O is for Aldehyde: Using Pyrrolysine Analogues to Introduce Reactive Carbonyls into Proteins for Bioconjugation

Robin Louis Brabham

Doctor of Philosophy

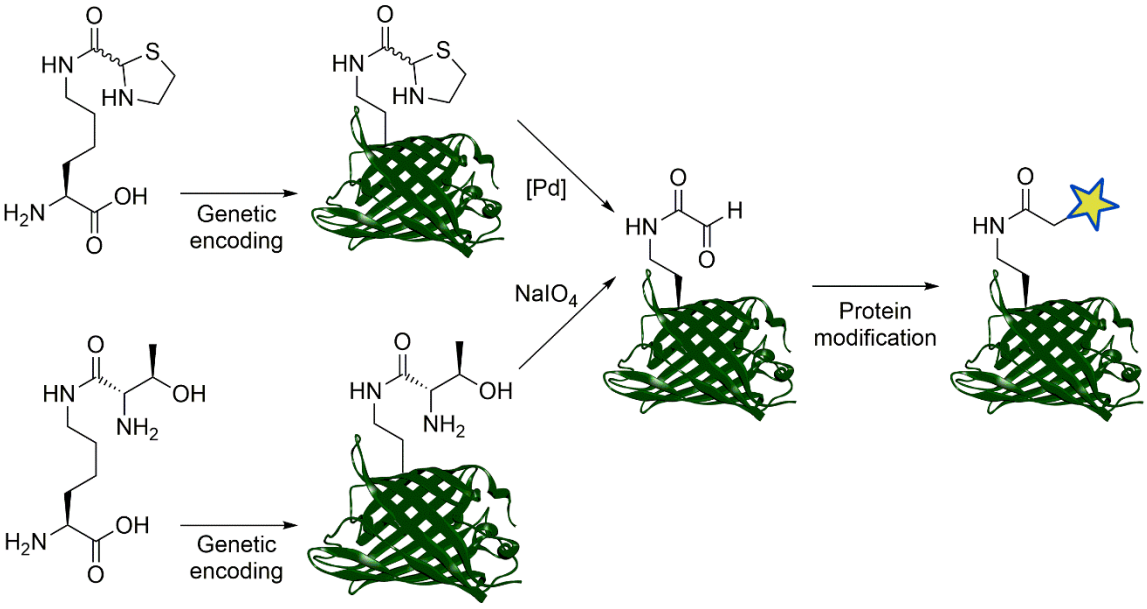
University of York

Chemistry

September 2019

Abstract

Chemical modification of proteins is a rapidly expanding area of interest with tangible effects upon society. Numerous therapeutics consisting of chemically functionalised proteins see mainstream use in clinical settings, such as proteins enhanced with polyethylene glycol or antibodies linked to anti-cancer drugs. Such modifications are no mean feats however, requiring powerful chemistry to modify the protein with the desired tag in a selective and controlled manner whilst retaining the native protein function. Some contemporary strategies achieve this through the introduction of functionality not commonly found in nature, with a key area of interest being protein aldehydes. This work reveals new methods for the installation of such aldehydes into proteins and their suitability as handles for site-specific chemical protein modification.



Amber stop codon suppression has been used to incorporate a wide range of unnatural functionality into proteins through the reassignment of the amber stop codon, adding a non-canonical amino acid into the protein exclusively at the site of the amber stop codon. Two lysine derivatives, one bearing a thiazolidine and the other a 1,2-aminoalcohol moiety, were found to be suitable for amber stop codon suppression. The thiazolidine-containing protein was treated with a palladium complex to decage a reactive glyoxyl aldehyde under mild conditions, whilst the same protein glyoxyl aldehyde was also exposed through oxidative cleavage of the 1,2-aminoalcohol-containing amino acid. The glyoxyl aldehyde was then used as a handle for site-specific chemical modification of proteins using oxime ligation, organocatalytic protein aldol ligation and strain-promoted alkyne-nitrone cycloaddition. This work enables the installation of the reactive glyoxyl aldehyde at a greater range of positions within proteins, no longer restricted to the *N*-terminus, and presents exciting new opportunities for aldehyde modification chemistry.

Table of Contents

List of Figures

List of Accompanying Material

This thesis, as an electronic copy, forms part of a package which also contains the following material:

- Raw NMR data for compounds characterised by NMR described in this thesis
- Raw ESI-FTICR-MS data for all protein species described in this thesis

Acknowledgments

During the almost four years it has taken to produce this thesis, I have been fortunate to have worked amongst a host of talented, conscientious and kind researchers in the B block lower laboratory with the Fascione, Parkin and Willems groups and I am truly grateful to every single one. Our laboratory has always worked such that we have been friends as well as colleagues, supporting each other without a second thought, and I know the groups will continue in that spirit. My thanks go to Dr Martin Fascione and Prof. Rod Hubbard for their diligent supervision and guidance on this project, ensuring that this thesis did indeed reach fruition, and to Prof. Ian Fairlamb for his warmly welcomed wise words at many points along the way.

As the final member of the Fascione Four to produce a thesis, I must pay thanks to the three other members of this quartet, having spent almost four years together and without you all I really would not have got this far. It will be a big change working without you all. Dr Emily Flack, HRH Queen of Biochemistry, you have been a kindred spirit throughout our time together in so many ways, whether in the lab throwing around “robust” DNA gels amid an atmosphere of destain or making positive life choices in the sanctuary that was Willow, and I could not have asked for someone better to go through this PhD with. Dr Harriet Chidwick, you have been a beacon of kindness, always giving out the best hugs and advice, whilst being an inspirational part of the Team Hamily Pse pioneers with the ability to conjure up litres of LB from nowhere in a flash. Dr Richard Spears, I have been proud to follow in many of your footsteps on protein modification whilst providing you with regular briefings on politics, gaming, and various aspects of the zeitgeist, and it has been a pleasure to work together on our shared interests.

As part of the “grown-ups’ office”, I am also highly thankful to Dr Tessa Keenan, who has provided me with so many fresh perspectives on how to be a better researcher in countless ways, and Dr Clare Mahon, who has been a joy to work alongside and witness her kicking the right backsides. The lovely Dr Darshita Budhadev, whilst you did forsake our office, nevertheless I cannot thank you enough for your kindness and encouragement (thankfully not as a Useless Boy) and for your immense lab knowledge, as well as our many chats on the walk home from swimming or the lab. Mark Dowsett, as you are pretty much already in the “grown-ups’ office”, it has been a pleasure to work and teach alongside you so much over the past three years, and we have made a great team together. Of course I must thank Julia Walton for everything has done in keeping the lab not just functioning but succeeding, for teaching me so much of the protein prep work with far more patience than I was due, and for being the hero who can always swoop down and make things right again. Beyond the “grown-ups’ office”, I have been lucky to

have shared the lab with many wonderful PhD students: Nick, Jenny, Tasha, Lewis, Tom, Katie, Sol, Alexandra, Chris, Sophie, Hope, and Lindsey, to whom I wish the best of success. My thanks also go to the teaching labs staff, a place in which I ended up far too frequently, for making that an enjoyable and memorable experience, and Dr Ed Bergström for all his support and brainpower with uncooperative mass spectrometers.

I am also grateful to many people outside of the lab and their efforts in helping me to retain some elements of sanity. Tom, Matt and Scott, you three have been alongside me for this time in its entirety, witnessing many highs and even more lows, and your friendship has meant a lot to me through those difficult moments. There are many additional queers, politicos, and queer politicos to whom I am also thankful for the multitude of distractions provided. As probably the worst Green possible, it has been a pleasure to have worked with so many of York's Liberal Democrats over the past few years, including the Remain campaign and finally being able to enjoy some positive election results together for once. I am particularly grateful to Erin Yarrow, fellow Old Person and dear muse, and someone who has challenged my thinking with her sharp, highly astute yet always good-natured discussions. I am also thankful to so many past and present people at the Students' Union, from my time chairing UoY Greens through to my time as a Trustee, who have given me many significant opportunities and helped me shape the university I hold dear. My friends from high school- Ste, Kate, George, Matt, Adam, Sophie, Dave, Jake, Alex, Jonathan, Phil, and others- have provided much-needed breaths of fresh air, be that at the Woodlands every Christmas Eve or whenever we get the time to meet up, and coming back to you all is always a particular joy.

Over most of the past seven years I have been fortunate enough to live with Tasha Trainor, to the extent of becoming an extended member of the Trainor family. We have seen each other at our weirdest- the regular speakers wars or fighting over the privilege of vacuuming the stairs in order to put off exam prep- and been there for each other during many difficult times. Finishing this PhD unfortunately means moving out as well and I will miss your companionship: I doubt I will find as wonderful a housemate as you.

Finally, I am deeply grateful to my family for their love and support through out this PhD, my studies and far beyond. My mother Ceri and my father Nigel have been on my side unwaveringly, never doubting my abilities and always being there as my biggest supporters. I am where I am in no small part due to you both, and my gratitude extends well beyond this acknowledgments section. That love and support has also come from Oma Jo, fortunately still with us and demonstrating a Teutonic fortitude that I hope to have acquired, as well as Opa Basil, Grandma Pat, step-grandfather Ron and aunt Andrea. Whilst they did not get to see this point of completion, they had every confidence that it would be reached, and I hope to continue with their faith in me always in mind.

Author's Declaration

I, Robin Louis Brabham, declare that this thesis is a presentation of original work and that I am the sole author. This work has not been previously presented for an award at this, or any other, University. All sources are acknowledged as References. The work submitted for this thesis is my own, with the exception of work that has formed part of jointly authored publications. The contributions made by the other authors towards this work are explicitly stated below. I can confirm that appropriate credit has been given within this thesis where reference has been made to work carried out by others and, above all, I am resoundingly grateful to those colleagues and friends listed below whose contributions have had a positive impact upon the direction of this thesis.

- Chapter 1: section 1.2 contains content from the publication "Pyrrolysine Amber Stop Codon Suppression: Development and Applications", ChemBioChem. The manuscript was prepared by R. L. Brabham and M. A. Fascione.

- Chapter 2: the concepts of the respective glycine and serine peptides **96** and **101** were initially explored to a preliminary extent by BSc student A. Hüsken, supervised by M. A. Fascione with assistance from R. L. Brabham; aminooxy peptides **123** & **124** were synthesised and purified by M. A. Fascione; phenacetaldehyde precursor **125** was synthesised by D. Budhadev; peptides **128** & **129** were oxidised to form OPAL probes **130** & **131** by R. J. Spears.

- Chapter 3: the plasmids pBAD-EGFP(Y39TAG)-His_6, pEVOL-pylT-$pylRS^{wt}$ and pEVOL-pylT-$pylRS^{AF}$ were received from E. Lemke abiding by an MTA; the plasmid pBAD-sfGFP(N150TAG)-His_6 was received from R. Mehl *via* Addgene.

- Chapter 4: section 4.2 contains content from the publication "Palladium-unleashed proteins: gentle aldehyde decgaing for site-selective protein modification", Chemical Communications. The manuscript was prepared by R. L. Brabham, R. J. Spears, J. Walton, S. Tyagi, E. A. Lemke and M. A. Fascione. Section 4.2 also forms part of patent PCT/GB2017/052896, with inventors M. A. Fascione, R. J. Spears, D. Budhadev and T. Keenan, filed through Secerna LLP. In section 4.2, R. J. Spears trialled the use of one equivalent of palladium reagent **142** for decaging, which was independently reproduced and further optimised by R. L. Brabham. In section 4.3, the stock of sfGFP(N150Threonyl-lysine)-His_6 used to trial oxidation in the absence of potassium chloride was prepared by T. Keenan. In section 4.4, the plasmid pET22b-BiGalK-His_6 plasmid was provided by T. Keenan.

- Chapter 5: J. Walton provided support with Western blots; R. Spears performed biotinylation using OPAL upon various GFPs *in vitro* and on cell lysate material using respective protein stocks or cell pellets of **136** or **137**, prepared by R. L. Brabham.
- Chapter 7: all small-molecule HRMS data were recorded by K. Heaton; all analytical HPLC traces of peptide probes were recorded by A. Dixon or S. Hicks; trypsin digests were performed by A. Dowle.

Chapter 1: Introduction

A review has been published based on section 1.2.[1]

1.1 Chemical protein modification in chemical biology: an overview

1.1.1 Moving beyond the central dogma

A paradigm shift in the understanding and philosophy of life itself came in 1957 when Francis Crick proposed the Central Dogma as a framework linking genetic information with its supposed function in the form of protein synthesis, stating “the main function of the genetic material is to control (not necessarily direct) the synthesis of proteins”. An early graphical representation (Figure 1) depicted these processes with arrows, implying that a protein was the ultimate vessel for information from which no information could be back-transferred, crucially defining information as “the sequence of amino acid residues, or other sequences related to it”.[2] The boldness of this grand conjecture was not lost upon Crick, sternly advocating for this proposition despite the recognised lack of supportive experimental data.

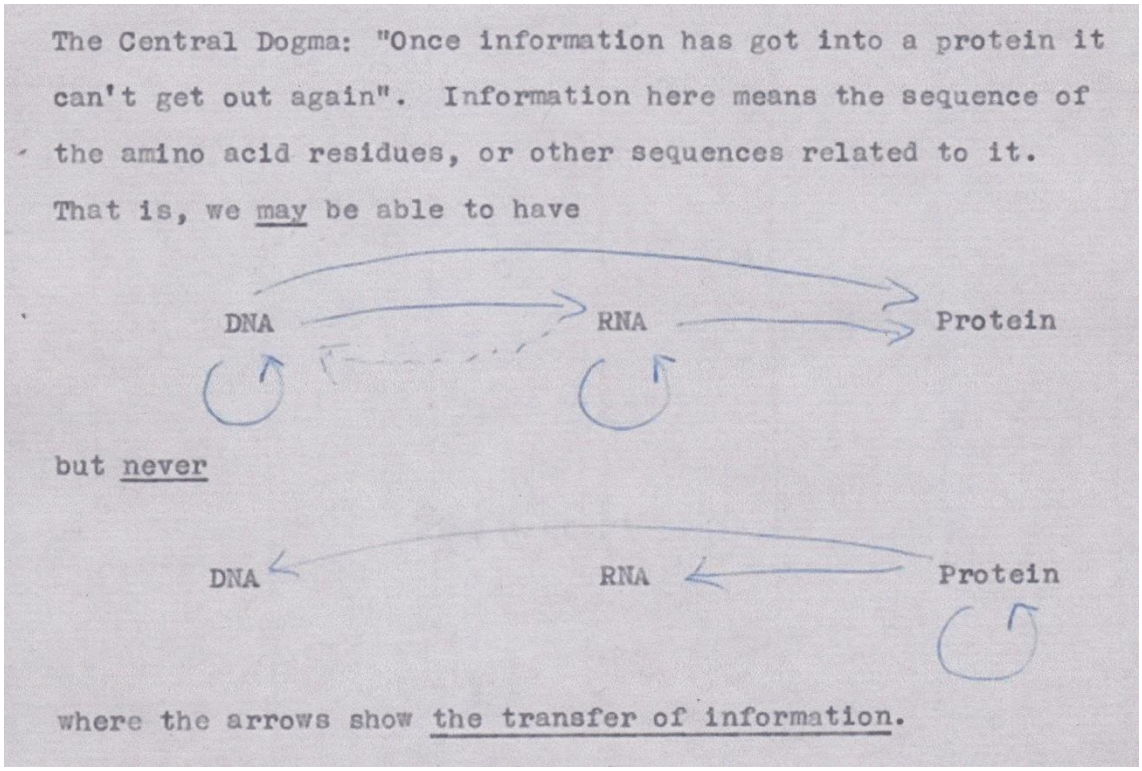


Figure 1: Crick’s original imagining of the central dogma, an unpublished note from 1956.[2]

The lack of supportive experimental data led to a slight redesign of the central dogma. Information could in fact be transferred from RNA to DNA through the use of reverse transcriptase, although this process was far less common than other forms of information transfer. Crick made a minor reconfiguration of the topology of the central dogma graph, producing the elegantly simple triangular network (Figure 2).[3] Of the nine possible arrows, three are omitted due to originating from proteins, three dashed as special cases, and three solid as the main information transfer processes.

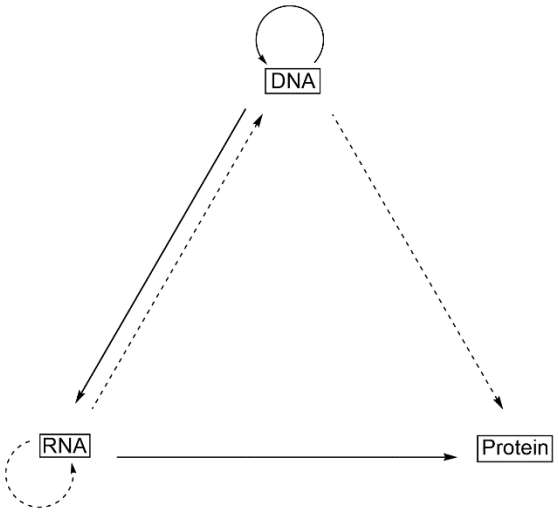


Figure 2: The familiar triangular schematic of the Central Dogma, adapted from Crick's 1970 design.[3]

Like many pieces of art, the subjectivity of the central dogma opens up a range of interpretations and misinterpretations. The common "DNA goes to RNA that then goes to proteins" statement fails to grasp the information transfer aspect of the central dogma, particularly the irreversibility implications. Indeed, the definition of information is arguably the most significant self-imposed limitation of the central dogma. Essentially restricted to primary sequence, this definition does not include the conformation of the resultant protein, the quantity expressed of the resultant protein, and the variation over time in these and other factors (with the exception of primary sequence). The fields of epigenetics and prions have together created problematic scenarios which challenge this central dogma and its underlying definition, where a protein product may be considered to change not only expression levels or conformation but also, to a debatable extent, primary sequence.[4]

A further consequence of this limiting definition is the omission of a vast network of downstream processes occurring beyond nucleic acid replication, transcription and translation. A translated protein is by no means the final product, with folding and subsequent conformational changes determining protein function, but this definition omits any structural organisation beyond the primary sequence. An additional limitation is that the information as a primary sequence contained within nucleic acids is not necessarily representative of the information as the primary sequence forming the framework of a protein. If a protein is a language, then whilst nucleic acids may be considered to contain only the basic English alphabet, nature has introduced diacritical marks such as circumflexes, çédilles, ligatures and tildes into proteins, conferring new meaning through altered functionality. Just as the tilde introduces a profound change of meaning between the words *años* and *anos*, so too can modifications of the primary sequence of a protein, sometimes more profound than the basic letters alone.

1.1.2 Side-chain post-translational modifications in nature

DNA in most organisms uses just four distinct bases organised as a triplet code, with the 64 possible triplet combinations encoding 20 standard amino acids and three terminations. As higher life forms with more complex functions and demands have evolved, nature has been compelled to keep up by developing alternative pathways to add functionality to the limited canon of 20 amino acids. Side-chain post-translational modifications (PTMs) most commonly involve nucleophilic moieties (Figure 3), where hydroxyl groups are typically modified with phosphate or glycans and amino groups often form isopeptide bonds such as acetylations and acylations with ubiquitin.[5] The redox activity of cysteine also sees use through the formation of disulphide bonds between cysteine residues or with *S*-nitrosylation.[6]

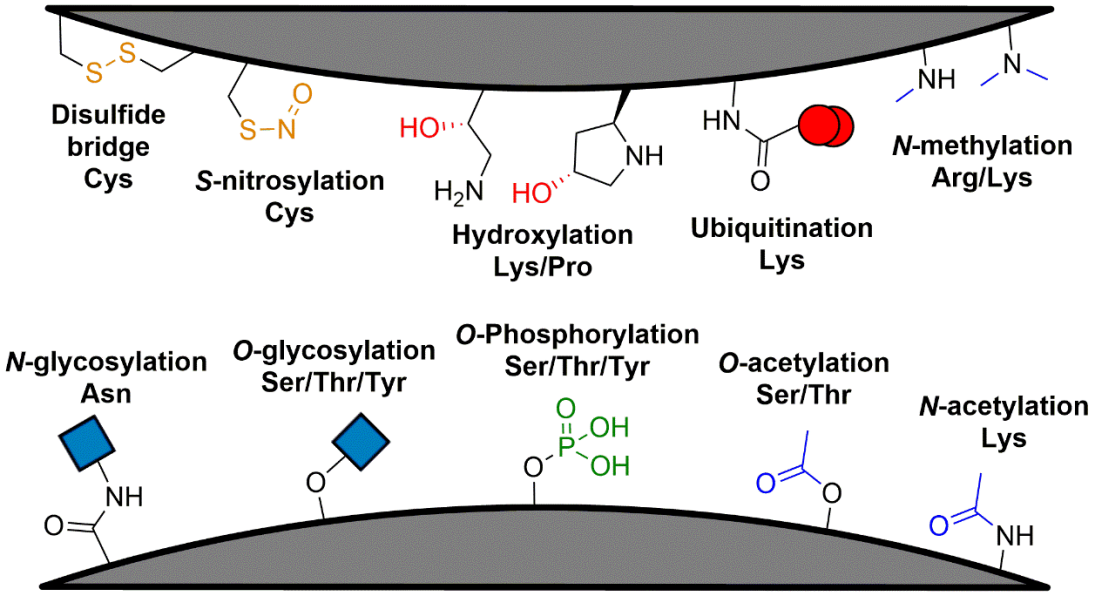


Figure 3: Common post-translational modifications of protein side chains.[5]

The crucial aspect of PTMs is the ability to transfer additional information, often reversibly, beyond that contained within the primary sequence, i.e. information not considered by the central dogma. Phosphorylation is a key example, acting as a switch to alter the protein conformation that can lead to activation or inactivation of protein function, permitting or blocking the transfer of a signal.[7] Ubiquitination is additional information itself, often acting as a signal that the labelled protein is to be degraded.[8] Glycosylation bestows some of the most intricate detail found on proteins, a secondary language beyond that of the primary sequence, such that an organism making use of glycosylation patterns may discriminate between its own matter and that of foreign and potentially harmful bodies to produce a selective defensive response.[9] Similarly, an invasive organism such as the trypanosome, well versed in nature's argot of glycosylation, can dupe a host into recognising the trypanosome as a native species and

not producing a defensive response.[10] Information itself can be considered to be an abstract selection pressure,[11] and the incorporation of post-translational modifications is one way devised by nature to continue to produce viable organisms fit for purpose in their complex, fluxional and hazardous environment, despite the apparent metabolic expense of devising such additional and elaborate pathways.

1.1.3 Early chemical protein modification: catching up with nature

Decoding the information in proteins has usually mandated destructive processes, from early Kjeldahl digestion to the paradigm-shifting Edman degradation.[12] Whilst proper use of Edman degradation could provide the primary sequence of a protein in many cases, little further could be gleaned: no information on secondary or higher structure, no information on function, and no information on the importance of various residues. In isolation, the results are a monotonous soliloquy, devoid of intonation and meaning. Whilst overall a destructive process, the initial step of Edman degradation is in fact a largely chemoselective and regioselective covalent protein modification (Figure 4). By controlling pH, only α-amino groups are sufficiently nucleophilic to attack the isothiocyanate reagent, with protonated lysine and arginine amino groups rendered largely inert. This represented an important early step in the use of chemical protein modification to understand the true significance of proteins and demonstrated that proteins can transfer information to a body not represented in the central dogma.

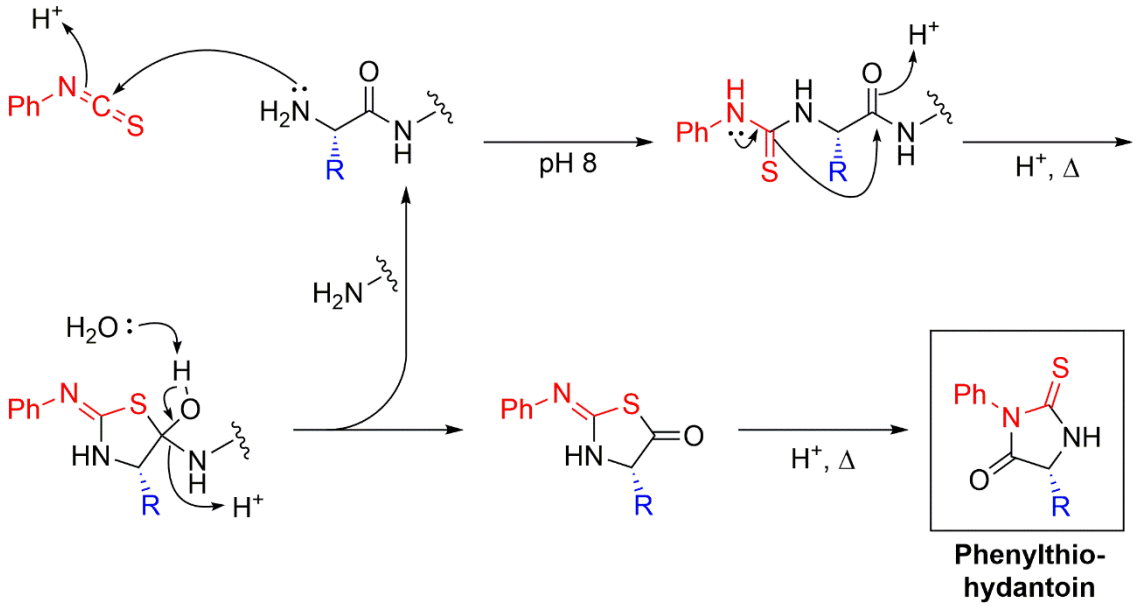


Figure 4: Nucleophilic attack of a peptide α-amino group upon phenylisothiocyanate under initially basic conditions forms an adduct which can degrade under subsequent acidic conditions, releasing the rest of the peptide, for another cycle of Edman degradation, and a phenylhydantoin after rearrangement. The phenylhydantoin bears the side-chain from the degraded residue and can be characterised by chromatography to identify the residue in question.

Whilst the sequence of amino acids is certainly important information, knowing the role of each residue provides more valuable insight into the function and context of a protein. Following on from reactions at the α-amino terminus, several reagents were found to selectively react with particular residues. If the chemical perturbation of a particular residue significantly altered protein function, then somehow this residue must be involved in protein function. Just as with post-translational modifications, nucleophilic residues proved to be suitable quarries for such chemical methods, offering good reactivity distinct from other residues to reduce off-target modification and provide more convincing results. The ε-amino group of lysine residues can be alkylated by reductive amination,[13] mimicking post-translational methylation,[14] or acylated with succinic anhydride[15] or acetic anhydride, used to imply the presence of a lysine residue in the binding site of an antibody.[16] Cysteine thiols are alkylated by iodoacetate or iodoacetamide in a reaction still enjoying ubiquitous usage in contemporary protein analysis.[17] The information conveyed in this way is still limited in utility, however. Unambiguously pinpointing which particular reactive residue participates in protein function may not be possible where multiple residues occur in a region of interest as the modifying reagent will likely modify all such residues. Furthermore, the modifying reagent may also denature the protein, such as where the modified residue may not be an active participant but may be crucial to maintaining the protein fold required for activity, presenting an additional confounding variable in this analysis. Arguably the most logically robust application of this method is to rule out the participation of one amino acid, e.g. cysteine, such that if modification with iodoacetamide does not significantly affect protein function, then the protein has not been denatured and the modified cysteine residues do not participate in protein activity. Interrogating proteins using these early modification strategies required a highly judicious selection of modification methods and the fortune of working on a well-behaved protein substrate.[15-17]

An important advance in protein modification came through the consideration of residue selectivity. The differences in pK_A values of various amino groups had already been exploited through Edman degradation; use of an analogous yet non-destructive reagent at a suitable pH could also lead to selective modification of residues, such as where nucleophilic residues are rendered largely inert due to protonation at low pH (Figure 5). The α-amino residue again proved to be a useful starting point for this venture, with alkylation using fluorodinitrobenzene used to determine the pK_A of the α-amino residue of human haemoglobin as 6.7 and finding a classic sigmoidal relationship between the rate and the pH.[18] The same reaction with bovine pancreatic RNAse A demonstrated some reactivity at the α-amino group with an estimated pK_A of 7.8, but the predominant product was instead modified at lysine-41, with the rate constant for the latter

modification an order of magnitude greater than that of the α-amino group rate constant.[19] Unusual lack of reactivity and high pK_A values of α-amino groups was also observed in porcine elastase, with a pK_A of 9.7, adding a note of caution to the reliance upon reactivity of α-amino groups.[20] This strategy was subsequently refined using radiolabelled fluorodinitrobenzene in the form of a pH-dependence assay for particular reactive residues, such as the single histidine residue in hen egg-white lysozyme and the α-amino group of a *Streptomyces* trypsin.[21]

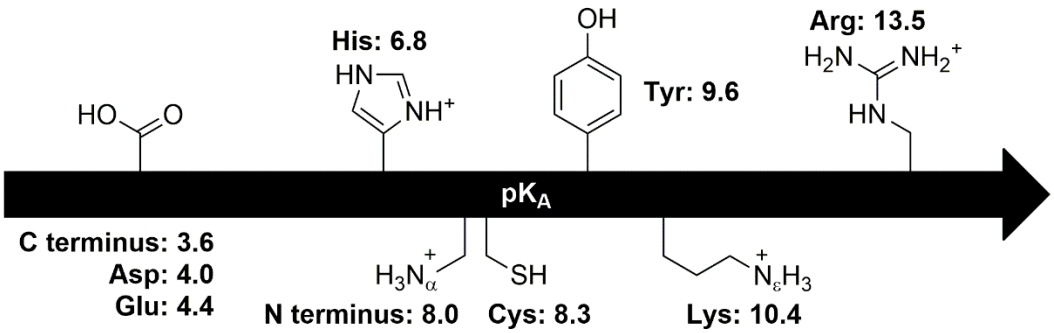


Figure 5: Literature pK_A constants of protic protein residues.[22]

Beyond the immediate kinetic and thermodynamic parameters obtained, this collective work highlighted a number of practical and philosophical considerations in the chemical protein modification field. Firstly, the variations in the proteome are manifested in the success (or lack thereof) of a modification strategy. A modification performed on one protein with one particular reagent may not perturb functionality, whilst the same process may render a different protein inert; one particular residue may be targeted in one protein, whilst a different amino acid is modified in another. Secondly, a consequence of the previous point, a protein modification strategy is likely to require some level of optimisation of conditions for each individual protein to achieve the desired outcome. It is not inconceivable, however, that a general modification strategy could be designed that minimises dependence upon variation amongst protein substrates and hence requires minimal tailoring. Thirdly, a protein is more than the sum of its parts: residue environment factors into residue reactivity, alongside which functional groups are present. At the time of this work, protein structure visualisation through methods such as crystallography and NMR was still not commonplace, meaning the information in the secondary, tertiary and quaternary structures of proteins could seldom be unambiguously considered. The modulation of residue pK_A values or the surprising lack of reactivity from some residues demonstrated the need to consider the microenvironment of reactive residues, their positioning on the surface of the protein and distance from interfering bulk protein regions. A synthetic chemist's perception of a particular amino acid may be wildly different from the role in which nature has deployed the residue in a protein. Finally, site selectivity is a crucial aspect of chemical protein

modification. Punctuating a sentence with a circumflex, when deployed correctly, can confer new, useful meaning; peppering the sentence with circumflexes will at best be of no use or hindrance, whilst at worst can completely diminish and obscure all of the information and meaning. Modification at one particular lysine residue may be innocuous, whilst modification at another, as observed with bovine pancreatic RNAse A, can leave the protein useless, whilst the generation of complex mixtures prevents unambiguous inference of the mixture composition. An ideal method would allow exclusive modification at the residue or residues desired with total control over reactivity- as nature has, of course, been doing.

1.1.4 Introducing "unnatural" modifications

The value of protein modification is not restricted to merely understanding the existing information bound inside a protein. Just as nature evolved PTMs to expand and augment the protein language beyond the initial canon of amino acids, so too can chemical protein modification be used to introduce new meaning or even an entirely new context. A major development was the conjugation of polyethylene glycol (PEG) polymers to proteins, with PEGylation attributed to improving protein solubility across a range of solvents, such as for industrial use of biocatalysis,[23] and increasing the stability and half-life of a therapeutic protein *in vivo*.[24] Initial PEGylation methods exploited the predictable nucleophilic character of lysine side chains with activated electrophilic derivatives of PEG, forming stable amine, amide or carbamate protein-PEG linkages (Figure 6).[25-27] The use of PEG activated with cyanuric chloride led to somewhat predictable off-target modification of nucleophilic active site cysteine thiols to inactivate enzyme functionality, but *N*-hydroxysuccinimidyl-PEG modification was found to offer minimal perturbation of functionality, presumably due to the milder activation method leading to fewer off-target modification of such thiol groups.[28] Cognisant of the shotgun-like nature of these methods, one approach made use of protecting group chemistry to add a further level of control over site selectivity. More nucleophilic α- and ε-amino groups were first protected with a water-soluble alkyoxycarbonyl protecting group, ensuring that mixed anhydride PEGylation only acylated two amino groups in the insulin protein used, with a final saponification treatment cleaving the remaining protecting groups.[29] Without such an approach, and when more complex proteins are used, PEGylation typically afforded a heterogenous mixture of proteins of varying molecular weights, with optimisation of conditions (usually varying the excess equivalents of the activated PEG reagent) required to target a particular average proportion of lysine residues targeted. Nevertheless, in this instance the end often justified the means, with even crudely PEGylated proteins displaying enhanced half-life *in vivo* as a direct result of PEGylation.[30, 31]

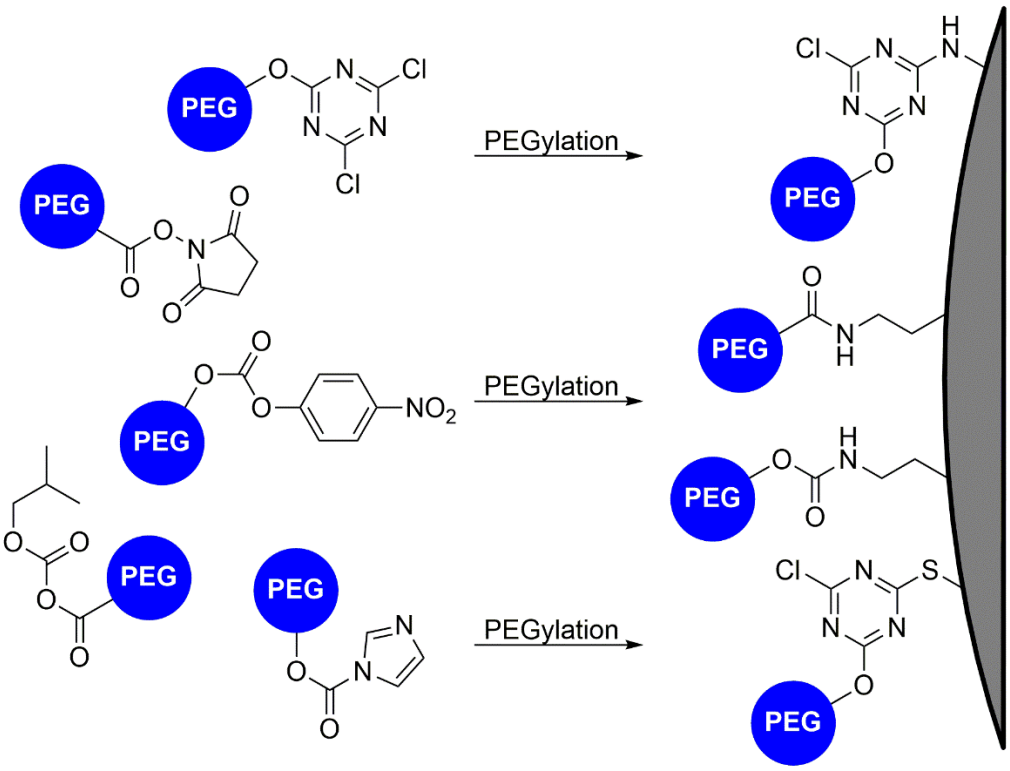


Figure 6: Activated PEG derivatives used to acylate or alkylate nucleophilic residues on protein surfaces for PEG attachment: primarily lysine residues, but also α-amino groups and thiols.[25-29]

An inspiration from nature came in the form of the use of the biotin-avidin pair, exhibiting one of the strongest binding constants known ($K_D \approx 10^{-15}$ mol dm^{-3}).[32] The initial discovery of this pair came through the onset of biotin deficiency in rats fed predominantly on raw albumen.[33] This was attributed to the high avidin content of this diet, with biotin, the so-called vitamin H,[34] being sequestered by the protein avidin and *Streptomyces* analogue streptavidin.[35] Biotin primarily acts as a cofactor in carbon transfer during metabolism, attached to and cleaved from the ε-amino group of a lysine residue of carboxylase proteins as a PTM where the ureido group captures carbonate to be attached as a carboxylate group to substrates such as pyruvate.[36] Avidin and analogues are suggested to have evolved as a defense mechanism: the decreased availability of biotin reduces the viability of bacterial pathogens in such environments by inhibiting metabolism, a conjecture supported by the localisation of avidin in avian eggs and seldom elsewhere in the same organisms.[37] More recently a further role of biotin was discovered as a PTM of lysine residues on histone proteins;[38] the precise nature and mechanism of this PTM is not yet fully understood, although various postulates consider biotinylation as a process involved in the regulation of gene expression or as a signal for DNA damage rather than a process involving an interaction with avidin or analogues.[39] Nevertheless, the power of the biotin-avidin pair was swiftly exploited as a biotechnological tool, intuitively used to demonstrate that pyruvate carboxylase was biotinylated through its retention on an avidin-Sepharose column.[40] It was quickly realised that the utility of biotin

lay in its role as the quarry of avidin, acting as a superb affinity tag for biomolecule substrates through avidin affinity chromatography or Western blotting (Figure 7).

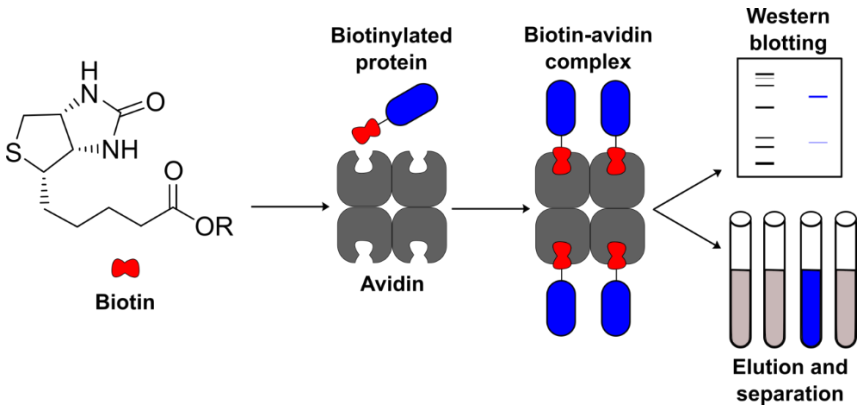


Figure 7: Biotinylated proteins can be separated from complex mixtures using avidin affinity chromatography or selectively detected using Western blotting with avidin-horseradish peroxidase or avidin-alkaline phosphatase conjugates.

Methods for the installation of biotin were quickly developed, following those used for chemical protein PEGylation. The free carboxylate group served as a helpful handle to prepare biotin activated with *para*-nitrophenyl, bromoacetyl and diazo groups for reactions with lysine, cysteine, and tyrosine/histidine residues respectively, with the same considerations and drawbacks as the analogous methods used for PEGylation.[41] An early limitation of the biotin-avidin pair was the occasional finding that the complexes of biotinylated proteins with avidin were less stable than anticipated, attributed to protein steric bulk impeding the biotin from accessing the binding site on avidin.[42] To ameliorate this issue, a spacer group was installed between the biotin and protein of interest. Whilst six-carbon alkyl linkers offered some improvement,[43] the best results were observed when short PEG spacers were used, owing to the enhanced aqueous solubility and reduced immugenicity conferred by conjugation to this polymer.[44] Optimisation of the affinity purification protocol has led to the use of monomeric avidin, rather than wild-type tetrameric avidin, which displays the same specificity for biotin but with a far lower binding constant ($K_D \approx 10^{-7}$ mol dm^{-3}) so that biotin-tagged proteins can be eluted under milder conditions compared to the harsh denaturing treatment required when using tetrameric avidin.[45] A frequent usage has been deploying biotin-tagged antibodies or hormones to bind with unknown antigens and receptors, with avidin affinity purification allowing isolation and elucidation of the binding partner.[46, 47] Through the repackaging of this evolutionary curiosity, biotinylation has become a staple of chemical protein modification, owing to the ease and reliability of avidin affinity as a method of detecting conjugated biotin and to the powerful biological probing abilities of the biotin tag.

Perhaps the most visually striking chemical protein modification is the introduction of fluorescence. In nature, fluorescence is commonly confined to small molecules rather

than proteins, bar a comparatively tiny family of fluorescent proteins. This paucity of protein fluorophores can be attributed to two shortcomings: the often-hydrophobic nature of fluorophores, hence poorly suited for aqueous milieu, and the common quenching or photocatalysed degradation of fluorophore excited states, to the extent that biosynthesis of the fluorophore is too inefficient a strategy compared to others in nature's possession. Development of synthetic protein fluorophores has also been sculpted by these two challenges and the two main fluorescent dyes used are dansyl and fluorescein derivatives,[48] where the problem of hydrophobicity is mitigated by polar substituents attached to the fused ring systems. The initial utility of this modification is the ease of verification, where SDS-PAGE shows a fluorescent protein band to qualitatively confirm successful bioconjugation. More intricate applications make use of donor-acceptor dye pairs through Förster resonance energy transfer (FRET), where the distance-dependent quenching of fluorescence of the donor dye by an acceptor dye can be used to estimate the distance between fluorophores.[49] In one such strategy, a surface-exposed tyrosine residue of liposome-bound gramicidin C peptides was labelled with either a donor dansyl dye or an acceptor diazo dye, with FRET-mediated quenching of dansyl fluorescence demonstrating that gramidicin C peptides are organised as dimers within membrane channels, separated by 33 Å.[50] This work highlights how chemical protein modification need not be particularly elaborate or sophisticated, but can still offer great insight when given a suitable biological application.

1.1.5 Contemporary methods for chemical protein modifcation

Chemical protein modification has been a fortunate beneficiary of the advances seen in analytical biochemistry. The structures obtained from protein crystallography are almost indispensible when designing a modification strategy for a protein of interest; molecular biology techniques allow access to target proteins produced reliably in house, rather than crudely extracting a protein of interest from a biological mixture; and mass spectrometry offers rapid access to high-resolution protein mass data which can quantify the extent of modification and characterise the bioconjugate through top-down or bottom-up proteomics. The sophistication of modern chemical protein modification is, at least in part, attributable to the power of these analytical tools which have enabled chemical biologists to ask more probing questions and to receive more unequivocal answers. This has, in turn, guided the current strategies in chemical protein modification. Site selectivity and specificity have become key characteristics of modification methods, with two main directions used to achieve this: tuning the chemistry to modify the native protein, or introducing a reactive, unnatural handle into the protein to open up a greater range of chemical methods.

Where native proteins are to be used for chemical protein modification, the same constraints seen for decades in this field pervade and strategies generally centre on the two most reliably reactive functional groups of α- or ε-amines and cysteinyl thiols. The latter functional group has received particular attention, where the rarity of this residue amongst proteins in nature goes some way towards solving the site selectivity problem and with the sulfur atom still offering a wealth of chemical reactivity. Sulfa-Michael addition reactions have received particular attention for this purpose, where the softness of the thiol compared to amino or hydroxyl nucleophiles affords a useful level of chemoselectivity. Maleimide-type chemistry (Figure 8) stands out as the favoured method for cysteine modifications: initially developed to quantify free thiol groups in proteins,[51] followed by as a cross-linking reagent,[52] maleimides were soon exploited as a superb handle for bioconjugations owing to the mild conjugation conditions, chemoselectivity, lack of side reactions, and the versatility and stability of the ensuing linkage. Wide-ranging applications of maleimide chemistry speak to the reliability and success of this work: fluorescent protein labelling,[53] designing biologically active protein-polymer conjugates,[54] immobilising proteins on quantum dots as part of a sensitive fluoroimmunoassay,[55] and preparing therapeutic antibody-drug conjugates.[56]

A known, minor limitation of maleimide chemistry is the tendency for a retro-Michael addition to occur, cleaving the bioconjugate in an uncontrolled and potentially undesirable manner.[57] This has been circumvented through either purposefully hydrolysing the thiosuccinimide ring,[58] which will subsequently not undergo retro-Michael addition,[59] or by tuning the maleimide reagent such that cleavage can be induced in a controlled manner. To this end, bromomaleimides have been developed, where the thiomaleimide formed following cysteine alkylation with bromomaleimide can be cleaved completely by treatment with a large excess of competing thiol.[60] Alternatively, the thiomaleimide can undergo a further sulfa-Michael addition with stoichiometric quantities of a second thiol, forming a dithiosuccinimide which is as resistant to thiol-mediated degradation as the thiosuccinimide formed when regular maleimides are used. The lability of dithiomaleimides was exploited using dibromomaleimides: a protein thiol can be alkylated and then linked *via* the maleimide core to a thiol tag, all of which can be cleaved by treatment with excess thiol or hydrolysed to stabilise the linkage.[61] Similarly, a dibromomaleimide or a dithiomaleimide tagging reagent can be inserted into disulfide bridges to retain native structural integrity whilst modifying with a useful tag, such as labelling an antibody with a drug payload at the interchain disulfide bridges.[62, 63] The development of maleimide chemistry has offered a high level of control particularly well suited to therapeutic applications, where payload delivery may necessitate reversible or irreversible conjugation to a protein depending on the circumstance.

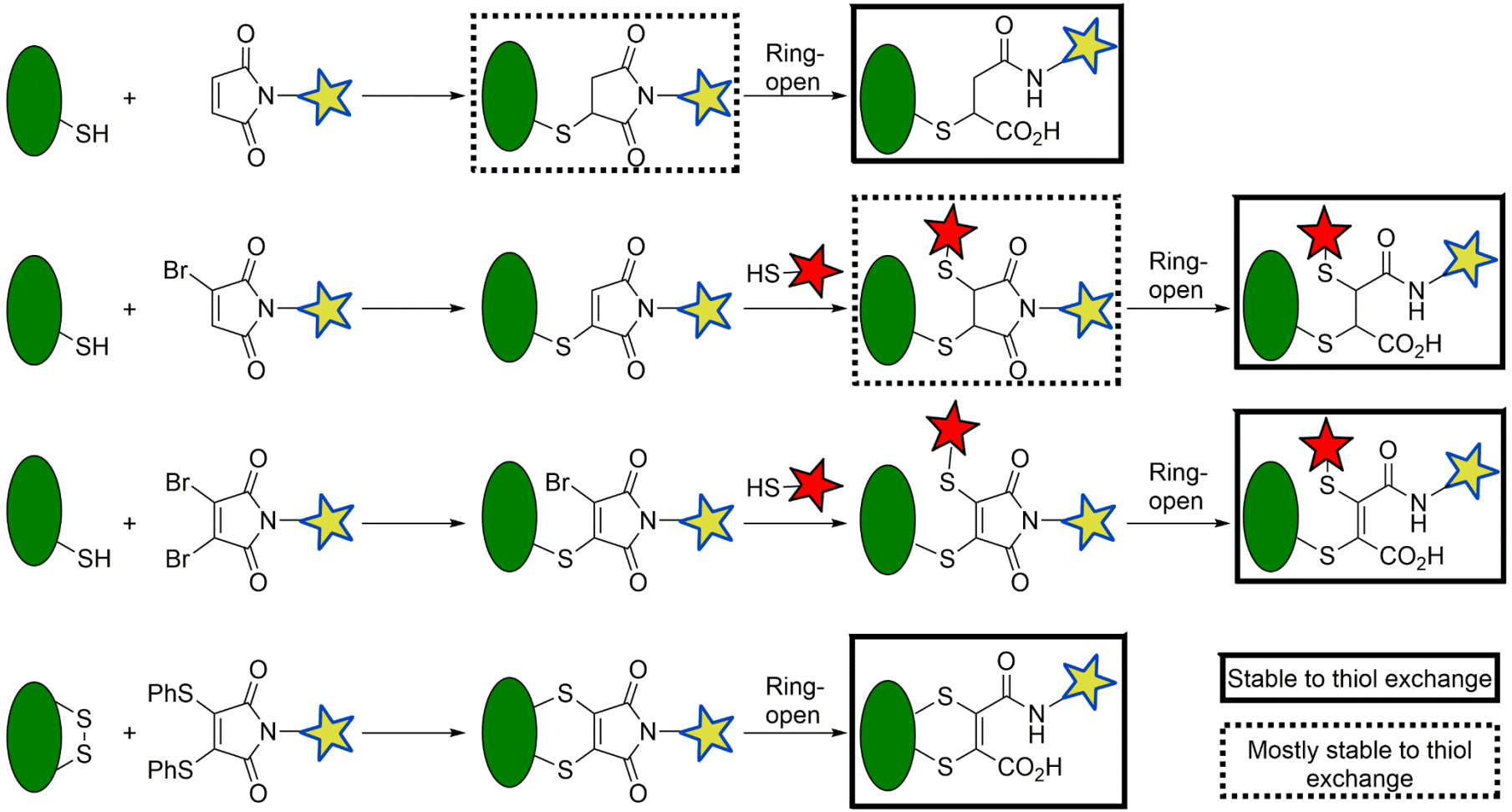


Figure 8: (from top to bottom) Protein modification of cysteine thiols using maleimides, bromomaleimides and dibromomaleimides to form thiosuccinimides (slightly prone to retro-Michael addition in the presence of thiols), thiomaleimides (reversibly exchange with excess thiol) and maleic/succinic acids (stable in the presence of thiols). A cystine disulfide bridge can, once reduced, react with a dithiomaleimide and subsequently ring-open to form a stable linkage where the cysteine residues are still joined with the addition of a tag such as a small-molecule drug.

Several strategies for modification of native *N*-termini *via* the α-amino group have emerged, most commonly relying on the same pK_A difference to establish site selectivity. The use of low pH favours modifications exclusively at the α-amino group, such as acylations using *N*-hydroxysuccinimide (NHS) esters and reductive amination, although such acidic conditions may not be suitable for all proteins and are often outside the optimal pH range for a given chemical modification strategy, leading to reduced bioconjugation yields.[64] Fortunately, the side chain functionality of the *N*-terminal residue can be drafted in to assist with such reactivity. Seryl and threonyl 1,2-aminoalcohols can form oxazolidines with aldehydes (Figure 9),[64] whilst cysteinyl 1,2-aminothiols can similarly trap aldehydes to form thiazolidines,[65] react with cyanobenzothiazolines,[66] or partake in the powerful native chemical ligation reaction.[67] The limitations of *N*-terminal strategies do hinder these strategies, however: the *N*-terminal residue may be buried within the protein or rendered unavailable for chemistry due to participation in active site function,[68] the modification can only be applied once to the protein, and the restriction of the modification to the *N*-terminus may not be desirable for the purpose of the bioconjugation, e.g. too close to the active site.

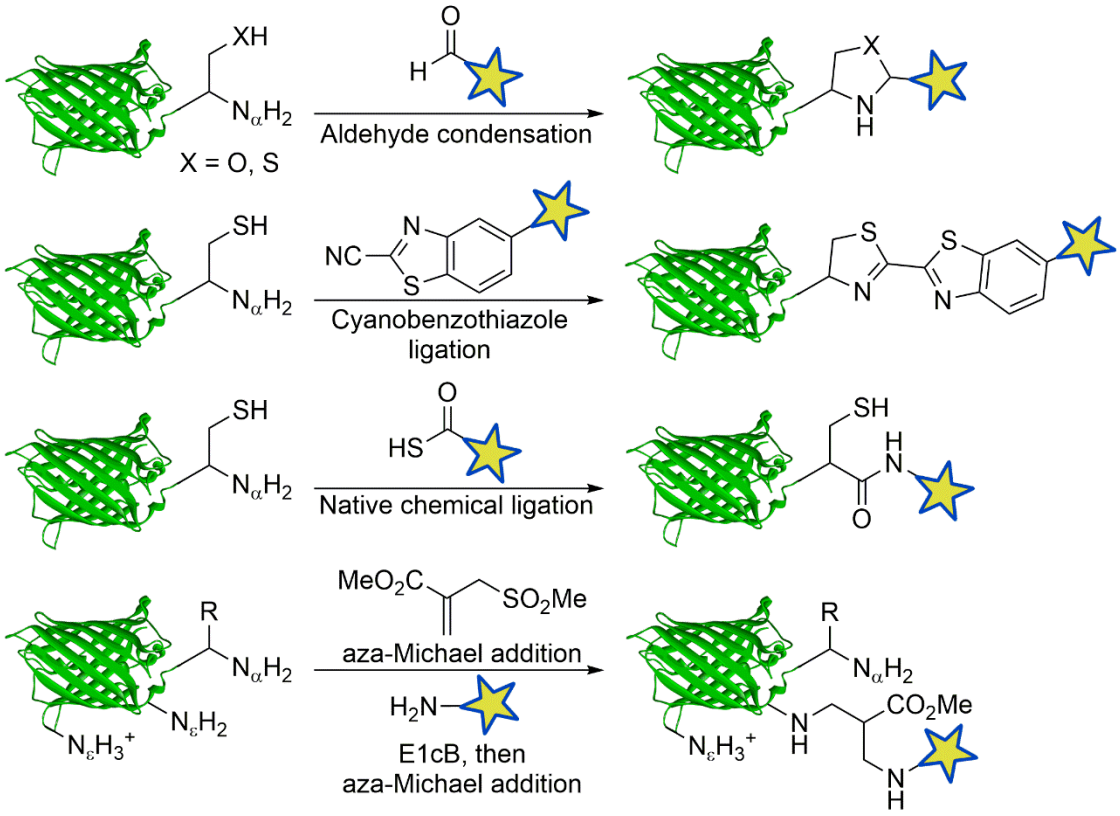


Figure 9: Modification of native amino groups at the N-terminus and at selective lysine side chains.

Modification of lysine amino groups is still held back by the distribution of multiple reactive ε-amino groups across the protein surface. In one instance, extensive kinetic optimisation of acylation using NHS esters afforded labelling at a single residue, but the

use of low equivalents (0.5) of labelling reagent lead to incomplete labelling and mandated affinity purification using an installed biotin tag.[69] A more sophisticated evolution of this strategy used computational modelling of the pK_A constants and microenvironments of ε-amino groups along with the design of various electrophilic aza-Michael acceptors to devise a labelling strategy modifying only a single lysine residue to completion (Figure 10).[70] From initial crystal structures to complex molecular dynamics calculations, the use of computational modelling to inform labelling of native proteins is likely to continue its expansion over the coming years, countering the diversity of the proteome with methods tailored both *in silico* and *in vitro* to every protein modification scenario imaginable.

Constrained by the limitations of early chemical modification methods for native substrates, attention turned upon the alteration of native residues to install non-native chemical functionality as a handle for new chemistries. This was initially discovered through harsh chemical methods used to understand protein structure, where proteins containing *N*-terminal serine or threonine residues exhibited a greater level of reactivity towards periodic acid[71] or periodate than expected, with the detection of formaldehyde byproduct used to confirm the presence of an *N*-terminal serine or threonine residue.[72] The residue left behind was identified as a glyoxylamide (Figure 10) owing in part to its reduction by borohydride,[73] the utility of which was then exploited through the transamination of this reactive aldehyde in the presence of Cu^{2+} and an amine donor to form a native glycine residue: a “chemical mutation”.[74] Intuitively the reverse reaction was also discovered, transaminating a native *N*-terminal glycine residue using Cu^{2+} and glyoxylic acid as an amine acceptor and subsequently exploiting the reactivity of the protein glyoxyl aldehyde with anilines to form Schiff base bioconjugates.[75] This reaction has now been optimised to work under milder conditions without any metals through the biomimetic use of pyridoxal-5-phosphate (PLP), a staple of biosynthetic transaminations, although the reaction remains an equilibrium which can hamper conversions.[76] A rich vocabulary of aldehyde chemistry has hence been translated for use on protein aldehyde substrates,[77] as the presence of a single aldehyde in a protein means that bioconjugations are inherently site-specific if the chemistry is chemospecific for aldehydes. Oxime/hydrazone ligations have surpassed Schiff base formation as the premier strategy for conjugating to protein aldehydes, where the hydroxylamine or hydrazine reagent is not only more nucleophilic, leading to greater rates of reaction, but the resulting bioconjugate, especially the oxime linkage, is also less electrophilic and therefore less prone to hydrolysis.[78-80]

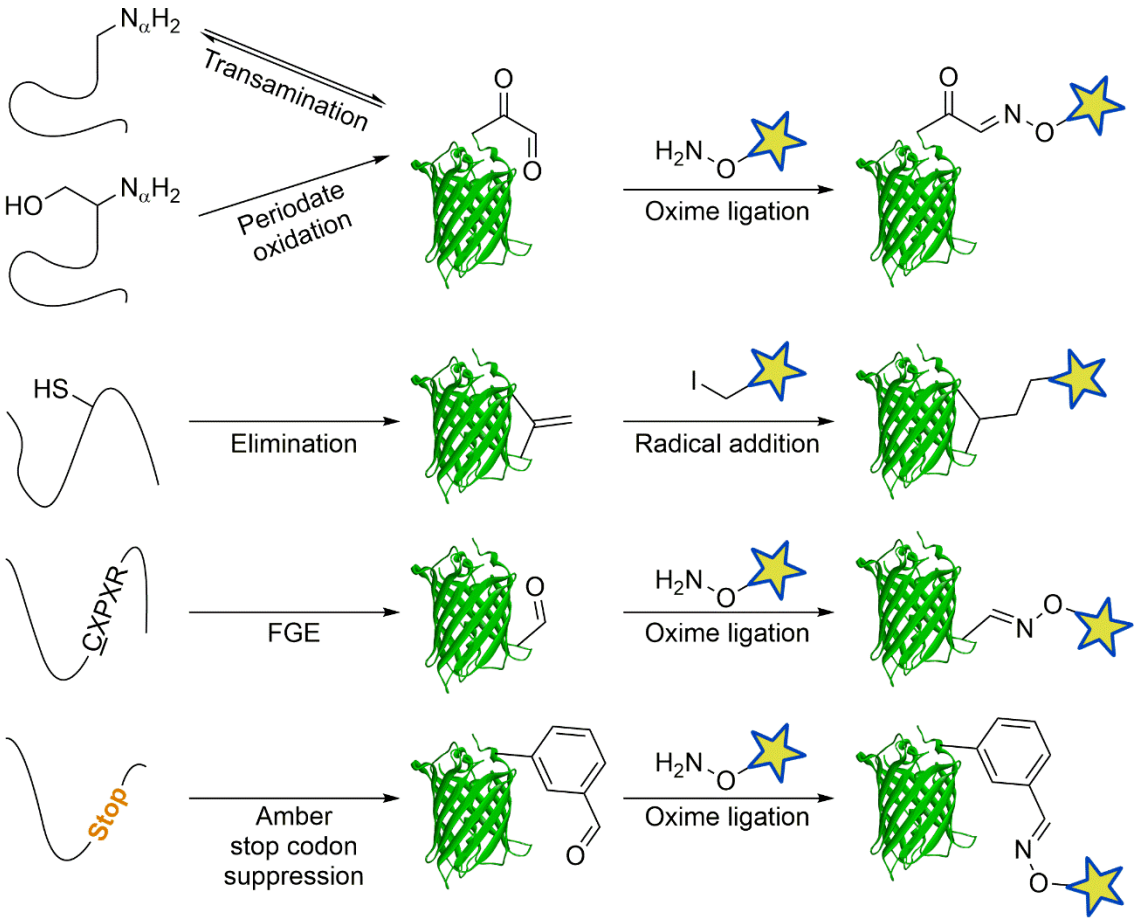


Figure 10: Selected methods for the incorporation of aldehydes and dehydroalanine into proteins for subsequent chemical modification.

A further demonstration of "chemical mutation" was observed through reactions at a nucleophilic active site serine residue: following activation with an acyl halide, the leaving group could either be substituted by a thioacetate nucleophile and hydrolysed to yield a native cysteine residue,[81] or eliminated with a base to form the residue dehydroalanine.[82] Whilst initially a novelty perturbing enzymatic performance, the authors noted the opportunities opened up by dehydroalanine as a means through which additional functionality could be installed in proteins.[83] Of course nature had already exploited this tool, such as generating dehydroalanine in lantibiotics through the enzymatic dehydration of specific serine or threonine residues and forming cyclic peptides through intramolecular sulfa-Michael additions with cysteine thiols.[84] Not wishing to be outdone by nature, intense research on the generation and reactivity of dehydroalanine ensued. Multiple methods have been devised for the installation of dehydroalanine, with perhaps the most useful method being the elimination of cysteine thiols (Figure 10), often activating the thiol as a sulfonium ion leaving group *via* bis-alkylation.[85] A notable consideration is the scrambling of chirality at the dehydroalanine residue, where the lack of chiral induction in subsequent chemical modification methods leads to a roughly 1:1 mixture of epimers.[86] In addition to afore-mentioned hetero-Michael additions,

dehydroalanine residues can undergo radical additions to form stable carbon-carbon bonds with a range of small molecule tags, including synthetically challenging post-translational modification mimics.[87]

A final boost to chemical protein modification has come directly from nature itself. As molecular biology techniques progressed, many useful enzymes suddenly became available in laboratories with very little effort required for production and some of nature's own protein modification strategies were uncovered. The active sites across a range of sulfatases contained the unusual aldehyde residue formylglycine, where the hydrated aldehyde diol played a mechanistic role in sulfate ester hydrolysis.[88] This residue was found to arise in a highly conserved CXPXR recognition sequence in which the cysteine residue is converted to formylglycine by formylglycine-generating enzyme (FGE). The biotechnological applications of this quirk were quickly realised: the installation of a short CXPXR sequence within a protein led to the exposure of a reactive aldehyde following treatment with FGE (Figure 10),[89] opening up aldehyde chemistry beyond the *N* terminus and even enjoying use for preparing antibody-drug conjugates.[90] A booming area of investigation is the use of amber stop codon suppression, in which an amber stop codon is reassigned as a proteinogenic amino acid with the assistance of an amber stop codon tRNA and the corresponding tRNA synthetase. Through the use of a *Methanocaldococcus janaschii* tRNA-synthetase pair, many tyrosine/phenylalanine analogues, including one bearing a reactive aldehyde,[91] have been incorporated into proteins and subsequently functionalised, with the single amino acid mutation representing a small structural perturbation with a big potential for chemical protein modification.[92]

1.1.6 The real-life impact of chemical protein modification

It is far too easy to continuously ponder and philosophise over chemical protein modification, and indeed most of science, when the repercussions of this technology on daily life are absolutely worth consideration. For chemical protein modification, or indeed any area of research, to be truly useful, the positive findings espoused in myriads of publications must be translated into beneficial tools for society. This has been the case with chemically PEGylated proteins, such as Pegasys. As a PEGylated derivative of interferon-2α, Pegasys was first approved in Switzerland in 2001 for the treatment of chronic hepatitis B and C infections,[93] with PEGylation responsible for the superior antiviral activity and longer half-life compared to standard interferon-2α. The PEGylation process uses a standard acylation method between a free interferon-2α amino group and a 40 kDa PEG NHS ester, with conditions optimised for monoacylation at the side chains of Lys-31, Lys-121, Lys-131 or Lys-134.[94] Pegasys is now listed on the World

Health Organisation List of Essential Medicines as a clear testament to its efficacy and thus to its impact on public health.

Perhaps the most celebrated example of chemical protein modification is the antibody-drug conjugate (ADC) trastuzumab emtansine, marketed as Kadcyla. Trastuzumab, marketed as Herceptin, is an antibody therapeutic which targets HER2 receptors, overexpressed on certain types of breast cancer, blocking this signalling pathway attributed to cell proliferation.[95] Combined with chemotherapy, trastuzumab just as the antibody was found to increase median survival time by 25%, approximately five months of extra time alive.[96] Kadcyla combines the targeting and HER2 binding of the trastuzumab antibody with the small molecule drug DM1 *via* a stable covalent linkage. To achieve this, first a bifunctional linker is attached *via* the NHS-activated carboxylate end to amino groups on trastuzumab, followed by sulfa-Michael addition of the DM1 thiol to the maleimide portion of the linker.[97] Every trastuzumab antibody is loaded with approximately 3.5 molecules of DM1 on average, as the lack of control in the NHS acylation step leads to some heterogeneity.[98] Nevertheless Kadcyla is approved for use against challenging cases of advanced HER2-positive breast cancer, with a significant phase III trial finding a 20% increase in median survival rates in patients receiving Kadcyla treatment,[99] and surely represents the boldest move yet for chemical protein modification out of the lab and into society.

1.1.7 Conclusions

The rise of chemical protein modification has been an evolution against evolution itself. This powerful technique has aided with the unravelling of nature's protein language, offered entirely new words for the protein vocabulary, and has been translated into forms that widely benefit humankind in the unwinnable fight against nature's planned obsolescence mechanisms, just as nature developed post-translational modifications to protract this struggle. Formerly an analytical tool primarily used to translate the protein language into human knowledge, powerful mass spectrometry and other methods now serve the needs of chemical protein modification, unlocking its full potential as a means to introduce new meaning into the protein language. The current directions of the field can be summarised by two main objectives: developing the most diverse array of methods, so that the toolbox contains a tool for every one of nature's diverse proteins, and developing further applications of this technology which further improve daily life, which primarily leads in a therapeutic direction. Given the rapidity with which this field moves, certainly over the past decade, there is no doubt that many inspired and imaginative strategies will be devised to meet these objectives.

1.2 Amber stop codon suppression: a chemical biology tool

1.2.1 Pyrrolysine, an unusual canonical amino acid

Arising from peculiarities observed in the proteome and genome of an otherwise innocuous methanogenic microorganism, an entirely new canonical amino acid was added to the genetic code. The archaeal species *Methanosarcina barkeri*, capable of utilising a range of alkylamines for methanogenesis,[100] was found to utilise three methyltransferase enzymes at the start of three analogous converging catabolic pathways in which mono-, di- and trimethylamine substrates were demethylated and reduced to ammonia and methane. Unusually the genes encoding the three methyltransferase enzymes were found to contain in-frame amber stop codons, TAG, which did not appear to terminate translation and were posited to genetically encode a lysine derivative based upon trypsin digests of the methyltransferases.[101] Around the gene cluster encoding the methyltransferases were two further oddities: the gene *pylT*, encoding a tRNA with a CUA anticodon, which would be complimentary to a UAG mRNA codon, and the gene *pylS*, encoding an aminoacyl-tRNA synthetase, pylRS, which was found to aminoacylate $tRNA_{CUA}$.[102] Mass spectrometry and crystallisation of a TAG-containing methyltransferase permitted elucidation of the recondite residue, assigned as pyrrolysine **1** (Pyl; single letter code O) in which the methylpyrroline ring, appended to a lysine backbone, was proposed as playing a crucial role as an electrophilic catalytic residue, trapping methylamines and activating the methyl group towards capture by the cobalt(I) centre of a corrinoid metalloenzyme.[103] Addition of the methylpyrroline ring was postulated to not be a PTM in the same vein as other acylated/alkylated lysines **2**-**4**, but instead the product of a biosynthetic pathway.[104] This was later demonstrated to be correct, as **1** was found to be biosynthesised from two units of lysine **5** *via* **6** and **7**[105] by pylB,[106] pylC,[107] and pylD (Figure 11).[108-110]

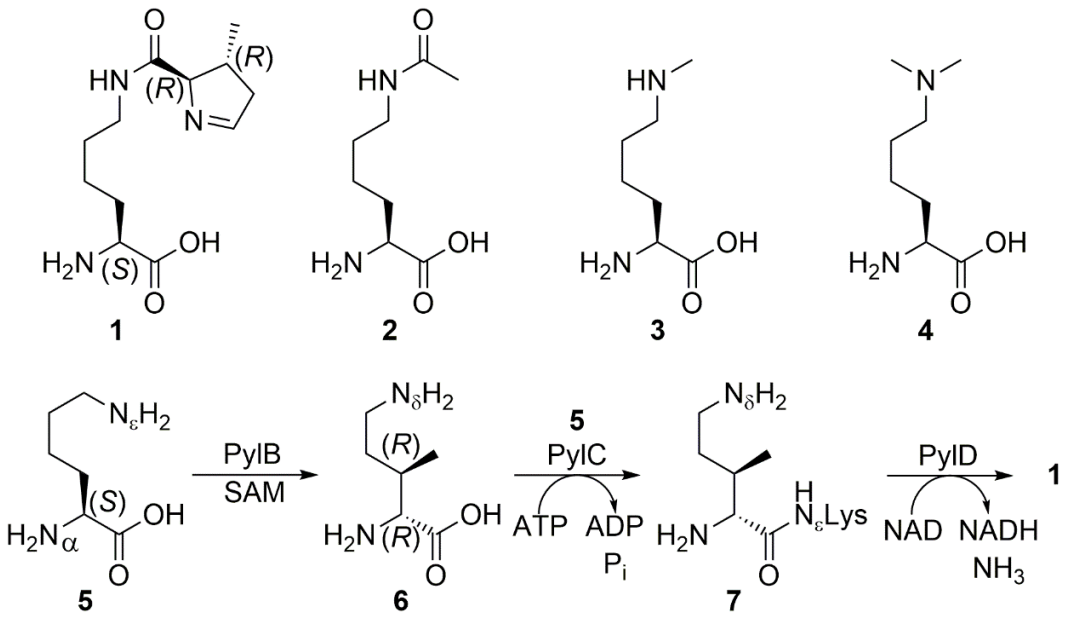


Figure 11: Pyrrolysine **1**, unlike post-translational modifications **2**-**4**, is biosynthesised prior to translation from L-lysine **5** *via* **6** and **7**.

With this reasoning established, the pathway to pyrrolysine-containing proteins can be posited. During translation, $tRNA_{CUA}$ is produced and charged with pyrrolysine by pylRS, affording pyrrolysyl-$tRNA_{CUA}$ which can be utilised in the ribosome to add pyrrolysine to the growing polypeptide chain at UAG codons. Thus, for amber stop codon suppression of a UAG-containing gene, three components are necessary: the amino acid pyrrolysine, the $tRNA_{CUA}$, and the pyrrolysyl tRNA synthetase, which nature has gifted *M. barkeri* and *Methanosarcina mazei* (*inter alia*) in the form of the *pylBCDST* genes. The orthogonality of the pyrrolysyl tRNA-RS pair to canonical amino acids is a further blessing, affording reliable pyrrolysine incorporation: only pyrrolysine, and no other canonical amino acids such as lysine, are suitable substrates for pylRS, whilst no canonical tRNA synthetases can charge any tRNA with pyrrolysine.[111] Given the criticality of pyrrolysine to the survival of *M. barkeri* in such inhospitable anaerobic environments, it is no wonder that nature has evolved such a disciplined mechanism for the utilisation of this privileged amino acid.

1.2.2 A Promiscuous tRNA-RS Pair for Protein Production

As mentioned, the pyrrolysyl tRNA-RS pair is orthogonal to canonical amino acids other than pyrrolysine. This orthogonality can be rationalised through nature's careful engineering of the structure and active site of the pylRS protein. A crystal structure of the wild type *M. mazei* in complex with Pyl-AMP, an intermediate in the formation of the aminoacyl-tRNA complex, exposed the presence of an unusually hydrophobic and caverous pocket (Figure 12).[112] This additional binding space accommodates the pyrroline ring of pyrrolysine, otherwise too bulky to be accommodated in the active site of canonical RNA synthetases. The overall structure of this RS and its propensity to exist

as a homodimer, with a homotetrameric complex considered possible, led to its classification as a class IIc RS with its closest relative being an ancestral form of phenylalanyl RS, putatively existing as a homotetramer.[112]

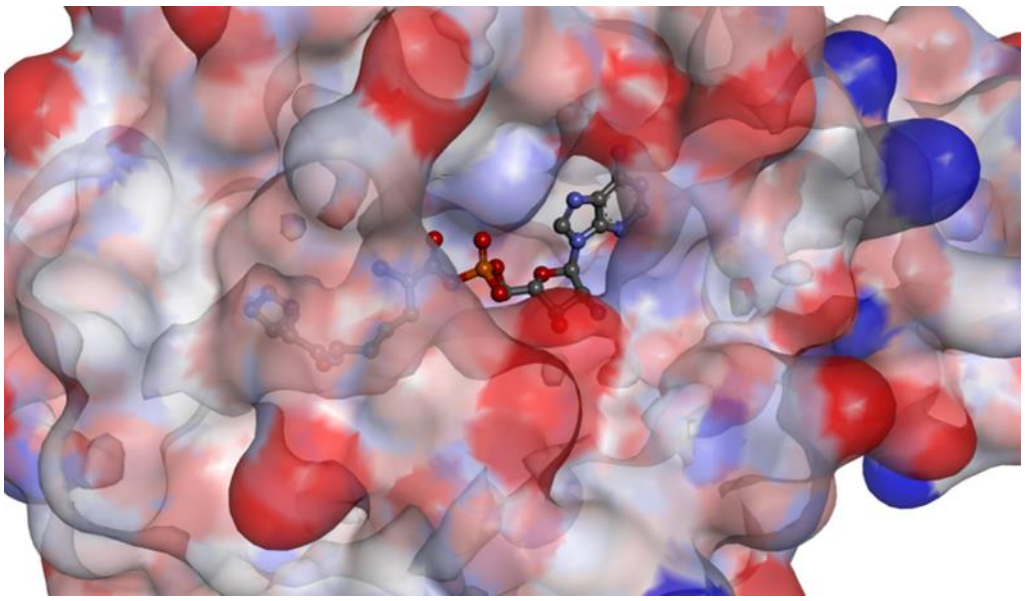


Figure 12: The hydrophobic methylpyrroline ring of **1** occupies a hydrophobic cleft deep within PylRS (PDB: 2Q7H).

Closer examination of the hydrophobic binding pocket affords a more complete rationalisation of the specificity of the pyrrolysyl RS. Tyr384 and Asn346 are arguably the most crucial residues, given both their function and their highly conserved nature:[112] Tyr384 acts as a kingpin, hydrogen-bonding to both the α-amino and α'-imine groups of pyrrolysine through the phenolic hydroxyl group,[113] whilst Asn346 hydrogen bonds directly with the amide carbonyl and through a water molecule with the α-amino group (Figure 13, left). Beyond these two gatekeeper residues, the outline of the deep hydrophobic pocket can be seen defined by the non-polar side chains of residues Ala302, Leu305, Tyr306, Leu309, Cys348 and Trp417 (Figure 13, right), accommodating the hydrophobic region of the methylpyrroline ring. The arrangement of this binding pocket clearly permits comfortable binding of pyrrolysine whilst deterring all other canonical amino acids. Obtaining a crystal structure of pylRS certainly offered a greater level of insight into the workings of this enzyme. However, this process directly yet perhaps unintentionally led to further discoveries which are arguably of far greater magnitude and consequence.

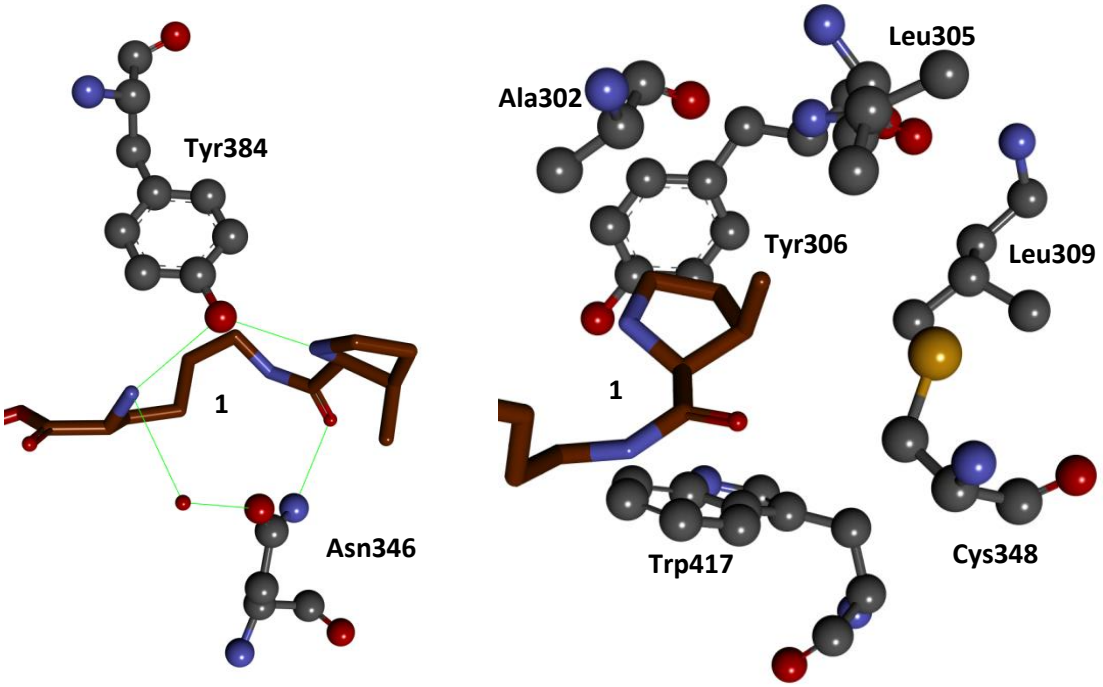


Figure 13: (left) Hydrogen bonding interactions with Tyr384 and Asn346 hold **1** in place; (right) various hydrophobic residues line the cavity occupied by the methylpyrroline ring (PDB: 2Q7H).

Attempts to obtain holo structures of pylRS were impaired by limited access to pyrrolysine itself. The methylpyrroline ring acts as a highly uncooperative moiety of the ligand in this sense, hydrolysing and ring-opening in a range of conditions and, for the same reason and the presence of two chiral centres, presenting a menacing synthetic challenge. The first reported synthesis of pyrrolysine constructed the methylpyrroline ring through a complex-directed asymmetric Michael addition between a nickel-coordinated glycine ligand and crotonaldehyde, followed by hydrolysis and cyclisation (Figure 14).[114] This latter step, achieving 43% yield, hampers the utility of this synthesis, affording only 9% final product over eight steps.

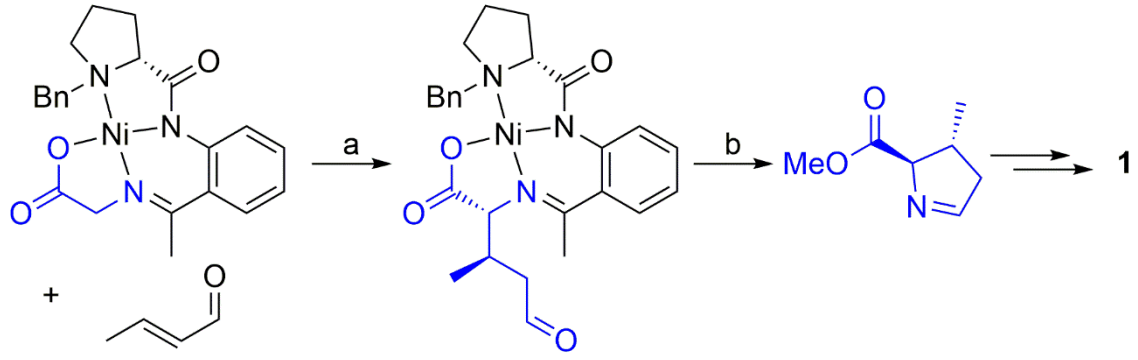


Figure 14: Asymmetric Michael addition of glycine to crotonaldehyde followed by hydrolysis and cyclisation forms the crucial methylpyrroline ring (blue) of **1**. a: DBU, DCM, r.t., 97%. b: HCl, MeOH, Δ, then TMSCl, MeOH, r.t., 43%.

