Novel routes to defined post translational modifications using non-canonical amino acids.

Harley Luke Worthy

A thesis submitted to Cardiff University for the degree Doctor of Philosophy

School of Biosciences

Cardiff University

September 2018

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed………………………………………(candidate)   Date…………………………

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of   **PhD**   (insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed………………………………………(candidate)   Date…………………………

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated.

Other sources are acknowledged by explicit references. The views expressed are my own.

Signed………………………………………(candidate)   Date…………………………

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed………………………………………(candidate)   Date…………………………

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.**

Signed………………………………………(candidate)   Date…………………………

Abstract

Proteins are inherently limited by the properties of their constituent amino acids and attempt to overcome this by using post translational modifications (PTMs). PTMs are highly specific and can effectively modulate protein function faster than simple up or down regulation of protein production. However, PTMs often require a suite of other proteins to regulate and perform the modification to ensure accuracy, which can be hard to engineer into synthetic proteins. By introducing new chemistry into proteins *via* non-canonical amino acids (ncAAs) we can expand the range of new non-native PTMs that we can explore.

Non-native PTMs (nnPTMs), have the potential to be both bioorthogonal and easily transferable between proteins. This thesis examines the effects of engineering nnPTMs into superfolder Green Fluorescent Protein (sfGFP) to study the effects on fluorescence of: **1)** modification with small molecules (Chapter 3), **2)** Creation of covalent protein dimers (Chapter4), **3)** Interfacing proteins to carbon nanomaterials (Chapter 5), and **4)** Look at the effects of engineering cooperativity using ncAAs (Chapter6). Most of this work focused on the ncAA, p-azido-L-phenylalanine (azF) as it has several properties that would be desirable for use in proteins such as photo reactivity and selective reactivity with alkynes. Moreover, as azF can be incorporated into any target protein in a range of hosts, it is an ideal starting point to engineer nnPTMs that are easily transferable.

Throughout this thesis the importance of intricate hydrogen bonding networks and water channels, to the function of a protein, is made apparent through a range of *in silico*, structural and biophysical techniques. *In silico* modelling is used throughout to predict; the effects of nnPTMs on sfGFP structure (Chapter 3 and Chapter 6), dimer interfaces in Chapter 4, and show functional linking between sfGFP and carbon nanotubes in Chapter 5.

Table of Contents

1 Introduction

1.1 Amino acid chemistry and protein function

Since Anfinsen's work with ribonuclease A,[1] it has long been established that in general amino acid sequence encodes all the information for a protein to fold to its functionally active form. This key principle is the driving force for protein engineering: changes to the amino acid sequence will thus change the structure and function of a protein. Classically, protein engineering has focused on utilising the natural amino acid repertoire to explore the protein structure-function relationship. However, by looking at the structure-function relationship of proteins more closely, it is clear that, the chemistry of this native amino acid set is not enough to facilitate all the functional requirements needed by protein. For example, nearly 50% of proteins bind a metal ion to function.[2] Many proteins undergo chemical modification (e.g. GFP as described below or serine proteases[3]) or require cofactors (e.g. haem) in order to function.[4] Scheme 1.1 depicts some of the more common post translational modifications (PTMs). PTMs alter the function of protein by introducing a new cofactor (e.g. biotin and lipoic acid in various multienzyme complexes[5]) or induce conformational changes such as phosphorylation of innate hydroxyl residues of proteins involved in signalling pathways.[6] Methylation and acetylation of lysines in histones has been shown to affect transcription and DNA repair.[7,8] Other examples of covalent PTMs include; the modification of the 3 chromophore residues in Green Fluorescent Protein, to form a functional protein (Section 1.5)[9] and the formation of disulphide bonds, to stabilise tertiary protein structures[6] and contribute to quaternary structure (e.g. IgGs REF). One of the advantages of PTMs aside from an increased chemical toolkit is that it allows cells to quickly respond to changes in environment by switching proteins on and off without large scale increases in protein production/degradation that would be energetically costly to the cell.[6]
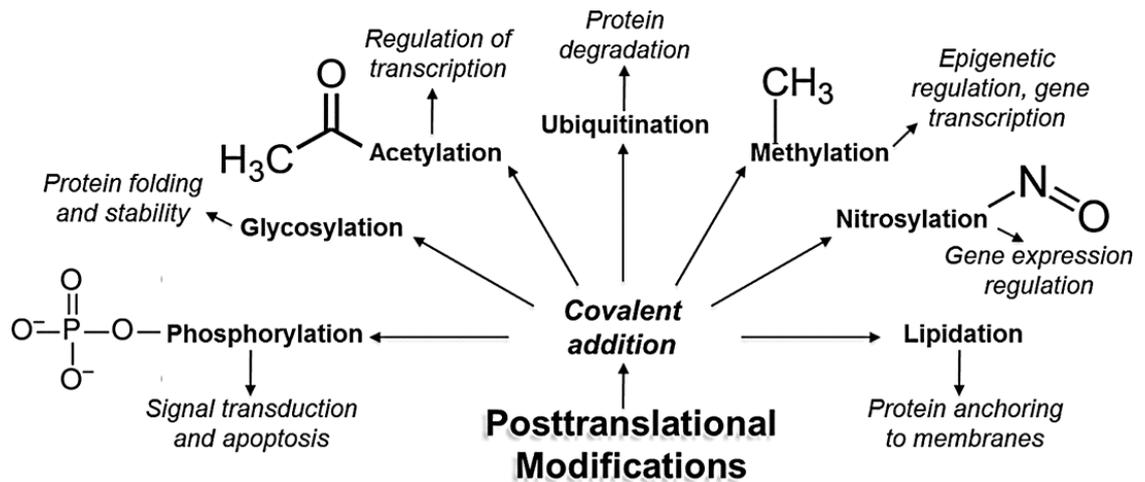
Scheme 1.1 Representation of some of the most common post-translational modifications.

Through evolution PTMs have become highly accurate, specific and reliable at modulating the function of proteins.[6] However, to get this level of control is very complex, often requiring associated protein machinery to ensure that the protein is correctly modified.[6] Such machinery is very specific to a target protein and sequence within that protein and limited additional modification options are available. For example, acetylation and deacetylation of specific lysines are regulated by protein families of acetylases and deacetylases. Also specific sequence motifs are often required to ensure correct targeting of a modification.[6] This makes engineering useful PTM events into a wide variety of proteins generally very challenging using a natural-based approach. It also limits adapting proteins for other approaches outside of biology such as material interfacing for use in nanotechnology (see Section 1.3). A potential solution to this problem is to introduce new and useful chemistry not present in biology into proteins to allow for precise and targeted modification.

1.2 Protein dimerisation

The fourth level of protein structure, quaternary structure, is where a functional protein unit comprises more than one polypeptide chain; this is known as protein

oligomerisations. In fact, the most common final structural form of proteins is oligomeric not monomeric, with symmetrical homodimers representing the most common observed form in the Protein Data Bank (PDB; https://www.ebi.ac.uk/pdbe/node/1).[10,11] Oligomerisation has many purposes from creating larger complex structures, such as the cytoskeleton and collagen,[11,12] and multienzyme complexes such as RNA polymerase and pyruvate dehydrogenase,[13] to small functional dimers such as cell receptors and transcription factors.[12-14] Protein oligomers are formed and stabilised by a variety of mechanisms, the most common of which, is the formation of a mutually compatible non-covalent interface comprised of hydrophobic interactions, hydrogen bonds (H-bonds) and electrostatic attraction.[13,15] Other less common mechanisms include domain swapping (usually swapped β-strands like with dimeric cytokines, and also seen in some cytochromes),[4] binding of shared metal cofactors (e.g. insulin)[13,16] and formation of inter protein disulphide bridges connecting monomers, which is common in antibodies to connect light and heavy chains.[10,11,13,17,18]

Given the abundance of oligomeric proteins, the questions to be asked are why they are so common and why are they typically symmetrical?[13] This question has been asked by many people including Klotz[19] and Monod[20] in the 1960s and more recently reviewed by Goodsell and Olson at the turn of the century.[13] All conclude that larger proteins are preferential to smaller proteins, because they are more resistant to denaturation and degradation through reduced solvent exposure from a reduced surface area.[13,15,19,20] There is also the advantage of having multiple active sites allowing for cooperative functionality between them. For example large multi-enzyme complexes, such as RNA polymerase, have an increased turnover rate compared to each of the individual subunits acting independently.[13]

These benefits could be achieved by creating single large proteins, however in nature normally many smaller subunits are used to generate the same effect. One of the benefits of using multiple protein subunits is with the control of translation errors. By

creating a large protein complex of small protein monomers, subunits with an error can be quickly discarded without great strain on the cell's resources.[13] In prokaryotes ~25% of proteins of 500 amino acids contain an amino acid substitution and 1/7 of proteins are released from the ribosome before the full length protein is created.[21,22] This means that proteins greater than 2000 amino acids are rarely fully translated and when they are translated the protein contains at least one error.[21,22] There is also an increased risk of misfolding with longer polypeptide chains having a large folding energy landscape and as such would require a large number of chaperone proteins.[21,22] Another benefit of multiple subunits as opposed to single large proteins is increased coding efficiency at the DNA level, *i.e.* a protein with 1000 amino acids could be coded by a single gene 3 kb long (not including regulatory DNA) or could be made up of 4, 250 amino acid subunits requiring a gene of only 750 bp and hence saving genetic space and space within the cell as well as saving energy on replicating and transcribing the longer gene.[11,13,23]

1.2.1 Design and creation of artificial protein oligomers

Designing and producing artificial self-assembling protein complexes is of great interest to protein engineering, because of the potential to create new functional ensembles and 'protein factories'.[24-26] Previous work has used a variety of techniques[4,25,27] including; fusion domains (e.g. Spycatcher[28] and split luciferase domains[29]),[24] helix-helix interactions,[30-32] metal ion bridging[33,34] and disulphide bridging[17] to name a few to create protein oligomers.[35-38]

The main reason for developing these artificial oligomers is to explore new structure and function space not currently present in nature and to use these new designs to try and solve technological, medical and scientific problems.[27,39,40] For example using domain insertion of split luciferase domains has been used to create Rho GTPase biosensors, where one half of the luciferase is genetically incorporated into the

GTPase and the second part of luciferase is attached to potential ligands.[29] If the ligand binds to the GTPase then the two pieces of luciferase combine to form a functionally active bioluminescent protein.[29] The Spycatcher-Spytag system is used for localisation and assisted purification of proteins of interest.[28,41] Since Bailey *et al.*[42] developed a method for metal-directed assembly of protein oligomers, metal coordination has been used to create variable diameter protein nanotubes and lattices[34,43] and Song and Tezcan[33] have successfully engineered rudimentary β-lactamase activity using metal ion coordination into a cytochrome based scaffold.[33]

Most of the techniques mentioned above rely on using structural information gleaned from the protein databank based on oligomers found in nature.[34,37,40] This information is then used to engineer proteins of interest to conform to a set of rules to generate dimerisation interfaces.[14,31,44] These interfaces multimerise because they either form hydrophobic patches that associate to escape solvent, hydrogen bonding networks where many weak interactions create a stable interface,[11] or the halves of an interface coordinate a metal ion[11,12] or heme.[11,12,45] A potential downside to these methodologies is the requirement for an extensive design process and significant engineering of proteins to incorporate multimerisation into new systems. Relying solely on symmetry can also limit the shapes and lattices that can be explored.[26]

A potential escape from this limitation is to use disulphide crosslinking *via* engineered cysteine residues. As mentioned above, engineering cysteines into proteins to form inter protein crosslinks is a common method for engineering dimerisation into proteins. It is a popular technique because there is very little modification of the target protein required (usually only a single residue). Functional covalent dimers of azurin were created by mutating residue Asn42 to cysteine allowing the formation of a disulphide linkage between the two monomers.[46] However, the dimers showed a decrease in electron transfer, attributed to reduced flexibility from the short disulphide linker.[46] Another issue with using this method includes increased risk of misfolded

proteins from mutated cysteines by forming incorrect disulphide bridges with existing cysteine residues.[47] This technique is also unsuitable for work *in vivo* due to the reducing environment of most living cells.[47] Introducing new chemistries into proteins may be a possible solution to the issues described above.

1.3 Material interfacing with non-biological systems

There are a wide variety of functions that proteins perform including molecular recognition (precise and defined binding to other moieties) catalysis and energy transfer/conversion that have potential uses outside of biology in areas such as synthetic biology, to create biosensors and 'enzyme factories'.[48] Generally, there is a need to couple the function of a protein to a secondary material (e.g. conductive for electronic output or photonic for an optical output) so that events in the protein can be observed by a measurable signal. For example, in glucose sensing the enzyme glucose oxidase is coupled to an electrical output, to monitor glucose levels.[49] In more recent advanced approaches glucose oxidase is attached to a gold nanoparticle *via* a mediator with a cysteine linkage, which in turn is linked to a semiconducting surface. Upon binding glucose, glucose oxidase generates electrons that are transmitted through the gold nanoparticle mediator to the semiconductor.[49]

Proteins have been interfaced with a variety of materials (>25),[48] of which gold,[12,50,51] graphene and carbon nanotubes[52-56] are of particular interest due to the material's properties, especially conductance.[57-60] Gold is a useful material to conjugate proteins due to its ability to conduct electrons across great distances (several nanometres).[61] Proteins can be attached to gold by simple adsorption (mediated through several non-covalent interactions), however this can lead to the unfolding of proteins and is undesirable.[61] Other approaches include attachment of native or introduced cysteine residues either directly onto gold or by using a linking mediator.[45,49,61,62] It has

been used extensively for the improvement of electrochemical response of a variety of enzymes including, Cytochrome *bd* oxidase,[63] and both glucose oxidase (see above) and dehydrogenase.[64] Gold has also been used to create biosensors to detect DNA (Using DNA Ligase) and viruses (peroxidase).[65,66] This lab has previously used gold to demonstrate the electron transfer capabilities of cytochrome b_{562} at the single molecule level.[50] Cytochrome b_{562} essentially acts as a transistor with conductance related to the applied redox potential.

Electronically active sp^2 materials like graphene and single-walled carbon nanotubes (swCNTs) have an extended π electron network.[67,68] Perturbation of these systems by, for example, introducing a local electrostatic surface (such as proteins) has been shown to alter the conductive properties of swCNTs.[67,68,69] For example Chen *et al*,[70] bound several different proteins including, human IgG, BSA amongst others to create a sensor for detecting proteins in serum. The sensor worked by detecting changes in electrical resistance on binding of proteins to the protein/swCNT hybrids.[70] CNTs were also used to monitor the activity of methyltransferases by bridging two CNTs with DNA and monitoring changes in conductance through the tubes.[69] In seminal work on single enzyme molecule analysis. Choi *et al*,[71] showed how even small local perturbations can generate a measurable output . They attached a single lysozyme to a SWCNT and monitored enzyme turnover through changes in enzyme conformation that resulted in local changes in the enzyme's surface electrostatics.

A variety of covalent and non-covalent approaches have been used to interface proteins to swCNTs.[72-75] For example, direct adsorption onto the swCNT sidewalls has been used in antibody detection by binding human IgG to swCNTs. Attachment *via* the amine group of lysines onto carboxyl groups of swCNTs has also been used to interface Microperoxidase-11 with CNTs to allow efficient electron transfer to a semiconducting membrane.[73]

a
Classical routes (e.g) lysine labelling

Non-optimal, heterogeneous, irregular signal

b
Site-specific interfacing

Optimal communication, homogenous, regular signal
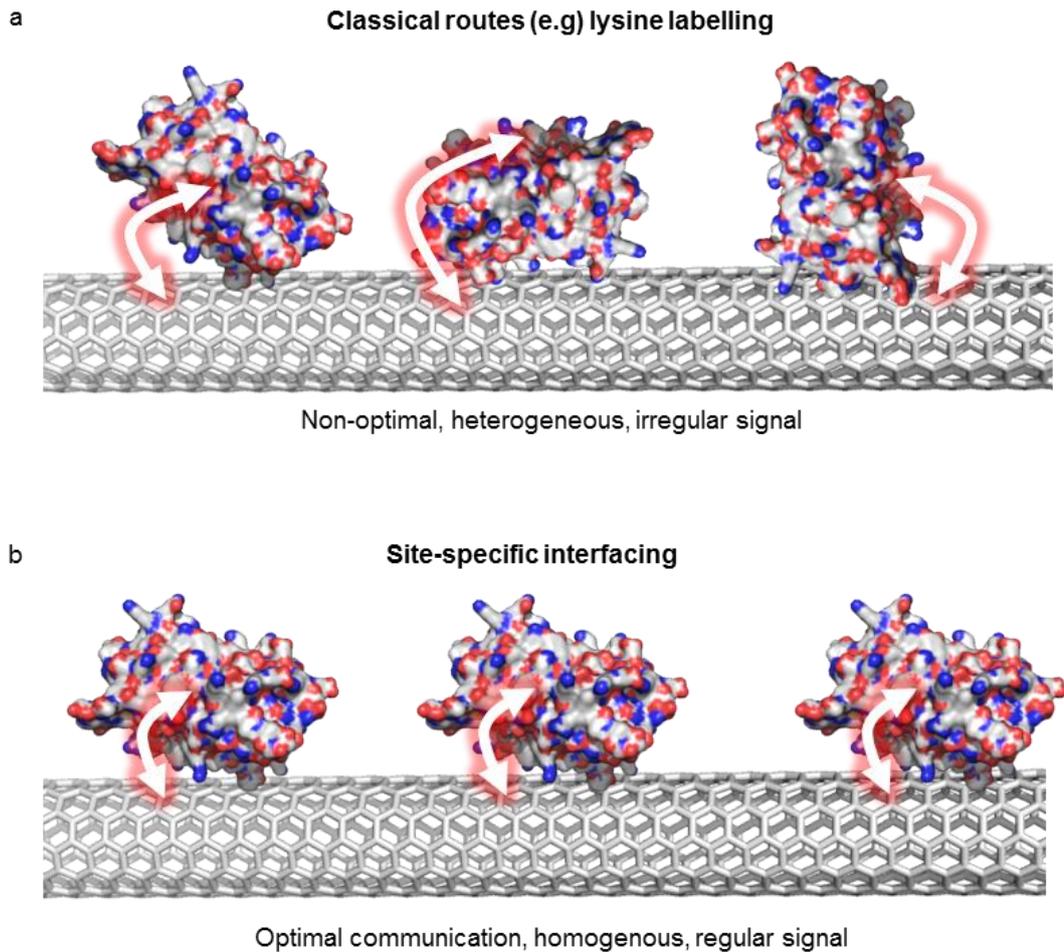
Figure 1.1 Comparison of specific and non-specific attachment onto swCNTs. a) classical non defined attachment to swCNTs e.g. attachment via lysine amine chemistry/hydrophobic adsorption. b) Site-specific attachment giving proteins same orientation relative to swCNT and single unified output.

The approaches outlined above are generally non-specific in terms of their protein attachment site leading to heterogenous and non-optimal interface locations, *i.e.* uncontrolled numbers of proteins on the surface in undefined orientations (Figure 1.1).[76] Or, they are restricted to a single set position which may not be optimal for protein function and the coupling of the protein output to the material.[49,76] For bulk studies, this is not a significant issue as heterogenous output signal is averaged out and weak coupling is amplified by the sheer number of molecules present. However, if we are to truly use proteins as nanoscale devices and components, there is a need to work at the single or few molecule level. This requires more precise molecular construction for

consistency and optimal coupling for maximal signal transfer from protein to material or *vice versa* (Figure 1.1).[50,51,59,77,78] Heterogenous attachment will give rise to inconsistent and variable outputs and lack of control over attachment may lead to loss of coupling between protein and material.

1.4 Non-Canonical amino acids and PTMs

One solution to overcome the current limitations with the native PTM approaches, mentioned above so as to facilitate the construction of novel protein oligomer (Section 1.2) and interface with non-biological materials (Section 1.3), is to genetically encode new and useful chemistry into proteins via non-canonical amino acids (ncAAs)[79,80]. These non-biological amino acids can be used to introduce many different functional groups to incorporate additional features into proteins, including but not limited to metal binding,[81,82] photo-reactive crosslinkers, improved phase resolution for crystal structure determination[80,83] and introducing unique bioorthogonal reaction handles such as alkynes, azides and tetrazines.[84-87]

1.4.1 Incorporation of ncAAs into proteins

Since the pioneering work of the Schultz lab in 1989,[88] to recode the genetic code to incorporate phenylalanine analogues including D-phenylalanine and p-nitro-L-phenylalanine into β-lactamase, the number of ncAAs capable of being incorporated has risen to well over 150.[89] Schultz's initial work focused on repurposing an archaeal tyrosyl-tRNA synthase tRNA to incorporate analogues of tyrosine into proteins. More recently natural machinery for incorporating pyrrolysine was been used for incorporating lysine derivatives[84,90-93] and is proving more adaptable for use in eukaryotic systems.[92]

As ncAAs are not part of the existing genetic code, a method is required to expand the genetic code to allow their incorporations. To incorporate a ncAA, there are two key requirements:[80,94,95] **(1)** an engineered tRNA that recognises a unique codon

specific to encode ncAA incorporation and **(2)** an engineered aminoacyl tRNA synthatase (aaRS) that adds the ncAA to the new cognate tRNA[80,94,95] (Figure 1.2a). Furthermore, the new tRNA and aaRS need to be orthogonal and not interact with any of the endogenous tRNA/aaRS. The main way this has been achieved is by using amber codon suppression. The amber stop codon (TAG) is the least prevalent of all codons in the standard genetic code in both prokaryotes and eukaryotes alike. [80,94] As it is naturally also non-encoding it is the simplest codon to adapt for ncAA incorporation and has been used to successfully incorporate a whole plethora of different ncAAs into a wide scope of expression systems from *Escherichia coli* to mammalian cell lines.[79,80] To maximise the efficiency of ncAA incorporations *E. coli* stains have been engineered with release factor-1 knocked out, [96] and all TAG codons removed.[97]

The first modified tRNA-aaRS pair was originally sourced from *Methanococcus jannaschii*[79,80]. The native tyrosine aaRS was engineered by directed evolution to aminoacylate its cognate tRNA with a tyrosine derivative, such as p-azido-L-phenylalanine[79,95,98-100]. The tRNA was also directly evolved to recognise the TAG codon (termed tRNA[CUA]) in the context of *E. coli*, primarily through engineering the anticodon loop.[79,80] The tRNA-ncAA complex recognises the amber stop codon (UAG for RNA) on the mRNA in the ribosome, which allows for the amino acid to be successfully incorporated into the protein (Figure 1.2), preventing termination by release factor-1.[79,80,95,96]
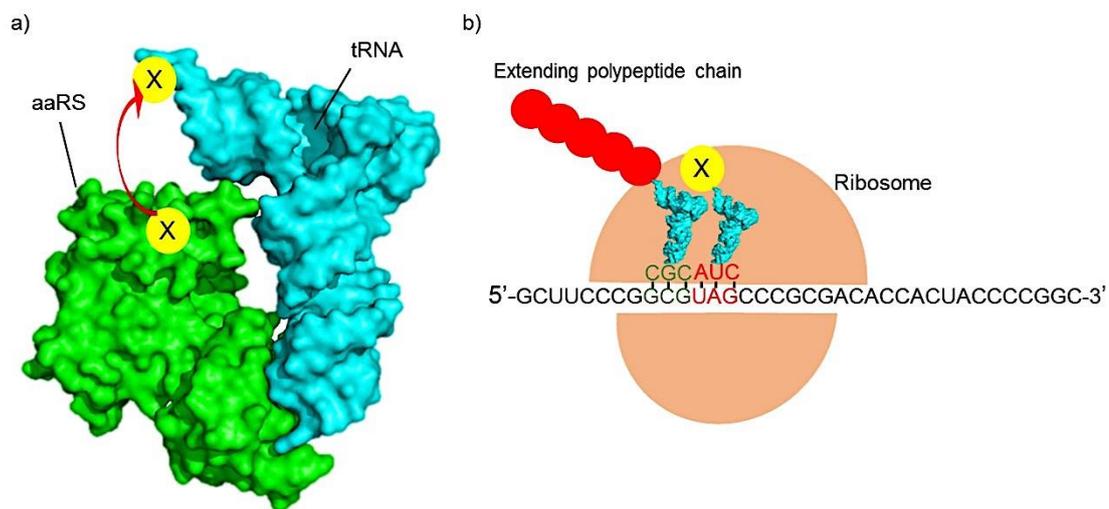
Figure 1.2 Incorporation of non-canonical amino acids into proteins. a) aminoacyl tRNA synthase (green) – tRNA (cyan) complex showing transfer of ncAA (yellow circle) b) Incorporation of ncAA, to propagating polypeptide chain of canonical amino acids (red circles) at amber stop codon (UAG) in the ribosome (labelled). X represents any non-canonical amino acid

A similar approach has been used to modify the tRNA/aaRS pair that incorporates pyrrolysine in *Methanosarcina mazei* and *Methanosarcina barkeri* to specifically incorporate pyrrolysine derived ncAAs at the amber stop codon.[84,91,92,101,102] Pyrrolysine (Figure 1.3) is a lysine derived ncAA incorporated at TAG codons in several methyltransferase genes of methanogenic bacteria.[103,104] Neumann *et al*[105] mutated the MbPylRS/MbtRNA$_{CUA}$ pair to incorporate N$^\varepsilon$-methyl-L-lysine into histones to mimic histone acetylation.[105,106] Subsequent work has further engineered the pyrrolysine system to incorporating other ncAA such as strained alkynes including s-cyclooctyne-L-lysine (SCO, Figure 1.3).[91,107]

This project will focus mainly on the ncAA p-azido-L-phenylalanine (azF, Figure 1.3), a photoreactive ncAA historically used for its photo-chemistry.[108] Chapter 4 will also use the pyrrolysine derivative s-cyclooctyne-L-lysine (SCO, Figure 1.3). SCO is a ncAA containing a strained alkyne group which reacts with azF to create a stable triazole linkage (Section 1.4.3).
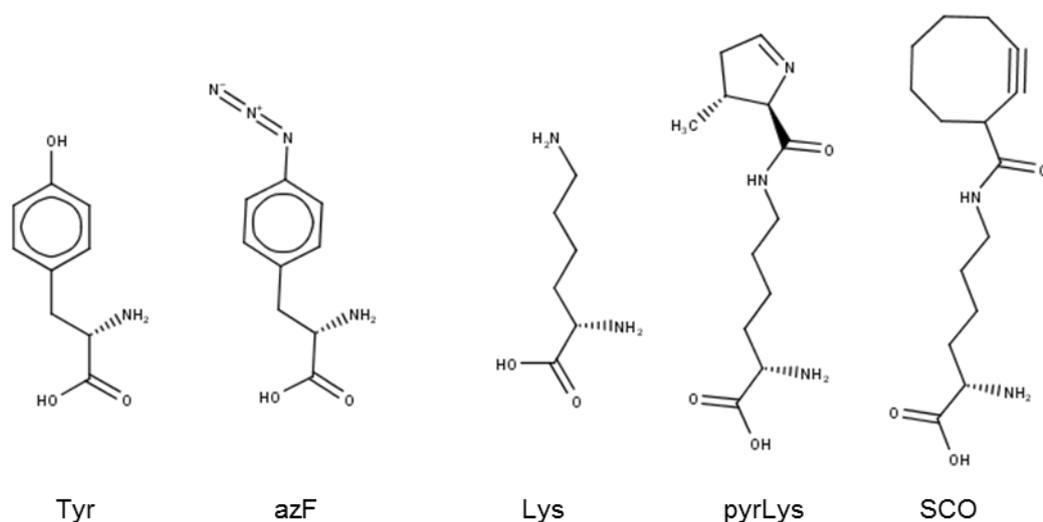
Figure 1.3 Line drawings of tyrosine (Tyr) and lysine (Lys) and their ncAA derivatives; p-azido-L-phenylalanine (azF), pyrrolysine (pyrLys), and s-cyclooctyne-L-lysine (SCO).

1.4.2 Phenylazide photochemistry

Phenyl azides (sometimes known as aryl azides) are an important class of organic compounds used in both chemistry and biology. Classically phenyl azide photochemistry has been used in biochemistry since the 1960s to identify protein-protein interactions by crosslinking.[108] The utility of phenyl azide chemistry in biology was realised early on in genetic code reprogramming as azF was one of the first ncAAs to be incorporated into proteins for useful applications.[94] Since then the photochemistry has been used to; modulate protein function; [83] label proteins with fluorophores;[77,83] attach proteins to DNA and nanomaterials[59] and investigating the interactome of *E. coli*.[59,78,84,85,109,110]

Phenyl azide photochemistry involves UV irradiation inducing the usual formation a singlet nitrene radical with the concomitant loss of molecular N_2 (Scheme 1.2). It should be noted that triplet nitrene can be formed although this usually requires extreme low temperatures (~77K).[111] From this point there are several pathways the nitrene radical can undertake depending on local environment.[108,112] The nitrene can be reduced to a phenylamine. In the context of the protein this has the potential to generate an

aromatic amine that can be utilised as a new hydrogen bond donor/acceptor group (Scheme 1.2, Route A). The nitrene radical can also undergo a reversible ring internalisation and expansion forming a dehydroazepine. Presence of a nucleophile leads to formation of an azepinone moiety (Scheme 1.2, Route D). It is also this route that classical protein crosslinking to lysine (via the primary amine) occurs.[108,111,113,114] The nitrene radical is also electrophilic so can attack regions of high electron density such as insertion across a C=C double such amino acids within aromatic groups in proteins and direct insertion across sp^2 nanocarbon materials such as graphene and carbon nanotubes (Scheme 1.2, Routes B and C respectively).[59,78]
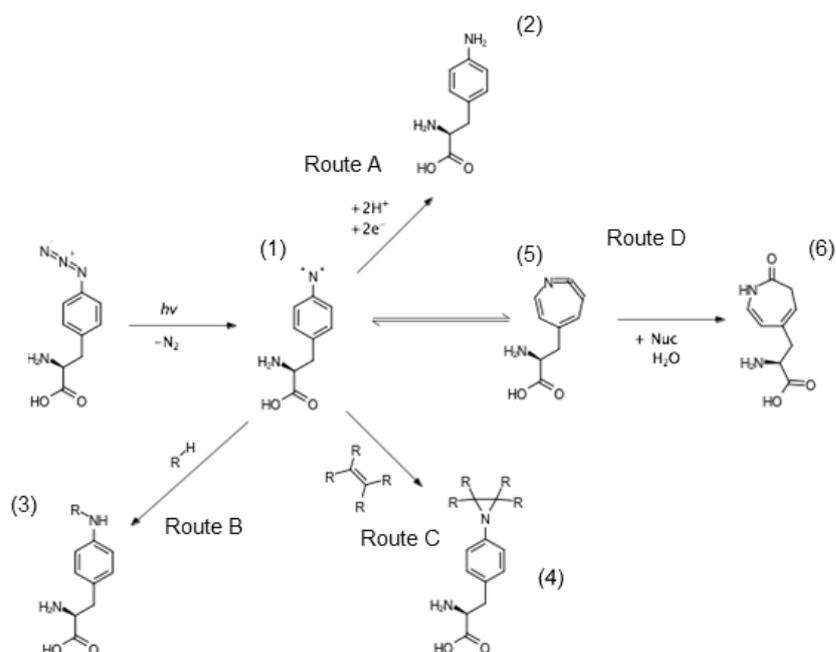
Scheme 1.2 Photochemical properties of p-azido-L-phenylalanine showing possible fates of azF after UV radiation induced formation of the nitrene radical (1). Route A: Reduction to p-amino-L-phenylalanine (2). Route B: Insertion in hydrocarbon sidechain forming amide crosslink (3). Route C: Incorporation across conjugated double bond system (4) e.g. graphene and carbon nanotubes. Route D: ring internalisation of nitrene to form dehydroazepine (5) then incorporation of a nucleophile to form an azepinone moiety (6).

1.4.3 Strain Promoted Azide Alkyne Cycloaddition

Phenyl azides, more correctly azides groups (N_3), can also undergo an orthogonal click reaction with alkynes via a cycloaddition reaction (Scheme 1.3). Click chemistry is an umbrella term for a wide range of chemical reactions that generate non-toxic products in high yields.[107] In addition, click reactions are compatible with physiological conditions and hence can be performed *in vivo*.[78,101,107,115] The incorporation of azF into proteins provides one half of the reaction component for Click chemistry. Initially, azides were reacted with aliphatic alkynes but this required the presence of Cu^+ that can be cytotoxic, including leading to the breakdown of proteins.[107] The mechanism for Cu mediated azide alkyne cycloaddition is shown in Scheme 1.3 and discussed extensively in Worrel *et al.*[116] In brief, two Cu^+ coordinate to the π and σ bonds of the alkyne to form an acetylide. Reversible coordination of the copper to the organic azide allows for nucleophilic attack of the N^3 atom of the azide by the β-carbon of the acetylide. Ligand exchange leads to closure of the ring structure creating a 6 membered metallacycle. Heterocycle reduction leads to release of one of the copper atoms creating a triazolide. Subsequent protonation of the triazole regenerates second copper atom and completes the reaction mechanism.

This initially made azide-alkyne cycloaddition largely unworkable in living cells and undertaken with caution *in vitro*. The other potential long term drawback was that only the 1,4 regioisomer around the new triazole linkage will form. To circumvent these issue Bertozzi introduced the use of strained cyclic alkynes to replace the aliphatic alkyne.[107,117,118] Using ring strained alkynes such as in SCO (Figure 1.3) and dibenzocyclooctyne (DBCO), the azide alkyne reaction (Scheme 1.4) can occur without need of a catalyst, and is known as strain-promoted azide alkyne cycloaddition (SPAAC).[107,117,118] The DBCO functional group has been used to link a variety of useful moieties including fluorescent probes such as Texas-red, and short aliphatic amine groups (DBCO-amine) that can be used as extra hydrogen-bond donor/acceptors and

further modification of the amine group or to simply induce local conformational changes.[78,85,109,115,119]
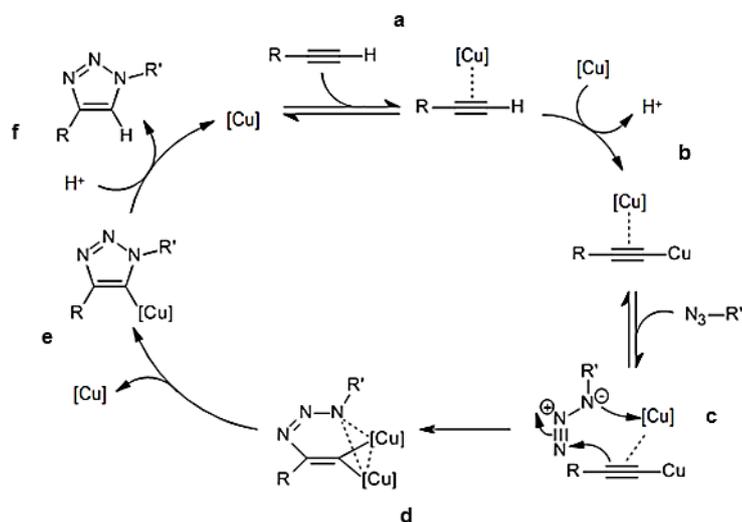
Scheme 1.3 Copper catalysed azide-alkyne cycloaddition. a) *in situ* formation of σ-bound copper(I) acetylide, b) recruitment of second π-bound copper(I), c) reversible coordination of organic azide to π-bound complex, d) nucleophilic attack at N3 of azide by β-carbon of alkyne on acetylide forms first N-C bond creating 6 membered metallacycle e) Ring contraction and release of copper to triazole-copper derivative f) protonation of triazole releases second copper molecule completing catalytic cycle (Adapted from Worrell *et al.*)[116]

The use of the pyrrolysine system to incorporate ncAAs, recently led to the genetic encoding of a lysine derived strained cyclooctyne (termed SCO; Figure 1.3) by the Lemke group.[84] The incorporation of a strained alkynes into proteins allows for labelling with fluorogenic azides. These are dyes that are quenched in azide form, but fluorescence strongly increases on conjugation *via* SPAAC.[84,120] Genetically encoded strained alkynes are not limited to click reactions with azides. In fact, strained alkynes can also react with tetrazines and nitrones.[92,93,121,122]
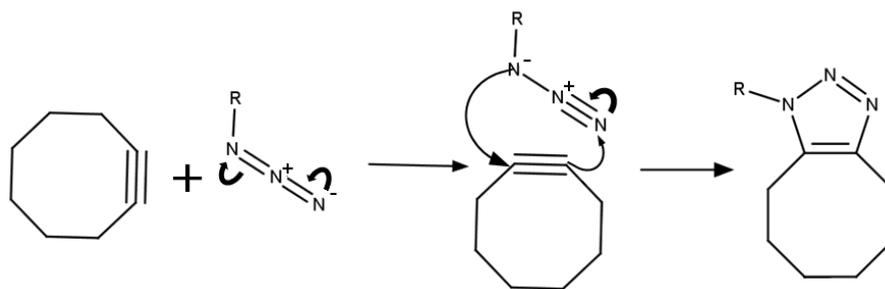
Scheme 1.4 Strain promoted azide alkyne cycloaddition. Electron transfer between alkyne and azide cause formation of triazole moiety.

1.5 Green Fluorescent Protein

Green Fluorescent Protein (GFP) originally found in *Aequorea victoria,*[9] is an autofluorescent protein that has since been used extensively as a genetically encoded fluorescent probe. Since the isolation of the gene encoding GFP, it has been engineered for many uses including, monitoring gene expression[123,124], cell imaging[125,126] and aiding in folding and solubility of proteins.[127] The first engineering efforts were aimed at making GFP more stable and fluorescent for production in cells cultured at 37°C[128,129] and soon moved to generating proteins with new fluorescent properties, notably shifting excitation and emission wavelengths to broaden the spectral range.[9,124,130,131] It is now used not only as a passive labelling component for cell biology but as active sensors in which fluorescence output can be modulated by specific inputs ranging from Ca^{2+} signalling[132,133] to redox conditions,[134] and even as components in nanoscience.[119,128,135-139]

The functional centre of GFP is the chromophore (p-hydroxybenzylideneimidazolinone) that resides in the middle of a helix within the centre of the β-barrel structure. The chromophore (CRO) is shielded from the solvent by a β-barrel containing 11 β-strands (Figure 1.4a). The chromophore is formed by the spontaneous rearrangement of Ser65, Tyr66 and Gly67 in the presence of O_2. Firstly, folding of the protein to form the barrel structure changes the torsional angles that bring

Ser65 and Gly67 into proximity. Secondly, water is released from nucleophilic attack on the carbonyl group of Ser65 by the amide group of Gly67, resulting in cyclisation. Finally oxidation of the α-β carbon bond in Tyr66, by molecular oxygen, gives GFP its fluorescence by forming an extended, delocalised π-electron system (Figure 1.4b).[9] The fluorescence of CRO is extremely sensitive to quenching and as such GFP is only fluorescent when the β-barrel is intact.
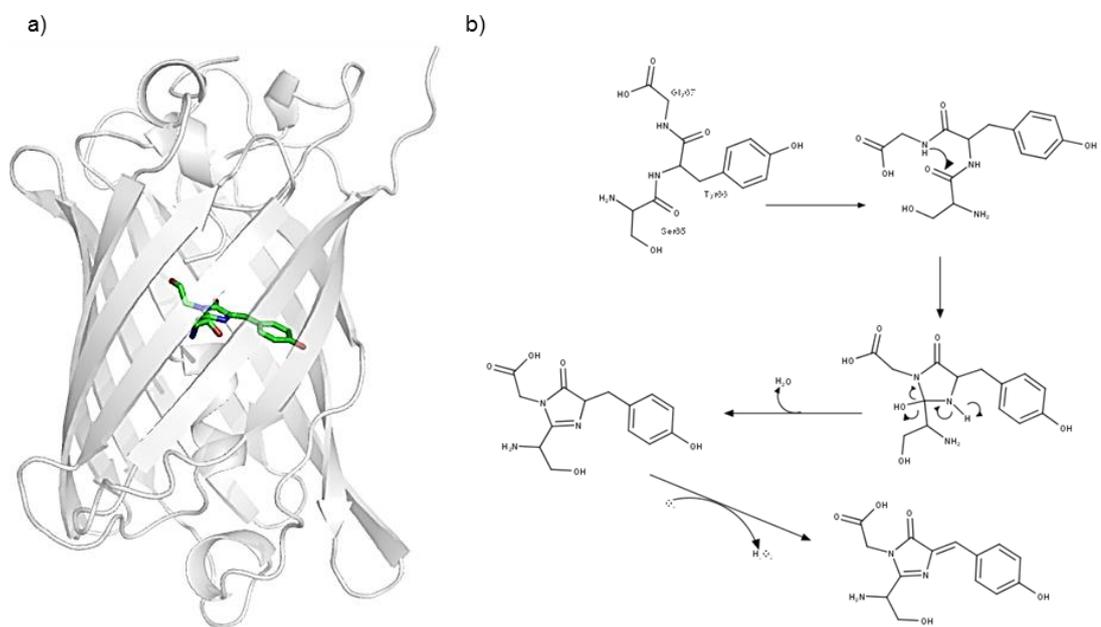
Figure 1.4 Structure of sfGFP a) Representation of sfGFP β-barrel in grey with chromophore shown as sticks highlighted in green, (PDB 2B3P) b) Scheme depicting chromophore maturation.

Tyr66 of the chromophore can have two potential protonation states of the ionisable phenol OH group: a neutral protonated (CroA) and as a charged phenolate (CroB) (Figure 1.7a). Dependent on the protonation state of the chromophore, GFP can be excited at either 395 nm (CroA) or 475 nm (CroB).[9,124,130] In the original GFP, the major excitation is at 395 nm due to the chromophore existing predominantly in the neutral CroA form. Brejc et al[140] calculated there to be a 6:1 ratio of neutral to anionic chromophores. However, under intense UV irradiation there is an increase in 475 nm

excitation.[124,141,142] This is through UV irradiation promoted deprotonation of the chromophore, most probably from proton transfer *via* hydrogen bonds of a buried water and S205 to E222.[140] The side chain of T203 is crucial to maintaining the phenolate anion. In the crystal structure of the original GFP, T203 exists in two conformations with the side chain OH group facing either towards or away from the chromophore in an approximate ratio of 15%-85% respectively.[9,140] This proportion is in agreement with the spectroscopic estimate.[9]

1.5.1 Fluorescence

The fluorescence of GFP (and other fluorophores) arises from the absorption of energy (usually light) by electrons in the chromophore. This causes the electrons to "jump" from their ground state to a higher energy excited state. The excited electrons relax into a lower energy radiation state by non-radiative energy transitions. Fluorescence emission occurs when an electron relaxes back towards the ground state where the wavelength of emission is inversely proportional to the energy lost. This can be visualised *via* a Jablonski diagram[143] as shown in Figure 1.5.

The environment of the chromophore determines these intermediate energy levels and hence tune the emission wavelength. When GFP is unfolded it is non-fluorescent, but the absorbance spectra stays the same. This is due to quenching by the surrounding solvent.[9,144]
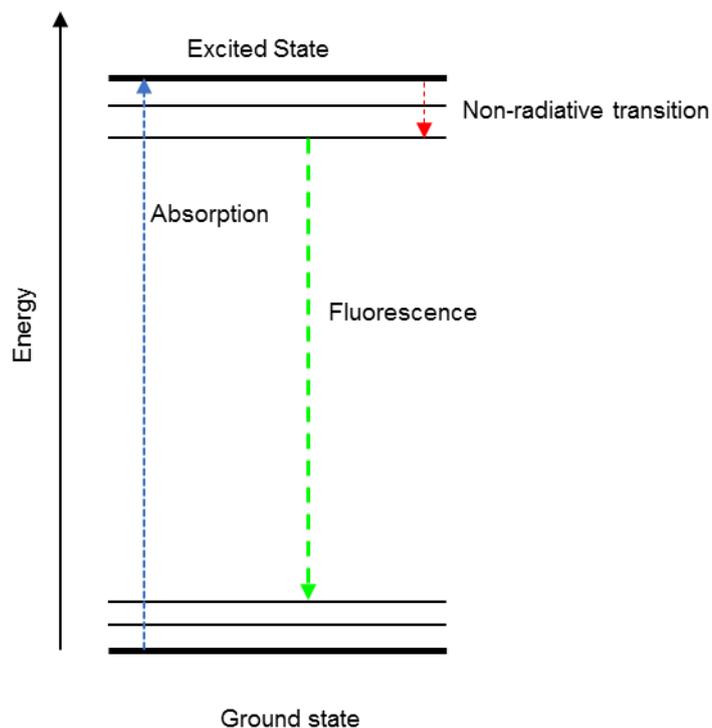
Figure 1.5 Jablonski diagram showing how fluorescence in GFP occurs. Electrons are excited from the ground state to higher energy "excited state" *via* absorption (Blue arrow). The electron relaxes to a lower energy excited state via non-radiative transition (Red Arrow). Further relaxation of electron to the ground state causes fluorescent emission.

1.5.1.1 Excited state proton transfer

It is arguable that excitation of GFP in the CroA form should emit in the blue region, but in fact emits green light as with excitation of CroB.[145,146] The Boxer group showed that excitation of GFP with 397 nm light, emission at 460 nm is briefly observed but decays at a rate matched by an increase in emission at 510nm.[141] The accepted reason for this is a process called, excited state proton transfer (ESPT). When GFP is excited in the CroA form, the proton in the hydroxyl group from Tyr66 is lost by shuttling through a 'proton wire',[140] carried through a water molecule, Ser205 and finally accepted by Glu222 forming the excited ionic intermediate (CroI*,Figure 1.6).[140,145,146] Upon the returning of GFP to the ground state the proton is restored to the chromophore. In GFPs with the S65T mutation in chromophore and hence having a dominant CroB form, the

hydrogen bonding network of Glu222 is changed, suppressing the negative charge which in turn allows the chromophore to be deprotonated.[147]
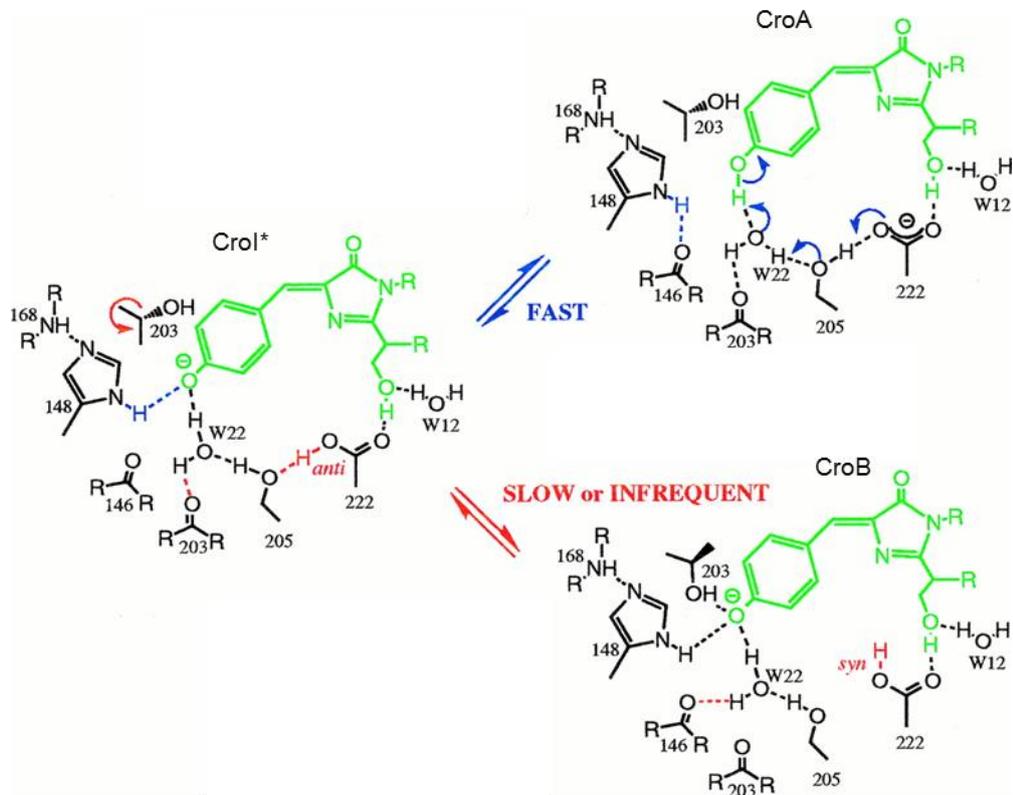
Figure 1.6 Suggested mechanism for the excited state protein transfer (ESPT) of GFP based on the structural data and the spectroscopic work.[141] CroA is the predominant form seen in the WT structure, CroB is the predominant form seen in the S65T structure, and CroI* is the intermediate not seen in either of the two structures. Changes from A to I are indicated in blue, and differences between I and B are in red. Modified from[140]

1.5.2 Engineering GFP

GFP has been extensively engineered and mutated since its discovery to increase both stability *in vivo*, and overall fluorescence.[9,127,148,149] One of the first useful variants of GFP, known as enhanced GFP (EGFP), which was generated by mutating Phe64 and Ser65 to leucine and threonine, respectively.[129] This results in a protein with increased fluorescence (due to an improved quantum yield) and improved folding at 37°C.[128,130] Later, EGFP was improved further, through the use of directed evolution[101,124,128,150] eventually leading to the generation of super-folder GFP (sfGFP)[127]. While the fluorescence properties of sfGFP were on a par with EGFP in terms of

quantum yield and brightness, sfGFP folded and matured more rapidly. It should be noted here that the reported molar absorbance coefficient by Pedelacq et al[127] for sfGFP is much higher than that repeatedly calculated by the Jones group[115,149,151,152] (80,000 $M^{-1}cm^{-1}$ *versus* 49,000 $M^{-1}cm^{-1}$) giving the protein a much higher reported brightness than is actually observed (Figure 1.7b).
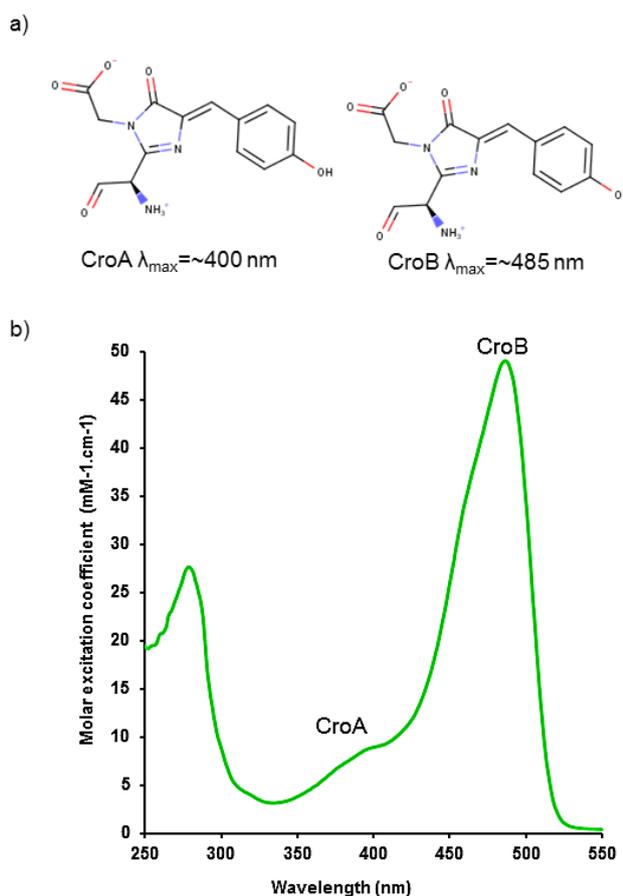
Figure 1.7 Absorbance properties of sfGFP a) Line drawings of hydroxylated and phenolate chromophores (CroA and CroB respectively). b) spectra of sfGFP[WT] highlighting which peaks are responsible for CroA and CroB in a sample.

sfGFP emits at 511nm with a quantum yield (ratio of photons emitted to photons absorbed) of 0.75. A hydrogen bond network between residues and ordered water molecules around the CRO of sfGFP ensure that the chromophore is predominantly in the CroB state. The key residues of this network include Thr203, His148 and Glu222. His148, and Thr203 form hydrogen bonds with the hydroxyl group of Tyr66, playing a

key role in deprotonating the chromophore.[115,147,153] E222 is also a key residue in deprotonation as well as ensuring correct protein folding and chromophore maturation of GFP.[77]

1.5.2.1 GFP and non-natural amino acid engineering.

As the fluorescence of GFPs is highly sensitive to changes in the local environment, it has been used extensively by this lab[59,109,110,115,149,151,152] and others[91,92,107,112,154-156] in the study of ncAA incorporation producing a number of variants with interesting and useful properties. Some of the most noticeable in terms of varying protein function were generated by Reddington *et al*.[152] Mutation of Tyr66, Phe145, and His148 with azF leads to variants that can switch on, off and change dominant excitation wavelength respectively.[152] Mutation of Tyr66 (the key chromophore residue) to azF, yields a non-fluorescent protein. Upon irradiation with UV light the azide reduces to an amine causing the protein to become fluorescent *i.e.* 'switches on'. However, mutation at residue 145 has the opposite effect. sfGFP[145azF] is fluorescent until being irradiated with UV irradiation, at which point the protein loses fluorescence and 'switches off'. The mechanism for the switching off of sfGFP[145azF], is by the azide directly crosslinking to the chromophore disrupting the conjugated double bond system hence removing the chromophore's ability to emit light.[152] The effects of His148 mutation to azF will be explored in more detail in Chapter 3. Other ncAA variants have been used to demonstrate *in vivo* labelling of cells[92] and interfacing with DNA origami tiles.[109]

1.5.2.2 GFP and nanoscience.

Although classically GFP and other fluorescent proteins are used for imaging and cell biology,[123,128,145] they have also been shown to be effective in nanotechnology as organic LEDs and lasers.[128,136,137,157,158] GFP can also act as an electron transfer system as demonstrated by.[159] This opens up extra applications for GFP to be used to create optically gated transistors and as a light capturing system for the creation of protein based solar cell systems. Korpany *et al*,[160] showed using Dronpa (a photo-

isomerisable GFP variant from coral)[161] attached to gold *via* non-covalent interactions with the His tag, that the conductance across the protein changes dependent on whether or not the chromophore is in the non-fluorescent '*trans*' isomer, or the bright 'cis' isomer. The conductance is higher in the dark form and reduces when Dronpa is activated.[160]

1.6 Aims and objectives

The central aim of this work is to show how novel chemistry incorporated at designed positions via non-canonical amino acids can be used to facilitate various non-native post-translational modifications. This will include new approaches to mimic natural processes such as modulation of protein function (Chapter 3) and engineering functionally linked protein-protein dimers (Chapter 4) that can be difficult introduce into proteins by classical protein engineering. I will also aim to use ncAA chemistry to facilitate the use of proteins in a non-natural context so adapting them for use in nanoscience. This will be demonstrated through the generation of defined protein-CNT nanohybrids for single molecule studies (Chapter 5) and ultimately for proteins as useful integrated nanoscale components. To achieve these goals, sfGFP will be used as a model protein. This is a good model, due to the well characterised and measurable fluorescence properties of sfGFP, the need for photo-controlled fluorescent proteins for high resolution imaging techniques[131,162,163] and the emerging area of autofluorescent proteins as nanocomponents. The initial stages of this work will focus on residue His148 firstly by modifying with small molecules and then by creating artificial covalent dimers. The work will then progress into the remit of interfacing sfGFP with swCNTs *via* two different approaches: photochemical and Click chemistry. Finally, sfGFP will be used to explore the effects of incorporating an ncAA at multiple residues to assess the idea of network interactions (i.e. cooperativity) between individual ncAA sites in the same protein molecule (Chapter 6).

2 Materials and Methods

Below is a description of general methods and materials used throughout this work. Any deviations from these methods will be noted in the main text.

2.1 Materials

2.1.1 Chemicals

Final working concentrations of antibiotic were as follows Carbenicillin (Melford, Suffolk, UK) 50 µg/mL, Chloramphenicol (Melford) 35 µg/mL and Tetracycline 25 µg/mL. All antibiotics were filtered through 0.2 µL syringe filter unit (VRW) and stored at -20 °C until required.

Non-canonical amino acids (ncAAs) were prepared as follows: p-azido-L-phenylalanine (azF, Bachem, Weil am Rhein, Germany) was dissolved in a minimal volume of 1M NaOH and added to growth media at a final concentration of 1 mM. S-cyclooctyne-L-lysine (SCO, Sichem, Bremen, Germany) was dissolved into 0.2 M NaOH, 15% DMSO to 100 mM and stored at -80 °C until required. Before adding to cells SCO was diluted 1:4 with 1M HEPES [pH 7.0] and added to media at a final concentration of 0.5 mM.

2.1.2 Bacterial strains

Propagation of DNA plasmids was undertaken by transforming plasmids (Section 2.2.6) into *Escherichia coli* NEB® 5-alpha (New England Biolabs (NEB), Hertfordshire, UK) cells due to their reduced DNase activity. Protein production was performed in *E. coli* TOP10™ cells (Thermo Fisher). Genotypes for both cell strains are described in Table 2.1.

Table 2.1 List of cell strains and Genotypes

E. coli cell strain	Genotype
NEB® 5-alpha	*fhuA2 D(argF-lacZ) U169 phoA glnV44 f80D(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*
TOP10™	*F- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80lacZΔM15 Δ lacX74 recA1 araD139 Δ(araleu)7697 galU galK rpsL (StrR) endA1 nupG*

2.1.3 Bacterial growth media

Super optimal condition (SOC) broth for cell recovery post DNA transformation was prepared by dissolving 31 g of Super optimal broth (SOB) powder (Melford) per 1 L of ultra-pure water. The mixture was then autoclaved and left to cool before addition of 4 % (w/v) glucose sterilised by passing through a 0.22 μm syringe filter. Liquid media Lysogeny broth (LB) and solid medium LB agar for plates were prepared by dissolving 20 g of granulated LB-Broth (Melford) or 35 g of LB Agar (Melford) respectively in 1L of ultra-pure water. Auto-induction medium ZYM-5052 [164], used to produce sfGFP variants, comprised 1% (w/v) tryptone; 0.5 %(w/v) yeast extract; 0.5% (v/v) glycerol; 0.05 % (w/v) glucose; 0.2% lactose; 25 mM Na_2HPO_4; 25 mM KH_2PO_4; 50 mM NH_4Cl; 5 mM $NaSO_4$; 2 mM $MgSO_4$; 1 x trace metals (4 μM $CaCl_2$; 2μM, $MnCl_2$, 2μM $ZnSO_4$, 0.4μM $CoCl_2$, 0.4μM $CuCl_2$, 0.4 $NiCl_2$, 0.4uM Na_2MoO_4, 0.4μM H_3BO_3 and 10μM $FeCl_3$ in ultra-pure water) and 0.05 % (w/v) L-arabinose. LB, LB-agar, and SOB were sterilised by autoclaving. For auto induction medium, powdered tryptone and yeast extract was dissolved in ultra-pure water and autoclaved. Other components of auto induction medium were sterilised by passing solutions through a 0.22 μm filter membrane and added to the yeast extract solution.

