Positive functional synergy of structurally integrated artificial protein dimers assembled by Click chemistry

Harley L Worthy‡, Husam Sabah Auhim‡, W. David Jamieson‡, Jacob R. Pope, Aaron Wall, Robert Batchlor, Rachel L. Johnson, Daniel W. Watkins, Pierre Rizkallah, Oliver Castell, D. Dafydd Jones*.

Supplementary Information.

Supplementary Methods.
Gene Sequences

> sfGFP^{E132TAG}

ATG GTT AGC AAA GGT GAA GAA CTG TTT ACC GGC GTT GTG CCG ATT CTG GTG GAA CTG GAT
GGT GAT GTG AAT GGC CAT AAA TTT AGC GTT CGT GGC GAA GGC GAA GGT GAT GCG ACC AAC
GGT AAA CTG ACC CTG AAA TTT ATT TGC ACC ACC GGT AAA CTG CCG GTT CCG TGG CCG ACC
CTG GTG ACC ACC CTG ACC TAT GGC GTT CAG TGC TTT AGC CGC TAT CCG GAT CAT ATG AAA
CGC CAT GAT TTC TTT AAA AGC GCG ATG CCG GAA GGC TAT GTG CAG GAA CGT ACC ATT AGC
TTC AAA GAT GAT GGC ACC TAT AAA ACC CGT GCG GAA GTT AAA TTT GAA GGC GAT ACC CTG
GTG AAC CGC ATT GAA CTG AAA GGT ATT GAT TTT AAA **TAG** GAT GGC AAC ATT CTG GGT CAT
AAA CTG GAA TAT AAT TTC AAC AGC CAT AAT GTG TAT ATT ACC GCC GAT AAA CAG AAA AAT
GGC ATC AAA GCG AAC TTT AAA ATC CGT CAC AAC GTG GAA GAT GGT AGC GTG CAG CTG GCG
GAT CAT TAT CAG CAG AAT ACC CCG ATT GGT GAT GGC CCG GTG CTG CTG CCG GAT AAT CAT
TAT CTG AGC ACC CAG AGC GTT CTG AGC AAA GAT CCG AAT GAA AAA CGT GAT CAT ATG GTG
CTG CTG GAA TTT GTT ACC GCC GCG GGC ATT ACC CAC GGT ATG GAT GAA CTG TAT AAA GGC
AGC CAC CAT CAT CAT CAC CAT TAA

> sfGFP^{H148TAG}

ATG GTT AGC AAA GGT GAA GAA CTG TTT ACC GGC GTT GTG CCG ATT CTG GTG GAA CTG GAT
GGT GAT GTG AAT GGC CAT AAA TTT AGC GTT CGT GGC GAA GGC GAA GGT GAT GCG ACC AAC
GGT AAA CTG ACC CTG AAA TTT ATT TGC ACC ACC GGT AAA CTG CCG GTT CCG TGG CCG ACC
CTG GTG ACC ACC CTG ACC TAT GGC GTT CAG TGC TTT AGC CGC TAT CCG GAT CAT ATG AAA
CGC CAT GAT TTC TTT AAA AGC GCG ATG CCG GAA GGC TAT GTG CAG GAA CGT ACC ATT AGC
TTC AAA GAT GAT GGC ACC TAT AAA ACC CGT GCG GAA GTT AAA TTT GAA GGC GAT ACC CTG
GTG AAC CGC ATT GAA CTG AAA GGT ATT GAT TTT AAA GAA GAT GGC AAC ATT CTG GGT CAT
AAA CTG GAA TAT AAT TTC AAC AGC **TAG** AAT GTG TAT ATT ACC GCC GAT AAA CAG AAA AAT
GGC ATC AAA GCG AAC TTT AAA ATC CGT CAC AAC GTG GAA GAT GGT AGC GTG CAG CTG GCG
GAT CAT TAT CAG CAG AAT ACC CCG ATT GGT GAT GGC CCG GTG CTG CTG CCG GAT AAT CAT
TAT CTG AGC ACC CAG AGC GTT CTG AGC AAA GAT CCG AAT GAA AAA CGT GAT CAT ATG GTG
CTG CTG GAA TTT GTT ACC GCC GCG GGC ATT ACC CAC GGT ATG GAT GAA CTG TAT AAA GGC
AGC CAC CAT CAT CAT CAC CAT TAA

> sfGFP^{H148C}

ATG GTT AGC AAA GGT GAA GAA CTG TTT ACC GGC GTT GTG CCG ATT CTG GTG GAA CTG GAT
GGT GAT GTG AAT GGC CAT AAA TTT AGC GTT CGT GGC GAA GGC GAA GGT GAT GCG ACC AAC
GGT AAA CTG ACC CTG AAA TTT ATT TGC ACC ACC GGT AAA CTG CCG GTT CCG TGG CCG ACC
CTG GTG ACC ACC CTG ACC TAT GGC GTT CAG TGC TTT AGC CGC TAT CCG GAT CAT ATG AAA
CGC CAT GAT TTC TTT AAA AGC GCG ATG CCG GAA GGC TAT GTG CAG GAA CGT ACC ATT AGC

TTC AAA GAT GAT GGC ACC TAT AAA ACC CGT GCG GAA GTT AAA TTT GAA GGC GAT ACC CTG
GTG AAC CGC ATT GAA CTG AAA GGT ATT GAT TTT AAA GAA GAT GGC AAC ATT CTG GGT CAT
AAA CTG GAA TAT AAT TTC AAC AGC **TGT** AAT GTG TAT ATT ACC GCC GAT AAA CAG AAA AAT
GGC ATC AAA GCG AAC TTT AAA ATC CGT CAC AAC GTG GAA GAT GGT AGC GTG CAG CTG GCG
GAT CAT TAT CAG CAG AAT ACC CCG ATT GGT GAT GGC CCG GTG CTG CTG CCG GAT AAT CAT
TAT CTG AGC ACC CAG AGC GTT CTG AGC AAA GAT CCG AAT GAA AAA CGT GAT CAT ATG GTG
CTG CTG GAA TTT GTT ACC GCC GCG GGC ATT ACC CAC GGT ATG GAT GAA CTG TAT AAA GGC
AGC CAC CAT CAT CAT CAC CAT TAA