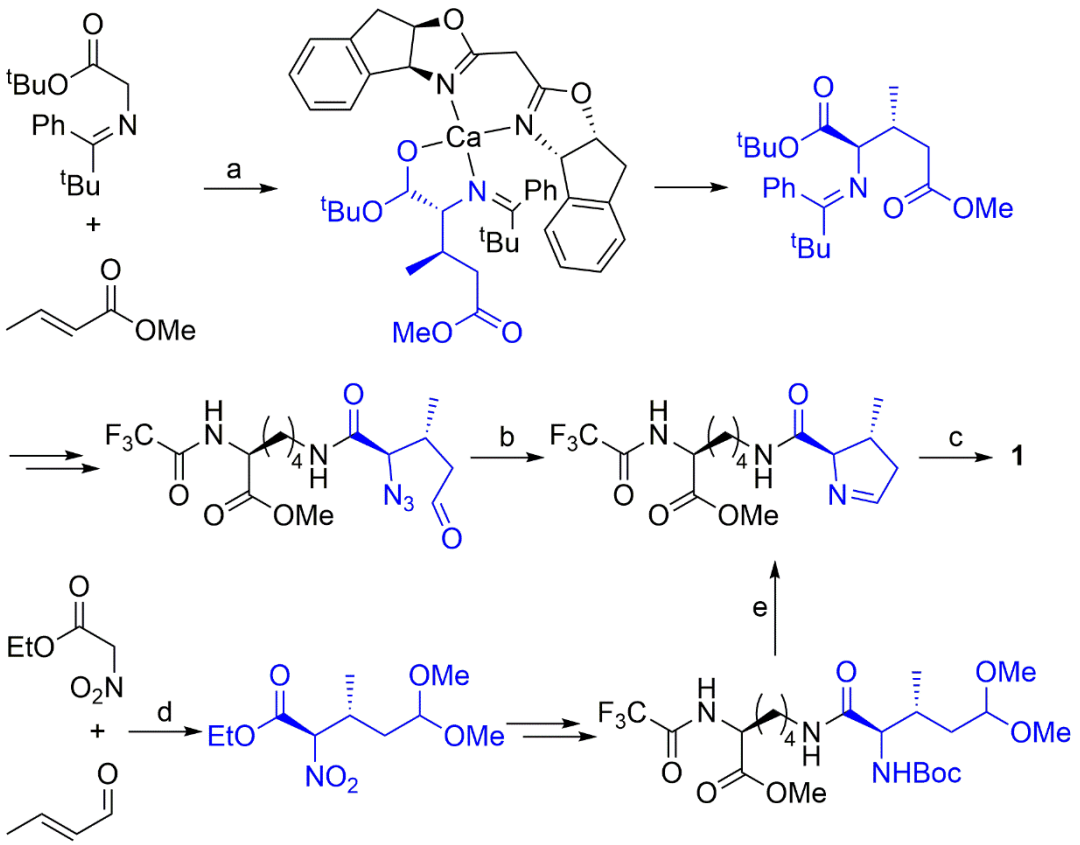


Figure 15: Two recent synthetic routes to **1**: (upper) metal-mediated asymmetric Michael addition of a glycine imine to methyl crotonoate prepares the stereocentres and a tandem Staudinger/aza-Wittig cyclisation closes the ring; (lower) organocatalytic asymmetric Michael addition of ethyl nitroacetate to crotonaldehyde affords the correct stereochemistry and subsequent deprotection and cyclisation forms the methylpyrroline ring of **1**. a: $Ca(O^iPr)_2$, 4 Å molecular sieves, THF, -20 °C, 99% e.e., >98% d.e., 92%. b: polystyrene-supported PPh_3, THF, Δ, quantitative. c: LiOH, MeOH, THF, H_2O, quantitative. d: Jørgensen catalyst, BzOH, then MeOH, *p*-TsOH, $CH(OMe)_3$, 90% e.e., 1:1 mixture of C-2 diasteromers, 96%. e: *p*-TsOH, MeOH, 32%.

To date, few strategies have improved upon these results. One strategy makes use of another metal-mediated asymmetric Michael addition to prepare both chiral centres with the correct configuration, with the final step of imine cyclisation *via* a tandem Staudinger/aza-Wittig reaction and subsequent deprotection affording pyrrolysine in 20% yield over 13 steps (Figure 15).[115] The shortest strategy, six steps, affords pyrrolysine in 19% yield: starting again with an asymmetric Michael addition, this time directed by a Jørgensen organocatalyst,[116] albeit affording a 1:1 mixture of diastereomers at C-2. Fortunately the "incorrect" diasteromer was found to isomerise to the correct diastereomer under strongly basic conditions, leading to the conclusion of the strategy with tandem deprotection and imine formation.[117] Nature's neat four-step biosynthesis easily outpaces these strategies, although pyrrolysine synthesised in this way would suffer from the same impractical degradation predicament impeding crystal preparation.

Evidently sufficient quantities of pyrrolysine were scraped out to lead to successful crystallisation of pylRS with Pyl-AMP. However a fortuitous discovery demonstrated that analogues **8** and **9** were suitable substrates for pylRS,[118] and an alternative holo structure of pylRS with ATP and **9** ligands was obtained (Figure 16).[112] Both pyrrolysine analogues also demonstrated suitability for proteinogenic amber stop codon suppression, allowing the production of active β-galactosidase despite the presence of an otherwise destructive Trp220-TAG mutation in the *lacZ* gene.[118] By this time, amber stop codon suppression had already been used to prepare useful protein constructs using leucyl[119] and tyrosyl $tRNA_{CUA}$-RS pairs from *Eschericia coli*[120] and *M. janaschii,*[121] all of which exhibited a broad tolerance to various non-canonical amino acids (NCAAs). Hence the greater utility of pylRS was theorised: as monogamous for pyrrolysine as pylRS may be amongst canonical amino acids, perhaps pylRS exhibits a more open relationship with non-canonical pyrrolysine analogues and can introduce a range of non-canonical lysine analogues into the genetic code, complementing existing non-canonical tyrosine and leucine analogues.

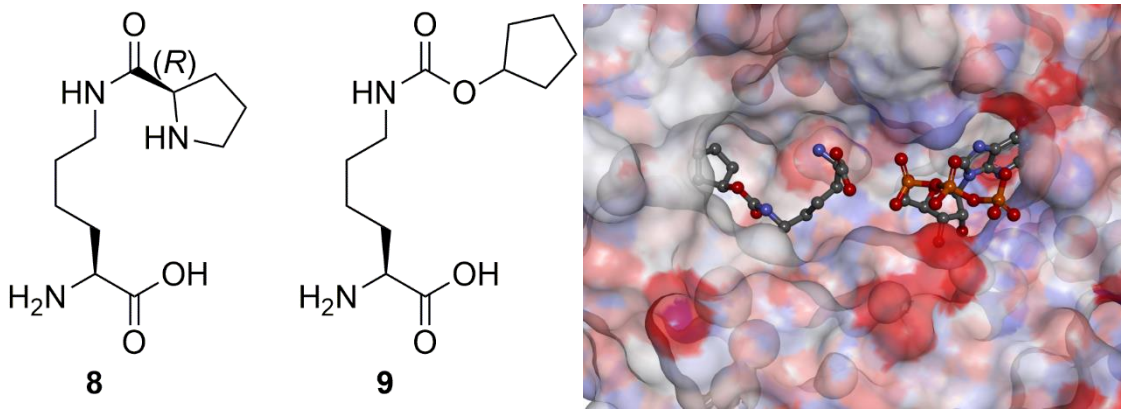


Figure 16: (left) Pyrrolysine analogues **8** and **9** were found to be suitable substrates for *M. mazei* pylRS, with a crystal structure of **9** bound to pylRS (right; PDB: 2Q7G).

The limits of the promiscuity of pylRS were further tested by lysine derivatives **10-13** (Figure 17). Both **10** and **11** were found to be suitable substrates for wild-type *M. mazei* pylRS[122] but **12** and **13** were not.[123] Through the use of enzyme engineering, key active site residues were randomised and, through multiple selection rounds, pylRS mutants were identified which could recognise **12** and **13** as substrates. The three mutations L309A, C348V and Y384F permitted turnover of **12**,[123] whilst with only two mutations, a Y306A Y384F pylRS variant ($pylRS^{AF}$) recognised both **12** and **13**, with the *ortho* substituent seemingly well-tolerated owing to the greater space in the enlargened binding pocket.[122] Furthermore, incorporation of **13** into a protein, in this case GST, allowed chemical modification using Bertozzi-Staudinger ligation,[124] affording a fluorescently labelled GST bioconjugate. It was also discovered that the single Y384 mutation enhanced turnover of **1**, **10** and **11**, albeit at the expense of introducing some undesirable

turnover of canonical amino acids. A higher rate in the generation of aminoacyl-tRNA was ascribed to a secondary role of the Y384 residue: as a physical gatekeeper, in addition to a figurative gatekeeper. Removal of the Y384 hydroxyl group putatively opens up access for the $tRNA_{CUA}$ terminal base to be acylated by the carboxylate phosphoester, lowering the activation energy for this pathway, although this kinetic boost is accompanied by the unwanted thermodynamic boost in non-specific amino acid binding. Nevertheless, this work demonstrated two crucial findings: not only the possibilities and limitations of enzyme engineering in broadening the range of pyrrolysine analogues used with pylRS, but also the utility of the expanded range of pyrrolysine analogues for protein modification and beyond.

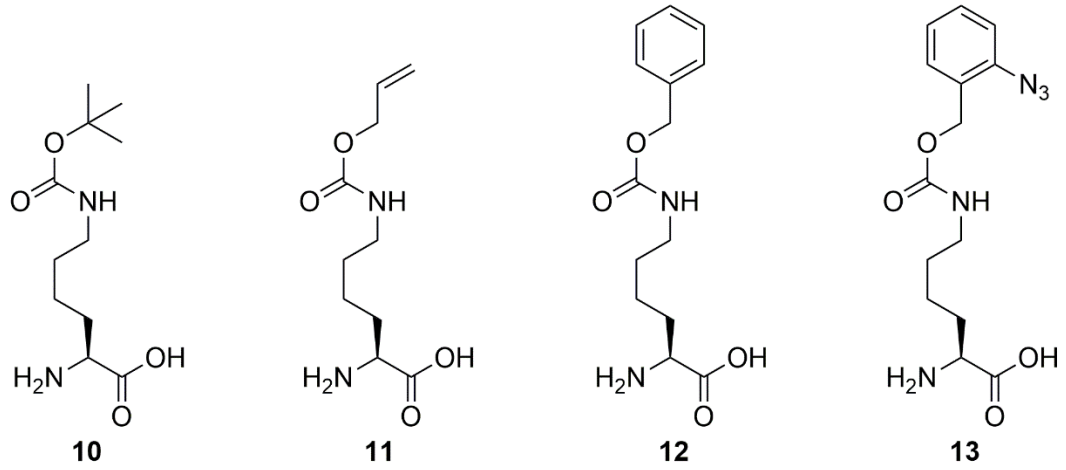


Figure 17: Protected derivatives of lysine recognised by wild-type or mutated pylRS.

For a tRNA-RS pair to be useful for protein modification, the protein production process must first be established (Figure 18). Owing to the absence of $tRNA_{CUA}$ and pylRS genes, host organisms must be transfected not only with the gene of interest but also with the suppressor pair genes. Multiple vectors have been designed which combine various genes on the same vector,[125] utilise evolved $tRNA_{CUA}$ species for superior stop codon suppression,[126] permit suppression of multiple stop codons[127] using multiple NCAAs,[128] and can be successfully used in eukaryotic host organisms.[129-131] During protein production, growth medium must also be supplanted with the NCAA(s) of choice, which can be through chemical synthesis or host biosynthesis.[132] Yields of the protein of interest are generally lower when amber stop codon suppression is used, in part due to premature termination of translation at the reassigned stop codon. Efforts to circumvent this problem have involved the generation of a viable host species of *E. coli* with all native amber stop codons mutated to opal or ochre[133] and through the gene knock-out of translation-terminating release factor one protein,[134] rendered redundant under these conditions.

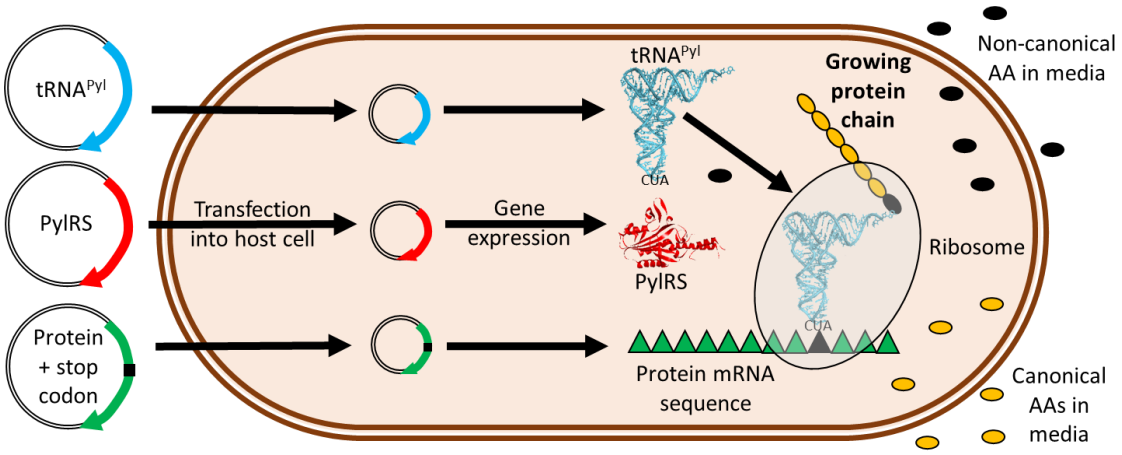


Figure 18: An overview of amber stop codon suppression.

1.2.3 Protecting groups and post-translational mimicry

One of the immediate priorities in the development of the pyrrolysyl tRNA-RS pair was the ability to genetically incorporate acetylated lysine **2**, as this residue is highly abundant in the proteome as a PTM. One set of high-value targets of this nature are histone proteins, in which the acylation and alkylation of various lysine residues has implications for various disease states and processes. An early method of preparing histone H4 with K16 acetylation required the use of native chemical ligation, stitching together a 22-mer synthetic peptide containing **2** with the remaining 80-mer peptide produced using recombinant methods.[135] As innovative and powerful a use of semisynthetic chemistry this may have been, it remains a laborious, challenging process and offers no guarantee that the final protein adopts the native fold and retains full function.

Amber stop codon suppression offered a superior solution, producing full-length recombinant proteins far more efficiently, yet was limited by the poor turnover of **2** by pylRS. Rational engineering of both *M. barkeri* and *M. mazei* pylRS proteins led to successful recognition of **2**, with analogous mutations possible due to the substantial level of homology between the two variants.[136] The four key mutations of L305I, Y306F, L309A and C348F applied to the *M. mazei* pylRS closed off the deepest portion of the binding pocket, allowing the comparatively small **2** to comfortably occupy the now-shrunken space (Figure 19).[123] The binding mode adopted is vastly different from that of **1** in the wild-type *M. mazei* pylRS: N346 mostly interacts with the α-amino and carboxylate groups whilst Y384 plays little role, appearing highly disordered. The carbon chain of **2** has straightened out: the α' centre lies deeper within the binding pocket, potentially to place the carbonyl oxygen atom in reach of aromatic residues for n→π* interactions, although this binding mode is overall suboptimal both upon inspection and by the measured amber stop codon suppression efficiency.[137] Nevertheless, with success demonstrated by the production of **2**-containing myoglobin, this methodology

was then applied to histone H3 in order to understand the effect of K56 acetylation, as the acetylation at this site in the middle of the protein rendered many synthetic methods unsuitable. Homogeneous samples of H3 containing native or acetylated lysine residues at position 56 were both prepared, the latter using amber stop codon suppression and **2**, permitting a thorough investigation into the function of H3 K56 acetylation in processes such as DNA breathing,[138] alongside the preparation of a variety of other monoacetylated H2A, H2B and H3 histones.

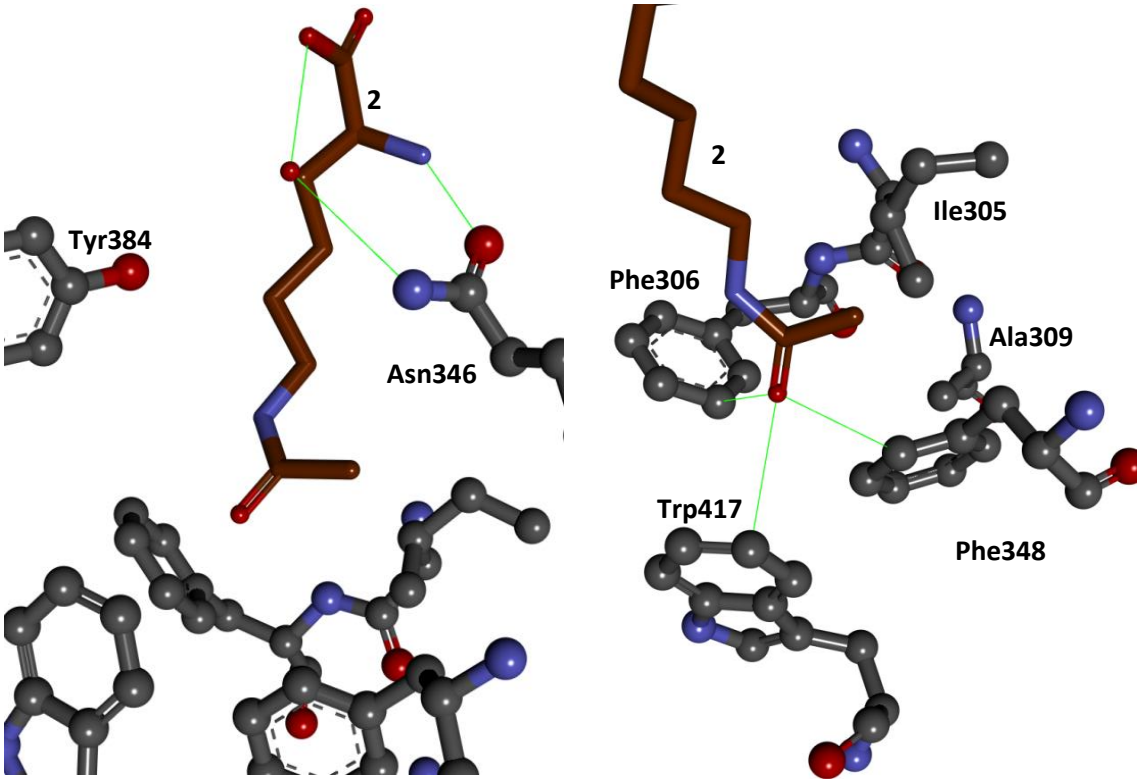


Figure 19: (left) In the pylRS mutant recognising **2**, hydrogen bonds with Asn346 still hold the substrate **2** in place but with a different geometry and no interactions between **2** and Tyr384; (right) the shrunken binding pocket is capped by three aromatic rings, two installed by mutations, all of which appear to orientated in such a way as to interact with the carbonyl oxygen of **2** (PDB: 4Q6G).

Given that acylation of histones frequently occurs at multiple lysine residues, the next logical development was the ability to incorporate more than one **2** residue into a protein. Through optimisation of the amber stop suppression protocol, homogenous samples of histone H3 acylated at up to four specific lysine residues could be produced in quantities sufficient for mononucleosome preparation and assaying.[139] Further development of a highly optimised cassette facilitated the production of histone H3 with as many as six acetylated lysine residues in mammalian cells, elucidating the effect of histone acylation in gene expression.[140]

However, acylation is not the only commonplace post-translational modification for histone proteins: alkylation is also often observed. Attempts to incorporate

monomethylated lysine **3** through the use of rationally guided directed evolution strategies failed to reproduce the success observed with **2**:[141] the methyl group is simply too small to sufficiently occupy the hydrophobic pocket, whilst discrimination between lysine and **3** also becomes challenging.

An astute strategy proposed a combination of traditional deprotection chemistry with amber stop codon suppression to solve this problem. Incorporation of **10** was known,[123] as was its lability under even mildly acidic conditions, so perhaps a methylated derivative **14** could still be deprotected but affording a methylated lysine residue rather than native lysine. Indeed **14** was found to remain a suitable substrate for the wild-type *M. barkeri* pylRS, with the extra methyl group seemingly causing little perturbation of substrate recognition with the result of full-length histone H3 containing **14** at position K9. Deprotection under conditions reported as mild, 2% TFA (Figure 20), completely cleaved the Boc group within four hours to afford the desired protein, H3K9Me1,[141] with methylation demonstrated not only by MS but also by the binding of heterochromatin protein 1, a protein specific for the full-length monomethylated histone. This work was subsequently expanded to produce histone H3 containing four **3** residues through the same general strategy, albeit making use of even more forcing conditions (50-60% TFA).[142]

An alternative strategy made use of a different protecting group: the Alloc-containing **15**, given that **11** was also a known substrate for pylRS and given the tolerance exhibited for the additional methyl group in amber stop codon suppression (Figure 20).[122] The primary advantage of this strategy is the less harsh manner in which the Alloc group can be removed, requiring a ruthenium complex in place of strong acid, rendering this strategy somewhat more biocompatible.[143] The introduction of **3** into proteins remains in essence a post-translational modification; now the modification is not an enzymatic methylation, but a chemical deprotection.

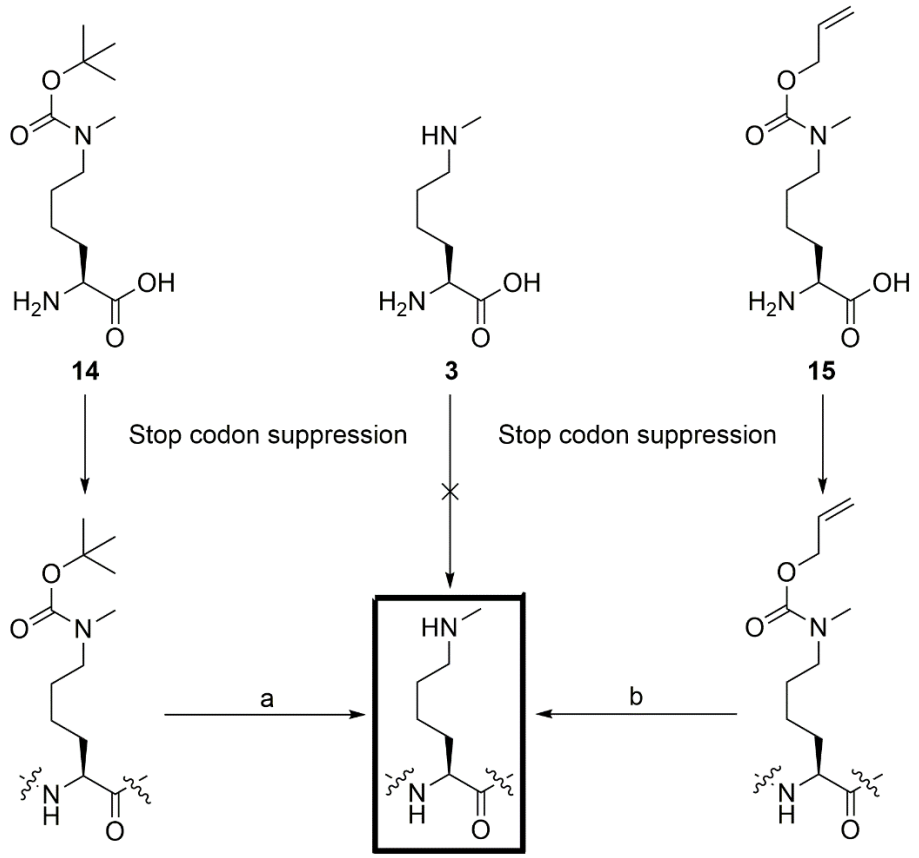


Figure 20: Genetic incorporation of **3** can only proceed indirectly *via* initial incorporation of **14** or **15** followed by chemical deprotection. a: 2% (v/v) TFA, 4 h, 37 °C. b: PhSH, RuCl(cod)Cp*, 3 h, 37 °C.

An even greater challenge is the addition of dimethylated lysine **4** to the genetic code. Discrimination between this amino acid and native lysine remains an issue, worsened by the inability to add a further protecting group to the ε-amino group due to the limited valency of nitrogen and the instability of suitable quaternary ammonium salts. The first strategy to produce proteins containing **4** more resembled a classical synthetic procedure, relying heavily on protecting group manipulation, than a recombinant method. Once again **10** was reliably incorporated into the protein, followed by chemical protection of all 12 other lysine residues in histone H3. Upon deprotection of the Boc group in 60% TFA, the single exposed lysine can undergo reductive methylation twice to form the desired dimethylated lysine residue, followed by global deprotection (Figure 21).[144]

Whilst certainly an impressive effort, the multiple harsh chemical protection and deprotection steps limit the effectiveness and scope of a methodology which ought to be broadly applicable and efficient. Later work sought to remedy these deficiencies by designing a simpler and less harsh route for the installation of **4**. This time only a single deprotection step was required: reductive decomposition of **16**, using an analogue of the protecting group seen with **13**, exposed an enamine on the residue side-chain. Upon spontaneous hydrolysis, the resulting aldehyde underwent reductive amination to form the desired dimethylated lysine residue (Figure 21).[145] With the exception of the lengthy

synthesis required to obtain **13**, this new methodology offered a far simpler route to proteins containing **4** merely through reversing the location of the required aldehyde and amine.

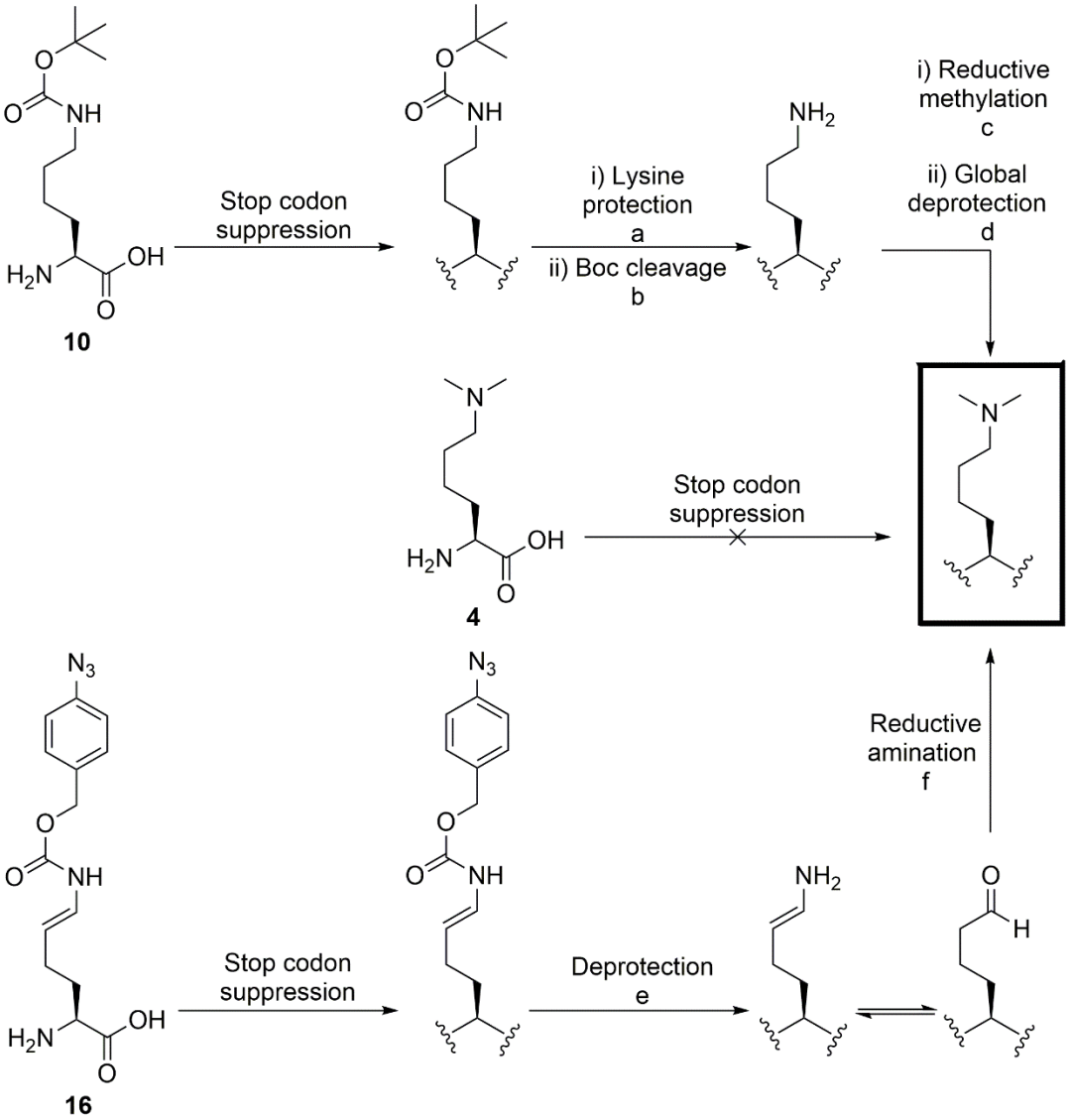


Figure 21: Protection/deprotection and alkylation strategies required to circumvent the inherent difficulties encountered when seeking to incorporate **4** using stop codon suppression. a: Cbz-OSu, DIPEA, DMSO. b: TFA, H_2O. c: CH_2O, DMAB. d: TfOH, DMS, TFA. e: TCEP, PBS, pH 7. f: $NaBH_3CN$, dimethylamine, PBS, pH 7.

Other various lysine acylations have been observed in the proteome and attempts to mimic these post-translational modifications using amber stop codon suppression have been met with some success. Short fatty acyl lysines **17, 18** and **19** (Figure 22) were recognised by engineered *M. barkeri* pylRS mutants[146, 147] and even wild type *M. mazei* pylRS,[148, 149] with yields eventually sufficient to allow the assembly and isolation of homogeneous histone H4 octamer samples. Two further lysine post-translational modifications, both found in endogenous histones, were also introduced into recombinant histones using NCAAs **20**[150] and **21,**[151] with an engineered pylRS variant

recognising not only the latter derivative but also analogues **22** and **23**, likely due to the inability of pylRS to discriminate between such comparatively small differences at the isobutyl group.

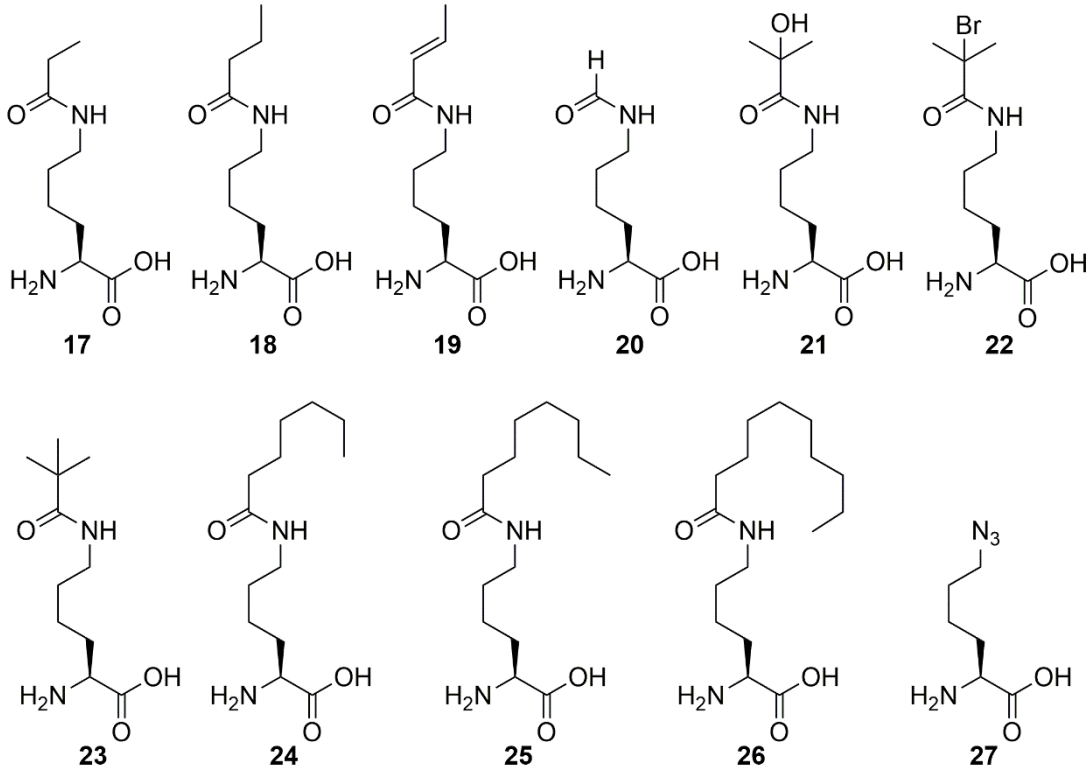


Figure 22: Acyl lysine derivatives and precursors suitable for use in amber stop codon suppression.

The medium-length fatty acyl lysine **24** proved to be amenable to amber stop codon suppression as well. The standard directed evolution of *M. mazei* pylRS facilitated access to recombinant heptanoylated glucagon-like peptide 1, with the heptanoylation reportedly improving binding to human serum albumin and increasing the half-life of this therapeutic protein.[152] Long-chain fatty acyl lysine derivatives eventually established the limits of amber stop codon suppression in this direction due to solubility problems: whilst **25** was soluble at the required concentrations in aqueous media for directed evolution of *M. mazei* pylRS and subsequent incorporation into histones, the longer-chain **26** was simply too hydrophobic and insoluble to be of any utility.[153] Whilst more labour-intensive, a more biomimetic strategy sought to use a genetically encoded reactive lysine derivative **27** which could undergo post-translational chemical modification, with the latter step allowing a greater range of possible acylations due to the lack of enzyme specificity limitations. Through the use of a traceless Staudinger ligation between the azide and a phosphinothioester bearing the desired acyl group, acetylation and novel succinylation of test proteins and histone H3 was observed.[154] A limitation of this method is the competition between Staudinger reduction and Staudinger ligation, producing an approximately 1:1 mixture of acylated:non-acylated protein.

The dovetailing of genetic incorporation of traditional synthetic chemistry protecting groups with bioorthogonal deprotection procedures has not only benefitted the study of histone proteins. An evolution in protecting strategies was adapted to the protein modification paradigm: as lysine derivatives **12**[123] and **28**[131] were an established part of the expanded genetic code, more mild deprotection of such lysine residues using palladium/H_2 and long-wave UV irradiation respectively was feasible in place of highly acidic conditions (Figure 23). Almost concomitantly, two different groups jumped upon this idea: whilst on-protein hydrogenation of the *N*-methyl derivative of **12** failed due to protein aggregation and poor yield,[155] both groups were able to photodecage the *N*-methyl derivative of **28** successfully. Later use of palladium, this time as a complex rather than elemental, fully showcased its excellent potential as a reagent in bioorthogonal deprotection with **11** and **29**.[156] The latter protecting group was shown to be easier to cleave in aqueous conditions than the former using two readily available, biologically compatible palladium complexes, culminating in the metal-mediated activation of a bacterial toxin through this palladium decaging of a key lysine residue.

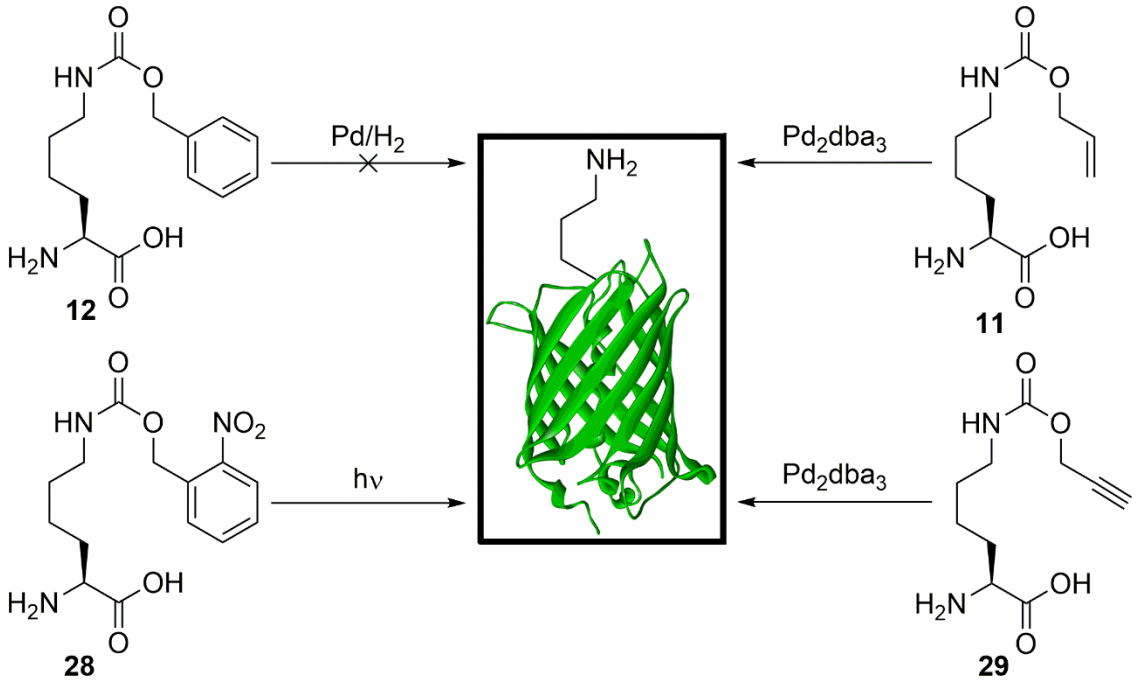


Figure 23: Protecting groups on lysine surrogates facilitates lysine activation using biocompatible deprotection chemistry.

1.2.4 Genetically encoded bioorthogonal chemistry

Click chemistry has quickly become the first thought at every mention of the phrase “bioorthogonal chemistry”: versatility, biocompatibility, coupling partner stability and comparatively rapid kinetics have all contributed to this well-earned reputation. An early use of click chemistry for protein modification saw a viral coat protein non-selectively

modified with an azido-acyl linker followed by ligation with an alkynyl fluorescein and an alkynyl dansyl-BSA bioconjugate.[157] Whilst the utility of the copper-catalysed azide-alkyne cycloaddition (CuAAC) had indeed been demonstrated, this work also revealed a limitation of this chemistry: the uncontrollable off-target labelling, as installation methods for both coupling partners were severely lacking in site specificity. Amber stop codon suppression has proven to be a near-perfect solution to this problem and its synergistic relationship with CuAAC, together offering unparalleled site specificity of coupling partner incorporation and therefore reactivity, has facilitated an unfettered infatuation with CuAAC amongst the chemical biology research community. Interest quickly honed in on suitable lysine derivatives such as alkynes **29-31** and azides **32** and **33** (Figure 24), generating newly CuAAC-modified bioconjugates such as proteins bearing fluorescent, affinity or spectroscopy tags;[158-161] dually labelled proteins for FRET studies both *in vitro*[162] and *in vivo;*[163] SUMOylated proteins;[164] and the site-specific furnishing of active hepatitis D virus surface proteins with purification tags whilst retaining infectivity of the viral particle.[165]

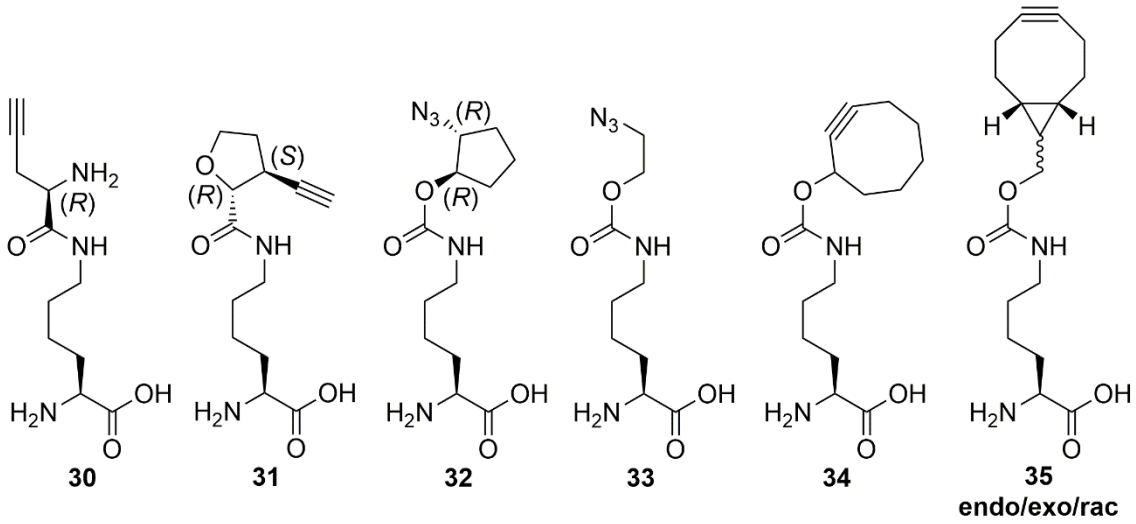


Figure 24: Pyrrolysine surrogates used in genetically encoded copper-catalysed and strain-promoted azide-alkyne cycloaddition bioconjugations.

Even as the field of bioorthogonal chemistry advanced, amber stop codon suppression maintained pace. CuAAC was superseded by the strain-promoted azide-alkyne cycloaddition (SPAAC), foregoing an often cytotoxic Cu^{I} catalyst in place of a more reactive alkyne[166] and hence marketed as “copper-free click chemistry” with the selling point of greater compatibility with *in vivo* applications.[167, 168] The shift from small linear alkynes to large, hydrophobic cyclooctynes was in fact well tolerated, with the $pylRS^{AF}$ mutant readily accepting both cyclooctyne **34**[169] and the more reactive bicyclononyne (BCN) **35**.[170] The lack of cytotoxic copper reagent required for SPAAC ligation, in conjunction with the other properties of azide-alkyne cycloaddition, led to a range of

applications of this site-specific protein bioconjugation strategy in challenging *in vivo* and therapeutic settings requiring strict homogeneity and minimal contamination: production of ADCs using mammalian[171] or insect expression systems,[172] where prokaryotic expression systems often fail; stapling together of antibody fragments to modulate biological activity;[173] PEGylation of adeno-associated viruses[174] and interferon,[175] increasing the stability and half-life respectively of the two therapeutics; and protein labelling with a ^{64}Cu-chelator to produce a positron-emitting antibody for localised positron emission tomography diagnostics.[176]

One of the shortcomings of SPAAC is the relatively slow kinetic profile of this ligation, significantly outpaced by CuAAC.[177] Whilst some tailoring of the cyclooctynes with electron-withdrawing substituents, such as difluorocyclooctyne compounds,[178] or with additional ring strain, such as dibenzocyclooctyne compounds,[179] significantly reduced this rate gap, the large structural perturbations required precluded the usage of such cyclooctyne derivatives in amber stop codon suppression, too bulky or polar to fit in even the most cavernous of hydrophobic binding pockets present among pylRS mutants. Another cycloaddition emerged to improve upon SPANC: strain-promoted inverse electron demand Diels-Alder cycloaddition (SPIEDAC), offering even faster kinetics than SPAAC or CuAAC[180] under metal-free conditions compatible with *in vivo* and therapeutic applications.[181, 182] Instead of an azide, a tetrazine acts as the diene to undergo a cycloaddition with a strained dienophile: not just BCNs, but also cyclopropenes, cyclobutenes, norbornenes and *trans*-cyclooctenes (TCOs). Just as with cyclooctynes, these three strained cycloalkenes were also highly amenable to use in amber stop codon suppression. Lysine derivatives **36**-**41** (Figure 25), as well as **35**, were rapidly developed to exploit SPIEDAC, exactly as predecessors SPANC and CuAAC had been, with TCOs such as **40** and **41** exhibiting the highest rates of SPIEDAC of all the dienophiles studied.[183] Some protein glycosylation has taken advantage of this genetically encoded SPIEDAC tool,[184] but the primary application of **36-40** has been fluorescent protein labelling, both *in vitro*[185-187] and on the surfaces of live prokaryotic[188, 189] and mammalian cells,[190] where the SPIEDAC properties of mildness, rapidity and fluorogenicity are highly germane. Notably **41** decomposes upon undergoing SPIEDAC to liberate the free amine group of the lysine backbone,[191] permitting lysine decaging in live cells observable over a time frame of mere minutes.[192]

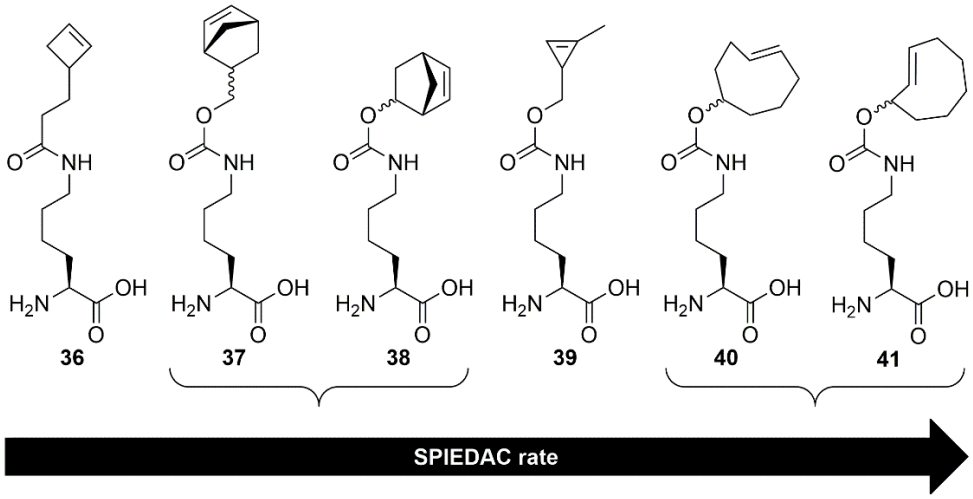


Figure 25: Lysine surrogates **36**-**41** developed for SPIEDAC bioconjugations in increasing rate order.[181, 183, 186, 188]

The ability to incorporate reactive azides, alkenes and alkynes into proteins using genetic code expansion has opened up a wide range of site-specific chemical bioconjugation strategies, where such incorporation had been the main bottleneck in the use of these strategies. Beyond click and Diels-Alder chemistry, a range of other bioorthogonal chemical modification strategies have been developed to take advantage of such reactivity enabled by amber stop codon suppression (Figure 26).

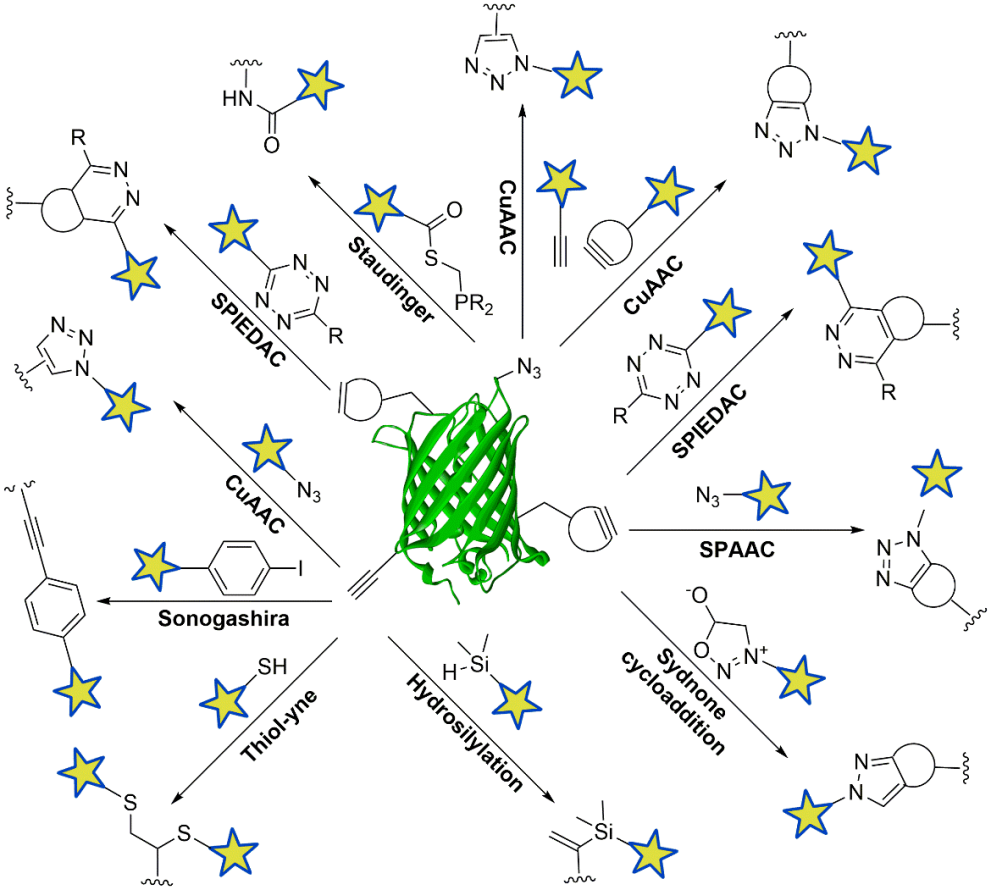


Figure 26: Bioorthogonal chemistry using genetically encoded functionality.[154, 193-196]

1.2.5 Diverging from Lysine to Other Amino Acid Analogues

Evidently a great range of lysine derivatives can be genetically encoded using the pylT/pylRS pair. Quickly it was realised that if pylRS mutants can recognise a diverse mélange of lysine derivatives, perhaps derivatives of other canonical amino acids, those with similar long alkyl or aryl groups, would also be suitable substrates. Through the mutation of six active site residues- L305, Y306, L309, N346, C348 and W417- a library of PylRS mutants were screened against canonical amino acid L-phenylalanine **42** and non-canonical derivative **43** (Figure 27). Only the mutations N346A and C348L were required to recognise **42** as a substrate, whilst uptake of **43** and analogue **44** required mutations to five out of the six key active site residues.[197] The utility of this discovery was quickly seized upon by using the aryl halide **44** as an on-protein handle for a Suzuki-Miyaura coupling with a dansylated boronic acid for fluorescent labelling. **44** had already enjoyed use both as a heavy atom source for protein crystal structure determination[198] and as a handle for Mizoroki-Heck, Sonogashira and Suzuki-Miyaura palladium-mediated cross-coupling bioconjugations both *in vitro*[199] and *in vivo*,[200] incorporated into proteins through genetic code expansion using a tyrosyl $tRNA_{CUA}$ from *Methanocaldococcus janaschii*[198] or a phenylalanyl $tRNA_{CUA}$ from *E. coli*.[201] At this point the *M. janaschii* tyrosyl $tRNA_{CUA}$ was renowned for its diverse established library of tyrosyl analogues used in genetic code expansion, including other reactive handles such as azides, alkynes and ketones.[202] Knowledge of this promiscuity whet the appetites of many keen to exploit the pyrrolysyl $tRNA_{CUA}$ system in a similar fashion.

Indeed a flurry of publications ensued. The pylRS mutant N346A C348A ($pylRS^{AA}$) was found to recognise unsaturated tyrosyl ethers **45**-**48** well and methyl ether **49** slightly;[203] further mutations were required for greater recognition of **49**.[204] Beyond *para*-substituted phenylalanyl/tyrosyl analogues, the $pylRS^{AA}$ mutant was found to encode an array of *meta*-substituted phenylalanine derivatives **50-60**, including the reactive handles of ketones, halides, azides and alkynes.[205] With these findings came the caveat that poorly recognised substrates such as **50** were outcompeted by canonical **42**, demonstrating that the promiscuity of the $pylRS^{AA}$ mutant was offset by a diminished level of orthogonality, analogous to mutations applied to Y384 (*vide supra)*.

Aldehyde **61** was also found to be a substrate for $pylRS^{AA}$, facilitating protein modification *via* oxime ligation.[91] This finding breathed fresh life into well-established protein carbonyl chemical modification, which had been hampered by the limitations on the positioning of the required reactive aldehydes. Previous methods to install protein aldehydes required an enzymatic tag, such as use of FGE or an exposed *N*-terminal serine, threonine or glycine residue; through **62**, site-specific protein aldehyde modification could now move beyond such sequence limitations with far greater freedom with an electronically distinct

aldehyde substrate.[77] Whilst pyrrolysine **1** and its demethylated analogue did find some use as protein aldehydes for modification,[206] the use of **61** was far more practical due to its stability and straightforward synthesis compared to the need to utilise the pyrrolysyl biosynthetic pathway in order to incorporate **1**.[132]

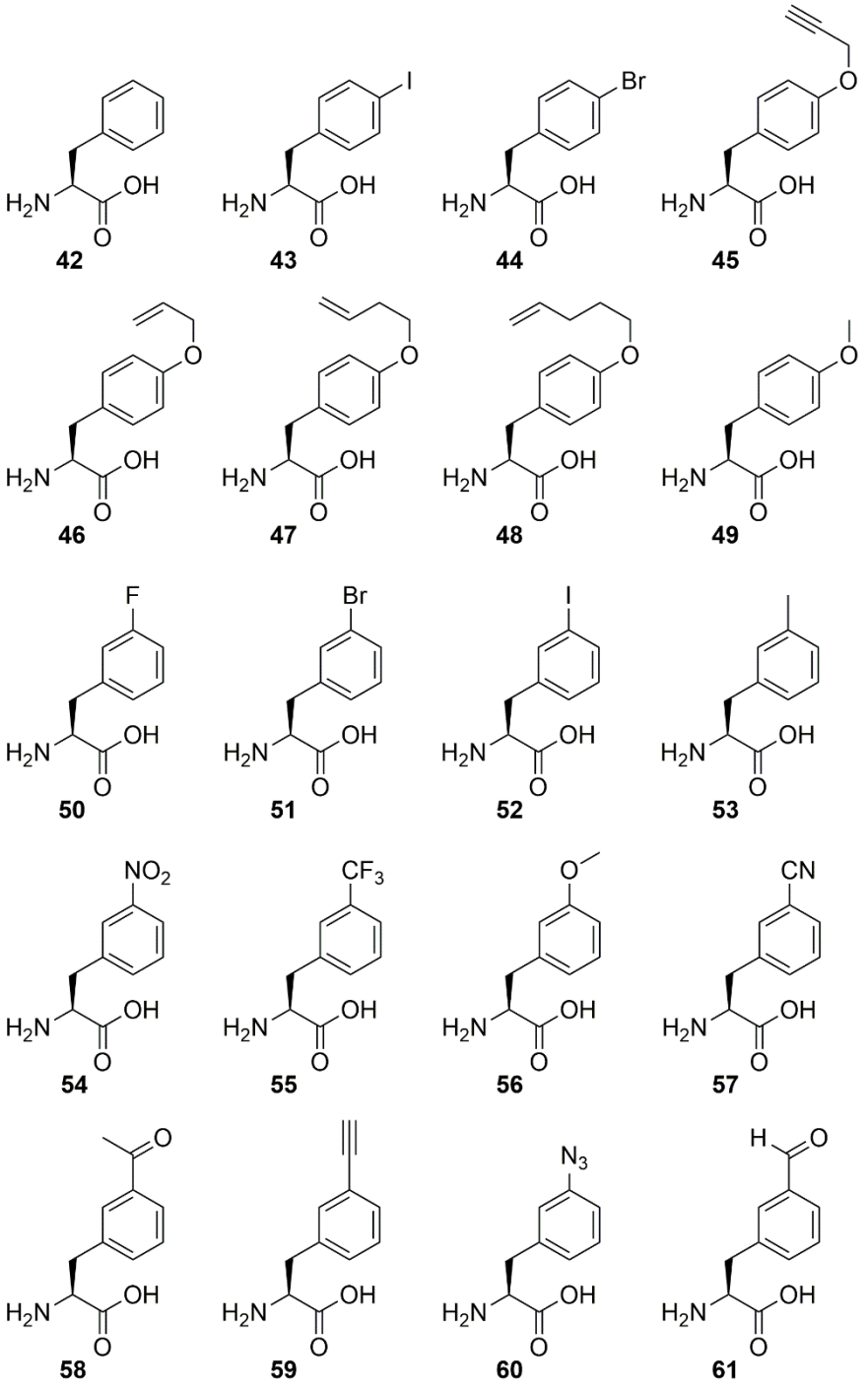


Figure 27: Genetically encoded phenylalanine and tyrosine analogues.

Further probing of the pylRS^{AA} mutant revealed that in addition to further *meta*-substituted hydrophobic phenylalanine derivatives **62-65**,[207] *ortho*-substituted compounds **66-71** were also recognised by pylRS^{AA} (Figure 28).[208] Given that hydrophobic ether substituents seemed broadly compatible with this mutant, SPAAC and

SPIEDAC substrates **72-74** were investigated as substrates for pylRS^{AA} but disappointingly no stop codon suppression was detected. Two further mutations were required: the same Y306A and Y384F mutations required, producing quadruple mutant pylRS^{AAAF} which proved to recognise **72**-**74** well and opened up an alternative route to genetically encode click chemistry.[209] The same mutant was also able to recognise fluorinated compounds **75**-**77**, with the fluorination pattern of **75** able to perturb and hence elucidate a key cation-pi interaction between a histone reader protein containing **75** and a histone containing a cationic trimethylysine residue.[210]

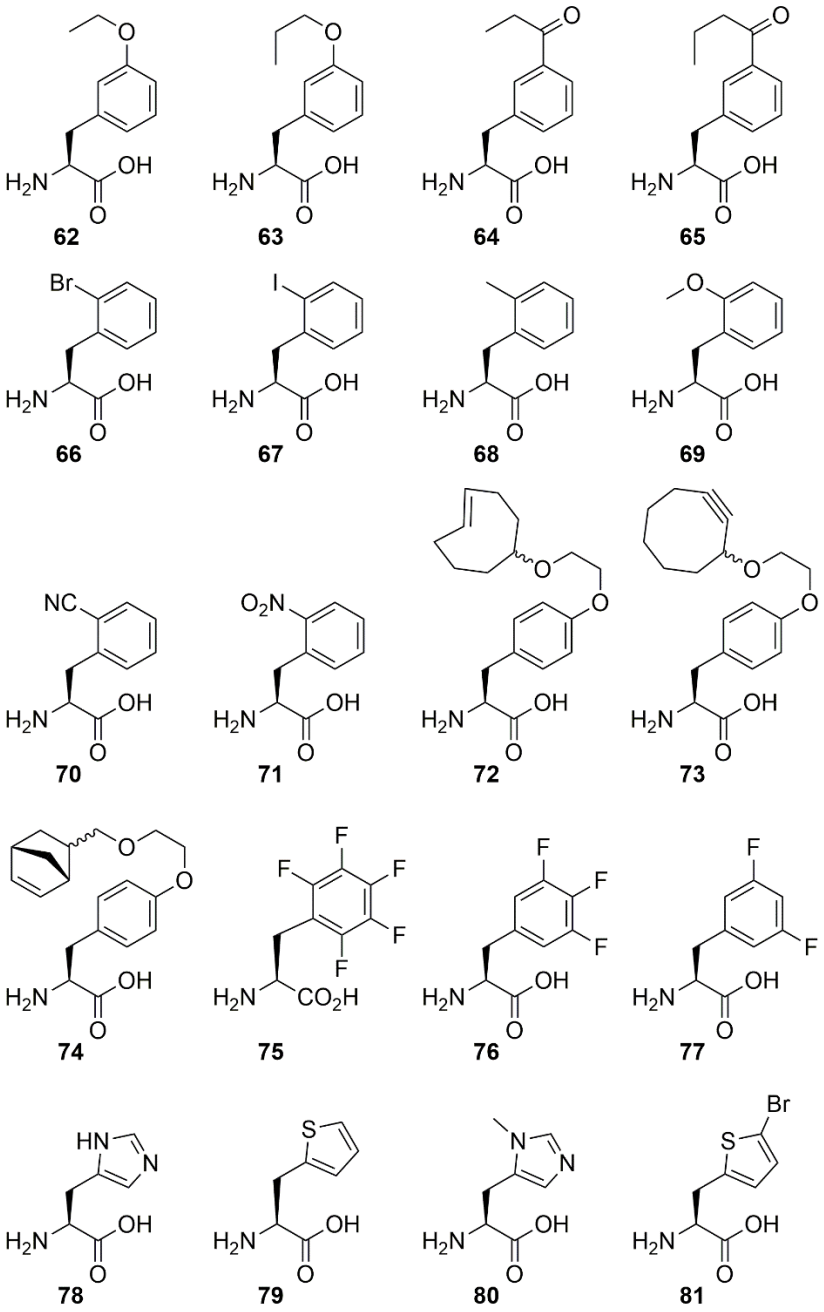


Figure 28: Further genetically encoded phenylalanine and histidine analogues.

Given that phenylalanyl and tyrosyl analogues were suitable for genetic incorporation, the next logical step turned towards analogues of another small, aromatic canonical amino acid: histidine **78**. Shrinking the pocket size and knocking out key hydrogen bonding interactions through respective Y306F and C348F key mutations led to recognition of histidine analogues **79**-**81**.[211] The pocket remained sufficiently hydrophobic to recognise only these hydrophobic histidine derivatives, whilst histidine **78** proved to be too polar to nestle inside thus retaining the orthogonality of this pylRS mutant to canonical amino acids. The size of the aromatic ring substituent had a clear effect on binding and hence TAG-containing protein yields: from the thiophene substituent of **79** to the *N*-methylimidazole ring of **80** yields double, whilst the additional bromine atom on **81** affords protein in three times the yield compared to **79**. Arguably the most useful analogue incorporated here is **80**, as *N*-methylation of histidine is a known post-translational modification. Substitution of **80** in place of **78** perturbs hydrogen bonding, prevents tautomerisation and can impede free rotation of the ring; archaeal *M. barkeri* methyl-coenzyme R reductase containing **80** is postulated to do so on account of this latter property, leading to greater binding of coenzyme B.[212] Subsequent use of **80** in amber stop codon suppression made use of these differing properties to alter the performance of a biotechnological enzyme. Ascorbate peroxidase containing **80** in place of a catalytic histidine residue was found to have a far greater turnover and mildly higher catalytic efficiency in the oxidation of guaiacol than the wild type enzyme, ascribed to disruption of active site hydrogen bonding and hyperconjugation stabilising radical intermediates.[213] This work neatly demonstrated the suitability of amber stop codon suppression in biotechnological applications where more chemical functionality is required than the basic canon of amino acids can provide.

Perhaps the most attractive canonical amino acid to a chemical biologist is cysteine **82** (Figure 29). A wealth of uses exists for this particular residue through the depth of known sulfur chemistry, sometimes to the extent of unwanted participation in side reactions. The ability to tame this wild beast is paramount to capitalising on its functionality, so genetic code expansion seemed a natural technique for this purpose. Other tRNA/RS pairs had been shown to successfully genetically encode cysteine derivatives,[119, 214] setting a high bar for success. Lysine-cysteine dipeptides **83*R*** and **83*S*** were the first breakthroughs, enabling incorporation of a 1,2-aminothiol at a location beyond the *N* terminus.[215] This moiety is highly useful due to its reactivity in native chemical ligation (NCL), with both isomers of **83** used to ubiquitinate proteins *via* NCL as ubiquitin is typically conjugated to lysine residues *via* an amide bond. In combination with a *C*-terminal thioester, this strategy was also used for protein cyclisation.[216] NCL was further exploited using **84** which, upon spontaneous cleavage of the nitrobenzyloxycarbonyl group after translation,

exposed the reactive 1,2-aminothiol for protein ubiquitination.[217] The leftover thiol was then removed by desulfurisation, pleasingly generating a native protein-ubiquitin linkage matching that found in nature.

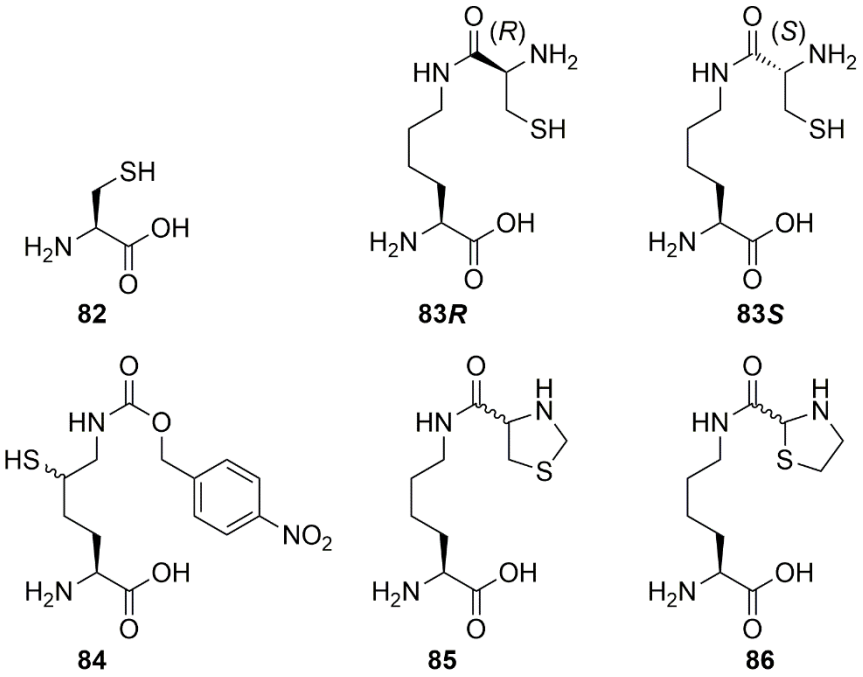


Figure 29: Cysteine and genetically encoded analogues.

Whilst the 1,2-aminothiol motif is indeed useful in protein modification on account of its reactivity, this property often becomes a hindrance. *N*-terminal cysteine residues, harbouring the 1,2-aminothiol motif, often trap electrophilic aldehydes and ketones such as metabolite pyruvate to form thiazolidine adducts,[218] with this undesired process affording a heterogeneous mixture of proteins with reduced reactivity. 4-substituted thiazolidine **85** was used to control this process, ensuring homogeneity in the protection of the 1,2-aminothiol group in this residue. The thiazolidine was then opened through the use of a stronger nucleophile, methoxyamine, to trap the electrophilic formaldehyde and liberate the 1,2-aminothiol.[219] This reactive residue was then used as a handle for cyanobenzothiazoline condensation ligation and, somewhat ironically, for thiazolidine ligation, where a further aldehyde bearing a useful substituent condensed again with the 1,2-aminothiol, reforming the stable thiazolidine linkage with the addition of added functionality.[220] The thiazolidine had demonstrated its utility as a protecting group for nucleophilic 1,2-aminothiols by electrophilic aldehydes; reversal of this arrangement would allow for the genetic incorporation of an aldehyde rather than a 1,2-aminothiol. 2-substituted thiazolidine **86** was designed with this exact strategy in mind, where treatment with a suitable electrophile exposed a highly reactive glyoxylamide group capable of undergoing oxime ligation.[221] The limitations of this methodology lay in the harsh deprotection conditions: incubation with silver acetate in 10% (v/v) acetic acid

certainly opened the thiazolidine, but these conditions are likely to denature the vast majority of protein samples.

A notable attribute of **83*R*/*S*** and **85** is the ease with which these compounds can be genetically encoded. Using the wild type pylRS from *M. barkeri*, all three compounds are successfully incorporated into proteins, albeit to a lesser extent than positive control **10**. This is perhaps somewhat obvious: the thiazolidine ring of **85** closely resembles the pyrroline ring of **1**, whilst the methylene thiol side chain of **83*R*/*S*** is seemingly sufficiently hydrophobic to fit in the binding pocket. However the wild type pylRS from *M. mazei* only recognises **83*S*** and neither of the other two close analogues.[219] The reasons behind this rather important difference in substrate recognition between the two variants, both commonly used and seen as almost interchangeable in this field, are largely unknown and here lies something of a gap in the literature. Both pylRS variants possess a conserved *C*-terminal catalytic domain with sequence differences only apparent in the *N*-terminal domain. Many crystal structures of the *M. mazei* pylRS variant have been obtained, often omitting the *N*-terminal domain yet still providing a helpful level of insight. Sadly, no crystal structures of the *M. barkeri* pylRS variant have been obtained to date, rendering as mere conjecture any attempted rationalisation of the differences in substrate recognition between the two variants.

Successful incorporation of cysteine so far relied upon the presence of a lysine backbone for recognition by pylRS. This seems logical given the small size of cysteine compared to lysine and the size of the hydrophobic pocket in the active site of pylRS. Genetic incorporation of cysteine derivatives without a lysine backbone would likely need some additional bulk in order to fit the cavernous binding site. The solution was again found in the use of protecting groups. Compound **87**, bearing the photolabile *o*-nitrobenzyl protecting group previously seen in **28**, offered the ability to genetically encode a native cysteine residue exposed by photolysis (Figure 30).[222] However biological use of the *o*-nitrobenzyl protecting group is limited by the deprotection conditions required, as only shortwave UV irradiation affords complete deprotection at the expense of damaging the protein or other biological structures. An alternative protecting group was deployed and **88** was shown to be more suitable, with photolysis requiring only minutes of irradiation with less harmful longwave UV. This photodecaging procedure was shown to be amenable to both *in vitro* and *in vivo* settings, restoring activity to catalytic cysteine residues cages as **88** in proteases and luciferases.[223]

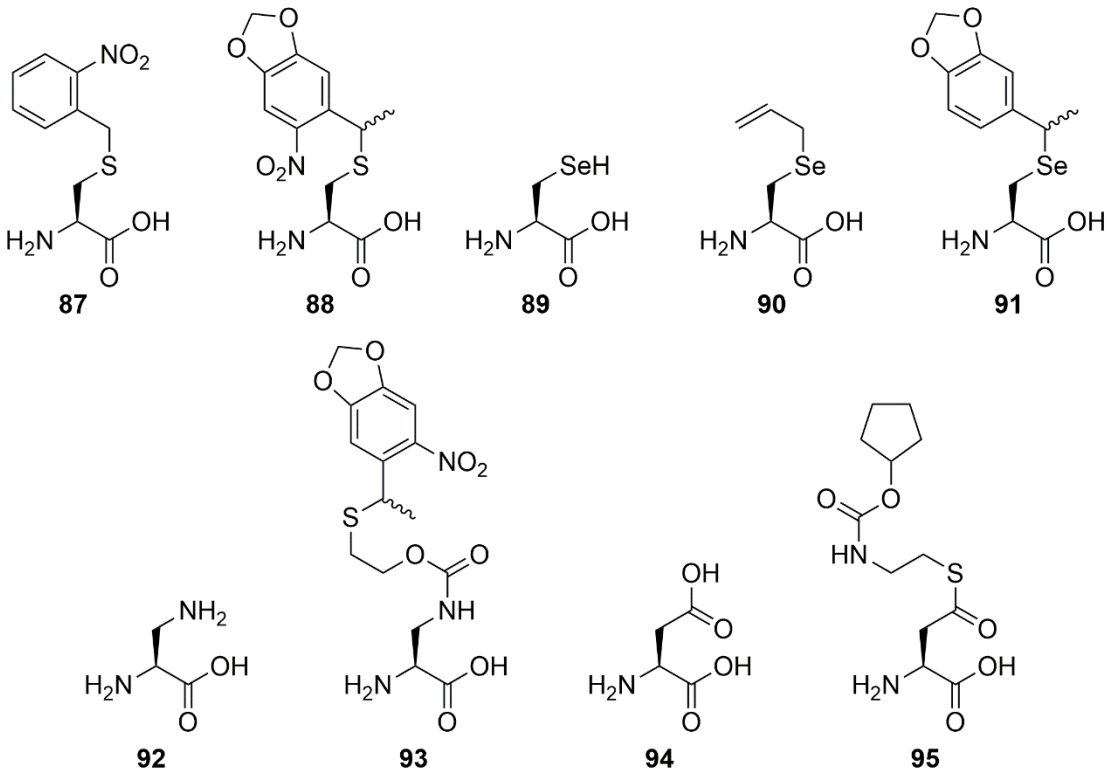


Figure 30: Further cysteine, selenocysteine and aspartate derivatives used in amber stop codon suppression.

One of the rarest canonical amino acids is cysteine analogue selenocysteine **89**. Given the successful incorporation of cysteine derivatives using amber stop codon suppression, selenocysteine derivatives would be a logical next step. Two similar strategies emerged concomitantly, both making use of protecting groups so that the NCAA is sufficiently bulky to be recognised by a pylRS mutant which retains orthogonality to smaller canonical amino acids. The first method made use of the allyl protecting group, seen previously with **11**, given the mild conditions required for deprotection.[224] Engineering of a pylRS mutant from *M. barkeri* led to the incorporation of **90** in native selenoproteins, with activity fully restored following overnight incubation with a palladium complex.[225] A second effort utilised photocaged derivative **91**, similarly affording active selenoproteins following longwave UV irradiation over a few minutes.[226] A further cysteine analogue is **92**, bearing an amino group in place of the thiol and useful as a tool to probe the mechanistic roles played by active site cysteine or serine residues. A quadruple mutant *M. barkeri* pylRS was suitable for incorporating **93** in place of a catalytic cysteine residue in a protease, exposing **92** following mild photolysis. This led to trapping of a covalent intermediate, where a stable amide product was formed in place of the regular transient thioester intermediate. Further exploitation led to the capture and crystallisation of analogous covalent intermediates in the biosynthesis of cyclic peptides, providing exceptional insight into the mechanistic pathways underpinning this process.[227] A final example demonstrated that even derivatives of aspartate **94** could be genetically

encoded. Compound **95**, bearing the cyclopentyloxycarbonyl group to aid recognition by pylRS, was incorporated into proteins as a handle for NCL, where the ability to generate protein thioesters had previously been restricted to the *C* terminus.[228] Complementary to the use of **84**, following NCL and desulfurisation the modification would resemble that of a native asparagine modification, showing clear further utility in the mimicry of such PTMS. Unfortunately, the thioester was found to possess some off-target reactivity, being trapped by reactive thiols such as glutathione or being hydrolysed to revert to canonical **94**, meaning that attempts to perform NCL on proteins containing **95** afforded heterogenous protein samples bearing varying modifications at the location of the amber stop codon.

1.2.6 Conclusions

The field of chemical biology has been truly enriched by the emergence and success of the pyrrolysyl tRNA/RS pair. Protein modification has been a major beneficiary, with the expanded amino acid alphabet providing a range of new reactive handles for site-specific bioorthogonal chemistry, translating reactions into the protein domain both *in vitro* and *in vivo*. This has enabled the generation of scores of bioconjugates bearing remarkable new functionality such as fluorescence, post-translational modifications, and cytotoxicity through linked drugs. Over 100 NCAAs stand in the repertoire of this system, permitting interrogation of biological systems impervious to previous human intervention to expand the understanding and development of modern therapeutics. Excitingly, new pyrrolysyl tRNA synthetases have been discovered in other methanogenic species of archaea,[229] offering further understanding into the promiscuous tRNA synthetase and opportunities to enhance the utility of this system.[230] The development of the pylT/RS pair has armed chemical biologists with an expansive array of tools which will continue to reward astute, creative utilisation.

1.3 Project Outline

The project was an investigation into the development of new strategies for chemical modification of proteins. The success of the field depends on the presence of a well-stocked toolbox of chemical methods for linking small molecules to proteins, with every tool possessing its own niches and drawbacks, and it is to this collection that this work sought to add.

This project focussed on the introduction of non-canonical reactivity into proteins through the installation of aldehydes. The glyoxyl aldehyde formed at the *N* terminal seryl, threonyl or glycyl residue of a protein has seen a range of uses in chemical protein modification, owing to the site specificity and general reliability of the chemistry developed for this purpose. [77, 231] However, this functionality is currently restricted to the *N*-terminus of the protein which may be unavailable for a variety of reasons, such as acylation, active site involvement, or burial within the protein. Furthermore, only one aldehyde is available for this purpose, a problem in the design of antibody-drug conjugates when often multiple drugs are attached to the same antibody scaffold.[56]

Amber stop codon suppression has quickly found its feet as a versatile method for the introduction of non-canonical functionality into proteins.[92] Use of the pyrrolysyl tRNA-RS pair has proven to be a highly suitable platform for this purpose, as the previous section indicates. With sufficient optimisation, multiple NCAAs can be genetically encoded into a protein.[147] As each NCAA represents only a single addition or substitution within the primary sequence, the structural perturbation is minimal and thus the positioning of each NCAA is highly flexible. The pyrrolysyl tRNA-RS appeared to be the perfect partner for the glyoxyl aldehyde to enable site-specific aldehyde modification beyond the protein *N* terminus.

The first objective was to design a glyoxyl aldehyde derivative suitable for genetic incorporation using the pyrrolysyl tRNA-RS pair. Due to the reactivity of the glyoxyl aldehyde, a “caged” analogue was considered in order to survive the aqueous milieu during protein biosynthesis. This analogue must be a suitable substrate for a pyrrolysyl tRNA synthetase, or genetic incorporation would not succeed. Hence the aldehyde derivative would likely need to contain a lysine backbone in order to sufficiently mimic pyrrolsyine, with the caged aldehyde moiety similarly not too distinct from the native tRNA synthetase substrate.

The second objective was to develop a suitable method for decaging the aldehyde on the protein, essentially combining traditional deprotection group chemistry and functional group conversions with the demands of biological chemistry. Ultimately the protein must “survive” the decaging process, retaining fold and function, but the decaging process

also ought to be rapid and lead to maximum conversion to produce a homogenous construct. Furthermore, the decaging chemistry must also work in aqueous milieu at a suitable pH for the protein without any side reactions with the native functionality of the protein.

The third and final objective was to demonstrate the functionality of the decaged glyoxyl aldehyde as a handle for chemical protein modification beyond the *N* terminus. The exposed aldehyde must be able to perform the repertoire of modification chemistry already showcased at the *N* terminus if it is to be of any use. To achieve this would be a symbiotic process: demonstrating the utility of the internal glyoxyl aldehyde and adding a new perspective to the established aldehyde modification methods, liberated from *N*-terminal shackles and free to roam more extensively about the protein domain.

Chapter 2: Synthesis of Peptides as Aldehyde Precursors & Reactive Probes

2.1 Introduction

The promiscuity of the pylT-S pair has been shown to extend to lysine dipeptides such as **83*R*** and **83*S***, with the utility of these canonical amino acid dipeptides demonstrated through the use of the additional cysteine residue as a handle for native chemical ligation.[215] This reaction requires the presence of an α-amino group, which is retained through the "pseudo-*N*-terminal" or "internal" (i.e. not at a terminal location) positioning of the cysteine residues on the lysine backbone. In this way, it was posited that native chemical ligation could now be used for protein functionalisation at a broad range of positions rather than just the N-terminus; theoretically, the genetically encoded cysteine derivative need only be at an accessible surface position for successful modification. In further developments to this system, the corresponding thiazolidine derivatives **85**[220] and **86**[221] were found to be similarly acceptable substrates for various wild-type pylT-S pairs and acted as "caged" cysteine handles for bioorthogonal modification, circumventing problematic trapping of the 1,2-aminothiol group by electrophilic carbonyl compounds such as pyruvate, omnipresent in the aqueous milieu of biological systems.[219] The limits of the pylT-S promiscuity, however, have arguably not yet been fully reached or pushed; attractive opportunities to incorporate further lysine dipeptides with "caged" reactivity for bioorthogonal purposes lie unexplored. Furthermore, various small molecules are required in order to perform the subsequent chemical modification of the exposed aldehyde and demonstrate its utility. In this chapter, the underpinning rationale and synthesis of such target compounds will be outlined.

2.1.1 Chapter overview

This chapter is composed of two main areas. Firstly, this chapter discusses the small molecules prepared using solution-phase synthesis: analogues of pyrrolysine as candidates for non-canonical amino acid mutagenesis with potential to be transformed into reactive aldehyde species. Secondly, this chapter discusses the larger peptide probes synthesised using solid-phase methods, combining reactivity towards a protein aldehyde with a useful label. This can be considered as the starting points of working forwards following a retrosynthesis of a protein bioconjugate (Figure 31): disconnecting between the protein and label leads to a reactive peptide probe and a protein aldehyde. The protein aldehyde can be obtained through a functional group interconversion from an aldehyde precursor residue, a genetically encoded pyrrolysine analogue.

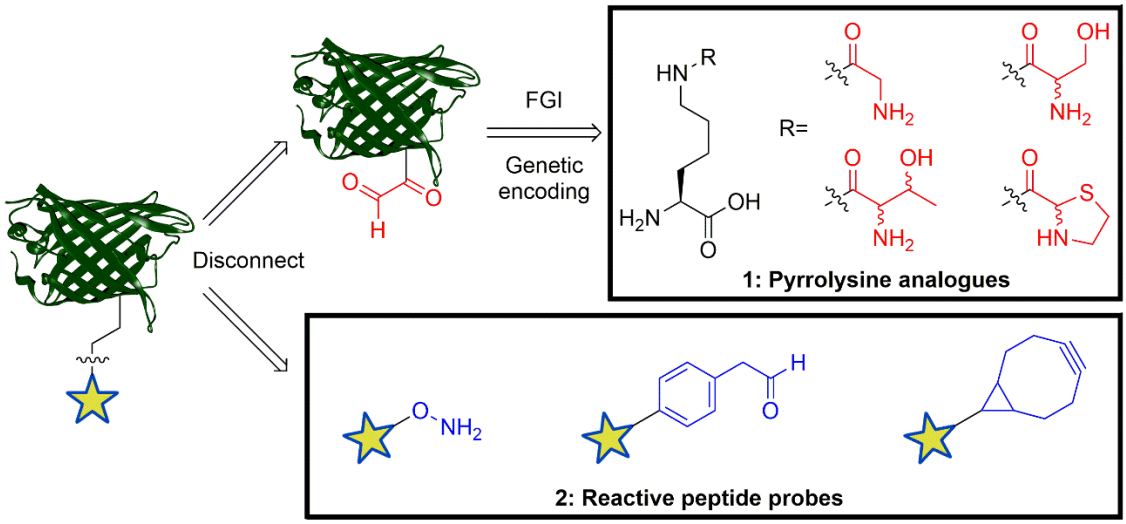


Figure 31: Retrosynthetic analysis of a protein bioconjugate, where disconnection affords 1) a protein aldehyde, prepared *via* amber stop codon suppression and functional group interconversion using small molecule pyrrolysine analogues, prepared in in the first half of this chapter, and 2) reactive peptide probes, harbouring both a useful label, denoted by a star, and functionality to react with the protein aldehyde, with the second half of this chapter focussing on the modular solid-phase synthesis used to prepare these probes.

2.2 Canonical amino acid dipeptides

Cysteine is not, however, the only canonical amino acid which would be of use when appended to the ε-amino group of a lysine residue. The pseudo-*N*-terminal location of an amino acid in such a location would present the α-amino group for reactivity; a useful property, given that the α-amino group of certain canonical amino acids is mechanistically vital for the transformation of these residues into reactive aldehyde species to be used in chemical modification strategies.

