are reported as follows: s (singlet), d (doublet), t (triplet), q (quartet), p (quintet), or br s (broad singlet). Coupling constants are reported as a J value in Hertz (Hz). The number of protons (n) for a given resonance is indicated as nH and is based on spectral integration values. ^{13}C NMR chemical shifts are reported as δ in units of parts per million (ppm) relative to $CDCl_3$ (δ 77.16, triplet) or DMSO-d_6 (δ 39.52, septet).

High Performance Liquid Chromatography (HPLC). HPLC was performed on Agilent 1100 series HPLC systems (Agilent Technologies, USA) equipped with in-line diode array detector (DAD) and fluorescence detector (FLD). Size exclusion chromatography (SEC) was accomplished on a TSKgel G3000SWXL column fitted with a TSKgel SWXL guard column (Tosoh Bioscience LLC, King of Prussia, PA) using an aqueous mobile phase (100 mM sodium phosphate, 200 mM NaCl, pH 7.6) at a flow rate of 0.55 mL/min. Column integrity was confirmed by analyzing bovine serum albumin (Sigma-Aldrich, St. Louis, MO) and CRM_{197} (Pfizer, St. Louis, MO) analytical standards, and a 1,350–670,000 Da gel filtration standard mixture (Bio-Rad, Hercules, CA). To integrate partially overlapping SEC peaks accurately, multiple Gaussian fits were performed using OriginPro 9.0 (OriginLab Corp., Northampton, MA).

Mass Spectrometry. Proteins and protein conjugates were analyzed on an Agilent 6224

Time-of-Flight (TOF) mass spectrometer with a dual electrospray source (ESI) connected in-line with an Agilent 1200 series HPLC (Agilent Technologies, USA). Chromatography was performed using a Proswift RP-4H (Thermo Scientific, USA) column with a H_2O/MeCN gradient mobile phase containing 0.1% formic acid. Mass spectra of proteins and protein conjugates were deconvoluted with MassHunter Qualitative Analysis Suite B.05 (Agilent Technologies, USA).

1.5.2 Experimental procedures

General procedure for preparing CRM_{197} conjugates. An Eppendorf tube was pre-chilled on ice, charged with a stock solution of CRM_{197} in storage buffer (33 μM final concentration, 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5), and diluted with ice-cold reaction buffer (DPBS, pH 8.0). To this solution was added a freshly prepared stock solution of the bifunctional linker (generally 10, 20, 50, or 100 equiv. from a 200 mM stock in DMF). The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant activated CRM_{197} was analyzed by ESI-TOF-MS; the sample was kept on ice until $\leq$2 min prior to injection. activated-CRM_{197} in reaction buffer (around 33 μM) was treated with either stock peptide solution (1.5 mg peptide/mg protein 15 μL peptide stock solution per 100 μL of 33 μM activated-CRM_{197}, 20 mg/mL stock concentration, in 0.6 M $NaHCO_3$, pH 9.2), or neat m-$dPEG_{12}$-SH followed by addition of 0.6 M $NaHCO_3$, pH 9.2. The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 3 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant CRM_{197} conjugate material was analyzed by ESI-TOF-MS and SEC; the sample was kept on ice until $\leq$2 min prior to injection.

Analytical Ultracentrifugation. The isolated fractions of CRM_{197} (M1 and M2) from size exclusion chromatography were diluted to 0.2-0.3 mg/mL with matched formulation buffer (25 mM HEPES buffer, 150 mM NaCl, pH 7.5, 10% sucrose). A total of 420 μL of sample was loaded into a 12 mm AUC epon centerpiece and subjected to 45,000 rpm at 25 °C until the completed depletion of boundary. The data were analyzed using Sedfit (version 14.1) with a bimodal model to determine the frictional coefficient (f/f0) and sedimentation coefficient for the main species. A blank CRM_{197} sample was evaluated similarly as a control. Buffer density and viscosity were calculated using Sednterp (version 20130813 beta).

General procedure for the preparation of (NAC-HPBIA)-CRM_{197}. An Eppendorf tube was pre-chilled on ice, charged with a stock solution of CRM_{197} in storage buffer (33 μM final concentration, 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5), and diluted with ice-cold reaction buffer (DPBS, pH 8.0). To this solution was added a freshly prepared stock solution of HPBIA (200 mM in DMF, 137 equiv.). The resulting mixture

was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. Additional HPBIA was added (137 equiv.), the solution was mixed by pipetting, and incubated for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. The resultant HPBIA-CRM_{197} (33 μM) was treated with neat *N*-acetylcysteamine (NAC) (0.15 μL per 100 μL of 33 μM HPBIA-CRM_{197}) or m-$dPEG_{12}$-SH. The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant (NAC-HPBIA)-CRM_{197} was analyzed by ESI-TOF-MS and SEC; the sample was kept on ice until $\leq$2 min prior to injection. General procedures for activation and conjugation were used to further elaborate (NAC-HPBIA)-CRM_{197}.

Peptide mapping of isolated M1 and M2. All samples were diluted to 1 mg/mL in PBS buffer pH 7.4. Samples were then digested with trypsin (1:20 ratio), under non-reducing and non-alkylating conditions, at 37 °C for 4 h. The resulting peptide mixture was injected for LC-MS/MS analysis on an Agilent HPLC connected in-line to a Q Exactive Plus mass spectrometer with a HESI source (Thermo Fischer Scientific). The HPLC diode-array detector was set to collect 214 and 280 nm wavelength traces. The peptides analyzed with a Waters C-SH C18 column (2.1 x 150 mm, 2.5 μm particle size) at 60 °C, using a gradient of 3-40% CH_3CN in H_2O (0.1% formic acid) over 50 min followed by a gradient of 40-90% CH_3CN in H_2O (0.1% formic acid) over 10 min, with a flow rate of 0.2 mL/min, monitored at 214 and 280 nm. The MS was operated in data-dependent mode for the top-10 most abundant ions present in the MS spectrum. The spray voltage was set to 3.5 kV, a capillary temperature of 250 °C, sheath gas of 35 arbitrary units, auxiliary gas of 10 arbitrary units and a probe temperature of 350 °C. MS1 spectra were collected in profile mode from 200 to 2000 m/z with a resolving power of 70,000. The injection time (IT) was set to 100 ms and the target automatic gain control (AGC) was set to $3x10^6$. MS2 spectra were collected in profile mode with a resolving power of 17,500 for the top-10 MS1 precursors above a threshold value of $2.5x10^4$ with normalized collision energy (NCE) of 27 and an isolation window of + 2 m/z. The IT was set to 200 ms and the target AGC was set to $5x10^5$. Precursor ions containing unassigned, +1 or >+8 charge states were excluded from fragmentation. Dynamic exclusion was not enabled. The MS data for each sample were collected from 3 to 68 min using a divert valve. Data analysis was performed using Proteome Discoverer (Thermo Fisher Scientific, version 1.4) using the Mascot (Matrix Science) search engine. A "No Enzyme" specificity was selected for all searches and a 1% false discovery rate (FDR) was enforced for peptide and protein results. The mass tolerance for precursor ions was set to 10 ppm and for fragment ions at 0.8 Da. Variable modifications were set to methionine oxidation, BAANS, and hydrolyzed BAANS modifications on His, Lys, Tyr and the N terminus. Extracted-ion chromatograms from the MS2 spectra were generated in QualBrowser (Xcalibur, Thermo Fisher Scientific) by selecting 702.33, 588.29, 459.25 and 322.19 m/z as diagnostic y5 through y2-ions of the CTNEHFRG peptide marker, assuming trypsin cleavage of the C-terminal Gly residue.

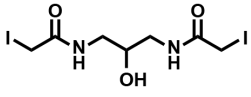
Direct PEGylation of CRM_{197}. An Eppendorf tube was pre-chilled on ice, charged with a stock solution of CRM_{197} in storage buffer (33 μM final concentration, 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5), and diluted with ice-cold reaction buffer (DPBS, pH 8.0). To this solution was added a stock solution of m-$dPEG_{12}$-NHS ester (25–100 equiv. from a 250 mM stock in DMF). The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 2 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant PEGylated CRM_{197} was analyzed by ESI-TOF-MS and SEC; the sample was kept on ice until $\leq$2 min prior to injection.

Determination of solvent-accessible surface area (SASA) of lysines in CRM_{197}. Total SASA of lysine residues was determined using BioLuminate (Schrödinger, LLC, USA). The crystal structure of monomeric DT (PDB ID: 1MDT) was analyzed as an analog of closed form CRM_{197}. Total SASA of lysines in the open form of CRM_{197} was determined by characterization of an isolated monomer from the crystal structure of CRM_{197} dimer (PDB ID: 1AE0). Change in SASA was not calculated for lysine residues that were either incomplete or missing from one or both crystal structures.

Preparation of aldehyde-CRM_{197} via reductive amination. Protocol adapted from [34]. In a 1.5 mL Eppendorf tube on ice, 10 μM CRM_{197} (166 μM stock in 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5) and 1 mM aldehyde (500 mM stock in DMSO) were added to 200 μL DPBS pH 8.0 buffer. To this was added 20 mM of freshly prepared $NaCNBH_3$ (100 mM in methanol). The solution was incubated at 22 °C for 20 h, and then purified via five buffer exchanges using a 30 kDa MWCO spin concentrator. A portion of the resulting mixture was diluted ten-fold with H_2O and promptly analyzed by ESI-TOF-MS.

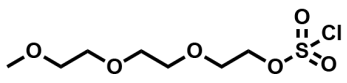
Preparation of benzaldehyde-CRM_{197} conjugates via reductive alkylation. Protocol adapted from [31]. In a 1.5 mL Eppendorf tube on ice, 50 μM CRM_{197} (166 μM stock in 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5) and 1 mM aldehyde (500 mM stock in DMSO) were added to 300 μL 50 mM phosphate with 100 mM formate, pH 7.6. To this was added 3 μL of the Ir-catalyst solution (5 mM in dH_2O). The solution was incubated at 22 °C for 20 h, and then purified via five buffer exchanges using a 30 kDa MWCO spin concentrator. A portion of the resulting mixture was diluted ten-fold with H_2O and promptly analyzed by ESI-TOF-MS.

1.5.3 Small molecule synthesis

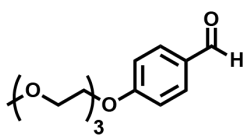


***N,N'*-(2-hydroxypropane-1,3-diyl)bis(2-iodoacetamide) (HP-BIA, 1)**. Protocol from [27]. To a solution of 1,3-diamino-2-propanol (0.47 g, 5.22 mmol, 1 equiv.) in 50% H_2O/MeOH (15 mL) was added 4-nitrophenyl iodoacetate (3.08 g, 10.03 mmol, 1.92 equiv.) portion-wise at room temperature under N_2. After 80 min, MeCN was added and the resulting yellow precipitate was

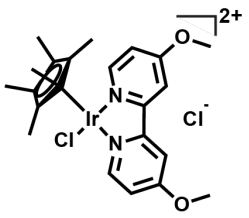
collected by filtration and washed with MeCN to afford the desired product as a white solid (960 mg, 43%). ^{1}H NMR (400 MHz, DMSO-d$_6$) δ 8.22 (t, J = 5.7 Hz, 2H), 5.05 (d, J = 5.0 Hz, 1H), 3.66 (s, 4H), 3.58–3.48 (m, 1H), 3.09 (dt, J = 13.6, 5.4 Hz, 2H), 2.99 (dt, J = 13.1, 6.1 Hz, 2H). MS calculated for $C_7H_{12}I_2N_2O_3$ is 425.89, found 426.5 $[M+H]^+$.



Methane sulfonyl tri(ethylene glycol) monomethyl ether (2a). Protocol adapted from [31]. To a 25-mL round-bottom flask equipped with a magnetic stirring bar was added tri(ethylene glycol) monomethyl ether (780 μL, 5.0 mmol), *N,N*-diisopropylethylamine (960 μL, 5.5 mmol), and dichloromethane (12.5 mL). The mixture was cooled to 0 °C and then methane sulfonyl chloride (430 μL, 5.5 mmol) was added via syringe. The reaction mixture was stirred for 1.5 h at 0 °C. The reaction solution was concentrated under reduced pressure and then residue was washed with brine, and partitioned between hexanes in a separatory funnel. After isolating the aqueous phase, the solution was extracted with three 5.0 mL portions of dichloromethane, and the combined organic phases were dried over anhydrous Na_2SO_4 and concentrated under reduced pressure to afford the product as a yellow oil (900 mg, 68% yield) which was used without further purification. ^{1}H NMR (500 MHz, $CDCl_3$): δ 4.35 (t, 2H, J = 4.5), 3.73 (t, 2H, J = 4.5), 3.63 (m, 6H), 3.52 (m, 2H), 3.34 (s, 3H), 3.05 (s, 3H).

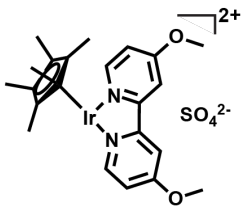


4-[tri(ethylene glycol) monomethyl ether] benzaldehyde (2). Protocol adapted from [31]. In a 10-mL scintillation vial equipped with a magnetic stirring bar were combined **2a** (262 mg, 1.0 mmol), 4-hydroxybenzaldehyde (134 mg, 1.1 mmol), cesium carbonate (358 mg, 1.1 mmol), and THF (2.5 mL). The mixture was heated to 70 °C for 18 h. The mixture was cooled to room temperature, and filtered to remove any solids. The filtrate was concentrated under reduced pressure and the resulting oil was purified by silica gel chromatography eluting with 1:1 ethyl acetate/hexanes. Fractions containing product were combined and concentrated under reduced pressure and dried *in vacuo*. The product was obtained as a clear oil (170 mg, 62% yield). ^{1}H NMR (500 MHz, $CDCl_3$): δ 9.96 (s, 1H), 7.45 (m , 2H), 7.39 (app. s, 1H), 7.21 (m, 1H), 4.19 (t, 2H, J = 4.6), 3.88 (t, 2H, J = 4.5), 3.73 (m, 2H), 3.68 (m, 2H), 3.64 (m, 2H), 3.54 (m, 2H), 3.37 (s, 3H). MS calculated for $C_{14}H_{20}O_5$ $[M]^+$ 268.13, found 269.13 $[M+H]^+$.



Cp*Ir(4,4'-dimethoxy-2,2'-bipyridine)Cl$_2$ (3a). Protocol from [31]. In a 10-mL scintillation vial equipped with a magnetic stirring bar were combined dichloro(pentamethylcyclopentadienyl)iridium (III) dimer (16.0 mg, 20.1 μmol), 4,4'-dimethoxy2,2'-bipyridine (8.7 mg, 40.2 µmol), and 2 mL of methanol. The heterogeneous mixture was stirred at room temperature until it became homogeneous (< 10 min). The solution was concentrated under reduced pressure and the residue was redissolved in a minimum amount of methylene chloride. The product was then precipitated by the dropwise addition

of hexanes until no more precipitate appeared. The precipitate was collected by filtration, washed with three 1 mL portions of hexanes and dried *in vacuo* to yield the product as a light yellow solid (21 mg, 81% yield). ^{1}H NMR (500 MHz, $CDCl_3$): δ 9.13 (d, 2H, J = 2.5), 8.38 (d, 2H, J = 6.5), 7.09 (dd, 2H, J = 6.5, 2.7), 4.39 (s, 6H), 1.65 (s, 15H). MALDI-MS calculated for $C_{22}H_{27}ClIrN_2O_2$ $[M\text{-}Cl]^+$ 577.1367, found 577.15.



Cp*Ir(4,4'- dimethoxy-2,2'-bipyridine)SO_4 (3). Protocol from [31]. In a 10-mL scintillation vail equipped with a magnetic sitr bar were combined **3a** (16.4 mg, 26.7 μmol), silver (I) sulfate (8.4 mg, 26.9 µmol), and 2 mL of ddH_2O. The heterogeneous mixture was stirred overnight at room temperature. The mixture was filtered to remove the precipitate, and the collected material was washed with three 1 mL portions of ddH_2O. The filtrate and washings were combined and the solvent was removed under reduced pressure. The product was obtained as a dull yellow solid (16.3 mg, 95% yield). ^{1}H NMR (500 MHz, D_3COD): δ 8.80 (d, 2H, J = 6.5), 7.94 (s, 2H), 7.21 (d, 2H, J = 6.2), 3.96 (s, 6H), 1.56 (s, 15H). MALDI-MS calculated for $C_{22}H_{28}IrN_2O_6$ $[M+H]^+$ 639.1274, found 639.15.

1.6 References

[1] S. A. Plotkin, "Vaccines: past, present and future", Nature Medicine **11**, S5–S11 (2005).

[2] B. Greenwood, "The contribution of vaccination to global health: past, present and future", Philosophical Transactions of the Royal Society B: Biological Sciences **369**, 20130433 (2014).

[3] L. H. Jones, "Recent advances in the molecular design of synthetic vaccines", Nature Chemistry **7**, 952–960 (2015).

[4] P. Costantino, R. Rappuoli, and F. Berti, "The design of semi-synthetic and synthetic glycoconjugate vaccines", Expert Opinion on Drug Discovery **6**, 1045–1066 (2011).

[5] E. DeGregorio, and R. Rappuoli, "From empiricism to rational design: a personal perspective of the evolution of vaccine development", Nature Reviews Immunology **14**, 505–514 (2014).

[6] J. Zhu, J. D. Warren, and S. J. Danishefsky, "Synthetic carbohydrate-based anticancer vaccines: the Memorial Sloan-Kettering experience", Expert Review of Vaccines **8**, 1399–1413 (2009).

[7] A. Fettelschoss, F. Zabel, and M. F. Bachmann, "Vaccination against Alzheimer disease", Human Vaccines & Immunotherapeutics **10**, 847–851 (2014).

[8] T. H. Do, Y. Chen, V. T. Nguyen, and S. Phisitkul, "Vaccines in the management of hypertension", Expert Opinion on Biological Therapy **10**, 1077–1087 (2010).

[9] J. Thorn, K. Bhattacharya, R. Crutcher, J. Sperry, C. Isele, B. Kelly, L. Yates, J. Zobel, N. Zhang, H. Davis, and M. McCluskie, "The effect of physicochemical modification on the function of antibodies induced by anti-nicotine vaccine in mice", Vaccines **5**, 11 (2017).

[10] P. T. Bremer, J. E. Schlosburg, M. L. Banks, F. F. Steele, B. Zhou, J. L. Poklis, and K. D. Janda, "Development of a clinically viable heroin vaccine", Journal of the American Chemical Society **139**, 8601–8611 (2017).

[11] P. T. Bremer, A. Kimishima, J. E. Schlosburg, B. Zhou, K. C. Collins, and K. D. Janda, "Combatting synthetic designer opioids: a conjugate vaccine ablates lethal doses of fentanyl class drugs", Angewandte Chemie **128**, 3836–3839 (2016).

[12] M. Bröker, F. Berti, J. Schneider, and I. Vojtek, "Polysaccharide conjugate vaccine protein carriers as a "neglected valency" – potential and limitations", Vaccine **35**, 3286–3294 (2017).

[13] E. Malito, B. Bursulaya, C. Chen, P. L. Surdo, M. Picchianti, E. Balducci, M. Biancucci, A. Brock, F. Berti, M. J. Bottomley, M. Nissum, P. Costantino, R. Rappuoli, and G. Spraggon, "Structural basis for lack of toxicity of the diphtheria toxin mutant CRM197", Proceedings of the National Academy of Sciences **109**, 5229–5234 (2012).

[14] M. Tontini, F. Berti, M. Romano, D. Proietti, C. Zambonelli, M. Bottomley, E. D. Gregorio, G. D. Giudice, R. Rappuoli, P. Costantino, G. Brogioni, C. Balocchi, M. Biancucci, and E. Malito, "Comparison of CRM197, diphtheria toxoid and tetanus toxoid as protein carriers for meningococcal glycoconjugate vaccines", Vaccine **31**, 4827–4833 (2013).

[15] A. K. Prasad, J.-h. Kim, and J. Gu, "Design and development of glycoconjugate vaccines", in *Carbohydrate-based vaccines: from concept to clinic* (American Chemical Society, Jan. 2018), pp. 75–100.

[16] Q.-Y. Hu, F. Berti, and R. Adamo, "Towards the next generation of biomedicines by site-selective conjugation", Chemical Society Reviews **45**, 1691–1719 (2016).

[17] V. A. Ferro, M. A. H. Khan, E. R. Earl, M. J. A. Harvey, A. Colston, and W. H. Stimson, "Influence of carrier protein conjugation site and terminal modification of a GnRH-i peptide sequence in the development of a highly specific anti-fertility vaccine. part i", American Journal of Reproductive Immunology **48**, 361–371 (2002).

[18] J. Peeters, T. Hazendonk, E. Beuvery, and G. Tesser, "Comparison of four bifunctional reagents for coupling peptides to proteins and the effect of the three moieties on the immunogenicity of the conjugates", Journal of Immunological Methods **120**, 133–143 (1989).

[19] J. Briand, S. Muller, and M. V. Regenmortel, "Synthetic peptides as antigens: pitfalls of conjugation methods", Journal of Immunological Methods **78**, 59–69 (1985).

[20] R. G. Arumugham, A. K. Prasad, and M. Hagen, "A-b immunogenic peptide carrier conjugates and methods of producing same", US8227403 (July 24, 2012).

[21] M. Kent, H. Yim, J. Murton, S. Satija, J. Majewski, and I. Kuzmenko, "Oligomerization of membrane-bound diphtheria toxin (CRM197) facilitates a transition to the open form and deep insertion", Biophysical Journal **94**, 2115–2127 (2008).

[22] B. Steere, and D. Eisenberg, "Characterization of high-order diphtheria toxin oligomers", Biochemistry **39**, 15901–15909 (2000).

[23] D. T. Crane, B. Bolgiano, and C. Jones, "Comparison of the diphtheria mutant toxin, crm197, with a haemophilus influenzae type-b polysaccharide-crm197 conjugate by optical spectroscopy", European Journal of Biochemistry **246**, 320–327 (1997).

[24] S. Pecetta, P. L. Surdo, M. Tontini, D. Proietti, C. Zambonelli, M. Bottomley, M. Biagini, F. Berti, P. Costantino, and M. Romano, "Carrier priming with CRM197 or diphtheria toxoid has a different impact on the immunogenicity of the respective glycoconjugates: biophysical and immunochemical interpretation", Vaccine **33**, 314–320 (2015).

[25] S. Crotti, H. Zhai, J. Zhou, M. Allan, D. Proietti, W. Pansegrau, Q.-Y. Hu, F. Berti, and R. Adamo, "Defined conjugation of glycans to the lysines of CRM197 guided by their reactivity mapping", ChemBioChem **15**, 836–843 (2014).

[26] G. T. Hermanson, *Bioconjugate techniques*, 3rd ed. (Academic Press, 2013) Chap. 3, pp. 127–228.

[27] J.-Y. Chang, U. Ramseier, T. Hawthorne, T. O'Reilly, and J. van Oostrum, "Unique chemical reactivity of His-21 of CRM-197, a mutated diphtheria toxin", FEBS Letters **427**, 362–366 (1998).

[28] U. Möginger, A. Resemann, C. E. Martin, S. Parameswarappa, S. Govindan, E.-C. Wamhoff, F. Broecker, D. Suckau, C. L. Pereira, C. Anish, P. H. Seeberger, and D. Kolarich, "Cross reactive material 197 glycoconjugate vaccines contain privileged conjugation sites", Scientific Reports **6** (2016) `10.1038/srep20488`.

