# Miniaturizing GFP

**Jack Miller**

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
degree of Master of Science in Chemistry

Department of Chemistry and Biomolecular Sciences
Faculty of Science
University of Ottawa
Ottawa, Ontario, Canada

**Abstract**

The green fluorescent protein (GFP) has enabled researchers to visualize a wide range of cellular processes, from protein expression to metastasis. However, its size (27 kDa) can disrupt localization and association of GFP-tagged proteins. Here, we aim to address this limitation by designing a miniature GFP (<20 kDa) that conserves its chromophore-forming pocket within a shortened beta-barrel fold. Using machine learning-assisted protein design, we have produced nineteen miniature GFPs, averaging 19 kDa each, that display varying levels of expression. These small GFPs display similar excitation and emission wavelengths to wild-type GFP, albeit with fluorescence reduced by four orders of magnitude due to low quantum yield and inefficient chromophore maturation. To improve brightness, we utilized random mutagenesis but were unable to isolate improved variants due to the detection limit of our selection method (FACS) being too high to distinguish miniaturized GFP fluorescence from background cellular fluorescence at the desired wavelengths. Our results show that while machine learning can be used to miniaturize GFP, this process leads to impaired function.

**Acknowledgements**

The creation of this document and the research described herein wouldn't have been possible were it not for the support of a select group of people. For that reason, I would like to formally recognize them.

The first person I would like to thank is Dr. Roberto Chica. Roberto gave me the opportunity to pursue my Master's in his lab and has been there throughout the entire process, offering his mentorship, counsel, and pleasant conversation about the state of the world. Under his tutelage, I have experienced tremendous growth, not only professionally as my research skills were honed, but also personally, as he has provided an environment conducive to self-exploration outside of the lab. As I write this, I'm nestled snugly in the splendor of the Rocky Mountains, surrounded by elk, snowshoe hares, and powdery peaks ripe for the shredding, an experience that wouldn't have been possible without Roberto's ongoing support.

To the other members of the Chica Lab, both past and present, I owe a debt of gratitude for enriching the last two years. The many late nights and early mornings demanded by science were made all the more enjoyable by having such a great *ensemble* of labmates. A special note of thanks goes out to Drs. Rojo Rakotoharisoa and Niayesh Zarifi for showing me the ropes, Serena Hunt for all her experimental know-how and philosophical discussions, Ben Smith for being a voice of reason, Pavlo Ignatusha for bringing the jams, Dr. Cindy Klaus for her Hilfsbereitschaft, Ilya Dementyev for his tech wizardry, and lastly Dr. Hang Pham for being a constant source of morale and baked goods.

Of course, I would be nothing without my parents, Sheila and Randy Miller. They have been rocks throughout my entire life, and any achievement of mine belongs to them as well. They have made countless sacrifices that can never be fully repaid, but hopefully, this Master's is at least the next step on the path towards buying them a retirement villa in the south of France.

**Table of Contents**

**List of Tables**



**List of Figures**

**List of Equations**

**List of Schemes**

**List of Abbreviations**

AF2 – AlphaFold2

avGFP – GFP from *Aequorea victoria*

CASP – Critical Assessment of Structure Prediction

CD – Circular Dichroism

CPD – Computational Protein Design

D.I.T. – Digital Integration Time

DDPM – Denoising Diffusion Probabilistic Model

dNTP – Deoxynucleotide Triphosphate

FACS – Fluorescence-Activated Cell Sorting

epPCR – Error-Prone PCR

FP – Fluorescent Protein

FRET – Förster Resonance Energy Transfer

$F_U$ – Fraction Unfolded

GEF – Genetically Encoded Fluorophore

GFP – Green Fluorescent Protein

IMAC – Immobilized Metal Affinity Chromatography

$k_C$ – Rate Constant of Competing processes

$K_D$ – Dissociation Constant

$k_{EC}$ – Rate Constant of External Conversion

$k_F$ – Rate Constant of Fluorescence

$k_{IC}$ – Rate Constant of Internal Conversion

$k_{ISC}$ – Rate Constant of Intersystem Crossing

LB – Luria Bertani

MPNN – Message Passing Neural Network

MLAPD – Machine Learning-Assisted Protein Design

MSA – Multiple Sequence Alignment

MT1 – Mutant Trajectory 1

MT2 – Mutant Trajectory 2

MW – Molecular Weight

PBS – Phosphate Buffered Saline

PDB – Protein Data Bank

PI – Propidium Iodide

pLDDT – predicted Local Distance Difference Test

pLM – Protein Language Model

$R_g$ – Radius of Gyration

RMSD – Root Mean Squared Deviation

SEC – Size Exclusion Chromatography

TB – Terrific Broth

TICT – Twisted Intramolecular Charge Transfer

$T_M$ – Melting Temperature

QY – Quantum Yield

**List of Supplementary Information**

**Statement of Authorship**

All research described in this thesis was conducted by the author, except for the preparation of chromophore-containing structures and their flexible-backbone repacking in *Chapter 3.1 Iterative Computational Design of a Miniature GFP*, which was carried out by Dr. Roberto A. Chica, and the Flow Cytometry detailed in *Chapter 3.4.4 Miniaturized GFPs can be Enhanced through Mutagenesis*, which was performed by Dr. Vera Tang.

## Chapter 1. Introduction

### 1.1 Discovery of Green Fluorescent Protein and its Subsequent Applications

The capacity to effectively treat and prevent diseases in humans and other species under our stewardship is fundamentally reliant on our understanding of both the homeostatic and pathological states of organisms, encompassing not only the biological system but also the surrounding environment. Many genetic disorders are due to mutations in protein-coding regions of DNA, which when expressed as diseased protein products can cause misfolding, instability, perturbed active sites, and impaired interactions within protein complexes[1]. In cases such as these, a refined molecular understanding is required to understand the given phenotype. Since antiquity, humans have theorized about the sources of these illnesses but have been limited by the tools available to them at the time. This limitation has often contributed to faulty beliefs being perpetuated for years, if not millennia as is the case with Miasma[2], before the invention of the microscope or Louis Pasteur's swan neck flasks[3]. Fortunately, the evolution of scientific understanding is driven by the refinement or rejection of theories as new insights and technologies emerge. In this regard, the advent of Green Fluorescent Protein (GFP) has been a pivotal advancement, revolutionizing our ability to study molecular and cellular processes and enabling precise modulation of these mechanisms for both research and therapeutic applications.

GFP, or avGFP, is a bioluminescent protein found in *Aequorea Victoria*, a hydrozoan jellyfish native to the North American Pacific Coast[4]. It was the first of its kind, alongside aequorin, to be isolated and characterized, thanks to the pioneering work of Osamu Shimomura *et al*. in the 1960s[5], a discovery that ultimately earned him the Nobel Prize in Chemistry in 2008. While the discovery of GFP itself is remarkable, its true significance lies in its application as a genetically encoded fluorophore (GEF)[5], a groundbreaking technology that has transformed the study of molecular and

cellular processes, greatly advancing the field of molecular medicine[6]. GEFs enable the visualization of individual molecules within live cells with exceptional temporal resolution and on rapid timescales[7]. These fluorophores are incorporated either directly into an organism's genome[8] or into plasmid DNA[9], offering a versatile tool for a broad range of applications. GEFs can be used to monitor the successful integration of target genes or the localization of their products when tagged or placed under a common promoter[10,11], detect transient protein expression[12], visualize and quantify dynamic changes in cellular structures by targeting specific components thereof[9,11], serve as biosensors for monitoring parameters such as pH, calcium levels, and redox states[13], and facilitate high-throughput cellular population enrichment with techniques like fluorescence-activated cell sorting (FACS)[14,15].

Although GFP was initially isolated in the 1960s, its potential as a GEF was not fully recognized until the early 1990s[13], following its successful cloning by Prasher *et al*. in 1992[16]. In 1994, GFP's first recombinant application was demonstrated through the tracking of protein expression in *Escherichia coli* and *Caenorhabditis elegans*[17], the latter of which was immortalized on the cover of *Science* in February of that year[13]. Shortly thereafter, articles containing engineered GFP variants began to emerge in scientific literature. In August 1994, blue and cyan variants of GFP were reported[13,18,19], followed by an S65T variant in 1995, which exhibited a fourfold increase in emission intensity and distinct excitation and emission peaks, in contrast to the dual peaks observed with the wild-type GFP[20,21]. A yellow GFP variant was also introduced in 1996[22]. Beyond spectral modifications, GFP has been further optimized for use as a GEF through engineering efforts aimed at enhancing its monomericity[23–25], stability[26,27], and maturation rates[28,29]. Additionally, the p$K_a$ of GFP's intrinsic chromophore, a trio of cyclized residues responsible for its photoactivity, has been engineered for dissociation at various pH values[25–27], thereby

broadening its utility as a biosensor. As of the time of this writing, GFP continues to be modified and applied in innovative ways as researchers strive to address increasingly complex scientific questions[30–33], further validating the Nobel Committee's 2008 decision.

### 1.2 Structural Analysis of Green Fluorescent Protein

The structure of GFP has become iconic, characterized by 11 β-strands that fold cylindrically to form an elegant, antiparallel β-barrel, often referred to as a β-can[34] (**Figure 1a**), that is nearly symmetrical along any axis. The protein is comprised of 238 amino acids with a molecular weight (MW) of 27 kDa, and dimensions of 42 Å x 24 Å, with a radius of gyration ($R_g$) of 18.14 Å ± 0.82[5,35]. Each end of the β-barrel is capped by a short α-helix, with a third, kinked helix running coaxially through the structure. Three key residues, Ser65, Tyr66, and Gly67, are located centrally on this axial helix. Upon autocatalytic cyclization, these residues form the 4-(p-hydroxybenzylidene)imidazolidin-5-one chromophore (**Figure 1c**), which is responsible for GFP's fluorescence[36]. This thesis utilizes the S65T variant mentioned earlier, so Thr65 will be referenced going forward. The chromophore is securely housed within the protein, shielded from fluorescence-quenching solvent by the extensive hydrogen-bonding network of the β-barrel[37,38], and to a lesser extent by the capping helices.