2.2 Molecular Biology

2.2.1 Expression plasmids and ncAA incorporation plasmids

Genes encoding wild type and single TAG mutations in super-folder Green Fluorescent Protein (sfGFP) equivalent to residues 26, 34, 132 148 and 204, were kindly donated by S. Reddington. The genes resided within an arabinose inducible pBAD construct.

Non-canonical amino acid (ncAA) incorporation was facilitated using the plasmids pDULE (azF) and pEVOL (SCO). The pDULE plasmids contained the engineered orthogonal tRNA and tyrosyl-tRNA synthase from *Methanococcus jannaschii.* The pEVOL plasmid contain a modified tRNA^{pyl}/pylRS pair from *Methanosarcina mazei* along with a secondary pylRS under control of an arabinose inducible pBAD promoter to boost incorporation when expressing sfGFP^{TAG} proteins[84]. These plasmids contain modified tRNA sequences to recognise the TAG stop codon and modified amino-acyl tRNA-synthases to catalyse the attachment of the specific ncAA to the modified tRNA.
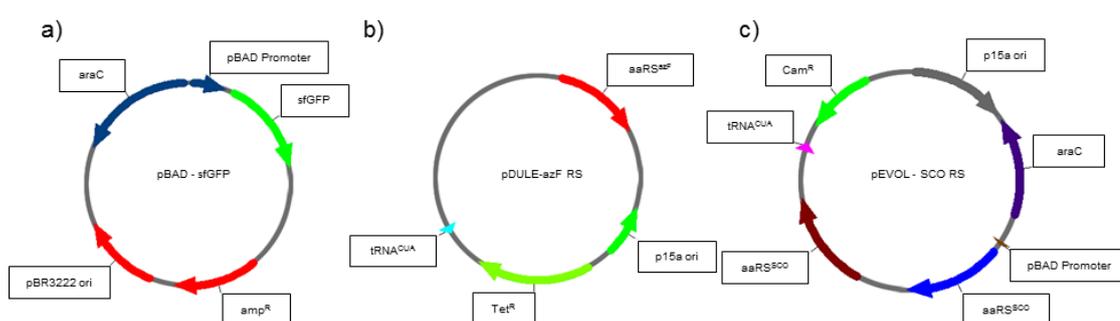
Figure 2.1: Graphical maps of plasmids used during this work. Expression plasmid a) pBAD-sfGFP and ncAA incorporation plasmids containing bioorthogonal tRNA/aaRS pairs b) pDULE-azF RS and c) pEVOL-SCO RS

2.2.2 DNA Oligonucleotides

DNA primers for mutagenesis were designed using Serial Cloner (http://serialbasics.free.fr/Serial_Cloner.html) and Amplifex (http://crn2m.univ-mrs.fr/recherche/equipe-t-brue/jullien-nicolas/programmation/amplifx/amplifx-575?lang=en) to optimise for %GC content and primer stability. Melting temperatures (Tms) were calculated using the New England Biolabs (NEB) online Tm calculator (http://tmcalculator.neb.com/#!/). DNA oligonucleotides were synthesised by Integrated DNA Technologies. A list of oligonucleotides is shown below (Table 2.2)

Table 2.2 sequences of oligonucleotides used in this work. Introduced mutations are highlighted in bold red text.

Primer Name	Sequence (5' > 3')	Tm (°C)
sfGFP26TAGF	**TAG** TTT AGC GTT CGT GGC GAA GCC	72
sfGFP26TAGR	ATG GCC ATT CAC ATC ACC ATC CAG TTC C	
sfGFP132TAGF	**TAG** GTA GGC AAC ATT CTG GGT CAT AAA CTG G	72
sfGFP132TAGR	TTT AAA ATC AAT ACC TTT CAG TTC AAT GCG GTT C	
pBAD_SeqF[1]	ATG CCA TAG CAT TTT TAT CC	N/A
pBAD_SeqR[1]	GAT TTA ATC TGT ATC AGG	

[1] Primers used for sequencing of mutated constructs. All primers were designed by Samuel C. Reddington[151]

2.2.3 DNA Purification

Purification was performed as described in the QIAgen manual. In brief, up to 800 µL of unpurified DNA from either cell lysate (using quick spin miniprep kit), or PCR mix (using quick spin PCR purification kit) was applied to the membrane. The tubes were then centrifuged at 15,000 rpm in a microfuge for 1 minute. The flow through was discarded and the sample was washed with an 80%(v/v) ethanol solution. A further two rounds of centrifugation at maximum speed (15,000 rpm; 1 minute and 3 minutes) were performed to ensure removal of any excess ethanol. DNA was eluted with 10-50 µL of molecular biology grade water. The tubes were incubated for at least 1 minute before

being eluted by centrifugation for 3 mins. The DNA was eluted into a fresh microfuge tube and kept.

2.2.4 DNA quantification

DNA concentration was determined using absorbance at 230, 260 and 280 nm. Absorbances were read using a NanoDrop® ND1000 spectrophotometer (Thermo Fisher Scientific). The NanoDrop® was blanked with molecular biology grade water and concentration of samples was determined using absorbance at 260 nm, according to the Beer-Lambert law. Absorbance at 230 and 280 nm was used to detect contamination of DNA with organic compounds, RNA or proteins.

2.2.5 Site Directed Mutagenesis

Site-directed mutagenesis was performed using whole plasmid PCR. PCR mixtures were composed of the following, 1x Q5 Reaction buffer (NEB), 200 µM dNTPs, 200 µM forward and reverse primers, 1 Unit of Q5 DNA polymerase and 0.5 ng of DNA template (at 1 ng/µL). A negative control was also performed, which contained 0.5 µL of molecular biology grade water instead of template DNA. The parameters for PCR reactions are shown below.

Step 1:   95 °C 30 s

Step 2:   95 °C 10 s

60-72 °C 30 s   x 30

72 °C 3 minutes

Step 3:   72 °C 5 minutes

PCR samples were analysed by 1% (w/v) agarose gels run in TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH8.8) containing 0.3 µg/mL ethidium bromide. Reactions which generated the expected products were purified using the QIAgen PCR purification

kit. Purified PCR product (usually 10 μl) was mixed with an equal volume of quick ligase buffer (NEB) and was phosphorylated by addition of 1 unit of polynucleotide kinase (NEB) and the mixture incubated at 37°C for 30 minutes. After phosphorylation 1 unit of Quick ligase (NEB) was added and samples incubated at room temperature for 5 mins. Ligated plasmid DNA was purified using the QIAgen reaction clean up protocol and DNA eluted in 10 μL for transformation (Section 2.2.6).

2.2.6 Bacterial transformations

E. coli NEB® 5-alpha cells were transformed with the required plasmids using the NEB heat shock protocol. In brief competent cells were incubated on ice with a total 1 μL of plasmid DNA (20-100 ng of DNA) for 30 minutes. Cells were then placed into a water bath set at 42 °C for 10 seconds before being returned to ice for 5 minutes. SOC (950 μl) was added to the mix, and the cells incubated at 37 °C in a shaking incubator (200 rpm) for one hour to recover. After incubation cells were spread onto LB Agar plates containing the appropriate selection antibiotic and incubated overnight at 37 °C.

Electroporation was used to transform plasmids into *E. coli* Top10™ (Thermo Fisher) cells; 50 μL of cells were mixed with plasmid DNA as above. Cells were transferred into an electroporation cuvette and applied with a short pulse of 2500 V. SOC was added to a total volume of 500 μL and cells left at 37 °C for 1 hour. After recovery cells were spread and incubated as above.

2.3 In silico molecular modelling

To predict the effects that modifying proteins with ncAAs and nnPTMs, sfGFP variants were modelled *in silico* using a suite of programs summed up in Table 2.3 These programs were used to design and incorporate ncAAs into proteins and then run molecular dynamics (MD) simulations.

Table 2.3 List of Software packages used during this project and their uses

Software package	Use
AMBER tools 14	File preparation for molecular dynamics
ACEPYPE	File preparation for adding to GROMACS database
Avogadro	Molecular design and input generator for GAMESS-US
GAMESS-US	Optimises bond geometries of small molecules to add to GROMACs database.
GROMACS	Energy minimisation and Molecular dynamics simulations
PyMOL	Protein viewer – inserting mutations into residues
RED	Calculates partial atomic charges on outputted structures of GAMESS
Swiss Sidechains	A PyMOL plugin used to mutate protein residues to ncAAs
VMD	Molecular visualisation program
XMGrace	Graph viewer for Molecular dynamics outputs

2.3.1 Design and parameterisation of ncAAs and modifications

Non-canonical amino acids and azide-alkyne linkages were constructed *de novo* in 3D using Avogadro molecular viewing and editing tool [165]. Atom names were assigned via usual naming conventions using a text editor. Capping groups N-methyl and acetyl were added to the Carboxyl and amino ends of the ncAA respectively using PyMOL [166]. Capping groups serve to fill all possible valences for geometry optimisation and partial charge derivation.

Geometry optimisation and partial charge derivation was performed using the RESP ESP Derive program (RED). RED integrates geometry optimisation using GAMESS-US with the RESP and ESP server allowing to assign partial charges to an optimised structure in a single step. The AnteChamber Python Parser interfacE (ACEPYPE [167]) was used to determine bond lengths, angle sizes and dihedrals to be imported into the Amber 99sb [168] forcefield to use ncAAs with GROMAC 4.5 [169].

Parameterised ncAAs were then incorporated into PyMOL via the Swiss sidechains plugin [166] [170]. Both DBCO-amine and SCO are asymmetrical across the alkyne bond. This means that there were two possible regioisomers '*anti*' and '*syn*' which had to be modelled for both linkages. A list of all ncAAs and modifications made for this work are listed in Table 2.3. Modelled residues were also used during molecular replacement of crystal X-ray structures (See Section 2.7.2.2) so where applicable, the residue ID accepted by the protein data bank (PDB) is also included in (Table 2.4).

	Residue code for GROMACS	PDB res ID	Image
p-azido-L-phenylalanine	AZF		
p-amino-L-phenylalanine	AMF	HOX	
s-cyclooctyne-L-lysine	SCO	SCP	
azF-DBCO-amine (*anti*)	AZA	66C	
azF-DBCO-amine (*syn*)	AZB		
azF-SCO (*syn*)	LNA	N/A[1]	
azF-SCO (*anti*)	LNB	N/A[1]	
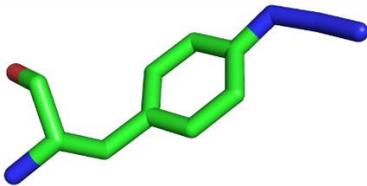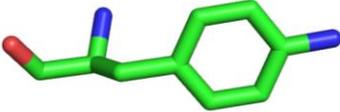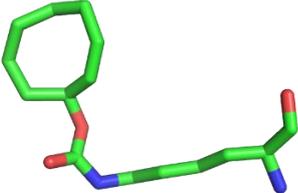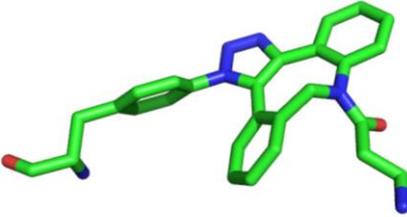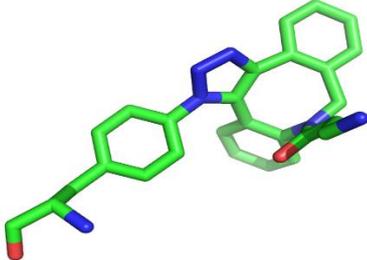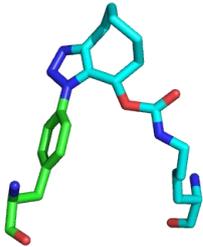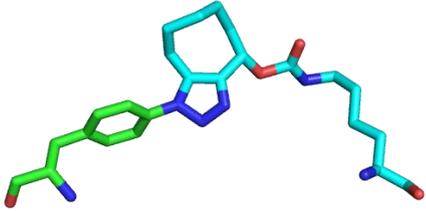
[1] Only applies to single residues for azF-[1]SCO linkages residues were split into their separate components azF (green) and SCP (cyan)

Table 2.4 List of residues used for molecular modelling with a 3D stick representation for each. Where applicable the PDB accession code is also provided.

2.3.2 Molecular dynamics with ncAAs

Models for sfGFP variants were initially based on a WT sfGFP crystal structure (2B3P), any missing residues were mutated using PyMOL mutagenesis script before the target residue was mutated to the chosen ncAA. Models were then prepared for MD via the following steps. Firstly, gmx pdb2gmx converts structures into gmx format. It also checks atoms names of residues according to the Amber99sb forcefield and adds hydrogens to all residues. The simulation box was then solvated using gmx solvate. Models were then charge balanced by adding sodium and chloride ions using gmx genion. Protein models were finally energy minimised to a lowest energy state using gmx mdrun. Energy minimisation was conducted by minimising energy to below 1000.0 kJ/mol/nm with and energy step size of 0.01 over a maximum of 50,000 steps, using the GROMACS steepest descent algorithm.MD simulations were run for a minimum of 1 ns under standard temperature and pressure (25 °C, 1 atmosphere). To assess the quality of the simulations, the RMSD across the simulation was checked to ensure they stayed in an acceptable range (0.15-0.20%).

2.3.3 Modelling of artificial dimers

To engineer protein oligomers, it is first prudent to determine which regions of proteins have a propensity to form an interface with each other. This was achieved *in silico* using a ClusPro server to samples the global interface potential (with one monomer kept constant [the receptor] and the other allowed to varied [the ligand]), combined with the Rosetta Dock application to generate final score functions [171-173]. The output sfGFP dimer models (sfGFP^{x2}) were then ranked based on total energy score. The model with the highest ClusPro cluster size in the top 5 lowest total score was used as the initial model.

ClusPro was used in multimer mode (set to dimers) using the structure of wild type sfGFP (2B3P) as a starting model. ClusPro generates ~100,000 structures and scores them using balanced energy coefficients as described by Kozakov *et al* [173] (Equation 2.1). Where E is the energy score of the complex; E_{rep} is the energy of the repulsive contribution of van der Walls interactions and E_{att} is the attractive interaction equivalent. E_{elec} is a term generated by electrostatic energy and *EDARS* is a term that mainly accounts for free energy change due to exclusion of water from the interface.

Equation 2.1: $$E = 0.40Erep + -0.40Eatt + 600Eelec + 1.00EDARS$$

The server then takes the 1000 models with the lowest scores and clusters them using pairwise to generate IRMSD (interface root mean squared deviation). Doing so creates clusters centred on the structure with the most neighbours within a 9 Å radius. Of the remaining models that do not fall within the first cluster the one with the most neighbours, is elected as the centre of the next cluster and so on until all models are part of a cluster. The centre models of each cluster are energy minimised using the CHARMM forcefield for 300 steps with fixed backbone to minimise steric clashes.

A model for each cluster was downloaded and run through ROSETTA's high resolution docking protocol. This added extra rotamers and subsequent minimisation of side chains. To docking protocol also rescores the models and adds an interface score [171,174]. The interface score is the total complex score minus the sum of the separate monomer energies and is used as a metric of how good a model is. Total score and interface score were plotted against IRMSD to highlight any outliers and remove them. The top 5 models were then used as a basis for determining which residues would be most suitable for crosslinking to form dimers.

2.4 Protein production and purification

2.4.1 Protein production and cell lysis

The two plasmids required for incorporation of ncAAs were: (1) pBAD-sfGFP^{XXXX} (Carbenicillin^R, high copy number; (Figure 2.1a), which carries desired gene mutation as denoted by the superscripted text; (2) a plasmid (pDULE or pEVOL; (Figure 2.1b +c), which carries an evolved tRNA$_{CUA}$/aaRS pair specific for the chosen ncAA. Incorporation of azF was carried out with pDULE (Tetracycline^R, medium copy number); SCO was incorporated using pEVOL (Chloramphenicol^R, medium copy number). One or the other was transformed into *E. coli* TOP10 cells as described in Section 2.1.1. A single colony from transformation was used to inoculate a 5 mL LB broth starter culture supplemented with carbenicillin (5- µg/mL) and either tetracycline (25 µg/mL) or chloramphenicol (35 µg/mL) for incorporating azF or SCO respectively. The starter culture was incubated overnight at 37 °C, 200 rpm. Expression cultures (100/500/1000 mL) of autoinduction medium (Section 2.1.3) were inoculated with starter culture (100 µL per 100 mL of expression culture) and incubated in the dark at 37 °C with shaking (200 rpm) overnight. After one hour of incubation azF or SCO was added to the culture as described in (Section 2.1.1). Negative controls for ncAA incorporation used an equivalent volume of resuspension buffer without ncAA.

Cells were pelleted by centrifugation at 5000 xg for 20 minutes at 10 °C in a Beckman JLA 16.250 rotor. Cells were then resuspended in 1/50th culture volume of lysis buffer. All lysis buffer contained protease inhibitors (cOmplete, Roche), in 50 mM Tris-HCl [pH8.0], 300 mM NaCl and 20 mM imidazole. Samples were then either lysed and purified the same day or stored at -80 °C until needed.

Before cell disruption cells were stored on ice (if stored at -80 °C cells were first thawed over ice) for 20 minutes whilst the chilled French pressure cell was assembled. A maximum of 35 mL of cell suspension was poured into the pressure cell per lysis. Cells were lysed under 1250 psi of pressure and collected into centrifuge tubes and the

soluble and insoluble fractions were separated by centrifugation, using a Beckman Coulter JA 25.5 rotor at 25,000 xg for one hour. Clarified cell lysate was then separated from the insoluble fraction.

2.4.2 Protein Purification

All protein purification columns were purchased from GE Healthcare and used with either an ÄKTA Prime Plus or ÄKTA Purifier FPLC systems. Elution from columns was monitored by absorption at 485 nm, to detect sfGFP where possible (485nm absorbance unavailable with AKTA Prime Plus), or 280 nm to detect total protein. 280 nm monitoring was switched off for elution of mutants containing an unmodified azF group to prevent premature photolysis.

Samples were concentrated using Vivaspin™ 10 kDa molecular weight cut-off spin filters (VWR). Samples were spun at 2800 x g until desired volume was reached. Buffer exchange between purification steps was done via PD10 desalting columns (GE Healthcare) according to manufacturer's instructions.

2.4.3 Metal Affinity Chromatography

The general principal of this technique involves passing crude cell lysate through a column containing sepharose beads coordinated to Ni^{2+} or in the case of Talon columns, Co^{2+}. Proteins containing a C-terminal hexa-histidine tag bind to the column via coordination with the cation. Other proteins will flow through the column. The bound proteins were eluted by applying imidazole which is the functional group of histidine and hence competes with proteins to bind to the cations. Proteins were eluted by gradually increasing the imidazole concentration from 20 mM to 250 mM.

Clarified sfGFP cell lysate was loaded onto a 5 mL His Trap™ HP column (binding capacity ~200 mg protein) equilibrated in binding buffer (50 mM Tris-HCl pH

8.0, 300 mM NaCl, 20 mM imidazole). After binding the column was washed with 20 column volumes of binding buffer. Bound protein was eluted using an imidazole gradient from 20 to 500 mM and 2 mL fractions were collected. Samples (40 µL) of each fraction were used for SDS-PAGE to check for purity.

2.4.4 Size exclusion chromatography (SEC)

A HiLoad™ 16/600Superdex™ S75 pg (preparative grade, 120 mL bed volume, separation range 3,000 Da to 75,000 Da), equilibrated in 50 mM Tris-HCl pH 8.0, was used for preparative SEC to remove final contaminants from sfGFP variants. For separation of dimers a HiLoad™ 16/600Superdex™ S200 pg (preparative grade, 120 mL bed volume, separation range 10,000 Da to 600,000 Da. Samples (0.5-2 mL) were loaded onto the column and elution was performed at a flow rate of 1 mL/min using 50 mM Tris-HCl pH 8.0. Fractions (2 mL) were collected according to absorbance at 485 nm if possible. When ÄKTA Prime Plus was used fractions were collected between elution volumes of 40 mL and 120 mL. This was to ensure that the full range of proteins were captured. Purity of fractions were assessed with SDS-PAGE and pure fractions containing target protein were pooled.

2.4.5 Ion exchange chromatography

Proteins were buffer exchanged into 50 mM Tris-HCl pH 8.0 and loaded onto a Mono Q™ anion exchange column. Proteins were eluted using a gradient of sodium chloride from 0-1 M over 20 column volumes. Fractions (2mL) were collected and analysed via SDS-PAGE.

2.5   Photolysis of sfGFP and sfGFP^{azF} variants

Photolysis was performed for 1 hour unless otherwise indicated using a 302 nm (range 275-380 nm) UVM-57 mid-range UV lamp (6 W; UVP, Cambridge, UK). Whole cell and cell lysate photolysis was performed in a petri dish on ice with UV lamp ~5 cm above dish. Resuspended cells expressing sfGFP variants were diluted with 50 mM Tris-HCl pH 8.0 to an equivalent of OD_{600} 10. Cells were poured onto a petri dish and irradiated for one hour. Aliquots (400 µl) were transferred into 5 mm x 5 mm QS Quartz cuvettes (Hellma). An aliquot prior to irradiation was also taken. Emission and excitation spectra were recorded as described in Section 2.6.4. Cell lysates of azF containing sfGFP variants were irradiated the same as with whole cells. For purified azF containing sfGFP variants, samples were diluted to either 10 µM or 0.5 µM for absorbance and fluorescence, respectively. Photolysis was performed in the appropriate Quartz cuvette held 1 cm from UV source. Fluorescence and absorbance spectra were taken before and after irradiation.

2.6   Protein analysis

2.6.1   SDS-PAGE

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to analyse protein samples using the mini PROTEAN 3 electrophoresis system (Bio-Rad). Composition of both stacking and separation gels are shown in Table 2.5. Samples were mixed with loading buffer (final concentrations: 2% (w/v) SDS, 0.2 M Tris-HCl pH 6.8, 0.04% (w/v) Bromophenol blue, 8% (w/v) glycerol and 10% (v/v) β-mercaptoethanol). Samples were heated to 95 °C for 10 mins. Whole-cell samples were centrifuged (>15,000 rpm, 5 mins) and resuspended to an equivalent OD_{600} of 10 in loading buffer. After boiling, samples were re-centrifuged to remove insoluble components. To visualise proteins, the gels were first stained in using Coomassie Blue

staining solution (40% (v/v) methanol, 10% (v/v) acetic acid and 0.1% (w/v) R250 Coomassie blue) followed by destaining with 40% (v/v) methanol and 10% (v/v) acetic acid.

Table 2.5 Components of SDS-PAGE separating and Stacking gels

	Separating gel	Stacking gel
Acrylamide/bis Acrylamide[1] (%, w/v)	15	5
Tris-HCl	0.375 mM (pH 8.8)	65 mM (pH6.8)
SDS (%, w/v)	0.1	0.2
APS (%, w/v)	0.05	0.1
TEMED (%, w/v)	0.02	0.2

[1] Acrylamide and N N'-methylene bis acrylamide in a 37.5:1 (w/w) ratio

2.6.2 Protein concentration determination: Bio-Rad DC method

Protein concentrations for new variants were calculated using the Bio-Rad *DC* Protein Assay. Aliquots of 5 µL of purified protein were pipetted in triplicate into a 96-well plate at 4 different dilutions (1x, 2x, 5x, and 10x dilution) a range of WT sfGFP samples at known concentrations (0.2-1 mg/mL) were used to calibrate the assay each time. DC assay reagents were added to each well as per instructions. The plate was gently mixed for 30 mins at room temperature and then the absorbance at 750 nm for each well was measured on a Tecan plate reader.

2.6.3 UV-visible absorption spectroscopy and calculation of extinction coefficients

UV-visible (UV-vis) absorption spectra were recorded on a Cary spectrophotometer in 1 cm pathlength cuvettes (Hellma, Müllhein, Germany). Spectra of samples were recorded from 250-700 nm at a rate of 300 nm/min. Extinction coefficients were calculated by diluting proteins down to 10 µM (as determined in Section 2.6.2) and recording full absorption spectra from 250-550 nm. Absorption and concentration values were then substituted into a rearranged version of the Beer-

Lambert law (Equation 2.2) to determine the molar extinction coefficient. Here, ε is the extinction coefficient ($M^{-1}cm^{-1}$), A is the absorbance value at λ_{max}, c is the protein concentration (M) and l is the pathlength (cm).

Equation 2.2:
$$\varepsilon = \frac{A}{cl}$$

2.6.4 Fluorescence spectroscopy

Emission and excitation spectra were determined using a Varian Cary Eclipse Fluorimeter. Samples (400 µL) were transferred into a 5 mm x 5 mm QS quartz cuvette (Hellma). Spectra were recorded at a rate of 300 nm/min with a 5 nm slit width. Emission spectra were recorded from a fixed excitation wavelength at the variant's excitation maximum (λ_{ex}) as determined from absorbance spectrum, up to 650 nm at 1 nm intervals. Excitation spectra were recorded by monitoring emission at a fixed wavelength corresponding to the wavelength at maximal emission (λ_{em}) over a range of wavelength down to 350 nm at 1 nm intervals. For whole cell fluorescence, cells were centrifuged at 5000 x g for 20 mins and resuspended to an equivalent of OD_{600} 0.1 in 50 mM Tris-HCl pH 8.0 (apart from WT sfGFP where OD_{600} = 0.025). For fluorescence analysis of cell lysate fluorescence, soluble lysate was separated from insoluble by centrifugation. The soluble fraction was then diluted as above to an equivalent of OD_{600} 0.1. For purified sfGFP and its variants, protein solution was diluted to 0.5 µM in 50 mM Tris HCl pH 8.0 for more detailed analysis (with exception of sfGFP dimers which were at 0.25 µM).

2.6.5 Quantum Yield determination

Quantum Yield (QY) is the ratio of photons emitted to photons absorbed with a theoretical maximum of 1.0. This is very difficult to determine directly so, QYs of variants

were determined by comparison with a standard molecule with a known QY using the below equation (Equation 2.2), where x is the unknown sample, s is the standard and η is the solvent refractive index. For most aqueous buffers $\eta \approx 1.33$ [115]. Fluorescein (QY = 0.75) was used as a standard due to its similar excitation and emission profiles to sfGFP. Variant samples (in 50 mM Tris-HCl pH8.0) and fluorescein dye (0.1 M NaOH) were diluted down to abs λ_{max} = 0.05. Emission spectra were then taken as described in Section 2.5.4 from λ_{ex} + 5 nm (to remove effect of incidental light scattering around λ_{ex}), to 650 nm. The integrals of the emission spectra ($\int Em$) were calculated and substituted into Equation 2.3. As there is negligible difference in refractive index between the two samples $\frac{\eta(x)}{\eta(s)} = 1$ and as such was ignored.

Equation 2.3:
$$\frac{QY(x)}{QY(s)} = \frac{\int Em(x)}{\int Em(s)} \cdot \frac{\eta(x)}{\eta(s)}$$

2.7 Structural analysis

2.7.1 Mass Spectrometry

Protein samples were diluted to 10 µM in 50 mM Tris-HCl (pH 8.0) and analysed at Cardiff University School of Chemistry's Mass Spectrometry suite. Samples were recorded by liquid chromatography time of flight mass spectrometry (LC/TOF-MS) using a Waters Synapt G2-Si QT in positive Electrospray ionisation mode. Mass peaks between 200-2,000 Da were recorded in positive Electrospray ionisation mode using Leucine Enkephalin as a calibrant. The data were processed using MassLynx 4.1 program using the Maximum entropy 1 add on. Proteins were passed through a Waters Acquity UPLC CSH 130 C18 (80°C) and eluted using a gradient of acetonitrile (5-95%) in 0.1% formic acid over 5 minutes. Mass predictions were made using ExPASy ProtParam tool [175] substituting residues mutated to azF with phenylalanine to get a

predicted mass based on the amino acid sequence. Predicted masses were altered to account for the following, chromophore maturation (-20 Da) and accounting for intact azide group in place of the H atom at *para* position (+41 Da) and modifications *e.g.* DBCO-amine (+276 Da). To calculate predicted masses for sfGFP dimers the azF component was calculated as above. The SCO monomer was calculated by substituting the primary structure of sfGFP with lysine at the mutation position and submitted ExPASy ProtParam. The predicted mass was then altered to account for chromophore maturation and the addition of the cyclooctyne group (+150 Da). The two monomer masses were then added together to generate a predicted dimer mass.

2.7.2 Crystallography

2.7.2.1 Crystal formation and harvesting

Purified protein samples in 50 mM Tris-HCl were concentrated to 10 mg/mL. Crystal formation was screened using sitting drop vapour diffusion across a wide variety of conditions as described by the PACT *premier*™ HT-96 broad crystallisation screen (Molecular Dimensions, Suffolk, UK). Drops were set up with equal volumes of protein and crystallisation buffer (0.2 µL each). Solutions for crystallisation were dispensed using an Art Robbins Pheonix robot (Alpha Biotech, UK). Crystallisation plates were sealed and stored at 25 °C. Images were taken of crystals at day 0, 1, 3, 7, and then once a week to monitor crystal growth. Any grown crystals were harvested by adding 1 mM ethylene glycol and freezing them in liquid nitrogen. Crystals remained in liquid nitrogen during transit to Diamond Light Source, Harwell, UK for X-ray diffraction.

2.7.2.2 Data acquisition and refinement

Data were collected by Dr. Pierre Rizkallah (Cardiff Medical School, Cardiff University) at Diamond Light Source (Harwell, UK), using beam lines IO2, and IO4.

Crystals were mounted via robot onto the beamline at 100 K and irradiated with X-rays (0.9795 Å). Diffracted X-rays were detected with a Pilatus 6M-F detector. Diffraction data were integrated and reduced using the XIA2 package[176]; POINTLESS[177] was used to determine the space group; and SCALA[177] and TRUNCATE[178] were used to scale and merge the data respectively. Structures were solved by molecular replacement with PHASER, using a previously defined structure as a model (5BT0). The structures were then manually edited using COOT[179] and then refined with TLS parameters using RefMac5. To insert non-standard residues and modifications into COOT and RefMAc5 parameter files had to be generated from the geometry optimised structures made in Section 2.3.1. Optimised structures were inputted into AceDRG to generate restraint parameters for use with COOT and RefMac[180]. All programs were accessed via the CCP4 package (http://www.ccp4.ac.uk/[178])

2.7.3 CAVER analysis

Tunnels and channels are very common in most proteins. These tunnels facilitate communication between the core of the protein and the bulk solvent. The PyMOL plugin CAVER 3.0[181] was used to analyse potential tunnels for transmission of protons or water molecules. CAVER finds pipeline like voids within protein 3D structures connecting a defined pocket (e.g. the sfGFP Chromophore) to the bulk solvent. Firstly, it creates a Voronoi diagram. It generates a representation of the protein using spheres of a set radius. The sphere radius is determined by the Van de Waal radius of the smallest atom. Atoms larger than this are represented by a user defined number of spheres. The program then probes the gaps between spheres that connect from the set point (sfGFP chromophore) to the bulk solvent. To prevent the software finding tunnels everywhere a minimum probe width is set. For this work a 0.75 Å probe radius was used. This diameter should be suitable to find potential water channels and proton shuttles.

2.7.4 Size exclusion chromatography-dynamic light scattering (SEC-DLS)

Samples were passed through a Superdex™ 200 10/100 GL column (GE, Healthcare) then passed through a Zetasizer MicroV dynamic light scattering (DLS) system (Malvern) to determine particle size and approximate molecular weight. Then it was finally passed through a Viscotek Refractive Index detector to determine an accurate protein concentration based on the difference between the sample and a reference cell filled with buffer. The whole system was powered using an ÄKTA Pure FPLC system (GE Healthcare). Access to the system was provided by the Protein Technology hub at Cardiff University School of Biosciences.

An aliquot (100 µL at ~5 mg/mL) of sample was injected onto the column equilibrated in 50 mM Tris-HCl [pH 8.0]. The sample was then flowed through the system at 0.75 mL.min^{-1}. Data were analysed using the OmniSEC software suite of packages provided with the system. With SEC-DLS the Mw of a protein is derived from the hydrodynamic radius which is directly derived from the observed light scattering. Therefore, to get a more accurate estimation of the Mw calibration is required. For this work sfGFP^{WT} was used as the calibrant.

2.7.5 Thermal denaturation

To test the thermostability of sfGFP variants, protein samples were diluted to 0.5 µm in 50 µL aliquots and placed in triplicate in a Rotor-Gene Q thermocycler (Qiagen). Samples were heated from 25 °C to 95 °C at a rate of 1°C/min. Fluorescence was monitored continuously throughout monitoring emission at 510nm (470 nm excitation).

The change in fluorescence over change in temperature (-ΔF/ΔT) was plotted against temperature. The temperature where the maximum change in fluorescence occurs was determined as the protein melting point. The higher this temperature is, the more stable to protein.

2.8 Strain-promoted azide-alkyne cycloaddition (SPAAC) reaction

2.8.1 Conjugation of small molecules

SPACC reactions were carried out with both whole cells and purified protein. Whole cells were diluted to an equivalent OD_{600} of 1 in 50 mM Tris-HCl (pH 8.0) and dibenzylcyclooctyne (DBCO)-amine was added to a final concentration of 100 µM. Reactions were left overnight at room temperature under dark conditions. Purified protein was diluted to 10 µM in 50 mM Tris-HCl (pH 8.0) and DBCO-amine (in DMSO) was added as above to 100 µM concentration and left for various lengths of time at room temperature with gentle mixing. After conjugation samples were spun at >15,000 xg for 5 mins to remove any insoluble particulates. Excess click reagents were removed by using a PD10 desalting column.

Click efficiency was determined by using a fluorescent DBCO dye; either DBCO-Texas Red (Figure 2.2a) or DBCO-TAMRA (Figure 2.2b). Dye was added to 10 µM of purified sfGFP protein to a final concentration of 50 µM. Reactions were left mixing overnight at room temperature. Following incubation, samples were run on an SDS-PAGE gel as described in Section 2.6.1. Gels were imaged on a GelDoc-It (Bio-Imaging Systems) before Coomassie staining to capture bands containing fluorescent dye and then after Coomassie staining to show total protein. The digital image of fluorescent dye was then analysed with Image J. The concentration of dye clicked to the protein was derived using the intensity ratio between Clicked and unclicked dye, with the later running in the dye front. A click efficiency was then determined by comparing the concentration of clicked protein to total protein.
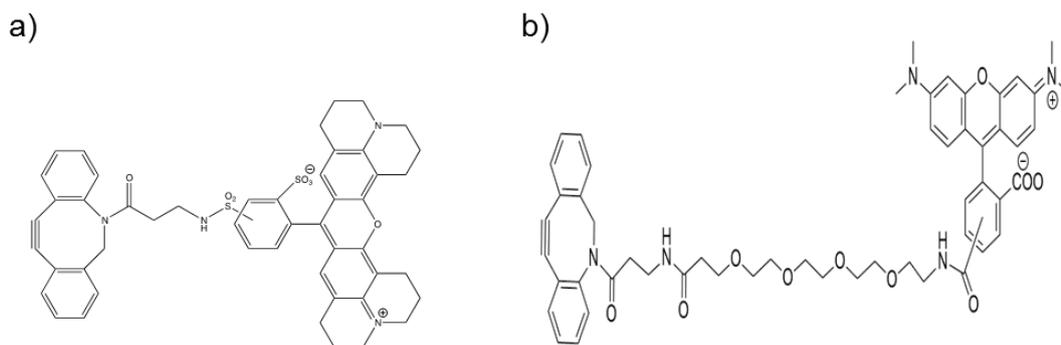
Figure 2.2 2D structures of a) DBCO-TEXAS RED and b) DBCO-TAMRA dyes used throughout this work.

2.8.2 Creation of artificial homo dimers

Purified sfGFP^{148azF} (100 µM) and sfGFP$^{148\ SCO}$ (100 µM) were mixed together in equal volume overnight at room temperature. Samples of monomer and dimer mix were analysed on SDS-PAGE to determine if the conjugation worked. Successful dimer reactions were purified via size exclusion as described in Section 2.4.4 using a HiLoad 26/600 Superdex S200 pg. After purification, absorbance and fluorescence spectra were taken of purified dimers (Sections 2.6.3, 2.6.4). Mass spec (Section 2.7.1) and crystallography trials (Section 2.7.2.1) were also undertaken.

3 Modulation of sfGFP structure and function *via* non-native post-translational modification

3.1 Introduction

Expansion of the genetic code to allow non-canonical amino acids (ncAAs) to be incorporated into proteins (reviewed extensively in,[122,182-184] and in Chapter 1) has allowed for novel approaches for precise protein post translational modifications (PTMs).[93,121] PTMs are commonly observed in nature, leading to addition of new chemistry that can modulate and supplement protein function.[115] Arguably the most important PTMs are covalent modifications *e.g.* phosphorylation of proteins, which acts as a highly responsive protein on/off switch,[6] or enzymatic cleavage of proteins in the clotting cascade.[3] The impact from these modifications varies from introducing conformational changes to the protein structure to addition of new chemical properties not available within the standard 20 amino acids.[115,185] The incorporation of ncAAs had a new facet to PTM through directed and defined modification of the protein through both the new chemistry itself[78,94,107,115,186-189] or covalent modification (termed bioorthogonal modification). Primarily, modification of ncAAs incorporated into proteins has been used to passively label proteins with useful reporter molecules, such as fluorescent probes,[91,121,190] as opposed to direct modulation of the protein. In this chapter, I will demonstrate how the use of ncAAs in conjunction with site-specific bioorthogonal modification can modulate protein function through induced conformational changes.

A versatile ncAA that has been used previously is p-azido-L-phenylalanine (azF). It introduces the phenyl azide reactive group into proteins,[94,155] which has classically been used as a tool in biochemistry as a means of photo-crosslinking useful moieties to proteins and other biologically relevant molecules.[85,191] As mentioned in section 1.4.3, there are two useful chemistries of phenyl azides that are of use to engineering new

PTM approaches: click chemistry and photoactivation. With regards to photochemistry, on irradiation with UV light a nitrene radical is formed following the loss of N_2. This radical can then follow various pathways (Section 1.4.2)[94,115,148] that can lead to functional changes within proteins.[44,110,149,152] An alternative approach is to use strain promoted azide-alkyne cycloaddition (SPAAC).[107,117] As described previously (Section 1.4.3) SPAAC is a reaction that is both biocompatible and bioorthogonal and allows for the labelling of proteins and other biomolecules, with a wide variety of adducts both *in vitro* and *in vivo*.[84,93,122,192] Genetic encoding of azF into proteins,[94] allows precise placement of phenyl azide chemistry within proteins, which enables opportunities to modulate protein function *via* either light or covalent modification.[115] There are several studies that have focussed on the molecular basis of photochemical control.[78,102,112,115] However, there is no data save for theoretical modelling[78] on how adding SPACC reagents such as cyclooctynes effects the structure of proteins.[115]
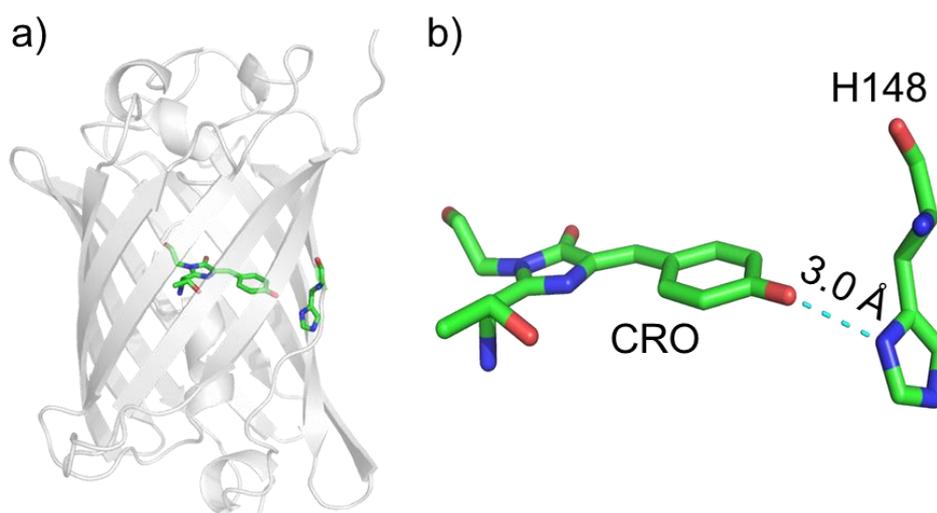
Figure 3.1 Structural position of His148 in GFP. a) Overall structure of sfGFP highlighting CRO and residue 148 as sticks. b) The proximity of sfGFP residue His148 to the Chromophore (CRO). Hydrogen bond between CRO hydroxyl group and Nitrogen of the imidazole group of His 148 shown in cyan.

As previously mentioned (Section 1.5) fluorescent proteins such as sfGFP have proven to be exemplar model systems for investigating and understanding the effects of ncAA incorporation on protein function.[183,193] Structural studies are crucial tools for understanding the molecular basis of action of ncAA engineering.[115] Such are still relatively rare for ncAA containing proteins but such information is now becoming available[85,87,110,154,194-196], especially recent work from the Jones group; structural studies of mutants that actively modulate function rarer still[115].

In this chapter, I will focus on how targeting residue His148 can lead to functional modulation of sfGFP fluorescence. His148 is proximal to the chromophore (~3.0 Å) and forms a hydrogen bond with the tyrosyl group (Figure 3.1). It has also been considered essential to the function and stability of sfGFP,[145,153] mutation of His148 to lysine and arginine (to promote deprotonation of the chromophore) significantly reduce yields of soluble folded protein.[153] Mutation of His148 also leads to an increase in 400 nm absorbance with subsequent decrease in absorbance at ~485 nm. Suggesting that the chromophore exists predominantly in the protonated CroA form as opposed to the usual phenolate CroB form, as discussed in Section 1.5 (Figure 1.5). His 148 is also partially surface accessible (117.1 Å^2) as such is a prime target for modification using SPAAC.
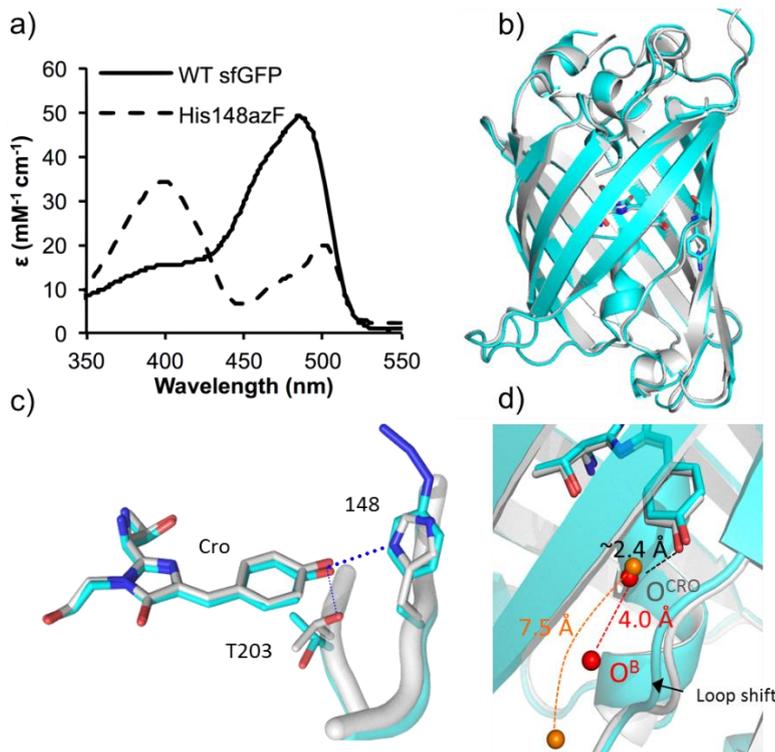
Figure 3.2 Functional and structural effects on mutating residue 148 in sfGFP to azF. a) Absorbance spectra of sfGFP ^{WT} (solid line) and sfGFP^{148azF} (dashed lines) showing photo-switching of dominant absorbance peak from 485 nm to 400 nm. b) Overlapping overall 3D structures of sfGFP^{WT} (grey) and sfGFP^{148azF} (cyan) chromophores and residue 148 are represented as sticks. c) Close-up of chromophore and residues 148 and T203, showing loss of hydrogen bonds between CRO, 148 and 203, when H148 is mutated to azF. d) Loop shift introduced by incorporation of azF at position 148 blocks a water channel from chromophore to bulk solvent hinted as cause of absorbance shift. Figure adapted from[115]

Previous work by this lab has established that azF incorporation at residue 148 alters the spectral properties of sfGFP[152]. The primary effect is a change in the ground state charged form of the chromophore (CRO) leading to change in the dominant absorbance [electronic excitation] blue shifting from 485 nm to 400 nm (Figure 3.2a). These effects are thought to be caused by the removal of a hydrogen bond between CRO and His148 leading CRO to exist predominantly in the CroA state (protonated, λ_{max} ~400 nm) with only the minority of CRO residues populating CroB (deprotonated, λ_{max} ~485 nm). A second integral hydrogen bond between the tyrosyl group of CRO and the hydroxyl group of T203, (Figure 3.2c) is also lost due to back bone shifting of β-strand 7.[115] Incorporation of azF also causes a back bone shift that brings the β-strand housing residue 148azF (7) closer to strand 6 (Figure 3.2b+d).

In this chapter I will present a structural and comparative insight into fluorescence switching by either UV irradiation or click addition of sfGFP His148 mutation to azF (sfGFP[148azF]). Firstly, I will confirm the photo-switching mechanism described by Hartley[115,186] after sfGFP[148azF] irradiation. I will also modify sfGFP[148azF] with DBCO-amine (Section 2.8.1) and investigate the effects on structure and function. Addition of DBCO-amine should introduce a new hydrogen bond donor/acceptor group via the terminal amine or the triazole formed by SPAAC linkage that can potentially lead to new interactions, networks, local conformational changes and ultimately modulate the fluorescence. Additionally, I will use computational modelling to predict the structural changes of SPAAC addition of DBCO-amine to sfGFP[148azF]. The accuracy of these predictions will be measured against the crystal structure of sfGFP[148azF]+DBCO-amine described towards the end of the chapter.

3.2 Results and Discussion

3.2.1 *In silico* modelling of sfGFP[148azF]+DBCO-amine

Computational modelling provides a relatively inexpensive and fast route to explore vast screens of mutations to proteins and their effect on protein structure. This affords researchers a prediction that they can design a hypothesis around. Theoretically, *in silico* modelling should be able to predict structural changes by reducing proteins to the lowest possible energy state then allowing them to relax using molecular dynamics.[169,197] By preforming mutations and modifications *in silico* before transferring to the wet lab allows labs to focus on a few residues saving time, labour and resources.
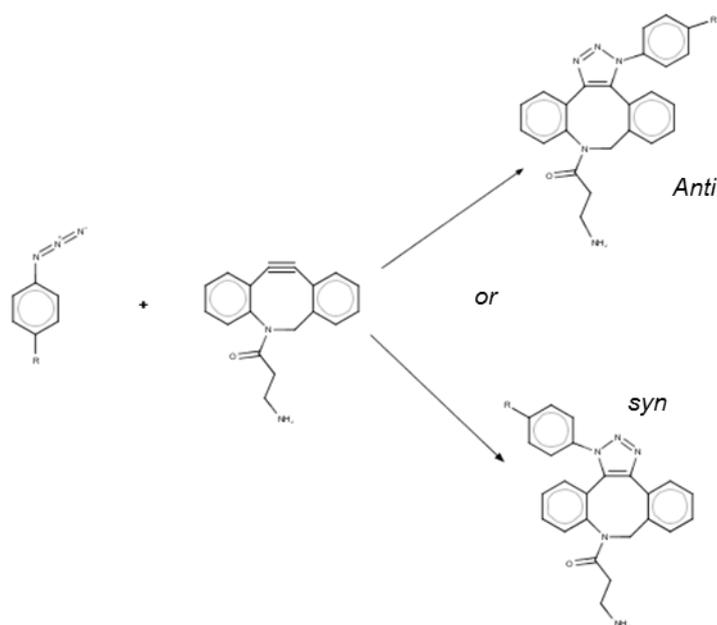
Figure 3.3 Click reaction scheme between phenyl azide and DBCO-amine creating two possible regioisomers of the complex, '*anti*' and '*syn*'

The first requirement is understanding the basic chemical outcomes and to generate the required geometry parameters for use in modelling. DBCO-amine is a heterocyclic cyclooctyne ring with two benzyl groups at C3-C4 and C7-C8. The amine side chain is connected to the molecule *via* a N atom at position 5 directly opposite the alkyne bond causing the molecule to be asymmetric. There are two possible regioisomers that can occur during the SPAAC reaction '*syn*' and '*anti*' (Figure 3.3), and hence both isoforms were modelled.

Before models of sfGFP could be generated with DBCO-amine attached, residues had to be designed and parameterised for mutation with PyMOL[166] and MD simulation simulations with GROMACS[169,198,199] as described in Section 2.3.1. The linkage residues for both '*syn*' and '*anti*' were designed, using Avogadro[165] chemical design software, as single ncAAs [*i.e.* azF-triazole-DBCO]. To ensure that all atom partial charges and bond lengths, angles, and dihedrals between linking components were energetically sensible, artificial residues were submitted to geometry optimisation

(GAMES-US),[200,201] and partial charge calculation (RED).[202-204] ACEPYPE[167] was then used to extract parameters for use with GROMACS.
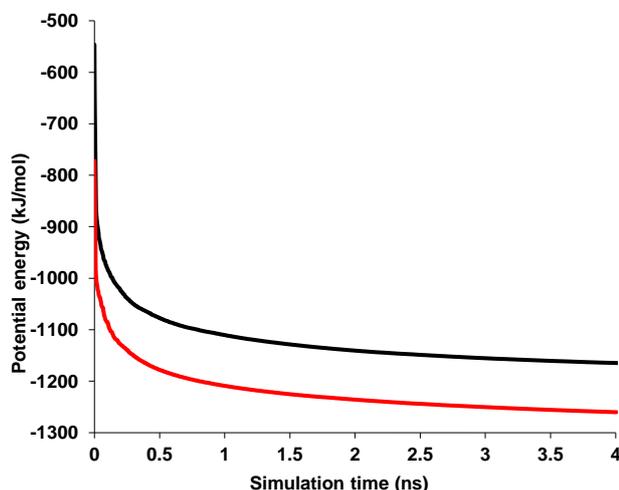
Figure 3.4 Energy minimisation curves over 4 ns for models of '*syn*' and '*anti*' regioisomers sfGFP[148azF] + DBCO-amine black and red respectively.

In silico models were generated of sfGFP148[azF] + DBCO-amine *(syn)* and sfGFP[148azF] + DBCO-amine(*anti*), using the crystal structure for sfGFP[148azF] (PDB ID: 5BT0), as a starting model as described in Section 2.3.2. Models were successfully energy minimised to convergence, *i.e.* they had reached a point where the change in potential energy approached zero. This is shown in the energy minimisation curves in Figure 3.4, where the change in potential energy tends towards zero. Of the two models sfGFP[148azF] DBCO-amine (*anti*), relaxed to a lower overall potential energy of *circa* -1250 kJ/mol than sfGFP[148azF]+DBCO-amine (*syn*), with a final potential energy of *circa* -1150 kJ/mol. This would suggest that the '*anti*' conformation is the more energetically favourable in this instance. Molecular dynamics was then run for 100 ps as outlined in Section 2.3.2 to 'relax' the protein structure and remove any steric clashes missed by energy minimisation. As described earlier in this chapter, incorporation of azF at residue 148 causes a backbone shift of β strand 7 that contributes to the reduction and blue shift

in absorbance [electronic excitation]. The '*syn'* and *'anti'* isomers were compared to the structure of sfGFP ^{WT} (2B3P), to determine if the changes caused by azF incorporation were reversed and any other structural alterations that might occur.
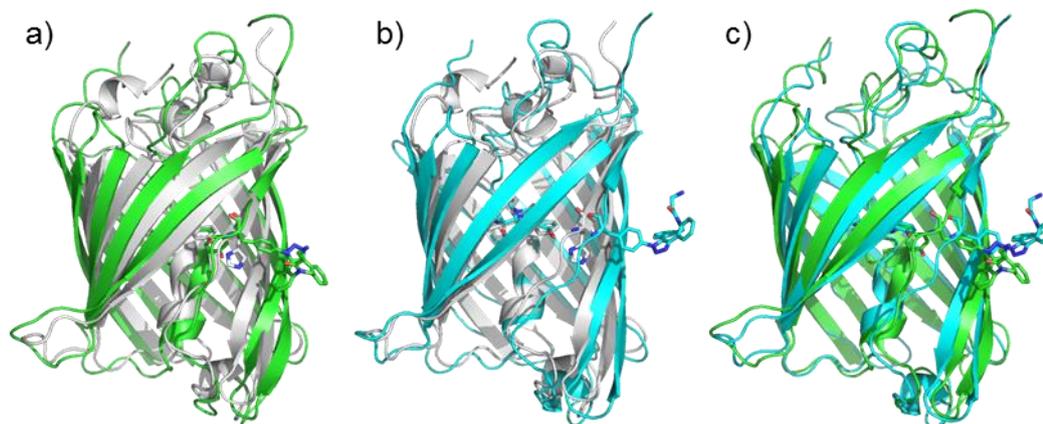
Figure 3.5 Backbone alignments of sfGFP^{WT} and *in silico* models. a) sfGFP^{WT} (grey) v sfGFP^{148azF}+DBCO-amine$^{'anti'}$ (green), b) sfGFP^{WT} (grey) v sfGFP^{148azF}+DBCO-amine$^{'syn'}$ (cyan), c) sfGFP^{148azF}+DBCO-amine$^{'anti'}$ v sfGFP^{148azF}+DBCO-amine$^{'syn'}$.

Overall, modelling suggested that attachment via either the *syn* or *anti* regioisomer had a large effect on the global structure of the protein. Alignment of the three structures (Figure 3.5), with sfGFP^{WT} shows that the backbone has changed. The RMSDs of the *'syn'* and *'anti'* regioisomers compared to sfGFP^{WT} being 1.51 Å and 1.92 Å respectively, suggesting differences between the models and the sfGFP^{WT} structure. One key difference between the '*syn'* and *'anti'* models is that β-strand 7 housing residue 148 is shifted in the *'anti'* form but the *'syn'* form is in a similar position to the WT (Figure 3.6b). Another potentially significant difference is at the chromophore. Backbone alignment (to sfGFP^{WT}) of the central strand containing the chromophore (T62 – C70) showed RMSDs of 0.96 Å and 0.85 Å *for* '*syn*' and '*trans*' models respectively, suggesting a large change in the functional core of the protein. This has the potential to significantly alter the hydrogen bonding network of sfGFP, causing alterations to fluorescence. The largest differences occur within the chromophore itself, (Figure 3.6a)

both models are distorted from the ideal planar structure with the most severe distortion arising from the *'syn'* model. If this chromophore conformation is accurate, it would disrupt the function of sfGFP^{148azF} +DBCO-amine by the conjugated double bond system. However, it is also possible that the simulation has caught the chromophore switching between cis and trans isomers, however this is unlikely as cis-trans isomerisation was not parameterised.
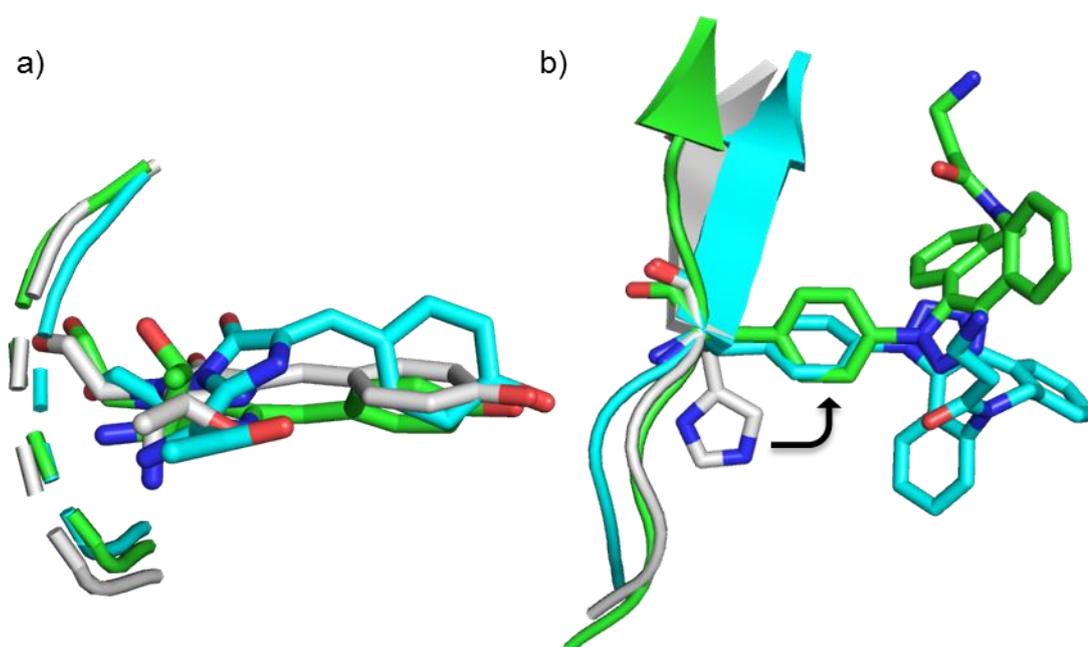
Figure 3.6 Structural comparison of sfGFP WT (grey) and 148 DBCO-amine models in '*syn*' and '*anti*' conformations (cyan and green respectively) a) Aligned chromophores and b) Alignment of β strand 7 with residue 148 represented as sticks.