[29] J. Qiao, K. Ghani, and M. Caruso, "Diphtheria toxin mutant CRM197 is an inhibitor of protein synthesis that induces cellular toxicity", Toxicon **51**, 473–477 (2008).

[30] G. L. Ellman, "Tissue sulfhydryl groups", Archives of Biochemistry and Biophysics **82**, 70–77 (1959).

[31] J. M. McFarland, and M. B. Francis, "Reductive alkylation of proteins using iridium catalyzed transfer hydrogenation", Journal of the American Chemical Society **127**, 13490–13491 (2005).

[32] J. M. Hooker, A. P. Esser-Kahn, and M. B. Francis, "Modification of aniline containing proteins using an oxidative coupling strategy", Journal of the American Chemical Society **128**, 15558–15559 (2006).

[33] A. M. ElSohly, J. I. MacDonald, N. B. Hentzen, I. L. Aanei, K. M. E. Muslemany, and M. B. Francis, "Ortho-methoxyphenols as convenient oxidative bioconjugation reagents with application to site-selective heterobifunctional cross-linkers", Journal of the American Chemical Society **139**, 3767–3773 (2017).

[34] N. Jentoft, and D. G. Dearborn, "Labeling proteins by reductive methylation using sodium cyanoborohydride", Journal of Biological Chemistry **254**, 4359–4365 (1978).

Chapter 2

Engineering CRM197 carrier protein for development of a structurally homogeneous conjugate material

ABSTRACT: The immunogenic properties of a synthetic vaccine are often a result of its structural features. The characterization of well-defined materials can enable the modularity of synthetic vaccine design. Carrier proteins such as cross-reactive material 197 (CRM_{197}), a non-toxic mutant of diphtheria toxin, have been characterized for use in clinically-relevant studies. We previously reported on a peptide-CRM_{197} conjugate material for use as a conjugate vaccine. Our synthetic methods resulted in two distinct populations of conjugate material, believed to be the result of a conformational change of CRM_{197} during preparation. In this chapter, we describe the use of protein engineering to produce CRM_{197} mutants that convey conformational stability for the conjugate material. A variety of bacterial expression and purification strategies were explored to obtain correctly folded recombinant CRM_{197}. One such method facilitated the successful soluble production of recombinant CRM_{197} and mutants, which will undergo structural analysis before and after peptide conjugation to determine the conformation of the engineered protein conjugate material.

2.1 Introduction

In recent years, the development of synthetic vaccines has become increasingly common due to advances in the characterization of immunogenetic materials and the potential for modularity of individual components [1, 2]. These vaccines, composed of synthetic antigens such as peptides, carbohydrates, or haptens (antigenic small molecules), can be designed to elicit a precise immune response to a specific target. As our understanding of the immune system increases, the number of potential targets also increases, which underscores the desire for modular approaches to drug design. However, one drawback is that these individual components often suffer from poor immunogenicity or delivery, and as such, conjugation to a carrier moiety is a commonly employed strategy. Carrier proteins, in particular, can boost an immune response to activate helper T-cells in a previously immunized individual [2, 3].

One extensively studied carrier protein is cross-reactive material 197 (CRM_{197}), a non-toxic variant derived from diphtheria toxin (DT), an exotoxin secreted by a pathogenic bacterium that causes diphtheria [4, 5]. Surprisingly, a single point mutation (Gly to Glu at position 52) renders the CRM_{197} non-toxic, while maintaining the structure and T-cell epitopes of the native toxin [6, 7]. Thus, CRM_{197} is a commonly studied and clinically relevant carrier protein, namely due to its apparent lower susceptibility than DT to pre- or co-commitment immunization with other vaccines [6, 8]. Efforts have been directed towards the synthesis of conjugate vaccines by decorating the surface of CRM_{197} with multiple copies of antigen via chemical linkers coupled to the many surface-available lysine residues on the protein [3, 9, 10]. Notably, the mapped T-cell epitopes of CRM_{197}, necessary for initial immune system recognition, lack lysine residues used for conjugation [7]. The resulting antigen-loaded conjugate material can elicit T-cell recognition that enables B-cell immune receptors to crosslink, which is important for initiating a specific antibody response to the presented antigen [2].

One application of CRM_{197} as a carrier protein is work towards the development of a peptide-CRM_{197} conjugate vaccine [11]. This material is developed from the activation of surface available amines (i.e. lysines or the N terminus) with an *N*-hydroxysuccinimide (NHS) ester bifunctional linker, followed by the conjugation of the cysteine-containing peptides, with the thiol reactive handle of the bifunctional linker (maleimide or haloacetamide). The identity of the peptide component can be diversified, provided that there is one maintained cysteine residue for successful conjugation.

In our studies it was discovered that the resulting peptide-CRM_{197} conjugate material is actually definitively conformationally heterogeneous. After structural analysis of CRM_{197}, and additional bifunctional linker screening, it was deduced that the conformational differences were in part due to non-specific modification of a particularly reactive hisidine-21 residue [11]. Additionally, it was noted that intraprotein crosslinking could occur at this His21 site and a neighboring lysine-24 residue, which subsequently altered the conforma-

tional distribution of the conjugate material further [11, 12].

As structure-immunogenicity relationships exist, we are interested in developing well-characterized, structurally homogeneous material to best optimize the therapeutic potential of this conjugate vaccine [13]. To do so, we seek to create a stable, uniform population of CRM_{197} protein monomers through protein engineering strategies. This conformational homogeneity of the carrier protein will facilitate development of well-characterized conjugate material, and enable therapeutic analysis for the structurally-defined synthetic vaccine.

2.2 Results and Discussion

Previous work from the lab examined the structural attributes of CRM_{197} in the context of the peptide conjugate material [11]. The uniquely reactive His21 is observed to be near a domain interface (C and R, **Figure 2.1ab**), which is hypothesized to contribute to the two different conformational populations observed within the resulting peptide-CRM_{197} conjugates. Blocking this site, by mutating His21 to an alanine or other non-reactive amino acid, could ablate the reported off-target bifunctional linker reactivity, thus maintaining the native "closed monomer" configuration. Alternatively, mutating His21 to a bulky residue, such as a tyrosine, could disrupt this domain interface, and promote a primed "open monomer" configuration for peptide conjugation. This could facilitate modification of the interfacial lysine residues, which is postulated to be another driving factor for the "open" population.

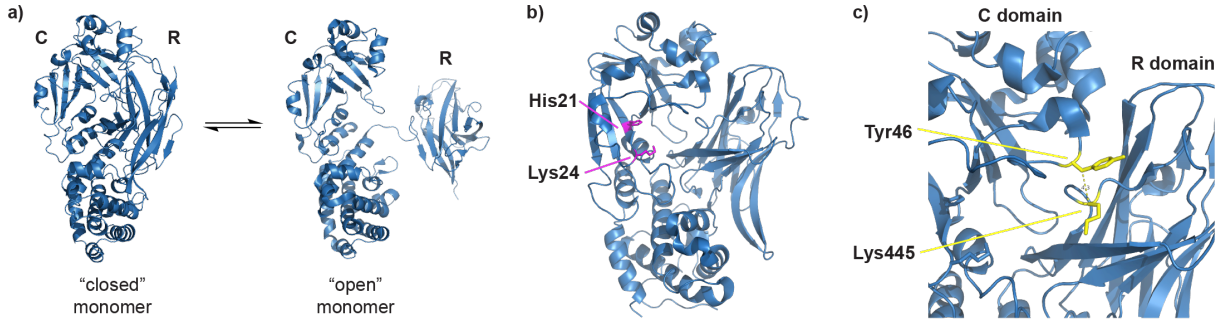


Figure 2.1: Conformational changes and proposed mutants of CRM_{197}. (a) CRM_{197} is found in the "closed monomer" conformation when expressed from its native host. Disruption of the interface between the catalytic (C) and receptor-binding (R) domains of monomeric CRM_{197} is hypothesized to result in the "open monomer" form. Structures modeled from DT monomer structure (PDB ID: 1MDT) and CRM_{197} dimer structure (PDB ID: 4AE0). (b) The His21 residue of interest sits in the domain interface of C and R, and when chemically activated, can react with neighboring Lys24. An alanine mutation is proposed to ablate reactivity at this site (His21Ala), and a tyrosine mutation is proposed to perturb the domain interface to promote the open conformation (His21Tyr). (c) To maintain the closed conformation, domain stapling via disulfide bond engineering is proposed. Two sites, Tyr46 and Lys445, are observed to be 4.3 Å apart (distance to/from the β-carbon of the side chain), and would permit cysteine mutations.

Another proposed mutation is the creation of a disulfide bond staple within the domain interface. Introducing a disulfide bond into a protein structure can be challenging, especially if the protein has pre-existing disulfide bonds. The sites of mutation must be chosen carefully

to not only allow the cysteine mutation, but also to promote bond formation [14]. Two residues, within 4 Å are located in the domain interface (**Figure 2.1c**). Tyrosine-46 and lysine-445 are within proximity and also should accept the mutation without much additional structural perturbation .

To introduce these mutations, we sought the use of a bacterial host to synthesize the recombinant CRM_{197} mutants. While CRM_{197} is robustly expressed and purified from its native host, *Corynebacterium diphtheriae*, reports of recombinant CRM_{197} from bacterial expression have proven challenging.

2.2.1 Inclusion body recovery

Expression of CRM_{197} from *Escherichia coli* has been reported using a variety of methods. Due to its two native disulfide bonds and potential host toxicity, bacterial expression of CRM_{197} is often reported to be recovered from insoluble inclusion bodies (IBs) [15]. Bacterial IB expression can be useful for therapeutically relevant proteins because it can facilitate endotoxin removal without hindering recovery of the protein of interest [16]. While overexpression of the desired protein is feasible, solubility and recovery of the aggregated protein can be quite challenging and occurs with limited success.

Stefan *et al.* reported the successful recovery and purification of a His-tagged and slightly truncated CRM_{197} construct from IBs, via expression from BL21AI *E. coli* cells [17]. The cell line used is particularly good for the expression of toxic proteins. The recombinant protein of Stefan *et al.* was solubilized using urea, and then refolded on an immobilized metal affinity chromatography (IMAC) column. In another report, Park *et al.* used detergent-mediated solubilization of a His-tagged CRM_{197} construct expressed from ClearColi BL21(DE3) *E. coli* cells [18]. This cell line is genetically modified not to trigger any endotoxin production. The recombinant protein of Park *et al.* was solubilized using a solution of *N*-Lauroylsarcosine sodium salt (sarkosyl) and then subsequently refolded using a drop-wise addition of 1% Triton X-100 and 10 mM 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS). Purification was conducted by IMAC and size-exclusion chromatography (SEC).

Thus, our initial efforts to obtain recombinant CRM_{197} were directed towards using a combination of the aforementioned IB recovery and purification methods. A bacterial codon-optimized *crm197* gene with a six-histidine tag and a factor ten A protease (fXa) cleavage site at the N terminus was synthesized and inserted into a pBAD vector (under an *araBAD* promoter), using standard BsaI Golden Gate cloning techniques [5]. The plasmid was transformed into several different *E. coli* cell lines and expression conditions were examined. Tuner(DE3) and MON105 cells produced the most robust expression of the CRM_{197} construct. Upon cellular lysis via sonication, the recombinant CRM_{197} aggregated and required solubilization, as anticipated.

First, solubilization by detergent was explored, following the protocol of Park and co-workers [18]. The transformed Tuner(DE3) cells were harvested after an overnight expression. The cells were lysed by sonication and then analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE, **Figure 2.2a**). The insoluble fraction was incubated with sarkosyl salt over several hours and gentle stirring, prior to refolding by the slow addition of more detergents (1% Triton X-100 and CHAPS) to induce protein folding. The solution was then applied to a nickel nitrilotriacetic acid (NiNTA) agarose resin column for further purification (**Figure 2.2b**). Characterization of the refolded protein by SEC and electrospray-ionization time-of-flight mass spectrometry (ESI-TOF-MS) did not suggest appropriate recovery and folding of the solubilized construct (**Figure 2.2bc**).

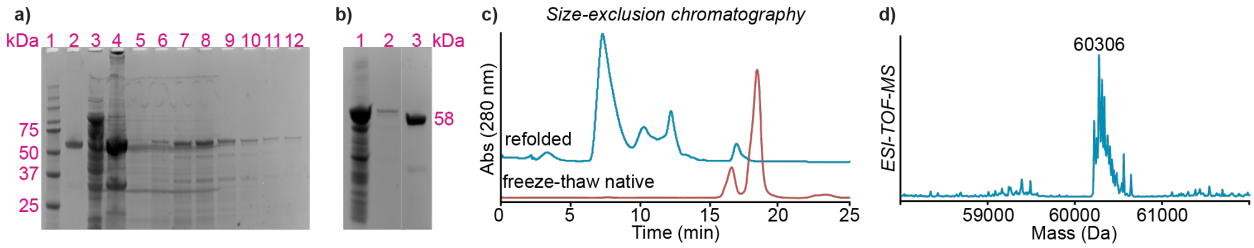


Figure 2.2: Detergent-mediated recovery of recombinant His-tagged CRM_{197}. (a) SDS-PAGE of Tuner(DE3) cellular lysis shows the His-tagged CRM_{197} in the insoluble fraction (lane 4). The aggregated fraction was solubilized with a two-hour incubation in 50 mM Tris-HCl pH 7.6 buffer with 1% sarkosyl salt. The resulting solution was then refolded by the drop-wise addition of 1% Triton X-100 and 10 mM CHAPS. The refolded solution was immediately applied to a 5 mL NiNTA column and washed with 50 mM sodium phosphate pH 7.6 with 150 mM NaCl. Fractions were analyzed by SDS-PAGE. Lane 1: protein standard, 2: native CRM_{197}, 3: lysed soluble fraction, 4: lysed insoluble fraction, 5-12: NiNTA fractions. (b) For further purification, CRM_{197} containing fractions were applied to a NiNTA spin column and buffer exchanged. Lane 1: combined NiNTA fractions containing His_6-CRM_{197}, 2: purified His_6-CRM_{197}, 3: native CRM_{197}. (c) Comparison by SEC of the refolded construct and natively expressed CRM_{197} (freeze-thawed to induce dimerization for peak comparison) shows little overlap. (d) ESI-TOF-MS analysis shows the correct mass of the recombinant CRM_{197} along with a substantial number of adducts, suggesting multiple protein states of questionable integrity.

One challenge experienced was the persistence of detergents, even after dialysis and buffer exchange. Furthermore, precipitation was observed after the chromatography steps, indicating that perhaps the solubilizing or refolding agents were not completely successful in recovering or refolding the recombinant CRM_{197}.

Next, the use of chaotropes to recover recombinant CRM_{197} from IBs was explored. Following Stefan *et al.*, recovery of the aggregated protein from MON105 *E. coli* cells was initiated with a 6 M urea incubation for 2 h (**Figure 2.3a**) [17]. In some cases, incomplete solubilization of the protein was observed by SDS-PAGE, and thus the incubation time and chaotrope were altered to a 4 h incubation at 16 °C using 6 M guanidinium as the chaotropic agent. The solubilized supernatant was then applied to a NiNTA agarose column for buffer exchange, refolding, and purification. As described in Stefan *et al.*, the sample was subjected to a slow flow of buffer of decreasing chaotrope concentration in order to promote protein folding on the column. An imidazole gradient was then applied to the column to release the recombinant His-tagged CRM_{197} (**Figure 2.3bc**).

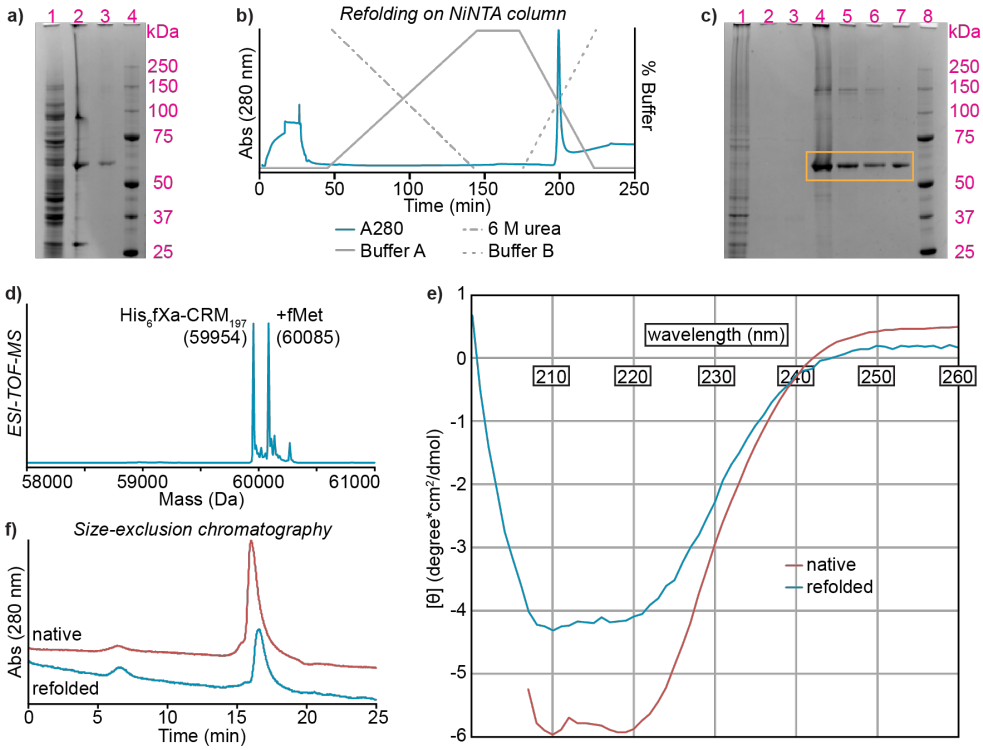


Figure 2.3: Detergent-mediated recovery of recombinant His-tagged CRM_{197}. (a) SDS-PAGE gel of expression from MON105 *E. coli* cells. The band in lane 3 indicates solubilization of the protein construct using buffer A (50 mM Tris HCl, 150 mM NaCl, pH 7.6) with 6 M urea. Lane 1: lysed soluble fraction, 2: lysed insoluble fraction, 3: solubilized supernatant, 4: protein standards. (b) FPLC trace of protein refolding and purification using IMAC. The solution was applied to NiNTA column, and washed with a gradient of 0-100% buffer A over 150 min. The construct was then eluted from the column using a gradient of buffer B (buffer A + 500 mM imidazole). (c) Fractions containing His_6-fXa-CRM_{197} were analyzed by SDS-PAGE gel. Lane 1-3: column flow-through (peak 1), 4-7: peak around 200 min, 8: protein standards. (d) Characterization by ESI-TOF-MS shows both the His_6-fXa-CRM_{197} construct with and without fMet. (e) CD spectroscopy was used to further analyze the refolded His_6-fXa-CRM_{197} construct (blue) versus the natively expressed CRM_{197} (red). (f) SEC analysis of refolded His_6-fXa-CRM_{197} construct (blue) compared to natively expressed CRM_{197} (red).

It should be noted that the imidazole concentration to elute protein was quite high, as the recombinant His-tagged CRM_{197} displayed a strong affinity for the NiNTA resin. Regardless, the refolded protein was eluted eventually and subjected to various characterization tests, including ESI-TOF-MS, circular dichroism (CD) spectroscopy, and SEC (**Figure 2.3d-f**). Some aggregation of the isolated CRM_{197} was observed, probably represented by the peak around five-minutes in the SEC. While the refolded recombinant protein did have similar structural characteristics as compared to the CRM_{197} from the native host, we noticed that the isolated His-tagged CRM_{197} was not stable, even in buffers with glycerol or sucrose additives. One challenge experienced was the incomplete removal of chaotrope and imidazole from the purified fractions. Furthermore, no DNAse activity was observed following reported literature protocols [17]. This led us to believe that the recombinant protein was perhaps not completely in the active or stable conformation.

It was hypothesized that perhaps extending or slowing the conditions during the refolding step could facilitate correct folding. Refolding via dialysis could allow for more gradual buffer condition manipulation. Furthermore, additives such as arginine, lysine, or glutathione (oxidize and reduced versions), could help facilitate proper refolding of the protein of interest with disulfide bonds [19]. While there is an infinite number of variations and combinations of additives to try, manipulation of oxidized/reduced glutathione was attempted. Unfortunately, characterization of recombinant CRM_{197} from dialysis-assisted refolding also did not suggest correct three-dimensional structure.

2.2.2 Cell-free protein synthesis

Cell-free protein synthesis (CFPS) has become a valuable and modular approach to obtaining recombinant proteins that are otherwise toxic to the bacterial host or result in insoluble bacterial expression [20, 21]. CFPS combines only the critical expression components of cellular machinery, avoiding the non-essential cellular components that often lead to recombinant expression issues, and lowers the time required to obtain pure protein. Further, kits and various optimized additives are easily purchased.

A CFPS kit with disulfide bond enhancer mixture was purchased from New England Biolabs Inc. and used for the expression of an untagged CRM_{197} construct in a pET28b vector under a T7 promoter (**Figure 2.4a**). According to manufacturer's instructions, the CFPS solution was combined with the disulfide bond enhancer and CRM_{197} construct, for a three-hour 37 °C incubation. Stain-free SDS-PAGE was employed, along with Western blotting, to analyze the resulting protein synthesis (**Figure 2.4b**).

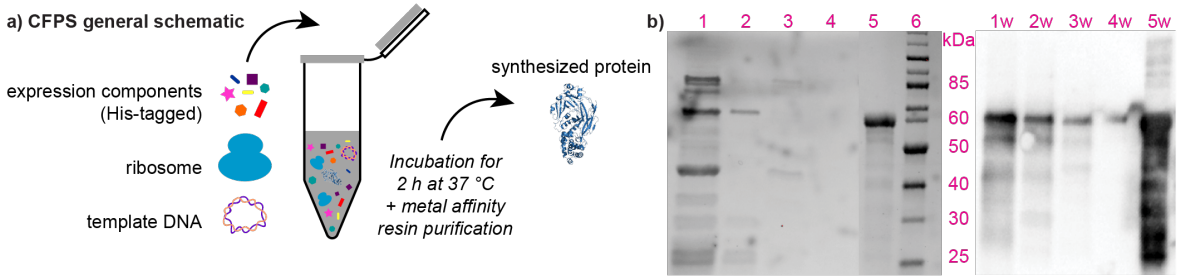


Figure 2.4: Cell-free protein synthesis (CFPS) to synthesize CRM_{197}. (a) Generalized schematic of CFPS, adapted from NEB PURExpress In Vitro Protein Synthesis Kit product information. Briefly, 25 μL reactions containing template DNA and His-tagged kit components (including disulfide bond enhancer solution) are incubated together at 37 °C for 2 h. The synthesized protein can be purified away from His-tagged components via IMAC. (b) SDS-PAGE and subsequent Western blot of optimized CFPS reaction shows successful recombinant CRM_{197} synthesis and purification. Lane 1: crude CFPS reaction, 2-3: washes of TALON resin, 4: 50 mM imidazole wash of TALON resin, 5: pure native CRM_{197}, 6: PageRuler Unstained Protein Ladder (Thermo). "w" represents corresponding Western blot (same lane assignments).

CRM_{197} expression was immediately successful with the disulfide bond enhancer; however, recovery of the protein was nontrivial. As previously observed, CRM_{197} appeared to

have strong affinity for NiNTA resin. Thus, while the kit components were all His-tagged and IMAC could be employed to collect the untagged CRM_{197}, the concentration of imidazole required to elute off CRM_{197} was higher than desired. Instead of NiNTA resin, TALON, a cobalt-based resin with lower His-tag affinity, was used with success (**Figure 2.4b**). However, as apparent in the SDS-PAGE image, the one downside of CFPS is the lower yield of protein; ultimately this was the limitation of this method.