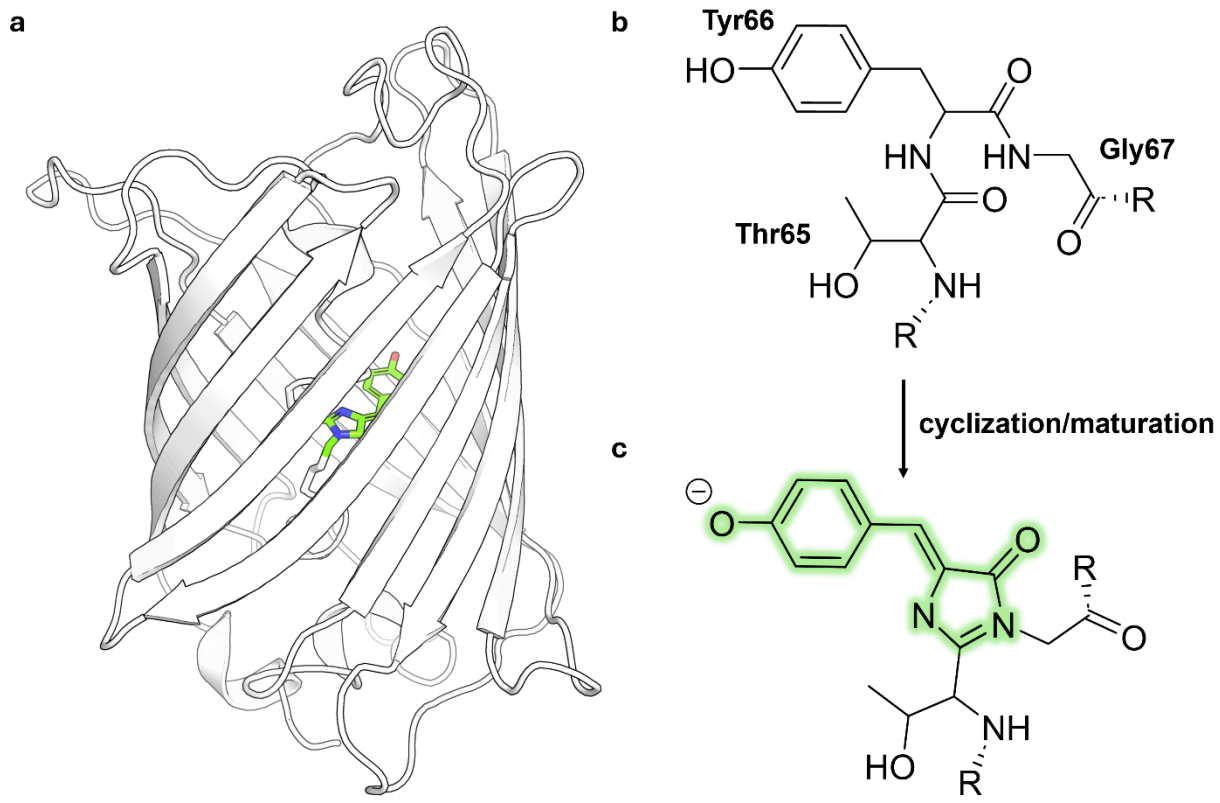



**Figure 1. The Green Fluorescent Protein and its chromophore.** (a) Crystal structure of GFP as seen with a fully formed chromophore, represented by sticks (PDB: 1EMA[34]). (b) The three chromophore-forming residues Thr65, Tyr66, and Gly67. (c) The mature 4-(p-hydroxybenzylidene)imidazolidin-5-one chromophore. The conjugated π electron system responsible for light absorbance at 488 nm can be seen highlighted in green.

While GFP is most useful in its monomeric form, wild-type GFP in *A. victoria* is thought to exist alongside aequorin as part of a heterotetramer[39], and exhibits a weak tendency to form homodimers (**Figure 2a**) with a dissociation constant ($K_D$) of 100 μM[36,40]. This characteristic of multimerization is common to nearly all fluorescent proteins (FPs)[39], regardless of their species of origin, as exemplified by the tetrameric DsRed (**Figure 2c**) found in corals of the *Discosoma* genus[41,42]. Interestingly, DsRed is one of many FPs that share a similar topology with GFP[43], yet are found in non-bioluminescent organisms[42]. This has led to the hypothesis that the GFP-like fold initially evolved to facilitate fluorescence for purposes other than bioluminescence, such as

enhancing resistance to solar radiation or increasing photosynthetic efficiency for endosymbionts through red-shifted light[42,44]. In these cases, the chromophore absorbs high-energy light from the sun and emits lower-energy photons, demonstrating the Stokes Shift, a phenomenon further discussed in *Chapter 1.4 On Fluorescence and its Properties*. The bioluminescence observed in *A. victoria*, facilitated by GFP, is a relatively recent adaptation[42] following the same principle, but with the chemiluminescent aequorin providing the high-energy photons instead of sunlight[45,46].

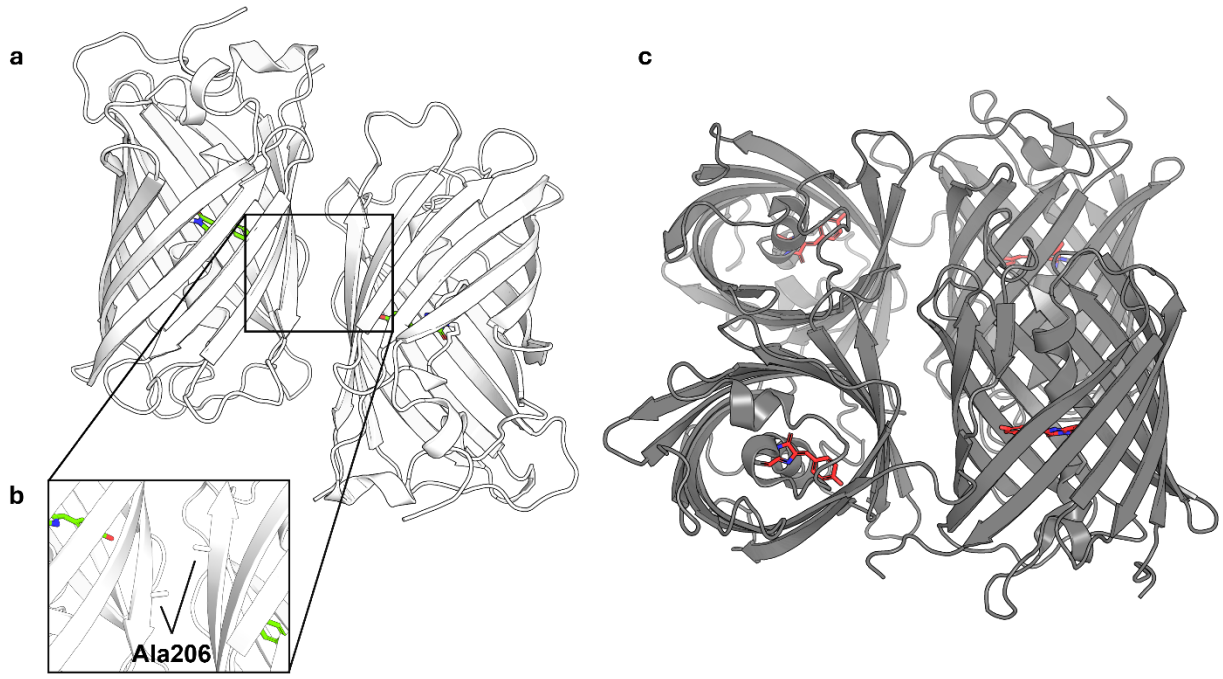



**Figure 2. Different Multimeric States of Fluorescent Proteins.** (a) Dimeric GFP (PDB: 1GFL[47]). (b) The dimerization interface of GFP. An A206K substitution increases the dissociation constant ($K_D$) of GFP and is used in monomeric variants, while also improving folding[48]. (c) The red fluorescent protein DsRed (PDB: 1G7K[49]) in its tetrameric oligomerization state. Red fluorescent proteins adopt the β-barrel fold, and are classified as GFP-like proteins.

GFP demonstrates remarkable stability, particularly for a protein sourced from the cold waters of the Pacific Ocean. It retains fluorescence up to temperatures of 65°C, with only a 50% reduction in fluorescence observed at 78°C[36,50]. While folding and chromophore maturation are negatively

affected by temperatures above 25°C, once the chromophore has matured, further increases in temperature have minimal impact on the protein's structure or fluorescence[36]. With that said, this temperature sensitivity at biologically relevant temperatures limits GFP's utility for *in vivo* studies. To address this limitation, numerous GFP variants have been engineered, typically by substituting bulky residues with smaller ones to enhance folding at higher temperatures[28,36]. Additionally, GFP fluorescence is sensitive to changes in pH, although it remains stable within a broad pH range, from 4.5 to 11[36]. The mechanism behind this sensitivity is tied to the chromophore maturation process, to be discussed in *Chapter 1.3 The Chromophore and its Role in Fluorescence*. At low pH, fluorescence is quenched due to protonation of key residues, whereas at high pH, most fluorescence loss results from protein denaturation, followed by quenching by the surrounding solvent[36]. In response, more acid-tolerant GFP variants have also been developed to expand the range of applications where GFP can be effectively utilized[51].

For fluorescence to occur, the preservation of GFP's archetypal fold is essential[52,53]. Numerous studies involving deletions, insertions, and circular permutations of GFP have demonstrated that residues 2–232 constitute the minimal domain required for fluorescence, with significant modifications being permissible only within the loop regions[54–56]. Alteration of the β-barrel structure therefore remains a significant challenge, but one that must be overcome for any substantial structural remodeling of GFP.

### 1.3 The Chromophore and its Role in Fluorescence

What makes GFP so useful is that its chromophore is formed as part of an autocatalytic post-translational modification. This reaction occurs spontaneously without the need for cofactors or molecular chaperones, requiring only molecular oxygen for complete maturation of residues Thr65, Tyr66, and Gly67 into the 4-(p-hydroxybenzylidene)imidazolidin-5-one chromophore[36] (**Figure 1c**). Although the precise mechanism of chromophore maturation remains unclear, it is widely accepted that protein folding and the reversible cyclization of the chromophore must occur prior to subsequent maturation via oxidation and dehydration reactions[57]. GFP becomes fluorescent in minutes[57], with the rate-limiting step being chromophore maturation[58] .

GFP's specific fold is critical to overcoming the entropic and enthalpic barriers to the cyclization of Thr65, Tyr66, and Gly67[57]. A tight-turn conformation on GFP's central helix aligns Gly67's lone electron pair with the $\pi^*$ orbital of Thr65's carbonyl carbon[57], positioning it for nucleophilic attack (**Figure 3**). This tight-turn conformation also precludes the formation of 9 out of 12 possible hydrogen bonds on the backbone of the central helix[57]. Together, these arrangements effectively prime the involved residues for cyclization[57].