> sfGFP^{Q204TAG}

ATG GTT AGC AAA GGT GAA GAA CTG TTT ACC GGC GTT GTG CCG ATT CTG GTG GAA CTG GAT
GGT GAT GTG AAT GGC CAT AAA TTT AGC GTT CGT GGC GAA GGC GAA GGT GAT GCG ACC AAC
GGT AAA CTG ACC CTG AAA TTT ATT TGC ACC ACC GGT AAA CTG CCG GTT CCG TGG CCG ACC
CTG GTG ACC ACC CTG ACC TAT GGC GTT CAG TGC TTT AGC CGC TAT CCG GAT CAT ATG AAA
CGC CAT GAT TTC TTT AAA AGC GCG ATG CCG GAA GGC TAT GTG CAG GAA CGT ACC ATT AGC
TTC AAA GAT GAT GGC ACC TAT AAA ACC CGT GCG GAA GTT AAA TTT GAA GGC GAT ACC CTG
GTG AAC CGC ATT GAA CTG AAA GGT ATT GAT TTT AAA GAA GAT GGC AAC ATT CTG GGT CAT
AAA CTG GAA TAT AAT TTC AAC AGC CAT AAT GTG TAT ATT ACC GCC GAT AAA CAG AAA AAT
GGC ATC AAA GCG AAC TTT AAA ATC CGT CAC AAC GTG GAA GAT GGT AGC GTG CAG CTG GCG
GAT CAT TAT CAG CAG AAT ACC CCG ATT GGT GAT GGC CCG GTG CTG CTG CCG GAT AAT CAT
TAT CTG AGC ACC **TAG** AGC GTT CTG AGC AAA GAT CCG AAT GAA AAA CGT GAT CAT ATG GTG
CTG CTG GAA TTT GTT ACC GCC GCG GGC ATT ACC CAC GGT ATG GAT GAA CTG TAT AAA GGC
AGC CAC CAT CAT CAT CAC CAT TAA

> Venus^{H148TAG}

ATG CGG GGT TCT CAT CAT CAT CAT CAT CAT GGT ATG GCT AGC ATG ACT GGT GGA CAG CAA
ATG GGT CGG GAT CTG TAC GAG AAC CTG TAC TTC CAG GGC TCG AGC ATG GTG AGC AAG GGC
GAG GAG CTG TTC ACC GGG GTG GTG CCC ATC CTG GTC GAG CTG GAC GGC GAC GTA AAC GGC
CAC AAG TTC AGC GTG TCC GGC GAG GGC GAG GGC GAT GCC ACC TAC GGC AAG CTG ACC CTG
AAG CTG ATC TGC ACC ACC GGC AAG CTG CCC GTG CCC TGG CCC ACC CTC GTG ACC ACC CTG
GGC TAC GGC CTG CAG TGC TTC GCC CGC TAC CCC GAC CAC ATG AAG CAG CAC GAC TTC TTC
AAG TCC GCC ATG CCC GAA GGC TAC GTC CAG GAG CGC ACC ATC TTC TTC AAG GAC GAC GGC
AAC TAC AAG ACC CGC GCC GAG GTG AAG TTC GAG GGC GAC ACC CTG GTG AAC CGC ATC GAG
CTG AAG GGC ATC GAC TTC AAG GAG GAC GGC AAC ATC CTG GGG CAC AAG CTG GAG TAC AAC
TAC AAC AGC **TAG** AAC GTC TAT ATC ACC GCC GAC AAG CAG AAG AAC GGC ATC AAG GCC AAC
TTC AAG ATC CGC CAC AAC ATC GAG GAC GGC GGC GTG CAG CTC GCC GAC CAC TAC CAG CAG
AAC ACC CCC ATC GGC GAC GGC CCC GTG CTG CTG CCC GAC AAC CAC TAC CTG AGC TAC CAG
TCC GCC CTG AGC AAA GAC CCC AAC GAG AAG CGC GAT CAC ATG GTC CTG CTG GAG TTC GTG
ACC GCC GCC GGG ATC ACT CTC GGC ATG GAC GAG CTG TAC AAG TAA

> Venus^{Q204TAG}

ATG CGG GGT TCT CAT CAT CAT CAT CAT CAT GGT ATG GCT AGC ATG ACT GGT GGA CAG CAA
ATG GGT CGG GAT CTG TAC GAG AAC CTG TAC TTC CAG GGC TCG AGC ATG GTG AGC AAG GGC
GAG GAG CTG TTC ACC GGG GTG GTG CCC ATC CTG GTC GAG CTG GAC GGC GAC GTA AAC GGC
CAC AAG TTC AGC GTG TCC GGC GAG GGC GAG GGC GAT GCC ACC TAC GGC AAG CTG ACC CTG
AAG CTG ATC TGC ACC ACC GGC AAG CTG CCC GTG CCC TGG CCC ACC CTC GTG ACC ACC CTG
GGC TAC GGC CTG CAG TGC TTC GCC CGC TAC CCC GAC CAC ATG AAG CAG CAC GAC TTC TTC
AAG TCC GCC ATG CCC GAA GGC TAC GTC CAG GAG CGC ACC ATC TTC TTC AAG GAC GAC GGC
AAC TAC AAG ACC CGC GCC GAG GTG AAG TTC GAG GGC GAC ACC CTG GTG AAC CGC ATC GAG
CTG AAG GGC ATC GAC TTC AAG GAG GAC GGC AAC ATC CTG GGG CAC AAG CTG GAG TAC AAC
TAC AAC AGC CAC AAC GTC TAT ATC ACC GCC GAC AAG CAG AAG AAC GGC ATC AAG GCC AAC
TTC AAG ATC CGC CAC AAC ATC GAG GAC GGC GGC GTG CAG CTC GCC GAC CAC TAC CAG CAG

AAC ACC CCC ATC GGC GAC GGC CCC GTG CTG CTG CCC GAC AAC CAC TAC CTG AGC TAC **TAG**
TCC GCC CTG AGC AAA GAC CCC AAC GAG AAG CGC GAT CAC ATG GTC CTG CTG GAG TTC GTG
ACC GCC GCC GGG ATC ACT CTC GGC ATG GAC GAG CTG TAC AAG TAA

Protein production: sfGFP variants.