While CFPS did not ultimately lead to optimal production of recombinant CRM_{197} for structural analysis, CFPS is an attractive method for non-canonical amino acid incorporation [22]. To obtain CRM_{197} with uniquely reactive handles for the purpose of developing novel peptide-CRM_{197} conjugate materials, CFPS could be employed to generate small amounts of CRM_{197} with orthogonal reactive handles for site-specific cargo loading.

2.2.3 Fusion protein constructs

Due to the difficulties in yield and refolding of insoluble recombinant CRM_{197}, efforts turned towards the production of soluble protein. The lack of an efficient, concise, recipe to express soluble protein from *E. coli* remains an issue within the industrial protein purification field [23, 24]. While screening for optimal expression conditions can lead to finding conditions for soluble expression, many alternative methods with more reliably positive outcomes exist.

In the case of recombinant CRM_{197} from *E. coli*, one technique reported involves the co-expression of the protein of interest with molecular chaperones [25]. Chaperone proteins are essential for the native expression of cellular proteins, as they prevent the aggregation and misfolding of newly synthesized proteins. Mahamad *et al.* screened co-expression conditions for a recombinant CRM_{197} construct with several common molecular chaperones and found the co-expression of trigger factor, in a particular cell line under specific expression conditions, led to the soluble production of recombinant CRM_{197} [25].

In a similar vein, the use of fusion tags, which are hyper-soluble proteins or peptides that are genetically fused to the protein of interest, can help enhance soluble expression and subsequent folding [23]. Many fusion partners have been characterized and optimized for use in generalized, high-throughput methodology. Some common partners include maltose-binding protein (MBP) [26], thioredoxin (Trx) [27], and small ubiquitin-related modifier (SUMO) [24]. These fusion partners can also be used for purification purposes, and can be cleaved from the protein of interest using a protease if a subsequent recognition site is genetically inserted.

Thus, a CRM_{197} construct was developed, with an N-terminal MBP fusion and a short linker region including a soluble FLAG tag and a Tobacco Etch Virus (TEV) protease cut site. This construct was cloned into pET28a vector using traditional Golden Gate cloning (**Figure 2.5a**). Though several cell lines were screened, Tuner(DE3) *E. coli* were ultimately

chosen due to the high protein yield. Additionally, following expression conditions were evaluated at lower temperatures; the construct with the best expression yield was incubated overnight at 16 °C (**Figure 2.5b**).

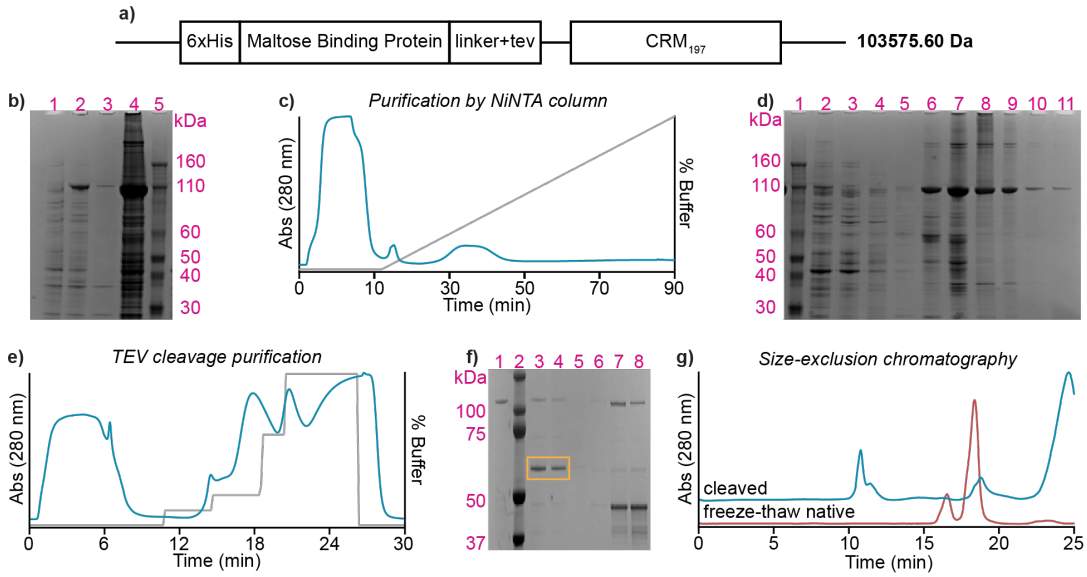


Figure 2.5: Expression, purification, and characterization of His_6-MBP-CRM_{197} construct. (a) Diagram of synthesized gene, where a His-tagged maltose binding protein (MBP) was appended to the N terminus of CRM_{197}, with a polyasparagine and FLAG peptide linker, prior to a TEV protease cleavage site. The construct is around 103 kDa. (b) The construct was expressed from Tuner(DE3) *E. coli* cells in the soluble fraction. Lane 1: pre-induction, 2: 16 h expression, 3: lysed insoluble fraction, 4: lysed soluble fraction, 5: protein standards. (c) Purification of the construct was accomplished over an NiNTA column and analyzed by SDS-PAGE (d). Lane 1: protein standards, 2-4: peak 1, 5: peak 2, 6-11: peak 3. (e) The recombinant CRM_{197} was liberated using AcTEV protease, following manufacturer's instructions. The cleaved mixture was applied to a NiNTA column for purification, as the TEV protease and MBP maintained His-tags, while the free CRM_{197} should not interact with the column. The fractions from a step-wise imidazole gradient were analyzed by SDS-PAGE (f). Lane 1: pre-cleavage construct, 2: protein standards, 3-4: flow-through (peak 1), 5: step one elution, 6: step two elution, 7: step 3 elution, 8: step 4 elution. (g) SEC was used to compare natively expressed CRM_{197} (red) and the recombinant CRM_{197} (blue). The native protein was freeze-thawed to induce dimer formation, but minimal overlap of the cleaved product was observed.

Purification of the lysed soluble fraction was initiated by applying the supernatant to a NiNTA column. The construct was eluted with an imidazole gradient and then analyzed by SDS-PAGE (**Figure 2.5cd**). Recovery of the construct was successful and subsequently treated with AcTEV protease to remove the His-tagged MBP-linker fusion. The liberated CRM_{197} was purified with another round of IMAC (**Figure 2.5ef**). The collected fractions were analyzed for structure similarity to the natively expressed CRM_{197} by comparing SEC traces (**Figure 2.5g**). Unfortunately, aggregate population, as well as a smaller fragment, was identified in the SEC. The lack of alignment with the natively expressed CRM_{197} SEC profile suggested incorrect folding or aggregation after MBP cleavage.

One common challenge with fusion protein cleavage is maintaining the stability of the liberated protein [23]. MBP is quite large in size (around 42 kDa) for a fusion tag, and

might interact with CRM_{197} (which is around 58 kDa) to promote a soluble aggregate of the construct, instead of the soluble, correctly folded species. Different buffer conditions during expression could be explored, or new constructs with different fusion proteins, such as a smaller fusion tag like SUMO (around 12 kDa), might promote correctly folded CRM_{197}.

2.2.4 Soluble expression

An alternative to fusion protein constructs was reported by Goffin *et al.*, who used a signal recognition particle pathway to direct periplasmic expression of recombinant CRM_{197}. The periplasm of *E. coli* provides an oxidizing environment where disulfide bond formation is facilitated. Through careful and thorough screening of periplasmic directing sequences and expression conditions, researchers were able to obtain soluble recombinant CRM_{197} in high yield [28].

While we were unable to reproduce the protocol and these conditions, the observations of Goffin *et al.* and knowledge from previous expression attempts guided our attempt for soluble expression of recombinant CRM_{197}. In each step, we sought to optimize for disulfide bond formation conditions. A new CRM_{197} construct was synthesized, which included a six-His tag and TEV protease cut site, cloned into a pET14b vector. One difference to the CRM_{197} sequence is that the engineered Golden Gate cloning site of the pET14b vector, which incorporates the TEV protease cut site, leaves a serine N-terminal residue (instead of the native CRM_{197} glycine). Two *E. coli* cell lines explored by Mahmad *et al.* were screened for soluble expression: OrigamiB(DE3) and SHuffle T7 Express pLysY. Further, strict expression conditions of lower density growth phase, cooler induction temperature, and minimal media were employed [29].

After an expression media screen of transfected OrigamiB(DE3) cells, much to our surprise, soluble His-tagged CRM_{197} was identified via SDS-PAGE (**Figure 2.6a**). The supernatant was applied to TALON resin for purification and further structural analysis by ESI-TOF-MS, CD spectroscopy, and SEC (**Figure 2.6b-d**). Not only did the recovered recombinant CRM_{197} appear to have similar structural characteristics as compared to CRM_{197} expressed from the native host, but cleavage of the His-tag by TEV protease did not seem to interfere with the liberated recombinant CRM_{197}.

Of note, the His-tagged CRM_{197} gene was cloned into a particular pET14b vector, which was discovered later in this expression endeavor. The pre-existing Golden Gate cloning site leaves a serine N-terminal residue (after TEV proteolysis), which, in our case, replaces the native CRM_{197} glycine. The vector also encodes for carbenicillin resistance, which was necessary for compatible protein expression from OrigamiB(DE3) *E. coli* cells, but no other notable differences to the pET28a vector (previously used, encoding for kanamycin resistance). It does seem fortuitous that this change, in addition to very careful manipulation of expression conditions, promoted such robust recombinant CRM_{197} expression, as seen in

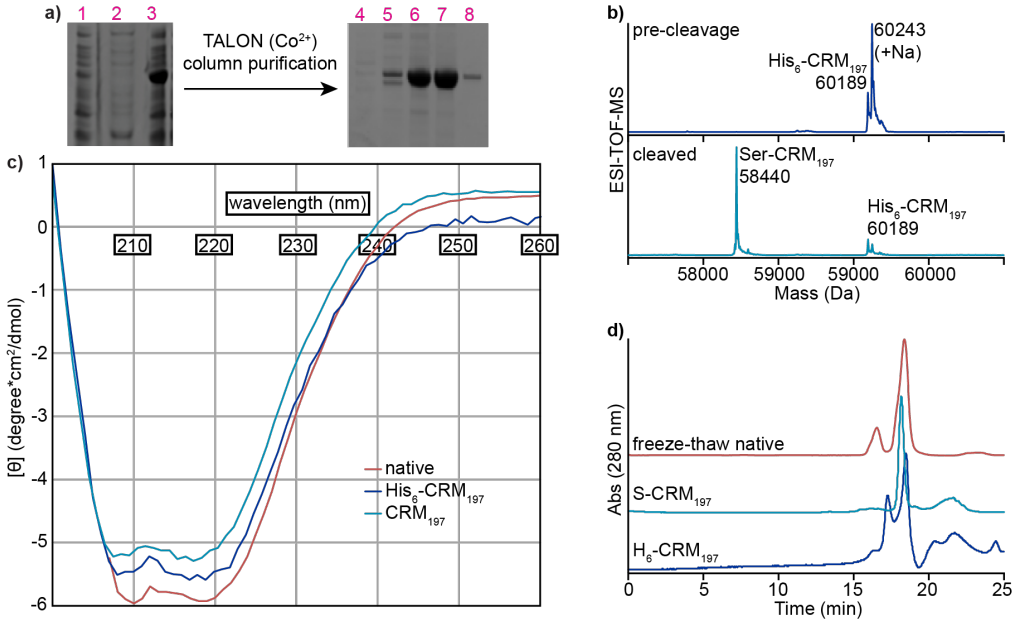


Figure 2.6: Characterization of His-tagged CRM_{197} expressed from OrigamiB(DE3) cells. (a) SDS-PAGE indicating the soluble expression and subsequent crude purification of His-tagged CRM_{197}. Purification was conducted using a TALON resin, according to manufacturer's direction. Lane 1: 0 h expression, 2: lysed insoluble fraction, 3: lysed soluble fraction, 4: wash after incubating resin with lysis supernatant, 5-7: 50 mM imidazole wash of resin to elute His-tagged CRM_{197}, 8: natively expressed CRM_{197}. Fractions containing the His-tagged CRM_{197} construct were subjected to TEV protease for His-tag cleavage. (b) ESI-TOF-MS traces pre- and post-cleavage of the His-tag. AcTEV protease (His-tagged) yielded approximately 85% un-tagged protein (lower trace), and the resulting mixture was subjected to a quick TALON resin spin column for purification (gel not shown). Note: the engineered TEV protease cut site leaves a serine N-terminal residue, replacing the native CRM_{197} glycine (c) CD spectroscopy results comparing His-tagged CRM_{197} (blue trace), un-tagged CRM_{197} (with serine N terminus, teal trace), and natively expressed CRM_{197} (red trace). Good overlap suggests correct folding of recombinant CRM_{197}. (d) SEC comparison of His-tagged CRM_{197} (blue trace), un-tagged CRM_{197} (with serine N terminus, teal trace), and natively expressed CRM_{197} (freeze-thawed to show dimer, red trace). Peak overlap further positively confirmed conformational similarity of recombinant CRM_{197} to the natively expressed version.

Figure 2.6a. This method might not be suitable for industrial scale up, which is often another challenge encountered in other CRM_{197} bacterial expression reports.

2.2.5 Peptide conjugation to recombinant CRM197

With the structural characterization results for the recombinant CRM_{197} ($rCRM_{197}$), peptide conjugate material was synthesized to ensure that the expression host did not play a role in the observed conformational changes. The same activation-conjugation protocol was followed [11]. As this was for confirmation of previous observations, we used the *N*-β-maleimidopropyl-oxysuccinimide ester (BMPS) bifunctional linker (native CRM_{197} peptide loading vs M2 graph found in **Supplemental Figure 2.8**), at both a lower (50 equiv. BMPS) and higher (150 equiv. BMPS) activation level. The material was monitored by ESI-TOF-MS prior to peptide conjugation to examine crosslinking (**Figure 2.7a**). The same

peptide (CTNEHFRG) used by Jaffe *et al.* was conjugated to $rCRM_{197}$, and then analyzed by ESI-TOF-MS and SEC (**Figure 2.7b**). The same activation levels were examined for a sample of $rCRM_{197}$ treated with bis(iodoacetamide) reagent, HPBIA, to cap His21 [11, 12].

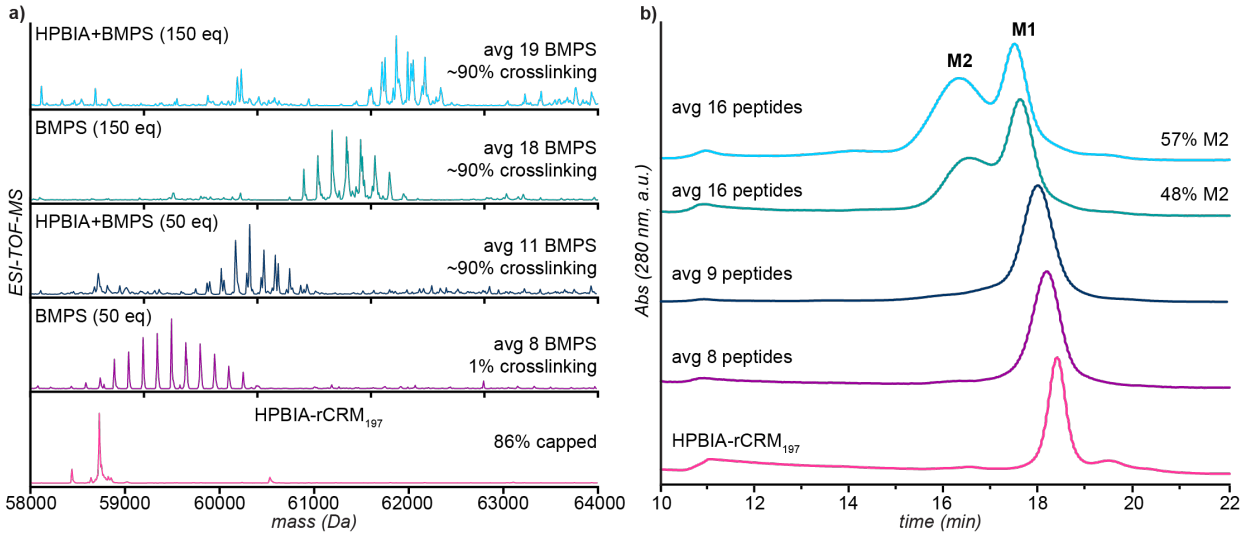


Figure 2.7: Analysis of peptide-$rCRM_{197}$ conjugate material. Using BMPS bifunctional linker, $rCRM_{197}$ and HPBIA-capped $rCRM_{197}$ was activated at both a lower (50 equiv. BMPS) and higher (150 equiv. BMPS) level. (a) Material was monitored by ESI-TOF-MS to examine crosslinking. Elevated levels of crosslinking were observed the 150 equiv. BMPS case, as well as the HPBIA-$rCRM_{197}$ samples. (b) After peptide conjugation, the material was analyzed by ESI-TOF-MS (not shown) and SEC. High levels of M2 are present in the samples at higher peptide loading (around 48% M2 for $rCRM_{197}$ and around 57% M2 for HPBIA-CRM_{197}.

Minimal crosslinking of BMPS-activated $rCRM_{197}$ was observed at the lower equivalents (purple trace, **Figure 2.7**). Interestingly, at higher BMPS equivalents and in the HPBIA-capped samples, nearly all examined material had crosslinking. More surprisingly, at lower equivalents of BMPS, in both the uncapped and HPBIA-capped case, very little M2 population was observed by SEC (purple and dark blue trace, **Figure 2.7b**). Significant M2 population was observed in the higher peptide loading cases, as predicted (teal and light blue SEC traces). As compared to the peptide conjugate material from the natively expressed CRM_{197}, there appears to be a more drastic increase in M2 population per peptide loaded. For example, **Supplemental Figure 2.8** shows that at an average of 16 peptides per protein, the conjugate material was around 40% M2, whereas 48% M2 was observed for $rCRM_{197}$ with the same average peptide loading. What is more curious is the absence of significant M2 population in the HPBIA-capped peptide-$rCRM_{197}$ conjugate material, as HPBIA-capping of natively expressed CRM_{197} resulted in an obvious M2 population at low peptide conjugation levels. Notably, the plotted trendline should not be directly and explicitly compared, as we used it to compare % M2 of uncapped CRM_{197} versus HPBIA-capping. Regardless, further examination of varying activation levels of $rCRM_{197}$ is ongoing. It was positive to see that conformational changes were still observed in peptide-$rCRM_{197}$ with protein derived from bacterial expression.

2.2.6 Future directions

With this recombinant CRM_{197} soluble expression protocol established, site-directed mutagenesis was used to successfully create the His21Ala, His21Tyr, and double mutant Tyr46Cys and Lys445Cys CRM_{197} mutants proposed in **Figure 2.1**. Expression of these mutants is ongoing. If the mutant purification and characterization is successful, then we intend to follow the same activation-conjugation protocol for the synthesis of peptide-CRM_{197} conjugate material. These mutant CRM_{197} conjugates will be analyzed by ESI-TOF-MS and SEC to determine peptide loading and conformational distribution.

2.3 Conclusions

Many creative bacterial expression methods for recombinant CRM_{197} have been reported, some of which were assessed. From these strategies, we have established a method for soluble expression of recombinant His-tagged CRM_{197} from OrigamiB(DE3) cells. Mutants proposed for the development of well-characterized peptide-CRM_{197} conjugate materials were cloned and are currently being expressed using the same protocol. We intend to characterize the mutants, with the hopes of structurally analyzing the resulting peptide-protein conjugate material. Conformational analysis will determine how the proposed mutations manipulate the structural landscape of CRM_{197}. As structure-immunogenicity relationships exist, findings presented herein will establish further structurally understanding of this valuable carrier protein, for future conjugate vaccine design and development.

2.3.1 Acknowledgements

Special acknowledgements to the Pfizer project collaborators, for material and helpful advice: Keshab Bhattacharya, Maire H. Caparon, Anouk Dirksen, and Mark A. Massa.

2.4 Supplemental Figure

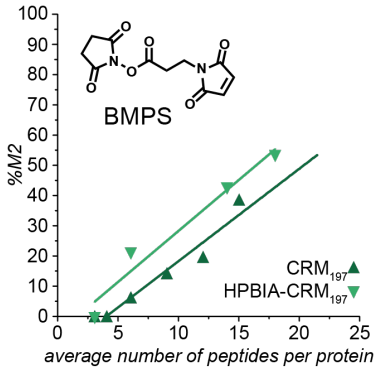


Figure 2.8: Peptide loading versus %M2 for CRM_{197} and HPBIA capped CRM_{197}. Data reproduced from [11].

2.5 Materials and Methods

2.5.1 General materials and instrumentation

Unless otherwise noted, all reagents and enzymes were obtained from commercial sources and used without any further purification. Water (dd-H_2O) used in all procedures was deionized using a NANOpure^{TM} purification system (Barnstead, USA). Centrifugations were performed with an Eppendorf 5424 R at 4 °C (Eppendorf, Hauppauge, NY). CRM_{197} from *Corynebacterium diphtheriae* was obtained from Pfizer, Inc. (St. Louis, MO). All samples of CRM_{197} were handled and stored at or below 4 °C.

Gel Analysis. For protein analysis, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was carried out on a Novex Mini-Protean apparatus using Novex precast 4-12% Bis-Tris polyacrylamide gels in MES buffer (Life Technologies, USA). Loading dye (Novex LDS Sample Buffer) and protein standard (BLUE2 protein standard, GoldBio Technologies) was purchased from commercial sources. Visualization of protein bands was accomplished by staining with Coomassie Brilliant Blue R-250 (Bio-Rad). Gel imaging was performed using Bio-Rad Gel Doc EZ molecular Imager and analyzed by quantitative analysis tool (Image Lab Version 5.2.1 build 11, Bio-Rad).

Fast Protein Liquid Chromatography (FPLC). FPLC was preformed on an AKTA Pure 25 L system, equipped with an in-line multiwavelength detector. Column used for IMAC was a HisTrap HP 5 mL (GE Life Sciences, PN 17524801), and separations were preformed using an buffered imidazole gradient (50 mM Tris HCl, 150 mM NaCl, 2 mM sodium azide, pH 7.6, with or without 500 mM imidazole). Size-exclusion chromatography was accomplished on a Superdex 75 Increase 10/300 GL (GE Life Sciences, PN 29148721), using an aqueous mobile phase (100 mM sodium phosphate, 200 mM NaCl, pH 8.0) at a flow rate of 1 mL/min.

High Performance Liquid Chromatography (HPLC). HPLC was performed on Agilent 1200 series HPLC systems (Agilent Technologies, USA) equipped with in-line diode array detector (DAD) and fluorescence detector (FLD). Size exclusion chromatography (SEC) was accomplished on a TSKgel G3000SWXL column fitted with a TSKgel SWXL guard column (Tosoh Bioscience LLC, King of Prussia, PA) using an aqueous mobile phase (100 mM sodium phosphate, 200 mM NaCl, pH 7.6) at a flow rate of 0.55 mL/min. Column integrity was confirmed by analyzing bovine serum albumin (Sigma-Aldrich, St. Louis, MO) and CRM_{197} (Pfizer, St. Louis, MO) analytical standards, and a 1,350–670,000 Da gel filtration standard mixture (Bio-Rad, Hercules, CA). To integrate partially overlapping SEC peaks accurately, multiple Gaussian fits were performed using OriginPro 9.0 (OriginLab Corp., Northampton, MA).