2.2.1 Transamination peptide

N-terminal glycine residues are well-known to undergo transamination in a biomimetic fashion using PLP.[76] Mechanistically, this reaction makes use of PLP as an amine sink to acquire the α-amino group from the residue in question, forming pyridoxamine-5-phosphate and the corresponding carbonyl (Figure 32). For all canonical amino acids with a β-carbon atom, a ketone is formed in this process, whilst glycine stands alone in forming a more reactive aldehyde: a glyoxyl aldehyde, with sufficient electrophilic character arising from the 1,2-dicarbonyl system to reliably undergo many carbonyl reactions with excellent conversions on reasonably short timescales.[77] Through both the well-established nature of glycine transamination[232] and the reactivity of the product, this procedure has led to the design and use of numerous bioconjugation methodologies;

myoglobin, being both readily available and harbouring an *N*-terminal glycine residue, has been superb fodder for these purposes.[233, 234]

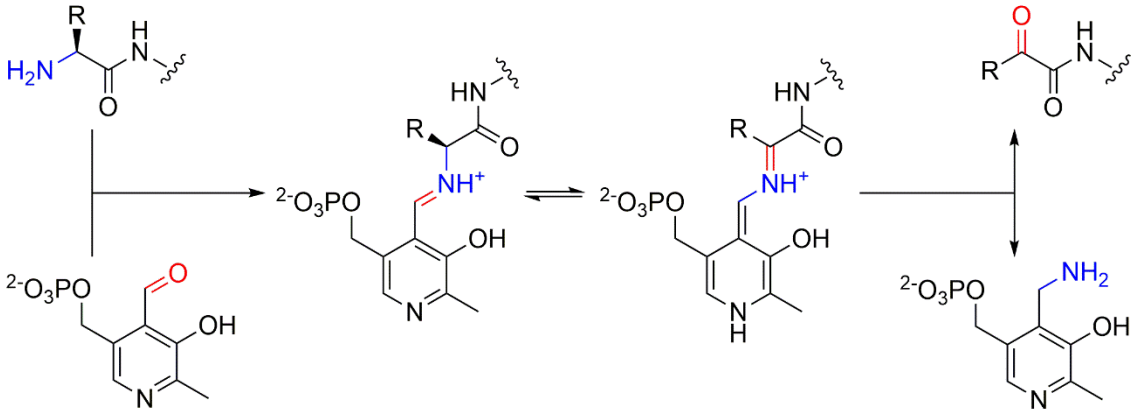


Figure 32: Transamination of amino acids using pyridoxal-5-phosphate.

It is the availability of the α-amino group which predisposes an *N*-terminal glycine residue to be transaminated. In this way, a pseudo-*N*-terminal glycine residue would retain this important moiety without the restriction of being located exclusively at the *N* terminus of a protein. Dipeptide **96** (Figure 33) putatively combines this aldehyde precursor with a lysine backbone, smuggling "caged" aldehyde reactivity into a protein sequence through the use of amber stop codon suppression.

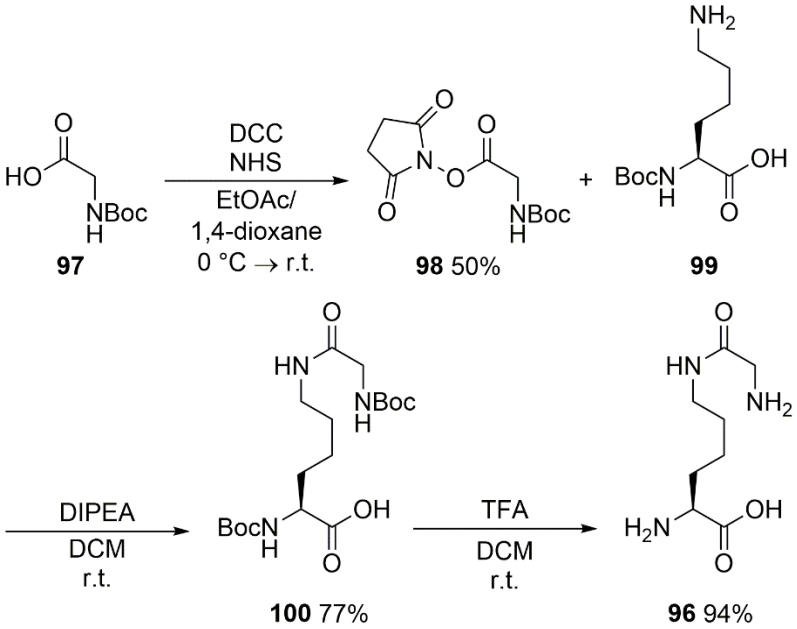


Figure 33: Synthesis of lysine-glycine dipeptide **96**.

The synthesis of **96** was achieved through an activation-coupling-deprotection strategy over three steps, following the route initially designed for and used by Annika Hüsken.[235] Protected glycine **97** was activated as the NHS ester **98** ahead of coupling to partially protected lysine **99** to afford dipeptide **100**. Deprotection and lyophilisation afforded **96**

in a final yield of 36% over three steps. This strategy was preferred to an *in situ* activation route due to the additional *C*-protection which would be required on the lysine partner, with this reagent being in short supply. The NHS activation step offers the least pleasing yield, with optimisation likely being required for further iterations. Nevertheless, through a straightforward synthetic route, **96** was obtained in sufficient quantity and of the required purity for later biological studies.

2.2.2 Periodate-oxidised peptides

Transamination of *N*-terminal glycine residues is not, however, the exclusive route to protein glyoxylamides. *N*-terminal serine and threonine residues can also be transformed into the reactive glyoxylamide through a simple oxidation using sodium metaperiodate.[78] This procedure requires vicinal nucleophiles, such as 1,2-diols and 1,2-aminoalcohols, which initially attack the electrophilic iodine(VII) centre. The resulting cyclic periodate ester decomposes to afford an iodine(V) salt and two aldehydes, of which one will be a glyoxyl species when serine and threonine are used (Figure 34). Whilst this reaction is generally fast (requiring only a few minutes to reach completion), periodate will also undergo various competing, slower side reactions with protein systems, such as oxidation of electron-rich aryl or sulfur(II)-containing residues,[236] necessitating careful control of reaction time, stoichiometry, and quenching.

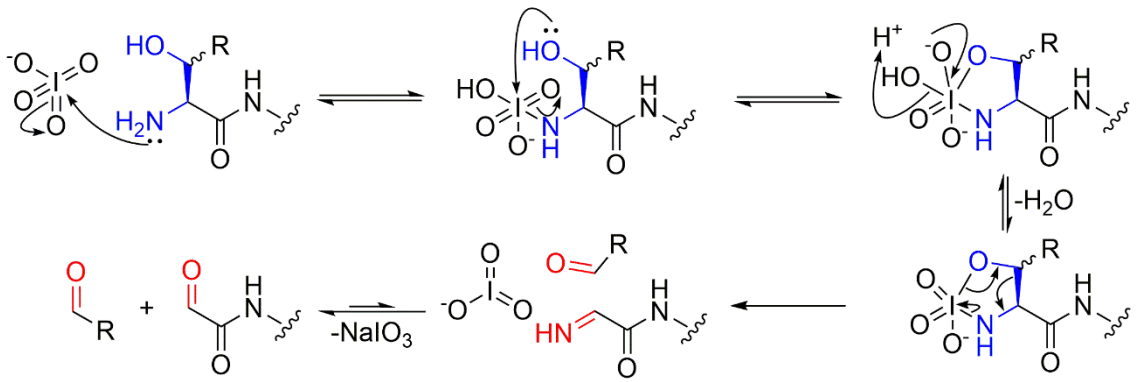


Figure 34: Sodium periodate oxidation of *N*-terminal serine and threonine residues.

Similar to transamination of *N*-terminal glycine residues, periodate oxidation requires a free α-amino group, in addition to a vicinal hydroxyl group; consequently, in native systems, only serine and threonine residues at a protein *N*-terminus undergo this oxidative cleavage reaction. As an alternative strategy to generate a protein glyoxylamide species, a putative route in which the periodate-mediated cleavage of a genetically encoded lysine dipeptide harbouring a 1,2-aminoalcohol motif, arising from a serine or threonine residue, was proposed.

In order to maximise the chances of discovering a suitable substrate for amber stop codon suppression, four dipeptides were synthesised, differing in the absence/presence of a β-methyl group (serine/threonine respectively) and the α-configuration (*S*/*R*, "natural" vs. "unnatural" respectively). For threonine derivatives, the β-configuration was unaltered and allothreonine derivatives not used on account of difficulties in obtaining the relevant starting materials. The modulation of α-configuration and β-methyl groups was considered to offer insight into the amber stop codon suppression process, with these changes subtly altering amino acid polarity, steric bulk, and positioning within the pylS active site.

An analogous activation-coupling-deprotection strategy was used to synthesise all four dipeptides.[235] First, the two "naturally" configured (2*S*) serine and threonine peptides were synthesised, with activation as the NHS esters followed by coupling to form protected dipeptides. Deprotection required the use of nucleophilic scavengers, namely water and triethylsilane, in order to intercept reversible *O*-alkylation commonly encountered during acidic deprotection of *t*-butyl ether groups.[237] Over all three steps, comparable cumulative yields of 39% and 32% of **101** *via* **102-104** and **105** *via* **106-108** were achieved (Figure 35), again in sufficient quantity and of sufficient purity for further work.

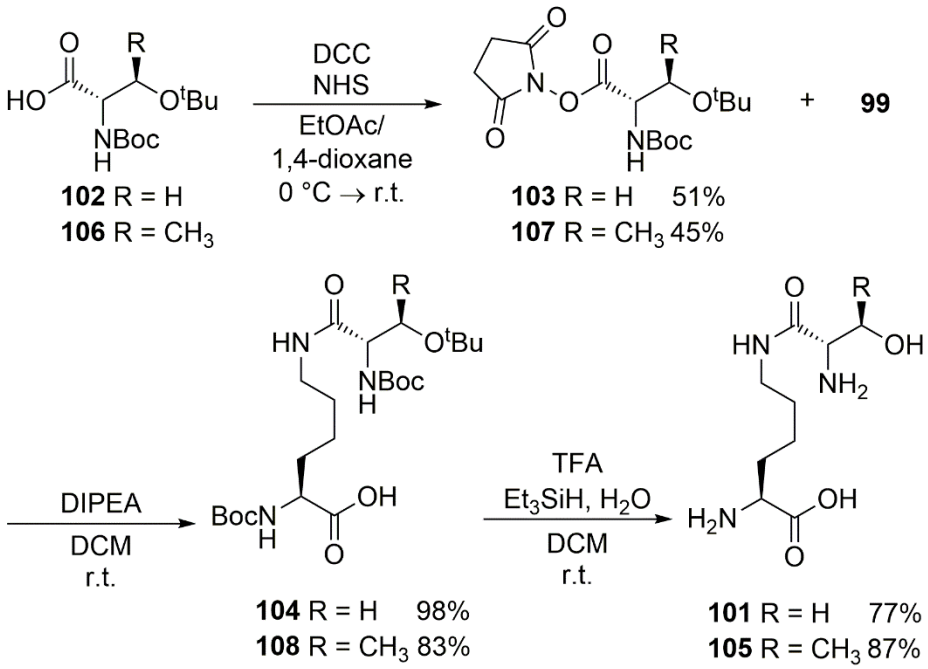


Figure 35: Synthesis of lysine-serine and lysine-threonine (L-configured) dipeptides.

The two "unnaturally" configured (2*R*) serine and threonine peptides **109** and **110** respectively were then synthesised in largely the same manner *via* **111**-**115** and **114**-**116** (Figure 36). An increase in the excess quantities of NHS and DCC raised the yields of NHS esters, followed by standard dipeptide coupling to form protected dipeptides.

Following deprotection and lyophilisation, dipeptides were obtained in improved cumulative yields of 50% and 56% respectively.

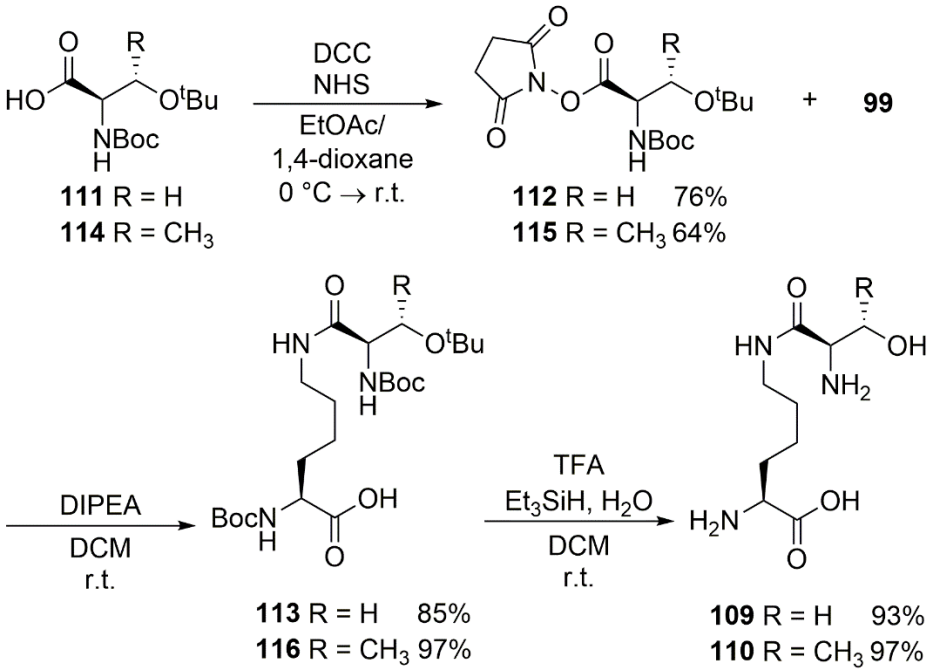


Figure 36: Synthesis of lysine-serine and lysine-threonine (D-configured) dipeptides.

2.3 Thiazolidine peptide synthesis

Whilst *N*-terminal glycine, serine and threonine residues can be considered to be canonical "caged" glyoxyls, a non-canonical precursor to a glyoxyl group may also prove to be useful. As mentioned, 1,2-aminothiols on cysteine derivatives have been protected with aldehyde-based protecting groups, with release of the cysteine derivative effected by a strong nucleophile such as methoxyamine.[219] The reverse scenario can be considered: release of the aldehyde effected by an electrophile. Furthermore, a thiazolidine carboxylate bears high structural similarity to a pyrroline carboxylate, the additional group attached to the lysine backbone of pyrrolysine. Indeed, 2- and 4-carboxythiazolidine derivatives have both been shown to be suitable substrates for amber stop codon suppression.[220, 221] Hence the third strategy for exposure of protein glyoxyl aldehydes presents itself: genetic incorporation of 2-carboxythiazolidine **86** followed by treatment with an appropriate, biocompatible electrophile.

A racemic synthesis of **86** has been reported;[221] whilst useful for the preparation of the thiazolidine group, the procedure made use of an *in situ* activation coupling strategy, unattractive for reasons already presented, and hence another two-step activation-coupling procedure was deployed (Figure 37). The condensation of cystamine **117** with glyoxylic acid **118** proceeded well as reported, although the absence of chiral induction affords **119** as a racemate. Boc protection was modified from the original procedure,

using reduced quantities of Boc anhydride to aid purification, and was almost quantitative in yield, with **119** obtained in 80% yield over two steps. Formation of NHS ester **120**, coupling to form protected dipeptide **121**, and standard TFA deprotection afforded **86** in 66% yield over three steps. This synthetic route was highly amenable to upscaling, with large quantities (20 mmol) of **86** synthesised in anticipation of the extensive work required in developing a suitable decaging method.

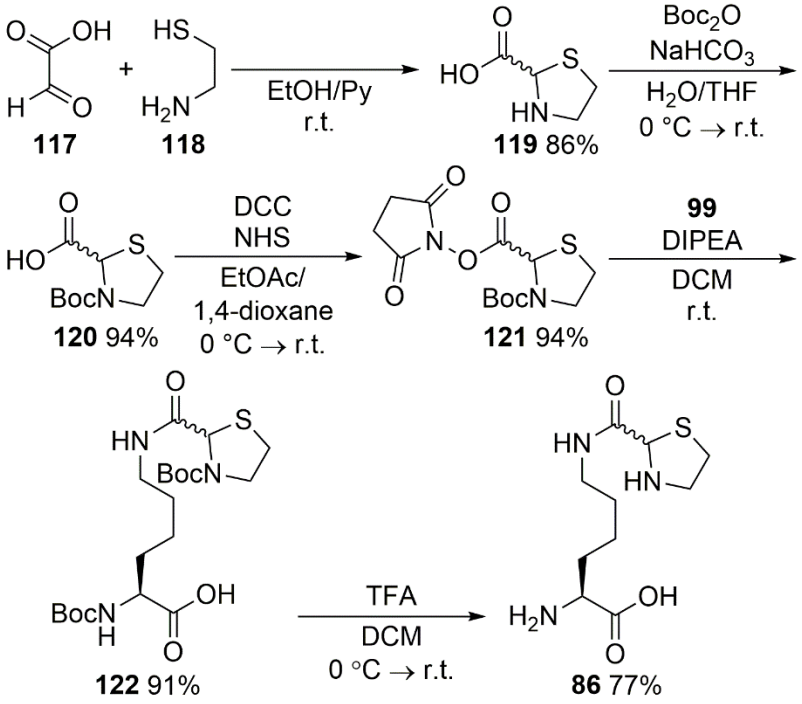


Figure 37: Synthesis of thiazolidine-lysine dipeptide **86**.

2.4 Reactive peptide probe synthesis

One of the virtues of a protein glyoxyl aldehyde is its reliable canon of reactivity. To this end, various small reactive probes are required in order to chemically modify the exposed protein aldehyde. Such probes must combine an aldehyde-reactive functional group with a group of utility, with fluorescent and affinity groups being prime examples, whilst maintaining biocompatibility such as solubility in aqueous solution.

Solid-phase peptide synthesis provided an elegant solution for the synthesis of these peptides, combining modularity with straightforward synthesis and minimal purification. Each peptide can be considered to follow the same "flag-like" template, from *N*-terminus to *C*-terminus: i) reactive functional group, for modification of the protein aldehyde, the "attachment point" of the flag; ii) a PEG-type spacer, increasing aqueous solubility and minimising steric bulk around the reactive functional group, effectively the "mast"; iii) either a fluorescent (dansyl) or affinity (biotin) tag, for further demonstration of successful

modification, the “flag” itself; iv) a resin attachment residue, for simplicity a glycine, the “cap”.

2.4.1 Aminooxy probes

A well-established bioorthogonal aldehyde modification strategy is oxime ligation, the condensation of an aldehyde with a highly nucleophilic aminooxy reagent.[79] This strategy typically requires some form of catalysis to be effective, which has generally taken the form of Brønsted-Lowry acid catalysis at pH 4-5 *via* a protonated carbonyl reactive intermediate.[238] Further developments have shown that nucleophilic catalysts such as aniline can be used to drive the reaction to completion at neutral pH,[239] forming a highly electrophilic iminium ion greatly susceptible to nucleophilic attack by the aminooxy reagent. The glyoxyl aldehyde has enjoyed much usage in this way, as the electrophilic character of this aldehyde to some extent balances out the sluggish conversions often observed with this reaction and has been used to generate a diverse portfolio of bioconjugates.[240-242]

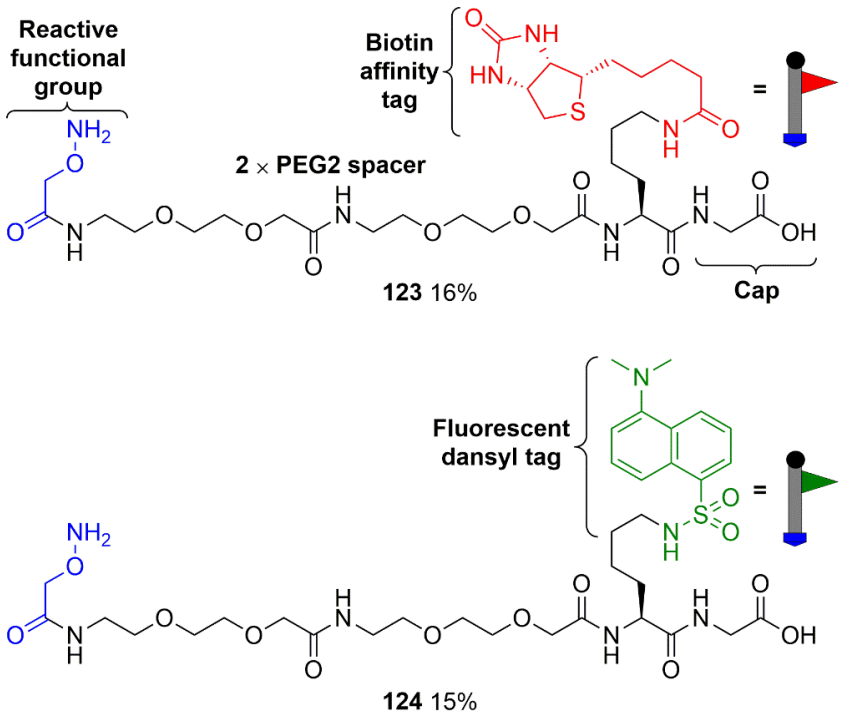


Figure 38: Structure and schematic of aminooxypeptides **123** and **124**.

Hence reactive aminooxy peptides **123** and **124** were synthesised by Dr Martin Fascione using Fmoc SPPS methodology (Figure 38). Whilst most residues were coupled by *in situ* activation using HCTU, the *N*-terminal aminooxyacetic acid unit was preactivated as the NHS ester. This is due to the ability of *N*-Boc-aminooxyacetic acid to be sufficiently nucleophilic, despite *N* protection, to intercept HCTU-activated carboxylates, whilst

preactivation circumvents this unwarranted side-reaction.[243] Both peptides were obtained in *ca*. 15% yield, with no purification required.

2.4.2 Aryl aldehyde probes

Work undertaken in the Fascione group has led to the development of the Organocatalytic Protein Aldol Ligation (OPAL): a cross-aldol reaction in which an enolisable aldehyde, typically in excess, reacts with a non-enolisable protein aldehyde in the presence of an organocatalytic proline derivative, forming a new carbon-carbon bond between aldehyde partners.[244] This reaction has been shown to be rapid and biocompatible, with full conversion typically taking one hour at neutral pH. Protein glyoxyl aldehydes are an ideal substrate for this reaction on account of the reliable reactivity and the inability to enolise. Furthermore, this reaction affords a product containing a further aldehyde, a β-hydroxyaldehyde, which is resistant to further OPAL modification but can be modified by oxime ligation. Hence OPAL is a second excellent ligation strategy to take advantage of an exposed protein aldehyde located at a non-terminal position.

The most reactive enolisable aldehydes have been shown to be arylacetaldehyde derivatives, whilst reactions with alkyl aldehydes tend to display much lower rate constants (two orders of magnitude smaller). This is likely due to the aryl group providing a thermodynamic boost to enamine formation, the putative rate-determining step in this reaction: the α-protons are more acidic (pK_A of phenylacetaldehyde vs. propanal are 15 vs. 17 respectively) and the p-orbitals of the amine and enamine π bond are conjugated to the aryl group to form an extended conjugated system. These two factors contribute to a greater concentration of enamine existing in solution, increasing the rate of the reaction. With this in mind, the reactive functional group to be installed on the appropriate OPAL probes needed to be carefully considered. Compound **125** was synthesised by Dr Darshita Budhadev for this purpose: a tyrosinol derivative containing an isopropylidene hemiaminal, which exposes a 1,2-aminoalcohol upon TFA deprotection/cleavage ready for periodate-mediated oxidation to form the phenylacetaldehyde-containing probe. **125** was attached to **126-resin** and **127-resin** (Figure 39), prepared using the same Fmoc SPPS protocol, whilst still on the resin, to afford probe precursors **128** and **129** in 84% and 50% yields respectively after cleavage, deprotection and isolation. Oxidation and purification were performed by Dr Richard Spears to afford probes **130** and **131** in quantitative conversion ready for bioconjugation using OPAL.

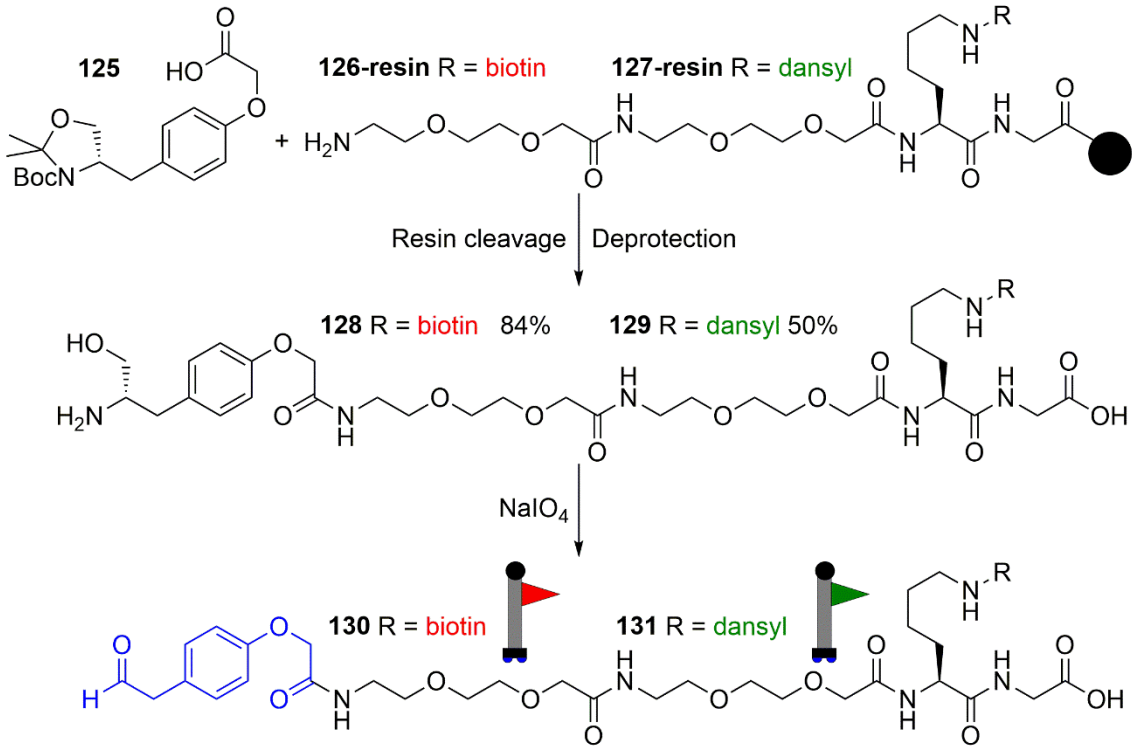


Figure 39: Synthesis of OPAL probes **130** and **131**.

2.4.3 Strained alkyne probes

A third bioconjugation strategy making use of the glyoxyl aldehyde is the strain-promoted alkyne-nitrone cycloaddition (SPANC).[245] In this reaction, a nitrone, formed *in situ* from a protein glyoxyl aldehyde and an *N*-substituted hydroxylamine reagent under *p*-anisidine nucleophilic catalysis, undergoes a [3+2] cycloaddition with a strained alkyne, typically a cyclooctyne or a BCN.[246] Arguably this procedure is a biocompatible translation of the so-called A^3 reaction,[247] in which an aldehyde and amine form an imine which can be attacked by a metallated alkyne, albeit using the alkyne as a dipolarophile rather than as a nucleophile and using ring strain instead of metals to activate the alkyne. Two new carbon-carbon bonds are generated under mild conditions to form a stable isoxazoline heterocycle linking the three components. SPANC is a further demonstration of the power of the glyoxyl aldehyde, capable of undergoing a three-component reaction with potential for multiple points of functionalisation.

Following the previous probe design, the SPANC probes required a strained alkyne on the *N*-terminus. A BCN was chosen for this purpose, on account of the greater reactivity of a BCN compared to a regular cyclooctyne. However, attempts to cap the *N*-terminus of precursor probes **126-resin** and **127-resin** on the solid phase using activated BCN **132** were largely unsuccessful, with no capping detected by LC-MS and with a significant quantity of **132** wasted due to the excess used for the solid phase. To circumvent this problem, the precursor probes were cleaved from the resin and the *N*-terminal capping

was performed in solution on **126-OH** and **127-OH** (Figure 40). Pleasingly, full conversion was observed by LC-MS within one hour. Subsequent purification by size-exclusion chromatography afforded **133** and **134** in 68% and 55% yields respectively.

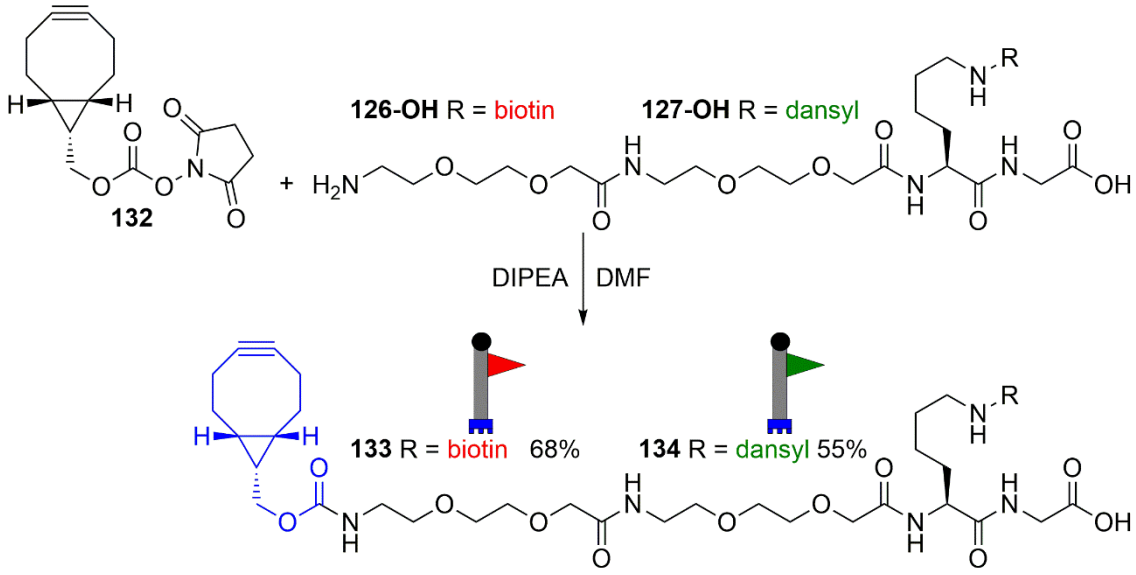


Figure 40: Synthesis of SPANC probes **133** and **134**.

2.5 Conclusions

Three strategies for the incorporation and unmasking of glyoxyl aldehyde groups in protein systems have been devised. Six lysine-based dipeptides required for these strategies have been synthesised in yields ranging from 32% to 66%, in the quantity and of the purity required for further biological studies.

Furthermore, three ligation strategies have been outlined to demonstrate the utility of the exposed glyoxyl aldehyde. Probes for oxime ligation and OPAL were synthesised using SPPS, whilst probes for SPANC were prepared using a combination of solid- and solution-phase synthesis, ahead of bioconjugation usage.

Chapter 3: Test Protein Design and Production

3.1 Introduction

In order to identify genetically encoded caged aldehydes and develop strategies for the genetic incorporation of caged aldehydes into proteins, a suitable protein scaffold is required. Ideally such a test protein would be readily produced and purified, observable by mass spectrometry and tolerant of a judiciously placed TAG mutation, whilst remaining a representative example of the wider proteome. Green fluorescent protein (GFP) is an excellent candidate for this role.

3.1.1 Wild-type GFP

The prototypical GFP hails from *Aequorea victoria*, a bioluminescent species of jellyfish, in this instance playing the role of a light transducer. In a species of *A. victoria,* the protein aequorin, a luciferin-luciferase pair, is activated and releases a flash of blue light in the absence of GFP and green light in the presence of GFP.[248] Why exactly this jellyfish evolved to produce this transducer protein, adding a second colour to the visual palette, is not known;[249] one hypothesis is that the ability to modulate the wavelength of bioluminescence reflects the two sea depths at which the organism lives and the wavelength most visible at that depth.[250]

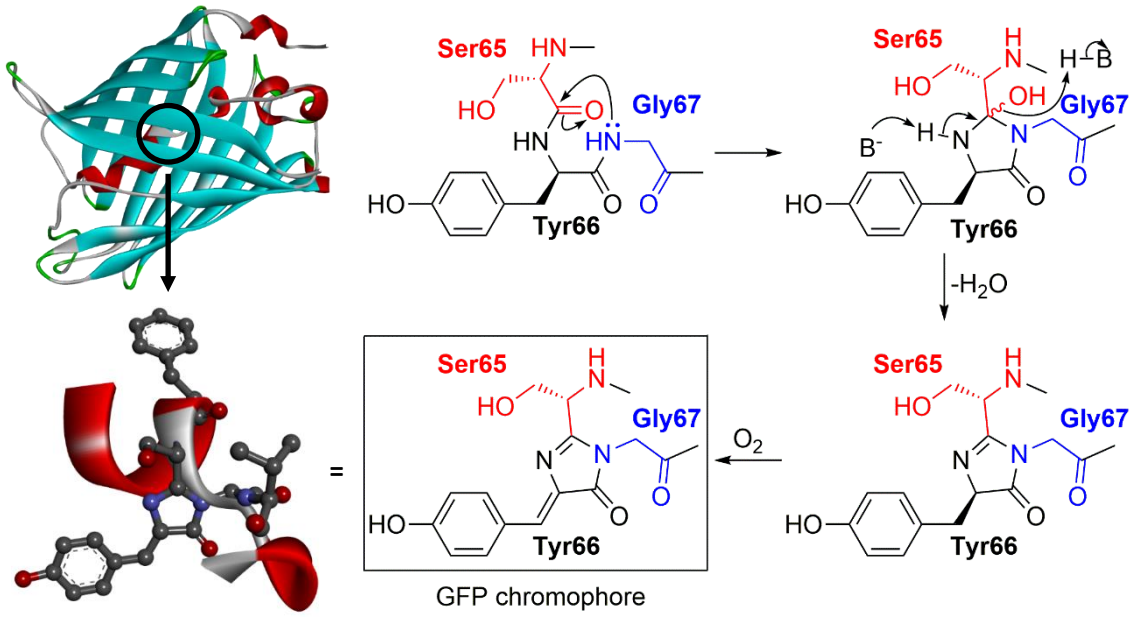


Figure 41: Inside the GFP β-can (top left, cyan) lies an α-helix (red) where the *p*-hydroxybenzylidene imidazolidinone chromophore can be found (bottom left), formed through an intramolecular cyclisation and elimination between Ser65 and Gly67 and the oxidation of Tyr66 (right).

The structure of GFP is highly optimised towards the function of fluorescence: a so-called "β-can", formed of a β-barrel capped by α helices threading through the centre, in the middle of which the chromophore is encapsulated (Figure 41).[251] Diligent toil by Shimomura,[252] subsequent Nobel laureate for his contribution to the body of work on

GFP, elucidated the structure of this chromophore as a *p*-hydroxybenzylidene imidazolidinone derivative,[253] later proposed to be formed by a spontaneous dehydrative cyclisation of residues Ser65, Tyr66 and Gly67 followed by oxidation by dioxygen.[254] The β-can acts as a stiff protective scaffold, enhancing the fluorescent properties of the chromophore: excluding dioxygen, a potential quenching or photobleaching agent;[255] reducing pathways for undesirable excited state quenching by vibrational relaxation, instead of fluorescence emission, by maintaining conformational rigidity and minimising collisions with solvent, presumably also reducing the Stokes' shift; and providing some shielding from fluctuations in pH, with fluorescence observed between the large range of pH 4 and pH 12.[256] With an absorption maximum at 476 nm, almost matching the emission maximum of aequorin,[257] and a high fluorescence quantum yield $\varphi_f = 0.79$ at an emission maximum of 504 nm,[258] clearly nature has painstakingly evolved GFP to complement and accentuate aequorin, affording the result of an elegant yet surreal example of bioluminescent resonant energy transfer noted even by Roman scholar Pliny the Elder almost 2000 years ago.

3.1.2 Engineering of GFP

The biotechnological potential of GFP was soon exploited, making the move from its original marine environment to its new home in the lab. Just as any living organism would be expected to adapt in order to better fit its updated ecological niche, so too was GFP forced to evolve to meet the demands of researchers through designed mutations and an entire evolutionary line of GFPs has been constructed.

One of the earliest mutants, and perhaps the most prominent, is enhanced GFP (EGFP), bearing two key mutations of F64L and S65T.[259] The S65T mutation has a profound impact upon the absorption profile of EGFP: a single absorption peak at 488 nm, emitting at 507 nm, with a molar absorption coefficient of 5.3×10^4 dm^3 mol^{-1} cm^{-1}.[258] A similar peak occurs in the absorption spectrum of wild-type GFP (wtGFP), centred at 476 nm but with a much smaller absorption coefficient of 9.5×10^3 dm^3 mol^{-1} cm^{-1}, the peak responsible for energy transfer from aequorin. However, an additional peak occurs at 395 nm with a molar absorption coefficient of 2.5×10^4 dm^3 mol^{-1} cm^{-1}, with both peaks emitting at 507 nm. The same GFP chromophore structure possesses two different absorption maxima and this has been attributed to the protonation state of the phenolic oxygen atom. The protonated species is responsible for the 395 nm absorption and the deprotonated species absorbs at 475 nm;[260] emission at 507 nm arises from fluorescent deexcitation of the excited state of the deprotonated chromophore, which can be accessed from the protonated chromophore excited state by a rapid proton exchange pathway. Whilst the wtGFP chromophore is mostly protonated at a pH below 12,[261] the

S65T mutation rearranges hydrogen bonding so that the chromophore exists exclusively in the deprotonated state, losing the absorption peak at 395 nm but greatly enhancing the molar absorption coefficient of the 475 nm absorption maximum.[262] On first glance, the S65T mutation appears to overcome one of nature's shortcomings in GFP design, producing a simpler absorption spectrum leading to greater fluorescence and hence a more efficient energy transfer from aequorin. However, the S65T mutation also leads to greater acid sensitivity of EGFP, increasing the protein pK_A by two units.[263] In nature's judgement, GFP design has been a balance between efficient energy transfer with aequorin and survival of GFP across cellular environments at varying pH; in the lab environment, where the aqueous milieu is generally pH-controlled, the latter factor loses relevance and chromophore efficiency becomes the main selection pressure.

The S65T mutation is synergistic with the F64L mutation in two ways: red-shifting the absorption maximum by 10 nm through stabilisation of the anionic excited state and improving folding efficiency at 37 °C, which had proven highly problematic with wtGFP.[258] The lab environment once again proves to come with its own selection pressures, where the production of protein at 37 °C becomes a highly desirable trait that had never been needed in the cold, cnidarian former habitat for wtGFP. From this same selection pressure arose a trio of different mutations: the "cycle3" mutations of F99S, M153T and V163A that, when installed in a wtGFP scaffold, led to a 40-fold increase in protein fluorescence.[264] This was attributed to a greater proportion of correctly folded GFP when the cycle3 mutations were present, as wtGFP exhibited a tendency to languish within inclusion bodies without the correct fold and chromophore when overexpressed in prokaryotic host systems. The side-chains of residues 99 and 153 are solvent-exposed, whilst the side chain of residue 163 is angled towards a hydrophilic internal pocket composed of Gln183, Tyr151, Lys162 and Asn164.[265] By increasing the hydrophilicity of the surface-exposed residues 99 and 153 and decreasing the hydrophobicity of the side-chain of residue 163 in a hydrophilic pocket, the cycle3 mutations were found to inhibit protein aggregation and hence inclusion body formation.[266] An unneeded variation in nature where GFP concentration is typically low, but a very practical solution to improve the yield of recombinant GFP.

Spurred by the success of the cycle3 and EGFP mutations, further optimisation of the GFP primary structure produced superfolder GFP (sfGFP) bearing six new key mutations of S30R, Y39N, N105T, Y145F, I171V and A206V in addition to the EGFP and cycle3 mutations.[267] sfGFP was found to exhibit some of the fastest folding kinetics of any GFP mutant, high tolerance of denaturing urea and superior in-cell fluorescence, owing to the stability of the fold and the reduced propensity for misfolded, aggregated, non-fluorescent sfGFP variants to form. Of the six mutations, S30R and Y39N contribute the

most to the enhanced sfGFP properties, with the former mutation forming a stabilising network of electrostatic interactions across the protein surface and the latter contorting a β-turn into a more favourable 3_{10} helix. Notably, the six mutations do not confer any additional pH stability or perturbation of spectroscopic properties. The pK_A of both EGFP and sfGFP is 5.9 and the absorbance/emission profiles overlap,[268] implying that the six new mutations in sfGFP are purely structural and do not alter the photochemistry of the chromophore. Owing to the enhancements arising from the respective mutations, EGFP and sfGFP have both seen extensive usage in chemical protein modification as practical test proteins for exploring new chemistries, easily produced in sufficient quantities *via* recombinant methods.[76, 196, 228] To this end, GFPs can serve as both a sensitive reporter protein in an amber stop codon suppression assay and also a useful test bed for new chemical protein modification strategies.

3.2 Constructs for amber stop codon suppression

3.2.1 Test protein design

Two GFPs were selected as test systems for amber stop codon suppression: EGFP and sfGFP. Two pBAD expression vectors harbouring the genes encoding the respective GFPs with *C*-terminal His_6 tags were received: EGFP(Y39TAG) from the Lemke group and sfGFP(N150TAG) from the Mehl group, where TAG represents a mutation of the native residue to the amber stop codon. Both constructs have been used in previous successful examples of amber stop codon suppression in prokaryotic systems.[169, 269]

For both proteins, the amber stop codon has been located in a position occupied by a surface-exposed residue in the wild-type protein, with a side chain extending towards the solvent rather than towards the centre of the protein (Figure 42). This rational design ensures that substitution with the NCAA offers minimal perturbation of the protein fold, as the side chain is not responsible for any key interactions with other residues which hold the structure together in the native fold. Furthermore, the side chain of the non-canonical residue will be available for reactions due to its placement on the surface of the β-barrel, minimising steric clashes which impair reactivity.

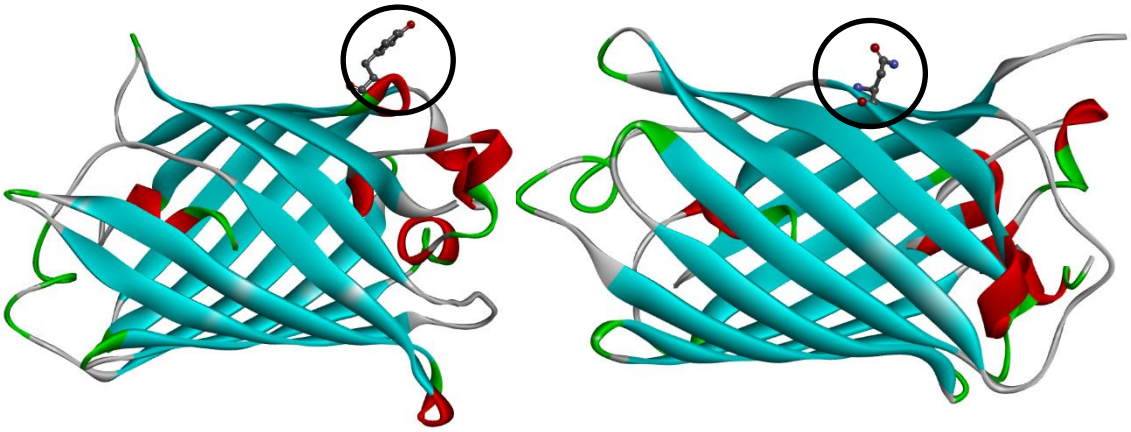


Figure 42: Positions of mutated residues: EGFP Y39 (left, PDB 2Y0G) and sfGFP N150 (right, PDB 2B3P).

The plasmid containing the gene for sfGFP(N150TAG) was used as received. The EGFP(Y39TAG) gene contains an additional *N*-terminal FLAG tag, however, and whilst this tag is generally functionally benign, the initial aspartate residue of the DYKDDDDK tag reduces the extent of *N*-terminal methionine cleavage.[270] In the absence of treatment with a more effective aminopeptidase, when observed by mass spectrometry a mixture of EGFP with and without the *N*-terminal methionine residue would be detected. No usage of the FLAG tag was desired thus this unnecessary additional complexity which would be seen in the protein mass spectrum arising from this tag could not be justified. Hence the *N*-terminal aspartate residue was mutated to become a serine residue. With a serine residue following the *N*-terminal methionine residue, complete cleavage of the *N*-terminal methionine residue would be expected,[271] with gene expression affording homogenous (by MS) EGFP. The choice of serine also permits oxidation with sodium periodate to afford an *N*-terminal protein glyoxyl, setting up the option of dual aldehyde decaging in tandem with a potential pyrrolysine surrogate incorporated at the position of the amber stop codon. Using site-directed mutagenesis and appropriate forward and reverse primers, the initial aspartate residue was mutated from the triplet GAT to the triplet TCT, a mutation involving the fewest possible base mismatches whilst still encoding a commonly utilised serine codon in *E. coli*, the desired host system for protein production. Sequencing confirmed successful installation of only the desired mutation in plasmid DNA isolated from five of eight colonies isolated, affording the desired construct ready for protein production.

3.2.2 Amber stop codon suppression system

The other two requisite components of amber stop suppression in this instance are the pyrrolysyl $tRNA_{CUA}$ and the corresponding pylRS. A pEVOL vector containing the pyrrolylyl $tRNA_{CUA}$ gene and two copies of the pylRS (*M. mazei*) gene, one constitutive and one under *araBAD* control, was received from the Lemke group (Figure 43). The pEVOL vector was designed specifically to maximise protein yields in amber stop codon

suppression.[125] The use of *araBAD* control over one copy of the pylRS gene allows controlled overexpression of the pylRS gene, improving the efficiency of amber stop codon suppression with minimal disruption to host growth through unintentional suppression of native amber stop codons.[272] Basal control over the other copy of the pylRS gene and the $tRNA_{CUA}$ gene also ensures that the concentrations are not high enough to be cytotoxic whilst also allows the production of a consistent yet moderate quantity of aminoacyl-$tRNA_{CUA}$, ensuring that translation of the protein of interest can begin immediately upon induction.

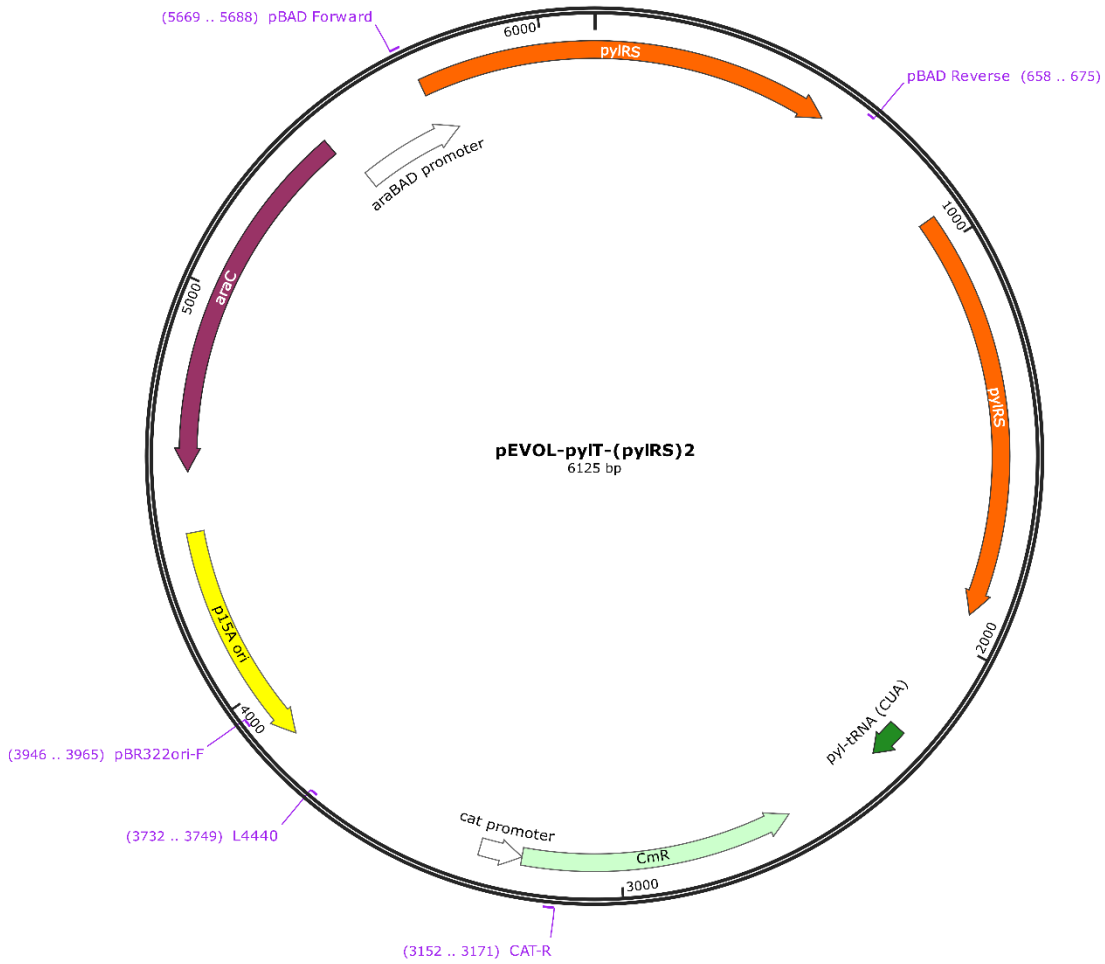


Figure 43: Plasmid map of the pEVOL vector harbouring one pyrrolysyl $tRNA_{CUA}$ gene (dark green) and two copies of the pylRS gene (orange), one under araBAD control (white) and one under constitutive control. Map prepared with SnapGene using the published description of the pEVOL vector.[125]

Two variants of this plasmid were received, differing in the sequence of the pylRS: one in which both copies correspond to that of the wild type ($pylRS^{wt}$), and one in which both copies correspond to the double mutant Y306A Y384F ($pylRS^{AF}$), known to tolerate bulkier, more hydrophobic moieties in the enlarged binding pocket.[122] Access to two pylRS variants will increase the likelihood of discovering a pyrrolysine analogue which can be successfully incorporated using amber stop codon suppression and allow for a

comparison and some level of insight into substrate recognition. Both derivatives of this plasmid have also seen successful use in amber stop codon suppression.[169]

3.3 Determining performance in amber stop codon suppression

With two pylRS mutants and six synthesised lysine derivatives to hand, a protein production-based assay was designed to ascertain whether each NCAA could indeed be genetically encoded by each pylRS variant. EGFP(Y39TAG)-His_6 was selected as the reporter protein on account of the highly optimised expression system available for use in amber stop codon suppression. Furthermore, the presence of EGFP can be determined not only through stained SDS-PAGE but also using the fluorescent properties of EGFP, visible in the cell pellet or cell lysate.[125] Together these properties ensure that the limit of detection of this assay is sufficiently low such that any successful amber stop codon suppression can be observed unequivocally, albeit in a qualitative rather than quantitative manner.

The assay takes the form of a miniaturised expression trial following a reported procedure for the expression of EGFP containing NCAAs.[169] In brief, Top10 *E. coli* cells transformed by vectors containing genes for EGFP(Y39TAG)-His_6, pylT, and either $pylRS^{wt}$ or $pyRS^{AF}$ were grown from a starter culture containing appropriate antibiotics and subsequently used to inoculate a separate preparative culture. The two preparative cultures were incubated until OD_{600} of 0.2 – 0.3 was reached, at which point the cultures were split into aliquots and a solution of NCAA (freshly prepared in aqueous 0.1 M NaOH) added to the culture aliquots to a final concentration of 2 mM. All aliquots were incubated for a further 30 minutes, induced at the same time and incubated overnight. Culture aliquots were subsequently centrifuged and cell pellets lysed chemically. Each centrifuged cell lysate aliquot was then analysed by Coomassie-stained SDS-PAGE. For negative control aliquots, a blank solution of aqueous 0.1 M NaOH was added. As a positive control for aliquots containing the $pylRS^{wt}$ variant, **29** was used; for aliquots containing the $pylRS^{wt}$ mutant, **34** was used (Figure 44). Synthesised NCAAs **86**, **96**, **101**, **105**, **109** and **110** were screened using this assay.

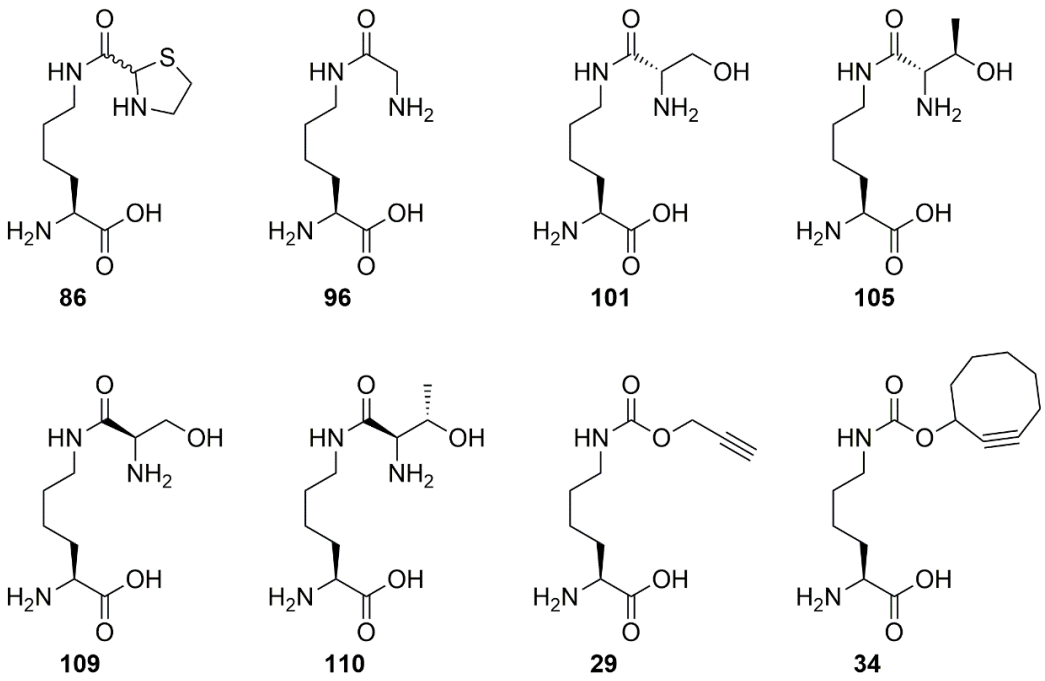


Figure 44: NCAAs **86**, **96**, **101**, **105**, **109** and **105** screened for uptake in amber stop codon suppression alongside positive controls **29** (pylRS^{wt}) and **34** (pylRS^{AF}).

3.3.1 Assaying with the wild-type pylRS

In the first assay, all six synthesised lysine analogues were studied alongside positive and negative controls in the presence of wild-type pylRS, affording eight samples to be analysed by SDS-PAGE. **86** is a known substrate of wild type *M. barkeri* pylRS;[221] given the generally overlapping set of substrates between analogous *M. barkeri* and *M. mazei* pylRS variants, **86** is expected to be a substrate for wild-type *M. mazei* pylRS.

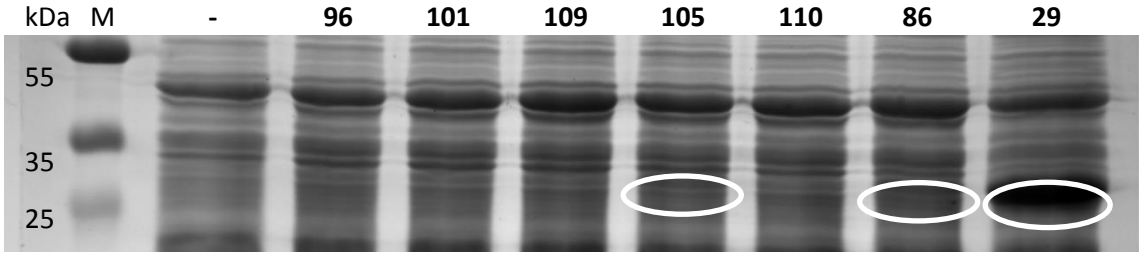


Figure 45: SDS-PAGE of cell lysate from the amber stop codon suppression expression trial of EGFP with pylRS^{wt} and NCAAs **86**, **96**, **101**, **105**, **109** and **110** with negative (-) and positive (**29**) controls. The circled bands in lanes **105**, **86** and **29** are attributed to full-length EGFP.

Analysis of the gel confirms presence of full-length EGFP in the positive control lane using **29**, with a clear, thick band corresponding to EGFP overexpression at *ca*. 28 kDa, the mass expected (Figure 45). No band at this position is observed in the negative control lane, confirming the orthogonality of the pylRS used to canonical amino acids. As expected, a band corresponding to EGFP can be seen in the **86** lane, albeit much thinner implying poorer recognition of **86** by pylRS^{wt} compared to **29**. The only other EGFP band

visible is found in the **105** lane, with this band also being far less pronounced than the corresponding band in the positive control lane, whilst lanes for **96, 101, 109** and **110** contained no bands for full-length EGFP.

To confirm these data, portions of the cell lysate were monitored by fluorescence. The results of the SDS-PAGE are reflected in the fluorescence observed here too: strong fluorescence in the positive control aliquot, mild fluorescence in the aliquots containing **86** and **105**, and only background visible for all other samples including the negative control (Figure 46). From this assay, amino acids **86** and **105** have been shown to recognised by wild type *M. mazei* pylRS to an appreciable extent, whilst all other NCAAs assayed showed no detectable extent of amber stop codon suppression.

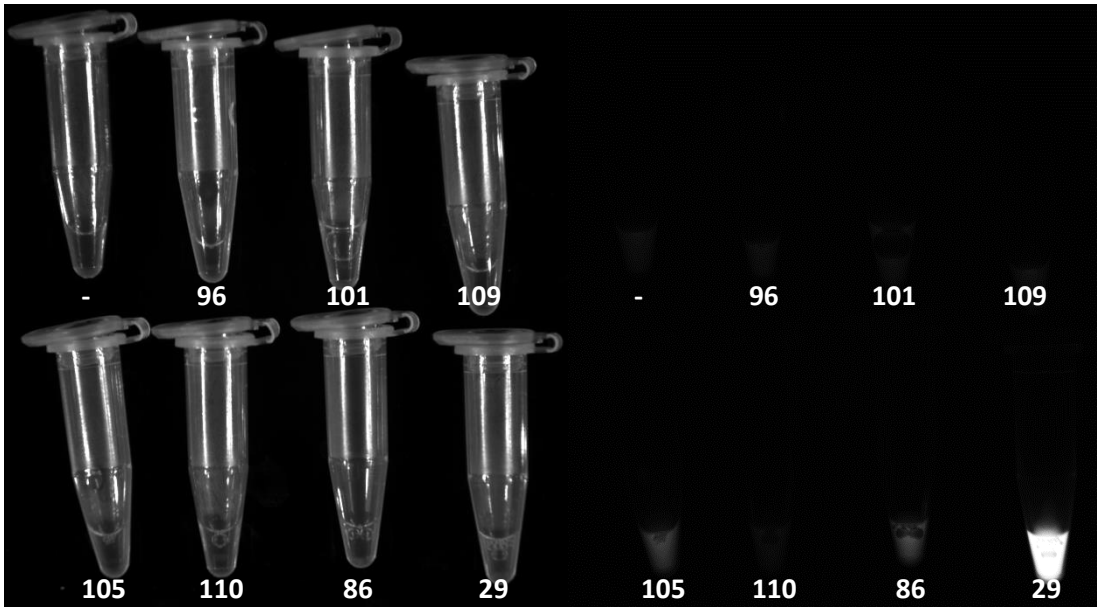


Figure 46: Visualisation by white light (left) and fluorescence (right) of cell lysate from the exp the amber stop codon suppression expression trial of EGFP with pylRS^{wt} and NCAAs **86**, **96**, **101**, **105**, **109** and **110** with negative (-) and positive (**29**) controls.

3.3.2 Assaying with the double mutant pylRS

The same assay was performed using the doubly mutated pylRS^{AF} with the same set of amino acids, bar the altered positive control of **34**, and analysed by SDS-PAGE (Figure 47).

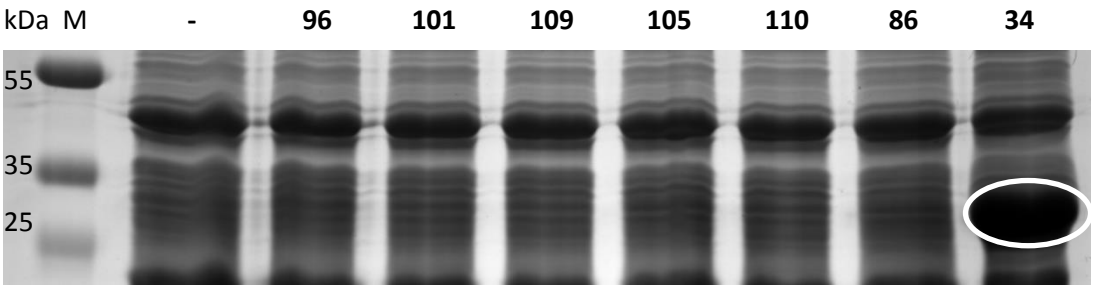


Figure 47: SDS-PAGE of cell lysate from the amber stop codon suppression expression trial of EGFP with pylRS^{AF} and NCAAs **86**, **96**, **101**, **105**, **109** and **110** with negative (-) and positive (**34**) controls. The circled band in lanes 34 is attributed to full-length EGFP.

This time only a single band corresponding to EGFP overexpression was observed in the positive control lanes, with all other lanes free of this band and resembling the negative control.

3.3.3 Discussion

Placing these results in the context of the active site understanding of pylRS^{wt} adds some justification. The small size of **96**, only one NH_3 unit larger than acetylated derivative **2**, presumably leads to poor occupancy of the large binding pocket. **2** could only be recognised by a pylRS variant following mutations to close off most of the binding pocket,[123] with the extra amino group of **96** projecting into a particularly hydrophobic portion of the binding pocket. Within the pylRS^{AF} binding pocket, the additional space opened up by the mutations is too deep to be accessed by **96**, suggesting even poorer binding in this synthetase mutant.

Recent work has found that **135**, an analogue of **96** following a similar objective of genetically encoding a pseudo-*N*-terminal glycine motif, was recognised by a pylRS derivative from *M. barkeri* (Figure 48).[273] Directed evolution led to just three mutations: L309A, N346Q and C346S (numbered according to the homologous *M. mazei* active site sequence), although no crystal structure displaying the exact binding mode was reported. Following protein production and Staudinger reduction, the pseudo-*N*-terminal glycine-glycine motif was used with sortase-mediated ligation to append ubiquitin to a protein of interest *via* a near-native linkage. This work suggests that the lack of recognition of **96** by pylRS^{wt} is not unsurprising, given the mutations required for recognition of **135**. The presence of the extra bulk provided by the azido group further supports the argument that the side chain of **96** is simply too small for adequate recognition by pylRS. Whilst the reduction of **135** to afford the *N*-terminal glycine residue is an innocuous step, it is still an additional task, implying that the azido group is a necessity to improve recognition by pylRS compared to a smaller amino group. This work

demonstrates both the possibility and utility of genetic encoding of pseudo-*N*-terminal glycine residues, albeit *via* different routes from those outlined within this thesis.

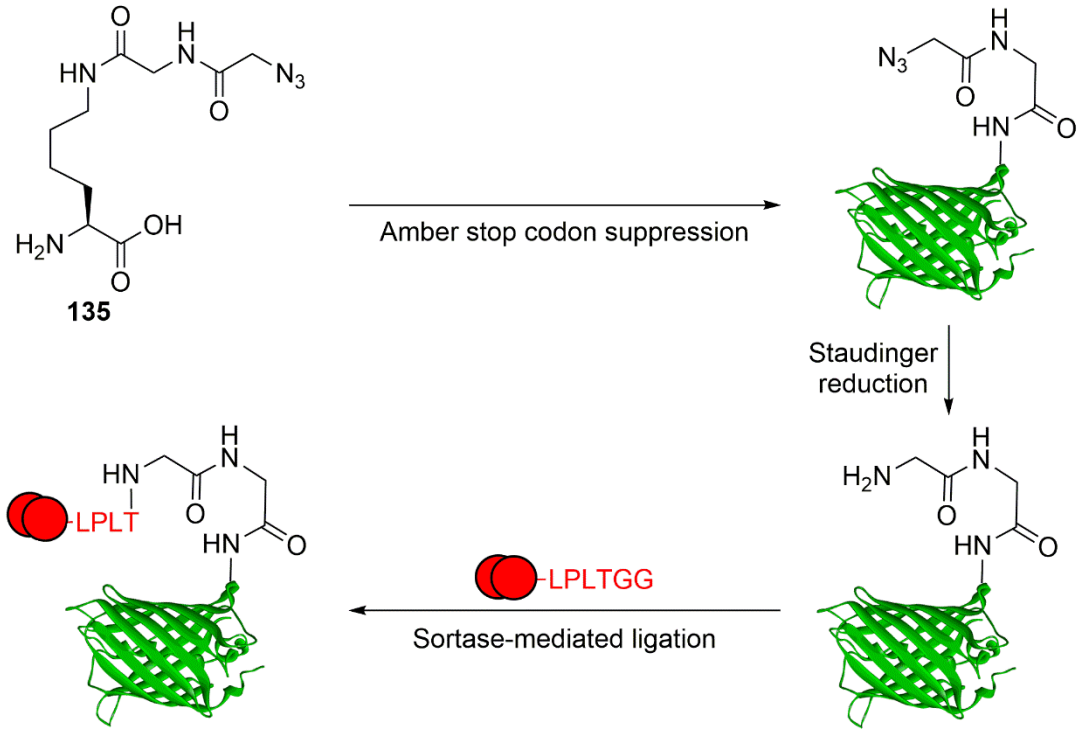


Figure 48: Genetic incorporation of **135** followed by Staudinger reduction affords a glycine-glycine-lysine motif amenable to transpeptidation with a tagged LPLTGG peptide, such as to install ubiquitin on the side chain of a lysine residue with a ubiquitin-LPLTGG-K linkage similar to native ubiquitin-LRLRGG-K linkages.

For the larger serine and threonine derivatives, the mode of binding is less simple to predict reliably due to the additional chiral centres, larger bulk and more hydrogen-bonding atoms. The likely explanation for a threonine derivative being a superior pylRS substrate to a serine derivative is that the extra methyl group can better occupy the hydrophobic space within the pocket. Even so, **105** is a far poorer substrate for pylRS^{wt} than positive control **29**, as seen from the relative size of the EGFP bands seen on the gel. The rationale behind the use of a Ser-Lys derivative arose from the genetic incorporation of Cys-Lys derivatives **83*R*** and **83*S***, where the *M. mazei* pylRS^{wt} was found to recognise both diastereomers.[215] Exchanging a sulfur atom for an oxygen atom seemed to be a sufficiently small structural perturbation such that some recognition by pylRS^{wt} would be retained; clearly this is not the case. The seryl oxygen atom in **101** or **109** is smaller and more polar than the sulfur atom in either diastereomer of **83**, suggesting poor occupancy of the large hydrophobic binding pocket. **86** combines the large, soft sulfur heteroatom of **83** with a somewhat hydrophobic saturated heterocycle analogous to the pyrroline ring of **1**, explaining the recognition of **86** observed with pylRS^{wt}.

Further work demonstrated a marked sensitivity of the *M. mazei* towards the α'-configuration of **83**, with **83*S*** using "unnatural" D-cysteine a superior substrate to **83*R***, just as the α' chiral centre in pyrrolysine is from "unnatural" D-methylornithine.[219] As neither Ser-Lys diastereomer **101** or **109** could be recognised by pylRS^{wt}, this implies that the loss of binding arising from the heteroatom change is not sufficiently compensated by the supposed boost offered by the "unnatural" configuration. The addition of a methyl group does clearly improve recognition, as can be seen comparing the assay results for **101** and **105**, although this is not the case between **109** and **110**.

105 appears to hit a sweet spot of the correct configuration and sufficient hydrophobicity from the methyl group. It is not clear which threonyl chiral centre in **110** is incorrectly configured. Comparisons with allo-Thr-Lys peptides would give some indication of the incorrectly configured centre or centres, as would a crystal structure of pylRS^{wt} bound to **105**. The reasons why configuration at the α' centre, and perhaps the β' centre as well, affect substrate recognition are not entirely clear. Assuming a similar mode of binding to pyrrolysine **1**, where the α' amino group is locked in place through hydrogen bonding to Y384, "unnatural" Thr-Lys **110** should be the best fit, with the α'-amino and β'-methyl groups of **110** matching the conformation adopted by **1** in the binding pocket. This conformation does, however, place the β' hydroxyl group pointing into a particularly hydrophobic region of the pocket composed of L309, C346 and W417. An alternative binding mode uses the β' hydroxyl group to hydrogen bond with Y384. For "natural" Thr-Lys **105** this would angle the β' methyl group at the side chains of W417 whilst the α' amino group lines up for hydrogen bonding with the side chain carbonyl oxygen atom of N346 (Figure 49), but for the "unnatural" **110** the α' amino group would be positioned in a way where no such hydrogen bonds could be established, implying that a "natural" α' configuration is key in this new binding mode. This binding mode, and arguably other conceivable modes too, still leads to poor occupancy of the rear of the pocket, likely a factor in the lower recognition of **105** by pylRS^{wt} compared to positive control NCAA **29**.

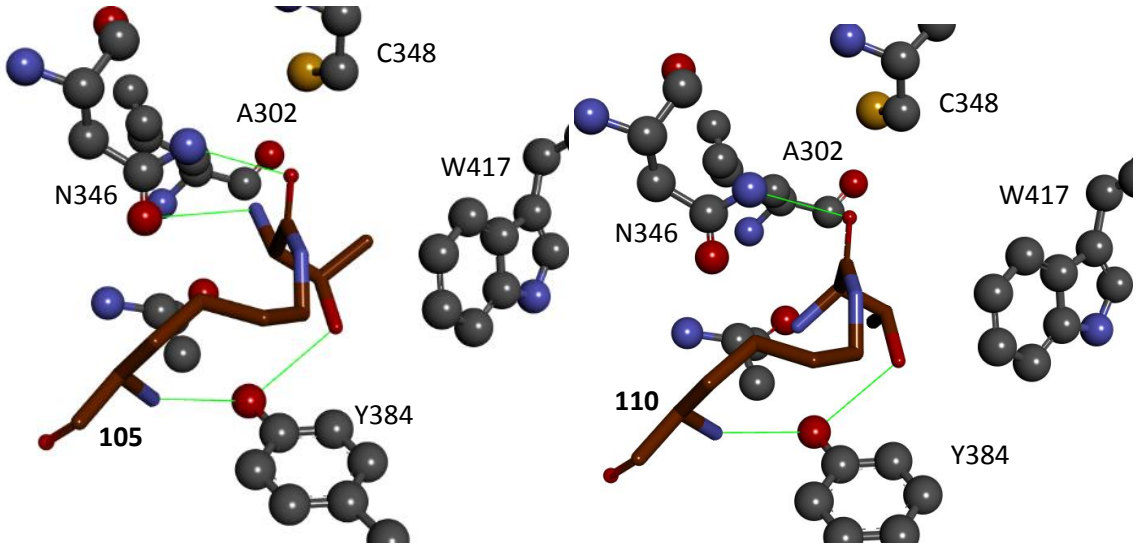


Figure 49: Proposed binding mode between pylRS and **105** (left) and **110** (right), with putative hydrogen bonding interactions highlighted in green.

Of course, these postulated binding modes are all based upon the assumption that the active site of pylRS is mostly rigid. In reality, this is likely not the case: whilst the interactions are retained, the precise geometry of the binding pocket will slightly distort in a different manner for each recognised substrate. Without a crystal structure to provide more definitive evidence of precise binding modes, establishing the precise reasons why **105** is recognised by pylRS^{wt} is not possible. Furthermore, this process exposes the flaws in designing NCAAs for amber stop codon suppression using a rigid *ab initio* rationale, unquestioning of the assumptions made about the flexibility of the active site of the tRNA synthetase. The organic chemist's assumptive 2D perspective is not fit for purpose in a complex 3D biological world.

Rationalising the results with pylRS^{AF} is more straightforward. As mentioned, the Y306A mutation opens the rear of the hydrophobic pocket and the Y384F mutation knocks out the gatekeeper residue. Whilst the Y384F mutation alone appears to increase promiscuity of pylRS, to the point that it is no longer completely orthogonal to canonical amino acids, with the Y306A mutation this pylRS mutant becomes highly selective for lysine derivatives bearing large, hydrophobic side chains such as **34**, the positive control. All of the studied lysine derivatives are far smaller and more polar than **34**, resulting in a poor fit in the hydrophobic binding pocket and negligible recognition by pylRS^{AF}.

3.4 Production of ncAA-containing proteins

With two novel NCAAs suitable for amber stop codon suppression, protein production was scaled up to obtain sufficient quantities for modification purposes. The EGFP(Y39TAG)-His$_6$ was selected as the first construct to be produced containing **86** as the *N*-terminal serine may prove to be of utility for dual modification purposes once a

method for decaging **86** has been developed. A second construct, sfGFP(N150TAG)-His_6, was also used to provide a second test subject. For the test decaging of **105**, the construct sfGFP(N150TAG)-His_6 was selected so that oxidation using sodium periodate would occur exclusively at the non-canonical residue without occurring additionally at the *N* terminus, which would impair attempts to quantify the extent of oxidation.

3.4.1 EGFP(Y39-(thiazolidine)lysine)-His_6

A scaled-up version of the protein production protocol used in expression trials was utilised for the preparation of EGFP(Y39-**86**)-His_6, **136**. A single culture was grown with the addition of **86**, induced and left to grow overnight. The harvested cell pellet was lysed by sonication and the lysate purified by Ni affinity chromatography. Fractions containing pure full-length protein, as determined by Coomassie-stained SDS-PAGE and the UV chromatographic trace (Figure 50), were pooled, desalted and concentrated. Whilst the fluorescence of fractions provides some indication of the presence and concentration of EGFP, this is of no use during SDS-PAGE. Boiling of the samples, required to homogenise the sample in a completely unfolded state and associate with SDS, ensures that any EGFP is no longer fluorescent, so any EGFP bands in SDS-PAGE can only be reliably detected with staining, e.g. Coomassie, rather than fluorescence. When the samples are not boiled, fluorescence is mostly retained, implying retention of the protein fold, but the resulting bands in SDS-PAGE do not correspond to the correct masses when visualised by fluorescence or by staining and the band is often more diffuse, likely due to the variance in folding states present in the sample. Approximately 40 mg protein was obtained per litre of culture. This protocol is facile, scalable and high-yielding, affording sufficient quantities of **136** for modification trials.

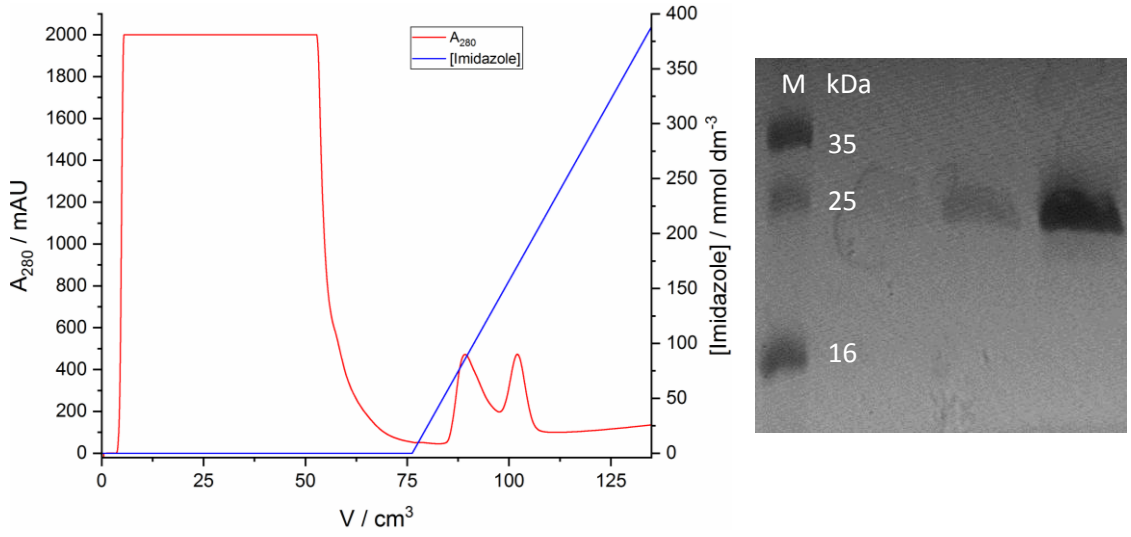


Figure 50: (left) 280 nm chromatogram of the purification of **136** by nickel affinity, eluting with an imidazole gradient; (right) SDS-PAGE of fractions containing **136**: a clear band around the expected 28.6 kDa mass.

ESI-FTICR-MS analysis also confirmed the purity and homogeneity of the sample, with the single peak in the deconvoluted mass spectrum displaying a mass consistent with the calculated mass of intact **136** (Figure 51). The mass of the intact protein should correspond to the mass of the amino acid sequence, factoring in the NCAA, as well as the mass difference associated with the post-translational modification required to form the chromophore, a loss of 20 Da. The protein mass is calculated as the molecular weight when considering the elemental composition of all residues and post-translational modifications, assuming no degradation. The protein is predominantly detected as [**136**+nH]$^{n+}$, although sodium adducts such as [**136**+(n-1)H+Na]$^{n+}$, 22 Da higher m/z, and phosphate (P_i) adducts such as [**136**+nH+H_3PO_4]$^{n+}$, 98 Da higher m/z, were also observed, likely due to incomplete removal of buffer salts. The deconvoluted spectrum depicts the protein as a neutral species, the mass of which is obtained from the peaks in the charge ladder of the raw spectrum assumed to be [M+nH]$^{n+}$ species, as the m/z and charge of such adducts are known. The base peak corresponds to the molecular weight of the protein species. The mass found of the neutral species was correct to within 2 Da, confirming the identity of **136**. A trypsin digest of **136**, performed by Dr Adam Dowle, confirmed that NCAA **86** had been installed only at position 39 (Figure 52) and confirmed the expected primary sequence of **136**.

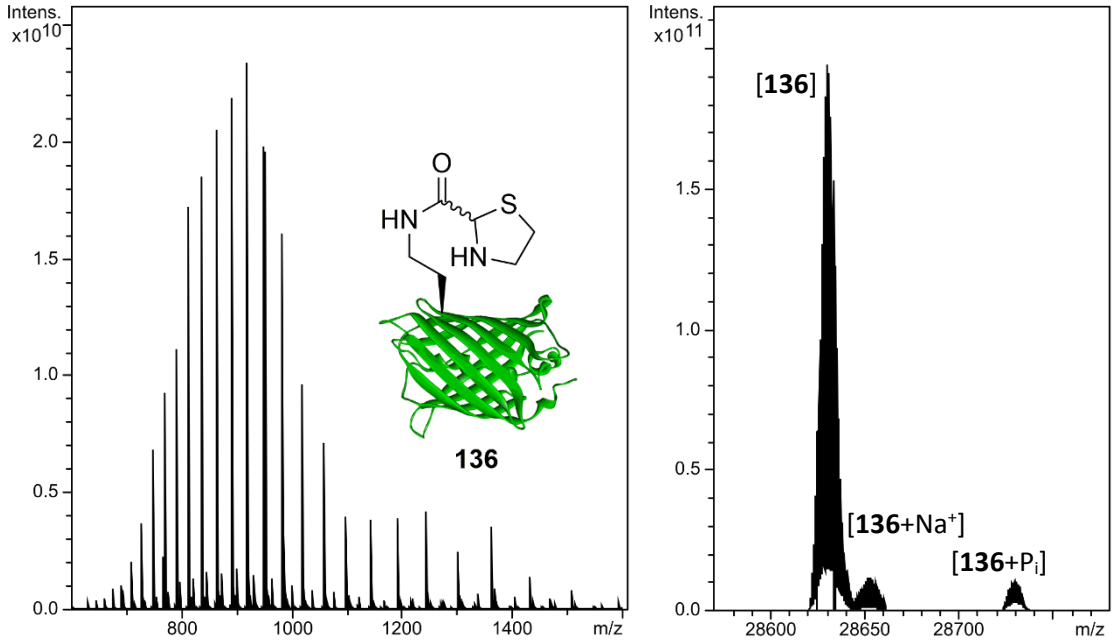


Figure 51: Raw (left) and deconvoluted (right) ESI-FTICR mass spectrum of **136**. For the neutral species, calc. 28632 Da; found 28633 Da.

