# Native Chemical Ligation in Protein Synthesis and Semi-Synthesis

Anne C. Conibear,[†] Emma E. Watson,[†] Richard J. Payne,* Christian F.W. Becker*

[†] these authors contributed equally

* corresponding authors: richard.payne@sydney.edu.au; christian.becker@univie.ac.at

**Abstract**

Native chemical ligation (NCL) provides a highly efficient and robust means to chemoselectively link unprotected peptide and protein segments to generate proteins. The ability to incorporate non-proteinogenic amino acids (e.g. D-amino acids or fluorescent labels) and post-translational modifications into proteins by stitching together peptide fragments has driven extremely important developments in peptide and protein science over the past 20 years. Extensions of the original NCL concept (including the development of thiol- and selenol-derived amino acids and desulfurisation and deselenisation methods), improved access to peptide thioesters, and the use of the methodology in combination with recombinantly expressed polypeptide fragments (termed Expressed Protein Ligation, EPL) have helped to further expand the utility of the methodology. Over the past five years, there has been a dramatic increase in the number of proteins that have been accessed by total chemical synthesis and semi-synthesis, including a large range of modified proteins; new records have also been set with regards to the size of proteins that can now be accessed *via* ligation chemistry. Together these efforts have not only contributed to a better understanding of protein structure and function, but have also driven innovations in protein science. In this tutorial review, we aim to provide the reader with the latest developments in NCL- and EPL-based ligation technologies as well as illustrated examples of using these methods, together with synthetic logic, to access proteins and modified proteins for biological study.

Key learning points

1. Basics of native chemical ligation (NCL) and limitations of the seminal methodology in chemical protein synthesis.

2. Development of thiolated and selenated amino acids to expand the capabilities of NCL to ligation sites other than cysteine.

3. Strategies for disconnecting protein targets for synthesis by NCL; review of currently available N → C and C → N directional ligation methods with selected examples.

4. Performing NCL with expressed protein segments *via* Expressed Protein Ligation (EPL) and Protein Trans-Splicing (PTS) to access larger protein targets and segmentally isotope-labelled proteins.

5. Post-translational modifications that can be incorporated into peptides and proteins *via* NCL and EPL.

6. Current limitations and unsolved challenges together with future opportunities in the field.

**The Advent of Native Chemical Ligation**

The seminal report of the native chemical ligation (NCL) methodology in 1994 marked a significant turning point in the way in which large polypeptides and proteins could be accessed by chemical synthesis.[1] While solid-phase peptide synthesis (SPPS) laid the groundwork for routine access to peptides, typically the method is ineffective for targets more than ~40-50 amino acids in length. The ability to conjugate unprotected peptide fragments to afford full-length proteins *via* NCL has led to a paradigm shift in the way in which chemists think about targeting these large biomolecules by total chemical synthesis. While recombinant expression technologies are an undeniably important means of accessing large, unmodified proteins, these methods lack the precision and scope that chemical synthesis provides for the incorporation of additional functionalities, be these native (post-translational modifications) or unnatural and non-proteinogenic (fluorescent, purification or NMR tags, etc.) modifications. It should be noted that NCL has inspired the development of a number of other ligation protocols (including KAHA, Serine/Threonine and sugar-assisted ligations to name a few) as a means of accessing proteins. However, this tutorial review is limited to the discussion of NCL, associated ligation strategies and applications of expressed protein ligation methods.

The NCL reaction, reported by Kent and co-workers in 1994, involves the chemoselective conjugation between a pair of unprotected peptide fragments, one functionalised as a C-terminal thioester and the other with an N-terminal cysteine (Cys) residue (Figure 1).[1] The reaction is carried out under aqueous conditions, usually at neutral pH, with a chaotropic agent, such as guanidine present to inhibit the formation of secondary structural elements that would

otherwise impede the reaction. The ligation proceeds through an initial trans-thioesterification step in which the nucleophilic thiol moiety of the Cys residue attacks the carbonyl carbon of the thioester, extruding a thiol and forming a new intermolecular thioester intermediate. This intermediate then rapidly rearranges *via* an S → N acyl shift, in which the terminal amine from the thiol fragment attacks the newly formed thioester to generate a native amide bond, thus restoring the thiol side chain of Cys. Since the original report of the methodology 24 years ago, NCL has been employed in the synthesis of hundreds of protein targets that has not only facilitated a critical understanding of structure and function, but has also driven innovations in protein science. One drawback of the original methodology was the requirement of a Cys residue, the least common of the proteinogenic amino acids (1.8% abundance) for the reaction to proceed. However, this limitation was overcome through subsequent extensions to the NCL method over the past decade.

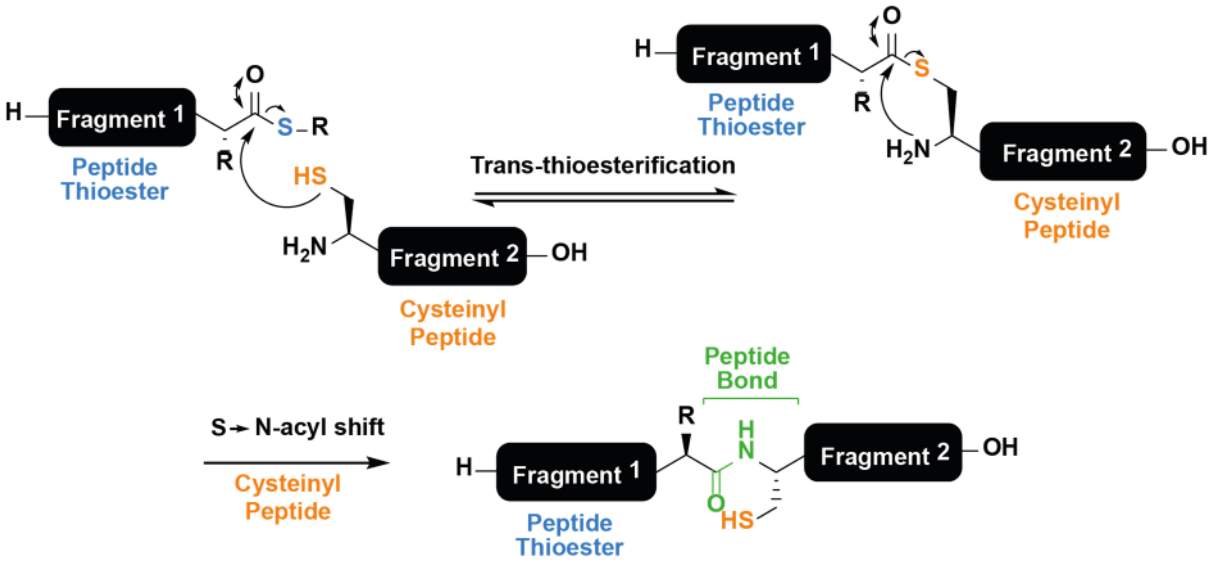



**Figure 1:** Mechanism of Native Chemical Ligation (NCL). The reaction between peptide thioester and cysteinyl peptide fragments *via* an initial trans-thioesterification followed by an S → N acyl shift to afford a native peptide (amide) bond.

## Extensions to Native Chemical Ligation

The single most significant addition to the NCL method was the development of chemistry to desulfurise the sulfhydryl side chain of the requisite Cys to the corresponding alanine (Ala) residue after completion of the ligation reaction; the high native abundance of alanine (8.9%) residues within proteins effectively increased the number of potential ligation junctions within

a target protein. Desulfurisation chemistry in combination with NCL was first reported by Dawson and co-workers in 2001, although this relied upon treatment with either Raney Ni or Pd on $Al_2O_3$ to effect the transformation.[2] While these conditions were capable of desulfurising products from NCL reactions, the large excesses of the metal catalysts required also led to undesirable side reactions with some side-chain moieties. The true utility of the NCL-desulfurisation manifold was realised in 2007 when Danishefsky and co-workers reported a modified radical desulfurisation protocol, facilitated by the radical initiator VA-044 (Figure 2).[3] This reaction is initiated by thermal decomposition of the water soluble VA-044 to generate two stable tertiary radicals, which are then proposed to effect homolysis of the thiolate bond to generate the corresponding sulfur centered radical. The thiolate radical can then trap a molecule of a phosphine additive, ultimately liberating the corresponding phosphoryl sulfide and an alanyl radical which abstracts a proton from a thiol H-atom donor (such as DTT) present in the reaction to afford a native Ala residue.

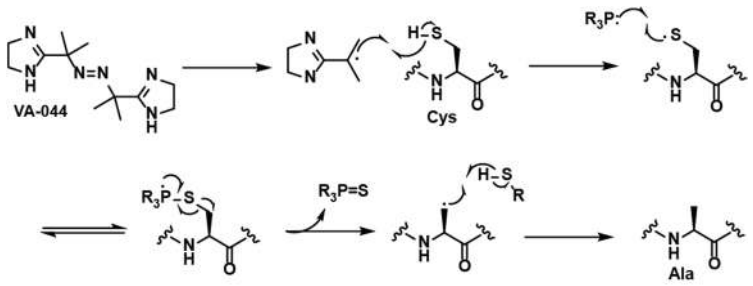



**Figure 2:** Proposed mechanism for the radical desulfurisation of Cys to Ala that can be performed following an NCL reaction.

The impact of this post-ligation desulfurisation chemistry in protein synthesis has been enormous. Specifically, the concept has inspired the design and synthesis of a number of thiolated amino acid building blocks that can be installed at the N-terminus of peptide fragments and can facilitate NCL reactions as Cys surrogates; following desulfurisation, these are converted to native amino acids, thus affording native polypeptides or proteins (Figure 3). Since the first thiolated amino acid was reported in 2006 (Phe), suitably protected variants of 15 of the 20 amino acids have been developed and used in ligation chemistry.[4] It should be noted that while each of these thiolated amino acids are competent in ligation chemistries they do possess a range of reactivities, especially in terms of the relative rates of desulfurisation. The reader is advised to investigate the specific behaviour of a candidate amino acid through

the primary literature prior to utilisation in ligation chemistry. Additionally, it should be noted that desulfurisation of larger protein targets (particularly when several sites of desulfursation are present) have been shown to be particularly sluggish, due to conformational effects from the target sequence preventing access to the appropriate reactive moieties and such factors should be considered when designing a ligation strategy. However, access to a large number of thiol-derivatised amino acids now provides the practitioner with several viable ligation strategies when disconnecting a target protein and subsequently the opportunity to circumvent some of the above issues should they arise. For comprehensive reviews of the synthesis and utility of thiolated amino acids the reader is directed to two recent and comprehensive reviews and the references therein.[4] The factors that influence and guide the selection of disconnection sites and therefore the synthetic strategy that can be used to assemble a given protein target will be discussed below.

Another key limitation of the original NCL method was the prolonged reaction times required for ligation at peptide thioesters possessing hindered amino acids on the C-terminus, such as valine and isoleucine.[5] Initial progress in accelerating the rate of the ligation reaction was achieved through the investigation of different thiol additives to enhance the reactivity of the thioester acyl donor component.[6, 7] However, an alternative approach has involved the replacement of the sulfur-based functionalities used in NCL with a selenium moiety to improve either the nucleophilicity of the Cys-containing fragment (through the inclusion of selenocysteine (Sec)) or the electrophilicity of the acyl donor fragment. For a comprehensive discussion of the utility of selenium in ligation chemistry, the reader is directed to the following reviews and references therein.[4, 8] Importantly, NCL at Sec can be coupled with chemoselective deselenisation (even in the presence of unprotected Cys residues) to afford either the corresponding alanine or, *via* a recently developed oxidative deselenisation transformation, a serine residue at the ligation junction.[4, 8] Additionally, the application of NCL and desulfurisation to flow-based systems has also shown to lead to significant rate enhancements.[9, 10] It is therefore tempting to speculate that flow chemistry could serve an important role in the efficient and automated chemical synthesis of proteins in the future.

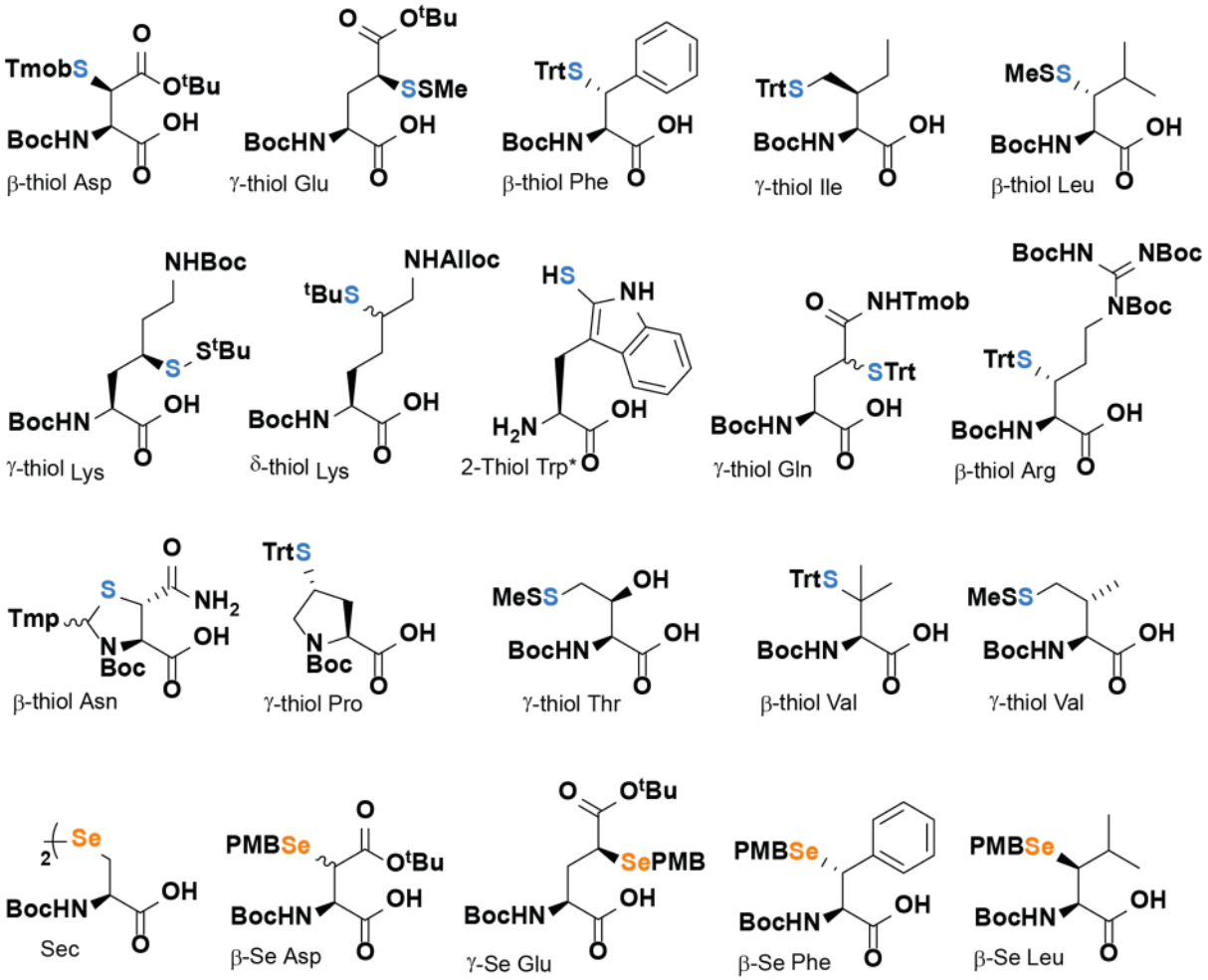



**Figure 3:** Thiolated and selenated amino acid auxiliaries for NCL. * 2-thiol moiety installed onto Trp on pre-assembled peptides. NB: All amino acids are Boc-protected as they are incorporated at the N-terminus of peptide fragments to reveal the a-amine after cleavage from solid support; two lysine derivatives have been prepared, one enables ligation on the ε-amine side chain to generate isopeptides.