Mass Spectrometry. Proteins were analyzed on an Agilent 6224 Time-of-Flight (TOF)

mass spectrometer with a dual electrospray source (ESI) connected in-line with an Agilent 1200 series HPLC (Agilent Technologies, USA). Chromatography was performed using a Proswift RP-4H (Thermo Scientific, USA) column with a H_2O/MeCN gradient mobile phase containing 0.1% formic acid. Mass spectra of proteins and protein conjugates were deconvoluted with MassHunter Qualitative Analysis Suite B.05 (Agilent Technologies, USA).

Circular Dichroism (CD) Spectroscopy. CD spectra were measured using an Aviv 410 CD spectrophotometer. The CD signal from 200 nm to 300 nm was collected in a 0.1-cm path length cuvette at 25 °C. Samples contained 0.3-0.5 mg/mL of protein in 25 mM potassium phosphate pH 8.0. All CD spectra were blanked with buffer in the absence of protein.

2.5.2 Experimental procedures

Gene cloning and expression. The synthetic *crm197* gene was optimized for *E. coli* codon usage and purchased as a gBlock (IDT). The gene was cloned into pET28a, pET14b, or pBAD vector using BsaI restriction enzyme sites. Genes encoding for tags, other proteins, or interchangable restriction enzyme sites were easily incorporated using traditional Golden Gate cloning [30]. One-step site-directed mutagenesis was conducted using Quikchange II Site-Directed Mutagenesis Kit (Agilent Technologies, CA). Primers used are as follows:

CRM_{197}
pBAD/pET28a fwd: aGGTCTCaCATGGGGGCGGACGATGTG
pBAD/pET28a rev: aGGTCTCaTTTAGGATTTGATCTCAAAGAACAGGGATAACTTGCTATTTAC
pET14b fwd: aGGTCTCaCGTCCGCGGACGATGTGGTAGACTCTTCAAAATC
pET14b rev: aGGTCTCaCGCTTTAGGATTTGATCTCAAAGAACAGGGATAAC

His_6-fXa
pET28a/pBAD fwd: aGGTCTCaCATGCATCATCATCATCATCATATCGAGGGAAGGGGCGCTGATGATGTTGTTGATT

His_6-TEV
pET28a/pBAD fwd: aGGTCTCaCATGCATCATCACCATCATCACGAAAACCTGTACTTCCAGGGTGGGGCGGACGATGTGGTAGA
pET14b fwd: aGGTCTCaCGCTTTAGGATTTGATCTCAAAGAACAGGGATAAC

His_6-MBP
pET28a/pBAD fwd: aGGTCTCaCATGCACCATCACCATCACCATAAAATCGAAGAAGGTAAACTGGTAATC

Polyarginine-FLAG-CRM_{197} (overlap with MBP)
aGGTCTCa CCCCGCCCTGGAAATAAAGATTTTCCTTGTCATCGTCATCTTTAT-

AATCGTTTTCCTCGATCCCGAGTCGTTTTCCTCGATCCCGAGG

H21A
fwd: CGTAATGGAAAACTTCTCCTCTTACGCTGGAACCAAGCC
rev: CGACATAGCCTGGCTTGGTTCCAGCGTAAGAGGAGAAG

H21Y
fwd: GGAAAACTTCTCCTCTTACTATGGAACCAAGCCAGGCTATG
rev: CATAGCCTGGCTTGGTTCCATAGTAAGAGGAGAAGTTTTCC

Y46C
fwd: GGCACACAGGGCAATTGCGATGACGACTGG
rev: CCAGTCGTCATCGCAATTGCCCTGTGTGCC

K445C
fwd: CCGTCCCGCCTACTCTCCGGGTCATGCTACGCAGCC
rev: GCCATCGTGAAGAAAAGGCTGCGTAGCATGACCCGG

Resulting plasmids were transformed into XL1Blu *E. coli* cells (Invitrogen), and were plated on LB agar plates containing kanamycin (50 μg/mL). Colonies were grown up in 4 mL overnight cultures, and the resulting plasmids were purified using QuickClean II Plasmid Miniprep Kit (GenScript). The incorporation of the desired sequences was confirmed via sequencing (Sequetech, Mountain View, CA).

For expression, the desired plasmid was transformed into either BL21AI (Invitrogen), BL21* (Invitrogen), MON105 (Pfizer, Inc), Tuner(DE3) (Novagen), OrigamiB(DE3) (Novagen), SHuffle T7 Express lysY (New England Biolabs) *E. coli* cells using the heat shock method, and transformants were selected on an LB agar plate (using the appropriate antibiotic). Recombinant cells harboring correct plasmid (confirmed by sequencing again) were grown in 100 mL of 2xYT media with shaking at 200 rpm, 30 °C, with appropriate antibiotic, for 16 h before subculturing 1/100 into an appropriate expression media (either Terrific broth, Luria Broth, or M9 salt [29]). The culture was grown at 37 °C until the OD600 reached 0.6, at which point the temperature was cooled to 16 °C. Once the OD600 reached 1.0, isopropyl-β-D-thiogalactopyranoside (IPTG) was added to the culture medium at 1 mM to induce protein expression, and cultures were incubated for an additional 4-16 h (depending on the cell line).

Detergent-mediated solubilization and protein recovery. Method adapted from original by Park *et al.* [18]. Tuner(DE3) cells were harvested from a 500 mL overnight expression and resuspended in 50 mL of 50 mM Tris-HCl buffer (pH 7.6) with protease inhibitor cocktail (Promega), and then disrupted using a sonicator (two-second pulses with two-second rest at 50 W) on ice for four-minutes, three times. The insoluble inclusion bodies were separated

from the cell lysate by centrifugation (13,000 $\times g$ for 20 min at 4 °C). The pellet was carefully resuspended in 20 mL of 1% *N*-Lauroylsarcosine sodium salt (sarkosyl) in 50 mM Tris-HCl buffer and then incubated at 4 °C until most of the pellet was solubilized with gentle shaking (approximately 1-2 h) to prevent foaming. Solubilized samples were centrifuged at 13,000 $\times g$ for 20 min at 4 °C, and the supernatant was collected (or stored at 4 °C). For the purification, 5 mL of tenfold folding solution (1% Triton X-100 and 10 mM CHAPS) was carefully added to the supernatant in a drop-wise manner, followed by the addition of 5 mL of 10x His-tag column equilibrating buffer (200 mM sodium phosphate buffer, 5 M NaCl) and the final pH was adjusted to 7.6. The resulting solution was immediately loaded into a NiNTA column (HisTrap HP 5 mL, GE Life Sciences). After sample injection, the column was washed with 10 column volumes of the same equilibrating buffer, and the bound protein was eluted in a step-wise manner with the same equilibrating buffer containing 250 mM imidazole. All chromatography steps were performed at a flow rate of 1 mL/min using FPLC system in a cold room. The resulting elution fractions of recombinant CRM_{197} were pooled and concentrated using 30 kDa MWCO AMICON ultracentrifigual filters (MiliporeSigma), with 100 mM NaPhos, 200 mM NaCl, pH 7.6 buffer. A second purification step was employed by using NiNTA spin columns (Qiagen). The presence of recombinant CRM_{197} and its purity level in eluted fractions was evaluated by SDS-PAGE. The detection of target CRM_{197} was achieved by western blot with murine monoclonal anti-diphtheria toxin (1:1000; Abcam) as primary antibody and goat polyclonal anti-mouse-lgG conjugated to horseradish peroxidase (HRP) (1:2500; Abcam) as secondary antibody. Additional characterization was done by ESI-TOF-MS and SEC-HPLC.

Detergent-mediated solubilization and protein recovery. Adapted from the original method by Stefan *et al.* [17]. Briefly, MON105 cells were harvested by centrifugation at 4000 $\times g$ for 20 min and the pellets were resuspended in lysis buffer (50 mM Tris–HCl pH 8.0, 500 mM NaCl, 1% Triton X-100, 1 mM phenylmethylsulfonylfluoride). Cellular lysis was obtained via sonication (two-second pulses with two-second rests at 50 W, five times) on ice, and centrifuged for 20 min at 10,000 $\times g$. The soluble fraction was removed and pellets were resuspended in a solubilization buffer containing 50 mM Tris–HCl pH 8, 500 mM NaCl, 1% Triton X-100 and 6 M urea at 30 °C for 3 h in a shaker. Solutions were centrifuged for 20 min at 10,000 $\times g$ and supernatants containing the insoluble fraction were collected and stored at 4 °C. Solubilized samples were loaded into a HisTrap 5 mL NiNTA column (GE Life Sciences). Non-specific binding was removed by washing with 5 column volumes of the solubilization buffer, then the urea was eliminated by a steady gradient with equilibration buffer (50 mM Tris–HCl pH 8.0, 500 mM NaCl, 1% Triton X-100) at a slow flow rate of 0.5 mL/min. His-tagged CRM_{197} was eluted with a second gradient involving imidazole (0-500 mM over 10 CV). Fractions were analyzed by SDS-PAGE; samples containing recombinant CRM_{197} were pooled and concentrated using 30 kDa MWCO AMICON ultracentrifigual filters (MiliporeSigma) against 50 mM Tris–HCl pH 8.0, 150 mM NaCl. Removal of the N-terminal synthetic tag was performed in a tube reaction containing the target protein, a proper cleavage buffer (50 mM Tris–HCl pH 8.0), and various units of AcTEV protease

(Invitrogen). Reactions were incubated at 4 °C over 16 h and then purified by IMAC on FPLC, and finally analyzed by SDS-PAGE, ESI-TOF-MS, and SEC-HPLC.

Cell-free protein expression. PURExpress In Vitro Protein Synthesis (New England Biolabs, PN E6800) was purchased and used as per manufacturer's instructions. Briefly, 200 ng of template DNA prepared by miniprep was used for a 25 μL reaction. In addition to the kit components, PURExpress Disulfide Bond Enhancer (New England Biolabs, PN E6820) was added (1 μL of PURExpress Disulfide Bond Enhancer A and 1 μL of PURExpress Disulfide Bond Enhancer B per 25 μL of PURExpress reaction). The reaction proceeded for 3 h at 37 °C, and was terminated by 4 °C incubation. Unless immediately used, the reaction was stored at -20 °C. Purification was conducted using a gravity IMAC column loaded with TALON Metal Affinity resin (Clontech Labs, Takara Bio). The presence of recombinant CRM_{197} and its purity level in eluted fractions was evaluated by SDS-PAGE, and the detection of target CRM_{197} was achieved by Western blot, with murine monoclonal anti-diphtheria toxin (1:1000; Abcam) as primary antibody and goat polyclonal anti-mouse-lgG conjugated to horseradish peroxidase (HRP) (1:2500; Abcam) as secondary antibody.

His_6-MBP-CRM_{197} purification and characterization. Tuner(DE3) cells were harvested by centrifugation and then resuspended in lysis buffer (50 mM Tris-HCl buffer, 150 mM NaCl, pH 7.6) with a protease inhibitor cocktail (Promega). Cellular lysis was obtained via sonication (two-second pulses with two-second rests at 50 W, five times) on ice, and centrifuged for 20 min at 10,000$\times g$ at 4 °C. The soluble fraction was applied to a NiNTA column on the FPLC, and purified using a 0-500 mM imidazole gradient. Eluted fractions were analyzed by SDS-PAGE and combined for buffer exchange to remove imidazole, using 30 kDa MWCO AMICON ultracentrifugation spin filters. Removal of the N-terminal His-tag and MBP was performed using AcTEV protease (Invitrogen), according to manufacturer's instruction. Reactions were incubated at 4 °C over 16 h and then purified by a second round of IMAC on FPLC, using the a stepwise gradient of imidazole (0-500 mM over 4 steps). After dialysis into 50 mM Tris-HCl buffer pH 7.6, the recombinant CRM_{197} was analyzed by SDS-PAGE, ESI-TOF-MS, and SEC-HPLC.

His_6-CRM_{197} purification and characterization. OrigamiB(DE3) cells were harvested by centrifugation and then resuspended in lysis buffer (50 mM NaCl buffer, 150 mM NaCl, pH 8.0) with protease inhibitor cocktail (Promega). Cellular lysis was obtained via sonication (two-second pulses with two-second rest at 50 W, five times) on ice, and centrifuged for 20 min at 10,000$\times g$. Purification of the construct was conducted using a gravity IMAC column loaded with TALON Metal Affinity resin (Clontech Labs, Takara Bio). The presence of His-tagged CRM_{197} was evaluated by SDS-PAGE and ESI-TOF-MS. Removal of the N-terminal His-tag was performed using AcTEV protease (Invitrogen), according to manufacturer's instruction. Reactions were incubated at 4 °C over 16 h and then purified by a second round of batch IMAC via TALON resin (Clontech Labs, Takara Bio). Recombinant CRM_{197} and its purity level was evaluated by SDS-PAGE, ESI-TOF-MS, SEC, and CD spectroscopy.

General procedure for preparing $rCRM_{197}$ conjugates from BMPS-$rCRM_{197}$. Protocol described previously [11]. An Eppendorf tube was pre-chilled on ice, charged with a stock solution of $rCRM_{197}$ in storage buffer (35 μM final concentration, 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5), and diluted with ice-cold reaction buffer (DPBS, pH 8.0). To this solution was added a freshly prepared stock solution of BMPS (50 or 150 equiv. from a 100 mM stock in DMF). The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant BMPS-$rCRM_{197}$ was analyzed by ESI-TOF-MS; the sample was kept on ice until $\leq$2 min prior to injection. BMPS-$rCRM_{197}$ in reaction buffer (33 μM) was treated with stock peptide solution (1.5 mg peptide/mg protein 15 μL peptide stock solution per 100 μL of 33 μM BMPS-$rCRM_{197}$, 20 mg/mL stock concentration, in 0.6 M $NaHCO_3$ pH 9.2). The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 3 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant $rCRM_{197}$-peptide conjugate was analyzed by ESI-TOF-MS and SEC; the sample was kept on ice until $\leq$2 min prior to injection.

Procedure for the preparation of (NAC-HPBIA)-$rCRM_{197}$. An Eppendorf tube was pre-chilled on ice, charged with a stock solution of $rCRM_{197}$ in storage buffer (35 μM final concentration, 25 mM HEPES, 150 mM NaCl, 10% sucrose, pH 7.5), and diluted with ice-cold reaction buffer (DPBS, pH 8.0). To this solution was added a freshly prepared stock solution of HPBIA (200 mM in DMF, 137 equiv.). The resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. Additional HPBIA was added (137 equiv.), the solution was mixed by pipetting, and incubated for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. The resultant HPBIA-$rCRM_{197}$ (35 μM) was treated with neat *N*-acetylcysteamine (NAC) (0.15 μL per 100 μL of 33 μM HPBIA-$rCRM_{197}$) and the resulting mixture was mixed thoroughly by gentle pipetting and incubated on ice for 1.5 h. The reaction mixture was then purified through five successive rounds of centrifugal filtration in 30 kDa MWCO filters with ice-cold reaction buffer. An aliquot of the resultant (NAC-HPBIA)-$rCRM_{197}$ was analyzed by ESI-TOF-MS and SEC; the sample was kept on ice until $\leq$2 min prior to injection. General procedures for activation and conjugation were used to further elaborate (NAC-HPBIA)-$rCRM_{197}$.

2.6 References

[1] E. DeGregorio, and R. Rappuoli, "From empiricism to rational design: a personal perspective of the evolution of vaccine development", Nature Reviews Immunology **14**, 505–514 (2014).

[2] L. H. Jones, "Recent advances in the molecular design of synthetic vaccines", Nature Chemistry **7**, 952–960 (2015).

[3] P. Costantino, R. Rappuoli, and F. Berti, "The design of semi-synthetic and synthetic glycoconjugate vaccines", Expert Opinion on Drug Discovery **6**, 1045–1066 (2011).

[4] J. G. Fitzgerald, "Diphtheria toxoid as an immunizing agent", The Canadian Medical Association Journal, 524–529 (1927).

[5] G. Giannini, R. Rappuoli, and G. Ratti, "The amino-acid sequence of two non-toxic mutants of diphtheria toxin: CRM45 and CRM197", Nucleic Acids Research **12**, 4063–4069 (1984).

[6] E. Malito, B. Bursulaya, C. Chen, P. L. Surdo, M. Picchianti, E. Balducci, M. Biancucci, A. Brock, F. Berti, M. J. Bottomley, M. Nissum, P. Costantino, R. Rappuoli, and G. Spraggon, "Structural basis for lack of toxicity of the diphtheria toxin mutant CRM197", Proceedings of the National Academy of Sciences **109**, 5229–5234 (2012).

[7] A. K. Prasad, J.-h. Kim, and J. Gu, "Design and development of glycoconjugate vaccines", in *Carbohydrate-based vaccines: from concept to clinic* (American Chemical Society, Jan. 2018), pp. 75–100.

[8] M. Tontini, F. Berti, M. Romano, D. Proietti, C. Zambonelli, M. Bottomley, E. D. Gregorio, G. D. Giudice, R. Rappuoli, P. Costantino, G. Brogioni, C. Balocchi, M. Biancucci, and E. Malito, "Comparison of CRM197, diphtheria toxoid and tetanus toxoid as protein carriers for meningococcal glycoconjugate vaccines", Vaccine **31**, 4827–4833 (2013).

[9] J. Briand, S. Muller, and M. V. Regenmortel, "Synthetic peptides as antigens: pitfalls of conjugation methods", Journal of Immunological Methods **78**, 59–69 (1985).

[10] Q.-Y. Hu, F. Berti, and R. Adamo, "Towards the next generation of biomedicines by site-selective conjugation", Chemical Society Reviews **45**, 1691–1719 (2016).

[11] J. Jaffe, K. Wucherer, J. Sperry, Q. Zou, Q. Chang, M. A. Massa, K. Bhattacharya, S. Kumar, M. Caparon, D. Stead, P. Wright, A. Dirksen, and M. B. Francis, "Effects of conformational changes in peptide–CRM197 conjugate vaccines", Bioconjugate Chemistry **30**, 47–53 (2018).

[12] J.-Y. Chang, U. Ramseier, T. Hawthorne, T. O'Reilly, and J. van Oostrum, "Unique chemical reactivity of His-21 of CRM-197, a mutated diphtheria toxin", FEBS Letters **427**, 362–366 (1998).

[13] J. Peeters, T. Hazendonk, E. Beuvery, and G. Tesser, "Comparison of four bifunctional reagents for coupling peptides to proteins and the effect of the three moieties on the immunogenicity of the conjugates", Journal of Immunological Methods **120**, 133–143 (1989).

[14] L. Perry, and R. Wetzel, "Disulfide bond engineered into t4 lysozyme: stabilization of the protein toward thermal inactivation", Science **226**, 555–557 (1984).

[15] R. P. Mishra, R. S. Yadav, C. Jones, S. Nocadello, G. Minasov, L. Shuvalova, W. Anderson, and A. Goel, "Structural and immunological characterization of E. coli derived recombinant CRM197 protein used as carrier in conjugate vaccines", Bioscience Reports **38** (2018) 10.1042/bsr20180238.

[16] U. Rinas, E. Garcia-Fruitós, J. L. Corchero, E. Vázquez, J. Seras-Franzoso, and A. Villaverde, "Bacterial inclusion bodies: discovering their better half", Trends in Biochemical Sciences **42**, 726–737 (2017).

[17] A. Stefan, M. Conti, D. Rubboli, L. Ravagli, E. Presta, and A. Hochkoeppler, "Overexpression and purification of the recombinant diphtheria toxin variant CRM197 in Escherichia coli", Journal of Biotechnology **156**, 245–252 (2011).

[18] A.-R. Park, S.-W. Jang, J.-S. Kim, Y.-G. Park, B.-S. Koo, and H.-C. Lee, "Efficient recovery of recombinant CRM197 expressed as inclusion bodies in E. coli", PLOS ONE **13**, edited by P. L. Ho, e0201060 (2018).

[19] Y. Maeda, H. Koga, H. Yamada, T. Ueda, and T. Imoto, "Effective renaturation of reduced lysozyme by gentle removal of urea", "Protein Engineering, Design and Selection" **8**, 201–205 (1995).

[20] F. Katzen, G. Chang, and W. Kudlicki, "The past, present and future of cell-free protein synthesis", Trends in Biotechnology **23**, 150–156 (2005).

[21] A. R. Goerke, and J. R. Swartz, "Development of cell-free protein synthesis platforms for disulfide bonded proteins", Biotechnology and Bioengineering **99**, 351–367 (2007).

[22] R. W. Martin, B. J. D. Soye, Y.-C. Kwon, J. Kay, R. G. Davis, P. M. Thomas, N. I. Majewska, C. X. Chen, R. D. Marcum, M. G. Weiss, A. E. Stoddart, M. Amiram, A. K. R. Charna, J. R. Patel, F. J. Isaacs, N. L. Kelleher, S. H. Hong, and M. C. Jewett, "Cell-free protein synthesis from genomically recoded bacteria enables multisite incorporation of noncanonical amino acids", Nature Communications **9** (2018) 10.1038/s41467-018-03469-5.

[23] D. Esposito, and D. K. Chatterjee, "Enhancement of soluble protein expression through the use of fusion tags", Current Opinion in Biotechnology **17**, 353–358 (2006).

[24] J. G. Marblestone, "Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO", Protein Science **15**, 182–189 (2006).

[25] P. Mahamad, C. Boonchird, and W. Panbangred, "High level accumulation of soluble diphtheria toxin mutant (CRM197) with co-expression of chaperones in recombinant Escherichia coli", Applied Microbiology and Biotechnology **100**, 6319–6330 (2016).

[26] R. B. Kapust, and D. S. Waugh, "Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused", Protein Science **8**, 1668–1674 (1999).

[27] E. R. LaVallie, E. A. DiBlasio, S. Kovacic, K. L. Grant, P. F. Schendel, and J. M. McCoy, "A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm", Nature Biotechnology **11**, 187–193 (1993).

[28] P. Goffin, M. Dewerchin, P. D. Rop, N. Blais, and P. Dehottay, "High-yield production of recombinant CRM197, a non-toxic mutant of diphtheria toxin, in the periplasm of Escherichia coli", Biotechnology Journal **12**, 1700168 (2017).

[29] P. O. Olins, and S. H. Rangwala, "Vector for enhanced translation of foreign genes in Escherichia coli", in *Methods in enzymology* (Elsevier, 1990), pp. 115–119.

[30] C. Engler, R. Kandzia, and S. Marillonnet, "A one pot, one step, precision cloning method with high throughput capability", PLoS ONE **3**, edited by H. A. El-Shemy, e3647 (2008).

Chapter 3

Protein N-terminal Modification Using 2-Pyridinecarboxaldehyde

ABSTRACT: Protein modification is a useful technique for the development of hybrid materials that can capitalize on the properties of individual components. We recently reported a single-step N-terminal modification with 2-pyridinecarboxaldehyde (2PCA), which proceeds under physiological conditions. Certain N-terminal residues on model peptides and proteins were found to have different reactivity and stability of 2PCA modification. We are characterizing the reaction mechanism in order to understand this relationship, and the key attributes promoting product formation. A few N-terminal protein sequences, as well as several 2PCA derivatives, were discovered to promote and/or stabilize protein modification. Through the presented 2PCA-protein analysis, and ongoing 2PCA-peptide and computational experimentation, we have gained a greater understanding on how to harness the potential tunability of this reaction.

3.1 Introduction

The chemical modification of proteins is a valuable tool for many applications, from generating modern biotherapeutics [1, 2] to studying cellular functions [3]. This ever-expanding reaction toolkit enables the attachment of synthetic moieties to proteins, which allows for the formation of conjugate materials that can capitalize on the properties of both components [4]. Synthesis of these bioconjugate constructs requires chemoselective reactions that are able to proceed in aqueous solutions under mild pH and temperature conditions to preserve the integrity of the biomolecule. Most importantly, reactions thus developed control the site of attachment and the number of modifications, such as the alkylation of genetically introduced cysteine residues [5], the targeting of artificial amino acids with distinct reactivity [6], native chemical ligations [7], and enzymatic labeling techniques [8]. In these cases, the frequency of the targeted site on a protein surface is either known or can be controlled; however, modification still requires varying levels of protein engineering.

