Simulation-guided engineering of split GFPs with efficient β-strand photodissociation

Yasmin Shamsudin[1,2], Alice R. Walker[1,3], Chey M. Jones[1], Todd J. Martínez[1] & Steven G. Boxer[1]**

Affiliations:

[1]Department of Chemistry, Stanford University, Stanford, CA 94305, USA.
[2]Department of Chemistry-BMC, Uppsala University, 752 37 Uppsala, Sweden
[3]Department of Chemistry, Wayne State University, Detroit, USA

*Corresponding authors

SGB: sboxer@stanford.edu and

YS: yasmin.shamsudin@kemi.uu.se

Keywords: Green Fluorescent Proteins, Protein Engineering, Multiscale modeling

Abbreviations: GFP Green Fluorescent Proteins, MD Molecular Dynamics

ABSTRACT

Green fluorescent proteins (GFPs) are ubiquitous for protein tagging and live-cell imaging. Split-GFPs are widely used to study protein-protein interactions by fusing proteins of interest to split GFP fragments that create a fluorophore upon complementation. Complementation is typically irreversible, and controlled dissociation of the fragments would be desirable. The quantum efficiency of light-induced photodissociation of split GFPs is low, with extensive mutagenesis and screening using traditional protein engineering approaches proving difficult to implement. To reduce the search space, key states in the dissociation process were modeled by combining classical and QM/MM molecular dynamics and enhanced sampling methods, enabling the rational design and engineering of split GFPs with up to 20-fold faster photodissociation rates using non-intuitive amino acid changes. This demonstrates the feasibility of modeling complex molecular processes using state-of-the-art computational methods, and the potential of integrating computational methods to increase the success rate in protein engineering projects.

INTRODUCTION

Green fluorescent proteins (GFPs) are the most widely used genetically encoded fluorescent reporters.[1] Since their discovery, GFPs have been the subject of exhaustive protein engineering efforts to enhance expression, stability, chromophore maturation rate, fluorescence quantum yield, color, and capacity for photoactivation, photoconversion, and photoswitching, the latter largely directed at applications for super-resolution imaging.[1–4] Split GFPs have been developed to probe protein-protein interactions by fusing fragments of the canonical GFP 11-stranded β-barrel to proteins whose interaction brings the fragments together, giving a fluorescence readout.[1,5]

A shortcoming of split GFP complementation assays is that they are generally irreversible because the binding of the split β-strand peptide to re-form the intact, albeit still split, GFP is irreversible. While studying the properties of split GFPs, we were surprised to observe that once cut, some versions of split GFPs can be photodissociated,[6,7] suggesting novel optogenetic applications of GFPs along with their well-studied role for imaging. Photodissociation of the best-characterized example, a circular permutant of super-folder GFP with strand 10 at the N-terminus and cut between strands 10 and 11, can be readily monitored by adding an excess of strand 10 containing the T203Y mutation that leads to a green-to-yellow color shift when it binds and replaces the photodissociated strand (Fig. 1a & b).[6,8] Detailed investigation of this and other circular permutants led to the general potential energy surface (PES) for the photodissociation process shown in Fig. 1c.[7] Strand photodissociation was shown to be a two-step process in which light activates chromophore *cis-trans* isomerization, followed by light-independent strand-dissociation. Unfortunately, the quantum efficiency for this process is too low for practical applications.[7,9] Improving the efficiency of strand photodissociation, while at the same time preserving the stability of the split GFP against spontaneous thermal dissociation is a challenging undertaking given the complexity of the steps involved in strand photodissociation (Fig. 1c). Previous attempts using rational low-throughput approaches such as site-specific mutagenesis produced only modest improvements,[7] while high-throughput methods using extensive mutagenesis and selection strategies proved very difficult to implement.[10]

As it is far from obvious what amino acid changes might enhance photodissociation while not adversely affecting spontaneous strand-dissociation in the dark, in the present work we aimed to limit the search space for mutagenesis by identifying key residues for mutation using computational modeling and simulations. Our simulations explore the structural landscape on the

PES (Fig. 1c) to gain insight on the structural changes associated with the split protein during and in response to chromophore *cis-trans* isomerization. The simulation strategy (Fig. 1d and Fig. 2) described in the following leads to predictions of sites for mutagenesis that were not expected by simple inspection of the crystal structure and guide the design of residues that are found to substantially increase the efficiency of photodissociation.

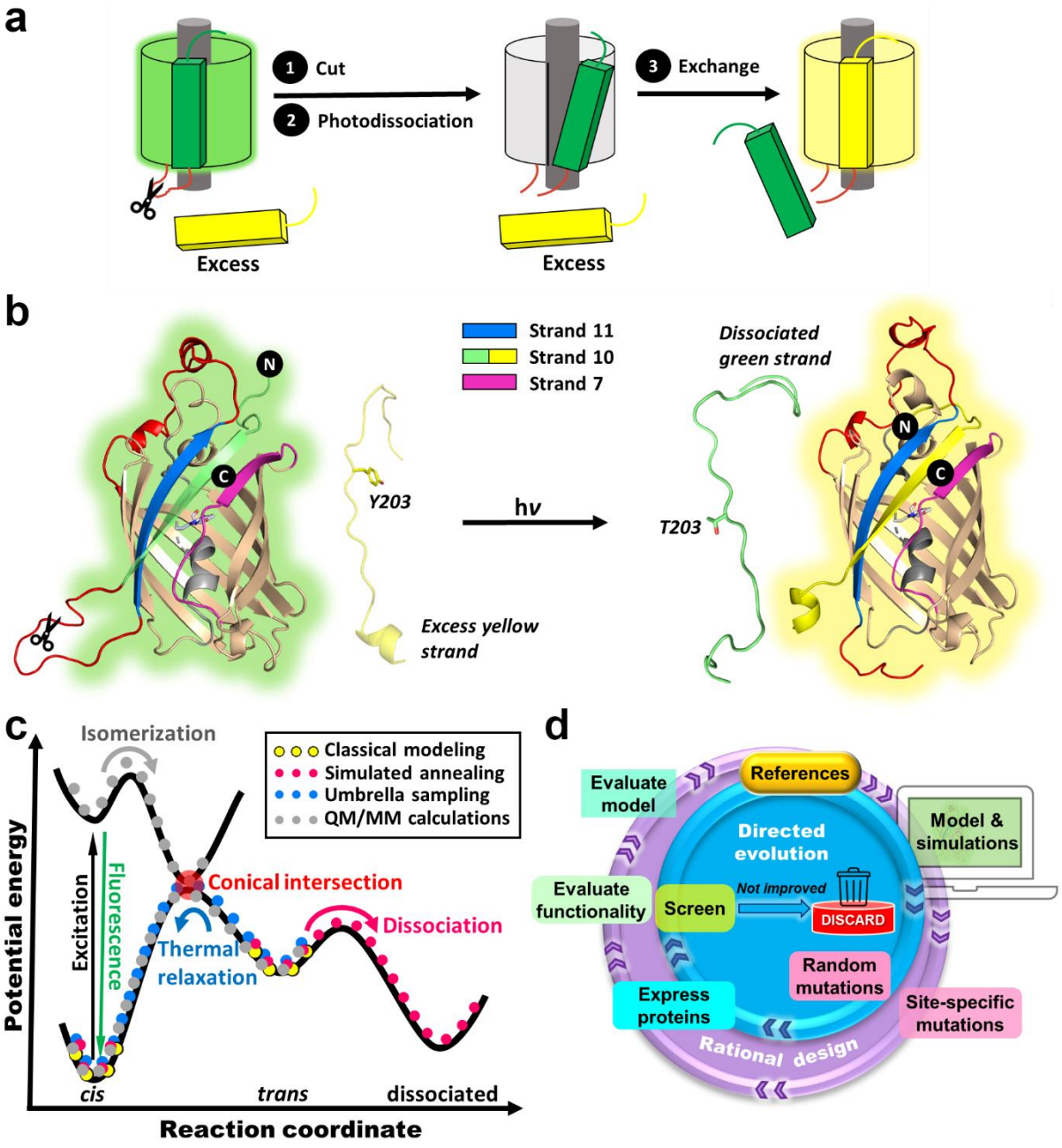


Fig. 1 a) Scheme of photodissociation and strand-exchange experiments. The reference protein is cleaved between strand 10 and 11 and irradiated with a 488nm laser, causing excitation, isomerization, and dissociation of the original

strand 10 in the presence of excess synthetic strands containing T203Y. Upon binding, the synthetic strand shifts the absorption from green to yellow, enabling measurement of pseudo-first order exchange rates that reflect the rate of photodissociation. **b)** Detailed view highlighting the N and C termini, internal alpha helix (gray), chromophore, the beta-strands 7 (purple), 10 (green/yellow), and 11 (blue), and the modeled loops on both ends of strand 11 (red). Scissors indicate the proteolytic cleavage site. **c)** Schematic of potential energy curves for photodissociation (black lines[7]), overlaid with the computational methods used to model the states (dots). The excited and ground states are degenerate at the conical intersection (red). **d**) Overview of the simulation-guided protein engineering scheme used, compared with traditional experimental-only methods.

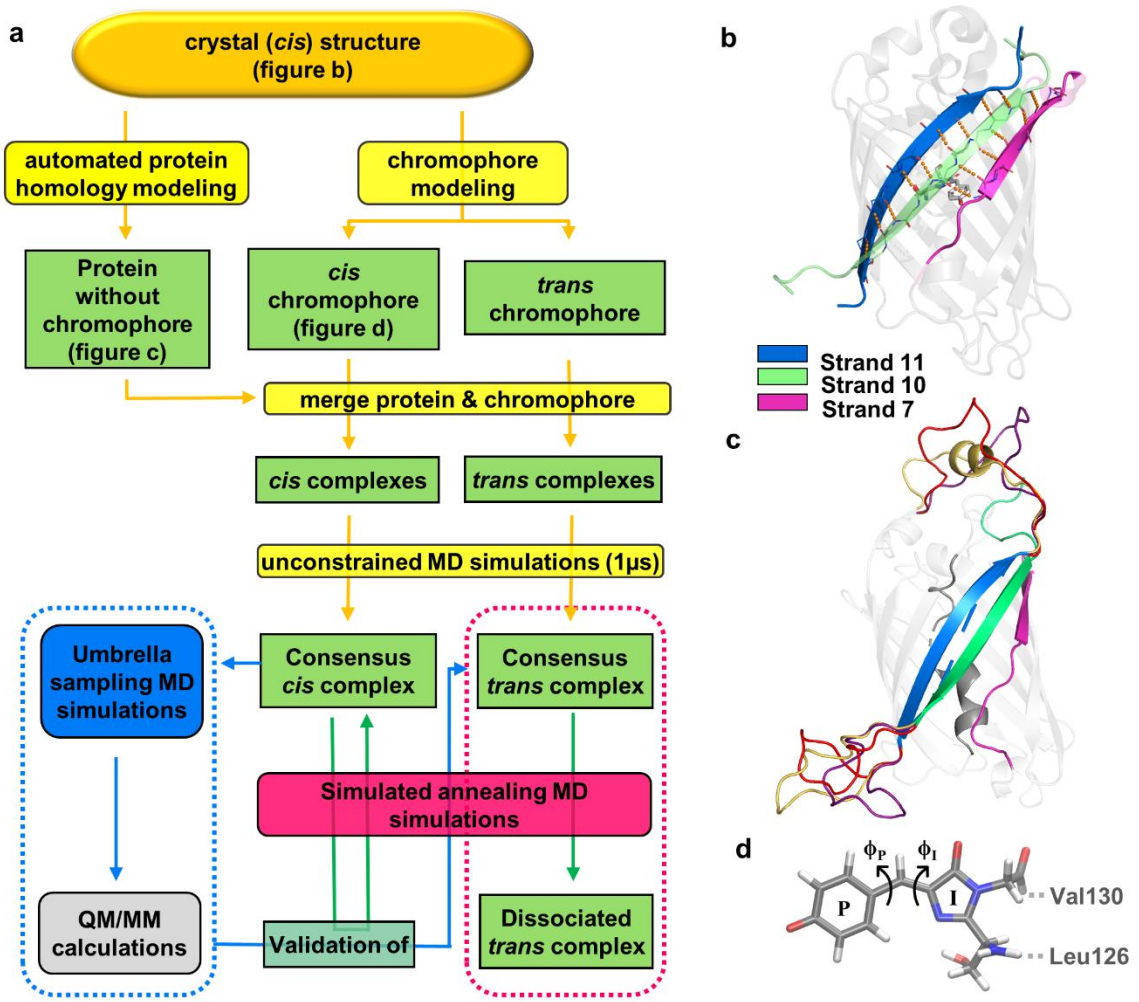


Figure 2 (/6 max). a) Computational protocol for generating models used for predicting dissociation and isomerization mechanisms leading to point mutations resulting in faster photodissociation or lowered isomerization barriers. **b)** The crystal structure from the PDB entry 6OFO, highlighting inter-strand hydrogen bonds. **c)** Homology models including loops (yellow, red, purple) missing in the crystal structure. The internal helix (gray) onto which the chromophores were modeled is highlighted. **d)** Structure of the anionic chromophore moiety (*cis* configuration), illustrating the two possible twisting dihedral angles ϕ_P and ϕ_I. Arrows illustrate clockwise rotation around the bonds. Dotted lines indicate capping atoms and neighboring residues. All protein images highlight the strands 7 (purple), 10 (green), and 11 (blue) for reference.