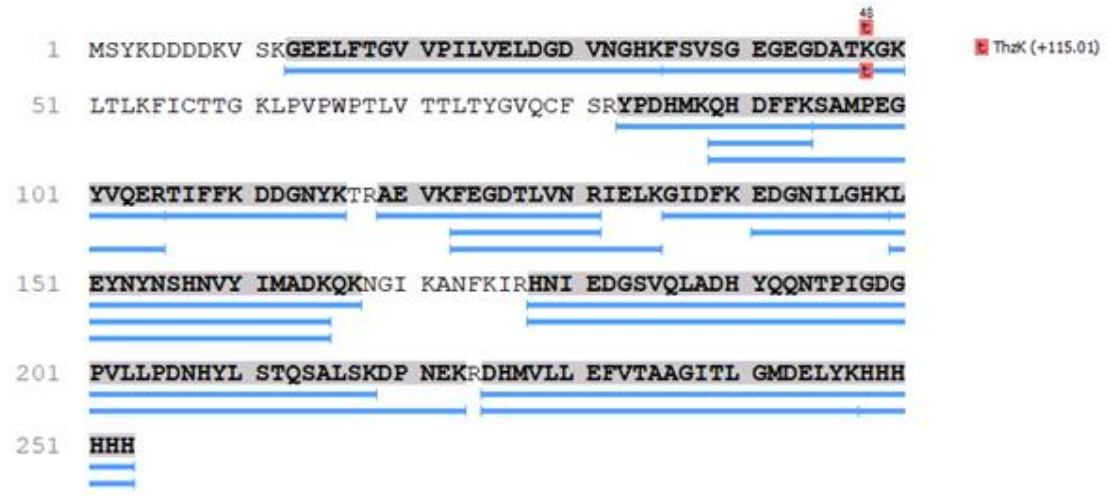


Figure 52: Trypsin digest of **136**, displaying found peptides in blue to map the sequence of the protein. The non-canonical amino acid is confirmed in the correct location, highlighted in red.

3.4.2 sfGFP(N150-(thiazolidine)lysine)-His$_6$

Due to the similarities between the constructs and the proteins, the same protocol was used for the preparation of sfGFP(N150-**86**)-His$_6$ **137**, obtained in a comparable yield of *ca*. 50 mg per litre of culture. Purification by nickel affinity afforded pure **137** as determined by SDS-PAGE (Figure 53).

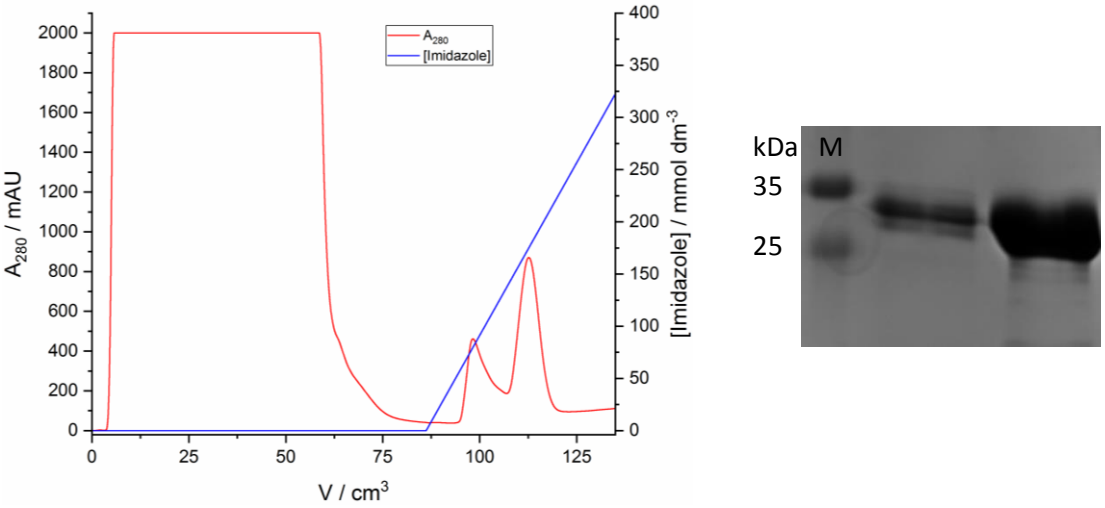


Figure 53: (left) 280 nm chromatogram of the purification of **137** by nickel affinity, eluting with an imidazole gradient; (right) SDS-PAGE of fractions containing **137**: a clear band around the expected 28.0 kDa mass.

Characterisation by ESI-FTICR-MS further confirmed the purity of **137** (Figure 54). Again, a phosphate adduct was observed, as well as an acetonitrile adduct arising from the use of 50:50:1 H_2O:MeCN:FA, but the principle species corresponds to neutral **137** within ± 2 Da. Trypsin digest again confirmed that NCAA **86** was installed exclusively at residue 150 whilst the rest of the protein corresponded to the expected native sequence (Figure 55).

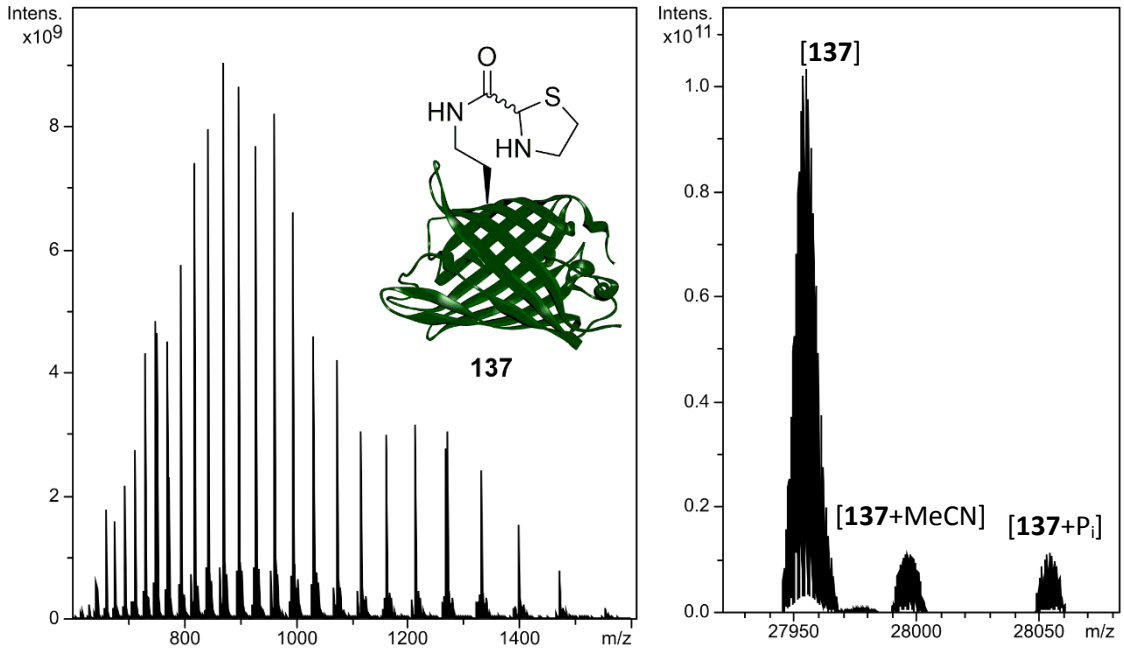


Figure 54: Raw (left) and deconvoluted (right) ESI-FTICR mass spectrum of **137**. For the neutral species, calc. 27957 Da; found 27957 Da.

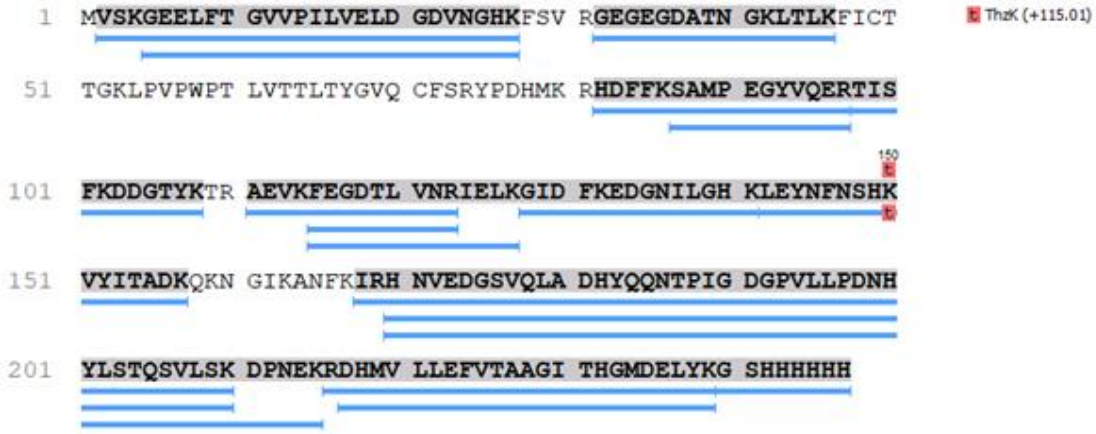


Figure 55: Trypsin digest of **137**, displaying found peptides in blue to map the sequence of the protein. The non-canonical amino acid is confirmed in the correct location, highlighted in red.

3.4.3 sfGFP(N150-(L-threonyl)lysine)-His_6

The same protocol was used for the preparation of sfGFP(N150-**105**)-His_6 **138**. The yield, ca. 18 mg per litre of culture, is strikingly lower than the yield of **137**, suggesting poorer recognition of **105** compared to **86** by $pylRS^{wt}$. Full-length, purified **138** was confirmed by SDS-PAGE (Figure 56).

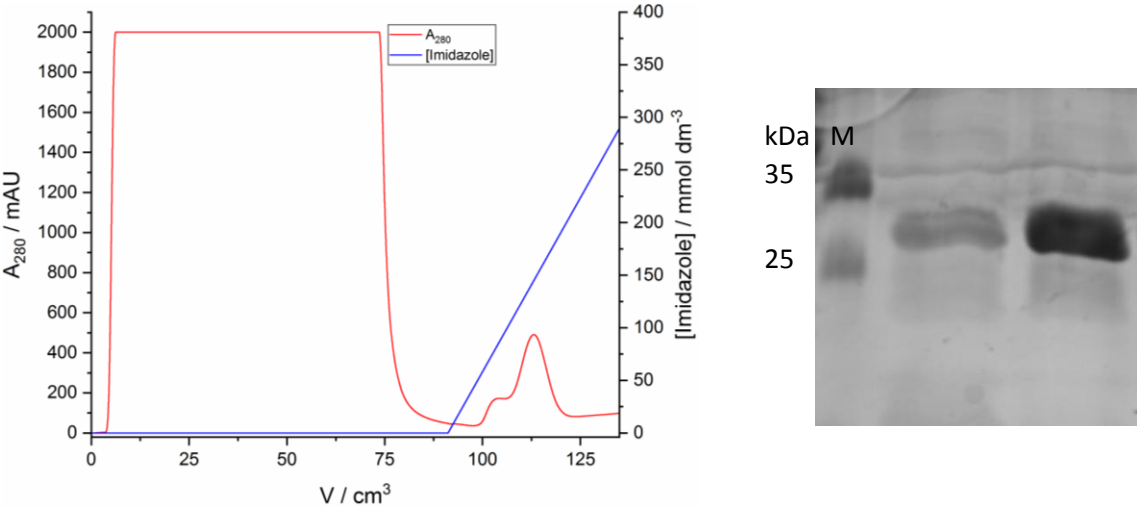


Figure 56: (left) 280 nm chromatogram of the purification of **138** by nickel affinity, eluting with an imidazole gradient; (right) SDS-PAGE of fractions containing **138**: a clear band around the expected 28.0 kDa mass.

ESI-FTICR-MS analysis similarly confirmed that the sample obtained contained exclusively full-length **138**, corresponding to the expected mass to within ± 2 Da (Figure 57). Trypsin digest confirmed that only position 150 corresponded to NCAA **105** (Figure 58). This is the first documented incorporation of **105** in a protein.

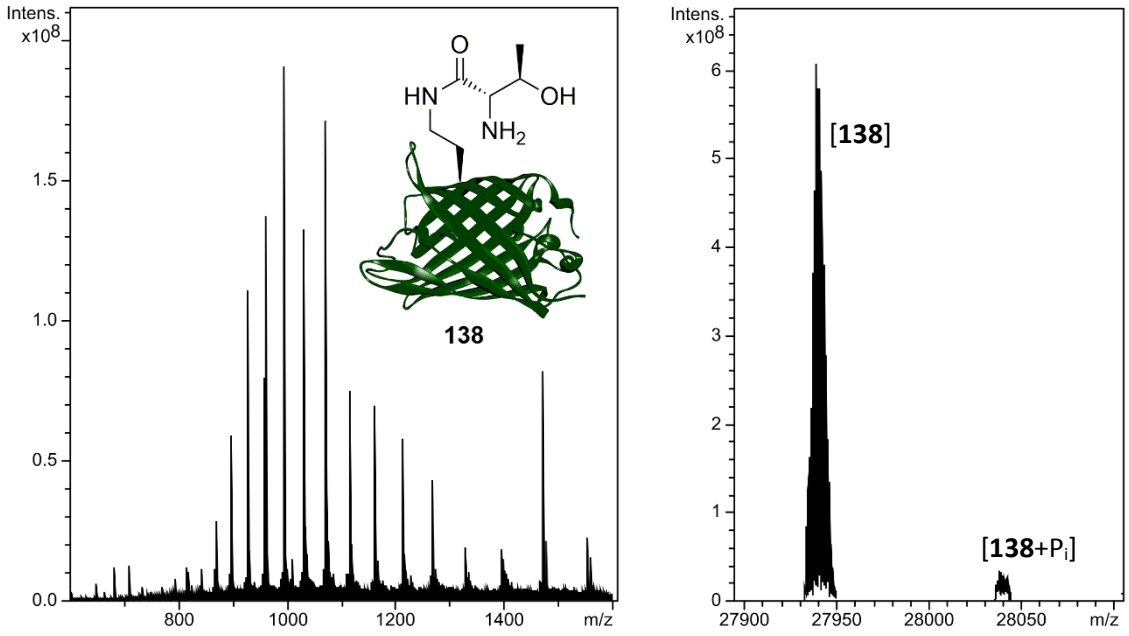


Figure 57: Raw (left) and deconvoluted (right) ESI-FTICR mass spectrum of **138**. For the neutral species, calc. 27943 Da; found 27942 Da.

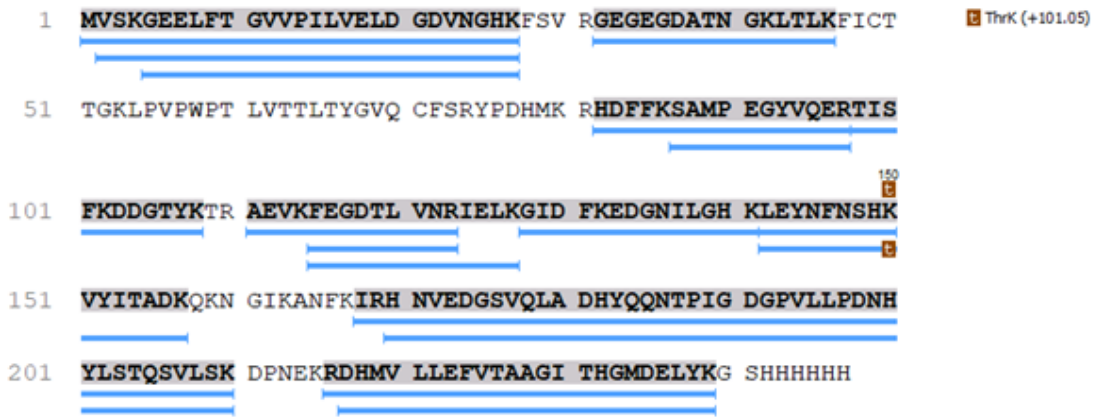


Figure 58: Trypsin digest of **138**, displaying found peptides in blue to map the sequence of the protein. The non-canonical amino acid is confirmed in the correct location, highlighted in brown.

3.5 Conclusions

All six synthesised lysine derivatives were assayed using EGFP as a reporter protein to ascertain their suitability in amber stop codon suppression with two different pylRS variants. The double mutant pylRS^{AF} proved to be unsuitable for use with any of the six lysine derivatives studied, whilst the wild type pylRS was found to turn over **86** and **105**. Both of these NCAAs were successfully incorporated into two variants of GFP containing amber stop codons, provisionally meeting the first of the three objectives of this project.

Chapter 4: Exploring Aldehyde Decaging

A communication has been published based upon Chapter 4.2.[274]

A patent has been approved containing material from Chapter 4.2.

4.1 Introduction

With two forms of caged aldehyde genetically encoded and incorporated into test proteins, attention turned to developing methodologies for exposing the protein aldehyde. 1,2-aminoalcohols and 1,1-aminothioethers do have a usable body of documented reactions within the organic chemistry domain,[78, 275, 276] acting as a helpful starting point. The principle challenge is adapting this reactivity for the biological world, as chemistry offering minimal perturbation outside of the moiety of interest. Careful tuning of conditions is required to balance the rapidity and conversion of the desired reaction with that of any side reactions or other unwanted processes.

4.2 Thiazolidine decaging

One method has been published for decaging thiazolidines to unmask glyoxyl aldehydes on proteins,[221] in which the protein thiazolidine is treated with 50 eq. silver acetate in 10% (v/v) aqueous acetic acid. The proposed mechanism begins with coordination of the thiazolidine sulfur atom to a silver(I) ion, followed by departure of the silver-sulfur complex with imine formation stabilising C-2 (Figure 59). In aqueous solution, this species hydrolyses rapidly to form the glyoxyl aldehyde. The fatal flaw in this method is the use of 10% (v/v) acetic acid, presumably to accelerate imine hydrolysis and prevent reformation of the imine through protonation of the resulting silver-cysteamine complex. The working pH, below 3, is unsuitable for a large proportion of the proteome, demonstrating the need for a more biologically compatible method. Furthermore, this method is incompatible with buffers containing chloride, such as NaCl commonly used to stabilise proteins, due to the precipitation of AgCl. When this procedure was performed on thiazolidine **136**, the EGFP precipitated instantly and no portion remained soluble, preventing any detection by mass spectrometry. The thiazolidine may have been decaged, but the inability to detect or use the subsequent protein aldehyde limits the utility of this method.

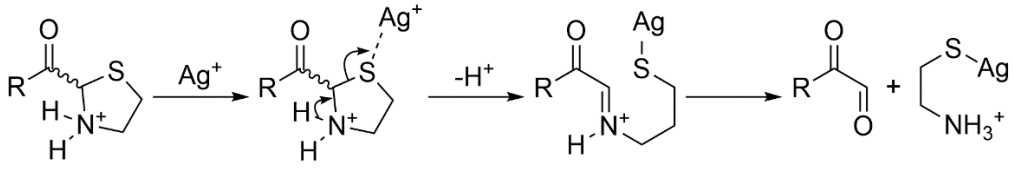


Figure 59: Proposed mechanism for the silver- and acid-mediated ring opening of a thiazolidine to afford a glyoxyl aldehyde and silver-cysteamine complex.

The combination of a Lewis acid, Ag^+, and a Brønsted-Lowry acid, AcOH, initially seems intuitive: unmasking the electrophilic aldehyde by capturing the nucleophilic heteroatoms with which it has condensed to form the thiazolidine, pairing the soft sulfur atom with soft Ag^+ and the hard nitrogen atom with H^+. The use of a superior Lewis acid would putatively

make the Brønsted-Lowry acid redundant, preventing thiazolidine reformation through tight coordination of the cysteamine by-product to a Lewis acid whilst foregoing the use of protein-denaturing low pH conditions. Whilst mercury salts have seen use for synthetic applications of thiazolidine ring opening without the need for acids,[277] their use was ruled out owing to toxicity and potential off-target reactions with other protein sulfur atoms. A similar strategy was trialled on thiazolidine peptides, where the glyoxyl aldehyde peptide was liberated using a palladium complex at neutral pH.[278] Palladium has been used for other decaging strategies, such as with propargyl and allyl protecting groups as previously discussed (Figure 23),[156] including use with live cells to demonstrate a low level of cytotoxicity. Hence palladium reagents stood out as candidates to be effective thiazolidine decaging reagents.

4.2.1 Palladium reagent screen

For the decaging reaction of thiazolidine **136** to afford protein glyoxyl **139**, palladium reagents **140**-**144** were screened (Figure 60). These reagents were selected following previous documented use in aqueous conditions or bioconjugations, demonstrating biocompatibility and reactivity to an appreciable extent.[156, 199, 279, 280]

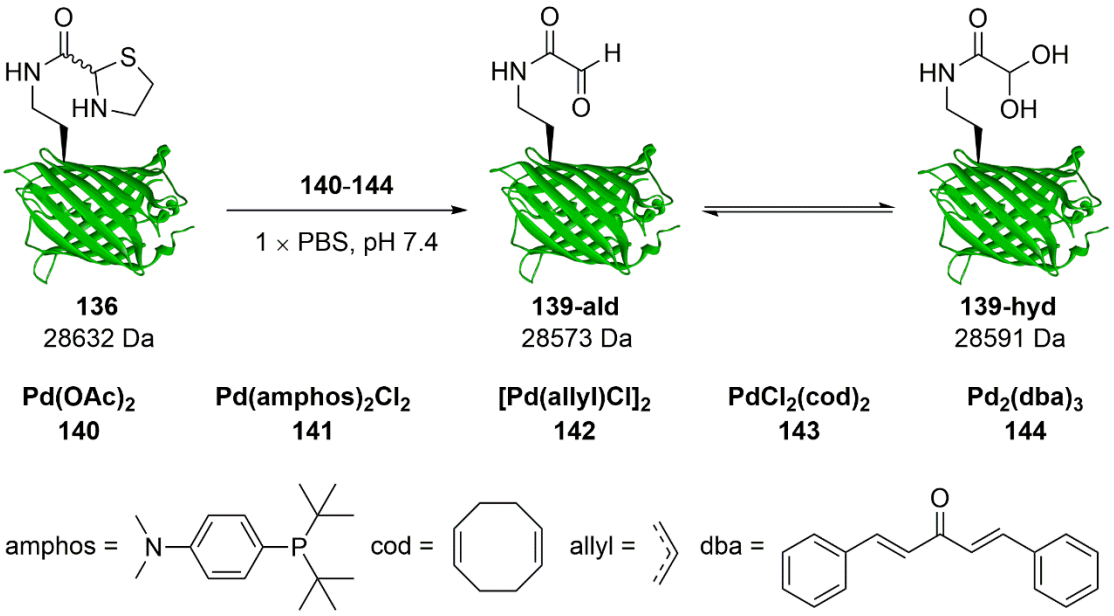


Figure 60: Palladium complexes screened for the decaging of protein thiazolidine **136** to afford a protein glyoxyl as aldehyde **139-ald** or hydrate **139-hyd**.

Each reagent was trialled with **136**, varying only the equivalents of each complex and the reaction time, after which the reaction mixture was quenched with DTT and purified by desalting. The extent of decaging was quantified by ESI-FTICR-MS (Table 1), comparing the deconvoluted peaks corresponding to **136** and **139** (Figure 61). Where a palladium reagent has led to protein precipitation, quantification data are based on any

remaining soluble protein, as the protein precipitate is removed from solution during the desalting process so no protein signal can be observed. This has led to some missing values in the table, where insufficient protein remains in solution for detection by MS. Conversion is estimated through a comparison of the deconvoluted peak cluster areas corresponding to protein starting materials and products on the assumption that all protein species ionise to comparable extents. Owing to sources of random error in the measurement of the peak intensities, problems with peak resolution, and the algorithm used for peak deconvolution, estimated conversions are quoted to the nearest 10% to cautiously take account of the limit of precision of this estimation method. Conversions are based upon single data points rather than replicates, as the purpose of this screen is to qualitatively identify promising palladium reagents for thiazolidine decaging, with subsequent optimisation to increase the conversion where needed, rather than to precisely quantify the decaging ability of each palladium reagent screened.

Table 1: Palladium reagents **140**-**144** at the indicated stoichiometric equivalents screened for the decaging of protein thiazolidine **136** (final protein concentration 310 μM in 1 × PBS) over the indicated time interval at 37 °C, before quenching with DTT and isolating the protein by centrifugation and desalting. Decaging conversion quantified by ESI-FTICR-MS. Protein precipitation judged by visual observation. “nd”: not determined.

Palladium complex	Equivalents of complex	t / h	% decaged	No protein precipitation
140	100	6	nd	
140	100	24	nd	
141	100	6	0	✓
141	100	24	0	✓
142	100	6	30	
142	100	24	nd	
142	1	1	100	✓
143	1	1	10	✓
143	100	1	nd	
144	100	1	0	✓
144	100	6	40	✓
144	200	6	40	✓
144	100	24	80	✓
144	100	36	90	✓
144	100	40	100	✓

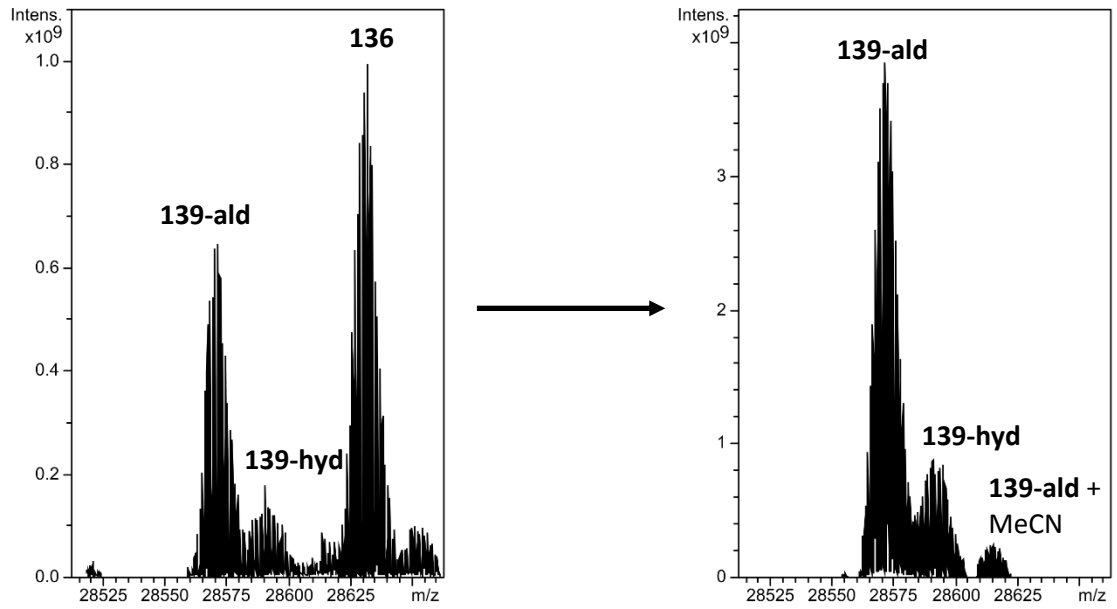


Figure 61: From partial to full decaging using **142**. Decaging with 100 eq. for 6 h (left) led to only *ca*. 30% decaging of **136** due, at least in part, to protein precipitation, whilst a single equivalent for one hour (right) led to complete decaging to form **139-ald** (calc. 28573 Da, found 28573 Da) and **139-hyd** (calc. 28591 Da, found 28592 Da), predominantly existing as the aldehyde rather than the hydrated diol.

Incubation of **136** with **140** led to instant protein precipitation, whilst **141** appeared to have no effect upon either the thiazolidine or the protein. This is likely due to the size of the bulky amphos ligands, conferring some aqueous solubility but hindering the palladium centre from reaching the thiazolidine given the surrounding protein mass. Precipitation was initially observed with **142**, although this must have only been partial precipitation as sufficient quantities remained in solution to quantify the partial decaging of the thiazolidine. Lowering the equivalents of **142** not only avoided protein precipitation but also led to 100% conversion within one hour. A similar strategy achieved a modicum of success with **143**: lowering the equivalents from 100 to 1 led to no protein precipitation, but the decaging observed was only minor. For reagent **144**, altering the time proved to be more effective than altering the equivalents used. Doubling the equivalents offered no improvement to decaging, whilst increasing the time to nearly two days afforded complete conversion (albeit at a very slow pace). Whilst **144** can be used to decage the thiazolidine to completion, use of **142** will lead to the same positive result on a much shorter timescale. For this reason, further thiazolidine decaging work makes use of **142**.

Reagent **144** stands alone in this table as the only Pd^0 species, whilst reagents **140**-**143** are sources of Pd^{II}. This is likely a factor in the unusual behaviour of this reagent for decaging compared to the other reagents screened. Whilst soluble in DMSO, **144** is known to be almost completely insoluble in water.[281] When prepared as a stock in an organic solvent, e.g. THF, and diluted into water, microcrystals of **144** are known to form spontaneously,[282] forming aggregates due to the poor aqueous solubility; Pd

nanoparticles are also formed, but to a much lesser extent and this arises from impurities in **144** rather than the reagent itself.[283] The decaging activity likely arises from exposed palladium centres on the surface of such microparticles, which is slowly lost as the exposed surface area decreases due to aggregation. This heterogeneous mechanism is profoundly less efficient than the homogeneous mode afforded by **142**, a far more active reagent for the decaging of **136**.

4.2.2 Optimised procedure

With the palladium reagent selected, further work sought to probe the optimal conditions required for this decaging: retaining complete conversion in the shortest time through a streamlined procedure. The DTT quench had been adapted from peptide chemistry and, whilst successful in removing palladium from the solution as judged by mass spectra, required a large excess of DTT and sometimes extensive centrifugation to clarify the reaction mixture.[278] An alternative quench, reported for use with palladium decaging of protein propargyl carbamate groups,[156] required only addition of 3-mercaptopropanoic acid as a stock solution followed by desalting, alleviating the need for centrifugation. This procedure is reported as a highly effective quench in combination with desalting, justified with ICP-MS quantification of the palladium retained in solution, and is practically more straightforward and reliable than the use of DTT. This quench was used in subsequent decaging trials.

Initial reagent screens had been conducted at 37 °C to maximise the chance of discovering any reactivity from the decaging reagents trialled. Lower temperatures would broaden the scope of this method, with exposure to lower temperatures generally preferred when handling to inhibit denaturation. At 7 °C, the *ca*. 10% conversion observed was markedly poorer than at 37 °C, but at 19 °C, 100% conversion was still observed. As the higher temperature served no functional purpose, further decaging work was performed at carefully monitored room temperature.

The palladium reagent **142** had initially been delivered as a 100 × stock in DMSO during reagent screening, where the solubility of all complexes in DMSO allowed for a consistent method of delivery. The role of this cosolvent was explored by comparison with DMF, 1,4-dioxane and MeCN (Figure 62,

Table 2), with water being unsuitable due to the insolubility of **142** at the required 30 mM concentration. The use of DMF and 1,4-dioxane solutions led to slightly poorer decaging, whilst MeCN led to almost the same decaging activity exhibited using DMSO. Nevertheless, DMSO was still shown to be the most effective solvent for the aldehyde decaging procedure.

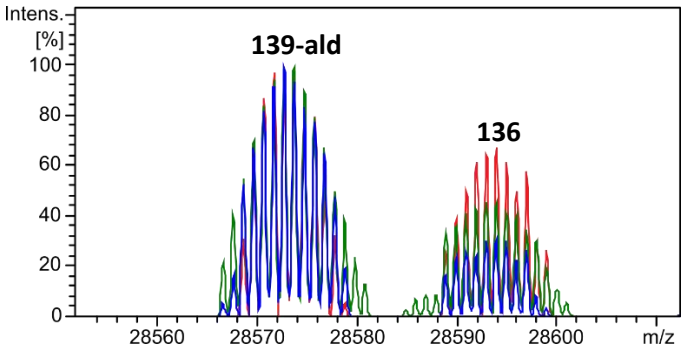


Figure 62: Decaging of **136**, as observed by MS, using **142** prepared in either MeCN (blue), DMF (green) or 1,4-dioxane (red). Spectra are normalised relative to the peak for **139-ald**.

Table 2: Solvent effect in the decaging of protein thiazolidine **136**.

Solvent	% decaged
DMSO	100
MeCN	90
DMF	80
1,4-dioxane	60

With the stock solution, temperature and quench improved, the optimised procedure was trialled using the second thiazolidine construct **137** to afford protein glyoxyl **145** (Figure 63). Pleasingly, the same conditions led to complete decaging as expected following a straightforward protocol. One limitation of this method of monitoring decaging is the presence of MeCN adducts, where the mass of **145-hyd** + MeCN is indistinguishable (within 1 Da) from that of starting material **137**. This is problematic for protein substrates where the glyoxyl aldehyde exists mostly as the hydrate, as is the case with **145** but not with **139**, perhaps fortuitously given the difficulties this would have posed when attempting to quantify decaging during reagent screens. In this instance, the putative MeCN adduct is so low in abundance relative to the main peaks of interest that decaging does appear to have reached completion. Further verification of this will take place through chemical modification of this site.

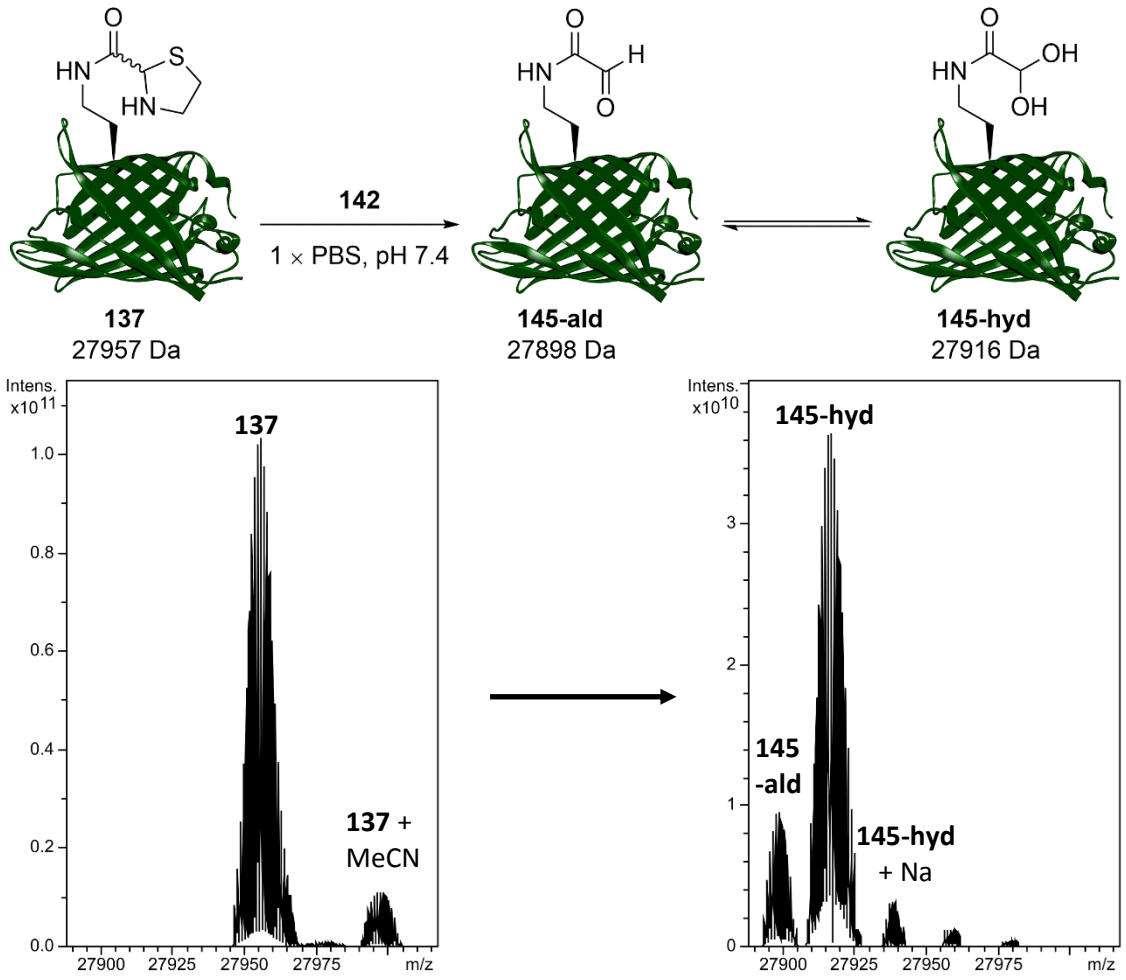


Figure 63: Decaging of protein thiazolidine **137** to afford protein glyoxyl **145** (top). ESI-FTICR-MS data show the consumption of **137** (lower left; calc. 27957 Da, found 27956 Da) to afford **145-ald** (lower right; calc. 27898 Da, found 27900 Da) and **145-hyd** (calc. 27916 Da, found 27918 Da).

4.2.3 Aldehyde versus hydrate

One notable difference is observed between the mass spectra of protein glyoxyl species **139** and **145**: the balance of the aldehyde versus the hydrate. In aqueous solution, glyoxyl aldehydes are so electrophilic that nucleophilic water attacks the glyoxyl to form a more stable hydrate species. This is observed with **145**, where the glyoxyl exists predominantly as the hydrate, in line with expectations.[231] However, the mass spectrum of **139** depicts the opposite scenario, where the aldehyde is predominantly not hydrated. Given the high degree of similarity between the EGFP and sfGFP scaffolds, the most likely explanation is the residue microenvironment.

In both sfGFP and EGFP, residue N150 is retained and is positioned in the middle of a β-sheet. At this location, the principle interactions responsible for maintaining protein fold involve the main chain of this residue and not the side chain, which is exposed to solvent and has a reasonable degree of freedom. This is why position 150 is a suitable site for installation of an NCAA, where the side chain has little involvement with other residues

145-hyd + MeCN

and can be altered with minimal structural perturbation whilst being available for chemical reactivity. Position 39 is markedly different, however. This residue was one of the key mutations leading to the sfGFP scaffold owing to the increased stability of the protein upon the Y39N mutation, as mentioned earlier. In EGFP, position 39 lies on a β-turn, but in sfGFP the effect of the single Y39N mutation is to tighten this β-turn into a more stable 3_{10}-helix, leading to a substantial improvement in fold stability (Figure 64).[284] Here, the side-chain carboxylate of D36, at the *N*-terminal end of the 3_{10}-helix, interacts with the N39 side chain amide. This has the effect of bringing these two residues closer together compared to the same site in EGFP, with the effect of forming a tighter turn at the end of the β-sheet. Residue 39 is indeed exposed to solvent, as seen in the crystal structure and through the documented reactivity of NCAAs installed at this site,[169, 170] but the participation of its side chain in protein-stabilising interactions need also be considered.

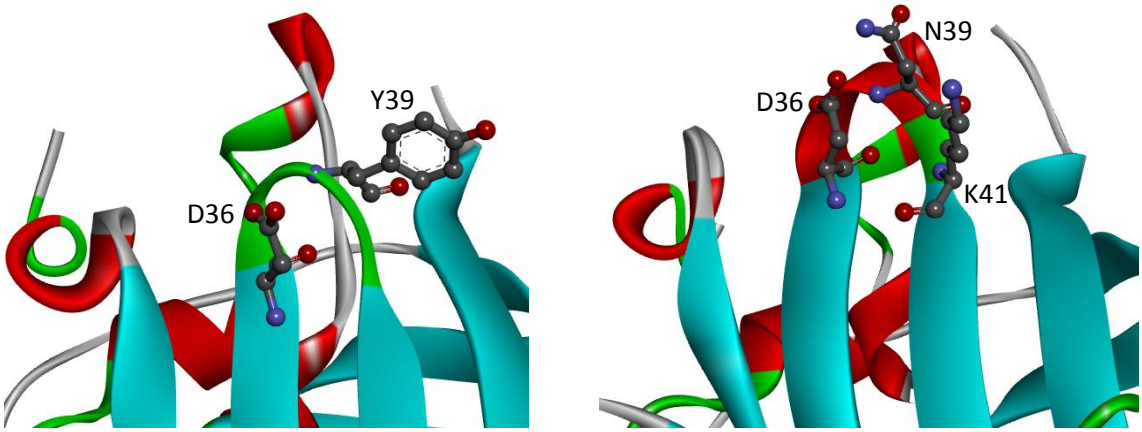


Figure 64: EGFP residue Y39 forms part of a β-turn (green) with little side-chain interaction (left, PDB: 2Y0G). This contrasts with sfGFP residue N39, where the residue has rotated to position the side chain amide to hydrogen bond with the side chain carboxylate of residue D36 (right, PDB: 2B3P). This tighter geometry takes the form of a 3_{10} helix rather than a β-turn, bringing together two strands of the β-barrel and stabilising the protein fold.

Whilst this difference in the mass spectra between **139** and **145** initially appears to be merely superficial, there is an important consideration to be made when attempting to modify **139**. Is this difference a limitation of the analytical method or are the two glyoxyl aldehydes chemically different? Has the microenvironment around **139** residue 39 stabilised the glyoxyl aldehyde to the point of diminished reactivity? From a broader perspective, this acts as a reminder of the considerations required for judicious placement of NCAAs using amber stop codon suppression. The primary concern is typically whether the NCAA may the protein, but clearly there is a need to consider how the rest of the protein may perturb the NCAA. This is not always obvious from crystal structures alone; had sfGFP not been developed, the full extent of the involvement of the side chain at position 39 in GFP fold stabilisation would not have been known.

4.2.4 Mechanism

Whilst the primary objective of the decaging procedure is liberation of a reactive glyoxyl aldehyde, some discussion of the mechanism is worthwhile for the purposes of method optimisation. Given that palladium reagent **142** can effect thiazolidine decaging without the need for a low pH, its superiority over the previous AgOAc strategy seems clear. This is likely due to the different geometries of complexation available to Ag^+, typically forming linear complexes, versus Pd^{2+}, capable of forming linear and square planar complexes. When cysteamine is released from the thiazolidine, the amino group can also coordinate to the Pd^{2+} centre to form a more stable chelated complex, tightly trapping the cysteamine. This is not possible with Ag^+, requiring high H^+ concentrations to protonate the cysteamine and prevent thiazolidine reformation.

The use of palladium can be something of a double-edged sword, however. The reagent must possess a careful balance between affinity for the thiazolidine versus affinity for the rest of the protein, where the metal may facilitate pathways such as protein denaturation or aggregation through coordination to cysteine thiols or methionine thioethers. For the palladium reagents screened, this is reflected in the number of equivalents required. For reagent **142**, an affinity for decaging is clear, but the protein precipitation seen at high equivalents clearly demonstrates the need to tailor the decaging method to the protein to achieve the correct balance, a common theme in biorthogonal chemistry.

The ligands of the palladium complex are evidently a factor in the reactivity of each complex. **140** only led to protein precipitation and its effect on decaging is unknown, due to the inability to quantify the decaged protein by mass spectrometry. **141** appeared to be inert, putatively due to the bulky ligands as mentioned. At low equivalents, very little decaging was observed with **143**, likely owing to the ligand: the 1,5-cyclooctadiene ligand chelates to the Pd^{2+} centre and serves to block off a face of the complex, providing steric bulk that inhibits nucleophilic attack by the thiazolidine sulfur atom.

The success of **142** is likely due to the dinuclear complex dissociating in the stock solution, on account of the differences in decaging observed when altering the solvent used for this purpose. Mononuclear complexes of palladium with DMSO, MeCN and DMF are well known, where the coordinating solvent acts as a suitable ligand.[285] The trend seen in

Table 2 suggests some influence of the coordinating atom on this process, in the order S > N (free) > N (hindered) > O. Whilst **142** is prepared as a dinuclear species and delivered as one equivalent, in the reaction mixture two equivalents of mononuclear Pd^{2+} are available to react with the thiazolidine. The mononuclear complex formed from **142** is likely composed of small ligands that do not chelate, such as a $PdCl_2L_2$-type complex

where L = coordinating solvent, minimising the bulk around the Pd^{2+} centre. This is one factor behind the enhanced reactivity of this $PdCl_2L_2$ complex towards protein thiazolidines compared to analogues **141** and **143**, in addition to the extra equivalent of Pd^{2+} present.

With this considered, a mechanism for this decaging reaction is proposed based upon Figure 59. A Pd^{2+} complex coordinates to the thiazolidine sulfur atom and causing ring opening,[286] followed by imine hydrolysis and coordination of the cysteamine amino group to form a Pd(cysteamine)L_2 chelate (Figure 65).

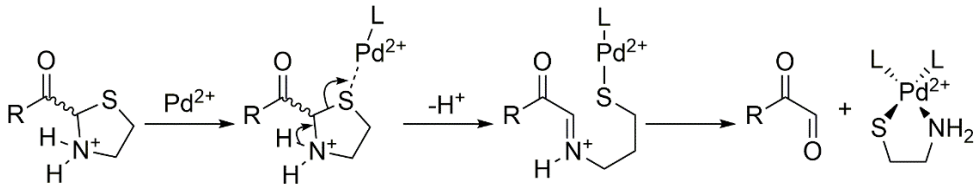


Figure 65: Proposed mechanism for the palladium-mediated ring opening of a thiazolidine to afford a glyoxyl aldehyde and silver-cysteamine complex. L = coordinating solvent ligand, e.g. DMSO.

4.3 Threonine oxidation

With one aldehyde-decaging method in hand with the thiazolidine **86**, attention turned towards the threonine derivative **105**. The 1,2-aminoalcohol motif led to periodate oxidation as the most obvious route due to the general reliability and body of work behind this methodology.

4.3.1 Condition screen

A procedure for oxidising the mutated *N*-terminal serine residue of EGFP has been optimised and reported by Dr Richard Spears.[244] These conditions were used as a starting point for the oxidation of **138** to form protein glyoxyl **145**, using six eq. $NaIO_4$ and 18 eq. L-methionine as a sacrificial reducing agent to prevent off-target oxidation (Figure 66). These conditions led to 60% conversion within four minutes of oxidation: a promising start, but apparently a more challenging substrate to oxidise than an *N*-terminal serine or threonine residue.

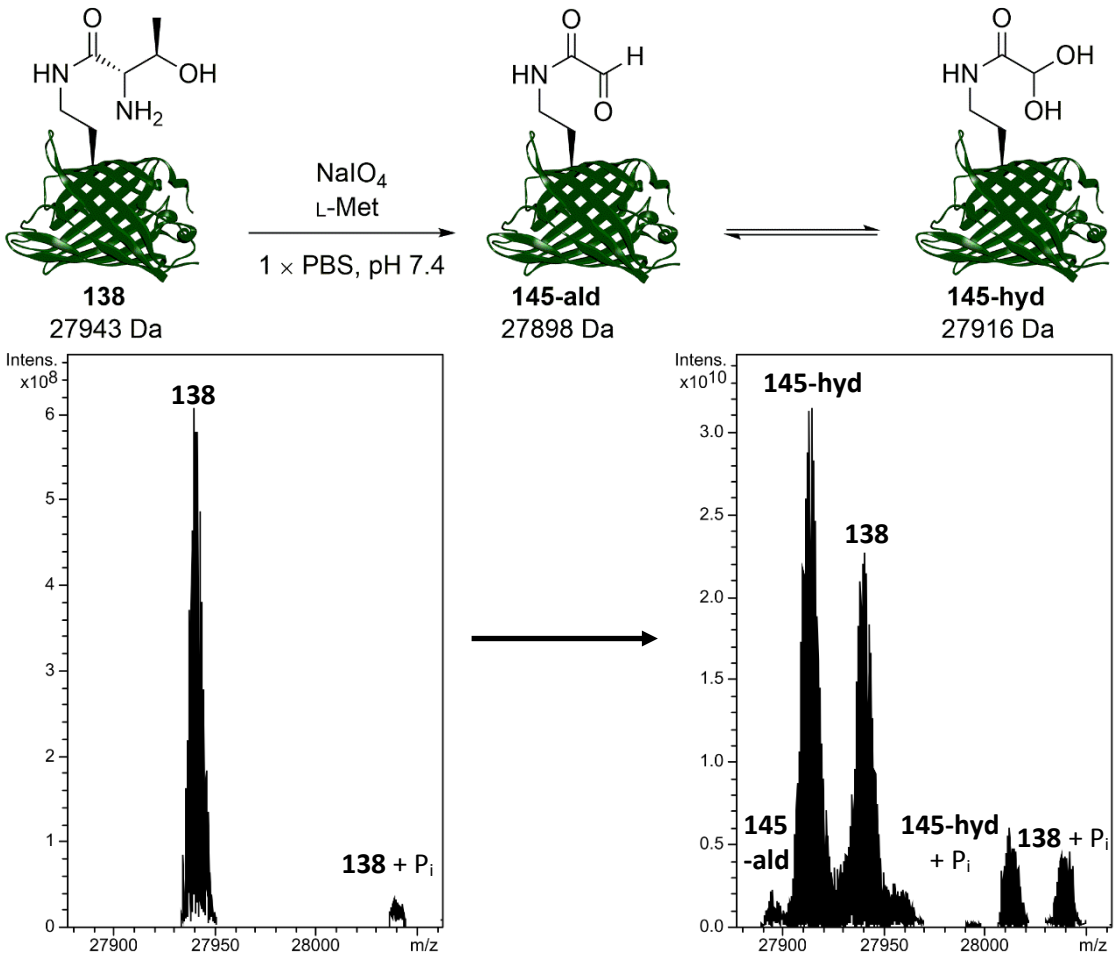


Figure 66: (top) Oxidation of protein 1,2-aminoalcohol **138** to afford protein glyoxyl **145**. ESI-FTICR-MS data show some consumption of **137** (lower left; calc. 27943 Da, found 27942 Da) to afford **145-ald** (lower right; calc. 27898 Da, found 27897 Da) and **145-hyd** (calc. 27916 Da, found 27916 Da) to approximately 60% conversion.

Attempts to increase this conversion were largely unsuccessful. Longer reaction times, from four to eight or twelve minutes, led to no improvement, with conversion stalling at 40-60%. Increasing the quantity of $NaIO_4$ to ten eq. (and correspondingly L-methionine to 20 eq.) led to overoxidation of the protein after eight minutes, with additional +16 peaks appearing in the mass spectra of the oxidised proteins. One consideration was whether a by-product of periodate oxidation, in this case acetaldehyde, was reacting with the exposed protein glyoxyl aldehyde **145** in a crossed-aldol reaction to afford product **146**, differing in mass from **138** by only 1 Da (Figure 67). Indeed this "accidental aldol" led to the reported OPAL method for modifying protein aldehydes.[287] This was investigated through trypsin digest, as the peptides produced can be characterised to within 0.1 Da, sufficient to discriminate between **145** and **146**, as well as through the use of peptide fragmentation. Through this peptide mapping, the presence of **145-ald** and **145-hyd** was confirmed in a qualitative manner, as was the presence of starting material **139**, but no

evidence of the aldol by-product **146** was found. Oxidation was indeed proceeding but hampered by some unknown factor.

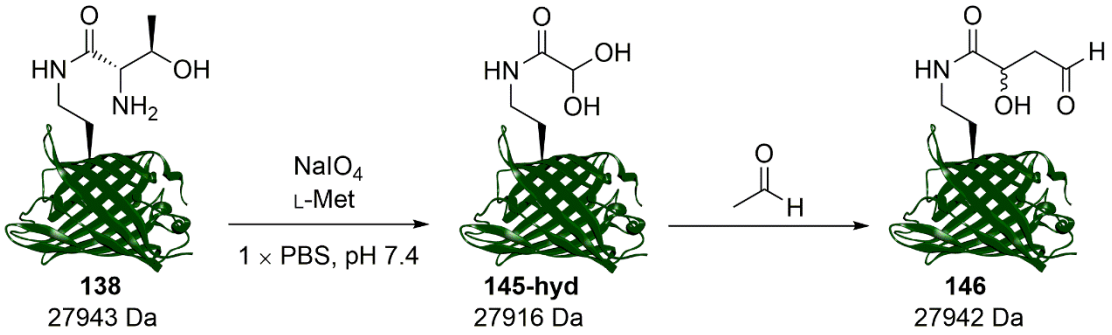


Figure 67: Putative aldol interception of **145** by the acetaldehyde released from periodate oxidation, leading to unwanted side product **146**, indistinguishable from **138** by intact protein mass spectrometry alone.

4.3.2 Optimised conditions

Reviewing literature protocols led to closer inspection of the buffer composition used. Sodium phosphate buffers are commonplace, as is the use of NaCl, but potassium ions are seldom encountered. This was the one component of PBS that stood out as suspicious. One hypothesis is that potassium cations interact with the periodate anion and can precipitate as potassium periodate, poorly soluble in water (*ca*. 8 mM) at the typical 0 °C of periodate oxidation, as was observed in early applications of periodate oxidation.[288] Given that periodate was only used in 5 eq., even minor precipitation of periodate could noticeably affect the extent and rate of oxidation.

A batch of **138** was prepared by Dr Tessa Keenan under potassium-free conditions, substituting PBS with 20 mM sodium phosphate buffer, 150 mM NaCl, pH 7.4. This composition approximates the buffering capacity and ionic strength of PBS but without any potassium. With conditions of 100 µM **138**, 5 eq. $NaIO_4$, and 10 eq. L-methionine, oxidation reached completion in four minutes (Figure 68). **145** exists almost completely as the hydrate and a minor extent of *N*-terminal methionine cleavage is also seen accompanying oxidative cleavage. The effect of removing potassium, seemingly an innocuous buffer component, is easily overlooked yet remarkably profound here.

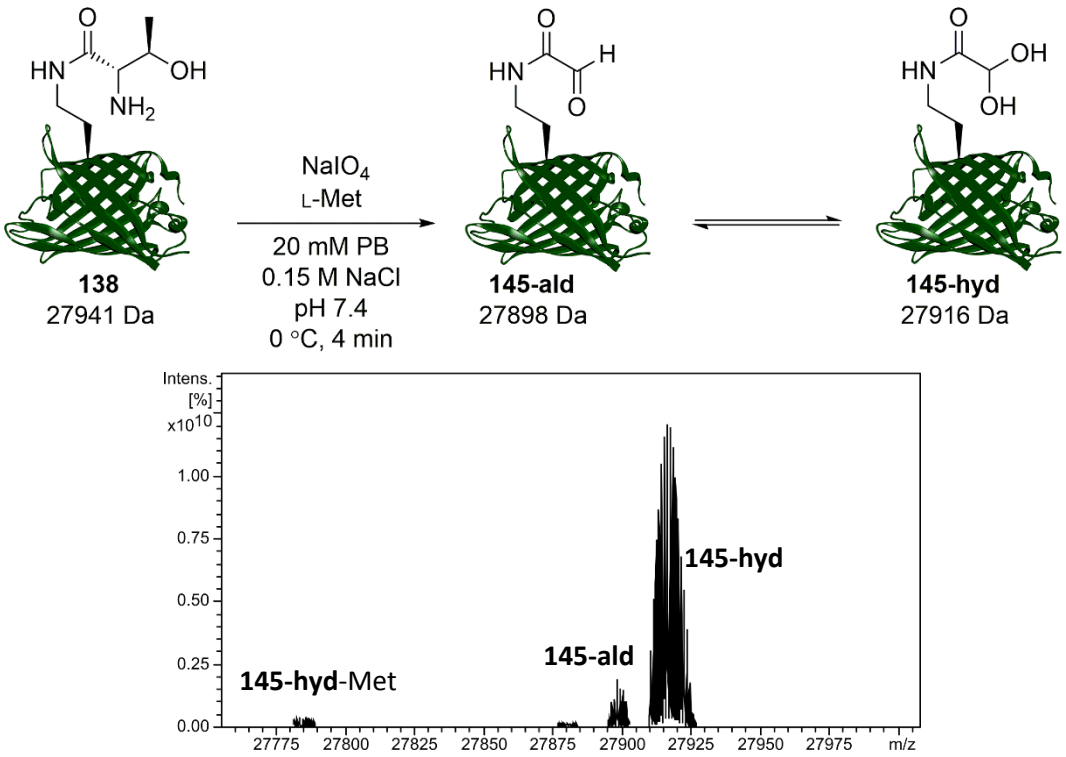


Figure 68: Optimised periodate oxidation of **138** affords **139**, predominantly as the hydrated aldehyde (calc. 27916 Da, found 27918 Da).

4.4 Beyond the GFP scaffold

Thus far GFPs have served as handy test proteins for trialling new methodologies for aldehyde decaging. To be truly of use however, these new strategies must be applicable to as substantial a proportion of the proteome as possible; not all proteins are as well behaved as GFP. A further substrate was sought to serve as a testbed and an enhanced challenge for aldehyde decaging.

4.4.1 Introducing BiGalK

Whilst this project was underway, in the same laboratory Dr Tessa Keenan was undertaking extensive work with enzymatic glycosylation. One particular protein enjoying particular usage was *Bifidobacterium infantis* galactokinase, BiGalK, which catalyses the phosphorylation of galactose to form galactose-1-phosphate and can be produced using recombinant methods in *E. coli* in appetisingly high yields.[289] Furthermore, when a sample was submitted for ESI-FTICR-MS analysis, this protein was remarkably straightforward to observe using this technique. One previous protein substrate, *Bacillus circulans* xylanase (Bcx), had proven to be far less amenable to study by ESI-MS at concentrations below 1 mM and required MALDI-MS, which is unsuitable for detecting intact protein mass changes of 40-60 Da from decaging. Given these positive practical points, BiGalK was taken on as a further test protein to broaden the scope of the newly

developed aldehyde decaging methods. Notably, the *N*-terminal residue of BiGalK is a threonine residue, presenting an opportunity for multiple modifications using orthogonal aldehyde strategies.

Whilst the kinetics, sequence and expression conditions of BiGalK were reported, at that time no crystal structure of BiGalK had been obtained. Structural understanding of BiGalK, including active site composition, was largely restricted to sequence analysis and comparison with homologues. This was an impediment when considering where exactly the amber stop codon should be installed in BiGalK, with little information on which residues would be exposed on the protein surface or potentially involved in enzymatic activity. Residue K417, located almost at the *C*-terminus of this 423-residue protein, was selected as its position and hydrophilic properties suggested that this residue would be exposed on the surface of the protein. More recently, Dr Tessa Keenan has obtained crystal structures of BiGalK that substantiate this claim. The *C*-terminus of BiGalK is highly disordered: the final residue with defined electron density is K417, and even the side chain of this residue is too disordered to be defined (Figure 69). Despite this, the *C*-terminal tail appears to be exposed to the solvent rather than buried within the protein. Whilst the zeal for using an unexplored protein was a helpful motivator in driving this research, having access to this structure would have greatly informed the rational mutation design process and would perhaps have permitted a more ambitiously located residue to have been selected for mutation. Nevertheless, using site-directed mutagenesis residue K417 was mutated to the amber stop codon to afford the desired BiGalK(K417TAG)-His_6 construct ready for production.

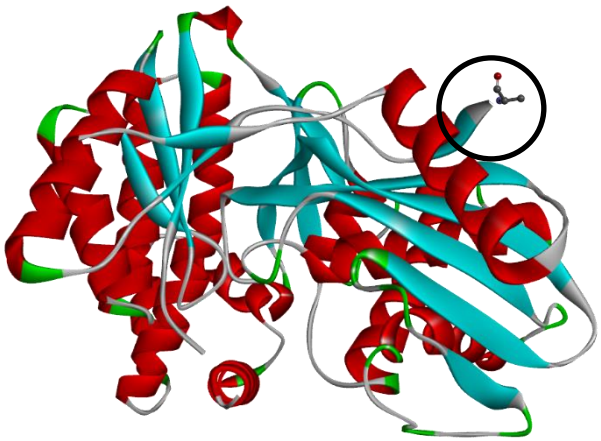


Figure 69: Preliminary crystal structure of BiGalK, provided by Dr Tessa Keenan (unpublished), with residue K417 highlighted.

As the *N*-terminal residue of BiGalK is threonine, the thiazolidine was first selected for installation by amber stop codon suppression, in order to unequivocally demonstrate single aldehyde decaging and obtain a homogeneous protein scaffold. The production

of thiazolidine-containing BiGalK(K417TAG)-His_6 **147** required little alteration from the protocol for the production of wild-type BiGalK, with only an additional induction for the pBAD vector and addition of NCAA **86** required. An expression trial found comparable high yields of BiGalK across several timings for these protocol additions, demonstrating little sensitivity to when the pEVOL plasmid was induced or when NCAA was added. A preparative scale protocol to produce **147** was thus designed from the reported GFP(TAG) and wild-type BiGalK protocols. Purification by Ni affinity afforded *ca*. 30 mg of **147** per litre of culture, lower than the 80 mg / L culture for wild-type BiGalK or 40 mg / L culture of thiazolidine-containing EGFP **136** but still a high yield providing a usable quantity of protein.

Characterisation by ESI-FTICR-MS confirmed the presence of full-length **147** containing **86** in place of a lysine residue (Figure 70). This spectrum was complicated by the presence of multiple sodium, MeCN and phosphate adducts, but the masses found match those expected. Disappointingly, the *N*-terminal methionine residue was present in over 90% of the protein, complicating the MS data and precluding the possibility of a dual aldehyde decaging strategy as Thr1 was not available for oxidation. Nevertheless, a new protein thiazolidine was at hand ready to trial aldehyde decaging.

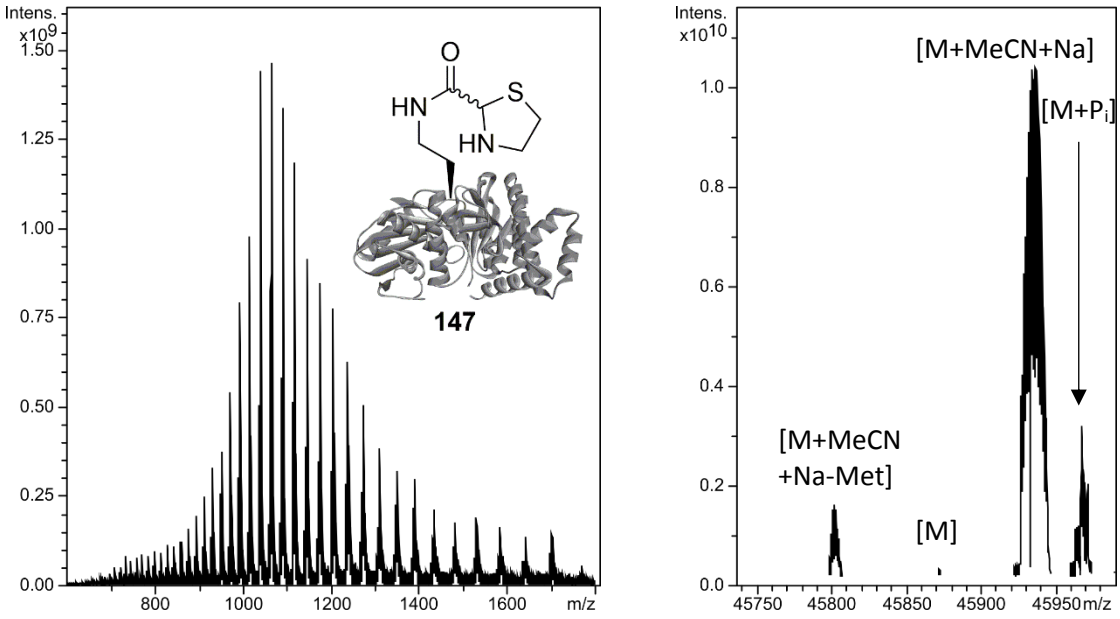


Figure 70: Raw (left) and deconvoluted (right) ESI-FTICR mass spectrum of **147**. For the dominant species, a sodiated MeCN adduct, calc. 45936 Da; found 45936 Da. For the methionine-cleaved species, calc. 45805 Da; found 45804 Da.

4.4.2 Decaging BiGalK

The first trial of decaging protein thiazolidine **147** to afford protein glyoxyl **148** was a reproduction of the conditions established using GFPs, with 300 μM of both **147** and palladium reagent **142**. Disappointingly, negligible conversion was observed. Increasing

the concentration of **142** led to protein precipitation, so instead the protein concentration was decreased, effectively raising the equivalents of **142** without raising the working concentration to a point where protein denaturation occurred (Table 3, Figure 71).

Table 3: Optimisation of concentrations of protein thiazolidine **147** and palladium reagent **142** required to achieve full decaging.

[**147**] / μM	[**142**] / μM	Eq. **142**	% conversion
300	300	1.0	0
200	300	1.5	50
100	150	1.5	0
75	300	4.0	100

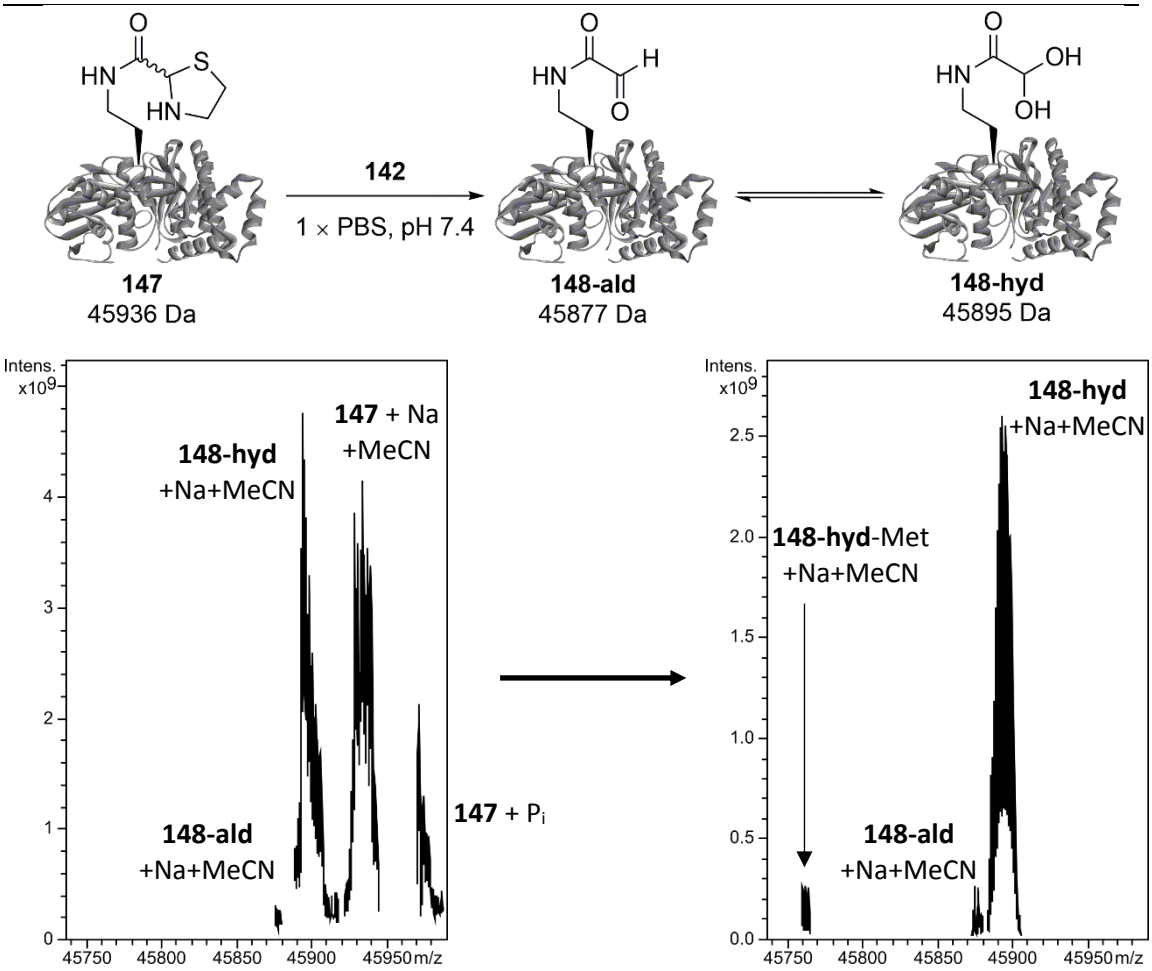


Figure 71: Decaging of protein thiazolidine **147** to afford protein glyoxyl **148** (top, masses shown correspond to the predominant sodiated acetonitrile adduct). ESI-FTICR-MS data show *ca*. 50% decaging at 1.5 eq. **142** at 300 μM (lower left), increasing to 100% conversion when **142** is added as 4 eq. at 300 μM (lower right), affording **148-hyd**; with Met1, calc. 45895 Da, found 45894 Da; without Met1, calc. 45764 Da, found 45765 Da.

At 1.5 eq. **142** at 300 μM, conversion increased to 50%, whilst the same number of equivalents but halved concentration of **142** and **147** led to no conversion. Pleasingly, complete decaging was observed with 4 eq. **142** at 300 μM. These results suggest that

a delicate balance exists between protein precipitation and thiazolidine decaging, where maximising the extent of decaging must be carefully weighed against any protein precipitation. Just as with periodate oxidation, each protein construct requires careful optimisation of palladium decaging conditions. Residue microenvironment, surrounding protein bulk and palladium-induced precipitation affinity are all likely factors responsible for achieving this balance. In this situation, a balance of both equivalents and concentration of palladium is required to achieve full decaging without protein precipitation. This work does show, however, that thiazolidine decaging is suitable for proteins beyond GFPs, broadening the scope of applicable substrates.

4.5 Conclusions

A procedure for protein thiazolidine decaging to expose glyoxyl aldehydes has been developed using palladium reagent **142**. For GFPs, this protocol requires only one equivalent of **142** and reaches full decaging within one hour at room temperature without protein precipitation. This procedure was then optimised for use on another protein scaffold, BiGalK, which again led to full decaging within one hour.

A second route to protein glyoxyls was optimised through the use of protein 1,2-aminoalcohols. Oxidation of this pseudo-*N*-terminal threonine residue with $NaIO_4$ affords the same glyoxyl revealed through palladium decaging of thiazolidines. Possessing two separate routes to the same reactive aldehyde species increases the utility of the glyoxyl aldehyde, raising the likelihood of at least one suitable decaging method being found for the protein of interest. Both routes reach full conversion rapidly under mild, biocompatible conditions, meeting the second objective of this project.

Chapter 5: Modification of the Decaged Aldehyde Handle

5.1 Introduction

The final objective of this project was to demonstrate the reactivity of the decaged protein aldehyde through the use of established aldehyde modification chemistry. The previous chapter focussed upon the aldehyde from an analytical perspective: can it be observed and quantified? This chapter considers the aldehyde from a more practical perspective: what does it do and how well? Herein lie the opportunities for showcasing the aldehyde not just as a peak in a mass spectrum, but as a handle for adding function to proteins in exciting and useful ways.

5.2 Oxime ligation

5.2.1 Overview

Oxime ligation is arguably the staple ligation method of protein aldehydes,[77] taking advantage of the aldehyde electrophilicity by using a strong nucleophile to form a more stable condensation product. This nucleophile typically benefits from the α-effect, wherein the nucleophilicity of a heteroatom can be enhanced by the presence of an adjacent heteroatom, such as hydrazine or aminooxy nucleophiles to form hydrazones and oximes respectively. Even with this enhanced nucleophilicity, some form of catalysis is required to activate the aldehyde towards electrophilic attack. This initially took the form of acid catalysis, where pH 4 was required to tag a glyoxyl-containing antibody with an aminooxy reagent containing an ^{125}I radiolabel.[290] Such a low pH is not suitable for all proteins- the EGFP and sfGFP scaffolds are denatured at pH 4- so nucleophilic catalysis was developed as an alternative. Aniline, and other substituted aniline derivatives, allow modification within 20 h at neutral pH, with more electron-rich anilines such as *p*-anisidine affording the greatest rate enhancements.[291] The protonated aniline Schiff base intermediate is even more electrophilic than a protonated glyoxyl aldehyde, leading to a much higher rate of ligation under compatible conditions with reliably high conversion (Figure 72). However, the resulting bioconjugate is not particularly stable, as every mechanistic step is a reversible equilibrium. At neutral pH in aqueous conditions, the half-life of a hydrazone is on the order of one hour, compared to a fortnight for an oxime linkage (the temperature for this is not reported),[80] although this reversibility need not always be a disadvantage depending upon the application. Nevertheless, oxime ligation remains the first-choice method for modifying biological aldehydes on proteins and carbohydrates.

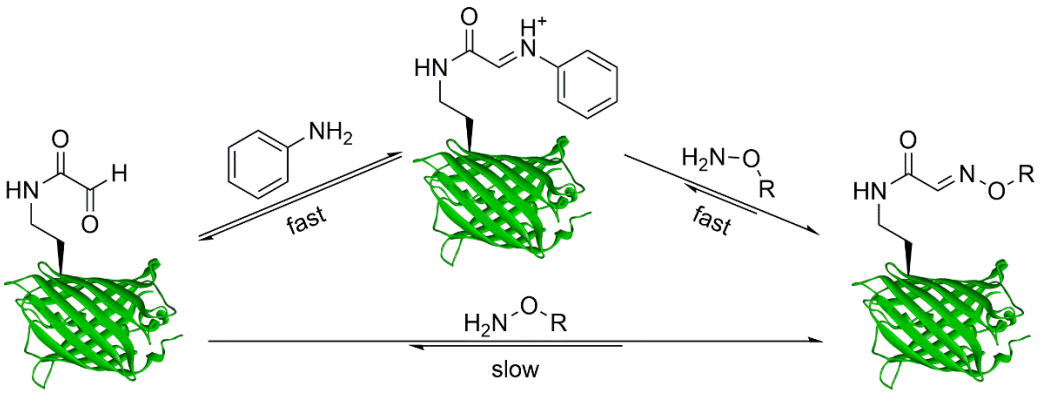


Figure 72: The rate of oxime ligation is greatly enhanced by aniline catalysis *via* the protonated aniline Schiff base intermediate.

5.2.2 Protein modification using oxime ligation

A variety of aniline-catalysed protocols have been reported;[89, 221, 240] one of the selling points of this ligation strategy is the scope for altering the substrates and conditions as the situation demands. Aniline typically remains at a working concentration of 0.1 M in order to drive formation of the reactive Schiff base intermediate. Oxime ligation was trialled upon protein glyoxyl **139** using *ca*. 35 eq. of biotin- and dansyl-aminooxy probes **123** and **124** to form protein oximes **149** and **150** (Figure 73). Full conversion was observed by ESI-FTICR-MS after 24 h, demonstrating that the decaged glyoxyls not only look like aldehydes, but also react like aldehydes. The mass spectra are complicated by the presence of unreacted probe, falling in the mass range of the protein charge ladder and not being sufficiently removed by the desalting process. Some efforts have been made to remove the signals from these mass spectra during data acquisition where the probe signals were masking the protein signals, simplifying the spectra for more reliable deconvolution. This process was then repeated using protein glyoxyl **145**, where once again full conversion to protein oximes **151** and **152** was observed (Figure 74).

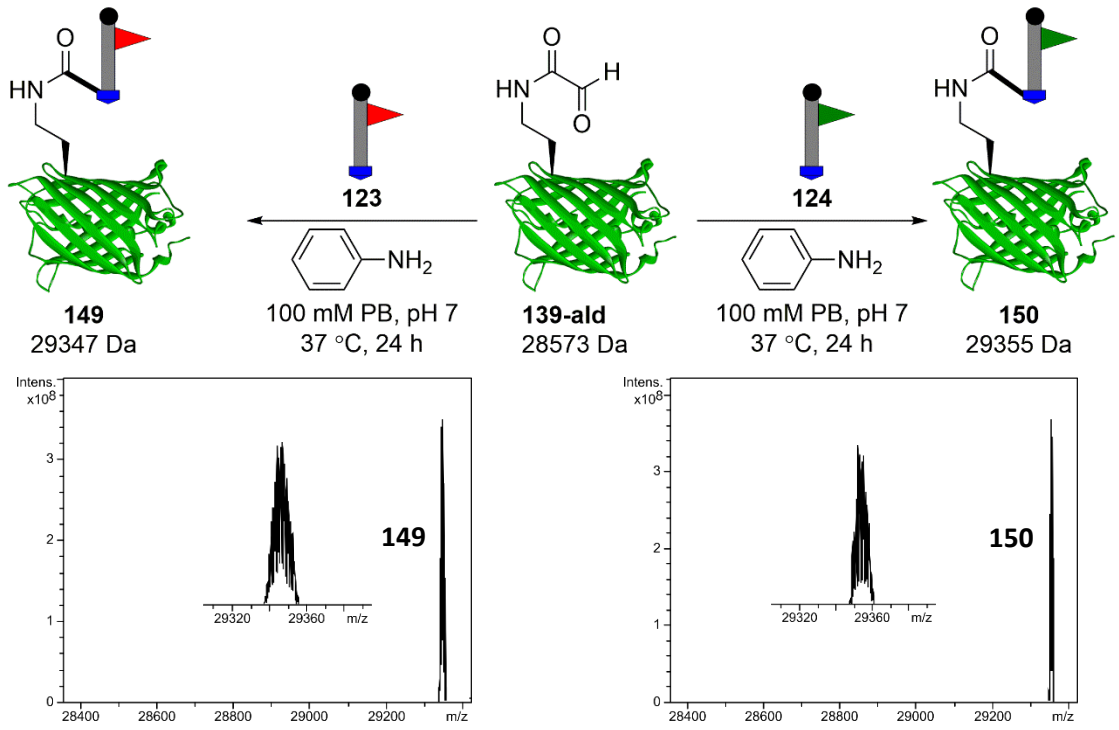


Figure 73: Oxime ligation of protein aldehyde **139** with aminooxy biotin probe **123** or aminooxy dansyl probe **124** (top). In both cases, ESI-FTICR-MS data show the complete consumption of **139-ald** to afford biotinylated protein **149** (lower left; calc. 29347 Da, found 29348 Da) and dansylated protein **150** (lower right, calc. 29355 Da, found 29356 Da) with the protein-probe linkage schematically depicted as a bold bond.

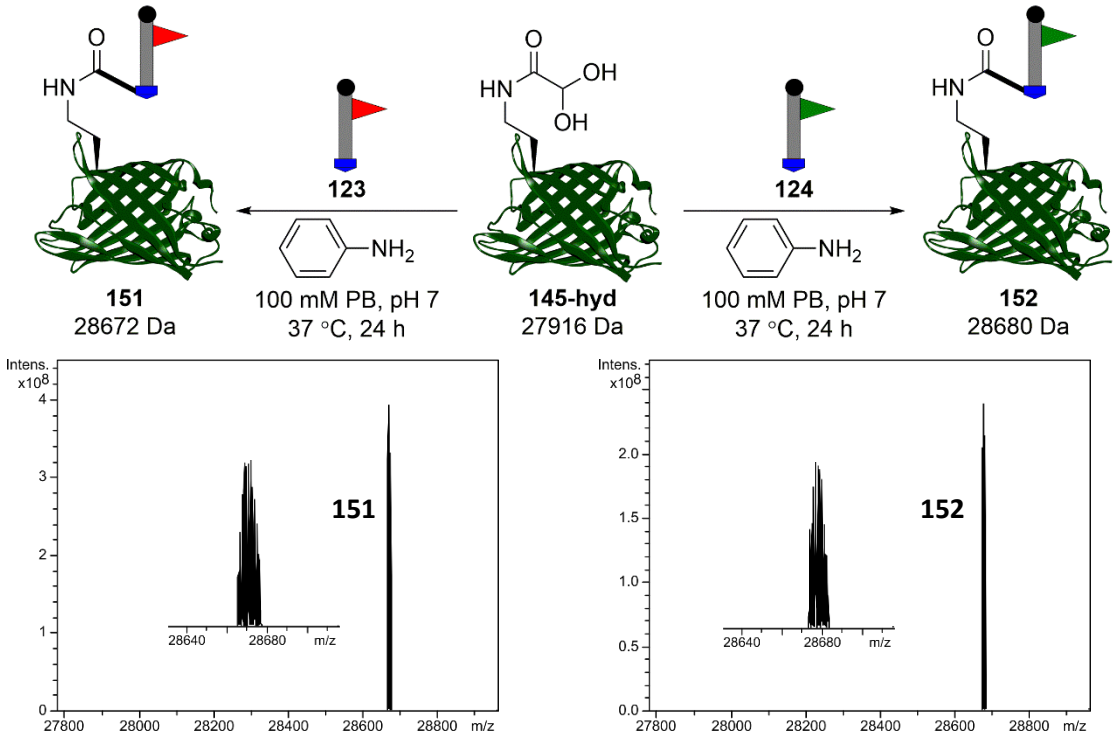


Figure 74: Oxime ligation of protein aldehyde **145** with aminooxy biotin probe **123** or aminooxy dansyl probe **124** (top). In both cases, ESI-FTICR-MS data show the complete consumption of **145-hyd** to afford biotinylated protein **151** (lower left; calc. 28672 Da, found 28673 Da) and dansylated protein **152** (lower right, calc. 28680 Da, found 28679 Da) with the protein-probe linkage schematically depicted as a bold bond.

As further verification and a demonstration of the bioconjugate utility, dansylated bioconjugate **153** was visualised by fluorescence following denaturing SDS-PAGE, ensuring that any fluorescent signals only arise from the dansyl group appended to a protein (Figure 75). The fluorescent gel confirms successful bioconjugation, with a band at *ca*. 29 kDa for **153** fluorescing as expected whilst negative control **145** could only be seen by Coomassie staining. Analogously, biotinylated protein **149** was visualised using colorimetric Western blotting. Following incubation with anti-biotin substrates, a band was detected for **149** at the expected mass of *ca*. 30 kDa, whilst no band was observed for negative control **139**. These data conclusively demonstrate the reactivity and functional labelling of the decaged protein glyoxyl aldehydes using oxime ligation.

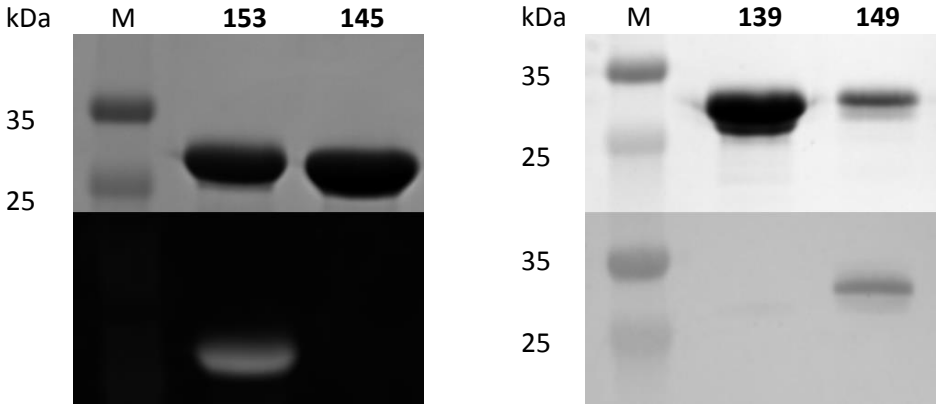


Figure 75: (left) Coomassie staining (upper) and fluorescence (lower) of unmodified negative control **145** and dansylated protein **153** following denaturing SDS-PAGE; (right) Coomassie staining (upper) and Western blotting with an anti-biotin substrate (lower) of unmodified negative control **139** and biotinylated protein **149** following denaturing SDS-PAGE.