Recently the use of selenocystine (the oxidised form of Sec) and an aryl selenoester has led to the development of a new ligation method dubbed the additive-free diselenide-selenoester ligation (DSL).[11] The improved reactivity of the Sec and the aryl selenoester led to an enormous improvement in ligation rates with reactions proceeding to completion in minutes even at the most sterically hindered junctions (that normally require >24 h reaction time *via* NCL). This methodology has been further extended by the development of selenol-derived amino acids that can mediate DSL or selenol-mediated NCL at other sites (Figure 3).[4] Most significantly, selenated aspartate and glutamate auxiliaries have been shown to undergo clean and

chemoselective deselenisation reactions in less than 1 minute (including in the presence of unprotected Sec residues). This rate enhancement now raises the exciting possibility of generating native proteins within minutes under a DSL-deselenisation manifold.

With such a variety of potential ligation strategies and disconnection points now available, the retrosynthetic analysis of a given protein target can be a serious undertaking. Added to this complexity is the recent development of methods that enable large peptides and proteins to be assembled in both N → C and C → N terminal directions, as well as through one-pot multi-component assembly strategies. Several recent authoritative reviews have been written in the area of NCL and this tutorial review does not seek to duplicate those efforts.[4, 12-15] Instead, this review aims to outline the major directional and multicomponent ligation strategies available as well as some guidance for selecting the most applicable strategy for a given synthetic target, including selection of sites for ligation.

**General Considerations in Choosing Ligation Junctions**

NCL and SPPS work hand in hand to facilitate the chemical synthesis of peptides and proteins. It is therefore unsurprising that the effectiveness of fragment synthesis by SPPS is a key determining factor in choosing appropriate ligation sites. Indeed, the NCL technology emerged as a result of the inability of SPPS alone to reproducibly afford peptides above a certain length threshold. Typically, the limit of SPPS is quoted as 40-50 amino acids however, this is highly dependent on whether Boc- or Fmoc-strategy SPPS is used to make a given fragment, as well as the difficulty of the sequence. It is therefore usual to choose ligation sites approximately 30-40 residues apart to provide the best opportunity of accessing the requisite fragments by SPPS in suitable purity (>97%) to be used in ligation chemistry. Despite the availability of turn inducing elements and solubility tags to improve the quality of fragments prepared by SPPS, the bottleneck in chemical protein synthesis is still the preparation of the peptide fragments on suitable scale and in suitable purity. The separation of truncated peptides generated during SPPS (usually by HPLC) becomes more difficult as sequences get longer, which must play into the choice of disconnection strategy. The type of amino acid residues found within regions of a given protein sequence should be considered as well. Shorter fragments may need to be chosen if difficult sequences (such as stretches of β-branched or hydrophobic amino acids) are present or disconnection sites may be chosen to break up difficult stretches between fragments to aid in their synthesis.

Another key factor to consider when choosing ligation junctions is the inherent chemical reactivity of the two amino acids that are being ligated and the effect this may have on the efficiency of the ensuing reaction. Traditionally, these choices were entirely dependent upon the positioning of appropriate Cys residues within the target peptide however, desulfurisation chemistry in concert with thiolated (and selenated) amino acids have largely overcome this problem.[4] Despite this, the choice of available thiolated amino acids remains an important consideration in choosing a junction as some such auxiliaries possess unique properties, which may aid or inhibit the ligation or desulfurisation processes. For example, β-thiolated aspartate has unique reactivity which enables its chemoselective desulfurisation in the presence of other unprotected thiol residues, e.g. free Cys residues, which may be an important consideration if the target sequence possesses native Cys residues.[16] Potential reactive functional groups within the target peptide should also be considered when determining applicable ligation sites as regioselectivity and chemoselectivity will ultimately be a critical determinant in the success of the synthesis.

In addition to considering the nucleophilic ligation partner, it is also important to consider the reactivity profile of the acyl donor-containing fragment. The steric and electronic properties of the acyl donor remain one of the key determining factors for the rate of NCL and DSL. Sterically hindered residues on the C-terminus of peptide thioesters and selenoesters typically react 20-fold slower in NCL and DSL, which may cause acyl donor hydrolysis to compete with productive ligation pathways, limiting the yield of the reaction.[5, 11] This effect can be mitigated by conducting the ligation at slightly lower pH and/or by increasing the equivalents of the acyl donor to account for any losses incurred through hydrolysis. Conversely, donors with unhindered or electron deficient C-terminal residues (e.g. Gly, Ala, Phe) typically undergo ligation much faster, which is important to consider if planning to utilise a kinetically controlled ligation manifold (*vide infra*).[5, 11, 17] Finally, some amino acid residues make poor acyl donors due to their proclivity to cyclise and eliminate the C-terminal activating group. Key residues of this type include Lys, Asp and Glu. The nucleophilicity of the Lys side-chain can be moderated by conducting the ligation at slightly acidic pH to prevent cyclisation, however this is not possible for the carboxylate group of the Asp and Glu residues, which require side chain protection for use as acyl donors in NCL.[12] Having assessed the potential reactivities of ligation partners, we must next consider the available NCL-based strategies for the fusion of peptide fragments to prepare a protein in the most efficient manner.

## Strategies for N → C Protein Assembly *via* NCL

### - Kinetically Controlled Ligation Strategies

One of the most powerful strategies for the assembly of multiple fragments into proteins *via* NCL chemistry is the use of a kinetically controlled ligation manifold. Such a strategy was first reported by Kent and coworkers in 2006[17] and enables the assembly of three peptide fragments in a single pot in the N → C direction to afford access to the desired peptide or protein target, usually in excellent yield. The strategy is underpinned by the ability to activate an unreactive acyl donor species *in situ* (Figure 4A). More specifically, a bifunctional fragment that contains an N-terminal Cys (or Cys surrogate) and the inert functionality on the C-terminus undergoes an initial chemoselective ligation at the N-terminus with another peptide fragment with a more reactive acyl donor. Following this first reaction, and often without intermediary purification, the inert acyl donor can be activated through the addition of an exogeneous thiol additive in order to participate in a second ligation step. Alkyl thioesters are commonly used as the inert acyl donors, as they react very slowly under an NCL manifold but readily undergo exchange (trans-thioesterfication) with thiol additives to form more reactive thioesters that are competent in subsequent ligation reactions. A variety of aryl thiol additives have been explored as additives for kinetically controlled ligation reactions, including 4-mercaptophenylacetic acid (MPAA) and thiophenol, however each of these additives suffers from the drawback that they prevent *in situ* desulfurisation due to their potent radical scavenging ability.[4, 12-14]

An advance in this area was the utilisation of the reactive alkyl thiol trifluoroethanethiol (TFET) to activate alkyl thioesters in kinetically controlled ligation reactions.[7] As an alkyl thiol, TFET has the advantage that it does not impede subsequent radical desulfurisation steps and thus facilitates more rapid protein assembly through one-pot kinetically-controlled ligation-desulfurisation manifolds. This strategy has recently been exploited for the synthesis of several proteins,[4, 12] including the anti-thrombotic mosquito protein anophelin from *Anopheles gambiae* (Figure 4B).[18] This latter example serves as a suitable exemplar to outline the synthetic strategy: Firstly the length of the anophelin protein (82 residues) and the presence of a labile post-translational modification (tyrosine sulfation) at Y12 (which would ideally be installed in a shorter fragment) led the authors to select a three-component ligation strategy. However, this was dependent upon the availability of tractable ligation junctions. While an appropriate T50-A51 junction exists (note H45-A46 was not used as His thioesters are prone to hydrolysis), disconnection in the N-terminal portion of the protein required the use of a

thiolated amino acid instead of Cys. Although a C → N ligation strategy would usually be the most appropriate for targets bearing N-terminal post-translational modifications (to protect the labile moiety and facilitate late-stage diversification), in this case it does not represent the best choice due to the requirement to use an orthogonally protected variant of the thiolated amino acid, which would require significant modifications to the synthesis of the building block. Thus, an N → C strategy, relying on differential acyl donor reactivity (rather than a C → N strategy relying on differential protection) was deemed the most appropriate for this target. More specifically, a kinetically controlled manifold mediated by thioester exchange was selected due to the mild conditions required for acyl donor activation in the presence of the labile tyrosine sulfate ester.

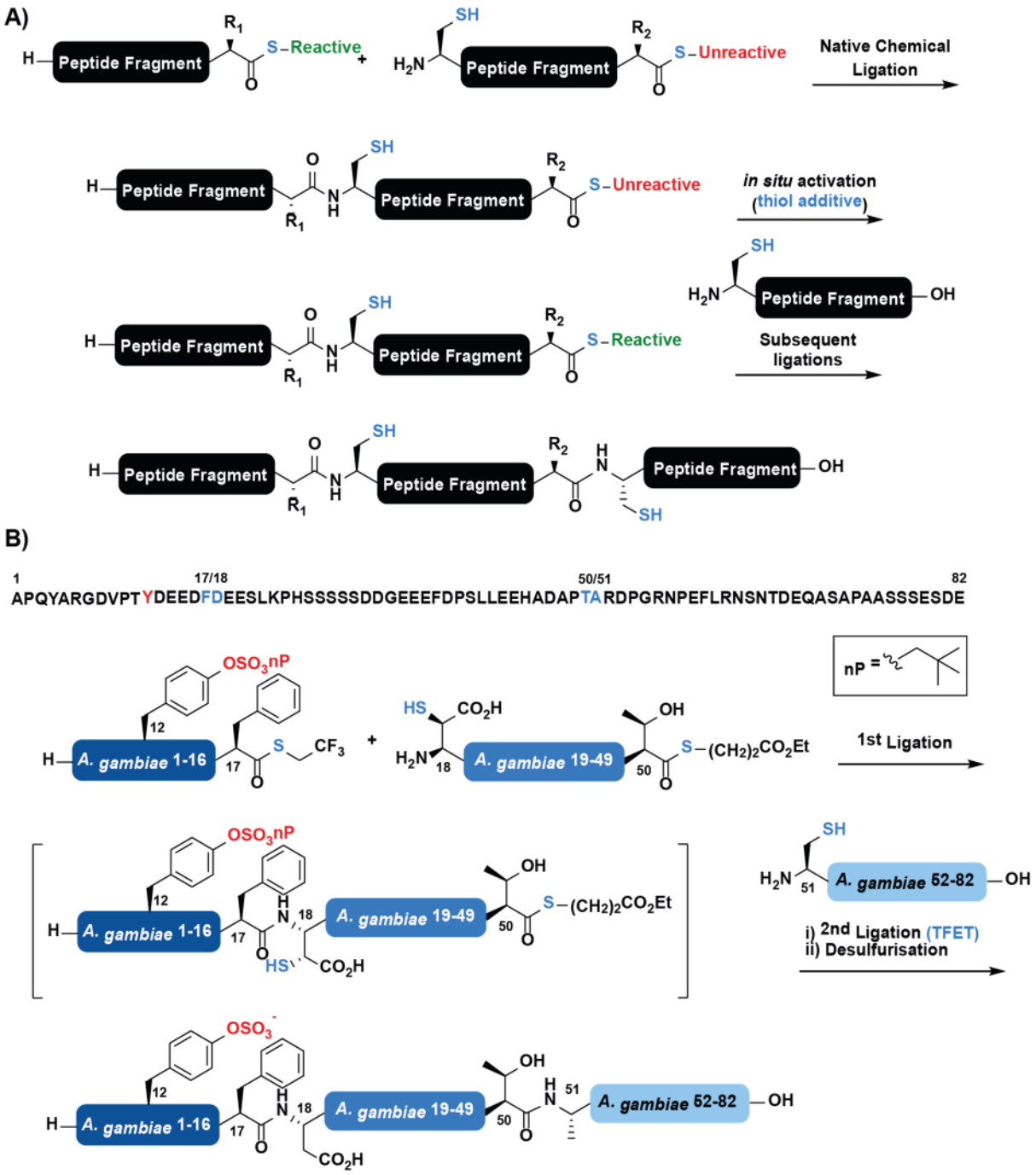



**Figure 4: A)** Schematic depicting N → C kinetically-controlled ligation; **B)** Synthesis of anticoagulant anophelin sulfoprotein from *A. gambiae via* one-pot kinetically-controlled ligation-desulfurisation in the N → C direction.

Following a decision to exploit a kinetically controlled ligation manifold, appropriate ligation junctions need to be selected that reinforce the inherent reactivities of the thioesters. Therefore, the choice of N-terminal disconnection site was selected with a view to enhancing the already improved reactivity of the TFET-thioester over the less reactive alkyl thioester that adorns the

middle bifunctional fragment. As such, in this case the F17-D18 junction was selected as the ligation junction as Phe thioesters are known to exhibit high reactivity in NCL reactions; the D18 was replaced with a thiolated Asp residue to facilitate ligation.[16] Importantly, by selecting T50-A51 as the other ligation site for the three component ligation, the regioselectivity of the protein assembly was reinforced owing to the hindered and therefore less reactive nature of the T50 thioester. Using this approach ligation between the N-terminal fragment bearing a C-terminal TFET thioester and the middle bifunctional fragment generated only the desired ligation product (Figure 4B). Supplementation of the reaction with TFET to facilitate exchange of the unreactive thioester at T50 to a TFET thioester and addition of the C-terminal fragment then led to successful assembly of the protein sequence (together with deprotection of the neopentyl ester moiety used to protect the sulfated Tyr residue). Without purification, the reaction mixture was subjected to radical desulfurisation conditions to generate the target anophelin sulfoprotein after HPLC purification. It should be noted that alternatives to TFET exist for this one-pot approach, e.g. methyl thioglycolate or 2-mercaptoethanesulfonate (MESNa),[4] although the latter additive generates relatively unreactive thioesters. It is also possible to exploit the increased reactivity of selenoesters as a replacement for the pre-activated thioester, reacting it with either Cys or Sec[11] using a similar ligation manifold.