RESULTS AND DISCUSSION

Modelling and validating GFP structures

The computational procedure used to obtain models for predicting mutation sites for faster strand photodissociation and lowered isomerization barriers is outlined in Fig. 2a, with each step detailed in SI Methods. As the crystal structure (PDB entry 6OFO)[9] has two unresolved loops and multiple partially resolved sidechains, mainly on the outside of the barrel, homology modelling was required for further simulations. Using the crystal structure (Fig. 2b) as the template, three homology models (Fig. 2c) of the split-GFP complex with the chromophore (Fig. 2d) in the *cis* conformation (*cis* complex) and three homology models with the chromophore in the *trans* conformation (*trans* complex) were created. Each cleaved complex was subjected to three replicates of 1 μs unrestrained MD simulations to obtain consensus structures that could be used as starting structures for umbrella sampling and simulated annealing simulations. Upon visual inspection, we found that the barrels and associated sidechains were stable throughout the simulations, without showing signs of unfolding or strand-dissociation. Furthermore, although the loops of the three homology models differed at the beginning of the simulations, they adopted similar conformations at the end. After 1 μs of simulations, the average heavy atom RMSD was 3.0 Å both for the nine simulations of the *cis* complexes, and for the corresponding nine simulations of the *trans* complexes. However, the average heavy atom RMSD of the peptide backbone in the β-barrel was < 1.0 Å, with the main differences observed in the disordered loops (see Fig. S1), indicating that all simulations had converged towards similar barrel structures. When comparing all 18 simulations, the average heavy atom RMSD was 3.6 Å. The average heavy atom RMSD of the peptide backbone in the β-barrel was still < 1.0 Å, with the main differences in the loop regions, apart from the introduced changes in the chromophore. Thus, one *cis* and one *trans* consensus complex was created using the final frame from one of the above trajectories of each chromophore conformation, and these *cis* and *trans* consensus complexes were then used as the basis for subsequent point mutations and enhanced sampling simulations.

The *cis* consensus complex was aligned with the template *cis* crystal structure, resulting in an average heavy atom RMSD of 1.0 Å, confirming a stable and conserved structure. Since no crystal structure exists for the corresponding *trans* complex, we used another computational method to validate the *trans* consensus complex; the *cis* consensus complex was used as the starting

point for classical umbrella sampling, gradually rotating and sampling the chromophore along its ϕ_I coordinate (Fig. 2d), thus modelling the *cis–trans* isomerization of the GFP chromophore in its protein environment albeit in the ground state. The heavy atom RMSD of the resulting *trans* complex at 160 degrees and the *trans* consensus complex was 2.0 Å, indicating high similarity despite different starting structures and computational approaches.

Apart from validating the *trans* consensus complex, the umbrella sampling simulations were used as starting structures for QM/MM calculations on the ground and excited states (see SI Methods). Although force field parameters such as atomic charges around the conical intersection cannot be well captured classically,[11,12] our QM/MM calculations based on these classically sampled trajectories revealed that substantial differences in atomic charges were only found within a 10-degree window at the conical intersection (ϕ_I=90-100°) (Fig. S2). Beyond this window, the umbrella sampling gave a plausible approximation for the description of the chromophore in the ground and excited states, as it moves from *cis* to *trans* and, therefore, the motion of protein residues around it. These results further validate the force field parameters used to simulate the *trans* consensus complex, while indicating the potential of analyzing the ground state umbrella sampling dynamics to assess how the protein environment changes around the chromophore as it isomerizes, which could guide the identification of mutation sites for decreasing rotational barriers for isomerization.

Finally, we assessed the stability of both the *cis* and *trans* consensus complexes through simulated annealing MD simulations (see SI Methods). Except for the rotation of the chromophore, only minor structural differences can be observed between the *cis* and *trans* complexes (overall average heteroatom RMSD: 3.6 Å). However, the simulated annealing simulations revealed how the differences in structure affect the stability of the split complexes. Notably, the barrel of the *cis* complex was intact even at high temperatures during the timescale of our simulations. By contrast, although the barrel in the *trans* complex was also very stable before heating, increased disorder, followed by strand-dissociation and unfolding, was observed during the high-temperature simulations (see discussion below). These results suggest that both consensus structures are stable for MD simulations at room temperatures. Experimentally, strand-exchange in the *cis* form occurs very slowly over the course of weeks but upon exposure to light, strand-exchange occurs in minutes to hours depending on incident power and the limited mutants that have been explored.[7] The

observed different behaviors at higher simulated temperatures are consistent with these experimental observations, indicating that the models are useful for comparing the dynamics of both the *cis* and *trans* complexes over time.

Simulating strand-dissociation and hydrogen bond analysis

The crystal structure of circularly permuted split GFP[9] reveals that the nearly ideal β-strand 10 is kept in place between the neighboring β-strands 7 and 11 (Fig. 2b). The main interactions between the β-strands in the barrel are inter-strand backbone hydrogen bonds formed between opposing main chain amides and carbonyl groups, especially between strands 10 and 11, and to a lesser degree, between strands 10 and 7. On strand 10, odd-numbered residues have their sidechains pointing into the barrel, while even-numbered residues have their sidechains pointing out of the barrel in the ground state structure (Fig. 3a). Besides the chromophore–sidechain interactions with Thr203 and Thr205 on strand 10 and His148 on strand 7, an inter-strand hydrogen bond involving a sidechain is seen only between Lys209 on strand 10 and the main chain of His217 on strand 11 (Fig. 3a). This observation was the basis of creating the Lys209Gln mutation to break the hydrogen bond, which resulted in up to two times faster photodissociation.[7] Unfortunately, no more obvious mutation sites can be inferred from the crystal structure, motivating the use of molecular modeling and simulations to obtain further structural information to guide additional mutations.

Through our combination of computational methods, we obtained models for studying the dynamic motions and structural rearrangements before, during and following isomerization, leading to disorder in the barrel and strand-dissociation. Interestingly, analyses of the simulated annealing simulations of the consensus *cis* complex revealed additional sidechain–sidechain interactions, mainly involving the outward-facing residues (Fig. 3b). On strand 10, Tyr200 stacks with Tyr151 on strand 7, while Ser202 alternatingly interacts with Asn229 on strand 11 and Asn149 on strand 7. Furthermore, Lys209 forms a salt-bridge with Asp216 on strand 11 while maintaining the interaction with His217 discussed above. Interactions between sidechains on strands 7 and 11 and mainchains on strand 10 can also be observed (Fig. 3b).

To understand how isomerization affects these interactions, and the protein in general, we analyzed the structural changes and changes in hydrogen bond networks induced during the umbrella sampling simulations (Fig. S3). Initially, as the chromophore isomerizes, the hydrogen bonds with Thr205 and His148 are broken. As the chromophore continues to rotate, interactions

with the Asn121 sidechain and the Ser147 and Tyr151 mainchains are briefly observed before the chromophore reaches the *trans* state where it is partially solvated. Thus, mutation of Thr205, His148, and Asn121 could reduce steric hinderance along the isomerization pathway, resulting in the lowering of the rotational barrier, and thus an increase of the isomerization quantum yield. Whereas His148 maintains a hydrogen bond with Asn146 on the same strand, rotation of Thr205 results in the breaking of hydrogen bonds between the mainchain and the sidechain of Ser147 on strand 7. Several hydrogen bonds are broken and formed during the isomerization process. However, it is notable that several of the broken hydrogen bonds are between strand 10 residues and residues on strand 7, including between Thr203 and Ser147, but no new hydrogen bonds are formed between the two strands (Fig. 3c). This could explain the decreased stability of the *trans* complex compared with the *cis* complex, and the observed increased probability of strand-dissociation.

Finally, to elucidate the strand-dissociation pathway following isomerization, we analyzed the simulated annealing trajectories of the consensus *trans* complexes. At the start of these simulations, most inter-strand sidechain–sidechain and sidechain–mainchain interactions seen in Fig. 3a are broken. Particularly, the strand 11 sidechains Lys214, Arg215, and His217 near the cleavage site between strands 10 and 11 rapidly separate from strand 10; however, the complex remains stable until the subsequent separation of the Lys209 sidechain on strand 10 from both the His217 main chain and Asp216 side chain on the neighboring strand 11. The separation process follows the complete solvation of the cleaved end of strand 10 (Fig. 3c) and is only observed during heating and simulations at 600 K (see SI Methods). Once Lys209 separates, the complex becomes increasingly disordered, especially around strands 10 and 7, and at the cleaved end of strand 11 (Fig. 3c). Notably, before the strand becomes completely disordered and detached, we can still observe the alternating sidechain–sidechain interactions between Ser202 and the neighboring Asn225 and Asn149. Meanwhile, Tyr200 loses the stacking interaction with Tyr151, although new interactions are formed with Asn149. At this point, mainchain–mainchain interactions can still be seen around Tyr200 and Thr205 (Fig. 3c). Although the preceding mechanisms and structures can be observed in all replicates of the high-temperature simulations of the *trans* complex, different scenarios emerge as the simulations progress from the structures equivalent to that illustrated in Fig. 3c. In some simulations, the gap between strands 10 and 7 increases while mainchain–mainchain interactions between strands 10 and 11 remain, ahead of strand-dissociation. In others,

the strand 10 moves in between strand 7 and 11, shifting the positions of strand 10 residues relative to the neighboring strands in a stepwise manner. Thus, as Asn225 and Asn129 lose their interactions with Ser202, these are replaced by the corresponding interactions with the next outward-facing residue on strand 10, Tyr200.

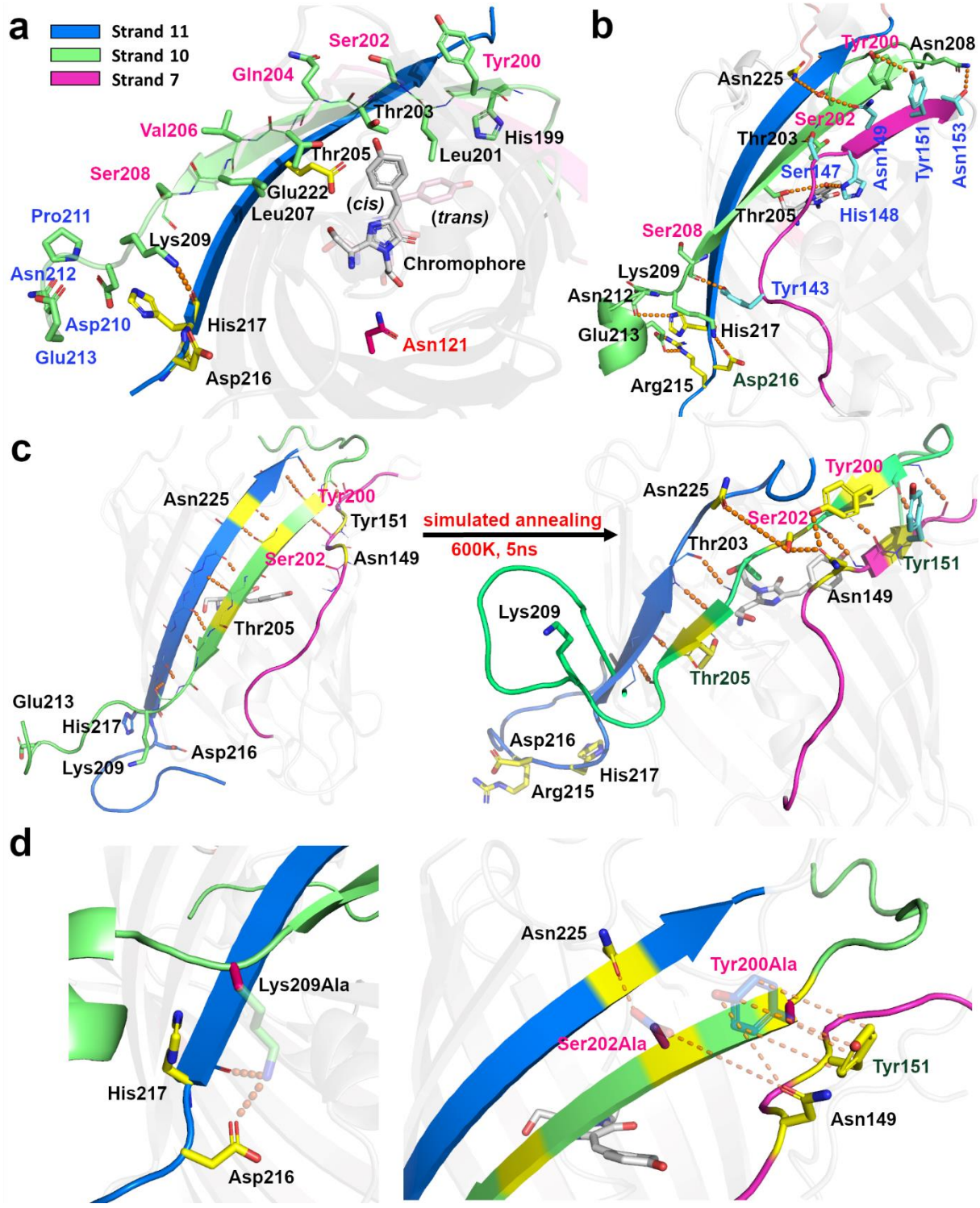


Figure 3 (/6 max). Overview of the hydrogen bonding (orange dashes) network between strand 10 (green) and the neighboring strands 7 (purple) and 11 (blue) in the **a)** consensus *cis* complex highlighting the inter-strand hydrogen bond between Lys209 and His21. Sidechains on strand 10 color-coded as follow: pointing out of the barrel (pink), into the barrel (black) or fully solvent-exposed (blue). Glu222 (yellow) and Asn121 (magenta) shown for reference. **b)** cis complex following simulated annealing highlighting inter-strand side-chain interactions. **c)** consensus *trans* complex illustrating the separation of Lys209 from His217 and Asp216, and the fully solvated cleaved end of the strand 10