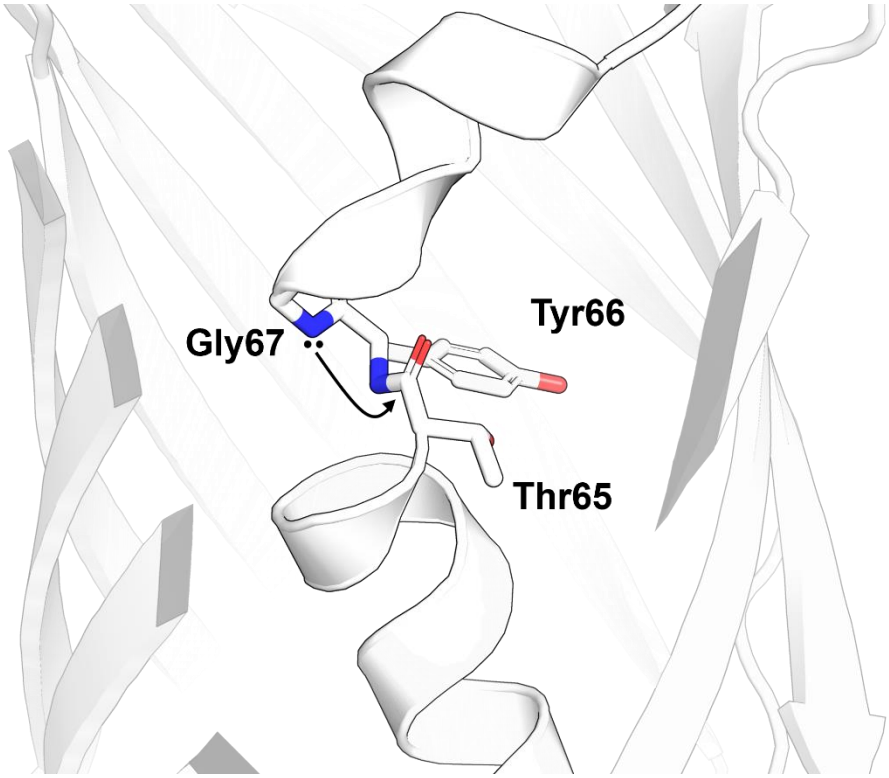
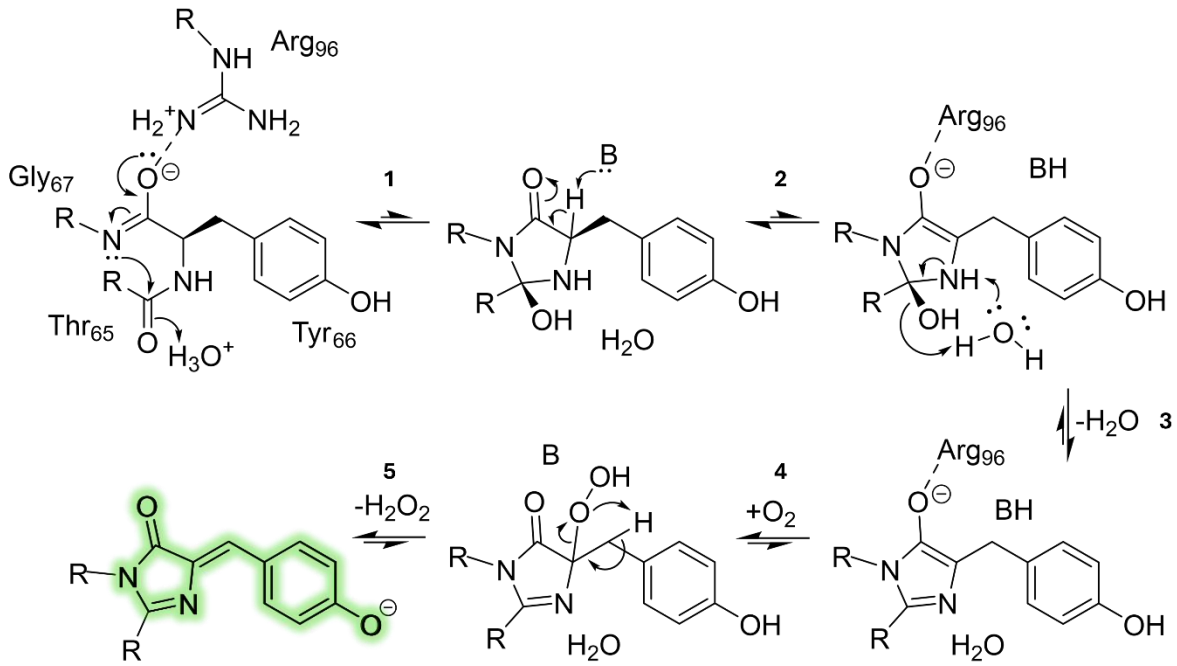


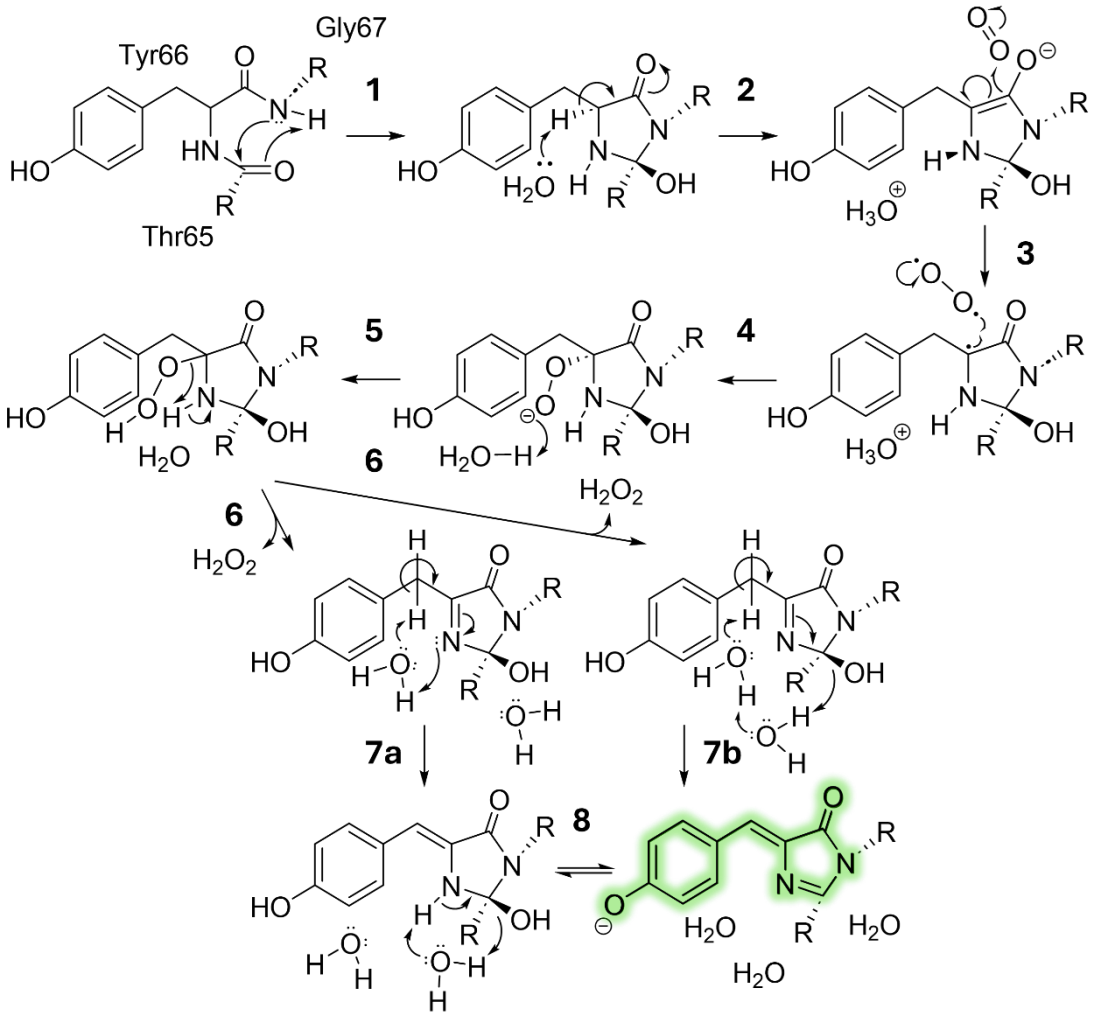

**Figure 3. Tight-Turn Conformation of GFP's Inner Helix.** The tight-turn conformation of GFPs inner helix demonstrated by the uncyclized R96M variant (PDB: 2AWJ[38]). This tight-turn aligns the lone pair of Gly67's amide electrons with Thr65's carbonyl, priming it for attack. Substitution of Gly67 with any other amino acid abrogates fluorescence, likely by hindering cyclization with the introduction of steric hinderance through non-hydrogen R groups[59]. The tight-turn also precludes the formation of 9 out of 12 possible hydrogen bonds, circumventing the need for them to be broken during cyclization, thereby lowering the enthalpic barrier[57].

After cyclization, two potential pathways for maturation are proposed: either maturation follows a dehydration-oxidation order (mechanism A)[60](**Scheme 1**), or it occurs in the reverse order with oxidation preceding dehydration (mechanism B)[61](**Scheme 2**). In mechanism A, the thermodynamically unfavourable cyclized intermediate is thought to be trapped by the dehydration step, whereas with mechanism B oxidation is responsible for this stabilization[57]. For this reason, both pathways are understood as conjugation-trapping mechanisms[57,62,63]. There is substantial evidence that under aerobic conditions mechanism B occurs predominantly in wild-type GFP[57,63], though work done in more recent years supports mechanism A for the S65T variant[64]. Despite the

competing theories, it is generally accepted that both mechanisms may occur in parallel, with the flux of each pathway depending on molecular oxygen availability[57,63] or specific mutations in the chromophore's environment[65]. For this thesis, mechanism A is more relevant, since we will be utilizing the S65T variant mentioned above.




**Scheme 1. The Proposed Mechanism A of Chromophore Maturation in GFP.** Proposed Cyclization-Dehydration-Oxidation mechanism (mechanism A) of GFP's chromophore maturation[60]. Gly67's amide nitrogen is believed to deprotonated by Glu222 either prior to or after cyclization[57], the former is shown here. (1) Thermodynamically unfavourable cyclization of Thr65/Tyr66/Gly67 backbones. (2) Deprotonation of Tyr66$C_\alpha$ and enolate formation, stabilized by electrostatic interactions with Arg96. (3) Reversible dehydration reaction of the main-chain, with Arg96 again providing stabilizing electrostatics. (4) Addition of molecular oxygen to Tyr66$C_\alpha$, forming a hydroperoxide adduct. (5) Deprotonation of Tyr66$C_\beta$ leading to oxidation through the generation of hydrogen peroxide. Adapted from Barondeau *et al.* (2005).




**Scheme 2. The Proposed Mechanism B of Chromophore Maturation in GFP, with Alternative Dehydration Reactions.** Proposed Cyclization-Oxidation-Dehydration mechanism (mechanism B) of GFP's chromophore maturation[61]. (1) Nucleophilic attack of Thr65's carbonyl by Gly67's amine lone electrons. (2) Proton abstraction of $Tyr66_{C\alpha}$ and enolate formation, which would be stabilized by Arg96. (3,4) Addition of molecular oxygen to $Tyr66_{C\alpha}$, forming a hydroperoxide adduct. The complex forms an open-shell singlet diradical complex before forming the closed-shell singlet hydroperoxyl adduct[66]. (5,6) Deprotonation of $Tyr66_N$ leading to oxidation through the generation of hydrogen peroxide. Intermediate 2 is formed here, which has two proposed mechanisms of dehydration. (7a) Formation of the enamine tautomer of Intermediate 2's imine. (7b) Dehydration is initiated by deprotonation of $Tyr66_{C\beta}$ by a water molecule. (8) Proton transfer from $Tyr66_N$ to $Thr65_O$ via an ordered water molecule, leading to reversible dehydration and the mature chromophore. Adapted from Rosenow *et al.* (2004).

In both proposed mechanisms of chromophore maturation, the conserved residues Arg96 and Glu222 play critical roles. Arg96 is essential for protein folding, stability and cyclization, but its most crucial function lies in facilitating chromophore maturation through electrostatic interactions[57]. Substitution of Arg96 with any other residue significantly prolongs the maturation time, though it does not completely prevent it[57]. It is thought that Arg96 stabilizes the chromophore's enolate intermediate by complementing a buildup of negative charge on Tyr66's main-chain oxygen[57]. This lowers the $pK_a$ of Tyr66's Cα proton, allowing it to be easily abstracted by a water molecule, which is polarized by Glu222[57]. This hypothesis is further strengthened by the fact that in an R96A variant, wild-type maturation rates can be restored by reintroducing a positive charge to the region via a Q183R substitution[57]. Beyond polarizing the water molecule for proton abstraction from Tyr66, experiments show that Glu222 serves as a general base in several other proton transfer reactions during the maturation process[67,68]. For example, it is involved in the deprotonation of Gly67's amine nitrogen[57], activating it for nucleophilic attack on Thr65's carbonyl group, thereby initiating cyclization. For Glu222 to act in this role it must mediate its effects through a network of water molecules, many of which can be seen in numerous crystal structures of GFP[57](**Figure 4**).