The sfGFP H148TAG, E132TAG and Q204TAG [2] mutants were constructed previously[3, 4]. The sfGFP^{H148C} mutant was constructed in a similar manner by PCR PCR-based site-directed mutagenesis of the sfGFP template gene using the primers 5'-TTCAACAGC**TGT**AATGTGTATATTACCG-3' and 5'-ATTATATTCCAGTTTATGACCCAGAATGTTGC-3'

Super-folder GFP (sfGFP) mutant plasmids (based on the pBAD vector) sfGFP^{Q204TAG} and sfGFP^{H148TAG} (gene sequence above) were co-transformed by electroporation into *E. coli* Top10 cells (Invitrogen) with either pDULE-cyanoRS (p-azido-L-phenylalanine [azF] incorporation) [5] or pEVOL-SCO (s-cyclooctyne-L-lysine [SCO] incorporation) [6]. The transformed cells were used to inoculate 1L flasks of autoinduction media according to the recipe defined in Studier *et al.* [7] and supplemented with 50 µg/mL carbenicillin and either, 25 µg/mL tetracycline or 35 µg/mL chloramphenicol dependant on whether expressing protein incorporating azF or SCO, respectively. Cultures were grown overnight at 37 °C in a shaking incubator. After 1 hour of growth cultures were inoculated with appropriate non-canonical amino acid to a final concentration of 0.5 mM. Cultures containing azF were kept in the dark until after dimerisation with the SCO-containing protein. A similar procedure was used to produce sfGFP^{H148C} but without transformation with the non-canonical amino acids incorporation plasmids or growth in the presence of the non-canonical amino acids.

Cells were harvested via centrifugation at 5000 xg for 20 mins. The supernatant was discarded and cells resuspended in 20 mL of 50 mM Tris-HCl pH8.0, 300 mM NaCl, 20 mM imidazole. The cells were lysed using a French press and the resulting lysate was clarified by centrifugation at 25,000 xg for at least 30 minutes. Cell lysates were then loaded onto a 5 mL HisTrapHP™ (GE Healthcare) equilibrated in lysis buffer. Bound GFP was eluted by washing the column in 250 mM Imidazole. Samples were then loaded onto a Superdex 75 column equilibrated in 50 mM Tris-HCl pH8.0 and purity was checked via SDS-PAGE analysis. Concentrations of monomer variants were determined using the Bio-RAD DC Protein Assay using wild type sfGFP^{WT} as a standard and correlated to the 280 nm absorbance.

Protein production: Venus variants.

The plasmid housing Venus (based on the pBAD vector and procured from Addgene) was used to prepare the Venus variant H148TAG (gene sequence above) via site-directed mutagenesis using Phusion HF polymerase (Finnzymes, Loughborough, Leicestershire). The primer pair, Venus148 F(AACAGCTAGAACGTCTATATCACC) and Venus148 R(GTAGTTGTACTCCAGCTTGTGC) were used. Venus was co-transformed by electroporation into *E. coli* Top10 cells with pDULE-cyanoRS (p-azido-L-phenylalanine [azF] incorporation). The transformed cells were used to inoculate 1L flasks of LB media supplemented with 100 µg/mL ampicillin, 25 µg/mL tetracycline and 0.1 mM of azF. Cultures were grown for 1 hour at 37 °C in a shaking incubator before

Deleted: ¶
In silico modelling of Super-folder GFP (**sfGFP) dimer interfaces.** ¶
ClustPro is a global docking rigid-body approach that requires no prior information on interface regions, and has been shown to be a good predictor of dimer interfaces [1]. ClusPro generates ~100,000 structures and scores them using balanced energy coefficients as described by Kozakov et al [1](Eq 1). 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.¶
¶
Eq 1: $E = 0.40E_{rep} + -0.40E_{att} + 600Ee_{lec} + 1.00E_{DARS}$¶
¶
The server then takes the 1000 models with the lowest scores and clusters them using pairwise to generate I-RMSD (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 becomes 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 force-field for 300 steps with fixed backbone to minimise steric clashes.¶

Deleted:

Deleted: GFP

Deleted:

Formatted: Superscript

Deleted: wt GFP

Formatted: Superscript

expression was induced by addition of 0.1% of arabinose and incubated for 24 hours at 25°C. Cultures were kept in the dark until after dimerisation with SCO.

Cells were harvested via centrifugation at 5000 xg for 20 mins. The supernatant was discarded and cells resuspended in 20 mL of 50 mM Tris-HCl pH8.0, 1 mM EDTA. The cells were lysed using a French press and the resulting lysate was clarified by centrifugation at 25,000 xg for at least 30 minutes. Cell lysates were then loaded onto a Protino^{R} Ni-TED 2000 Packed Columns (Machery-Nagel, Germany) equilibrated in equilibration-wash buffer (50 mM Na H_2PO_4, 300 mM NaCl, pH 8) then allowed to drain by gravity. Bound Venus was eluted with 3 bed volumes of elution buffer (50 mM Na H_2PO_4, 300 mM NaCl, 250 mM imidazole, pH 8). Samples were then loaded onto a Superdex 75 column equilibrated in 50 mM Tris-HCl pH8.0 and purity was checked via SDS-PAGE analysis. Concentrations of monomer variants were determined using the Bio-RAD DC Protein Assay using wild type sfGFP (sfGFP^{WT}) as a standard and correlated to the 280 nm absorbance.

Molar Extinction coefficient determination

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 250-600nm at a rate of 300 nm/min at 1 nm intervals. Extinction coefficients were calculated by diluting proteins down to 10 µM (5 µM dimers) and recording full absorption spectra from 250-600 nm. Absorption and concentration values were then substituted into a rearranged version of the Beer-Lambert law (Eq 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).

Eq 2: $\varepsilon = \frac{A}{cl}$

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 (λex) 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 (λem) over a range of wavelength down to 350 nm at 1 nm intervals. For purified sfGFP and Venus variants, protein solution was diluted to 0.5 µM in 50 mM Tris HCl pH 8.0 with exception of dimer fluorescence which were recorded at 0.25 µM.