before heating, and the partially disordered state where inter-strand sidechain-sidechain interactions involving strand 10-residues have been reduced to only Tyr200 and Ser202 after heating. Labeled residues (yellow) shown for reference, and **d)** Our identified mutation sites Lys209, Ser202, and Tyr200. Transparent sidechains show original amino acids, while dashes indicate removed polar interactions following mutations to alanine.

Protein engineering predictions.

As shown schematically in Fig. 1a, photodissociation of split strand 10 can be readily measured by irradiating the split-GFP complex in the presence of an excess of a strand 10 peptide containing the T203Y mutation that converts the protein with an absorption at 470 nm (GFP) into one that absorbs at 505 nm (YFP-like). To simplify comparison of strand-exchange rates and validate the predictability of our models, we introduced three criteria for our protein design aimed at faster strand photodissociation. First, mutations are only introduced on the dissociating strand to ensure identical products when the excess peptide binds to form YFP. Second, to minimize interference with the chromophore on the interior of the barrel, we considered only solvent-exposed residues. Although counterintuitive, the simulations suggest that such residues could be important along the dissociation pathway. Third, to separate and assess the reliability and predictiveness of our models for isomerization and strand photodissociation, single-point mutations should be identified as either affecting isomerization or strand photodissociation, but not both.

To identify point mutations that could affect dissociation separately from isomerization, Fig. 3b highlights several interactions between strand 10 and the neighboring strands. Dissociation occurs in the *trans* complex, shown in Fig. 3c, and here the number of interactions between strand 7 and neighboring strands has been drastically reduced, significantly decreasing the number of potential mutation sites. Our simulations indicate that the hydrogen bond between Lys209 and His217, and the salt bridge between Lys209 and Asp216 are stable at 300 K. Since the separation of Lys209 from both precedes the disordering of strand 10, mutation of Lys209 could enhance photodissociation rates; however, the single-point Lys209Gln mutation resulted in only modest improvements.[9] Although Gln209 would not form a salt-bridge with Asp216, it can still form a hydrogen bond, thus likely limiting the effect of the mutation. Meanwhile, a non-polar residue, such as isoleucine, would prevent such interactions, while an alanine would also reduce non-polar sidechain interactions.

Apart from the Lys209, we identified Ser202 and Tyr200 as two additional candidates for mutation. In the simulated annealing simulations, Ser202 forms alternating hydrogen bonds with Asn225 on strand 11 and Asn149 on strand 7. Meanwhile, Tyr200 is either stacked with Tyr151 on strand 7 or hydrogen bonded to Asn149. Although these interactions are also present in the *cis* complex, the *cis* complex has additional inter-strand interactions (Fig. 2b and 3b) that stabilize the protein and could reduce the effects of the mutations at these sites. Since the *trans* complex has fewer stable inter-strand interactions (Fig. 3c), the relative effect of these mutations on strand-dissociation should be larger. Thus, while Tyr200, Ser202, and Lys209 could be independently mutated, our simulations suggest that a combination of mutations of all three would produce the largest impact on strand-dissociation (Fig. 3d). To reduce their polar sidechain interactions, we considered two alternatives – eliminating sidechain interactions through alanine mutations, and introducing steric clashes using bulky and nonpolar residues, such as isoleucine, valine, leucine, and tryptophan (Fig. 4a).

Although improving the efficiency of strand-dissociation is the focus of this work, it could also be desirable to introduce mutations that reduce the isomerization barrier to increase the yield of the dissociating *trans* complex (Fig. 1c). The umbrella sampling simulations and the subsequent hydrogen bond analysis indicate three sites where mutations could affect isomerization: Thr205, His148, and Asn121. Thr205 on strand 10 and His148 on the neighboring strand 7 could affect both isomerization and strand-dissociation as isomerization precedes strand-dissociation, so it would be difficult to untangle the contribution to each step at such mutation sites. Meanwhile, Asn121 is situated on the side of the chromophore opposite the dissociating strand (Fig. 3a), making it a suitable candidate for investigating the potential impact on isomerization separately from strand-dissociation.

Experiments

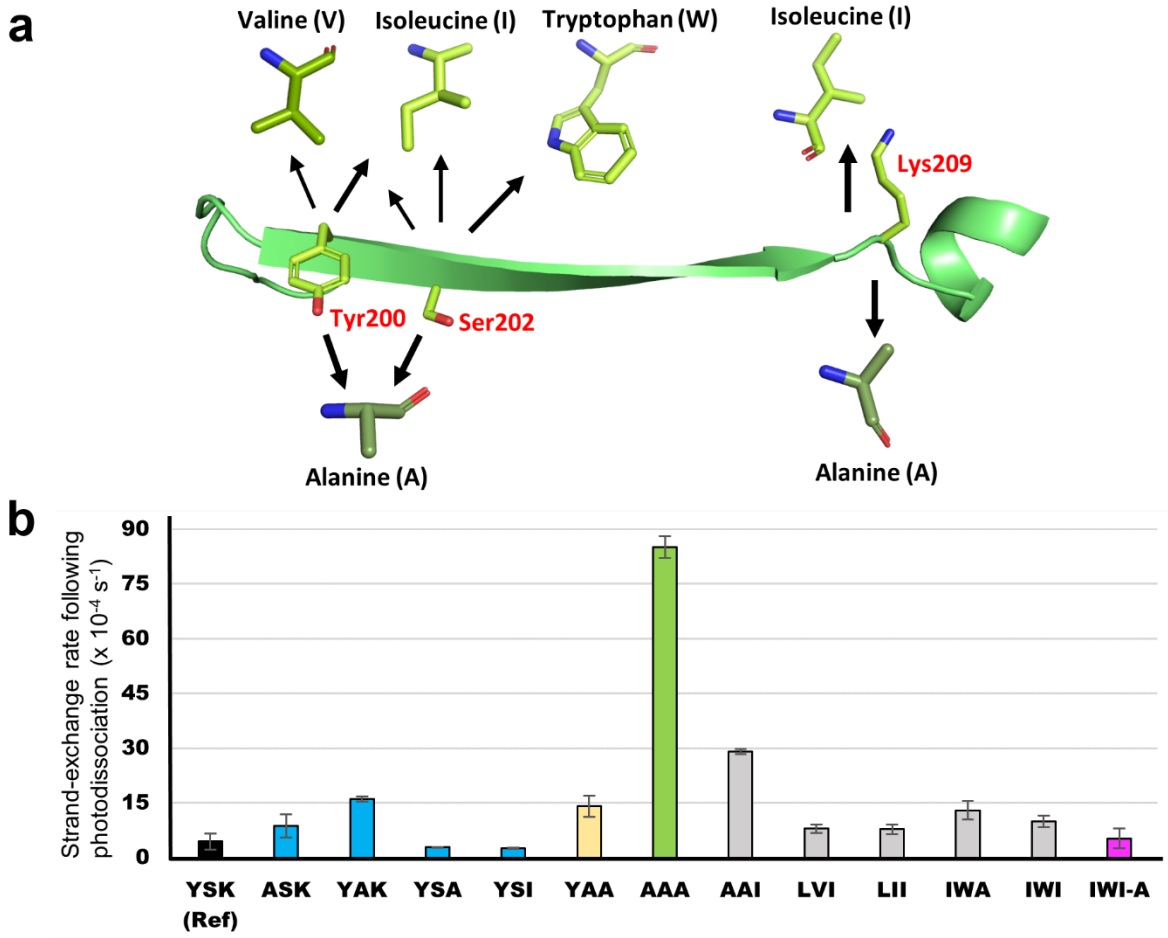


Figure 4 (/6 max). a) Mutant combinations considered in this work. Our reference structure containing Tyr200 (Y), Ser202 (S), and Lys209 (K) is referred to as YSK and shown on the strand. The mutants are denoted by the amino acids in these positions and orders. **b)** Pseudo-first order rates of strand-exchange (Fig. 1a) following photodissociation for the YSK (black) and new mutants. Single-point mutants are highlighted in blue, the double mutant in yellow, triple mutants in grey, with the fastest photodissociating triple mutant AAA highlighted in green, and the quadruple mutant in pink, which includes the IWI mutations on the strand, and the additional Asn121Ala mutation. Error bars show standard deviations between replicates. The spontaneous strand off rate in the dark is in the baseline. 488 nm laser irradiation at 26 mW was used for all experiments.

As Tyr (Y) 200, Ser (S) 202, and Lys (K) 209 are not directly interacting with the chromophore, and are not expected to directly affect isomerization, changes in kinetics resulting from mutations at these sites are expected to mainly affect strand-dissociation. Henceforth, our reference structure is referred to as "YSK" (the mutations are listed by the amino acids at the mutated positions from the N- to C-termini, Fig. 4a). To test our predictions, four single-point mutants, one double mutant, and six triple mutants (Fig. 4b), as well as the reference structure were expressed, cleaved, and subjected to laser-induced strand-exchange experiments (Fig. 1a) (SI

Methods). Our results show that among the three mutation sites, the Ser202Ala mutation caused the largest improvement for the single-point mutants. Additionally, the inclusion of the Ser202Ala mutation in double and triple mutants consistently improved strand-exchange rates. Furthermore, all our triple-mutants display faster rates compared with our reference protein YSK, indicating the usefulness of our models. Most notably, the AAA mutant displays the largest rate increase, up to 20 times faster than YSK (Fig. 4b, Table S1). To ensure that the mutation did not also equivalently increase spontaneous dissociation in the *cis* form, we also monitored strand-exchange rates of samples not exposed to light (dark controls). The dark exchange rate was 85 times slower than the light-induced exchange rate.

Fluorescence quantum yields (FQYs) of GFP mutants have previously been measured to estimate *cis–trans* isomerization efficiency. Although changes in fluorescence are not necessarily correlated with isomerization, FQYs can nevertheless be used to probe the excited-state behavior of mutants, as excited-state isomerization competes with fluorescence emission. For the triple mutants AAA and IWI, we observed 2-4% increases in FQY (Table S2), suggesting that the point mutations on the dissociating strand do not have a substantial effect on the isomerization efficiency. Thus, we can deduce that changes in observed exchange rates in our single, double, and triple mutants are directly correlated to strand-dissociation rates.

To assess our model of the isomerization pathway, two quadruple mutant were expressed, based on the AAA and IWI mutants, adding the Asn121Ala mutation, and referred to as AAA-A and IWI-A. However, only IWI-A yielded enough protein for further experiments. To investigate if the Asn121Ala mutation would affect the isomerization barrier, we measured the FQYs for it. Interestingly, the experiments revealed a decrease by 8% compared with the corresponding IWI mutant and a 4% decrease compared with the YSK reference structure. Although this result does not prove that the mutation increases the *cis–trans* isomerization efficiency, it could be one explanation for the results. To validate this assumption, we performed umbrella sampling simulations of the IWI-A mutant, which showed lowered isomerization barriers of 5 kcal/mol compared with YSK (see Fig. S4), indicating that increased *cis–trans* isomerization is at least partially the reason for the decreased FQY. Interestingly, despite this increase in the population of the *trans* complex, the strand photodissociation of IWI-A did not increase relative to IWI, and the exchange rate for IWI-A decreased compared with IWI.