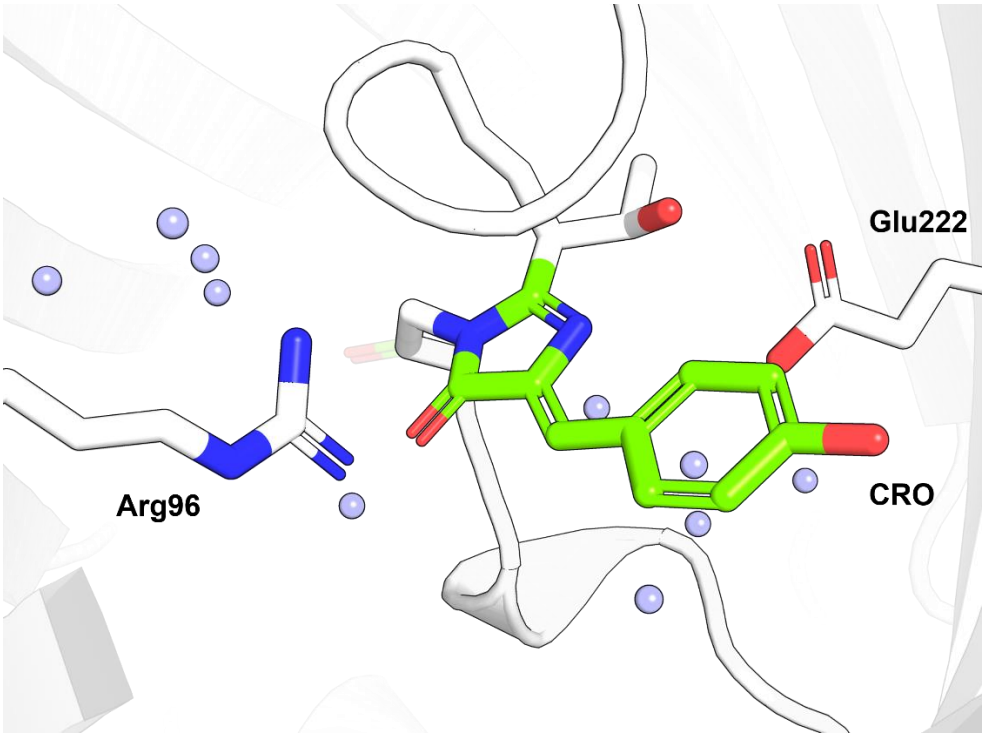



**Figure 4. Structural Water in X-ray Crystallographically Determined Structure of GFP.** Structured water molecules (blue spheres) are critical for chromophore maturation, taking part in a hydrogen-bonding network with Glu222 that's responsible for many deprotonation events. In most crystal structures such as this (PDB: 1EMA[34]), water molecules can be found within close proximity to the chromophore and catalytic Arg96 and Glu222 residues. Molecular Dynamics simulations show that many of these water molecules stay within hydrogen-bonding distance of Arg96, Glu222 and chromophore atoms involved in maturation[69].

One potential drawback of chromophore maturation is the production of hydrogen peroxide, which is integral to the process[57]. The molecular oxygen that is added to Tyr66's Cα must leave as $H_2O_2$ for the imidazolidin-5-one ring to become conjugated[57]. This conjugation enhances the chromophore's ability to absorb light, thereby increasing its fluorescence. Although this $H_2O_2$ doesn't typically lead to cell death, as evidenced by mature GFP being found naturally in *A. victoria*, it does contribute to oxidative stress within the cell[70]. Whether or not this poses a problem experimentally, such as with overexpression of a GFP-tagged protein, is context-dependent and may vary based on the specific experimental conditions.

### 1.4 On Fluorescence and its Properties

Fluorescence is a physical property exhibited by certain materials that allows them to absorb light at one wavelength (colour) and re-emit it at another, generally at a lower energy and longer wavelength, which as mentioned above, is a phenomenon known as the Stokes Shift. Although light can theoretically be re-emitted at higher energy levels, known as an Anti-Stokes Shift[71], this process falls outside the scope of this thesis and will not be explored further. There are various reasons why light is generally re-emitted at lower energy levels, and they can be better understood using a Jablonski diagram (**Figure 5**).

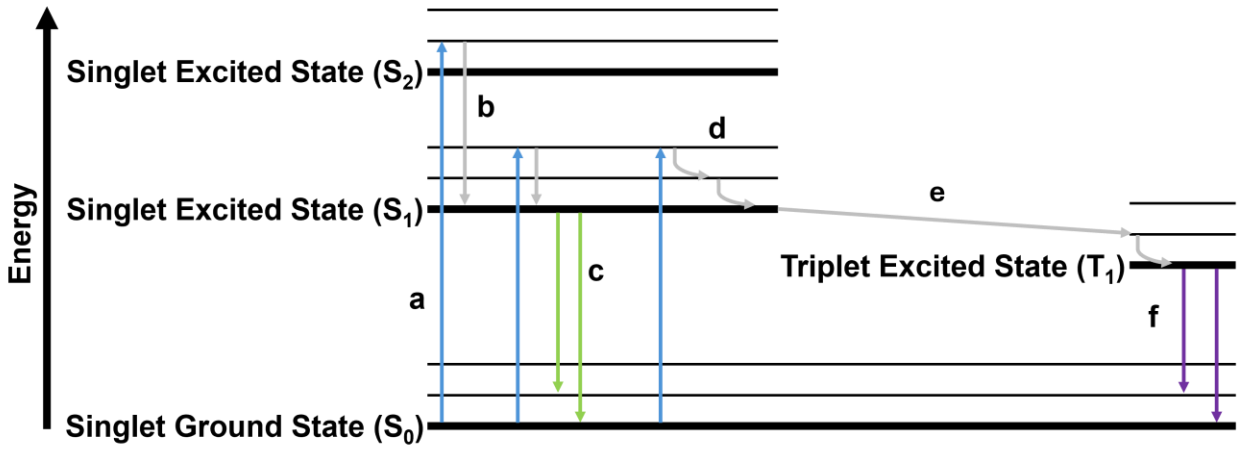



**Figure 5. Different Routes of Energy Decay as Seen in a Jablonski Diagram.** (a) A fluorescent molecule can be excited from the ground state to the excited singlet state upon absorbance of light. (b) If the molecule is excited to a singlet state higher than $S_1$, it must undergo Internal Conversion via non-radiative decay until reaching $S_1$. (c) Once at $S_1$, the molecule can emit a photon and return to the ground state with an equivalent loss of energy, so called fluorescence. (d) If the molecule is excited to a higher vibronic substate (thin lines), it must relax down to the lowest vibronic substate before fluorescing, or in the case of (e) undergoing Intersystem Crossing from an excited singlet state ($S_1$, $S_2$ or higher) to an excited triplet state. (f) If the molecule is in an excited triplet state, it must relax down to its lowest vibronic substate before releasing a photon in what is known as phosphorescence. Adapted from Geddes (2016)[72].

When a fluorescent molecule absorbs a photon (light) it transitions from a ground state ($S_0$) to an excited singlet state ($S_1$ or $S_2$), where all electrons are paired[72]. The energy absorbed can be determined by **Equation 1**, where (E) represents energy, ($h$) is Planck's constant, ($c$) is the speed of light, and ($\lambda$) represents wavelength[73]. For this transition to occur, the energy of the absorbed photon must match the energy difference between the ground state and the excited state. An excited molecule in the $S_1$ state can return to the ground state by releasing a photon[72]. Each state has multiple vibronic substates (thin lines in **Figure 5**), corresponding to different vibrational energy levels that slightly change the electronic state of the molecule[72]. Depending on the energy of the absorbed photon, a fluorescent molecule can be excited from $S_0$ to any vibronic substate of $S_1$ or $S_2$[72]. If a molecule is excited to a higher $S_1$ vibronic substrate, the molecule can relax down vibronic substates to the bottom of $S_1$ through a variety of mechanisms, such as vibrational relaxation, before releasing a photon and returning the molecule to any vibronic substate of $S_0$ that corresponds to the loss of the photon's energy[72].

$$\mathrm{E} = \frac{hc}{\lambda} \qquad \text{eq. [1]}$$

If a molecule is excited to the higher $S_2$ or $S_3$ electronic states, it must first undergo internal conversion before transitioning to $S_1$[72]. This means that energy must be transferred in a non-radiative process from higher to lower electronic states. Once at $S_1$, the molecule can release a photon and return to the ground state as described above[72]. It may also undergo what is known as intersystem crossing, which occurs when the lowest vibrational energy level of $S_1$ overlaps with a higher vibrational energy level of the triplet state ($T_1$), which contains unpaired electrons[72]. From $T_1$, the molecule can return to $S_0$ by releasing a photon in a process known as phosphorescence[72].

Fluorescence typically occurs within $10^{-9}$ to $10^{-6}$ seconds, whereas phosphorescence occurs on a much longer timescale, ranging from $10^{-4}$ to $10^{2}$ seconds[72], making it less useful for GEF applications[72].

One useful fluorescent phenomenon is that of förster resonance energy transfer (FRET). FRET is the radiationless transmission of energy from an excited chromophore (donor) to another region within the chromophore-containing molecule, to other molecules, or more relevantly, to other chromophores (acceptors)[74,75]. This collision-free transfer occurs without thermal energy conversion and in a distance-dependent manner, via a long range dipole-dipole coupling mechanism[74,75]. The transmission can occur if there is sufficient overlap between the donor's emission and acceptor's molar extinction spectra (**Figure 6**), what is termed the Overlap Integral (OLI), and calculated using **Equation 2**, where ($\lambda$) is the given wavelength, $\varepsilon_A(\lambda)$ is the molar absorption coefficient of the acceptor at that wavelength, and $F_D(\lambda)$ is the donor's normalized fluorescence emission at the same wavelength[74,75]. The FRET mechanism is only valid when a donor-acceptor pair is physically separated by 1-10 nm[74,75]. At distances below 1 nm collisions and complexation would prevail, while beyond 10 nm photo emission from the donor would occur[74,75], which while having the possibility to excite neighbouring molecules via reabsorption, wouldn't constitute FRET. This distance dependency makes FRET an indispensable research tool as a "Spectroscopic Ruler", as the efficiency of its energy transfer follows **Equation 3**[74], where (E) is the efficiency of energy transfer, where (R) is the donor-acceptor distance and ($R_0$) is a characteristic value where there is a 50% probability of energy transfer, termed the Förster distance[75]. By tagging molecules with GEFs, researchers can utilize FRET to probe their spatial relationships.