Single molecule imaging and data processing

Single molecule imaging was performed using a custom built total internal reflection fluorescence (TIRF) microscope based on a Nikon Ti-U inverted microscope and Andor iXon ultra 897 EMCCD camera. Illumination was provided by a Venus 473nm DPSS laser with a power output of 100mW. Laser coupling into the microscope was

achieved via a custom built optical circuit (components were sourced from Thorlabs, Chroma and Semrock) followed by a single mode fibre-optic launch. Laser power at the microscope stage averaged at 5.8μW/μm^2. The total internal reflection illumination angle was generated using a combination of fibre-optic micro-positioning and a high numerical aperture TIRF objective (Nikon, CFI Apochromat TIRF 60X oil, NA1.49). The Excitation and fluorescence emission wavelengths were separated using a dichroic mirror with a 488nm edge (Chroma zt488rdc-xr). Emitted wavelengths were further filtered using a 500nm edge long pass filter (Chroma hhq500lp) and a 500-550nm band pass filter (Chroma et525/50m). Acquisitions were controlled using the Andor Solis software package. Frame exposure times were set to 60ms and an EM gain of 250 was used. Coverslips used for TIRF imaging underwent oxygen plasma treated to remove fluorescent contaminants prior to use. Protein solutions were diluted to concentrations suitable for single molecule measurements before droplets were placed onto coverslips for imaging.
Single molecule imaging data was processed and analysed using ImageJ[8] and Matlab (R2017a) (MathWorks U.S.A.). 32 bit floating point TIFF image stacks were used throughout. The first acquisition frame was removed from all image sequences to account for latency of shutter opening by the camera TTL trigger. All images were processed to normalise for spatial variation in intensity profile of the laser illumination using a reference image look-up of relative spatial illumination intensity, mapping the laser illumination created from a Gaussian blurred (20 pixel radius) median z-projection of a fluorescent image stack. The resulting image stack was then corrected for temporal laser intensity fluctuations to minimise the noise in extracted traces. This was achieved by quantifying fluctuations in the global image background and scaling the corresponding frame accordingly, relative to the mean. Practically, this was achieved by removing bright fluorescent spots, defined as any pixel with an intensity greater than 0.05 standard deviations above the median pixel intensity of that frame. Identified pixels were assigned a value equal to the median pixel intensity, effectively erasing them to give a background only image stack. Each frame was scaled relative to the mean intensity of all frames (all pixels) and used to create a temporal lookup table of relative frame to frame laser power fluctuations. This enabled correction of the main image stack. Background counts were subtracted by the pixel-wise subtraction of time averaged median pixel intensity of a background region of interest. Spots were detected using the ImageJ plugin trackmate [9], integrated in the FIJI[10] distribution of ImageJ. Detection was used as a means of automatically identifying spots and removing distinct "off" states which due to their abundance can mask peaks within intensity distributions. These dark states occur either as a result of photo-bleaching or as part of a natural fluorophore blinking phenomenon. Trackmate detects spots occurring above a background threshold thus spots which are either photobleached or existing in a dark state for the total duration of any given frame are not included in the detection process. An estimated spot diameter of 4 pixels was applied with a difference of Gaussian (DoG) detection routine. This applies differently sized Gaussian blurs (greater or lesser than the estimated spot diameter) to two copies of each frame which are then subtracted from one another. This process acts as a spatial bandpass filter enhancing features in the range of the estimated spot diameter enabling detection. As spots were static, linking was performed using spot linking and gap closing distances of 1 pixel. A frame gap closing distance of 3 was also used to link spots displaying long "off" states. Data was exported to matlab where a Gaussian mixture model was fitted to the logarithm of resultant frequency of intensity values for all spots of a given dimer. To assess the

suitability of fitting, four gaussian mixture models with components ranging from 1-4 were fit to both sfGFP^{148x2} and sfGFP^{WT} (1000 replicates/model). The mean Akaike information criteria was calculated for each model with the minimal value indicative of the most probable fit.In addition spatial coordinates of tracked spots were used to generate representative traces from corrected stacks. For each dimer, data sets consisted of two separate acquisitions 24 seconds (400 frames) in length amassing information from ~200 dimer pairs each.

Protein structure determination

Samples of sfGFP^{148x2} were concentrated to 10 mg/mL using spin concentrators (10,000 Da mW cut-off). Crystallisation trays were set up using either JCB or PACT pre-made crystallography screens. Trays were monitored regularly to check for crystal formation.

Several crystals grew in condition C6 of the PACtT crystallography screen (0.1M PCTP [pH 9.0], 25% PEG 1500).These crystals were collected and taken to Diamond light Source (Harwell, UK) for X-ray diffraction measurement. Data were reduced using the XIA2 package [11] assigned a space group using POINTLESS [12], scaled using SCALA [12] and merged using TRUNCATE [13]. Structures were solved by molecular replacement with PHASER, using a previously determined sfGFP structure (PDB code 2B3P). Structures were then adjusted manually using COOT [14] and refined by TLS restrained refinement using RefMac [15]. All the above programs were accessed via the CCP4 package (http://www.ccp4.ac.uk/) [13].

Mass Spectrometry

Protein samples were buffer exchanged into fresh 50 mM Tris-HCl pH 8.0 and diluted to 10 µM, for mass spectrometry analysis. 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 was processed using MassLynx 4.1 programme 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.

Supplementary Tables

Table S1. Statistics for *in silico* modelling of sfGFP dimer interfaces.

Model [a]	Total energy (kJ/mole)	Interface Energy (kJ/mole)	I-RSMD ($Å^2$)	RMSD GFP^{148x2} ($Å^2$) [b]
Model 5	-503.94	-13.434	1.755	**4.72**
Model 1	-501.966	-4.192	0.278	9.53
Model 4	-497.071	-6.993	0.397	18.96
Model 2	-497.112	-6.375	0.56	11.95
Model 3	-494.492	-3.079	0.195	8.31

[a] models ordered according to their rank. [b] compared to the determined structure

Table S2: Crystallographic statistics for sfGFP^{148x2}

	GFP^{148x2}
PDB ID	5NHN
Wavelength (Å)	0.979
Beamline	Diamond IO4
Space group	$P6_5$
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)
CC1/2	1.000 (0.680)
R(merge)[a] (%) (last shell)	7.9 (68.8)
B(iso) from Wilson (Å^2)	41.03
B(iso) from refinement	50.8
Log Likelihood Coordinate rms	0.126
Non-H atoms	3877
Solvent molecules	226
R-factor[b] (%)	18.2
R-free[c] (%)	21.1
Rmsd bond lengths (Å)	0.015
Rmsd bond angles (°)	1.868
Core region (%)	98.42
Allowed region (%)	1.13
Additionally allowed region (%)	0
Disallowed Region (%)	0.46

Deleted: x

Deleted: 2

Table S3. Spectral properties of sfGFP and Venus variants.