The backbone of β-strand 7 (N144 – T153) for both models were aligned to sfGFP^{WT} to observe any shifts in the strand containing the modified azF residue at 148 compared to the wild type. The RMSDs for *'syn'* and *'anti'* were 0.89 Å and 0.94 Å respectively, showing some significant changes to β-strand 7. Residue 148-DBCOamine in both *'syn'* and *'anti'* models is rotated about 90° anticlockwise about the C$_\alpha$ χ1 bond

compared to the orientation of His148 in the sfGFP^WT structure. This would imply that the rotation of this residue would be responsible for the shift of β strand 7.

The above structures have provided us with a hypothesis as to how DBCO-amine modification would affect sfGFP^148azF. It also shows that DBCO-amine will more likely attach in the 'anti' conformation as this conformation had the lowest potential energy. To test our ability to predict the structural outcomes of modifications we must create sfGFP^148azF+DBCO-amine.

3.2.2 Expression of sfGFP^148azF in *E. coli* cells

To probe how modification of sfGFP^148azF with DBCO amine affects the structure-function relationship, we need to produce the basic starting protein: sfGFP^148azF. To do this I will use the plasmid encoding the gene for sfGFP^H148TAG described in [152]. We also need to determine that only azF is being incorporated at residue 148. To do this we expressed sfGFP^148azF in presence and absence of azF and undertook mass analysis.
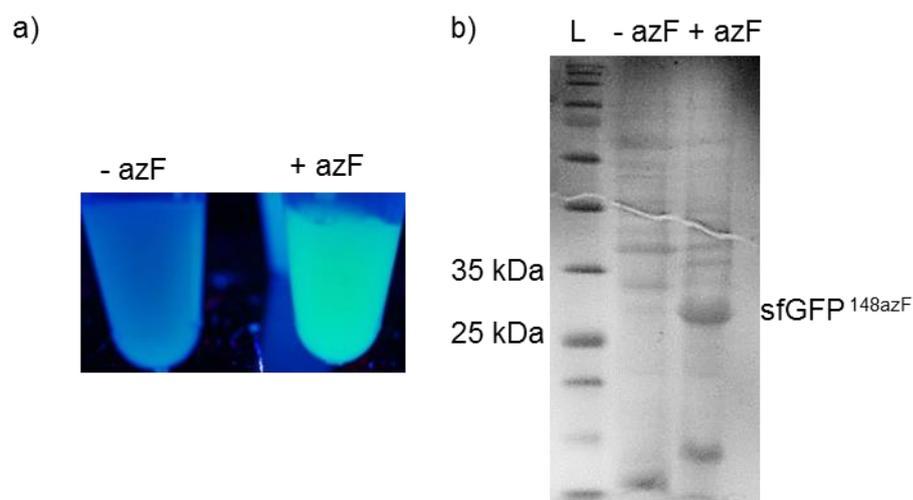
Figure 3.7 Expression test of sfGFP^148azF in presence and absence of azF (+ve and -ve respectively). a) Resuspended cell cultures under UV light showing green fluorescence in presence of azF but not without. b) SDS-PAGE gel showing expression of a protein ~27 kDa in + azF sample only.

SfGFP^{148azF} was successfully produced only in the presence of azF. As shown in Figure 3.7a, sfGFP^{148azF} cells had an observable green colour only in the presence of azF, indicative of the production of functional sfGFP; cells cultured without azF exhibited no observable colour beyond that of the natural colour of *E. coli* cells. Analysis of cell contents by SDS-PAGE (Figure 3.7b) shows a band in the + azF cell culture at about 27 kDa that was not present in the negative control (- azF). This is enough evidence to conclude that cells expressing sfGFP^{148azF} can produce functional protein and confirms that only azF is being incorporated at the 148TAG position.

Fluorescence in whole cells prior to lysis, shows a significant change in the excitation profile of sfGFP^{148azF} compared to sfGFP^{WT} (Figure 3.8). The most significant difference is sfGFP^{148azF} has a dominant excitation peak at 400 nm and a secondary peak at 500 nm, which is *circa* 70% of the intensity of the 400 nm peak. This suggests that the protonated CroA form is dominant as opposed to sfGFP^{WT}, which shows that the CroB form is dominant, with the excitation peak at 485 nm dominant compared to 400 nm excitation. Upon UV irradiation, the 400 nm peak is significantly reduced compared to the unirradiated form suggesting that there is a decrease in the population of deprotonated chromophores. There is no major change in the ~485 nm (CroB) peak intensity suggesting an overall loss in fluorescence potentially caused by protein breakdown. There is a slight blue shift (~10 nm) from 500 nm to 490 nm with no significant effect on intensity.

In whole cells, it appears that sfGFP^{148azF} is significantly less fluorescent than sfGFP^{WT}. This could be accounted by a lower production level of azF incorporated proteins in *E. coli* compared to wild type proteins in relation to cell density. For an appropriate comparison between sfGFP^{WT} and sfGFP^{148azF} both need to be purified and compared at a standardised concentration.
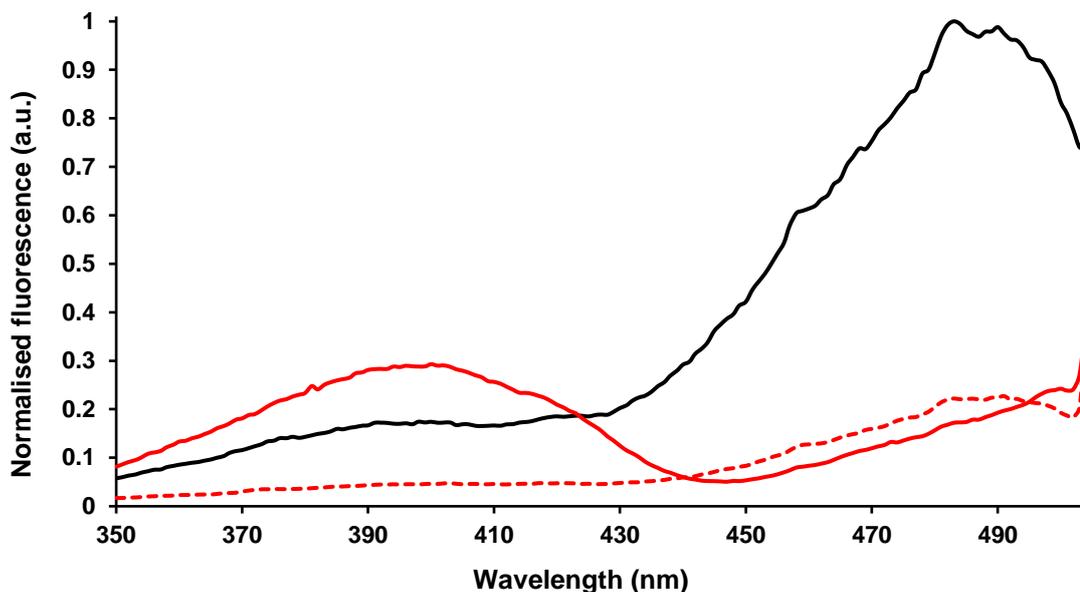
Figure 3.8 Excitation spectra monitoring emission at 511 nm of whole *E. coli* cells expressing sfGFP ^{WT} (Black) and sfGFP ^{148azF} (Red) before and after UV irradiation (solid and dashed lines respectively). UV irradiation of sfGFP ^{148azF} causes a decrease in fluorescence at the 400 nm peak and for the 500nm peak to blue shift to ~490 nm with no major change in intensity. Spectra normalised to sfGFP ^{WT}.

3.2.3 Purification and characterisation of sfGFP ^{148azF}

3.2.3.1 Purification

Before detailed analysis and modification of sfGFP ^{148azF}, the protein needs to be purified and compared to sfGFP ^{WT}. The sfGFP constructs used in this thesis have a hexahistidine tag at the C-terminal (GSHHHHHH-Cterm) for nickel affinity purification. The first step was nickel affinity chromatography, with elution monitored by UV-vis absorbance at 485 nm (Figure 3.9a). Two elution peaks with absorbance at 485 nm were observed: a minor peak at around 50 mM and the major peak at 150 mM Imidazole. Fractions (2 mL) were collected around the peaks and run on SDS PAGE along with the column flow through to assess the level of protein that didn't bind the column.
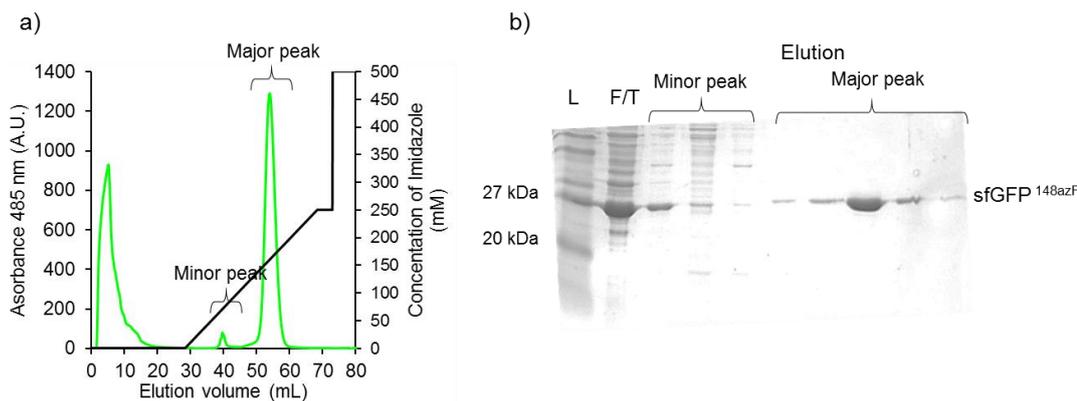
Figure 3.9 Purification of sfGFP[148azF]. a) Elution profile from nickel purification of sfGFP[148azF]. Absorbance at 485 nm in green and concentration of imidazole (mM) in Black. b) SDS-PAGE analysis.

Analysis by SDS-PAGE (Figure 3.9b) revealed that the first elution peak contained sfGFP[148azF] was contaminated with other proteins. This is due to non-specific binding of non-target proteins and sfGFP[148azF] with partially cleaved His-tags. The second elution peak contained pure sfGFP[148azF]. There also appears to be a large concentration of protein at ~27 kDa in the column flow through suggesting that the sample exceeded the column's binding capacity. The fractions representing this peak were pooled, and buffer exchanged into fresh 50 mM Tris-HCl pH 8.0 as described in Section 2.4.2. Although there is a significant band at ~27kDa in the flow through, it was deemed unnecessary to expend the time and energy on purifying the sample further.

3.2.3.2 Characterisation of sfGFP[148azF]

To study the effects of modification with alkyne adducts we need to characterise the spectral properties of the unmodified proteins. Samples of sfGFP[148azF] at 0.5 μM (fluorescence) and 10 µM (absorbance) were prepared and analysed before and after irradiation with UV light as described in Section 2.5. Purified sfGFP[148azF] shows a bimodal absorbance spectrum in the 350 nm - 550 nm range (Figure 3.10a), with λ_{max} at 400 nm and 500 nm, with the molar extinction coefficients calculated to be 34,000

and 28,000 M^{-1}.cm^{-1} respectively. This is a different peak ratio compared to the 2:1 ratio of 400 nm:500nm as described previously [152]. This suggests that some photolysis of azF has occurred as this results in a reduction in the 400 nm: 500 nm ratio.
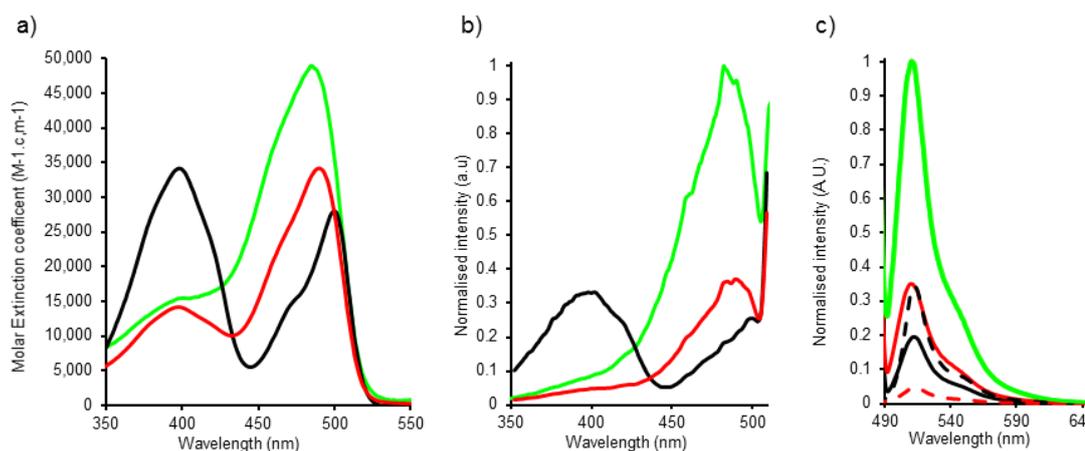
Figure 3.10 Absorbance and fluorescence spectra of sfGFP ^{WT} (green), sfGFP^{148azF} (black) and sfGFP^{148UV} (red). a) Absorbance spectra shown as Molar extinction coefficient. b) Excitation spectra showing sfGFP^{148azF} photo-switching to a dominant 490 nm excitation after irradiation. c) Emission spectra of above variants excitation was at 485, 500 and 490 nm for sfGFP ^{WT}, sfGFP^{148azF} and sfGFP^{148UV} respectively (solid lines). sfGFP^{148azF} and sfGFP^{148UV} were also excited at 400 nm (dashed lines). Fluorescence spectra were normalised to the maximum emission value of sfGFP^{WT}.

After irradiating the sample with UV light, photo-switching was observed as evidenced by the change in the 400 nm:500 nm peak ratio. The 400 nm absorbance peak was reduced by ~42% whilst the 500 nm peak blue shifted ~10 nm to 490 nm and increased to a molar extinction coefficient of ~35,000 M^{-1}.cm^{-1}. A similar trend was observed for the fluorescence excitation spectra (on monitoring emission at 511 nm) (Figure 3.10b): before irradiation sfGFP^{148azF} has a dominant excitation peak at 400 nm a secondary excitation peak at 500 nm. The photo switching between CroA and CroB forms upon UV irradiation indicates that the azide group in sfGFP^{148azF} was intact and functional. It also indicates that the protein is sensitive to UV irradiation. UV irradiation of sfGFP^{148azF} also agree with previous work showing a photo-switching between the major electronic excitation peaks of 400 nm and 490 nm.

3.2.3.3 Mass spectrometry of sfGFP[148azF]

To further confirm that the purified protein had incorporated azF at residue H148 the mass was determined by LC/MS-TOF mass spectrometry as described in Section 2.7.1. Aliquots of sfGFP[WT] and irradiated sfGFP[148azF] were also run as described. Observed masses are displayed with expected masses calculated (as described Section 2.7.1) in Table 3.1.

Table 3.1 Mass analysis of sfGFP[148azF]

Species	Calculated Mw [1] (Da)	Measured Mw (Da)
sfGFP [WT]	27827	27,826
sfGFP[148azF]	27,879	27,877
Photolysis product (no crosslink)		
Phenyl amine	27,852	
Dehydroazepine	27,850	27,865
Azepinone	27,867	

[1] Masses calculated as described in Section 2.7.1 using ExPASy MW ProtParam tool for native protein sfGFP148[F] and then adjusted for chromophore maturation (-20) and replacement of F for azF (+41).

Mass spectrometry confirmed that both sfGFP[WT] and sfGFP[148azF] (Figure 3.11a+b) matched the expected masses (27800 and 27877 Da respectively) for a full length protein, with the expected sequence and CRO maturation. Irradiated sfGFP[148azF] (Figure 3.11c) had a lower mass by 12 Da compared to the unirradiated sample. The decrease is too small to be explained by reduction of the azide group to an amine (-26 Da) (see scheme in Figure 3.11d). It is also too small to explain the formation of the dehydroazepine observed in the irradiated sfGFP[148azF] crystal structure (-28). However, the dehydroazepine is an unstable intermediate which incorporates a water molecule forming an azepine moiety. The expected mass of the azepine moiety of 27,867 Da closely matches the observed mass of 27,865 Da. This result confirms that the previous crystal structure of irradiated sfGFP[148azF] (5BTT) had captured an intermediate during the photolysis reaction. Crystals were grown under dark conditions and then were

irradiated before X-ray diffraction, as such the water molecule required for the conversion of the dehydroazepine to an azepine moiety was unable to move close enough to the residue and hence trapping the intermediate state.
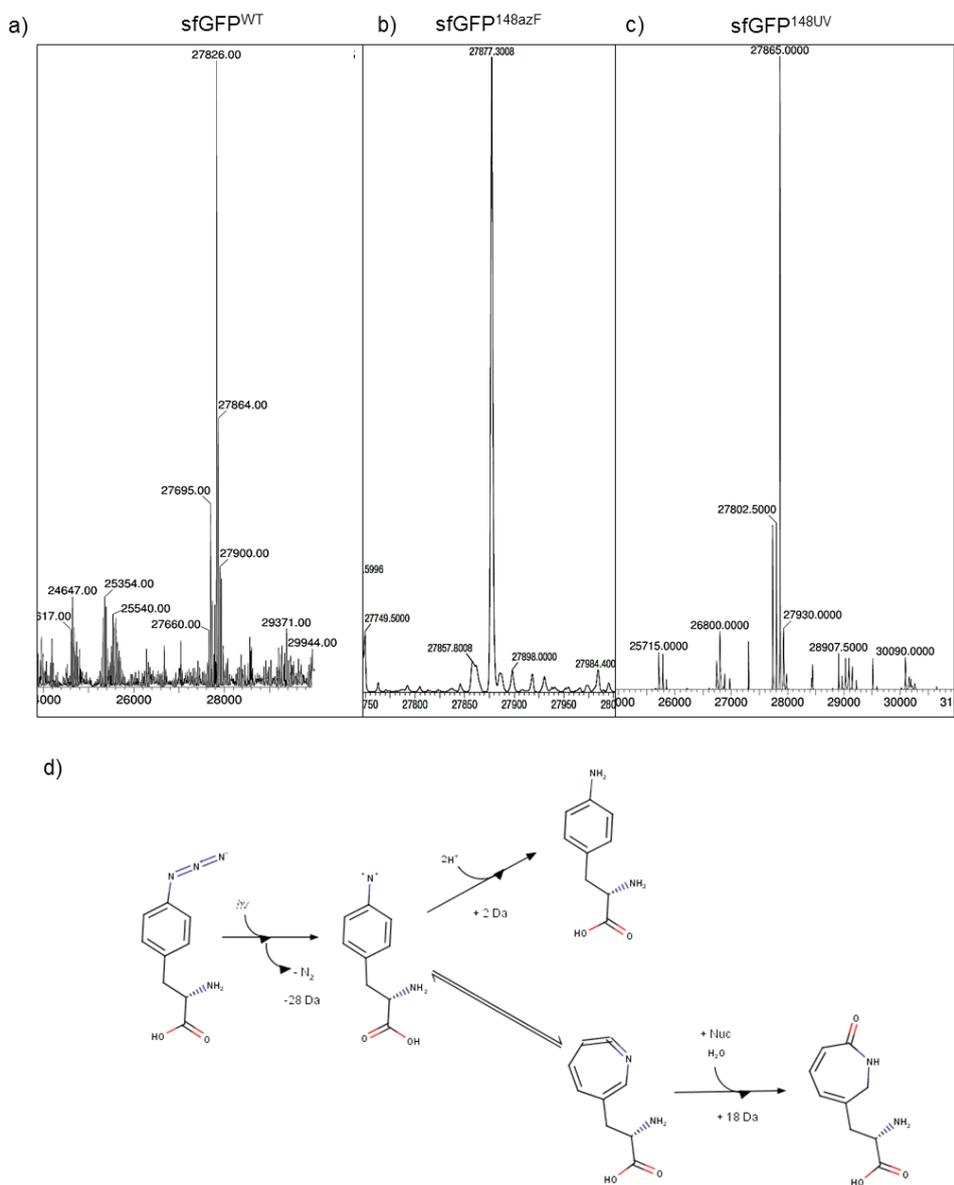
Figure 3.11 Mass Spectrometry of sfGFP variants. Chromatograms of a sfGFP[WT] b) sfGF[P148azF] and c) sfGFP[148UV]. d) Scheme showing two possible paths after formation of nitrene radical that could result in mass change in c).

3.2.4 Production, purification and properties of sfGFP^{148azF}+DBCO-amine

3.2.4.1 Copper free azide-alkyne modification of sfGFP^{148azF}

As described previously (Section 1.4.1), there are various alkyne adducts that can be attached to an azide group *via* strain promoted azide alkyne cycloaddition (SPAAC). Di-benzyl-cyclooctyne (DBCO) adducts were chosen as they react with azides without the presence of a copper catalyst. This is amenable to working *in vivo* as Cu$^+$ ions may negatively interfere with the functions of living cells. DBCO-amine was selected as the amine group introduces a potential hydrogen bond donor, which has the potential to alter to the hydrogen bonding network of sfGFP.

Fluorescent proteins including sfGFP are often used *for in vivo* imaging. Therefore, it would useful to be able to modify proteins whilst they are in cells. To determine if is possible, DBCO-amine was added to *E. coli* cells expressing sfGFP^{148azF} and left overnight at 4°C as described in Section 2.8.1. Fluorescence spectra were taken as soon as DBCO-amine was added and after overnight incubation as described (Section 2.6.4). After incubation with DBCO-amine there is a *circa* 75% decrease in fluorescence on excitation at 400 nm and a concomitant increase in fluorescence on excitation at 490 nm (Figure 3.12a). Such a change in fluorescence strongly indicates that DBCO-amine modifies protein function in the context of whole cells. This strongly suggests that *E. coli* cell membranes are permeable to DBCO-amine, allowing DBCO-amine to modify target proteins containing the bioorthogonal reaction handle within cells.

For detailed characterisation of the modified protein, purified sfGFP^{148azF} was clicked with DBCO-amine overnight as outlined in Section 2.8.1 with a tenfold molar excess of DBCO-amine (100 µM) compared to sfGFP^{148azF} (10 µM). Fluorescence spectra were taken at the start of the reaction and after incubation overnight (Figure 3.12a). Modification of isolated sfGFP^{148azF}, yields a similar change in excitation spectra as in whole cells (Figure 3.12a). There is a circa 80% decrease in fluorescence at 400 nm excitation and corresponding blue shift of the 500 nm peak to 490nm (~10 nm).

Fluorescence at the new 490nm peak increased over 3-fold in comparison to unmodified sfGFP[148azF] (Figure 3.12b).
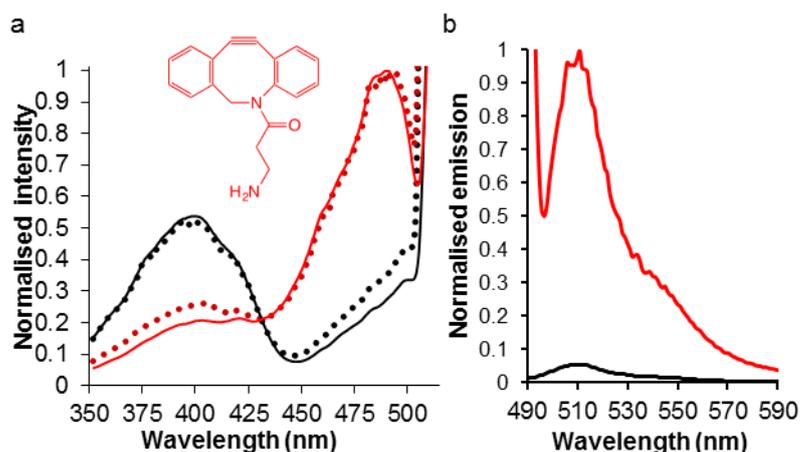
Figure 3.12 Alteration of sfGFP[148azF] fluorescence using SPAAC. a) change in fluorescence excitation *in vitro* (solid lines) and *in vivo* dashed lines in absence (black) and presence (red) of DBCO-amine (inset). b) Emission spectra of *in* vitro of sfGFP[148azF] + DBCO-amine at 400 nm excitation (black) and 490 nm excitation (red) Data are normalised to sfGFP[148azF] + DBCO-amine.

To determine the efficiency of sfGFP[148azF] modification, SPAAC was performed with a fluorescent dye DBCO-Texas Red (Figure 3.13a) as described in (Section 2.8.1). The probe was added in a 10-fold molar excess to protein and left to incubate overnight at room temperature. Efficiency of the reaction was assessed by loading the incubated sample onto an SDS-PAGE and then imaging the gel. The concentration of protein that had conjugated with dye was calculated as described (Section 2.8.1) and divided by total protein concentration to yield and efficiency percentage. According to this method, sfGFP[148azF] clicked to the dye with an efficiency of ~75%.

To determine if the described photo-switching (above) is a result of DBCO-amine attachment or modification of the protein in general, excess dye from the (sfGFP[148azF] + DBCO-Texas Red) reaction mix, was removed using a spin concentrator column and absorbance and emission spectra (λ_{ex}=400 nm) were taken. Modification of sfGFP[148azF] with DBCO-Texas red did not alter the absorbance of sfGFP[148azF] (Figure 3.13b). There

is no significant change in the 400:490 nm absorbance ratio, suggesting that it is specifically the amine side chain that causes the absorbance and fluorescence change when sfGFP[148azF] is modified with DBCO-amine. Alternatively, the bulk of the dye may block access to conformations open to the amine derivative that are responsible for the change in absorbance.
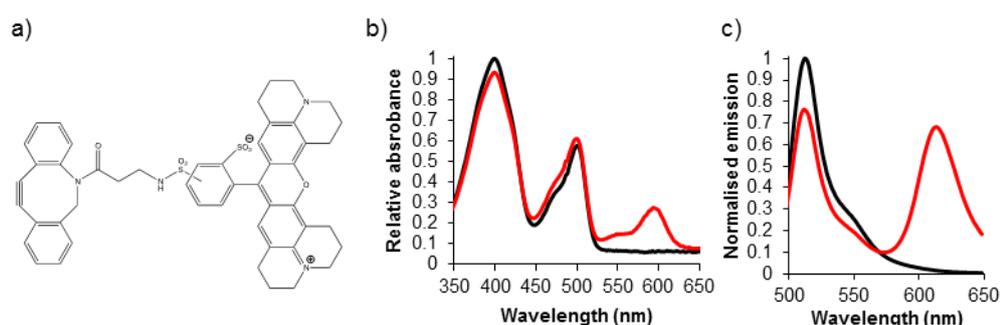
Figure 3.13 Modification with DBCO-Texas red. a) Chemical drawing of DBCO-Texas red b) Absorbance spectra of sfGFP[148azF] Unmodified (black) and modified with DBCO-Texas red c) Fluorescence emission (λ_{ex}=400 nm) coloured as in b).

3.2.4.2 Separation of modified and unmodified sfGFP[148azF]

Modified and unmodified sfGFP[148azF] protein was separated by ion exchange chromatography using a Mono Q column (GE Healthcare) and eluted via a gradient of NaCl from 0-1M. Elution of GFP was monitored by 485 nm absorbance. The chromatogram showed two distinct peaks (Figure 3.14a). Due to the small size of DBCO-amine (276 Da) there is little difference between modified and unmodified protein on an SDS-PAGE gel. Therefore, fractions of both peaks were collected and pooled separately as peak 1 and peak 2, then submitted for mass spectrometry.

The expected mass for sfGFP[148azF] modified with DBCO-amine (sfGFP[148azF] + DBCO-amine), is 28,155 Da. The first peak gave a single major mass product of 28,154 Da (Figure 3.14b). Hence, the sample contained pure sfGFP[148azF] + DBCO-amine. The second sample from ion exchange, showed two mass populations at 27,878 Da and

28,154 Da (Figure 3.14c). This meant the sample contained a mixture of both modified and unmodified protein and was discarded.
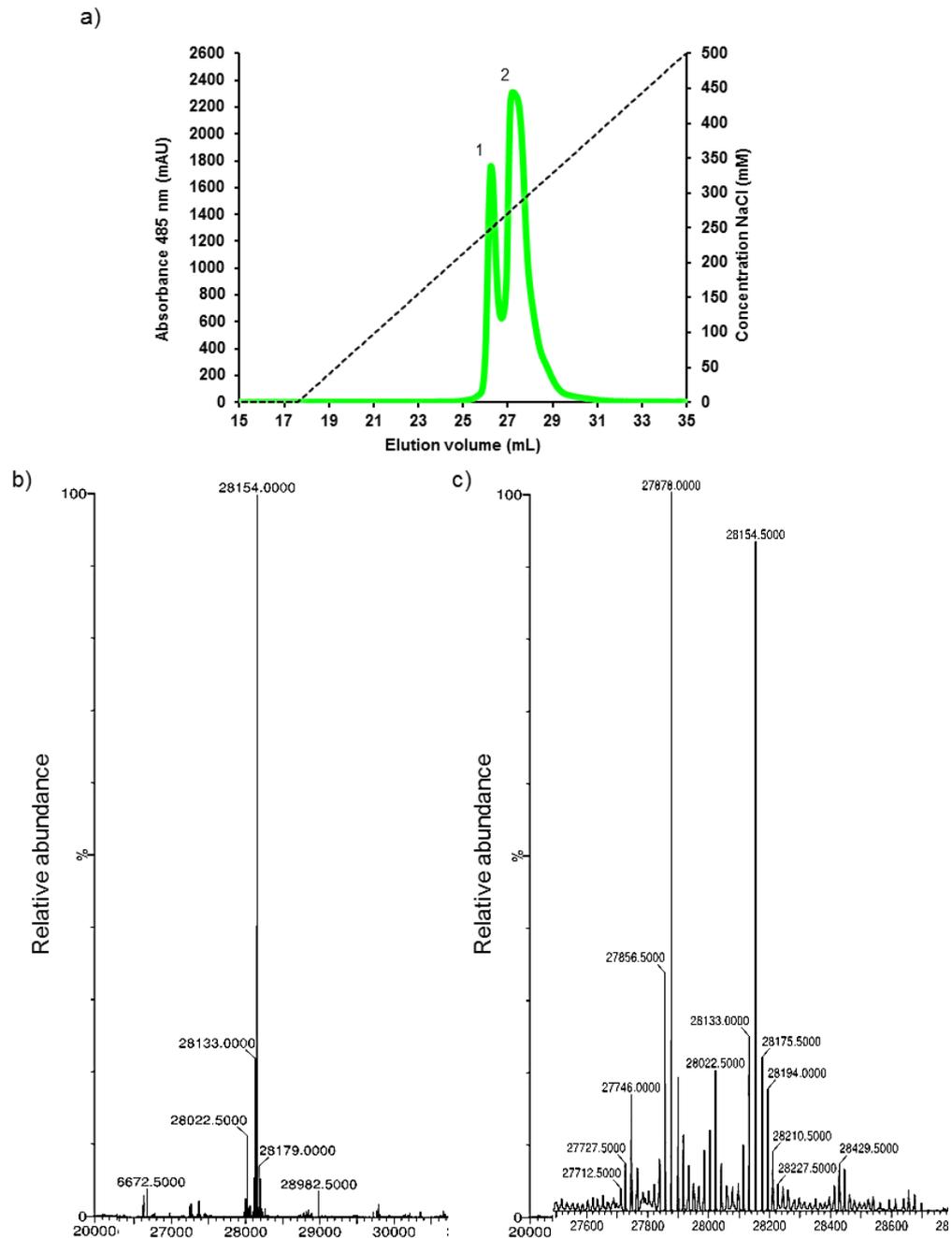
Figure 3.14 Separation of modified and unmodified sfGFP[148azF] a) Elution profile after ion exchange separation of modified and unmodified sfGFP[148azF]. Absorbance (485 nm) in green and concentration of NaCl in black. Two peaks were eluted at ~140 mM and 170 mM NaCl (peaks one and two). b) mass spectra of peak 1 from a showing single peak at 28,154 Da. c) Mass spectra of peak 2 from a showing two major peaks, at 27,878 and 28,154.

3.2.5 Spectral properties of purified sfGFP^{148azF}+DBCO-amine

The data from *in vivo* fluorescence suggests that modification of sfGFP^{148azF} with DBCO-amine reverses the effects of azF incorporation at residue 148 on sfGFP function. To determine the extent of this reversal the Absorbance and fluorescence emission spectra of purified sfGFP^{148azF}+DBCO-amine were taken. Absorbance spectra of the purified sfGFP^{148azF}+DBCO-amine shows a significant change in function with the major absorbance peaks shifting from the CroA (λ_{max} ~400 nm) to CroB (~490 nm) dominant absorbance peak (Figure 3.15a). The molar extinction coefficient at CroA (ε_{400}) decreased from 34,000 to 15,000 M^{-1}cm^{-1}. Conversely the secondary peak at 500 nm blue shifted 10 nm and ε was increased from ~25,000 to ~35,000 M^{-1}cm^{-1}. This change is greater than that observed after UV irradiation of unmodified sfGFP^{148azF} (Figure 3.15).
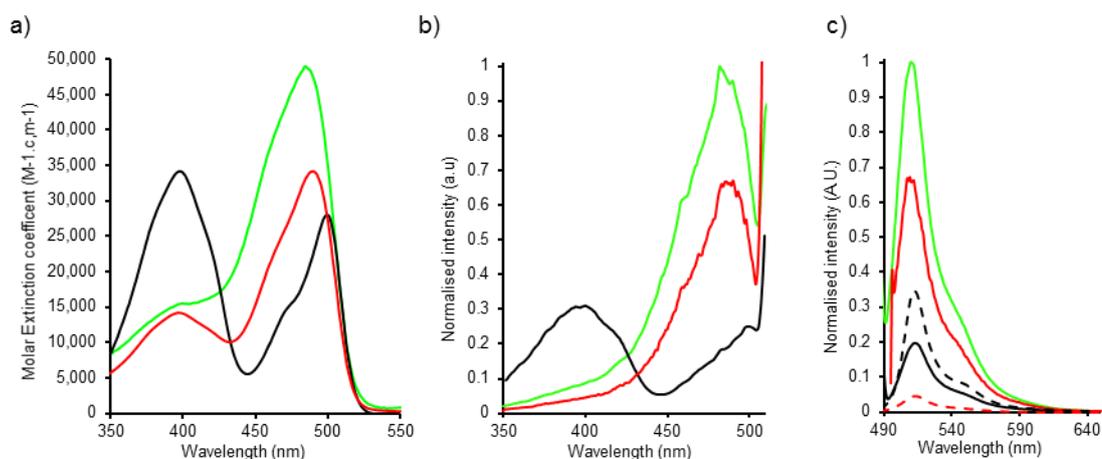
Figure 3.15 Absorbance and fluorescence spectra of modified and unmodified sfGFP148azF in vitro a) Absorbance spectra of purified sfGFP WT (green) sfGFP148azF (black) and sfGFP148azF + DBCO-amine (red). b) Excitation spectra (emission 511nm) c) Emission spectra of sfGFPWT, sfGFP148azF + DBCO-amine (Ex 485 nm, solid line; Ex 400, dashed line) and sfGFP148azF (Ex 500nm, solid line; ex 400 nm, dashed line). Fluorescence spectra were normalised to sfGFPWT.

Despite the presence of a 400 nm absorbance peak, excitation at this wavelength results in negligible fluorescence emission. This represents a ~95% loss of emission at the less favourable 400 nm excitation and a concurrent ~3.5 fold increase

in emission at ~490 nm excitation (Figure 3.15c). The quantum yield and brightness of sfGFP[148azF] at 490 nm excitation improved after modification with DBCO-amine (~3 fold and ~6 fold respectively; Table 3.2). The quantum yield was also higher (85%) than sfGFP[WT] (75%) and other commonly used fluorescent proteins (*e.g.* EGFP 60%,[77] and EYFP 61%) on excitation at the preferred 490 nm.

Table 3.2 Spectral Properties of sfGFP148[azF]

sfGFP[148azF]	λ_{max} (nm)	λ_{em} (nm)	ε $(M^{-1}cm^{-1})$	Φ (%)	Brightness ($\varepsilon \times \Phi$)	
					$M^{-1}cm^{-1}$	% sfGFP
DARK[a]	400	511	34246	69	23630	64
	500	511	15769	32	5046	14
UV[a]	400	511	24024	30	7207	20
	490	511	34262	51	17474	48
+DBCO-amine	400	511	23600	N/A		
	490	511	36900	85	31365	85
sfGFP[a]	485	511	49036[c]	75	36777	100

[a] Data derived previously[115,186] [b] Excitation at 400 nm was negligible and so QY was not determined. [c] Value differs from described in Pedelacq et al[127]

3.2.6 Crystallography

As mentioned previously (Section 3.1), little structural work, save theoretical modelling,[78] has been done to elucidate the mechanisms of how nnPTMs would affect the structure of the protein. Given the interesting observations in terms of modulating protein function on attachment and the lack of molecular information to understand the underlying mechanism of action, sfGFP[148azF]+DBCO-amine was crystallised, and the structure determined using X-ray diffraction at Diamond light source (Section 2.7.2). At time of publication (2016),[115] the described structure below was the first such structure of a SPAAC modified protein.

3.2.6.1 Crystal formation and refinement

SfGFP148azF+DBCO-amine crystals appeared in 0.1M PCTP Buffer, pH 6.0, 25% PEG 1500 after 7 days of incubation at 20°C . Of the crystals submitted for X-ray diffraction the best data set diffracted to 2.66 Å. Full diffraction and refinement statistics are listed in full in Table 3.3. The crystal had a space group of $P2_12_12_1$ which is different to both light and dark states of sfGFP^{148azF} (PDB 5BT0 and 5BTT) which both have a space group of $P4_32_12_1$. This suggests that there is a significant difference with how the crystals formed. Analysis of the contents of the crystal cell suggests that there were 2 molecules in the unit cell.

The structure was successfully solved with molecular replacement using 5BT0 as a model. After initial refinement steps, residue 148 was replaced with both '*syn'* and '*anti'* regioisomers of azF + DBCO-amine triazole linkage and refined to determine which form fitted the density best. For both molecules in the unit cell, the electron density fitted the '*anti'* regioisomer only. Once the triazole linkage orientation was established, the model went through several more rounds of refinement until the R factor and R_{free} values reached convergence at 19.4% and 29.0% respectively. The structure was deposited onto the PDB database under accession code 5DY6.

<div align="center">**sfGFP^{148azF} + DBCO-amine**</div>

Data collection/reduction statistics	
Wavelength (Å)	0.97623
Beamline	Diamond I03
Space group	$P2_12_12_1$
a (Å)	42.98
b (Å)	89.38
c (Å)	122.37
Resolution range (Å)	72.18-2.66
Total reflections measured	99,142
Unique reflections	14,111
Completeness (%) (last shell)	99.2 (99.4)
I/σ (last shell)	15.0 (1.1)
R(merge)[a] (%) (last shell)	5.8 (199.6)
B(iso) from Wilson (Å^2)	100.4
Refinement statistics	
Non-H atoms	3,658
Solvent molecules	2
R-factor [b] (%)	19.4
R-free [c] (%)	29.0
RMSD bond lengths (Å)	0.016
RMSD bond angles (°)	2.113
Ramachandran Plot Statistics	
Core region (%)	94
Allowed region (%)	5
Additionally allowed region (%)	0
Disallowed Region (%)	1

Table 3.3 Statistics from X-ray diffraction refinement and final bond parameters for crystal structure of sfGFP^{148azF} + DBCO-amine

3.2.6.2 Structure and overall topology

As with the crystal structures of sfGFP^{148azF} and sfGFP^{148UV}, two molecules are observed in the unit cell. sfGFP^{148azF} and sfGFP^{148UV} show a side by side arrangement, however the protein monomers in the sfGFP^{148azF}+DBCO-amine structure, are arranged in a top to bottom conformation (Figure 3.16a). Both structures for sfGFP^{148azF} and

sfGFP[148UV] (Figure 3.16b+c respectively), have a crystal interface around residue 148 (β strands 7 and 8). This has potential to be disrupted by the increased surface area of residue 148 after modification with DBCO-amine and gives one explanation of the different monomer configuration.
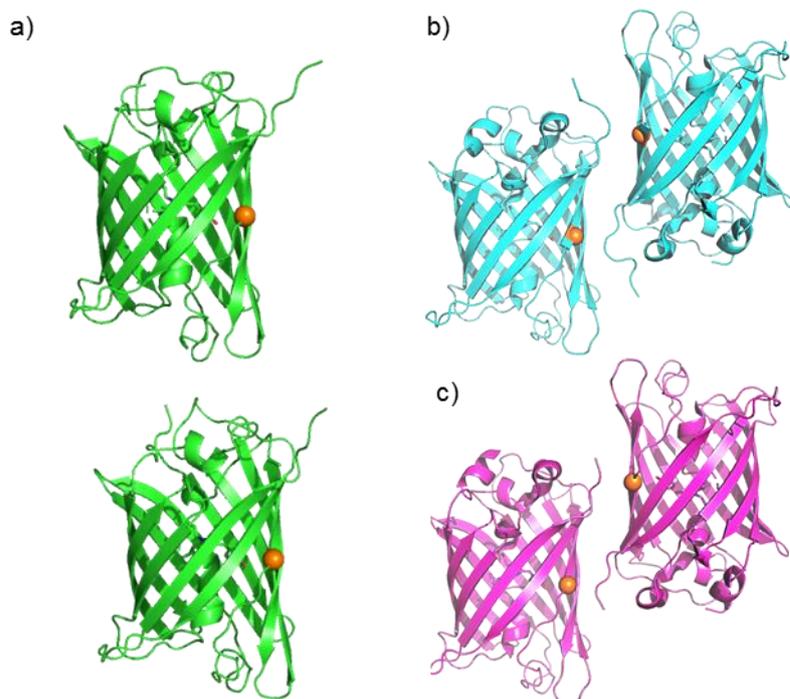
Figure 3.16 Unit cell arrangements for sfGFP[148azF] variant structures a) sfGFP[148azF] + DBCO-amine b) sfGFP[148azF] c) sfGFP[148UV] Residue 148 is shown as orange spheres.

Overall the backbone RMSD between both monomers is at 0.09 Å with very few visible differences between both chains. Both chain A and chain B monomers maintain the central β-barrel structure with no major differences to the overall protein scaffold (Figure 3.17a). Both chromophores align perfectly showing no perturbations within the core of the protein (Figure 3.17b). There are differences between the termini of both chains however these are inherently flexible regions and have no bearing on the main structure of the protein.
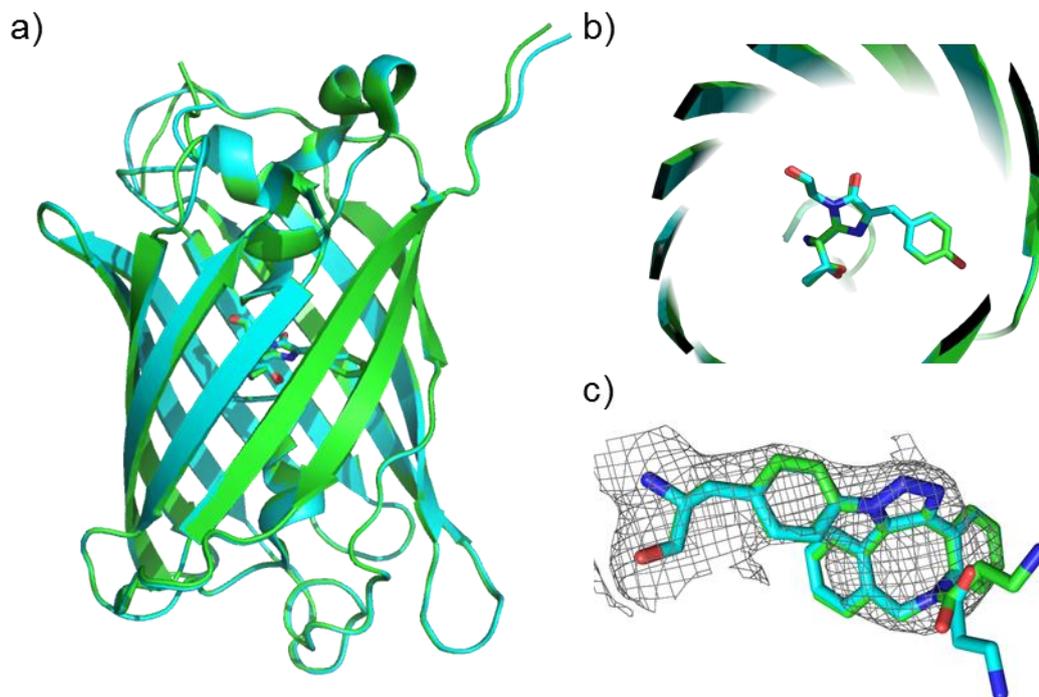
Figure 3.17 Alignment of Chain A (green) and B (cyan) of sfGFP^{148azF} + DBCO-amine crystal structure (5DY6) a) Overlay of chain A and chain B highlighting structural differences. b) Cross-section of the β-barrel showing both chromophores in perfect alignment. d) Residue 148 of both chains shows modification with DBCO-amine with little difference apart from amine side chain.

The azF–DBCO-amine triazole linkages of both chains align perfectly and fit within the electron density (Figure 3.17c). However, there was no clear density observed for the amine sidechain. This implies that there is inherent flexibility at this part of the molecule. The structure exhibits two possible rotamers for the amine side-chain. The octynyl ring forms a boat conformation and the puckered triazole linkage is rotated ~45° with respect to the plane of the aromatic side chain. The other noticeable difference between both chains is Loop P187-V193. The P187-V193 loop has an RMSD of 1.39 Å between chain A and B. This loop shift is caused by different orientations of the carbonyl group of I188 and of P192, with neither Pro187 or Pro192 changing from trans to cis peptide bonds (Figure 3.18).
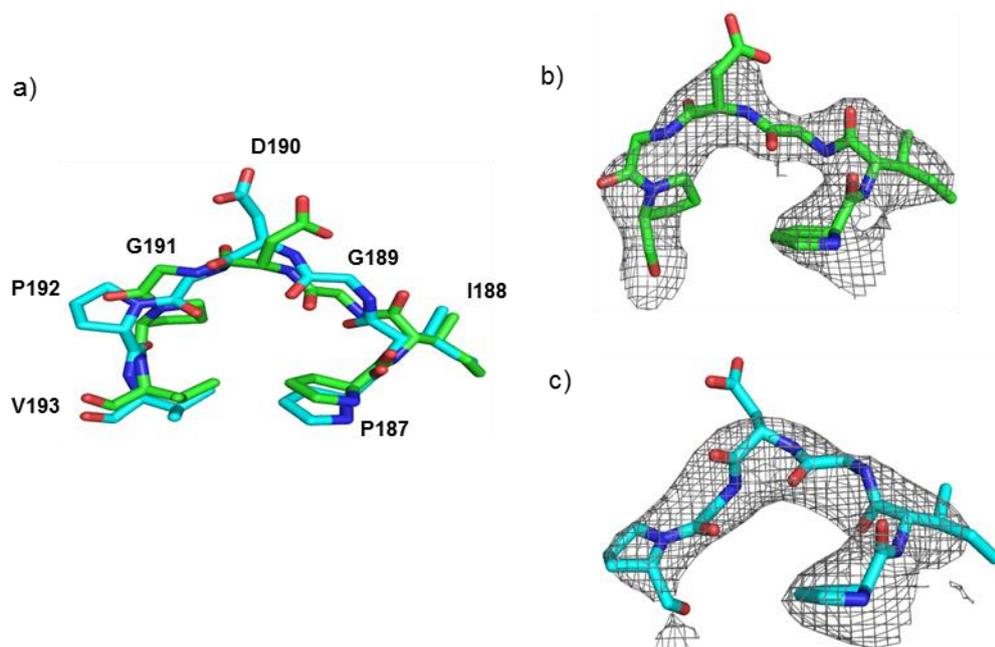
Figure 3.18 Comparison of loop P187-V193 in chain A (green) and chain B (cyan). a) overlay of both chains. Electron density maps of b) chain A and c) chain B.

3.2.6.3 Comparisons with other structures

The overall structure of DBCO-amine modified sfGFP^{148azF} is similar in terms of its global structure to the original sfGFP^{WT} (2B3P) and the unmodified sfGFP^{148azF} (5BT0). However, subtle yet significant conformational changes around residue 148 are observed that are likely to drive the functional changes observed on modification with DBCO-amine. When compared to sfGFP^{WT} (2B3P) and sfGFP^{148azF} (5BT0), the overall structures are very similar (Figure 3.19a). After backbone alignment the RMSD between sfGFP^{WT} and sfGFP^{148azF}+DBCO-amine is 0.44 Å and 0.29 Å between the modified and unmodified structures. There are however important differences with regards to local structure, especially around residue 148.
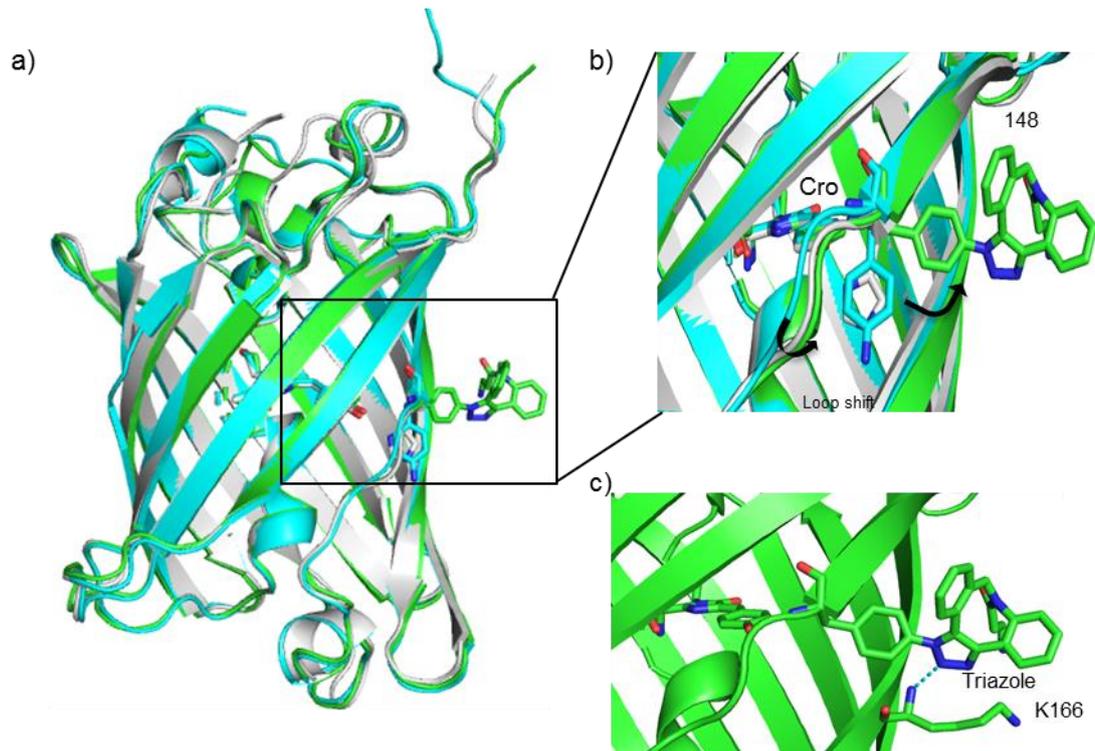
Figure 3.19 Structural overlay of DBCO-amine modified sfGFP[148azF]. a) Backbone alignment of sfGFP WT (white) and sfGFP[148azF] before and after modification with DBCO-amine (cyan and green respectively). b) Zoomed in region around residue 148. Chromophore and 148 shown as sticks. c) Potential hydrogen bond between the triazole linkage and the backbone amide of K166, stabilising a backbone shift. (Figure derived from[115]).

For both sfGFP[WT] and unmodified sfGFP[148azF] residue 148 is oriented to occupy similar positions close to the CRO (see Figure 3.19b, above), and point away from the N and C termini. The main difference between the two is that sfGFP[148azF] can no longer form a H-bond to the tyrosyl group of the CRO and the local water molecules are rearranged disrupting the conserved hydrogen bond network throughout sfGFP. Hence, disrupting the principle route of escape for the tyrosyl hydroxyl proton in the Chromophore. Modification with DBCO-amine causes a rotation of residue 148 through approximately 90° around the χ1 dihedral (Figure 3.19b). This causes a shift in the backbone β-strand 7 that houses residue 148. The strand shifts by 1.4 Å reverting to a similar configuration observed for sfGFP[WT]. A putative hydrogen bond between the triazole linkage and the backbone amide of residue K166 appears to stabilise the structural shift (Figure 3.19c). Therefore the general mechanism of action is akin to

PTMs such as phosphorylation,[115] *i.e.* small conformational changes result in alterations to a critical non-covalent interaction network.
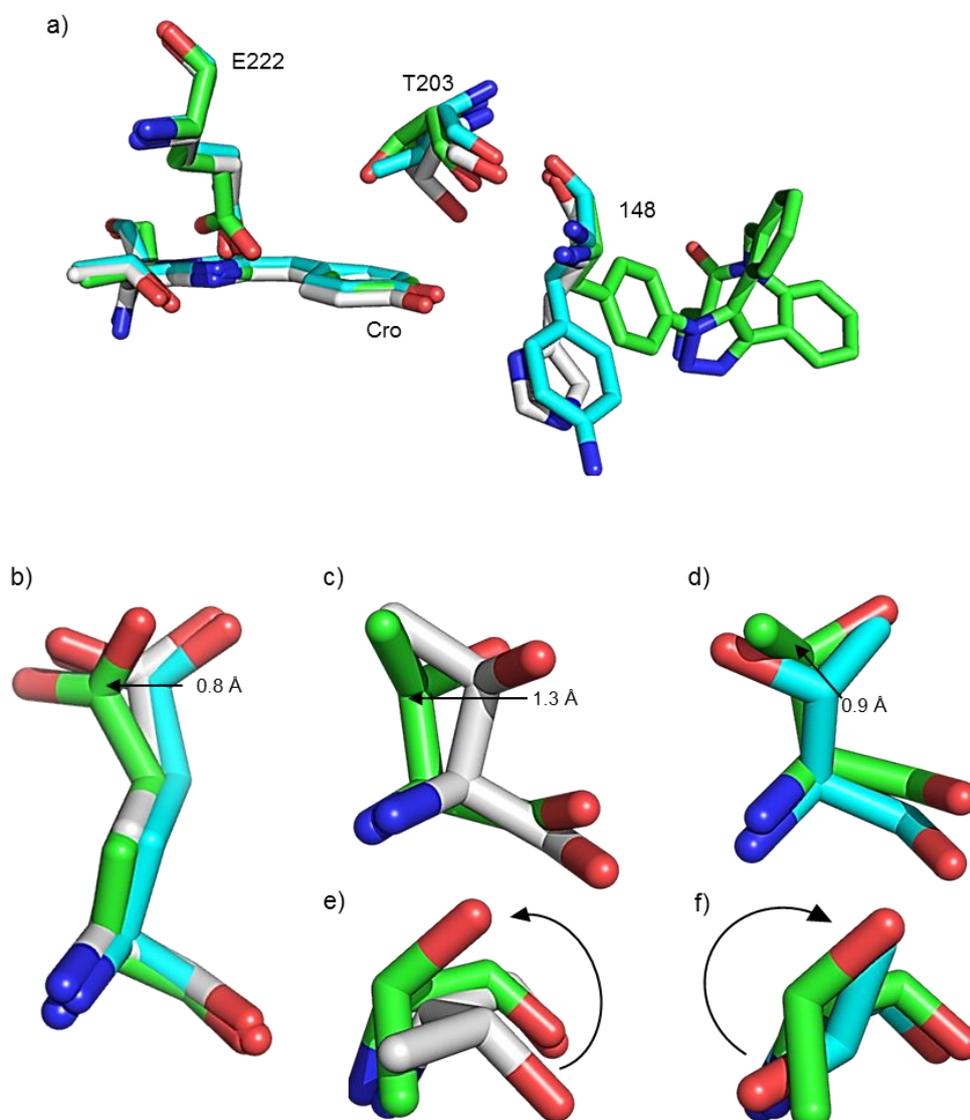
Figure 3.20 Variation of residues T203 and E222 in sfGFP[148azF] structures. a) alignment of sfGFP[WT] (grey) sfGFP[148azF] (cyan) sfGFP[148azF]+ DBCO-amine (green), highlighting T203 and E222 with CRO and residue 148 shown for reference. b) close-up view of E222 showing a 0.8 Å shift of the side chain carboxyl group between modified and unmodified sfGFP[148azF]. c+d) Side views of T203 alignment showing distance shift of C_Δ between sfGFP[WT], sfGFP[148azF] and safGFP[148azF] + DBCO-amine. Side view of T203 alignments (e+f) showing rotation of hydroxyl group.

Residues T203 and E222 are proximal to the chromophore and residue 148 (Figure 3.20a) and have both been shown to be important for the protonation state of

the chromophore.[9,77] Both residues show subtle changes in their conformation. There is little alteration in the conformation of E222 between the structures of sfGFP^{WT} and sfGFP^{148azF}. However, when sfGFP^{148azF} is modified with DBCO-amine there is a 0.8 Å shift in the delta carbon (C$_\Delta$; Figure 3.20b). As mentioned previously in this chapter T203 is shifted after mutating 148 to a phenylazide.[115] Modification with DBCO-amine induces more conformational changes of T203. The sidechain shifts ~1.3 Å and ~0.9 Å with respect to sfGFP^{WT} and sfGFP^{148azF} (respectively; Figure 3.20c+d). There is also a rotation of the side chain with respect to both structures of ~90° anticlockwise (sfGFP^{WT}; Figure 3.20e) and ~180° clockwise (sfGFP^{148azF}; Figure 3.20f). It is likely that these changes to hydrogen bonding in the DBCO-amine structure, re-establish the hydrogen bonding network around the chromophore aiding in deprotonation of the chromophore. The shift of T203 to a position near that of sfGFP^{WT} re-establishes the direct hydrogen bond between T203 and CRO directly aiding in deprotonation.