- **Unmasking Thioester Precursors *via* N→S Acyl Shift Chemistry**

Unreactive alkyl thioesters are not the only means to introduce an unreactive moiety that can be activated *in situ* for protein synthesis in the N → C direction; a popular strategy involves exploitation of amide bonds as a latent functionality that can be unveiled to generate a reactive acyl donor. In an analogous manner to that exploited by inteins in nature, these approaches rely on an N → S acyl shift to convert an amide to the corresponding thioester which can either directly participate in a ligation or can undergo further activation by exchange with an exogenous thiol. Such an approach to generate thioesters from amides is challenged by the inherent reversibility (i.e. an S → N acyl shift) to regenerate the more stable amide bond. As such, significant effort has centred on the manipulation of this equilibrium to drive it towards thioester formation. A number of different approaches have been successfully exploited to date, including activated resin bound precursors, alkylated Cys moieties and cysteinyl prolyl esters to promote thioester formation.[12, 15] Thioester generation through N → S acyl shift chemistry has also been facilitated by specifically designed auxiliaries such as the *bis*(2-

sulfanylethyl)amido (SEA) group, which can be readily incorporated during SPPS and possesses tunable reactivity. This auxiliary was first reported by Melnyk and co-workers in 2010 and relies upon the controlled reduction of the auxiliary disulfide to the corresponding *bis*-thiol which can then undergo an N → S acyl shift to yield the desired thioester (Figure 5A).[19] The authors have also extended this technology with the creation of an additional *bis*(2-selenylethyl)amido (SeEA) auxiliary which can be activated in the presence of the SEA auxiliary, adding another level of orthogonality to the technology.

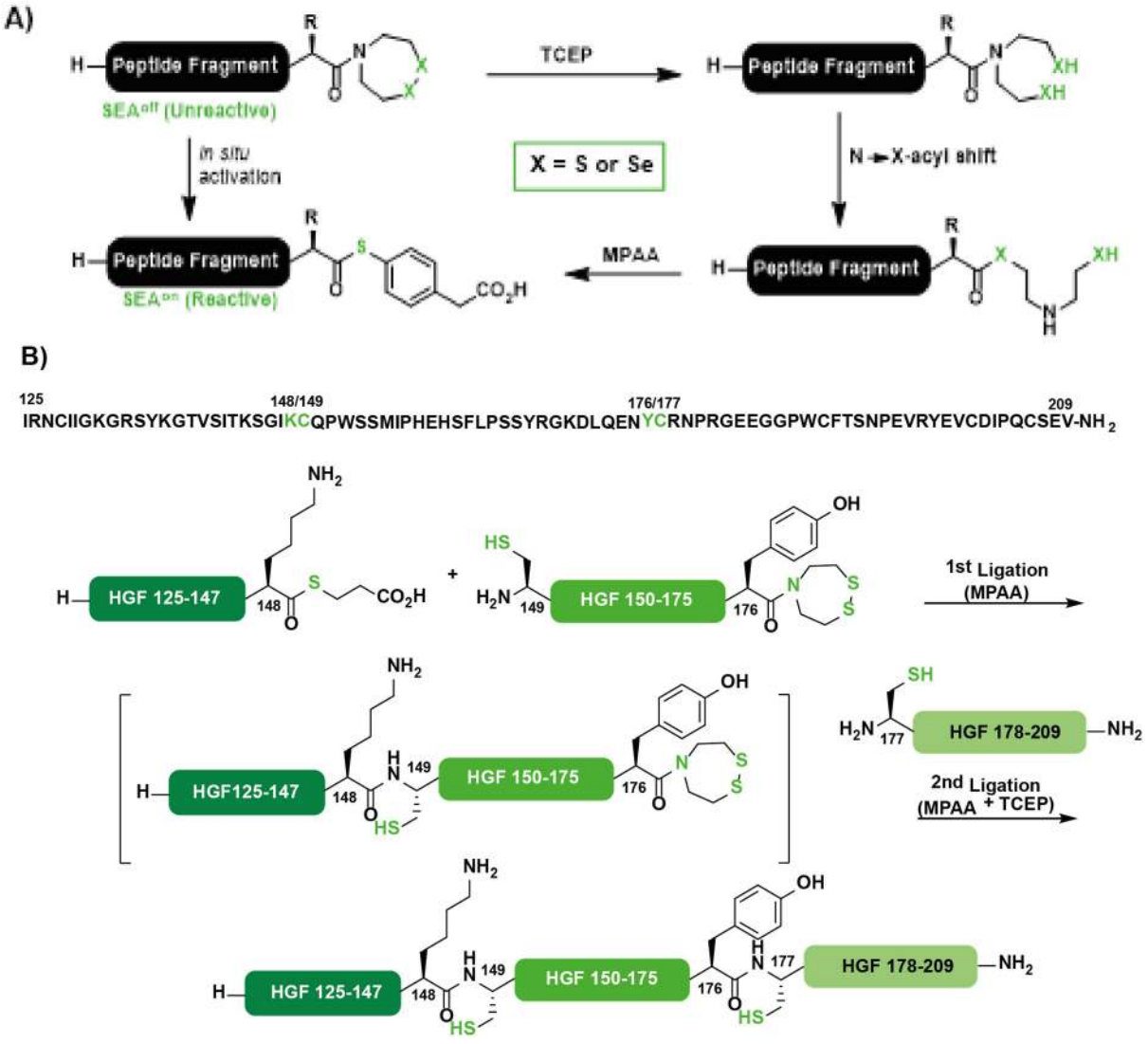



**Figure 5: A)** Schematic of the SEA auxiliary approach for the generation of thioesters; **B)** Synthesis of the K1 domain of Human Growth Factor (HGF) *via* an SEA mediated NCL strategy in the N → C direction.

The advantage of these modern auxiliary approaches is that the activation step is dependent on the addition of an exogenous phosphine reducing agent such as *tris*(2-carboxyethyl)phosphine (TCEP) to unmask the reactive thiol/selenol moieties for subsequent thioester generation, thus affording tight control over the activation of the latent acyl donor for multi-component ligation-based assembly of proteins.[15] This benefit was showcased by the one-pot, three-component assembly of the 85 amino acid Kringle 1 (K1) domain of hepatocyte growth factor (HGF) (Figure 5B).[20] Following examination of the HGF sequence, the authors selected two appropriately positioned Cys residues to serve as ligation sites (K148-C149 and Y176-C177) thus necessitating the synthesis of three suitably functionalised fragments of roughly equal length (24, 28 and 33 residues). Assembly of the HGF domain *via* the SEA based strategy relied upon an initial ligation between the N-terminal fragment and a bifunctional middle fragment, bearing a free Cys residue on the N-terminus and a C-terminal SEA auxiliary. The N-terminal fragment was synthesised as an unreactive alkyl thioester, which could be activated *in situ via* thioester exchange with exogeneous MPAA without affecting the SEA auxiliary in the oxidised (off) state on the C-terminus. Upon completion of the first ligation, the C-terminal fragment bearing a free Cys residue could be added to the reaction mixture with TCEP to facilitate reduction of the auxiliary, subsequent N → S rearrangement to the thioester, and ligation to afford the desired target protein in one pot.

- **Acyl Hydrazides as Masked Thioesters for NCL**

A variety of latent functionalities exist in addition to those of that facilitate an N → S rearrangement. This includes *N*-acyl-benzimidazolinone (Nbz) functionalised peptides first reported by Blanco-Canosa and Dawson that can be converted to C-terminal thioesters when required.[21] However, perhaps the most widely exploited alternatemasked thioester functionalities in NCL are C-terminal acyl hydrazides, first reported by Liu and co-workers in 2011.[22] The utility of acyl hydrazides arises from simple installation of the functionality on the C-terminus of peptides by prefunctionalising solid supports. This provides the practitioner the opportunity of using peptide acyl hydrazides to synthesise preformed peptide thioesters as an alternative to other SPPS methods or to generate the thioester by *in situ* activation for the execution of multi-component ligation protocols in the N → C direction.[4] Such reactivity stems from the ability to convert the acyl hydrazide into the corresponding acyl azide *via* nitrite oxidation, which can be reacted with a suitable thiol to generate an active thioester species (Figure 6A). The ability to conduct this transformation *in situ* represents a crucial step in the

extension of this technology and is elegantly demonstrated in the synthesis of α-synuclein by Zheng *et al*.[23]

α-Synuclein is a 140 amino acid protein which has been implicated in the pathogenesis of neurodegenerative disease, and therefore serves as a topical target for synthetic investigation.[23] Examination of the protein sequence by the authors revealed three appropriately positioned alanine residues that could be used as ligation junctions through the substitution with Cys residues that could later be desulfurised (A29-A30, G68-A69 and G106-A107, Figure 6B). These ligation junctions are particularly attractive as they divide the target protein into four roughly equal length fragments (29-38 residues in length) and all exploit unhindered acyl donor residues (Ala or Gly) which facilitate robust and rapid ligation reactions. Through three rounds of *in situ* oxidation with $NaNO_2$ and displacement with MPAA as a thiol additive followed by NCL, the authors were able to obtain the 140-mer in good overall yield of 11% after a final desulfurisation reaction that simultaneously converted three Cys residues to three Ala residues at the ligation junctions. This protocol demonstrates the utility of the acyl hydrazide strategy for the generation of large peptides and proteins *via* chemical synthesis through iterative N → C ligations.

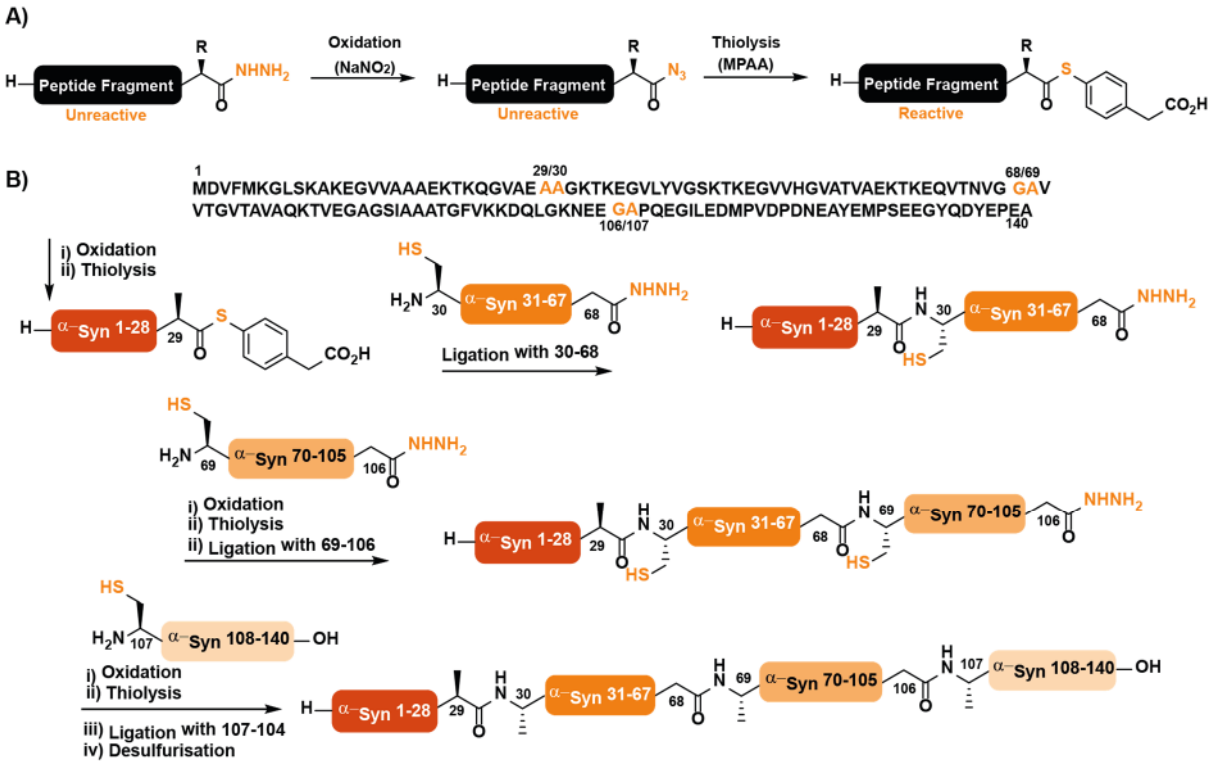



**Figure 6: A)** Activation of acyl hydrazides through oxidation and thiolysis steps; **B)** Synthesis of α-Synuclein *via* an acyl hydrazide mediated ligation strategy in the N → C direction.

## Strategies for C → N Protein Assembly *via* NCL

- **Acm as a Protecting Group for NCL**

Chemical synthesis of peptides and proteins *via* NCL can also be accomplished in a C → N direction, with this strategy reliant on the use of suitable protecting groups for N-terminal thiol moieties to facilitate controlled assembly of multiple peptide fragments. C → N ligation strategies are particularly attractive for peptides and proteins bearing N-terminal modifications such as PTMs or functionalities that may affect the solubility or chromatography of the target proteins, as it enables the late stage installation of these moieties.

In considering potential thiol protecting groups for such an application one must identify functionalities which can be installed and remain intact throughout standard NCL manipulations, but can be removed in a chemoselective manner when desired. One of the early examples of such a protecting group is the acetamidomethyl (Acm) protecting group that is commonly employed for the protection of N-terminal Cys residues during NCL,[24] and also enables chemoselective desulfurisation at other unprotected Cys residues within a given sequence (Figure 7A).[4, 12-14]

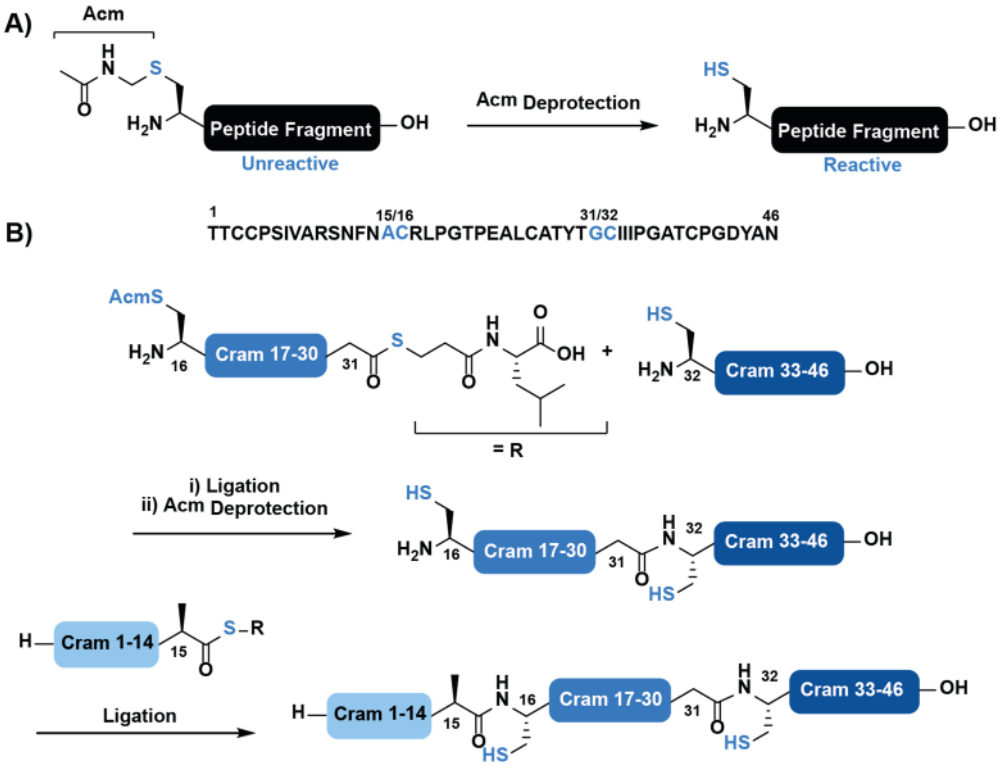


**Figure 7: A)** Acm deprotection; **B)** Synthesis of Crambin[V15A] *via* a C → N NCL strategy using Acm protection of Cys.