$$J = \int_0^\infty F_D(\lambda)\varepsilon_A(\lambda)\lambda^4 d\lambda \qquad \text{eq. [2]}$$

$$E = \frac{R_0^6}{R^6+R_0^6} \qquad \text{eq. [3]}$$

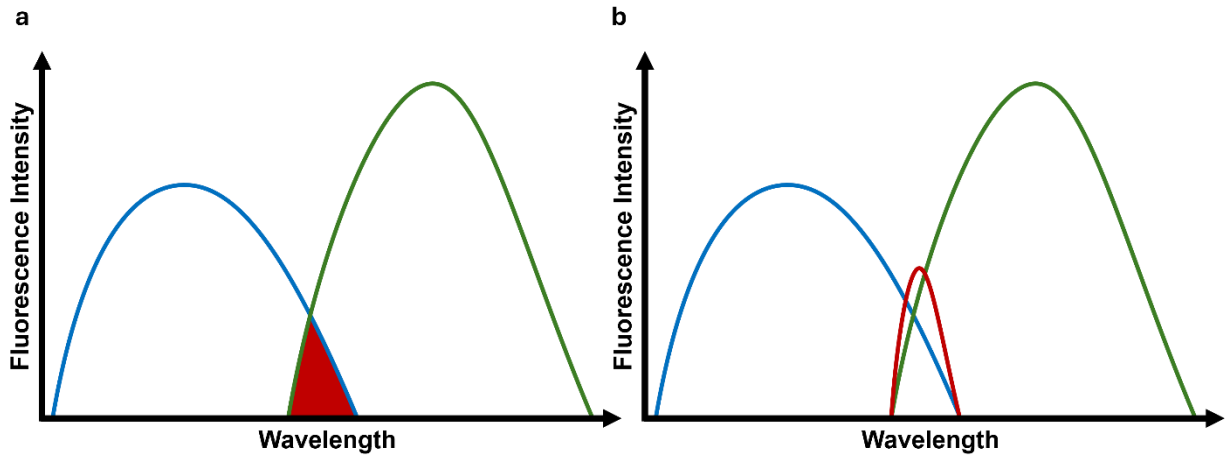



**Figure 6. Illustration of Overlap Required for Förster Resonance Energy Transfer (FRET).** (a) Simplified diagram of the overlap (red) required of a donor chromophore's emission spectrum (blue) and an acceptor molecule's molar extinction spectrum (green) for FRET to occur. In reality, the Overlap Integral (OLI) that describes this spectral region is seen in (b), where the Overlap curve containing the integral (red) is in units of OLI $nm^{-1}$, the donor chromophore's emission spectrum (blue) is in units of $nm^{-1}$, and the acceptor molecule's molar extinction spectrum (green) is in units of $M^{-1}$ $cm^{-1}$. Figure adapted from Medintz and Hildebrandt (2014)[74].

As previously mentioned, excited fluorescent molecules can lose energy through non-radiative decay, where no photon emission occurs. One such mechanism that is particularly relevant and detrimental to GFP fluorescence is twisted intramolecular charge transfer (TICT). When a molecule undergoes TICT, conformational changes in the excited molecule can lead to an intersecting landscape of potential energy surfaces for the excited and ground states, which can promote fast non-radiative decay[76,77]. In the case of GFP, these conformational changes typically occur in the chromophores phenolate moiety. Specifically, the phenol ring undergoes rotation around the β and γ dihedral angles, formed by atoms $N_1$-$C_1$-$C_2$-$C_3$ and $C_1$-$C_2$-$C_3$-$C_4$, respectively[76]

(**Figure 7**). Due to the $sp^2$ planarity of the β dihedral, and bulwarking of the imidazolidin-5-one ring by Phe165 and Thr62, the majority of this effect is due to rotation around the γ dihedral[76]. This rotation leads to the formation of a twisted configuration that favors non-radiative decay, also disrupting the planarity required for fluorescence[78]. TICT is a significant source of fluorescence quenching, where the molecule returns to the ground state without emitting a photon, and it can substantially reduce a molecule's Quantum Yield (QY).

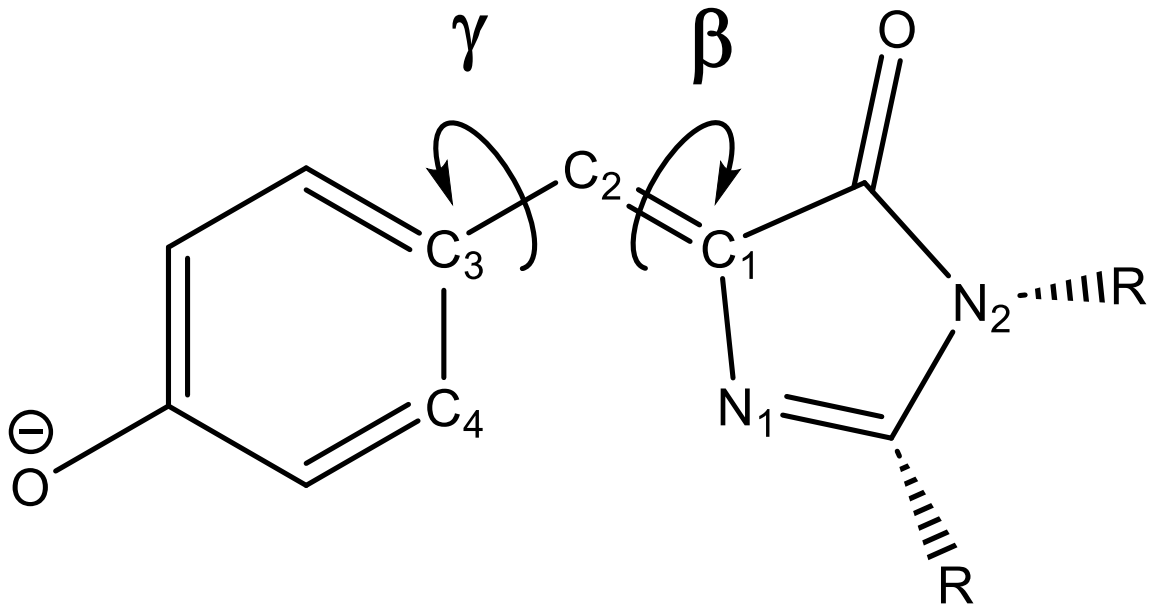



**Figure 7. Hula-Twist Motion Seen in GFP's Chromophore.** The Hula-Twist is a form of twisted intramolecular charge transfer (TICT), in which the excited chromophore undergoes a conformational change that creates an intersecting landscape of potential energy surfaces between the excited and ground states[76]. Due to the overlapping states, rapid non-radiative decay can occur. In GFP, most of the twisting occurs around the γ dihedral angle, since the β dihedral is restricted by $sp^2$ planarity and ther steric hinderance of neighbouring residues. Tight packing of the chromophore-containing pocket can reduce this effect. Adapted from Baffour-Awuah and Zimmer (2004)[76].

QY is defined as the number of photons emitted per 100 absorbed photons[72] (**Equation 4**). This makes it a percentage value that is conveniently represented between 0 and 1, where 0

represents no photon emission and 1 represents 100% photon emission. A more rigorous definition of QY can be expressed by **Equation 5**, where $k_F$ is the fluorescence rate constant, $k_{IC}$ is the rate constant for internal conversion (non-radiative decay to lower vibrational states), $k_{EC}$ is the rate constant of external conversion (energy transferred to solvent), $k_{ISC}$ represents intersystem crossing (singlet to triplet state transition), and $k_C$ is the rate constant for any other competing process, including photodecomposition[79]. In this equation, $k_F$ is the only rate constant that contributes to fluorescent emission, while all the other rate constants contribute to non-radiative decay and reduce the overall QY.

$$\Phi = \frac{\text{n}_{\text{photons emitted}}}{\text{100 photons absorbed}} \qquad \text{eq. [4]}$$

$$\Phi = \frac{k_F}{k_F + k_{IC} + k_{EC} + k_{ISC} + k_C} \qquad \text{eq. [5]}$$

Fluorescent proteins, such as GFP, are typically ranked and reported based on their brightness, which is a value derived from the product of the protein's molar extinction coefficient (ε) and its QY (Φ), as seen in **Equation 6**. The Molar Extinction Coefficient reflects a molecule's ability to absorb light at a specific wavelength and is measured in units of $M^{-1}$ $cm^{-1}$, while QY was discussed earlier and is unitless. To determine molar extinction coefficients, the Beer-Lambert law is generally used (**Equation 7**), where $A$ is the absorbance (unitless), $c$ is the concentration of the protein in solution (M), and $l$ is the pathlength of light travelled through the solution during measurement (cm)[80]. By engineering proteins to increase either their molar extinction coefficient or QY, one can enhance the overall brightness of the fluorescent protein.

$$\text{Brightness} = \varepsilon \times \Phi \qquad \text{eq. [6]}$$

$$A = \varepsilon cl \qquad \text{eq. [7]}$$

### 1.5 Previous Examples of Engineering Smaller GFPs

As discussed in *Chapter 1.2 Structural Analysis of Green Fluorescent Protein*, GFP is quite large, with a molecular weight of 27 kDa. This large size can present challenges when using GFP as a GEF. GFP's length of 238 amino acids (714 nucleotides) can crowd viral vectors, such as commonly used adeno-associated viruses[81], and limit the size of genetic payloads it can be tagged to. Additionally, once expressed, GFP can interfere with the normal functioning of the molecule it is tagging, such as disrupting processes like viral capsid assembly[82], or hindering neuronal axonal movement[55]. For these reasons, as well as the potential for unforeseen negative effects in certain experimental contexts, there is a strong desire to develop smaller GFP variants to minimize these issues while retaining their fluorescent properties.

While GFP has been the subject of extensive engineering efforts[13], these have primarily focused on modifying or improving its spectral characteristics, folding and maturation processes, and stability, as discussed in *Chapter 1.1 Discovery of Green Fluorescent Protein and its Subsequent Applications*. Traditional protein engineering approaches aimed at miniaturizing GFP have been explored, but it has been demonstrated that only fifteen residues can be deleted from GFP without compromising its fluorescence, and these deletions are primarily restricted to the terminal and loop regions[55]. Given this limitation, researchers have at times been forced to develop or utilize techniques to circumvent this size dependency of GFP's fluorescence. Common examples of this include Circular Permutation, to which GFP is incredibly robust, and Protein Splitting (Split GFP), a method that fragments GFP into separate peptides which can be reconstituted into a functional, fluorescent complex through noncovalent heterodimerization[83].

Perhaps the most well-known example of a smaller GFP variant is the miniGFP[84], which has a molecular weight of 13 kDa. While miniGFP is indeed smaller than the original GFP, it requires exogenous flavin to function as its chromophore. One important note is that this protein is an evolved mutant of phiLOV3, a novel Light-, Oxygen-, and Voltage-based flavin-binding protein[84]. Since the miniGFP requires an exogenous chromophore and shares little to no homology with GFP, they may have developed a small green fluorescent protein, but they haven't miniaturized GFP.

More recent and apt examples of GFP miniaturization, which use GFP as a starting template, are the XFPs developed using the Raygun method[85]. Raygun is an example of MLAPD which purports to enable the design of insertions and deletions into a template protein, thereby allowing for miniaturization. Raygun achieves this by employing a single shot Protein Language Model, which uses a probability function to generate candidate sequences. Experimental testing of eight Raygun designs revealed that five exhibited dim fluorescence. The smallest green XFP had a 10.5% reduction in length, with most of this reduction occurring in the loop and distal β-strand regions, unsurprisingly. This outcome aligns with GFP's incompatibility with deletions in more structured regions of the protein, and with Raygun's bias towards deletions in unstructured regions[85].

While Raygun is a successful example of MLAPD being used to miniaturize GFP, many such models exist which can facilitate this aim. What's more exciting is that these models can be combined in novel ways, creating protein engineering pipelines which can be fine-tuned to tackle specific design objectives.