Variant	λ_{max} (nm)	λ_{EM} (nm)	ε ($M^{-1}cm^{-1}$)	QY	Brightness
$sfGFP^{WT}$	485	511	49000[a]	0.75[a]	36750
$Venus^{WT}$	515	528	92200[b]	0.65	59930
$sfGFP^{148SCO}$	395	511	31000	0.52	16120
	492	511	17300	0.84	14532
$Venus^{148azF}$	517	525	30100	0.45	13545
$GFVen^{148}$	400	517	24000	0.42	10080
	505	517	96000	0.46	44160
$Venus^{204azF}$	515	528	87600	0.42	36792
$sfGFP^{204SCO}$	485	511	39800	0.66	26268
$GFVen^{204}$	492	530	102000	0.70	71400
	514	530	125000	0.60	75000

[a] We have reported previously a significant shortfall in the molar absorbance coefficient we routinely calculate (here and [2-4]) and that published by Pedelacq et al [16]. [b] Published previously by Nagai et al [17] and measured in the current study as 95000 $M^{-1}cm^{-1}$. Given that our value is close to the reported value, we have used the reported value.

Supplementary Figures

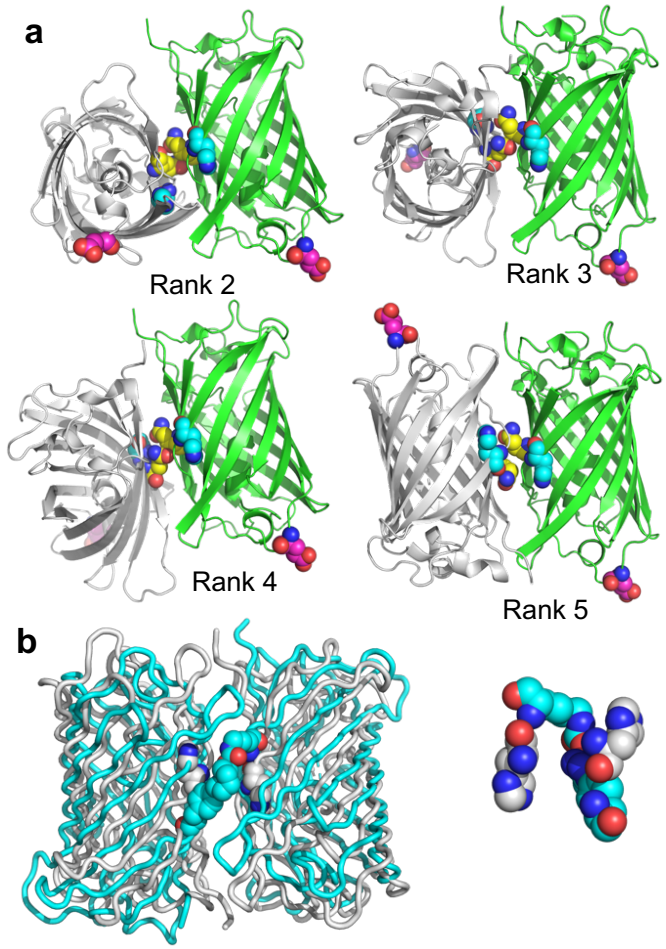

Figure S1. Models of sfGFP dimerisation. (a) The 2nd to 5th ranked models of sfGFP dimerisation. The top ranked model is shown in the main text. Ranking was performed as described in the main text experimental section. Statistics are shown in Table S1. The Glu132, His148 and Gln204 are coloured magenta, cyan and yellow, respectively. The reference sfGFP structure is coloured green. (b) Overlay of sfGFP^{148x2} structure (cyan) with the closest model (grey, model rank 1st), with a calculated RMSD of 4.7 Å. Residue 148 for both chains of the model and sfGFP^{148x2} are also shown separated as spheres.

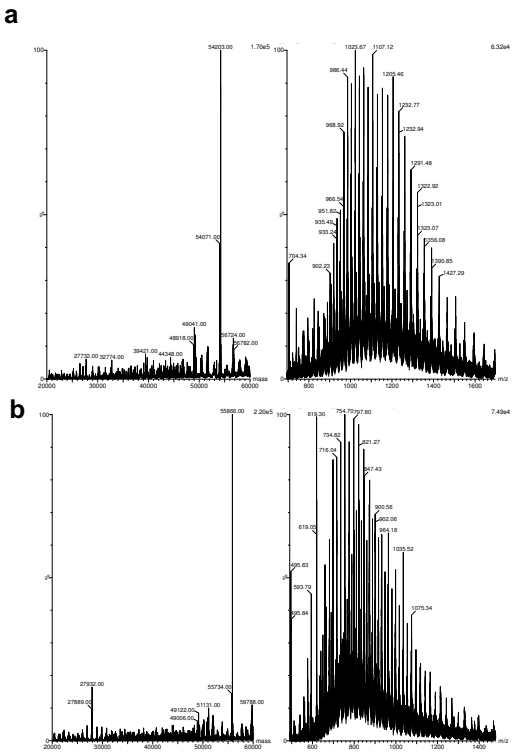

Figure S2. sfGFP dimer formation. (a) Mass spectrum of the sfGFP^{148x2} dimer. The theoretical molecular weight for full length dimerised protein is 55846 Da. The observed mass (54203 Da) matches the loss of the His tag from each monomer (823 Da x 2 = 1646 Da; 54200 Da). (b) Mass spectrum of the sfGFP^{204x2} dimer. The theoretical molecular weight for full length dimerised protein is 55864 Da, with a mass of 55866 Da observed.

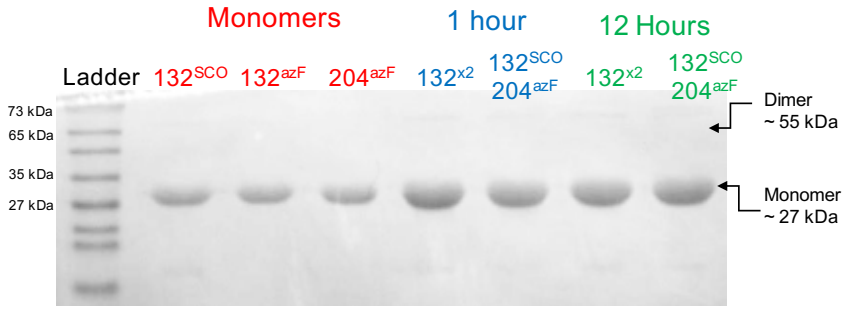

Figure S3. Dimerisation potential of a non-dimer interface residue, as predicted by *in silico* modelling. The residue predicted not to form part of the interface is Glu132 of sfGFP. The SCO ncAA was incorporated at residue 132 (132^{SCO}) of GFP and dimerisation with azF incorporated at either residue 132 (132^{azF}) or 204 (204^{azF}) of sfGFP. No clear dimerisation product was observed for the 132^{azF}-132^{SCO} (132^{x2}) or 132^{SCO}-204^{azF} by gel mobility shift assay.