As a result, favorable single-site protein modification techniques target uniquely reactive sites; one such site is the N-terminal amine, due to its unique environment and pKa value [9]. N-terminal protein modification strategies offer significant advantages for bioconjugate preparation as they can be used for a wide range of protein targets produced by a variety of expression systems. Some of these strategies target specific amino acid residues at the N terminus; for example, tryptophan residues can be modified by Pictet-Spengler reactions [10], serine and threonine residues can yield reactive ketones or aldehydes after periodate oxidation [11], and cysteine residues can react with thioesters (native chemical ligation [12]), or aldehydes to form stable thiazolidines [13]. Our group has reported a site-specific transamination reaction that introduces reactive ketones or aldehydes at the N terminus [14, 15], as well as an oxidative coupling reaction between aminophenols and N-terminal proline residues [16]. Powerful as these reactions are, many of these methods place constraints on the specific N-terminal amino acid that is present.

Recently, our group demonstrated that 2-pyridinecarboxaldehyde (2PCA) can be utilized in a simple, one-step method to modify the N terminus of a broad scope of structurally and chemically varied proteins [17]. We believe this reaction to proceed through the initial reactivity of the 2PCA aldehyde with amine nucelophiles (i.e. lysine side chains or the N terminus), forming an imine intermediate. This transient intermediate is acted upon by the penultimate nitrogen of the amide backbone, cyclizing to form the stable imidazolidinone product (**Figure 3.1a**).

While this modification reaction features mild reaction conditions and excellent site-specificity, we observed that imidazolidinone formation reaction is reversible over extended time periods, after removal of excess 2PCA reagent. Shown in **Figure 3.1b**, the of Tobacco Mosaic Virus monomer protein is modified by 2PCA, but the product reverts to the unmodified product after reaction purification, as monitored by electrospray ionization time-of-flight

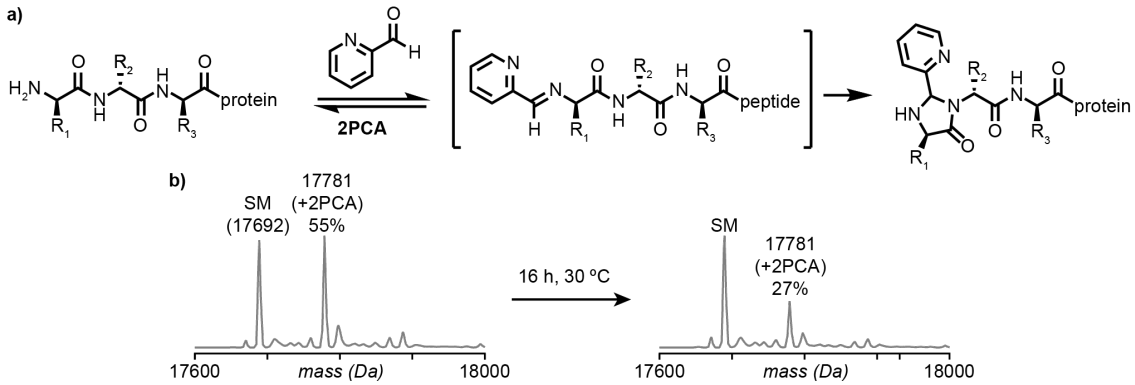


Figure 3.1: N-terminal protein modification with 2-pyridinecarboxaldehyde (2PCA). (a) 2PCA selectively reacts with the N-terminal amine of protein substrates to form a stable cyclic product. The reaction proceeds through an imine intermediate, which is reacted upon by the penultimate nitrogen of the amide backbone to form the imidazolidinone product. (b) Tobacco Mosaic Virus monomer protein (TMV; Ala-Gly-Ser N terminus; 17692 Da) is singly modified by 2PCA, to 55% conversion. Upon removal of unreacted 2PCA and incubation for 16 h at 37 °C, approximately 35% decrease in modified product (down to 27% conversion) is observed by ESI-TOF-MS. Reaction conditions: 25 μM TMV, 10 mM 2PCA, 50 mM phosphate buffer pH 7.5, 16 h, 37 °C.

mass spectrometry (ESI-TOF-MS). Thus, it is recommended that conjugates be stored in the presence of 2PCA, at $\leq$ 4 °C, or used promptly after preparation.

Though the 2PCA modification is an excellent reaction for single-site, chemoselective bioconjugation, the potential tunability of this reaction can be further investigated and utilized. Reversibility of the 2PCA modification returns the original, unmodified protein, thus making this a "traceless" modification technique that could be useful for delivery applications. In other contexts, however, such as fluorophore labeling or enzyme immobilization, non-reversible modification would be preferable. In order to control this aspect of the chemistry more thoroughly, we embarked on a combined experimental and computational study of the reaction to understand the overall energetics of imidazolidinone formation, the likely mechanisms through which the forward and reverse reactions proceed, and the affects of differing amino acid side chains on reaction performance. As a result of these ongoing studies, we have emerged with a clearer picture of how this reaction proceed and design criteria for tuning 2PCA-biomolecule bioconjugation.

3.2 Results and Discussion

3.2.1 Identification of the major product isomers

In our initial report, imidazolinone formation was postulated to introduce a new stereogenic center into the product [17]. Assuming there are energetic consequences for one product isomer (both in the forward and reverse direction), we sought to characterize the initial product isomer ratio, and to identify if one product offered better stability. To deter-

mine the inherent diastereoselectivity of the reaction, we analyzed the 2PCA modification of an Ala-Gly-Gly (AGG) tripeptide with ^{1}H NMR, after HPLC purification of the reaction product (**Figure 3.2a**). The key 2-pyridylmethine proton signals were distinctly identified between 5.5 and 5.75 ppm using COSY and HSQC (see **Supplemental Figures 3.15 & 3.16**). The zoom-in of this region indicates two peaks with approx. 3:1 ratio.

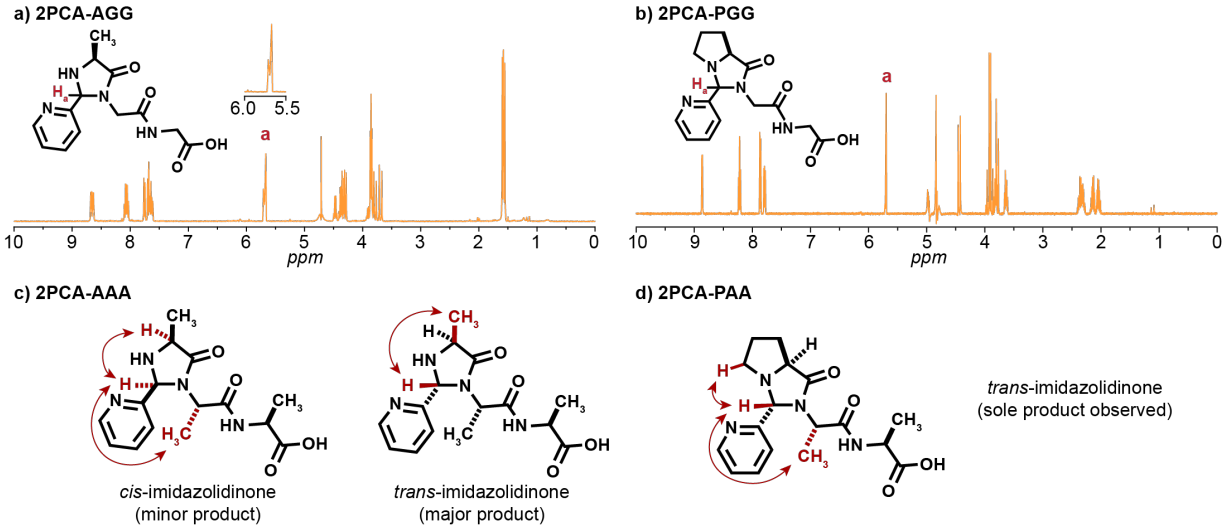


Figure 3.2: ^{1}H NMR characterization of 2PCA-peptide. (a) 2PCA modification of Ala-Gly-Gly tripeptide yields around 3:1 ratio of *trans*-to-*cis* imidazolidinone isomers. (b) In contrast, N-terminal proline tripeptides (PGG shown) yield only the *trans* imidazolidinone isomer. (c) 2PCA modified Ala-Ala-Ala tripeptide was analyzed by 2D NOESY and the major and minor product conformations were elucidated. Relevant correlations are shown in red. (d) 2PCA modified Pro-Ala-Ala tripeptide was analyzed by 2D NOESY and the single product conformation (*trans*) was determined, based on the depicted correlations. NOESY data were collected and analyzed by Nicholas Dolan.

To determine the identity of the major diastereomer, a 2D NMR experiment was particularly diagnostic. Nuclear overhauser effect spectroscopy (NOESY) is a sensitive NMR technique used to identify interactions between nuclei through-space; a NOESY spectrum will depict through-space correlations via spin-spin relaxation. Nicholas Dolan, a project co-worker in the Francis Group, was able to identify the product isomer conformations of with this method. Analysis of a 2PCA modified Ala-Ala-Ala (AAA) tripeptide indicated that the major 2-pyridylmethine proton peak correlated with a methyl group substituent of the imidazolidinone. This transannular interaction was elucidated to be the *trans* isomer. The minor product peak showed correlation of the 2-pyridylmethine proton to both an alpha proton of the tripeptide and a methyl group. From this, we determined this was the *cis* isomer; correlations arise from the syn interaction of the two imidazolidinone ring protons, as well as from an interaction with the methyl side chain of the second amino acid.

A similar set of NMR experiments was conducted for proline terminated tripeptides. A Pro-Gly-Gly (PGG) tripeptide modified by 2PCA surprisingly exhibited a single product diastereomer (**Figure 3.2b**). Analysis of a Pro-Ala-Ala tripeptide modified by 2PCA using

NOESY showed correlations between the 2-pyridylmethine signal of the imidazolidinone and two of the protons in the proline ring side chain, as well as to one of the methyl groups. This suggests that the single observed product is the *trans* isomer, where the imidazolidinone peak and the alpha proton of the ring are in an anti configuration.

With the product conformations assigned and respective energetics analyzed, we sought to elucidate the mechanistic pathway of the 2PCA modification. Density functional theory was chosen as the method for determining the reaction energetics and the identification of transition state species, using 2PCA-Ala-Ala capped with a C-terminal methyl group as a model substrate. One significant challenge for these studies was the high degree of conformational flexibility in the substrates and many of the intermediates. Molecular mechanics methods (Macromodel, OPLS3e force field) were therefore used to generate and minimize large populations of conformers to identify the lowest energy candidates. For each compound under study, the geometries of all conformers within 3 kcal/mol of the global minimum were optimized using at the B3LYP-D3/6-31G** level. At this stage, the resulting global minima were subjected to a second geometry optimization and vibrational spectrum calculation (B3LYP-D3/6-31G**) to determine the zero-point energies and the internal entropy values. Finally, refined electronic energy calculations were performed on each optimized geometry using an improved functional and an expanded basis set (ωB97M-V/6-311G-3df-3pd^{++}). These values were used to determine enthalpy values that can be compared among isomers (**Figure 3.3**). Using this approach, the overall reaction was found to be enthalpically favorable, with $\Delta H^{\circ}rxn = -11.5$ kcal/mol. The imine intermediate was calculated to be higher in energy, with $\Delta H^{\circ}imine = -3.8$ kcal/mol starting from the free peptide and 2PCA. Thus, the enthalpic advantage of cyclizing the imine species is –7.7 kcal/mol.

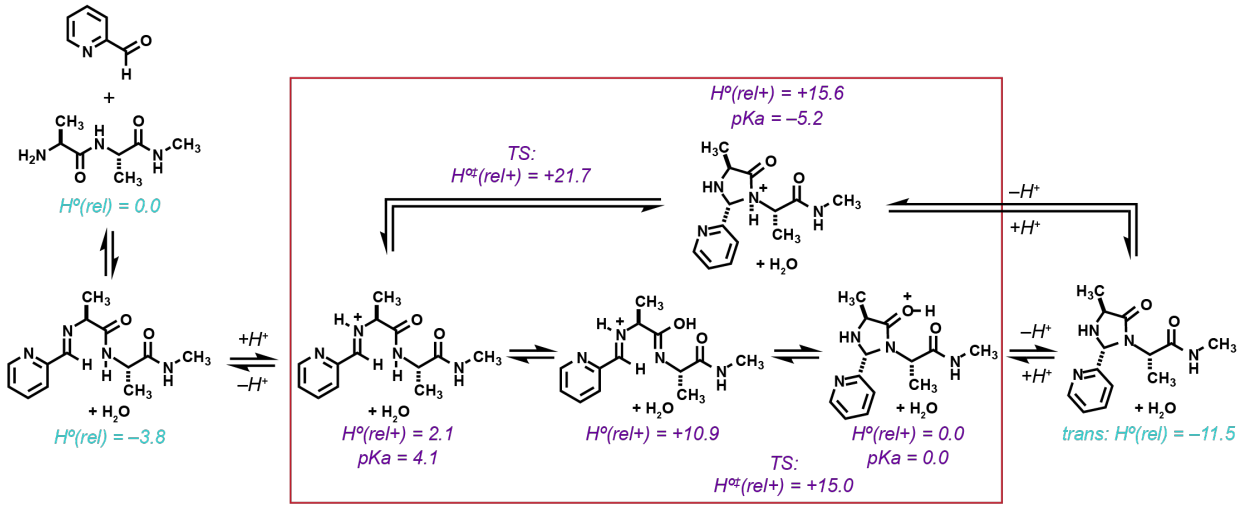


Figure 3.3: Computational analysis of the 2PCA modification reaction mechanism. Two potential routes for product formation were examined using density functional theory, boxed in red. One method evolves through the direct attack of the backbone amide nitrogen lone pair (upper route) and the other through the attack of an amidate tautomer (lower route). Entropy values are color-coded; calculations were conducted and analyzed by Prof. Matthew Francis.

Protonation of the imine species leads to a series of isomeric cationic compounds that were compared to determine the most likely reaction pathway. These species appear in the red box. It is likely that the nitrogen atom of the imine species must be protonated for the reaction to occur, both to activate the resulting iminium carbon for nucleophilic attack and to prevent the formation of a nitrogen anion as the addition occurs. What is less clear is whether the cyclization proceeds through the direct attack of the amide nitrogen lone pair on the iminium carbon (upper route, **Figure 3.3**), or through the attack of an amidate tautomer (lower route, **Figure 3.3**). Although this species would be expected to be higher in energy than the amide, it would likely be lower in energy than the transition state for the cyclization, and thus could still be a viable intermediate. One can use similar considerations to determine how amide protons exchange with the bulk solvent, either through protonation and subsequent protonation of the amide nitrogen, or through a tautomerization mechanism. Evidence for both of these pathways have been reported in the literature [18, 19], with the former pathway predominating at lower pH conditions and the tautomerization occurring preferentially at higher pH.

It was calculated that relative transition state $\Delta H^{\circ\ddagger}$ for cyclization for the amidate tautomer was lower than that of the backbone amide nitrogen (red box, **Figure 3.3**). Calculated pKa values suggest that the protonated amide nitrogen is significantly less favorable, in comparison to the protonated amide carbonyl. These ΔH° and pKa analysis, in addition to considering slightly basic reaction conditions, would suggest this tautomerization pathway is the most likely mechanism of N-terminal 2PCA modification.

From this point, we were able to examine imidazolidinone formation of 2PCA-dipeptides with alternative amino acids. Using the protonated imidazoldinone product as a standard, the relative $\Delta H^{\circ\ddagger}$ values of an energetically favorable transition state and subsequent ring-opening conformation were analyzed for three dipeptides (**Figure 3.4**). In the case of Ala-Ala and Ala-Gly, an identified transition state involves a stabilizing hydrogen bond between the nitrogen of the pyridine ring and and the hydrogen of the peptide N terminus. Interestingly, simply having a Gly in the second position related to a higher enthalpy value, indicating a more exothermic transition and potentially more stable cyclic product (**Figure 3.4b**). This was also noted in the 2PCA-Pro-Gly calculation (**Figure 3.4b**), and thus there is evidence that a second position Gly could offer some product stability. Further analysis of this phenomenon is undergoing computational analysis.

Another consideration of the 2PCA-Pro-Gly product is that the transition state $\Delta H^{\circ\ddagger}$ was quite high, indicating slower kinetics (**Figure 3.4c**). This could further indicate a more stable 2PCA modification, as the energy barrier to ring opening of a proline terminated species is more unfavorable than other N-terminal amino acids. Thus, perhaps a Pro-Gly N-terminal sequence would promote the stable 2PCA modification of protein substrates.

This transition state analysis grants us understanding of the reaction mechanism on small

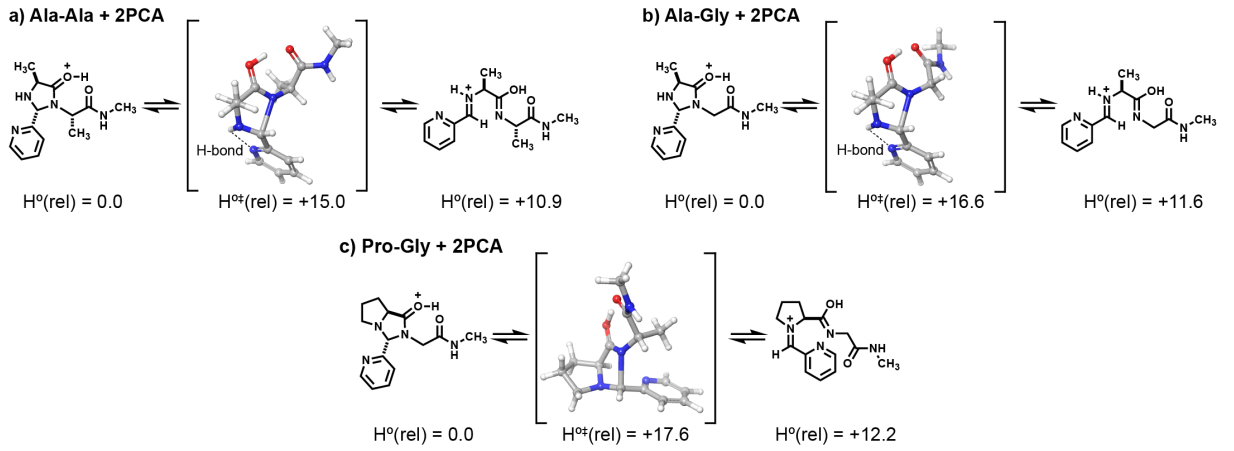


Figure 3.4: Computed 2PCA transition state conformations for various N-terminal residues. The $\Delta H^{\circ\ddagger}$ of ring opening of protonated imidazolidinone products was calculated. (a) 2PCA-Ala-Ala amidate tautomer is about 10.9 kcal/mol higher than the protonated imidazolidinone. (b) Interestingly, the calculated 2PCA-Ala-Gly ring opening is higher (11.6 kcal/mol), indicating a more exothermic conversion. The 2PCA-Pro-Gly ring opening afforded the largest ΔH° value (12.2 kcal/mol). Calculations conducted and analyzed by Prof. Matthew Francis.

peptides, and can facilitate the development of second generation 2PCA derivatives with additional transition state $\Delta H^{\circ\ddagger}$ and $\Delta G^{\circ\ddagger}$ calculations (which are ongoing). However, analysis of 2PCA modification on proteins can further develop our understanding of the relationship between N terminal identity and modification yield and stability. In particular, from the transition state analysis, we are interested in exploring the 2PCA modification of proteins with N-terminal proline residues and second position glycine residues. The remainder of the chapter will focus on protein modification analysis, with mention to ongoing computational reaction and peptide NMR analysis.

3.2.2 Analysis of 2PCA modification on variable N-terminal residues

Ubiquitin is a small (8.6 kDa), robust protein important for cellular degradation pathways. Single, site-specific modification is highly desired for its potential biophysical applications. As multiple surface available lysine residues are necessary for the cellular function of ubiquitin, modification at the N terminus is a viable strategy for non-disruptive site-specific labelling. The 2PCA modification reaction is a viable option in this case, not only because it would presumably be non-disruptive, but also because 2PCA can be easily functionalized with dyes of interest [17]. One caveat of the 2PCA reaction is that N-terminal flexibility is necessary for product formation, and unfortunately, ubiquitin is highly compact in structure, especially at the N terminus (sequence M-Q-I-F-, **Figure 3.5a**). Predictably, we were unable to modify ubiquitin with 2PCA under the published conditions. Single (and a proposed double) modification was observed only when the reaction was incubated at 58 °C; however, the integrity of the ubiquitin was probably compromised during this process (**Figure 3.5b**).

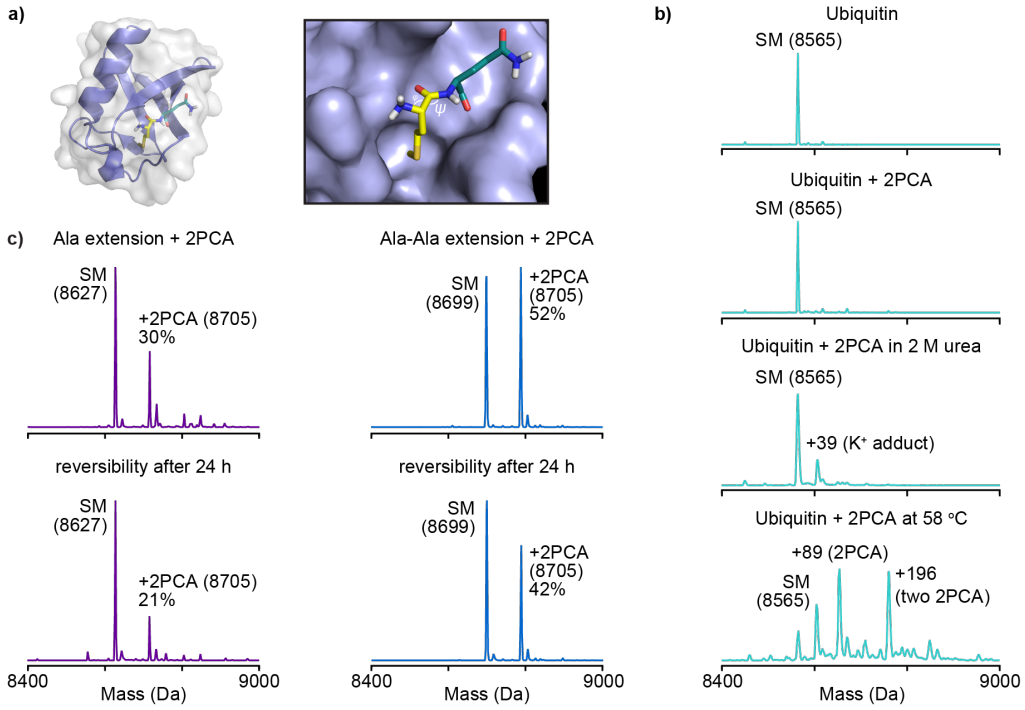


Figure 3.5: 2PCA modification of ubiquitin. (a) Ubiquitin protein (8565 Da) is compact in structure (PDB ID: 1UBQ). The N terminus of the protein is folded into the secondary structure of the protein itself, thus restricting the rotational freedom. (b) ESI-TOF-MS of 2PCA modification attempts on ubiquitin were generally unsuccessful (starting material "SM" peak is labeled). Partial denaturation of the protein was explored by the addition of 2 M urea, or incubating the reaction at 58 °C. Increased heat resulted in successful 2PCA modification (68% single addition, 48% double addition). (c) By extending the N terminus of ubiquitin by a single alanine, 2PCA modification was successful with a 30% yield as analyzed by ESI-TOF-MS. Reversibility was observed after 24 h (21% yield). Further, a double Ala-Ala extension increased 2PCA modification yield to 52%. Reversibility was observed by after 24 h (42% yield). Reaction conditions: 25 μM ubiquitin, 10 mM 2PCA, 50 mM phosphate buffer pH 7.5, 16 h, 37 °C.