3.2.6.4   Comparison to sfGFP^{148UV}

The structural mechanism for the basis of modulating the function of sfGFP^{148azF} with UV irradiation has been previously established (see Section 3.2.3.3 and [115,186]). Given both UV irradiation and modification with DBCO-amine induce similar changes in sfGFP^{148azF} function (*i.e.* shift from 400 nm to 490 nm major excitation) through changes in ground state protonation of the CRO, the structures for both are compared below. Alignment shows that visually both structures are very similar, which is backed up by a very low backbone RMSD (Figure 3.21a; 0.26 Å).
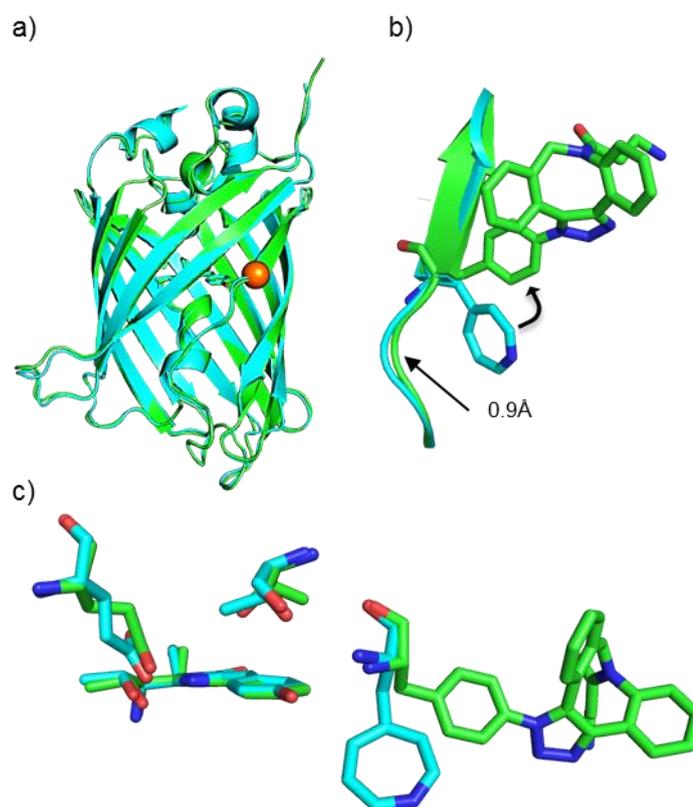
Figure 3.21 Structural comparison of sfGFP[148UV] (cyan) and sfGFP[148azF] + DBCO-amine (green). a) whole backbone alignment with residue 148 displayed as orange sphere. b) closeup of β-strand 7 with residue 148 shown as sticks. c) Chromophore region showing differences between E222, T203 and residue 148.

Alignment of β-strand 7 of both sfGFP[148UV] and sfGFP[148azF]+DBCO-amine revealed an RMSD of 0.43 Å (Figure 3.21b). The two key differences are a 0.9 Å shift in the backbone of residues N-terminal to residue 148. Residue 148 is rotated ~90° anticlockwise in the structure for sfGFP[148azF]+DBCO-amine with respect to sfGFP[148UV], which occupies a similar orientation to sfGFP[WT] and sfGFP[148azF]. The chromophore region also has a couple of differences namely alternate rotamers of T203 and E222 (Figure 3.21c). Although both T203 and E222 sidechains occupy different rotamers in the sfGFP[148UV] and sfGFP[148azF]+DBCO-amine, the hydroxyl (T203) and carboxyl (E222) occupy the same space in both structures and so there would be no difference in hydrogen bonding networks between both structures.

3.2.6.5 Caver analysis

As described earlier it is believed that the observed fluorescence changes, from incorporating azF at residue 148 occur because of disruption to hydrogen bonding around the chromophore. The most noticeable of which is a removal of the hydrogen bond between the chromophore tyrosyl group and the imidazole group of His148 when histidine is mutated to azF. When sfGFP[148azF] is modified with UV light or DBCO-amine, it is assumed that there is a rearrangement of the conserved water network causing a water molecule to replace the position that His148 occupies in sfGFP[WT] and hence promotes the CroB form of the chromophore. However, this is only an assumption because of the low resolution of the sfGFP[148azF]+DBCO-amine structure there aren't many structured water molecules.
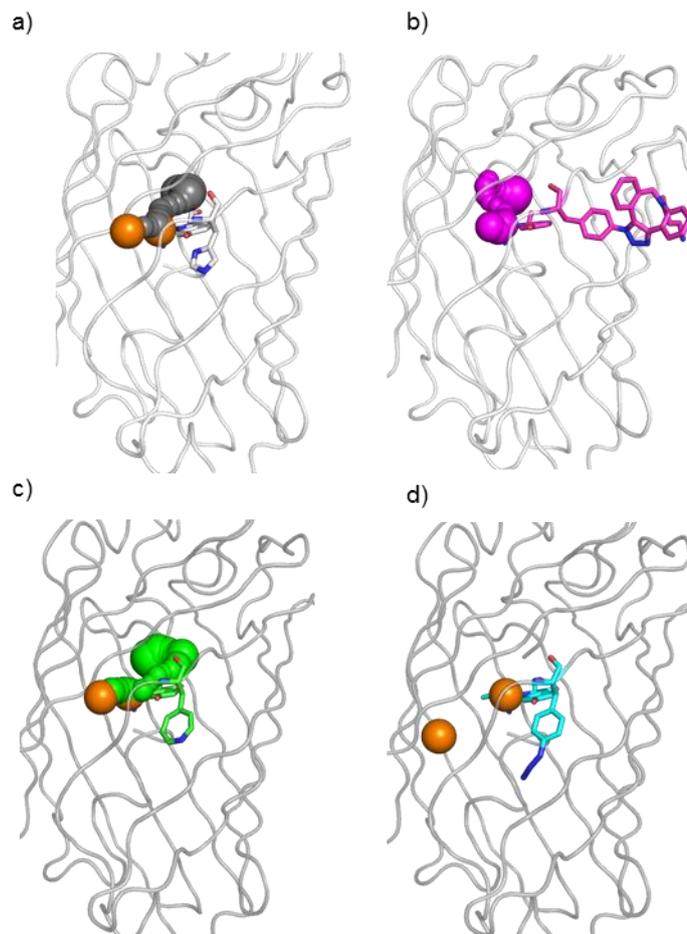
Figure 3.22 Predicted channels from caver 3.0 program exiting between β strands 6 and 7. a) Wild type sfGFP b) sfGFP[148azF] +DBCO-amine, c) sfGFP[148UV] and d) sfGFP[148azF]. Structural waters are represented by gold spheres, Tunnels are represented as chain of spheres.

The resolution for the modified structure (2.66 Å), is significantly lower than the dark and light structures of sfGFP^{148azF} (2.03 Å and 2.14 Å respectively), and over 1 Å lower resolution than sfGFP^{WT} (2B3P; 1.4 Å). Due to the lower resolution of sfGFP^{148azF}+DBCO-amine, there are few structural waters observed compared to the other structures. Therefore, it is not possible to directly determine if a water has replaced the function of His148. However, with use of the program Caver[181], it is possible to visualise a conserved channel between sfGFP^{WT,} sfGFP^{148UV}, and sfGFP^{148azF}+DBCO-amine (Figure 3.22a-c) and an implied water network. These features are absent in the sfGFP^{148azF} structure (Figure 3.22d). The region around β-strand 7 between residues 146-148 is believed to be a key proton "escape route" within the charge transfer network.[138,185] The disruption of the CRO-148 hydrogen bond, shown in the sfGFP^{148azF} structure, promotes the dominance of the CroA form of the chromophore. The reestablishment of the network by introduction of a structural water molecule promotes the CroB form. The lack of such a channel (and thus potentially water-contributing proton wire) in sfGFP^{148azF} compared to the other structures may be a significant contributing factor to its CRO being predominantly in the protonated CroA form in the ground state. Modification of the azF148 by either light or Click attachment of DBCO-amine restores the channel and thus at least in part restores the phenolate ground state of CRO.

3.2.6.6 Comparison of sfGFP^{148azF}+DBCO-amine to the '*anti*' model

As the structure for sfGFP^{148azF}+DBCO-amine shows that the triazole formed in the '*anti*' conformation it was compared against the model for '*anti*' conformation. Whole backbone alignment showed a RMSD of 1.8 Å between the two structures, which is a significant difference. Visually most of the differences occur between flexible loops at the ends of the β-barrel and at the N and C termini (Figure 3.23a). The most significant backbone perturbation is on β-strand 7, where there is a separation of 1.7 Å between both strands N terminal from Residue 148 (Figure 3.23b). Residue 148 for both

sfGFP^{148azF} + DBCO-amine structure and the *in silico* model are rotated about 90° away from the position of the WT (Figure 3.23c). This suggests that the model was able to predict the orientation of the phenylalanine side chain. However, there is not a perfect alignment between the two residues.
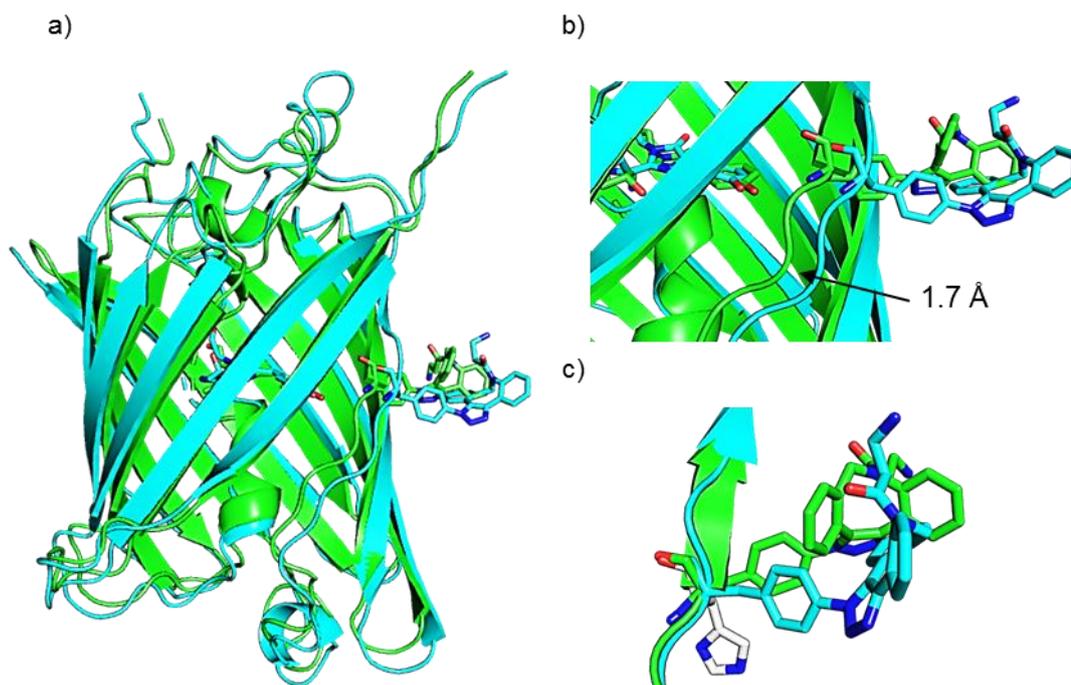
Figure 3.23 Comparison of sfGFP^{148azF} + DBCO-amine structure (Green) and *in silico* model (cyan) a) overall backbone alignment. b) Zoomed in region of β-strand 7. c) Alignment of β strand 7 His148 for sfGFP^{WT} shown in grey.

The chromophore region (T62 – C70) of the '*anti*' model was aligned to the equivalent region of the sfGFP^{148azF}+DBCO-amine and the backbone aligns comparatively well compared to the overall protein structure (RMSD 0.9 Å) but is still quite high compared to the crystal structures of sfGFP^{148azF} and sfGFP^{148UV} (Figure 3.24a). The chromophores do not align properly with an offset of *circa* 1.3 Å (Figure 3.24b), suggesting that both chromophores occupy the centre of the β-barrel differently. The chromophores also do not occupy the same conformation, side chain of Thr65 for the '*anti*' model is rotated ~120° anticlockwise compared to the crystal structure (Figure

3.24c) and the tyrosyl group of the model chromophore is also rotated clockwise ~45° from the plane of the sfGFP[148azF]+DBCO-amine structure (Figure 3.24d). Although some these perturbations are energetically feasible and could be explained as alternate conformations. The most plausible cause is due to incorrectly parameterising the restraints within the GROMACS program. Firstly, the sfGFP chromophore is an extended conjugated double bond system and is hence required to be planar. Secondly, no other sfGFP structure to the author's knowledge shows the same orientation of the T65 component of the chromophore.
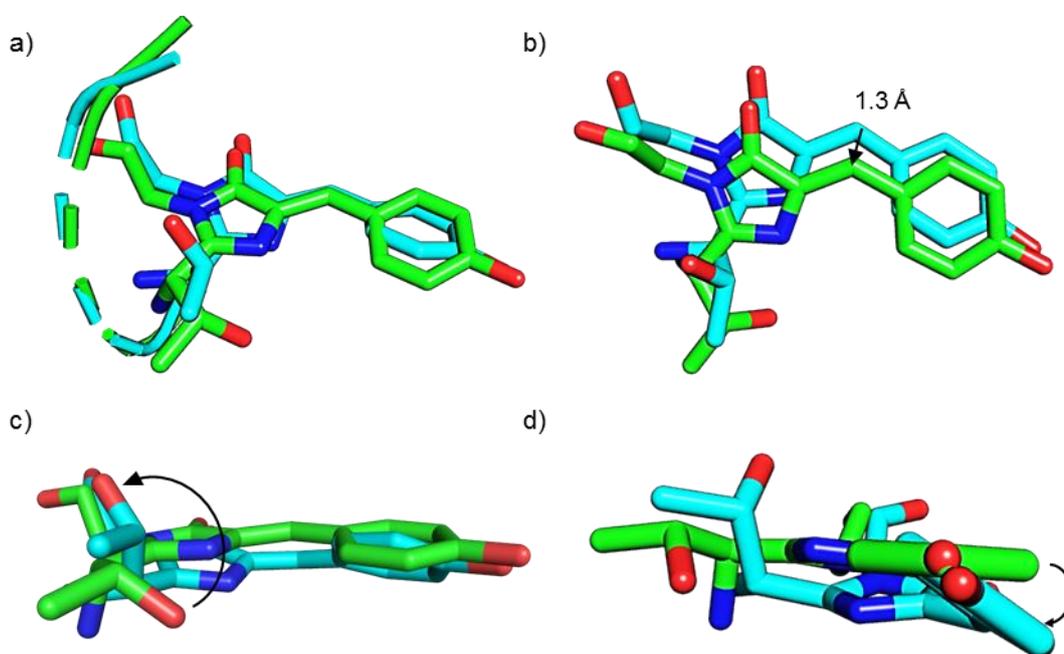
Figure 3.24 Comparison of the chromophore region of sfGFP[148azF] + DBCO amine structure (Green) and *in silico* model (cyan). a) Top view b) Top view c) Side view d) End view from the tyrosyl group.

3.3 Conclusions

As discussed at the start of this chapter, covalent PTMS are important for the fine tune regulation of protein function allowing for a more instant response to changing cellular conditions.[6] PTMs can be either; reversible such as phosphorylation[6] and acetylation[7,8] for rapid on/off regulation of proteins; or they can be irreversible such as

enzymatic cleavage in thrombin[3] and the clotting cascade, which allows a large concentration of the protein to be safely stored in the inactive form until required.[3,6] The ncAA, N^{ε}-acetyl-lysine has been used to mimic acetylation in histones to establish its role in nucleosome positioning and transcription.[106] However, this is a very specific example and is hard to transfer, with incorporation of azF we are able to transplant non-canonical PTMs into any protein system.

The versatility of genetically encoded phenyl azide chemistry has been demonstrated above to switch protein function through two different nnPTM events, bioorthogonal conjugation and photolysis with UV irradiation. In both examples, the changes in dominant protonation state of the chromophore occurred as an indirect result of altered local interactions. Namely, changes to hydrogen bonding networks around the chromophore. Local interaction networks are key to the function of all proteins and so, genetically encoded phenylazide chemistry can be used to functionally modulate proteins outside of classical photo-crosslinking pathways.[85,94,102,117,152] The structural knowledge of the mechanisms of action for both routes, will help to increase understanding of molecular pathways.

Part of this work was to see if using *in silico* modelling would be able to predict the structure of a modified protein, which was achieved to an extent as two models were successfully created. However, there were significant differences compared to both the existing sfGFP[WT] crystal structure and the newly generated structure for sfGFP[148azF]+DBCO-amine. There were also issues with the parameterisation of the chromophore *in silico, i.e.* planarity of the chromophore was not conserved. The models did however, correctly predict that the side chain for residue 148 would be rotated away from the protein after modification with DBCO-amine. Use of the models also suggested the correct regioisomer of the modification (*'anti'*). Most biological processes occur on a microsecond scale therefore it should be noted here that 1 ns may not necessarily be enough time for a simulation to properly equilibrate. Therefore, *in silico* modelling can

provide useful predictions for the molecular mechanisms of nnPTMs, but greater care is necessary to ensure a real result is outputted and not a product of user bias/error. The information provided above, about the structural effects of modifying PTMs can be used to improve *in silico* design of similar modifications in the future.[78] By inputting the experimental data of the sfGFP[148azF]+DBCO-amine crystal structure into GROMACS or other MD software, we are able to improve the parameterisation of the ncAAs.[25,169,197,205-209] This is an underdeveloped field so there is still a lot of work to be done.

Since publishing this work,[115] I have improved use of molecular modelling to help explain improved bioluminescence and stability in deletion mutants of firefly luciferase (fluc),[210] furthering the work of Arpnio *et al.*[148] By modelling individual deletions on the binding loop of fluc we were able to explain how changes to the structure lead to various changes in function.

4   Engineering functionally linked protein dimers.

4.1   Introduction

The design and production of self-assembling protein complexes is currently of great interest, because it provides new routes to construct new functional architectures with properties not present in the simple monomeric form.[24-26] Oligomerisation is common place in nature; most proteins exist as oligomers either as part of transient or permanent complexes.[12-14] The most commonly observed oligomers are symmetrically organised homodimers.[10,11] As such oligomerisation has emerged as an alternative route to engineering new, useful sampling of new functional-structural space and higher order complexes from a limited selection of monomers. Creation of new, emergent systems will require engineered protein oligomers to have the ability to self-assemble and show cooperativity between subunits so making complexes more than the sum of their parts. Ideally this means inter subunit communication and creation of new structural features. There have been great strides towards creating self-assembling protein oligomers using a variety of approaches.[4,25,27] Previous work has used, helix-helix interactions,[30-32] metal ion coordination,[33,34,211] fusion domains,[24] disulphide bridging[17] and using naturally inspired protein-protein interfaces to remodel proteins to successfully create self-assembling oligomers[35-38]. However, there is little to no functional synergy observed between individual subunits, due to long-range interactions outside of the oligomerization interface that need to be considered for connecting functional centres. Both disulphide and metal coordination have been used to create a range of GFP oligomers but with little evidence of communication.[17] Additionally to this, work by Edwards *et al*. where domain insertion was used to successfully link TEM β-lactamase and cytochrome b562 (cyt$_{b562}$) to generate haem dependent ampicillin resistance in bacteria.[212,213] When cyt$_{b562}$ is saturated with haem then TEM is active. However, in the absence of haem TEM is non-functional hence functional modulation was introduced into the protein.[212,213] Cyt$_{b562}$ has also been successfully linked to GFP to create an

energy transfer system.[134] The issue with domain insertion is that it requires the secondary protein domain to be encoded into the primary domain at the genetic level.[134,212,213] This requires extra linker sequences between the two protein domains. Another requirement for this work is that the protein needs to have both N and C terminal domains at the linkage region or for the protein to be circularly permutated to accommodate this requirement and as such is not ideal for many proteins.[134,212,213]

Having previously demonstrated how proteins can be modulated through modification with small molecules, in this chapter I will describe the design and creation of communicating, artificial fluorescent protein dimer networks using click chemistry (Figure 4.1a, Section 1.4.3).[187,188,214] There are many benefits to this approach including a simpler design that is reliable and predictable. Mutually compatible sites can be placed at various points throughout the constituent monomer proteins to determine the optimal crosslink position. Classical cystine-based disulphide bridges, struggle to achieve this due to the lack of reaction orthogonality and hence can lead to mixed oligomer populations (A + B = A-A, B-B and A-B). In addition, the presence of cysteines already present in the protein's sequence could lead to unintended disulphide bond formation or additional engineering is required to remove native cysteines. Disulphide bonds are also susceptible to reduction and the short side chain of cysteine could lead to steric clashes that may inhibit dimer formation. NcAAs such as tyrosine or lysine derivatives[182,183,187,194] (*e.g.* Figure 4.1a) reduce the risk of clashes by having longer sidechains, whilst maintaining the structural intimacy required for inter-monomer communication.

Strain promoted azide alkyne cycloaddition (SPAAC)[118], as used in Chapter 3, was utilised to assemble the dimers due to its simplicity and that it doesn't require any toxic cofactors or catalysts. Unlike copper catalysed azide-alkyne reaction SPAAC allows for the formation of both '*syn*' and '*anti*' regioisomers[118,215] (Figure 4.1b).
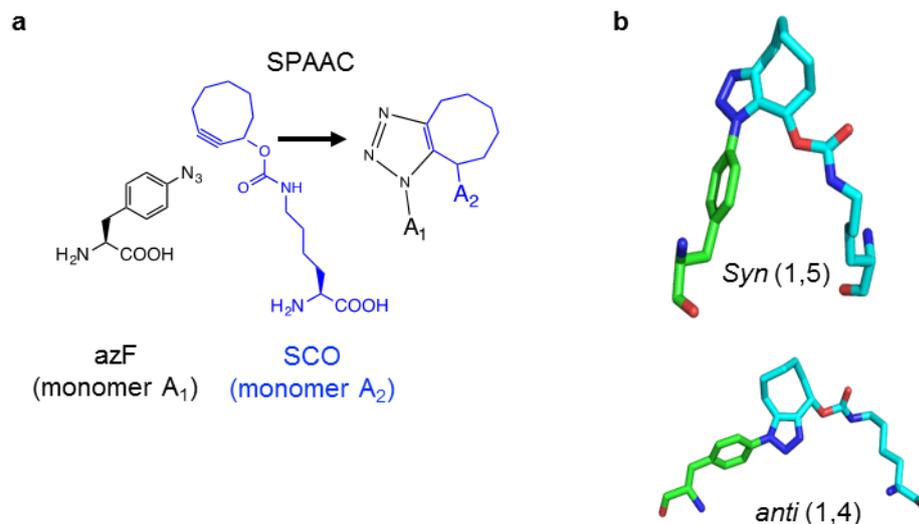
Figure 4.1 SPAAC with genetically encoded ncAA, azF (azide) and SCO (strained alkyne). a) Scheme showing SPAAC reaction between two genetically coded ncAAs. b) Regioisomers of SPAAC reaction (Section 2.3).

Super-folder green fluorescent protein (sfGFP)[127] was chosen as the target protein due to its track record of being a good model for understanding the molecular influence of ncAA incorporation on protein samples, [152,193,194] in the context of click reactions[115] and biohybrid assemblies.[59,109,119] Autofluorescent proteins also have important fundamental molecular mechanism associated with the fluorescent process, and have thus proved to be an ideal system for studying the importance of long range interaction networks, long range proton wires, electron transfer and coupled photochemistry.[101,138,145,216-218] This chapter will present how *in silico* design can be used to generate protein homodimers with enhanced and positively modulated function. Experimentally determined structures of the dimers show that novel long range interaction networks are formed connecting the chromophores of the homodimer monomers. The data described below suggest that communication between the chromophores is determined by symmetry, as non-symmetrical homodimers showed little evidence of synergy.

4.2 Results and Discussion

4.2.1 *In silico* design of click chemistry sites

It is predicted that regions of a protein's surface with a higher tendency to associate with each other, *i.e.* hydrophobic pockets and regions with large numbers of hydrogen bond donors and acceptors which can stabilise the interface, are more likely to yield a productive click reaction. This is because the more stable the interface, the greater time that the click compatible residues are in proximity to each other and hence facilitate the reaction. While monomeric proteins like sfGFP[9,77,127] do not readily self-associate weak interactions can be manifested and potentially exploited in the context of Click chemistry. Previous work from the Jones group has shown that sites for optimal Click chemistry efficiency do not always coincide with higher degree of surface accessibility,[151] which is against the normal dogma associated with optimal sites on proteins for modification.

To predict potential optimal coupling regions and interaction surfaces, *in silico* docking was used to assist dimer design. Initially, the ClusPro server[173] (cluspro.org) was used to generate potential dimer configurations as described (Section 2.3.3). ClusPro creates large numbers of possible dimer interfaces, energy minimises them and then takes the top 1000 models with the lowest energy and clusters them to other structures within a RMSD of 9 Å then ranks them according to the number of members in a cluster. Of the 18 models output from ClusPro, the 5 largest clusters (Models 1-5, Figure 4.2) were refined and analysed using RosettaDock[171,174] to rank the models based on energy and interface score (Table 4.1). The advantage of using Rosetta is that it minimises the side chains of the protein as opposed to the rigid-body minimisation used by ClusPro and hence reduces steric clashes at the interface.[171,172] Whilst there were several different orientations of one sfGFP to the other observed, docking revealed residues 145-148, 202-207 and 221-224 were present at the interface of the models and so were used as a basis for selecting residues for dimerisation.
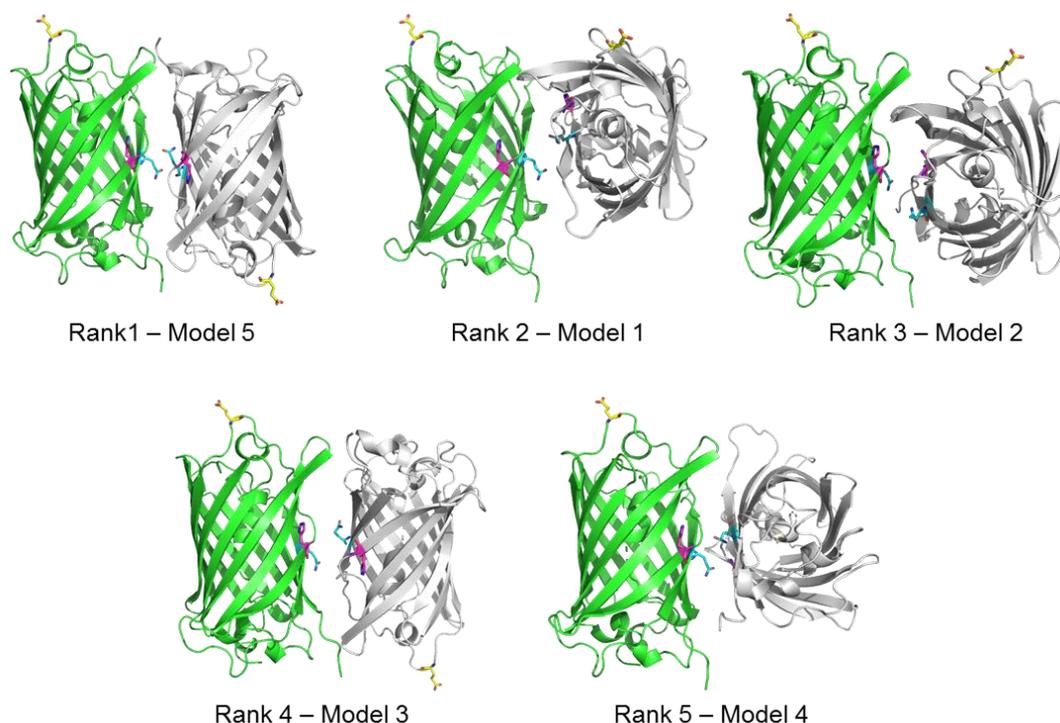
Figure 4.2 Top 5 lowest energy models after Rosetta dock, with reference chain in green and second chain in white. Residues 132, 148 and 204 shown as sticks (yellow, magenta and cyan respectively).

Table 4.1 Statistics for *in silico* modelling of sfGFP dimer interfaces. Models are ranked from lowest energy to highest according to RosettaDock.

Model	Total energy (kJ/mole)	Interface Energy (kJ/mole)	I-RMSD (Å^2)
Model 5	-503.94	-13.434	1.755
Model 1	-501.966	-4.192	0.278
Model 4	-497.071	-6.993	0.397
Model 2	-497.112	-6.375	0.56
Model 3	-494.492	-3.079	0.195

4.2.2 Rationale for selection of residue 148 as a dimerisation point.

Three residues were selected, based on the dimer models, for replacement with click compatible ncAAs, azF[85,94] (azide) and SCO[84,92] (alkyne) (Figure 4.1a). H148 and Q204 were selected based on their location at the predicted interface (Figure 4.3a) and their association with the chromophore (Figure 4.3b). The association with the

chromophore of both residues influences the charged state of the chromophore,[9,146] that exists in two states the protonated A form (CroA) that excites at ~400 nm and the deprotonated phenolate B form (CroB) that excites at ~490 nm, as described earlier in the introduction (Section 1.5). Both residues also have been shown to contribute towards the formation of an extended proton and water wire tunnel from the chromophore to the solvent.[138,139] Previous work shows that residues H148 (Chapter 3) and Q204 have also can be efficiently modified with small cyclooctyne adducts after being substituted with azF (Section 3.2.4.1).[109,151]
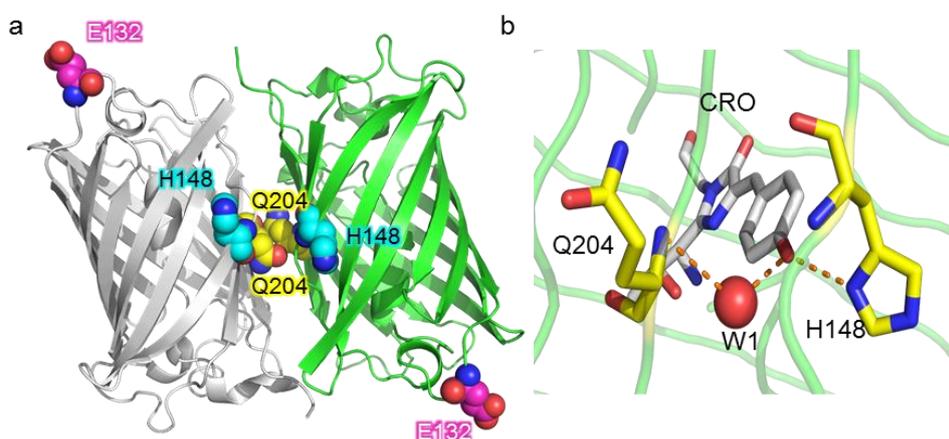
Figure 4.3 Residue selection for creating homodimers a) Top ranked model from ClusPro and RosettaDock, residues 132, 148 and 204 shown as spheres and labelled. b) Molecular interactions between residues 148 and 204 (yellow), the sfGFP chromophore (grey) and water molecule W1 (red sphere).

4.2.3 Production, purification and characterisation of sfGFP^{SCO} variants

4.2.3.1 Expression of sfGFP^{148SCO}

The production of the following proteins has been described elsewhere but followed the same procedure as outlined in Section 2.4: sfGFP^{148azF} (Chapter 3);[115,186,194] sfGFP^{132azF} and sfGFP^{204azF} (Chapter 5 and [151]) The non-canonical amino acid s-cyclooctyne-L-lysine (SCO) was chosen as the Click partner to azF for dimerisation because it can be genetically encoded using the a reprogramd codon (TAG) through an

engineered *Methanosarcina mazei* pyrrolysine-tRNA/amino acyl tRNA synthase in a manner similar to the azF incorporation (Chapter 3). Genetic encoding was possible through use of the plasmid pEVOL[SCO],[84] which was a kind gift from the Lemke group (EMBL, Heidelberg, Germany). As all three SCO containing proteins (sfGFP[132SCO], sfGFP[148SCO], and sfGFP[204SCO]) were produced in the same way only the production and purification of sfGFP[148SCO] is described below.
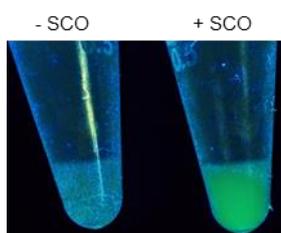
Figure 4.4 Comparison of expression of sfGFP[148SCO] in the presence (+SCO) and absence (-SCO) of SCO. a) Resuspended cell pellets irradiated by UV light shows fluorescence in the presence of SCO but no fluorescence in the absence. b) SDS-PAGE gel of cell pellets showing increased expression of a protein at about 27 kDa in the +ve control.

The pBAD-GFP plasmid containing the sfGFP gene with the codon equivalent to residue 148 replaced by the amber stop codon TAG (termed pBAD-GFP[148TAG]) and pEVOL[SCO] for incorporating SCO in response to the TAG stop codon were transformed into *E. coli* TOP10 cells (Section 2.2.6). After overnight incubation in autoinduction medum, cells had a light green colour (Figure 4.4). Thus, SCO was successfully incorporated into sfGFP at residue 148 to generate a functional [coloured and fluorescent] full length sfGFP[148SCO].

Cells were harvested and lysed as described previously (Section 2.4.1). SfGFP[148SCO] was purified from cell lysates using a 5 mL cobalt-NTA 5 column (Generon) equilibrated in 50mM Tris-HCl [pH 8.0], 300mM NaCl, 20 mM Imidazole. The protein was eluted with a gradient of 20 mM to 500 mM Imidazole and fractions showing a significant peak at 400nm absorbance were collected (Figure 4.5). Subsequent analysis

by SDS-PAGE proved that fractions containing sfGFP[148SCO] were at least 95% pure (Figure 4.5b).
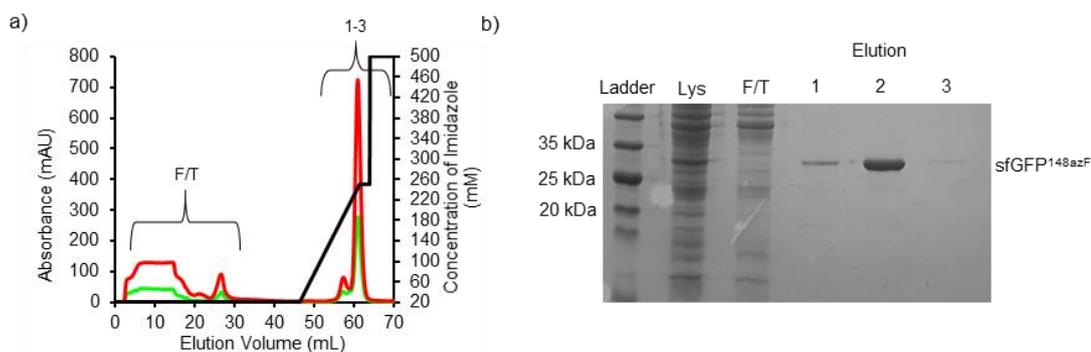
Figure 4.5 Cobalt affinity purification of sfGFP [148 SCO] a) chromatogram of purification run showing 400 nm and 485 nm absobance in red and green respectively. Concentration of imidazole is also shown in black. b) SDS-PAGE analysis of samples from purification lanes from left to right are as follows L: Ladder, Lys: cell lysate before purification F/T: column flow through before imidazole elution; elution fractions 1-3.

4.2.3.2 Characterisation of sfGFP[148SCO] and sfGFP[204SCO]

Previous work has shown that incorporating ncAAs into proteins can have a wide variety of effects on the function of the protein (For examples see Chapter 3 and [78,110,115,151,186]). Therefore, before studying what effects are caused on dimerisation with proteins containing azF, there is a need to characterise effects SCO incorporation has on sfGFP. As has been described above and in the previous chapter His148 is required as part of the process involved in generating a deprotonated phenolate chromophore in the ground state via a hydrogen bond network. It is predicted that incorporation of SCO at residue 148 will disrupt this network similarly to the effect shown by azF incorporation (see Figure 3.2 in Chapter 3).[115]

Similarly to sfGFP[148azF], the major population of sfGFP[148SCO] predominantly exists in the CroA form as evidenced by the major absorbance peak at 395 nm. The presence of an absorbance peak associated with CroB (492 nm), suggests that a small population of chromophores occupy the CroB state (Figure 4.6a). The molar extinction coefficients

for sfGFP^{148SCO} are like sfGFP^{148azF} (See Chapter 3 and below), at 31,000 M^{-1}.cm^{-1} and 17,300 M^{-1}.cm^{-1} for CroA and CroB respectively. Incorporation of SCO at 148 also reduced the emission of sfGFP to the same level as that on incorporation of azF. Excitation at either 400 nm or 492 nm showed fluorescence had dropped by 75% and 85% (respectively compared to sfGFP^{WT} (Figure 4.6b, Table 4.2), effectively switching the protein 'off'.

Incorporating azF at residue 204 has limited impact on the absorbance of sfGFP, save for a minor increase in molar extinction coefficient (~4%, Figure 4.6c). There is however a decrease in fluorescence of about 20% (Figure 4.6d). Incorporation of SCO at position 204 leads to a loss in molar extinction coefficient (down to ~40,000 M^{-1}.cm^{-1}). However, there is no difference in emission in comparison to sfGFP^{WT}.

Table 4.2 Spectral characteristics of sfGFP variants

Variant	λ_{max} (nm)	λem (nm)	ε (M^{-1}.cm^{-1})	QY	Brightness
sfGFP^{WT}	485	511	49,000	0.75	36750
sfGFP^{148azF}	400	511	34,200	0.69	23,598
	500	511	19,800	0.32	6,336
sfGFP^{148SCO}	395	511	31,000	0.52	16,120
	492	511	17,300	0.84	14,532
sfGFP^{204azF}	485	511	51,000	0.68	34,680
sfGFP^{204SCO}	485	511	39,800	0.6	26,268
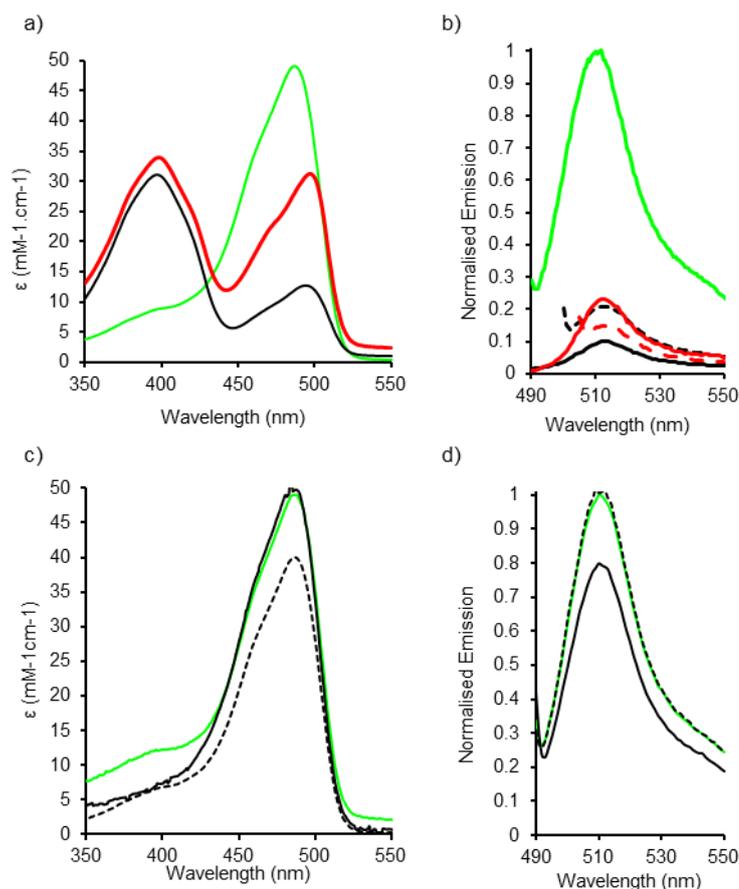
Figure 4.6 Spectral properties of sfGFP variants. a) Molar extinction coefficient of sfGFP^{WT} (green), sfGFP^{148azF} (red) and sfGFP^{148SCO} (black). b) Emission spectra of sfGFP^{WT} (Ex = 485 nm). and sfGFP^{148azF} and sfGFP^{148SCO} excited at 400 nm (solid lines) and at 492nm (dashed lines). c) Molar extinction coefficient spectra of sfGFP^{WT} (green), sfGFP^{204azF} (black solid lines) and sfGFP^{204SCO} (Black dashed lines). d) Normalised emission spectra of sfGFP^{204} variants (Ex = 495 nm). Absorbance spectra were measured at 10 µM protein. Fluorescence spectra were measured at 0.5 µM protein and normalised to sfGFP^{WT}

4.2.4   Creation and purification of sfGFP dimers

4.2.4.1   Preliminary dimerisation.

Table 4.3 Table listing dimer variations created by click chemistry and constituent monomers.

Dimer	Monomer 1	Monomer 2	Symmetrical/Non symmetrical
sfGFP^{148x2}	sfGFP^{148azF}	sfGFP^{148SCO}	Symmetrical
sfGFP^{204x2}	sfGFP^{204azF}	sfGFP^{204SCO}	Symmetrical
sfGFP$^{148SCO-204azF}$	sfGFP^{204azF}	sfGFP^{148SCO}	Non symmetrical
sfGFP$^{148azF-204SCO}$	sfGFP^{148azF}	sfGFP^{204SCO}	Non symmetrical

Now that the properties of the monomers have been determined, testing of dimer formation can begin. Initially, the credibility of predicting a dimer interface was tested by analysing products of dimerisation through gel mobility shift assays. The combinations of mutants tested are shown in Table 4.3, and are split into: symmetrical dimers, comprising two monomers with a ncAA at the same position; and non-symmetrical dimers where both monomers have ncAAs at different residues. The 204x2 dimer data were kindly supplied by Jacob Pope.

To create dimeric sfGFP, azido (sfGFP^{148azF} or sfGFP^{204azF}) and alkyne (sfGFP^{148SCO} or sfGFP^{204SCO}) variants of sfGFP were added in a 1:1 molar ratio. Each monomer was at a concentration of 100 µM as described in Section 2.8.2. Samples of monomers and the click reaction mix were run on SDS-PAGE gel to determine the extent of dimerisation. If monomers have successfully dimerised with each other than there should be an obvious band shift in the dimerisation lane. An example of a mobility shift assay for the formation of sfGFP^{148x2} is shown below (Figure 4.7). All four SDS-PAGE mobility shift experiments showed dimer formation was possible. The band on SDS-PAGE dimer ranged from an apparent mass of 56 kDa – 65 kDa. Dimers were separated from unconjugated monomers by size exclusion chromatography as described in Section 2.4.4. It should be noted several rounds of SEC were required to remove the last traces of monomeric protein, but all four dimers were purified.
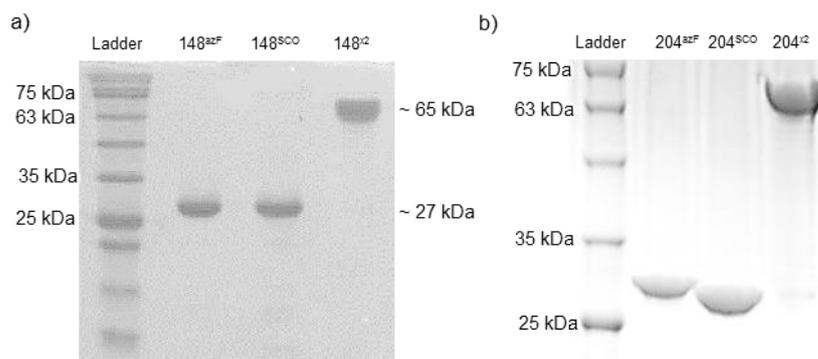
Figure 4.7 SDS-PAGE of purified sfGFP Dimers compared to monomers. a) sfGFP[148] variants. Lanes from left to right: Gene Flow BLUeye protein marker (Ladder), sfGFP[148azF] (148[azF]), sfGFP[148SCO](148[SCO]), Dimer mixture (148[x2]). b) sfGFP[204] variants Lanes from left to right: Gene Flow BLUeye protein marker (Ladder), sfGFP[204azF] (204[azF]), sfGFP[204SCO](204[SCO]), Dimer mixture (204[x2]).

4.2.4.2 Non-dimerising residues

Residue 132 was not predicted to be at the potential dimer interface (Figure 4.3a) but like residues 148 and 204, has been shown to be Click chemistry compatible with a range of cyclooctyne based moieties from fluorescent dyes (Chapter 5 and [151]), to short single DNA strands to carbon nanotubes (Chapter 5 and [109,119]). As such it should provide an ideal residue to test our ability to predict viable protein-protein Click chemistry reactions. Adding sfGFP[132SCO] to sfGFP[132azF] (Chapter 5 and [151]) and to sfGFP[204SCO] showed no evidence of dimerisation (Figure 4.8), as predicted by *in silico* modelling. This suggests that, our approach successfully selects regions for creating dimers. It is possible that residue 132 does not dimerise efficiently as it is on a flexible loop and hence dynamic, reducing the likelihood of the azide and alkyne moieties coming into contact long enough for the reaction to occur. Another possible explanation is that the residues surrounding 132 are unable to form enough weak intermolecular bonds to stabilise the interface.
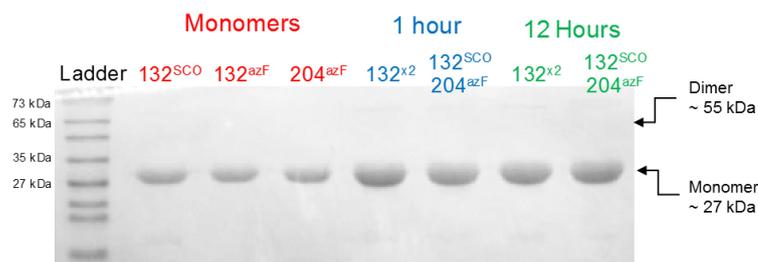
Figure 4.8 Dimerisation potential of non-dimer interface residue E132. SCO was incorporated at residue 132 (132^{SCO}) of GFP and dimerisation was attempted with azF incorporated at either 132 (132^{azF}) or 204 (204^{azF}) of GFP. No clear dimerisation product was observed for either symmetrical of non-symmetrical dimer (132^{x2} and 132^{SCO}-204^{azF} was observed by gel mobility shift assay.

4.2.4.3   Mass spectrometry

To prove that dimers were covalently linked, aliquots (10 µM) of sfGFP^{148x2} and sfGFP^{204x2} were submitted for liquid chromatography time of flight mass spectrometry (LC-/TOF-MS) as described in Section 2.7.1. The Mass spectra yields dominant peaks at 54,203 Da and 55,866 Da for sfGFP^{148x2} (Figure 4.9a) and sfGFP^{204x2} (Figure 4.9b) respectively. For sfGFP^{204x2} the observed mass is 2 Da higher than predicted (55,864) by the method described (Section 2.7.1). This minor difference could be explained by the deconvolution spectra calling a $^{13}C_2$ peak, *i.e.* a mass peak where two of the carbon atoms are of the rare but stable isotope ^{13}C,[219] however it is most likely to do with the resolution of the mass spectrometer. As such, we can assume that sfGFP^{204x2} is what we believe it to be and has successfully formed a covalent dimer.
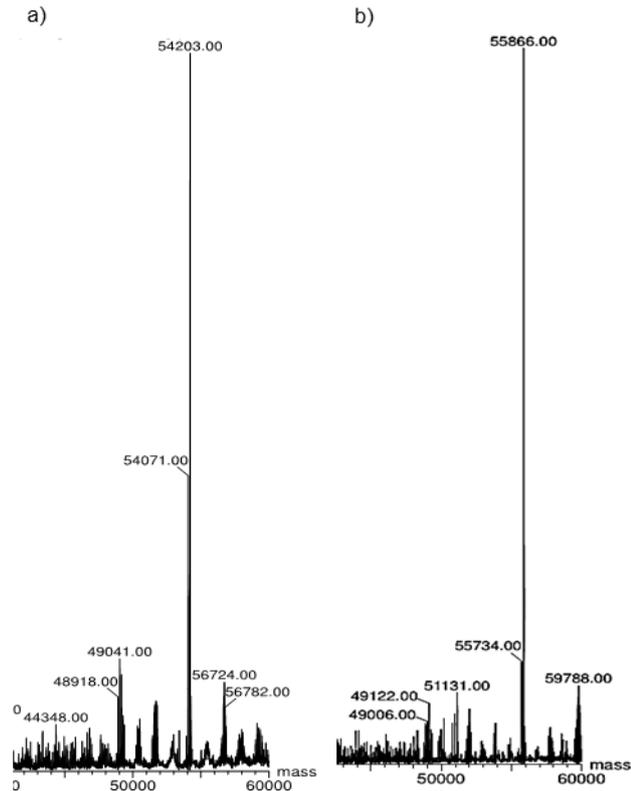
Figure 4.9 Determining the size of symmetrical sfGFP homodimers. Mass spectra of a) sfGFP[148x2] and b) sfGFP[204x2].

Things become more complex with sfGFP[148x2], as the observed mass is significantly lower (~1640 Da) than expected mass of 55,846 Da. The protein is too large to be either of the monomers (azF = 27,878 Da, SCO = 27,968 Da), yet is too small for the predicted full length dimer. The mass loss could be explained by the partial loss of both His tags (2xHHHHHH) which theoretically would be a mass loss of 1646 Da. This would make the observed mass 3 Da heavier than the predicted. This mass difference could again be explained by the deconvolution software picking out a peak with multiple Carbon-13 atoms *i.e.* $^{13}C_3$. Therefore, we can conclude that the sample is sfGFP[148x2] and that we have successfully produced both symmetrical homodimers and they have crosslinked as we expected *via* SPAAC.

4.2.5 Fluorescence properties of symmetrical and non-symmetrical sfGFP dimers

By using bioorthogonal reaction handles, both symmetrical and non-symmetrical dimers were generated. This was achieved by using a common crosslink residue to mimic symmetry, or different cross-link residues to mimic non-symmetry. The fluorescent properties of both symmetrical (sfGFP[148x2] and sfGFP[204x2]) and non-symmetrical dimers (sfGFP[148azF-204SCO] and sfGFP[148SCO-204azF]) were investigated.

4.2.5.1 sfGFP[148x2]

Regulated activity is a common feature of protein oligomers,[13] ranging from active complex formation, to functional switching through interaction network modulation. As described previously in this chapter (also in Chapter 3 and [115,152,220]), mutation of H148 removes a key H-bond with the chromophore that disrupts a proton shuttling network used to deprotonate the chromophore. Hence, sfGFP[148] mutants tend to show dominant CroA (neutral form) absorbance as opposed to CroB (phenolate form). This is true of sfGFP[148azF] and sfGFP[148SCO].

Conjugation of the two proteins via SPAAC has the potential to restore the networks responsible for shifting the population of the CroA state to CroB for at least one protein in the dimer [sfGFP[148azF]]. As can be seen in Figure 4.10a, this appears to be the case with sfGFP[148x2]. The dominant absorbance peak is in the CroB region, although it is slightly red shifted from sfGFP[WT] (λ_{max} = 485 nm) to 492 nm. There is a 3-fold increase (1.5-fold per chromophore) in brightness (Table 4.4) due to an equivalent increase in molar absorbance coefficient from 49,000 $M^{-1}.cm^{-1}$ (sfGFP[WT]) to over 150,000 $M^{-1}.cm^{-1}$ (Table 4.4, Figure 4.10a). This increase equates to 4x the combined value of both constituent monomers, *i.e.* the sum absorbance at ~490 nm of sfGFP[148azF] and sfGFP[148SCO] (~37,100 $M^{-1}.cm^{-1}$). The CroB: CroA absorbance ratio shifts 10-fold from ~0.5 for the monomers to ~5 for the dimer (Figure 4.10a). This order of magnitude shift in absorbance is mirrored in fluorescence emission. Excitation at the CroB

wavelength (492 nm) results in a ~10-fold increase compared to either monomer (Figure 4.10b, Table 4.4).

Table 4.4 Spectral properties of sfGFP variants

Variant	λ_{max} (nm)	λem (nm)	ε ($M^{-1}.cm^{-1}$)	QY	Brightness ($M^{-1}.cm^{-1}$)
sfGFP^{WT}	485	511	49,000	0.75	36,750
sfGFP^{148azF}	400	511	34,200	0.69	23,598
	500	511	19,800	0.32	6,336
sfGFP^{148SCO}	395	511	31,000	0.52	16,120
	492	511	17,300	0.84	14,532
sfGFP^{148x2}	492	511	150,200	0.80	120,160
sfGFP^{204azF}	485	511	51,000	0.68	34,680
sfGFP^{204SCO}	485	511	39,800	0.6	26,268
sfGFP^{204x2}	490	511	160,000	0.71	113,600

Despite both azF and SCO being chemically very different to each other and the original imidazole group of H148, the absorbance spectrum implies that both chromophores in sfGFP^{148x2} exist predominantly in the CroB state. Additionally, the dimer exhibits overall enhanced function sfGFP^{148x2} when compared to sfGFP^{WT} suggesting further synergistic benefits of dimer formation (Table 4.4, Figure 4.10b). Molar excitation and brightness increase of ~320% for sfGFP^{148x2} (160% per chromophore) compared to the WT, higher than expected for a simple additive effect where both monomers are acting independently of each other. The change is more dramatic when compared to the monomers, where a 8-20-fold increase in brightness observed.
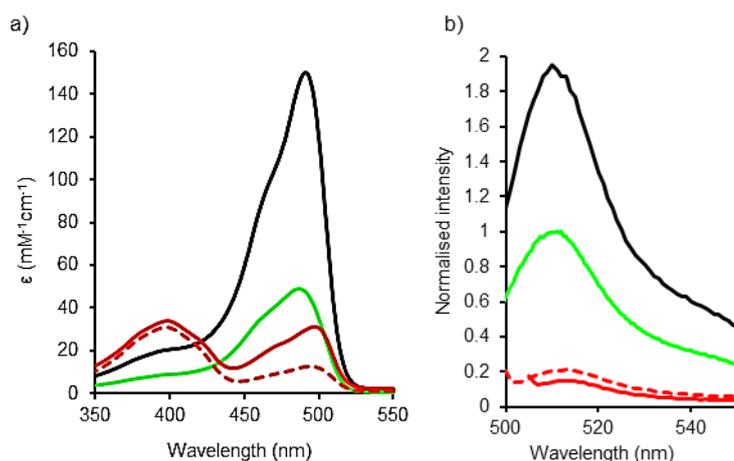
Figure 4.10 Spectral properties of sfGFP[148] variants before and after dimerisation. a) Absorbance spectra of sfGFP[WT] (green), sfGFP[148x2] (black), sfGFP[148azF] (red), and sfGFP[148SCO] (red dashed). b) Emission spectra (Excitation at 492 nm for sfGFP[148] variants, excitation 485 nm for sfGFP[WT]. Emission spectra are normalised to sfGFP[WT]. Absorbance spectra were carried out at 10 µM and Fluorescence emission spectra were carried out at 0.5 µM.

4.2.5.2 sfGFP[204x2]

Dimerisation at residue 204 also enhances protein function beyond simple monomer addition. As described previously, incorporation of either azF or SCO at 204 has very little effect on sfGFP fluorescence and absorbance[151] (Section 4.2.3.2, Figure 4.11a) with the CroB form being predominant in both sfGFP[204azF] and sfGFP[204SCO]. Formation of sfGFP[204x2] causes major enhancement of both absorbance and fluorescence. Molar absorbance coefficient (ε) increased ~330% compared to sfGFP[WT] from ~49,000 $M^{-1}.cm^{-1}$ to 160,000 $M^{-1}.cm^{-1}$ (Figure 4.11a). This equates to an ε of 80,000 $M^{-1}.cm^{-1}$ per chromophore and an increase of 200% and 160% compared to sfGFP[204SCO] and sfGFP[204azF] respectively. Fluorescence emission also increases in line with the increased light absorbance (Figure 4.11b). After dimerisation sfGFP[204x2] increases ~3-fold compared to sfGFP[WT] and sfGFP[204SCO], and ~3.75-fold greater than sfGFP[204azF] (Table 4.4). As with sfGFP[148x2], sfGFP[204x2] has enhanced brightness both in comparison to the combination of two monomers and to sfGFP[WT], a long established benchmark for fluorescent protein performance. Showing that the modelling was not only successful in predicting interface regions that both help to promote the formation of a permanent

crosslink but when dimers are formed create functional synergy between the two monomers.
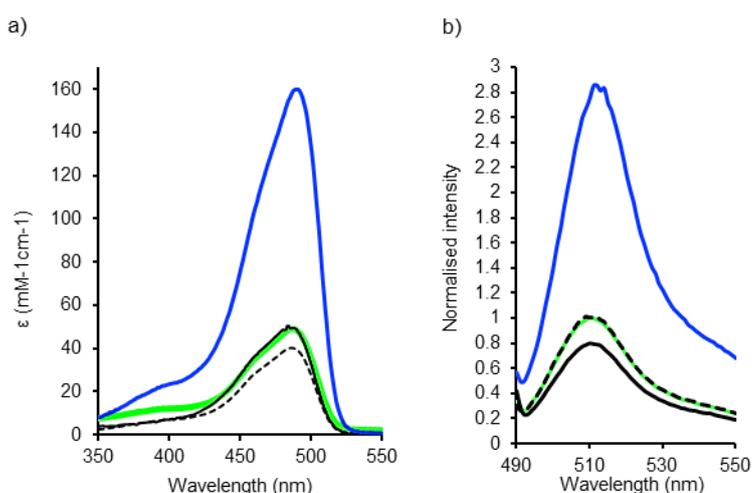
Figure 4.11 Spectral properties of sfGFP^{204x2} variants. a) Absorbance spectra of sfGFP^{WT} (green), sfGFP^{204x2} (blue), sfGFP^{204azF} (black), and sfGFP^{204SCO} (black dashed). b) Emission spectra of sfGFP^{204} variants (Excitation at 487 nm) and sfGFP^{WT} (excitation at 485 nm). Emission spectra are normalised to sfGFP^{WT}. Absorbance spectra were carried out at 10 µM and Fluorescence emission spectra were carried out at 0.5 µM.