5.3 OPAL

5.3.1 Overview

The organocatalytic protein aldol ligation, OPAL, is a relatively new addition to the field of protein aldehyde modification, arising from a deep interest in the opportunities for protein modification presented by aldehyde handles within the Fascione laboratory. As mentioned, this reaction was first discovered when a protein glyoxyl reacted with acetaldehyde, released by the periodate oxidation exposing the protein glyoxyl. Diligent development of OPAL by Dr Richard Spears appropriated some of the most powerful developments in organocatalytic aldol reactions and crossed-aldol reactions to produce a remarkably simple yet effective protein modification strategy.[244]

Only three reagents are required for OPAL: an acceptor aldehyde, a donor aldehyde and an organocatalyst. The acceptor aldehyde must be non-enolisable and highly electrophilic, with the protein glyoxyl aldehyde proving to be an ideal candidate for this

role. The most reactive donor aldehydes, the nucleophilic species, were found to be phenacetaldehyde derivatives. Proline tetrazole **154** acts as an organocatalyst, condensing with the donor aldehyde to form a highly nucleophilic enamine that attacks the acceptor aldehyde (Figure 76).[244] The product links together both aldehydes through a stable carbon-carbon bond linkage and contains a β-hydroxy aldehyde group that can subsequently undergo oxime ligation, opening up routes for dual modification of proteins. OPAL has already enjoyed a creative array of applications, including folate labelling for potential use in cancer-targeting therapeutics and the mimicry of two post-translational modifications installed on a protein implicated in *Leishmania* virulence.

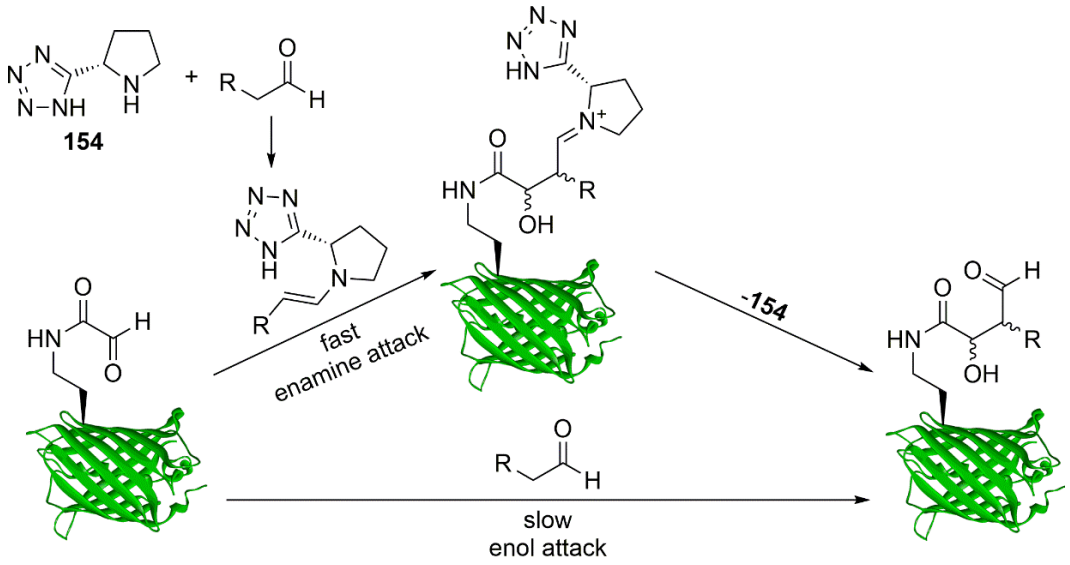


Figure 76: Donor aldehydes slowly attack the acceptor aldehyde *via* the enol tautomer, but the use of organocatalyst **154** opens up a much faster pathway through the formation of a more nucleophilic enamine.

5.3.2 Protein modifiation using OPAL

As with oxime ligation, a helpful degree of flexibility surrounds OPAL, with the consequence that little optimisation is required to achieve full conversion. Protein and probe conditions are largely the same as those used in oxime ligation, but far fewer equivalents are required of the organocatalytic proline tetrazole **154** compared to the vast excess of catalytic aniline required for oxime ligation. Protein glyoxyls **139** and **145** were reacted with dansyl probe **131** at neutral pH for just one hour, after which 100% conversion to dansylated proteins **155** and **156** was observed by ESI-FTICR-MS (Figure 77). The β-hydroxyaldehyde does not appear to be hydrated to an appreciated extent, a symptom of its diminished reactivity that prevents further OPAL from happening.

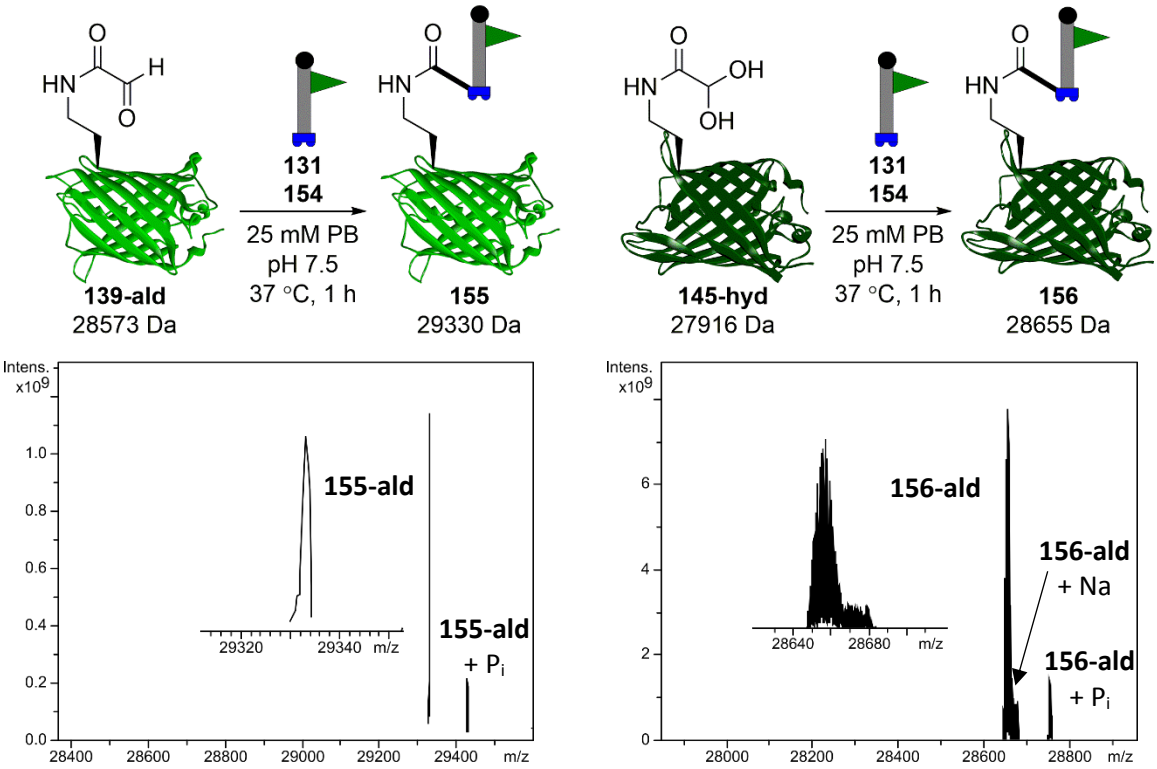


Figure 77: OPAL using aminooxy biotin probe **123** with protein aldehydes **139** (upper left) and **145** (upper right). In both cases, ESI-FTICR-MS data show the complete consumption of the protein aldehyde to afford dansylated proteins protein **155** (lower left; calc. 29330 Da, found 29332 Da) and **156** (lower right, calc. 28655 Da, found 28657 Da) with the protein-probe linkage schematically depicted as a bold bond. The β-hydroxyaldehyde is observed only as the aldehyde form without hydration.

Fluorescence imaging of **155** and **156** following denaturing SDS-PAGE further confirmed successful labelling (Figure 78).

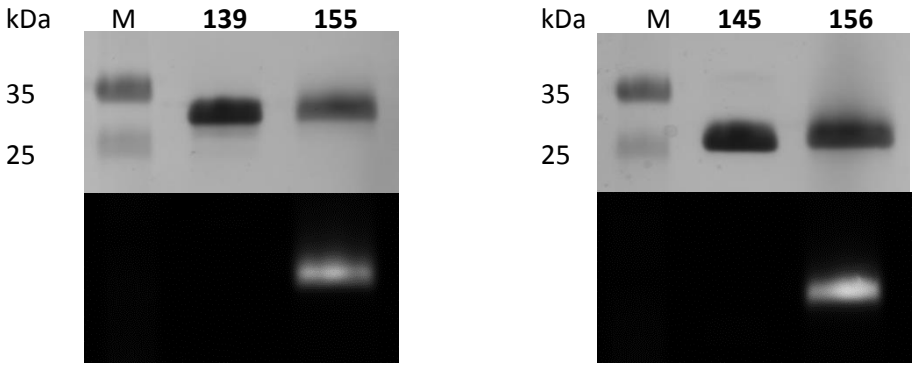


Figure 78: (left) Coomassie staining (upper) and fluorescence (lower) of unmodified negative control **139** and dansylated protein **155** following denaturing SDS-PAGE; (left) Coomassie staining (upper) and fluorescence (lower) of unmodified negative control **145** and dansylated protein **156** following denaturing SDS-PAGE.

Labelling of protein glyoxyls **139** and **145** with biotin probe **130** was performed by Dr Richard Spears and the same success was reported.[244] As a further demonstration of the synergy of thiazolidine decaging and OPAL in an experiment performed by Dr

Richard Spears, cell lysate material from the preparation of thiazolidine-containing **136** was treated with palladium reagent **142** and then biotin probe **130** under OPAL conditions, with biotin affinity chromatography using monomeric avidin successfully capturing and then eluting the biotinylated EGFP. This protein pull-down work is an elegant demonstration of the biocompatibility and utility of the palladium-OPAL pair, delivering even in the challenging milieu of cell lysate material with exquisite selectivity.

5.4 SPANC

5.4.1 Overview

Strain-promoted alkyne-nitrone cycloaddition, SPANC, has seen use not only through *in vitro* protein modification, [245] but also live-cell labelling,[292] nanoparticle loading,[293] and functionalisation of polymers.[294] The electrophilicity of the aldehyde partner is exploited as a means by which a nitrone can be formed: first through the formation of a Schiff base with organocatalytic *p*-anisidine, then through nucleophilic attack by an *N*-substituted hydroxylamine compound. The resulting nitrone acts as a 1,3-dipole ready to undergo a cycloaddition with a strained alkyne such as a BCN (Figure 79). Varying the substituent on the hydroxylamine has allowed the pairing of SPANC with CuAAC, where *N*-propargylhydroxylamine was used to form the nitrone. The resulting isoxazoline linkage is generally stable, although some highly reactive cycloalkynes are documented to undergo rearrangements after SPANC.[295] Strain-promoted cycloadditions are generally regarded as bioorthogonal on account of the paucity of suitable 1,3-dipoles such as azides, nitrones and nitrile oxides found in nature, as well as the limited reactivity of strained alkyne dipolarophiles towards functional groups commonly found in cellular milieu.

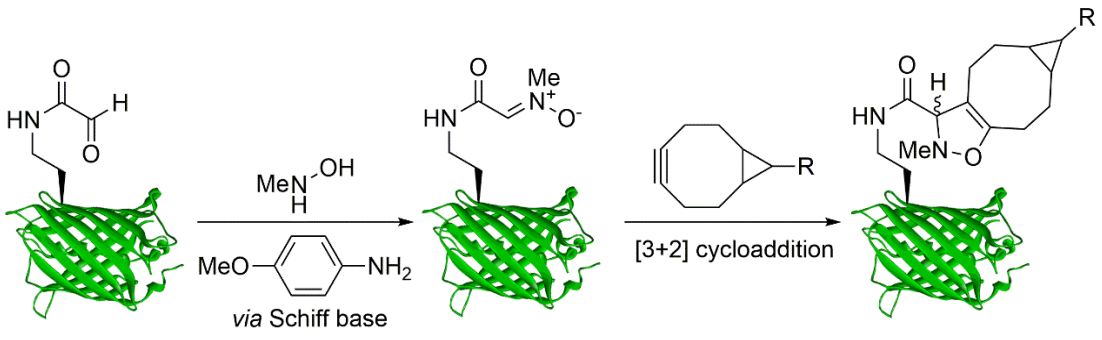


Figure 79: Promoted by *p*-anisidine catalysis, the protein glyoxyl condenses with *N*-methylhydroxylamine to form an oxime, becoming a nitrone upon deprotonation. This protein nitrone undergoes a [3+2] cycloaddition with a strained alkyne to furnish the isoxazoline bioconjugate.

5.4.2 Protein modification using SPANC

As SPANC had seen previous use on protein glyoxyl aldehydes, the conditions reported served as a helpful starting point for trialling conditions. The *p*-anisidine organocatalyst is required in large excess, as is the case with oxime ligation, and 50 equivalents of *N*-methylhydroxylamine serve to drive nitrone formation to completion, but only 20 equivalents of strained alkyne are needed to completely consume the protein nitrone. Hence protein glyoxyls **139** (Figure 80) and **145** (Figure 81) were reacted with BCN probes **133** and **134** under SPANC conditions to form labelled proteins **157**-**160** at pH 6.8 for 18 h. Once again complete conversion was observed by ESI-FTICR-MS for all four trial reactions.

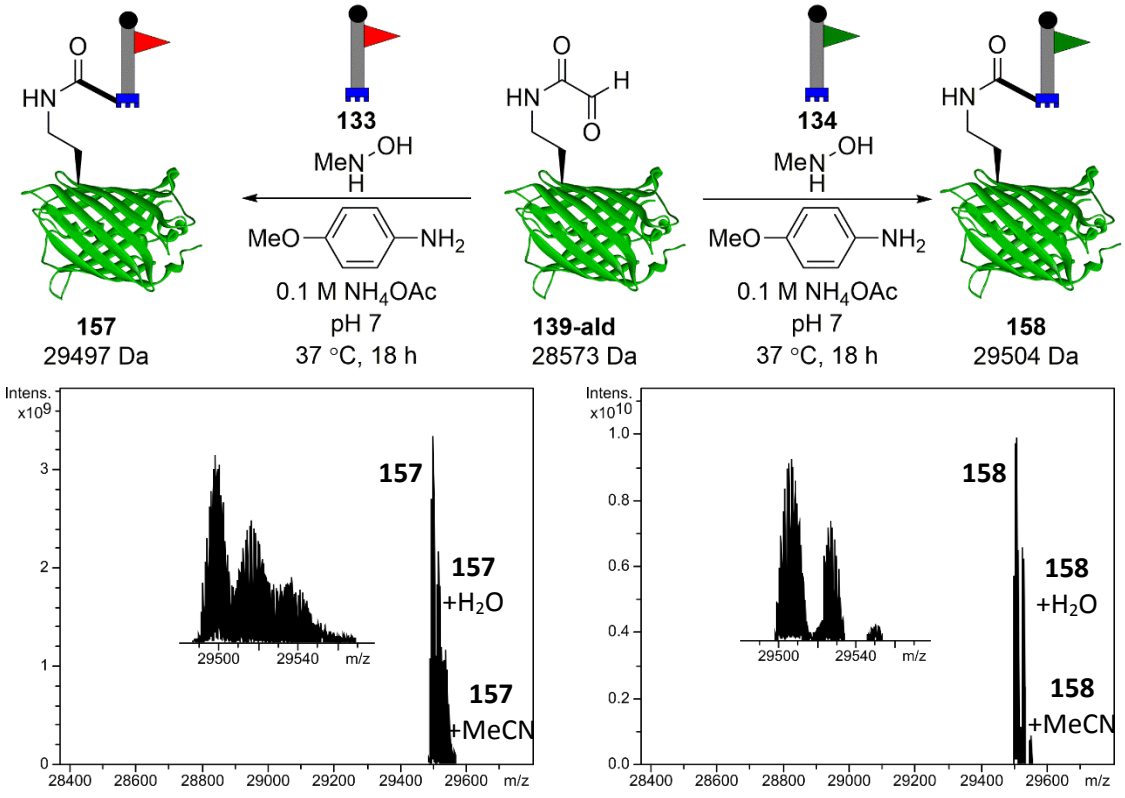


Figure 80: SPANC of protein aldehyde **139** with BCN biotin probe **133** or BCN dansyl probe **134** (top). In both cases, ESI-FTICR-MS data show the complete consumption of **139-ald** to afford biotinylated protein **157** (lower left; calc. 29497 Da, found 29499 Da) and dansylated protein **158** (lower right, calc. 29504 Da, found 29506 Da) with the protein-probe linkage schematically depicted as a bold bond.

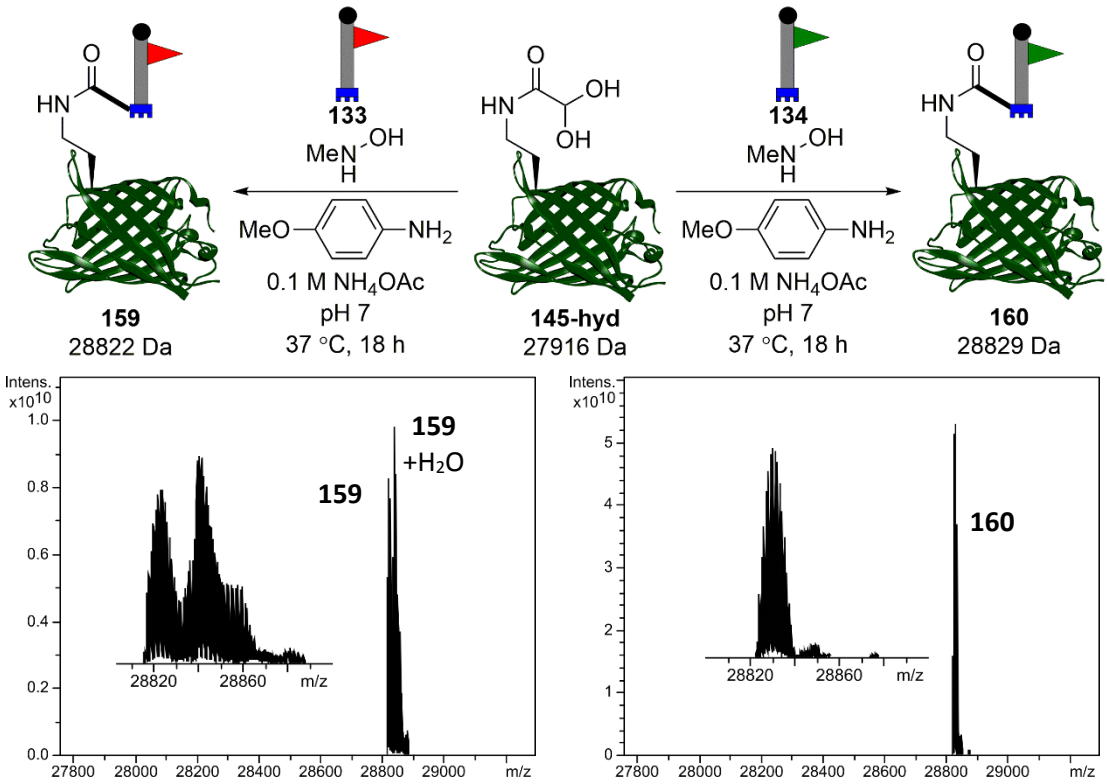


Figure 81: SPANC of protein aldehyde **145** with BCN biotin probe **133** or BCN dansyl probe **134** (top). In both cases, ESI-FTICR-MS data show the complete consumption of **139-ald** to afford biotinylated protein **159** (lower left; calc. 28822 Da, found 28824 Da) and dansylated protein **160** (lower right, calc. 28829 Da, found 28831 Da) with the protein-probe linkage schematically depicted as a bold bond.

Labelled proteins **157**-**160** were then visualised by fluorescence and Western blotting following denaturing SDS-PAGE as a further confirmation of successful modification (Figure 82).

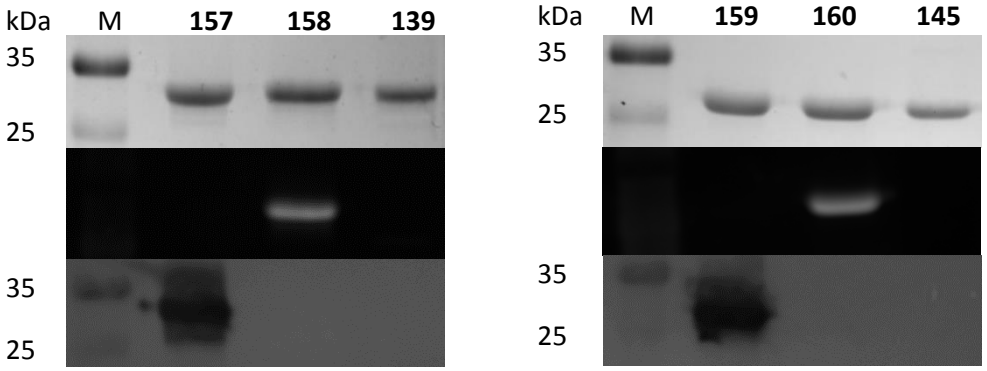


Figure 82: (left) Coomassie staining (upper), fluorescence (middle) and Western blotting with an anti-biotin substrate (lower) of unmodified negative control **139**, biotinylated protein **157** and dansylated protein **158** following denaturing SDS-PAGE; (right) Coomassie staining (upper), fluorescence (middle) and Western blot with an anti-biotin substrate (lower) of unmodified negative control **145**, biotinylated protein **159** and dansylated protein **160** following denaturing SDS-PAGE.

5.5 Conclusions

The decaged protein glyoxyl aldehydes have been shown to undergo three standard and useful bioconjugation reactions of oxime ligation, OPAL and SPANC. Using this chemistry, proteins have been tagged with biotin and dansyl groups, as demonstrated by Western blots and fluorescence. For *in vitro* protein modification, OPAL and SPANC had only seen usage on *N*-terminal glyoxyl aldehydes. This work has shown that these bioconjugations are not limited in scope to the *N*-terminus, removing a heavy restriction on the use of this chemistry. The “internal aldehyde” exhibits comparable reactivity to an *N*-terminal glyoxyl aldehyde and presents new opportunities for site-selective protein modification beyond the protein *N*-terminus.

Chapter 6: Conclusions and Future Directions

6.1 Summary

Two non-canonical amino acids harbouring caged glyoxyl aldehydes, **86** and **105**, have been found to be suitable for amber stop codon suppression using the *M. mazei* pyrrolysyl tRNA-synthetase (wild type) pair, producing full-length protein containing the caged aldehyde. Four other potential caged aldehyde NCAAs were not recognised to an appreciable extent by either wild-type tRNA synthetase or the Y306A-Y384F mutant.

Proteins containing **86** were found to liberate the glyoxyl aldehyde upon treatment with palladium complex **142**, the palladium complex most active in decaging without inducing protein precipitation. This procedure requires some optimisation for individual proteins, in order to balance the reactivity of the palladium complex to open the thiazolidine with potential side-reactions including protein precipitation. Alternatively, proteins containing **105** underwent oxidative cleavage using $NaIO_4$ to afford the same protein glyoxyl aldehyde. Together, **86** and **105** present two different routes for accessing the glyoxyl aldehyde, offering useful flexibility when a particular protein of interest may be intolerant of one decaging procedure but not another.

Most crucially of all, the exposed protein glyoxyl serves as an excellent handle for site-specific protein modification. Bioconjugations using oxime ligation, OPAL and SPANC proceeded to full conversion, conferring additional avidin affinity or fluorescent functionality through the use of reactive probes bearing biotin or dansyl groups. As judged by the methods used here, the “internal aldehyde” exposed from **86** or **105** appears to exhibit the same reactivity as its *N*-terminal counterpart, functioning as a useful bioorthogonal handle for chemical protein modification.

6.2 Future work

6.2.1 Optimised amber stop codon suppression

Whilst **86** and **105** are substrates for wild-type *M. mazei* pylRS, the recognition is not optimal. The assay used here to determine successful amber stop codon suppression is only qualitative, but clear differences in protein yields can be seen from both the SDS-PAGE and fluorescence when comparing **86** and **105** to positive control **29**. For proteins such as GFPs and BiGalK, where yields are typically high, this is not a major issue. However, the utility of **86** and **105** would be increased were a tRNA synthetase evolved specifically to recognise these NCAAs, as the increased protein yields would make this strategy more practical for use with proteins with challenging production protocols. Indeed, this strategy has become commonplace when developing new NCAAs for amber stop codon suppression: one synthetase out of many variants is selected to work with a

particular NCAA, whilst the approach outlined in this thesis is the converse. This work has not shown that NCAAs **96**, **99**, **101** and **110** are unsuitable for amber stop codon suppression, but that no suitable tRNA synthetase of those screened was found to recognise these NCAAs. A directed evolution approach, screening a library of synthetases with mutations in key locations, would not only conclusively rule in or rule out **96**, **99**, **101** or **110** from amber stop codon suppression, but also find the optimal mutants for the recognition of **86** and **105**, leading to superior protein yields.

6.2.2 Applications of this work: *in vitro*

As previously discussed, the field of chemical protein modification operates with two distinct objectives. The first centres on chemistry: finding the widest array of strategies for chemical modification of proteins. The work in this thesis has focussed on this direction, seeking to add a new instrument to the toolbox of methods available. It is not the intention of this work to appear as superior or inferior to any other method; its utility lies in being an option, one of many, such that for even the most challenging protein or situation a suitable modification strategy exists.

The second objective in this field of more broad than pure chemistry: developing applications for chemical protein modification, showcasing the powerful chemistry through situations with a greater impact than test beds. This has primarily focussed on therapeutics, although biophysical applications are also under consideration.[296] The prominence of this objective is increasing over time, where impact is valued over novelty. One of the key tests of contemporary chemical protein modification strategies is its suitability for the generation of ADCs, pushing the chemistry well beyond the comfort of well-behaved test proteins tagged with fluorophores and often requiring multiple modifications. Palladium decaging, paired with suitable ligation such as OPAL, does seem to be a suitable platform for this purpose, given the site selectivity of both methods and the stability of the resulting bioconjugate. One common drawback of amber stop codon suppression, however, is the low yield of proteins produced this way. Ameliorating this issue requires extensive protein production optimisation and likely the mutation of a superior pylRS, with greater recognition of **86**. Nevertheless, antibody modification is the next logical step for developing this work, adding impact to the novelty.

A further application of the palladium decaging is post-translational mimicry. Prior to this work, glyoxyl aldehydes had been restricted to protein *N*-termini, but PTMs face no such restriction and can be found in a far greater range of locations on the protein. Given that this work has lifted the site restriction, glyoxyl aldehydes are now more suitable as handles for post-translational mimicry. The dual modification of hydrophilic surface protein A, HASPA, using OPAL on an *N*-terminal glyoxyl aldehyde followed by oxime

ligation on the ensuing β-hydroxyaldehyde, has demonstrated the suitability of protein aldehydes for this task.[244] Using palladium decaging and subsequent bioconjugation, this post-translation mimicry can extend beyond the *N*-terminus into multiple locations, slowly advancing on nature's own methods for post-translational modification.

6.2.3 Applications of this work: live cell labelling

Applications of chemical protein modification *in vitro*, such as ADCs, still have a central place in the field. Increasing in prominence is the need for demonstrations of chemical methods in living systems towards *in vivo* applications, where site specificity is truly tested by a more diverse milieu and side reactions with a greater range of substrates are possible- in addition to maintaining the health of the organism under investigation. Palladium complex **142** has seen use for live-cell protein deprotection, with negligible cytotoxic effects observed through careful control of the concentration of palladium.[156] Furthermore, palladium decaging and OPAL biotinylation has been shown to succeed in cell lysate material, owing to the selectivity of both methodologies.[244] This suggests the suitability of palladium-mediated thiazolidine decaging for live cell labelling in tandem with a bioconjugation strategy. Cellular incorporation of a thiazolidine through promiscuous metabolic pathways, such as the GlcNAc pathway,[297] would serve to smuggle the caged aldehyde into live cells ready for decaging and bioconjugation, such as for labelling in flow cytometry experiments.

Chapter 7: Experimental

7.1 Chemical synthesis

7.1.1 General methods

Solvents and Starting Materials

All solvents were dried prior to use according to standard methods, with the exception of solvents used for flash chromatography purposes, where GPR-grade solvents were used. All commercially-available reagents were used as received. Analytical grade reagents were supplied by Sigma-Aldrich, Fisher Scientific, VWR International, and TCI. **29** and **34** were purchased from Sirius Fine Chemicals.

General Procedures

All solution-phase reactions were carried out under a dry nitrogen atmosphere using oven-dried glassware unless otherwise stated. All concentrations were performed *in vacuo* unless stated otherwise. Analytical TLC was performed on silica gel 60-F^{254} with detection by fluorescence and/or charring following immersion in a solution of ninhydrin (1.5 g in 100 mL *n*-butanol and 3 mL acetic acid). All compounds reported in this section have been synthesised for the first time unless explicitly stated otherwise.

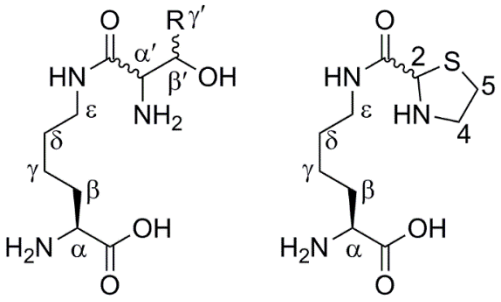
Instrumentation

Small-molecule HRMS data were obtained at room temperature on a Bruker Daltonics micrOTOF. Analytical HPLC of peptides was performed on a Shimadzu Prominence HPLC equipped with a Shimadzu photodiode array using an Accucore C18 2.6 µm column, 2.1 × 150 mm**.** All samples were run using gradients of HPLC-grade H_2O and MeCN, spiked with 0.1% (v/v) FA. Infra-red spectra were recorded on a PerkinElmer UATR2 spectrometer. Optical rotations were recorded at 20 °C using a Bellingham & Stanley ADP450 series polarimeter and values reported in units of 10^{-1} deg cm^2 g^{-1}.

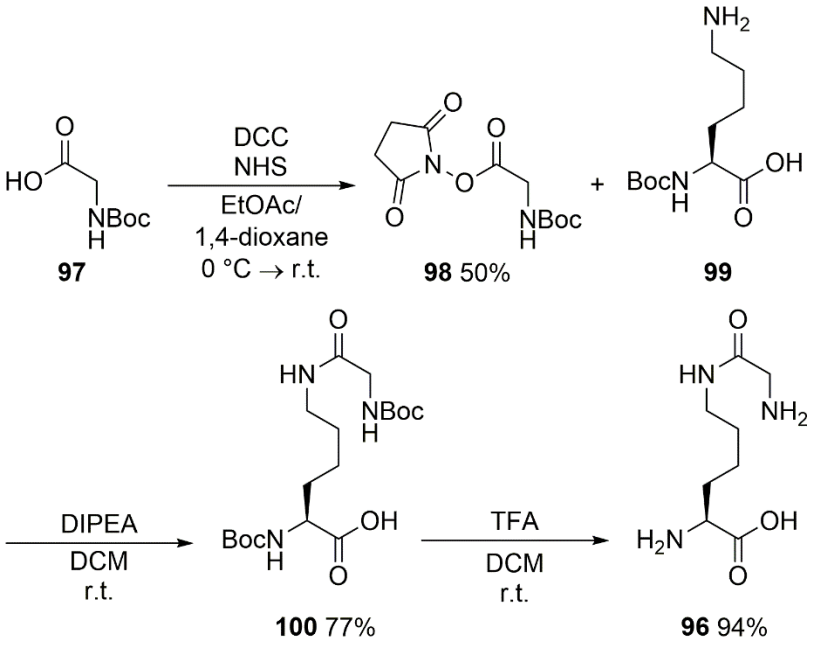
^{1}H and ^{13}C NMR spectra were recorded at 500 MHz and 126 MHz respectively on a Bruker AV500 instrument using an internal deuterium lock at room temperature. Signals were assigned using additional DEPT135, COSY, HSQC and HMBC experiments. Chemical shifts are reported in ppm according to the following references: DMSO-d_6: δ_H 2.50; δ_C 39.52 (centre of septet).

The following abbreviations were used to describe signal multiplicities or appearances: s, singlet; d, doublet; t, triplet; dt, doublet of triplets; q, quartet; qd, quartet of doublets; quint, quintet; dq, doublet of quintets; m, multiplet.

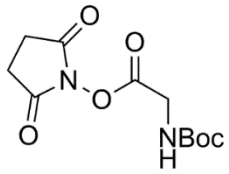
Acyclic amino acid carbon atoms are labelled with Greek letters (in subscript) in accordance with IUPAC conventions, with hydrogen and nitrogen atoms labelled with the Greek letter of the carbon atom to which they are attached. Prime symbols (') are used to label atoms of the second residue in a dipeptide. Thiazolidine carbon and nitrogen atoms are labelled with numbers in accordance with IUPAC conventions, with hydrogen atoms labelled with the number of the carbon or nitrogen atom to which they are bonded.



7.1.2 Solution-phase synthesis



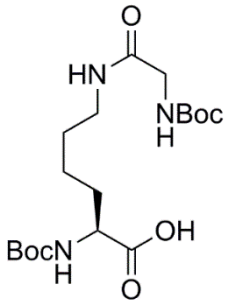
N-hydroxysuccinimidyl 2-(*tert*-butoxycarbonylamino)-acetate 98



N-hydroxysuccinimide (0.73 g, 6.3 mmol, 1.05 eq.) was added to a solution of 2-(*tert*-butoxycarbonylamino)-acetic acid **97** (1.05 g, 6.0 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (18 mL) at 0 °C. Dicyclohexylcarbodiimide (1.30 g, 6.3 mmol, 1.05 eq.) was added in one portion and the reaction left to warm to r.t. and stirred for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (20 mL) and washed sequentially with 5% (w/w) $NaHCO_3$ (10 mL), H_2O (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude *N*-hydroxysuccinimidyl 2-(*tert*-butoxycarbonylamino)-acetate **98** as a white powder of sufficient purity for further manipulation (0.82 g, 50%); δ_H (500 MHz, DMSO-d_6): 7.47 (major rotamer) and 7.09 (minor) (t, 1H, ^3J 6.2 Hz, NH), 4.09 (major) and 4.02 (minor) (d, 2H, ^3J 6.2 Hz, H_α), 2.81 (s, 4H, Su CH_2), 1.39 (major) and 1.37 (minor) (s, 9H, Boc CH_3); δ_C (126 MHz, DMSO-d_6): 170.0 (Su CO), 167.1 (major) and 166.9 (minor) (CO_2Su, mixture of rotamers), 155.6 (major) and 154.4 (minor) (CO_2NH), 79.2 (minor) and 78.8 (major) ($\underline{C}(CH_3)_3$), 41.2 (minor) and 39.7 (major) (C_α), 28.1 (major) and 27.6 (minor) (Boc CH_3), 25.4 (Su CH_2); IR (ATR): 3296, 2984, 2935, 1733, 1704, 1678, 1531, 1368, 1290, 1210, 1167, 1083, 1047, 993, 948, 902, 866, 814,

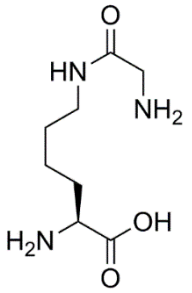
764, 746, 644, 616; HRMS: found $[M+Na]^+$ 295.0897; $C_{11}H_{16}N_2O_6Na$ requires 295.0901 (Δ = 1.2 ppm). Characterisation data in agreement with previous literature reports.[298]

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-(2-(*tert*-butoxycarbonylamino)-acetamido)-hexanoic acid 100

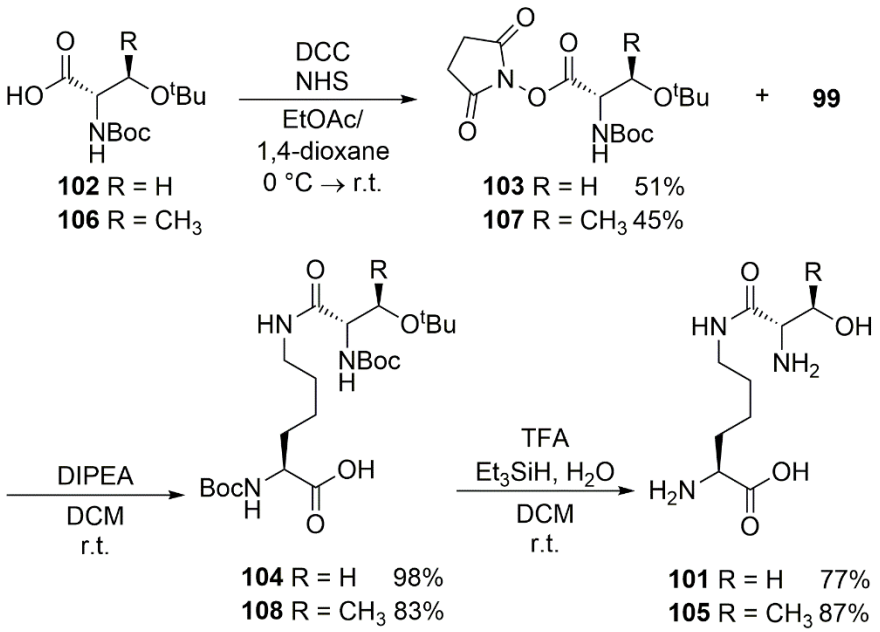


DIPEA (1.75 mL, 10 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl 2-(*tert*-butoxycarbonylamino)-acetate **98** (0.68 g, 2.5 mmol, 1.0 eq.) in DCM (9 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (0.62 g, 2.5 mmol, 1.0 eq.) in DCM (12 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (30 mL) and washed with sat. citric acid (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude (2*S*)-2-(*tert*-butoxycarbonylamino)-6-(2-(*tert*-butoxycarbonylamino)-acetamido)-hexanoic acid **100** as an off-white powder of sufficient purity for further manipulation (0.78 g, 77%); δ_H (500 MHz, DMSO-d_6) 7.72 (t, 1H, ^3J 5.4 Hz, HN_ε), 7.01 (d, 1H, ^3J 8.0 Hz, HN_α), 6.85 (t, 1H, ^3J 6.1 Hz, $HN_{\alpha'}$), 3.78-3.84 (m, 1H, H_α), 3.48 (d, 2H, ^3J 6.1 Hz, $H_{\alpha'}$), 2.97-3.07 (m, 2H, H_ε), 1.22-1.65 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6) 174.2 (CO_2H), 169.0 (CONH), 155.7, 155.6 (Boc CO_2N), 78.0, 77.9 ($\underline{C}(CH_3)_3$), 53.4 (C_α), 43.2 ($C_{\alpha'}$), 38.2 (C_ε), 30.4 (C_β), 28.7 (C_γ), 28.2 (Boc CH_3), 23.0 (C_δ); IR (ATR): 3321, 2978, 2933, 1691, 1516, 1453, 1392, 1366, 1247, 1158, 1050, 1024, 946, 860, 780, 562; HRMS: found $[M+Na]^+$ 426.2214; $C_{18}H_{33}N_3O_7Na$ requires 426.2211 (Δ = -0.7 ppm).

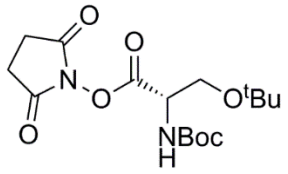
(2*S*)-2-amino-6-(2-aminoacetamido)-hexanoic acid 96



Trifluoroacetic acid (4 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-(2-(*tert*-butoxycarbonylamino)-acetamido)-hexanoic acid **100** (0.63 g, 1.55 mmol) in DCM (8 mL) at 0 °C and left to reach rt. Once the reaction was determined complete by TLC, usually within 3 h, the reaction mixture was triturated with ice-cold Et_2O (30 mL), filtered, and washed with ice-cold Et_2O (4 × 20 mL). The resulting off-white powder was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford the product (2*S*)-2-amino-6-(2-aminoacetamido)-hexanoic acid **96** as a fluffy off-white powder (470 mg, diacetate salt, 94%); $[\alpha]_D$ +7.2 (c 0.9, H_2O); δ_H (500 MHz, DMSO-d_6): 8.40 (t, 1H, ^3J 5.5 Hz, HN_ε), 3.78 (t, 1H, ^3J 6.2 Hz, H_α), 3.52 (s, 2H, $H_{\alpha'}$), 3.08-3.12 (m, 2H, H_ε), 1.70-1.82 (m, 2H, H_β), 1.25-1.46 (m, 4H, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6): 171.1 (CO_2H), 165.7 (CONH), 52.1 (C_α), 40.1 ($C_{\alpha'}$), 38.6 (C_ε), 29.7 (C_β), 28.4 (C_γ), 21.8 (C_δ); IR (ATR): 2945, 1663, 1508, 1435, 1257, 1183, 1132, 1014, 914, 837, 798, 722, 599, 518; HRMS: found $[M+H]^+$ 204.1344; $C_8H_{18}N_3O_3$ requires 204.1343 (Δ = -0.9 ppm).



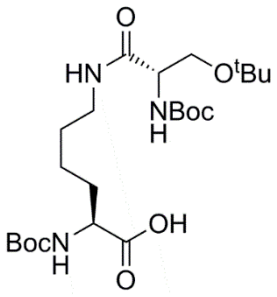
N-hydroxysuccinimidyl (2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate 103



N-hydroxysuccinimide (0.73 g, 6.3 mmol, 1.05 eq.) was added to a solution of (2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoic acid **102** (1.57 g, 6.0 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (18 mL) at 0 °C. Dicyclohexylcarbodiimide (1.30 g, 6.3 mmol, 1.05 eq.) was added in one portion and the reaction left to warm to r.t. and stirred for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (20 mL) and washed sequentially with 5% (w/w) $NaHCO_3$ (10 mL), H_2O (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude *N*-hydroxysuccinimidyl (2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate **103** as a white powder of sufficient purity for further manipulation (1.10 g, 51%); δ_H (500 MHz, DMSO-d_6): 7.32 (major rotamer) and 6.92 (minor) (d, 1H, ^3J 8.3 Hz, NH), 4.45-4.52 (major) and 4.32-4.37 (minor) (m, 1H, H_α), 3.62-3.68 (m, 2H, H_β), 2.80 (s, 4H, Su CH_2), 1.39 (major) and 1.37 (minor) (s, 9H, Boc CH_3), 1.14 (major) and 1.08 (minor) (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6): 170.2 (minor) and 169.8 (major) (Su CO, mixture of rotamers), 167.0 (major) and 166.3 (minor) (CO_2Su), 155.1 (major) and 154.7 (minor) (CO_2NH), 79.4 (minor) and 78.8 (major) (Boc $\underline{C}(CH_3)_3$), 73.2 (ether $\underline{C}(CH_3)_3$), 61.0 (major) and 60.7 (minor) (C_β) 54.3 (minor) and 53.0 (major) (C_α), 28.1 (major) and 27.6 (minor) (Boc CH_3), 27.1 (minor) and

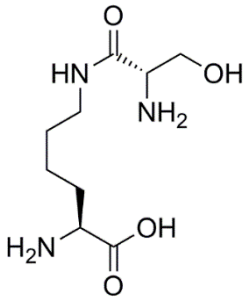
27.0 (major) (ether CH_3), 25.4 (Su CH_2); IR (ATR): 3410, 2982, 2935, 1820, 1783, 1735, 1714, 1475, 1444, 1427, 1364, 1196, 1170, 1102, 1083, 1064, 1046, 1036, 995, 912, 867, 835, 810, 774, 749, 647, 576, 553; HRMS: found $[M+Na]^+$ 381.1633; $C_{16}H_{26}N_2O_7Na$ requires 381.1632 (Δ = -1.0 ppm). Characterisation data in agreement with previous literature reports.[299]

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid 104



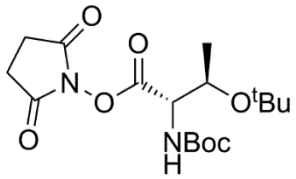
DIPEA (1.75 mL, 10 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl (2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate **103** (0.90 g, 2.5 mmol, 1.0 eq.) in DCM (8 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (0.62 g, 2.5 mmol, 1.0 eq.) in DCM (12 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (30 mL) and washed with sat. citric acid (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid **104** as an off-white powder of sufficient purity for further manipulation (1.20 g, 98%); δ_H (500 MHz, DMSO-d_6) 7.80 (t, 1H, ^3J 5.0 Hz, HN_ε), 6.98 (d, 1H, ^3J 8.0 Hz, HN_α), 6.85 (d, 1H, ^3J 8.3 Hz, $HN_{\alpha'}$), 3.95 (dt, 1H, ^3J 8.4 Hz, $^3J_{\alpha'\text{-}\beta'}$ 5.6 Hz, $H_{\alpha'}$), 3.80 (dt, 1H, ^3J 7.9 Hz, $^3J_{\alpha\text{-}\beta}$ 7.0 Hz, H_α), 3.38-3.43 (m, 2H, $H_{\beta'}$), 2.94-3.12 (m, 2H, H_ε), 1.20-1.65 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ), 1.09 (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6) 174.3 (CO_2H), 169.9 (CONH), 155.6, 155.0 (Boc CO_2N), 78.1, 77.9 (Boc $\underline{C}(CH_3)_3$), 72.6 (ether $\underline{C}(CH_3)_3$), 62.0 ($C_{\beta'}$), 54.9 ($C_{\alpha'}$), 53.4 (C_α), 38.2 (C_ε), 30.3 (C_β), 28.6 (C_γ), 28.2, 28.1 (Boc CH_3), 27.2 (ether CH_3), 22.9 (C_δ); IR (ATR): 3321, 2976, 2933, 1705, 1661, 1501, 1392, 1365, 1247, 1161, 1088, 1019, 860, 779, 735; HRMS: found $[M+Na]^+$ 512.2959; $C_{23}H_{43}N_3O_8Na$ requires 512.2942 (Δ = -3.5 ppm).

(2*S*)-2-amino-6-((2*S*)-2-amino-3-hydroxypropanamido)hexanoic acid 101



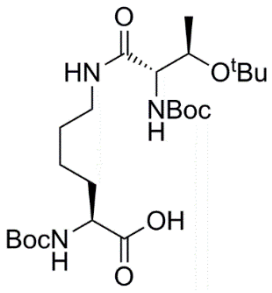
Deprotection cocktail 95:2.5:2.5 (v/v) TFA:TIS:H_2O (4 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid **104** (1.01 g, 2.06 mmol) in DCM (8 mL) at 0 °C and left to reach rt. Once the reaction was determined complete by TLC, usually within 3 h, the reaction mixture was triturated with ice-cold Et_2O (30 mL. The suspension was centrifuged at 4 000 × *g* for 5 min, supernatant decanted, and the peptide pellet resuspended in ice-cold Et_2O (20 mL), with this process repeated 5 times in total, after which the pellet was left to air dry. The resulting off-white powder was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford (2*S*)-2-amino-6-((2*S*)-2-amino-3-hydroxypropanamido)-hexanoic acid **101** as an off-white foam (560 mg, diacetate salt, 77%); $[\alpha]_D$ +21.1 (c 1.0, H_2O); δ_H (500 MHz, DMSO-d_6): 8.41 (t, 1H, ^3J 5.5 Hz, HN_ε), 3.55-3.63 (m, 4H, H_α, $H_{\alpha'}$, $H_{\beta'}$), 3.08-3.12 (m, 2H, H_ε), 1.69-1.79 (m, 2H, H_β), 1.28-1.47 (m, 4H, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6): 171.1 (CO_2H), 166.6 (CONH), 60.3 ($C_{\beta'}$), 54.4 ($C_{\alpha'}$), 52.1 (C_α), 38.5 (C_ε), 29.7 (C_β), 28.3 (C_γ), 21.7 (C_δ); IR (ATR): 3316, 2944, 1661, 1572, 1521, 1432, 1273, 1251, 1184, 1127, 1037, 836, 799, 721, 600, 542; HRMS: found $[M+H]^+$ 234.1451; $C_9H_{20}N_3O_4$ requires 234.1448 (Δ = -0.6 ppm).

N-hydroxysuccinimidyl (2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonyl-amino)-butanoate 107



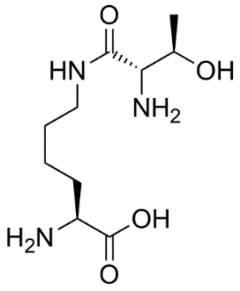
N-hydroxysuccinimide (0.91 g, 7.9 mmol, 1.05 eq.) was added to a solution of (2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoic acid **106** (2.06 g, 7.5 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (22 mL) at 0 °C. Dicyclohexylcarbodiimide (1.62 g, 7.9 mmol, 1.05 eq.) was added in one portion and the reaction left to warm to r.t. and stirred for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (20 mL) and washed sequentially with 5% (w/w) $NaHCO_3$ (10 mL), H_2O (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude *N*-hydroxysuccinimidyl (2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoate **107** as a white powder of sufficient purity for further manipulation (1.25 g, 45%); δ_H (500 MHz, DMSO-d_6): 6.97 (d, 1H, ^3J 9.3 Hz, $N_\alpha H$), 4.42 (dd, 1H, ^3J 9.3 Hz, $^3J_{\alpha\text{-}\beta}$ 4.1 Hz, H_α), 4.02-4.07 (m, 1H, H_β), 2.80 (s, 4H, Su H), 1.40 (s, 9H, Boc CH_3), 1.19 (d, 3H, $^3J_{\beta\text{-}\gamma}$ 6.2 Hz, H_γ), 1.14 (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6): 169.9 (Su CO), 166.8 (CO_2H), 155.2 (Boc CO_2NH), 78.9 (Boc $\underline{C}(CH_3)_3$), 73.9 (ether $\underline{C}(CH_3)_3$), 67.2 (C_β), 57.9 (C_α), 28.2 (Boc CH_3), 28.1 (ether CH_3), 25.4 (Su CH_2), 19.6 (C_γ); IR (ATR): 3351, 2978, 2935, 2119, 1819, 1781, 1732, 1711, 1512, 1456, 1368, 1241, 1205, 1151, 1120, 1090, 1059, 957, 914, 868, 825, 792, 644, 605, 546; HRMS: found $[M+Na]^+$ 395.1793; $C_{17}H_{28}N_2O_7Na$ requires 395.1789 (Δ = -1.0 ppm). This compound has been prepared previously but the characterisation data were not reported.[300]

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid 108

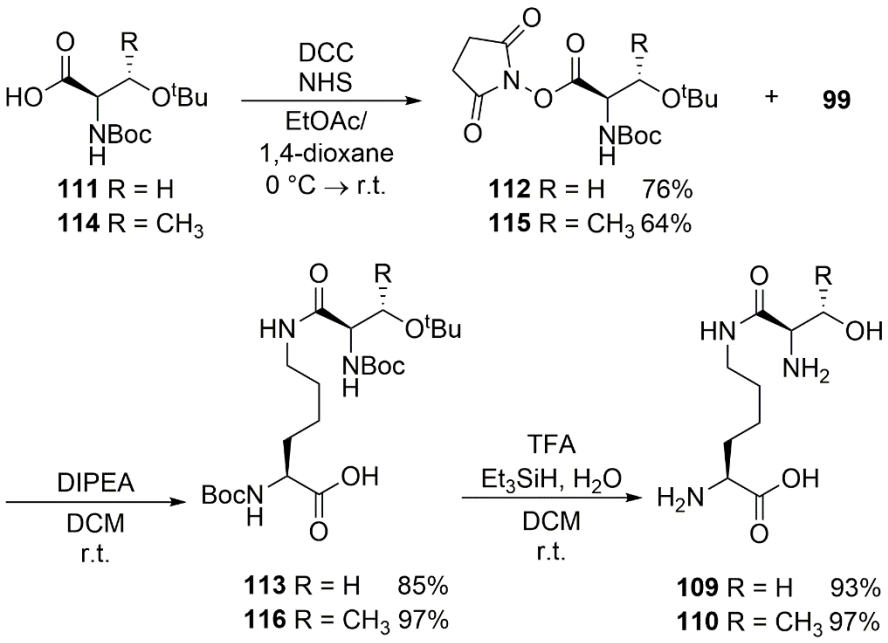


DIPEA (2.1 mL, 12 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl (2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoate **107** (1.12 g, 3.0 mmol, 1.0 eq.) in DCM (9 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (738 mg g, 3.0 mmol, 1.0 eq.) in DCM (15 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (30 mL) and washed with sat. citric acid (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid **108** as a white fluffy powder of sufficient purity for further manipulation (1.25 g, 83%); δ_H (500 MHz, DMSO-d_6) 7.71 (t, 1H, ^3J 5.6 Hz, HN_ε), 6.95 (d, 1H, ^3J 8.0 Hz, HN_α), 5.95 (t, 1H, ^3J 9.1 Hz, $HN_{\alpha'}$), 3.79-3.90 (m, 3H, H_α, $H_{\alpha'}$, $H_{\beta'}$), 2.93-3.12 (m, 2H, H_ε), 1.23-1.65 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ), 1.08 (s, 9H, ether CH_3), 1.00 (d, 3H, $^3J_{\beta'\text{-}\gamma'}$ 6.2 Hz, $H_{\gamma'}$); δ_C (126 MHz, DMSO-d_6) 174.2 (CO_2H), 169.8 (CONH), 155.5, 155.1 (Boc CO_2N), 78.3, 77.8 (Boc $\underline{C}(CH_3)_3$), 73.3 (ether $\underline{C}(CH_3)_3$), 67.4 ($C_{\beta'}$), 59.4 ($C_{\alpha'}$), 53.4 (C_α), 38.4 (C_ε), 30.4 (C_β), 28.5 (C_γ), 28.2, 28.1 (Boc CH_3), 28.0 (ether CH_3), 23.1 (C_δ), 19.9 ($C_{\gamma'}$); IR (ATR): 3436, 2980, 1722, 1706, 1493, 1453, 1424, 1391, 1364, 1312, 1248, 1216, 1160, 1091, 1068, 1038, 946, 864, 836, 793, 775, 754, 683; HRMS: found $[M+Na]^+$ 526.3089; $C_{24}H_{45}N_3O_8Na$ requires 526.3099 (Δ = 3.2 ppm).

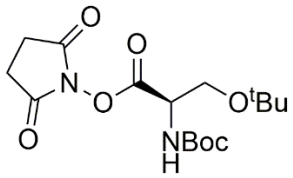
(2*S*)-2-amino-6-((2*S*,3*R*)-2-amino-3-hydroxybutanamido)-hexanoic acid 105



Deprotection cocktail 95:2.5:2.5 (v/v) TFA:TIS:H_2O (4 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid **108** (1.11 g, 2.2 mmol) in DCM (6 mL) at 0 °C and left to reach rt. Once the reaction was determined complete by TLC, *ca.* 3 h, the reaction mixture was split into two portions and each portion triturated with ice-cold Et_2O (30 mL). The suspension was centrifuged at 4 000 × *g* for 5 min, supernatant decanted, and the peptide pellet resuspended in ice-cold Et_2O (20 mL), with this process repeated 5 times in total, after which the pellet was left to air dry. The resulting off-white solid was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford (2*S*)-2-amino-6-((2*S*,3*R*)-2-amino-3-hydroxybutanamido)-hexanoic acid **105** as a fluffy white foam (706 mg, diacetate salt, 87%); $[\alpha]_D$ +9.5 (c 1.1, H_2O); δ_H (500 MHz, DMSO-d_6): 8.53 (t, 1H, ^3J 5.6 Hz, HN_ε), 3.85 (app. quint., 1H, $^3J_{\alpha'\text{-}\beta',\ \beta'\text{-}\gamma'}$ 6.5 Hz, $H_{\beta'}$), 3.79 (t, 1H, $^3J_{\alpha\text{-}\beta}$ 6.3 Hz, H_α), 3.48 (d, 1H, $^3J_{\alpha'\text{-}\beta'}$ 6.8 Hz, $H_{\alpha'}$), 3.14 (app. dq, 1H, ^2J 12.7 Hz, ^3J 6.6 Hz, H_ε), 3.05 (app. dq, 1H, ^2J 12.9 Hz, ^3J 6.5 Hz, H'_ε), 1.70-1.82 (m, 2H, H_β), 1.24-1.47 (m, 4H, H_γ, H_δ) 1.12 (d, 3H, $^3J_{\beta'\text{-}\gamma'}$ 6.4 Hz, $H_{\gamma'}$); δ_C (126 MHz, DMSO-d_6): 171.1 (CO_2H), 166.8 (CONH), 65.8 ($C_{\beta'}$), 58.5 ($C_{\alpha'}$), 52.1 (C_α), 38.5 (C_ε), 29.7 (C_β), 28.3 (C_γ), 21.8 (C_δ), 20.0 ($C_{\gamma'}$); IR (ATR): 2944, 1660, 1512, 1432, 1250, 1182, 1131, 918, 838, 798, 721, 599, 517; HRMS: found $[M+H]^+$ 248.1604; $C_{10}H_{22}N_3O_4$ requires 248.1605 (Δ = 1.0 ppm).



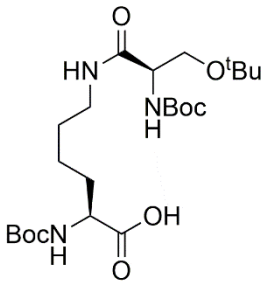
N-hydroxysuccinimidyl (2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate 112



N-hydroxysuccinimide (0.44 g, 3.9 mmol, 1.1 eq.) was added to a solution of (2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoic acid **111** (0.91 g, 3.5 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (10 mL) at 0 °C. Dicyclohexylcarbodiimide (0.79 g, 3.9 mmol, 1.1 eq.) was added in one portion and the reaction left to warm to r.t. and stirred for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (20 mL) and washed sequentially with 5% (w/w) $NaHCO_3$ (10 mL), H_2O (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude *N*-hydroxysuccinimidyl (2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate **112** as a white powder of sufficient purity for further manipulation (0.96 g, 76%); δ_H (500 MHz, DMSO-d_6): 7.32 (major rotamer) and 6.92 (minor) (d, 1H, ^3J 8.3 Hz, NH), 4.48-4.52 (major) and 4.33-4.37 (minor) (m, 1H, H_α), 3.62-3.68 (m, 2H, H_β), 2.80 (s, 4H, Su CH_2), 1.40 (major) and 1.37 (minor) (s, 9H, Boc CH_3), 1.14 (major) and 1.08 (minor) (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6): 169.8 (Su CO), 167.0 (major) and 166.3 (minor) (CO_2Su, mixture of rotamers), 155.1 (major) and 154.7 (minor) (CO_2NH), 79.4 (minor) and 78.8 (major) (Boc $\underline{C}(CH_3)_3$), 73.2 (ether $\underline{C}(CH_3)_3$), 61.0 (major) and 60.7 (minor) (C_β) 54.3 (minor) and 53.0 (major) (C_α), 28.1 (major) and 27.6 (minor) (Boc CH_3), 27.0 (ether CH_3), 25.4 (Su CH_2); IR (ATR):

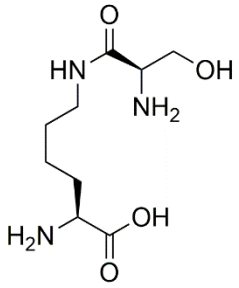
3410, 2980, 2934, 1819, 1783, 1735, 1713, 1475, 1445, 1426, 1365, 1195, 1165, 1102, 1082, 1065, 1046, 1035, 995, 912, 866, 810, 775, 749, 647, 577, 553; HRMS: found $[M+Na]^+$ 381.1638; $C_{16}H_{26}N_2O_7Na$ requires 381.1632 (Δ = -1.3 ppm).

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid 113



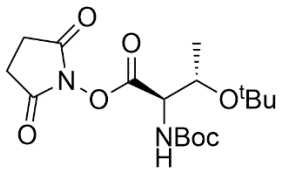
DIPEA (1.75 mL, 10 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl (2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanoate **112** (0.90 g, 2.5 mmol, 1.0 eq.) in DCM (8 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (0.62 g, 2.5 mmol, 1.0 eq.) in DCM (12 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (30 mL) and washed with sat. citric acid (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid **113** as a white foam of sufficient purity for further manipulation (1.04 g, 85%); δ_H (500 MHz, DMSO-d_6) 7.77 (t, 1H, ^3J 5.7 Hz, HN_ε), 6.97 (d, 1H, ^3J 8.0 Hz, HN_α), 6.42 (d, 1H, ^3J 8.2 Hz, $HN_{\alpha'}$), 3.78-3.97 (m, 2H, $H_{\alpha'}$, H_α), 3.38-3.44 (m, 2H, $H_{\beta'}$), 2.96-3.09 (m, 2H, H_ε), 1.20-1.65 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ), 1.09 (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6) 174.3 (CO_2H), 169.8 (CONH), 155.6, 155.0 (Boc CO_2N), 78.1, 77.9 (Boc $\underline{C}(CH_3)_3$), 72.6 (ether $\underline{C}(CH_3)_3$), 62.0 ($C_{\beta'}$), 54.9 ($C_{\alpha'}$), 53.4 (C_α), 38.2 (C_ε), 30.3 (C_β), 28.6 (C_γ), 28.2, 28.1 (Boc CH_3), 27.2 (ether CH_3), 22.9 (C_δ); IR (ATR): 3331, 2976, 1705, 1501, 1392, 1365, 1247, 1161, 1088, 1019, 860, 779, 735; HRMS: found $[M+Na]^+$ 512.2941; $C_{23}H_{43}N_3O_8Na$ requires 512.2942 (Δ = 0.3 ppm).

(2*S*)-2-amino-6-((2*R*)-2-amino-3-hydroxypropanamido)-hexanoic acid 109



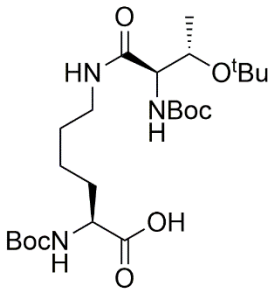
Deprotection cocktail 95:2.5:2.5 (v/v) TFA:TIS:H_2O (4 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid **113** (1.01 g, 2.06 mmol) in DCM (6 mL) at 0 °C and left to reach rt. Once the reaction was determined complete by TLC, usually within 3 h, the reaction mixture was triturated with ice-cold Et_2O (30 mL). The suspension was centrifuged at 4 000 × *g* for 5 min, supernatant decanted, and the peptide pellet resuspended in ice-cold Et_2O (20 mL), with this process repeated 5 times in total, after which the pellet was left to air dry. The resulting off-white powder was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford (2*S*)-2-amino-6-((2*R*)-2-amino-3-hydroxypropanamido)-hexanoic acid **109** as a fluffy off-white foam (657 mg, diacetate salt, 93%); $[\alpha]_D$ +31.8 (c 1.0, H_2O); δ_H (500 MHz, DMSO-d_6): 8.45 (t, 1H, ^3J 5.6 Hz, HN_ε), 3.66-3.82 (m, 4H, H_α, $H_{\alpha'}$, $H_{\beta'}$), 3.07-3.11 (m, 2H, H_ε), 1.71-1.82 (m, 2H, H_β), 1.28-1.46 (m, 4H, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6): 171.2 (CO_2H), 166.8 (CONH), 60.5 ($C_{\beta'}$), 54.6 ($C_{\alpha'}$), 52.2 (C_α), 38.6 (C_ε), 29.8, (C_β), 28.4 (C_γ), 21.8 (C_δ); IR (ATR): 3317, 2946, 1661, 1581, 1519, 1428, 1278, 1255, 1184, 1129, 1039, 840, 802, 720, 600, 518; HRMS: found $[M+H]^+$ 234.1448; $C_9H_{20}N_3O_4$ requires 234.1448 (Δ = 0.3 ppm).

N-hydroxysuccinimidyl (2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonyl-amino)-butanoate 115



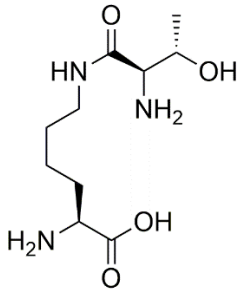
N-hydroxysuccinimide (257 mg, 2.2 mmol, 1.1 eq.) was added to a solution of (2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoic acid **114** (561 mg, 2.0 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (6 mL) at 0 °C. Dicyclohexylcarbodiimide (460 mg, 2.2 mmol, 1.1 eq.) was added in one portion and the reaction left to warm to r.t. and stirred for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (10 mL) and washed sequentially with 5% (w/w) $NaHCO_3$ (10 mL), H_2O (5 mL) and brine (5 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude *N*-hydroxysuccinimidyl (2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoate **115** as a white foam of sufficient purity for further manipulation (483 mg, 64%); δ_H (500 MHz, DMSO-d_6): 6.97 (d, 1H, ^3J 9.3 Hz, $N_\alpha H$), 4.42 (dd, 1H, ^3J 9.3 Hz, $^3J_{\alpha\text{-}\beta}$ 4.1 Hz, H_α), 4.02-4.07 (m, 1H, H_β), 2.80 (s, 4H, Su H), 1.40 (s, 9H, Boc CH_3), 1.19 (d, 3H, $^3J_{\beta\text{-}\gamma}$ 6.2 Hz, H_γ), 1.14 (s, 9H, ether CH_3); δ_C (126 MHz, DMSO-d_6): 169.9 (Su CO), 166.8 (CO_2H), 155.2 (Boc CO_2NH), 78.9 (Boc $\underline{C}(CH_3)_3$), 73.9 (ether $\underline{C}(CH_3)_3$), 67.2 (C_β), 57.9 (C_α), 28.2 (Boc CH_3), 28.1 (ether CH_3), 25.4 (Su CH_2), 19.6 (C_γ); IR (ATR): 3350, 2978, 2933, 1819, 1781, 1731, 1710, 1512, 1456, 1367, 1242, 1205, 1152, 1120, 1090, 1059, 957, 914, 868, 825, 791, 644, 605, 546; HRMS: found $[M+Na]^+$ 395.1792; $C_{17}H_{28}N_2O_7Na$ requires 395.1789 (Δ = -0.5 ppm).

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid 116

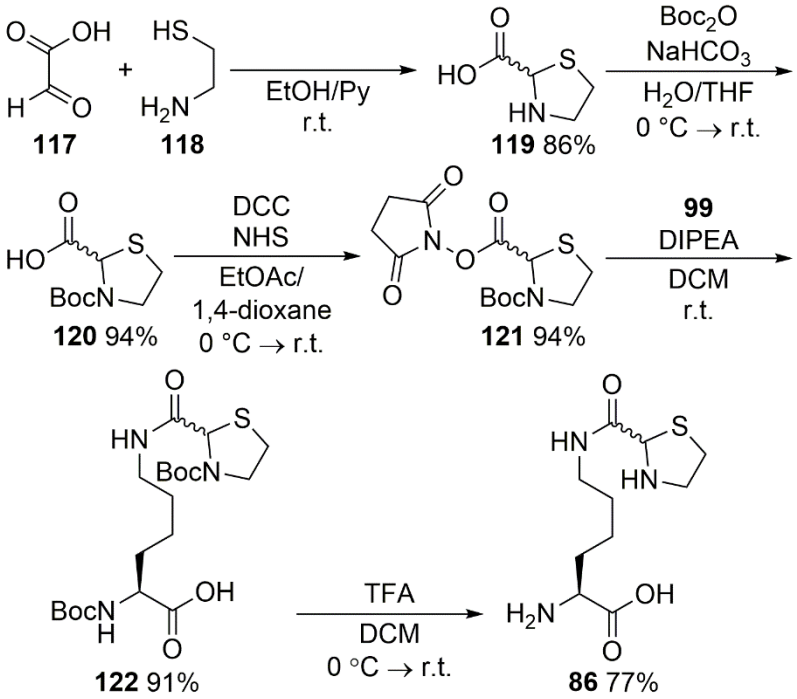


DIPEA (0.6 mL, 5.2 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl (2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanoate **115** (483 mg, 1.3 mmol, 1.0 eq.) in DCM (6 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (320 mg g, 1.3 mmol, 1.0 eq.) in DCM (9 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (15 mL) and washed with sat. citric acid (5 mL) and brine (5 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford crude (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)hexanoic acid **116** as a white foam of sufficient purity for further manipulation (635 mg, 97%); δ_H (500 MHz, DMSO-d_6) 7.71 (t, 1H, ^3J 5.6 Hz, HN_ε), 6.98 (d, 1H, ^3J 8.0 Hz, HN_α), 5.95 (t, 1H, ^3J 9.2 Hz, $HN_{\alpha'}$), 3.70-3.90 (m, 3H, H_α, $H_{\alpha'}$, $H_{\beta'}$), 2.97-3.08 (m, 2H, H_ε), 1.21-1.73 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ), 1.08 (s, 9H, ether CH_3), 1.00 (d, 3H, $^3J_{\beta'\text{-}\gamma'}$ 6.2 Hz, $H_{\gamma'}$); δ_C (126 MHz, DMSO-d_6) 174.2 (CO_2H), 169.9 (CONH), 155.6, 155.1 (Boc CO_2N), 78.3, 77.9 (Boc $\underline{C}(CH_3)_3$), 73.3 (ether $\underline{C}(CH_3)_3$), 67.3 ($C_{\beta'}$), 59.4 ($C_{\alpha'}$), 53.4 (C_α), 38.5 (C_ε), 30.4 (C_β), 28.5 (C_γ), 28.2, 28.1 (Boc CH_3), 28.0 (ether CH_3), 23.1 (C_δ), 20.0 ($C_{\gamma'}$); IR (ATR): 3320, 2979, 1708, 1663, 1496, 1392, 1366, 1249, 1161, 1065, 1022, 959, 857, 779; HRMS: found $[M+Na]^+$ 526.3108; $C_{24}H_{45}N_3O_8Na$ requires 526.3099 (Δ = -0.3 ppm).

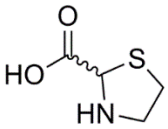
(2*S*)-2-amino-6-((2*S*,3*R*)-2-amino-3-hydroxybutanamido)-hexanoic acid 110



Deprotection cocktail 95:2.5:2.5 (v/v) TFA:TIS:H_2O (2 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid **116** (588 mg, 1.2 mmol) in DCM (3 mL) at 0 °C and left to reach rt. Once the reaction was determined complete by TLC, *ca.* 3 h, the reaction mixture was triturated with ice-cold Et_2O (30 mL). The suspension was centrifuged at 4 000 × *g* for 5 min, supernatant decanted, and the peptide pellet resuspended in ice-cold Et_2O (20 mL), with this process repeated 5 times in total, after which the pellet was left to air dry. The resulting off-white solid was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford (2*S*)-2-amino-6-((2*R*,3*S*)-2-amino-3-hydroxybutanamido)hexanoic acid **110** as a fluffy off-white foam (539 mg, di(trifluoroacetate) salt, 97%); $[\alpha]_D$ -52.2 (c 1.0, H_2O); δ_H (500 MHz, DMSO-d_6): 8.50 (t, 1H, ^3J 5.6 Hz, HN_ε), 3.85 (app. quint., 1H, $^3J_{\alpha'\text{-}\beta',\,\beta'\text{-}\gamma'}$ 6.4 Hz, $H_{\beta'}$), 3.77 (t, 1H, $^3J_{\alpha\text{-}\beta}$ 6.3 Hz, H_α), 3.46 (d, 1H, $^3J_{\alpha'\text{-}\beta'}$ 6.8 Hz, $H_{\alpha'}$), 3.15 (app. dq, 1H, ^2J 12.6 Hz, ^3J 6.5 Hz, H_ε), 3.05 (app. dq, 1H, ^2J 13.0 Hz, ^3J 6.5 Hz, H'_ε), 1.70-1.82 (m, 2H, H_β), 1.24-1.47 (m, 4H, H_γ, H_δ) 1.12 (d, 3H, $^3J_{\beta'\text{-}\gamma'}$ 6.3 Hz, $H_{\gamma'}$); δ_C (126 MHz, DMSO-d_6): 171.0 (CO_2H), 166.7 (CONH), 65.8 ($C_{\beta'}$), 58.5 ($C_{\alpha'}$), 52.1 (C_α), 38.4 (C_ε), 29.7 (C_β), 28.2 (C_γ), 21.8 (C_δ), 20.0 ($C_{\gamma'}$); IR (ATR): 2980, 1662, 1581, 1519, 1429, 1250, 1184, 1130, 1038, 932, 838, 799, 721, 599, 517; HRMS: found $[M+H]^+$ 248.1606; $C_{10}H_{22}N_3O_4$ requires 248.1605 (Δ = 0.0 ppm).

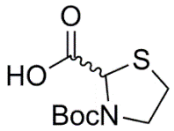


Thiazolidine 2-carboxylic acid 119



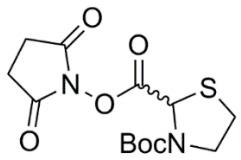
A solution of 2-mercaptoethylamine hydrochloride **118** (15.2 g, 134 mmol, 1.0 eq) in 5:2 (v/v) ethanol:pyridine (70 mL) was added to a stirred solution of glyoxylic acid **117** (50% solution in water, 21.8 g, 147 mmol, 1.1 eq.) in ethanol (25 mL). The reaction was left to stir for 2 h at r.t., after which the off-white precipitate was isolated by filtration and washed with ethanol, yielding racemic thiazolidine-2-carboxylic acid **119** as a white powder (15.4 g, 86%), used without further purification; δ_H (500 MHz, DMSO-d_6): 4.83 (s, 1H, H-2), 3.33-3.38 (m, 1H H-5), 2.97-3.02 (m, 1H, H-5'), 2.75-2.83 (m, 2H, H-4, H-4'); δ_C (126 MHz, DMSO-d_6): 172.0 (CO_2H), 66.3 (C-2), 52.8 (C-5), 34.5 (C-4); IR (ATR): 3110, 1622, 1590, 1376, 1354, 1323, 1300, 1277, 998, 881, 866, 721, 643; HRMS: Found $[M+H]^+$ 134.0270; $C_4H_8NO_2S$ requires 134.0270 (Δ = 0.2 ppm). Characterisation data in agreement with previous literature reports.[221]

3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylic acid 120



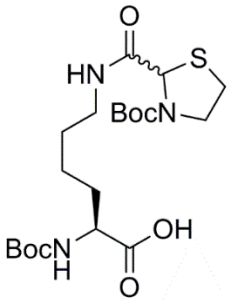
A stirred suspension of thiazolidine-2-carboxylic acid **119** (5.2 g, 39 mmol, 1.0 eq.) in 2:1 (v/v) H_2O:THF (120 mL) was cooled to 0 °C with stirring and $NaHCO_3$ (9.9 g, 118 mmol, 3.0 eq.) added in portions. Di-*tert*-butyl dicarbonate (9.0 g, 41 mmol, 1.05 eq.) was added in one portion and the mixture was left stirring for 18 h at r.t.. The reaction mixture was cooled to 0 °C, diluted with EtOAc (80 mL), and acidified to pH 3 using 12 M HCl. The organic layer was extracted, washed with brine (80 mL), dried over $MgSO_4$ and concentrated to yield racemic 3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylic acid **120** as a white solid (8.6 g, 94%); δ_H (500 MHz, DMSO-d_6): 5.10 and 5.03 (s, 1H, H-2, mixture of rotamers), 3.70 (br app. s, 2 H, H-5, H-5'), 3.06 (br app. s, 2H, H-4, H-4'), 1.40 and 1.36 (s, 9H, Boc CH_3, mixture of rotamers); δ_C (126 MHz, DMSO-d_6): 172.0 and 171.6 (CO_2H, mixture of rotamers), 152.9 and 152.1 (Boc CO_2NH, mixture of rotamers), 80.0 ($\underline{C}(CH_3)_3$), 59.5 and 59.0 (C-2, mixture of rotamers), 49.7 and 49.3 (C-5, mixture of rotamers), 30.3 and 29.1 (C-4, mixture of rotamers), 28.0 and 27.8 (Boc CH_3, mixture of rotamers); IR (ATR): 2982, 1804, 1756, 1371, 1211, 1113, 1062, 843, 774; HRMS: found $[M+Na]^+$ 256.0610; $C_9H_{15}NO_4SNa$ requires 256.0614 (Δ = 1.9 ppm). Characterisation data in agreement with previous literature reports.[221]

_N_-hydroxysuccinimidyl 3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylate 121



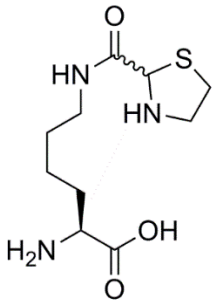
N-hydroxysuccinimide (0.61 g, 5.3 mmol, 1.05 eq.) was added to a solution of 3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylic acid **120** (1.17 g, 5.0 mmol, 1.0 eq.) in 1:1 (v/v) EtOAc:1,4-dioxane (16 mL) at 0 °C. Dicyclohexylcarbodiimide (1.09 g, 5.3 mmol, 1.05 eq.) was added in one portion and the reaction left to warm to r.t. and stir for 3 h. The reaction mixture was filtered through a Celite pad and concentrated. The crude product was redissolved in EtOAc (20 mL) and washed sequentially with 5% (w/v) $NaHCO_3$ (10 mL), H_2O (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford racemic *N*-hydroxysuccinimidyl 3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylate **121** as an off-white powder (1.56 g, 94%); δ_H (500 MHz, DMSO-d_6): 5.72 and 5.66 (s, 1H, H-2, mixture of rotamers), 3.81 (br app. s, 1H, H-5), 3.69 (br app. s, 1H, H-5'), 3.20 (br app. s, 2H, H-4, H-4'), 2.81 (s, 4H, OSu CH_2), 1.42 (s, 9H, Boc CH_3); δ_C (126 MHz, DMSO-d_6): 169.7 (Su CO), 167.0 (CO_2Su), 151.5 (CO_2NH), 81.1 ($\underline{C}(CH_3)_3$), 58.1 and 57.4 (C-2, mixture of rotamers), 49.7 and 49.5 (C-5, mixture of rotamers**),** 29.4 (C-4), 27.8 and 27.5 (Boc CH_3), 25.5 and 25.0 (Su CH_2); IR (ATR): 2980, 1810, 1774, 1732, 1699, 1455, 1386, 1367, 1205, 1162, 1140, 1114, 1062, 1046, 991, 793, 648; HRMS: found $[M+Na]^+$ 353.0781; $C_{13}H_{18}N_2O_6SNa$ requires 353.0778.0614 (Δ = -0.7 ppm).

(2*S*)-2-(*tert*-butoxycarbonylamino)-6-(thiazolidine-2-carboxamido-3-(*tert*-butoxycarbonyl))-hexanoic acid 122



DIPEA (2.8 mL, 16 mmol, 4.0 eq.) and a solution of *N*-hydroxysuccinimidyl 3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylate **121** (1.32 g, 4.0 mmol, 1.0 eq.) in DCM (12 mL) were added sequentially at r.t. to a stirred suspension of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-aminohexanoic acid **99** (0.98 g, 4.0 mmol, 1.0 eq.) in DCM (20 mL). The reaction mixture was stirred at r.t. for 3 h, filtered and concentrated. The residue was redissolved in DCM (30 mL) and washed with sat. citric acid (10 mL) and brine (10 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated to afford a diastereomeric mixture of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-(thiazolidine-2-carboxamido-3-(*tert*-butoxycarbonyl))-hexanoic acid **122** as an off-white foam (1.68 g, 91%); δ_H (500 MHz, DMSO-d_6) 7.96 (s, 1H, HN_ε), 6.97 (d, 1H, ^3J 7.9 Hz, HN_α), 5.14 and 5.02 (s, 1H, H-2, mixture of rotamers), 3.67-3.81 (m, 3H, H_α, H-5, H-5'), 2.99-3.14 (m, 4H, H_ε, H-4, H-4'), 1.22-1.64 (m, 24H, Boc CH_3, Boc CH_3, H_β, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6) 174.2 (CO_2H), 170.1 (CONH), 155.6 and 152.23 (Boc CO_2N, mixture of rotamers), 79.6 and 77.9 ($\underline{C}(CH_3)_3$, mixture of rotamers), 60.2 (C-2), 53.4 (C_α), 50.0 (C-5), 38.3 (C_ε), 30.4 (C_β), 29.0 and 28.6 (C_γ, mixture of rotamers), 28.2 and 27.9 (Boc CH_3), 22.9 (C_δ);); IR (ATR): 2930, 1668, 1528, 1390, 1365, 1255, 1157, 1047, 891, 857, 778, 731, 637; HRMS: found $[M+Na]^+$ 484.2095; $C_{20}H_{35}N_3O_7SNa$ requires 484.2088 (Δ = -1.1 ppm).