### 1.6 Machine Learning-Assisted Protein Design

In recent years, there has been a rapid expansion in the availability of machine learning models for protein design[86,87], enabling access to regions of sequence and structure that were previously unreachable without labor- and resource-intensive techniques such as chimeragenesis and directed evolution[88]. MLAPD has matured enough that it can effectively tackle three cruxes in the field: novel backbone generation[89,90], inverse folding[91,92], and structure prediction[93,94]. While the design of functional proteins using MLAPD is becoming increasingly feasible[95,96], its effectiveness in this area is less explored. A computational protein design pipeline leveraging these advancements, which is becoming a standard methodology, typically follows three steps: 1) the generation of a backbone using methods such as Flow Matching or denoising diffusion probabilistic models (DDPM), either unconditionally or by scaffolding functional motifs, 2) design of sequences capable of adopting the desired fold, with protein language models (pLM) and message passing neural networks (MPNN) being frequently used, and 3) validation of the design using a deep learning–based structure prediction model (**Figure 8)**.

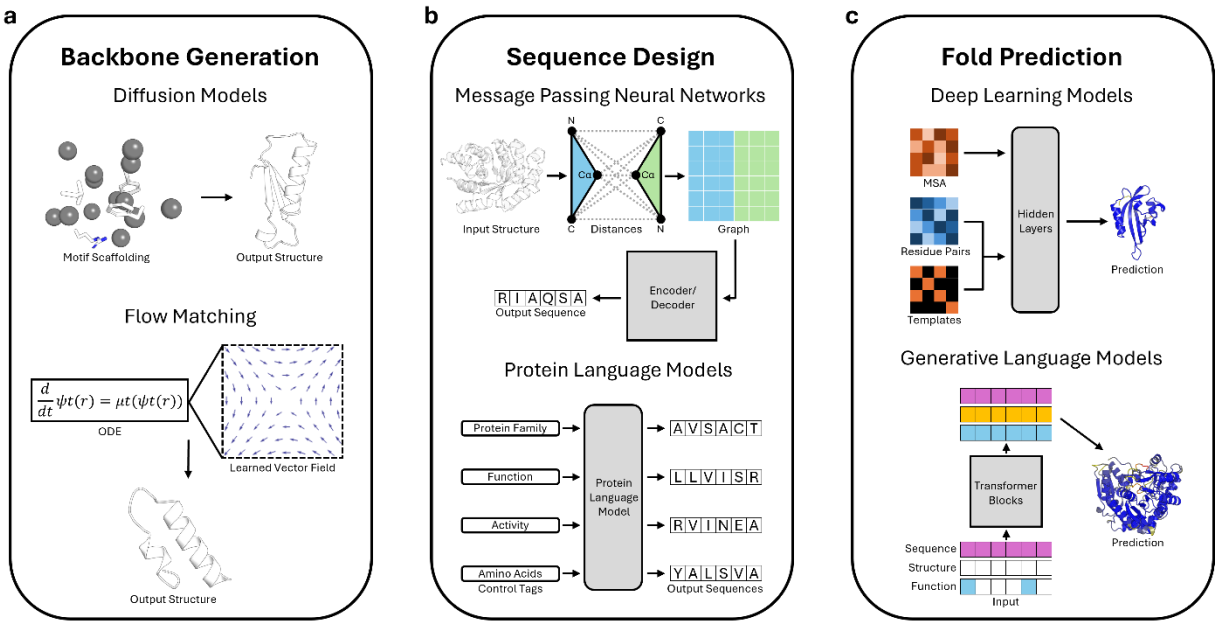



**Figure 8. General Protein Design Pipeline Using Machine Learning.** A typical machine learning–assisted protein design pipeline consists of three main stages: (a) Backbone generation using models such as denoising diffusion probabilistic models or flow matching. (b) Design of amino acid sequences capable of adopting the generated backbone. Message passing neural networks are commonly employed, though alternative approaches such as protein language models are also effective. (c) Sequence validation to assess the likelihood of folding into the target structure. Deep learning models have become the standard for this step, although generative language models and other methods may also be used. Redrawn from Dauparas *et al.* (2022), Jumper *et al.* (2021), Madani *et al.* (2023), and Hayes *et al.* (2025).

In this pipeline, novel protein structures are often generated using Flow Matching methods like FoldFlow-2[97], or DDPMs such as RFdiffusion[89] or Protpardelle[90]. Flow Matching models generate backbones by integrating ordinary differential equations over a learned vector field, having been trained on 2.8 billion proteins from numerous sequence and structure databases[96,97]. DDPMs for protein design are trained to denoise structures from the Protein Data Bank (PDB) which have been corrupted with Gaussian noise[89]. Once trained, DDPMs can iteratively refine initially random noise into realistic protein structures[87]. While earlier DDPMs like RFdiffusion used a rigid N-Cα-C frame to represent residue backbones, more recent models like Protpardelle are capable of incorporating sidechain information, enabling the co-generation of both structure

and sequence[87]. With RFdiffusion All-Atom[98] and RFdiffusion2[99], it's even possible to model ligand molecules. Using this pipeline, structures are either generated unconditionally or by scaffolding around a desired motif. Unconditional structure generation requires minimal user input, typically just the desired sequence length, while motif-scaffolding requires initial structural information to serve as nucleation points. Since models like RFdiffusion only generate structures consisting of backbone atoms, it's necessary to use another method to design the corresponding sequence. Co-generative models such as Protpardelle can bypass this step, although designing structure and sequence separately has been shown to yield comparable, and in some cases superior[100], results in terms of designability and diversity[87].

While pLMs like ProGen[101] are able to successfully design protein sequence, the use of MPNNs for this task has become ubiquitous. ProteinMPNN[91] and LigandMPNN[92] are two examples of MPNNs which have proven to be effective at designing foldable sequences. Both models were trained on datasets derived from the PDB. To determine sequences, ProteinMPNN encodes N, Cα, C, O, and virtual Cβ atoms as nodes, with the pairwise distances between them as edges in a graphical representation of the structure[91], whereas LigandMPNN similarly encodes these atoms, along with distances for ligand atoms involved in protein-ligand interfaces[92]. Once the structure is encoded, both ProteinMPNN and LigandMPNN decode sequences in an order-agnostic, autoregressive manner. By allowing for random sampling of the decoding order, these models can utilize contextual information from previously generated residues, which has been shown to modestly enhance sequence recovery[91]. A major advantage of using these language models and neural networks is that they provide an alternative to traditional physics-based sequence design methods, which often rely on expensive Monte Carlo simulations to evaluate the effects of individual amino acid substitutions on a given design[102].

Neural networks have demonstrated their true potential in the realm of structure prediction. The protein-folding problem, which has stymied researchers since the first protein structure determination with atomic-resolution[103], has largely been addressed by deep learning algorithms such as AlphaFold2 and the generative language model ESM3[93,96]. Similar to the methods previously described, AlphaFold2 was trained using data from the PDB, as well as sequence data from UniProt[104] and metagenomic data from MGnify[105]. ESM3, on the other hand, was trained using data from the PDB, UniProt, and several other sequence and structural databases[96]. While both models have proven to be effective in structure prediction, AlphaFold2 has demonstrated superior accuracy at critical assessments of structure prediction (CASP) events, although versions of ESM3 return predictions much more quickly[106]. The accuracy of AlphaFold2, and the speed of ESM3, can be attributed to key differences in their approaches, AlphaFold2 performs a multiple sequence alignment (MSA), whereas ESM3 does not. This MSA enables AlphaFold2 to incorporate structural data from homologous proteins and pairwise features of input sequences to refine its predictions[93]. The omission of this step in ESM3 not only accelerates predictions but also makes it particularly effective when predicting “orphan” proteins with few homologs[107]. When it comes to protein design, the choice of structure-prediction method depends on the specific experimental conditions and objectives. For instance, for the redesign of a well-characterized protein like GFP, AlphaFold2 might be the preferred tool, provided sufficient computational resources are available for its implementation.

Significant progress has been made in the development of miniaturized GFPs, yet there remains considerable room for improvement. The ultimate goal, or "holy grail", is the creation of a significantly smaller GFP that retains the ability to intrinsically form a chromophore. Achieving this milestone would represent a major breakthrough in GFP engineering and could

unlock novel applications in biological imaging. While still in its early stages, MLAPD and pipelines like the one outlined above show great potential in advancing this objective.

### 1.7 Overview of Project

As discussed earlier in this chapter, a truly miniature GFP could be of great use to researchers in a variety of disciplines. While progress has been made in this area, we believe there is still room for further advancement. Therefore, this project aimed to miniaturize GFP to less than 20 kDa (a 26% reduction), while maintaining the ability of the canonical tripeptide (Thr65/Tyr66/Gly67) to autocatalytically cyclize and mature into the endogenous 4-(p-hydroxybenzylidene)imidazolidin-5-one chromophore. Chapter 2 will explore how this project leveraged recent developments in MLAPD to design candidate sequences, with Chapter 3 exploring the attempted improvement of designs using an iterative design process as well as traditional mutagenesis techniques. Chapter 4 will cover the materials and methods used in this project. Chapter 5 will provide a summary of the findings and suggest future directions for advancing this project.

In Chapter 2 we hypothesized that GFP could be miniaturized to at least 75% of its original size while still retaining spontaneous chromophore maturation. To achieve this, we deleted all but the residues critical for chromophore maturation from an x-ray crystallography-generated *in silico* model of GFP. The remaining key residues were used as the starting chassis for GFP miniaturization, and connecting residues were diffused in place using a generative diffusion model. The result was an assortment of 3D structures, which were then ran through a message passing neural network to design amino acid sequences which could adopt the associated folds. Of the four designs experimentally tested, all were positive for fluorescence, with spectral properties resembling those of GFP, albeit with the best design having a QY 337-fold lower than that of GFP.

In a second round of design, we revisited the first-round designs, optimizing and validating their sequences further while incorporating new developments. Importantly, the second round incorporated a fully formed chromophore during the design process, an approach that was not possible in the first round but was hypothesized to enhance fluorescence. Out of the twelve second round designs experimentally tested, all showed fluorescence with spectral properties similar to those of GFP. The best of these second-round designs had a QY 320-fold lower than that of GFP, but due to a lack of biological replicates this value should be viewed critically.

Chapter 3 concludes with an attempt to enhance our initial designs using Directed Evolution. Despite generating a mutant population with enhanced green fluorescence, this approach did not yield the expected results. The reason for this was that the detection limit of our selection method (FACS) was too high to distinguish the fluorescence of designed FPs from background cellular fluorescence at the desired wavelengths, which hindered our ability to select improved variants and ultimately prevented us from building on beneficial mutations with subsequent rounds of mutagenesis. This was disappointing, but an instrument has been recently identified that appears to have the required sensitivity to evolve these designs, discussed in Chapter 5, offering new hope for future optimization efforts.

## Chapter 2. Designing a Miniaturized Green Fluorescent Protein

### 2.1 Non-Iterative Computational Design of Miniature GFP

Rather than starting from scratch by utilizing or designing a *de novo* protein scaffold and subsequently introducing function, as is typical in computational protein design (CPD), we chose to design a protein that would facilitate spontaneous chromophore maturation by leveraging existing GFP structures (**Figure 9**). This approach was primarily driven by the incomplete understanding of the protein folding and chromophore maturation pathway of GFP, as well as the

limitations of current rational CPD methods, which are not yet sufficiently advanced to address such a dynamic task.