4.2.5.3 Non-symmetrical dimers sfGFP$^{148azF-204SCO}$ and sfGFP$^{148SCO-204azF}$

Dimers were generated that linked residue 148 and 204 in the two available combinations (148azF+204SCO and 148SCO+204azF). The absorbance and fluorescence spectra of the non-symmetrical dimers (sfGFP$^{148azF-204SCO}$ and sfGFP$^{148SCO-204azF}$) are shown below with a predictive additive absorbance spectrum for each dimer (Figure 4.12). The additive absorbance spectra are the summation of each individual monomer absorbance spectra (*i.e.* sfGFP^{148azF} + sfGFP^{204SCO} and sfGFP^{148SCO} + sfGFP^{204azF}). After dimerisation there is no obvious functional synergy observed between individual monomers. In both dimeric forms, the CroA and CroB states are present suggesting a mixture of the individual properties of each monomer *i.e.* CroA state (from sfGFP^{148}) and CroB state (sfGFP^{204}). The retention of the significant CroA form in the hetero-linked dimers highlights the importance of the symmetry to the formation the inter-CRO interaction network formed by sfGFP^{148x2}. The absorbance spectrum of

sfGFP[148azF-204SCO] shows a degree of synergy but there is still a significant CroA form in the dimer below. This change is likely due to the effect of modifying 148azF as shown by modification with DBCO-amine or UV irradiation (Chapter 3).[115] Dimerisation of sfGFP[148SCO] with sfGFP[204azF] has an overall negative effect on protein function as molar extinction coefficients for both CroA and CroB forms are less than predicted in the monomer additive spectrum (Figure 4.12b). This suggests that constituent monomers of both sfGFP[148azF-204SCO] and sfGFP[148SCO-204azF] are acting largely as individual proteins with little effect on each other.
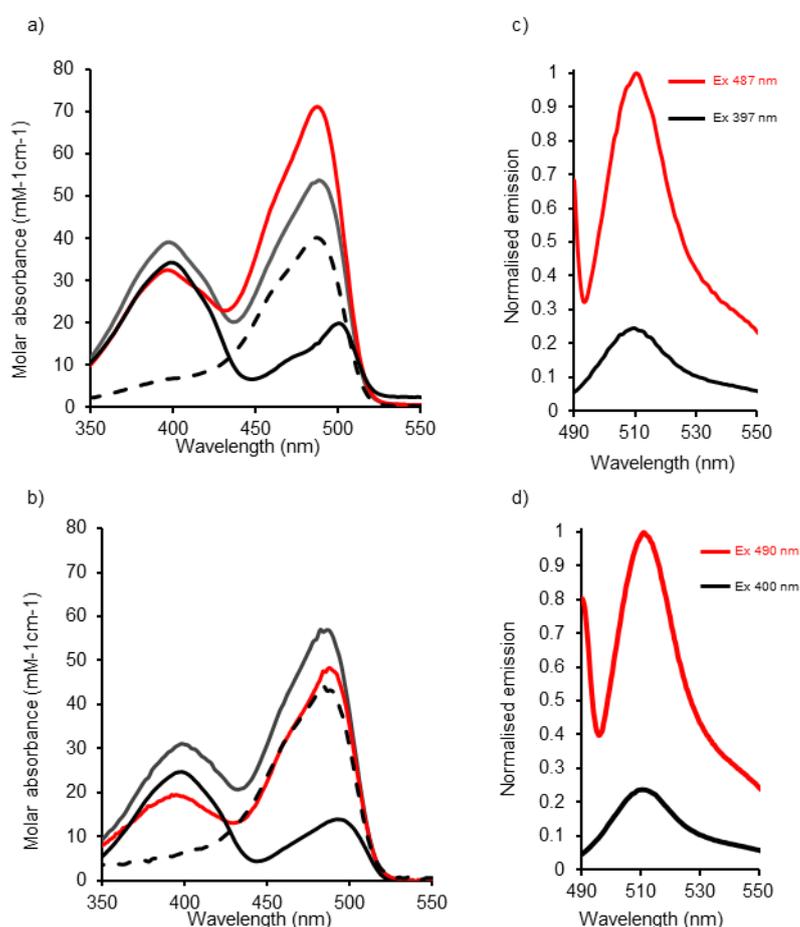
Figure 4.12 Spectra of non-symmetrical dimers sfGFP[148azF-204SCO] (a-b) and sfGFP[148SCO-204azF] (c-d). a) Measured absorbance spectrum of sfGFP[148azF-204SCO] (red) compared to the individual monomers sfGFP[148azF] (black, solid line), sfGFP[204SCO] (black, dashed line), and the additive spectrum of both monomers (solid grey line). b) Emission spectra at excitation wavelengths shown for sfGFP[148azF-204SCO]. a c) Measured absorbance spectrum of sfGFP[148SCO-204azF] (red) compared to the individual monomers sfGFP[148SCO] (black, solid line), sfGFP[204azF] (black, dashed line), and the additive spectrum of both monomers (solid grey line). d) Emission spectra at excitation wavelengths shown for sfGFP[148SCO-204azF]. Emission spectra were normalised to sfGFP[WT] (Excitation at 485 nm).

4.2.6 Thermal denaturation

It has been reported previously, that dimerisation can result in a more stable folded protein structure.[10,11,221] Strong interaction networks are formed between hydrophobic dimerisation domains, which are now buried from the bulk solvent and hence, reduce the energy state of the complex, meaning increased stability of the complex over the individual monomers.[11,221] Multiple techniques have been used to show the stability of proteins usually involving either chemical or thermal denaturation,[11,134,148,222] for example Pellegioni *et al*[222] used a combination of techniques to show that dimerisation of a flavoprotein increased thermal stability.
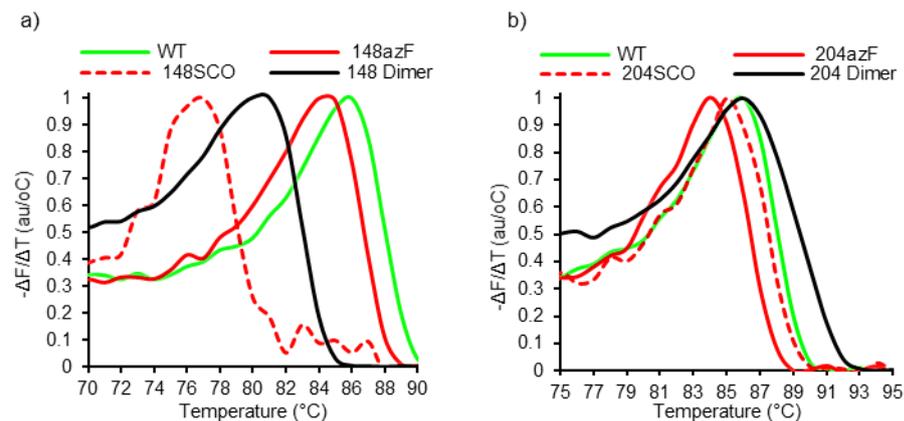
Figure 4.13 Thermal denaturation curves of sfGFP variants showing the transition from folded to unfolded protein. a) sfGFP[WT] (green), sfGFP[148x2] (black) and sfGFP[148azF] and sfGFP[148SCO] (red solid and dashed lines respectively). b) sfGFP[WT] (green), sfGFP[204x2] (black) and sfGFP[204azF] and sfGFP[204SCO] (red solid and dashed lines respectively). Curves represent change in fluorescence against temperature and are normalised to the transition point for each sample.

To determine if dimerisation has any effect on the stability, sfGFP[WT], sfGFP[148x2] and sfGFP[204x2], and constituent monomers were thermally denatured as described in Section 2.7.5. The overall thermal denaturation characteristics are shown above in Figure 4.13. The more stable the protein, the more resistant it should be to thermal unfolding. GFP fluorescence is known to be dependent on the folded state of the protein [9,127,134] so unfolding of GFP leads to the loss of inherent fluorescence due to exposure

of the CRO to the solvent leading to rapid quenching. Thus, fluorescence is a sensitive way of monitoring thermal unfolding and thus general overall stability.

It should be noted that sfGFP is a very robust and stable protein, more so than most proteins. This is evident from the thermal unfolding of sfGFP^{WT}, which has an observed denaturation midpoint of 86°C. Incorporation of azF at both residues 148 and 204, and incorporation of SCO at residue 204, have a minimal effect on stability, with the denaturation midpoints *circa* 84-86°C. Incorporation of SCO at residue 148 had a more significant effect, with sfGFP^{148SCO} having an observed midpoint of 77°C, suggesting reduced stability of the protein. The reason for such change is not known. From the structure of sfGFP^{148azF} compared to sfGFP^{WT}, little change is seen in side chain packing on mutations but azF and histidine can occupy similar volume (Figure 4.14). SCO on the other hand, is an extended aliphatic side chain so local contacts that are possible when aromatic groups are present at residue 148 may be lost so leading to a local destabilisation around the CRO. It should be noted that while we expect fluorescence to be a good reporter on the global structural integrity of GFP, in some of the mutants, loss of fluorescence could be due to local perturbations in the CRO environment that may proceed global unfolding.
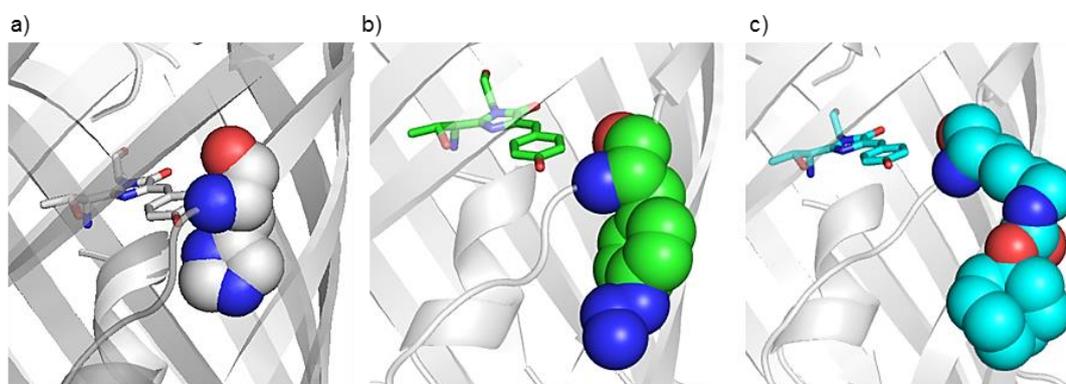
Figure 4.14 Respective volumes of residue 148 (spheres) of a) sfGFPWT (grey), b) sfGFP148azF (green) and c) sfGFP^{148SCO} (cyan). His and azF occupy similar volumes whereas the extended aliphatic side chain of SCO occupies a different area of space.

The sfGFP^{148x2} dimer had an observed denaturation midpoint of 80°C, halfway between the midpoints of sfGFP^{148azf} and sfGFP^{148SCO} (Figure 4.13a). It is quite noticeable that the peak for sfGFP^{148x2} is broader than the peaks for the monomers suggesting that there is independent unfolding of each monomer within the dimer causing a longer transition point from folded to unfolded protein. A similar broadening is observed with sfGFP^{204x2}. However, the midpoint is 86 °C which is marginally higher than either monomer (sfGFP^{204azF} = 84 °C, sfGFP^{204SCO} = 85 °C) and the same as sfGFP^{WT}. The lack of significant increase in midpoint suggests that there is no increased stability of sfGFP upon forming dimers, but as mentioned the monomers are already very stable compared to most proteins. Therefore, to better investigate the effects of this dimerisation on stability future work should start with a less stable starting point.

4.2.7 Crystallography

Given that both symmetrical dimers have synergy in terms of functional switching and enhanced spectral properties, the structures of sfGFP^{148x2} and sfGFP^{204x2} were determined to provide an insight into the molecular basis of action. To do this both dimers were crystallised, and the structures solved. How the structure of sfGFP^{148x2} was solved is described below.

4.2.7.1 Crystal formation, Diffraction and refinement

Purified sfGFP^{148x2} was concentrated to approximately 10 mg/mL and used to set up crystal trials. Aliquots (0.2 µL) of protein were tested for crystal formation across 96 crystallisation conditions using a broad screen as described in Section 2.7.2. Crystals were left to grow at 25 °C for 2 months. Crystal formation was monitored as described (Section 2.7.2). Several conditions yielded large rod like crystals with a distinct green colour. These were harvested as described and taken to Diamond Light Source for X-ray diffraction. Of the successful conditions the one that yielded the best crystal, *i.e.* the one with the highest resolution, was condition C6 (0.1M PCTP Buffer [pH 9.0], 25% PEG

1500). The crystal formed as a long hexagonal prism, with a distinct green colour (Figure 4.15a). The crystal was over 570 μm long and on average 44 μm in diameter. The crystal diffracted to a resolution of 1.96 Å and had a space group of $P6_5$ (Figure 4.15b). This is different to most sfGFP crystal structures determined by this lab including sfGFP[148azF] and sfGFP[148azF] + DBCO-amine, which had space groups of $P\ 4_32_12_1$ and $P\ 2_12_12_1$ respectively. This suggests that dimer formation has disrupted the traditional crystal contact points of sfGFP.
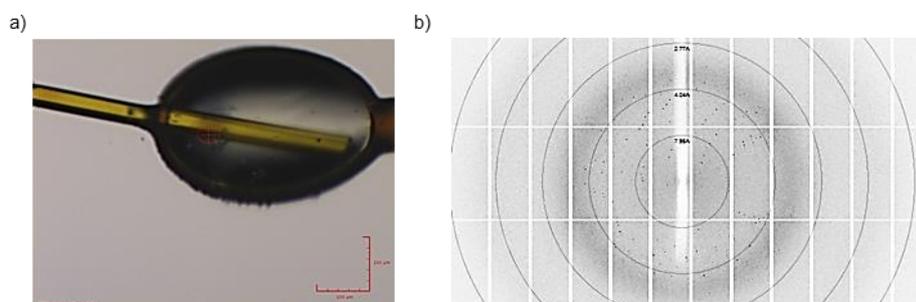
Figure 4.15 sfGFP[148x2] crystal and diffraction pattern. a) Beam line image of sfGFP[148x2] crystal C6. With 100 μm scale bar for size reference. b) X-ray diffraction pattern of sfGFP[148x2] crystal

Full diffraction statistics are shown in Table 4.5. In brief there were over 964,841 reflections of which 41,916 were unique giving 21.8x coverage of the crystal (multiplicity). There was 100% completeness of data across all resolution shells and 99% in the highest resolution shell. The correlation coefficient ($CC_{1/2}$) was 1.0 (0.68 in highest resolution shell). These statistics suggest that the data are of sufficient quality for refinement. Using the structure of sfGFP[148azF] (5BT0) as a search model, cell component analysis (Mathews test, Section 2.7.2.2) determined that there were two monomer units (*i.e.* one dimer) in the unit cell. Using 5BT0 as a search model, PHASER was able to resolve the phasing of the diffraction data *via* molecular replacement to generate a model for refinement. The phased model contained two monomer proteins forming a single dimer in the unit cell, chain A and chain B. After substituting azF and

SCO in both possible combinations and refining, it was deemed that chain A was the azF containing monomer and that chain B contained SCO at residue 148. At convergence the R $_{factor}$ and R $_{free}$ were at 18.2% and 21.1% and the RMSD of bond lengths and angles were 0.015 Å and 1.868° respectively (Table 4.5). This suggests that the refinement has generated a good model and so was submitted to the protein data bank and assigned the accession number 5NHN. It should be noted that for simplicity, Protein Data Bank in Europe (PDBe, http://www.ebi.ac.uk/pdbe/node/1) have since split the linkage into lysine and phenylalanine within chains A and B respectively, and a new ligand containing the triazole linkage and the SCO renamed DB5. However, in the context of this chapter we will consider the linkage as azF and SCO.

4.2.7.2 Overall structure of sfGFP[148x2]

Upon examining the refined structure for sfGFP[148x2], there is a head-to-tail monomer arrangement with both pairs of termini facing away from each other (Figure 4.16a). The two chains are linked by an elongated azide-alkyne triazole link in the '*anti*' conformation at the centre of the dimer (Figure 4.16b). Both chromophores are 15 Å apart and facing in towards each other across the interface (Figure 4.16c). The surface area of the dimer interface was calculated by PISA (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html),[223] and the interface has a buried area of ~1,300 Å^2, similarly to naturally occurring dimer interfaces,[15] with largely the same residues from each monomer contributing. Hydrogen bonding plays an important role with residues E142, N146, S147, N149 and N170 from both monomers contributing to form 8 inter-subunit hydrogen bonds (Figure 4.16d).

Table 4.5 Crystallographic statistics for sfGFP ^{x2} crystal.

sfGFP^{148x2}

PDB ID	5NHN
Refinement statistics	
Wavelength (Å)	0.979
Beamline	Diamond IO4
Space group	$P\,6_5$
a (Å)	99.80
b (Å)	99.8
c (Å)	108.92
Resolution range (Å)	67.71-1.96
Total reflections measured	964841
Unique reflections	41,916
Completeness (%) (last shell)	100 (99.9)
Multiplicity (last shell)	21.8 (14.1)
I/σ (last shell)	22.9 (4.0)
CC$_{1/2}$	1.000 (0.680)
R(merge) (%) (last shell)	7.9 (68.8)
B(iso) from Wilson (Å^2)	41.03
Refinement statistics	
B(iso) from refinement	50.8
Log Likelihood Coordinate rms	0.126
Non-H atoms	3877
Solvent molecules	226
R-factor (%)	18.2
R-free (%)	21.1
RMSD bond lengths (Å)	0.015
RMSD bond angles (°)	1.868
Ramachandran Plot Statistics	
Core region (%)	98.42
Allowed region (%)	1.13
Additionally allowed region (%)	0
Disallowed Region (%)	0.46
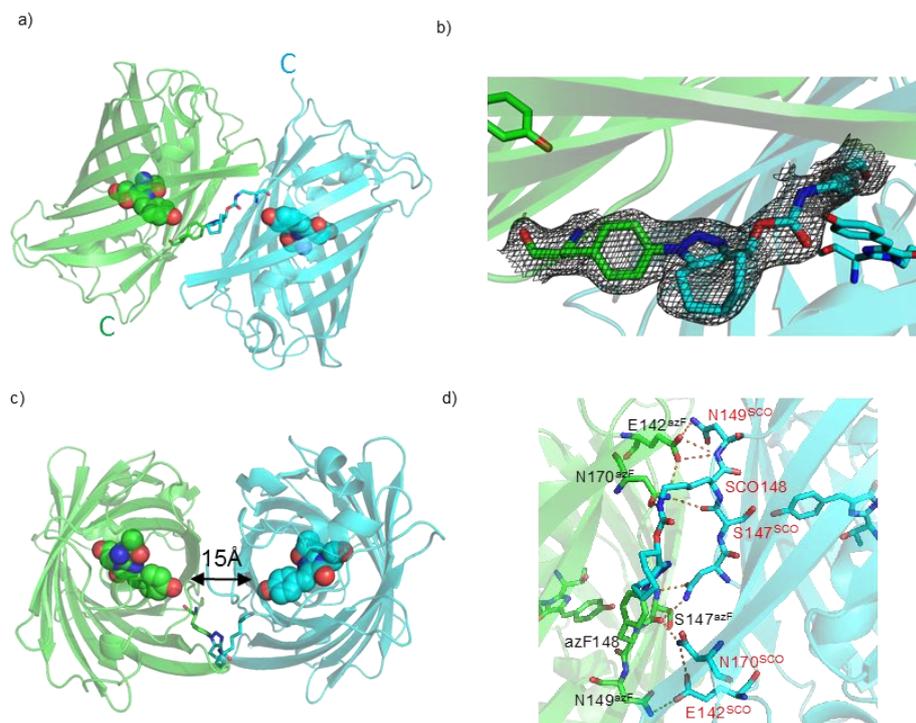
Figure 4.16 Structure of sfGFP^{148x2}. Chain A (azF) is coloured green and Chain B (SCO) is coloured cyan. a) Dimer arrangement of sfGFP148x2 with triazole linkage shown as sticks. b) azF-SCO triazole linkage showing electron density as grey mesh. c) Top down view of sfGFP^{148x2} showing distance between chromophores. d) Stick representation of hydrogen bonding network at sfGFP^{148x2} interface.

Formation of the cross link causes residue 148 of the azF monomer to shift away from the chromophore. This creates a cavity that is occupied by a water (W1^{azF} in Figure 4.17a), that forms hydrogen bonds with the chromophore and 148azF. In the sfGFP^{148SCO} monomer unit these interactions are mirrored with CRO^{SCO} and 148SCO (W1^{SCO}). Both these structured waters replace the disrupted hydrogen bond from mutating His148 to either azF or SCO. They also form key components of an extended hydrogen bonding network, predominantly composed of water molecules, spanning the dimer interface and connecting both chromophores (Figure 4.17b). Each monomer (sfGFP^{148azF} and sfGFP^{148SCO}) contributes three symmetrical water molecules (W1, W2, and W3) with the central water W4 and the backbone of F145^{SCO} provide the bridge across the interface, of the two monomers, that links the two networks together. Hence, dimerisation forms an extended, inter-monomer proton wire formed mostly by structurally managed waters linked by hydrogen bonds and allowing for the

deprotonation of both chromophores causing the switch from CroA state to the CroB form.
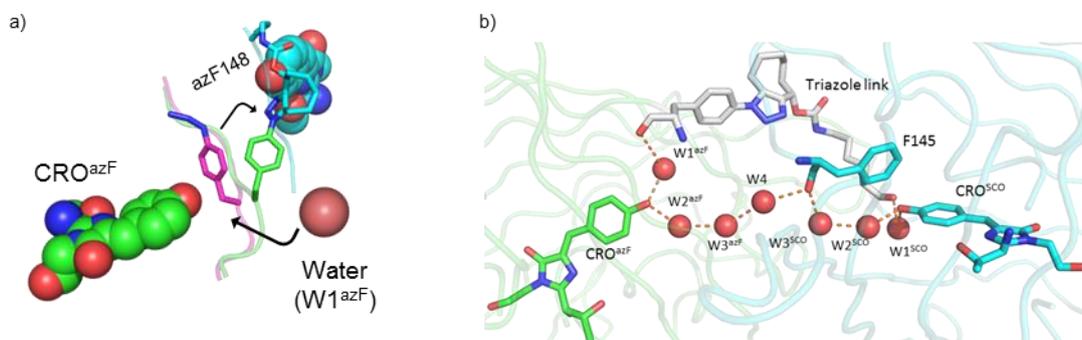
Figure 4.17 Structural change caused by dimerisation of sfGFP[148] a) Conformational change to azF148 on dimerisation. Residue 148azF from 5BT0[115] coloured magenta. b) Water network linking both chromophores of sfGFP[148x2] (CRO[azF] and CRO[SCO] for azF and SCO monomers respectively), forming a proton wire.

4.2.7.3   Comparisons between sfGFP[148x2] chains and sfGFP monomers

To determine if there are any major structural changes, both chains of the sfGFP[148x2] structure were aligned with each other and with the structure for sfGFP[148azF]+DBCO-amine described in Chapter 3. Full backbone alignment of the A and B chains of sfGFP[148x2] yielded an RMSD of 0.12 Å. This suggests that there is very little structural difference between both chains. Upon visual inspection we see no significant differences to the overall protein structure (Figure 4.18a). The β-strand containing

residue 148 also showed no significant distances, save for the rotation of the χ1 dihedral

of residue 148 of ~30° and a 2.2 Å backbone shift of residue Ser147 (Figure 4.18b).
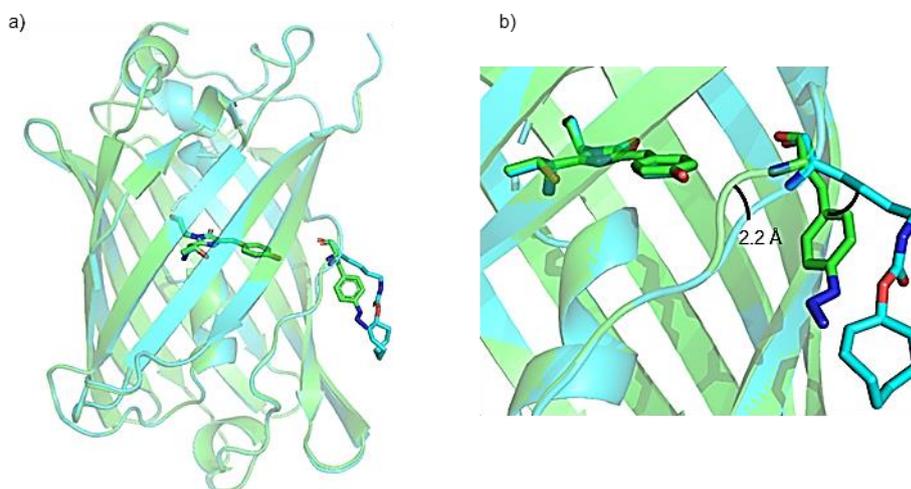
Figure 4.18 Structural comparison between Chain A (green) and Chain B (cyan) of sfGFP[148x2]. a) overall structure, b) comparison of region around residue 148. Residue 148 and CRO of both chains shown as sticks.

Alignment of chains A and B to the structure of sfGFP[148azF]+DBCO-amine (5DY6

yields back bone RMSDs of 0.27 Å for both chains. As with alignment between chains

A and B, alignment to the DBCO-amine modified sfGFP shows no perturbations of note

(Figure 4.19a), save on β-strand 7, where residue 148 occupied different χ1 dihedrals

(Figure 4.19b). Figure 4.19 shows that β-strand 7 of chain A and sfGFP[148azF]+DBCO-

amine align very closely. However, β-strand 7 of chain B does not align with

sfGFP[148azF]+DBCO-amine with a similar shift described when compared to chain A

above.
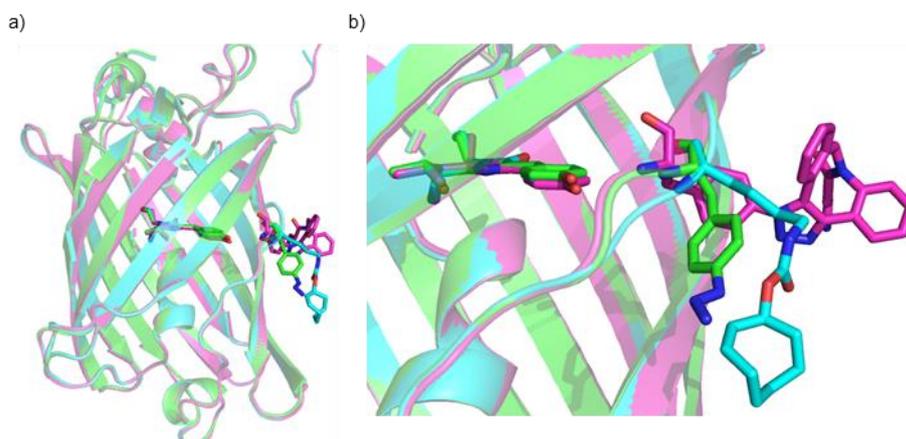
Figure 4.19 Structural comparison between Chain A (green) and Chain B (cyan) of sfGFP^{148x2} and sfGFP^{148azF}+DBCO-amine (magenta). a) overall structure, b) comparison of region around residue 148. Residue 148 and CRO shown as sticks.

Backbone alignment to sfGFP^{WT}, had RMSDs of 0.31 Å and 0.42 Å for sfGFP^{148x2} chains A and B respectively, with most variation occurring in flexible loop regions and at the N and C termini (Figure 4.20). One major difference between both chains A and B compared to sfGFP^{WT} was at the loop containing residues 187-195 which showed the same conformation differences as described in Chapter3 between the WT structure and sfGFP^{148azF}+DBCO-amine, which reaffirms the suggestion that this loop is flexible, and structure varies dependent on crystal packing (Figure 4.20b). There is also a 2.9 Å shift of β-strand 7 in Chain B compared to sfGFP^{WT} (Figure 4.20c), which could explain why there is still a significant CroA population in sfGFP^{148x2} with the strand occupying a similar position to the crystal structure of sfGFP^{148azF}.
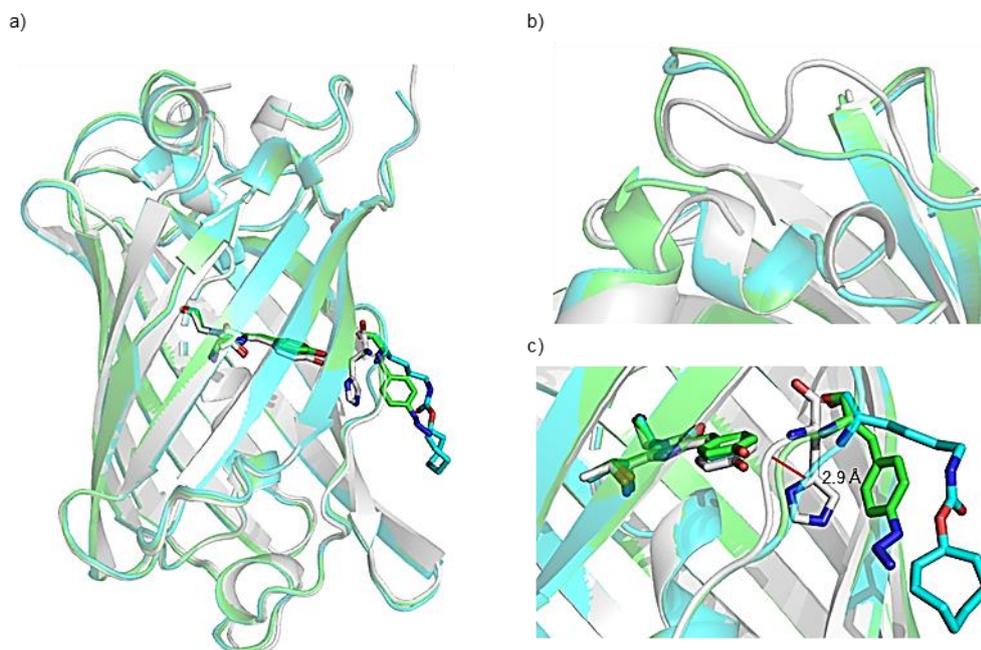
Figure 4.20 Structural comparison of sfGFP^{WT} (Grey) and sfGFP^{148x2} Chain A (green) and B (cyan). a) overall protein structure showing minor deviations at loops and termini. b) Closeup of loop 187-195 region showing large variations between WT and dimer structures. c) Close up of 148 region showing 2.9 Å shift in β-strand 7 for Chain B in comparison to sfGFP^{WT}

4.2.7.4   Comparison with sfGFP^{204x2}

The structure of sfGFP^{204x2} was solved by Dr Pierre Rizkallah and Jacob Pope. The crystal diffracted to a resolution of 1.28 Å, with a $P2_12_12_1$ space group. There were 4 molecules in the unit cell *i.e.* two dimers, both showing density around the linkage site. sfGFP^{204x2} dimers form a quasi-symmetrical "head-to-head" alignment with an ~70° offset between the two monomers (Figure 4.21a). The protein monomers are arranged such that the C termini are close together at one end, as are the loops linking β-strands 10 and 11 at the opposite end of the protein, forming a mirror image between monomers (Figure 4.21a) This arrangement is stabilised by the formation of the azF-SCO triazole linkage in the *'syn'* conformation, causing a reverse turn type structure (Figure 4.21b). The two chromophores are ~22 Å apart and are not in plane with each other (Figure 4.21c).
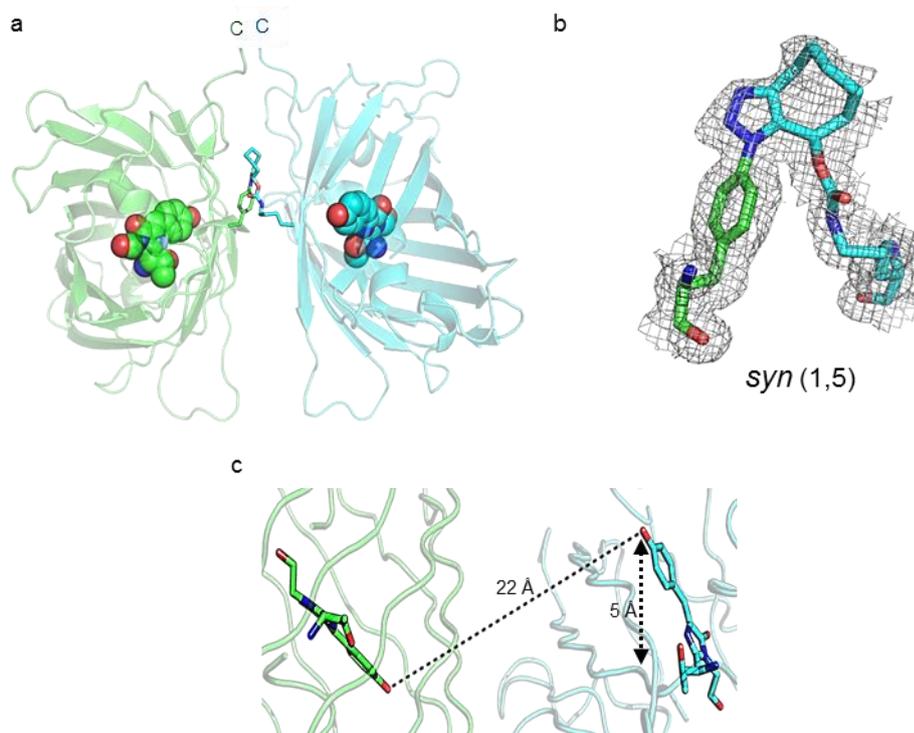
Figure 4.21 Structure of sfGFP^{204x2}. sfGFP^{204azF} coloured green and sfGFP204SCO is coloured cyan. a) overall monomer arrangement and interface with azF-SCO crosslink shown as sticks. and chromophores shown as spheres. b) Close up of triazole linkage with electron density map overlaid. c) Chromophore arrangement within sfGFP^{204x2}

The sfGFP^{204x2} dimer interface is divided into two halves; a directly contacting interface, and a narrow (~5-6 Å wide) water filled cavity coordinating several organised water molecules (Figure 4.22a). The interface is maintained by a non-symmetrical network of hydrogen bonds between residues, N146, S147, T225 of sfGFP^{204azF} and residues N39, and R73 of sfGFP^{204SCO}, with the triazole linkage also contributing (Figure 4.22b). The hydrophobic interactions are more symmetrical, with residues V206, L221 and F223 from both monomers contributing (Figure 4.22c). The two chromophores of sfGFP^{204x2} are linked *via* an indirect hydrogen bonding network involving: 204SCO, the triazole linkage, various side chain residues and the organised waters in the interface cavity (Figure 4.23).
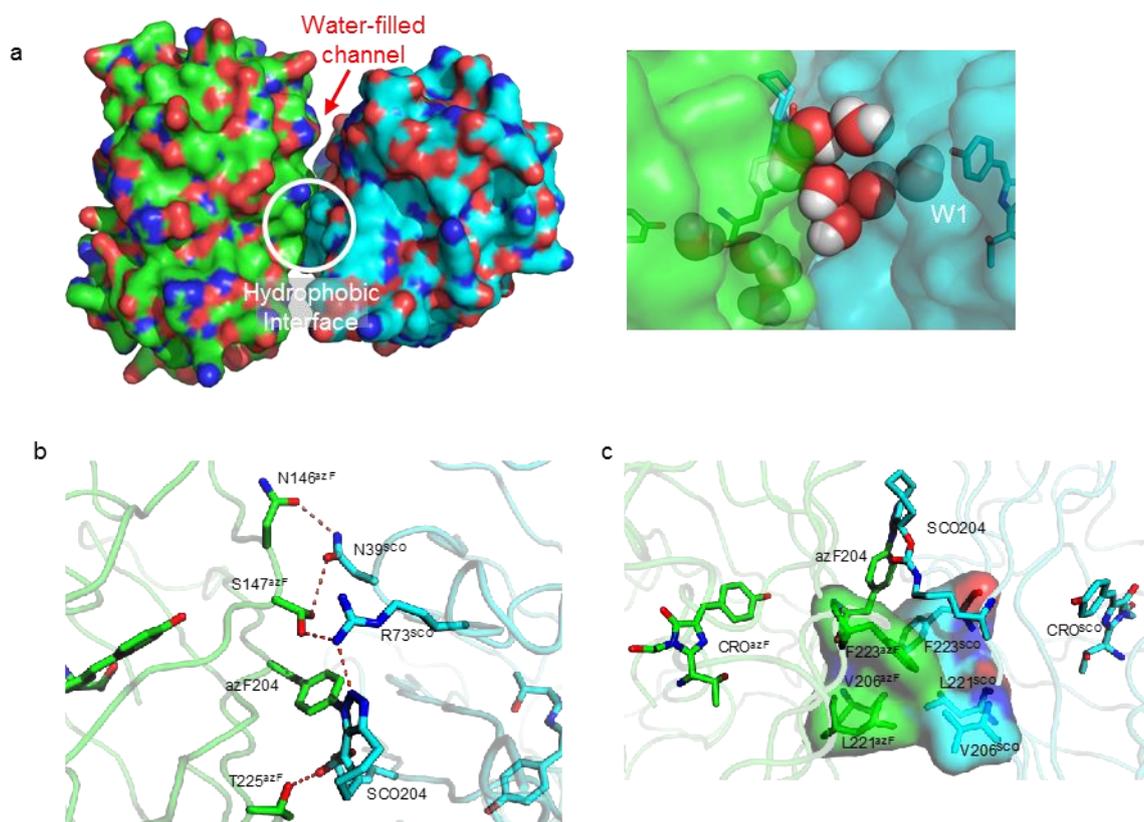
Figure 4.22 sfGFP[204x2] dimer interface. a) Dimer interface including presence of water molecules within inter-monomer cavity (right). Conserved water W1 is highlighted. b) Interface hydrogen bonding between both monomers. c) Hydrophobic interface below the SPAAC linkage site.

Although residues 148 and 204 are close together on adjacent β-strands (7 and 8 respectively), there are several key differences between the two dimer structures. Firstly, monomer arrangement in sfGFP[204x2] is very different compared to sfGFP[148x2]. Compared to the "head-to-tail" monomer arrangement found in sfGFP[148x2], sfGFP[204x2] forms a "head-to-head" alignment with an ~70° offset between the two monomers. This arrangement is stabilised by the formation of the azF-SCO triazole linkage in the *'syn'* conformation, causing a reverse turn type structure, as opposed to the elongated *'anti'* conformation found in sfGFP[148x2]. It should also be noted that the density around the link for sfGFP[204x2] was not as well defined as in sfGFP[148x2], suggesting that the link was dynamic or there was a small population of unit cells where the monomer arrangement was reversed. The two chromophores are ~22 Å apart and are not in plane with each other, as is the case with sfGFP[148x2], suggesting that potential communication between

monomers would be more direct in sfGFP[148x2] than in sfGFP[204x2]. The interface surface area, as calculated by PISA (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html),[223] of sfGFP[204x2] is almost half the area of sfGFP[148x2] (900 Å^2 versus 1600 Å^2), suggesting a weaker dimerisation interface as there are fewer contacts formed between the two monomer units.[221]
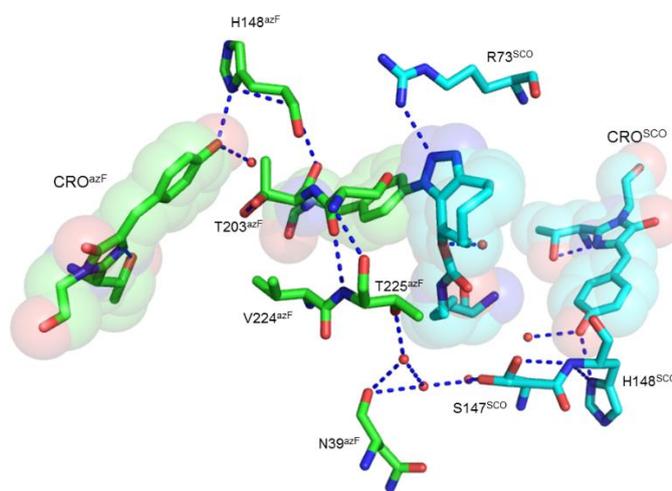
Figure 4.23 The hydrogen bonding network connecting CRO[azF] and CRO[SCO] of sfGFP[204x2]. Hydrogen bonds shown as blue dashed lines, water molecules shown as small red spheres.

4.2.7.5 The role of water networks in sfGFP dimers

Both sfGFP[148x2] and sfGFP[204x2], like other sfGFP structures,[77,115,127] show evidence of defined water channels thought to act as proton shuttles from the chromophore to the bulk solvent. There is also another water channel that interconnects both chromophores within the dimer. The classical ball and stick representation used above shows the precise residues and water molecules involved in these channels at a high resolution, however it is a very complex representation, and it can be hard to visualise the direct path and size of potential channels. Using CAVER,[181] which calculates both the path of potential channels and the size of channels, can help to find which residues and water molecules are involved in forming a channel. CAVER overlays

a simple representation of potential channels within a structure as shown below for sfGFP[148x2] (Figure 4.24a) and sfGFP[204x2] (Figure 4.24b). CAVER is also able to determine the length of the tunnels and the width at any given point throughout the tunnel, highlighting any bottlenecking points. Of the two structures, sfGFP[148x2] has the shortest and most direct route at ~41 Å long as opposed to 94 Å between the two monomers of sfGFP[204x2]. The tunnel of sfGFP[148x2] is also wider than sfGFP[204x2], at its narrowest the channel is 0.91 Å in diameter whereas sfGFP[204x2] is narrower at 0.76 Å diameter at the channel bottleneck. This has the potential to slow down dynamic exchange of water molecules and hence can impede proton shuttling transfer within the dimer system.[181]
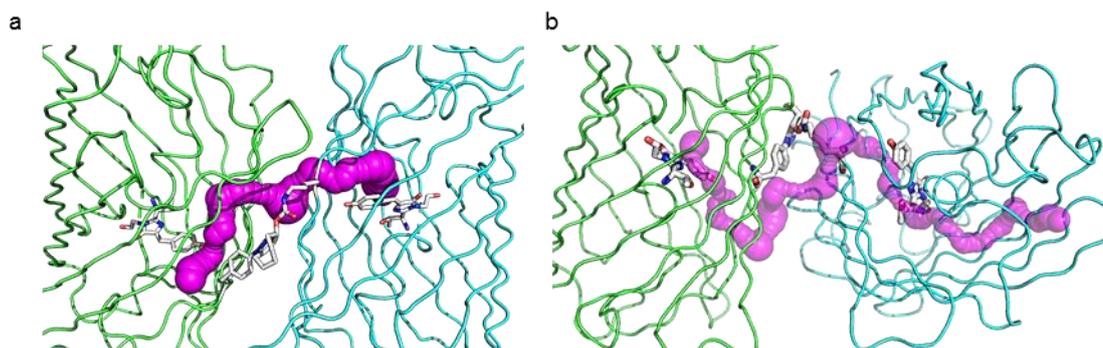
Figure 4.24 Caver analysis of sfGFP dimers. Monomers containing azide coloured green and monomers containing SCO coloured cyan. Predicted channels for a) sfGFP[148x2] and b) sfGFP[204x2] are displayed in magenta.

It has been proposed that water networks play a vital role in the function of fluorescent proteins.[110,115,138,139,185,217] In WT GFP and other monomeric FPs from *A. victoria*, water molecules other than the directly bonded CRO water molecule (W1,Figure 4.17), that contribute to the network are largely exposed to bulk solvent and hence subject to dynamic exchange. However, in the sfGFP dimer structures above these key waters are held within the dimeric interface and hence, buried either wholly or partially forming hydrogen bonds with the protein, slowing down their exchange with bulk solvent.

Surface burial and reduced dynamics may only be simple effects, but they may make an important contribution to the observed functional effects shown above. As such the structural insights into how proteins can communicate between each other may be a foundation to transfer the combination of *in silico* design and SPAAC to create protein dimers into other protein systems. In addition, the reduced dynamics observed due to partial burying of solvent molecules, is likely to have a key role in the formation of the more concerted, organised, and permanent proton wire network as observed in sfGFP[148x2].

4.2.7.6 Comparison to the *in silico* model

One of the central aims of this work was to design functionally enhanced protein dimers using *in silico* predictions. As described above (Section 4.2.1), five models were generated and to test their validity they need to be compared to the solved structures. A table of RMSDs after backbone alignment of each model vs both dimers is displayed below (Table 4.5). The best model to align with sfGFP[148x2] was the highest ranked model (Model 5) with an RMSD of 4.72 Å, whilst sfGFP[204x2] aligned best with Model 4 (Ranked 3rd) with an RMSD of 5.62 Å, which are much higher compared to backbone alignments with sfGFP[WT] with RMSDs of 0.4 and 0.2 Å for sfGFP[148x2] and sfGFP[204x2] respectively.
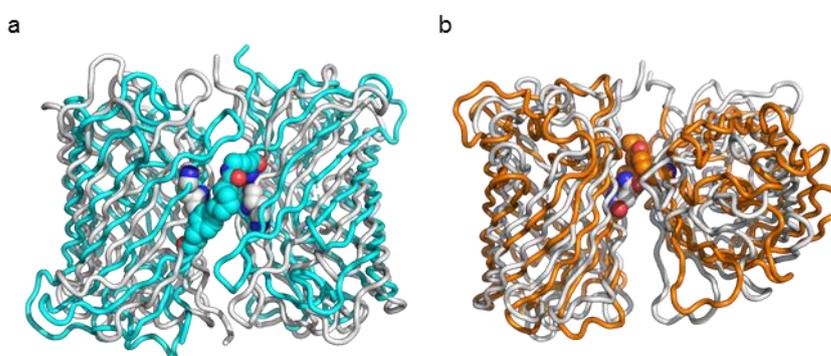
Figure 4.25 Comparison of *in silico* (grey) v structurally determined sfGFP dimers a) sfGFP[148x2] (cyan) v Top ranked model. b) sfGFP[204x2] (orange) v 3rd ranked model.

Visual inspection of the alignment of both sfGFP148x2 v Model5 (Figure 4.25a) and sfGFP204x2 v Model 4 (Figure 4.25b) highlights how poorly the molecules align with very little visible backbone overlap. Therefore, the models are not like the solved structure. However, when taking into consideration that the initial clustering for each model included all structures within a 9 Å RMSD, it is possible that the precise monomer conformations observed in both crystal structures are included in these clusters but without viewing all the members of each cluster and aligning to each we cannot be sure. It should also be noted that ClusPro generates dimers using a rigid backbone based algorithm and as such does not account for subtle permutations of side chains. Despite these limitations *in silico* modelling was able to generate viable models for providing a starting point to the engineering process limiting target selection to a narrow region of a protein as opposed to 'blindly' selecting residues from the whole protein. As such the aim of using *in silico* techniques to further the design of dimers has been successful, not only in showing which residues to select but in which ones not to select as shown in the failure to generate any dimeric protein by incubating sfGFP[132SCO] with either sfGFP[132azF] or sfGFP[204azF].

Table 4.6 Comparison of RMSDs of backbone alignment of *in silico* models v determined structures.

Model	Energy score (kJ/mol)	RMSD v sfGFP148x2 (Å)	RMSD v sfGFP204x2 (Å)
Model 5	-503.94	4.72	12.77
Model 1	-501.966	9.53	20.55
Model 4	-497.071	18.96	5.62
Model 2	-497.112	11.95	20.83
Model 3	-494.492	8.31	8.54

4.3 Conclusions

The most commonly found protein units in nature is symmetrical protein dimers. Synergistic effects are common, with inter-monomer communication occurring through shared interaction networks. This chapter has shown that, monomeric proteins units can successfully generate dimers through *in silico* prediction of potential dimeric interface regions and use of bioorthogonal ncAA chemistries particularly SPAAC. We have also shown that symmetry is a key factor in creating functional synergy between monomers, through the formation of continuous transport networks. Generating new protein oligomerisation systems is a topic of great interest in protein engineering. [14,25,26,35,224] It allows a route to understanding this commonly observed biological process, explores new functional and structural space, and expands use of proteins in nanoscience. One of the key challenges to creating functionally linked dimers, has been to generate functional links between separate subunits beyond the region of the interface, which has been achieved in this chapter. This methodology is not limited to the chemistry and ncAAs described above, with other strained alkyne regioisomers[92,109] and linking chemistries available,[121,122,214] and hence open to a vast variety of dimer conformations not explored here. Coupling this new knowledge with advancements in ncAA incorporation such as, codon reprogramming to allow multiple different ncAAs to be incorporated[225,226] and cell lines that have had reprogrammed codons replaced to reduce non-specific ncAA incorporation,[227-229] we are able to design higher order functionally linked oligomers for use in synthetic biology and in nanoscience. This chapter focused on creating functionally linked homodimers, the next step is to design and create heterodimers between different proteins to study how their functions would integrate/enhance. This area of research along with generating oligomers above dimers is currently being undertaken at the Jones group.

5 Precise interfacing of proteins with single-walled Carbon Nanotubes (swCNTs) using genetically encoded phenyl azide chemistry

5.1 Introduction

Proteins perform a wide variety of functions ranging from catalysis and molecular recognition to charge transfer and energy conversion.[119] The dimensions of many proteins are in the nanometer range: from 2-10 nm for monomeric proteins to >100 nm for complexes. Thus, given the variety of functions performed by proteins and the scale they work at, proteins can be considered nature's nanomachines.[119] To exploit these biomolecular interactions with single molecule specificity and control, is a central challenge for biotechnology, as is the assembly of platforms to monitor these events.[230-234] Studying proteins at the single molecule level allows for measurement of dynamics and distributions of molecular properties and population states, usually masked by ensemble measurements that average the signal over all the population. Utilising proteins at the single molecule level can also lead to increased sensitivity for many applications, from biosensing to molecular electronics.[54,71,231,235] The increased sensitivity arises by being able to look at differences between bulk effects and differing behaviours of individual molecules. By using orientation specific, directed attachment, we can observe many single molecules acting in the same way, hence amplifying the overall signal. Therefore, there is an increased interest is developing strategies that permit single biomolecule interfaces to be generated that are directly linked to nanoelectronic systems such as nano carbon.

The aim of this chapter is to create stable protein-swCNT interfaces that can be used for downstream applications such as, nanoelectronics and biosensors.[52,54,57,60,69,235-240] A key issue in development of protein-swCNT interfaces, is ensuring site-specific attachment to a nanoelectrode. Ideally, in solution assembly is

required to develop solution processing single-molecule bioelectronic devices. However, with pristine swCNTs this is not possible due to the low solubility of swCNTs.

There has been great interest in using one-dimensional nanostructured materials as nanoelectronic systems for biological interfacing, and single walled carbon nanotubes (swCNTs) emerged as an important and viable option.[52-56] Previous work has reported that biomolecules near swCNTs can alter the electronic properties of the nanotubes.[241] Various mechanisms, such as direct charge transfer or local perturbations of the pi electron network by introducing charged residues to the swCNT surface, cause these changes, with even minute changes in the local charge environment showing an obvious effect in swCNT properties, such as conductance.[52,54,71,238] Appropriate size compatibility is ensured as swCNTs are on the same size scale as biological molecules.

Both covalent[236,242-245] and non-covalent[246-249] chemical strategies have successfully been employed to interface biomolecules (predominantly DNA) with swCNTs, and interactions have been monitored by electronic and optical techniques.[52,250,251] With regards to proteins however, many of these approaches are non-specific in terms of their attachment site and leading to heterogenous and non-optimal interface locations.[119] This can result in the attachment of multiple uncontrolled molecules onto individual swCNTs in a non-specific manner. As a result, there is currently little scope for measuring individual proteins interfaced in defined orientations. Hence it hampers the development of single-molecule interfaces for device implementation.

The importance of orientation and site-specific attachment has been demonstrated by Holland et $al.$[49] for glucose oxidase attached to gold nanoparticle using mutations at different residues. Previous work from the Jones lab[51] demonstrated with cytochrome b_{562} that engineered thiol groups and pairs of thiols can be used to assemble the protein in different orientations on gold. This showed how different orientations lead to different conductance profiles. However, despite this precise control at the single

molecule level, over orientation and number of proteins attached to nanoelectronic systems such as swCNTs is still an issue.[49,51]
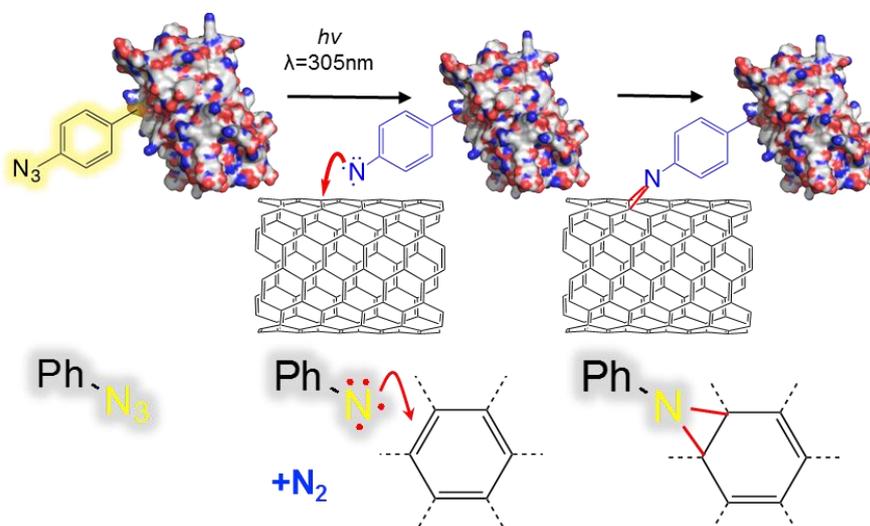
Scheme 5.1 Photochemical insertion of a protein *via* genetically encoding phenylazide chemistry. sfGFP (PDBid = 2B3P) is used as a model protein.

In this chapter, the interfacing of proteins to swCNTs in an orientation specific manner is explored, using phenyl azide chemistry. Phenylazide photochemistry is a powerful approach to attaching molecules to swCNTs; irradiation with ultraviolet light (UV) results in the formation of a highly reactive nitrene radical that can react with the electron rich SP^2 carbon bonds in swCNT side walls,[113,252] creating a permanent crosslink (Scheme 5.1). However, phenyl azide photochemistry has typically only been used for attaching small molecules to swCNTs, as biomolecules typically don't have the required chemistry. As described in the previous chapters, I have shown that using a reprogramd genetic codon,[183] it is possible incorporate phenyl azides into proteins using the non-canonical amino acid (ncAA) p-azido-L-phenylalanine (azF).[94,152,182] UV irradiation of azF containing proteins will result in covalent attachment of proteins to swCNTs, at a defined residue in a specific orientation, creating an intimate interface between protein and swCNT.

The use of phenyl azide chemistry opens up additional avenues of protein-swCNT interfacing through the use of bioorthogonal Click reaction; here I describe how such bioorthogonal Click reactions can be used to generate 1:1 end-wall protein conjugates to swCNTs. The attachment of these proteins is shown to be specifically directed in orientation of attachment. The swCNTs were functionalised at Queen Mary University (London, UK) by Dr Mark Freeley (Palma group). The process involves wrapping short single stranded DNA oligonucleotides (ssDNA) around swCNTs.[249] This process solubilised the tubes, due to tight binding from π–π stacking interactions.[249] DNA wrapping of tubes also protects the sidewalls leaving only the ends of tubes available to functionalisation *via* carboxylic acid groups.[253-255] DBCO-amine groups were linked to the swCNTs end walls, for attachment with azF containing proteins, using bioorthogonal "1+1" click chemistry, known as strain promoted azide-alkyne cycloaddition (SPAAC; Figure 5.1a).[115,117,118]

Autofluorescent proteins such as super-folder green fluorescent protein (sfGFP)[127] are very useful for nano electronics. They come in a wide range of different colours[9,127,256] creating a large range of absorption spectrum coverage that can be used for photo capture systems.[159] They also have been shown to exhibit light modulated conductance and stimulated emission by forming complex energy transfer systems.[136,137] sfGFP was used to demonstrate the importance of linkage site for communication between proteins and swCNTs. Two swCNT-protein configurations were investigated, where sfGFP mutants were engineered with azF at residues proximal (Residue 204) and distant (Residue 132) to the chromophore of sfGFP (Figure 5.1b). The effects of mutations on sfGFP have been explored previously (Chapter 4 and [115,149,151,152]). Effects on GFP fluorescence on interfacing with swCNTs were monitored by fluorescence both in solution and on surfaces at single protein level, where evidence of coupling between swCNTs and sfGFP was observed. Part of this work has been published in Freeley *et al*.[119]
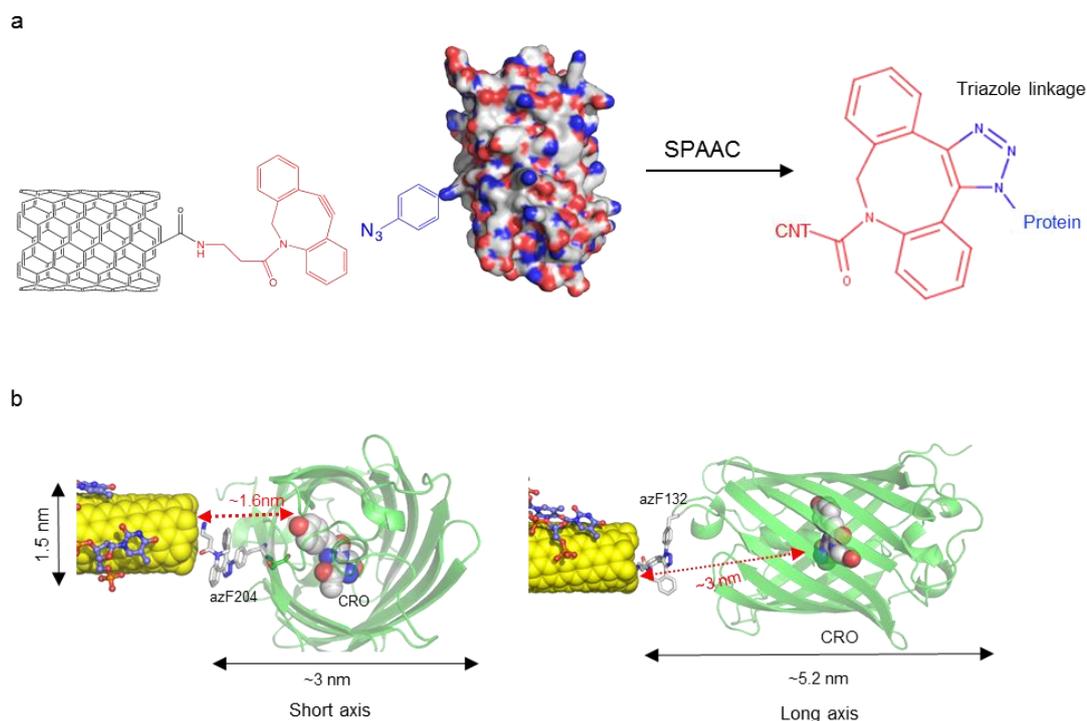
Figure 5.1 Single protein attachment of sfGFP to swCNTs. a) Scheme showing conjugation of sfGFP to functionalised swCNT using SPAAC. b) Schematics of the short axis sfGFP[204azF] (sfGFP[SA]) and long axis sfGFP[132azF] (sfGFP[LA]) swCNT interface points and orientations. Linkage residues are shown as grey sticks and swCNT shown as gold spheres. Approximate distances between swCNT end wall and sfGFP CRO are shown in red.