(2*S*)-2-amino-6-(thiazolidine-2-carboxamido)-hexanoic acid 86



Trifluoroacetic acid (60 mL) was added dropwise to a stirred solution of (2*S*)-2-(*tert*-butoxycarbonylamino)-6-(thiazolidine-2-carboxamido-3-(*tert*-butoxycarbonyl))-hexanoic acid **122** (26.3 g, 57.0 mmol) in DCM (120 mL) at 0 °C and left to reach r.t.. Once the reaction was determined complete by TLC, usually within 3 h, the reaction mixture was triturated with ice-cold Et_2O (240 mL), filtered, and washed with ice-cold Et_2O (4 × 50 mL). The resulting off-white powder was redissolved in 10% (v/v) aqueous acetic acid and lyophilised to afford a diastereomeric mixture of (2*S*)-2-amino-6-(thiazolidine-2-carboxamido)-hexanoic acid **86** as a fluffy off-white powder (19.6 g, acetate/trifluoroacetate salt, 90%); δ_H (500 MHz, DMSO-d_6): 8.46 (t, 1H, ^3J 5.4 Hz, HN_ε), 8.29 (br s, 3H, H_2N_α, HN-2), 5.10 (s, 1H, H-2), 3.86 (t, 1H, ^3J 5.8 Hz, H_α), 3.53 (dt, 1H, ^2J 11.7 Hz, ^3J 6.0 Hz, ^3J 6.0 Hz, H-5), 3.38 (dt, 1H, ^2J 11.7 Hz, ^3J 6.5 Hz, ^3J 6.5 Hz, H-5'), 3.01-3.10 (m, 4H, H-4, H-4', H_ε), 1.73-1.80 (m, 2H, H_β), 1.29-1.45 (m, 4H, H_γ, H_δ); δ_C (126 MHz, DMSO-d_6): 171.0 (CO_2H), 167.5 (CONH), 62.2 (C-2), 51.9 (C_α), 50.4 (C-5), 38.6 (C_ε), 31.5 (C-4), 29.6 (C_β), 28.3 (C_γ), 21.7 (C_δ); IR (ATR): 2944, 1660, 1430, 1180, 1128, 836, 797, 720; HRMS: found $[M+H]^+$ 262.1220; $C_{10}H_{20}N_3O_3S$ requires 262.1220 (Δ = 1.6 ppm).

7.1.3 Solid-phase synthesis

General Methods

Preloaded resin preparation

The preloaded 2-chlorotrityl resin was weighed out into a 2 mL SPPS cartridge fitted with a PTFE stopcock, swollen in DCM for 30 min and then filtered.

Fmoc deprotection

A solution of 20% piperidine in DMF was added to the resin and gently agitated by rotation for 2 minutes. The resin was filtered off and repeated four more times, followed by washes with DMF (5 × 2 min with rotation).

Amino acid coupling

DIPEA (11 eq.) was added to a solution of amino acid (5 eq.) and HCTU (5 eq.) dissolved in the minimum volume of DMF and the solution added to the resin. The reaction mixture was gently agitated by rotation for 1 h, and the resin filtered off and washed with DMF (3 × 2 min with rotation).

Cleavage Cocktails

Deprotection and resin cleavage: 95:2.5:2.5 (v/v) TFA:H_2O:triisopropylsilane. Cleavage only: 4:1 (v/v) DCM: 1,1,1,3,3,3-hexafluoroisopropanol.

Cleavage and Isolation

The resin was washed with DCM (3 × 2 min with rotation) and MeOH (3 × 2 min with rotation). The resin was dried on a vacuum manifold and further dried on a high vacuum line overnight. A solution of cleavage cocktail was added to the resin and gently agitated by rotation for 60 min.

For peptides cleaved and deprotected: the reaction mixture was drained into ice-cold Et_2O and centrifuged at 5000 × *g* at 4 °C until pelleted (5-10 min). The supernatant was carefully decanted and subsequently resuspended, centrifuged and supernatant decanted three more times. The peptide pellet was dissolved in 10% (v/v) aq. AcOH and lyophilised.

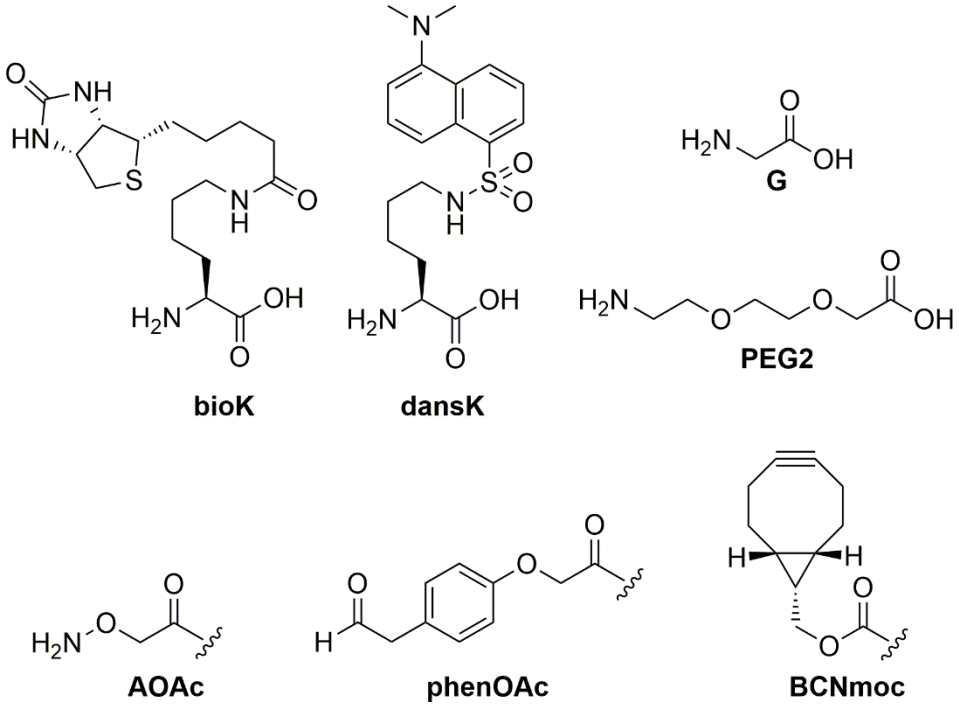
For proteins undergoing further couplings: the reaction mixture was drained and concentrated *in vacuo*.

Nomenclature

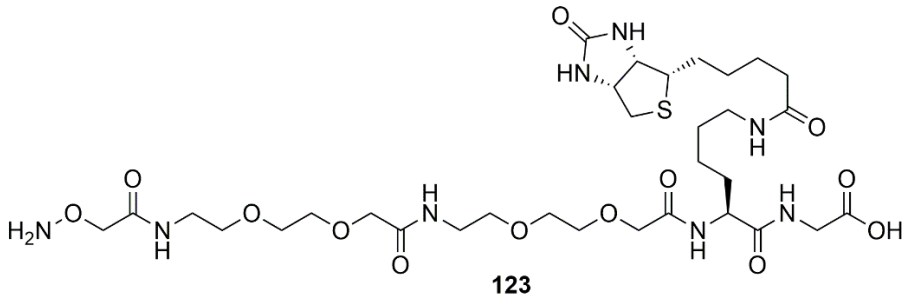
The names and structures of all standard and non-standard residues and *N*-terminal capping moieties featuring in the peptides synthesised are listed below.

Residues: bioK, N_ε-(+)-biotinyl L-lysine; dansK, N_ε-dansyl L-lysine; G, glycine; PEG2, 8-amino-2,6-dioxaoctanoic acid.

Caps: AOAc, aminooxyacetyl; phenOAc, phenacetaldehyde-4-oxyacetyl; BCNmoc, *endo*-bicyclo[6.1.0]non-4-yn-9-ylmethoxycarbonyl.



AOAc-PEG2-PEG2-bioK-G-OH 123

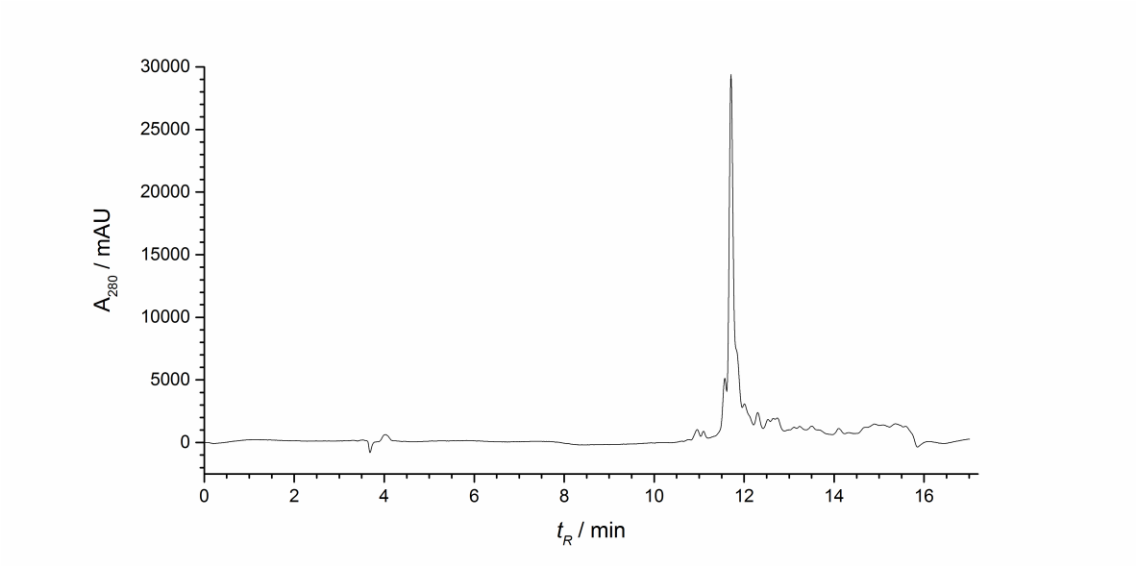


123

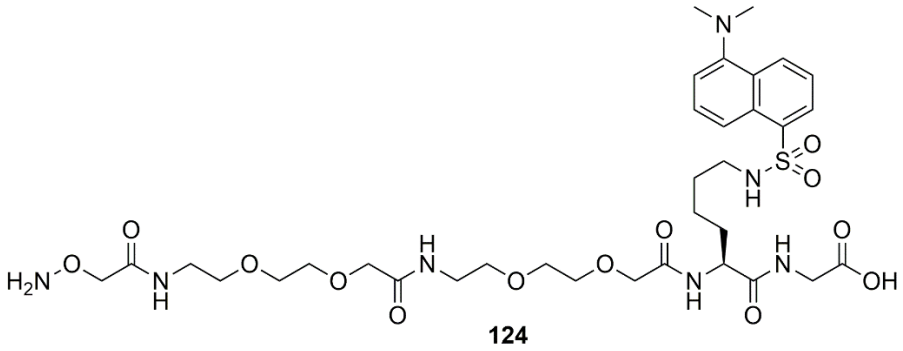
This peptide was synthesised by Dr Martin Fascione.

Peptide **123** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (168 mg, loading 0.53 mmol g^{-1}, 0.089 mmol), Fmoc-bioK-OH and Fmoc-PEG2-OH and capped with 3 eq. Boc-NHOAc-OSu, no HCTU. Resin cleavage and deprotection, purification and lyophilisation afforded the peptide as a fluffy white powder (10.7 mg, 16%).

HRMS: found $[M+H]^+$ 793.3784; $C_{32}H_{57}N_8O_{13}S$ requires 793.3760 (Δ = -2.4 ppm); HPLC: t_R = 11.70 min.



AOAc-PEG2-PEG2-dansK-G-OH 124

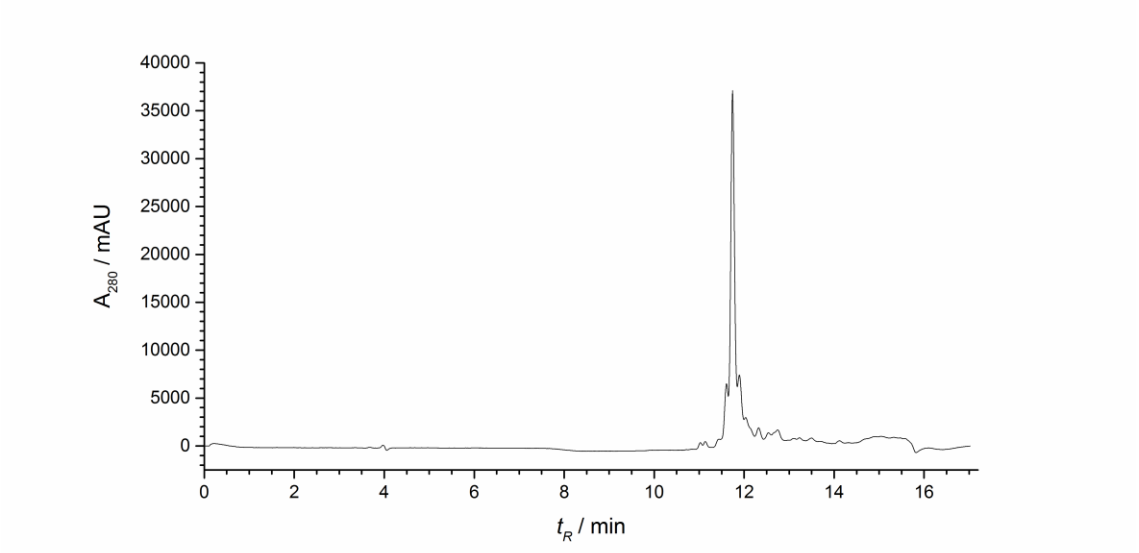


124

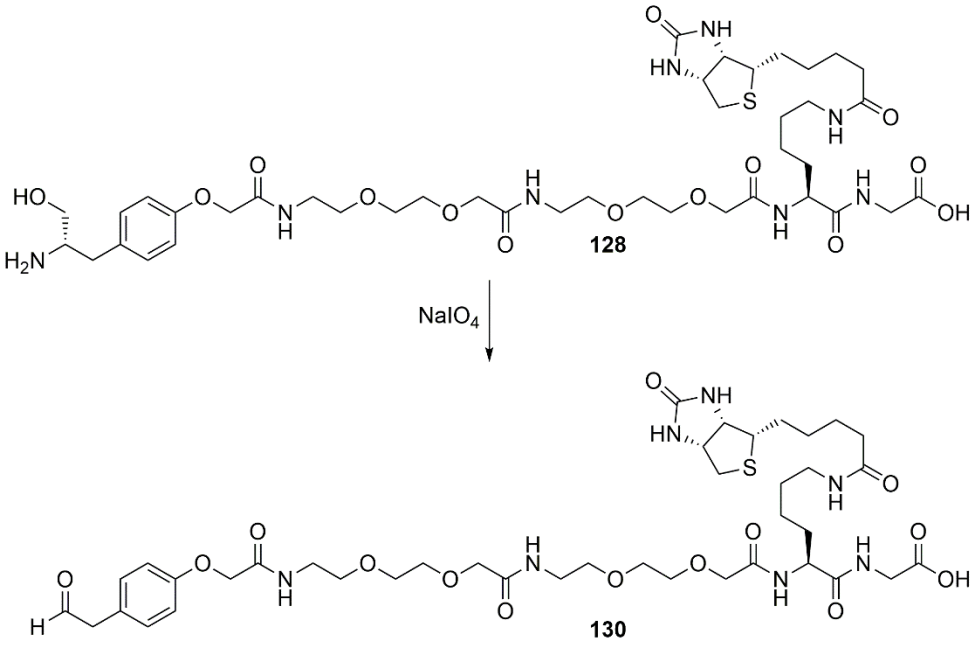
This peptide was synthesised by Dr Martin Fascione.

Peptide **124** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (100 mg, loading 0.53 mmol g^{-1}, 0.053 mmol), Fmoc-dansK-OH and Fmoc-PEG2-OH and capped with 3 eq. Boc-NHOAc-OSu, no HCTU. Resin cleavage and deprotection, purification and lyophilisation afforded the peptide as a fluffy yellow powder (6.8 mg, 15%).

HRMS: found $[M+H]^+$ 800.3504; $C_{34}H_{54}N_7O_{13}S$ requires 800.3495 (Δ = -1.4 ppm); HPLC: t_R = 11.74 min.

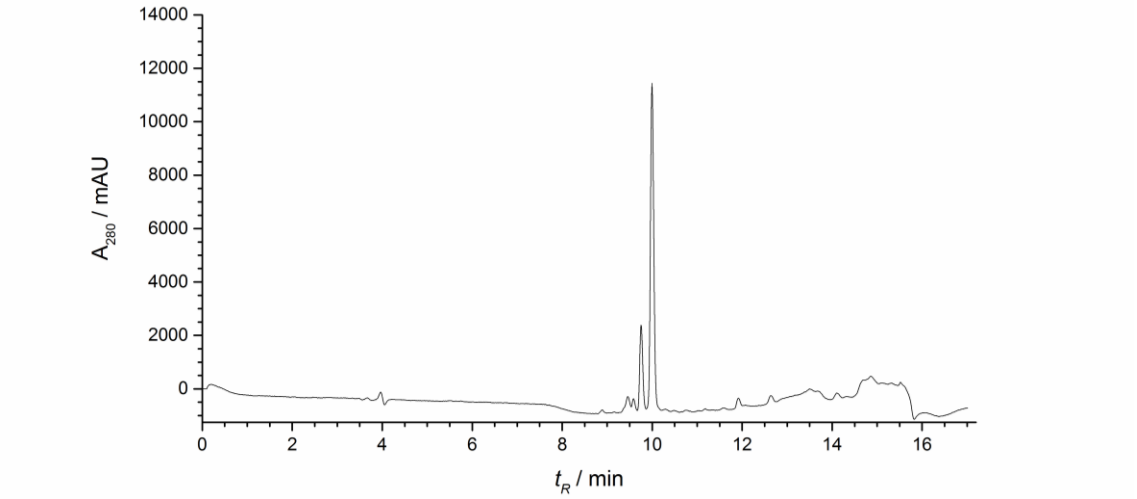


phenOAc-PEG2-PEG2-bioK-G-OH 130



Precursor peptide **128** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (100 mg, loading 0.54 mmol g^{-1}, 0.054 mmol), Fmoc-bioK-OH and Fmoc-PEG2-OH and capped with **125**. Resin cleavage and deprotection, purification and lyophilisation afforded the peptide as a fluffy white powder (41 mg, 98%).

HRMS: found $[M+H]^+$ 782.3781; $C_{35}H_{56}N_7O_{11}S$ requires 782.3753; HPLC: t_R = 9.99 min.

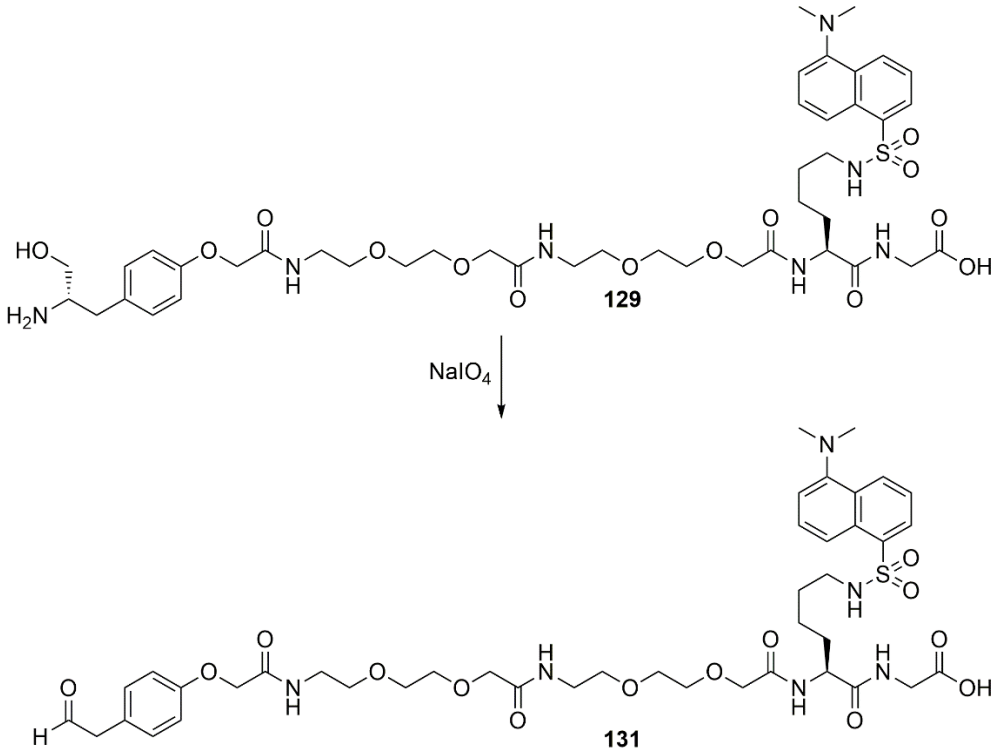


This peptide was oxidised and purified by Dr Richard Spears.

Solutions of L-methionine (stock 200 mM in 0.1 M phosphate buffer, 0.1 M NaCl, pH 7.0, 250 µL, final concentration 55 mM) and $NaIO_4$ (stock 112 mM in 0.1 M phosphate buffer, 0.1 M NaCl, pH 7.0, 210 µL, final concentration 26 mM) were added to peptide **128** (stock 26 mM in 0.1 M phosphate buffer, 0.1 M NaCl, pH 7.0, 500 µL, final concentration 14 mM). The reaction was mixed thoroughly and left in the dark at 0 °C for 2 min. The

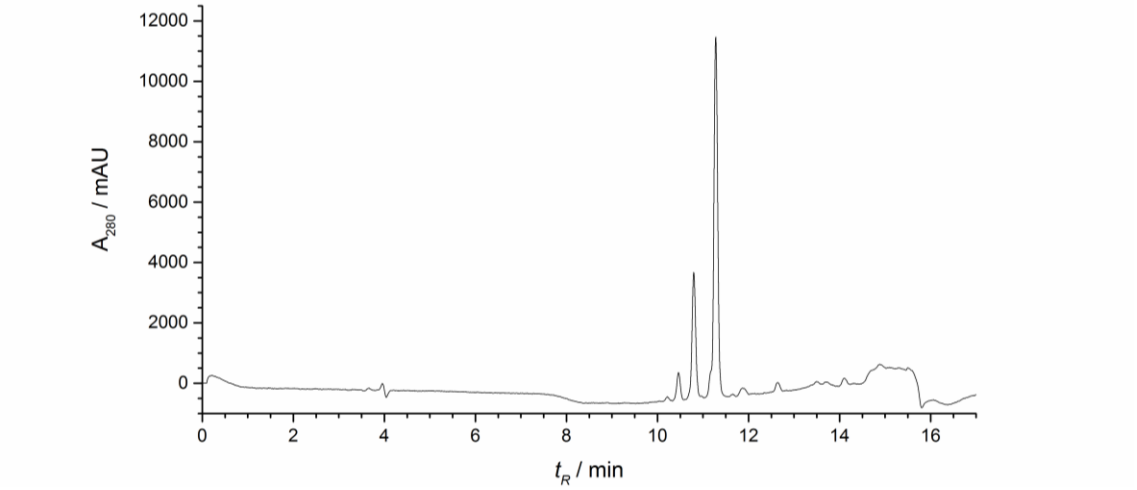
peptide was then purified by reversed-phase chromatography (Grace Davison Extract Clean, equilibrated in 1:1 H_2O:MeCN, eluted with 0-100% MeCN in H_2O gradient), diluted with water and lyophilised to afford **130** as a white fluffy powder (9 mg, 84%).

phenOAc-PEG2-PEG2-dansK-G-OH 131



Precursor peptide **129** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (85 mg, loading 0.54 mmol g^{-1}, 0.046 mmol), Fmoc-dansK-OH and Fmoc-PEG2-OH and capped with **125**. Resin cleavage and deprotection, purification and lyophilisation afforded the peptide as a fluffy yellow powder (33 mg, 92%).

HRMS: found $[M+H]^+$ 789.3511; $C_{37}H_{53}N_6O_{11}S$ requires 789.3488 ; HPLC: t_R = 11.28 min.

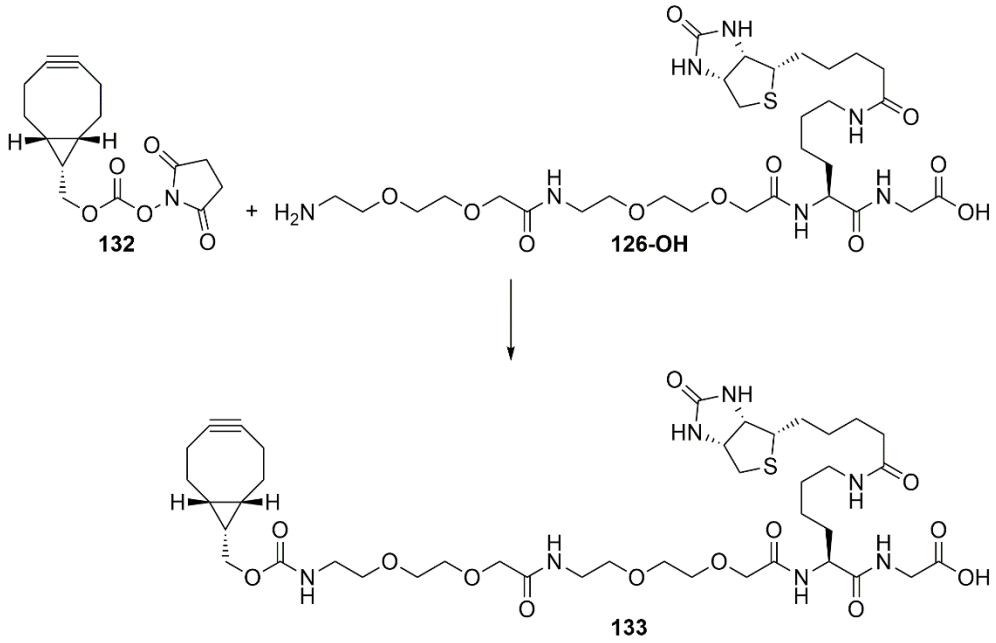


This peptide was oxidised and purified by Dr Richard Spears.

Solutions of L-methionine (stock 200 mM in 0.1 M phosphate buffer, 0.1 M NaCl, pH 7.0, 250 µL, final concentration 55 mM) and $NaIO_4$ (stock 112 mM in 0.1 M phosphate buffer,

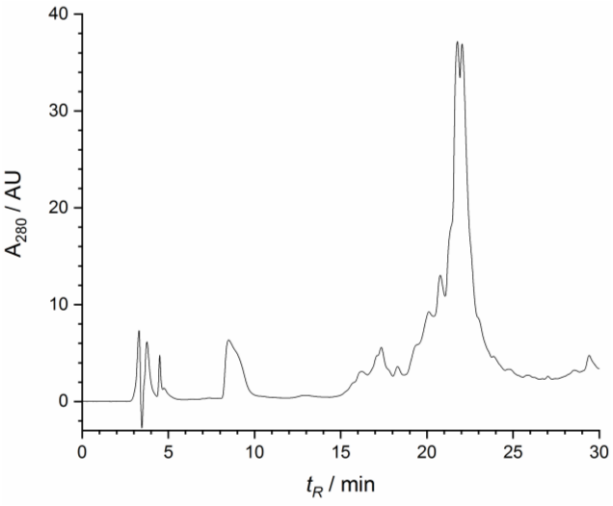
0.1 M NaCl, pH 7.0, 210 µL, final concentration 26 mM) were added to peptide **129** (stock 26 mM in 0.1 M phosphate buffer, 0.1 M NaCl, pH 7.0, 500 µL, final concentration 14 mM). The reaction was mixed thoroughly and left in the dark at 0 °C for 2 min. The peptide was then purified by reversed-phase chromatography (Grace Davison Extract Clean, equilibrated in 1:1 H_2O:MeCN, eluted with 0-100% MeCN in H_2O gradient), diluted with water and lyophilised to afford **131** as a white fluffy powder (4 mg, 40%).

BCNmoc-PEG2-PEG2-bioK-G-OH 133

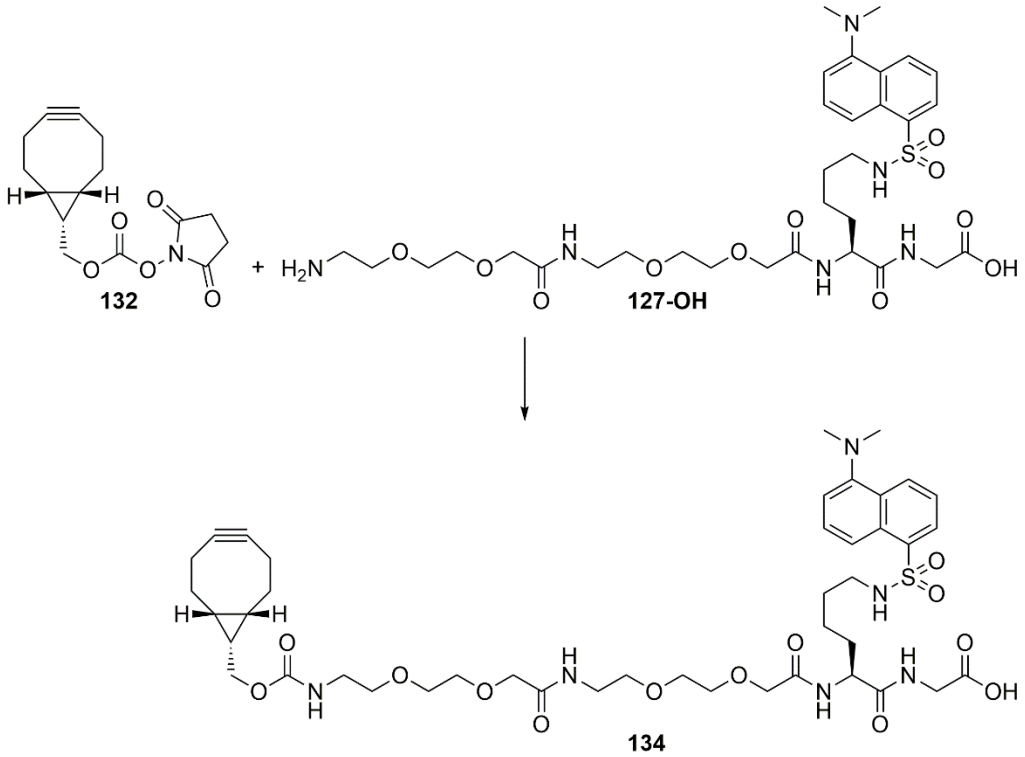


Precursor peptide **126-OH** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (75 mg, loading 0.52 mmol g^{-1}, 0.039 mmol), Fmoc-bioK-OH and Fmoc-PEG2-OH and isolated following resin cleavage (no deprotection), purification and concentration. The residue was dissolved in DMF (1 mL) and BCNmoc-OSu **132** (13 mg, 0.044 mmol, 1.1 eq.) and DIPEA (9 µL, 0.049 mmol, 1.2 eq.) added. The reaction was left stirring at r.t. for 16 h, then concentrated, redissolved in MeOH and purified by size-exclusion chromatography (Sephadex LH-20, equilibrated in MeOH, eluted with isocratic MeOH) to afford **133** as a fluffy white powder (24 mg, 68%).

HRMS: found $[M+Na]^+$ 918.4283; $C_{41}H_{65}N_7O_{13}S$ requires 918.4253 (Δ = -2.5 ppm); HPLC: t_R = 21.91 min.

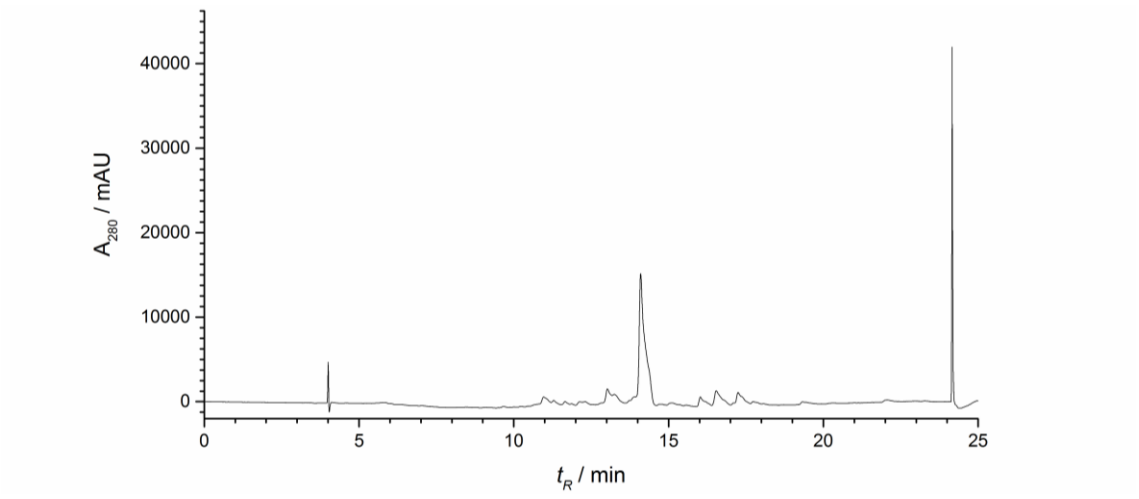


BCNmoc-PEG2-PEG2-dansK-G-OH 134



Precursor peptide **127-OH** was synthesised according to the general SPPS protocol using preloaded H-Gly-2ClTrt resin (127 mg, loading 0.52 mmol g^{-1}, 0.066 mmol), Fmoc-bioK-OH and Fmoc-PEG2-OH and isolated following resin cleavage (no deprotection), purification and concentration. The residue was dissolved in DMF (2 mL) and BCNmoc-OSu **132** (21 mg, 0.072 mmol, 1.1 eq.) and DIPEA (14 µL, 0.079 mmol, 1.2 eq.) added. The reaction was left stirring at r.t. for 16 h, then concentrated, redissolved in MeOH and purified by size-exclusion chromatography (Sephadex LH-20, equilibrated in MeOH, eluted with isocratic MeOH) to afford **134** as a fluffy yellow powder (33 mg, 55%).

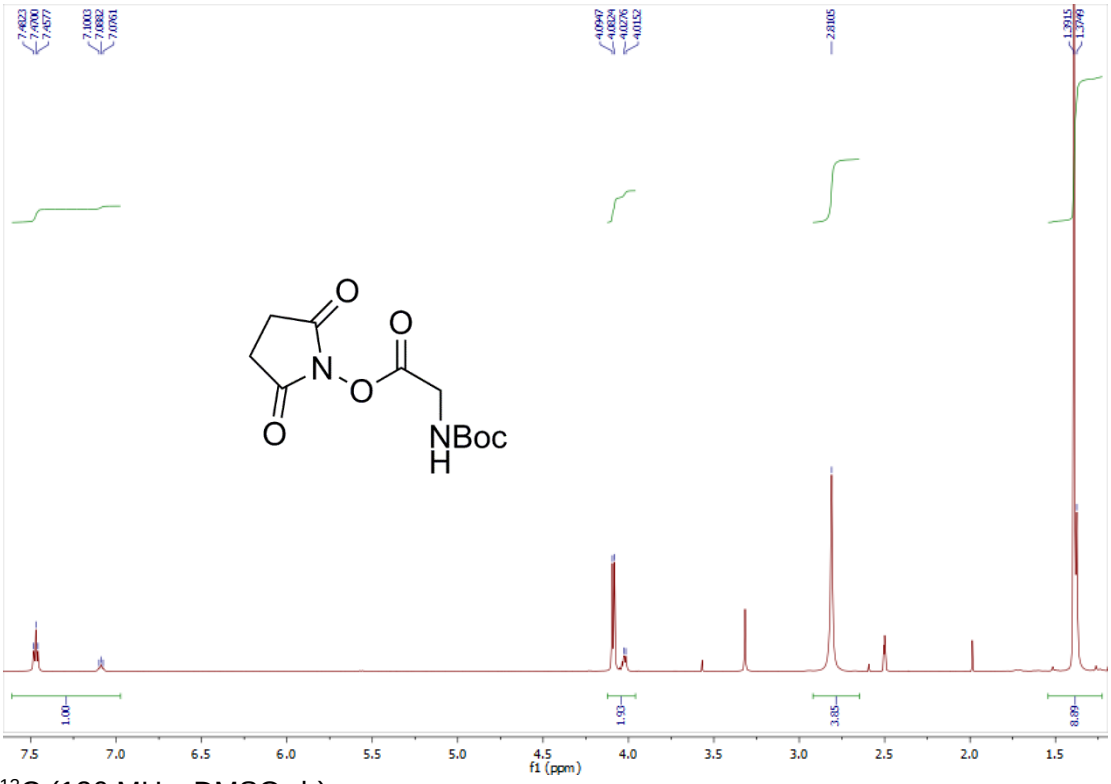
HRMS: found $[M+Na]^+$ 925.4010; $C_{43}H_{62}N_6O_{13}SNa$ requires 925.3988 (Δ = -1.1 ppm); HPLC: t_R = 14.12 min.



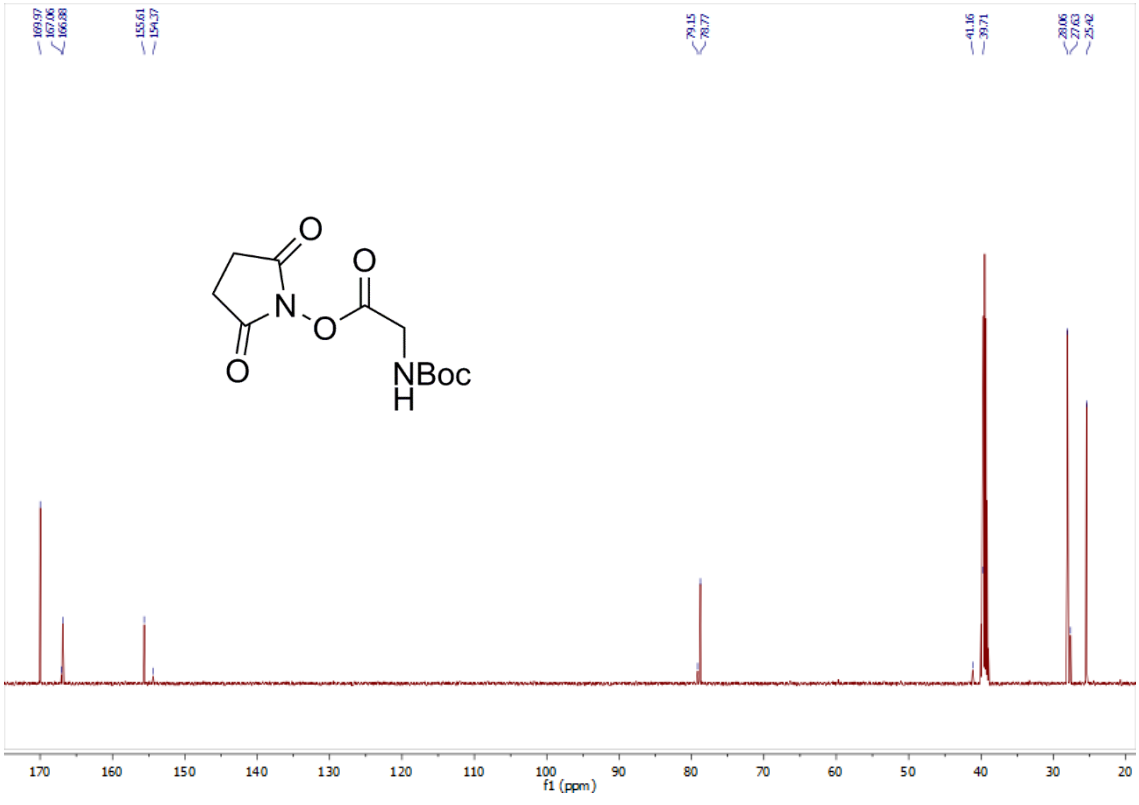
7.1.4 NMR spectra

N-hydroxysuccinimidyl 2-(*tert*-butoxycarbonylamino)-acetate 98

^{1}H (500 MHz, DMSO-d_6):

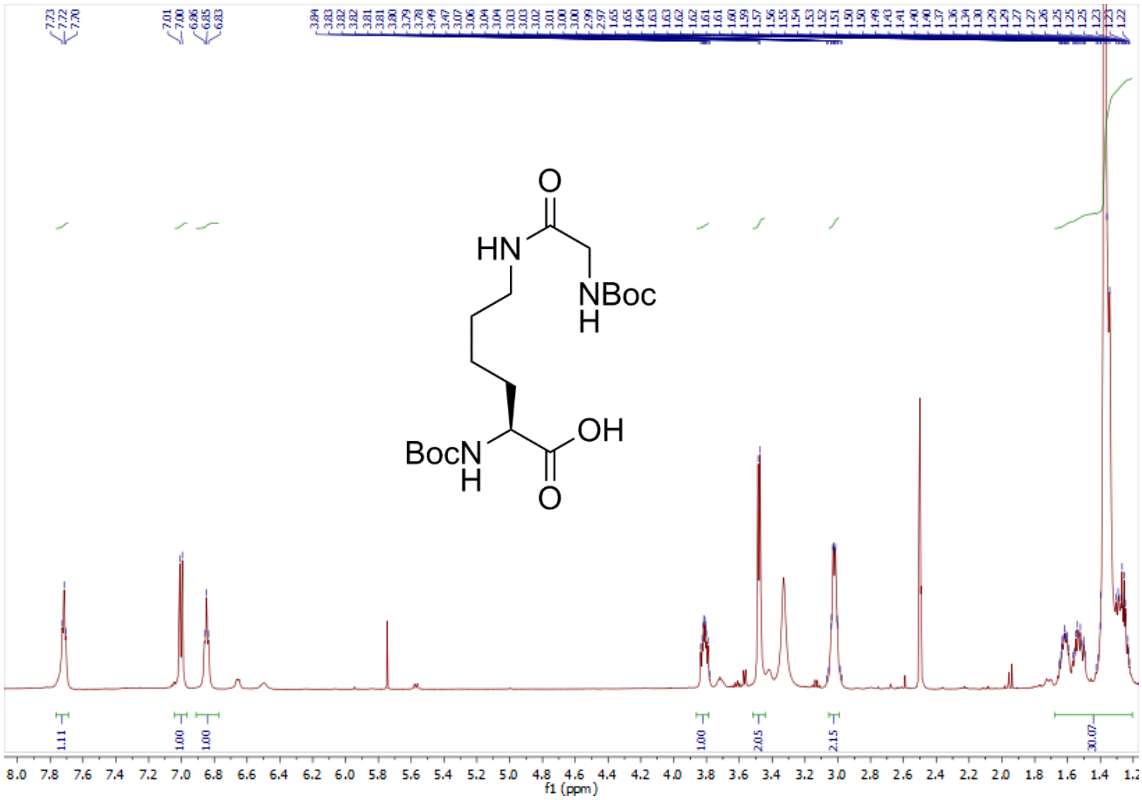


^{13}C (126 MHz, DMSO-d_6):

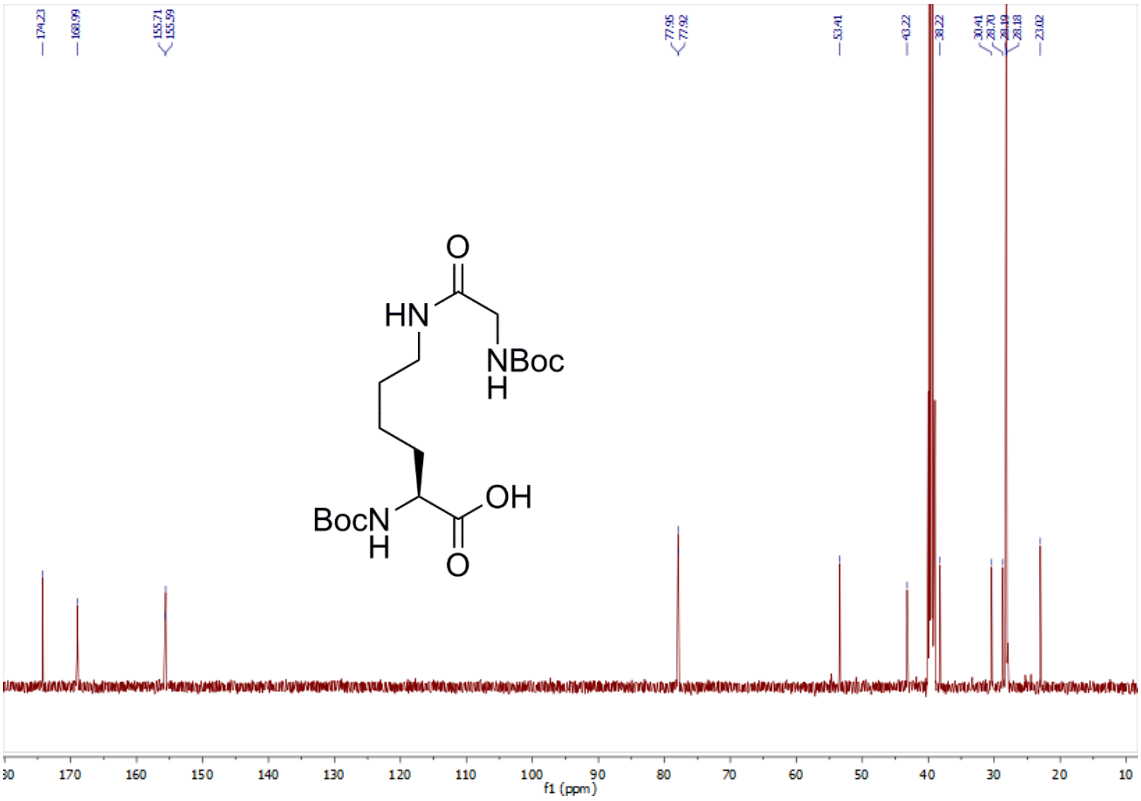


(2S)-2-(tert-butoxycarbonylamino)-6-(2-(tert-butoxycarbonylamino)-acetamido)hexanoic acid 100

^{1}H (500 MHz, DMSO-d_6):

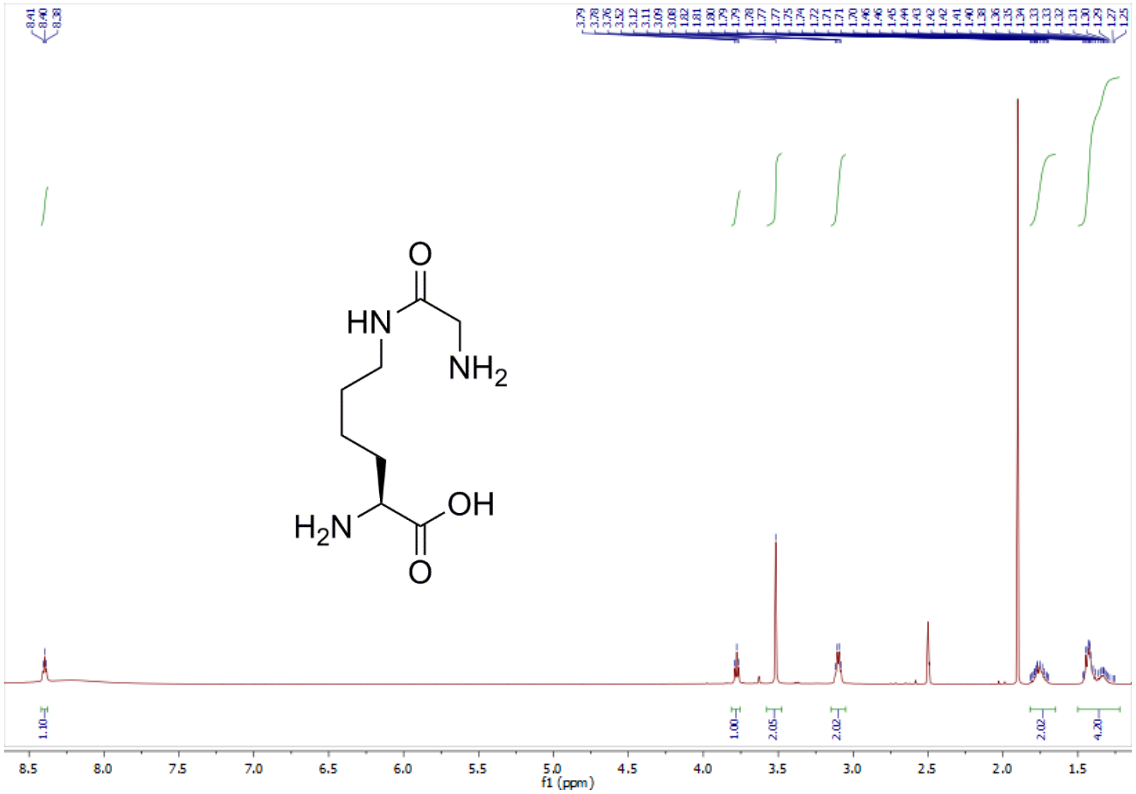


^{13}C (126 MHz, DMSO-d_6):

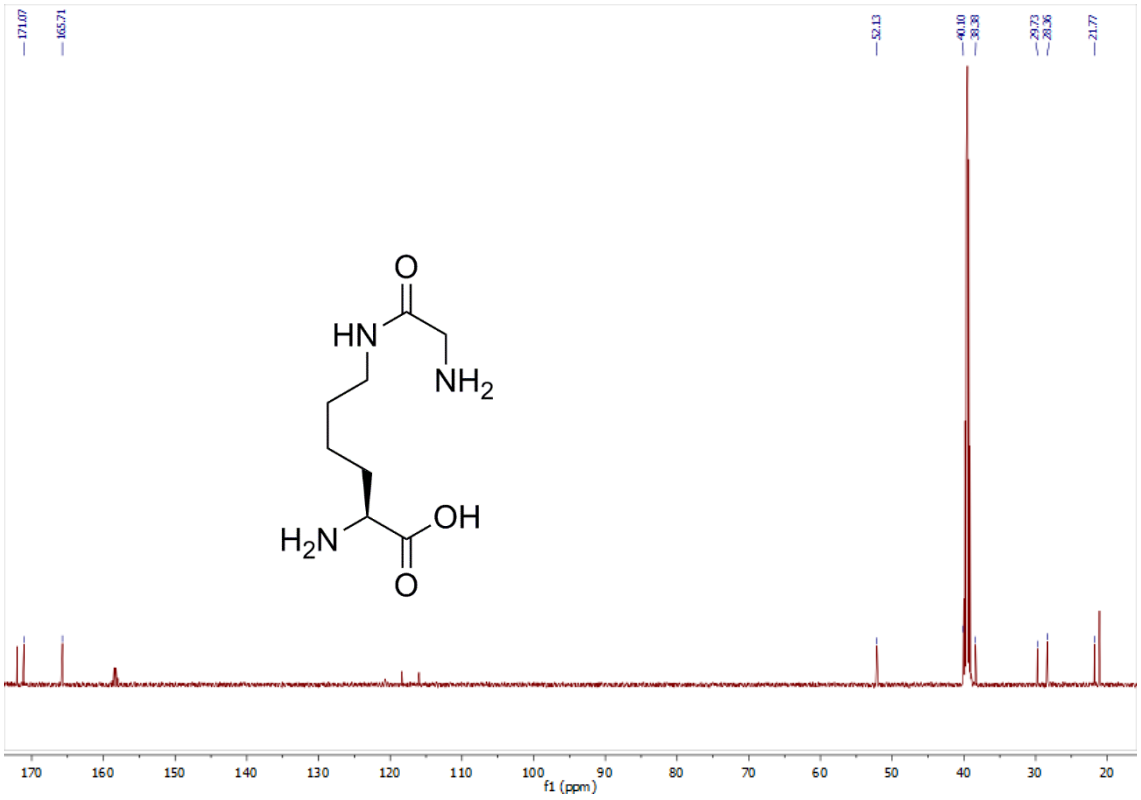


(2*S*)-2-amino-6-(2-aminoacetamido)hexanoic acid 96

^{1}H (500 MHz, DMSO-d_6):

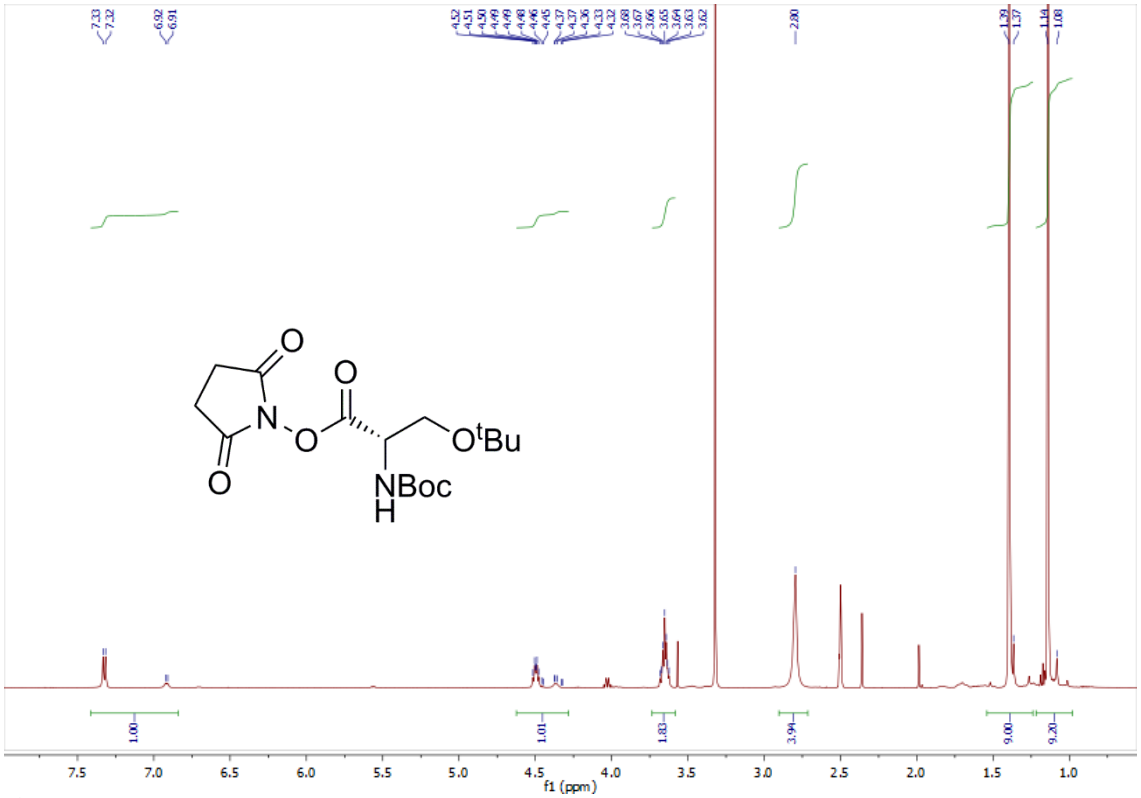


^{13}C (126 MHz, DMSO-d_6):

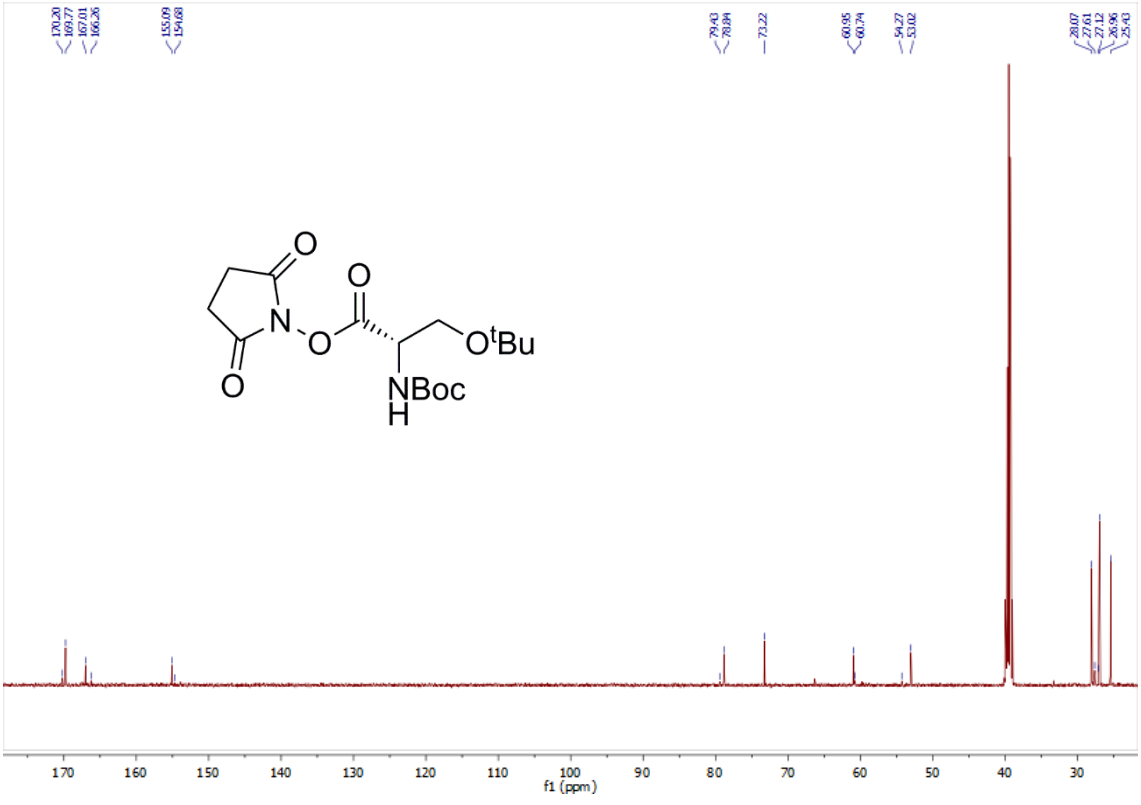


_N_-hydroxysuccinimidyl (2_S_)-3-(_tert_-butoxy)-2-(_tert_-butoxycarbonyl-amino)-propanoate 103

^{1}H (500 MHz, DMSO-d_6):

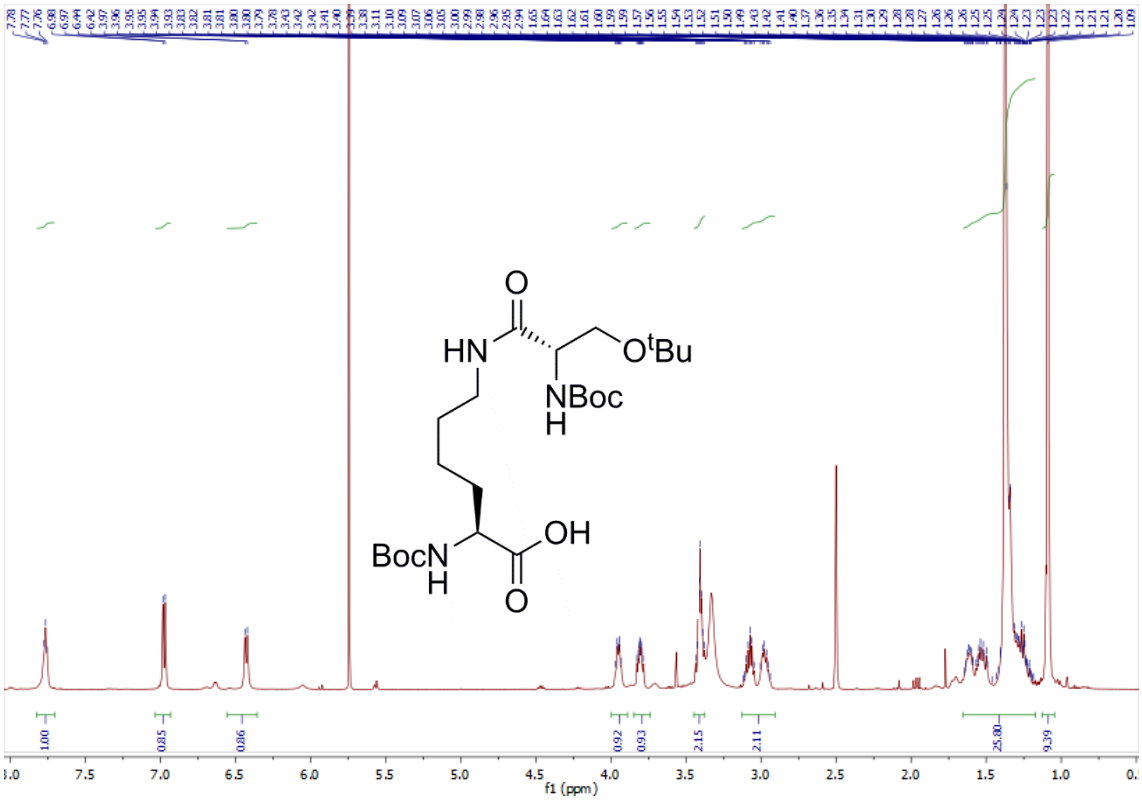


^{13}C (126 MHz, DMSO-d_6):

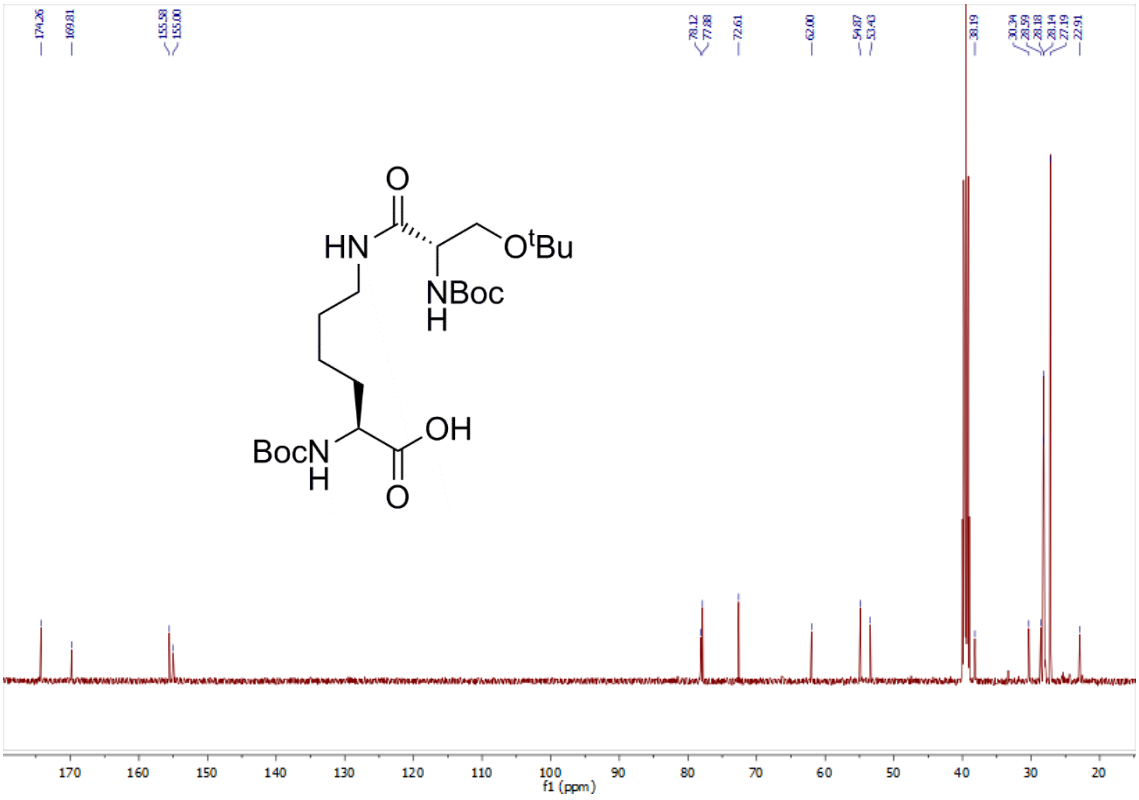


(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid 104

^{1}H (500 MHz, DMSO-d_6):

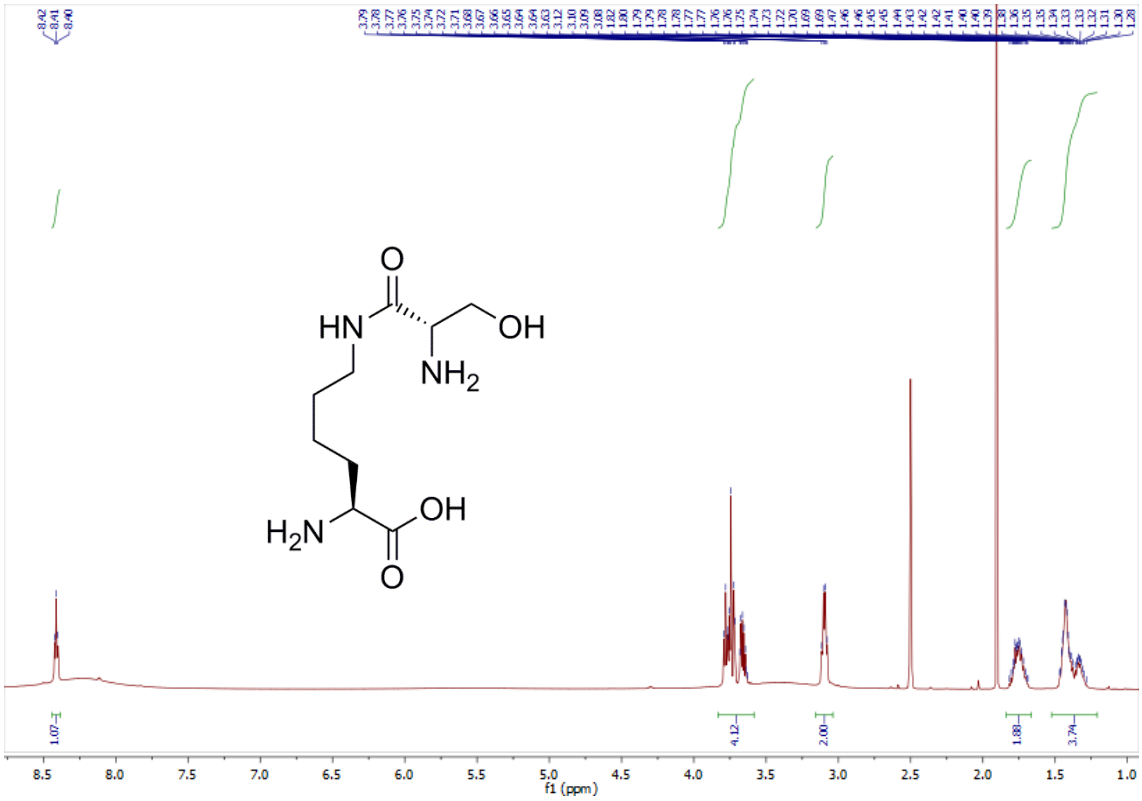


^{13}C (126 MHz, DMSO-d_6):

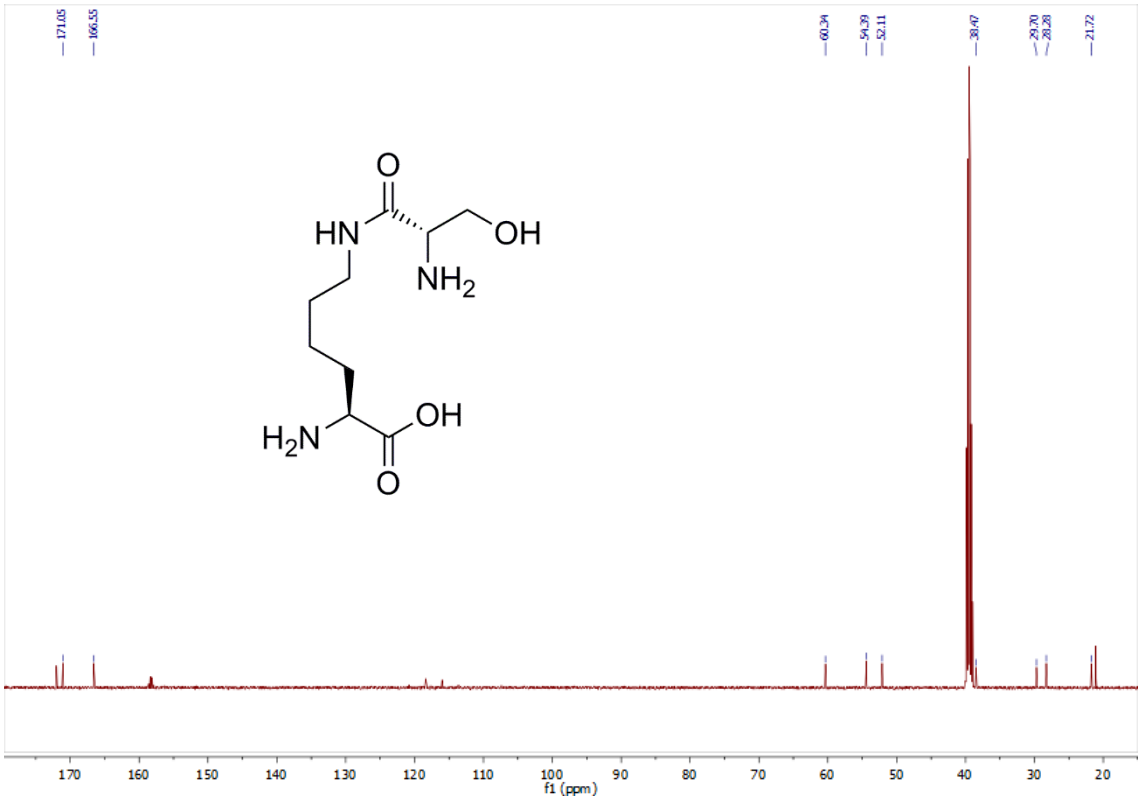


(2*S*)-2-amino-6-((2*S*)-2-amino-3-hydroxypropanamido)-hexanoic acid 101

^{1}H (500 MHz, DMSO-d_6):

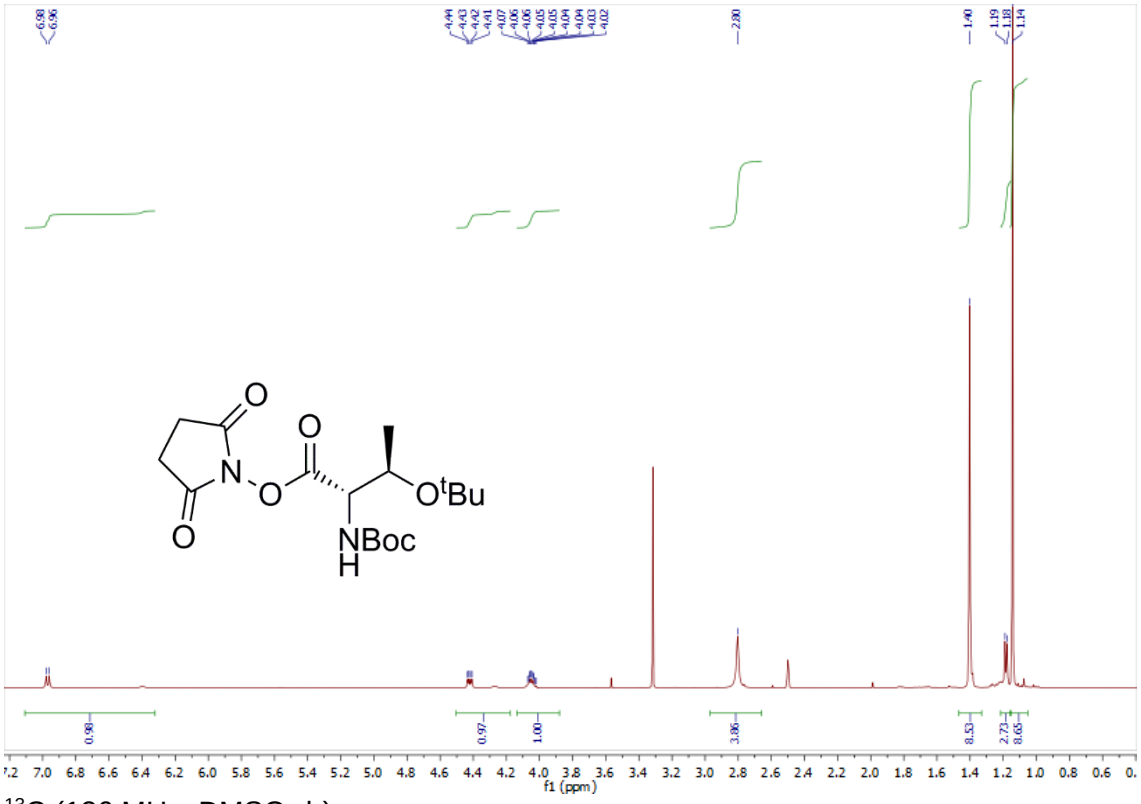


^{13}C (126 MHz, DMSO-d_6):

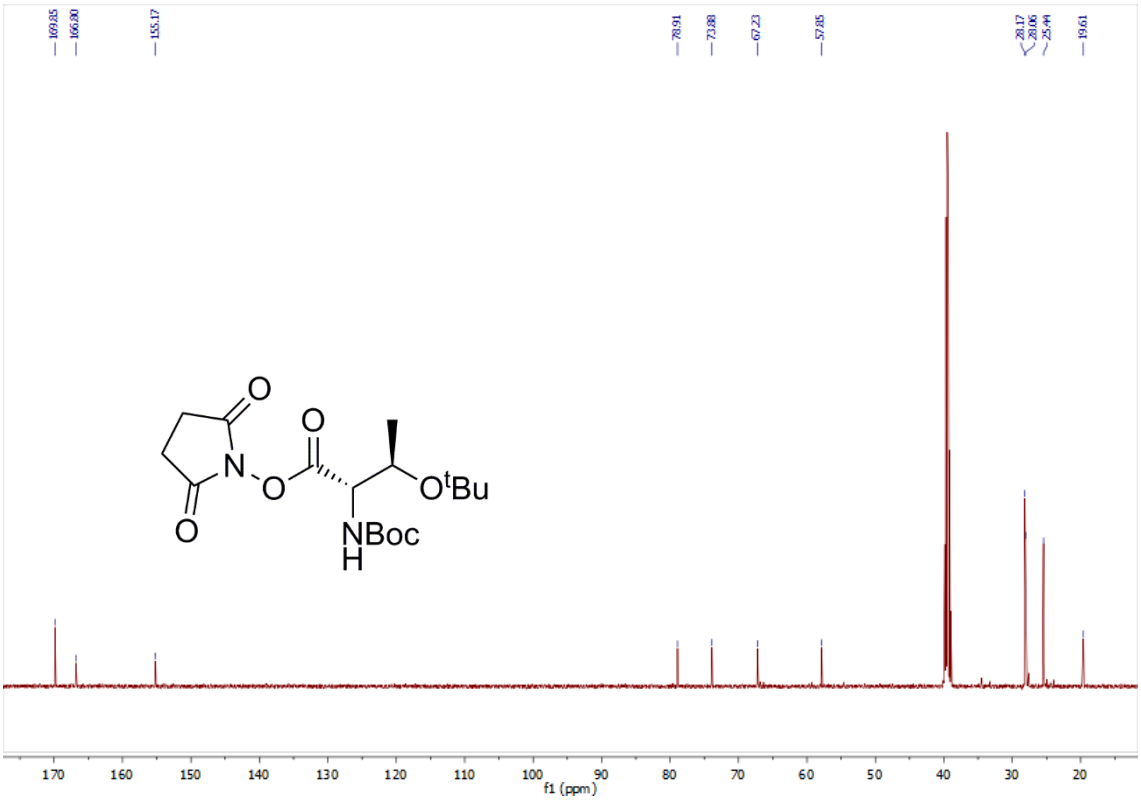


_N_-hydroxysuccinimidyl (2_S_,3_R_)-3-(_tert_-butoxy)-2-(_tert_-butoxycarbonyl-amino)-butanoate 107

^{1}H (500 MHz, DMSO-d_6):

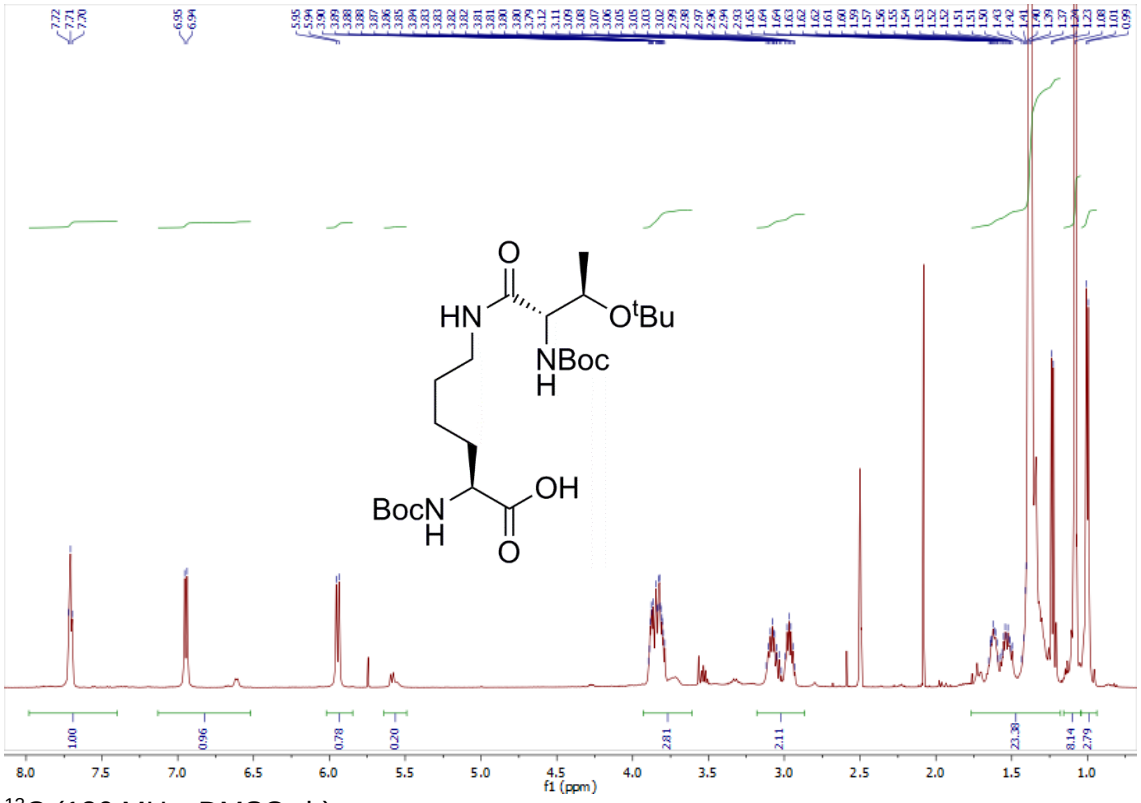


^{13}C (126 MHz, DMSO-d_6):

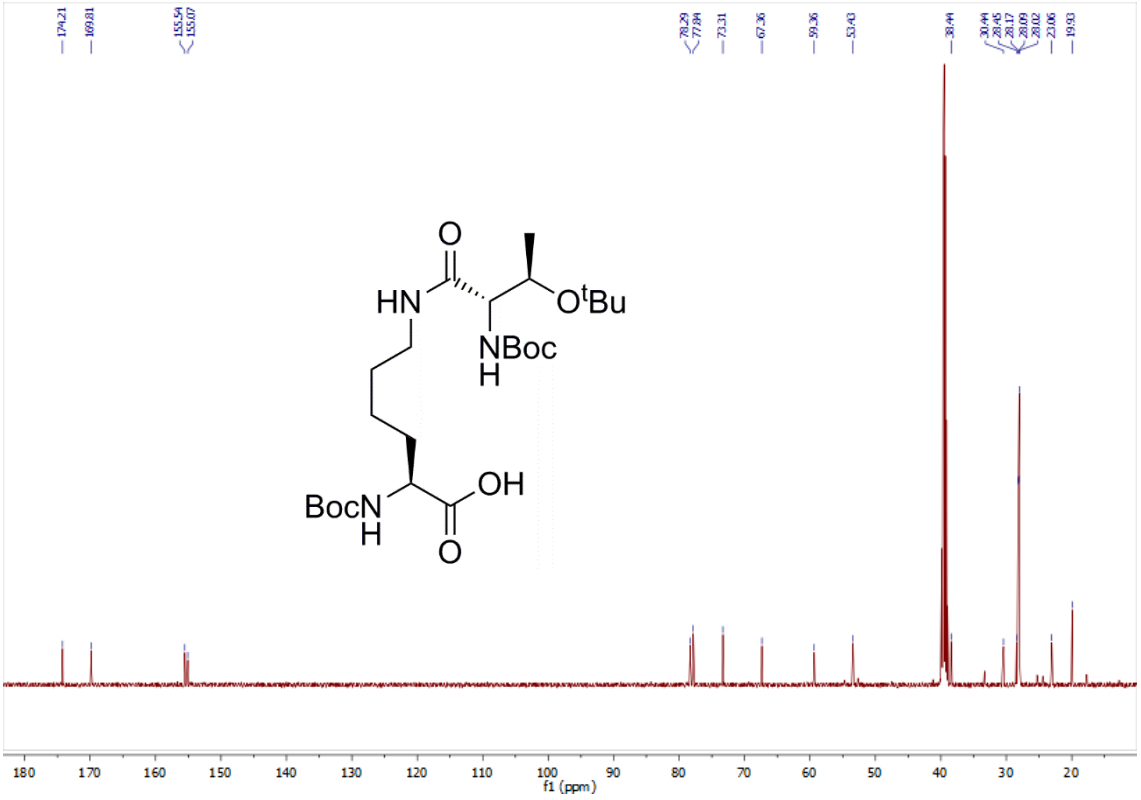


(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*S*,3*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid 108

^{1}H (500 MHz, DMSO-d_6):