An early example of the utility of this protecting group is the synthesis of the model peptide Crambin[V15A] by Kent and co-workers (Figure 7B). The synthesis of Crambin[V15A] has become something of a methodological benchmark in ligation chemistry, with several syntheses reported, making for interesting synthetic comparison. On examination of the peptide sequence there are several Cys residues present that could, in theory, be used as ligation sites, however, the most central of these was not utilised in this work. The authors chose not to exploit the L25-C26 junction found in this 46-mer, even though it would have yielded two fragments of approximately equal length, due to the sterically encumbered nature of the requisite donor which would have made for a difficult ligation. It should be noted that with more recent advances to ligation methodology (such as the use of preformed selenoesters) this junction would now be routinely accessible and highlights the impact of contemporary technologies on the field. In lieu of this hindered ligation site, the authors instead considered two other potential junctions (A15-C16 and G31-C32), which each yield two requisite fragments of differing lengths (15 and 31 residues in each case). Interestingly, the authors chose to compare the efficacy of ligation at each of these junctions and discovered that access to even the moderately long 31-mers proved challenging, inspiring them to access the protein *via* a three-component ligation approach. This serves as an important example in the important relationship between ligations and the synthesis of their requisite fragments. When assessing ligation junctions it is always important to consider the synthesis of the fragments required to effect the final ligation and that the best-looking ligation junction may not be the most effective overall if the SPPS required to reach the desired fragments is problematic. In order to achieve the three-component assembly of Crambin[V15A], the authors chose a C → N ligation strategy using Acm protection of C16. This altered strategy significantly improved access to the target fragments and enabled their assembly by first ligating the C-terminal fragment with a bifunctional middle fragment bearing an alkyl thioester (which was activated *in situ* with thiophenol) followed by deprotection of the Acm group using $Ag(OAc)_2$. The purified Cram16-46 fragment could then be submitted to a second NCL reaction with the Cram1-15 thioester to afford the full-length protein. A key drawback to the early use of the Acm protecting group were difficulties in removing the protecting group following the ligation and the harsh reaction conditions required. Modern adaptations whereby the Acm group can be rapidly removed by palladium

catalysis under milder conditions[25] have reinvigorated the use of this protecting group (and its derivatives, e.g. trifluoroacetamidyl methyl (Tfacm)) in NCL.

- **Thiazolidine and Selenazolidine Protecting Groups for NCL**

In addition to the Acm group, several alternatives exist for N-terminal Cys protection to facilitate the C → N based assembly of proteins *via* NCL. Amongst the most widely adopted has been the thiazolidine (Thz) protecting group in which the Cys thiol and the N-terminal amine are condensed with formaldehyde to effect protection (Figure 8A). This functionality, like the Acm group, is stable to standard ligation conditions, but can be readily opened by nucleophiles such as $MeONH_2$ to facilitate subsequent ligations in the same pot and, as such, has found tremendous utility in peptide and protein synthesis.[4, 12-15]

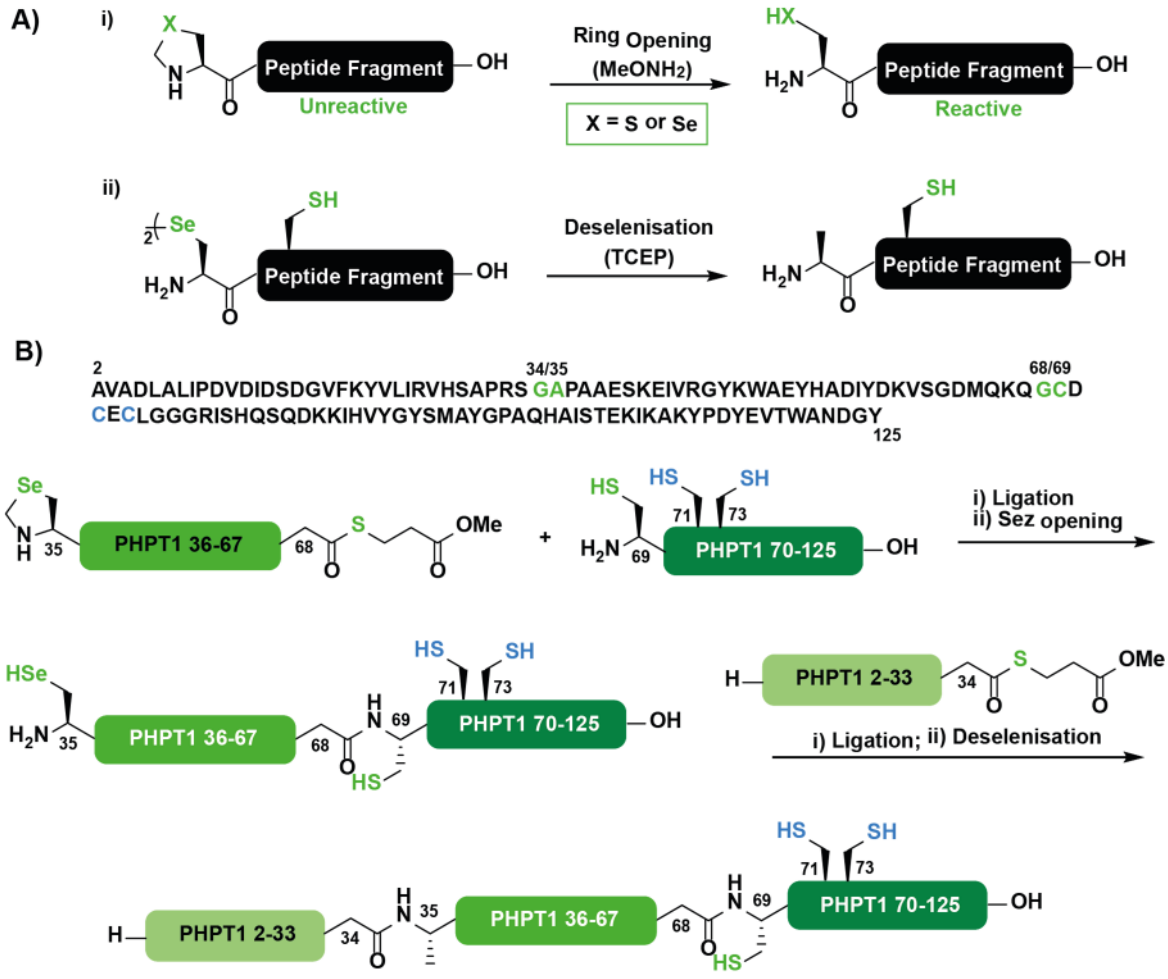


**Figure 8: A)** i) Thz and Sez deprotection, (ii) chemoselective deselenisation in the presence of unprotected Cys residues; **B)** Synthesis of PHPT1 *via* a C → N ligation strategy employing key selenazolidine and deselenisation steps.

A selenated version of this functionality has also been developed and was elegantly exploited by Metanis and co-workers in the synthesis of phosphohistidine phosphatase 1 (PHPT1).[26] Examination of the sequence of PHPT1 highlights an immediate problem that there are Cys residues present only in the C-terminal portion of the 124 residue protein, leaving a 68 residue N-terminal segment without an appropriate Cys for ligation (Figure 8B). Of course this problem can be overcome through the use of a thiolated auxiliary but this also creates another potential problem of how to remove this auxiliary without desulfurizing the native Cys residues. While orthogonal protection could facilitate this, Metanis and coworkers took a far more elegant approach, exploiting the chemoselective deselenisation of Sec in the presence of Cys. With this strategy in mind, the authors chose to utilise the most N-terminal of the available Cys residues (G68-C69) as one of the ligation junctions and split the remaining N-terminal portion roughly in half through a G34-A35 junction, where the Ala residue would be mutated to Sec that could be deselenised following the ligation event to restore the native sequence. With the requisite fragments in hand, the authors were then able to assemble PHPT1 *via* an initial NCL between the C-terminal fragment and the bifunctional middle fragment bearing the N-terminal Sez. The initial ligation was facilitated by MPAA as a thiol auxiliary and afforded the C-terminal portion of the target protein within 5 h, followed by opening of the selenazolidine moiety with $MeONH_2$. NCL between a thioester and a selenol is notoriously slow, even in the presence of thiol additives such as MPAA, as both are prone to oxidation by atmospheric oxygen. This can be avoided by the addition of a reducing agent such as TCEP, however this can also cause deselenisation prior to the ligation. The authors managed to circumvent this by the addition of sodium ascorbate as a radical quencher in addition to the TCEP to prevent deselenization and facilitate the desired ligation. Following purification, chemoselective deselenisation of the Sec residue afforded the native sequence without desulfurisation of any of the Cys residues.

As demonstrated above, numerous strategies exist for the synthesis of peptides and proteins *via* an NCL manifold, either in the N → C or C → N direction as well as convergent assembly of larger fragments generated by these strategies. Significant advances have been made since the

initial report of NCL, which have opened up previously inaccessible ligation junctions and thereby facilitated access to a broader range of proteins. Through improved technologies, multicomponent and iterative ligation strategies have allowed the synthesis of proteins of increasing length, the longest to date being the 358-residue all D-amino acid Dpo4 enzyme, which was accessed through a combination of (modified)Thz, Acm and acyl hydrazide based strategies.[27] While access to proteins of this length is feasible, it requires tremendous synthetic effort and therefore total chemical synthesis is not always the method of choice for accessing larger protein targets. Techniques have therefore been developed which enable the fusion of expressed proteins with chemically synthesised peptide fragments (which often include modifications inaccessible *via* expression techniques) via an NCL manifold. These approaches have been termed Expressed Protein Ligation (EPL) and Protein Trans Splicing (PTS) and will be the second focus of discussion for this tutorial review.

**Extending Chemical Protein Synthesis to Semi-synthesis by use of Expressed Proteins**

Semi-synthesis of proteins *via* Expressed Protein Ligation (EPL) and Protein Trans-Splicing (PTS) is an area that has developed tremendously over the past 20 years. Herein, we provide a brief general background and then highlight several recent developments that already have or could have significant impact in the areas of production of homogeneous, modified semi-synthetic proteins for their application in structural and cell biology as well as in medicinally relevant areas, such as for generating therapeutic proteins. As navigating the world of inteins and selecting the most appropriate types for a given application is challenging, we will briefly discuss the advantages and disadvantages of several intein and split intein systems.

More than 20 years ago the first example of protein semi-synthesis was described, in which fragments of cytochrome c spontaneously re-fused with synthetic peptides after cleavage with cyanogen bromide (CNBr).[28] Even more successful were approaches relying on reverse proteolysis in which the amide bond cleavage reaction typically performed by proteases was driven backwards (mainly in organic solvents) to provide target protein and/or peptide segments. Under these conditions aminolysis of an acyl-enzyme intermediate is preferred over hydrolysis and two polypeptide chains are connected through a native amide bond. Significant improvements to this approach have been achieved *via* genetic engineering of protease active sites, as demonstrated with the synthesis of modified methionyl-extended human growth hormone using subtiligase.[29] However, it should be noted that the use of N-terminal segments

with suitable C-terminal leaving groups, typically esters, limits this approach to N-terminal modification of larger proteins.

As we will see, combining strategies originating from purely chemical synthesis with recombinant expression methods can lead to a number of advantages in the types of biomolecules that can be accessed. This was demonstrated early on by selectively functionalising biologically produced proteins using established chemistry such as periodate oxidation leading to N-terminal aldehydes that selectively react with hydrazine- or aminooxy-containing moieties. However, a drawback of this approach, at least from a biological standpoint was the non-native bonds generated within the protein backbone using this chemistry. Even though there are several examples of synthetic and semi-synthetic proteins in which these non-native structures do not interfere with structure and function, strategies providing native amide bonds at the ligation site have been more readily accepted and taken up by the community.[14] To fit these requirements and to overcome a long-standing challenge in the chemical synthesis of larger polypeptide chains, new chemical ligation strategies have been developed that allow unprotected peptide segments to be linked in aqueous solvents. The use of aqueous solvents solved the historic challenges related to low solubility encountered by peptide chemists when trying to link protected peptides by condensation reactions. We have discussed the most widely used method – NCL – in the prior section, with a specific focus on its use in generating homogeneous, modified proteins *via* several consecutive ligation steps. Despite the ability to combine multiple ligation steps in a bi-directional manner (*vide supra*), the number of NCL reaction steps that can be combined and still generate sufficient quantities of the target protein is limited. Indeed, despite decades of effort the largest protein that has been produced to date (358 residues) is still significantly smaller than the average size of human proteins (ca. 480 residues). In order to overcome the challenges of assembling larger proteins by NCL-based chemical protein synthesis, it was recognised early on that expressed proteins can be generated with an N-terminal cysteine residue and are therefore directly compatible with the NCL reaction manifold. A semi-synthetic version of the transcription factor protein NFAT-1 was obtained this way and impressively showcased the complementarity of chemical synthesis, which can flexibly generate peptides bearing modified and/or non-proteinogenic amino acids, and protein expression, which has no inherent limitations with respect to the size of proteins generated.[30] The discovery and detailed study of protein splicing, an autocatalytic process that excises a polypeptide (the intein, derived from internal or intervening protein) from a given ribosomally produced protein and joins the N- and C-terminally flanking polypeptides

(the exteins, derived from external proteins) together *via* an amide bond (Figure 10A), provided another valuable link between chemistry and biology, enabling expressed proteins with C-terminal α-thioesters to be obtained.[31] These tools transitioned the NCL method into the realm of protein semi-synthesis in a process termed Expressed Protein Ligation (EPL, Figure 9B).[32] This process was subsequently extended to Protein Trans-Splicing (PTS, Figure 9C), a process whereby two discontinuous protein domains, termed split intein segments, associate to form functional intein units. Figure 9 provides an overview of the reaction sequences applied in EPL and PTS approaches. Selected applications of these methods to the generation of target proteins are also outlined below.