5.2 Results and discussion

5.2.1 Creation and purification of sfGFP[SA] and sfGFP[LA]

Plasmids for sfGFP[204azF] (sfGFP[SA], Chapter 4) and sfGPF[132azF] (sfGFP[LA] , Chapter 4) were kindly donated by Reddington[151] and were co-transformed into TOP10 (Invitrogen) *E. coli* cells with pDULE[cyanoRS], expressed, and purified as described in Section 2.4.3 and [115,151,152] by nickel affinity chromatography. SDS-PAGE analysis of sfGFP[LA] and sfGFP[SA] showed a single protein band at ~27 kDa with >95% purity (Figure 5.2). Purified protein was buffer exchanged into fresh 50 mM Tris-HCl [pH 8.0] and the concentration determined.
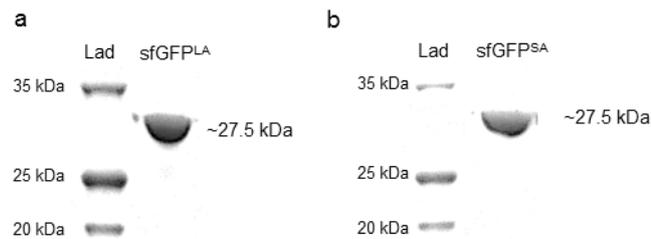
Figure 5.2 SDS-PAGE analysis showing purified a) sfGFP^{LA} b) sfGFP^{SA}. Weights of molecular marker (Lad) are labelled.

To ensure that modification of the proteins with either UV irradiation or DBCO-amine had no effect on sfGFP fluorescence, aliquots of sfGFP^{SA} and sfGFP^{LA} (10 µM) were modified with either irradiation by UV light (305nm) for 15 minutes, or addition of DBCO-amine (50 µM) overnight at room temperature. Absorbance and fluorescence spectra of sfGFP^{SA} and sfGFP^{LA} were taken before and after modification. As shown in Figure 5.3a, modification of sfGFP^{LA} yields no discernible loss of absorbance with either UV irradiation or DBCO-amine modification. However, modification by either route does instigate a loss of ~10% fluorescence emission (Figure 5.3a). As with sfGFP^{LA} modification of sfGFP^{SA} with UV irradiation and DBCO-amine shows no discernible effect on absorbance with only a circa 10% drop in emission (Figure 5.3b). The loss of fluorescence observed by modifying sfGFP^{SA} with either proposed route is to be expected given the proximity of residue 204 to the chromophore (0.57 nm) and its role in sfGFP water networks.[115,134,185] However, the loss of fluorescence in sfGFP^{LA} is surprising given the distance of residue 132 from the sfGFP chromophore (2.59 nm). However, the loop where residue 132 resides, is at the end of a predicted channel, which may play a role in water networking and charge transfer within the core of GFP and thus influence fluorescence. This will be covered in more detail in Chapter 6.
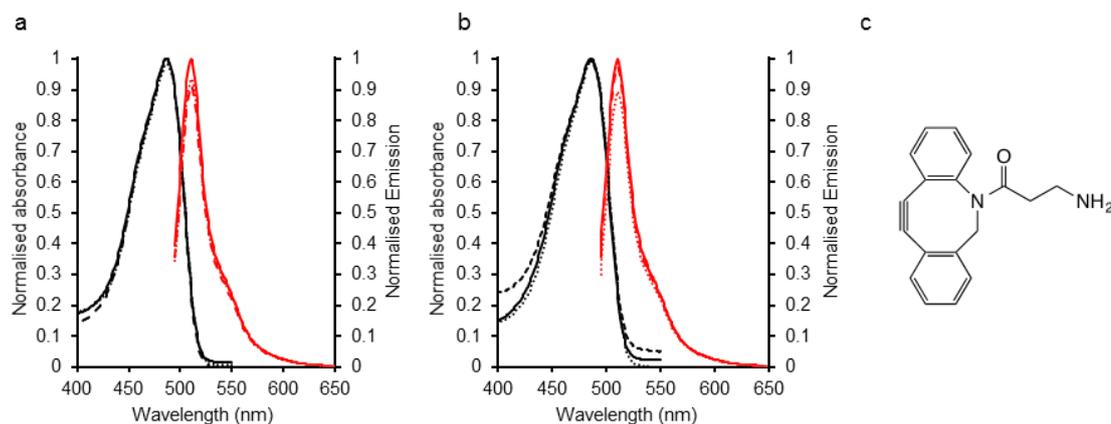
Figure 5.3 Absorbance (Black) and emission spectra after excitation at 485 nm (Red) of a) sfGFP[LA] and b) sfGFP[SA]. Before modification (solid lines), after 15 mins UV irradiation (dotted lines) and after overnight incubation with DBCO-amine (dashed lines) (c) Structure of DBCO-amine. Spectra were normalised to the unmodified protein (black). Click reactions were performed at 10 µM protein to 50 µM DBCO-amine in 50 mM Tris-HCl [pH 8.0] 0.5% DMSO. For fluorescence spectra samples were diluted to 0.5 µM in 50 mM Tris-HCl [pH 8.0].

5.2.2 Attachment of sfGFP[LA] and sfGFP[SA] to swCNT sidewalls.

To attach proteins to the sidewalls of CNTs as described in Scheme 5.1. Aliquots of sfGFP[SA] and sfGFP[LA] (100 nm) were drop casted onto glass coverslips with swCNTs deposited, followed by irradiation with UV (305 nm LED light) and washing with distilled water and ethanol. AFM analysis was performed by Dr Adam Beachey and Miss Suzanne Thomas (School of Physics and Astronomy, Cardiff University) using Gwyddion image analysis software (http://gwiddion.net).[257] The overall morphology of the AFM images for the two variants suggests that they attach in regular and defined orientations across the swCNT based on the location of the azF residue within the protein. sfGFP[LA] is generally circular with a narrow diameter while sfGFP[SA] appears to bind with its long axis almost perpendicular to the swCNT (Figure 5.4).

The measured heights are lower than expected based on the crystal structure.[127] Both variants measured *circa* 1 nm high as opposed to ~3.1 nm and ~5 nm for sfGFP[SA] and sfGFP[LA] respectively. It should be noted that the proteins are unlikely to be uniformly attached at the apex of the tube and the azF is placed in a flexible loop,[77] so a true height profile of the protein alone relative to the tube will be difficult to measure. Furthermore,

AFM is known to underestimate heights of soft matter such as biomolecules.[159,258] The overall difference in width of the protein spots compared to swCNT is an also important independent validation of protein binding.
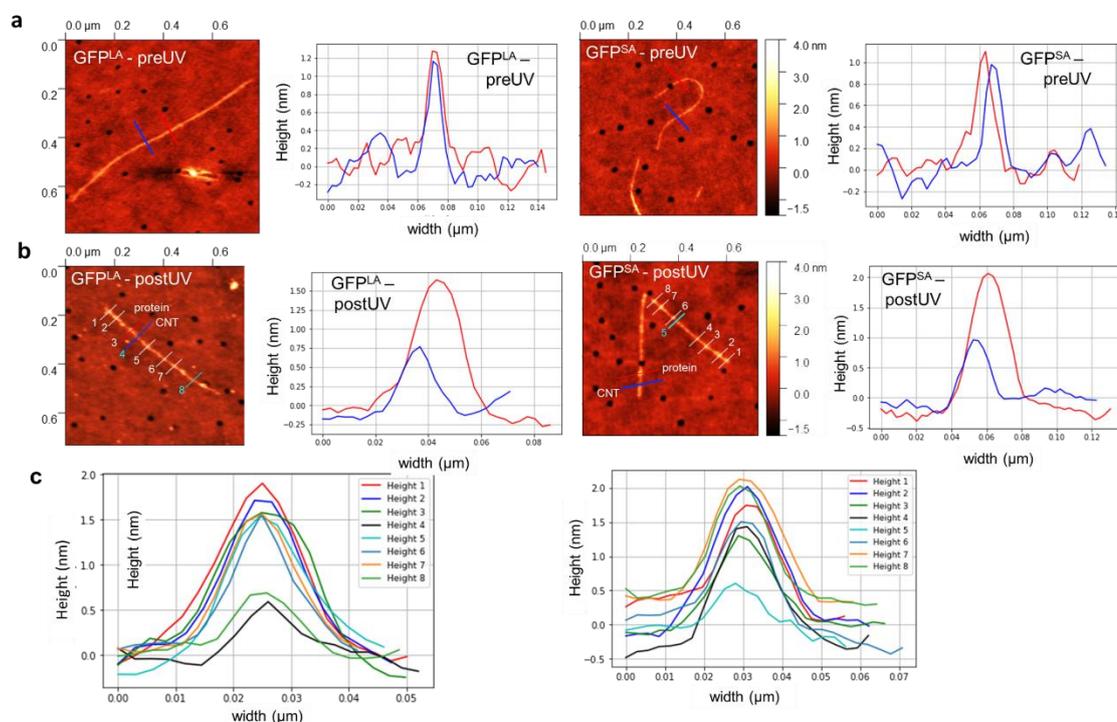
Figure 5.4 AFM analysis of sfGFP[LA] (Left hand panels) and sfGFP[SA (Right hand panels]. Each panel presents a representative AFM image of protein (100 nM) deposited on coverslips containing swCNTs together with height analysis of representative areas. (a) Control samples without illumination (preUV) and after washing; (b) samples after illumination at 305 nm (postUV) and washing. There are clear differences between sample pre- and post-illumination, including distinctive spots with heights and widths larger than of the swCNT, indicative of protein attachment only on irradiation. The height analysis in (a) and (b) corresponds to areas of the equivalent colours in the AFM image. In (b), blue represents a bare CNT region (labelled CNT on AFM image) and red a sfGFP attachment area (labelled protein on AFM image). (c) Height analysis across the length of the tube shown in (b). The heights in the plot key correspond the equivalent areas numbered on the AFM image. Heights 4 and 8 (cyan) in the AFM image in (b) correspond to bare tubes between protein attached areas (white).

When the azF sfGFP variants were incubated in the dark with swCNTs, little permanent binding was observed (Figure 5.4a). Initial AFM scans suggested that there may be some residual GFP[LA] bound to swCNTs in the dark but the peaks thought to correspond to protein disappeared on repeat scanning and the tip quickly became fouled. sfGFP[WT] did not display any apparent binding prior to and after illumination with

UV light confirming the requirement of the azF to generate a permanently attached protein (Figure 5.5).
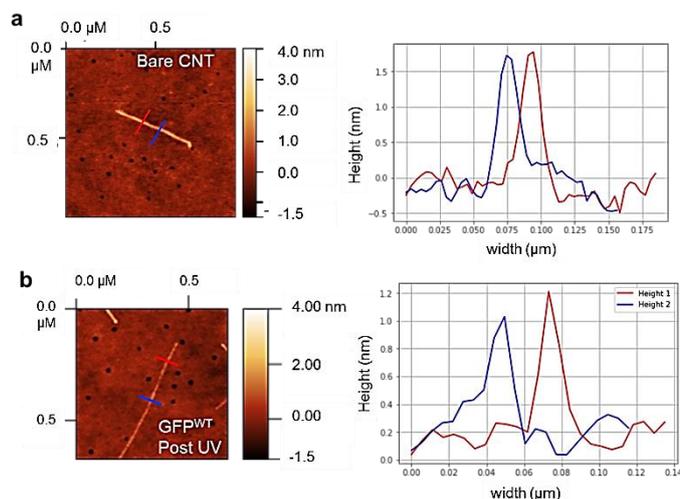
Figure 5.5 Representative AFM images and height analysis of (a) bare swCNT and (b) wild-type sfGFP[WT] (no azF incorporated) after illumination and washing. The height profiles shown in the right-hand panels represent the equivalent coloured the lines in the AFM images. In the sfGFP[WT] sample, protein was applied (100 nM) and subjected to 305 nm illumination as for the azF containing proteins. There is no evidence of permanent protein binding given both the lack of defined spots on the swCNTs together with heights and width profiles.

5.2.3 Functional effects of sfGFP sidewall attachment

GFP provides an ideal model to study the effect attachment to swCNTs has on protein function, using the inherent fluorescence of the protein as a useful output for monitoring function down to the single molecule level. As fluorescence is sensitive to even small changes to the local environment and protein structure, it allows impact on function and thus potentially communication between the two systems to be monitored.[9] Total internal reflection fluorescence microscopy (TIRF) was used to investigate surface bound molecular fluorescence at single molecule resolution. For this sfGFP variants were cast onto swCNTs fixed onto glass coverslips. TIRF was carried about by Dr David Jamieson (School of Pharmacy). Comparative images covering the same area following AFM then TIRF of swCNT bound sfGFP[SA] are shown in Figure 5.6 showing a correlation,

between areas of high fluorescence with areas with a high density of proteins attached to swCNTs. This confirms that that sfGFP^{SA} is functional after attachment to swCNTs.
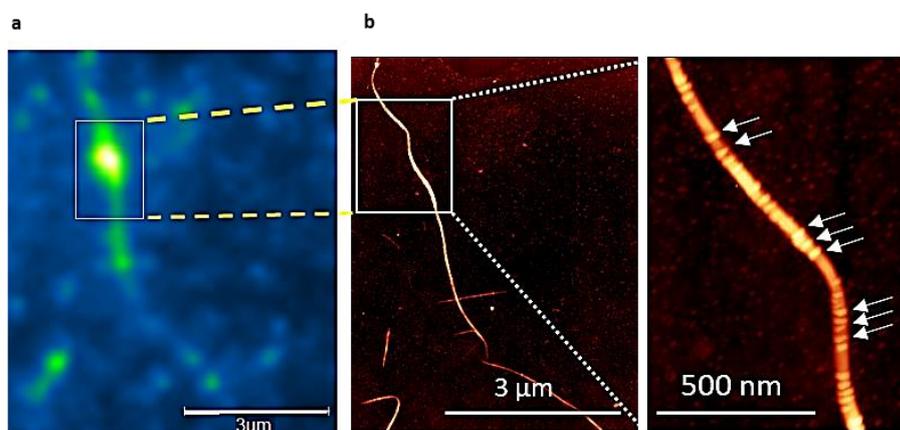
Figure 5.6 Functional analysis of sfGFP^{SA} interfaced with swCNTs. A) TIRF image with approximate corresponding area b) imaged by AFM (magnified area shown to right of main AFM image). Presence of some protein molecules are indicated by white arrows.

For both sfGFP^{SA} and sfGFP^{LA}, linear structures were observed by TIRF that equate to protein bound to swCNTs while an irradiated mix of sfGFP^{WT} and swCNTs yielded no such obvious structures (Figure 5.7). In comparison to free protein, reduced emission intensity was observed (Figure 5.8a-b). Analysis of TIRF data reveals the function of GFP is significantly altered on attachment to swCNTs, with the position of attachment having an additional influence. Single molecule imaging shows that sfGFP^{LA} has increased rate of blinking when bound to tubes in comparison to free protein (Figure 5.8d),[101,216] while sfGFP^{SA} displays more classical single state characteristics. Both proteins free in solution exhibit single step photobleaching, confirming that single proteins were observed for both variants (Figure 5.8c-d).
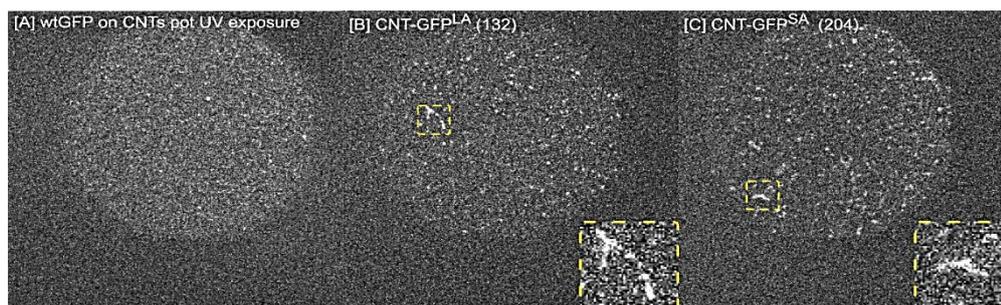
Figure 5.7 Single molecule TIRF imaging of a) sfGFP^{WT} (wtGFP), b) sfGFP^{LA} and c) sfGFP^{SA} with CNTs following UV treatment. sfGFP^{WT} is not observed to undergo UV initiated binding to CNTs, with no fluorescently decorated tube-like structures visible. sfGFP^{LA} and sfGFP^{SA} variants are observed to decorate CNTs, with fluorescent tube-like structures visible via TIRF imaging.

After attachment to swCNTs, both proteins show a significant reduction in fluorescence intensity compared to the free protein, and each exhibited temporal fluctuations in addition to on/off blinking. TIRF images (Figure 5.8a-b) illustrate this change in behaviour upon sfGFP^{SA} coupling to swCNTs; single spots are observable until they are stochastically photobleached for sfGFP^{SA} free in solution (Figure 5.8a). In contrast, a much weaker fluorescence is observed for swCNT-sfGFP^{SA} with low intensity spots fluctuating and blinking (Figure 5.8b). The addition of successive frames reveals the overall sfGFP^{SA} decorated swCNT structure (Figure 5.8b).

Intensity traces from individual spots of swCNT coupled proteins (Figure 5.8e-f) shows a decrease in fluorescence intensity when both sfGFP^{LA} and sfGFP^{SA} are attached to swCNTs. However, the traces of sfGFP-swCNT hybrids also show an increase in the fluorescence lifetime, *i.e.* the time it takes for sfGFP to photobleach (switch off), of individual protein spots and an increase in blinking frequency (rapid on/off rates), compared to the free protein (Figure 5.9a). This suggests a dynamic photophysical interaction between sfGFP proteins and the swCNT. Whilst swCNT coupled sfGFP^{SA} photobleaches to background levels, swCNT-sfGFP^{LA}, although less bright, has an increased resistance to photobleaching (Figure 5.9b). Both variants exhibit dynamic fluctuations of small magnitude, rarely showing discrete steps, unlike the free protein counterparts. The observed fluctuations are not readily attributed to a

small number of discrete states in the data, suggesting highly dynamic processes, likely the result of continuous variations in events including possible charge or energy transfer between protein and swCNT. Due to length-scales of swCNTs and the diffraction limit, it cannot be ruled out that co-localisation and simultaneous observation of more than one protein are contributing to some of the added complexity of TIRF traces. Given the attachment of multiple proteins to swCNTs it's possible that multiple simultaneous energy/charge transfer events to a single tube create perturbations and feedback to individual proteins, resulting in the more continuous, rather than discrete, energy behaviour observed. This would be consistent with discrete step-like events observed being a result of swCNTs with single or low numbers of decorated protein.
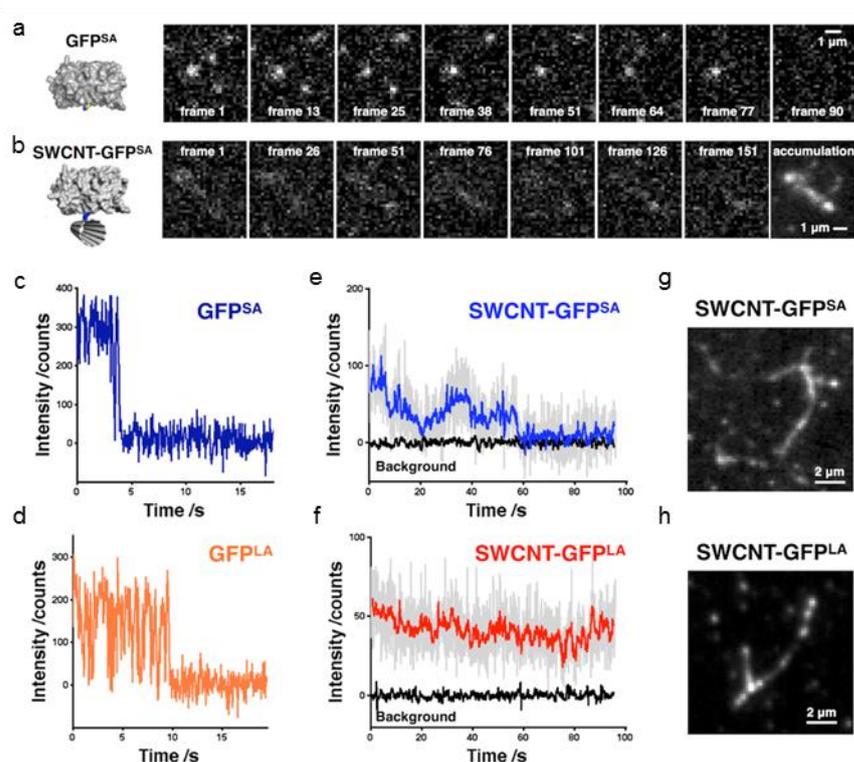
Figure 5.8 Single molecule imaging of sfGFP[SA] and sfGFP[LA] free in solution and CNT bound. a-b) TIRF imaging time course of sfGFP[SA] free and swCNT attached. a) Single free sfGFP[SA] are observed that stochastically photobleach. b) swCNT-sfGFP[SA] is less bright and undergoes significant intensity fluctuations. Summing frames reveals the linear structure of decorated swCNTs. c-f) representative intensity traces of sfGFP[SA] and sfGFP[LA] free in solution (c-d) and swCNT bound (e-f). Attachment alters the photo-physics of the fluorescent properties; reducing fluorescence intensity, prolonging lifetime and inducing fluctuation. Raw data (grey), Chung Kennedy filtered data (blue/red) and Chung Kennedy filtered background intensity traces are shown). g-h) sfGFP[SA] and sfGFP[LA] decorated swCNTs revealed by TIRF.

The major fluorescence changes seen above suggests that there is some degree of interaction between the swCNTs and the attached proteins. The different fluorescent behaviours of sfGFP[LA] and sfGFP[SA] indicate that the attachment position (with respect to the protein) has a role in this apparent communication between protein and swCNT. The classic explanation for this communication is energy between sfGFP and swCNT, which is known to occur in sfGFP when close to a suitable acceptor.[134,259] Simple energy transfer however, cannot explain the longer decay times observed alongside the lower intensities, which may suggest that the photocycling between absorption and emission events is slower. Simultaneous energy transfer events of multiple proteins on a single swCNT could feedback to individual proteins reexciting the chromophore for extended 'on time'. The attachment of multiple proteins to single swCNTs could result in many simultaneous energy/charge transfer events to a single tube creating the feedback to individual fluorescence proteins, resulting in a more continuous behaviour as opposed to the classical discrete stepwise behaviour observed with single molecules. This would be consistent with data of swCNTs with the small step-like events observed with individual tubes with single or few proteins attached to it, limiting the influence of simultaneous energy transfer and feedback.
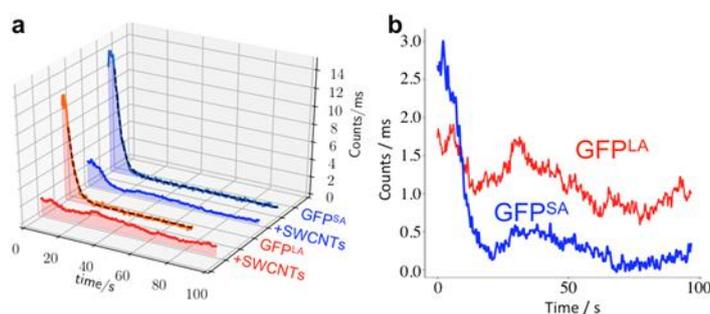
Figure 5.9 Comparison of ensemble single molecule TIRFM photobleaching kinetics of GFP[SA] (blue) and GFP[LA] (red). A reduction in fluorescence intensity and prolongation of photobleaching lifetime is observed for both GFP attachment orientations (a). This change in fluorescence characteristics manifests as different ensemble photobleaching behaviour between GFP[SA] (blue) and GFP[LA] (red) attachment (b). SWCNT attachment ensemble decay deviates from the simple exponential decay of the proteins in solution.

The fluorescence behaviour of sfGFP[LA] was affected to a greater extent than sfGFP[SA], with the former exhibiting reduced intensity but little photobleaching even after ~90s. This is unexpected given the differences in distance and positions between the azF residue and the sfGFP Chromophore. Analysis of the sfGFP structure with CAVER,[181] which detects possible channels including water tunnels, show possible channels between protein and swCNT (Figure 5.10). This observation together with previous work on water-based charge transfer networks in sfGFP,[138,139,185] suggests that residue 132 is close to the exit of an internal water coupled directly to the chromophore. Coupling proteins to the electron rich pi bond network of the swCNT side wall, close to the tunnel exit may provide a mechanism for the influence of the swCNT on fluorescence. While the exact nature of coupling between swCNT and sfGFP and the associated networks needs to be investigated further, it does not appear to be a simple distance dependent relationship highlighting the potential importance of coupling position in terms of bond networks commonly found in protein.
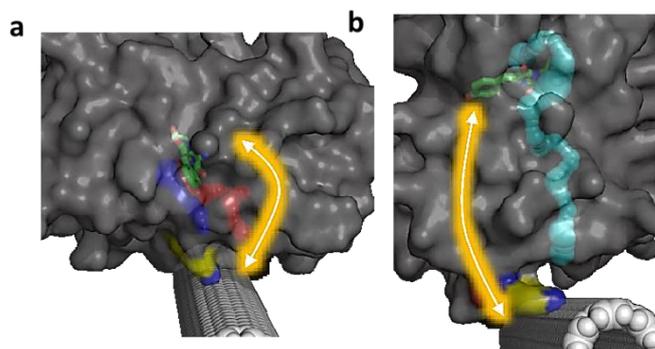
Figure 5.10 Internal tunnels (calculated using CAVER[181]) for GFPSA (a, blue and red) and GFPLA (b, cyan) from the functional centre, the chromophore, to the SWCNT sidewall. The chromophore is coloured green and shown in the stick representation. The surface of the protein is shown in grey and as transparent so that the internal tunnels and chromophore can be visualised.

5.2.4 Attachment of sfGFP^{LA} and sfGFP^{SA} to ssDNA wrapped swCNTs

As shown in the previous section and described by others,[119,236,243-245] it is hard to investigate single molecule interactions with swCNTs due to relatively large areas for attachment, *i.e.* it's hard to control the number and spacing of proteins that attach to the tubes, which means that it is hard to isolate individual proteins. In this section I describe the attachment of single protein molecules the ends of single swCNTs. This is achieved by blocking the side walls of swCNTs from attachment by coating the tubes in single stranded DNA (ssDNA) as described in Tu *et al.*[249] The ssDNA binds to the swCNT side-wall *via π-π* stacking interactions. This allows the tubes to disperse in aqueous solutions as well as protecting the side walls from functionalisation. The ssDNA/swCNTs were functionalised by attaching DBCO-amine to free carboxyl groups at the CNT endwalls. The DBCO-amine linker was chosen as it allowed for proteins to be anchored close to the swCNT end. The short nature of the linker sterically hinders binding of other protein molecules.

Purified sfGFP^{SA} and sfGFP^{LA} were sent to Queen Mary University (London, UK) for attachment to DNA-functionalised swCNTs, as described in[119,249] by Dr Mark Freeley. Protein (1 μM) was mixed with ssDNA/swCNTs (200 nm) and left to incubate overnight. Covalent attachment of sfGFP^{SA} and sfGFP^{LA} mutants to DBCO-functionalized ssDNA/SWCNTs was monitored by casting the hybrid solutions on muscovite mica and imaging the substrate surface *via* atomic force microscopy (AFM).

Representative AFM images of sfGFP-swCNT nanohybrids are shown below (Figure 5.11a +b). AFM reveals both sfGFP^{SA} and sfGFP^{LA}, are tethered uniquely to the terminal ends of individual ssDNA/swCNTs as shown by individual round spots ~3 nm high at the end of each of the modified tubes. Both sfGFP variants linked to swCNTs with similar efficiencies of ~30% (measured on AFM by visually counting tubes with and without proteins). AFM analysis also revealed that of the obtained nanohybrids ~88% of swCNT-sfGFP^{SA} and ~82% of swCNT-sfGFP^{LA} nanostructures exhibited single protein

attachment to only one end of the nanotube. It is believed that protein-swCNT ratios greater than one (*i.e.* 2:1 protein-swCNT hybrid formation) were minimised by the mild sonication process required for forming the ssDNA-swCNTs.[254,260,261] The sonication reduced the number of free carboxyl groups at the swCNT termini and hence reduced DBCO-protein anchoring sites.
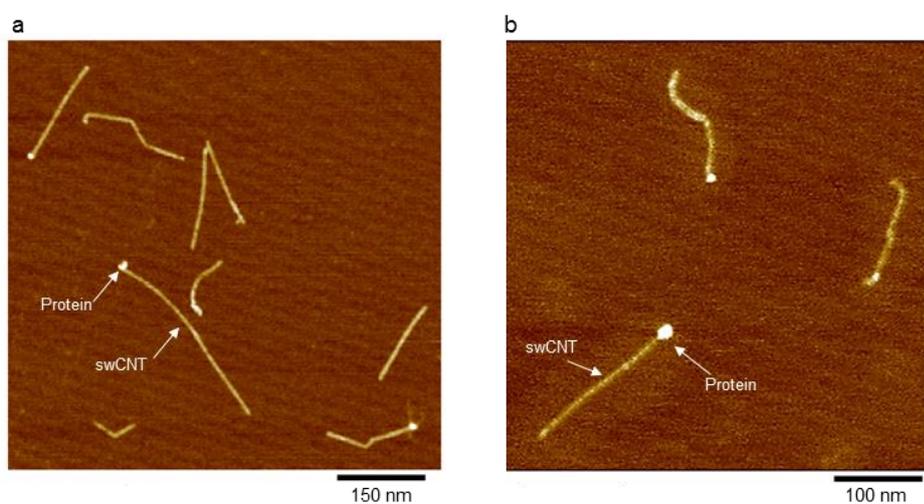
Figure 5.11 AFM images of sfGFP-swCNT hybrids, a) ssDNA/swCNT-sfGFP[SA] b) ssDNA/swCNT-sfGFP[LA]. Z-scale (height) = 6 nm. Images adapted from[119]

5.2.5 Fluorescent Characterization of ssDNA/swCNT-sfGFP nanohybrids.

To demonstrate communication within the ssDNA/swCNT-sfGFP nanohybrids, steady-state fluorescence spectroscopy was performed (Dr Mark Freely, QMU). sfGFPs have been shown to act as light-induced electron donors in photochemical reactions with various electron acceptors,[254,262,263] and changes in sfGFP fluorescence are a good indicator of communication through other coupled processes such as energy transfer.[134] Therefore, the emission intensity of sfGFP was compared before and after conjugation to ssDNA/SWCNTs.

In the case of sfGFP[SA], fluorescence emission is reduced presumably *via* quenching, by ~80% after conjugation with ssDNA/swCNTs (Figure 5.12a). This change

is significant as modification with DBCO-amine alone has no observable effect on fluorescence (Figure 5.12b), and there is only a ~20% reduction when mixing sfGFP^{SA} with ssDNA/swCNTs in the absence of linking molecules (Figure 5.12a). This indicates that the employed coupling strategy gives rise to an enhanced quenching beyond that caused by passive and transient physisorption of proteins and CNTs, or due to the chemical functionalization of the proteins at the bioengineered anchoring point.
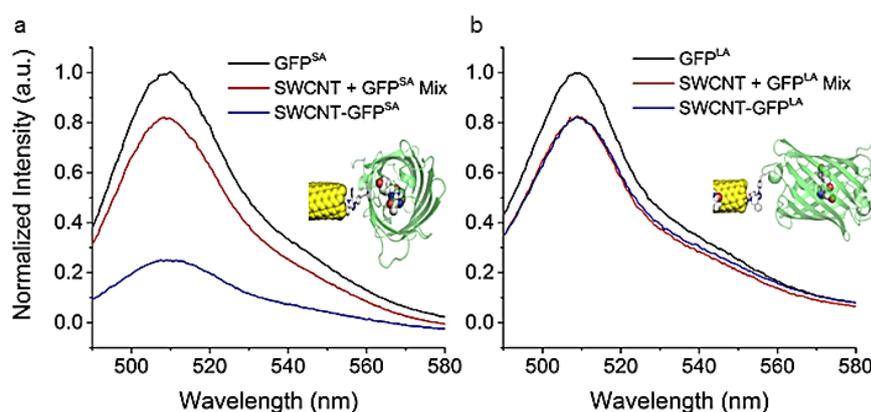
Figure 5.12 Steady state fluorescence spectra of a) sfGFP^{SA} and b) sfGFP^{LA}. Emission spectra of free protein (black), ssDNA/swCNT-sfGFP hybrids (blue) and ssDNA/swCNTs mixed with free protein in the absence of DBCO-amine linker (red). representative models of swCNT-sfGFP nanohybrids are displayed (inset).

In comparison, fluorescence of sfGFP^{LA} in the long axis monofunctionalized hybrids was only quenched by ∼20%. This is the same reduction as observed by mixing sfGFP^{LA} and ssDNA/swCNTs (Figure 5.12b), however it is ~10% more than addition of DBCO-amine to sfGFP^{LA} alone (Figure 5.3a). These results strongly suggest that direct communication between the proteins and the CNTs only occurs for the SWCNT-GFP^{SA} nanohybrids: the precise nature of the interface position plays an important role in the coupling due to the predicted shorter distance between the CRO and the nanotube in the sfGFP^{SA} hybrids as compared to the sfGFP^{LA} hybrids.

5.2.6 Single-Molecule fluorescence of ssDNA/swCNT-sfGFP nanohybrids.

Nanotube-sfGFP coupling was monitored, to single molecule resolution, *via* total internal reflection fluorescence (TIRF) microscopy (performed by Dr Mark Freeley), by casting dilute solutions of sfGFP-swCNT nanohybrids onto glass coverslips. sfGFP has characteristic blinking behaviour which can be influenced by the local environment.[101,216,264] Representative intensity versus time single molecule traces showing 'on' and 'off' for sfGFP^{SA} and sfGFP^{LA} both in free form and conjugated to swCNTs can be shown in Figure 5.13. Single-molecule fluorescence dynamic studies of sfGFP^{WT} have shown that off-time periods are independent of excitation intensity.[264] Individual sfGFP^{SA} showed high blinking activity with shorter 'off' states, *i.e.* faster switching between "on" and "off" states, while the swCNT-GFP^{SA} nanohybrids exhibited longer off time periods (Figure 5.13a and b). This is good evidence that there is functional coupling between the swCNT and sfGFP^{SA}, at the single molecule level. The swCNT-sfGFP^{LA} heterostructures however, exhibited almost identical blinking rates when compared to the sfGFP^{LA} alone (Figure 5.13c+d).
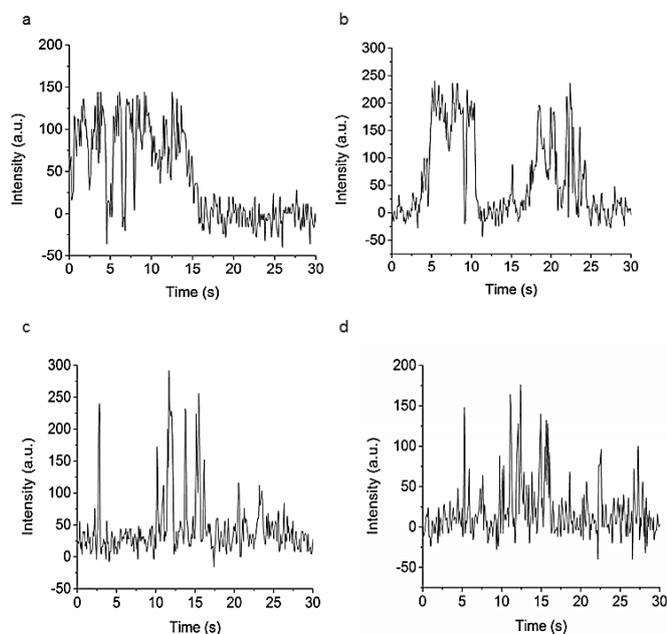
Figure 5.13 Representative single-molecule fluorescence traces for: a) sfGFP^{SA} b) swCNT-sfGFP^{SA} hybrid c) sfGFP^{LA} and d) swCNT- sfGFP^{LA} hybrid. Adapted from [119]

This behaviour was confirmed and quantified by constructing histograms of single-molecule off-times for the hybrid structures and both sfGFP variants (Figure 5.14). The histograms were fitted with monoexponential decay curves as described in previous single-molecule sfGFP studies;[216,264] from these, the decay time for each sample was calculated. sfGFPSA alone exhibited a decay time of 9.2 s, while the SWCNT-GFP[SA] nanohybrids had a decay time of 7.7 s (Figure 5.14a). The shortening of the decay time corresponds to an increase in the length of the off times, corroborating the longer off times seen in the intensity versus time plots displayed in Figure 5.13. sfGFP[LA] has a shorter decay time (6.6s) compared to sfGFP[SA], however differences in decay times between different protein variants is to be expected, as has been observed previously.[216,265] This variance is due to mutations altering the chromophore environment (*i.e.* H-bonding networks), and the ability of the chromophore to return to a ground state to be reexcited.[61] As the hydrogen bonding network around CRO extends across the whole protein (Chapter 6) even mutations far from the chromophore such as with the E132azF mutation in sfGFP[LA], causes alterations in long range interactions and hence subtly affects the lifetime of sfGFP fluorescence. In contrast to sfGFP[SA], the decay time of sfGFP[LA] single-molecule off-events is unchanged in the swCNT-sfGFP[LA] hybrids (Figure 5.14b).
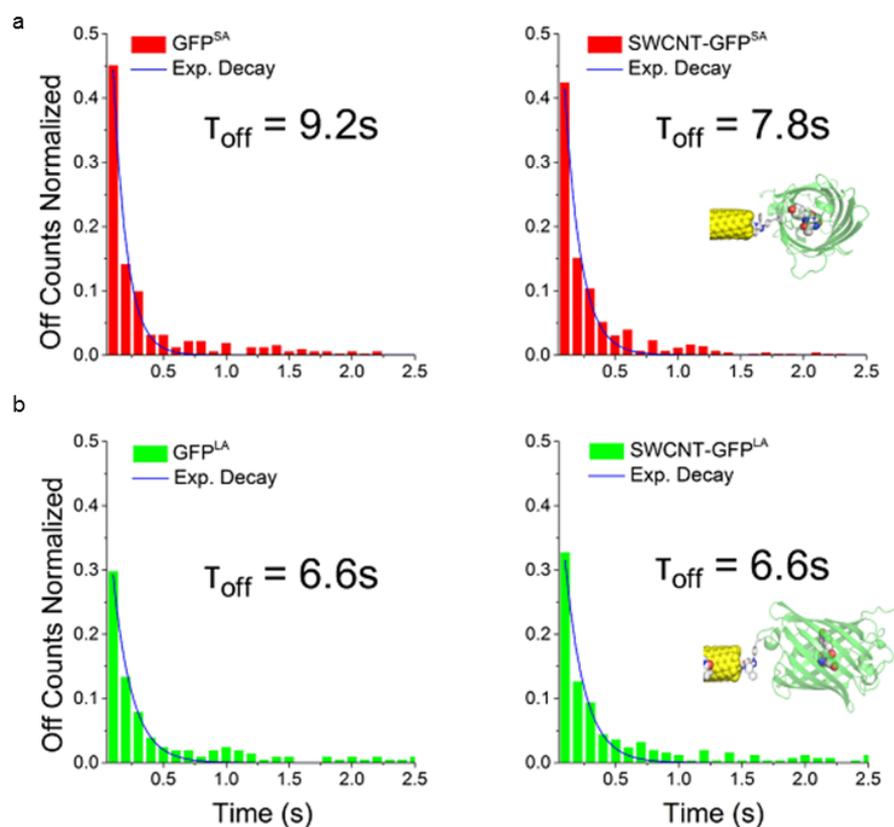
Figure 5.14 Histograms and corresponding exponential fits of single molecule off times of a) sfGFP[SA] and swCNT- sfGFP[SA] hybrids, and b) sfGFP[LA] and swCNT-sfGFP[LA] hybrids. Adapted from [119]

Overall, only the short axis (swCNT-GFP[SA]) nanohybrids showed any single-molecule alterations to the optical properties of sfGFP. It is believed that the close spatial proximity of the sfGFP[SA] chromophore in relation to the end of the swCNT is the driver of the changes in lifetime; this demonstrates the importance of appropriately designing linkage positions for functional communication. The observed electronic coupling is likely to occur through energy or charge transfer. Previous work has shown that energy transfer between GFP and a local acceptor can result in quenching in a distance dependent manner.[134,259] Additionally, photoinduced charge transfer between GFP, acting as the donor,[135,262,263,266] and the CNT as the acceptor[253,255,267-272] is a particularly interesting alternative to classical energy transfer given the differences between sfGFP[SA] and sfGFP[LA] in terms of the attachment position. Given that graphene and swCNTs have been shown can have an x^{-4} distance dependence on energy transfer

efficiency,[273] and that the chromophore is twice as far away from the swCNT in sfGFP[LA] as it is in sfGFP[SA] (3.1 Å versus 1.6 Å respectively, Figure 5.1). There is a 16 fold decrease in predicted energy transfer efficiency and hence why attachment of sfGFP[LA] to swCNT end walls seemingly has no effect. This also helps to explain why both sfGFP variants showed altered fluorescence properties upon direct attachment to swCNTs (Figure 5.8), as without the DBCO-amine linker the chromophore is brought into closer contact with the π electron system.

To further explain why there is apparent communication between sfGFP[SA] and the swCNT, a model was created of sfGFP[SA] +DBCO-amine, (as described in Chapter2) and then attached to a swCNT *in silico*. Analysis of the model swCNT-sfGFP hybrid with CAVER[181] (Figure 5.15a), showed a proton exit/entry point close (1.28 Å) to the attachment site in sfGFP[SA] (residue 204) linked to a charge transfer network back to CRO (Figure 5.15b).[217,218] Changes to these residues are known to affect the charge transfer network resulting in changes in the charged form of the CRO ground state.[9,274] This is likely how communication between swCNT and sfGFP would occur, with the azF + DBCO-amine triazole linkage providing a bridge between the two existing charge transfer systems. Thus, the interface position with respect to the protein influences the degree of this coupling, where only the correctly bioengineered system exhibits the expected direct protein − nanotube communication.
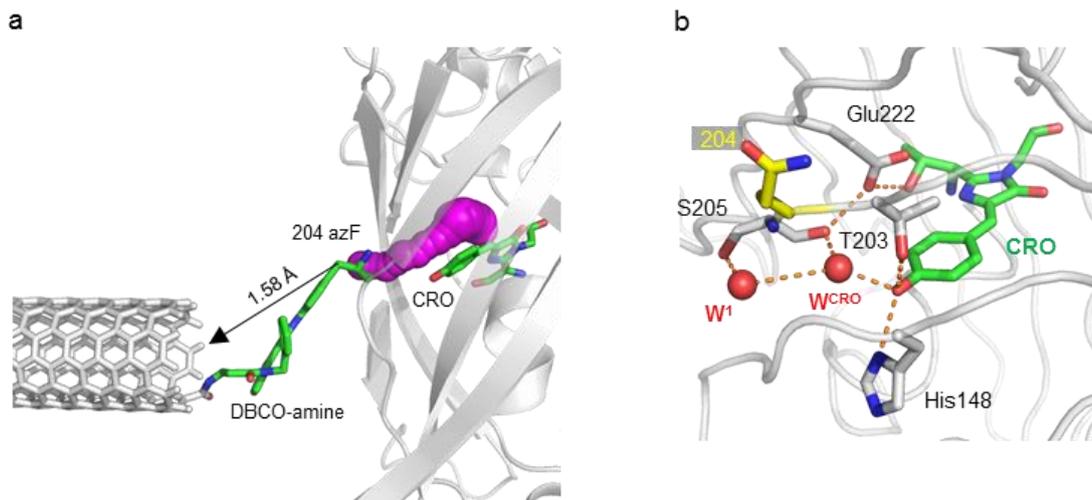
Figure 5.15 Interaction network of residue 204 a) Predicted caver channel (magenta spheres) from theoretical model of swCNT-sfGFP^{SA} (grey). CRO and 204azF + DBCO-amine linkage (green sticks) highlighted. b) The local interaction network of Gln204 (yellow sticks) with respect to CRO (green sticks), other residues (grey sticks, annotated) and water molecules (red spheres and annotated W^X.

5.2.7 Generation of a sfGFP – CNT protein junction

The swCNT-sfGFP assemblies created above, hold great potential for the development of solution processable single-molecule optobioelectronic systems and devices including light gated (biotransistors) and sfGFP-based stimulated emission (BioLEDS).[275-277] Biomolecular function (e.g., sfGFP electronic excitation and light induced charge transfer [134,160,275-277]) can be used to modulate conductance, and proteins have been observed to demonstrate transistor-like behaviour at the single molecule level.[45,50,51] Ideally, any protein component will be fully integrated into circuit system much in the same manner as a solid-state device. Attaching proteins to the side walls of materials such as CNTs means proteins are on the periphery and not directly integrated. In other words, a protein should be part of the circuit and act a junction between two conducting elements i.e. the SWCNTs (Figure 5.16a). To facilitate this, a sfGFP variant was engineered with two CNT-anchoring azide handles on opposite faces of the protein along the short axis (sfGFP^{SAx2}). The two azF groups on a single molecule of sfGFP^{SAx2}

142

allows the protein to bridge two swCNT endwalls, which then form the basis of a nanoelectronic circuit component.
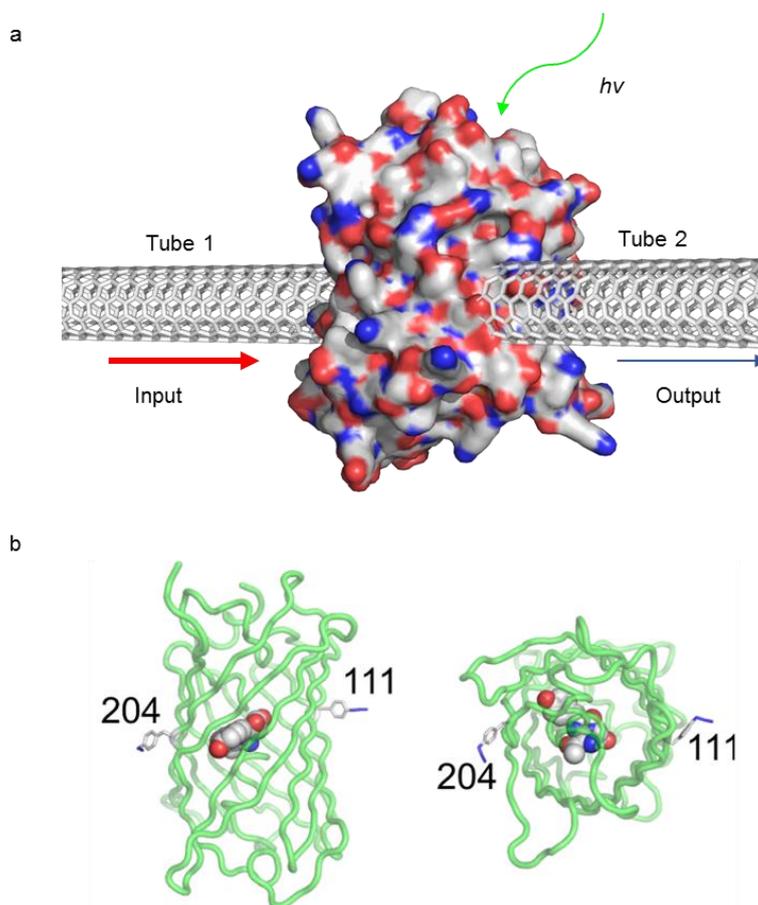
Figure 5.16 Proposed model for protein based CNT bridges. a) Model of sfGFP^{SAx2} between two swCNTs. Demonstrating change in input current as light (*hv*) is absorbed by protein molecule. b) *In silico* model (Side-on and top-down views) of sfGFP^{SAx2} with residues 111 and 204 mutated to azF.

Residue Glu111 was selected as the second residue for generating the molecular bridge. Glu111 was selected as it is on the opposite face of the sfGFP β-barrel to residue 204 (Figure 5.16b). The E111TAG mutation was introduced into the existing pBAD-sfGFP^{SA} using site directed ligase independent mutagenesis (SLIM) to create sfGFP^{SAx2}.[278] The plasmid for sfGFP^{SAx2} was co transformed into *E. coli* cells with the pDULE azF incorporation plasmid. Transformed cells expressing were visibly green in

presence of azF but not in its absence suggesting that azF was successfully incorporated into sfGFP^{SAx2} (Figure 5.17a). SDS-PAGE analysis shows an over-expressed band at *circa* 27.5 kDa in the presence of azF, which isn't present when azF is not added (Figure 5.17b). Cells expressing sfGFP^{SAx2} were harvested, lysed, and purified as described previously (Section 2.4.3) by nickel affinity chromatography. SDS-PAGE analysis showed the protein had greater than 95% purity (Figure 5.18), and so was buffer exchanged into fresh 50 mM Tris-HCl [pH 8.0] and the concentration determined.
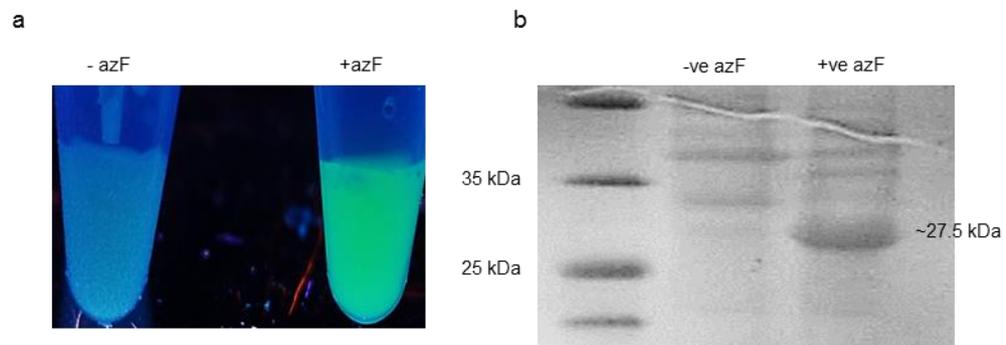
Figure 5.17 Comparison of expression of sfGFP^{SA}x2 in the presence (+azF) and absence (-azF) of azF. a) Resuspended cell pellets irradiated by UV light shows visible fluorescence in the presence of azF but no fluorescence in the absence. b) SDS-PAGE gel of cell pellets showing increased expression of a protein at about 27 kDa in the +azF sample.
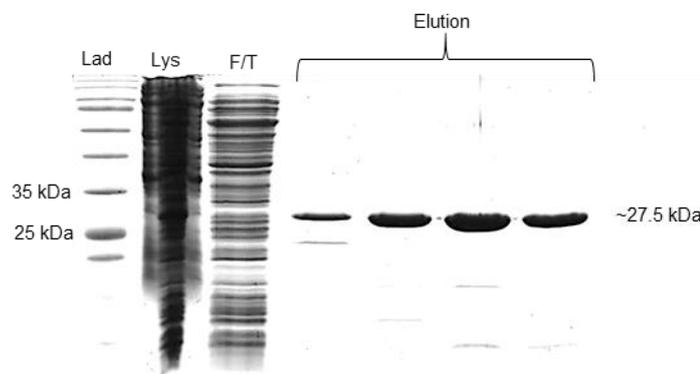
Figure 5.18 SDS-PAGE showing purification of sfGFP^{SAx2}. Lanes as follows: Ladder (Lad), Cell lysate (Lys), column flow through (F/T) and elution fractions. Elution shows single protein band at ~27.5 kDa.

To ensure that modification of sfGFP^{SAx2} would not affect protein functionality, sfGFP^{SAx2} (10 μM) was incubated for 1 hour with DBCO-amine as described above. However, DBCO-amine was added in a 10x molar excess (100μM) instead of 5x as with sfGFP^{SA} and sfGFP^{LA}, because of the presence of an extra phenylazide moiety in sfGFP^{SAx2}. Absorbance and emission spectra were recorded both before and after DBCO-amine addition (Figure 5.19). Absorbance spectra of purified sfGFP^{SAx2} showed no change in $λ_{max}$ (485 nm) but there was a slight reduction in molar extinction coefficient from 51 mM^{-1}.cm^{-1} (sfGFP^{SA}) to ~41 mM^{-1}.cm^{-1}. Modification of sfGFP^{SAx2} with DBCO-amine resulted in a slight reduction in molar excitation coefficient ~40 mM^{-1}.cm^{-1} and emission (excitation 485 nm) ~10%. As the reduction is only slight, it can be assumed that modification with DBCO-amine does not have a significant effect on protein function. Therefore, sfGFP^{SAx2} is a suitable candidate for creating a swCNT-protein-swCNT bridge.
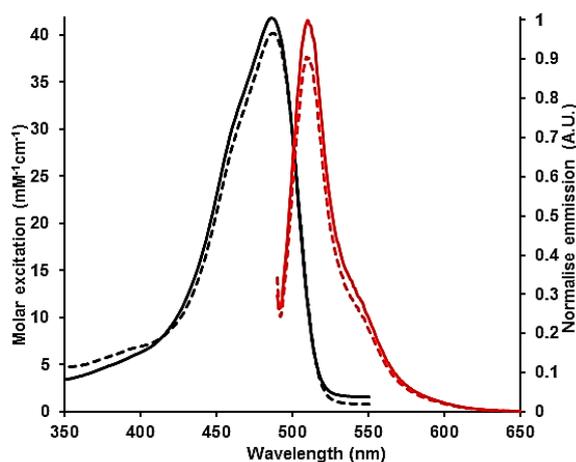
Figure 5.19 Absorbance (Black) and emission spectra of sfGFP^{SAx2} (Red, Ex=485nm), before and after 1 hour incubation with DBCO-amine (solid and dashed lines respectively). Spectra were recorded at 10 μM and 0.5 μM for absorbance and emission spectra respectively. Fluorescence emission normalised relative to unmodified sfGFP^{SAx2}.

Purified sfGFP^{SAx2} was conjugated with swCNTs as above and deposited onto silica by Dr Mark Freeley as previously described. AFM showed that sfGFP^{SAx2} was able to bridge two swCNTs together as shown in Figure 5.20. Height profile analysis from

AFM images showed a clear difference in height between protein junctions (1.5 nm) and miscellaneous adsorbates along the swCNTs (0.7nm), confirming that the aim to generate protein based junctions between sfGFP^{SAx2} and swCNTs. Further investigations will focus on optimizing the yield of formation of such systems and investigating their electrical properties in device configurations. In future, it would also be useful to explore other orientations of sfGFP including the long axial variant (sfGFP^{LAx2}). Investigating non-linear protein bridging (e.g. right angle variants) and multiway junctions (3-4 nanotubes connected to a single protein).
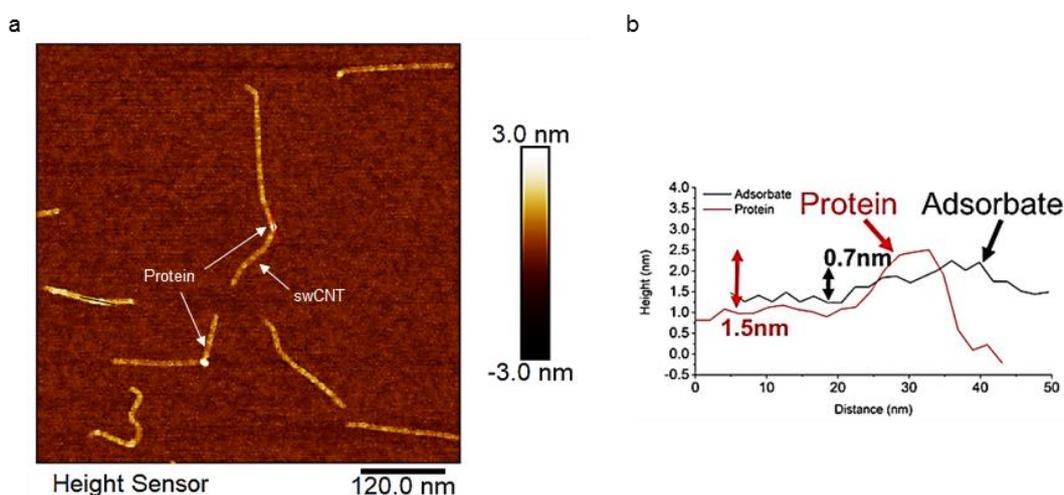
Figure 5.20 Analysis of swCNT-sfGFP^{SAx2} hybrids a) Representative AFM image of sfGFP^{SAx2}-swCNT junction labelled. b) height profiles along tube axis demonstrating the difference between random adsorbates on CNT walls (black) and a protein junction (red).