Thus, genetic modification is necessary to modify ubiquitin with 2PCA. A single amino acid (alanine) N-terminal extension of ubiquitin was developed and successfully modified with 2PCA (**Figure 3.5c**). Interestingly, a two amino acid (double alanine) N-terminal extension of ubiquitin also promoted modification, but to a larger extent (**Figure 3.5c**). Reversibility of the modified bioconjugate was still observed in both cases.

Like in the 2PCA-peptide product analysis, we were interested in monitoring the 2PCA modification and reversibility on a protein. Fortuitously, ubiquitin is a great subject for protein NMR, due to its small compact size. Unfortunately, the 2-pyridylmethine signal could not be identified from a normal, non-isotopic, 2PCA modification of Ala-Ala-ubiquitin (**Figure 3.6a**). The water solvent signal around 4 ppm obscured any definative signal identification. By using an isotopic labelled 2PCA, 2D HSQC NMR could be used to correlate the 2-pyridylmethine proton. The synthesis of a ^{13}C derivative of 2PCA was proposed but not yet successfully synthesized (**Figure 3.6b**).

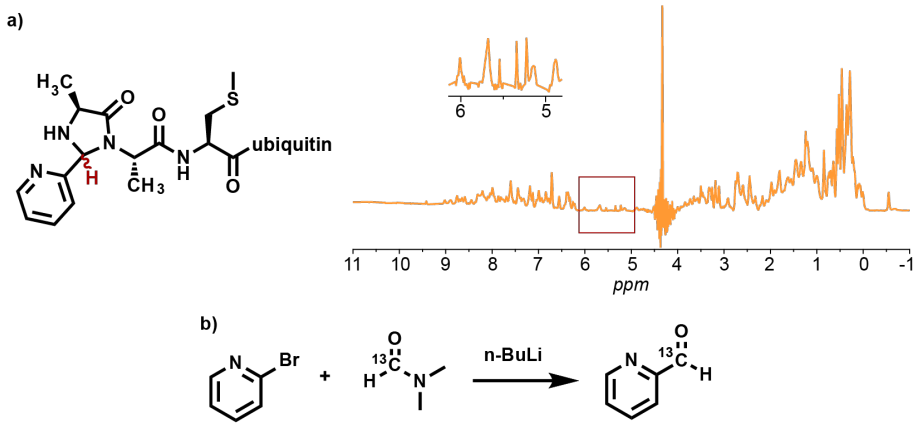


Figure 3.6: Protein NMR of 2PCA-ubiquitin mutant and proposed synthesis of ^{13}C-2PCA. (a) ^{1}H-NMR of Ala-Ala-ubiquitin was obtained. NMR was taken in 10% D_2O:buffer (50 mM KPhos pH 7.5), and the solvent peak was suppressed. (b) Proposed synthetic route to obtain ^{13}C-2PCA could proceed with halogen-lithium exchange of 6-bromo-2-pyridinecarboxaldeyde by *n*-BuLi and subsequent reaction with a ^{13}C-dimethylformamide.

Intrigued by the success of the N-terminal alanine extensions, a small panel of single and double amino acid N-terminal extensions of ubiquitin was designed. We sought to analyze how the identity of the N-terminal amino acid affects both the efficiency of the forward reaction and the stability of the imidazolidinone product. Following insight from **Figure 3.2**, proline and proline-alanine N-terminal extensions of ubiquitin were created. The forward and reverse 2PCA modification was monitored by ESI-TOF-MS. The initial modification of the single proline extension mutant protein was observed to be generally low (**Figure 3.7a**). This could be explained by the secondary nature of the N-terminal amine, which inherently reduces the degrees of freedom of proline. Predictably, extending the N terminus with an additional alanine (Pro-Ala-ubiquitin), increased the modification yield (**Figure 3.7b**). In both situations, the reversibility was decreased with a proline N terminus, as compared to the alanine N-terminal mutants (**Figure 3.5c**, see **Supplemental Figure 3.17** for compiled data comparisons).

Taking into account **Figure 3.4c**, the relative energetics of 2PCA modification of a proline N-terminal residue seem to promote a more stable modification. However, it should be noted that the proline N terminus in bacterial proteins (native or genetically modified) is not always accessible [20]. As observed in **Figure 3.7a**, N-terminal proline does not always promote the cleavage of the formyl-methionine start codon. Unfortunately from the presented data, it is not fully clear if proline terminated proteins promote a more stable 2PCA modification.

We were also interested in other N-terminal residues that have unique reactive side chains. As shown in **Figure 3.1a**, formation of the imidazolidinone product occurs via the intramolecular cyclization of the penultimate amino acid into the imine intermediate.

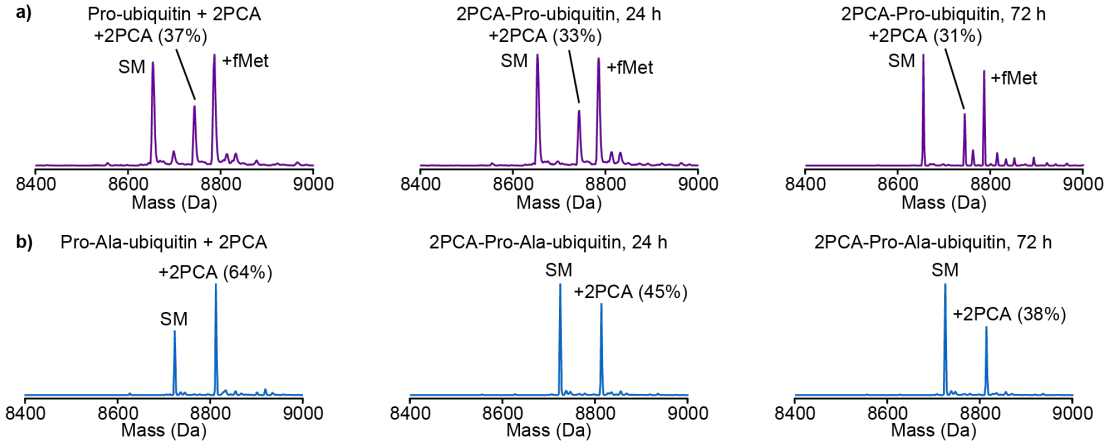


Figure 3.7: 2PCA modification of N-terminal proline ubiquitin mutants. (a) A Proline-ubiquitin mutant (8654 Da) was modified with 2PCA (8743 Da) to 37% yield. After excess 2PCA removal and 24 h incubation at 37 °C, the product was re-analyzed. A slight decrease was observed (33%). After two additional days at the same conditions, analysis showed another slight decrease modification (31%). (b) Another proline terminated mutant, Pro-Ala-ubiquitin (8725 Da), was modified with 2PCA (8814 Da) to 64% yield. After the purified product was left a 37 °C for 24 h, the product was re-analyzed, resulting in a decrease of modified protein (45%). After two additional days at the same conditions, analysis showed another decrease in modification (38%). Reaction conditions: 25 μM protein and 50 mM 2PCA in 50 mM phosphate buffer, pH 8.0 (note - slightly higher reaction buffer pH was used to promote the deprotonation of N-terminal proline for reactivity with 2PCA). Modifications were analyzed using ESI-TOF-MS.

However, some amino acids bearing beta-functional groups offer additional reaction pathways that could compete with imidazolidinone formation. One that almost certainly occurs is a Pictet-Spengler reaction with N-terminal tryptophan residues, as has been reported in previous studies [10]. The successful modification of 2PCA with Trp-Gly-Gly was confirmed by LC/MS, but we did not ascertain which reaction pathway was functioning due to the complexity of the NMR spectrum. While a tryptophan N terminus is useful method for peptide modification, it is difficult to obtain protein substrates with N-terminal tryptophan resides because the formyl-methionine starting amino acid is retained when tryptophan is present in the second position [20]. Thus, we did not further deliberate on the product identity of 2PCA modification of N-terminal tryptophan residues.

Two other N-terminal residues that have unique properties are serine and cysteine. Both amino acids have nucleophilic side chains, which can theoretically compete with the penultimate nitrogen of the amide backbone for product cyclization (**Figure 3.8**). Instead of the imidazolidinone product, the hydroxymethyl group of serine could form an oxazolidinone product, and the thiol group of cysteine could form a thiazolidinone product. The resulting product rings have different chemical properties, as compared to the imidazoldinone, and can be identified by NMR. Thus, we characterized 2PCA modification and reversibility of serine and cysteine N-terminal tripeptides, as well as protein modification with serine and cysteine N-terminal extensions of ubiquitin.

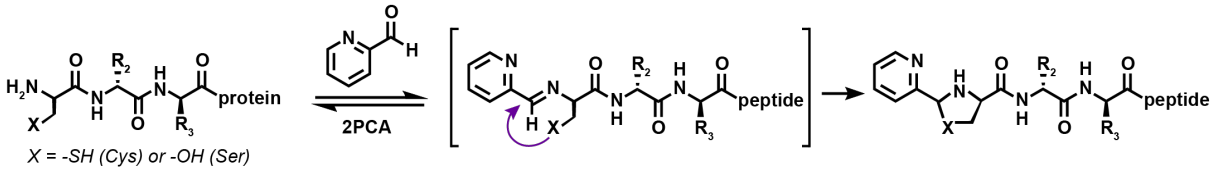


Figure 3.8: Alternative cyclic product of 2PCA modification with serine or cysteine N-terminal residues. Serine and cysteine have nucleophilic side chains, which could compete with the amide nitrogen for product formation. Instead of an imidazolidinone product, the hydroxymethyl group of serine could form an oxazolidinone product, and the thiol group of cysteine could form a thiazolidinone product.

First, we analyzed 2PCA modification of two serine N-terminal extensions of ubiquitin by ESI-TOF-MS. Interestingly, single and double modification of both mutants was observed, at low yields (**Figure 3.9**). Overall, the modification reversed more quickly in the single extension case, though modification was low initially (**Figure 3.9a**). This was not a robust modification strategy, nor a stable one. On the other hand, Ser-Ala-ubiquitin showed rapid decrease of the double 2PCA modification, but relative stability of the single modification, even over 72 h (**Figure 3.9b**).

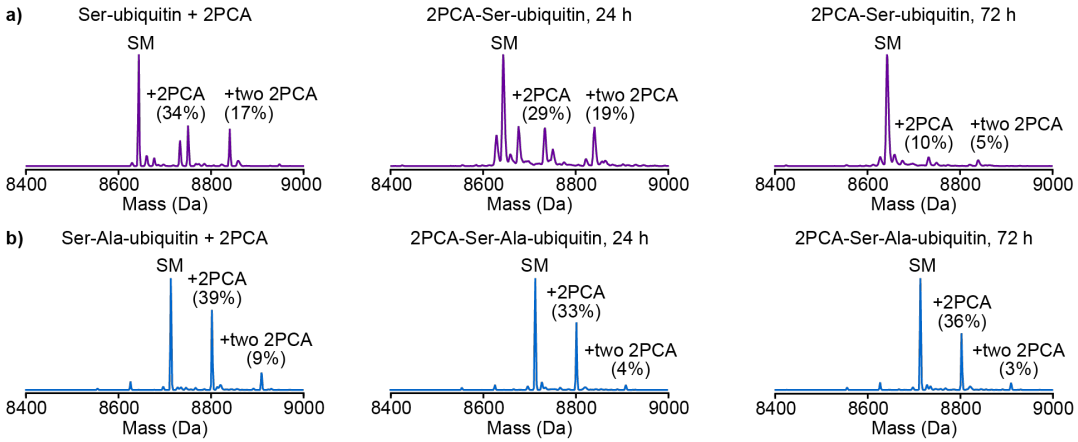


Figure 3.9: 2PCA modification of serine N-terminal ubiquitin mutants. (a) A Ser-ubiquitin mutant (8644 Da) was created and modified with 2PCA. Both single (8733 Da) and double (8840 Da) modifications were observed, at 34% and 17% yield respectively. Reversibility was monitored 24 h after excess 2PCA removal, which resulted in a decrease of singly modified product (29%) and a slight increase in doubly modified yield (19%). After 72 h modification, the reaction nearly completely reversed (10% single and 5% double modification). (b) Ser-Ala-ubiquitin mutant (8714 Da) was created and modified by 2PCA. Both the single (8804 Da) and double (8911 Da) modification were observed, at 39% and 9% yield respectively. Reversibility was monitored at 24 h after excess 2PCA removal, which resulted in a slight decrease in both populations (33% single and 4% double modification). After 72 h modification, very little doubly modified product remained (3%), but the singly modified product yield slightly increased (36%). Reaction conditions: 25 μM protein and 50 mM 2PCA in 50 mM phosphate buffer, pH 7.5. Modifications were analyzed using ESI-TOF-MS.

While the results of serine terminated ubiquitin mutants were convoluted, namely due to the double modification and low product yield, we were still interested in elucidating the structure of the cyclic product. Serine terminated tripeptides modified with 2PCA were an-

alyzed via NMR.[1] The reactions of Ser-Gly-Gly (SGG) and Ser-Pro-Gly (SPG) with 2PCA were compared, in order to confirm product identity. When proline is the penultimate amino acid (as with SPA) cyclization to form the imidazolidinone product is not possible, so any product formed must instead involve participation of the hydroxymethyl group. By ^{1}H NMR, the SGG-2PCA conjugate displayed a single set of diastereomers with 2-pyridylmethine protons, and the chemical shift was virtually identical to that obtained for the imidazolidinone with AGG. In contrast, no product peaks were observed by ^{1}H NMR when SPG was reacted with 2PCA. This suggests that imidazolidinone formation is the operating pathway when serine is the N-terminal amino acid.

In the case of a cysteine N terminus, a remarkably different modification situation was observed. We postulate that the imidazolidinone formation competes with thiazolidine formation, which is a known method for N-terminal peptide and protein modification [13]. We created single and double amino acid N-terminal extensions of ubiquitin (Cys and Cys-Ala) and subjected them to 2PCA modification (**Figure 3.10**). Not only was the near 100% modification yield outstanding, but even after incubation of the product for 72 h at 37 °C, little reversibility was observed.

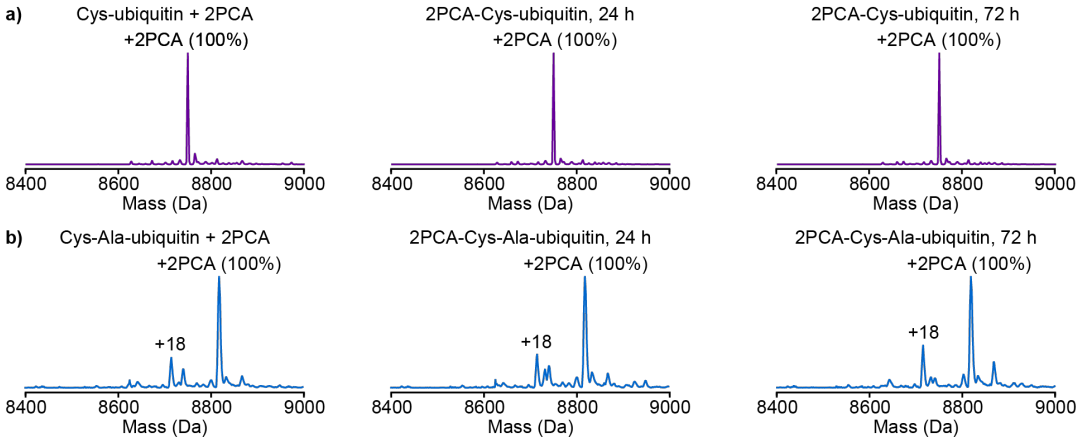


Figure 3.10: 2PCA modification of cysteine N-terminal ubiquitin mutants. (a) A Cys-ubiquitin mutant (8660 Da) was modified with 2PCA (8749 Da) to near full conversion. After removal of excess 2PCA and 72 h incubation at 37 °C, no reversibility was observed. (b) Cys-Ala-ubiquitin mutant (8730 Da) was created and modified by 2PCA (8819 Da) to essentially full conversion. Likewise, no reversibility was observed, even after 72 h. Reaction conditions: 25 μM protein and 50 mM 2PCA in 50 mM phosphate buffer, pH 7.5. Modifications were analyzed using ESI-TOF-MS.

Intrigued, we analyzed 2PCA modification of cysteine terminated tripeptides by NMR, which also showed remarkable stability.[1] Upon exposure of Cys-Gly-Gly (CGG) to 2PCA, a pair of diastereomeric signals were observed in a similar region to those arising from the imidazolidinone 2-pyridylmethine protons. To elucidate the structural identity of these chemical shifts, we designed an experiment in which the sulfhydryl group of CGG was first alkylated

[1] Experiment conducted and analyzed by Nicholas Dolan

with iodoacetic acid, forcing imidazolidinone formation if interacting with 2PCA. Contrastingly, a Cys-Pro-Gly (CPG) tripeptide, which cannot form the imidazolidinone product due to the second position proline, was modified by 2PCA to from the thiazolidine product. Both species were characterized by ^{1}H NMR and the chemical shifts were compared. It was determined that signals from 2PCA-CGG in the region of interest were similar to 2PCA-CPG, and thus due to a thiazolidine product. In an additional proof of thiazolidine product, when the sulfhydryl group of CPG was alkylated with iodoacetic acid prior to the reaction, no further products were observed upon 2PCA exposure.

Considering both the protein and the peptide modification results, these results strongly suggest that the thiazolidine product predominates when cysteine is in the N-terminal position. Furthermore, this combination promotes a stable N-terminal modification product. The formation of a thiazolidine product ring upon modification are being examined for synthetic 2PCA derivatives.

A third N-terminal sequence pattern that was analyzed stemmed from **Figure 3.4**. A second position glycine is suggested to lower the transition state energetic barrier, for the formation of the imidazolidinone product. Thus, we created three two amino acid N-terminal extensions of ubiquitin, Ala-Gly, Pro-Gly, and Ser-Gly (**Figure 3.11**). In all cases, the initial modification yield was higher than in the corresponding single extension or X-Ala cases (see **Supplemental Figure 3.17** for compiled data). Ala-Gly and Pro-Gly N-terminal extensions resulted in low reversibility (**Figure 3.10ab**), even after 72 h. With Ser-Gly-ubiquitin, initial double modification disappeared prior to the 72 h analysis; the single modification yield is reported to increase over time, relative to the double modification (**Figure 3.11c**). Just taking the single modification into account, the initial modification yield was observed at 64%, with 58% yield stabilizing after 24 h reversibility. Notably, the 2PCA modification yield was significantly higher for the Ser-Gly than the other serine N-terminal ubiquitin mutants.

Overall, this second position glycine pattern facilitates a robust and stable modification of the substrate by 2PCA. Additional comprehensive comparative supplemental data of the modification and reversibility yield each of ubiquitin N-terminal extension mutants can be found in **Supplemental Figure 3.17**.

3.2.3 N-terminal modification with second generation 2PCA derivatives

One of the most outstanding things of the 2PCA modification reaction is that, in most cases, there is no need for genetic modification of the protein substrate. However, with ubiquitin, modification by 2PCA was facilitated by amino acid extension at the N terminus due to the structurally constrained nature of the protein. As noted in the ESI-TOF-MS traces,

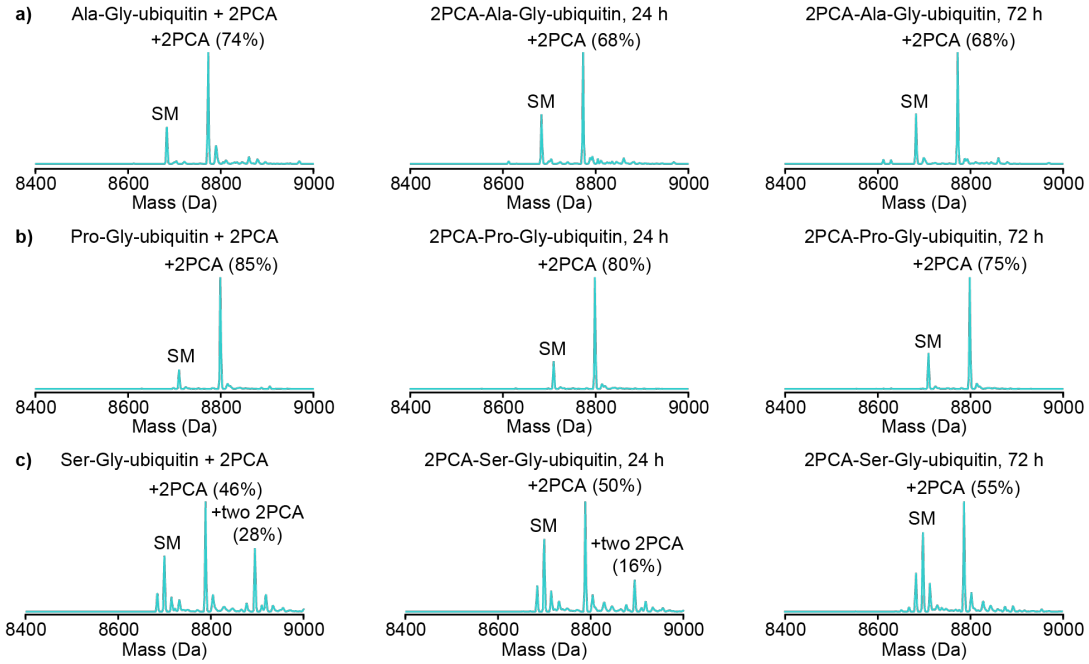


Figure 3.11: Effect of an X-Gly- N-terminal sequence on 2PCA conjugate reversibility. (a) Ala-Gly-ubiquitin mutant (8685 Da) was created and modified by 2PCA (8774 Da), to a 74% yield. Reversibility was monitored after 24 h and 72 h after excess 2PCA removal, which resulted in a slight decrease of modification yield (stable at 68%). (b) Pro-Gly-ubiquitin mutant (8711 Da) was created and modified by 2PCA (8800 Da), with a 85% yield. Reversibility was monitored 24 h after excess 2PCA removal, which resulted in a slight product decrease (80%). After 72 h, 75% modification yield was observed. (c) Ser-Gly-ubiquitin mutant (8700 Da) was created and modified by 2PCA. Both the single (8789 Da) and double (8889 Da) modification were observed, at 46% and 28% yield respectively. Reversibility was monitored at 24 h after excess 2PCA removal, which resulted in a slight increase in singly modified product yield (50%), but a significant decrease in doubly modified yield (28%). After 72 h, no doubly modified product was observed, but an increase in singly modified product was observed (55%). Reaction conditions: 25 μM protein and 50 mM 2PCA in 50 mM phosphate buffer, pH 7.5 or pH 8.0 for N-terminal proline proteins. Modifications were analyzed using ESI-TOF-MS.

simply extending the N terminus by one or two amino acids did not result in consistent modification. As we observed in 2PCA-peptide NMR, modification and reversibility rates were dependant on the amino acid identity at the N terminus. In addition to supplementing our interesting 2PCA modification correlation to N-terminal identity, a library of heteroaromatic aldehyde compounds were and are being thoroughly analyzed to understand the effect of various substituent on product yields and stability.