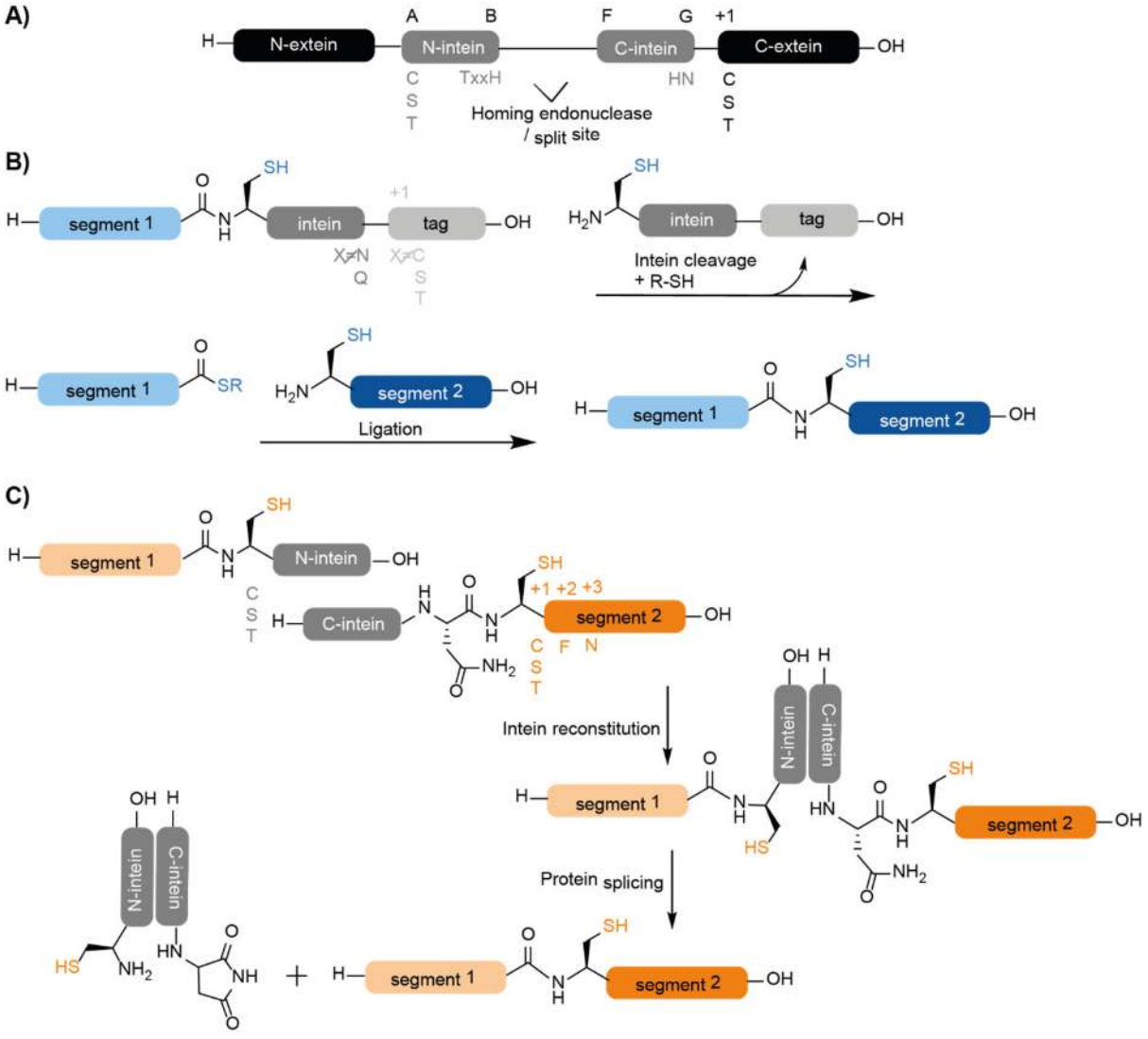



**Figure 9:** Inteins for protein semi-synthesis. **A)** General overview of an extein-intein fusion construct with critical residues for intein function shown below; **B)** Expressed protein ligation reaction with an intein fusion construct that cannot undergo the complete splicing process due to mutation within the intein and C-extein sequence shown in grey. The N-terminal cleavage

reactions generates the desired protein-thioester for subsequent NCL with a peptide/protein segment carrying an N-terminal cysteine residues; **C)** Protein trans splicing bringing the two split intein domains, fused to the respective N- and C-terminal exteins, together for a complete splicing reaction.

**Inteins for Expressed Protein Ligation (EPL)**

Inteins occur throughout the phylogenetic tree and typically contain a Hint (hedgehog/intein) domain that is designed to break and make peptide bonds in a controlled fashion by bringing the cleavage and ligation junction into close proximity.[33] Additional functional domains can be inserted within the Hint domain, e.g. homing endonucleases and DNA binding domains. These domains are selfish genetic elements helping to spread the intein throughout the genome but are not relevant for intein function and when removed can lead to artificially split inteins (see below).[34] It has become increasingly obvious that inteins are used in posttranslational regulation by being dependent on certain environmental conditions, such as redox status or temperature. As protein splicing is a highly precise process, the relevant amino acid residues are arranged in four splicing motifs, generally found in all inteins (Figure 9A). The N-terminal motif, also named block A or N1 contains the nucleophilic amino acid (Cys, Ser or Thr) for the initial attack on the amide bond between extein and intein, block B (N2) contains two highly conserved residues (Thr and His), whereas the C-terminal motifs F (C2) and G (C1) contain several more conserved residues, e.g. a C-terminal asparagine often preceded by a histidine (Figure 9A). Detailed investigations of the splicing mechanism have revealed varying pathways, however, these will not be discussed further in this review.[34]

Some of the alternative splicing pathways of inteins are also of interest for protein engineering applications, for example if no N-terminal nucleophile is required on the intein, and instead the N-terminal nucleophile of the C-extein or an internal nucleophile of the intein is used. In a similar context, bacterial intein-like (BIL) proteins that can undergo different cleavage reactions between the intein and exteins could be exploited for specific protein engineering and semi-synthesis reactions (Figure 10). Fortunately for protein semi-synthesis, inteins are very flexible in accepting different C-terminal sequences of the N-extein, however they always require an amino acid residue with a nucleophilic sidechain at the N-terminus of the C-extein to undergo splicing or cleavage. This flexibility within the N-extein sequence paved the way for the initial use of inteins as cleavable purification tags and shortly after as tools to obtain

protein α-thioesters that interface so elegantly with NCL.[32] The latter is easily achieved by enforcing off-pathway reactions of inteins that can lead to N- or C-terminal cleavage (Figure 10B-D). If the critical residues on the intein C-terminus and the nucleophile of the C-extein are mutated to unreactive amino acids (in many cases to Ala), only the initial thioester forming step takes place and will lead to N-terminal cleavage over time. If this cleavage reaction is enforced by addition of highly water-soluble small molecules containing sufficiently nucleophilic thiol functionalities, such as the sodium salt of mercaptoethane sulfonate (MESNa), the N-extein is cleaved off carrying a C-terminal α-thioester. This modification is stable at neutral pH and the resulting protein α-thioester (much like their peptide counterparts) can be stored and handled under these conditions for extended periods.[14] It should be noted however that due to the fact that inteins are single turnover enzymes, they are highly optimised with respect to the order of reaction but not necessarily for kinetics, which can slow down processing. Our current understanding of the conformational changes in inteins upon splicing is based on a few well-studied examples. While the lack of consensus between these studies has hampered the development of more efficient inteins by rational design, generally speaking, asparagine cyclisation is the limiting step of this process, which is only induced if a branched intermediate is formed to prevent premature C-terminal cleavage (Figure 10C).[35]

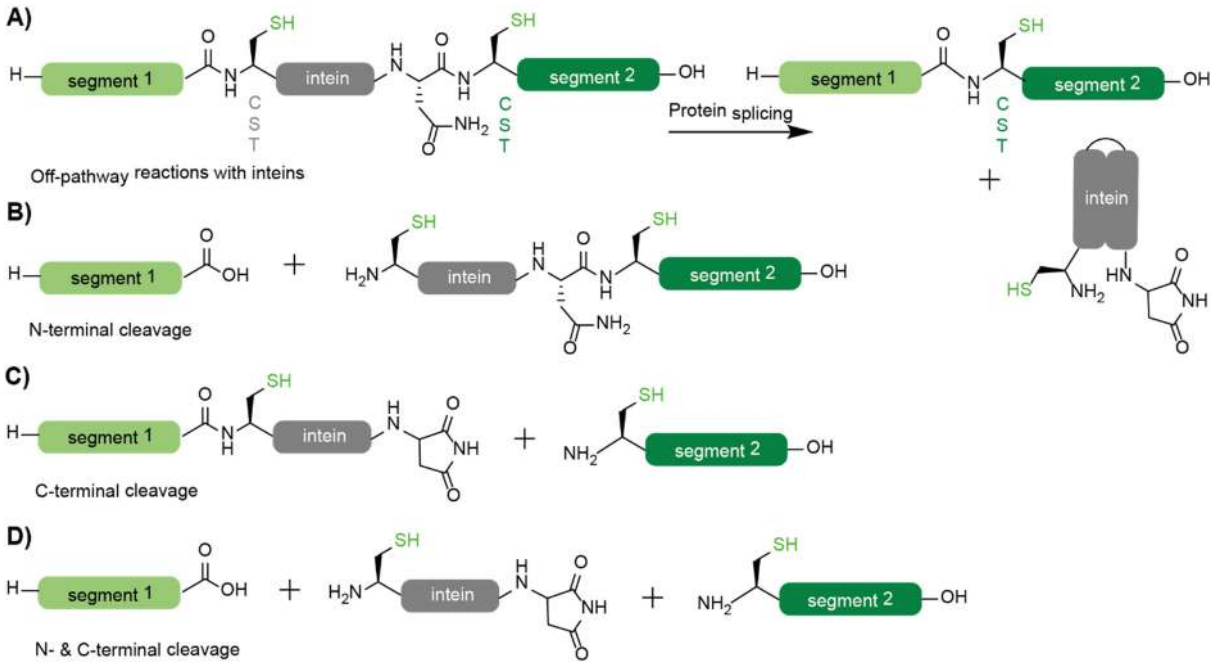



**Figure 10:** Protein splicing and side reaction. **A)** General protein splicing reaction; **B)** N-terminal cleavage and hydrolysis generating an N-terminal protein segment and the intein linked to the C-terminal extein; **C)** C-terminal cleavage giving access to the C-terminal extein

with an N-terminal cysteine residue and a fusion construct comprising the N-terminal extein and the intein with a C-terminal succinimide; **D)** N- and C-terminal cleavage dividing the fusion construct in three segments.

To facilitate the purification of a given extein-intein fusion protein, a number of different purification tags have been C-terminally linked to inteins, with the chitin-binding domain (CBD) and histidine-tag (His-tag) being the most frequently employed (Figure 9A). A marked difference between these two purification tags is that the CBD needs to be folded to recognise chitin beads, but that the His-tag allows purification of denatured proteins. The latter property is often useful as many extein-intein fusion proteins that are designed to generate protein α-thioesters for protein engineering are expressed in insoluble inclusion bodies.[36] If this is the case, solubilizing the fusion protein using high concentrations of denaturants such as 6 M Gdn-HCl or 8 M urea in aqueous buffers also unfolds the intein, rendering it splice incompetent. At least partial refolding of the intein is required to cleave off the desired protein thioester successfully, as demonstrated early on for the GyrA intein from *Mycobacterium xenopi* (*Mxe*), which leads to sufficient splicing activity at 4 M urea concentrations and below. This robust refolding behaviour makes the *Mxe* GyrA intein a highly useful tool in protein semi-synthesis even though several inteins and split inteins with similar or faster cleavage kinetics have been described over the past ten years. Our ability to generate protein and peptide thioesters by chemical synthesis and by intein cleavage has undergone dramatic improvements over the past two decades, however it can still be advantageous if the isolation of such thioester species can be avoided. One way to achieve this is the use of protein trans splicing (PTS), in which the functional intein generates thio- (or oxo-) esters and can be followed with an *in situ* ligation reaction. Thereby all pitfalls related to their generation and handling can be avoided.

**Split Inteins for Protein Trans Splicing (PTS)**

The majority of naturally occurring inteins exists as continuous polypeptide chains embedded between the N- and C-exteins as described above. However, several inteins have been found as discontinuous domains (by comparison on the genetic level) expressed from independent genes that spontaneously assemble with high affinities ($K_D$ values in the nM range) via a "capture and collapse" mechanism that leads to fast splicing intein complexes.[37] The best characterised split inteins for PTS so far stem from cyanobacteria such as *Nostoc punctiforme* (*Npu*) and *Synechocystis sp* PCC6803 (*Ssp*), in which they assemble the catalytically active subunit of polymerase III (DnaE). In contrast to contiguous inteins, naturally occurring split

inteins such as the *Npu* and *Ssp* DnaE inteins exhibit extremely fast splicing reactions upon assembly, thereby allowing modified, functional proteins to be generated from split precursors at high speed and very low concentrations; these two points are not typically achievable when carrying out EPL with peptide/protein α-thioesters generated from synthesis or intein cleavage. A major drawback of these inteins however, is their dependence on the N-terminal sequence of the C-extein (positions +2 and +3 in Figure 9C), in addition to the necessary nucleophilic residues in +1 positions of inteins. Recent advances in engineering split inteins have yielded even faster splicing split inteins but also variants that are more flexible with respect to the flanking sequences of the C-extein, especially the large hydrophobic residue in the +2 position in the *Npu* DnaE and artificial split inteins such as the Cfa variant (Figure 9A).[38] For protein semi-synthesis, a specific feature of naturally occurring split inteins has been tremendously helpful, namely the fact that the C-terminal segment of the split intein is typically short (~36-50 amino acids) and therefore accessible by SPPS. This allows for chemical manipulation and attachment of synthetic C-terminal extein sequences carrying posttranslational modifications (PTMs). Taken together, the availability of very fast split inteins and variants with increased amino acid tolerance in the +2/+3 position provides new opportunities for the fast assembly of semi-synthetic proteins with modified C-termini at low concentrations.

Whereas split inteins with short C-terminal segments exist, and have been optimised over the past 15 years, a general lack of their counterparts with short, chemically accessible N-terminal segments became obvious. To remedy this situation and to allow protein assembly with synthetic N-terminal segments, artificially split inteins such as the *Ssp* DnaB intein with an 11 amino acid N-terminal segment were generated and used for N- terminal protein modification.[39] Interestingly, the *Ssp* DnaB split intein is also one of the few artificial split inteins that does not require a denaturation step to allow functional assembly of the intein complex. Artificially split inteins have been created in the past mostly by splitting the intein sequence in a position where the homing endonuclease domain was inserted (Figure 9A). This has led to split inteins that, in contrast to the naturally split variants described above, either require a denaturation step before assembly of a functional intein complex, or the addition of an external trigger to bring the two intein domains together. The latter type of inteins has been successfully employed for different variations of conditional protein splicing *in vitro* and *in vivo*.[37] *In vivo* applications of split inteins for the assembly of semi-synthetic proteins in cells provide a unique tool that has been modified to be dependent on different external triggers such as small molecules, light, temperature, pH or redox status.[37]

## Applications of EPL and PTS for Generating Semi-synthetic Modified Proteins

Recent developments in methods for generating C-terminal thioesters and N-terminal Cys residues or their respective surrogates from both recombinant and synthetic protein segments have facilitated the semi-synthesis of numerous proteins for a variety of applications that are illustrated below with select recent examples. Protein semi-synthesis is ideally suited to the study of protein posttranslational modifications (PTMs), which can be introduced site-selectively on a synthetic protein segment and ligated to a recombinantly expressed segment. Here we review examples in which PTMs are installed at the N-terminus, C-terminus and in the middle of proteins, to illustrate how the techniques described in the previous sections can be applied to the synthesis of homogeneous posttranslationally modified proteins, including cyclic proteins. Unnatural modifications can also be installed in proteins using EPL and PTS techniques and we give selected recent examples of how NCL has been used to attach chemical handles, labels and tags to recombinant proteins. A further application of EPL and PTS is for segmental labelling of proteins for structural studies using nuclear magnetic resonance (NMR) and scattering techniques and we give examples of how this has provided insights into protein structures and interactions. Finally, we highlight exciting recent developments in extending EPL and PTS to new areas, illustrating the potential of this tool for pushing the frontiers of chemical biology.