5.2.8 Comparisons of both approaches

Above is described two routes that the ncAA azF can be used to interface sfGFP with swCNTs. Firstly, by using classical phenylazide photochemistry, to directly link proteins to swCNT sidewalls, and secondly, using click chemistry to attach proteins to ssDNA wrapped swCNT endwalls that have been functionalised with DBCO groups. Using a variety of techniques, we have been able to show specific, functional linkage that shows evidence of communication using either approach. The communication has been shown to be orientation specific *i.e.* when the CRO of sfGFP is orientated closer

to the swCNT (sfGFP[SA]) there is a greater effect than when CRO is orientated further away (sfGFP[LA]).

Both approaches, have advantages and disadvantages for example: the direct linkage approach allows for a very direct and intimate linkage allowing for efficient communication between attached proteins and swCNT. However, it is hard to control the number and location of proteins attached to swCNT endwalls.[119] By using the ssDNA wrapped swCNTs this issue is avoided allowing for precise individual attachment of proteins to swCNT endwalls allowing for proteins to be directly integrated into a circuit containing swCNTs. This is useful for when a protein is needed to be directly integrated into a system *i.e.* for creating a protein transistor.[160] However, if a larger number of proteins is required to make for example a biosensor[57] or light harvesting system,[159] then it is an advantage to use the direct linkage to side walls allowing for multiple proteins to bind the swCNTs. Another advantage the ssDNA/swCNTS has over regular swCNTs is that proteins can be attached in solution and hence proteins aren't concentrated onto tubes by drying effects. However, work in this lab is begining to explore the possibility of in solution UV attachment of proteins onto regular swCNTs.

5.3 Conclusions

In summary, functional sfGFP-swCNT nanohybrids were created, using two orientations of sfGFP to create functionally linked nanohybrids using both phenylazide photochemistry and click chemistry. Attachment of sfGFP[SA] and sfGFP[LA], to the side and end walls of swCNTs was explored and showed that it is possible to functionally link proteins peripherally (side wall attachment) and directly (end wall attachment), with respect to a swCNT based nanodevice. Both approaches have the potential to be used as nanoelectronic components. However, there is still scope for refinement with both techniques.

Using phenylazide photochemistry we demonstrated that the intimate linkage at defined residues to the sidewalls of swCNTs, using genetically incorporated phenylazides, is both robust and leads to functionally linked communication between bound proteins and the swCNT leading to extended fluorescence lifetimes. The degree of influence on the fluorescence is, in part dependent on the attachment position of the protein to the swCNT, *i.e.* there is a more noticeable change when sfGFP[SA] is attached to the tubes (CRO close to swCNT) than when sfGFP[LA] is linked to the tube (CRO further away). Therefore, defining protein attachment allows for both consistent fabrication of nanohybrids, and orientation based functional linkage. As such this method has shown its potential to be transferred to other protein systems.[85]

Protein-swCNT nanohybrids were also created by coupling single proteins to individual swCNTs via an in solution methodology. A monofunctionalization yield of more than 80% was obtained in the formation of ssDNA/swCNT-sfGFP heterostructures with controlled protein orientation. Fluorescence investigations showed evidence of site-specific functional communication between the proteins and the nanotubes. Additionally, single molecule fluorescence studies confirmed the importance of bioengineering optimal protein attachment sites: evidence of coupling was observed when sfGFP was attached *via* the 204 residue, close to the chromophore, while in the case of sfGFP attached via the 132 residue, which is at a larger distance from the chromophore, no protein communication with the nanotube was detected. Finally, the incorporation of two bioorthogonal reaction handles into individual sfGFPs allowed the construction of single protein bridging systems. As ncAA incorporation with bioorthogonal reaction handles becomes more accessible,[122,183] including azF,[85] this approach provides a general route for protein attachment to the ends of carbon nanotubes. The approach is by and large, a powerful approach to generate tailored and optimal single protein−CNT hybrids that hold great potential for the development of solution processable single-molecule bioelectronic systems and devices based on the use of carbon nanoelectrodes.

6 Cooperativity between multiple ncAAs in the same protein

6.1 Introduction

The interlinked nature of the interactions that comprise protein 3D structure means that cooperativity is prevalent on varying degrees of scale. Cooperativity can take the form of stabilising protein secondary and tertiary structure though extended hydrogen bond donor/acceptor networks or hydrophobic interactions allowing the protein to fold robustly to withstand changes in environmental conditions without undergoing gross unfolding events. Cooperativity can also be critical for function as outlined by catalytic triads and O_2 transport by haemoglobin. Even residues apparently far away from each other in the molecular structure can be linked through bond networks so be structurally and functionally linked. Haemoglobin is an excellent example and so are the artificial fluorescent proteins described in Chapter 4. Despite the obvious importance of cooperativity, little research has investigated the effects multiple ncAAs incorporated into a structure can impact on function. In this chapter I will address this by looking at how incorporating 2 or 3 azF residues can alter the light sensitivity of GFP. It is important to iterate here that cooperativity will be effects observed on incorporation of multiple azF that are not present when only single azF residues are present at the selected positions.

As described previously (Chapter 3) incorporating non-canonical amino acids (ncAAs) such as p-azido-L-phenylalanine into proteins has numerous benefits including introducing new chemistries into proteins, instil new functions and directed post translational modifications. There are many studies investigating the incorporation of ncAAs,[78,85,86,94,95,107,115,119,155,279] which introduce a wide variety of ncAAs for many different uses such as infrared probes,[280] aiding in solving crystal structures,[83] introducing photocontrol with photocaged and light reactive residues,[85,110,115,149,152] and introducing defined post translational modifications.[78,115,119,186] Most of this research has focused on incorporating single ncAA into proteins.[84,92,94,96,155,194] Although there has

been research on incorporating a ncAA at multiple positions in a protein and indeed work where, two distinct non-natural amino acids have been incorporated (e.g. for the purpose of attaching two different fluorescent probes to monitor protein dynamics[225]), very little work has been done to investigate the potential of incorporating multiple amino acids at strategic regions within a protein to function synergistically.
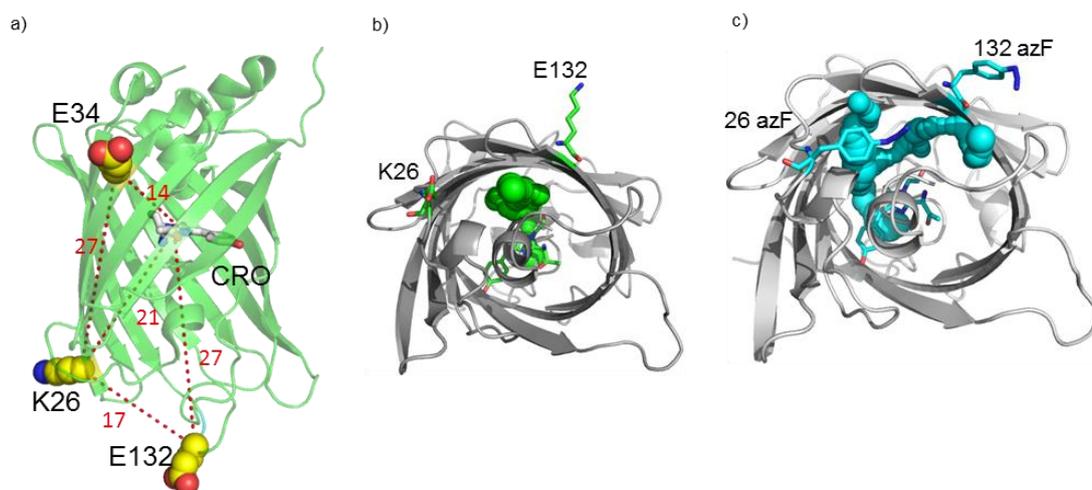
Figure 6.1 a) Structure of sfGFP highlighting residues K26, E34 and E132 and the distances between each other and the chromophore. b) Bottom up view of sfGFP[WT] showing a predicted CAVER tunnel in cyan. c) same view as b) of a sfGFP [26/132 azF] model showing that previous tunnel has split into two similar tunnels.

This chapter will focus on incorporation of 2 or 3 azF residues into a single sfGFP molecule to determine what, if any, effect there will be on function, especially sensitivity to irradiation with UV light (see Chapter 1). The three residues chosen are K26, E34 and E132. Previous work indicates that when azF is incorporated individually at each of these positions, there is no significant shift in fluorescence either before or after UV irradiation.[152] Residues Lys26 and Glu132 are relatively close to each other (17 Å, α carbon to α carbon, Figure 6.1a), but do not make any apparent direct interactions and are distant from the chromophore (21 Å and 27 Å for K26 and E132 respectively), so should not affect the function of the chromophore. However, using CAVER 3.0 we find a potential channel from the chromophore that exits between these two residues (Figure

6.1b), which may influence the charge transfer process.[139,185,218,262,265] When azF is modelled into residues 26 and 132 we see that this predicted channel splits into two channels highlighting the potential for these positions to generate a profound effect on the function of the protein (Figure 6.1c). As described in Chapter 5, interfacing sfGFP at the 132 position with swCNTs leads to major changes in sfGFP fluorescence (increased resistance to photobleaching with lower intensities), suggesting communication between bound sfGFP molecules and the swCNT. Thus, residue 132 and its neighbouring residue 26 may have more significant effects on GFP structure-function relationship than might be anticipated.

6.2 Results

6.2.1 Construction of double and triple mutants

Plasmids for sfGFP^{26TAG}, sfGFP^{34TAG}, sfGFP^{132TAG} and sfGFP$^{34/132TAG}$ were already available from previous work by Reddington[85,152] and Hartley (Unpublished data). However, the remaining 2 double mutants and the triple mutant had to be created. This was undertaken using whole plasmid site directed mutagenesis as described in Section 2.2. Subsequent agarose gel analysis of the PCR products showed single bands at *circa* 4.5 kb for all three PCR reactions indicating successful amplification of the plasmid (Figure *6.2*). There was no amplification of the negative controls (PCR reactions with no templates) suggesting that amplification was of the desired PCR product. PCR products were purified and recircularised as described in Section 2.2.3), then transformed into *E. coli* NEB 5α cells for DNA propagation and sequencing as described in Section 2.2.6. Sequencing confirmed successful mutation of all three plasmids.
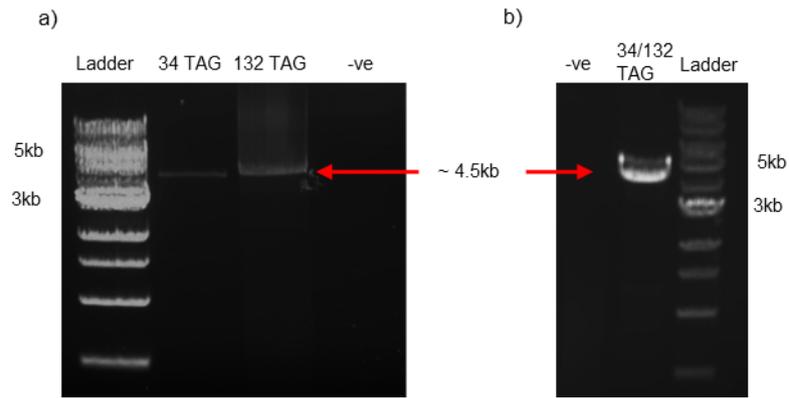
Figure 6.2 Agarose gels showing amplification of a) sfGFP ^{34TAG} and sfGFP ^{132TAG} and b) sfGFP $^{34/132}$ plasmids with sfGFP ^{26TAG} mutation primers. There is no amplification in either negative control.

6.2.2 *In situ* fluorescence changes of cell lysates from sfGFP variant expressing *E. coli*.

Plasmids encoding the production of sfGFP$^{26/34azF}$, sfGFP$^{26/132azF}$, sfGFP$^{34/132azF}$ and sfGFP$^{26/34/132azF}$ were cotransformed into *E. coli* cells with pDULE^{cyanoRS} (Section 2.2.6 and [94]). To determine which if any of the sfGFP double/triple mutants affect the fluorescence after UV irradiation, cell cultures expressing each variant were harvested, resuspended in 50 mM Tris-HCl [pH 8.0] to an OD$_{600}$ of 0.1 and lysed (Section 2.5). Fluorescence emission spectra (on excitation at 485 nm) were measured before and after 1 hour UV irradiation. Of the double mutants sfGFP$^{26/34azF}$ sfGFP$^{34/132azF}$ showed no discernable change in emission after UV treatment and so were discarded from future experiments (Figure 6.3a and b).
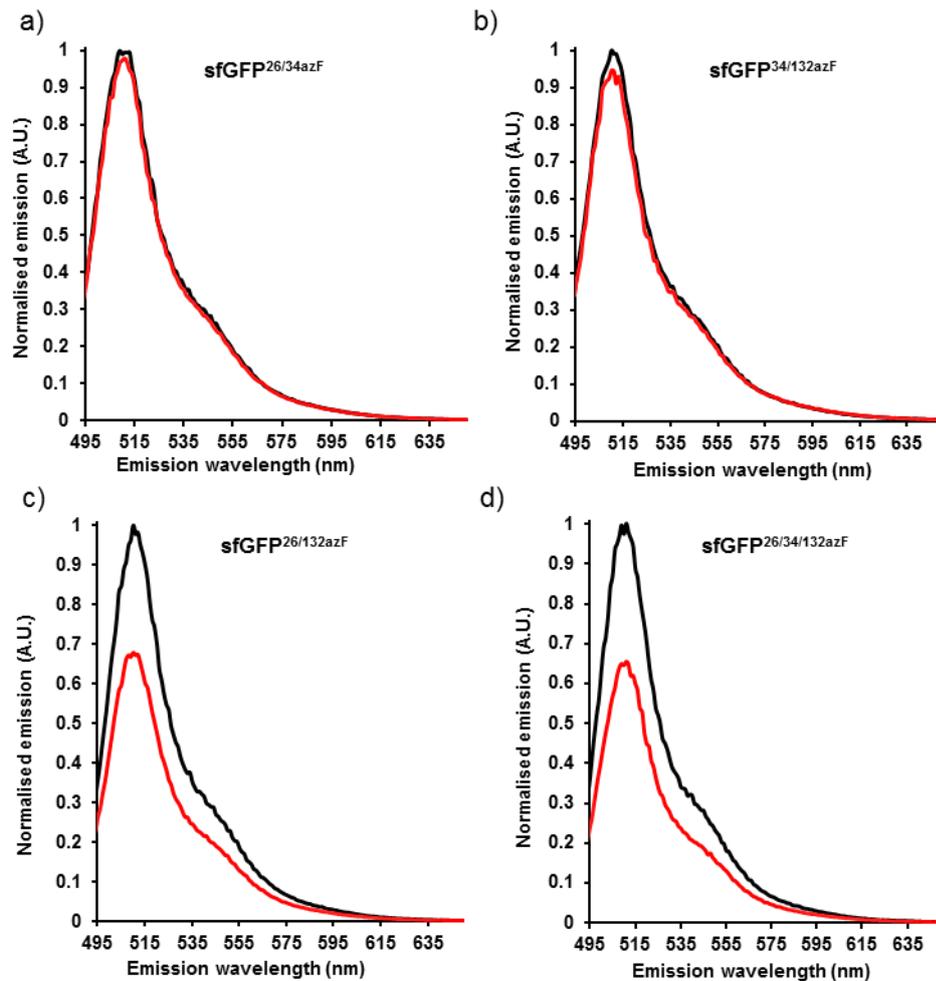
Figure 6.3 Emission spectra of sfGFP variants straight after cell lysis. Cell lysate emission spectra for a) sfGFP[26/34azF]; b) sfGFP[34/132azF]; c) sfGFP[26/132azF] and d) sfGFP[26/34/132azF] before (Black) and after (red) UV irradiation. There is no discernable difference in emission of the cell lysates after UV exposure in the 26/34 and 34/132 samples. There is however a decrease in emission after UV exposure in both the 26/132 and triple azF samples. Cell lysates were excited at 485 nm.

The emission of sfGFP[26/132azF] decreased by about 35% (Figure 6.3c). As single azF incorporation at either residue 26 or 132 has little effect on fluorescence intensity on exposure to UV light (Figure 6.4), suggesting that there is an interaction, either directly or indirectly, between residues 26 and 132 that caused the reduction in fluorescence beyond a simple additive effect of having multiple ncAAs in a protein. The triple azF containing variant, sfGFP[26/34/132azF] shows no further decrease in fluorescence (Figure 6.3d), when an additional azide is incorporated at residue 34. This suggests that that the loss of fluorescence is due to cooperation between azF26 and azF132. To investigate the potential basis for the observed effects on fluorescence, sfGFP[26/132azF]

and sfGFP$^{26/34/132azF}$ were produced on a larger scale and purified. From here on in sfGFP$^{26/132azF}$ and sfGFP$^{26/34/132azF}$ will be referred to as sfGFP^{azFx2} and sfGFP^{azFx3} respectively.
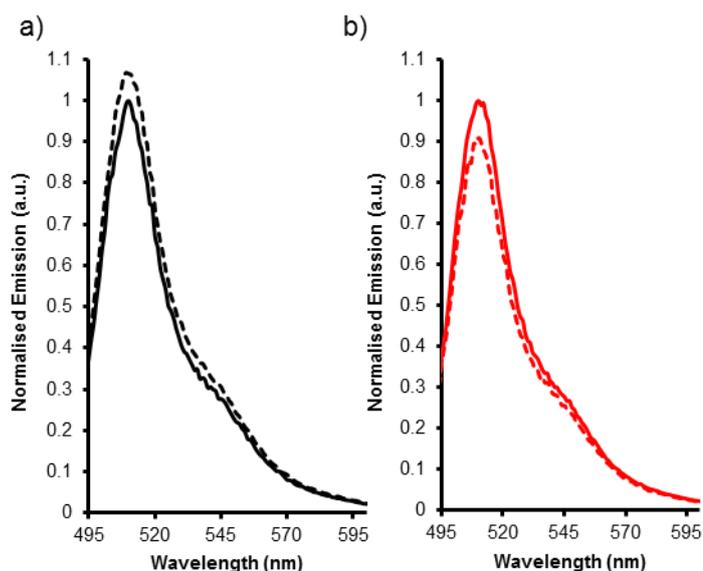
Figure 6.4 Fluorescence of cell lysates before (solid lines) and after (dashed lines) UV irradiation. a) sfGFP^{26azF} and b) sfGFP1^{32azF}

6.2.3 Production and purification of sfGFP^{azFx2} and sfGFP^{azFx3}

To ensure azF incorporation at TAG stop codons as opposed to non-specific incorporation of tyrosine, sfGFP^{azFx2} and sfGFP^{azFx3} were expressed in presence and absence of azF (Section 2.4.1). As shown in Figure 6.5a, we see the presence of a ~27 kDa protein at high yield only with the presence of azF. There is also a similar expression between both the double and triple mutants. This means that there is no significant drop in expression by incorporating a third azF residue. Both cultures with azF added were visibly green and fluorescent (Figure 6.5b).
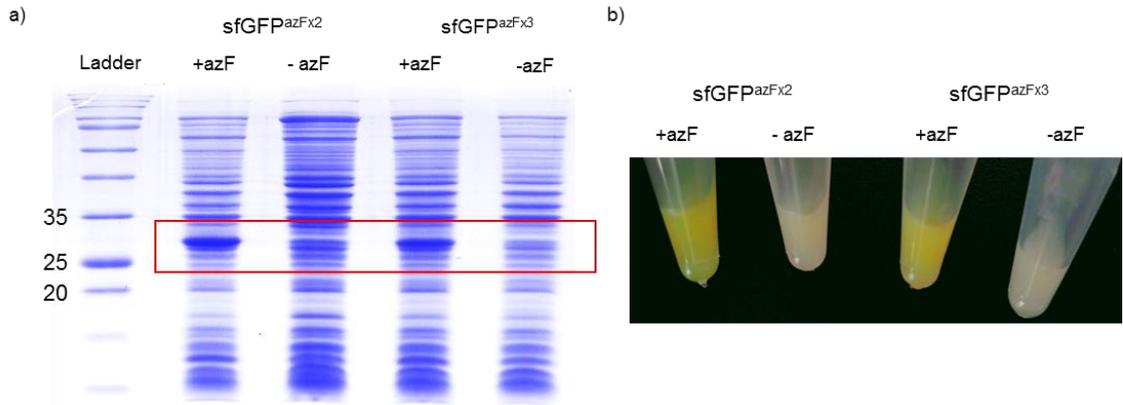
Figure 6.5 a) Whole-cell SDS-PAGE gel showing expression of sfGFP x^2 and sfGFP x^3 in presence (+azF) and absence (-azF) of azF. b) Whole cells expressing sfGFP x^2 or sfGFP x^3 resuspended in 50 mM Tris-HCl [pH 8.0] in presence (+azF) and absence (-azF) of azF showing functional sfGFP in azF positive samples.

Cells from both sfGFP^{azFx2} and sfGFP^{azFx3} were harvested and lysed as previously described (Section 2.4.1). Cell lysates were then passed through a HiTrap TALON® column (5 mL) under the same conditions as in Section 2.4.3. Elution of sfGFP^{azFx3}, yielded a single band at ~27 kDa in the elution. However, sfGFP^{azFx2} showed evidence of a secondary band ~56kDa as well as the 27 kDa band, suggesting the possibility of a sfGFP dimer or a contamination (Figure 6.6). The pure fractions of sfGFP^{azFx2} and sfGFP^{azFx3} with only a 27 kDa band were pooled, and buffer exchanged into fresh 50mM Tris-HCl [pH 8.0] for spectrophotometry (Section 2.6.3).
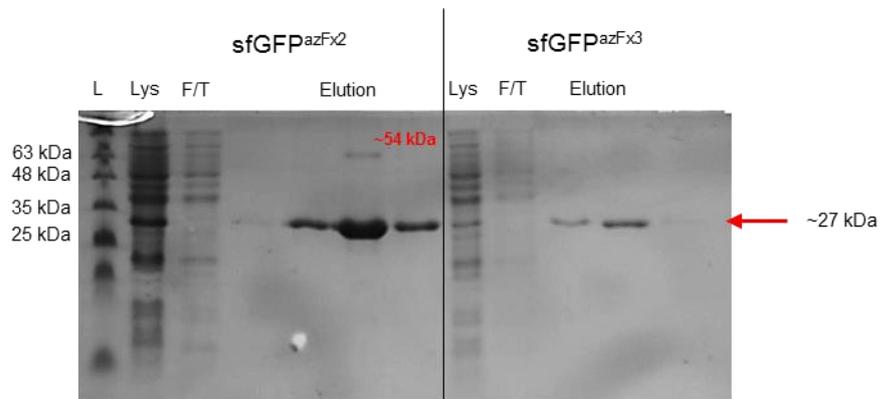
Figure 6.6 SDS Page analysis after cobalt affinity chromatography of sfGFP^{azFx2} and sfGFP^{azFx3} showing in both cases a band ~27 kDa in the lysate (Lys) and elution fractions but not in the flow through (F/T). A second band of ~54 kDa is shown in the elution of sfGFP^{azFx2}.

6.2.4 UV irradiation induced spectral changes of sfGFP^{azFx2} and sfGFP^{azFx3}

To confirm the emission data shown by the cell lysates and determine if the fluorescence loss is due to azF residue, the absorbance and emission spectra were taken at 10 µM and 0.5 µM respectively before and after UV irradiation. The absorbance spectra of both sfGFP^{azFx2} and sfGFP^{azFx3} dark forms shows a dominant absorbance at 485 nm with molar excitation coefficients of *circa* 51 mM^{-1}cm^{-1} and 54 mM^{-1}cm^{-1} for sfGFP^{azFx2} and sfGFP^{azFx3} (respectively, Figure 6.7, Table 6.1). Upon UV irradiation absorbance reduces for both variants to ~41 mM^{-1}cm^{-1}.
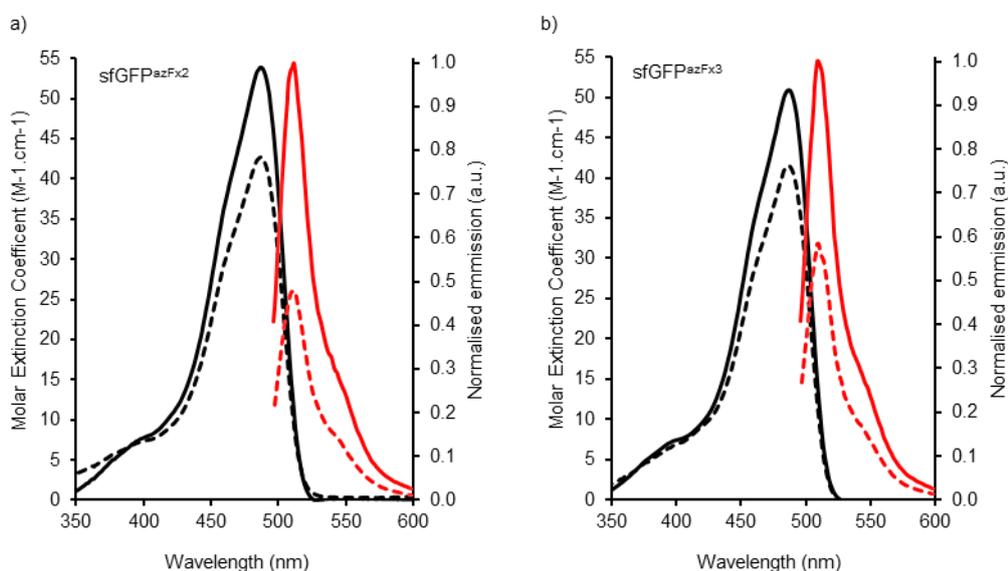
Figure 6.7 Absorbance (black) and emission (red) spectra of poly azF variants before and after 1 hour of UV irradiation (solid and dashed lines respectively of a) sfGFP^{azFx2} and b) sfGFP^{azFx3} (red)). Absorbance spectra normalised to millimolar excitation coefficient. Emission spectra (excitation at 485 nm) are normalised to the dark sample.

The emission spectra (485 nm excitation) of sfGFP^{azFx2} (Figure 6.7a) and sfGFP^{azFx3} (Figure 6.7b) show a solitary emission peak at 511nm for both dark and irradiated samples. After 1 hour of UV irradiation of sfGFP^{azFx2} there is a 50% reduction in fluorescence intensity. The fluorescence intensity of sfGFP^{azFx3} decreased *circa* 40% after UV treatment, suggesting that the additional azF residue at residue 34 slightly reduces the impact of irradiation on fluorescence. The emission spectra of the purified

protein confirms the original observation in cell lysates and helps to confirm that there is some level of cooperativity between 26 and 132.

Table 6.1 Spectral characteristics of sfGFP variants

		λ_{max} nm	λ_{em} nm	ε mM^{-1}cm^{-1}
sfGFP^{WT}		485	511	49
sfGFP^{azFx2}	Dark	485	511	51
	UV	485	511	41
sfGFP^{azFx3}	Dark	485	511	54
	UV	485	511	41

6.2.5 SDS-PAGE analysis shows multimerisation of sfGFP^{azFx2}

Both the cell lysate and *in vitro* purified protein data confirmed that the combination of azF at residues 26 and 132 causes significant loss of fluorescence after UV irradiation. Given that the residues are far from the chromophore, it was not clear how photochemical conversion of the two azF residues elicited their functional effect. One possibility is photo-crosslinking. Classically, azF has been used as a photo-crosslinker,[12,85,112,214] so it is possible that the residues have crosslinked to another part of the same protein as observed by Reddington *et al,*[152] where azF incorporation at residue 145 of sfGFP, creates a fluorescent protein that 'switches off' after exposure to UV irradiation through an internal crosslinking mechanism: azF145 crosslinks to the chromophore. Although direct crosslinking to the chromophore is unlikely due to the distance (> 20 Å) between CRO and residues 26 and 132, crosslinking around the end of the sfGFP β-barrel structure could lead to a disruption of the hydrogen bonding network around the chromophore. The potential for sfGFP^{azFx2} to crosslink was analysed by SDS-PAGE. Any crosslinking would show as a band shift on SDS-PAGE. Samples of purified sfGFP^{azFx2} at varying concentrations, from 10µM to 150µM, were irradiated

with UV light for 1 hour. After Irradiation, samples were diluted to 10µM and loaded onto an SDS-PAGE gel with a 10 µM sample of unirradiated sfGFP^{azFx2}. This was repeated with sfGFP^{26azF} and sfGFP^{132azF} as controls.
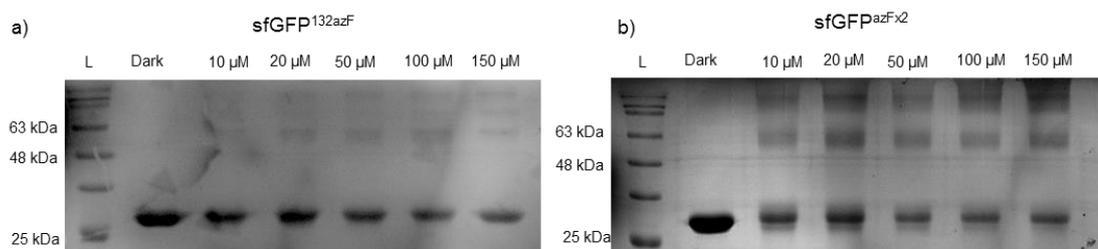
Figure 6.8 SDS-PAGE analysis of, a) sfGFP^{132azF}, and b) sfGFP^{azFx2}, in the dark state (10 µM), and after irradiation at 10, 20 50, 100 and 150 µM. The gels show bands at ~27 kDa throughout and in the irradiated samples proteins at higher molecular weights are also observed.

There was no obvious crosslinking observed for sfGFP^{26azF} (not shown) as the mobility of the protein sample did not change on irradiation. However, sfGFP^{132azF} displayed a band laddering pattern of increasing molecular weight species. The laddering effect was determined to be a minor component <20% determined using ImageJ software) and only at higher protein concentrations (Figure 6.8a). This suggests that an inter-protein crosslink event has occurred. The higher molecular weight bands appear at ~56 kDa, the equivalent of a sfGFP dimer, and in discrete steps that matched trimers (~83 kDa) and higher order oligomers. We cannot be certain that they are oligomers of sfGFP and not band shifts due to self-crosslinking, however the regularity of the ladder effect strongly suggest it. The presence of discrete higher molecular weight protein bands was more evident on UV treatment of sfGFP^{azFx2}. The banding shows regular intervals in steps of approximately 27 kDa. This suggests that crosslinking and thus UV induced protein oligomerisation is more prevalent in the double azF mutant (>70% at 100 µM and higher, Figure 6.8b). Even at low concentrations of Protein (10 µM) significant multimerisation (~40%) was observed. This suggests that although the

presence of an azF at residue 26 has no inherent crosslinking potential it enhances the ability of azF132 to crosslink to other sfGFP molecules.

6.2.6   Determination of oligomerisation states with SEC-DLS

To determine the extent of oligomerisation and the approximate ratios of monomeric and oligomeric protein. Irradiated sfGFP^{azFx2} was concentrated to 5 mg/mL and 100 μL was run in series through size exclusion chromatography followed by analysis of the eluent by dynamic light scattering (SEC-DLS) as described in (Section 2.7.4). Both absorbance and refractive index (RI) show a shoulder and small peak eluting at about 10 mL. There are then 2 more peaks at ~12 mL and 14 mL (Figure 6.9a). All peaks show strong right angled scattering and hence indicates that they are protein (Figure 6.9b). Due to the use of an SEC column the smallest species comes off last and so peaks will now be discussed from right to left according to the numbering in Figure 6.9.

Peak 1 had a hydrodynamic radius of 2.98 nm which equated to a Mw of 27.9 kDa was determined to be monomeric protein. Peak 2 had a 4.1 nm hydrodynamic radius with a Mw of 56.8 kDa, which equates to dimeric protein (theoretical value 55.9 kDa). Peak 3 gave a radius of 4.69 nm and a Mw of ~86 kDa, indicative of a trimer (theoretical Mw 83.8 kDa). The shoulder contains a mixture of different sized particles that couldn't be isolated with the available resolution of the column and suggests that it is populated by higher order oligomers. The total protein recovered from the column was 3.77 mg/mL (75.4 % recovery). Of this ~33% (1.25 mg/mL) was within the monomeric peak. 21% (0.78 mg/mL) was dimer and 12% (0.46 mg/mL) trimer. The remaining protein is within the initial shoulder of all the other order oligomers. This proves that the extra bands seen on SDS-PAGE are indeed multimers of sfGFP^{azFx2} rather than any effects that intramolecular crosslinking may have on electrophoretic mobility. However, the data do not suggest where crosslinking may be occurring.
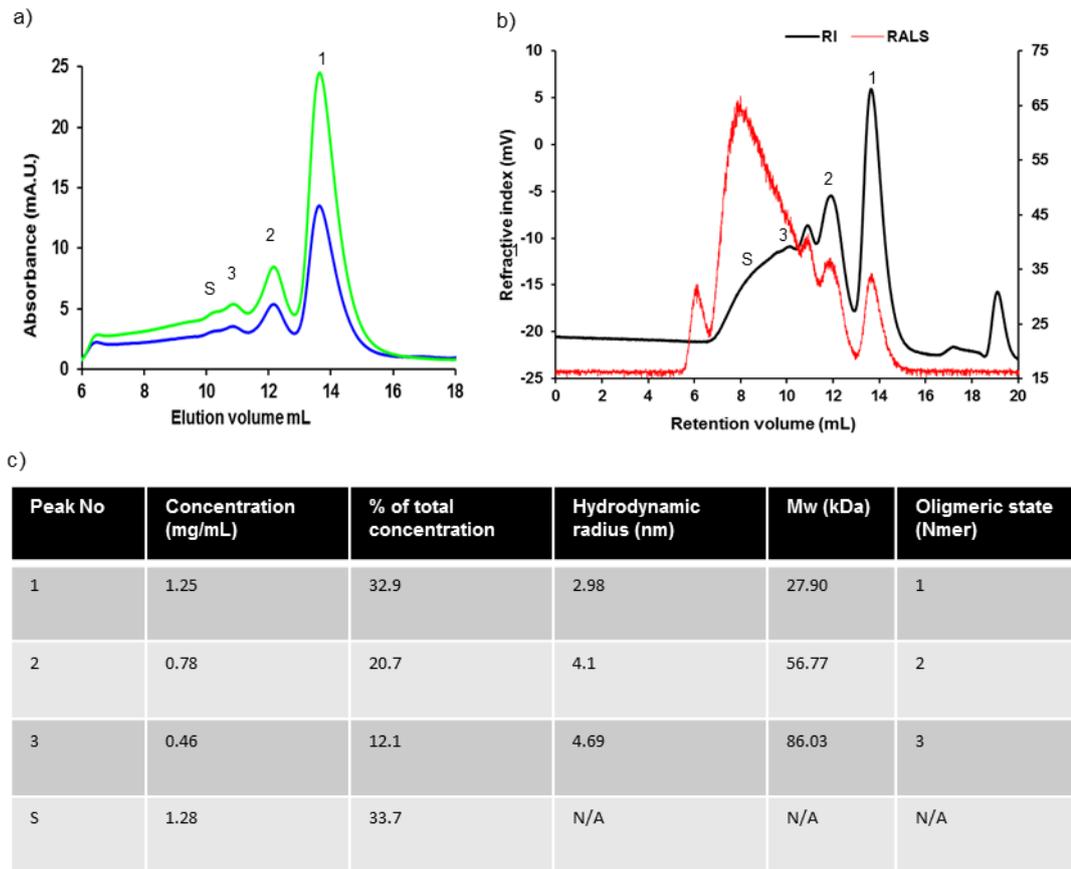
Figure 6.9 Analysis of SEC – DLS output. a) Absorbance traces at 280 nm (blue) and 485 nm (green) Showing protein eluting between 6 and 16 mL elution. Plot of refractive index (RI) and right angled light scattering (RALS) against elution volume. RI matches absorbance profile closely but RALS is higher at lower elution volume. In both instances monomeric, dimeric and trimeric protein peaks are labelled as 1, 2 and 3 respectively. c) Peak table highlighting concentration, molecular weight and predicted oligomeric state. DLS was calibrated to sfGFP^{WT}

6.2.7 Fragmentation mass spec using endopeptidases

To determine where cross links were occurring it was decided to use fragmentation mass spectrometry to identify cross links. In most instances, proteins can be digested directly with proteases such as trypsin and chymotrypsin. However, sfGFP is very thermodynamically stable (Chapter 4) and has been shown to be highly resistant to proteolysis.[281] In order to cleave sfGFP effectively it needs to be denatured to allow easy access to cleavage sites, as such, sfGFP^{WT}, dark and irradiated forms of sfGFP^{azFx2} (~1 mg/mL) were denatured with 6M Guanidinium hydrochloride and cleaved using cyanogen bromide. The advantage of this form of chemical cleavage is that it can be

performed under denaturing conditions,[282] and as such can be used to digest sfGFP. The digests were left overnight and then buffer exchanged into water and concentrated down in a speed vac. ExPASy peptide cutter was used to predict the fragment sizes which are listed in (Table 6.2).
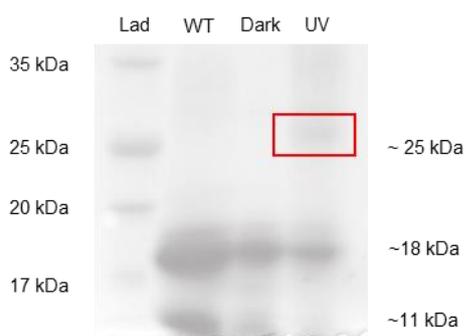
Figure 6.10 SDS-PAGE analysis of cyanogen bromide cleavage of sfGFP[WT] (WT), and dark and irradiated forms of sfGFP[azFx2] (Dark and UV). All three samples successfully cleaved into 2 visible bands at 11 kDa and ~18 kDa, with the irradiated sample showing a minor band at ~25.5 kDa.

Before sending samples off for mass spectrometry they were first analysed by SDS-PAGE to determine if the cleavage was successful (Figure 6.10). Both sfGFP[WT] and the dark state of sfGFP[azFx2] showed two bands on SDS-PAGE at *circa* 11 kDa and 18 kDa. The irradiated form showed an additional faint band at ~25 kDa which could be crosslinked protein. Mass spec didn't show any peaks that match the predicted masses from ExPASy Peptide cutter (Table 6.2). Subsequent attempts to digest the protein including extracting bands from SDS-PAGE and digesting with trypsin or chymotrypsin have also yielded no clear results and so the mechanism of oligomerisation is still unknown. The bands on SDS-PAGE at ~11 kDa and ~18 kDa could possibly be fragments 1-78 and 89-218 respectively, whilst the faint band at ~25 kDa could potentially be a crosslink between two 89-218 fragments. However, without further investigation it is impossible to be certain of this.

Table 6.2 Predicted fragments and lengths form cyanogen bromide cleavage.

Fragment range	Fragment Mw (Da)		Comments
	sfGFP^{WT}	sfGFP^{azFx2}	
0	149	149	GFP numbering starts with N terminal Met as Residue 0
1-78	8,449	8,509	Chromophore and residue 26
79-88	1,266	1,266	
89-218	14,860	14,920	Residue 132
219-233	1,559	1,559	
234-246	1,634	1,634	

6.3 Conclusions

In summary, a mutant sfGFP was successfully engineered that contained two azF residues that work cooperatively, leading to a reduction in sfGFP fluorescence. This is to my knowledge the first instance of engineered cooperativity with ncAAs. By incorporating azF into sfGFP at residues 26 and 132 the fluorescence intensity reduces by ~50%. The apparent mechanism for the fluorescence reduction, is *via* sfGFP^{azFx2} crosslinking to other sfGFP monomers. Although the exact linkage site was not known, it is likely that residue 132 is the residue responsible for crosslinking and azF26 enhances the crosslinking as shown by SDS-PAGE. Given more time I would further explore the linkage and image the oligomers with electron microscopy to determine the alignment.

Given that we see higher order oligomers it is obviously not as simple as residue 132 cross-linking to 132. Although this is only speculation, I believe that 132 crosslinks to one of the β-strands of another sfGFP molecule obscuring the water channels between residues 26 and 132 and reducing fluorescence. Should this be the case each crosslink would leave an exposed azF at the end of the β-barrel allowing for attachment to another monomer in continuous pattern possibly a spiral like arrangement which should account for the small increases in hydrodynamic radius observed by DLS. However, as previously stated without details on the precise location of the crosslink(s)

and information of how monomers are arranged with respect to each other, exact

conclusions cannot be drawn.

7 Discussion

7.1 Personal overview

The common theme throughout this thesis is to utilise the new chemistry available through the use of a reprogramd genetic code to introduce new functional features into a protein and to understand mechanism of action through the use of molecular techniques, especially structural biology. This epitomises the "Design-Build-Test-Learn" process central to synthetic biology. For synthetic biology to flourish as an area, we still need to understand the molecular basis of action – the "Learn" phase. There is still a limited amount of information on the structural impact of ncAA incorporation and downstream effect.

In addition, I feel that this thesis provides new insights into GFP mechanism. For such an important protein used in a variety of contexts, and relatively well studied structural and functionally, some new perspectives on function have been gained. Not least the importance of structured water molecules and their dynamics. Structured water molecules are normally ignored in structural analysis. Here, I have shown their importance in terms of controlling the charged state of CRO and general fluorescent properties such as molar absorbance and quantum yield. Importantly, I have shown that they can be used as part of the protein engineering process.

7.2 General overview

Non-canonical amino acids (ncAAs) are an ideal method to introduce new chemistry into proteins, potentially introducing new functionality. They also allow for new defined routes for post-translational modification. These can be used to control protein function and interface proteins with other proteins and various nanomaterials. One of the key post-translational events in biology is covalent modification that induces changes in protein function through concerted conformational changes. Chapter 3 explored a new

approach using a non-biological, non-native PTM (nnPTM) process. At about the time as work from this thesis was published, several other papers were published utilising the same idea,[78,283,284] *i.e.* introduce nnPTM into proteins by design that modulate function. In this thesis, p-azido-L-phenylalanine (azF) was incorporated into super-folder Green Fluorescent Protein (sfGFP)[127] to positively modulate protein function (fluorescence) using strain promoted azide alkyne cycloaddition (SPAAC) click chemistry (Chapter 3).[107] I also demonstrated how sfGPF148azF is successfully modified with DBCO-amine in *E. coli* showing that the click chemistry approach is amenable for some *in vivo* applications. However, this is limited to the size of the molecule used as DBCO-Texas red was too big to label sfGFP in cells.

Another key post-translational event is protein assembly, either with other proteins or biomolecules. The majority of cellular proteins interact with other proteins either transiently or permanently in form of quaternary structure. Protein oligomerisation routinely results in synergy whereby the assembled complex is greater than the sum of its parts. While a lot of effort has gone into designing protein assemblies, few display this critical feature of synergy. In this thesis, a second ncAA, s-cyclooctyne-L-lysine was used in conjunction with azF to covalently link two sfGFP monomers *via* SPAAC to form artificial protein dimers (Chapter 4). The dimers showed functional enhancement *i.e.* there was a significant increase in fluorescence. This was extended upon in Chapter 6 whereby the photochemistry of azF was used to form higher order assemblies of GFP. An additional interesting facet of this chapter was that combined azF mutations were required to improve assembly efficiency suggesting that the ncAA were acting cooperatively. In most cases, people use ncAA for their new chemistry alone and not for eliciting cooperative effects, despite such cooperativity and interlinking network being a common feature of protein structure.

Finally, I used PTMs outside a biological context in the area of nanoscience. For proteins to be used as single/few molecule components in nanomaterials, we need

defined and controlling coupling with the nanomaterial, so a homogenous interaction is made for consistent bottom-up construction. NcAAs, especially azF provide an ideal route for such coupling and I show in Chapter 5 how both Click chemistry and photochemistry can be used to achieve designed coupling to single walled carbon nanotubes (SWCNTs). Crucially, we find that interfacing is not just a benign process and coupling site influences the communication.

7.3 Benefits of non-native post translational modifications for protein engineering

In this thesis I introduced non-native post translational modifications (nnPTMs) into sfGFP. These nnPTMs were used to modulate sfGFP fluorescence, induce dimerisation and anchor sfGFP to a surface in the form of carbon nanotubes. These modifications are all present in nature as discussed in Chapter 1, but the difference is that the modifications have been introduced by incorporation of a ncAA (azF) into a protein that is not normally modified post translationally in the ways described in this thesis. There are several key points that need to be discussed to determine the effectiveness of nnPTMs, including efficiency, reaction kinetics, ability to work *in vivo*, and determining the locations for modification. The use of azF as a route for nnPTMs will be discussed in the next section.

Natural PTMs evolved to respond to rapid changes in environment and as such need to be both fast and label to a high efficiency. The efficiency of nnPTMs is, in part at least, determined by the reaction kinetics of the modification, with slower reactions leading to lower efficiencies. For example, attaching DBCO-amine to sfGFP[148azF] reached its peak modification level after an hour with little increase in modification afterwards (*in vitro*) and there was ~70% modification efficiency. However, sfGFP dimers were left almost 24 hours with a much lower efficiency, *i.e.* only about ~25% of sfGFP

monomers formed dimers. This is attributed to the much slower reaction kinetics of using SCO compared to DBCO.[285,286] Using UV irradiation however has a much faster rate of reaction with effects noticeable after minutes (seconds when using confocal lasers)[110,152] and reactions reaching near completion however it is hard to determine which reaction the nitrene moiety will undergo. Therefore, in terms of reaction kinetics and efficiency these approaches are not comparable to natural PTMs. owever, this is not necessarily an issue as they are still faster and more efficient than trying to engineer the modifications using canonical PTMs.

7.4 Use of azF as a route to non-canonical PTMS

Phenylazides have many characteristics that are desirable for engineering PTMs into proteins.[112,114] The azide group is very stable in most chemical environments, save for irradiation with UV light, and presence of strong reducing agents,[111,112,114] and can crosslink to a variety of functional groups including primary amines, and aromatic groups, *via* UV irradiation forming a nitrene radical.[111,112,114] Although this can be an issue as it is hard to predict the route that the nitrene will undertake in a protein environment as the residues around the mutation site help to determine the fate of the nitrene, *e.g.* if there is a reducing local protein environment,[85] then the nitrene is more likely to reduce to the phenyl amine as shown with sfGFP[66azF].[152] As azF, and by extension proteins with azF incorporated, are photosensitive it requires that, any work done on the protein (expression, purification, and analysis) must be performed under dark conditions. In the short term, this can be an inconvenience, however once the laboratory is set up appropriately this is easily managed.

Alternatively, azF can undergo cycloaddition with strained alkynes (SPAAC).[107] This is a biorthogonal reaction, meaning that it can be used *in vivo* and is specific with no unwanted biproducts. This is particularly useful for imaging within living cells[91,107,162]

allowing proteins to be labelled with a small fluorophore without the need to genetical encode a fluorescent protein or domain which can cause folding issues.[91,107,162] SPAAC is an improvement on the copper-mediated azide-alkyne cycloaddition reaction, which required a copper catalyst that is toxic to living systems.[107,116,118] A potential issue with using azF in this manner is the relatively slow kinetics of the SPAAC reaction, which can reduce the efficiency of the conjugation.[285] With the development of cycloalkynes with increased stain by adding electron withdrawing groups such as benzyl groups (dibenzyl-cyclooctyne),[286] additional ring structures (bicyclononyne) and substituting hydrogens with fluorine atoms (e.g. difluoro-cyclooctyne), the rate and the efficiency of the reaction has increased.[285,286] Using phenylazides also reduces the rate of SPAAC ~7-fold compared to an aliphatic azide.[287] This can be avoided by using a ncAA amino acid with an aliphatic azide group such as the pyrrolysine derivative azido-ethoxycarbonyl-L-lysine.[288]

Given the success of incorporating SCO and other strained alkyne derivatives it is perhaps better to use these as opposed to an azide. There are several reasons for this including that, alkyne based ncAAs do not require to be kept in the dark and there is a greater variety of adducts *e.g.* fluorescent probes commercially available designed to react with alkynes than for reacting with azides.[121,122,288] An additional benefit of incorporating strained alkynes into proteins as opposed to azides is that strained alkynes can also react with tetrazines and nitrones expanding the range of new nnPTMs that can be explored. Alkyne reactions with tetrazines and nitrones are also bioorthogonal and proceed at a much faster rate than with azides.[84,195]

7.5 Usefulness of *in silico* modelling

In silico modelling has been a key component in rational protein design since the pioneering work by Shakhnovich *et al*[289] in 1994, where they used computational design to create protein based heteropolymer lattices. Applications for *in silico* modelling range from simple molecular dynamic simulations based on existing protein

structures,[167,169,171,198,199,206] to predicted protein small molecule binding in drug design to *ab initio* protein folding of novel protein chains, to prediction of oligomerisation.[14,32,172,181] To meet the high demand for effective techniques to develop *in silico* predictions many groups have created several software packages that aid researchers in adopting a computational approach, delivered in both server and downloadable formats to suit the user's requirements. These include but by no means limited to, GROMACS[167,169,198,199] (MD and QMMM simulations), ROSETTA[171,206,290] (*ab initio* protein folding, protein-protein, protein ligand docking), and ClusPro[172,173,291] (protein-protein, protein-ligand docking software).

A small part of this thesis was to use *in silico* predictions to guide the direction of research. Much research is being geared towards computational approaches in an effort to streamline engineering by reducing workload and time by exploring many possibilities, e.g. amino acid substitutions, and narrowing down the most likely mutants to explore. In part this was possible, as use of ClusPro[172] successfully predicted an interface region on sfGFP for dimerisation. With the prediction it was possible to successfully create functionally linked sfGFP protein dimers. Not only was the technique able to correctly predict where dimerisation would occur but it demonstrated that it can correctly predict where it cannot occur. This was proven by trying to create a dimerisation interface at residue 132 which was ultimately unsuccessful.

To generate the dual ncAA mutant, CAVER[181] was used to predict the presence of water channels in the sfGFP crystal structure. A long-range channel was found leading from the chromophore to the end the β-barrel exiting between two loops containing residues 26 and 132. *In silico* modelling substituting residues 132 and 26 showed a change in the predicted tunnels meaning that these mutations could potentially alter the function of sfGFP. Upon expression, purification and UV irradiation of sfGFP[26/132azF], it was discovered that the protein loses fluorescence. As such correctly

predicted that mutations at both residues would have an effect, however modelling did not predict the mechanism for the change.

One of the more ambitious uses of *in silico* modelling during this thesis is the work described in Chapter 3, where I attempted to use MD simulations to predict the structural effects of modification of sfGFP[148azF] with DBCO-amine. In order to perform MD simulations parameters have to be created for every type of atom, bond and long range interaction including lengths, angles and dihedrals, for all residues in a system.[101,169,292,293] This was not an issue for the 20 canonical amino acids, because there are parameters available for use in GROMACS with my choice being the AMBERsb99 forcefield.[168,292] However, to model sfGFP[148azF]+DBCO-amine it was necessary to create parameters for the chromophore (considered a ncAA for *in silico*), azF both intact and modified with DBCO-amine. Creating the residues *in silico* is described in Chapter 2, and for the most part successful in creating representations for demonstrative purposes and seen in Chapters 3,4,5 and 6 and in.[115,119] However, after molecular dynamics several issues became apparent such as the chromophore losing the correct conformation. This was attributed to incorrect parameterisation and is an area for further work to perfect the technique. It should be noted that generation of stable forcefields for molecular dynamics is very much an ongoing project involving large international collaborations and requires dedicated time and focus greater than that allowed as a small part of a single PhD. For example, the collaboration that maintains and improves the AMBER suite of forcefields has over 50 individual authors and is currently on the 18th iteration.[294]

7.6 Significance of small structural perturbations on local conditions

As mentioned throughout this thesis, hydrogen bonding networks around the chromophore (CRO), and throughout the structure of sfGFP are integral to the function of sfGFP. Indeed, H-bond networks are pivotal through the proteome and have roles beyond acting to simply maintain structure. I have shown here that using ncAA and associated downstream process, such networks can be manipulated to drive functional changes. In terms of sfGFP, small changes to this network can lead to large changes in the fluorescence properties of sfGFP.[9,127,135,146,147,153] A prime example of this is with mutation of residue H148 to azF, which as described in Chapters 1 and 3, and an important residue in maintaining the ionic state of the chromophore by directly hydrogen bonding to the tyrosyl group of the chromophore. The mutation to azF removes a key hydrogen bond between CRO and 148 and causes a backbone shift of ~1.4 Å on β-strand 7 further disrupting the hydrogen bonding network. The effect of this is that sfGPF[148azF] electronically excites primarily at 400 nm suggesting that CRO exists primarily in the protonated CroA form in the ground state and reduces the overall fluorescence. Modification with DBCO-amine reverses the shift in β-strand 7 and re-establishes the hydrogen bond network, presumably by coordinating a water molecule to hydrogen bond with the tyrosyl group of CRO promoting deprotonation. This shifts the predominant excitation peak to ~485 nm with little observable 400nm peak suggesting that CRO exists almost exclusively in the anionic CroB form with very few CROs in the neutral CroA. This is analogous to other native PTMs such as phosphorylation, where a subtle conformational change activates the protein.[295] However, unlike phosphorylation SPAAC is not easily reversible so is no use for reversible functional switching.

When modifying sfGPF[148azF] with sfGPF[148SCO] to form a covalently bound dimer (sfGFP[148x2]), similar small structural changes, *i.e.* a ~1.4 Å shift in β-strand 7 allowing for the entry of a structural water molecule, create a continuous hydrogen bonding

network connecting the two chromophores of the dimer. It is believed that this new network between monomers is what gives the enhanced fluorescence observed in Chapter4. The dominant excitation has again been shifted from CroA excitation to CroB for both monomers and again proves how small changes to the structure of a protein can lead to great effects in function.

The hydrogen bonding network around the chromophore has been shown to act as a proton shuttling system described as a network of 'proton wires',[185] which has been shown to act as light induced proton and electron donors.[262] In Chapter 5, sfGFP is interfaced with swCNTs creating bio-nanohybrids. These bio-nanohybrids show evidence of communication between protein and CNT. It is believed that there is electronic communication between sfGFP and swCNT, which allows for the energy transfer displayed as a reduction in fluorescence intensity but an increase in time until photobleaching.

Knowledge of these hydrogen bonding networks can be used to rationally engineer proteins to allow for specific control of a protein. In Chapter 6 we use the knowledge gained about the hydrogen bonding networks in sfGFP, to rationally design a dual azF mutant (sfGFP[azFx2]) with azF incorporated around the exit of one of these hydrogen channels connecting the chromophore to solvent at the end of the β-barrel opposite to the N and C termini. UV irradiation of sfGFPazFx2 causes a fluorescence decrease due to individual protein monomers crosslinking with each other. Although we were unable to elucidate the exact location of the crosslinking it is likely to block the exit of this network and hence affect the ability of sfGFP to fluoresce.

7.7 Future Work

This thesis has presented a large body of work including the first demonstrations (to the author's knowledge) of functional linked dimers using SPAAC and the first crystal structure of sfGFP containing a ncAA modified post translationally using SPAAC.

However, there is still more work to be done. Firstly, in Chapter 6 I began to look at the effect of incorporating multiple ncAAs in a protein to induce functional cooperativity between the two residues. The resulting protein (sfGFP^{azFx2}), showed a significant loss of fluorescence predicted to be due to multimerisation of the protein disrupting a water channel between sfGFP chromophore and the bulk solvent. However, the precise locations of the cross links were not determined, and this should be rectified before progressing this avenue further.

In Chapter 4 artificial covalently linked protein dimers were generated that showed enhanced function beyond the simple addition of the two proteins together. To further this work, we need to look beyond simple homodimers *i.e.* generating dimers (and multimers) that contain different proteins. Work in lab has already progressed to linking sfGFP, with other fluorescent proteins such as mVenus and mCherry to better demonstrate the energy transfer potential of these oligomers. Linkage of sfGFP with the electron transfer protein cytochrome b_{562}, which has previously been shown to undergo energy transfer with GFP in the form of quenching[134] is being explored. Another route that should be explored is linking a series of enzymes in a reaction pathway to engineer simple protein factories. For example, fusing a superoxide dismutase with a peroxidase to create a detoxification system. One route that is of personal interest to me would be to create a pair of fluorescent proteins that are non-fluorescent until dimerisation where they then activate. This has potential applications in the study of protein-protein interactions, which could be applied *in vivo* to monitor binding events in real time. To do this I would recreate the basis for systems like nano-bit (Promega), which uses a split luciferase system to monitor protein-protein interactions. Firstly, I would generate a split sfGFP, *e.g.* fragment 1 – β-strands 1-10, fragment 2 β-strand 11. The fragment at linked to two proteins of interest, for example two different GPCR monomers and dimerisation can be monitored by an increase in fluorescence. The split sfGFP fragments can either be genetically introduced into the proteins of interest or can be attached *via* SPAAC with

the individual proteins having one ncAA and the split sfGFP having the second. However, this would require a large amount of optimising to ensure correct labelling of proteins.

Towards the end of Chapter 5, the possibility of creating protein based junctions was explored by creating a protein with azF at opposite ends of the protein. This allowed for attachment of two swCNTs to the protein forming a junction. Unfortunately, the effects of this dual attachment on the protein were not explored. Looking at the effect of this attachment on conductance through the tubes before and after excitation of sfGFP would demonstrate the viability of using sfGFP as a potential optically gated transistor in nanoelectronics. This part of the research only focussed on the 'short axis' of sfGFP and hence future work should also explore the use of the 'long axis' as a protein junction as well as looking at attachment other than directly opposite and 3/4 way junctions.

Refinement of the approaches described in Chapter 5, to interface proteins and swCNTs to form functional electronic components, is perhaps one of the most exciting future directions to take this research. Firstly, adapting the UV linkage approach to a solution based method should improve the uniformity in density and number of proteins along the swCNT. To reach the full potential of the sfGFP-swCNT nanohybrids they need to be reliably and robustly attached to electrodes so their use as potential light harvesting systems and as optical transistors can be explored. On the note of light harvesting systems, as one of the big issues with early solar cells was not being able to absorb the full light spectrum available, it would be useful to expand the range of fluorescent proteins (FPs) across the colour range to maximise the range of wavelengths absorbed by the systems to capture as much energy as possible.