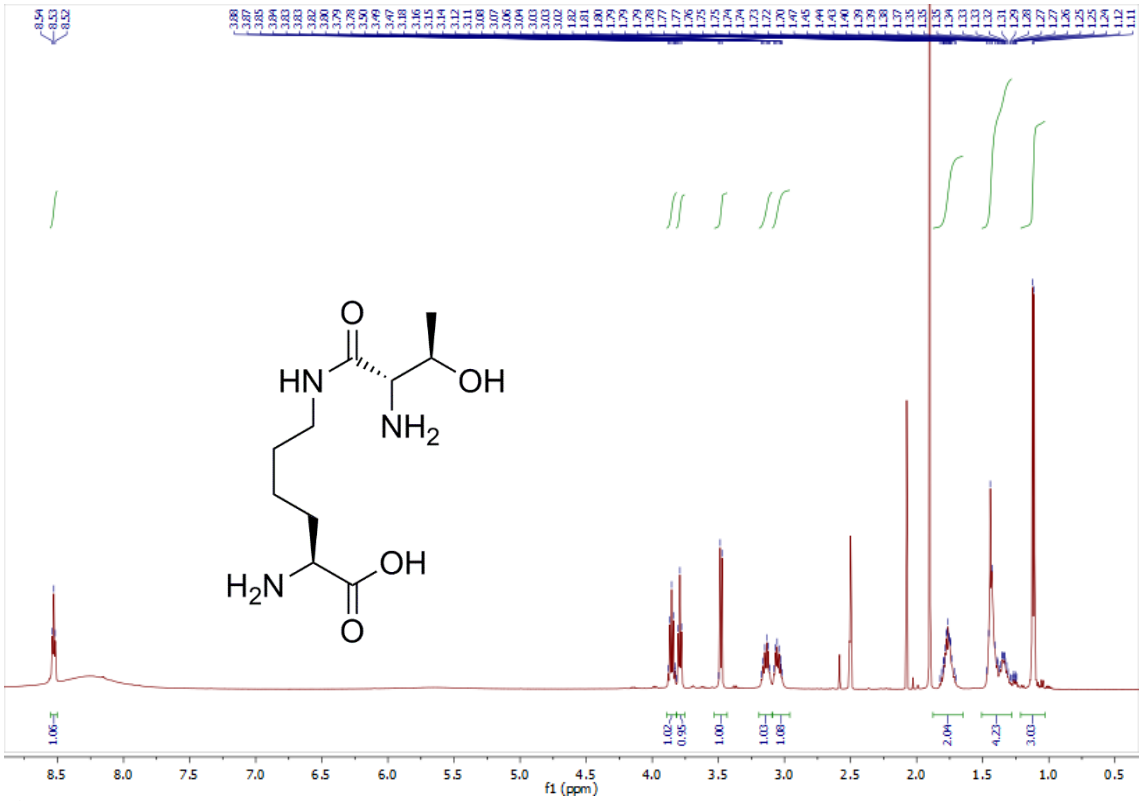


^{13}C (126 MHz, DMSO-d_6):

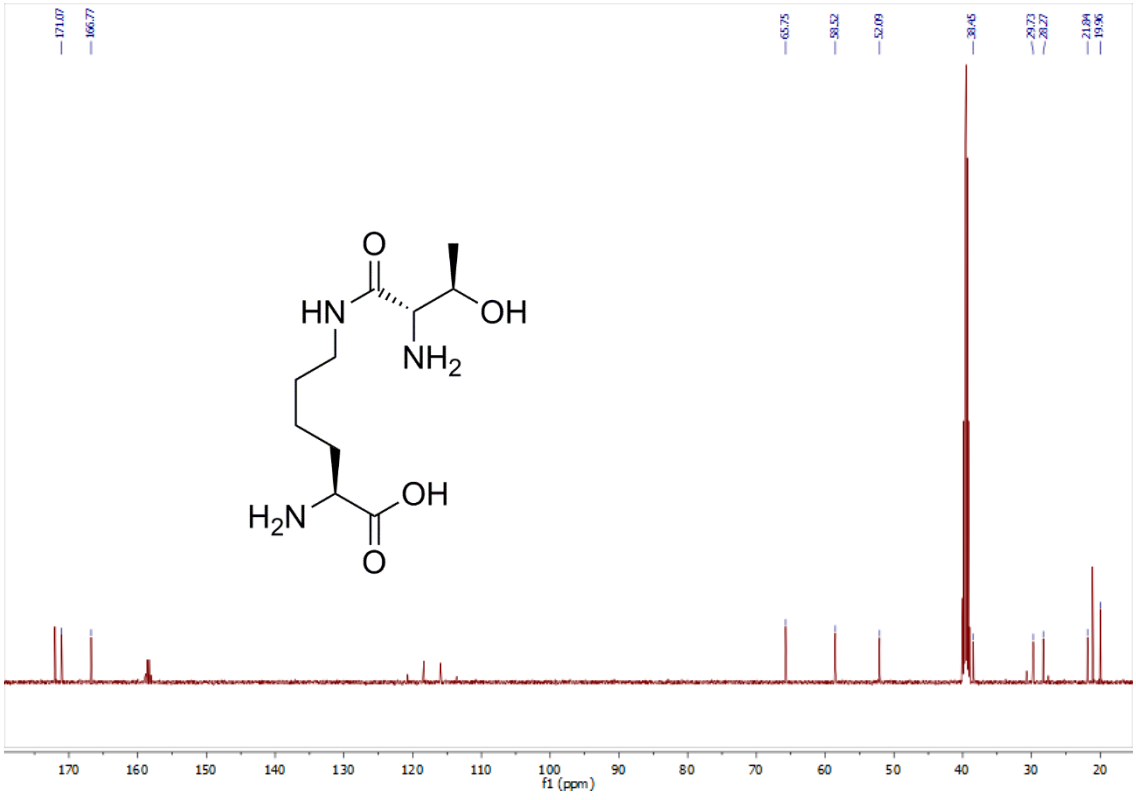


(2*S*)-2-amino-6-((2*S*,3*R*)-2-amino-3-hydroxybutanamido)-hexanoic acid 105

^{1}H (500 MHz, DMSO-d$_6$):

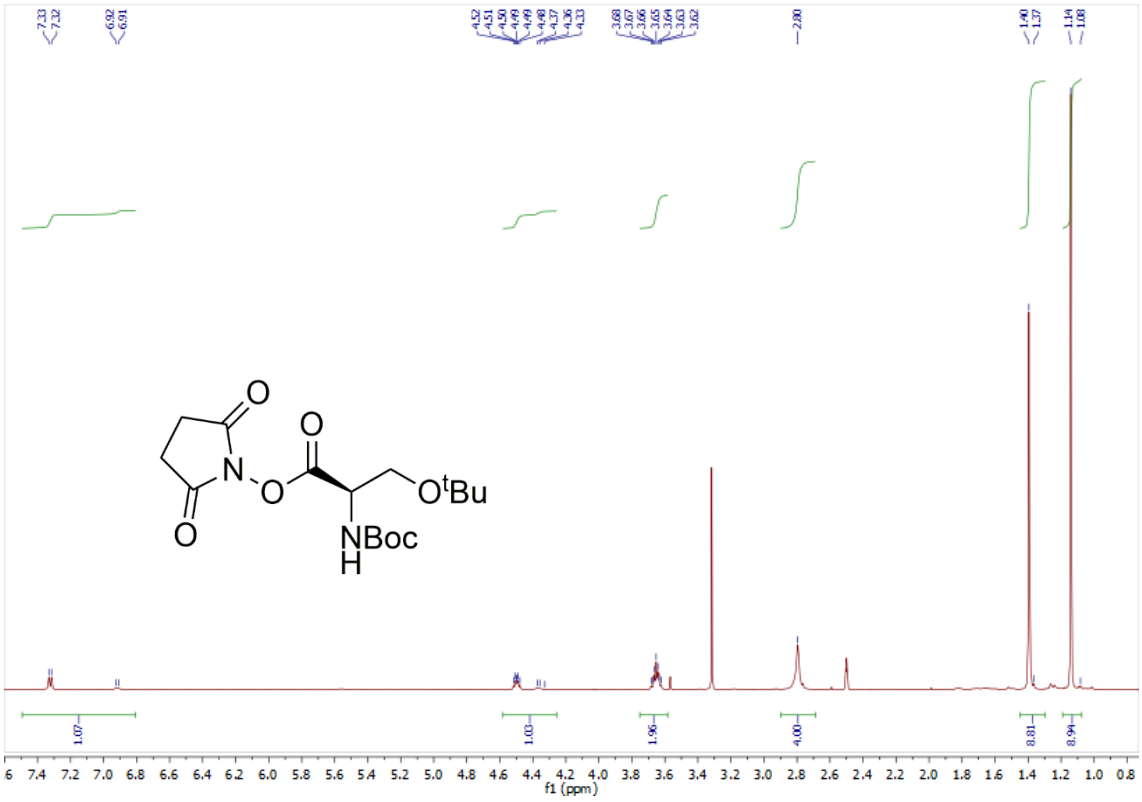


^{13}C (126 MHz, DMSO-d$_6$):

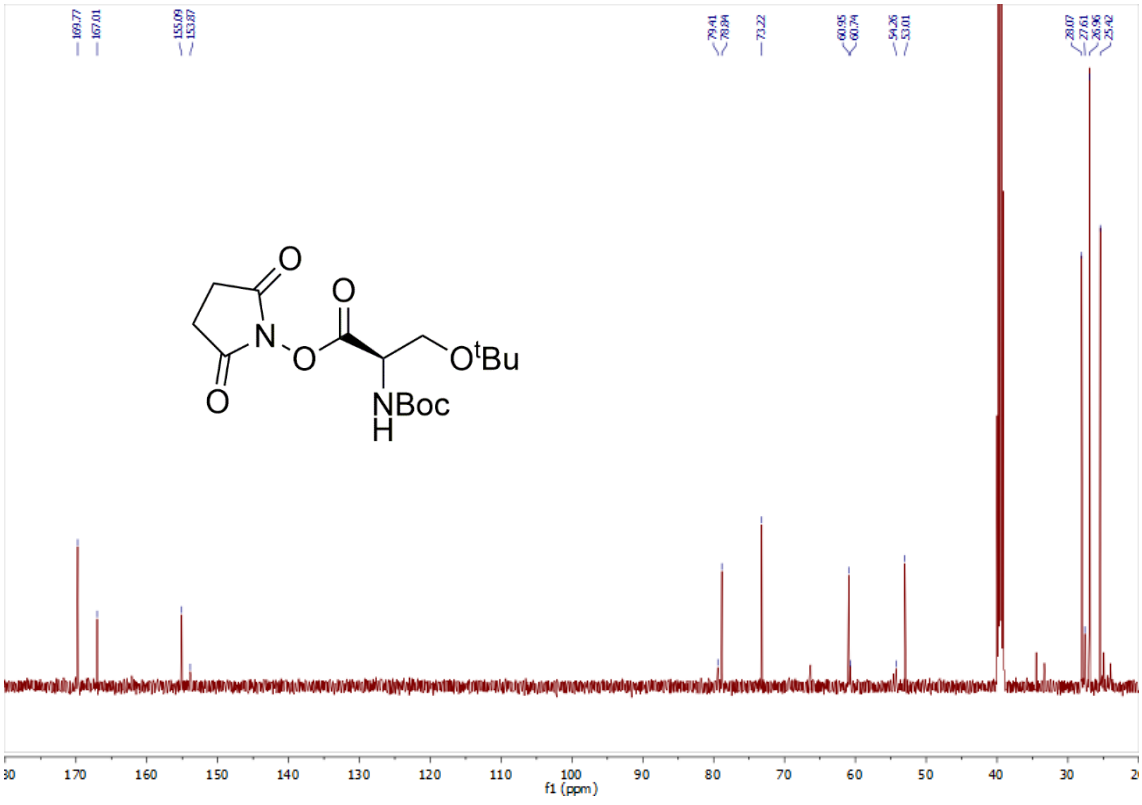


_N_-hydroxysuccinimidyl (2_R_)-3-(_tert_-butoxy)-2-(_tert_-butoxycarbonylamino)-propanoate 112

^{1}H (500 MHz, DMSO-d_6):

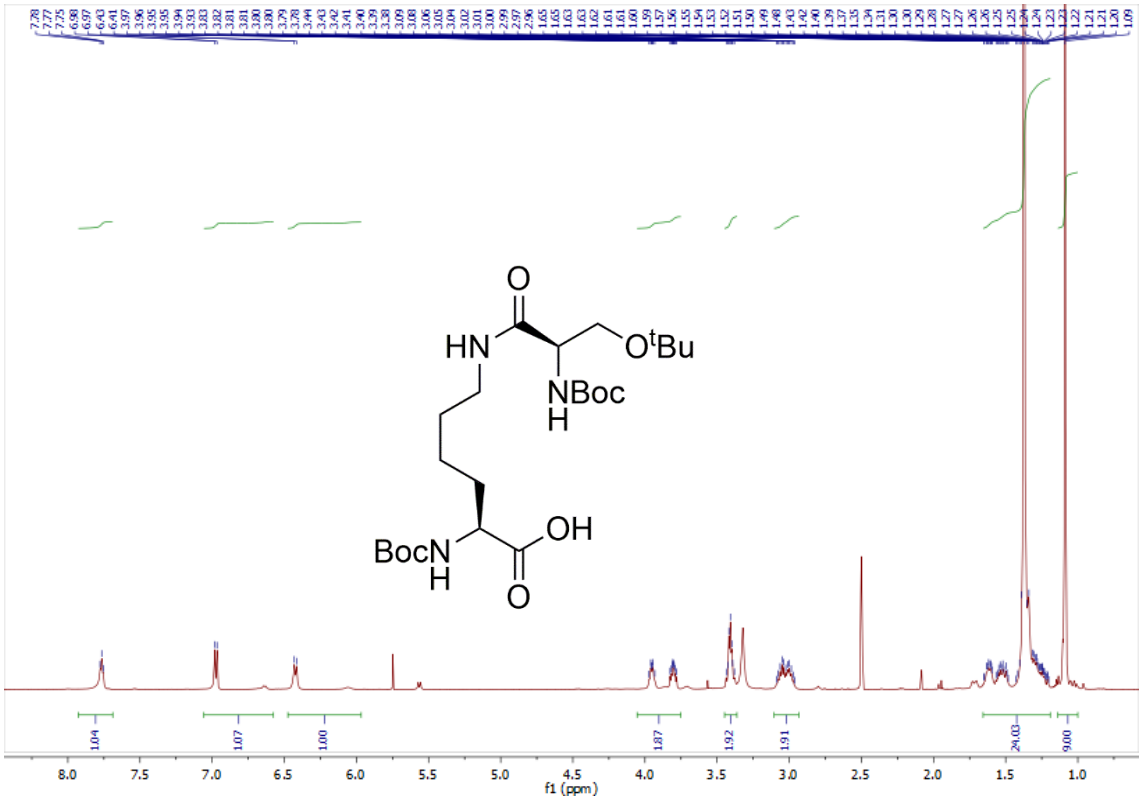


^{13}C (126 MHz, DMSO-d_6):

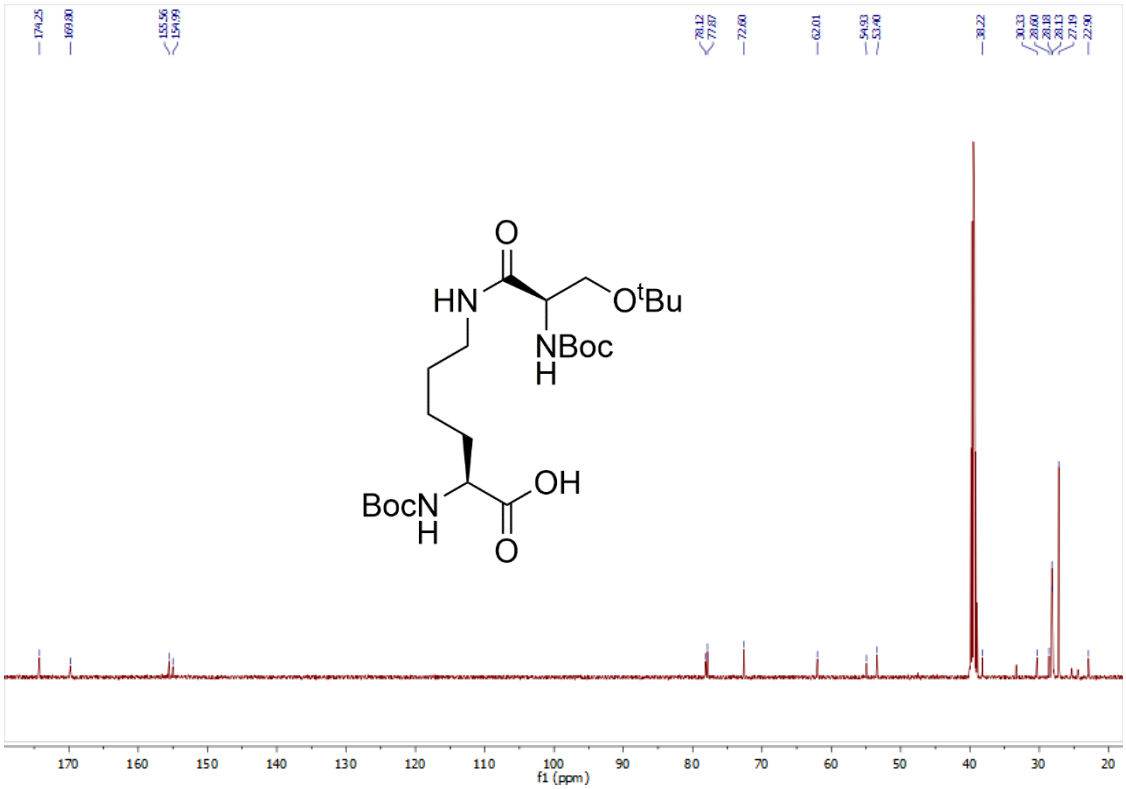


(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-propanamido)-hexanoic acid 113

^{1}H (500 MHz, DMSO-d_6):

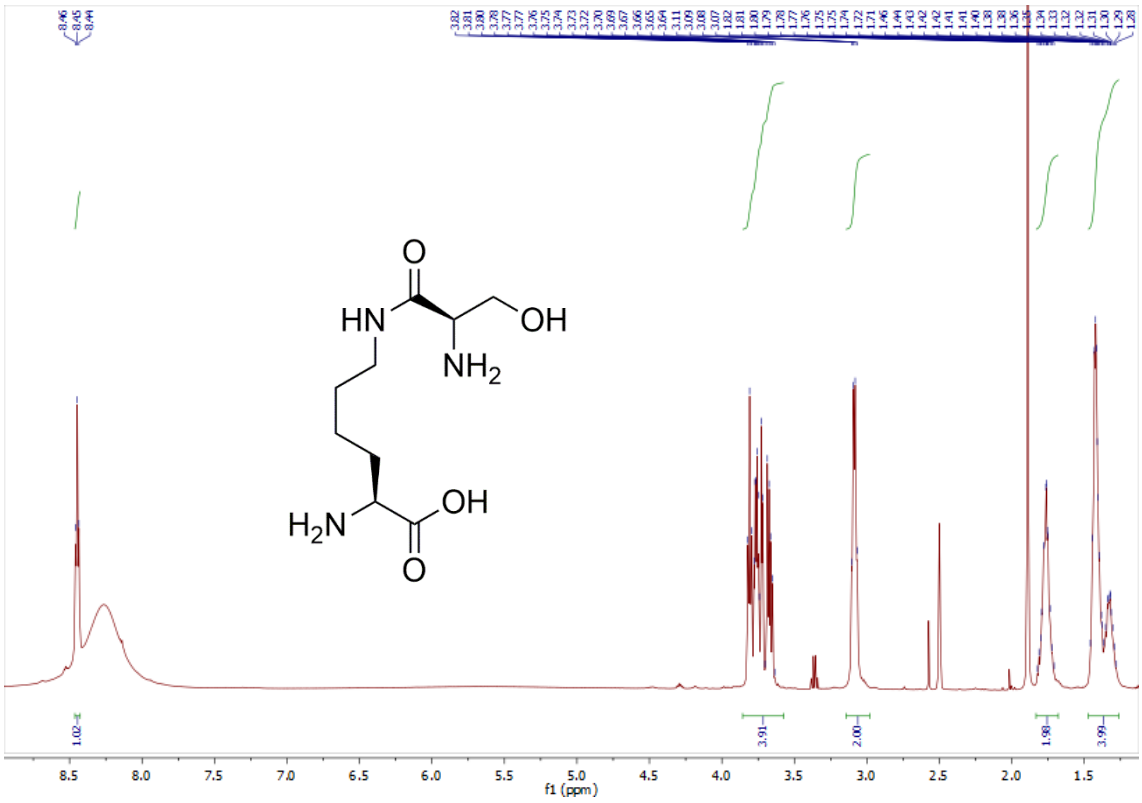


^{13}C (126 MHz, DMSO-d_6):

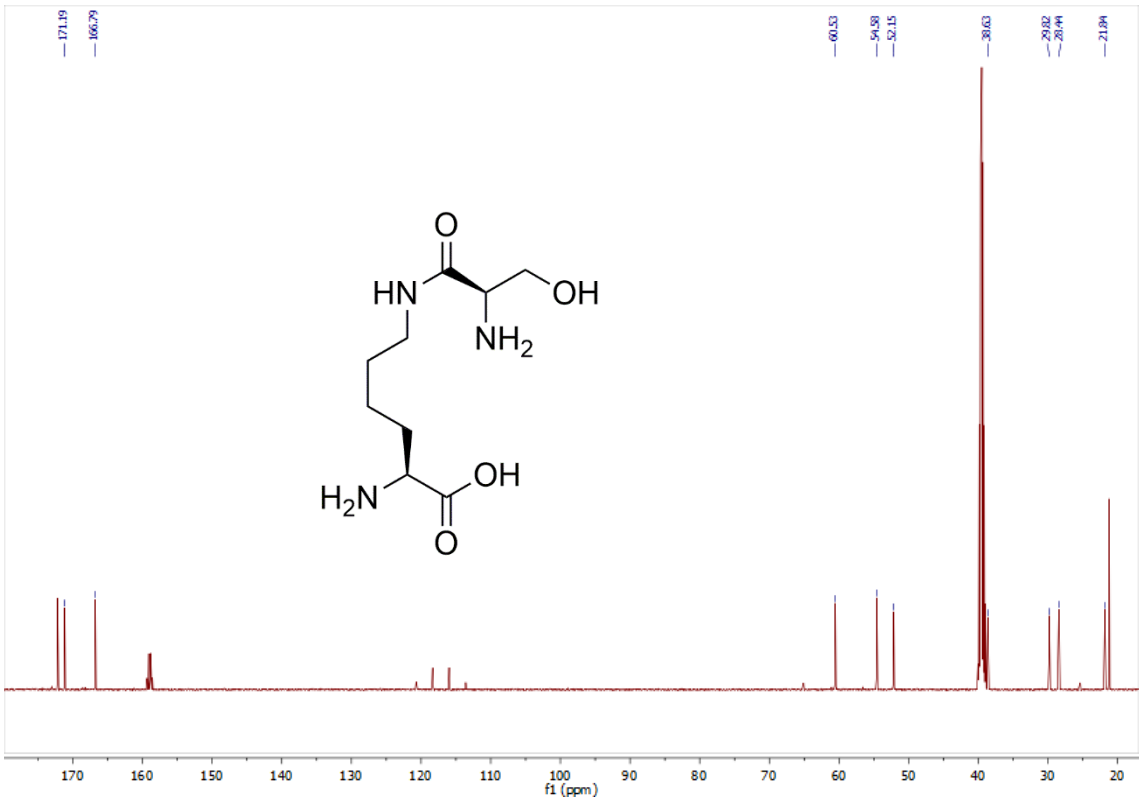


(2*S*)-2-amino-6-((2*R*)-2-amino-3-hydroxypropanamido)-hexanoic acid 109

^{1}H (500 MHz, DMSO-d_6):

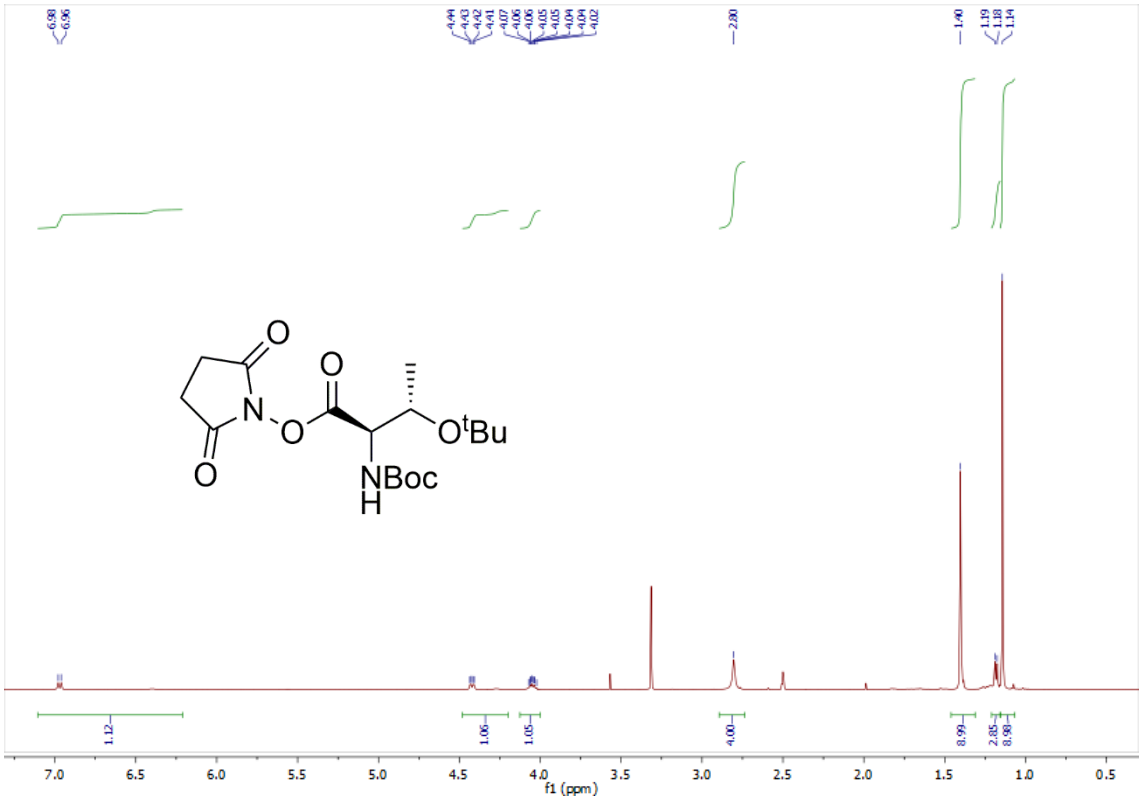


^{13}C (126 MHz, DMSO-d_6):

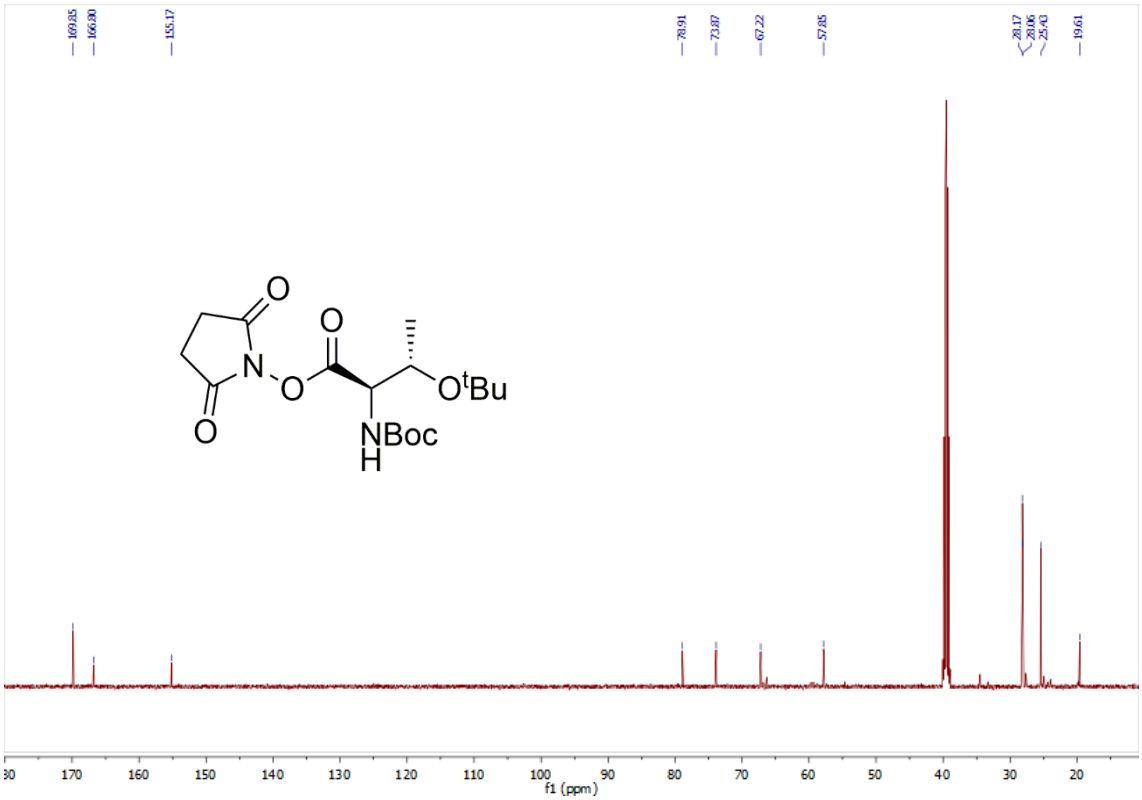


_N_-hydroxysuccinimidyl (2_R_,3_S_)-3-(_tert_-butoxy)-2-(_tert_-butoxycarbonyl-amino)-butanoate 115

^{1}H (500 MHz, DMSO-d_6):

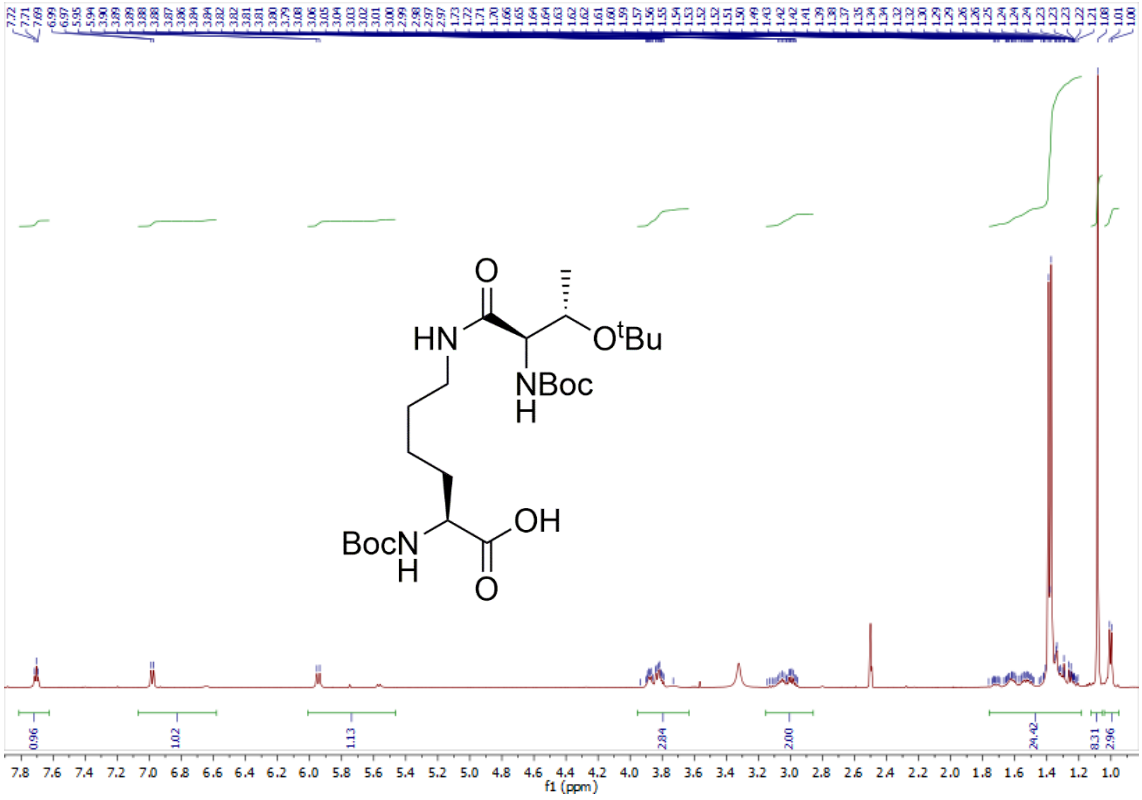


^{13}C (126 MHz, DMSO-d_6):

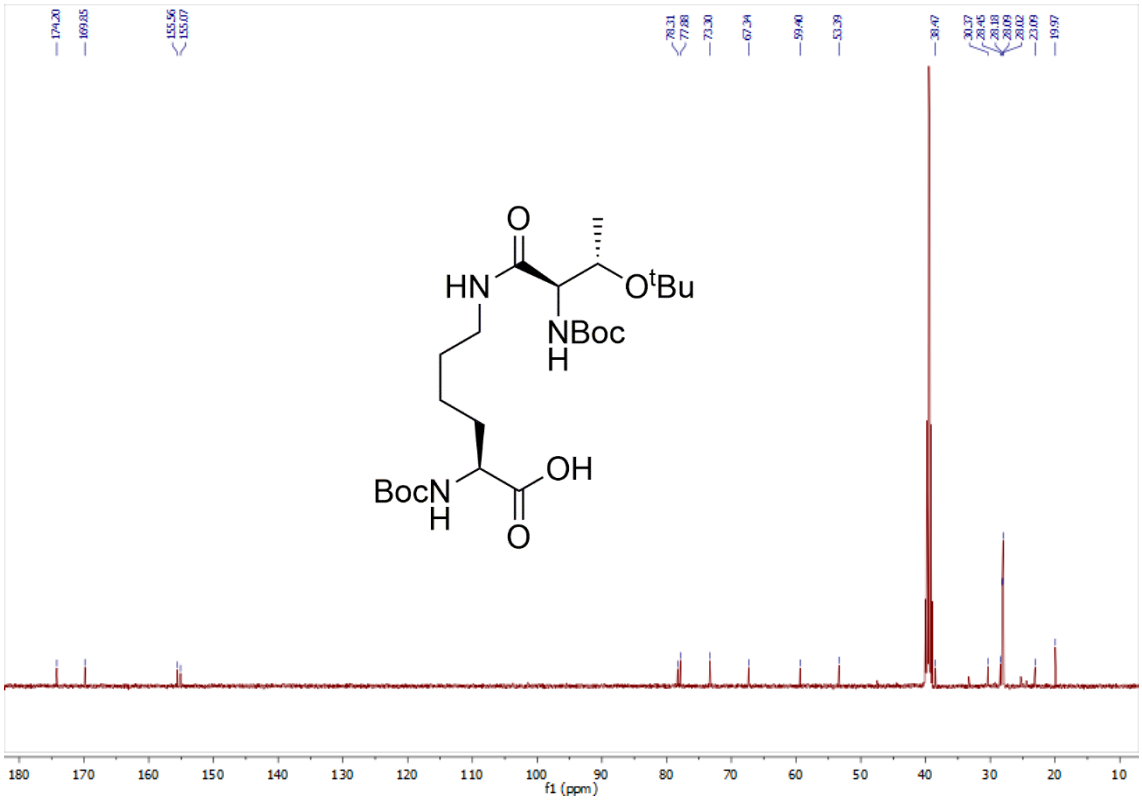


(2*S*)-2-(*tert*-butoxycarbonylamino)-6-((2*R*,3*S*)-3-(*tert*-butoxy)-2-(*tert*-butoxycarbonylamino)-butanamido)-hexanoic acid 116

^{1}H (500 MHz, DMSO-d_6):

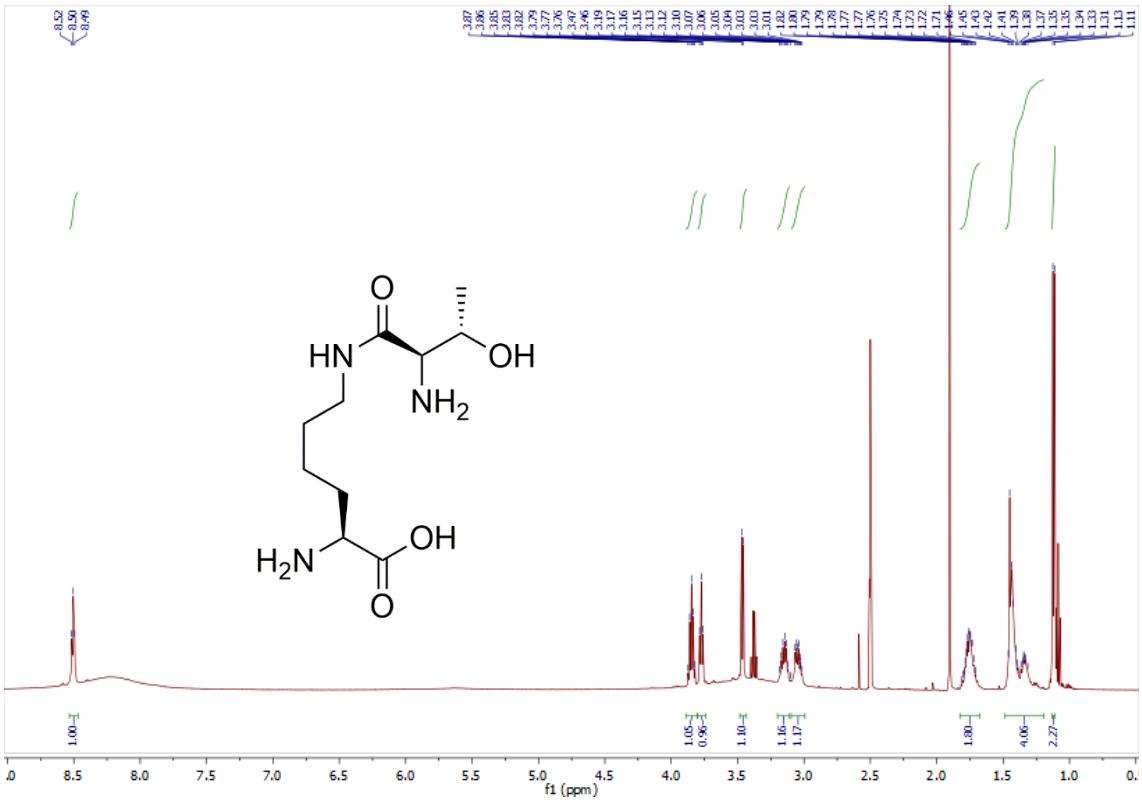


^{13}C (126 MHz, DMSO-d_6):

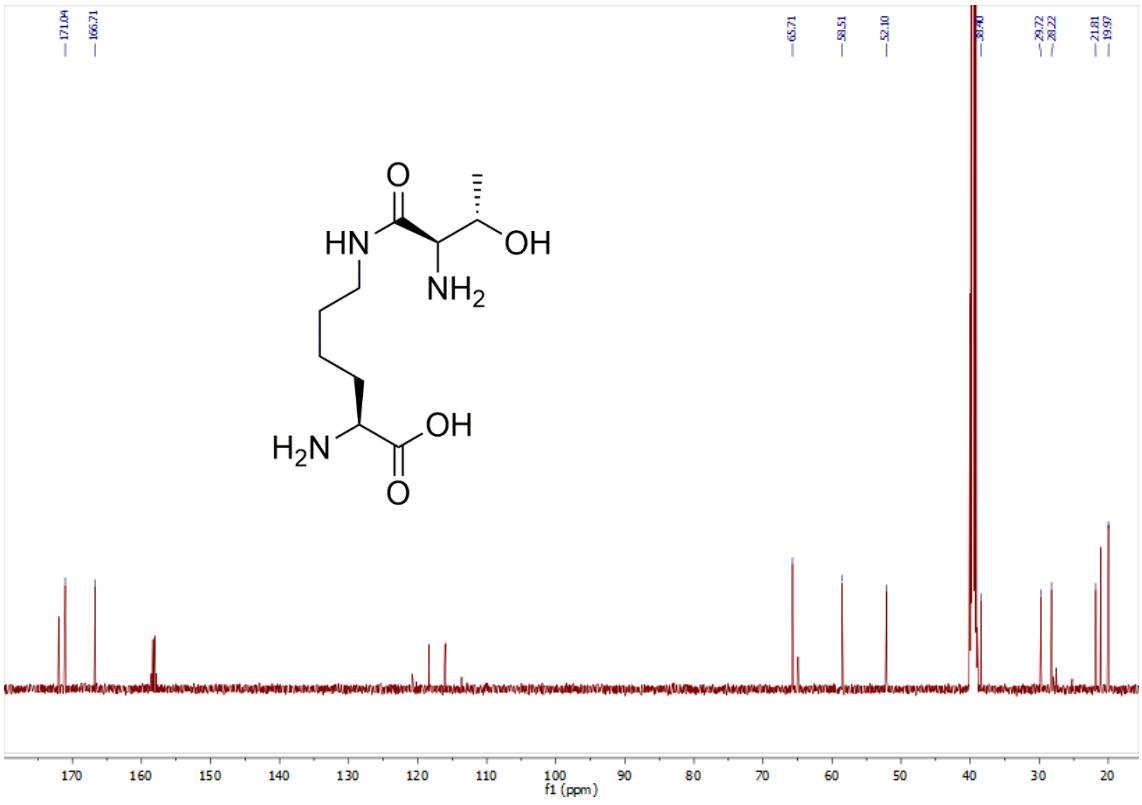


(2*S*)-2-amino-6-((2*S*,3*R*)-2-amino-3-hydroxybutanamido)-hexanoic acid 110

^{1}H (500 MHz, DMSO-d_6):

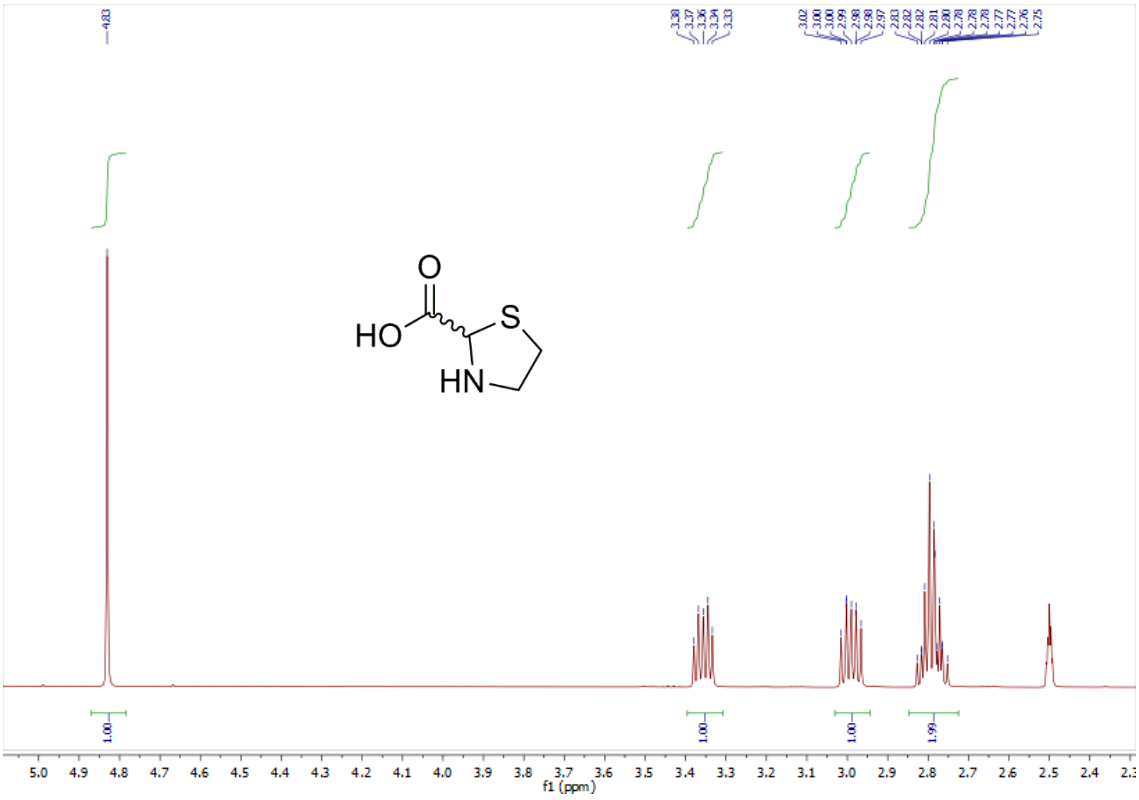


^{13}C (126 MHz, DMSO-d_6):

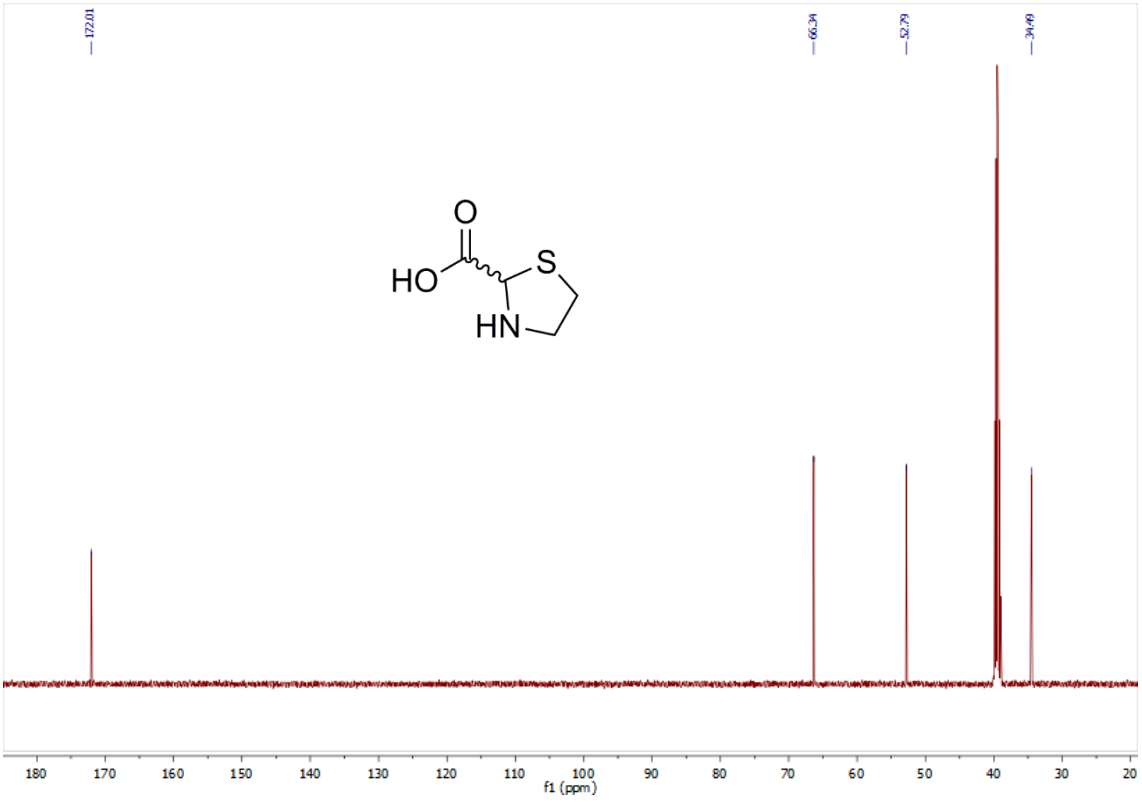


Thiazolidine 2-carboxylic acid (racemate) 119

^{1}H NMR (500 MHz, DMSO-d_6):

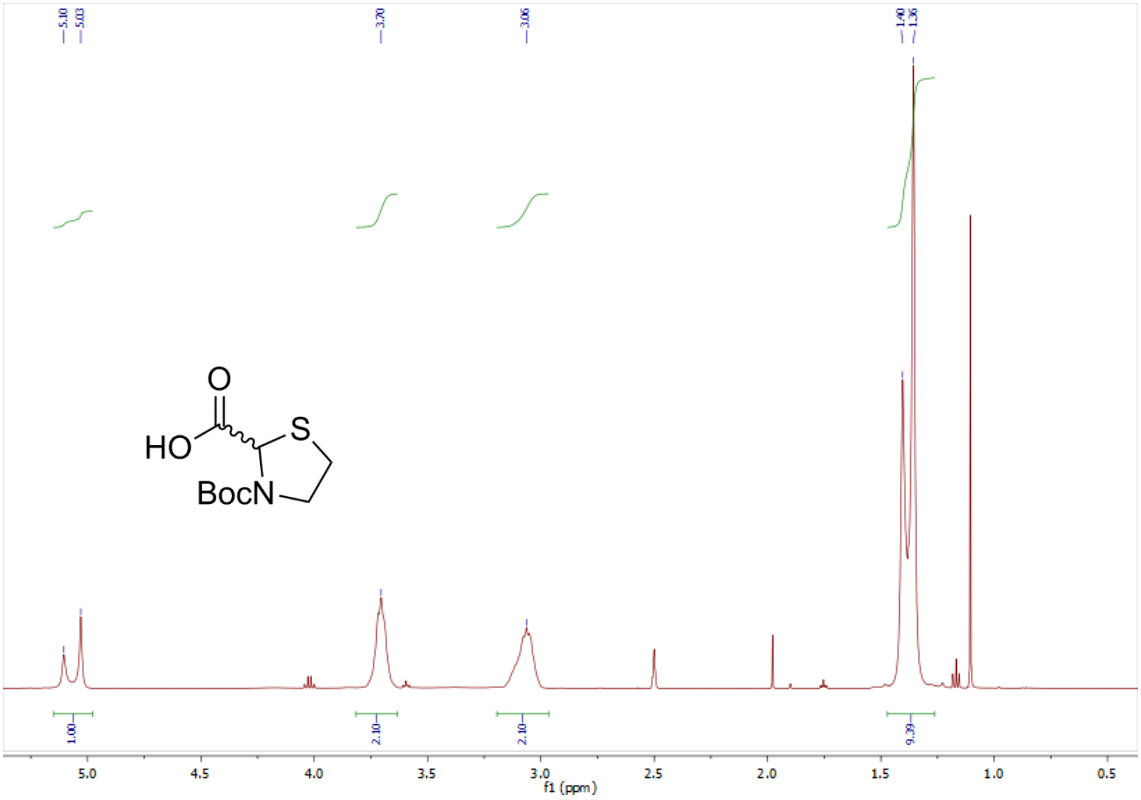


^{13}C (126 MHz, DMSO-d_6):

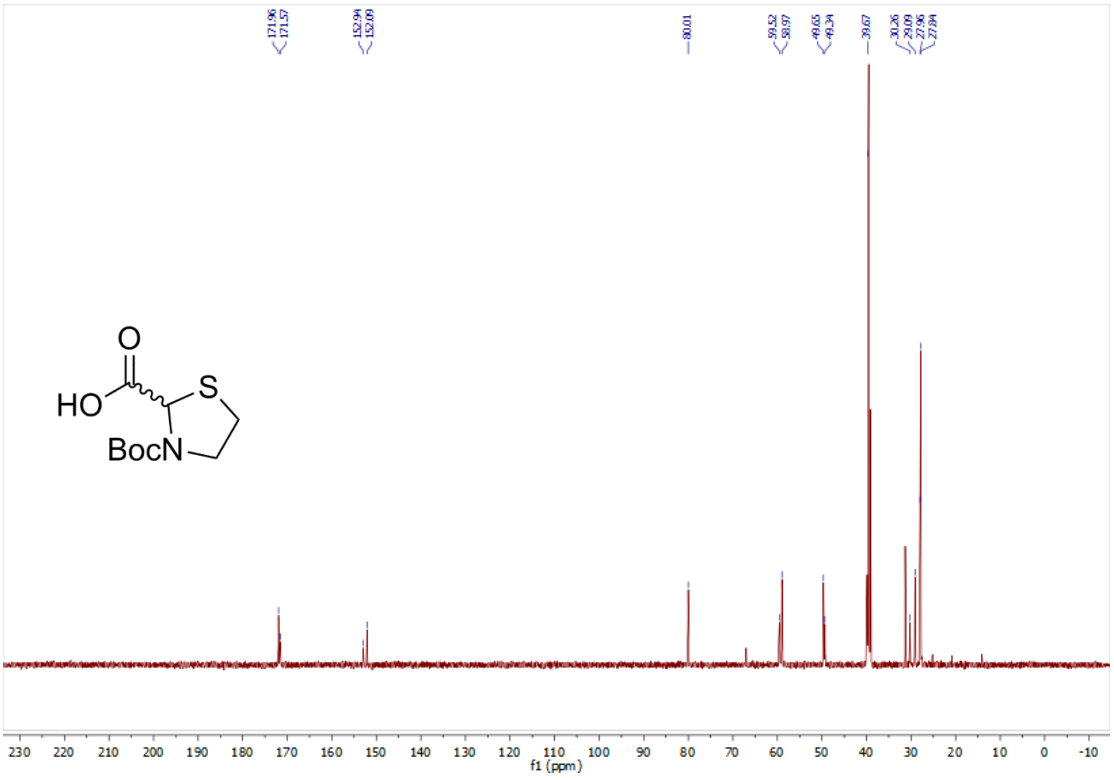


3-(*tert*-butoxycarbonyl)thiazolidine-2-carboxylic acid (racemate) 120

^{1}H (500 MHz, DMSO-d_6):



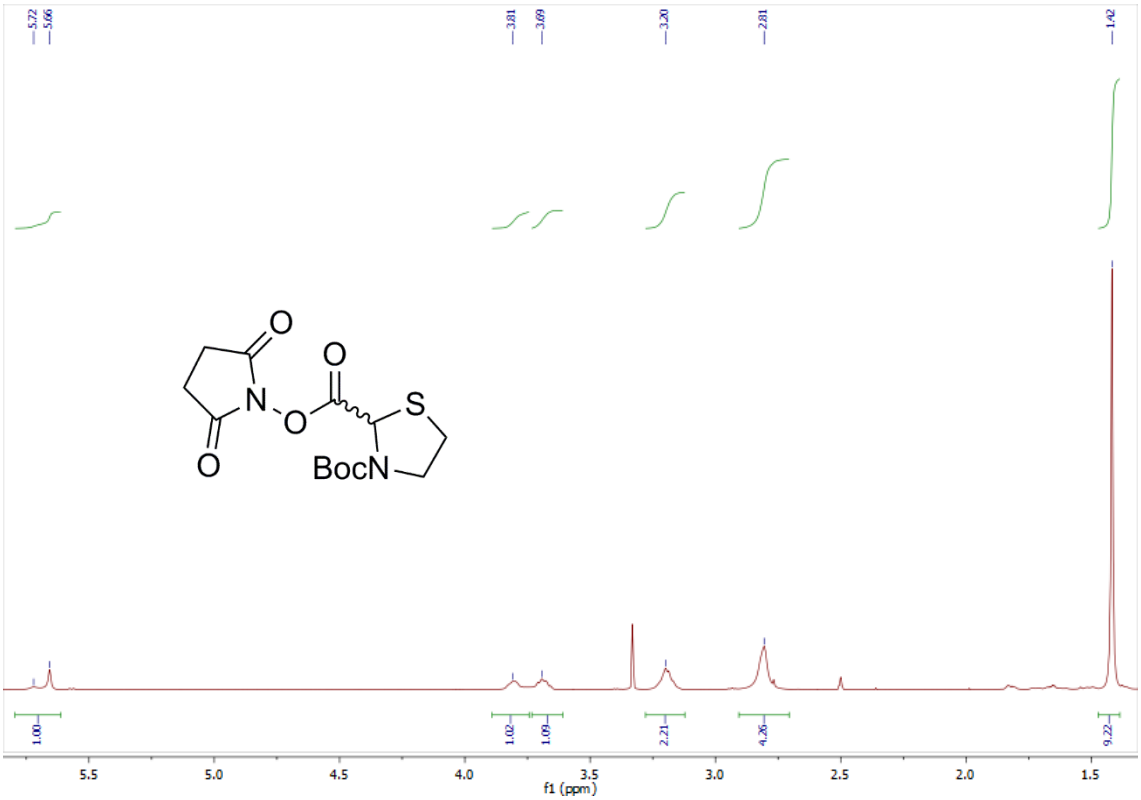
^{13}C (126 MHz, DMSO-d_6):



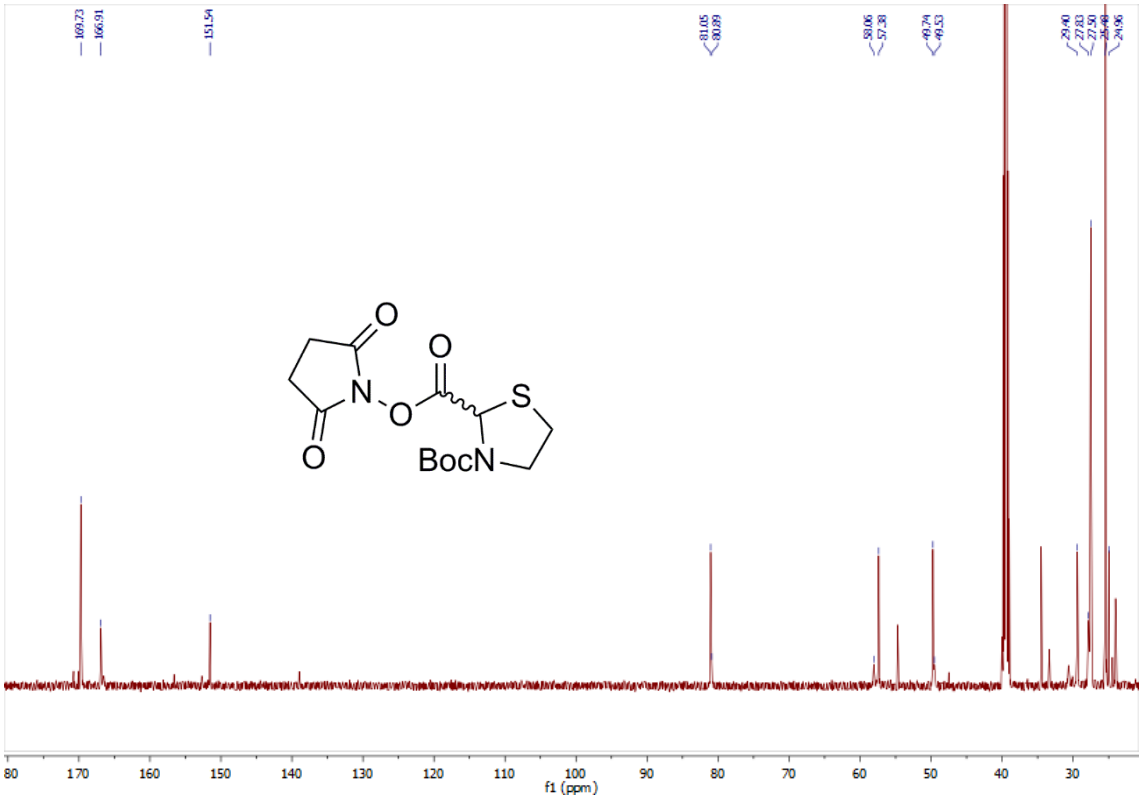
_N_-hydroxysuccinimidyl 3-(_tert_-butoxycarbonyl)thiazolidine-2-carboxylate

121

^{1}H (500 MHz, DMSO-d_6):

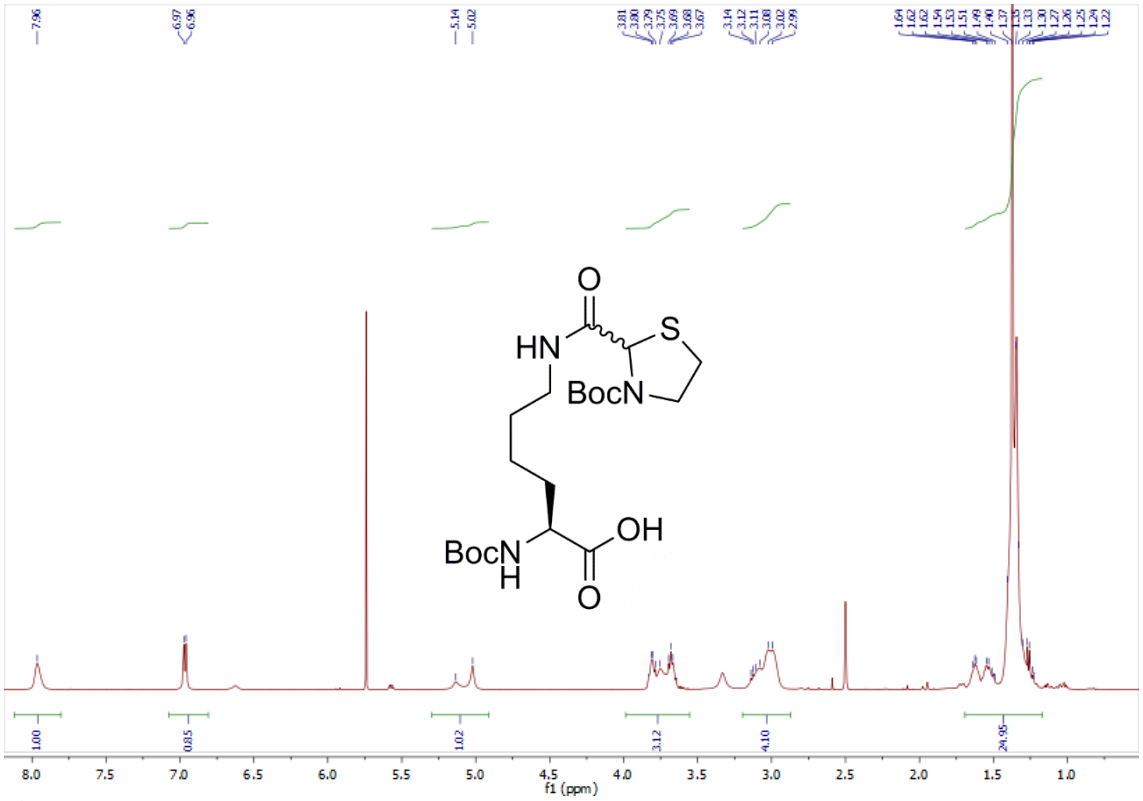


^{13}C (126 MHz, DMSO-d_6):

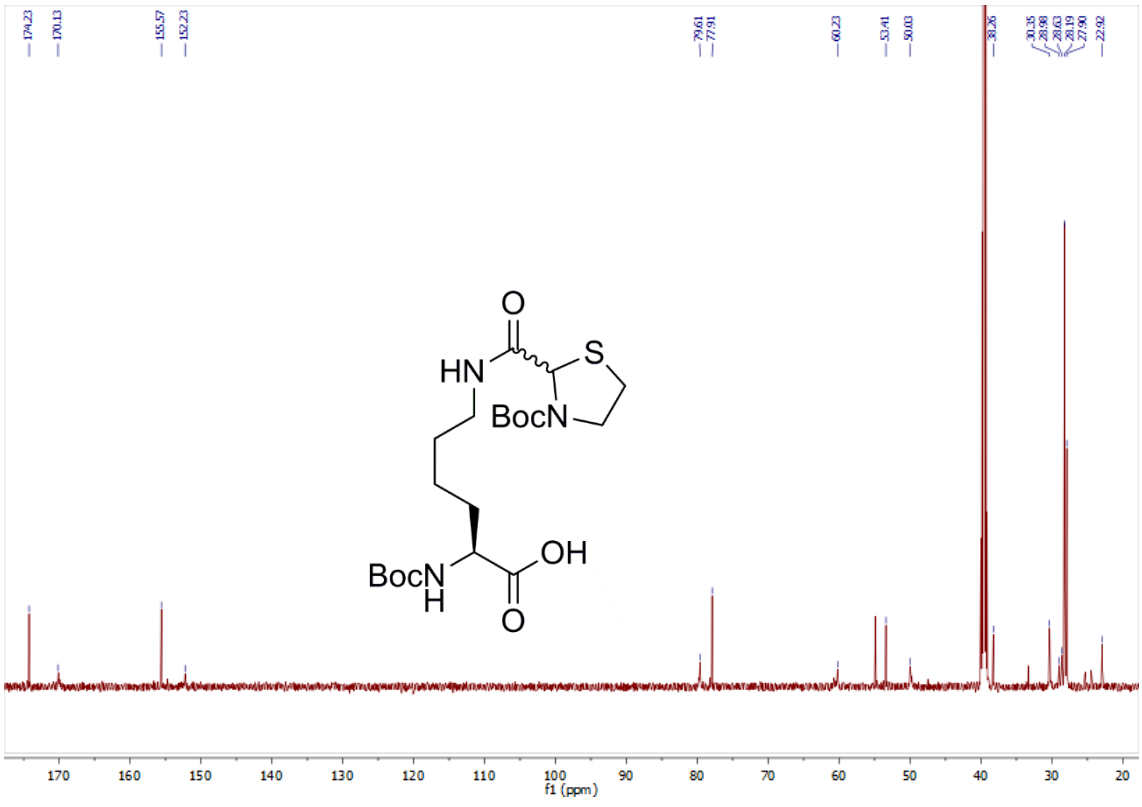


(2*S*)-2-(*tert*-butoxycarbonylamino)-6-(thiazolidine-2-carboxamido-3-(*tert*-butoxycarbonyl))-hexanoic acid 122

^{1}H (500 MHz, DMSO-d_6):

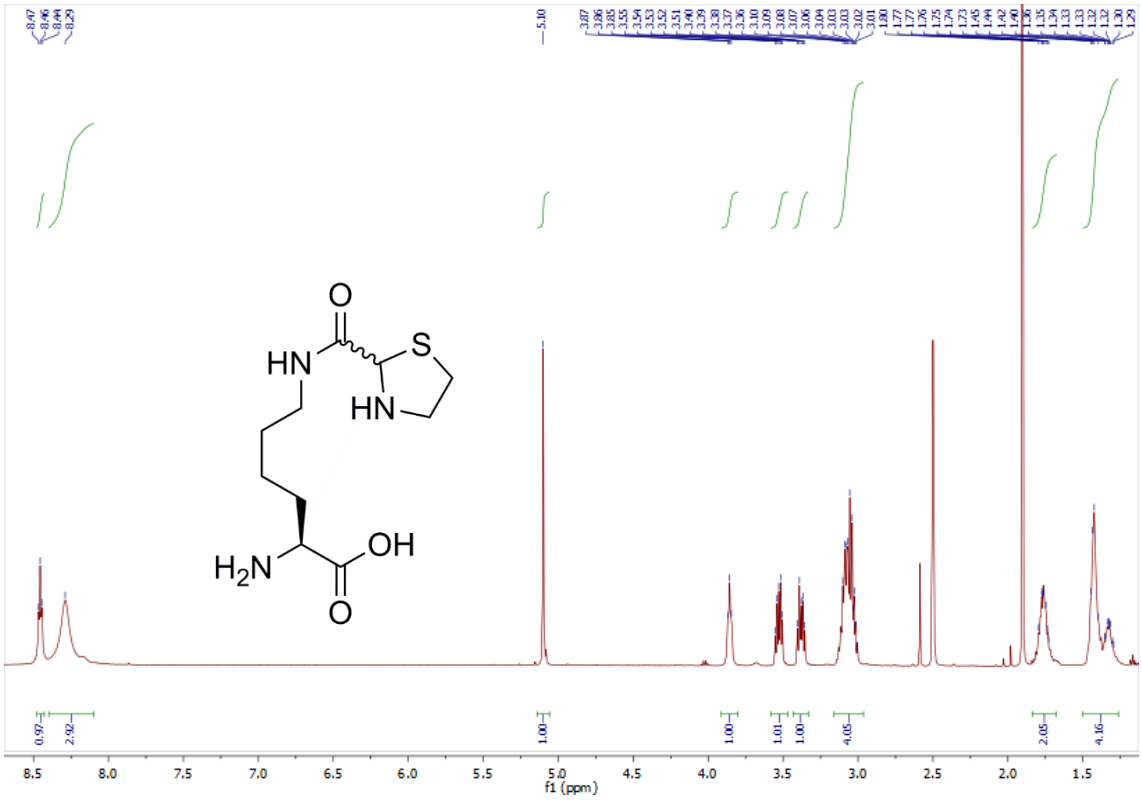


^{13}C (126 MHz, DMSO-d_6):

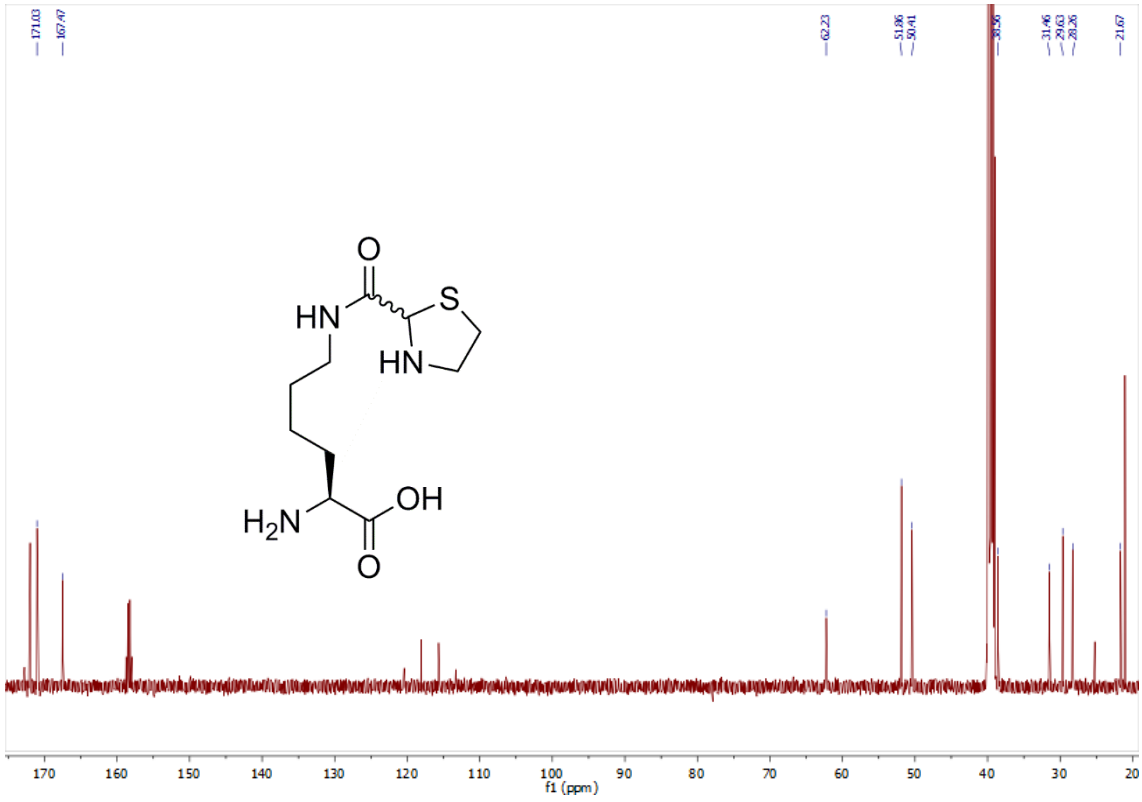


(2*S*)-2-amino-6-(thiazolidine-2-carboxamido) hexanoic acid 86

^{1}H (500 MHz, DMSO-d_6):



^{13}C (126 MHz, DMSO-d_6):



7.2 Molecular biology and protein expression

7.2.1 Plasmid information

The vectors pEVOL, harbouring pylT and pylRS (M. mazei, wild type or Y306A Y384F double mutant) genes, and pBAD, harbouring the EGFP(Y39TAG) gene, were provided by Edward Lemke following an MTA and have been previously described. The pBAD vector harbouring the sfGFP(N150TAG) gene was a gift from Ryan Mehl (Addgene plasmid #85483). The pET22b plasmid harbouring the BiGalK gene was supplied by Dr Tessa Keenan.

7.2.2 Site-directed mutagenesis

Mutation of the *N*-terminal residue of GFP(Y39TAG) to serine (from aspartate as part of a FLAG tag originally encoded in pBAD-GFP(Y39TAG)) was performed using an Agilent Quikchange Multi Site-Directed Mutagenesis Kit according to the provided protocol using the following primers (Sigma):

5' - CACTTTATCATCATCATCTTTGTAAGACATGGTTAATTCCTCCTGTTAGCCC - 3'
5' - GGGCTAACAGGAGGAATTAACCATGTCTTACAAAGATGATGATGATAAAGTG - 3'

Mutation of BiGalK-His_6 K417 to the amber stop codon was performed using an Agilent Quikchange Multi Site-Directed Mutagenesis Kit according to the provided protocol using the following primers (Sigma):

5' - GCCGCGAGGCCTAGCTTGCGGCC - 3'
5' - GGCCGCAAGCTAGGCCTCGCGGC - 3'

7.2.3 tRNA synthetase substrate recognition assay

The pBAD vector containing ampicillin resistance and Ser-EGFP(Y39TAG) genes, together with the pEVOL vector containing $tRNA^{Pyl}$, pylRS (*M. mazei*, either wild type or Y306A Y384F double mutant) and chloramphenicol resistance genes, were co-transformed into electrocompetent *E. coli* Top10 cells and selected on LB agar plates containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml). Two starter cultures were prepared by picking a single colony from one of the two transformants (pBAD-EGFP(Y39TAG)-His_6 and either pEVOL-pylT-$pylRS^{wt}$ or pEVOL-pylT-$pylRS^{AF}$, subsequently referred to as $pylRS^{wt}$ and $pylRS^{AF}$ cultures).

100 mL Terrific Broth Medium containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml) in a 250 mL baffled conical flask was inoculated with 1.0 mL of either the $pylRS^{wt}$

or pylRS^{AF} overnight culture. The preparative cultures were grown at 37 °C with shaking (220 rpm). Typically within 3 h, the cultures reached an OD_{600} of 0.2-0.3. At this point, 10 mL from each preparative culture was aliquoted into a 50 mL Falcon tube, to which a solution of test NCAA (**86**, **96**, **101**, **105**, **109** or **110**) or positive control (**29** for pylRS^{wt} and **34** for pylRS^{AF}) was added (stock concentration 50 mM in 0.1 M NaOH, 400 µL, final concentration 2 mM). For negative controls, a solution of 0.1 M NaOH (400 µL) was added. Each aliquot was left at 37 °C with shaking (220 rpm) for 30 min, at which point L-arabinose (stock 20% (w/v), 10 µL, final concentration 0.02% (w/v)) was added to induce gene expression and cultures left to grow for 16-18 h at 37 °C with shaking (220 rpm).

1.5 mL of each aliquot culture was taken and clarified by centrifugation (7000 × *g*, 10 min). The supernatant was decanted and each pellet resuspended in 75 µL Bugbuster with 0.75 u benzonase and protease inhibitors. The resuspended cell pellets were left to lyse chemically at r.t. with gentle rocking (1 h) and then clarified by centrifugation (17 000 × *g*, 30 min). The supernatant was collected and either visualised by fluorescence or analysed by Coomassie-stained SDS-PAGE.

7.2.4 Expression and purification of green fluorescent proteins

The pBAD vector containing ampicillin resistance and either Ser-GFP(Y39TAG) or sfGFP(N10TAG) genes, together with the pEVOL vector containing tRNA^{Pyl}, pylRS (*M. mazei*, wild type) and chloramphenicol resistance genes, were co-transformed into electrocompetent *E. coli* Top10 cells and selected on LB agar plates containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml).

For small-scale expression, 0.5 mL of an overnight culture grown from a single colony was inoculated into 50 mL Terrific Broth Medium containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml) in a 250 mL baffled conical flask. At 37 °C with shaking (220 rpm), cells typically grew within 3 h to an OD_{600} of 0.2-0.3, at which point NCAA (stock solution 80 mM in 0.1 M NaOH (aq.)) was added to a final concentration of 1.5 mM. The cultures were allowed to grow until an OD_{600} of 0.4-0.6, at which point protein expression was induced by addition of L-arabinose (stock solution 20% (w/w)) at a final concentration of 0.02% (w/w) and left to grow for 16-18 h at 37 °C with shaking (220 rpm).

The cultures were harvested by centrifugation (6 000 × g, 4 °C, 20 min). Pellets were resuspended in 4 × PBS, 10 mM imidazole, pH 8.0 with a Pierce Protease Inhibitor (EDTA-free) tablet and then lysed by sonication on ice for 6 × 30 s with 30 s intervals. The lysate was clarified by centrifugation (20 000 × g, 4 °C, 20 min) and loaded onto a Ni HiTrap Chelating HP column (1 ml, GE Healthcare) pre-equilibrated in 4 × PBS, 10

mM imidazole, pH 8.0. The column was washed with 10 column volumes of this buffer and then eluted using a gradient of 0-100% 4 × PBS, 500 mM imidazole, pH 8.0 over 7.5 column volumes, taking 0.5 mL fractions, and the column washed with 7.5 column volumes of 4 × PBS, 500 mM imidazole, pH 8.0, taking 0.5 mL fractions. Fractions containing full-length protein (as determined by SDS-PAGE) were pooled, dialysed into 1 × PBS, pH 7.4 and concentrated (Vivaspin centrifugal concentrator, 10000 MWCO) to a final concentration of 330 µM (as determined by UV-visible spectroscopy, ε_{280} = 2.0 × 10^4 dm^3 mol^{-1} cm^{-1}) and stored at -80 °C.

For large-scale expressions, the procedure was followed largely as above with 1 L cultures inoculated with 10 mL of an overnight culture, to which NCAA and L-arabinose were added to the same final concentrations. For purification, a larger Ni HiTrap Chelating HP column (5 mL, GE Healthcare) was used, taking 2.5 mL fractions. This procedure is identical whether using NCAAs **86** or **105**.

7.2.5 Expression and purification of BiGalK

The pET22b vector containing ampicillin resistance and BiGalK(K417TAG) genes, together with the pEVOL vector containing $tRNA^{Pyl}$, pylRS (*M. mazei*, wild type) and chloramphenicol resistance genes, were co-transformed into electrocompetent *E. coli* BL21-DE3 cells and selected on LB agar plates containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml).

For small-scale expression, 0.5 mL of an overnight culture grown from a single colony was inoculated into 50 mL LB medium containing ampicillin (100 µg/ml) and chloramphenicol (35 µg/ml) in a 250 mL baffled conical flask. At 37 °C with shaking (220 rpm), cells typically grew within 3 h to an OD_{600} of 0.2-0.3, at which point NCAA (stock solution 80 mM in 0.1 M NaOH (aq.), final concentration 1.5 mM) and L-ara (stock solution 20% (w/v) final concentration 0.02% (w/v)), were added. The cultures were allowed to grow until an OD_{600} of 0.8, at which point protein expression was induced by addition of IPTG (stock solution 1 M in H_2O, final concentration 0.2 mM) and left to grow for 16-18 h at 16 °C with shaking (220 rpm).

The cultures were harvested by centrifugation (8 000 × g, 4 °C, 20 min). Pellets were resuspended in 50 mM Tris buffer, 20 mM imidazole, pH 8.0 with a Pierce Protease Inhibitor (EDTA-free) tablet and then lysed by sonication on ice for 6 × 30 s with 30 s intervals. The lysate was clarified by centrifugation (20 000 × g, 4 °C, 20 min) and loaded onto a Ni HiTrap Chelating HP column (1 mL, GE Healthcare) pre-equilibrated in 50 mM Tris buffer, 20 mM imidazole, pH 8.0. The column was washed with 10 column volumes of this buffer and then eluted using a gradient of 0-100% 50 mM Tris, 250 mM imidazole,

pH 8.0 over 7.5 column volumes, taking 0.5 mL fractions, and the column washed with 7.5 column volumes of 50 mM Tris, 250 mM imidazole, pH 8.0, taking 0.5 mL fractions. Fractions containing full-length protein (as determined by SDS-PAGE) were pooled, dialysed into 1 × PBS, pH 7.4 and concentrated (Vivaspin centrifugal concentrator, 10000 MWCO) to a final concentration of 330 μM (as determined by UV-visible spectroscopy, $\varepsilon_{280} = 3.0 \times 10^4$ dm^3 mol^{-1} cm^{-1}) and stored at -80 °C.

7.2.6 Protein characterisation

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analysis and staining

All SDS-PAGE analysis was performed using handcast 12% polyacrylamide gels. Samples were reduced by boiling for 10 min (2% SDS, 2 mM 2-mercaptoethanol, 4% glycerol, 40 mM Tris-HCl pH 6.8, 0.01% bromophenol blue). Molecular weight markers used were PageRuler Plus Prestained Protein Ladder (Thermo Scientific). Each gel was run at 200 volts for 45-80 min. For Coomassie stain experiments, the gel was washed with fixing solution (40% (v/v) MeOH, 10% AcOH), stained with 0.1% (w/v) Coomassie Brilliant Blue R-250 (50% (v/v) MeOH, 10% AcOH), and finally repeatedly washed with destain solution (50% (v/v) MeOH, 10% AcOH). Images of the resulting gels were captured and analysed using a Syngene G:BOX Chemi XRQ equipped with a Synoptics 4.0 MP camera, with GeneSys software (Version 1.5.7.0).