CONCLUSION

β-strand photodissociation of split GFP is a complex process, and conventional directed evolution methods to increase the photodissociation yield while not enhancing spontaneous dissociation in the dark have proven difficult.[10] Therefore, we turned to simulations, even though the overall process being simulated requires multiple approaches, including multiscale modeling and enhanced sampling methods, and pushes the limits of what is possible. Based on molecular modelling and simulations, we have made predictions and then engineered a series of GFP mutants with specific properties, namely split GFPs with substantially more efficient β–strand-dissociation following photoexcitation.

Here, various computational methods were used to model the entire pathway of *cis*-to-*trans* isomerization and the subsequent strand-dissociation process. Although there are clear limits to the usefulness of the methods used here, there are also obvious advantages. For instance, the complexity of the excited-state isomerization makes simulations in the ground state at best an indirect approach to a true excited-state calculation; however, the latter are computationally very expensive. Meanwhile, although the protein movement caused by the chromophore isomerization in the excited state can be approximated by simulations in the ground state, the corresponding calculated barriers of isomerization in the ground state are likely not accurate. The goal here, however, was to visualize the trajectory of strand-dissociation, inspect the structures for interactions that are broken during strand-dissociation, and then make and test predictions based on this model. Similarly, although the simulated annealing simulations proved to be a useful method for visualizing the dissociation mechanism and identifying key residues for point mutations, using such simulations to predict how much faster a mutant would dissociate compared to the reference structure is challenging. Nonetheless, through the combination of theoretical, computational, and experimental methods, we have demonstrated an approach to engineering and modifying a complex protein function, producing what could be a form of split GFP that should be useful both as an optogenetic and imaging tools. The ability to efficiently dissociate the strand with light opens possibilities of site-specific reversible macromolecular interactions, genetically encoded caged enzymes, and site-specific cargo delivery in cells.[13]

SI METHODS

MODELLING THE SPLIT GFP COMPLEX

Our reference protein is a circularly permuted variant of the superfolder GFP (PDB ID: 2B3P)[14], with an added sacrificial loop for enzymatic cleavage and dissociation of β-strand 10 (see complete DNA and amino acid sequences in the SI). A closely homologous protein structure with the chromophore in the *cis* conformation has been determined (PDB ID: 6OFO)[9]. It contains two mutations (Cys48Ser and Cys70Ala), unresolved loops on both ends of strand 11, and multiple partially unresolved sidechains, mainly on loops and the solvent-exposed face of the barrel.[10] Thus, further modeling was required to obtain starting structures suitable for computational simulations, corresponding to the split reference protein. Therefore, the 6OFO[9] crystal structure was used as the template for homology modeling of the reference protein with both cysteines restored, and Glu222 in the protonated form. The two models generated using Modeller[15] and one model created using Prime[16] (see Fig. 2b) contained the β-barrel, internal α-helix, and the missing loops, excluding the sacrificial loop (see SI for full sequence) and the chromophore, which was separately modelled (see below).

The *cis* and *trans* chromophores (Fig. 2d and S1) were built using the Build menu in Maestro (Bioluminate 4.3[17]) and minimized using MacroModel[17] before optimizing the structures with GAUSSIAN16[18] using B3LYP/6-31G(d). The obtained RESP charges were fitted with Antechamber.[19] The chromophores were manually merged with the internal α-helix of each homology model, and the charges at the interface were manually corrected to create systems with integer charges. Additional parameters were obtained from the GAFF2[19] and FF14SB[20] force fields.

Three complexes were created with the *cis* chromophore, and three models with the *trans* chromophore. We will refer to complexes containing the chromophore modeled in the *cis* conformation as *cis* complexes and the corresponding complexes with the chromophore modeled in the *trans* conformation as *trans* complexes.

To maintain consistency with previous works on GFP, we use the superfolder GFP (sfGFP) nomenclature in our discussions, despite differences in our complexes due to the circular permutation placing the N-terminal on strand 10 and added residues due to the sacrificial loop (see Fig. S5 for comparison). However, since the cleavage site of the loops between strands 10 and 11

do not have corresponding amino acids in sfGFP (see Table S3 for conversion table), we will refer to the amino acids by their computational model numbers. Thus, the six complexes were manually cleaved between Arg26(comp) and His27(comp) and protons were automatically added to the new terminals. The terminal sidechains were then selectively minimized using MacroModel[17] to prevent steric clashes between terminal protons in the subsequent simulations.

Molecular dynamics simulations

For all simulations aiming to create starting consensus and mutant structures for isomerization or enhanced sampling simulations, the minimization, heating, and equilibration protocol was as follows: The complex was prepared using the Antechamber *tleap* module, which assigned protein parameters using the FF14SB and GAFF2 forcefields, and the complex was placed in an octahedral simulation box filled with SPC/E waters[21] and counterions. The starting structure was minimized, followed by heating to 300 K (NVT) over 250 ps, and 5 ns of equilibration (NPT). The *pmemd* module in AMBER 2018[22] was used for all simulations. The step size during heating and equilibration and unconstrained MD (NPT) simulations was 2 fs, using SHAKE and RATTLE constraints. The Langevin thermostat[23] was used, with a collision frequency of 2 ps^{-1}, and the cutoff for non-bonded interactions was set at 9 Å.

Consensus structure modelling

Following equilibration, the proteins were subjected to three replicates of unconstrained 1 μs MD simulations. The trajectories of the final 5 ns for each simulation were manually inspected. The simulations were considered converged as the loops looked similar (average heteroatom RMSD: 3.6 Å) regardless of starting point. The consensus model was created by merging the barrel with representative loop regions extracted from the last frame of one simulation with each chromophore conformation.

Modelling mutants

Point mutations of Tyr200, Ser202, Lys209, and Asn121 were introduced on the consensus *cis* structures using the Mutate Residue option in the Build menu in Maestro. Rotamers of bulky ligands were chosen to minimize steric clashes. The triple mutations AAA (Tyr200Ala, Ser202Ala, and Lys209Ala) and IWI (Tyr200Ile, Ser202Trp, and Lys209Ile), and quadruple mutations with

Asn121Ala, AAA-A and IWI-A, all with the chromophore modeled in the cis conformation, were computationally modelled. Following equilibration, the mutated structures were subjected to three replicates of 20 ns of unconstrained MD simulations at 300 K. Frames from these simulations were used as starting structures for umbrella sampling simulations (see below).

MAPPING THE ISOMERIZATION PATHWAY

Isomerization

To investigate factors that promote or inhibit *cis–trans* isomerization of the chromophore in the superfolder strand 10 circular permutant, we used the *cis* consensus model as a basis for classical umbrella sampling on the ground state. The *cis* structures from the end of unconstrained, relaxed classical MD trajectories at 300 K were used as the starting structures for umbrella sampling simulations using the *pmemd.cuda* module in AMBER[22] with classical molecular dynamics. Specifically, we use biased MD simulations to model *cis–trans* isomerization in split GFP by gradually rotating and sampling the chromophore along its ϕ_I coordinate. Because the protein environment surrounding the chromophore is asymmetric, we generated potential of mean force (PMF) profiles associated with clockwise and counter-clockwise rotations along ϕ_I (Fig. 2d). The ϕ_I dihedral angle of the chromophore was rotated in increments of 10° in both directions using a harmonic constraint of 200 kcal/mol. Each window began with the final snapshot of the previous window, equilibrated for 50 ps, and run for 1 ns in the NPT ensemble with a 1 fs timestep. The classical sampling was performed with Langevin thermo/barostat with a collision frequency of 1 ps^{-1}. The resulting PMFs (Fig. S4) were constructed using the weighted histogram analysis method (WHAM).

QM/MM methods

The QM/MM calculations were performed with the TeraChem/OpenMM interface.[24–29] The QM region was defined as the chromophore and adjacent capping residues (Fig. S6a) while the MM region was the remaining protein and solvent. As justified in previous gas-phase and solution-phase studies of the anionic GFP chromophore,[11,12] the α-corrected implementation of the state-averaged complete active space self-consistent field (SA-CASSCF) method (known as α-CASSCF) was used to obtain optimized structures on the ground (S_0) and excited (S_1) electronic states.[11,12] As illustrated in Fig. S6b, we use an active space of four electrons and three orbitals,

optimizing the orbitals to minimize the average energy of the first three singlet states, i.e. SA3-CAS(4/3)SCF. Following our previous work[11,12], α(0.64)-SA3-CASSCF(4,3)/6-31G* with a D3 dispersion correction was used in our split GFP system.

Optimizations for each electronic state were initiated from the same starting structures, which were sampled from classical umbrella sampling along ϕ_I. The ϕ_I dihedral angle was constrained for each optimization. Structures were optimized with QM/MM using the α(0.64)-SA3-CASSCF(4,3)/6-31G* level of theory. Explicit water molecules more than 10 Å from the chromophore were frozen in the optimizations. Geometries were obtained using GPU-accelerated TeraChem[28] and the DL-FIND[25] optimization library. These geometries were used to validate the choice of using the one-bond-flip mechanism, which is a rotation around the ϕ_I dihedral angle, while leaving the ϕ_P dihedral angle free or unchanging (Fig. 2d) and to determine sections of the umbrella sampling that were the most distorted compared to the excited state charges and geometry.

Using geometries sampled from the classical umbrella sampling, the potential energy surface of split GFP was mapped along ϕ_I (10° intervals) using QM/MM optimizations at the α(0.64)-SA3-CASSCF(4,3)/6-31G* level. S_0-optimized structures were obtained by constraining ϕ_I and allowing the remaining degrees of freedom of chromophore and protein to relax (Fig. S2b). This method was benchmarked against extended multistate multireference second-order perturbation theory (XMS-CASPT2). Single point calculations at the SA3-XMS-CASPT2(4,3)/6-31G* level were run on α(0.64)-SA3-CASSCF(4,3)/6-31G* optimized geometries to ensure the reliability of the α parameter (Fig. S6c).

Hydrogen bond analysis

To limit the search space for potential mutagenesis, we first determined residues that had broken hydrogen bonds at least 30% of the simulation time in the umbrella sampling after moving the chromophore from *cis* to *trans* (Fig. S3). Hydrogen bond analysis was performed with AMBER's cpptraj module. The standard cutoffs of 3 Å / 135° were used to compute hydrogen bond occupancy times for each possible protein atom. Residues that changed in hydrogen bond occupancy by ± 50% between the clockwise or counterclockwise rotation from ϕ_I= 0 to 90° and 90° to 180° were extracted as a metric of residues that substantially change during isomerization.

SIMULATING STRAND-DISSOCIATION

Simulated annealing

Equilibrated consensus *cis* and *trans* structures were subjected to three replicates of enhanced sampling simulations using the *pmemd* module of AMBER as follows: the proteins were heated from 300K to 400 K over 1 ns, simulated at 400 K for 4 ns, heated to 600 K over 1 ns, simulated at 600 K for 4 ns, then cooled to 500 K over 1 ns before finally being simulated at 500 K for up to 150 ns. The trajectories were analyzed using VMD.

PROTEIN ENGINEERING

Experimental mutagenesis

The reference protein was obtained from gene synthesis (see SI for DNA sequence).[7] Point mutations were performed with the QuikChange Lightning Mutagenesis kit (Agilent) according to the manufacturer's protocol. Multiple point mutations were done in stages where Tyr200 and Ser202 were first mutated using one primer, followed by Lys209, and finally Asn121. The correct introduction of mutations was verified through DNA sequencing performed by ELIM Biopharm.

Protein expression

The pET-15b vectors containing the genes of interest were transformed into chemically competent BL21(DE3) *Escherichia coli* (Invitrogen). Baffled 3-L flasks containing 1 L of 47.6 g/L modified Terrific Broth (Fisher BioReagents), 8 g/L glycerol (Fisher, CAS 56-81-5) and 100 mg/L ampicillin (Sigma-Aldrich, CAS 69-52-3) were inoculated with single colonies of *E. coli* and grown at 37ºC with shaking at 180 rpm until reaching OD 0.6 to 0.8 at 600 nm. Then, 0.25 g/L of isopropyl β-D-1-thiogalactopyranoside (Fisher, CAS 367-93-1) were added to the cultures to induce protein expression. The cultures were incubated for an additional 20 h at 20ºC while shaking at 180 rpm.

Protein purification

E. coli containing the proteins of interest were pelleted by centrifugation at 6500 rcf for 30 min. The cell pellets were suspended in lysis buffer, an aqueous buffer at pH 8.0 containing 50 mM Tris-HCl (Fisher, CAS 1185-53-1) and 250 mM NaCl (Fisher, CAS 7647-14-5). They were then lysed with a high-pressure homogenizer (Avestin EmulsiFlex-C3). The lysate was centrifuged at 25,000 rcf for 90 min. The resulting supernatant was added to a column of Ni-NTA Agarose resin

(QIAGEN) pre-equilibrated with Buffer A (10m M NaCl, 20 mM Tris-HCl, pH 8.0 aqueous buffer). The column was rinsed with 2 column volumes of a wash buffer of 20 mM imidazole (Aldrich, CAS 288-32-4) in Buffer A before being rinsed with 2 column volumes of 200 mM imidazole in Buffer A. The fractions of eluate judged by visual inspection to contain GFP were pooled and exchanged by spin-filtration into anion-exchange Buffer A and stored at 4ºC overnight. The GFP was then purified by anion-exchange chromatography (HiTrap 5 mL Q HP; GE Healthcare) with a gradient of Buffer A and B (1 M NaCl, 80 mM Tris-HCl, pH 8.0 aqueous buffer) and stored at 4ºC.