Previous work from the lab [21], in additional to recent and ongoing computational analysis of the 2PCA-peptide adduct, have suggested additional 2PCA derivatives that could promote tunability of this reaction. In a prior screen, we noted that the electronic properties of the heteroaromatic ring had a large effect on the overall conversion; it is postulated that an electron deficient ring most likely activates the attached aldehyde for nucleophilic attack and facilitates the cyclization step. Interestingly, it was found that compounds with electron withdrawing substituents resulted in lower modification yield as compared to compounds

with electron donating substituents, regardless of the position on the ring.

Taking Le Châtelier's principle into consideration, the relative equilibrium of transient imine intermediate to free 2PCA is a suggested driving force of cyclic product formation. However, in water, activated aldehydes are known to establish a rapid equilibrium with their hydrate form [22]. This equilibrium is important for the 2PCA reaction because a higher aldehyde to hydrate ratio in water should increase the reactivity of the reagent, as the hydrate form will not react with the N terminus. In a previous study from the lab, when substituents were incubated in deutrated water and analyzed by ^{1}H NMR, the properties of the substituent were found to affect the ratio of aldehyde to hydrate in solution [21]. Compounds with electron donating substituents were observed to have a higher ratio of aldehyde to hydrate compared to 2PCA, while electron withdrawing substituents, such as halogens, had a lower ratio of aldehyde to hydrate.

Two particular 2PCA derivatives were examined in the modification of protein substrates (**Figure 3.12**). One compound is 2PCA with an electron donating methoxy substituent, at position six of the pyridine ring. Contrasting, the other compound examined is 2PCA with an electron withdrawing chloro substituent, at position four of the pyridine ring.

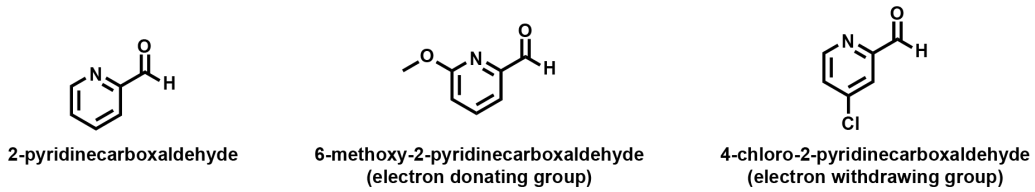


Figure 3.12: Alternative 2PCA compounds for tunable N-terminal modification

We modified various proteins with the 6-methoxy-2-pyridinecarboxaldehyde (6MeO-2PCA) and observed high conversion yield; modification of four ubiquitin mutants are reported in **Figure 3.13**. We also analyzed the reversibility of 6MeO-2PCA. We noted that the substituents may also affect the stability of the product, considering our computational analysis of the transition state, and the principle of microscopic reversibility. In all cases, we noted a steep decrease in product yield 24 and 72 h after excess 6MeO-2PCA removal (**Figure 3.13**). So, while the reactivity with the N terminus increased, the stability of the 6MeO-2PCA-protein product decreased.

The second 2PCA-derivative we examined was 4-chloro-2-pyridinecarboxaldehyde (4Cl-2PCA). As mentioned before, electron withdrawing substituents, such as halogens, promote a lower ratio of aldehyde to hydrate compared to 2PCA. Thus, we anticipated and observed the protein modification yield with 4Cl-2PCA to be lower than that of 2PCA, at the same reaction conditions. In this comparison, we analyzed 2PCA modification of a Gly-ubiquitin mutant and compared that to 4Cl-2PCA modification of the same Gly-ubiquitin (**Figure 3.14**). Modification by 4Cl-2PCA resulted in lower initial modification yield (66% versus

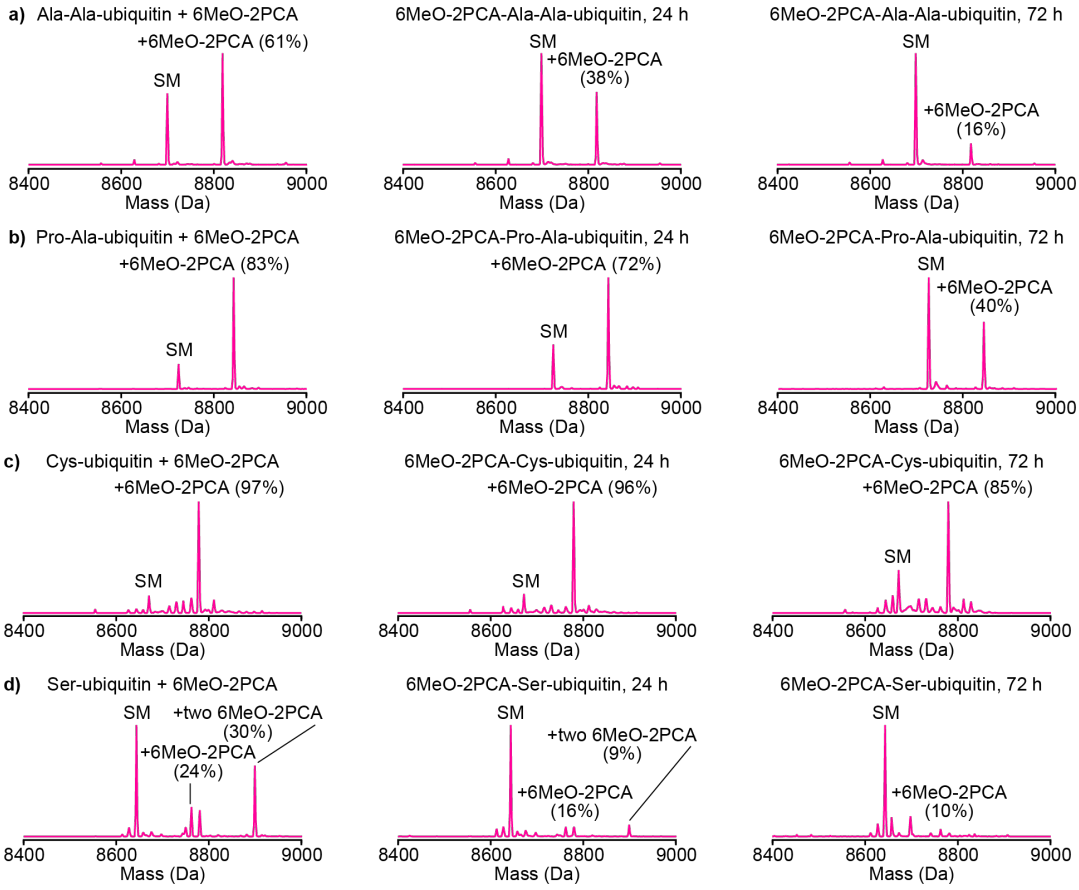


Figure 3.13: Protein modification 6-methoxy-2-pyridinecarboxaldehyde (6MeO-2PCA). The electron donating properties of the methoxy substitutent contribute to higher rates of hydrolysis of the 2PCA adduct, thus destabilizing the product.(a) The Ala-Ala-ubiquitin mutant was modified with 2PCA to 78% yield. Reversibility was monitored over 24 h (reduced to 26% product) and 72 h (reduced further to 6% product). Drastic reversibility is observed. (b) Modification with 4Cl-2PCA was resulted in slightly lower yield (66%), but remarkably improved reversibility stability was observed (36% product remaining after 24 h, and 20% product remaining after 72 h).

78%). However, we did note that the stability of the 4Cl-2PCA was higher over the 24 and 72 h analysis.

Both compounds show the potential for tunability of this N-terminal modification. One core issue of the reaction is that the aldehyde substituent of a pyridine ring is highly activated and can readily form hydrate, which is ultimately problematic for robust N-terminal modification. However, as the ubiquitin mutant modification results highlight, and with ongoing analysis, the stability of the N-terminal modification product can be easily tuned by the addition of substituents on the 2PCA ring. This offers several interesting possibilities towards the development of second generation 2PCA derivatives.

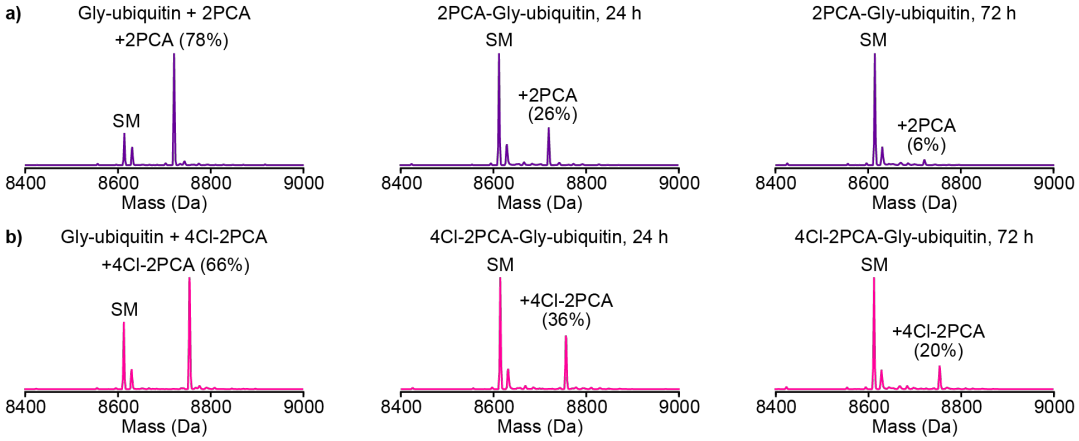


Figure 3.14: Comparison of 2PCA and 4-chloro-2-pyridinecarboxaldehyde (4Cl-2PCA) protein modification. Inductively, the chloro group meta to the aldehyde (and subsequent imine intermediate) is electron withdrawing, which could contribute to increased stability of the 2PCA adduct. (a) A Gly-ubiquitin mutant was created and modified with 2PCA to 78% yield. Reversibility was monitored over 24 h (reduced to 26% product) and 72 h (reduced further to 6% product). Drastic reversibility is observed. (b) Modification with 4Cl-2PCA was resulted in slightly lower yield (66%), but remarkably improved reversibility stability was observed (36% product remaining after 24 h, and 20% product remaining after 72 h).

3.3 Conclusions

The 2PCA N-terminal modification is a highly useful reaction that facilitates the single, site-specific chemical modification of a biomolecule of interest. As a result of these ongoing studies, we have emerged with a clearer picture of how this reaction proceeds and design criteria for tuning the modification of the synthesized conjugate material. Furthermore, as site-specific protein modification techniques are of utmost interest, especially those that are easily controlled, applications of this bioconjugation reaction have shown great potential, from drug delivery and enzyme remediation of pollutants, to novel heterobifunctional linkers [23–26]. Because chemical modification of biomolecules is an important set of techniques, especially for therapeutic applications, second generation 2PCA molecules will help enhance the ever-expanding bioconjugation field.

3.3.1 Acknowledgements

Special acknowledgements to Nicholas Dolan, Jim MacDonald, Diomedes Dieppa-Matos, Anneliese Gest, Jane Honda, and Prof. Matt Francis for their intellectual contributions and/or work included in this chapter.

3.4 Supplemental Figures

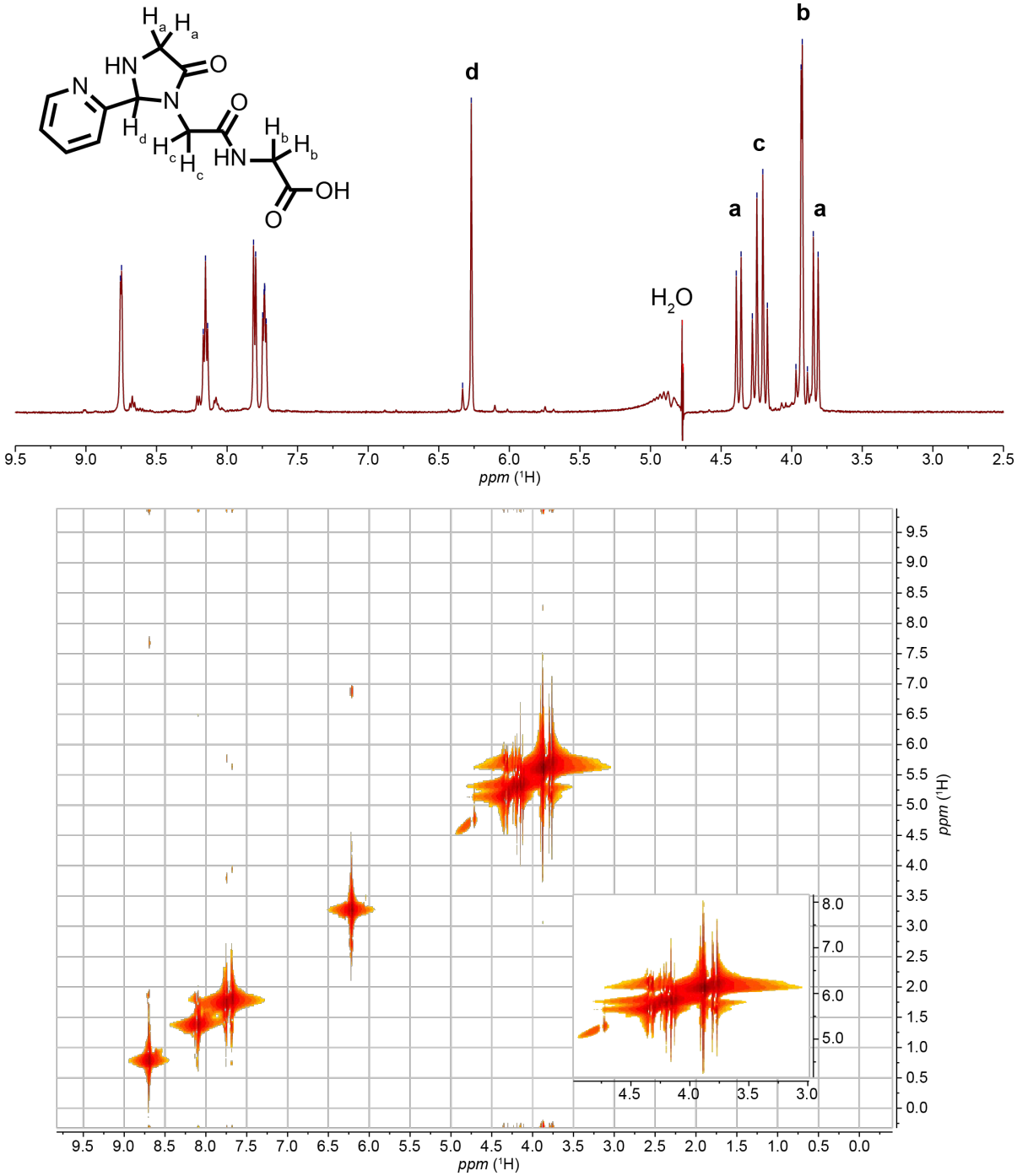


Figure 3.15: ^{1}H NMR and ^{1}H-^{1}H COSY of 2PCA-GGG. Relevant protons are assigned, according to the associated 2D NMR correlated spectroscopy (COSY). COSY is a method used to determine signals that arise from neighboring protons, through bonds. Notably, large splitting values were observed from the GGG protons, but can be distinguished using COSY.

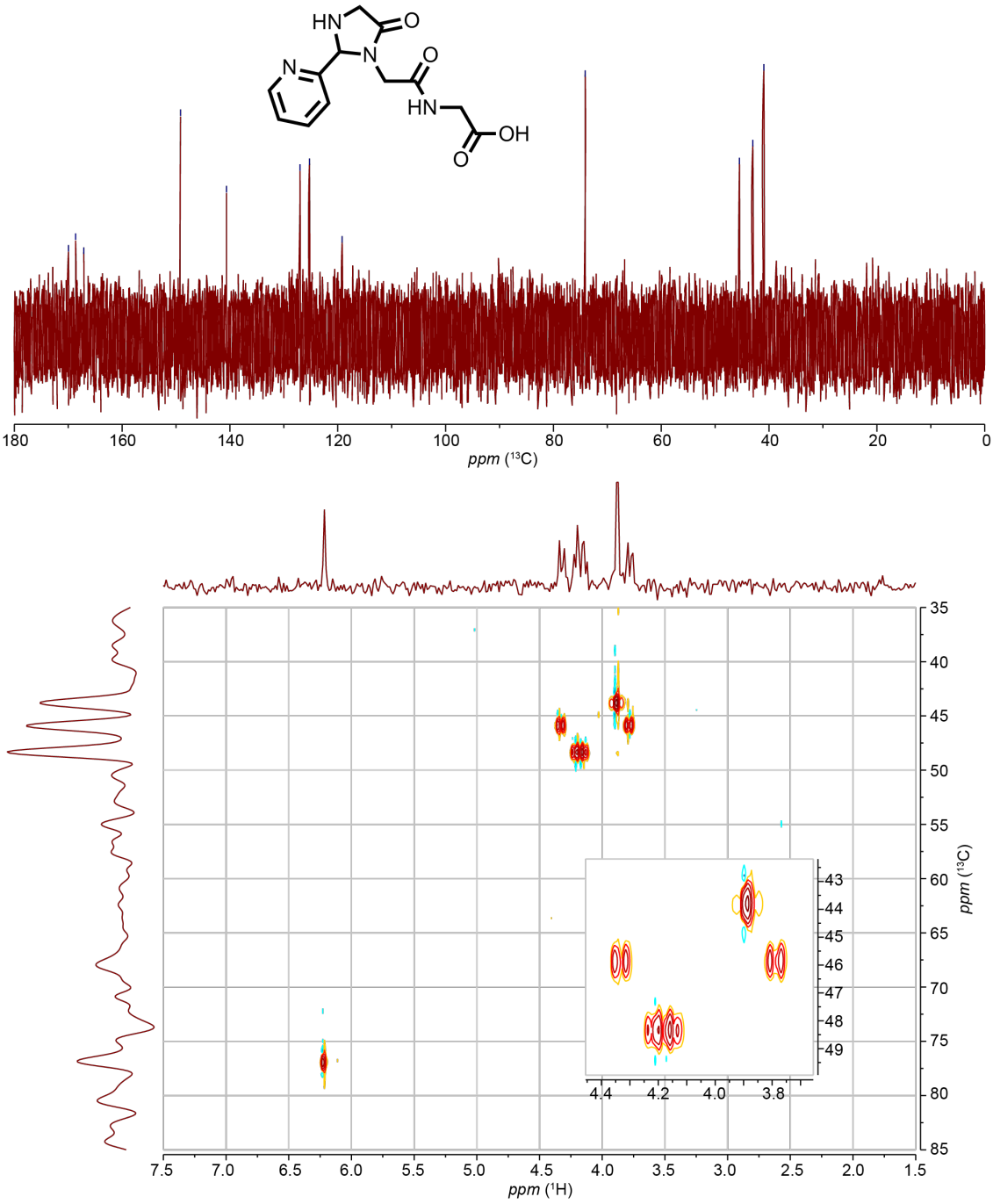
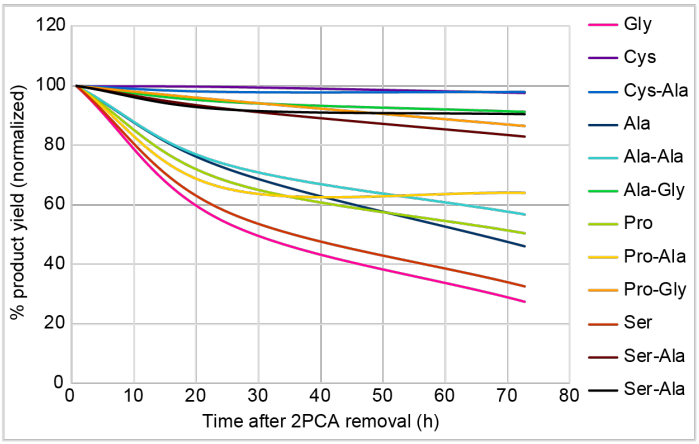


Figure 3.16: ^{13}C NMR and ^{13}C-^{1}H HSQC of 2PCA-GGG. Proton-carbon single bond correlations are determined from the ^{13}C NMR and ^{1}H NMR using ^{13}C-^{1}H Heteronuclear Single Quantum Coherence (HSQC) NMR. This experiment is used to identify the 2-pyridyl methine carbon and associated 2-pyridylmethine proton.

a)

Average % modification of ESI-TOF-MS analyzed 2PCA-ubiquitin						
N terminus	*MW (Da)*	*initial yield %*	*% after 24 h*	*% after 72 h*	*% reversibility, 24 h*	*% reversibility, 72 h*
Gly	8616	60	32.5	16.5	45.83	72.50
Cys	8660	99.25	99	97	0.25	2.27
Cys-Ala	8730	100	98	98	2.00	2.00
Ala	8630	33.5	24.25	15.5	27.61	53.73
Ala-Ala	8699	65.2	48	37	26.38	43.25
Ala-Gly	8685	75.5	71.5	69	5.30	8.61
Pro	8654	35.75	24.33	18	31.94	49.65
Pro-Ala	8725	38.5	25.25	24.67	34.42	35.92
Pro-Gly	8711	88.3	84	76.33	4.87	13.56
Ser	8644	33.67	19.5	11	42.08	67.33
Ser-Ala	8714	39.33	36.33	32.67	7.63	16.93
Ser-Gly	8700	64	59	58	7.81	9.38

b)



c)

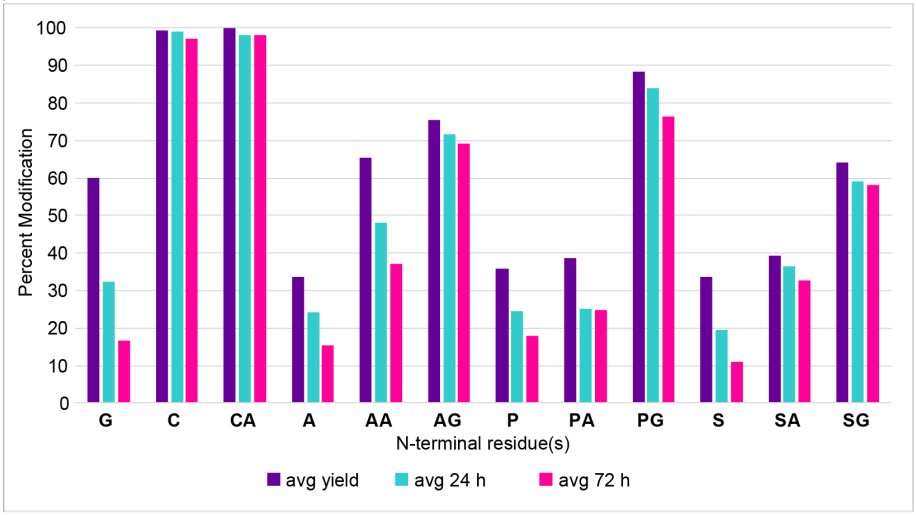


Figure 3.17: Compiled data on 2PCA modification of ubiquitin mutants. (a) Yields of 2PCA modification are tabulated, averaged from at least three different modification experiments, under the same reaction conditions. Reversibility over 24 h and 72 h are also averaged, as well as the percent decrease of the product. (b) Graph displaying the normalized percent decrease of 2PCA-protein product over 72 h, for all ubiquitin mutants. (c) Graph displaying the tabulated average 2PCA modification yield, and subsequent reversibility yield.