Determination of protein concentration

Protein concentration was estimated using UV-visible absorption spectroscopy and the predicted molar absorption coefficient at 280 nm (ExPASy ProtParam). This predicted ε_{280} value is based on empirical data relating to the molar absorption coefficients of tyrosine, tryptophan and cystine, respectively 1490, 5500 and 125 dm^3 mol^{-1} cm^{-1},[301] with other canonical residues demonstrating negligible absorbance at 280 nm. The predicted molar absorption coefficient is therefore only based upon the primary sequence of the protein and is an estimate, as secondary and tertiary protein structure may influence individual residue microenvironments and perturb the value of the molar absorption coefficient. This method also encounters difficulties with post-translational modifications, such as the chromophore of GFP and derivatives. However, the ε_{280} of the GFP chromophore is known to be low[253] and can be ignored in an estimate of GFP concentration to within 16% error.[302]

Fluorescent imaging

For fluorescent imaging of fluorescently modified proteins, the SDS PAGE gel was washed with fixing solution (40% MeOH, 10% AcOH). Visualisation of protein fluorescence, and images of the resulting gels, were captured by excitation at 302 nm and analysed using a Syngene G:BOX Chemi XRQ equipped with a Synoptics 4.0 MP camera in line with GeneSys software (Version 1.5.7.0).

Western blot analysis

For Western blot analysis, 2.5 µg of biotinylated protein samples or negative controls were run on 12% SDS–PAGE and transferred onto a nitrocellulose membrane filter (0.45 µm, Amersham Protran Sandwich, GE Healthcare) using an electroblot apparatus (Bio-Rad, Hercules, CA) at 100 V, 350 mA for 1 h in cooled transfer buffer (25 mM Tris–HCl pH 8.3, 192 mM glycine, 0.1% SDS, 20% (v/v) methanol). The membrane was incubated in blocking solution (Phosphate-buffered saline (PBS) tablets, Sigma)) containing 5% (w/v) non-fat dry milk powder for 16 h at 4 °C. The membrane was processed through sequential incubations with primary antibody, alkaline phosphatase anti-biotin (goat, Vector Labs, CA) 1:1000 dilution in PBS for 1 hour at room temperature, followed by washing in PBS, 0.01% Tween-20, and then incubation with visualising substrate BCIP/NBT Alkaline Phosphatase Substrate Kit (Vector Labs, CA) until immunoreactive proteins on the membrane were visible (*ca*. 20 min). The reaction was stopped by washing the membrane in distilled water. The membranes were imaged using a Syngene G:BOX Chemi XRQ equipped with a Synoptics 4.0 MP camera, with GeneSys software (Version 1.5.7.0).

Protein mass spectrometry

Protein ESI mass spectra were obtained on a Bruker Solarix XR 9.4 T FTICR instrument. Samples were desalted and analysed at a final concentration of 0.3-10 µM in 50:50:1 (v/v) H_2O:MeCN:FA. Mass spectra were analysed and deconvoluted using Bruker DataAnalysis 4.4.

Trypsin digest

All trypsin digests were performed by Dr Adam Dowle.

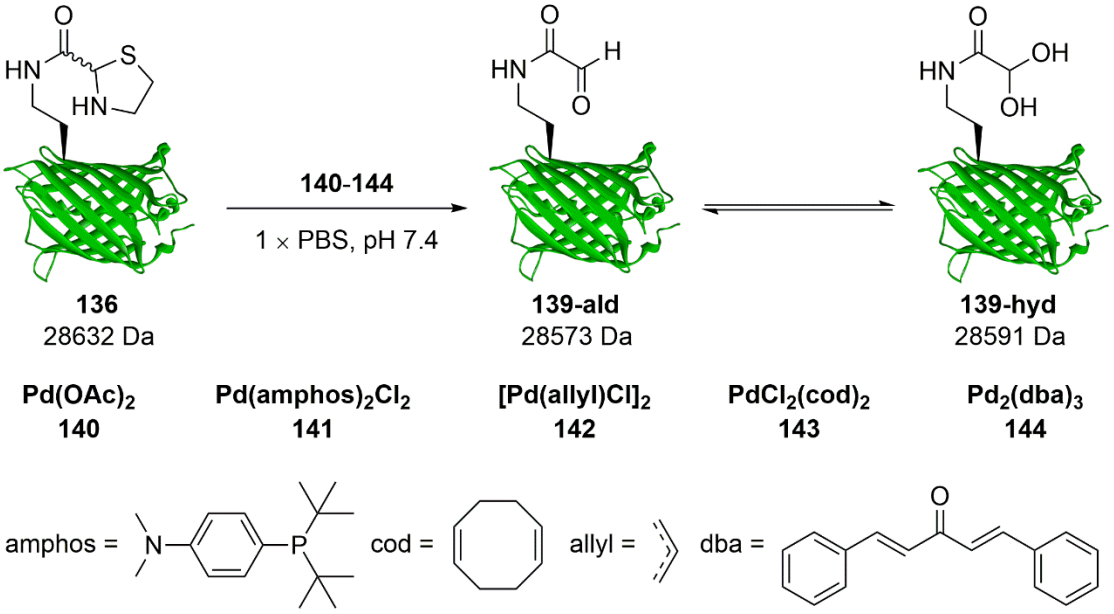
A solution of protein (5 µg in 100 µL 100 mM ammonium bicarbonate) was reduced with 5 mM DTT for 30 min at 50 °C and alkylated with 15 mM iodoacetamide for 30 min in the dark at r.t., followed by digestion by the addition of 0.2 µg trypsin (sequencing grade, Promega) and incubated overnight at 37 °C. Protease activity was stopped with aqueous trifluoroacetic acid (1% (v/v)).

The resulting mixture of peptides was loaded onto a Pepmap (50 cm × 2 μm, 100 Å) and a Thermo C18 EasyNano nanocapillary column (15 cm × 75 μm) and eluted over a gradient of aqueous 3-35% (v/v) MeCN over 35 min into a Thermo Orbitrap Fusion hybrid mass spectrometer. MS1 and MS2 spectra were acquired in the Orbitrap mass analyser with Easy-IC internal calibration. Data-dependent acquisition was performed in top speed mode using a fixed 1 s cycle, selecting the most intense precursors with charge states 2-5. HCD was used for peptide fragmentation with 32% activation energy.

Resulting spectral data were searched against the expected protein sequences using the PEAKS-DB search program. Search criteria specified: enzyme, trypsin; peptide tolerance, 3 ppm; MS/MS tolerance, 0.01 Da. Carbamidomethylation (C) was set as a fixed modification. Peptide matches were filtered to achieve a global false discovery method of < 5 %.

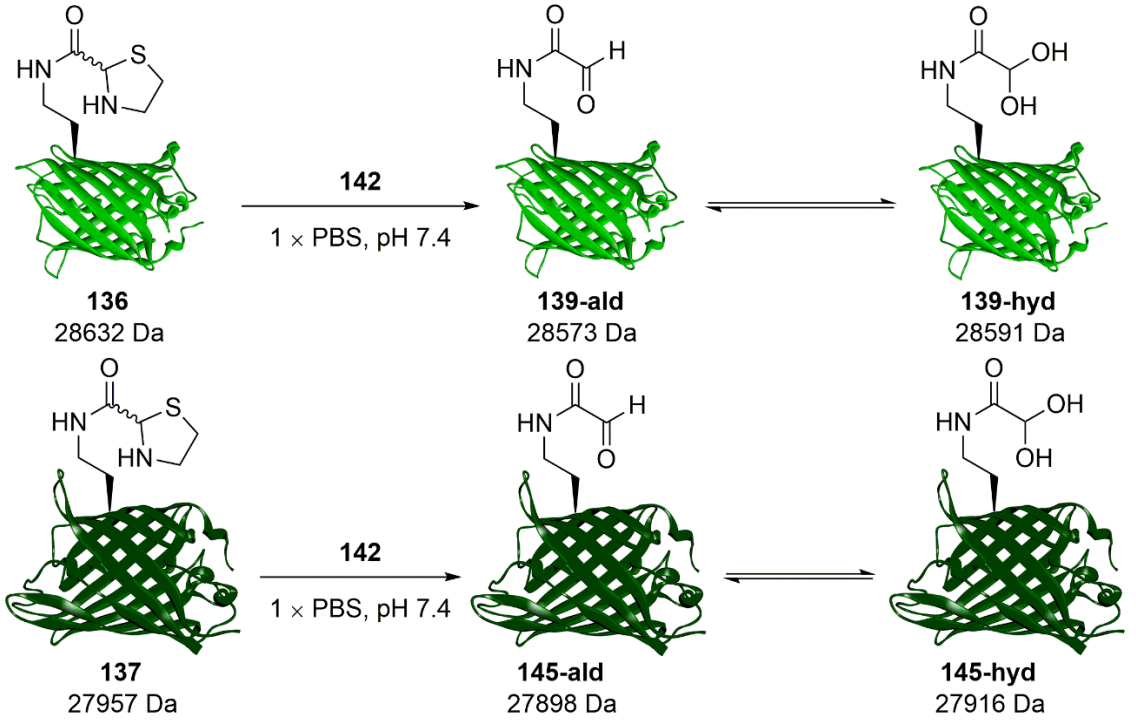
7.3 Protein modification protocols

7.3.1 Palladium reagent screen



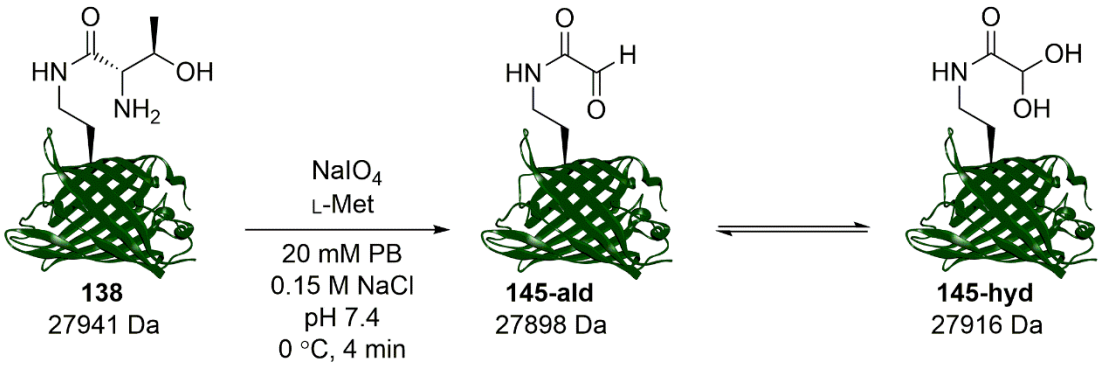
For a typical 100 µL scale reaction, palladium reagent **140**-**144** (stock 0.55 M in DMSO for 100 eq., 6 µL, final concentration 31 mM) was added to **136** (stock 330 µM in 1 × PBS, pH 7.4, 100 µL, final concentration 310 µM) in a 0.5 mL Eppendorf tube. The reaction mixture was vortex mixed (1 s) and incubated in a water bath at 37 °C. The reaction was quenched by the addition of DTT (a spatula tip, *ca*. 5 mg) and clarified by centrifuged (5 000 × *g*, 2 min or longer as needed). The supernatant was decanted and desalted using a PD SpinTrap G25 column (GE Healthcare Life Sciences) into H_2O for mass spectrometry analysis.

7.3.2 Optimised palladium decaging procedure



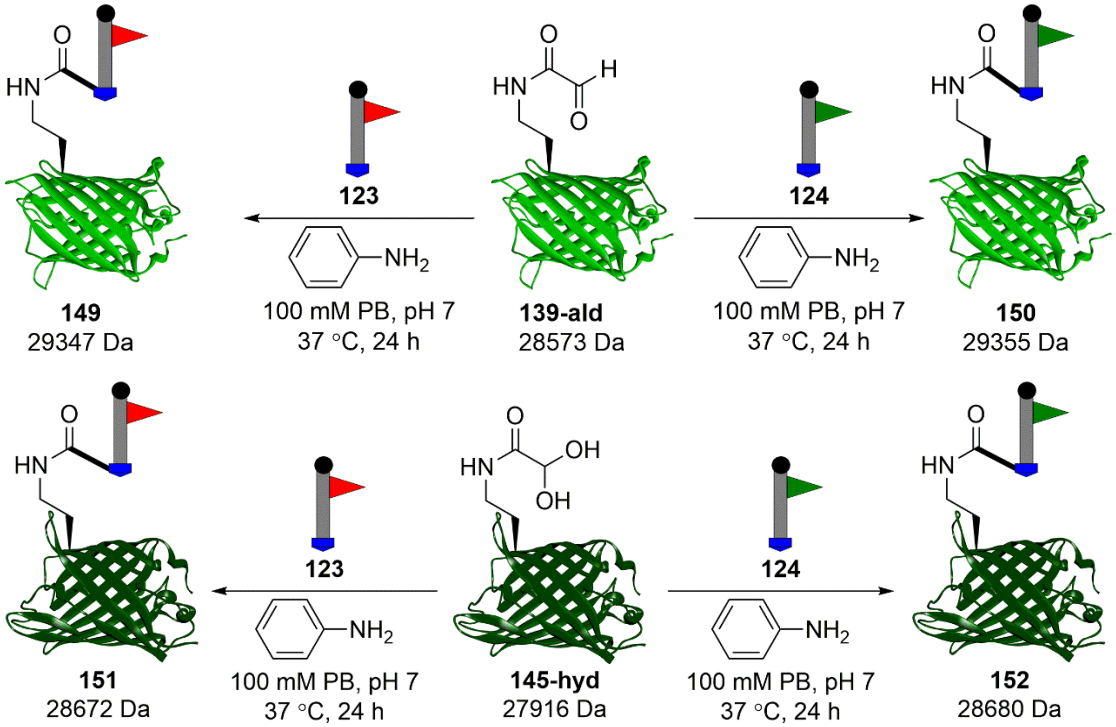
For a typical 100 µL scale reaction, palladium complex **142** (stock 30 mM in DMSO, 1 µL, final concentration 0.3 mM, 1 eq.) was added to protein thiazolidine **136** or **137** (stock 0.3 mM in 1 × PBS, pH 7.4, 100 µL, final concentration 0.3 mM, 1 eq.) in a 0.5 mL Eppendorf tube, mixed by pipette tip swirling, and left at rt for 1 h. The reaction was quenched by the addition of 3-mercaptopropanoic acid (stock 1% (v/v) in 10 × PBS, pH 7.4, 1 µL, final concentration 0.01% (v/v)) and left at rt for 15 min. The reaction mixture was diluted up to 500 µL with H_2O and desalted using a PD Minitrap G-25 (GE Healthcare), gravity method, into H_2O for analysis and further manipulation, affording protein glyoxyl **139** or **145**.

7.3.3 Optimised periodate oxidation procedure



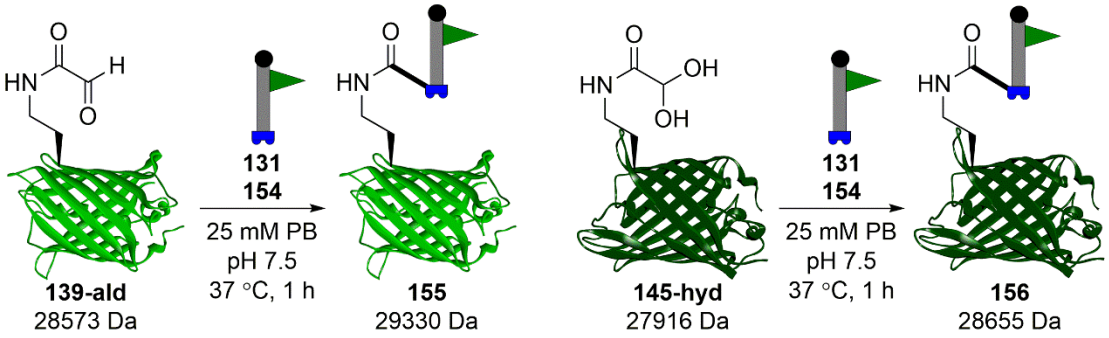
For a typical 100 µL scale reaction, solutions of L-methionine (stock 102 mM in 20 mM phosphate buffer, 150 mM NaCl, pH 7.4, 1 µL, final concentration 1 mM, 10 eq.) and $NaIO_4$ (stock 51 mM in 20 mM phosphate buffer, 150 mM NaCl, pH 7.4, 1 µL, final concentration 0.5 mM, 5 eq.) were added to protein 1,2-aminoalcohol **138** (stock 100 µM in 20 mM phosphate buffer, 150 mM NaCl, pH 7.4, 100 µL, final concentration 100 µM, 1 eq.). After mixing by gentle pipetting, the reaction mixture was incubated at 0 °C in the dark for 4 min, after which the oxidised protein was purified by desalting using a PD Minitrap G-25 (GE Healthcare), gravity method, into H_2O for analysis and further manipulation.

7.3.4 Oxime ligation



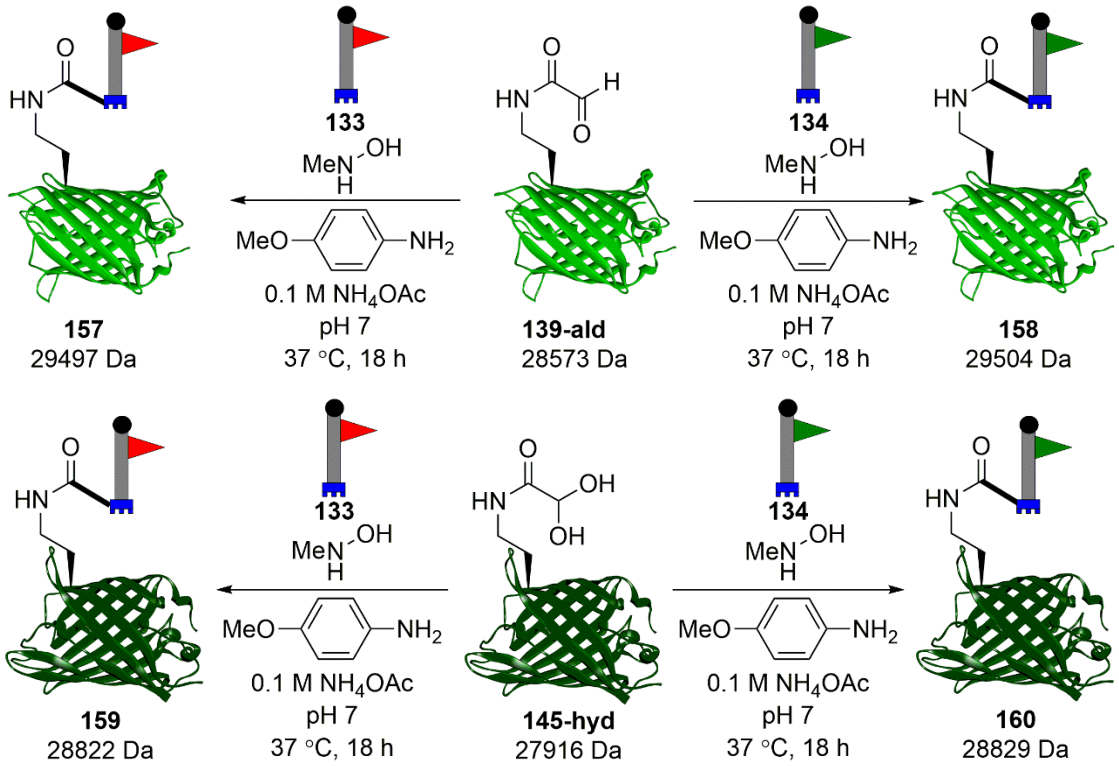
For a typical 100 µL scale reaction, aniline (neat, 1 µL, final concentration 0.1 M) and aminooxy probe **123** or **124** (stock 0.1 M in H_2O, 2 µL, final concentration 2.0 mM) were added to protein glyoxyl **139** or **145** (stock 60 µM in 100 mM sodium phosphate buffer, pH 7.0, 97 µL, final concentration 58 µM) and mixed by pipette tip swirling. After incubation at 37 °C in a water bath for 24 h, the reaction was diluted up to 500 µL with H_2O and desalted using a PD Minitrap G-25 (GE Healthcare), gravity method, into H_2O for analysis and further manipulation.

7.3.5 Organocatalytic protein aldol ligation



For a typical 100 µL scale reaction, solutions of phenacetaldehyde probe **131** (stock 50 mM in H_2O, 4 µL, final concentration 2 mM, 33 eq.) and proline tetrazole **154** (stock 400 mM in 25 mM phosphate buffer, pH 7.5, 6 µL, final concentration 2.5 mM, 26 eq.) were added to protein glyoxyl **139** or **145** (stock 300 µM in H_2O, 25 µL, final concentration 75 µM, 1 eq.) and the solution made up to 100 µL with buffer (65 µL, 25 mM phosphate buffer, pH 7.5). After mixing by gentle pipetting, the reaction mixture was incubated at 37 °C for 1 h, after which the modified protein was purified by desalting using a PD Minitrap G-25 (GE Healthcare), gravity method, into H_2O for analysis and further manipulation.

7.3.6 SPANC ligation



For a typical 100 µL scale reaction, solutions of *p*-anisidine (stock 1 M in 50% aq. MeCN, 10 µL, final concentration 100 mM), *N*-methylhydroxylamine hydrochloride (stock 75 mM in 0.1 M NH_4OAc, pH 6.8, 2 µL, final concentration 1.5 mM, 50 eq.) and BCN probe **133** or **134** (stock 3 mM in 50% aq. MeCN, 20 µL, final concentration 600 µM, 20 eq.) were added to protein aldehyde **139** or **145** (stock 100 µM in H_2O, 30 µL, final concentration 30 µM, 1 eq.) and the solution made up to 100 µL with buffer (38 µL, 0.1 M NH_4OAc, pH 6.8). After mixing by gentle pipetting, the reaction mixture was incubated at 37 °C for 16 h, after which the modified protein was purified by desalting using a PD Minitrap G-25 (GE Healthcare), gravity method, into H_2O for analysis and further manipulation.

Chapter 8: Appendix

8.1 Protein sequences

The locations of amber stop codons are denoted by X. The initial methionine residue has been assigned the residue number one for all protein constructs found to contain this residue after production, i.e. sfGFP and BiGalK. For the EGFP construct, literature examples have excluded the eight-residue FLAG tag from numbering and only the native sequence has been numbered.[169] Following this example, the mutated FLAG tag has not been included in residue numbering and has been highlighted in gray to indicate this. Owing to the mutated FLAG tag, the EGFP proteins produced do not contain the initial methionine residue and this has been omitted.

Ser-EGFP(Y39TAG)-His_6

```
         1          11          21          31          41          51
 SYKDDDDKV SKGEELFTGV VPILVELDGD VNGHKFSVSG EGEGDATXGK LTLKFICTTG
        61          71          81          91         101         111
KLPVPWPTLV TTLTYGVQCF SRYPDHMKQH DFFKSAMPEG YVQERTIFFK DDGNYKTRAE
       121         131         141         151         161         171
VKFEGDTLVN RIELKGIDFK EDGNILGHKL EYNYNSHNVY IMADKQKNGI KANFKIRHNI
       181         191         201         211         221         231
EDGSVQLADH YQQNTPIGDG PVLLPDNHYL STQSALSKDP NEKRDHMVLL EFVTAAGITL
       240
GMDELYKHHH HHH
```

sfGFP(N150TAG)-His_6

```
1       10         20         30         40         50         60
MVSKGEELFT GVVPILVELD GDVNGHKFSV RGEGEGDATN GKLTLKFICT TGKLPVPWPT
        70         80         90        100        110        120
LVTTLTYGVQ CFSRYPDHMK RHDFFKSAMP EGYVQERTIS FKDDGTYKTR AEVKFEGDTL
       130        140        150        160        170        180
VNRIELKGID FKEDGNILGH KLEYNFNSHX VYITADKQKN GIKANFKIRH NVEDGSVQLA
       190        200        210        220        230        240
DHYQQNTPIG DGPVLLPDNH YLSTQSVLSK DPNEKRDHMV LLEFVTAAGI THGMDELYKG


SHHHHHH
```

BiGalK(K417TAG)-His_6

```
1       10         20         30         40         50         60
MTAVEFIEPL THEEGVSQAT KLFVDTYGAA PEGVWAAPGR VNLIGEHTDY NAGLCLPIAL
        70         80         90        100        110        120
PHRTFIALKP REDTKVRVVS GVAPDKVAEA DLDGLKARGV DGWSAYPTGV AWALRQAGFD
       130        140        150        160        170        180
KVKGFDAAFV SCVPLGSGLS SSAAMTCSTA LALDDVYGLG YGDSDAGRVT LINAAIKSEN
       190        200        210        220        230        240
EMAGASTGGL DQNASMRCTE GHALLLDCRP ELTPLENVSQ QEFDLDKYNL ELLVVDTQAP
       250        260        270        280        290        300
HQLNDGQYAQ RRATCEEAAK ILGVANLRVT ADGISKADDQ FQALKETLDA LPDETMKKRV
       310        320        330        340        350        360
RHVVTEIERV RSFVRAFAQG DIKAAGRLFN ASHDSLAADY EVTVPELDIA VDVARKNGAY
       370        380        390        400        410        420
GARMTGGGFG GSIIALVDKG QGHEIAQKIA DRFEKEGFNA PRALPAFAAA SASREAXLAA


ALEHHHHHH
```

Abbreviations

Ac	Acetyl
ADC	Antibody-drug conjugate
ADP	Adenosine diphosphate
Alloc	Allyloxycarbonyl
ATP	Adenosine triphosphate
ATR	Attenuated total reflectance
BCN	Bicyclo[6.1.0]non-4-yne
Bcx	*Bacillus circulans* xylanase
BiGalK	*Bifidobacterium infantis* galactokinase
Bn	Benzyl
Boc	Tert-butoxycarbonyl
BSA	Bovine serum albumin
Bz	Benzoyl
Cbz	Benzyloxycarbonyl
cod	Cycloocta-1,5-diene
Cp*	1,2,3,4,5-pentamethylcyclopentadiene
CuAAC	Copper-catalysed alkyne-azide cycloaddition
dba	Dibenzylideneacetone
DBU	1,8-diazabicyclo[5.4.0]undec-7-ene
DCC	Dicyclohexylcarbodiimide
DCM	Dichloromethane
d.e.	Diastereomeric excess
DIPEA	*N*,*N*-diisopropylethylamine
DMAB	*para*-dimethylaminobenzaldehyde
DMF	*N*,*N*-dimethylformamide
DMS	Dimethylsulfide
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
e.e.	Enantiomeric excess
EGFP	Enhanced green fluorescent protein
eq.	Equivalents
ESI	Electrospray ionisation

FA	Formic acid
FGE	Formylglycine-generating enzyme
Fmoc	9*H*-fluoren-9-ylmethoxycarbonyl
FRET	Förster resonance energy transfer
FTICR	Fourier-transform ion cyclotron resonance
GFP	Green fluorescent protein
GlcNAc	*N*-acetylglucosamine
GST	Glutathione *S*-transferase
HASPA	Hydrophilic acylated surface protein A
HCTU	2-(6-chloro-1*H*-benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate
HPLC	High-performance liquid chromatography
HRMS	High-resolution mass spectrometry
ICP	Inductively coupled plasma
IR	Infrared
LB	Lysogeny broth
LC	Liquid chromatography
MALDI	Matrix-assisted laser desorption/ionisation
MeCN	Acetonitrile
MS	Mass spectrometry
NAD	Nicotinamide adenine dinucleotide
NADH	Reduced nicotinamide adenine dinucleotide
NCAA	Non-canonical amino acid
NCL	Native chemical ligation
NHS	*N*-hydroxysuccinimide
NMR	Nuclear magnetic resonance
OPAL	Organocatalytic protein aldol ligation
PB	Phosphate buffer
PBS	Phosphate-buffered saline
PDB	Protein Data Bank
PEG	Polyethylene glycol
P_i	Phosphate
PLP	Pyridoxal-5-phosphate
PTFE	Polytetrafluoroethylene
p-Ts	*para*-toluenesulfonic acid
Py	Pyridine
Pyl	Pyrrolysine

pylRS	Pyrrolysyl RNA synthetase
pylT	Pyrrolysyl tRNA
RNA	Ribonucleic acid
RS	tRNA synthetase
r.t.	Room temperature
SAM	*S*-adenosyl methionine
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
sfGFP	Superfolder green fluorescent protein
SPAAC	Strain-promoted alkyne-azide cycloaddition
SPANC	Strain-promoted alkyne-nitrone cycloaddition
SPIEDAC	Strain-promoted inverse electron-demand Diels-Alder cycloaddition
SPPS	Solid-phase peptide synthesis
Su	N-hydroxysuccinimidyl
SUMO	Small ubiquitin-like modifier
TCEP	Tris(2-carboxyethyl)phosphine
TCO	trans-cyclooctene
Tf	Trifluoromethylsulfonyl
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
TIS	Triisopropylsilane
TLC	Thin-layer chromatography
TMSCl	Trimethylsilyl chloride
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer ribonucleic acid
UV	Ultraviolet
wt	Wild type

References

1. R. Brabham and M. A. Fascione, *ChemBioChem*, 2017, **18**, 1973-1983.
2. M. Cobb, *PLoS Biol.*, 2017, **15**, e2003243.
3. F. Crick, *Nature*, 1970, **227**, 561-563.
4. E. V. Koonin, *Biol. Direct*, 2012, **7**, 27.
5. C. Walsh, *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, W. H. Freeman, 2006.
6. H. S. Chung, S.-B. Wang, V. Venkatraman, C. I. Murray and J. E. Van Eyk, *Circ. Res.*, 2013, **112**, 382-392.
7. F. Ardito, M. Giuliani, D. Perrone, G. Troiano and L. Lo Muzio, *Int. J. Mol. Med.*, 2017, **40**, 271-280.
8. L. Song and Z.-Q. Luo, *J. Cell Biol.*, 2019, **218**, 1776.
9. J. D. Marth and P. K. Grewal, *Nat. Rev. Immunol.*, 2008, **8**, 874-887.
10. M. Kashif, A. Moreno-Herrera, E. E. Lara-Ramirez, E. Ramírez-Moreno, V. Bocanegra-García, M. Ashfaq and G. Rivera, *J. Drug Targeting*, 2017, **25**, 485-498.
11. S. Prabakaran, G. Lippens, H. Steen and J. Gunawardena, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2012, **4**, 565-583.
12. P. Edman, *Acta Chem. Scand.*, 1950, **4**, 283-293.
13. G. E. Means and R. E. Feeney, *Biochemistry*, 1968, **7**, 2192-2201.
14. N. Jentoft and D. G. Dearborn, *J. Biol. Chem.*, 1979, **254**, 4359-4365.
15. H. Buttkus, J. R. Clark and R. E. Feeney, *Biochemistry*, 1965, **4**, 998-1005.
16. A. L. Grossberg and D. Pressman, *Biochemistry*, 1963, **2**, 90-96.
17. A. M. Crestfield, S. Moore and W. H. Stein, *J. Biol. Chem.*, 1963, **238**, 622-627.
18. R. J. Hill and R. W. Davis, *J. Biol. Chem.*, 1967, **242**, 2005-2012.
19. A. L. Murdock, K. L. Grist and C. H. W. Hirs, *Arch. Biochem. Biophys.*, 1966, **114**, 375-390.
20. H. Kaplan, K. J. Stevenson and B. S. Hartley, *Biochem. J.*, 1971, **124**, 289-299.
21. R. G. Duggleby and H. Kaplan, *Biochemistry*, 1975, **14**, 5168-5175.
22. M. A. Hass and F. A. Mulder, *Annu. Rev. Biophys.*, 2015, **44**, 53-75.
23. Y. Inada, K. Takahashi, T. Yoshimoto, A. Ajima, A. Matsushima and Y. Saito, *Trends Biotechnol*, 1986, **4**, 190-194.
24. E. M. Pelegri-O'Day, E. W. Lin and H. D. Maynard, *J. Am. Chem. Soc.*, 2014, **136**, 14323-14332.
25. A. Abuchowski, J. R. McCoy, N. C. Palczuk, T. van Es and F. F. Davis, *J. Biol. Chem.*, 1977, **252**, 3582-3586.

26. C. O. Beauchamp, S. L. Gonias, D. P. Menapace and S. V. Pizzo, *Anal. Biochem.*, 1983, **131**, 25-33.
27. F. M. Veronese, R. Largajolli, E. Boccu, C. A. Benassi and O. Schiavon, *Appl. Biochem. Biotechnol.*, 1985, **11**, 141-152.
28. E. Boccu, R. Largajolli and F. M. Veronese, *Z. Naturforsch., C: J. Biosci*, 1983, **38**, 94-99.
29. M. Ehrat and P. L. Luisi, *Biopolymers*, 1983, **22**, 569-573.
30. A. Abuchowski, T. van Es, N. C. Palczuk and F. F. Davis, *J. Biol. Chem.*, 1977, **252**, 3578-3581.
31. E. Boccu, G. P. Velo and F. M. Veronese, *Pharmacol. Res. Commun.*, 1982, **14**, 113-120.
32. N. M. Green, *Biochem. J.*, 1963, **89**, 585-591.
33. M. A. Fixsen, *Biochem. J.*, 1931, **25**, 596-605.
34. D. U. V. V, D. B. Melville, P. Gyorgy and C. S. Rose, *Science*, 1940, **92**, 62-63.
35. P. Gyorgy, C. S. Rose, R. E. Eakin, E. E. Snell and R. J. Williams, *Science*, 1941, **93**, 477-478.
36. D. Pacheco-Alvarez, R. S. Solorzano-Vargas and A. L. Del Rio, *Arch. Med. Res.*, 2002, **33**, 439-447.
37. N. M. Green, in *Advances in Protein Chemistry*, eds. C. B. Anfinsen, J. T. Edsall and F. M. Richards, Academic Press, 1975, vol. 29, pp. 85-133.
38. T. Kuroishi, L. Rios-Avila, V. Pestinger, S. S. Wijeratne and J. Zempleni, *Mol. Genet. Metab.*, 2011, **104**, 537-545.
39. N. Kothapalli, G. Camporeale, A. Kueh, Y. C. Chew, A. M. Oommen, J. B. Griffin and J. Zempleni, *J. Nutr. Biochem.*, 2005, **16**, 446-448.
40. R. E. Barden, B. L. Taylor, F. Isoashi, W. H. Frey, G. Zander, J. C. Lee and M. F. Utter, *Proc. Natl. Acad. Sci. USA*, 1975, **72**, 4308-4312.
41. E. A. Bayer and M. Wilchek, in *Methods of Biochemical Analysis*, ed. D. Glick, John Wiley & Sons, 1980, ch. The Use of the Avidin-Biotin Complex as a Tool in Molecular Biology, pp. 1-45.
42. K. Hofmann, G. Titus, J. A. Montibeller and F. M. Finn, *Biochemistry*, 1982, **21**, 978-984.
43. F. M. Finn, G. Titus, D. Horstman and K. Hofmann, *Proc. Natl. Acad. Sci. USA*, 1984, **81**, 7328-7332.
44. G. Elia, *Proteomics*, 2008, **8**, 4012-4024.
45. O. H. Laitinen, H. R. Nordlund, V. P. Hytonen, S. T. Uotila, A. T. Marttila, J. Savolainen, K. J. Airenne, O. Livnah, E. A. Bayer, M. Wilchek and M. S. Kulomaa, *J. Biol. Chem.*, 2003, **278**, 4010-4014.
46. D. R. Gretch, M. Suter and M. F. Stinski, *Anal. Biochem.*, 1987, **163**, 270-277.

47. E. Hazum, *J. Chromatogr.*, 1990, **510**, 233-238.
48. L. Stryer, *Annu. Rev. Biochem.*, 1978, **47**, 819-846.
49. D. W. Piston and G.-J. Kremers, *Trends Biochem. Sci.*, 2007, **32**, 407-414.
50. W. Veatch and L. Stryer, *J. Mol. Biol.*, 1977, **113**, 89-102.
51. T. C. Tsao and K. Bailey, *Biochim. Biophys. Acta*, 1953, **11**, 102-113.
52. J. E. Moore and W. H. Ward, *J. Am. Chem. Soc.*, 1956, **78**, 2414-2418.
53. C. W. Wu and L. R. Yarbrough, *Biochemistry*, 1976, **15**, 2863-2868.
54. R. N. Johnson, P. Kopeckova and J. Kopecek, *Biomacromolecules*, 2012, **13**, 727-735.
55. N. Kotagiri, Z. Li, X. Xu, S. Mondal, A. Nehorai and S. Achilefu, *Bioconjug. Chem.*, 2014, **25**, 1272-1281.
56. N. Jain, S. W. Smith, S. Ghone and B. Tomczuk, *Pharm. Res.*, 2015, **32**, 3526-3540.
57. S. C. Alley, D. R. Benjamin, S. C. Jeffrey, N. M. Okeley, D. L. Meyer, R. J. Sanderson and P. D. Senter, *Bioconjug. Chem.*, 2008, **19**, 759-765.
58. L. N. Tumey, M. Charati, T. He, E. Sousa, D. Ma, X. Han, T. Clark, J. Casavant, F. Loganzo, F. Barletta, J. Lucas and E. I. Graziani, *Bioconjug. Chem.*, 2014, **25**, 1871-1880.
59. A. D. Baldwin and K. L. Kiick, *Bioconjug. Chem.*, 2011, **22**, 1946-1953.
60. L. M. Tedaldi, M. E. Smith, R. I. Nathani and J. R. Baker, *Chem. Commun.*, 2009, DOI: 10.1039/b915136b, 6583-6585.
61. M. E. Smith, F. F. Schumacher, C. P. Ryan, L. M. Tedaldi, D. Papaioannou, G. Waksman, S. Caddick and J. R. Baker, *J. Am. Chem. Soc.*, 2010, **132**, 1960-1965.
62. J. P. Nunes, M. Morais, V. Vassileva, E. Robinson, V. S. Rajkumar, M. E. Smith, R. B. Pedley, S. Caddick, J. R. Baker and V. Chudasama, *Chem. Commun.*, 2015, **51**, 10624-10627.
63. M. Morais, J. P. M. Nunes, K. Karu, N. Forte, I. Benni, M. E. B. Smith, S. Caddick, V. Chudasama and J. R. Baker, *Org. Biomol. Chem.*, 2017, **15**, 2947-2952.
64. C. B. Rosen and M. B. Francis, *Nat. Chem. Biol.*, 2017, **13**, 697-705.
65. L. Zhang and J. P. Tam, *Anal. Biochem.*, 1996, **233**, 87-93.
66. H. Ren, F. Xiao, K. Zhan, Y. P. Kim, H. Xie, Z. Xia and J. Rao, *Angew. Chem. Int. Ed.*, 2009, **48**, 9658-9662.
67. P. E. Dawson, T. W. Muir, I. Clark-Lewis and S. B. Kent, *Science*, 1994, **266**, 776-779.
68. T. J. Simmons, K. E. H. Frandsen, L. Ciano, T. Tryfona, N. Lenfant, J. C. Poulsen, L. F. L. Wilson, T. Tandrup, M. Tovborg, K. Schnorr, K. S. Johansen,

B. Henrissat, P. H. Walton, L. Lo Leggio and P. Dupree, *Nat. Commun.*, 2017, **8**, 1064.

69. X. Chen, K. Muthoosamy, A. Pfisterer, B. Neumann and T. Weil, *Bioconjug. Chem.*, 2012, **23**, 500-508.
70. M. J. Matos, B. L. Oliveira, N. Martinez-Saez, A. Guerreiro, P. Cal, J. Bertoldo, M. Maneiro, E. Perkins, J. Howard, M. J. Deery, J. M. Chalker, F. Corzana, G. Jimenez-Oses and G. J. L. Bernardes, *J. Am. Chem. Soc.*, 2018, **140**, 4004-4017.
71. A. H. Gordon, A. J. Martin and R. L. Synge, *Biochem. J.*, 1941, **35**, 1369-1387.
72. Geschwind, II and C. H. Li, *Biochim. Biophys. Acta*, 1954, **15**, 442-443.
73. H. B. Dixon, *Biochem. J.*, 1962, **83**, 91-94.
74. H. B. Dixon and L. R. Weitkamp, *Biochem. J.*, 1962, **84**, 462-468.
75. H. B. Dixon and V. Moret, *Biochem. J.*, 1965, **94**, 463-469.
76. J. M. Gilmore, R. A. Scheck, A. P. Esser-Kahn, N. S. Joshi and M. B. Francis, *Angew. Chem. Int. Ed. Engl.*, 2006, **45**, 5307-5311.
77. R. J. Spears and M. A. Fascione, *Org. Biomol. Chem.*, 2016, **14**, 7622-7638.
78. K. F. Geoghegan and J. G. Stroh, *Bioconjugate Chem.*, 1992, **3**, 138-146.
79. K. Rose, *J. Am. Chem. Soc.*, 1994, **116**, 30-33.
80. J. Kalia and R. T. Raines, *Angew. Chem. Int. Ed.*, 2008, **47**, 7523-7526.
81. L. Polgar and M. L. Bender, *J. Am. Chem. Soc.*, 1966, **88**, 3153-3154.
82. D. H. Strumeyer, W. N. White and D. E. Koshland, Jr., *Proc. Natl. Acad. Sci. USA*, 1963, **50**, 931-935.
83. H. Weiner, W. N. White, D. G. Hoare and D. E. Koshland, Jr., *J. Am. Chem. Soc.*, 1966, **88**, 3851-3859.
84. R. W. Jack and G. Jung, *Curr. Opin. Chem. Biol.*, 2000, **4**, 310-317.
85. J. M. Chalker, S. B. Gunnoo, O. Boutureira, S. C. Gerstberger, M. Fernandez-Gonzalez, G. J. L. Bernardes, L. Griffin, H. Hailu, C. J. Schofield and B. G. Davis, *Chem. Sci.*, 2011, **2**, 1666-1676.
86. J. Dadova, S. R. Galan and B. G. Davis, *Curr. Opin. Chem. Biol.*, 2018, **46**, 71-81.
87. T. H. Wright, B. J. Bower, J. M. Chalker, G. J. L. Bernardes, R. Wiewiora, W.-L. Ng, R. Raj, S. Faulkner, M. R. J. Vallée, A. Phanumartwiwath, O. D. Coleman, M.-L. Thézénas, M. Khan, S. R. G. Galan, L. Lercher, M. W. Schombs, S. Gerstberger, M. E. Palm-Espling, A. J. Baldwin, B. M. Kessler, T. D. W. Claridge, S. Mohammed and B. G. Davis, *Science*, 2016, **354**, aag1465.
88. B. Schmidt, T. Selmer, A. Ingendoh and K. Vonfigura, *Cell*, 1995, **82**, 271-278.
89. I. S. Carrico, B. L. Carlson and C. R. Bertozzi, *Nat. Chem. Biol.*, 2007, **3**, 321-322.

90. R. A. Kudirka, R. M. Barfield, J. M. McFarland, P. M. Drake, A. Carlson, S. Banas, W. Zmolek, A. W. Garofalo and D. Rabuka, *ACS Med. Chem. Lett.*, 2016, **7**, 994-998.
91. A. Tuley, Y. J. Lee, B. Wu, Z. U. Wang and W. R. Liu, *Chem. Commun.*, 2014, **50**, 7424-7426.
92. A. Dumas, L. Lercher, C. D. Spicer and B. G. Davis, *Chem. Sci.*, 2015, **6**, 50-69.
93. D. L. Barnard, *Curr. Opin. Investig. Drugs*, 2001, **2**, 1530-1538.
94. P. Bailon, A. Palleroni, C. A. Schaffer, C. L. Spence, W.-J. Fung, J. E. Porter, G. K. Ehrlich, W. Pan, Z.-X. Xu, M. W. Modi, A. Farid, W. Berthold and M. Graves, *Bioconjugate Chem*, 2001, **12**, 195-202.
95. J. Bange, E. Zwick and A. Ullrich, *Nat. Med.*, 2001, **7**, 548-552.
96. D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga and L. Norton, *N. Engl. J. Med.*, 2001, **344**, 783-792.
97. S. Girish, M. Gupta, B. Wang, D. Lu, I. E. Krop, C. L. Vogel, H. A. Burris Iii, P. M. LoRusso, J. H. Yi, O. Saad, B. Tong, Y. W. Chu, S. Holden and A. Joshi, *Cancer Chemother. Pharmacol.*, 2012, **69**, 1229-1240.
98. G. D. Lewis Phillips, G. Li, D. L. Dugger, L. M. Crocker, K. L. Parsons, E. Mai, W. A. Blattler, J. M. Lambert, R. V. Chari, R. J. Lutz, W. L. Wong, F. S. Jacobson, H. Koeppen, R. H. Schwall, S. R. Kenkare-Mitra, S. D. Spencer and M. X. Sliwkowski, *Cancer Res.*, 2008, **68**, 9280-9290.
99. S. Verma, D. Miles, L. Gianni, I. E. Krop, M. Welslau, J. Baselga, M. Pegram, D. Y. Oh, V. Dieras, E. Guardino, L. Fang, M. W. Lu, S. Olsen, K. Blackwell and E. S. Group, *N. Engl. J. Med.*, 2012, **367**, 1783-1791.
100. H. Hippe, D. Caspari, K. Fiebig and G. Gottschalk, *Proc. Natl. Acad. Sci. USA*, 1979, **76**, 494-498.
101. C. M. James, T. K. Ferguson, J. F. Leykam and J. A. Krzycki, *J. Biol. Chem.*, 2001, **276**, 34252-34258.
102. G. Srinivasan, C. M. James and J. A. Krzycki, *Science*, 2002, **296**, 1459-1462.
103. B. Hao, W. Gong, T. K. Ferguson, C. M. James, J. A. Krzycki and M. K. Chan, *Science*, 2002, **296**, 1462-1466.
104. S. K. Blight, R. C. Larue, A. Mahapatra, D. G. Longstaff, E. Chang, G. Zhao, P. T. Kang, K. B. Green-Church, M. K. Chan and J. A. Krzycki, *Nature*, 2004, **431**, 333-335.
105. M. A. Gaston, L. Zhang, K. B. Green-Church and J. A. Krzycki, *Nature*, 2011, **471**, 647-650.

106. F. Quitterer, A. List, W. Eisenreich, A. Bacher and M. Groll, *Angew. Chem. Int. Ed.*, 2012, **51**, 1339-1342.
107. F. Quitterer, A. List, P. Beck, A. Bacher and M. Groll, *J. Mol. Biol.*, 2012, **424**, 270-282.
108. S. E. Cellitti, W. Ou, H. P. Chiu, J. Grunewald, D. H. Jones, X. Hao, Q. Fan, L. L. Quinn, K. Ng, A. T. Anfora, S. A. Lesley, T. Uno, A. Brock and B. H. Geierstanger, *Nat. Chem. Biol.*, 2011, **7**, 528-530.
109. F. Quitterer, P. Beck, A. Bacher and M. Groll, *Angew. Chem. Int. Ed.*, 2013, **52**, 7033-7037.
110. F. Quitterer, P. Beck, A. Bacher and M. Groll, *Angew. Chem. Int. Ed.*, 2014, **53**, 8150-8153.
111. C. Polycarpo, A. Ambrogelly, A. Berube, S. M. Winbush, J. A. McCloskey, P. F. Crain, J. L. Wood and D. Soll, *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 12450-12454.
112. J. M. Kavran, S. Gundllapalli, P. O'Donoghue, M. Englert, D. Soll and T. A. Steitz, *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 11268-11273.
113. W. T. Li, A. Mahapatra, D. G. Longstaff, J. Bechtel, G. Zhao, P. T. Kang, M. K. Chan and J. A. Krzycki, *J. Mol. Biol.*, 2009, **385**, 1156-1164.
114. B. Hao, G. Zhao, P. T. Kang, J. A. Soares, T. K. Ferguson, J. Gallucci, J. A. Krzycki and M. K. Chan, *Chem. Biol.*, 2004, **11**, 1317-1324.
115. M. L. Wong, I. A. Guzei and L. L. Kiessling, *Org. Lett.*, 2012, **14**, 1378-1381.
116. M. Marigo, T. C. Wabnitz, D. Fielenbach and K. A. Jorgensen, *Angew. Chem. Int. Ed. Engl.*, 2005, **44**, 794-797.
117. M. Y. Han, H. Z. Wang, W. K. An, J. Y. Jia, B. C. Ma, Y. Zhang and W. Wang, *Chem. Eur. J*, 2013, **19**, 8078-8081.
118. C. R. Polycarpo, S. Herring, A. Berube, J. L. Wood, D. Soll and A. Ambrogelly, *FEBS Lett.*, 2006, **580**, 6695-6700.
119. N. Wu, A. Deiters, T. A. Cropp, D. King and P. G. Schultz, *J. Am. Chem. Soc.*, 2004, **126**, 14306-14307.
120. K. Sakamoto, A. Hayashi, A. Sakamoto, D. Kiga, H. Nakayama, A. Soma, T. Kobayashi, M. Kitabatake, K. Takio, K. Saito, M. Shirouzu, I. Hirao and S. Yokoyama, *Nucleic Acids Res.*, 2002, **30**, 4692-4699.
121. L. Wang, A. Brock, B. Herberich and P. G. Schultz, *Science*, 2001, **292**, 498-500.
122. T. Yanagisawa, R. Ishii, R. Fukunaga, T. Kobayashi, K. Sakamoto and S. Yokoyama, *Chem. Biol.*, 2008, **15**, 1187-1197.
123. T. Mukai, T. Kobayashi, N. Hino, T. Yanagisawa, K. Sakamoto and S. Yokoyama, *Biochem. Biophys. Res. Commun.*, 2008, **371**, 818-822.

124. E. Saxon and C. R. Bertozzi, *Science*, 2000, **287**, 2007-2010.
125. T. S. Young, I. Ahmad, J. A. Yin and P. G. Schultz, *J. Mol. Biol.*, 2010, **395**, 361-374.
126. C. Fan, H. Xiong, N. M. Reynolds and D. Soll, *Nucleic Acids Res.*, 2015, **43**, e156.
127. A. Chatterjee, S. B. Sun, J. L. Furman, H. Xiao and P. G. Schultz, *Biochemistry*, 2013, **52**, 1828-1837.
128. H. Xiao, A. Chatterjee, S. H. Choi, K. M. Bajjuri, S. C. Sinha and P. G. Schultz, *Angew. Chem. Int. Ed.*, 2013, **52**, 14080-14083.
129. S. Cohen and E. Arbely, *ChemBioChem*, 2016, **17**, 1008-1011.
130. S. Greiss and J. W. Chin, *J. Am. Chem. Soc.*, 2011, **133**, 14196-14199.
131. P. R. Chen, D. Groff, J. Guo, W. Ou, S. Cellitti, B. H. Geierstanger and P. G. Schultz, *Angew. Chem. Int. Ed.*, 2009, **48**, 4052-4055.
132. W. Ou, T. Uno, H. P. Chiu, J. Grunewald, S. E. Cellitti, T. Crossgrove, X. Hao, Q. Fan, L. L. Quinn, P. Patterson, L. Okach, D. H. Jones, S. A. Lesley, A. Brock and B. H. Geierstanger, *Proc. Natl. Acad. Sci. USA*, 2011, **108**, 10437-10442.
133. M. J. Lajoie, A. J. Rovner, D. B. Goodman, H. R. Aerni, A. D. Haimovich, G. Kuznetsov, J. A. Mercer, H. H. Wang, P. A. Carr, J. A. Mosberg, N. Rohland, P. G. Schultz, J. M. Jacobson, J. Rinehart, G. M. Church and F. J. Isaacs, *Science*, 2013, **342**, 357-360.
134. D. B. Johnson, C. Wang, J. Xu, M. D. Schultz, R. J. Schmitz, J. R. Ecker and L. Wang, *ACS Chem. Biol.*, 2012, **7**, 1337-1344.
135. M. Shogren-Knaak, H. Ishii, J. M. Sun, M. J. Pazin, J. R. Davie and C. L. Peterson, *Science*, 2006, **311**, 844-847.
136. H. Neumann, S. Y. Peak-Chew and J. W. Chin, *Nat. Chem. Biol.*, 2008, **4**, 232-234.
137. L. T. Guo, Y. S. Wang, A. Nakamura, D. Eiler, J. M. Kavran, M. Wong, L. L. Kiessling, T. A. Steitz, P. O'Donoghue and D. Soll, *Proc. Natl. Acad. Sci. USA*, 2014, **111**, 16724-16729.
138. H. Neumann, S. M. Hancock, R. Buning, A. Routh, L. Chapman, J. Somers, T. Owen-Hughes, J. van Noort, D. Rhodes and J. W. Chin, *Mol. Cell*, 2009, **36**, 153-163.
139. I. A. Young, C. Mittal and M. A. Shogren-Knaak, *Protein Expr. Purif.*, 2016, **118**, 92-97.
140. S. J. Elsasser, R. J. Ernst, O. S. Walker and J. W. Chin, *Nat. Methods*, 2016, **13**, 158-164.
141. D. P. Nguyen, M. M. Garcia Alai, P. B. Kapadnis, H. Neumann and J. W. Chin, *J. Am. Chem. Soc.*, 2009, **131**, 14194-14195.

142. T. Yanagisawa, M. Takahashi, T. Mukai, S. Sato, M. Wakamori, M. Shirouzu, K. Sakamoto, T. Umehara and S. Yokoyama, *ChemBioChem*, 2014, **15**, 1830-1838.
143. H. W. Ai, J. W. Lee and P. G. Schultz, *Chem. Commun.*, 2010, **46**, 5506-5508.
144. D. P. Nguyen, M. M. Garcia Alai, S. Virdee and J. W. Chin, *Chem. Biol.*, 2010, **17**, 1072-1076.
145. Z. A. Wang, Y. Zeng, Y. Kurra, X. Wang, J. M. Tharp, E. C. Vatansever, W. W. Hsu, S. Dai, X. Fang and W. R. Liu, *Angew. Chem. Int. Ed.*, 2017, **56**, 212-216.
146. C. H. Kim, M. Kang, H. J. Kim, A. Chatterjee and P. G. Schultz, *Angew. Chem. Int. Ed.*, 2012, **51**, 7246-7249.
147. B. J. Wilkins, L. E. Hahn, S. Heitmuller, H. Frauendorf, O. Valerius, G. H. Braus and H. Neumann, *ACS Chem. Biol.*, 2015, **10**, 939-944.
148. M. J. Gattner, M. Vrabel and T. Carell, *Chem. Commun.*, 2013, **49**, 379-381.
149. V. Flugel, M. Vrabel and S. Schneider, *PLoS One*, 2014, **9**, e96198.
150. T. Wang, Q. Zhou, F. Li, Y. Yu, X. Yin and J. Wang, *ChemBioChem*, 2015, **16**, 1440-1442.
151. H. Xiao, W. Xuan, S. Shao, T. Liu and P. G. Schultz, *ACS Chem. Biol.*, 2015, **10**, 1599-1603.
152. C. Fu, Q. Chen, F. Zheng, L. Yang, H. Li, Q. Zhao, X. Wang, L. Wang and Q. Wang, *Angew. Chem. Int. Ed.*, 2019, **58**, 1392-1396.
153. W. W. Wang, Y. Zeng, B. Wu, A. Deiters and W. R. Liu, *ACS Chem. Biol.*, 2016, **11**, 1973-1981.
154. Z. A. Wang, Y. Kurra, X. Wang, Y. Zeng, Y. J. Lee, V. Sharma, H. Lin, S. Y. Dai and W. R. Liu, *Angew. Chem. Int. Ed.*, 2017, **56**, 1643-1647.
155. Y. S. Wang, B. Wu, Z. Wang, Y. Huang, W. Wan, W. K. Russell, P. J. Pai, Y. N. Moe, D. H. Russell and W. R. Liu, *Mol. Biosyst.*, 2010, **6**, 1557-1560.
156. J. Li, J. Yu, J. Zhao, J. Wang, S. Zheng, S. Lin, L. Chen, M. Yang, S. Jia, X. Zhang and P. R. Chen, *Nat. Chem.*, 2014, **6**, 352-361.
157. V. Hong, S. I. Presolski, C. Ma and M. G. Finn, *Angew. Chem. Int. Ed.*, 2009, **48**, 9879-9883.
158. X. Li, T. Fekner and M. K. Chan, *Chem. Asian J.*, 2010, **5**, 1765-1769.
159. D. P. Nguyen, H. Lusic, H. Neumann, P. B. Kapadnis, A. Deiters and J. W. Chin, *J. Am. Chem. Soc.*, 2009, **131**, 8720-8721.
160. Z. Hao, Y. Song, S. Lin, M. Yang, Y. Liang, J. Wang and P. R. Chen, *Chem. Commun.*, 2011, **47**, 4502-4504.
161. J. Zhang, S. Yan, Z. He, C. Ding, T. Zhai, Y. Chen, H. Li, G. Yang, X. Zhou and P. Wang, *J. Phys. Chem. Lett.*, 2018, **9**, 4679-4685.

162. T. Fekner, X. Li, M. M. Lee and M. K. Chan, *Angew. Chem. Int. Ed.*, 2009, **48**, 1633-1635.
163. Y. Yang, S. Lin, W. Lin and P. R. Chen, *ChemBioChem*, 2014, **15**, 1738-1743.
164. M. M. Lee, T. Fekner, T. H. Tang, L. Wang, A. H. Chan, P. H. Hsu, S. W. Au and M. K. Chan, *ChemBioChem*, 2013, **14**, 805-808.
165. S. Lin, H. Yan, L. Li, M. Yang, B. Peng, S. Chen, W. Li and P. R. Chen, *Angew. Chem. Int. Ed.*, 2013, **52**, 13970-13974.
166. N. J. Agard, J. A. Prescher and C. R. Bertozzi, *J. Am. Chem. Soc.*, 2004, **126**, 15046-15047.
167. J. M. Baskin, J. A. Prescher, S. T. Laughlin, N. J. Agard, P. V. Chang, I. A. Miller, A. Lo, J. A. Codelli and C. R. Bertozzi, *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 16793-16797.
168. P. V. Chang, J. A. Prescher, E. M. Sletten, J. M. Baskin, I. A. Miller, N. J. Agard, A. Lo and C. R. Bertozzi, *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 1821-1826.
169. T. Plass, S. Milles, C. Koehler, C. Schultz and E. A. Lemke, *Angew. Chem. Int. Ed.*, 2011, **50**, 3878-3881.
170. A. Borrmann, S. Milles, T. Plass, J. Dommerholt, J. M. Verkade, M. Wiessler, C. Schultz, J. C. van Hest, F. L. van Delft and E. A. Lemke, *Chembiochem*, 2012, **13**, 2094-2099.
171. M. P. VanBrunt, K. Shanebeck, Z. Caldwell, J. Johnson, P. Thompson, T. Martin, H. Dong, G. Li, H. Xu, F. D'Hooge, L. Masterson, P. Bariola, A. Tiberghien, E. Ezeadi, D. G. Williams, J. A. Hartley, P. W. Howard, K. H. Grabstein, M. A. Bowen and M. Marelli, *Bioconjug. Chem.*, 2015, **26**, 2249-2260.
172. C. Koehler, P. F. Sauter, M. Wawryszyn, G. E. Girona, K. Gupta, J. J. Landry, M. H. Fritz, K. Radic, J. E. Hoffmann, Z. A. Chen, J. Zou, P. S. Tan, B. Galik, S. Junttila, P. Stolt-Bergner, G. Pruneri, A. Gyenesei, C. Schultz, M. B. Biskup, H. Besir, V. Benes, J. Rappsilber, M. Jechlinger, J. O. Korbel, I. Berger, S. Braese and E. A. Lemke, *Nat. Methods*, 2016, **13**, 997-1000.
173. A. Kato, M. Kuratani, T. Yanagisawa, K. Ohtake, A. Hayashi, Y. Amano, K. Kimura, S. Yokoyama, K. Sakamoto and Y. Shiraishi, *Bioconjug. Chem.*, 2017, **28**, 2099-2108.
174. T. Yao, X. Zhou, C. Zhang, X. Yu, Z. Tian, L. Zhang and D. Zhou, *Molecules*, 2017, **22**.
175. B. Zhang, H. Xu, J. Chen, Y. Zheng, Y. Wu, L. Si, L. Wu, C. Zhang, G. Xia, L. Zhang and D. Zhou, *Acta Biomater.*, 2015, **19**, 100-111.

176. Y. Wu, H. Zhu, B. Zhang, F. Liu, J. Chen, Y. Wang, Y. Wang, Z. Zhang, L. Wu, L. Si, H. Xu, T. Yao, S. Xiao, Q. Xia, L. Zhang, Z. Yang and D. Zhou, *Bioconjug. Chem.*, 2016, **27**, 2460-2468.
177. N. J. Agard, J. M. Baskin, J. A. Prescher, A. Lo and C. R. Bertozzi, *ACS Chem. Biol.*, 2006, **1**, 644-648.
178. J. A. Codelli, J. M. Baskin, N. J. Agard and C. R. Bertozzi, *J. Am. Chem. Soc.*, 2008, **130**, 11486-11493.
179. X. Ning, J. Guo, M. A. Wolfert and G. J. Boons, *Angew. Chem. Int. Ed.*, 2008, **47**, 2253-2255.
180. M. L. Blackman, M. Royzen and J. M. Fox, *J. Am. Chem. Soc.*, 2008, **130**, 13518-13519.
181. N. K. Devaraj, R. Weissleder and S. A. Hilderbrand, *Bioconjug. Chem.*, 2008, **19**, 2297-2299.
182. N. K. Devaraj and R. Weissleder, *Acc. Chem. Res.*, 2011, **44**, 816-827.
183. K. Lang, L. Davis, S. Wallace, M. Mahesh, D. J. Cox, M. L. Blackman, J. M. Fox and J. W. Chin, *J. Am. Chem. Soc.*, 2012, **134**, 10317-10320.
184. T. Machida, K. Lang, L. Xue, J. W. Chin and N. Winssinger, *Bioconjug. Chem.*, 2015, **26**, 802-806.
185. E. Kaya, M. Vrabel, C. Deiml, S. Prill, V. S. Fluxa and T. Carell, *Angew. Chem. Int. Ed.*, 2012, **51**, 4466-4469.
186. A. Sachdeva, K. Wang, T. Elliott and J. W. Chin, *J. Am. Chem. Soc.*, 2014, **136**, 7785-7788.
187. S. Schneider, M. J. Gattner, M. Vrabel, V. Flugel, V. Lopez-Carrillo, S. Prill and T. Carell, *ChemBioChem*, 2013, **14**, 2114-2118.
188. K. Liu, B. Enns, B. Evans, N. Wang, X. Shang, W. Sittiwong, P. H. Dussault and J. Guo, *Chem. Commun.*, 2017, **53**, 10604-10607.
189. T. Plass, S. Milles, C. Koehler, J. Szymanski, R. Mueller, M. Wiessler, C. Schultz and E. A. Lemke, *Angew. Chem. Int. Ed.*, 2012, **51**, 4166-4170.
190. K. Lang, L. Davis, J. Torres-Kolbus, C. Chou, A. Deiters and J. W. Chin, *Nat. Chem.*, 2012, **4**, 298-304.
191. J. Li, S. Jia and P. R. Chen, *Nat. Chem. Biol.*, 2014, **10**, 1003-1005.
192. X. Fan, Y. Ge, F. Lin, Y. Yang, G. Zhang, W. S. Ngai, Z. Lin, S. Zheng, J. Wang, J. Zhao, J. Li and P. R. Chen, *Angew. Chem. Int. Ed.*, 2016, **55**, 14046-14050.
193. S. Wallace and J. W. Chin, *Chem. Sci.*, 2014, **5**, 1742-1744.
194. Y. Li, M. Pan, Y. Li, Y. Huang and Q. Guo, *Org. Biomol. Chem.*, 2013, **11**, 2624-2629.

195. J. Li, S. Lin, J. Wang, S. Jia, M. Yang, Z. Hao, X. Zhang and P. R. Chen, *J. Am. Chem. Soc.*, 2013, **135**, 7330-7338.
196. T. T. Kwan, O. Boutureira, E. C. Frye, S. J. Walsh, M. K. Gupta, S. Wallace, Y. Wu, F. Zhang, H. F. Sore, W. Galloway, J. W. Chin, M. Welch, G. J. L. Bernardes and D. R. Spring, *Chem. Sci.*, 2017, **8**, 3871-3878.
197. Y. S. Wang, W. K. Russell, Z. Wang, W. Wan, L. E. Dodd, P. J. Pai, D. H. Russell and W. R. Liu, *Mol. Biosyst.*, 2011, **7**, 714-717.
198. J. Xie, L. Wang, N. Wu, A. Brock, G. Spraggon and P. G. Schultz, *Nat. Biotechnol.*, 2004, **22**, 1297-1301.
199. K. Kodama, S. Fukuzawa, H. Nakayama, K. Sakamoto, T. Kigawa, T. Yabuki, N. Matsuda, M. Shirouzu, K. Takio, S. Yokoyama and K. Tachibana, *ChemBioChem*, 2007, **8**, 232-238.
200. C. D. Spicer, T. Triemer and B. G. Davis, *J. Am. Chem. Soc.*, 2012, **134**, 800-803.
201. K. Kodama, S. Fukuzawa, K. Sakamoto, H. Nakayama, T. Kigawa, T. Yabuki, N. Matsuda, M. Shirouzu, K. Takio, K. Tachibana and S. Yokoyama, *ChemBioChem*, 2006, **7**, 1577-1581.
202. W. Liu, A. Brock, S. Chen, S. Chen and P. G. Schultz, *Nat Methods*, 2007, **4**, 239-244.
203. Y. S. Wang, X. Fang, A. L. Wallace, B. Wu and W. R. Liu, *J. Am. Chem. Soc.*, 2012, **134**, 2950-2953.
204. J. K. Takimoto, N. Dellas, J. P. Noel and L. Wang, *ACS Chem. Biol.*, 2011, **6**, 733-743.
205. Y. S. Wang, X. Fang, H. Y. Chen, B. Wu, Z. U. Wang, C. Hilty and W. R. Liu, *ACS Chem. Biol.*, 2013, **8**, 405-415.
206. 2010.
207. A. Tuley, Y. S. Wang, X. Fang, Y. Kurra, Y. H. Rezenom and W. R. Liu, *Chem. Commun.*, 2014, **50**, 2673-2675.
208. J. M. Tharp, Y. S. Wang, Y. J. Lee, Y. Yang and W. R. Liu, *ACS Chem. Biol.*, 2014, **9**, 884-890.
209. Y. Kurra, K. A. Odoi, Y. J. Lee, Y. Yang, T. Lu, S. E. Wheeler, J. Torres-Kolbus, A. Deiters and W. R. Liu, *Bioconjug. Chem.*, 2014, **25**, 1730-1738.
210. Y. J. Lee, M. J. Schmidt, J. M. Tharp, A. Weber, A. L. Koenig, H. Zheng, J. Gao, M. L. Waters, D. Summerer and W. R. Liu, *Chem. Commun.*, 2016, **52**, 12606-12609.
211. H. Xiao, F. B. Peters, P. Y. Yang, S. Reed, J. R. Chittuluru and P. G. Schultz, *ACS Chem. Biol.*, 2014, **9**, 1092-1096.

212. J. Kahnt, B. Buchenau, F. Mahlert, M. Kruger, S. Shima and R. K. Thauer, *FEBS J.*, 2007, **274**, 4913-4921.
213. A. P. Green, T. Hayashi, P. R. Mittl and D. Hilvert, *J. Am. Chem. Soc.*, 2016, **138**, 11344-11352.
214. D. H. Jones, S. E. Cellitti, X. Hao, Q. Zhang, M. Jahnz, D. Summerer, P. G. Schultz, T. Uno and B. H. Geierstanger, *J. Biomol. NMR*, 2010, **46**, 89-100.
215. X. Li, T. Fekner, J. J. Ottesen and M. K. Chan, *Angew. Chem. Int. Ed.*, 2009, **48**, 9184-9187.
216. M. M. Lee, T. Fekner, J. Lu, B. S. Heater, E. J. Behrman, L. Zhang, P. H. Hsu and M. K. Chan, *ChemBioChem*, 2014, **15**, 1769-1772.
217. S. Virdee, P. B. Kapadnis, T. Elliott, K. Lang, J. Madrzak, D. P. Nguyen, L. Riechmann and J. W. Chin, *J. Am. Chem. Soc.*, 2011, **133**, 10708-10711.
218. I. E. Gentle, D. P. De Souza and M. Baca, *Bioconjug. Chem.*, 2004, **15**, 658-663.
219. D. P. Nguyen, T. Elliott, M. Holt, T. W. Muir and J. W. Chin, *J. Am. Chem. Soc.*, 2011, **133**, 11418-11421.
220. X. Bi, K. K. Pasunooti, A. H. Tareq, J. Takyi-Williams and C. F. Liu, *Org. Biomol. Chem.*, 2016, **14**, 5282-5285.
221. X. Bi, K. K. Pasunooti, J. Lescar and C. F. Liu, *Bioconjug. Chem.*, 2017, **28**, 325-329.
222. D. P. Nguyen, M. Mahesh, S. J. Elsasser, S. M. Hancock, C. Uttamapinant and J. W. Chin, *J. Am. Chem. Soc.*, 2014, **136**, 2240-2243.
223. R. Uprety, J. Luo, J. Liu, Y. Naro, S. Samanta and A. Deiters, *ChemBioChem*, 2014, **15**, 1793-1799.
224. M. P. Exner, T. Kuenzl, T. M. To, Z. Ouyang, S. Schwagerus, M. G. Hoesl, C. P. Hackenberger, M. C. Lensen, S. Panke and N. Budisa, *ChemBioChem*, 2017, **18**, 85-90.
225. J. Liu, F. Zheng, R. Cheng, S. Li, S. Rozovsky, Q. Wang and L. Wang, *J. Am. Chem. Soc.*, 2018, **140**, 8807-8816.
226. A. P. Welegedara, L. A. Adams, T. Huber, B. Graham and G. Otting, *Bioconjug. Chem.*, 2018, **29**, 2257-2264.
227. N. Huguenin-Dezot, D. A. Alonzo, G. W. Heberlig, M. Mahesh, D. P. Nguyen, M. H. Dornan, C. N. Boddy, T. M. Schmeing and J. W. Chin, *Nature*, 2019, **565**, 112-117.
228. W. Xuan, D. Collins, M. Koh, S. Shao, A. Yao, H. Xiao, P. Garner and P. G. Schultz, *ACS Chem. Biol.*, 2018, **13**, 578-581.
229. J. C. W. Willis and J. W. Chin, *Nat. Chem.*, 2018, **10**, 831-837.
230. V. Beránek, J. C. W. Willis and J. W. Chin, *Biochemistry*, 2019, **58**, 387-390.

231. O. El-Mahdi and O. Melnyk, *Bioconjug. Chem.*, 2013, **24**, 735-765.
232. R. A. Scheck, M. T. Dedeo, A. T. Iavarone and M. B. Francis, *J. Am. Chem. Soc.*, 2008, **130**, 11762-11770.
233. T. Sasaki, K. Kodama, H. Suzuki, S. Fukuzawa and K. Tachibana, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 4550-4553.
234. M. J. Han, D. C. Xiong and X. S. Ye, *Chem. Commun.*, 2012, **48**, 11079-11081.
235. A. Hüsken, BSc Thesis, University of York, 2016.
236. J. R. Clamp and L. Hough, *Biochem. J.*, 1965, **94**, 17-24.
237. A. Mehta, R. Jaouhari, T. J. Benson and K. T. Douglas, *Tetrahedron Lett.*, 1992, **33**, 5441-5444.
238. J. Chen, W. Zeng, R. Offord and K. Rose, *Bioconjug. Chem.*, 2003, **14**, 614-618.
239. A. Dirksen and P. E. Dawson, *Bioconjug. Chem.*, 2008, **19**, 2543-2548.
240. T. R. Branson, T. E. McAllister, J. Garcia-Hartjes, M. A. Fascione, J. F. Ross, S. L. Warriner, T. Wennekes, H. Zuilhof and W. B. Turnbull, *Angew. Chem. Int. Ed.*, 2014, **53**, 8323-8327.
241. M. Rashidian, M. M. Mahmoodi, R. Shah, J. K. Dozier, C. R. Wagner and M. D. Distefano, *Bioconjug. Chem.*, 2013, **24**, 333-342.
242. F. Tian, Y. Lu, A. Manibusan, A. Sellers, H. Tran, Y. Sun, T. Phuong, R. Barnett, B. Hehli, F. Song, M. J. DeGuzman, S. Ensari, J. K. Pinkstaff, L. M. Sullivan, S. L. Biroc, H. Cho, P. G. Schultz, J. DiJoseph, M. Dougher, D. Ma, R. Dushin, M. Leal, L. Tchistiakova, E. Feyfant, H. P. Gerber and P. Sapra, *Proc. Natl. Acad. Sci. USA*, 2014, **111**, 1766-1771.
243. S. Foillard, M. O. Rasmussen, J. Razkin, D. Boturyn and P. Dumy, *J. Org. Chem.*, 2008, **73**, 983-991.
244. R. J. Spears, R. L. Brabham, D. Budhadev, T. Keenan, S. McKenna, J. Walton, J. A. Brannigan, A. M. Brzozowski, A. J. Wilkinson, M. Plevin and M. A. Fascione, *Chem. Sci.*, 2018, **9**, 5585-5593.
245. X. Ning, R. P. Temming, J. Dommerholt, J. Guo, D. B. Ania, M. F. Debets, M. A. Wolfert, G. J. Boons and F. L. van Delft, *Angew. Chem. Int. Ed.*, 2010, **49**, 3065-3068.
246. R. P. Temming, L. Eggermont, M. B. van Eldijk, J. C. van Hest and F. L. van Delft, *Org. Biomol. Chem.*, 2013, **11**, 2772-2779.
247. V. A. Peshkov, O. P. Pereshivko and E. V. Van der Eycken, *Chem. Soc. Rev.*, 2012, **41**, 3790-3807.
248. O. Shimomura, F. H. Johnson and Y. Saiga, *J. Cell. Comp. Physiol.*, 1962, **59**, 223-239.
249. R. Y. Tsien, *Annu. Rev. Biochem.*, 1998, **67**, 509-544.

250. J. X. Yue, N. D. Holland, L. Z. Holland and D. D. Deheyn, *Sci. Rep.*, 2016, **6**, 28350.
251. F. Yang, L. G. Moss and G. N. Phillips, Jr., *Nat. Biotechnol.*, 1996, **14**, 1246-1251.
252. O. Shimomura, *J. Microsc.*, 2005, **217**, 1-15.
253. O. Shimomura, *FEBS Lett.*, 1979, **104**, 220-222.
254. A. B. Cubitt, R. Heim, S. R. Adams, A. E. Boyd, L. A. Gross and R. Y. Tsien, *Trends Biochem. Sci.*, 1995, **20**, 448-455.
255. B. D. Rao, M. D. Kemple and F. G. Prendergast, *Biophys. J.*, 1980, **32**, 630-632.
256. W. W. Ward and S. H. Bokman, *Biochemistry*, 1982, **21**, 4535-4540.
257. H. Morise, O. Shimomura, F. H. Johnson and J. Winant, *Biochemistry*, 1974, **13**, 2656-2662.
258. G. H. Patterson, S. M. Knobel, W. D. Sharif, S. R. Kain and D. W. Piston, *Biophys. J.*, 1997, **73**, 2782-2790.
259. B. P. Cormack, R. H. Valdivia and S. Falkow, *Gene*, 1996, **173**, 33-38.
260. M. Ormo, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien and S. J. Remington, *Science*, 1996, **273**, 1392-1395.
261. W. W. Ward, H. J. Prentice, A. F. Roth, C. W. Cody and S. C. Reeves, *Photochem. Photobiol.*, 1982, **35**, 803-808.
262. S. J. Remington, *Protein Sci.*, 2011, **20**, 1509-1519.
263. A. A. Heikal, S. T. Hess and W. W. Webb, *Chem. Phys.*, 2001, **274**, 37-55.
264. A. Crameri, E. A. Whitehorn, E. Tate and W. P. C. Stemmer, *Nat. Biotechnol.*, 1996, **14**, 315-319.
265. R. Battistutta, A. Negro and G. Zanotti, *Proteins: Struct., Funct., Bioinf.*, 2000, **41**, 429-437.
266. H. Fukuda, M. Arai and K. Kuwajima, *Biochemistry*, 2000, **39**, 12025-12032.
267. J.-D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger and G. S. Waldo, *Nat. Biotechnol.*, 2006, **24**, 79-88.
268. T. M. Roberts, F. Rudolf, A. Meyer, R. Pellaux, E. Whitehead, S. Panke and M. Held, *Sci. Rep.*, 2016, **6**, 28166.
269. S. J. Miyake-Stoner, C. A. Refakis, J. T. Hammill, H. Lusic, J. L. Hazen, A. Deiters and R. A. Mehl, *Biochemistry*, 2010, **49**, 1667-1677.
270. F. Sherman, J. W. Stewart and S. Tsunasawa, *Bioessays*, 1985, **3**, 27-31.
271. F. Frottin, A. Martinez, P. Peynot, S. Mitra, R. C. Holz, C. Giglione and T. Meinnel, *Mol. Cell Proteomics*, 2006, **5**, 2336-2349.

272. S. E. Cellitti, D. H. Jones, L. Lagpacan, X. Hao, Q. Zhang, H. Hu, S. M. Brittain, A. Brinker, J. Caldwell, B. Bursulaya, G. Spraggon, A. Brock, Y. Ryu, T. Uno, P. G. Schultz and B. H. Geierstanger, *J. Am. Chem. Soc.*, 2008, **130**, 9268-9281.
273. M. Fottner, A. D. Brunner, V. Bittl, D. Horn-Ghetko, A. Jussupow, V. R. I. Kaila, A. Bremm and K. Lang, *Nat. Chem. Biol.*, 2019, **15**, 276-284.
274. R. L. Brabham, R. J. Spears, J. Walton, S. Tyagi, E. A. Lemke and M. A. Fascione, *Chem. Commun.*, 2018, **54**, 1501-1504.
275. A. Dondoni and P. Merino, in *Comprehensive Heterocyclic Chemistry II*, eds. A. R. Katritzky, C. W. Rees and E. F. V. Scriven, Pergamon, Oxford, 1996, pp. 373-474.
276. J. V. Metzger, in *Comprehensive Heterocyclic Chemistry*, eds. A. R. Katritzky and C. W. Rees, Pergamon, Oxford, 1984, pp. 235-331.
277. A. Meyers and J. L. Durandetta, *J. Org. Chem.*, 1975, **40**, 2021-2025.
278. M. Jbara, S. Laps, S. K. Maity and A. Brik, *Chemistry*, 2016, **22**, 14851-14855.
279. B. J. Stenton, B. L. Oliveira, M. J. Matos, L. Sinatra and G. J. L. Bernardes, *Chem. Sci.*, 2018, **9**, 4185-4189.
280. J. Wang, S. Zheng, Y. Liu, Z. Zhang, Z. Lin, J. Li, G. Zhang, X. Wang, J. Li and P. R. Chen, *J. Am. Chem. Soc.*, 2016, **138**, 15118-15121.
281. D. L. Rodman, N. A. Carrington and Z.-L. Xue, *Talanta*, 2006, **70**, 426-431.
282. S. Franzen, M. Cerruti, D. N. Leonard and G. Duscher, *Journal of the American Chemical Society*, 2007, **129**, 15340-15346.
283. D. N. Leonard and S. Franzen, *The Journal of Physical Chemistry C*, 2009, **113**, 12706-12714.
284. J. D. Pedelacq, S. Cabantous, T. Tran, T. C. Terwilliger and G. S. Waldo, *Nat. Biotechnol.*, 2006, **24**, 79-88.
285. B. B. Wayland and R. F. Schramm, *Inorg. Chem.*, 1969, **8**, 971-976.
286. T. H. Fife, R. Natarajan, C. C. Shen and R. Bembi, *J. Am. Chem. Soc.*, 1991, **113**, 3071-3079.
287. R. J. Spears, PhD Thesis, University of York, 2018.
288. F. Brown, S. Dunstan, T. G. Halsall, E. L. Hirst and J. K. N. Jones, *Nature*, 1945, **156**, 785-786.
289. L. Li, Y. Liu, W. Wang, J. Cheng, W. Zhao and P. Wang, *Carbohydr. Res.*, 2012, **355**, 35-39.
290. M. Kurth, A. Pelegrin, K. Rose, R. E. Offord, S. Pochon, J. P. Mach and F. Buchegger, *J. Med. Chem.*, 1993, **36**, 1255-1261.
291. M. Wendeler, L. Grinberg, X. Wang, P. E. Dawson and M. Baca, *Bioconjug. Chem.*, 2014, **25**, 93-101.

292. C. S. McKay, J. A. Blake, J. Cheng, D. C. Danielson and J. P. Pezacki, *Chem. Commun.*, 2011, **47**, 10040-10042.
293. M. Colombo, S. Sommaruga, S. Mazzucchelli, L. Polito, P. Verderio, P. Galeffi, F. Corsi, P. Tortora and D. Prosperi, *Angew. Chem. Int. Ed.*, 2012, **51**, 496-499.
294. P. A. Ledin, N. Kolishetti and G.-J. Boons, *Macromolecules*, 2013, **46**, 7759-7768.
295. C. McKay, PhD Thesis, University of Ottawa, 2012.
296. F. Clow, J. D. Fraser and T. Proft, *Biotechnol Lett.*, 2008, **30**, 1603-1607.
297. D. J. Vocadlo, H. C. Hang, E.-J. Kim, J. A. Hanover and C. R. Bertozzi, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 9116.
298. Patent WO2013107820A1, 2013.
299. C. M. Haney, M. T. Loch and W. S. Horne, *Chemical Commun.*, 2011, **47**, 10915-10917.
300. S. Kasai, S. Konno, F. Ishikawa and H. Kakeya, *Chemical Commun.*, 2015, **51**, 15764-15767.
301. S. C. Gill and P. H. von Hippel, *Anal. Biochem.*, 1989, **182**, 319-326.
302. G. Jung, J. Wiehler and A. Zumbusch, *Biophys. J.*, 2005, **88**, 1932-1947.