## Semi-synthesis of PTM Proteins using EPL

A vast array of PTMs regulate gene transcription, signal transduction and a host of other core cellular processes in a complex interplay that we are only just beginning to appreciate. Access to proteins involved in these processes has been hindered by the difficulties of generating homogeneously modified proteins, especially using bacterial recombinant expression systems. Although techniques for introducing site-specific PTMs in recombinant systems have made significant progress, for example using amber codon suppression, site-directed mutagenesis and eukaryotic expression systems, several challenges still remain. In contrast, the relative ease with which multiple different modifications can be installed site-selectively by SPPS provides access to a wealth of modified peptide fragments that can be ligated to recombinant unmodified protein fragments produced in bacteria or other expression systems. The resulting homogeneously modified proteins can then be used to probe the roles of individual PTMs and combinations of PTMs on the properties and interactions of the proteins. Widely occurring PTMs such as phosphorylation, glycosylation, acetylation, methylation, ubiquitylation and

lipidation have been the main focus of protein semi-synthesis studies, however rarer PTMs have also been incorporated and studied using this method. Many of the proteins generated and studied to date fall into the categories of histone proteins and proteins involved in neurodegeneration; we will highlight examples of these and other classes of modified proteins bearing PTMs at the N- or C-termini, as well as in the middle of the sequence.

Intein-mediated protein cleavage (as described earlier) provides access to protein α-thioesters that can be ligated to synthetic peptides bearing PTMs as well as an N-terminal Cys residue. In a recent example, this method was applied to the semi-synthesis of lipidated microtubule-associated light chain 3 (LC3) proteins, which are involved in autophagy processes and responses to pathogens.[40] The recombinant segment (residues 1-114) was fused to maltose-binding protein to increase solubility during ligation and purification as well as to an intein, which was cleaved to yield a MESNa thioester as shown in Figure 11A. The synthetic peptide (residues 115-120) was lipidated at the C-terminal carboxylic acid with phosphatidylethanolamine or a 16-carbon hexanoyl chain and then ligated to the recombinant segment. The lipidated LC3 constructs were used to study the de-conjugation of lipidated LC3 by the cysteine protease RavZ, providing insights into the antiautophagy mechanism of *Legionella pneumophila*.[40]

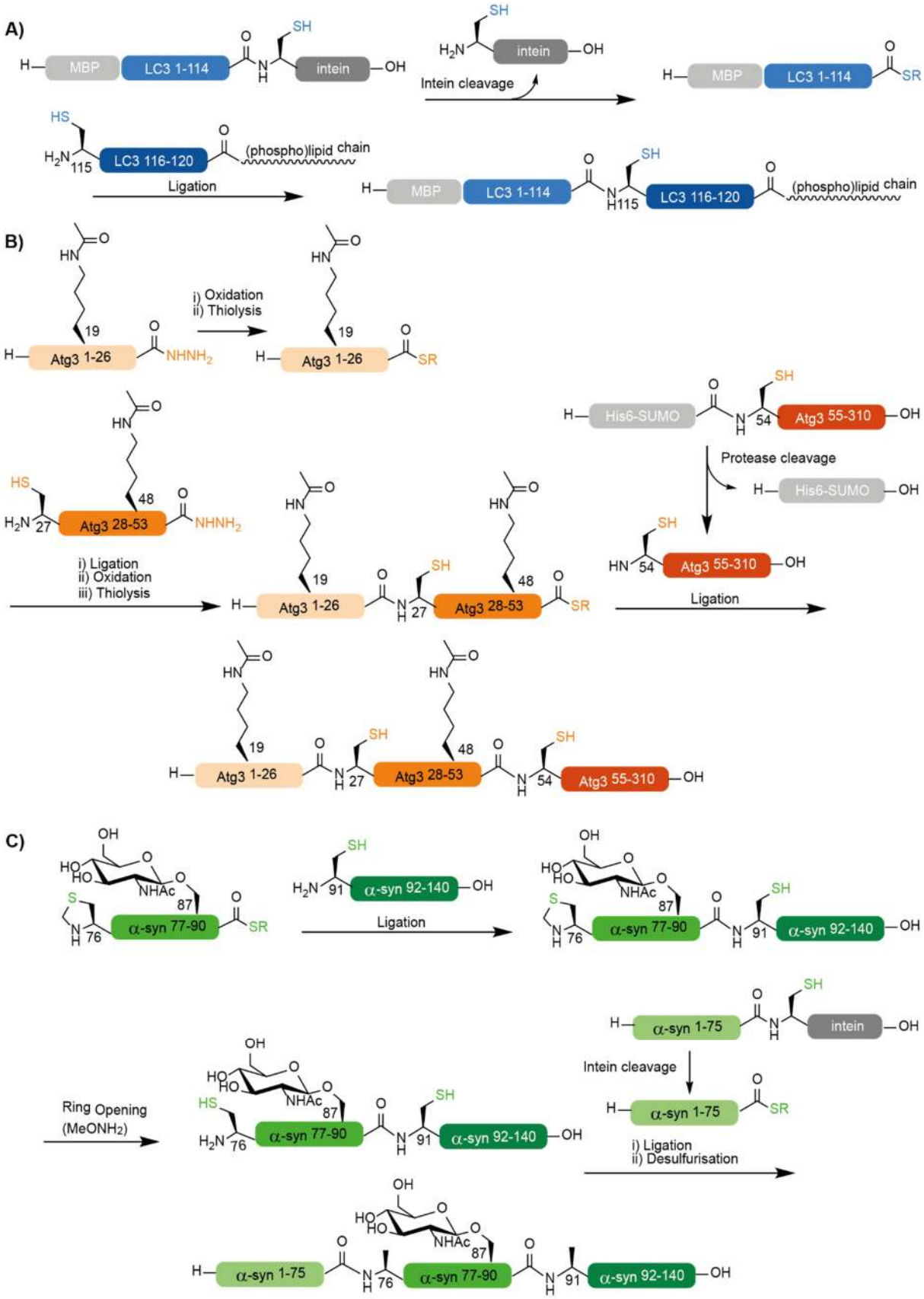

A)
MBP
LC3 1-114
intein
Intein cleavage
LC3 116-120
(phospho)lipid chain
Ligation
B)
Atg3 1-26
Atg3 28-53
Atg3 55-310
His6-SUMO
i) Oxidation
ii) Thiolysis
Protease cleavage
i) Ligation
ii) Oxidation
iii) Thiolysis
Ligation
C)
α-syn 77-90
α-syn 92-140
α-syn 1-75
Ligation
Ring Opening
(MeONH2)
i) Ligation
ii) Desulfurisation

**Figure 11:** Applications of EPL for generating modified proteins. **A)** Semi-synthesis of lipidated LC3. **B)** Semi-synthesis of acetylated Atg3. **C)** Semi-synthesis of glycosylated α-synuclein.

For the semi-synthesis of proteins bearing modifications at their N-termini, the synthetic methods described above are used to generate peptides bearing C-terminal thioesters *via* specialised linkers or thioester precursors. The remaining part of the protein is then obtained by recombinant expression of a precursor protein that is usually cleaved by a selective protease (e.g. TEV protease, thrombin, ubiquitinase, etc.) to give the requisite N-terminal Cys for EPL. This strategy was used to make semi-synthetic Atg protein with two N-terminal acetylation (Ac) sites in the N → C direction from one recombinant and two synthetic segments as outlined in Figure 11B.[41] The two synthetic segments (residues 1-26, K19Ac and residues 27-53, K48Ac) were generated as peptide hydrazides by SPPS. The hydrazide on the N-terminal segment (residues 1-26) was converted to a thioester for ligation *via* NCL to the second segment (residues 27-53). The C-terminal hydrazide on this segment was then converted to a thioester for ligation to the C-terminal recombinant segment (residues 54-310). A SUMO tag fusion approach was used to generate the N-terminal Cys residue on the recombinant segment after other approaches using a TEV protease cleavage site or direct expression of residues 54-310 proved inefficient, illustrating that several strategies may need to be explored to access the target fragments. The homogeneous diacetylated Atg3 bound to membranes and promoted lipidation of Atg8, a key event in autophagy.[41]

Proteins bearing PTMs that are more than 50 residues from either the N- or C-terminus are more challenging to generate *via* EPL because they require at least three segments and two ligation reactions. Strategies for protection of Cys residues and thioester precursors described above have nevertheless provided access to these molecules. These strategies have been used to explore the effects of N-glycosylation and O-glycosylation on α-synuclein, a protein implicated in Parkinson's disease.[42] In the latter study, several sites in the central 'non-amyloid component' domain are known to be O-glycosylated and α-synuclein glycosylated at Ser87 was prepared in the C → N direction from three fragments by EPL as shown in Figure 11C.[42] As α-synuclein contains no native Cys residues, Ala 76 and Ala 91 were mutated to Cys for ligation. The N-terminal segment (residues 1-75) was expressed recombinantly fused to an intein, which was cleaved in the presence of MESNa to give a C-terminal thioester. SPPS was used to synthesise the central glycosylated segment (residues 76-90) on Dawson 3-(Fmoc-amino)-4-aminobenzoyl AM resin, bearing an N-terminal Thz. The C-terminal segment

(residues 91-140) was expressed recombinantly with the N-terminal Cys generated by endogenous methionine aminopeptidase. Assembly of the segments was carried out in a C → N direction with ligation of the middle and C-terminal segments followed by conversion of the Thz into Cys using methoxyamine at acidic pH, thus enabling ligation to the N-terminal segment to generate the full-length sequence. Finally, the Cys residues at the ligation junctions were desulfurised to the corresponding native alanine moieties. The resulting full-length site-selectively O-glycosylated α-synuclein was then tested in aggregation and membrane-binding assays.[42] O-glycosylation of Thr72 or Ser87 decreased the aggregation of α-synuclein and inhibited its cleavage by the protease calpain, without affecting its association with membranes.[42]

An elegant N → C direction EPL strategy to assemble modified lysine acetylated histone H4 from two synthetic and one recombinant segment has also been described.[43] In this approach, instead of protecting the N-terminal Cys residues, peptide hydrazides were used as thioester precursors. The N-terminal synthetic segment (residues 1-14, bearing a peptidase-cleavable sidechain linker) was ligated to the middle synthetic peptide hydrazide (residues 15-37, acetylated at Lys-20). Following desulfurisation and thioester conversion, the unmodified recombinant C-terminal segment (residues 38-102) was ligated, followed by *in situ* desulfurisation. The cleavable linker was used to assemble asymmetric nucleosomes comprising two copies of histone H4, one unmodified, and one with an acetyl lysine residue.[43]

While synthetic peptide segments can be temporarily protected at the N- and C-termini using chemical strategies, the methods of intein cleavage and protease cleavage used to generate recombinant proteins with C-terminal α-thioesters and N-terminal Cys residues, respectively, are in themselves effective protection strategies for these functionalities. Attachment of small protein PTMs such as ubiquitin and small ubiquitin-like modifier (SUMO) to a side chain of a protein is also readily achieved using semi-synthesis techniques and many such studies have been used to elucidate the role of ubiquitylation in protein degradation. In a recent example that demonstrates temporary 'masking' of both thioesters and N-terminal cysteines, a semi-synthetic strategy to produce histone 2A (H2A) that was phosphorylated and ubiquitylated was designed (Figure 12A).[44] The central unmodified segment (residues 21-134) was generated recombinantly, fused to SUMO at the N-terminus (protecting the N-terminal Cys) and to the N-terminal split *Npu* intein at the C-terminus (acting as a masked thioester). Conversion to the C-terminal thioester was achieved by incubation with the C-terminal *Npu* intein and MESNa followed by ligation of the synthetic phosphorylated peptide (residues 135-142) in the presence

of methyl thioglycolate (an alternative additive to TFET) as a thiol catalyst. The N-terminal Cys on the central segment was then 'unmasked' by cleavage of the N-terminal SUMO by ubiquitin-like protease 1 (Ulp1). The N-terminal ubiquitin-bearing segment was itself generated by EPL of recombinant ubiquitin thioester to a synthetic peptide bearing a Cys on a Lys side chain. This Cys was desulfurised after ligation of ubiquitin and the C-terminal hydrazide on the synthetic peptide converted to a thioester for ligation of the N-terminal ubiquitinylated construct to the semi-synthetic phosphorylated segment (Figure 12A). Desulfurisation of the two remaining non-native cysteines yielded the final ubiquitylated and phosphorylated H2A, which was used to study the effects of these modifications on chromatin organisation.[44]

Illustrating the potential of semi-synthetic strategies for investigating interactions between molecules inaccessible by current molecular biology methods, the crosstalk between histone modifications was studied using two semi-synthetic modified histone proteins: sumoylated histone H4 (suH4) and dimethylated histone H3 (H3K4Me2).[45] As shown in Figure 12B, semi-synthetic H3K4Me2 was generated by ligation of a synthetic N-terminal segment (residues 1-6) bearing dimethyl-lysine and a C-terminal hydrazide with a recombinant segment (residues 7-135) possessing an N-terminal Cys, which was subsequently desulfurised. Sumoylated histone H4 was generated by ligation of the synthetic N-terminal H4(1-14) peptidyl hydrazide to recombinant SUMO thioester at the side-chain of Lys12 using a (2-(tritylthio)ethyl)hydroxylamine NCL auxiliary. The auxiliary was then cleaved and the hydrazide converted to a C-terminal thioester for ligation to recombinant H4(15-102) and desulfurisation.[45] Assembly of the two modified histones into nucleosomes enabled the effects of sumoylation of H4 on demethylation of H3 by the LSD1-CoREST complex to be examined; sumoylation of H4 stimulates intra-nucleosomal demethylation by LSD1 and possibly also decacetylation by histone deacetylase 1 (HDAC1).[45] Future studies should allow the crosstalk between other combinations of PTMs and the effects on gene activation to be elucidated.