Protein cleavage

Due to the light sensitivity, protein cleavage and subsequent purification of all mutants was done in the dark. The GFP was incubated at room temperature with 100 units of trypsin (Type III from bovine pancreas, ≥ 10,000 BAEE units per mg; Sigma) per 1 mg of GFP for 20 min while stirring constantly to cleave both the thrombin loops between the His-tags and the factor Xa loops after the N-terminal β-strands. The GFP was then purified by anion-exchange chromatography (HiTrap 5 mL Q HP; GE Healthcare) with a gradient of Buffers A and B (*vide supra*), followed by spin-filtration into anion-exchange Buffer A. Cleaved mutants were stored at 4ºC in anion-exchange Buffer A in the dark.

Mass spectrometry

Mutant identities before and after cleavage were verified by electrospray ionization mass spectrometry (ESI-MS) measured with LC–MS (Waters 2795 HPLC with ZQ single quadrupole MS in Stanford University Mass Spectrometry (SUMS) facility. Deconvoluted masses were calculated using Intact Mass (Protein Metrics). Reported expected masses (**Table S10**) are the average mass based on the sequence, calculated with the Peptide Mass Calculator provided by PeptideSynthetics.

UV-Vis absorbance measurements

UV-Vis kinetic measurements were performed with a PerkinElmer Lambda 25 UV-Vis spectrometer. Data acquisition was performed every 1.0 nm at a maximum scan rate of 480 nm/min. UV-Vis measurements not for kinetic measurements were performed with a PerkinElmer Lambda 365 UV-Vis spectrometer. Mutant concentrations and extinction coefficients were determined by measuring the UV-Vis absorbance at 447 nm in 0.1 M NaOH (Fisher BioReagents, CAS 497-19-

8) and scaling by the known extinction coefficient of the deprotonated chromophore in the denatured protein (44,100 M^{-1} cm^{-1})[30].

Laser-dissociation experiments

A thermally conductive metal cuvette holder was affixed to a magnetic stir plate stirring at 1200 rpm. 4-mL quartz cuvettes with magnetic stir bars containing 3 μM of cleaved protein and 70 μM of the YFP peptide strand, with the single point T203Y mutation compared with our reference strand 10 (see SI), solvated in 3 mL Buffer A (*vide supra*) were incubated in the holder in the dark while stirring for 15 minutes before a control UV-vis measurement was taken. A 30 mW diode laser (85-BCD-030-115, Melles Griot) was used for irradiation with 488 nm light. Laser power was measured with a stabilized thermal power sensor (Part No. S302C, Thorlabs) coupled to a digital optical power console (Part No. PM100D, Thorlabs). The reported power (26 ± 1 mW) was scaled to 96% of the measured power due to 4% external reflection at the air-quartz interface on the surface of the cuvette containing the sample. It was then normalized based on sample volume. Sample irradiation was performed for defined time intervals using an automated shutter with millisecond precision (Model 845HP Digital Shutter, Newport Research). Upon completion of each irradiation time interval the sample was stirred continuously in the dark for 15 minutes before UV-vis measurement. Each mutant was subjected to three replicates of laser-dissociation experiments.

Fluorescence measurements

Absolute fluorescence quantum yields were measured on uncleaved proteins by comparing GFP mutant fluorescence to fluorescence of the standard Fluorescein (Aldrich Chemicals, LOT 101F-0681, CAS 2321-07-5) in 0.1 M NaOH, which has a reported absolute fluorescence quantum yield of 0.90 at 488 nm.[31] Excitation was performed at 488 nm, and signal acquisition was performed every 0.5 nm at a scan speed of 120 nm/min with a slit width of 1.0 nm. Sample concentration was calibrated using a PerkinElmer Lambda 365 UV-Vis spectrometer. The fluorescence measurements were performed on a PerkinElmer LS 55 fluorescence spectrometer with emission and excitation slit widths of 2.5 nm.

SIMULATION FILES AVAILABILITY

Files associated with performed simulations, including starting structures, parameters, simulation input files, and short videos of simulation trajectories, can be retrieved from the Zenodo server using the following URL: https://doi.org/10.5281/zenodo.7674800.

REFERENCES

1. Romei, M. G. & Boxer, S. G. Split Green Fluorescent Proteins: Scope, Limitations, and Outlook. *Annu Rev Biophys* **48**, 19–44 (2019).
2. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: Pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences U.S.A* **106**, 9995–10000 (2009).
3. Misteli, T. & Spector, D. L. Applications of the green fluorescent protein in cell biology and biotechnology. *Nat Biotechnol* **15**, 961–964 (1997).
4. Remington, S. J. Green fluorescent protein: A perspective. *Protein Science* vol. 20 1509–1519 Preprint at https://doi.org/10.1002/pro.684 (2011).
5. Berríos, K. N. *et al.* Controllable genome editing with split-engineered base editors. *Nat Chem Biol* **17**, 1262–1270 (2021).
6. Do, K. & Boxer, S. G. Thermodynamics, kinetics, and photochemistry of β-strand association and dissociation in a split-GFP system. *J Am Chem Soc* **133**, 18078–18081 (2011).
7. Lin, C. Y., Both, J., Do, K. & Boxer, S. G. Mechanism and bottlenecks in strand photodissociation of split green fluorescent proteins (GFPs). *Proc Natl Acad Sci U S A* **114**, E2146–E2155 (2017).
8. Do, K. & Boxer, S. G. GFP variants with alternative β-strands and their application as light-driven protease sensors: A tale of two tails. *J Am Chem Soc* **135**, 10226–10229 (2013).
9. Lin, C. Y., Romei, M. G., Oltrogge, L. M., Mathews, I. I. & Boxer, S. G. Unified Model for Photophysical and Electro-Optical Properties of Green Fluorescent Proteins. *J Am Chem Soc* **141**, 15250–15265 (2019).
10. Romei, M. G. Controlling Color and Photoisomerization Pathways in Photoactive Proteins: The Role of the Protein Environment. (Stanford University, 2020).
11. Jones, C. M., List, N. H. & Martínez, T. J. Resolving the ultrafast dynamics of the anionic green fluorescent protein chromophore in water. *Chem Sci* **12**, 11347–11363 (2021).
12. List, N. H., Jones, C. M. & Martínez, T. J. Internal conversion of the anionic GFP chromophore: In and out of the I-twisted S1/S0conical intersection seam. *Chem Sci* **13**, 373–385 (2022).
13. Do, K. Green Fluorescent Protein and its 10th β-strand: Controlling Protein-Peptide Interactions with Light. (Stanford University, 2014).
14. Pédelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79–88 (2006).
15. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinformatics* **15**, (2006).
16. Jacobson, M. P. *et al.* A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins: Structure, Function and Genetics* **55**, 351–367 (2004).
17. Zhu, K. *et al.* Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins: Structure, Function and Bioinformatics* **82**, 1646–1655 (2014).
18. Frisch, M. J. *et al.* Gaussian16. Preprint at (2016).
19. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **25**, 247–260 (2006).

20. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).
21. Mark, P. & Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *Journal of Physical Chemistry A* **105**, 9954–9960 (2001).
22. Case, D. A. , *et al.* AMBER 18.
23. Davidchack, R. L., Handel, R. & Tretyakov, M. v. Langevin thermostat for rigid body dynamics. *Journal of Chemical Physics* **130**, (2009).
24. Isborn, C. M., Luehr, N., Ufimtsev, I. S. & Martínez, T. J. Excited-state electronic structure with configuration interaction singles and Tamm-Dancoff time-dependent density functional theory on graphical processing units. *J Chem Theory Comput* **7**, 1814–1823 (2011).
25. Kästner, J. *et al.* DL-FIND: An open-source geometry optimizer for atomistic simulations*. *Journal of Physical Chemistry A* **113**, 11856–11865 (2009).
26. Song, C., Wang, L. P. & Martínez, T. J. Automated Code Engine for Graphical Processing Units: Application to the Effective Core Potential Integrals and Gradients. *J Chem Theory Comput* **12**, 92–106 (2016).
27. Titov, A. v., Ufimtsev, I. S., Luehr, N. & Martinez, T. J. Generating efficient quantum chemistry codes for novel architectures. *J Chem Theory Comput* **9**, 213–221 (2013).
28. Ufimtsev, I. S. & Martínez, T. J. Quantum chemistry on graphical processing units. 1. strategies for two-electron integral evaluation. *J Chem Theory Comput* **4**, 222–231 (2008).
29. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, (2017).
30. Bokman, S. H. & Ward, W. W. RENATURATION OF AEQUOREA GREEN-FLUORESCENT PROTEIN. *Biochemical and Biophysical Research Communication* **101**, 1372–1380 (1981).
31. Gaigalas, A. K. & Wang, L. Measurement of the Fluorescence Quantum Yield Using a Spectrometer With an Integrating Sphere Detector. *J Res Natl Inst Stand Technol* **113**, 17–28 (2008).

Author Contributions: Y.S. designed the study, modeled the chromophore and protein complexes, performed classical and annealing simulations, and the experiments. A.R.W, and C.M.J performed the umbrella sampling and quantum chemical validation. Y.S. A.R.W, and C.M.J wrote the manuscript. Y.S., A.R.W., C.M.J., T.J.M., and S.G.B edited and revised the manuscript.

Funding: Y. S. was supported by the Knut and Alice Wallenberg Foundation. This work was supported in part by NIH Grant GM118044 (to S.G.B.) and by the U. S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through Advanced Computing (SciDAC) program and the Chemical Sciences, Geosciences, and Biosciences Division of the Office of Basic Energy Sciences, Office of Science, U. S. Department of Energy, AMOS program (to T. J. M.).

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgements: We thank the Stanford Research Computing Center for the computer resources used in this project. We thank Drs. Chi-Yun Lin and Matt G. Romei for valuable discussions. We thank Dr. Chu Zheng for helping us with expressing, mutating, and purifying proteins. We also thank Jacob Kirsh for assistance with protein purification and for valuable discussions.

Supporting Information for

Simulation-guided engineering of split GFPs with efficient β-strand photodissociation

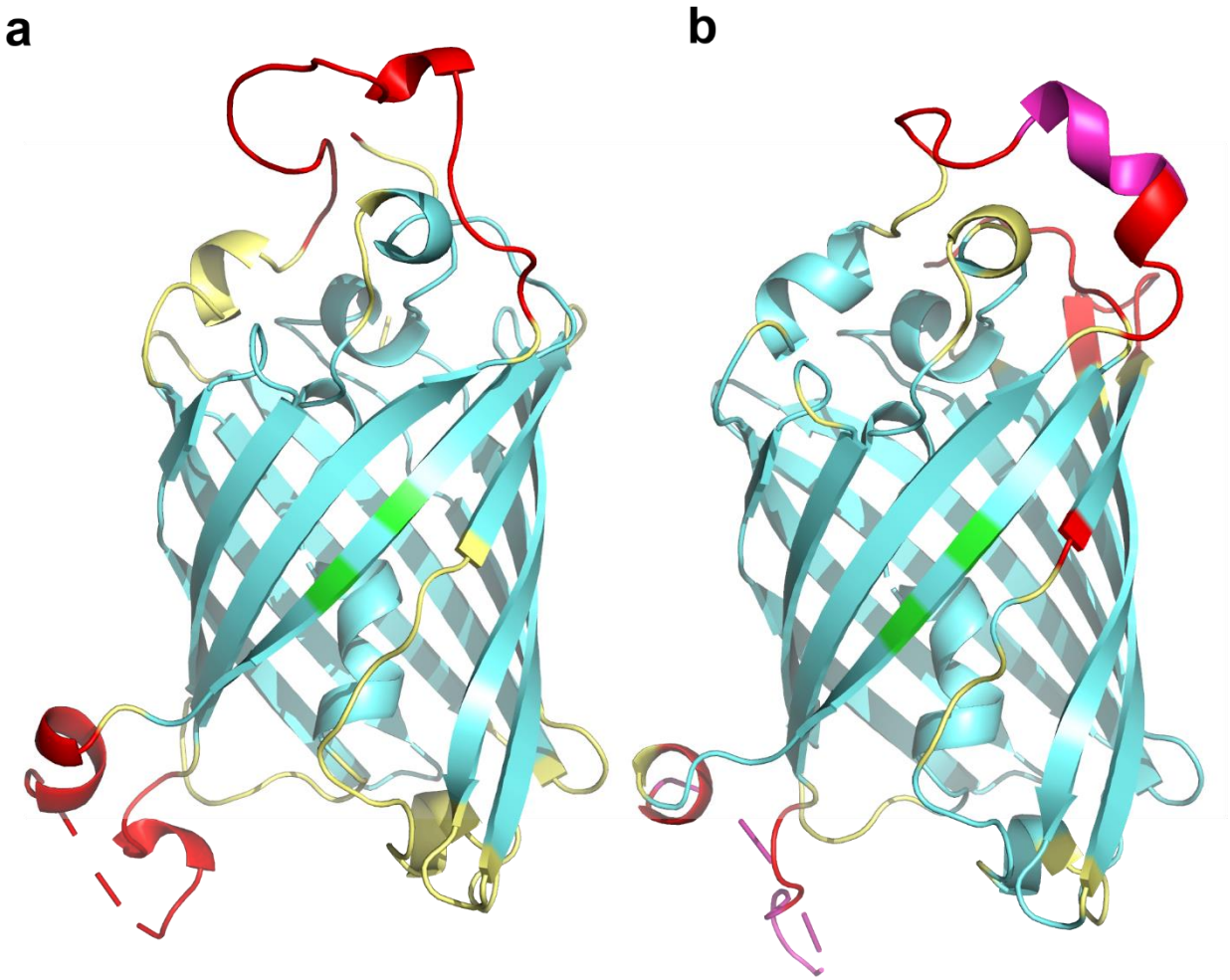
Yasmin Shamsudin[1,2], Alice R. Walker[1,3], Chey M. Jones[1], Todd J. Martínez[1] & Steven G. Boxer[1]**