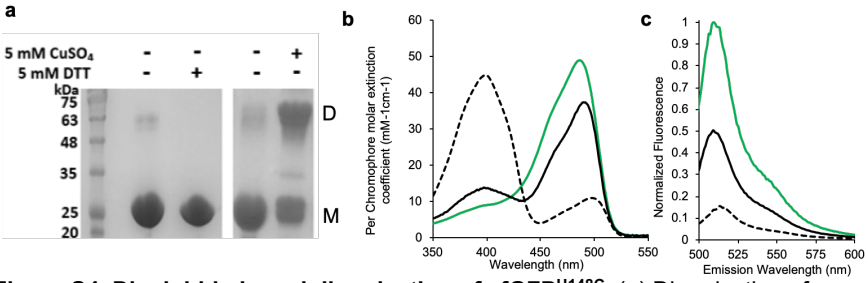

Figure S4. Disulphide-based dimerisation of $sfGFP^{H148C}$. (a) Dimerisation of GFP^{H148C} as analysed by non-reducing SDS PAGE gel mobility shift assay. The monomer has an estimated mass of ~27 kDa and the dimer ~ 55 kDa. (b) Absorbance spectra of $sfGFP^{H148C}$ monomer (dashed black line), GFP^{H148C} dimer (black line) and $sfGFP^{WT}$ (green line). The molar absorbance values have been normalised to a per chromophore basis for comparison. (c) Fluorescence emission spectra on excitation at 490 nm. Fluorescence spectra for monomeric GFP^{WT} (green) and $sfGFP^{H148C}$ (dashed black line) were measured using 0.5 μM protein and dimeric $sfGFP^{H148C}$ (solid black line) using 0.25 μM. Spectra were normalised to the GFP^{WT} values. **Technical note.** $sfGFP^{H148C}$ was initially purified as a monomer under reducing conditions maintained by the presence of 5 mM DTT. Dimerisation was carried out at a monomer concentration of 50 μM in 50 mM Tris buffer (pH8.0) at room temperature. Where stated above, 5 mM CuSO4 was added to the reaction buffer as an oxidizing agent to induce disulphide formation. Dimerisation was confirmed by SDS-PAGE (see above) and the dimers formed were purified from the monomeric species by size exclusion chromatography using a Superdex 75 column, equilibrated with 50 mM Tris (pH 8.0).

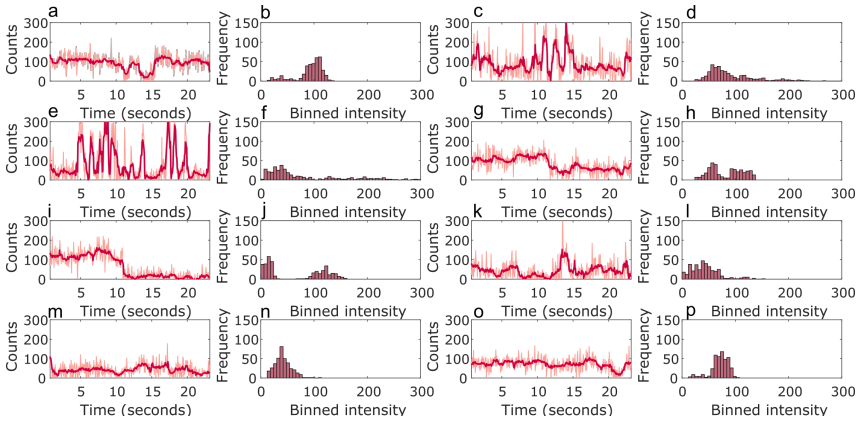

Figure S5. Representative sample of sfGFP^{148x2} single molecule time course traces (raw and Cheung-Kennedy filtered) coupled with paired intensity frequency histograms (generated from Cheung Kennedy filtered data) to the right of each trace. Traces highlight the complexity of behaviour demonstrated by dimers at the single molecule level, showing a range of fluorescence states, transitions and on times. Histograms show no well defined or recurring intensity peaks emphasising the inherent intensity variability of sfGFP^{148x2} in contrast to sfGFP^{WT}.