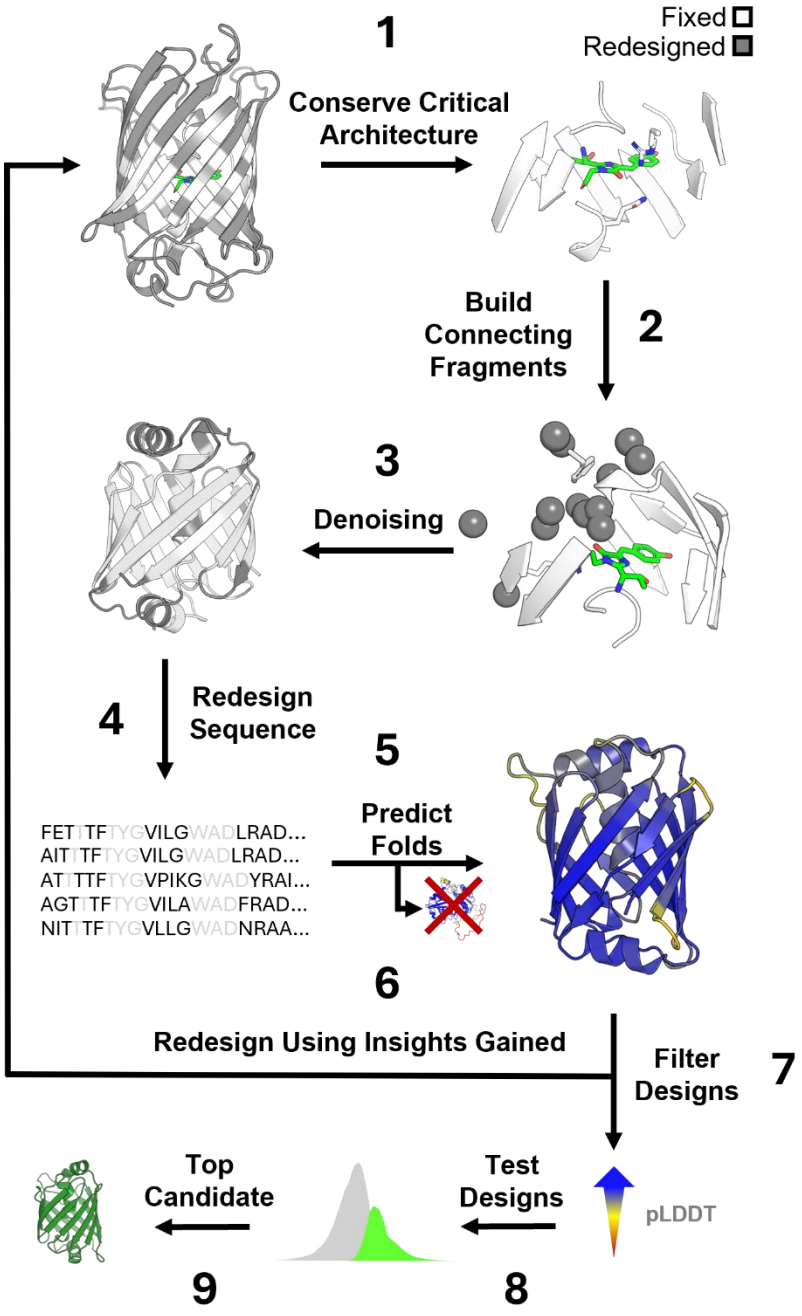



**Figure 9. Computational Design Process for GFP Miniaturization.** (1) Using the crystal structure of PDB: 1EMA[34], residues not within 8 Å of the chromophore were deleted, leaving a spheroid of disjointed residues. (2) Connecting fragments were diffused to connect the spheroid

in a GFP-like fold using RFdiffusion, which (3) denoises a structure from gaussian noise. (4) Sequences for diffused structures were designed using ProteinMPNN, fixing the identity of residues found in the original spheroid. (5) Sequences from ProteinMPNN were submitted to AlphaFold2 for structure prediction. (6) If needed, the design process was repeated. (7) Structures were filtered by AlphaFold2's residue-wise pLDDT scoring before being (8) experimentally characterized and (9) a top design was chosen.

To design a protein capable of facilitating spontaneous chromophore maturation we first turned to the well-characterized crystal structure of Enhanced GFP (EGFP, PDB: 1EMA[34]). We chose EGFP instead of wild-type GFP as a reference due to its simplified excitation spectrum with a peak at 488 nm resulting from the S65T substitution[34]. In contrast, wild-type GFP has two absorbance maxima at 395 and 475 nm[36].

With the 1EMA structure, we pruned away all residues that did not have atoms within 8 Å of the chromophore (**Figure 10**). This allowed us to retain key residues critical for chromophore maturation, including Thr62, Gln69, Arg96, His148, and Glu222, as well as those which likely play a minor role or which tune GFP's emission spectra, such as Thr203[108]. In GFP, the tightly packed chromophore pocket is essential for preventing non-radiative decay; therefore, this 8 Å cutoff was also applied to replicate this packing. By preserving these crucial residues and maintaining their spatial orientations, we hypothesized that we could preserve GFP's ability to fluoresce while minimizing the overall size of the protein.

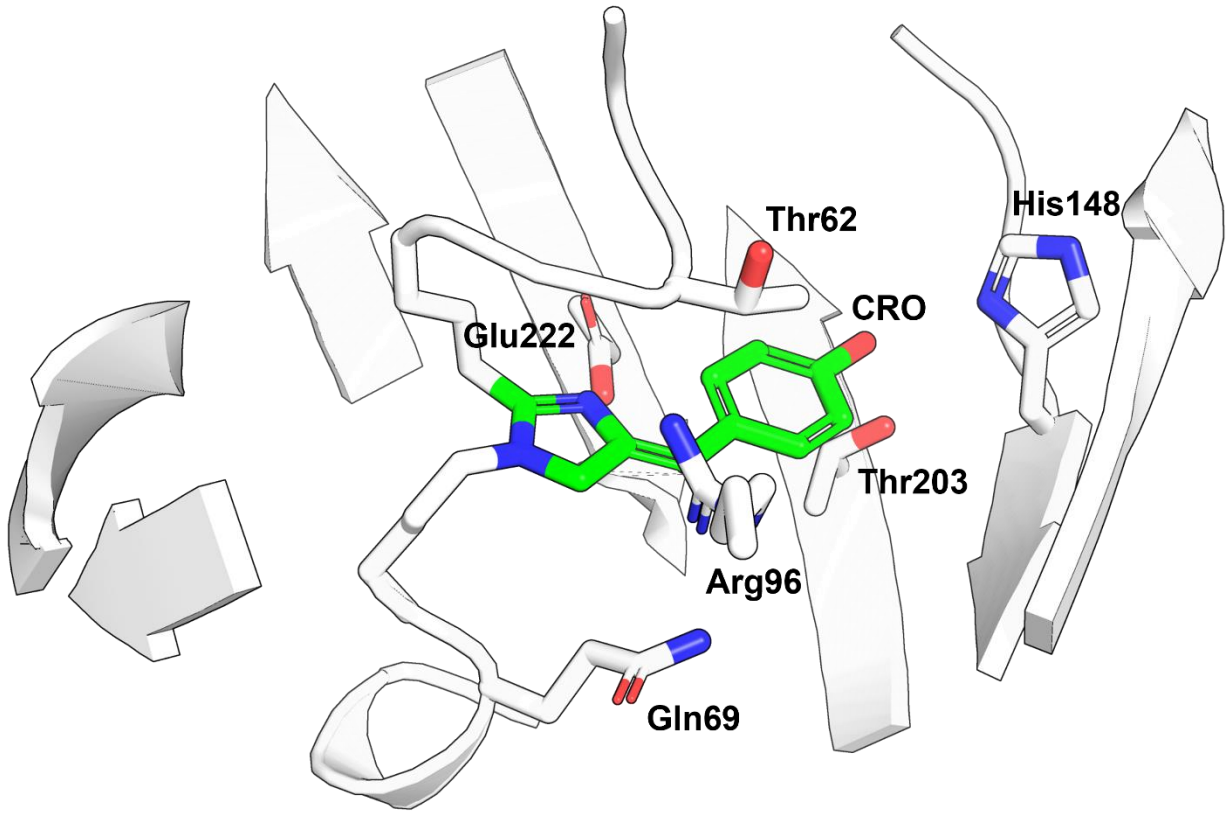



**Figure 10. Spheroid of Disjointed Fragments Preserved Within 8 Å of GFP's Chromophore During Miniaturization.** To maintain GFP's ability to autocatalytically cyclize and mature its chromophore (green) after miniaturization, residues with atoms within 8 Å of the chromophore were preserved. This ensured that not only catalytically active residues like Arg96 and Glu222 were kept, but also other residues which facilitate maturation or fluorescence, like His148 and Thr203. By keeping the entire spheroid, and not just critical residues, further design stages were able to be biased towards more easily achieving a β-barrel fold.

After creating the 8 Å spheroid of disjointed fragments, we needed to transform it back into a monomer. In doing so, this required a delicate balance between ensuring the design would fold properly and not adding excess residues that would compromise the goal of miniaturization. To build connecting structures we turned to the new machine learning software RFdiffusion[89] (**Figure 11**), a generative Denoising Diffusion Probabilistic Model trained to denoise corrupted protein structure representations. RFdiffusion was trained on samples from the Protein Data Bank (PDB) and diffuses a user-defined number of residues in 3D space until it converges on a structure. We

chose to use RFdiffusion instead of Protpardelle since we were planning to design sequences with ProteinMPNN, and this was shown to be as effective as using a co-generative technique[87]. By providing our 8 Å spheroid of fragments, we only needed to diffuse residues in the gaps, essentially guiding the software to complete the structure in a manner consistent with GFP's overall fold. When we began miniaturizing GFP, it wasn't possible to incorporate non-canonical amino acids at later stages in the design process. To circumvent this, we also diffused three residues as place holders for Thr65, Tyr66, and Gly67 on the central helix. RFdiffusion only diffuses rigid N-Cα-C main-chain atoms, meaning that while we were able to generate four potential structures with this software (**Figure 12**), we had to use additional methods to ensure that they would make sense chemically.

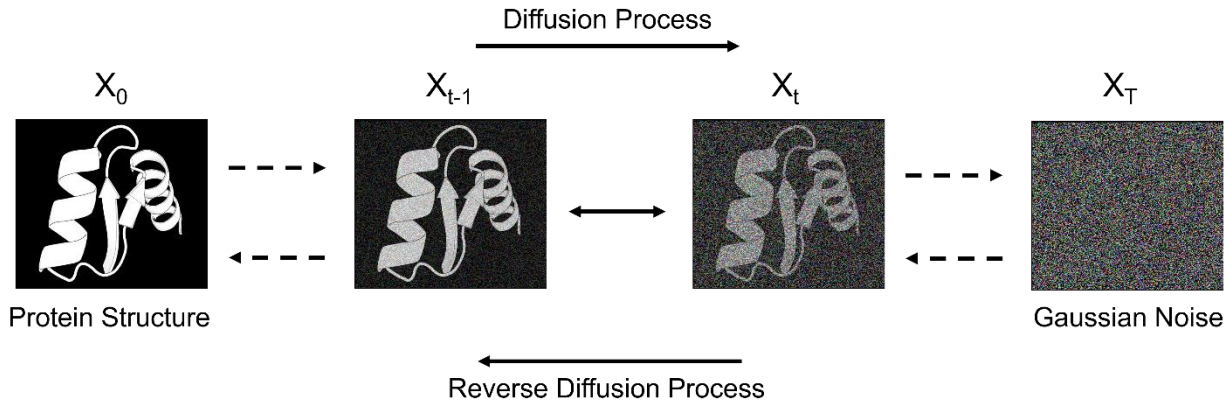



**Figure 11. Process of RFdiffusion Structure Generation.** RFdiffusion is a denoising diffusion probabilistic model that was trained on 3-dimensional structures in the Protein Database (PDB) that have been corrupted with Gaussian noise. RFdiffusion was trained to stochastically reverse this corruption until an output resembling the training data was generated. To generate protein structures RFdiffusion starts with a Gaussian distribution of backbone atoms ($X_T$) and makes a prediction towards a final protein structure ($X_0$). At each timestep ($X_t$), RFdiffusion self-conditions by using a noised version of its first prediction ($X_{t-1}$) as the input for the next prediction. Figure was reproduced with permission from Smith, B., unpublished.