Correspondence to: sboxer@stanford.edu and yasmin.shamsudin@kemi.uu.se

This PDF file includes:

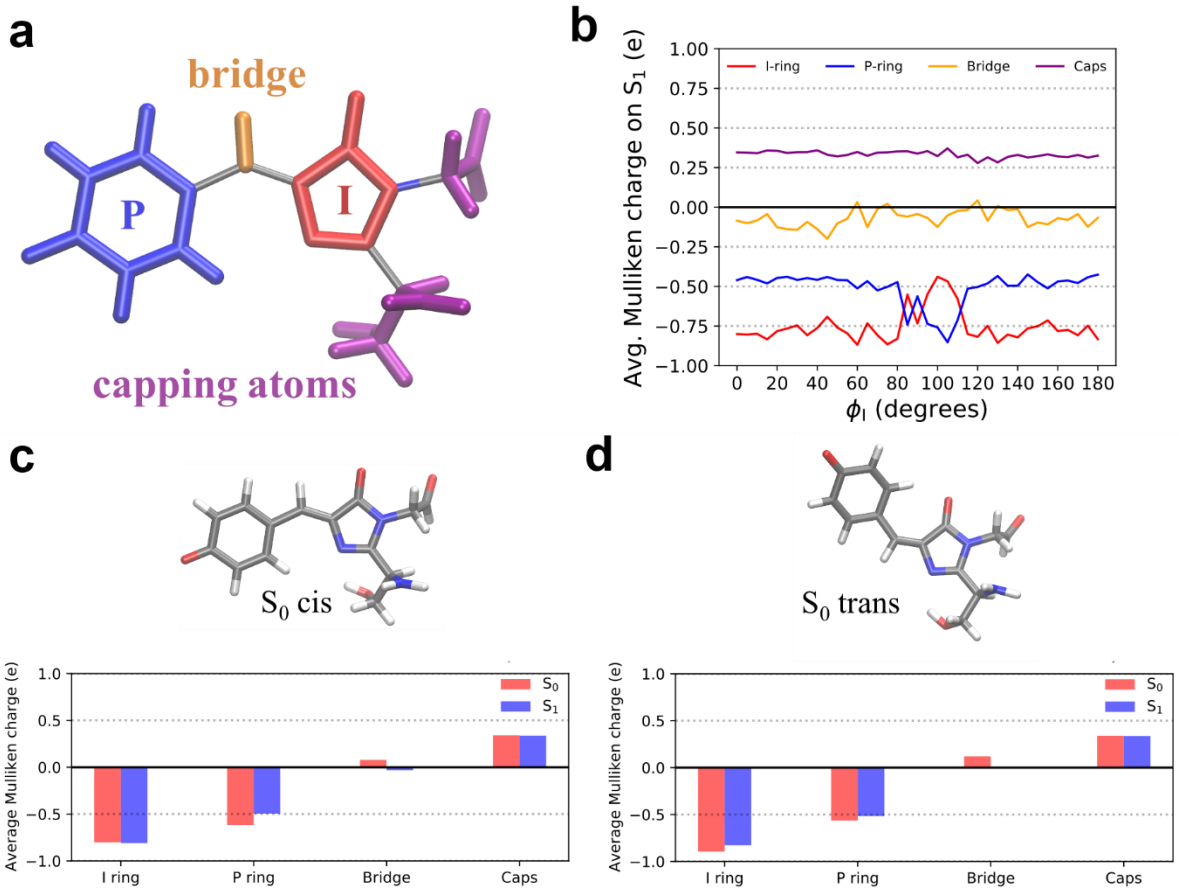
Table of Contents

Fig. S1 RMSD of replicated *cis* and *trans* simulations



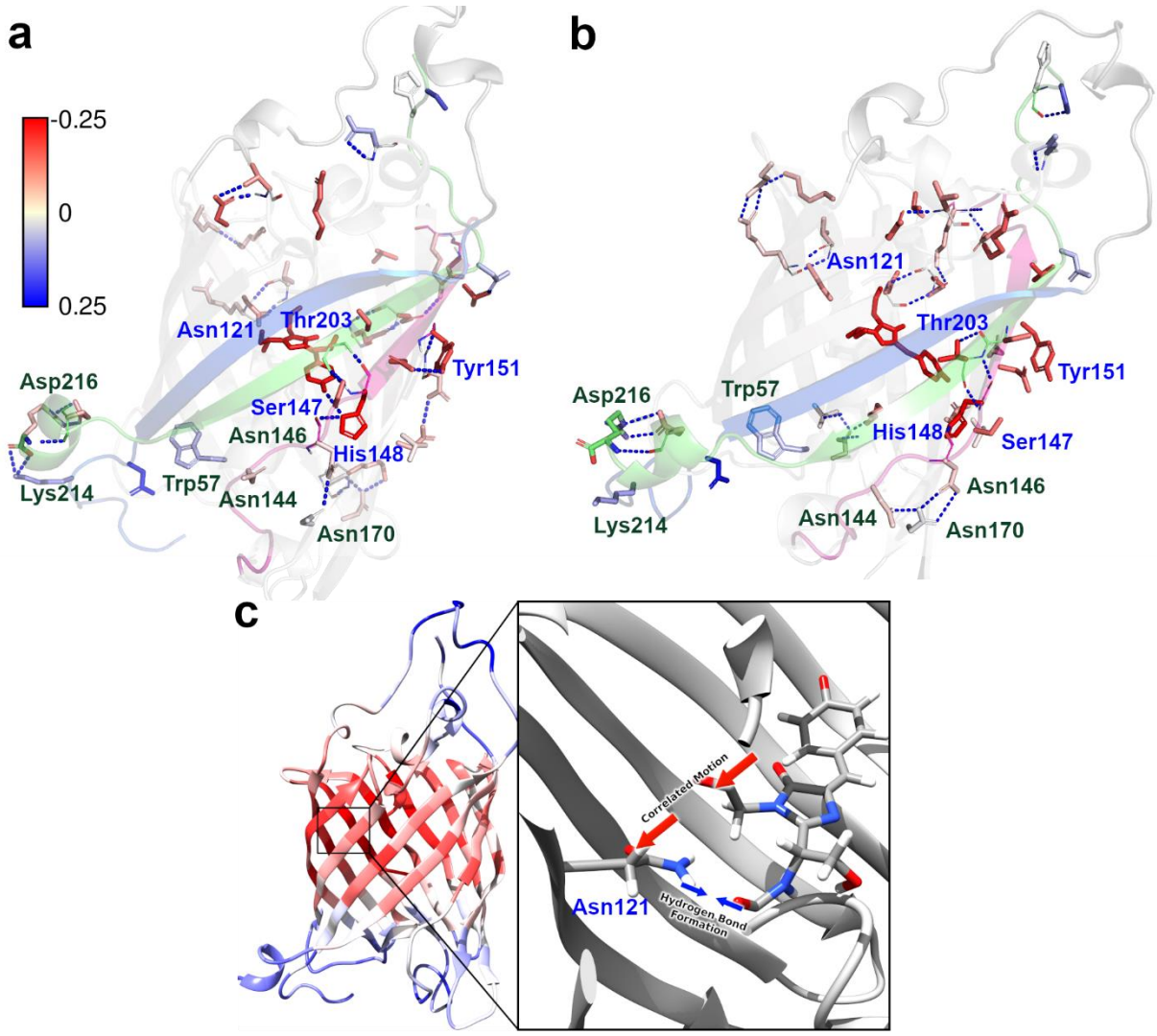
(a) *cis* and **(b)** trans cleaved GFPs after simulations at 300 K for 1μs, color-coded according to the RMSDs between replicates. Cyan: less than 1 Å. Yellow: 1 – 2 Å. Red: 2 – 15 Å. Magenta: 15 – 25 Å. The dotted lines indicate the cleaved loop (residues 45 – 73). Thr203 and Thr205 on strand 10 are shown in green for reference.

Fig. S2 Electronic states



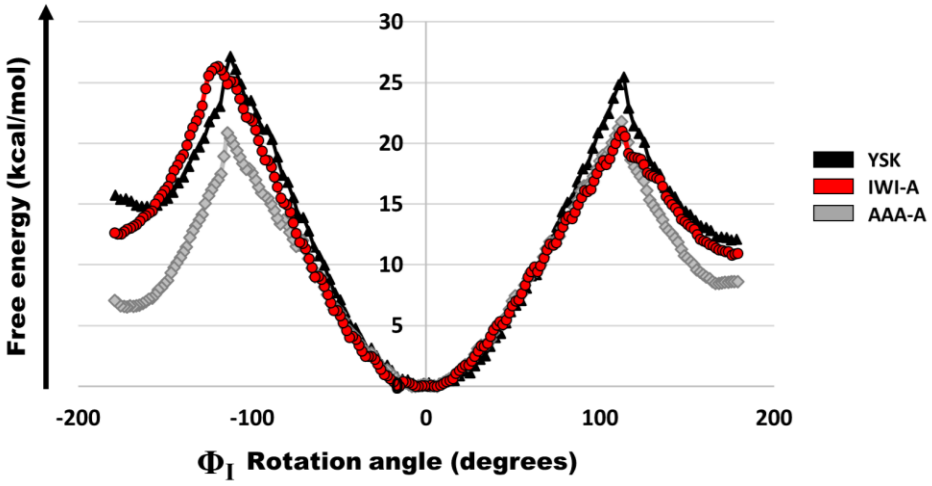
(a) *cis* chromophore colored by moiety: P-ring (blue), bridge (orange), I-ring (red). **(b)** Average Mulliken charge of atoms in each moiety as a function of ϕ_I. Structures were optimized on S_1 using the α(0.64)-SA3-CASSCF(4,3)/6-31G* level of theory. Mulliken charges for moieties of the **(c)** *cis* and **(d)** *trans* forms of the GFP chromophore. Shown structures and charges correspond to S_0-optimized *cis* and *trans* configurations in the GFP protein scaffold. Starting structures for *cis* and *trans* optimizations were obtained from ϕ_I=0° and ϕ_I=180° umbrella sampling windows, respectively. Average Mulliken charges were computed by averaging the atomic charges calculated across all atoms in their respective moiety. QM/MM optimizations were performed using the α(0.64)-SA3-CASSCF(4,3)/6-31G* level of theory.