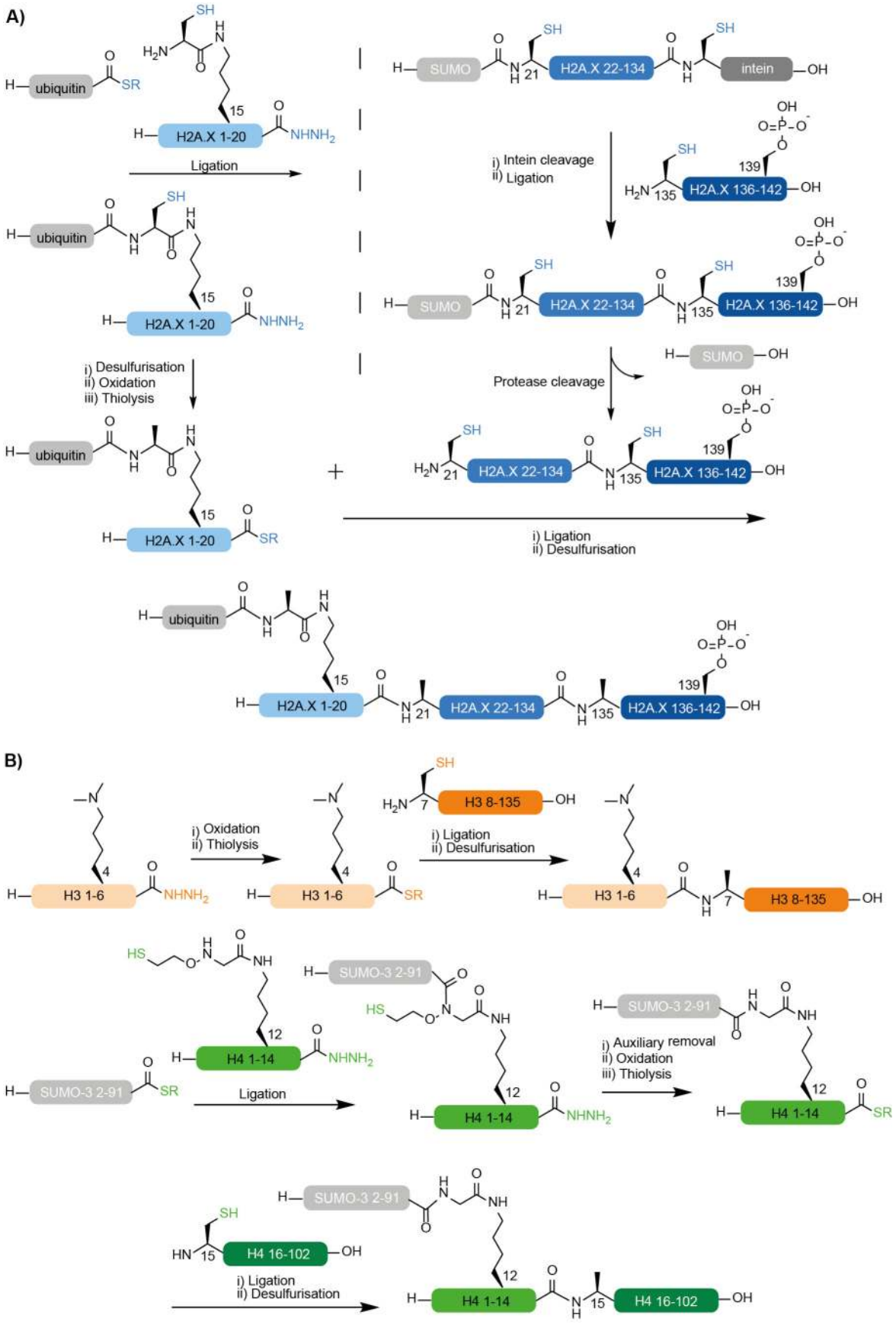

A)
ubiquitin
SR
H2A.X 1-20
NHNH2
Ligation
SUMO
H2A.X 22-134
intein
OH
i) Intein cleavage
ii) Ligation
H2A.X 136-142
i) Desulfurisation
ii) Oxidation
iii) Thiolysis
Protease cleavage
i) Ligation
ii) Desulfurisation
B)
H3 1-6
H3 8-135
i) Oxidation
ii) Thiolysis
i) Ligation
ii) Desulfurisation
SUMO-3 2-91
H4 1-14
Ligation
i) Auxiliary removal
ii) Oxidation
iii) Thiolysis
H4 16-102
i) Ligation
ii) Desulfurisation

**Figure 12:** Convergent EPL strategies for proteins bearing multiple modifications. **A)** Semi-synthesis of histone 2A. **B)** Semi-synthesis of histones 3 and 4 for assembly into nucleosomes to study crosstalk between modifications.

## Unnatural Protein Modifications

Just as naturally-occurring PTMs can be installed by EPL, the technique is equally applicable for the installation of unnatural chemoselective handles, labels, tags and a plethora of other functionalities accessible by synthetic chemistry. EPL has been used extensively for attaching fluorescent labels to recombinant proteins for imaging, both *in vitro* and in live cells. For example, *in vitro* modification of chromatin (histone H2B) with a fluorescent tag or ubiquitin was achieved by PTS between a peptide construct incorporating a C-intein, the desired modification tag and a cell-penetrating peptide, and a native histone fused to the N-intein.[46] Cell-permeable nanobodies have also been constructed by ligating a synthetic cell-penetrating peptide to a recombinantly-produced nanobody for intracellular immunolabelling and antigen manipulation.[47] Several split inteins have been explored for attaching a modified synthetic or recombinant protein segment to an unmodified segment *via* PTS, achieving both N- and C-terminal modifications. For example, an N-intein fused to a synthetic peptide bearing a fluorescent modification was ligated by PTS to a recombinant segment fused to the C-intein (Figure 13A).[39]

## EPL and PTS for Protein Cyclisation

Head-to-tail cyclic proteins are post-translationally modified by the formation of a peptide bond between the N- and C-termini of the polypeptide chain. They occur naturally and are characterised by chemical and enzymatic stability as well as potent biological activities. Non-native cyclisation of a given protein can also increase stability and activity and both NCL and EPL have been used to cyclise synthetic and recombinant precursors for such applications. For example, the naturally-occurring theta-defensin RTD-1 was generated *via* a modified protein splicing unit (Figure 13B).[48] An RTD-1 construct bearing an N-terminal Met and C-terminally fused to a modified *Mxe* Gyr intein was expressed in bacteria. Cleavage of the N-terminal Met by endogenous methionine endopeptidases yielded the N-terminal Cys and cleavage of the intein in the presence of glutathione yielded the C-terminal α-thioester for intramolecular EPL.

Although cyclic peptides are also easily accessible by SPPS followed by NCL, the EPL or PTS approach allows for uniform isotope labelling for NMR studies. In a recent example, a split-intein method was used to produce cyclic membrane scaffold protein (MSP), which was then used to assemble nanodiscs of various sizes with increased stability and homogeneity for NMR applications (Figure 13C).[49]

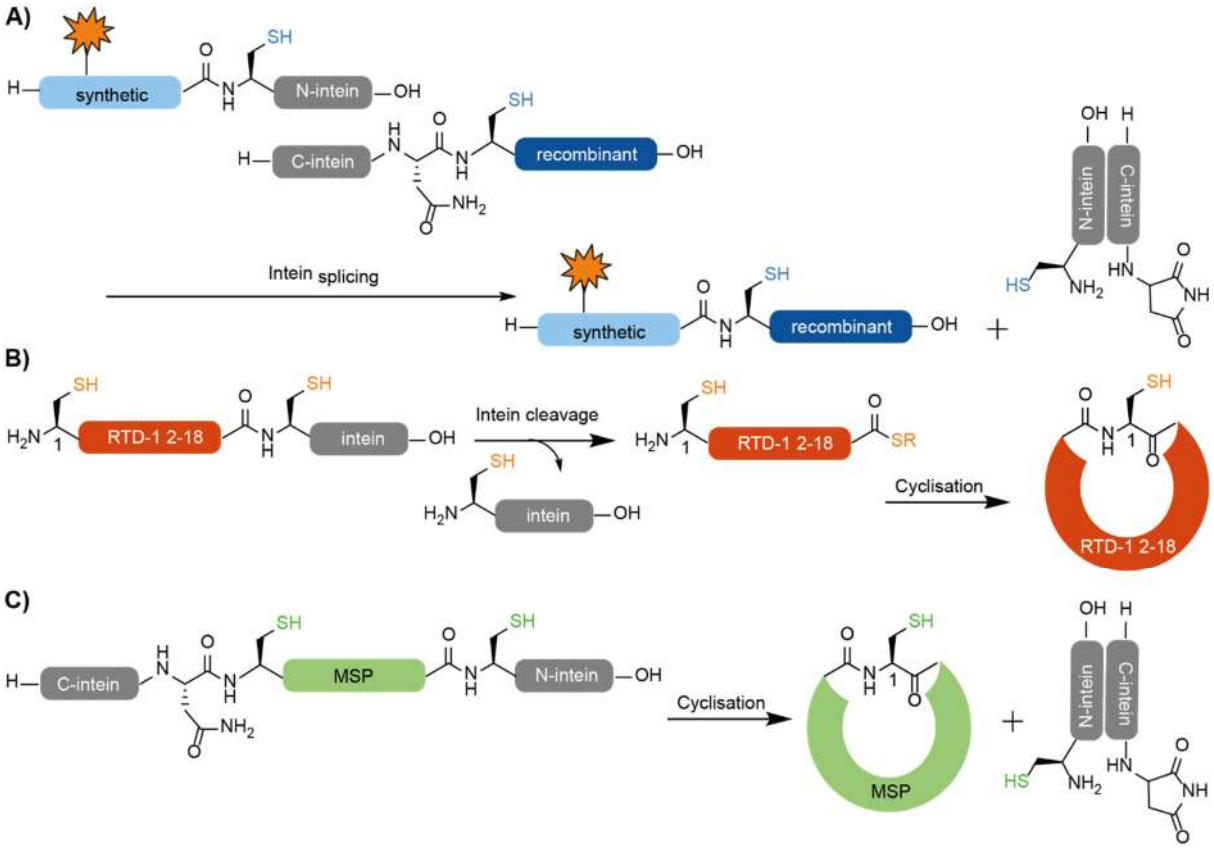



**Figure 13:** Intein splicing applications for generating **A)** fluorescently labelled; **B)** and **C)** cyclic proteins.

### Segmental Isotope Labelling of Proteins by EPL and PTS

The ability to generate a complete native protein by ligating two segments using EPL or PTS complements enzymatic ligation strategies and provides access to isotope labelled segments of a protein for structural biology studies. Segmental isotope labelling makes use of the ability to produce uniformly isotope labelled recombinant proteins; in contrast, chemical synthesis of uniformly isotope labelled peptides is prohibitively expensive due to the cost of purchasing isotopically labelled amino acid building blocks for use in SPPS. Segmental isotope labelling is useful for structural biology studies, particularly $^{15}N$- and $^{13}C$-labelling for NMR, because of the resultant decrease in spectral crowding as only the labelled segments are observed but

sequential resonance assignment from triple-resonance NMR experiments can still be carried out. Developments in protein deuteration and NMR pulse sequences such as methyl-TROSY (transverse relaxation optimised spectroscopy) experiments now enable structural and dynamic studies of large proteins in which the methyl groups of valine, leucine and isoleucine are $^{13}C$-labelled. In one recent study, several of these techniques were combined; the change in conformation and membrane-interactions upon phosphorylation of the β2-adrenoreceptor (β2AR) GPCR and its binding to arrestin were studied by NMR in the lipid bilayers of nanodiscs (Figure 14).[50] The intracellular C-terminal segment was isotope labelled and ligated to the transmembrane segments by PTS and then the complete β2AR was reconstituted in nanodiscs and enzymatically phosphorylated. The combination of segmental labelling with selective labelling of methyl groups in the C-terminal region and methionine residues in the transmembrane regions enabled the authors to propose that phosphorylation leads to adhesion of the C-terminal region to the intracellular parts of the transmembrane helices or lipids and allosteric rearrangement of the transmembrane helices. This results in the formation of the structural motif for β-arrestin binding, which modulates signal transduction.[50] Furthermore, interactions between labelled and unlabelled protein segments can be studied using $^{15}N$- or $^{13}C$-edited experiments and 'invisible' solubility or affinity tags can be ligated to isotope-labelled proteins. A segmental labelling approach has also recently been used to study the conformations of fibrils of the yeast prion protein Sup35NM domains by segmental and specific labelling coupled with dynamic nuclear polarisation and solid state NMR.[51] Small model systems do not faithfully represent the interactions of this full-length 253-residue protein but the spectra of the full-length uniformly labelled protein are too crowded for spectral resolution of long-range interactions. Segmental labelling of a 14-residue N-terminal segment of Sup35NM using PTS followed by DNP-enhanced NMR studies enabled the authors to determine that fibrils were not parallel in-register.

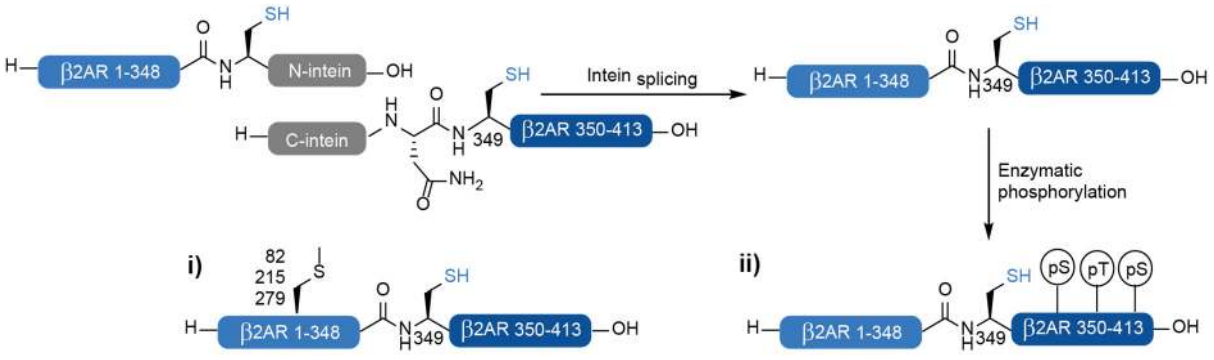


**Figure 14:** Segmental isotope labelling of β2AR using PTS. **i)** incorporation of labelled amino acids, **ii)** enzymatic phosphorylation.

**New Developments in Expanding the Applications of EPL and PTS**

New developments in EPL and PTS have resulted in an ever-widening diversity of new applications for these techniques. The size and diversity of protein targets is continually growing and includes examples of soluble, modified and membrane proteins. A novel application of EPL with nuclease-deficient Cas9 (dCas9) has allowed for the targeting of synthetic peptides to specific gene loci by performing in-cell EPL of the synthetic peptide to dCas9 and a cognate guide RNA.[52] This strategy offers the potential to deliver small molecules or epigenetic probes to target gene loci.