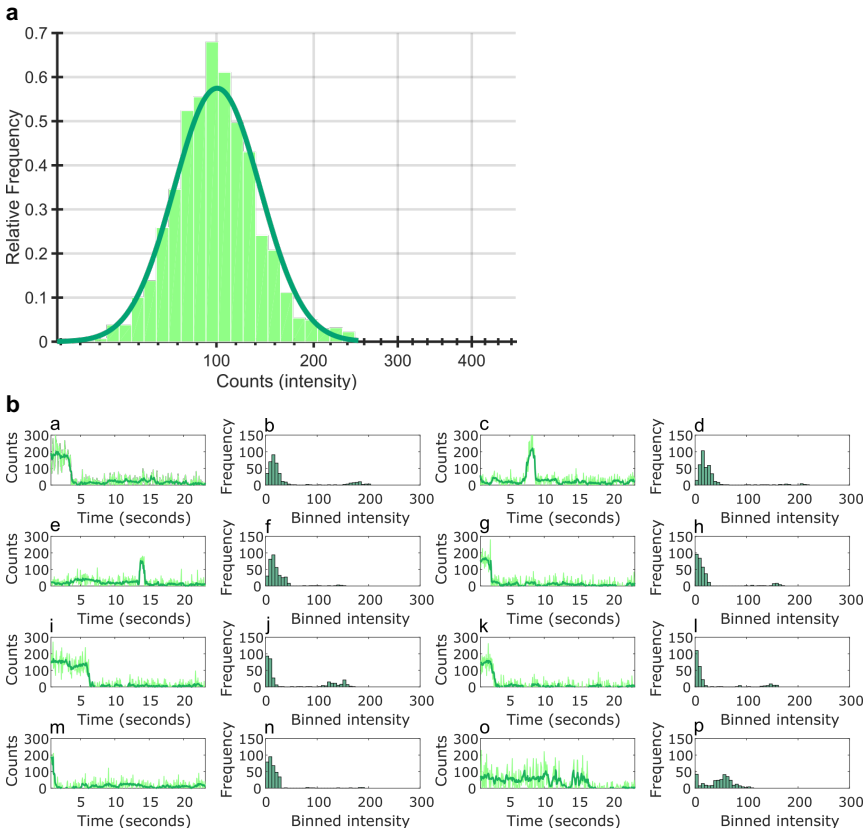

Figure S6. (a) A single molecule fluorescence intensity histogram for sfGFP^{WT} consisting of 204 trajectories (2244 spots). The histogram data fits to a single log normal distribution centred around 100 counts. (b) Representative sample of monomeric sfGFP^{WT} single molecule time course traces (raw and Cheung-Kennedy filtered) coupled with paired intensity frequency histograms (generated from Cheung Kennedy filtered data) to the right of each trace. Common intensity states are predominant in traces which often contain single clear-cut transitions alongside long lived dark states. In the majority of traces, transition to a dark state occurs after a relatively short period of time. Histograms generally show clear separation between baseline and intensity peaks which commonly arise between counts of 100-200.

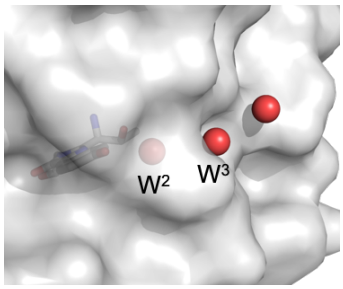

Figure S7. Extended water molecule (red balls) network from CRO (sticks) to surface in sfGFP^{WT}. Water molecule W2 and W3 that are also present in sfGFP^{148x2} (Figure 4 in main manuscript) are indicated on the figure together with an additional associated water molecule. Only W2 is partially buried while the other two water molecules are potentially free to exchange with the bulk solvent. GFP is shown in surface representation.

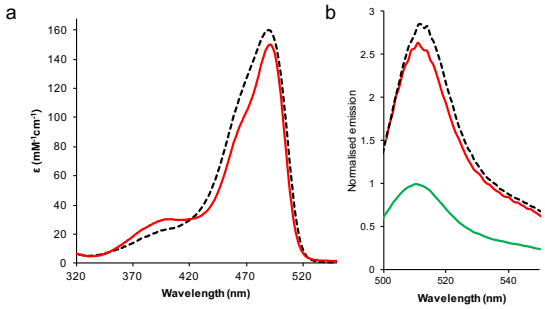

Figure S8. Comparison of the sfGFP^{148x2} (red line), sfGFP^{204x2} (dashed line) and sfGFP^{WT} (green line). (a) Absorbance spectra of the two dimer species. (b) Emission spectra of two dimers and monomeric sfGFP^{WT} (0.5 μM, excitation at λ_{max} (Table 1)). Fluorescence is normalised to the sfGFP^{WT} intensity.

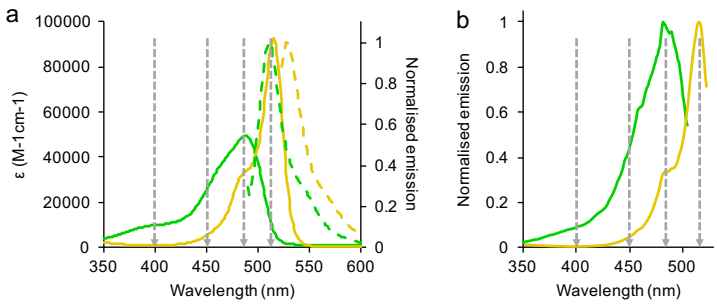

Figure S9. Comparison of the sfGFP (green) and Venus (gold) absorbance and fluorescence spectra. (a) absorbance (solid line) and emission spectra (dashed line). (b) Excitation spectra (on monitoring emission at λ_{EM}; Table S1). The grey dashed arrows indicate wavelengths used to monitor communication between two monomers in GFVen dimers.

```
Venus          -35 MRGSHHHHHHGMASMTGGQQMGRDLYENLYFQGSSMVSKGEELFTGVVPI     14
                                                   X|||||||||||||
sfGFP                                            1-MSKGEELFTGVVPI     14

Venus           15 LVELDGDVNGHKFSVSGEGEGDATYGKLTLKLICTTGKLPVPWPTLVTTL     64
                   ||||||||||||||X|||||||X||||||X||||||||||||||||||
sfGFP           15 LVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLVTTL     64

Venus           65 GYGLQCFARYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKF     114
                   X||X|||X|||||||X|||||||||||||||||||X|||||X|||||||||
sfGFP           65 TYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKF     114

Venus          115 EGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYITADKQKNGIKAN     164
                   ||||||||||||||||||||||||||||||X|||||||||||||||||||
sfGFP          115 EGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKAN     164
                                                   ^
Venus          165 FKIRHNIEDGGVQLADHYQQNTPIGDGPVLLPDNHYLSYQSALSKDPNEK     214
                   ||||||X|||X|||||||||||||||||||||||||X||X|||||||||
sfGFP          165 FKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEK     214

Venus          215 RDHMVLLEFVTAAGITLGMDELYK--------     274
                   ||||||||||||||X|||||||
sfGFP          215 RDHMVLLEFVTAAGITHGMDELYKLEHHHHHH     246
```

Figure S10. Sequence alignment between the versions of Venus and GFP used in the current study. The mutated H148 is shown in bold. Blue, red and green highlighted residues correspond to His tags, TEV protease cleavage motif and CRO forming residue, respectively. Residues that differ between the two are highlighted by a X.