Once the protocols for protein-CNT hybridization have been optimised with FPs the next step is to interface different proteins onto CNTs to create biosensors *e.g.* interfacing β-lactamase inhibitor proteins to CNTs has the potential for creating an antimicrobial resistance detector. By incorporating ligand binding domains of receptors,

we can adapt the technology for high throughput drug discovery to find compounds that bind receptors.

This research focussed on the use of azF and the associated chemistries (UV photochemistry and SPAAC) that azF introduces into a protein. As discussed previously (Chapter 1 and Section 7.4), there are other ncAAs that can bring other new chemistries into proteins. One potentially interesting ncAA is SCO, which has been used in this thesis as a counterpart for azF in the generation of sfGFP dimer. As discussed in Section 7.4, SCO has a strained alkyne functional group which undergoes click reactions with azides, tetrazines and phenylnitrones. The reaction rates with the latter two are much faster than the reaction between SCO and phenyl azides and so would be a preferable alternative to increase the efficiency of creating artificial protein oligomers and protein-CNT hybrids.

7.8 Summary

Chapter 3 demonstrated the modulation of sfGFP using click chemistry. Upon adding DBCO-amine to sfGFP[148azF], the dominant excitation was switched from ~400nm to ~485nm hinting that the chromophore was existing predominantly in the anionic CroB state. The crystal structure of sfGPF[148azF]+DBCO-amine was solved to reveal the likely cause for the switch in chromophore state. Using structural information and prediction from CAVER[181] software it was decided that the mechanism is from a backbone shift of β-strand 7 regenerating a water channel that is lost when H148 is substituted for azF.

Chapter 4 introduces a methodology for the design and creation of functional sfGFP protein dimers. *In silico* design was initially used to narrow down a potential interface site from the whole of sfGFP to a small region around β-strands 7 and 8. Residues 148 and 204 were selected to create two symmetrical protein dimers (sfgFP[148x2] and sfGFP[204x2]) with increased fluorescence beyond a simple additive effect. sfGFP[148x2] also showed the same switch in dominant excitation as seen in Chapter 3.

Crystallisation of the sfGFP^{148x2} dimer reveals that the enhanced function is from formation of an intricate network between both chromophores in the dimer. Chapter 4 also highlighted the importance of symmetry in these protein oligomeric systems by creation of non-symmetrical sfGFP dimers showing no enhancement of function.

Chapter 5 describes the creation of sfGFP-swCNT nanohybrids using two different approaches, UV irradiation and click chemistry. The importance of orientation of proteins was investigated by creating two sfGFP mutants with azF incorporated at either residue 132 or 204, termed sfGFP^{LA} and sfGFP^{SA} respectively. Attachment to swCNTS *via* direct UV linkage showed communication between proteins and swCNTs for both residues, however with the click-chemistry approach only sfGFP^{SA} showed any evidence of communication, reinforcing the importance of proximity in effective energy transfer. Chapter 5 also presented the creation of a sfGFP mutant with two azF residues incorporated dubbed sfGFP^{SAx2}. Although energy transfer in this mutant was not explored, it was successfully linked to the end of two swCNTs creating a CNT-protein junction.

Chapter 6 attempted to engineer cooperativity in sfGFP using two ncAAs. A sfGFP variant was created with azF incorporated at residues 26 and 132 (sfGFP^{azFx2}). After UV irradiation the fluorescence of sfGFP^{azFx2} decreased by ~50%. The mechanism for this fluorescence loss is due to crosslinking between sfGFP molecules creating ordered oligomers. Although unable to elucidate the exact crosslinking site for the oligomerisations it is likely that azF132 is the main crosslinking residue with azF26 promoting the crosslink, as shown by low level crosslinking in the sfGFP^{132azF}, which increases when azF is incorporated at residue 26.

7.9 Publications associated with this thesis

Chapter 3

Hartley, A.M., **Worthy, H.L.**,[1] Reddington, S.C., Rizkallah, P.J. & Jones, D.D. Molecular basis for functional switching of GFP by two disparate non-native post-translational modifications of a phenyl azide reaction handle. *Chemical Science* **7**, 6484-6491 (2016).

Chapter 4

Worthy, H.L.,[1] Auhim, H.S, Jamieson, W.D., Pope, J.R., Wall, A., Johnson, R.L., Watkins, D.W., Rizkallah, P.J., Castell, O.K., & Jones, D.D. Artificially assembled, structurally integrated protein dimers by designed Click chemistry: positive functional synergy through new inter-unit networks. (2018)[2]

Chapter 5

Freeley, M., **Worthy, H.L.**,[1] Ahmed, R., Bowen, B., Watkins, D., Macdonald, J.E., Zheng, M., Jones, D.D., & Palma, M. Site-Specific One-to-One Click Coupling of Single Proteins to Individual Carbon Nanotubes: A Single-Molecule Approach. *Journal of the American Chemical Society* **139**, 17834-17840 (2017)

Beachey, A., **Worthy, H.L.**,[1] Jamieson, W.D., Thomas, S.K., Craciun, M.F., Macdonald, J.E., Elliott, M., Castell, O.K., & Jones, D.D. Designed direct protein covalent nanotube attachment and its electronic impact on single molecule fluorescent protein function (2018)[2]

Other publications

Halliwell, L.M., Jathoul, A.P.m Bate, J.P., **Worthy, H.L.**, Anderson, J.C., Jones, D.D, & Murray, J.A.H. Delta Flucs: Brighter *Photinus pyralis* firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnology and Bioengineering* **115**, 50-59 (2018)

---

[1] Joint 1st author
[2] Publication Pending

8 Bibliography

1. Anfinsen, C.B., Redfield, R.R., Choate, W.L., Page, J. & Carroll, W.R. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *The Journal of Biological Chemistry* **207**, 201-210 (1954).
2. Lu, Y., Yeung, N., Sieracki, N. & Marshall, N.M. Design of functional metalloproteins. *Nature* **460**, 855-862 (2009).
3. Miah, M.F. & Boffa, M.B. Functional analysis of mutant variants of thrombin-activatable fibrinolysis inhibitor resistant to activation by thrombin or plasmin. *Journal of Thrombosis and Haemostasis* **7**, 665-672 (2009).
4. Oohora, K. & Hayashi, T. Hemoprotein-based supramolecular assembling systems. *Current Opinion in Chemical Biology* **19**, 154-161 (2014).
5. Perham, R.N. Swinging Arms and Swinging Domains in Multifunctional Enzymes: Catalytic Machines for Multistep Reactions. *Annual Review of Biochemistry* **69**, 961-1004 (2000).
6. Walsh, C.T., Garneau-Tsodikova, S. & Gatto, G.J. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angewandte Chemie-International Edition* **44**, 7342-7372 (2005).
7. Waters, R., van Eijk, P. & Reed, S. Histone modification and chromatin remodeling during NER. *DNA Repair* **36**, 105-113 (2015).
8. Yu, S.R. et al. Global genome nucleotide excision repair is organized into domains that promote efficient DNA repair in chromatin. *Genome Research* **26**, 1376-1387 (2016).
9. Tsien, R.Y. The green fluorescent protein. *Annual Review of Biochemistry* **67**, 509-544 (1998).
10. Marianayagam, N.J., Sunde, M. & Matthews, J.M. The power of two: protein dimerization in biology. *Trends in Biochemical Sciences* **29**, 618-625 (2004).
11. Mei, G., Di Venere, A., Rosato, N. & Finazzi-Agro, A. The Importance of Being Dimeric. *Febs Journal* **272**, 16-27 (2005).
12. Ali, M.H. & Imperiali, B. Protein oligomerization: How and why. *Bioorganic & Medicinal Chemistry* **13**, 5013-5020 (2005).
13. Goodsell, D.S. & Olson, A.J. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure* **29**, 105-153 (2000).
14. Ahnert, S.E., Marsh, J.A., Hernandez, H., Robinson, C.V. & Teichmann, S.A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**(2015).
15. Larsen, T.A., Olson, A.J. & Goodsell, D.S. Morphology of protein-protein interfaces. *Structure with Folding & Design* **6**, 421-427 (1998).
16. Adams, M.J. et al. Structure of Rhombohedral 2 Zinc Insulin Crystals. *Nature* **224**, 491-& (1969).
17. Leibly, D.J. et al. A Suite of Engineered GFP Molecules for Oligomeric Scaffolding. *Structure* **23**, 1754-1768 (2015).
18. Lewis, M.S. & Youle, R.J. Ricin Subunit Association - Thermodynamics and the Role of the Disulphide Bond in Toxicity. *Journal of Biological Chemistry* **261**, 1571-1577 (1986).
19. Klotz, I.M. PROTEIN SUBUNITS - A TABLE. *Science* **155**, 697-& (1967).
20. Monod, J., Wyman, J. & Changeux, J.P. On Nature of Allosteric Transitions - A Plausible Model. *Journal of Molecular Biology* **12**, 88-& (1965).
21. Labeit, S. & Kolmerer, B. Titins - Giant Proteins in Charge of Muscle Structure and Elasticity. *Science* **270**, 293-296 (1995).
22. Pauling, L. Protein Interactions. 4. Aggregation of Globular Proteins. *Discussions of the Faraday Society*, 170-176 (1953).
23. Crick, F.H.C. & Watson, J.D. Structure of Small Viruses. *Nature* **177**, 473-475 (1956).

24. Kobayashi, N. & Arai, R. Design and construction of self-assembling supramolecular protein complexes using artificial and fusion proteins as nanoscale building blocks. *Current Opinion in Biotechnology* **46**, 57-65 (2017).

25. Norn, C.H. & Andre, I. Computational design of protein self-assembly. *Current Opinion in Structural Biology* **39**, 39-45 (2016).

26. Yeates, T.O., Liu, Y.X. & Laniado, J. The design of symmetric protein nanomaterials comes of age in theory and practice. *Current Opinion in Structural Biology* **39**, 134-143 (2016).

27. Ljubetic, A., Gradisar, H. & Jerala, R. Advances in design of protein folds and assemblies. *Current Opinion in Chemical Biology* **40**, 65-71 (2017).

28. Zakeri, B. et al. Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E690-E697 (2012).

29. Leng, W.B. et al. Novel Split-Luciferase-Based Genetically Encoded Biosensors for Noninvasive Visualization of Rho GTPases. *Plos One* **8**(2013).

30. Huang, P.S. et al. High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481-485 (2014).

31. Mou, Y., Huang, P.S., Hsu, F.C., Huang, S.J. & Mayo, S.L. Computational design and experimental verification of a symmetric protein homodimer. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 10714-10719 (2015).

32. Thomson, A.R. et al. Computational design of water-soluble alpha-helical barrels. *Science* **346**, 485-488 (2014).

33. Song, W.J. & Tezcan, F.A. A designed supramolecular protein assembly with in vivo enzymatic activity. *Science* **346**, 1525-1528 (2014).

34. Brodin, J.D. et al. Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nature Chemistry* **4**, 375-382 (2012).

35. Fallas, J.A. et al. Computational design of self-assembling cyclic protein homo-oligomers. *Nature Chemistry* **9**, 353-360 (2017).

36. Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365-1368 (2015).

37. Sahasrabuddhe, A. et al. Confirmation of intersubunit connectivity and topology of designed protein complexes by native MS. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 1268-1273 (2018).

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

39. Aupic, J., Lapenta, F., Strmsek, Z. & Jerala, R. Towards designing new nano-scale protein architectures. *Synthetic Biology* **60**, 315-324 (2016).

40. Bale, J.B. et al. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389-394 (2016).

41. Pessino, V., Citron, Y.R., Feng, S.Y. & Huang, B. Covalent Protein Labeling by SpyTag-SpyCatcher in Fixed Cells for Super-Resolution Microscopy. *Chembiochem* **18**, 1492-1495 (2017).

42. Bailey, J.B., Subramanian, R.H., Churchfield, L.A. & Tezcan, F.A. Metal-Directed Design of Supramolecular Protein Assemblies. *Peptide, Protein and Enzyme Design* **580**, 223-250 (2016).

43. Brodin, J.D., Smith, S.J., Carr, J.R. & Tezcan, F.A. Designed, Helical Protein Nanotubes with Variable Diameters from a Single Building Block. *Journal of the American Chemical Society* **137**, 10468-10471 (2015).

44. Zhu, S. et al. Genetically encoding a light switch in an ionotropic glutamate receptor reveals subunit-specific interfaces. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 6081-6086 (2014).

45. Artes, J.M., Diez-Perez, I. & Gorostiza, P. Transistor-like Behavior of Single Metalloprotein Junctions. *Nano Letters* **12**, 2679-2684 (2012).

46. van Amsterdam, I.M.C. et al. Dramatic Modulation of Electron Transfer in Protein Complexes by Crosslinking. *Nature Structural Biology* **9**, 48-52 (2002).

47. van Vught, R., Pieters, R.J. & Breukink, E. Site-Specific Functionalization of Proteins and Their Applications to Therapeutic Antibodies. *Computational and Structural Biotechnology Journal* **9**, e201402001 (2014).

48. Chen, M., Zeng, G.M., Xu, P., Lai, C. & Tang, L. How Do Enzymes 'Meet' Nanoparticles and Nanomaterials? *Trends in Biochemical Sciences* **42**, 914-930 (2017).

49. Holland, J.T., Lau, C., Brozik, S., Atanassov, P. & Banta, S. Engineering of Glucose Oxidase for Direct Electron Transfer via Site-Specific Gold Nanoparticle Conjugation. *Journal of the American Chemical Society* **133**, 19262-19265 (2011).

50. Della Pia, E.A., Macdonald, J.E., Elliott, M. & Jones, D.D. Direct Binding of a Redox Protein for Single-Molecule Electron Transfer Measurements. *Small* **8**, 2341-2344 (2012).

51. Della Pia, E.A. et al. Fast electron transfer through a single molecule natively structured redox protein. *Nanoscale* **4**, 7106-7113 (2012).

52. Guo, X.F. Single-Molecule Electrical Biosensors Based on Single-Walled Carbon Nanotubes. *Advanced Materials* **25**, 3397-3408 (2013).

53. Ordinario, D.D. et al. Sequence Specific Detection of Restriction Enzymes at DNA-Modified Carbon Nanotube Field Effect Transistors. *Analytical Chemistry* **86**, 8628-8633 (2014).

54. Rosenstein, J.K., Lemay, S.G. & Shepard, K.L. Single-molecule bioelectronics. *Wiley Interdisciplinary Reviews-Nanomedicine and Nanobiotechnology* **7**, 475-493 (2015).

55. Sorgenfrei, S. et al. Label-free single-molecule detection of DNA-hybridization kinetics with a carbon nanotube field-effect transistor. *Nature Nanotechnology* **6**, 125-131 (2011).

56. So, H.M. et al. Single-walled carbon nanotube biosensors using aptamers as molecular recognition elements. *Journal of the American Chemical Society* **127**, 11906-11907 (2005).

57. De Leo, F., Magistrato, A. & Bonifazi, D. Interfacing proteins with graphitic nanomaterials: from spontaneous attraction to tailored assemblies. *Chemical Society Reviews* **44**, 6916-6953 (2015).

58. Jia, C.C., Ma, B.J., Xin, N. & Guo, X.F. Carbon Electrode-Molecule Junctions: A Reliable Platform for Molecular Electronics. *Accounts of Chemical Research* **48**, 2565-2575 (2015).

59. Zaki, A.J. et al. Defined covalent assembly of protein molecules on graphene using a genetically encoded photochemical reaction handle. *Rsc Advances* **8**, 5768-5775 (2018).

60. Zhang, A. & Lieber, C.M. Nano-Bioelectronics. *Chemical Reviews* **116**, 215-257 (2016).

61. Elliott, M. & Jones, D.D. Approaches to single-molecule studies of metalloprotein electron transfer using scanning probe-based techniques. *Biochemical Society Transactions* **46**, 1-9 (2018).

62. Fournier, E. et al. Creation of a gold nanoparticle based electrochemical assay for the detection of inhibitors of bacterial cytochrome bd oxidases. *Bioelectrochemistry* **111**, 109-114 (2016).

63. Hermanova, S., Zarevucka, M., Bousa, D., Pumera, M. & Sofer, Z. Graphene oxide immobilized enzymes show high thermal and solvent stability. *Nanoscale* **7**, 5852-5858 (2015).

64. Khaksarinejad, R., Mohsenifar, A., Rahmani-Cherati, T., Karami, R. & Tabatabaei, M. An Organophosphorus Hydrolase-Based Biosensor for Direct

Detection of Paraoxon Using Silica-Coated Magnetic Nanoparticles. *Applied Biochemistry and Biotechnology* **176**, 359-371 (2015).

65. Sachdeva, V. & Hooda, V. Immobilization of nitrate reductase onto epoxy affixed silver nanoparticles for determination of soil nitrates. *International Journal of Biological Macromolecules* **79**, 240-247 (2015).

66. Attar, A. et al. Amperometric inhibition biosensors based on horseradish peroxidase and gold sononanoparticles immobilized onto different electrodes for cyanide measurements. *Bioelectrochemistry* **101**, 84-91 (2015).

67. Kim, J., Grate, J.W. & Wang, P. Nanostructures for enzyme stabilization. *Chemical Engineering Science* **61**, 1017-1026 (2006).

68. Laurent, N., Haddoub, R. & Flitsch, S.L. Enzyme catalysis on solid surfaces. *Trends in Biotechnology* **26**, 328-337 (2008).

69. Wang, H. et al. Transducing methyltransferase activity into electrical signals in a carbon nanotube-DNA device. *Chemical Science* **3**, 62-65 (2012).

70. Chen, R.J. et al. Noncovalent functionalization of carbon nanotubes for highly specific electronic biosensors. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4984-4989 (2003).

71. Choi, Y. et al. Single-Molecule Lysozyme Dynamics Monitored by an Electronic Circuit. *Science* **335**, 319-324 (2012).

72. Asuri, P. et al. Water-soluble carbon nanotube-enzyme conjugates as functional biocatalytic formulations. *Biotechnology and Bioengineering* **95**, 804-811 (2006).

73. Gooding, J.J. et al. Protein electrochemistry using aligned carbon nanotube arrays. *Journal of the American Chemical Society* **125**, 9006-9007 (2003).

74. Karajanagi, S.S., Vertegel, A.A., Kane, R.S. & Dordick, J.S. Structure and function of enzymes adsorbed onto single-walled carbon nanotubes. *Langmuir* **20**, 11594-11599 (2004).

75. Yu, X., Chattopadhyay, D., Galeska, I., Papadimitrakopoulos, F. & Rusling, J.F. Peroxidase activity of enzymes bound to the ends of single-wall carbon nanotube forest electrodes. *Electrochemistry Communications* **5**, 408-411 (2003).

76. Redeker, E.S. et al. Protein Engineering For Directed Immobilization. *Bioconjugate Chemistry* **24**, 1761-1777 (2013).

77. Arpino, J.A.J., Rizkallah, P.J. & Jones, D.D. Crystal Structure of Enhanced Green Fluorescent Protein to 1.35 angstrom Resolution Reveals Alternative Conformations for Glu222. *Plos One* **7**(2012).

78. Hartley, A.M. et al. Functional modulation and directed assembly of an enzyme through designed non-natural post-translation modification. *Chemical Science* **6**, 3712-3717 (2015).

79. Chin, J. et al. Addition of p-Azido-L-phenylalanine to the Genetic Code of Escherichia coli. *J. Am. Chem. Soc.* **124**, 9026-27 (2002).

80. Xie, J. & Schultz, P.G. Innovation: A chemical toolkit for proteins - an expanded genetic code. *Nature Reviews Molecular Cell Biology* **7**, 775-782 (2006).

81. Xie, J.M., Liu, W.S. & Schultz, P.G. A genetically encoded bidentate, metal-binding amino acid. *Angewandte Chemie-International Edition* **46**, 9239-9242 (2007).

82. Wuttke, D.S., Gray, H.B., Fisher, S.L. & Imperiali, B. Semisynthesis of Bipyridyl Alanine Cytochrome-C Mutants - Novel Proteins with Enhanced Electron-Transfer Properties. *Journal of the American Chemical Society* **115**, 8455-8456 (1993).

83. Sakamoto, K. et al. Site-specific incorporation of an unnatural amino acid into proteins in mammalian cells. *Nucleic Acids Research* **30**, 4692-4699 (2002).

84. Plass, T., Milles, S., Koehler, C., Schultz, C. & Lemke, E.A. Genetically Encoded Copper-Free Click Chemistry. *Angewandte Chemie-International Edition* **50**, 3878-3881 (2011).

85. Reddington, S., Watson, P., Rizkallah, P., Tippmann, E. & Jones, D.D. Genetically encoding phenyl azide chemistry: new uses and ideas for classical biochemistry. *Biochemical Society Transactions* **41**, 1177-1182 (2013).

86. Tam, A. & Raines, R.T. Protein Engineering With the Traceless Staudinger Ligation. *Methods in Enzymology: Non-Natural Amino Acids* **462**, 25-44 (2009).

87. Tookmanian, E.M., Phillips-Piro, C.M., Fenlon, E.E. & Brewer, S.H. Azidoethoxyphenylalanine as a Vibrational Reporter and Click Chemistry Partner in Proteins. *Chemistry-a European Journal* **21**, 19096-19103 (2015).

88. Noren, C.J., Anthonycahill, S.J., Griffith, M.C. & Schultz, P.G. A General Method for Site-Specific Incorporation of Unnatural Amino-Acids into Proteins. *Science* **244**, 182-188 (1989).

89. Dumas, A., Lercher, L., Spicer, C.D. & Davis, B.G. Designing logical codon reassignment - Expanding the chemistry in biology. *Chemical Science* **6**, 50-69 (2015).

90. Nikic, I. et al. Minimal Tags for Rapid Dual-Color Live-Cell Labeling and Super-Resolution Microscopy. *Angewandte Chemie-International Edition* **53**, 2245-2249 (2014).

91. Nikic, I. & Lemke, E.A. Genetic code expansion enabled site-specific dual-color protein labeling: superresolution microscopy and beyond. *Current Opinion in Chemical Biology* **28**, 164-173 (2015).

92. Nikic, I., Kang, J.H., Girona, G.E., Aramburu, I.V. & Lemke, E.A. Labeling proteins on live mammalian cells using click chemistry. *Nature Protocols* **10**, 780-791 (2015).

93. Tyagi, S. & Lemke, E.A. Single-molecule FRET and crosslinking studies in structural biology enabled by noncanonical amino acids. *Current Opinion in Structural Biology* **32**, 66-73 (2015).

94. Chin, J.W. et al. Addition of p-azido-L-phenylaianine to the genetic code of Escherichia coli. *Journal of the American Chemical Society* **124**, 9026-9027 (2002).

95. Sun, R., Zheng, H., Fang, Z. & Yao, W. Rational design of aminoacyl-tRNA synthetase specific for p-acetyl-L-phenylalanine. *Biochem Biophys Res Commun* **391**, 709-15 (2010).

96. Mukai, T. et al. Codon reassignment in the Escherichia coli genetic code. *Nucleic Acids Research* **38**, 8188-8195 (2010).

97. Isaacs, F.J. et al. Precise Manipulation of Chromosomes in Vivo Enables Genome-Wide Codon Replacement. *Science* **333**, 348-353 (2011).

98. Iraha, F. et al. Functional replacement of the endogenous tyrosyl-tRNA synthetase-tRNATyr pair by the archaeal tyrosine pair in Escherichia coli for genetic code expansion. *Nucleic Acids Res* **38**, 3682-91 (2010).

99. Kobayashi, T. et al. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. *Nat Struct Biol* **10**, 425-32 (2003).

100. Kobayashi, T. et al. Structural basis of nonnatural amino acid recognition by an engineered aminoacyl-tRNA synthetase for genetic code expansion. *Proc Natl Acad Sci U S A* **102**, 1366-71 (2005).

101. Acharya, A. et al. Photoinduced Chemistry in Fluorescent Proteins: Curse or Blessing? *Chemical Reviews* **117**, 758-795 (2017).

102. Nguyen, D.P. et al. Genetic Encoding and Labeling of Aliphatic Azides and Alkynes in Recombinant Proteins via a Pyrrolysyl-tRNA Synthetase/tRNA(CUA) Pair and Click Chemistry. *Journal of the American Chemical Society* **131**, 8720-+ (2009).

103. Hao, B. et al. A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science* **296**, 1462-1466 (2002).

104. Srinivasan, G., James, C.M. & Krzycki, J.A. Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA. *Science* **296**, 1459-1462 (2002).

105. Neumann, H., Peak-Chew, S.Y. & Chin, J.W. Genetically encoding N-epsilon-acetyllysine in recombinant proteins. *Nature Chemical Biology* **4**, 232-234 (2008).

106. Nguyen, D.P., Alai, M.M.G., Kapadnis, P.B., Neumann, H. & Chin, J.W. Genetically Encoding N-epsilon-Methyl-L-lysine in Recombinant Histones. *Journal of the American Chemical Society* **131**, 14194-+ (2009).

107. Baskin, J.M. et al. Copper-free click chemistry for dynamic in vivo imaging. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 16793-16797 (2007).

108. Bräse, S., Gil, C., Knepper, K. & Zimmermann, V. Organic Azides: An Exploding Diversity of a Unique Class of Compounds. *Angew. Chem. Int. Ed.,* **44**, 5188–5240 (2005).

109. Marth, G. et al. Precision Templated Bottom-Up Multiprotein Nanoassembly through Defined Click Chemistry Linkage to DNA. *Acs Nano* **11**, 5003-5010 (2017).

110. Reddington, S.C. et al. Genetically encoded phenyl azide photochemistry drives positive and negative functional modulation of a red fluorescent protein. *Rsc Advances* **5**, 77734-77738 (2015).

111. Schnapp, K.A., Poe, R., Leyva, E., Soundararajan, N. & Platz, M.S. Exploratory Photochemistry of Fluorinated Aryl Azides - Implications for the Design of Photoaffinity-Labelling Reagents. *Bioconjugate Chemistry* **4**, 172-177 (1993).

112. Morris, J.L. et al. Aryl Azide Photochemistry in Defined Protein Environments. *Organic Letters* **15**, 728-731 (2013).

113. Liu, L.H. & Yan, M.D. Perfluorophenyl Azides: New Applications in Surface Functionalization and Nanomaterial Synthesis. *Accounts of Chemical Research* **43**, 1434-1443 (2010).

114. Schrock, A.K. & Schuster, G.B. Photochemistry of Phenyl Azide - Chemical Properties of the Transient Intermediates. *Journal of the American Chemical Society* **106**, 5228-5234 (1984).

115. Hartley, A.M., Worthy, H.L., Reddington, S.C., Rizkallah, P.J. & Jones, D.D. Molecular basis for functional switching of GFP by two disparate non-native post-translational modifications of a phenyl azide reaction handle. *Chemical Science* **7**, 6484-6491 (2016).

116. Worrell, B.T., Malik, J.A. & Fokin, V.V. Direct Evidence of a Dinuclear Copper Intermediate in Cu(I)-Catalyzed Azide-Alkyne Cycloadditions. *Science* **340**, 457-460 (2013).

117. Jewett, J.C., Sletten, E.M. & Bertozzi, C.R. Rapid Cu-Free Click Chemistry with Readily Synthesized Biarylazacyclooctynones. *Journal of the American Chemical Society* **132**, 3688-+ (2010).

118. Sletten, E.M. & Bertozzi, C.R. From Mechanism to Mouse: A Tale of Two Bioorthogonal Reactions. *Accounts of Chemical Research* **44**, 666-676 (2011).

119. Freeley, M. et al. Site-Specific One-to-One Click Coupling of Single Proteins to Individual Carbon Nanotubes: A Single-Molecule Approach. *Journal of the American Chemical Society* **139**, 17834-17840 (2017).

120. Sivakumar, K. et al. A fluorogenic 1,3-dipolar cycloaddition reaction of 3-azidocoumarins and acetylenes. *Organic Letters* **6**, 4603-4606 (2004).

121. Lang, K. & Chin, J.W. Bioorthogonal Reactions for Labeling Proteins. *Acs Chemical Biology* **9**, 16-20 (2014).

122. Lang, K. & Chin, J.W. Cellular Incorporation of Unnatural Amino Acids and Bioorthogonal Labeling of Proteins. *Chemical Reviews* **114**, 4764-4806 (2014).

123. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. Green Flourescent Protein as a Marker for Gene-Expression. *Science* **263**, 802-805 (1994).

124. Cubitt, A.B. et al. Understanding, Improving and Using Green FLuorescent Proteins. *Trends in Biochemical Sciences* **20**, 448-455 (1995).

125. Shelby, R.D., Hahn, K.M. & Sullivan, K.F. Dynamic elastic behavior of alpha-satellite DNA domains visualized in situ in living human cells. *Journal of Cell Biology* **135**, 545-557 (1996).

126. Zlokarnik, G. et al. Quantitation of Transcription and Clonal Selection of Single Living Cells With Beta-Lactamase as a Reporter. *Science* **279**, 84-88 (1998).

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

128. Cormack, B.P., Valdivia, R.H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33-38 (1996).

129. Yang, T.T., Cheng, L.Z. & Kain, S.R. Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein. *Nucleic Acids Research* **24**, 4592-4593 (1996).

130. Heim, R., Cubitt, A.B. & Tsien, R.Y. Improved Greene Fluorescence. *Nature* **373**, 663-664 (1995).

131. Shaner, N.C., Steinbach, P.A. & Tsien, R.Y. A Guide to Choosing Fluorescent Proteins. *Nature Methods* **2**, 905-909 (2005).

132. Miyawaki, A. et al. Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature* **388**, 882-887 (1997).

133. Romoser, V.A., Hinkle, P.M. & Persechini, A. Detection in living cells of Ca2+-dependent changes in the fluorescence emission of an indicator composed of two green fluorescent protein variants linked by a calmodulin-binding sequence - A new class of fluorescent indicators. *Journal of Biological Chemistry* **272**, 13270-13274 (1997).

134. Arpino, J.A.J. et al. Structural Basis for Efficient Chromophore Communication and Energy Transfer in a Constructed Didomain Protein Scaffold. *Journal of the American Chemical Society* **134**, 13632-13640 (2012).

135. Acikgoz, S. et al. Photoinduced electron transfer mechanism between green fluorescent protein molecules and metal oxide nanoparticles. *Ceramics International* **40**, 2943-2951 (2014).

136. Dietrich, C.P. et al. An exciton-polariton laser based on biologically produced fluorescent protein. *Science Advances* **2**(2016).

137. Fernandez-Luna, V., Coto, P.B. & Costa, R.D. When Fluorescent Proteins Meet White Light-Emitting Diodes. *Angewandte Chemie-International Edition* **57**, 8826-8836 (2018).

138. Shinobu, A., Palm, G.J., Schierbeek, A.J. & Agmon, N. Visualizing Proton Antenna in a High-Resolution Green Fluorescent Protein Structure. *Journal of the American Chemical Society* **132**, 11093-11102 (2010).

139. Shinobu, A. & Agmon, N. Proton Wire Dynamics in the Green Fluorescent Protein. *Journal of Chemical Theory and Computation* **13**, 353-369 (2017).

140. Brejc, K. et al. Structural basis for dual excitation and photoisomerization of the Aequorea victoria green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2306-2311 (1997).

141. Chattoraj, M., King, B.A., Bublitz, G.U. & Boxer, S.G. Ultra-fast excited state dynamics in green fluorescent protein: Multiple states and proton transfer. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 8362-8367 (1996).

142. Yokoe, H. & Meyer, T. Spatial dynamics of GFP-tagged proteins investigated by local fluorescence enhancement. *Nature Biotechnology* **14**, 1252-1256 (1996).

143. Jablonski, A. Efficiency of anti-stokes fluorescence in dyes. *Nature* **131**, 839-840 (1933).

144. Morandeira, A., Furstenberg, A. & Vauthey, E. Fluorescence quenching in electron-donating solvents. 2. Solvent dependence and product dynamics. *Journal of Physical Chemistry A* **108**, 8190-8200 (2004).

145. Remington, S.J. Fluorescent proteins: maturation, photochemistry and photophysics. *Current Opinion in Structural Biology* **16**, 714-721 (2006).

146. Remington, S.J. Green fluorescent protein: A perspective. *Protein Science* **20**, 1509-1519 (2011).

147. Ormo, M. et al. Crystal structure of the Aequorea victoria green fluorescent protein. *Science* **273**, 1392-1395 (1996).

148. Arpino, J.A.J., Reddington, S.C., Halliwell, L.M., Rizkallah, P.J. & Jones, D.D. Random Single Amino Acid Deletion Sampling Unveils Structural Tolerance and the Benefits of Helical Registry Shift on GFP Folding and Structure. *Structure* **22**, 889-898 (2014).

149. Reddington, S.C. et al. Directed evolution of GFP with non-natural amino acids identifies residues for augmenting and photoswitching fluorescence. *Chemical Science* **6**, 1159-1166 (2015).

150. Crameri, A., Whitehorn, E.A., Tate, E. & Stemmer, W.P.C. Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnology* **14**, 315-319 (1996).

151. Reddington, S.C., Tippmann, E.M. & Jones, D.D. Residue choice defines efficiency and influence of bioorthogonal protein modification via genetically encoded strain promoted Click chemistry. *Chemical Communications* **48**, 8419-8421 (2012).

152. Reddington, S.C. et al. Different Photochemical Events of a Genetically Encoded Phenyl Azide Define and Modulate GFP Fluorescence. *Angewandte Chemie-International Edition* **52**, 5974-5977 (2013).

153. Campanini, B. et al. Role of histidine 148 in stability and dynamics of a highly fluorescent GFP variant. *Biochimica Et Biophysica Acta-Proteins and Proteomics* **1834**, 770-779 (2013).

154. Budisa, N. et al. Probing the role of tryptophans in Aequorea victoria green fluorescent proteins with an expanded genetic code. *Biological Chemistry* **385**, 191-202 (2004).

155. Chin, J.W. et al. An expanded eukaryotic genetic code. *Science* **301**, 964-967 (2003).

156. Jang, S., Sachin, K., Lee, H.J., Kim, D.W. & Lee, H.S. Development of a Simple Method for Protein Conjugation by Copper-Free Click Reaction and Its Application to Antibody-Free Western Blot Analysis. *Bioconjugate Chemistry* **23**, 2256-2261 (2012).

157. Niklaus, L. et al. Micropatterned Down-Converting Coating for White Bio-Hybrid Light-Emitting Diodes. *Advanced Functional Materials* **27**(2017).

158. Press, D.A., Melikov, R., Conkar, D., Firat-Karalar, E.N. & Nizamoglu, S. Fluorescent protein integrated white LEDs for displays. *Nanotechnology* **27**(2016).

159. Zajac, J.M. et al. Time-Resolved Studies of Energy Transfer in Thin Films of Green and Red Fluorescent Proteins. *Advanced Functional Materials* **28**(2018).

160. Korpany, K.V. et al. Conductance Switching in the Photoswitchable Protein Dronpa. *Journal of the American Chemical Society* **134**, 16119-16122 (2012).

161. Ando, R., Mizuno, H. & Miyawaki, A. Regulated fast nucleocytoplasmic shuttling observed by reversible protein highlighting. *Science* **306**, 1370-1373 (2004).

162. Day, R.N. & Davidson, M.W. The fluorescent protein palette: tools for cellular imaging. *Chemical Society Reviews* **38**, 2887-2921 (2009).

163. Patterson, G.H., Knobel, S.M., Sharif, W.D., Kain, S.R. & Piston, D.W. Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophysical Journal* **73**, 2782-2790 (1997).

164. Studier, F.W. Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification* **41**, 207-234 (2005).

165. Hanwell, M.D. et al. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* **4**(2012).

166. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. (2010).

167. Sousa da Silva, A.W. & Vranken, W.F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **5**, 367 (2012).

168. Case, D. et al. AMBER 2015. (University of California, San Fransisco, 2015).

169. Pronk, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845-54 (2013).

170. Gfeller, D., Michielin, O. & Zoete, V. SwissSidechain: a molecular and structural database of non-natural sidechains. *Nucleic Acids Research* **41**, D327-D332 (2013).

171. Alford, R.F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13**, 3031-3048 (2017).

172. Comeau, S.R., Gatchell, D.W., Vajda, S. & Camacho, C.J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Research* **32**, W96-W99 (2004).

173. Kozakov, D. et al. The ClusPro web server for protein-protein docking. *Nature Protocols* **12**, 255-278 (2017).

174. Leaver-Fay, A., et al. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods in Enzymology* **487**, 545-74 (2011).

175. Gasteiger E. et al. *Protein Identification and Analysis Tools on the ExPASy Server*. in *The Proteomics Protocols Handbook* (ed. Walker, J.M.) 571-607 (Humana press, 2005).

176. Winter, G. xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* **43**, 186-190 (2010).

177. Evans, P. Scaling and assessment of data quality. *Acta Crystallographica Section D* **62**, 72-82 (2006).

178. Collaborative. The CCP4 suite: programs for protein crystallography. *Acta Crystallographica Section D* **50**, 760-763 (1994).

179. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D* **60**, 2126-2132 (2004).

180. Long, F. et al. AceDRG: a stereochemical description generator for ligands. *Acta Crystallographica Section D* **73**, 112-122 (2017).

181. Chovancova, E. et al. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. . *PLoS Comput Biol* **8**, e1002708 (2012).

182. Chin, J.W. Expanding and Reprogramming the Genetic Code of Cells and Animals. in *Annual Review of Biochemistry, Vol 83*, Vol. 83 (ed. Kornberg, R.D.) 379-408 (2014).

183. Liu, C.C. & Schultz, P.G. Adding New Chemistries to the Genetic Code. in *Annual Review of Biochemistry, Vol 79*, Vol. 79 (eds. Kornberg, R.D., Raetz, C.R.H., Rothman, J.E. & Thorner, J.W.) 413-444 (2010).

184. Zhang, W.H., Otting, G. & Jackson, C.J. Protein engineering with unnatural amino acids. *Current Opinion in Structural Biology* **23**, 581-587 (2013).

185. Agmon, N. Proton pathways in green fluorescence protein. *Biophysical Journal* **88**, 2452-2461 (2005).

186. Hartley, A.M. Cardiff University (2014).

187. Kim, C.H. et al. Synthesis of Bispecific Antibodies using Genetically Encoded Unnatural Amino Acids. *Journal of the American Chemical Society* **134**, 9918-9921 (2012).

188. Kim, S., Ko, W., Sung, B.H., Kim, S.C. & Lee, H.S. Direct protein-protein conjugation by genetically introducing bioorthogonal functional groups into proteins. *Bioorganic & Medicinal Chemistry* **24**, 5816-5822 (2016).

189. Kosiova I Fau - Kosiova, I., Janicova A Fau - Janicova, A. & Kois P Fau - Kois, P. Synthesis of coumarin or ferrocene labeled nucleosides via Staudinger ligation.

190. Haney, C.M., Wissner, R.F. & Petersson, E.J. Multiply labeling proteins for studies of folding and stability. *Current Opinion in Chemical Biology* **28**, 123-130 (2015).

191. Flemming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of *B. influenzae* (reprinted). *Reviews of Infectious Diseases* **2**, 129-139 (1980).

192. McKay, C.S. & Finn, M.G. Click Chemistry in Complex Mixtures: Bioorthogonal Bioconjugation. *Chemistry & Biology* **21**, 1075-1101 (2014).

193. Niu, W. & Guo, J. Expanding the chemistry of fluorescent protein biosensors through genetic incorporation of unnatural amino acids. *Molecular Biosystems* **9**, 2961-2970 (2013).

194. Bae, J.H. et al. Expansion of the genetic code enables design of a novel "gold" class of green fluorescent proteins. *Journal of Molecular Biology* **328**, 1071-1081 (2003).

195. Wang, F., Niu, W., Guo, J. & Schultz, P.G. Unnatural Amino Acid Mutagenesis of Fluorescent Proteins. *Angewandte Chemie-International Edition* **51**, 10132-10135 (2012).

196. Dippel, A.B. et al. Probing the effectiveness of spectroscopic reporter unnatural amino acids: a structural study. *Acta Crystallographica Section D-Structural Biology* **72**, 121-130 (2016).

197. Vanommeslaeghe, K. et al. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J Comput Chem* **31**(2010).

198. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* **4**(2008).

199. van der Spoel, D., van Maaren, P.J. & Caleman, C. GROMACS Molecule & Liquid Database. *Bioinformatics* **28**(2012).

200. Gordon, M. & Schmidt, M. Advances in electronic structure theory: GAMESS a decade later. in *Theory and Applications of Computational Chemistry: the first forty years* (eds. Dykstra, C., Frenking, G., Kim, K. & Scuseria, G.) 1167-1189 (Elsevier, Amsterdam, 2005).

201. Schmidt, M.W. et al. General Atomic and Molecular Electronic-Structure System. *Journal of Computational Chemistry* **14**, 1347-1363 (1993).

202. Dupradeau, F.-Y. et al. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Physical Chemistry Chemical Physics* **12**, 7821-7839 (2010).

203. Dupradeau, F.Y. et al. R.E.DD.B: A database for RESP and ESP atomic charges, and force field libraries. *Nucleic Acids Res* **36**(2008).

204. RESP ESP charge Derive Home Page. [http://www.q4md-forcefieldtools.org].

205. Brooks, B.R. et al. CHARMM: The Biomolecular Simulation Program. *J Comput Chem* **30**(2009).

206. Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annual Review of Biochemistry* **77**, 363-382 (2008).

207. Pérez, A. et al. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical Journal* **92**(2007).

208. Ribeiro, A.A.S.T., Horta, B.A.C. & de Alencastro, R.B. MKTOP: A program for automatic construction of molecular topologies. *J Braz Chem Soc* **19**(2008).

209. Ryckaert, J.P. & Bellemans, A. Molecular-Dynamics of liquid alkanes. *Faraday Discuss Chem Soc* **66**(1978).

210. Halliwell, L.M. et al. Delta Flucs: Brighter Photinus pyralis firefly luciferases identified by surveying consecutive single amino acid deletion mutations in a thermostable variant. *Biotechnology and Bioengineering* **115**, 50-59 (2018).

211. Der, B.S. et al. Metal-Mediated Affinity and Orientation Specificity in a Computationally Designed Protein Homodimer. *Journal of the American Chemical Society* **134**, 375-385 (2012).

212. Edwards, W.R., Busse, K., Allemann, R.K. & Jones, D.D. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Research* **36**(2008).

213. Edwards, W.R. et al. Regulation of beta-Lactamase Activity by Remote Binding of Heme: Functional Coupling of Unrelated Proteins through Domain Insertion. *Biochemistry* **49**, 6541-6549 (2010).

214. Patterson, D.M. & Prescher, J.A. Orthogonal bioorthogonal chemistries. *Current Opinion in Chemical Biology* **28**, 141-149 (2015).

215. Rostovtsev, V.V., Green, L.G., Fokin, V.V. & Sharpless, K.B. A stepwise Huisgen cycloaddition process: Copper(I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angewandte Chemie-International Edition* **41**, 2596-+ (2002).

216. Dickson, R.M., Cubitt, A.B., Tsien, R.Y. & Moerner, W.E. On/off blinking and switching behaviour of single molecules of green fluorescent protein. *Nature* **388**, 355-358 (1997).

217. Salna, B., Benabbas, A., Sage, J.T., van Thor, J. & Champion, P.M. Wide-dynamic-range kinetic investigations of deep proton tunnelling in proteins. *Nature Chemistry* **8**, 874-880 (2016).

218. van Thor, J.J. Photoreactions and Dynamics of the Green Fluorescent Protein. *Chemical Society Reviews* **38**, 2935-2950 (2009).

219. Cottrell, J.S. Protein identification using MS/MS data. *Journal of Proteomics* **74**, 1842-1851 (2011).

220. Seifert, M.H.J. et al. Backbone Dynamics of Green Fluorescent Protein and the Effect of Histidine 148 Substitution. *Biochemistry* **42**, 2500-2512 (2003).

221. Miller, S., Lesk, A.M., Janin, J. & Chothia, C. The Accessible Surface-Area and Stability of Oligomeric Proteins. *Nature* **328**, 834-836 (1987).

222. Pollegioni, L. et al. Contribution of the dimeric state to the thermal stability of the flavoprotein D-amino acid oxidase. *Protein Science : A Publication of the Protein Society* **12**, 1018-1029 (2003).

223. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**, 774-797 (2007).

224. Woolfson, D.N. et al. De novo protein design: how do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology* **33**, 16-26 (2015).

225. Wan, W. et al. A Facile System for Genetic Incorporation of Two Different Noncanonical Amino Acids into One Protein in Escherichia coli. *Angewandte Chemie-International Edition* **49**, 3211-3214 (2010).

226. Xiao, H. et al. Genetic Incorporation of Multiple Unnatural Amino Acids into Proteins in Mammalian Cells. *Angewandte Chemie-International Edition* **52**, 14080-14083 (2013).

227. Lajoie, M.J. et al. Genomically Recoded Organisms Expand Biological Functions. *Science* **342**, 357-360 (2013).

228. Mandell, D.J. et al. Biocontainment of genetically modified organisms by synthetic protein design. *Nature* **518**, 55-+ (2015).

229. Rovner, A.J. et al. Recoded organisms engineered to depend on synthetic amino acids. *Nature* **518**, 89-+ (2015).

230. Claridge, S.A., Schwartz, J.J. & Weiss, P.S. Electrons, Photons, and Force: Quantitative Single-Molecule Measurements from Physics to Biology. *Acs Nano* **5**, 693-729 (2011).

231. Ishijima, A. & Yanagida, T. Single molecule nanobioscience. *Trends in Biochemical Sciences* **26**, 438-444 (2001).

232. Min, W. et al. Fluctuating enzymes: Lessons from single-molecule studies. *Accounts of Chemical Research* **38**, 923-931 (2005).

233. Ritort, F. Single-molecule experiments in biological physics: methods and applications. *Journal of Physics-Condensed Matter* **18**, R531-R583 (2006).

234. Smiley, R.D. & Hammes, G.G. Single molecule studies of enzyme mechanisms. *Chemical Reviews* **106**, 3080-3094 (2006).

235. Guo, X.F. et al. Covalently bridging gaps in single-walled carbon nanotubes with conducting molecules. *Science* **311**, 356-359 (2006).

236. Guo, X., Gorodetsky, A.A., Hone, J., Barton, J.K. & Nuckolls, C. Conductivity of a single DNA duplex bridging a carbon nanotube gap. *Nature Nanotechnology* **3**, 163-167 (2008).

237. Liu, S. et al. Single-Molecule Detection of Proteins Using Aptamer-Functionalized Molecular Electronic Devices. *Angewandte Chemie-International Edition* **50**, 2496-2502 (2011).

238. Sims, P.C. et al. Electronic Measurements of Single-Molecule Catalysis by cAMP-Dependent Protein Kinase A. *Journal of the American Chemical Society* **135**, 7861-7868 (2013).

239. Willner, I. & Willner, B. Biomaterials integrated with electronic elements: en route to bioelectronics. *Trends in Biotechnology* **19**, 222-230 (2001).

240. Wen, J., Xu, Y., Li, H., Lu, A. & Sun, S. Recent applications of carbon nanomaterials in fluorescence biosensing and bioimaging. *Chemical Communications* **51**, 11346-11358 (2015).

241. Peng, X.H., Chen, J.Y., Misewich, J.A. & Wong, S.S. Carbon nanotube-nanocrystal heterostructures. *Chemical Society Reviews* **38**, 1076-1098 (2009).

242. Arpino, J.A.J., Baldwin, A.J., McGarrity, A.R., Tippmann, E.M. & Jones, D.D. In-Frame Amber Stop Codon Replacement Mutagenesis for the Directed Evolution of Proteins Containing Non-Canonical Amino Acids: Identification of Residues Open to Bio-Orthogonal Modification. *Plos One* **10**(2015).

243. Tan, B., Hodak, M., Lu, W. & Bernholc, J. Charge transport in DNA nanowires connected to carbon nanotubes. *Physical Review B* **92**(2015).

244. Vedala, H. et al. The effect of environmental factors on the electrical conductivity of a single oligo-DNA molecule measured using single-walled carbon nanotube nanoelectrodes. *Nanotechnology* **19**(2008).

245. Weizmann, Y., Chenoweth, D.M. & Swager, T.M. DNA-CNT Nanowire Networks for DNA Detection. *Journal of the American Chemical Society* **133**, 3238-3241 (2011).

246. Ito, M. et al. The Effect of DNA Adsorption on Optical Transitions in Single Walled Carbon Nanotubes. *Journal of Physical Chemistry C* **119**, 21141-21145 (2015).

247. Lyonnais, S. et al. DNA-carbon nanotube conjugates prepared by a versatile method using streptavidin-biotin recognition. *Small* **4**, 442-446 (2008).

248. Nepal, D. & Geckeler, K.E. Proteins and carbon nanotubes: Close encounter in water. *Small* **3**, 1259-1265 (2007).

249. Tu, X., Manohar, S., Jagota, A. & Zheng, M. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature* **460**, 250-253 (2009).

250. Budhathold-Uprety, J., Langenbacher, R.E., Jena, P.V., Roxbury, D. & Heller, D.A. A Carbon Nanotube Optical Sensor Reports Nuclear Entry via a Noncanonical Pathway. *Acs Nano* **11**, 3875-3882 (2017).

251. Landry, M.P. et al. Single-molecule detection of protein efflux from microorganisms using fluorescent single-walled carbon nanotube sensor arrays. *Nature Nanotechnology* **12**, 368-377 (2017).

252. Pastine, S.J. et al. A facile and patternable method for the surface modification of carbon nanotube forests using perfluoroarylazides. *Journal of the American Chemical Society* **130**, 4238-+ (2008).

253. Attanzio, A. et al. Carbon Nanotube-Quantum Dot Nanohybrids: Coupling with Single-Particle Control in Aqueous Solution. *Small* **13**(2017).

254. Palma, M. et al. Controlled Formation of Carbon Nanotube Junctions via Linker-Induced Assembly in Aqueous Solution. *Journal of the American Chemical Society* **135**, 8440-8443 (2013).

255. Zhu, J. et al. Solution-Processable Carbon Nanoelectrodes for Single-Molecule Investigations. *Journal of the American Chemical Society* **138**, 2905-2908 (2016).

256. Rekas, A., Alattia, J.R., Nagai, T., Miyawaki, A. & Ikura, M. Crystal structure of Venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity. *Journal of Biological Chemistry* **277**, 50573-50578 (2002).

257. Necas, D. & Klapetek, P. Gwyddion: an open-source software for SPM data analysis. *Central European Journal of Physics* **10**, 181-188 (2012).

258. Fuentes-Perez, M.E., Dillingham, M.S. & Moreno-Herrero, F. AFM volumetric methods for the characterization of proteins and nucleic acids. *Methods* **60**, 113-121 (2013).

259. Ganesan, S., Ameer-beg, S.M., Ng, T.T.C., Vojnovic, B. & Wouters, F.S. A dark yellow fluorescent protein (YFP)-based Resonance Energy-Accepting Chromoprotein (REACh) for Forster resonance energy transfer with GFP. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 4089-4094 (2006).

260. Kaempgen, M., Lebert, M., Haluska, M., Nicoloso, N. & Roth, S. Sonochemical optimization of the conductivity of single wall carbon nanotube networks. *Advanced Materials* **20**, 616-+ (2008).

261. Riesz, P. & Kondo, T. Free-Radical Fromation Induced by Ultrasound and its Biological Implications. *Free Radical Biology and Medicine* **13**, 247-270 (1992).

262. Bogdanov, A.M. et al. Green fluorescent proteins are light-induced electron donors. *Nature Chemical Biology* **5**, 459-461 (2009).

263. Choi, J.W. et al. Photoinduced electron transfer in GFP/viologen/TCNQ structured hetero-LB film. *Synthetic Metals* **126**, 159-163 (2002).

264. Garcia-Parajo, M.F., Segers-Nolten, G.M.J., Veerman, J.A., Greve, J. & van Hulst, N.F. Real-time light-driven dynamics of the fluorescence emission in single green fluorescent protein molecules. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7237-7242 (2000).

265. Peterman, E.J.G., Brasselet, S. & Moerner, W.E. The fluorescence dynamics of single molecules of green fluorescent protein. *Journal of Physical Chemistry A* **103**, 10553-10560 (1999).

266. Lv, X.X. et al. Ultrafast Photoinduced Electron Transfer in Green Fluorescent Protein Bearing a Genetically Encoded Electron Acceptor. *Journal of the American Chemical Society* **137**, 7270-7273 (2015).

267. Batmunkh, M., Biggs, M.J. & Shapter, J.G. Carbon Nanotubes for Dye-Sensitized Solar Cells. *Small* **11**, 2963-2989 (2015).

268. Hao, F. et al. High Electrocatalytic Activity of Vertically Aligned Single-Walled Carbon Nanotubes towards Sulfide Redox Shuttles. *Scientific Reports* **2**(2012).

269. Hu, L. et al. Light-induced charge transfer in Pyrene/CdSe-SWNT hybrids. *Advanced Materials* **20**, 939-+ (2008).

270. Jeong, S., Shim, H.C., Kim, S. & Han, C.-S. Efficient Electron Transfer in Functional Assemblies of Pyridine-Modified NQDs on SWNTs. *Acs Nano* **4**, 324-330 (2010).

271. Li, C. et al. Photovoltaic Property of a Vertically Aligned Carbon Nanotube Hexagonal Network Assembled with CdS Quantum Dots. *Acs Applied Materials & Interfaces* **5**, 7400-7404 (2013).
272. Sheeney-Haj-Khia, L., Basnar, B. & Willner, I. Efficient generation of photocurrents by using CdS/Carbon nanotube assemblies on electrodes. *Angewandte Chemie-International Edition* **44**, 78-83 (2005).
273. Federspiel, F. et al. Distance Dependence of the Energy Transfer Rate from a Single Semiconductor Nanostructure to Graphene. *Nano Letters* **15**, 1252-1258 (2015).
274. Pakhomov, A.A. & Martynov, V.I. GFP family: Structural insights into spectral tuning. *Chemistry & Biology* **15**, 755-764 (2008).
275. Choi, J.W., Nam, Y.S., Oh, B.K., Lee, W.H. & Fujihira, M. Artificial molecular photodiode consisting of GFP/viologen hetero-thin film. *Synthetic Metals* **117**, 241-243 (2001).
276. Choi, J.-W., Nam, Y.S., Jeong, S.-C., Lee, W.H. & Petty, M.C. Molecular rectifier consisting of cytochrome c/GFP heterolayer by using metal coated optical fiber tip. *Current Applied Physics* **6**, 839-843 (2006).
277. Lee, B. et al. Rectified photocurrent in a protein based molecular photo-diode consisting of a cytochrome b(562)-green fluorescent protein chimera self-assembled monolayer. *Biosensors & Bioelectronics* **19**, 1169-1174 (2004).
278. Chiu, J., March, P.E., Lee, R. & Tillett, D. Site-directed, Ligase-Independent Mutagenesis (SLIM): a single-tube methodology approaching 100% efficiency in 4 h. *Nucleic Acids Research* **32**(2004).
279. Agrawal, D. & Hackenberger, C.P.R. Site-specific chemical modifications of proteins. *Indian Journal of Chemistry Section a-Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry* **52**, 973-991 (2013).
280. Schultz, K.C. et al. A Genetically Encoded Infrared Probe. *Journal of the American Chemical Society* **128**, 13984-13985 (2006).
281. Chiang, C.F., Okou, D.T., Griffin, T.B., Verret, C.R. & Williams, M.N.V. Green fluorescent protein rendered susceptible to proteolysis: Positions for protease-sensitive insertions. *Archives of Biochemistry and Biophysics* **394**, 229-235 (2001).
282. Andreev, Y.A., Kozlov, S.A., Vassilevski, A.A. & Grishin, E.V. Cyanogen bromide cleavage of proteins in salt and buffer solutions. *Analytical Biochemistry* **407**, 144-146 (2010).
283. Tsai, Y.H., Essig, S., James, J.R., Lang, K. & Chin, J.W. Selective, rapid and optically switchable regulation of protein function in live mammalian cells. *Nature Chemistry* **7**, 554-561 (2015).
284. Warner, J.B., Muthusamy, A.K. & Petersson, E.J. Specific Modulation of Protein Activity by Using a Bioorthogonal Reaction. *Chembiochem* **15**, 2508-2514 (2014).
285. Dommerholt, J., Rutjes, F. & van Delft, F.L. Strain-Promoted 1,3-Dipolar Cycloaddition of Cycloalkynes and Organic Azides. *Topics in Current Chemistry* **374**(2016).
286. Ning, X., Guo, J., Wolfert, M.A. & Boons, G.-J. Visualizing metabolically labeled glycoconjugates of living cells by copper-free and fast huisgen cycloadditions. *Angewandte Chemie-International Edition* **47**, 2253-2255 (2008).
287. Zimmerman, E.S. et al. Production of Site-Specific Antibody-Drug Conjugates Using Optimized Non-Natural Amino Acids in a Cell-Free Expression System. *Bioconjugate Chemistry* **25**, 351-361 (2014).
288. Lang, K. et al. Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. *Nature Chemistry* **4**, 298-304 (2012).
289. Shakhnovich, E.I. Proteins With Selected Sequences Fold into Unique Native Conformation. *Physical Review Letters* **72**, 3907-3910 (1994).

290. Kaufmann, K.W., Lemmon, G.H., DeLuca, S.L., Sheehan, J.H. & Meiler, J. Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry* **49**, 2987-2998 (2010).

291. Kozakov, D. et al. How good is automated protein docking? *Proteins-Structure Function and Bioinformatics* **81**, 2159-2166 (2013).

292. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins-Structure Function and Bioinformatics* **78**, 1950-1958 (2010).

293. O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**(2011).

294. Case, D.A. et al. AMBER 2018. in *AMBER* 18 edn (University of California, San Francisco, 2018).

295. Beltrao, P., Bork, P., Krogan, N.J. & van Noort, V. Evolution and functional cross-talk of protein post-translational modifications. *Molecular Systems Biology* **9**, 13 (2013).