As described in earlier sections, many Cys surrogates and ligation auxiliaries have been developed to broaden the scope of NCL to non-Cys ligation sites on synthetic peptides. Recombinantly-expressed protein segments, in contrast, have been more limited to Cys residues at the N-terminus. Recently however, methods for recombinant incorporation of Sec have been described,[53] utilizing the more rapid ligations already established for synthetic peptides bearing N-terminal Sec.[4] This was illustrated recently by the synthesis of the native selenoprotein M *via* 'Sec-EPL'. A recombinant N-terminal Sec-containing segment was made in *E. coli* by misloading the cysteinyl-tRNA with Sec and cleaving the maltose-binding protein fusion construct with TEV protease before the Sec residue. The approach can also be used for Sec-EPL of non-selenoproteins by following the ligation with a deselenisation step to a native Ala or Ser, or for conjugation of PTMs by elimination of selenium to form dehydroalanine followed by electrophilic addition of the desired modification.[53] Although the Sec-EPL strategy provides access to proteins with a selenium ligation site more than about 50 residues from the termini, it has the potential drawback that the selenium-containing segment should not contain Cys residues as these would also be incorporated as Sec.[53] A strategy for generating selenoesters during intein cleavage would further increase potential EPL applications in the same way that DSL has expanded the scope and rate of synthetic peptide ligations. An alternative to using Sec or ligation auxiliaries for ligation at non-Cys sites involves enzyme-mediated ligation reactions.[29]. In this approach the catalytic residue (typically Cys, Ser or Thr) is used in the initial transfer reaction of a (thio)ester and the active site appears to be tolerant of a variety of residues at the P1-4 and P'1-P'2 sites.

**Critical Discussion / Outlook**

Proteins are the effector molecules of the vast majority of biological process and therefore are intimately involved in many diseases and pathologies. However, accessing sufficient quantities of homogeneous proteins from biological sources for in depth investigation is not always possible, and usually not feasible when they possess post-translational modifications. As such, chemical protein synthesis has a vital role in accessing homogenous samples of biologically and medicinally relevant proteins in sufficient quantities for their study. The development of NCL as a means of accessing proteins beyond the length limits of SPPS has been integral to this pursuit, especially when coupled with desulfurisation/deselenisation in concert with thiolated/selenated amino acids as a means of opening up a wider range of potential ligation junctions and thereby facilitating access to a broader range of target proteins. Subsequent advances in the NCL field have enabled multi-component fragment assembly in either the N → C or C → N directions (or a combination of these two strategies for a convergent approach) which has enabled the synthesis of proteins up to 358 residues in length by total chemical synthesis. However, this is still smaller than the average human protein and thus the combination of NCL with protein expression to allow access to longer (post-translationally modified) proteins through EPL and to a minor extent also through PTS has also been crucial. A key technology in the semi-synthesis of proteins *via* EPL involves modern recombinant technologies to generate protein fragments with N-terminal Cys residues or the use of inteins to produce proteins bearing C-terminal thioester functionalities that can be used in NCL and protein splicing reactions to form larger protein targets. The rapid evolution of NCL-related technologies in the 24 years since the initial report of the reaction has afforded the modern peptide scientist with a litany of options in terms of synthetically accessing a target protein and it can only be expected that the number of technologies centered on the NCL concept will continue to increase in the coming years. As a result, the size and complexity of proteins accessed through total chemical synthesis and semi-synthesis will continue to increase, which will further expand our knowledge on the features that control 3D protein structure, activity and stability. This deep understanding of the main functional molecules of life will help delineate the roles of proteins and their modifications in disease and provide new avenues for the development of protein therapeutics.

**Acknowledgements**

EEW was funded by an Australian Postgraduate Scholarship and a John A. Lamberton Research Scholarship. ACC and CFWB have been funded by the Vienna Science and Technology Fund (WWTF) through project LS17-008. RJP is supported through an Australian Research Council Future Fellowship (FT130100150).

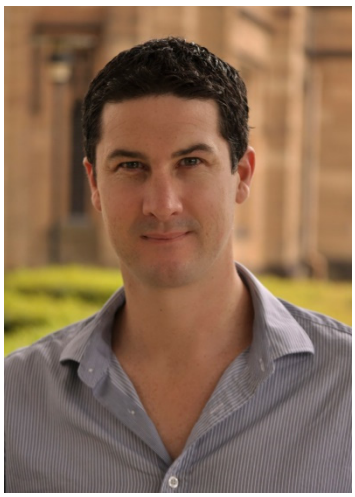
## References


1. P. E. Dawson, T. W. Muir, I. Clark-Lewis and S. B. H. Kent, *Science*, 1994, **266**, 776-779.
2. L. Z. Yan and P. E. Dawson, *J. Am. Chem. Soc.*, 2001, **123**, 526-533.
3. Q. Wan and S. J. Danishefsky, *Angew. Chem. Int. Ed.*, 2007, **46**, 9248-9252.
4. S. S. Kulkarni, J. Sayers, B. Premdjee and R. J. Payne, *Nat. Rev. Chem.*, 2018, **2**, 0122.
5. T. M. Hackeng, J. H. Griffin and P. E. Dawson, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 10068-10073.
6. E. C. B. Johnson and S. B. H. Kent, *J. Am. Chem. Soc.*, 2006, **128**, 6640-6646.
7. R. E. Thompson, X. Y. Liu, N. Alonso-Garcia, P. J. B. Pereira, K. A. Jolliffe and R. J. Payne, *J. Am. Chem. Soc.*, 2014, **136**, 8161-8164.
8. R. Mousa, R. Notis Dardashti and N. Metanis, *Angew. Chem. Int. Ed.*, 2017, **56**, 15818-15827.
9. T. S. Chisholm, D. Clayton, L. J. Dowman, J. Sayers and R. J. Payne, *J. Am. Chem. Soc.*, 2018, DOI: 10.1021/jacs.8b03115.
10. N. Ollivier, T. Toupy, R. C. Hartkoorn, R. Desmet, J.-C. M. Monbaliu and O. Melnyk, *Nature Communications*, 2018, **9**, 2847.
11. N. J. Mitchell, L. R. Malins, X. Y. Liu, R. E. Thompson, B. Chan, L. Radom and R. J. Payne, *J. Am. Chem. Soc.*, 2015, **137**, 14011-14014.
12. S. Bondalapati, M. Jbara and A. Brik, *Nat. Chem.*, 2016, **8**, 407-418.
13. L. R. Malins and R. J. Payne, *Curr. Opin. Chem. Biol.*, 2014, **22**, 70-78.
14. S. B. H. Kent, *Chem. Soc. Rev.*, 2009, **38**, 338-351.
15. L. Raibaut, N. Ollivier and O. Melnyk, *Chem. Soc. Rev.*, 2012, **41**, 7001-7015.
16. R. E. Thompson, B. Chan, L. Radom, K. A. Jolliffe and R. J. Payne, *Angew. Chem. Int. Ed.*, 2013, **52**, 9723-9727.
17. D. Bang, B. L. Pentelute and S. B. H. Kent, *Angew. Chem. Int. Ed.*, 2006, **45**, 3985-3988.
18. E. E. Watson, X. Liu, R. E. Thompson, J. Ripoll-Rozada, M. Wu, I. Alwis, A. Gori, C.-T. Loh, B. L. Parker and G. Otting, *ACS Cent. Sci.*, 2018, **4**, 468-476.
19. N. Ollivier, J. Dheur, R. Mhidia, A. Blanpain and O. Melnyk, *Org. Lett.*, 2010, **12**, 5238-5241.
20. N. Ollivier, J. Vicogne, A. Vallin, H. Drobecq, R. Desmet, O. El Mahdi, B. Leclercq, G. Goormachtigh, V. Fafeur and O. Melnyk, *Angew. Chem. Int. Ed.*, 2012, **51**, 209-213.
21. J. B. Blanco-Canosa and P. E. Dawson, *Angewandte Chemie (International ed. in English)*, 2008, **47**, 6851-6855.
22. G. M. Fang, Y. M. Li, F. Shen, Y. C. Huang, J. B. Li, Y. Lin, H. K. Cui and L. Liu, *Angew. Chem. Int. Ed.*, 2011, **50**, 7645-7649.
23. J. S. Zheng, S. Tang, Y. K. Qi, Z. P. Wang and L. Liu, *Nat. Protoc.*, 2013, **8**, 2483-2495.
24. D. Bang, N. Chopra and S. B. H. Kent, *J. Am. Chem. Soc.*, 2004, **126**, 1377-1383.
25. M. Jbara, S. K. Maity and A. Brik, *Angew. Chem. Int. Ed.*, 2017, **56**, 10644-10655.
26. S. Reddy Post, S. Dery and N. Metanis, *Angew. Chem. Int. Ed.*, 2015, **55**, 992-995.
27. W. Jiang, B. Zhang, C. Fan, M. Wang, J. Wang, Q. Deng, X. Liu, J. Chen, J. Zheng, L. Liu and T. F. Zhu, *Cell Discovery*, 2017, **3**, 17037.
28. A. C. Woods, J. G. Guillemette, J. C. Parrish, M. Smith and C. J. A. Wallace, *J. Biol. Chem.*, 1996, **271**, 32008-32015.
29. M. Schmidt, A. Toplak, P. J. Quaedflieg and T. Nuijens, *Curr. Opin. Chem. Biol.*, 2017, **38**, 1-7.
30. J. P. Pellois and T. W. Muir, *Curr. Opin. Chem. Biol.*, 2006, **10**, 487-491.
31. H. Paulus, *Annu. Rev. Biochem.*, 2000, **69**, 447-496.

32. T. W. Muir, D. Sondhi and P. A. Cole, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 6705-6710.
33. G. Volkmann and H. D. Mootz, *Cell Mol. Life Sci.*, 2013, **70**, 1185-1206.
34. K. V. Mills, M. A. Johnson and F. B. Perler, *J. Biol. Chem.*, 2014, **289**, 14498-14505.
35. S. Frutos, M. Goger, B. Giovani, D. Cowburn and T. W. Muir, *Nat. Chem. Biol.*, 2010, **6**, 527-533.
36. N. H. Shah and T. W. Muir, *Chem. Sci.*, 2014, **5**, 446-461.
37. H. D. Mootz, *ChemBioChem*, 2009, **10**, 2579-2589.
38. A. J. Stevens, G. Sekar, N. H. Shah, A. Z. Mostafavi, D. Cowburn and T. W. Muir, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 8538-8543.
39. C. Ludwig, M. Pfeiff, U. Linne and H. D. Mootz, *Angew. Chem. Int. Ed.*, 2006, **45**, 5218-5221.
40. A. Yang, I. Hacheney and Y. W. Wu, *Bioorg. Med. Chem.*, 2017, **25**, 4971-4976.
41. Y. T. Li, C. Yi, C. C. Chen, H. Lan, M. Pan, S. J. Zhang, Y. C. Huang, C. J. Guan, Y. M. Li, L. Yu and L. Liu, *Nat. Commun.*, 2017, **8**, 14846.
42. Y. E. Lewis, A. Galesic, P. M. Levine, C. A. De Leon, N. Lamiri, C. K. Brennan and M. R. Pratt, *ACS Chem. Biol.*, 2017, **12**, 1020-1027.
43. N. Guidotti, C. C. Lechner and B. Fierz, *Chem. Commun.*, 2017, **53**, 10267-10270.
44. S. Kilic, I. Boichenko, C. C. Lechner and B. Fierz, *Chem. Sci.*, 2018, **9**, 3704-3709.
45. A. Dhall, C. E. Weller, A. Chu, P. M. M. Shelton and C. Chatterjee, *ACS Chem. Biol.*, 2017, **12**, 2275-2280.
46. Y. David, M. Vila-Perello, S. Verma and T. W. Muir, *Nat. Chem.*, 2015, **7**, 394-402.
47. H. D. Herce, D. Schumacher, A. F. L. Schneider, A. K. Ludwig, F. A. Mann, M. Fillies, M. A. Kasper, S. Reinke, E. Krause, H. Leonhardt, M. C. Cardoso and C. P. R. Hackenberger, *Nat. Chem.*, 2017, **9**, 762-771.
48. A. Gould, Y. Li, S. Majumder, A. E. Garcia, P. Carlsson, A. Shekhtman and J. A. Camarero, *Mol. Biosyst.*, 2012, **8**, 1359-1365.
49. M. Jonas, D. Goricanec and F. Hagn, *ChemBioChem*, 2018, DOI: 10.1002/cbic.201800345.
50. Y. Shiraishi, M. Natsume, Y. Kofuku, S. Imai, K. Nakata, T. Mizukoshi, T. Ueda, H. Iwai and I. Shimada, *Nat. Commun.*, 2018, **9**, 194.
51. K. K. Frederick, V. K. Michaelis, M. A. Caporini, L. B. Andreas, G. T. Debelouchina, R. G. Griffin and S. Lindquist, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 3642-3647.
52. G. P. Liszczak, Z. Z. Brown, S. H. Kim, R. C. Oslund, Y. David and T. W. Muir, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 681-686.
53. J. Liu, Q. Chen and S. Rozovsky, *J. Am. Chem. Soc.*, 2017, **139**, 3430-3437.

**Author Biographies**

**Richard J. Payne**



Richard J. Payne graduated from the University of Canterbury, New Zealand with a BSc(Hons) degree (2002), a PhD from the University of Cambridge (2006) and was a Lindemann Fellow at The Scripps Research Institute (2006-2008). In 2008, he moved to the University of Sydney as a Lecturer where he is currently Professor of Organic Chemistry and Chemical Biology. Prof. Payne's research focuses on utilising the tools of synthetic chemistry to address problems of biological and medicinal significance. His lab has developed a number of synthetic technologies for the ligation-based assembly of large polypeptides and proteins. These methodologies have been employed in the chemical synthesis of a number of post-translationally modified proteins to understand structure-function and for the elucidation of new drug leads for a range of diseases.

**Emma E. Watson**

Emma E. Watson graduated from the University of Sydney, Australia in 2015 with a BSc(Hons) in Chemistry and Biochemistry. She is currently a final year PhD student in the lab of Prof. Payne at the University of Sydney in which her research has focused on the development of new methods to facilitate the synthesis of proteins as well as the application of these methods for the synthesis of post-translationally modified proteins of potential therapeutic interest.

**Christian F. W. Becker**

Christian F.W. Becker studied chemistry at the University of Dortmund (Germany) and obtained his diploma in 1998. After receiving his PhD in 2001 from the same university he was a postdoctoral fellow with Gryphon Therapeutics from 2002 to 2003. He became a group leader at the Max-Planck Institute in Dortmund, Germany in 2004 and was appointed as Professor for Protein Chemistry at the Technische Universität München in 2007. In 2011 he became Professor and Head of the Institute of Biological Chemistry at the University of Vienna. His group develops and uses chemical as well as biochemical means to generate peptides and proteins with otherwise unattainable (posttranslational) modifications to address biomedical and biotechnological challenges.

**Anne C. Conibear**

Anne C. Conibear completed her BSc(Hons) in Chemistry and Biochemistry (2008) and her MSc (2010) in Chemistry as a Beit Trust Scholar at Rhodes University, South Africa. She then completed her PhD (2014) with Prof. David Craik at the University of Queensland, Australia, focusing on structural characterisation and applications of cyclic disulfide-rich peptides. In 2014, she was awarded an Interdisciplinary Cancer Research (INDICAR) postdoctoral fellowship at the University of Vienna, Austria, where she was hosted by Prof. Christian Becker working on targeted immune-stimulating peptide conjugates for cancer therapy. She has recently been awarded a University of Queensland Development Fellowship to return to Australia and start an independent project focusing on the synthesis and structural properties of posttranslationally modified proteins.