Fig. S3. Hydrogen bond networks in the *cis* and *trans* complexes



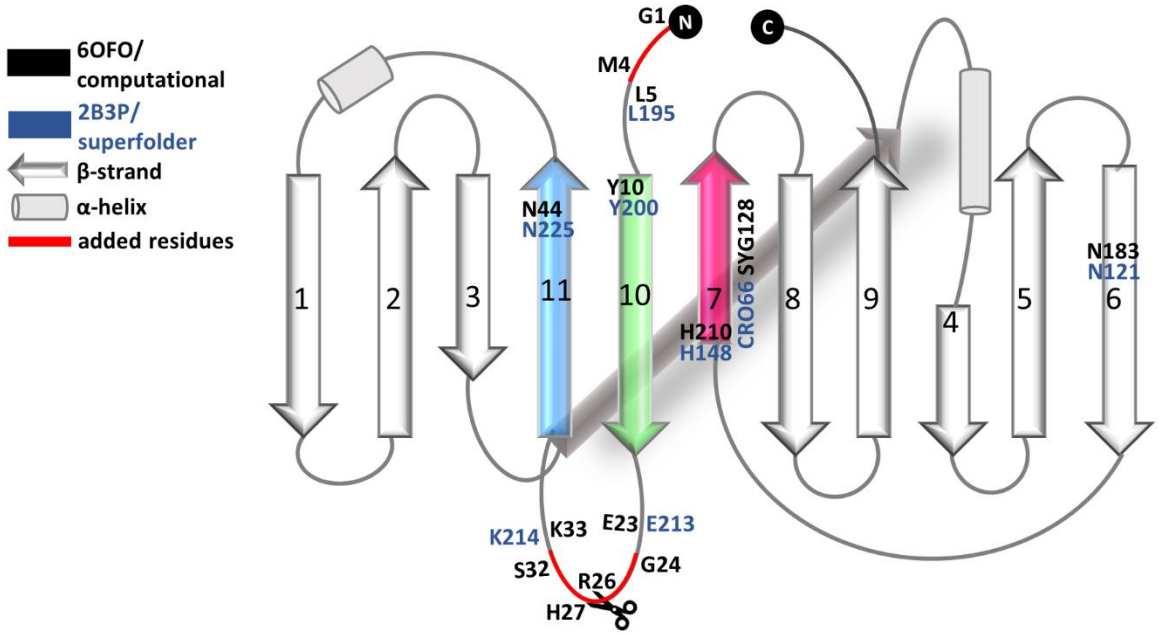
Summary of differences in hydrogen bonding and correlation of motion between the umbrella sampling from the chromophore $\phi_{I,,}$angle windows from 0-90° (top left), and from 90-170° (top right). Residues shown in stick are colored by their difference in correlation (blue, rotating with) or anticorrelation (red, rotating against) to the chromophore as the system moves from the midpoint of the isomerization to fully *trans*. Labeled residues are either directly associated with the chromophore (blue), or part of the allosteric network (green). The overall differences in correlated motion as the system moves from *cis* to *trans* is shown on the bottom left, mapped onto a representative protein structure. The specific mutation chosen from the blue (correlated) region, Asn 121, is shown as a zoom on the bottom right panel, indicating the direction of correlated motion in blue arrows and the competing transient hydrogen bond formation shown in red.

Fig. S4. PMF plots



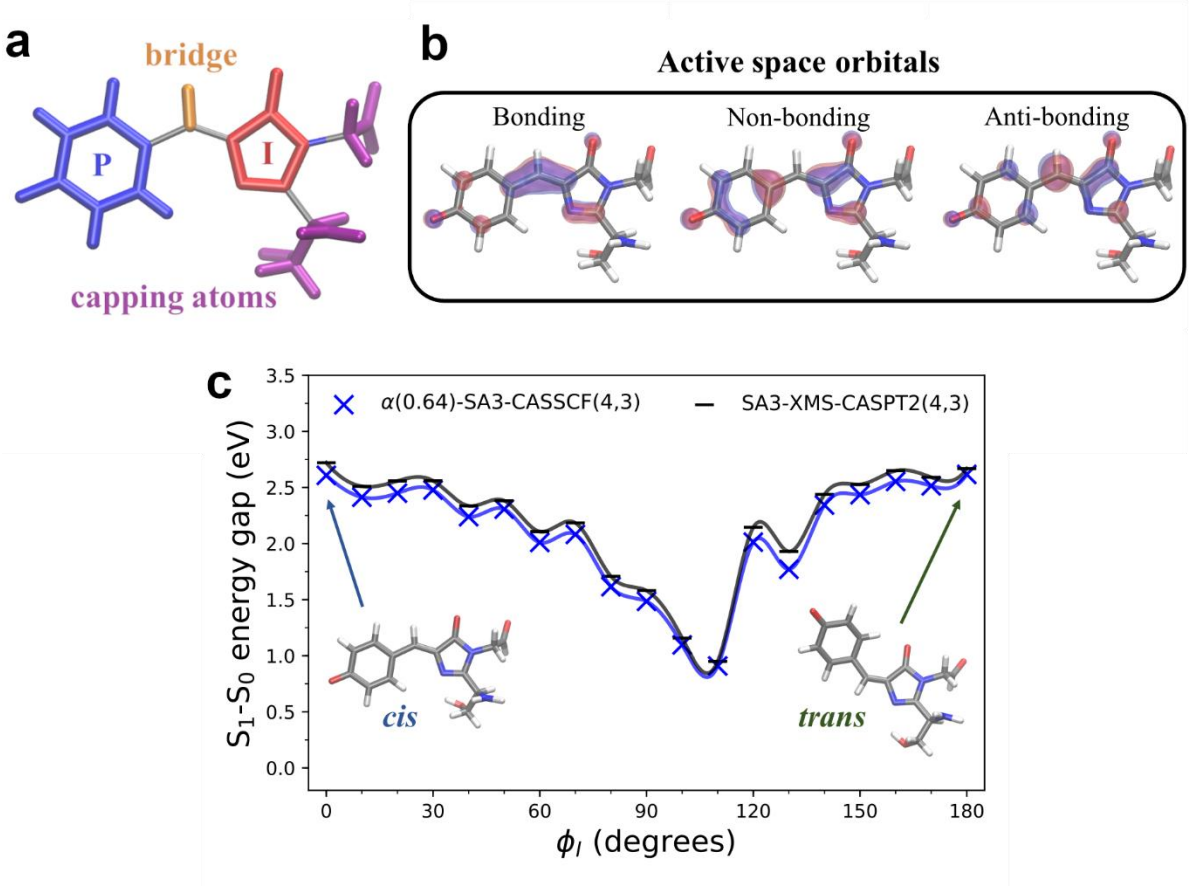
PMF of ground-state *cis*-to-*trans* isomerization of the chromophore in the reference structure YSK (black triangles), the expressed quadruple mutant IWI-A (red circles), and the AAA-A mutant (grey circles), along one-bond-flip coordinates, rotating ϕ_I. Errors computed by bootstrap analysis in WHAM via Monte Carlo (100 trials) in free energy range from 0.1 to 0.5 kcal/mol due to large force restraints.

Fig. S5. Schematic topology of split GFP



Schematic topology of split GFP, comparing the numbering of our computational model (and the 6OFO crystal structure) with the superfolder GFP (sfGFP) (PDB ID: 2B3P). Computational numbering is displayed in black while sfGFP numbering is shown in red (see Table S3). Red lines indicate residues added in the sacrificial loop or due to circular permutation, which are absent in sfGFP. Circles indicate the N and C terminals. Strand numbers are indicated at the center of the strand. The internal helix is shown in gray. Scissors indicate the cleavage site.

Fig. S6 Electronic state methods



(a) *Cis* chromophore colored by moiety: P-ring (blue), bridge (orange), I-ring (red). **(b)** Active space orbitals of the GFP chromophore at the α(0.64)-SA3-CASSCF(4,3)/6-31G* level of theory. **(c)** Benchmark calculations for the energy gap between the S_0 and S_1 electronic states as a function of twisting along ϕ_I. Energies using the α(0.64)-SA3-CASSCF(4,3) method were obtained from QM/MM optimizations on S_0, and energies using the SA3-XMS-CASPT2(4,3) method were obtained from single point energy calculations on α-CASSCF optimized structures. The 6-31G* basis set was used for both methods. The depicted moieties of the anionic GFP chromophore are used for charge analysis as a function of ϕ_I twisting.

Table S1. Strand exchange rates

Measured pseudo first order strand exchange rates ± standard deviation of replicates for laser-induced strand-exchange experiments using 488 nm laser irradiation at 26 mW in all experiments except the AAA (dark) where no external light source was used. Dark experiments were only done for AAA, as it was the fastest-photodissociating variant, and the best candidate for further exploration.

Mutants	**Rate** $(10^{-4} \cdot s^{-1})$
YSK (Reference)	4 ± 2
ASK	9 ± 3
YAK	16 ± 1
YSA	3 ± 1
YSI	3 ± 1
YAA	14 ± 3
AAI	29 ± 1
AAA	85 ± 3
AAA (dark)	1 ± 1
LII	8 ± 1
LVI	8 ± 1
IWA	13 ± 3
IWI	10 ± 2
IWI-A	8 ± 1

Table S2. Experimental fluorescence quantum yields

Protein	FQY [%]
YSK (Reference)	38 ± 1
AAA	40 ± 1
IWI	42 ± 1
IWI-A	34 ± 2

Table S3. Conversion table between computational models and sfGFP crystal structure

Since the sequence of superfolder GFP (sfGFP) is the standard nomenclature in GFP research papers, we refer to amino acids according to their sfGFP numbering in our discussions. However, since there are differences between sfGFP and our complexes due to the circular permutation, placing the N-terminal on strand 10 and added residues due to the sacrificial loop, this conversion table can be used to convert between the computational models (corresponding to the 6OFO crystal structure sequence) and the sfGFP crystal structure (see amino acid sequence comparison on page 4 in this SI).

strand 10	**sfGFP**
Gly1, Ser2, His3, Met4	N/A
Leu5	195
Tyr10	200
Ser12	202
Thr13	203
Thr15	Ser205
Lys19	209
Glu23	213
Gly24, Thr25, Arg26	N/A
strand 11	**sfGFP**
His27, Ser28, Gly29, Ser30, Gly31, Ser32	N/A
Lys33	214
Arg34	215
Asp35	216
His36	217
His40	Leu221
Glu41	222
Asn44	225
strand 7	**sfGFP**
His 210	148
Asn211	149
Tyr213	151
Other strands	
Cys110	48
Asn183	121
a-helix	
GYS128	CRO66
Cys132	70

N/A: Not available (extra inserted loop in the 6OFO and computational structures)

Table S4. Mass spectrometry

Expected mass was calculated as the average (av.) from the primary sequence using the online PeptideSynthetics Peptide Mass Calculator and reducing the mass to account for the maturation of the chromophore and the loss of the N-terminal methionine in the protein sequence (which is removed *in vivo*).

Mass of intact and cleaved GFP variants.

Mutants	**Expected mass (Da)**	**Observed mass[(a)] (Da)**
YSK (Reference)	30719	30730
ASK	30627	30638
YAK	30703	30716
YAA	30646	30657
IWI-A	30710	30722
AAA	30553	30565
AAI	30596	30607
IWA	30711	30723
IWI	30753	30764
LVI	30666	30694
LII	30680	30693
YSA	30662	28922[(b)]
YSI	30704	28966[(b)]
Truncated protein[(c)] of single, double, and triple mutants	25344	25356
Truncated quadruple mutant IWI-A	25301	25311

a) Proteins of ~30kDa have ± 15 Da deviations, depending on the protonation states.

b) Observed masses were consistent with the mass of the proteins, with the loss of the N-terminal HIS-tag

c) As all mutations were done on the strand that dissociates, the truncated protein after dissociation is the same for all single, double, and triple mutants.

Supplementary text

DNA sequences

Complete sequence for the reference structure, which is also referred to as YSK according to the mutation sites Tyr200 (Y), Ser202 (S), Lys209 (K). Codons for these sites, and Asn121 are highlighted and in bold. The (TAA) ochre stop codon was used:

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGC
TGCCGGATAACCAT**TAT**CTG**AGC**ACCCAGACCGTGCTGAGC**AAA**GATCCGAACGAAGGCACCCG
CGGCAGCGGCAGCATTGAAGGCCGCCATAGCGGCAGCGGCAGCAAACGCGATCACATGGTGCTG
CATGAATATGTGAACGCGGCGGGCATTACCCATGGCATGGATGAACTGTATGGCGGCACCGGCG
GCAGCGCGAGCCAGGGCGAAGAACTGTTTACCGGCGTGGTGCCGATTCTGGTGGAACTGGATGG
CGATGTGAACGGCCATAAATTTAGCGTGCGCGGCGAAGGCGAAGGCGATGCGACCATTGGCAAA
CTGACCCTGAAATTTATTTGCACCACCGGCAAACTGCCGGTGCCGTGGCCGACCCTGGTGACCA
CCCTGAGCTATGGCGTGCAGTGCTTTAGCCGCTATCCGGATCACATGAAACGCCATGATTTTTT
TAAAAGCGCGATGCCGGAAGGCTATGTGCAGGAACGCACCATTAGCTTTAAAGATGATGGCAAA
TATAAAACCCGCGCGGTGGTGAAATTTGAAGGCGATACCCTGGTG**AAC**CGCATTGAACTGAAAG
GCACCGATTTTAAAGAAGATGGCAACATTCTGGGCCATAAACTGGAATATAACTTTAACAGCCA
TAACGTGTATATTACCGCGGATAAACAGAAAAACGGCATTAAAGCGAACTTTACCGTGCGCCAT
AACGTGGAAGATGGCAGCGTGCAGCTGGCGGATCATTATCAGCAGAACACCCCGATTGGCGATG
GCCCGGTGCTGTAA

Note: The superfolder GFP nomenclature for Tyr200, Ser202, Lys209, and Asn121 correspond to Tyr10, Ser12, Lys19, and Asn183, respectively, in the computational models, in accordance with the 6OFO crystal structure (see Fig. S4 or conversion table S3).