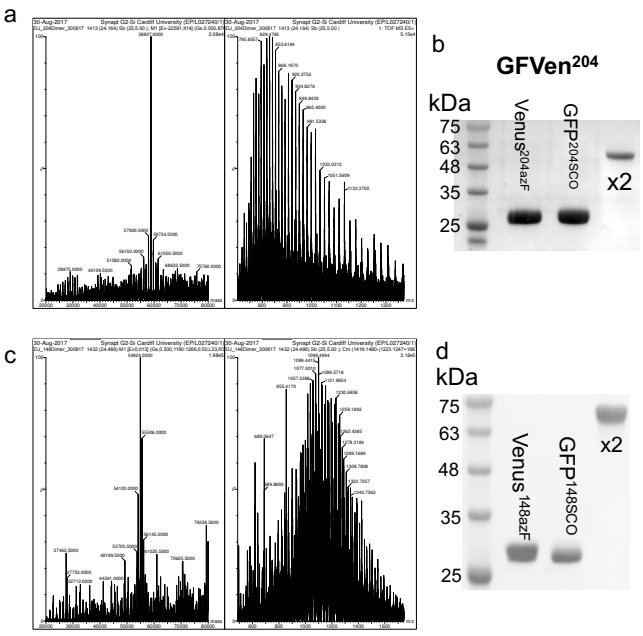

Figure S11. Dimerisation of GFVen dimers. (a) GFVen^{204}. The calculated mass of GFVen^{204} was 58820 Da (Venus^{204azF}, 30843 Da; sfGFP^{204SCO}, 27977 Da). The measured mass was 58827 Da, a difference of +7 Da (0.012% difference). (b) SDS PAGE analysis of GFVen^{204} dimerisation. (c) Mass spectra of GFVen^{148}. The major peak at 54924 Da corresponds to GFP^{204SCO} (27698 Da) and Venus^{148azF} with a truncated N-terminal extension (up to -4-GSSM in Figure S10; 26956 Da), with has a calculated molecular mass of 54924 Da. The second smaller peak at 55506 Da, corresponds to sfGFP^{204SCO} and Venus^{148azF} minus the N-terminal extension up (to -7-YFQG; 27539Da), with a calculated mass of 55507 Da. The third minor peak at 54103 Da corresponds to sfGFP^{148SCO} with the loss of its C-terminal His-tag (27145 Da) and Venus^{148azF} with the N-terminal extension truncated (up to -4-GSSM in Figure S13; 26956 Da), which has a calculated molecular mass of 54101 Da. The GFVen^{148} dimer was confirmed independently by SDS-PAGE, as shown in (d) of the main text. Terminal processing of the H148 variants seems to be a common theme (see Figure S2).

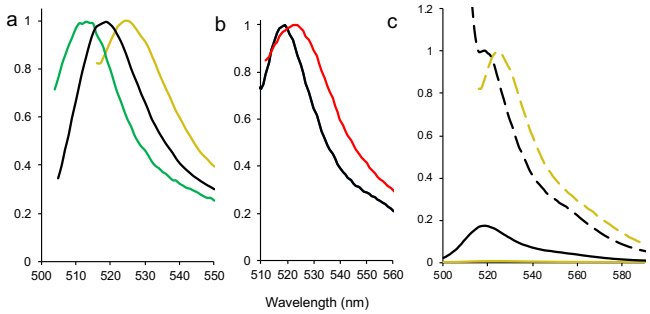

Figure S12. Comparison of emission spectra. (a) GFP^{148azF} (green; excitation 490 nm), Venus^{148azF} (gold; excitation 510 nm) and GFVen^{148} (black; excitation 490 nm). (b) Comparison of measured emission spectra (on excitation at 505 nm) of GFVen^{148} (black line) and additive emission spectrum of GFP^{148SCO} (excitation at 490 nm) and Venus^{148azF} (excitation at 505 nm). The molar absorbance coefficient for each monomer at their excitation wavelengths was similar (~17,200 and ~17,800 $M^{-1}cm^{-1}$, respectively). (c) Emission of Venus^{148azF} (gold) and GFVen^{148} (black) on excitation at 400 nm (solid line) or 510 nm (dashed line).

Supplementary References

1. D. Kozakov, D. R. Hall, B. Xia, K. A. Porter, D. Padhorny, C. Yueh, D. Beglov and S. Vajda, *Nat Protoc*, 2017, **12**, 255-278.
2. S. C. Reddington, E. M. Tippmann and D. D. Jones, *Chemical communications*, 2012, **48**, 8419-8421.
3. A. M. Hartley, H. L. Worthy, S. C. Reddington, P. J. Rizkallah and D. D. Jones, *Chemical Science*, 2016, DOI: 10.1039/C6SC00944A.
4. S. C. Reddington, P. J. Rizkallah, P. D. Watson, R. Pearson, E. M. Tippmann and D. D. Jones, *Angew Chem Int Ed Engl*, 2013, **52**, 5974-5977.
5. S. J. Miyake-Stoner, C. A. Refakis, J. T. Hammill, H. Lusic, J. L. Hazen, A. Deiters and R. A. Mehl, *Biochemistry*, 2010, **49**, 1667-1677.
6. T. Plass, S. Milles, C. Koehler, C. Schultz and E. A. Lemke, *Angew Chem Int Ed Engl*, 2011, **50**, 3878-3881.
7. F. W. Studier, *Protein Expr Purif*, 2005, **41**, 207-234.
8. C. A. Schneider, W. S. Rasband and K. W. Eliceiri, *Nat Methods*, 2012, **9**, 671-675.
9. J. Y. Tinevez, N. Perry, J. Schindelin, G. M. Hoopes, G. D. Reynolds, E. Laplantine, S. Y. Bednarek, S. L. Shorte and K. W. Eliceiri, *Methods*, 2017, **115**, 80-90.
10. J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak and A. Cardona, *Nat Methods*, 2012, **9**, 676-682.
11. G. Winter, *J Appl Crystallogr*, 2009, **43**, 196-190.
12. P. Evans, *Acta Crystallogr D Biol Crystallogr*, 2006, **62**, 72-82.
13. S. Bailey, *Acta Crystallographica Section D-Biological Crystallography*, 1994, **50**, 760-763.
14. P. Emsley and K. Cowtan, *Acta Crystallogr D Biol Crystallogr*, 2004, **60**, 2126-2132.
15. G. N. Murshudov, A. A. Vagin and E. J. Dodson, *Acta Crystallogr D Biol Crystallogr*, 1997, **53**, 240-255.
16. J. D. Pedelacq, S. Cabantous, T. Tran, T. C. Terwilliger and G. S. Waldo, *Nat Biotechnol*, 2006, **24**, 79-88.
17. T. Nagai, K. Ibata, E. S. Park, M. Kubota, K. Mikoshiba and A. Miyawaki, *Nat Biotechnol*, 2002, **20**, 87-90.
18. E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor and J. Damborsky, *PLoS Comput Biol*, 2012, **8**, e1002708.