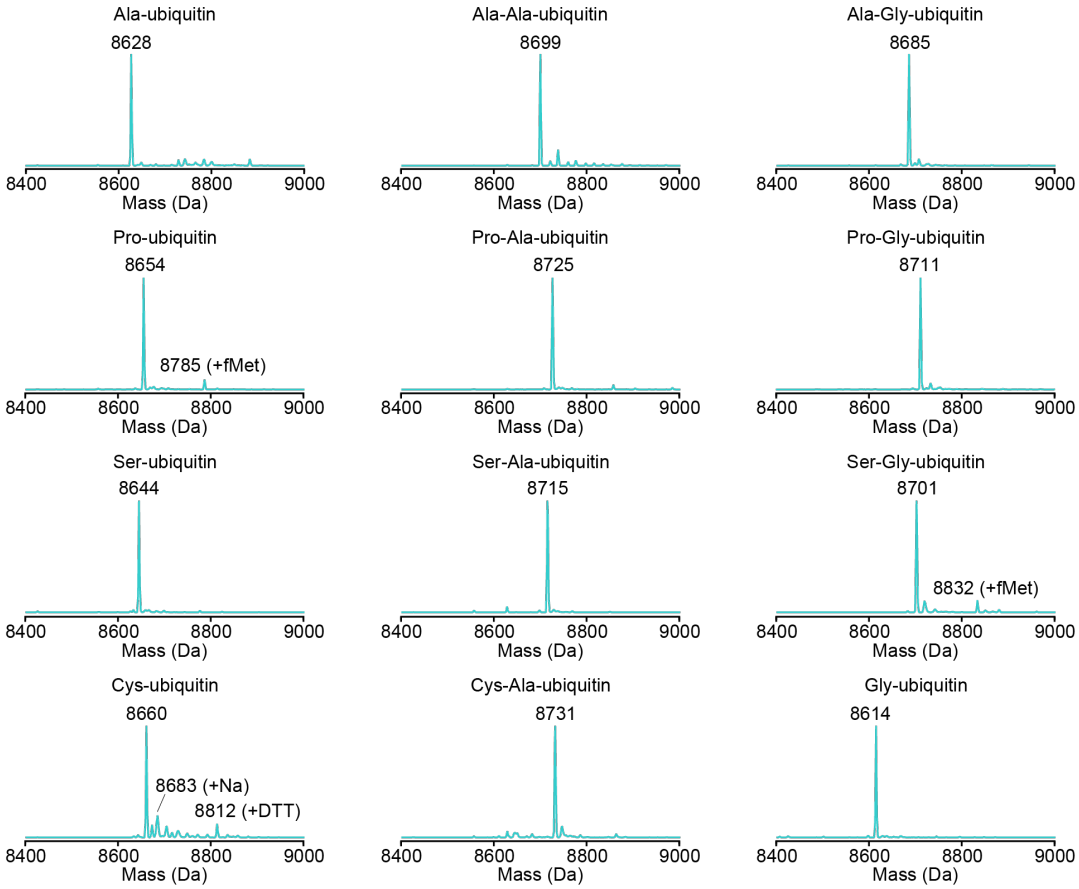


Figure 3.18: ESI-TOF-MS of ubiquitin mutants. Displayed are the N-terminal extension ubiquitin mutants that were expressed and purified for this study. Expression and purification protocol can be found in the Materials and Methods (experimental procedures) section.

3.5 Materials and Methods

3.5.1 General methods and instrumentation

Unless otherwise noted, all reagents were obtained from commercial sources and used without any further purification. Analytical thin layer chromatography (TLC) was performed on EM Reagent 0.25 mm silica gel 60-f254 plates and visualized by ultraviolet (UV) irradiation at 254 nm and/or staining with potassium permanganate. Purifications by flash silica gel chromatography were performed using EM silica gel 60 (230–400 mesh). All organic solvents were removed under reduced pressure using a rotary evaporator. Water (dd-H_2O) used in all procedures was deionized using a NANOpure^{TM} purification system (Barnstead, USA). Centrifugations were performed with an Eppendorf 5424 R at 4 °C (Eppendorf, Hauppauge, NY). Peptides were procured from GenScript (Piscataway, NJ).

Gel Electrophoresis. For protein analysis, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was carried out on a Novex Mini-Protean apparatus using Novex precast 4-12% Bis-Tris polyacrylamide gels in MES buffer (Life Technologies, USA). Loading dye (Novex LDS Sample Buffer) and protein standard (BLUE2 protein standard, GoldBio Technologies) was purchased from commercial sources. Visualization of protein bands was accomplished by staining with Coomassie Brilliant Blue R-250 (Bio-Rad). Gel imaging was performed using Bio-Rad Gel Doc EZ molecular Imager and analyzed by quantitative analysis tool (Image Lab Ver. 5.2.1 build 11, Bio-Rad).

Fast Protein Liquid Chromatography (FPLC). FPLC was preformed on an AKTA Pure 25 L system, equipped with an in-line multiwavelength detector. Column used for Ion Exchange Chromatography was a HiTrap Q HP 5 mL (GE Life Sciences, PN 17115401) at a flow rate of 2 mL/min. Purifications were preformed using an buffered (50 mM Ammonium Acetate, pH 4.5) salt gradient (50-500 mM NaCl).

Mass Spectrometry. Peptides, proteins, and protein conjugates were analyzed on an Agilent 6224 Time-of-Flight (TOF) mass spectrometer with a dual electrospray source (ESI) connected in-line with an Agilent 1200 series HPLC (Agilent Technologies, USA). Protein chromatography was performed using a Proswift RP-4H (Thermo Scientific, USA) column with a H_2O/MeCN gradient mobile phase containing 0.1% formic acid. Peptide analysis was preformed using an Acclaim 120 C18 (5 micron 100x2.1 mm, Thermo Scientific). Mass spectra of peptides, proteins, and protein conjugates were deconvoluted with MassHunter Qualitative Analysis Suite B.05 (Agilent Technologies, USA). Small molecule LC-MS analysis was performed on a C18 (1.7 micron, 150x2.5 mm, Gemini column, Phenomenex) with a gradient mobile phase containing 0.1% formic acid.

Nuclear Magnetic Resonance (NMR) Spectroscopy. ^{1}H and ^{13}C spectra were measured with a Burker AV-400 (400 MHz, 100 MHz), Burker DRX-500 (500 MHz, 150 MHz), or Burker AV-600 (600 MHz, 150 MHz) spectrometers. ^{1}H NMR chemical shifts are reported as δ in units of parts per million (ppm) relative to residual $CHCl_3$ (δ 7.26, singlet) or DMSO-d_6 (δ 2.50, pentet). Multiplicities are reported as follows: s (singlet), d (doublet), t (triplet), q (quartet), p (quintet), or br s (broad singlet). Coupling constants are reported as a J value in Hertz (Hz). The number of protons (n) for a given resonance is indicated as nH and is based on spectral integration values. ^{13}C NMR chemical shifts are reported as δ in units of parts per million (ppm) relative to $CDCl_3$ (δ 77.16, triplet) or DMSO-d_6 (δ 39.52, septet).

High Performance Liquid Chromatography (HPLC). HPLC was performed on Agilent 1200 series HPLC systems (Agilent Technologies, USA) equipped with an in-line diode array detector (DAD), fluorescence detector (FLD), and automatic fraction collector. Semi-preparative reverse-phase chromatography was achieved using a C8 stationary phase (5 micron, 250x10 mm Syncronis column, Thermo Scientific) and a H_2O/CH_3CN with 0.1% TFA gradient mobile phase at a flow rate of 3.0 mL/min.

3.5.2 Experimental procedures

2PCA-Peptide modification. To a 10 mM solution of tripeptide (10 mg, 0.05 mmol) in 50 mM phosphate buffer at pH 7.5 was added 2PCA (50 mg, 0.46 mmol). The reaction was briefly agitated to ensure proper mixing and then incubated at 37 °C without further agitation. After 16 h, the reaction was cooled to room temperature and concentrated under reduced pressure. The resulting material was purified by RP-HPLC.
Gly-Gly-Gly: MS (ESI) calculated for $C_{12}H_{14}N_4O_4$ ($[M+H]^+$) 279.10, found 279.2. ^{1}H NMR (500 MHz, D_2O) δ 8.75 (d, J = 3.9 Hz, 1H), 8.15 (t, J = 7.8 Hz, 1H), 7.81 (d, J = 7.7 Hz, 1H), 7.76 – 7.69 (m, 1H), 6.27 (s, 1H), 4.38 (d, J = 17.1 Hz, 1H), 4.31 – 4.15 (m, 2H), 3.93 (d, J = 4.3 Hz, 2H), 3.83 (d, J = 17.1 Hz, 1H). See **Supplemental Figure 3.15** for spectra. ^{13}C NMR (151 MHz, D_2O): δ 169.96, 168.64, 167.11, 149.17, 140.64, 127.00, 125.25, 119.19, 74.09, 45.48, 43.00, 40.96. See **Supplemental Figure 3.16** for spectra.
Ala-Gly-Gly: MS (ESI) calculated for $C_7H_{13}N_3O_4$ ($[M+H]^+$) 204.09, found 204.2. ^{1}H NMR (400 MHz, D_2O) δ 4.09 (q, J = 7.1 Hz, 1H), 4.04 – 3.91 (m, 2H), 3.88 (s, 2H), 1.50 (d, J = 7.1 Hz, 3H).
Pro-Gly-Gly: MS (ESI) calculated for $C_9H_{15}N_3O_4$ ($[M+H]^+$) 230.11, found 230.2. ^{1}H NMR (400 MHz, D_2O) δ 4.38 (d, J = 6.9 Hz, 2H), 4.07 – 3.91 (m, 8H), 3.36 (dt, J = 13.7, 7.0 Hz, 4H), 2.40 (d, J = 2.2 Hz, 1H), 2.09 – 1.97 (m, 8H).

Ubiquitin expression and purification. Ubiquitin, and all ubiquitin mutants, were expressed and purified as previously described [27]. Briefly, RosettaII(DE3)pLysS *E. coli* cells were transformed with a pET28a vector containing the ubiquitin gene from *S. cerevisiae* under control of a T7 promoter. Cells were grown in Terrific Broth supplemented with 1% glycerol at 37 °C until OD600 = 1.5-2.0 and were induced with 0.5 mM IPTG overnight at 16 °C. Cells were harvested by centrifugation, and pellets were frozen at -80 °C. For purification, the lysis buffer contained 50 mM Tris-HCl, pH 7.6, 0.02% NP-40, 2 mg/mL lysozyme, benzonase (Novagen), and protease inhibitors included aprotinin, pepstatin, leupeptin and PMSF. Cells were lysed by sonication (on ice) twice, followed by centrifugation to remove cellular debris. The supernatant was subjected to salting out; 60% perchloric acid was slowly added to a final concentration of 0.5%, and the solution was stirred on ice for a total of 20 min. A 5 mL HiTrap SP FF column (GE Life Sciences) was used for cation-exchange chromatography, and ubiquitin-containing fractions were pooled and exchanged into storage buffer (20 mM Tris-HCl, 150 mM NaCl, pH 7.6) by repeated dilution and concentration in Amicon Ultra 3 kDa MWCO spin concentrators (Millipore). See **Supplemental Figure 3.18** for ESI-TOF-MS of purified proteins.

Ubiquitin mutagenesis. QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies, USA) was used to extend the N terminus of ubiquitin. Mutated plasmids were transformed into XL1Blu *E. coli* cells, cultured overnight, and subsequently miniprepped to verify incorporation of the amino acid(s) was by sequencing. The following primers were used for the respective extensions:

Ala extension (M-A-M-Q-I-F)
Forward: 5'-GAAGGAGATATACCCATATGGCTATGCAGATTTTCG-3'
Reverse: 5'-GACGAAAATCTGCATAGCCATATGGGTATATCTCC-3'

Ala-Ala extension (M-A-A-M-Q-I-F)
Forward: 5'-GAAGGAGATATACCCATATGGCTGCTATGCAGATTTTCG-3'
Reverse: 5'-GACGAAAATCTGAGCAGCCATCATATGGGTATATCTCC-3'

Ala-Gly extension (M-A-G-M-Q-I-F)
Forward: 5'-CTTTAAGAAGGAGATATACCCATATGGCTGGTATGCAGATTTTCG-3'
Reverse: 5'-GTCTTGACGAAAATCTGCATACCAGCCATATGGGTATATC-3'

Pro extension (M-P-M-Q-I-F)
Forward: 5'-GATATACCCATATGCCGATGCAGATTTTCGTCAAGAC-3'
Reverse: 5'-GACGAAAATCTGCATACGCATATGGGTATATCTCC-3'

Pro-Ala extension (M-P-A-M-Q-I-F)
Forward: 5'-GGAGATATACCCATATGCCGGCTATGCAGATTTTCGTCAG-3'
Reverse: 5'-GACGAAAATCTGCATAGCCGGCATATGGGTATATCTCC-3'

Pro-Gly extension (M-P-G-M-Q-I-F)
Forward: 5'-CTTTAAGAAGGAGATATACCCATATGCCGGGTATGCAGATTTTCG-3'
Reverse: 5'-CAAAGTCTTGACGAAAATCTGCATACCCGGCATATGGGTATATC-3'

Ser extension (M-S-M-Q-I-F)
Forward: 5'-GATATACCCATATGTCTATGCAGATTTTCGTCAAGAC-3'
Reverse: 5'-GACGAAAATCTGCATAGACATATGGGTATATCTCC-3'

Ser-Ala extension (M-S-A-M-Q-I-F)
Forward: 5'-GGAGATATACCCATATGTCTGCTATGCAGATTTTCGTCAG-3'
Reverse: 5'-GACGAAAATCTGCATAGCAGACATATGGGTATATCTCC-3'

Ser-Gly extension (M-S-G-M-Q-I-F)
Forward: 5'-GAAGGAGATATACCCATATGGTCTGGATGCAGATTTTCGTCAAG-3'
Reverse: 5'-CCGGTCAAAGTCTTGACGAAAATCTGCATACCAGACATATGGG-3'

Cys extension (M-C-M-Q-I-F)
Forward: 5'-GATATACCCATATGTGTATGCAGATTTTCGTCAAGAC-3'
Reverse: 5'-GACGAAAATCTGCATACACATATGGGTATATCTCC-3'

Cys-Ala extension (M-C-A-M-Q-I-F)
Forward: 5'-GGAGATATACCCATATGTGTGCTATGCAGATTTTCGTCAG-3'

Reverse: 5'-GACGAAAATCTGCATAGCACACATATGGGTATATCTCC-3'

Gly extension (M-G-M-Q-I-F)
Forward: 5'-GAAGGAGATATACCCATATGGGTATGCAGATTTTCGTCA-3'
Reverse: 5'-GACGAAAATCTGCATACCCATATGGGTATATCTCC-3'

General method for the modification of proteins with 2PCA. Protocol adapted from MacDonald *et al.* [17]. The reaction was prepared in a 0.6 mL microcentrifuge tube. A 25 μL aliquot was taken from a 100 μM solution of protein (2.5 nmol, final concentration 50 μM) and added to 15-24.75 μL of various buffers at various pH values. To the resulting solution was added a 0.25 to 10 μL aliquot from a 100 mM solution of 2PCA in ddH_2O (25–1,000 nmol, final concentration 0.5–20 mM). The reaction was briefly agitated to ensure proper mixing and incubated at room temperature or 37 °C without further agitation. After various time points, the reaction was purified using repeated (five times) centrifugal filtration against a 0.5 mL Amicon Ultra centrifugal spin concentrator with an appropriate molecular weight cutoff (EMD Millipore, USA). Modification was monitored by ESI-TOF LC-MS.

General method for the modification of commercial and synthetic peptides with 2PCA. Method from [17]. A 2 μl aliquot was taken from a 1 mM solution of peptide (2 nmol, final concentration 100 μM) and added to 16 μl of 10 mM phosphate buffer at pH 7.5. To the resulting solution was added a 2 μL aliquot from a 100 mM solution of 2PCA (200 nmol, final concentration 10 mM). The reaction was briefly agitated to ensure proper mixing and incubated at room temperature or 37 °C without further agitation. After various time points, excess 2PCA was quenched by the addition of hydroxylamine and purified using a C18 cartridge (Waters, Sep-pak, 1 cc, 50 mg) following manufacturer's instructions. Modification was monitored by ESI-TOF LC-MS or RP-HPLC.

General method for NMR analysis of 2PCA-tripeptide modification. Tripeptide (2 mM) was mixed with 2PCA (10 mM) in phosphate buffer (50 mM) at pH 7.5. A solution of DMSO (1 μL) as internal standard and D_2O (10% v/v) were added, and the mixture was then transferred to an NMR tube. Spectra were acquired in a 500 or 600 MHz Advance series Bruker NMR spectrometer set to maintain a constant temperature. The noted reaction temperature refers to that calibrated with an ethylene glycol sample within one week of the experiment (either neat or 80% in DMSO-d_6). NMR spectra were acquired with a 90°pulse and 8 scans per timepoint at set intervals of 5-30 minutes (varied depending on the signal-to-noise ratio and half-life). Concentrations of starting materials and products were determined by integration against the internal standard.

General method for the computational analysis of 2PCA-peptide transition states. Molecular mechanics methods (Macromodel, OPLS3e force field) were used to generate and minimize large populations of conformers to identify the lowest energy candidates. For each compound under study, the geometries of all conformers within 3 kcal/mol of the global

minimum were optimized using at the B3LYP-D3/6-31G** level. At this stage, the resulting global minima were subjected to a second geometry optimization and vibrational spectrum calculation (B3LYP-D3/6-31G**) to determine the zero-point energies and the internal entropy values. Finally, refined electronic energy calculations were performed on each optimized geometry using an expanded basis set (ωB97M-V/6-311G-3df-3pd^{++}).

3.6 References

[1] S. B. Gunnoo, and A. Madder, "Bioconjugation using selective chemistry to enhance the properties of proteins and peptides as therapeutics and carriers", Organic & Biomolecular Chemistry **14**, 8002–8013 (2016).

[2] F. Li, and R. I. Mahato, "Bioconjugate therapeutics: current progress and future perspective", Molecular Pharmaceutics **14**, 1321–1324 (2017).

[3] M. Baalmann, M. J. Ziegler, P. Werther, J. Wilhelm, and R. Wombacher, "Enzymatic and site-specific ligation of minimal-size tetrazines and triazines to proteins for bioconjugation and live-cell imaging", Bioconjugate Chemistry **30**, 1405–1414 (2019).

[4] L. S. Witus, and M. B. Francis, "Using synthetically modified proteins to make new materials", Accounts of Chemical Research **44**, 774–783 (2011).

[5] J. M. Chalker, G. J. L. Bernardes, Y. A. Lin, and B. G. Davis, "Chemical modification of proteins at cysteine: opportunities in chemistry and biology", Chemistry - An Asian Journal **4**, 630–640 (2009).

[6] A. Deiters, T. A. Cropp, M. Mukherji, J. W. Chin, J. C. Anderson, and P. G. Schultz, "Adding amino acids with novel reactivity to the genetic code of saccharomyces cerevisiae", Journal of the American Chemical Society **125**, 11782–11783 (2003).

[7] T. W. Muir, "Semisynthesis of proteins by expressed protein ligation", Annual Review of Biochemistry **72**, 249–289 (2003).

[8] P. Wu, W. Shui, B. L. Carlson, N. Hu, D. Rabuka, J. Lee, and C. R. Bertozzi, "Site-specific chemical modification of recombinant proteins produced in mammalian cells by using the genetically encoded aldehyde tag", Proceedings of the National Academy of Sciences **106**, 3000–3005 (2009).

[9] J. M. Gilmore, R. A. Scheck, A. P. Esser-Kahn, N. S. Joshi, and M. B. Francis, "N-terminal protein modification through a biomimetic transamination reaction", Angewandte Chemie International Edition **45**, 5307–5311 (2006).

[10] X. Li, L. Zhang, S. E. Hall, and J. P. Tam, "A new ligation method for n-terminal tryptophan-containing peptides using the pictet–spengler reaction", Tetrahedron Letters **41**, 4069–4073 (2000).

[11] K. F. Geoghegan, and J. G. Stroh, "Site-directed conjugation of nonpeptide groups to peptides and proteins via periodate oxidation of a 2-amino alcohol: application to modification at n-terminal serine", Bioconjugate Chemistry **3**, 138–146 (1992).

[12] P. Dawson, T. Muir, I. Clark-Lewis, and S. Kent, "Synthesis of proteins by native chemical ligation", Science **266**, 776–779 (1994).

[13] J. P. Tam, Q. Yu, and Z. Miao, "Orthogonal ligation strategies for peptide and protein", Biopolymers **51**, 311–332 (1999).

[14] L. S. Witus, C. Netirojjanakul, K. S. Palla, E. M. Muehl, C.-H. Weng, A. T. Iavarone, and M. B. Francis, "Site-specific protein transamination using n-methylpyridinium-4-carboxaldehyde", Journal of the American Chemical Society **135**, 17223–17229 (2013).

[15] H. B. Dixon, and R. Fields, "Specific modification of NH2-terminal residues by transamination", in *Methods in enzymology* (Elsevier, 1972), pp. 409–419.

[16] A. C. Obermeyer, J. B. Jarman, and M. B. Francis, "N-terminal modification of proteins with o-aminophenols", Journal of the American Chemical Society **136**, 9572–9579 (2014).

[17] J. I. MacDonald, H. K. Munch, T. Moore, and M. B. Francis, "One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes", Nature Chemical Biology **11**, 326–331 (2015).

[18] C. L. Perrin, and E. R. Johnston, "Saturation-transfer study of the mechanism of proton exchange in amides", Journal of the American Chemical Society **101**, 4753–4754 (1979).

[19] C. L. Perrin, "Proton exchange in amides: surprises from simple systems", Accounts of Chemical Research **22**, 268–275 (1989).

[20] F. Frottin, A. Martinez, P. Peynot, S. Mitra, R. C. Holz, C. Giglione, and T. Meinnel, "The proteomics of n-terminal methionine cleavage", Molecular & Cellular Proteomics **5**, 2336–2349 (2006).

[21] J. I. MacDonald, "Site-specific modification of proteins with 2-pyridinecarboxaldehyde derivatives", PhD thesis (University of California, Berkeley, 2016).

[22] Y. Pocker, J. E. Meany, and B. J. Nist, "Reversible hydration of 2- and 4- pyridinecarboxaldehydes", The Journal of Physical Chemistry **71**, 4509–4513 (1967).

[23] A. M. ElSohly, J. I. MacDonald, N. B. Hentzen, I. L. Aanei, K. M. E. Muslemany, and M. B. Francis, "Ortho-methoxyphenols as convenient oxidative bioconjugation reagents with application to site-selective heterobifunctional cross-linkers", Journal of the American Chemical Society **139**, 3767–3773 (2017).

[24] J. P. Lee, E. Kassianidou, J. I. MacDonald, M. B. Francis, and S. Kumar, "N-terminal specific conjugation of extracellular matrix proteins to 2-pyridinecarboxaldehyde functionalized polyacrylamide hydrogels", Biomaterials **102**, 268–276 (2016).

[25] D. D. Brauer, E. C. Hartman, D. L. V. Bader, Z. N. Merz, D. Tullman-Ercek, and M. B. Francis, "Systematic engineering of a protein nanocage for high-yield, site-specific modification", Journal of the American Chemical Society **141**, 3875–3884 (2019).

[26] B. Koo, N. S. Dolan, K. Wucherer, H. K. Munch, and M. B. Francis, "Site-selective protein immobilization on polymeric supports through n-terminal imidazolidinone formation", Biomacromolecules **20**, 3933–3939 (2019).

[27] E. J. Worden, C. Padovani, and A. Martin, "Structure of the Rpn11–Rpn8 dimer reveals mechanisms of substrate deubiquitination during proteasomal degradation", Nature Structural & Molecular Biology **21**, 220–227 (2014).