Mutations

Codons used for single-point mutations.

Residue	**Codon**	**Residue**	**Codon**	**Residue**	**Codon**	**Residue**	**Codon**
Tyr200	*TAT*	*Ser202*	*AGC*	*Lys209*	*AAA*	*Asn121*	*AAC*
Ala200	GCT	Ala202	GCC	Ala209	GCA	Ala121	GCC
Ile200	ATT	Trp202	TGG	Ile209	ATT		
Leu200	CTT	Val202	GTC				
		Ile202	ATC				

Note: The superfolder GFP nomenclature for Tyr200, Ser202, Lys209, and Asn121 correspond to Tyr10, Ser12, Lys19, and Asn183, respectively, in the computational models, in accordance with the 6OFO crystal structure. (see Fig. S4 or conversion table S3).

Amino acid sequences

Complete sequence for the reference structure, which is also referred to as YSK according to the mutation sites Tyr200 (Y), Ser202 (S), and Lys209 (K). The crystal structure sequence of superfolder GFP (sfGFP) and the quadruple mutant IWI-A included for comparison. Note that the strand 10 (green) is at the beginning of the sequence in YSK and IWI-A due to circular permutation, compared with sfGFP. The sequences are color-coded as follows: added peptide sequence of the pET-15b N-terminal His-tag (yellow), strand 10 (green), spacer loop (red), sacrificed loop (red italics), strand 11 (blue) chromophore (pink), and Cys (orange). The four mutation sites are highlighted in bold, with the trypsin protease cleavage sites indicated by ▼.

Ref (YSK)	MGSSHHHHHHSSGLVPR▼GSHMLPDNH**Y**L**S**TQTVLS**K**DPNEGTR▼*GSGSIEGR*▼HS GSGSKRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDV NGHKFSVRGEGEGDATIGKLTLKFI**C**TTGKLPVPWPTLVTTL**SYG**VQ**C**FSRYPDHMK RHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLV**N**RIELKGTDFKEDGNIL GHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVL
sfGFP	MSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFI**C**TTGKLPVPW PTLVTTL**TYG**VQ**C**FSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKF EGDTLV**N**RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVE DGSVQLADHYQQNTPIGDGPVLLPDNH**Y**L**S**TQSVLS**K**DPNEKRDHMVLLEFVTAAGI THGMDELYK
IWI-A	MGSSHHHHHHSSGLVPR▼GSHMLPDNH**I**L**W**TQTVLS**I**DPNEGTR▼*GSGSIEGR*▼HS GSGSKRDHMVLHEYVNAAGITHGMDELYGGTGGSASQGEELFTGVVPILVELDGDV NGHKFSVRGEGEGDATIGKLTLKFI**C**TTGKLPVPWPTLVTTL**SYG**VQ**C**FSRYPDHMK RHDFFKSAMPEGYVQERTISFKDDGKYKTRAVVKFEGDTLV**A**RIELKGTDFKEDGNIL GHKLEYNFNSHNVYITADKQKNGIKANFTVRHNVEDGSVQLADHYQXNTPIGDGPVL

Note: The superfolder GFP nomenclature for Tyr200, Ser202, Lys209, and Asn121 correspond to Tyr10, Ser12, Lys19, and Asn183, respectively, in the computational models, in accordance with the 6OFO crystal structure. (see Fig. S4 or conversion table S3).

Peptide sequence of the yellow (YFP) strand

LPDNHYLS**Y**QTVLSKDPNE

The peptide exchanged for the dissociating β–strand, resulting in the formation of a yellow fluorescence protein (YFP), was designed to match the N-terminal β–strand 10 in our reference structure YSK (see below), except for the Thr203Tyr mutation, and was synthesized by Elim Biopharmaceuticals.

Peptide sequence of the reference (YSK) strand, highlighting the mutation sites and the Thr203 position (green).

LPDNH**Y**L**S**T**K**QTVLS**K**DPNE

Electronic states

In order to choose mutations to affect the isomerization pathway, we needed to both (1) validate that the classical umbrella sampling could approximate the protein's response to the chromophore along the path, and (2) that the particular reaction coordinate that we chose for the umbrella sampling led to a reasonable pathway on the excited state potential surface. We applied a combined quantum mechanics/molecular mechanics (QM/MM) partitioning scheme in order to describe electronic properties of split GFP on the ground and excited state surfaces, using a combination of the same forcefield parameters and protein setup as described in the molecular dynamics simulations section of the SI Methods and QM for the chromophore as described below. geometry was optimized from each window of the umbrella sampling twice, once on the ground state and once on the first excited state. In this way, we mapped out the potential energy surface that the classical umbrella sampling approximated, and found specific regions of difference between force field, QM/MM on the ground state, and QM/MM on the excited state. We found that the largest differences in charge and geometry were found near the conical intersection, but that generally the umbrella sampling gave a plausible approximation for the description of the chromophore and, therefore, the motion of protein residues around it.

To determine if our choice of isomerization reaction coordinate was reasonable, we compared our QM/MM split GFP calculations along the umbrella sampling *cis-trans* isomerization coordinate to the latest published reports on the different types of isomerization of HBDI, the GFP chromophore without a protein scaffold[12,13]. We assumed in this work that the chromophore in split GFP primarily undergoes a one bond flip (OBF) mechanism analogous to the most common isomerization mechanism in HBDI, where the chromophore twists about one central dihedral angle as it moves from *cis* to *trans* (**Fig. 2d**).[12,13] There are several possible mechanisms for the *cis* to *trans* isomerization to occur, including OBF about ϕ_I, OBF about ϕ_P and hula twist, among others. In all of these, once the chromophore is excited, the single and double bonds across the bridging carbons transition to a delocalized electronic structure, lengthening the double bond and creating flexibility in the bridge. As the different possible dihedral angles twist, and the energy gap between the ground and excited state decreases, the bridging carbon pyramidalizes with its hydrogen in the center of the bridge. At this point, the chromophore has reached a conical intersection, and can either relax back to its original *cis* form or to the *trans* form (**Fig. 1c**). It is well known that as the chromophore rotates along ϕ_I, the energy gap between the ground and excited state should decrease as the angle approaches 90° (from either the 0° or 180° directions).[12,13] Therefore, to validate our choice of OBF about ϕ_I to determine mutagenesis targets, we assumed that a reasonable umbrella sampling would approximate these features.

In our QM/MM optimizations based on the umbrella sampling, we observe twisted intramolecular charge transfer from the I-ring to the P-ring as ϕ_I approaches 90°. These optimizations indicate a small region close to the conical intersection with a substantial difference

in charge and geometry (ϕ_I=90-100), but minimal differences from ϕ_I=0-80 and 110-180 degrees (**Fig. S1**).

These findings are consistent with the known behavior of the chromophore, indicating that the structural sampling performed by umbrella sampling is sufficient for capturing these characteristics when analyzed at the QM/MM level despite the discrepancy in the charges on the chromophore close to the conical intersection. **Fig. S1** illustrates that these sampled structures capture this behavior on the (a) ground and (b) excited states, suggesting that the chromophore and its environment are behaving reasonably well during this sampling procedure.

Umbrella sampling simulations

Because the protein environment surrounding the chromophore is asymmetric with respect to bond isomerization, we generated potential of mean force (PMF) profiles associated with clockwise and counterclockwise rotations along ϕ_I to determine the energy barrier to isomerization (**Fig. S3**). We investigate both clockwise and counterclockwise rotation since the pocket around the chromophore is relatively open, and, as shown in **Fig. S3**, there is not a large difference in energetic barrier to rotation in either direction despite the protein scaffold. The free energy barrier between the *cis* and *trans* configurations in the reference structure was ~25 kcal/mol (**Fig. S3**), with the *trans* isomer ~15 kcal/mol above the *cis* isomer, and both energies significantly larger than the ~15 kcal/mol barrier and ~2 kcal/mol difference for the anionic chromophore in solution.[14] Because the chromophore is anchored to the protein, and a relatively rigid protein scaffold is more constricting than a mobile solvent environment, the large barrier associated with our PMF is reasonable. We recomputed the umbrella sampling pathways for the quadruple mutant IWI-A, and a theoretical AAA-A (**Fig S3**) mutant that includes the original AAA mutant plus the Asn121Ala mutation. AAA-A did not express measurable amounts and was only simulated. We found PMF barriers of around 21 kcal/mol in the positive direction and 26 kcal/mol in the negative direction for IWIA, and 22 kcal/mol in the positive direction/21 kcal/mol in the negative direction for AAA-A. Both indicate a substantial energetic lowering for the barrier to isomerization in the positive direction, supporting our hypothesis that the Asn121Ala mutation chosen from the umbrella sampling would have an effect on the isomerization. Higher level calculations are likely needed to move beyond a qualitative assessment of the excited state surface.

Rotational correlation analysis and hydrogen bond occupancies

To compute the relationship of motion between the protein residues and the chromophore and determine possible mutagenesis targets, the center of mass of each residue was calculated and the correlation averaged over the umbrella sampling simulation windows with cpptraj. This generated a matrix of values where a positive value indicated a correlated motion. For example, each residue is exactly correlated with itself, and has a value of +1. A negative value indicates anticorrelation, i.e., residues that are moving in the opposite direction. A value close to zero indicates no correlation between the residue motion. We posited that residues that are strongly associated with the motion of the chromophore as it twists could be involved with the higher energetic barrier to isomerization in the protein as compared to free in solution. To narrow our list further, we then extracted the correlation vector associated with the chromophore relative to each other residue, i.e., if residues moved with, or out of the way, as the chromophore rotates (highly correlated or anticorrelated, respectively), and computed the difference as the chromophore moves from *cis* to *trans* (when ϕ_I =0-90° (**Fig. S2a**) vs when ϕ_I =100-180° (**Fig. S2b**)). This uncovers protein residues that are associated strongly with or against the motion of the chromophore, and that change their motion substantially as the chromophore moves from *cis* to *trans*. We combined this analysis with computed changes in hydrogen bonding occupancies over the course of the umbrella sampling to generate a list of potential mutagenesis targets that could affect the isomerization. Our chosen point mutation, Asn121, shows both a correlated motion with the chromophore and a competing transient hydrogen bond that forms during isomerization (present for ~20% of the isomerization pathway) (**Fig. S2c**). The correlated motion is mitigated by the formation of this hydrogen bond—we therefore hypothesized that eliminating the hydrogen bond via point mutation would allow the isomerization to proceed more easily, since there would not be a competing hydrogen bond interaction that would interfere with this motion.

Mutations of Asn121Ala

Although this work focuses on the structural features of strand-dissociation, we also predicted lowered *cis*-to-*trans* isomerization barriers for the Asn121Ala mutation. Although we expressed both AAA-A and IWI-A, only the expression of IWI-A yielded enough protein for further experiments. Although it displayed modestly lower fluorescence quantum yields, we did not observe faster dissociation or higher yields of the YFP product, compared with the corresponding triple mutant IWI. With a larger population of the *trans* complex, more strand-exchange is expected, although the rate should be similar to IWI if the strand-exchange step is still rate limiting and unaffected by the Asn121Ala mutation. However, the similar final YFP product concentrations of IWI and IWI-A indicate that the mutation affects chromophore rotation in both directions, thus also affecting the rate of thermal relaxation (**Fig. 1c**) back to the *cis* conformation in the ground state, resulting in relatively similar strand-exchange rates. Nonetheless, the observed FQYs in combination with the umbrella sampling simulations suggest that our predictions were qualitatively correct and are useful for predicting sites that can affect the isomerization pathway. We therefore suggest that exploring the motion of surrounding residues as they relate to the

twisting of the chromophore during isomerization indicate promising directions for future mutagenesis targets to improve *trans* product yield. Further mutations could be chosen in a similar fashion to improve isomerization yield for photoswitchable protein applications, i.e., mutating sites that show changes in correlated motion and hydrogen bonding as the chromophore moves from *cis* to *trans*. Apart from the Asn121Ala mutation, our simulations indicated that mutations at Thr205 and His148 could also affect isomerization rates. Since we have demonstrated the ability of reliably predicting useful mutation sites, these sites could be good candidates for future investigations of single sites that could affect both isomerization and strand-dissociation. In the particular case of split GFP, while isomerization is not the rate limiting step, we find that we were able to affect the FQY and thus potentially the isomerization yield in this way. That said, this is more broadly applicable to FP photoswitching in general rather than primarily split GFP strand dissociation. That said, the motion of the dissociating strand is anticorrelated with the motion of the chromophore, meaning that as the chromophore twists, the dissociating strand is pushed away. This provides a possible link between the *cis* to *trans* isomerization motion, and the subsequent series of hydrogen bond changes that lead to eventual strand dissociation.