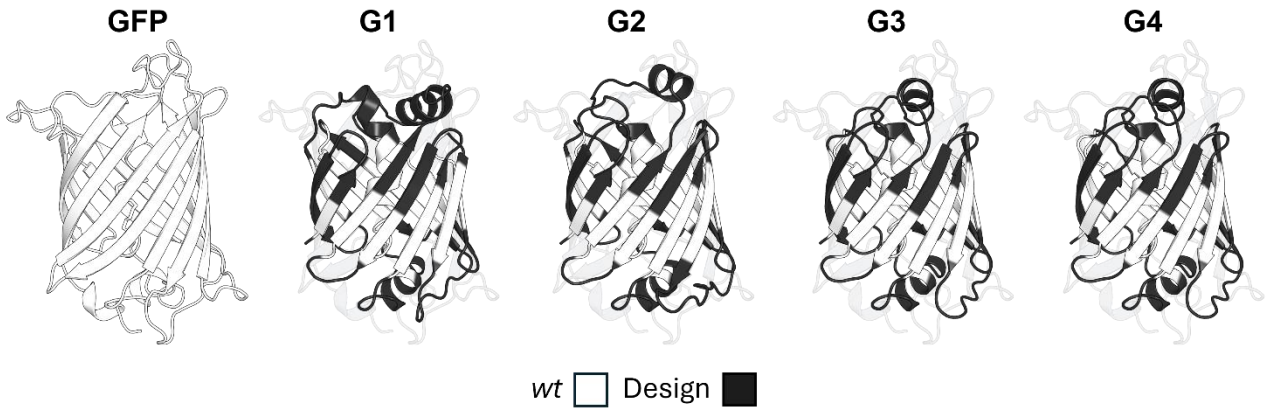



**Figure 12. RFdiffusion Generated Backbones G1-G4.** Four designed miniature GFPs created by diffusing residues (black) between the gaps of a spheroid of residues left over after pruning of the 1EMA[34] crystal structure (white). Residues were diffused using RFdiffusion, maintaining N-C directionality. Designs consisted of diffused N-Cα-C main-chain atoms, functioning only as backbones.

Once again trusting machine learning, we used ProteinMPNN[91] to design amino acid sequences which would adopt our desired folds. ProteinMPNN is a Message Passing Neural Network that was also trained on the PDB. ProteinMPNN takes the structure of a protein as input and predicts an array of amino acid sequences that are statistically likely to adopt the specified fold. Since we had opted out of designing structures and sequences simultaneously with Protpardelle, ProteinMPNN was the only viable method available to use at this time. Using this method, we designed sequences for each of our four RFdiffusion models. ProteinMPNN has the functionality to "fix" certain residues, which allowed us to conserve those originally found in our 8 Å spheroid of fragments, thereby largely preserving the architecture of the chromophore's wildtype environment. For each RFdiffusion model, we generated an array of one hundred sequences with ProteinMPNN.

To reduce the number of sequences for experimental characterization, we implemented a filtering step to validate the designs using AlphaFold2[93] (AF2). While there are other structure

prediction models available[109,110], AF2 has become the gold standard in the field. Although other prediction software outperform AF2 in some cases[111], AF2 has the unique advantage of informing its predictions with a multiple sequence alignment (MSA), thereby capturing co-evolutionary information and enhancing accuracy[106]. After inputting a sequence to AF2, it returns the predicted protein structure and a corresponding residue-wise predicted local distance difference test (pLDDT) score. This score ranges from 0 to 100 and indicates the confidence AF2 has in its prediction for each residue, with a score of 100 signifying the highest level of confidence. We used these pLDDT scores, both globally and locally, to guide our selection of the most promising designs (**Figure 13**). In addition to relying on the pLDDT scores, we also performed visual inspection of the predicted structures to identify designs with both reasonable and designable folds.

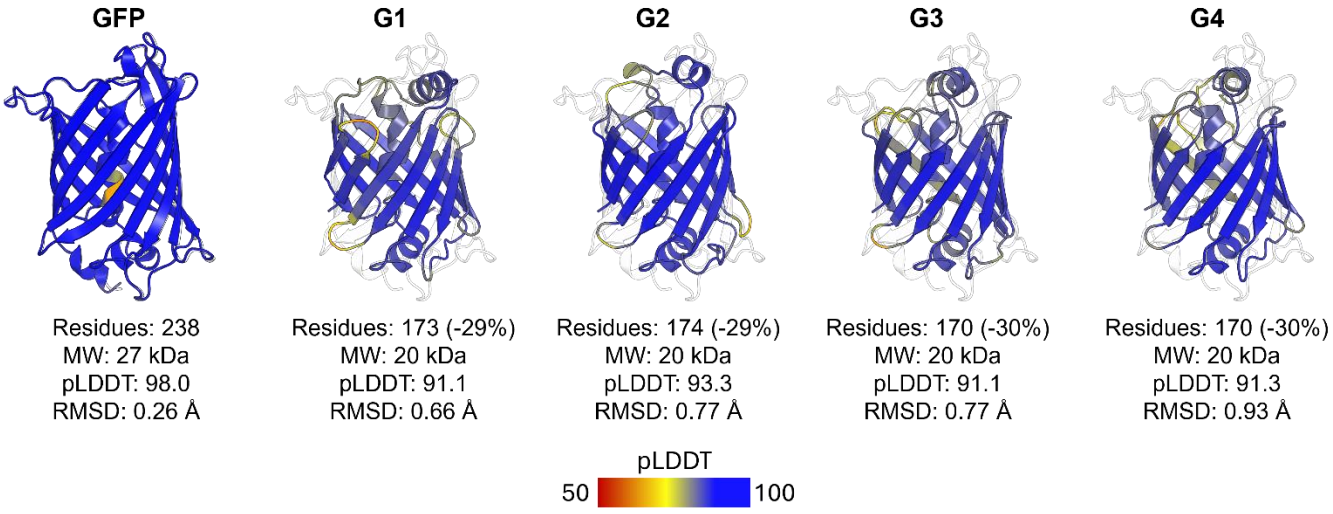



**Figure 13. AlphaFold Predictions of Designed Fluorescent Proteins.** Predicted structures for designed FPs G1-G4 (coloured according to AlphaFold2's per-reside pLDDT score, increasing from red-yellow-blue). G1-G4 backbones were designed using RFdiffusion to create a GFP-like fold around a spheroid of preserved residues from a pruned crystal structure of GFP (PDB: 1EMA[34]). Sequences were designed for G1-G4 using ProteinMPNN's soluble model and fixing residues found within 8 Å of 1EMA's chromophore. Designed FP sequences were submitted to AlphaFold2 for validation of their folds. Backbone RMSD calculations were done against PDB: 1EMA using the *align* algorithm in PyMOL (Schrödinger LLC).

There was good agreement between the AF2 predicted structures of first round designs and their respective RFdiffusion models (**Figure 14**), based on Cα root mean squared deviation (RMSD) analyses. Using the *cealign* algorithm in PyMOL (Schrödinger LLC), first round structures had the following RMSD values: G1: 0.84 Å, G2: 1.4 Å, G3: 1.6 Å, and G4: 1.1 Å. The Cα RMSD between the structure of PDB: 1EMA with diffused residues 65, 66, and 67 (in lieu of a chromophore) and its AF2 prediction was 0.28 Å. Importantly, the kinked conformation of the inner helices was maintained after the diffusion of these three residues. The authors of RFdiffusion considered a global RMSD compared to the AlphaFold structure below 2.0 Å to be a success[89]. Additionally, given that the RMSD between identical proteins from different experimental characterizations can range from 0 to 1.2 Å[112] and that an RMSD of less than 3.0 Å can be indicative of structural homologues[113], there was reason to believe that the designed folds were feasible.

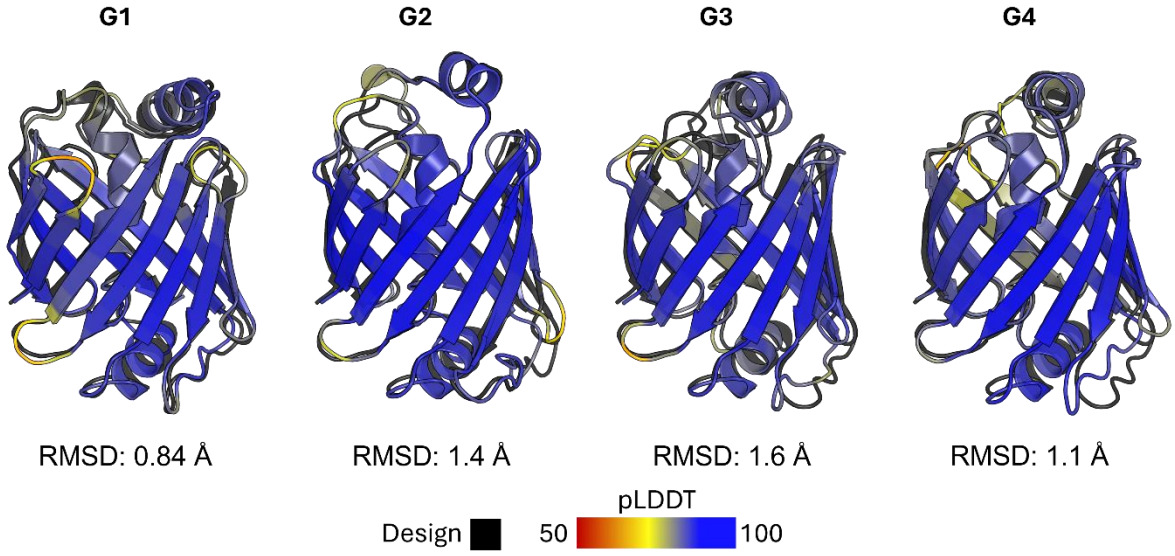



**Figure 14. Comparison of First Round RFdiffusion Models and their AlphaFold Predicted Structures.** AlphaFold2 predicted structures (red-yellow-blue) of first round FPs are aligned to their starting RFdiffusion designs (black). Structures were aligned in PyMOL (Schrödinger LLC) using the *cealign* algorithm. RMSD calculations were done using the *cealign* algorithm from PyMOL (Schrödinger LLC).

At this time, we lacked a specific metric to predict whether a design would successfully promote chromophore cyclization and maturation, so our approach was to ensure that key residues, particularly those involved in the chromophore formation, were correctly positioned relative to 1EMA (**Figure 15**). This visual comparison allowed us to confirm that the critical side-chains were placed appropriately, though it did not guarantee successful chromophore maturation. Two critical things to note is that AF2 doesn't perform any quantum mechanical calculations or predictions, and as such, the predicted structures did not contain an actual chromophore, and secondly, AlphaFold2 is not very accurate at predicting side-chain conformations[114]. Due to these limitations, we could only ensure that Thr65, Tyr66, and Gly67 residues were generally positioned in agreement with 1EMA's chromophore, and that the same was true for catalytically critical residues like Arg96 and Glu222.

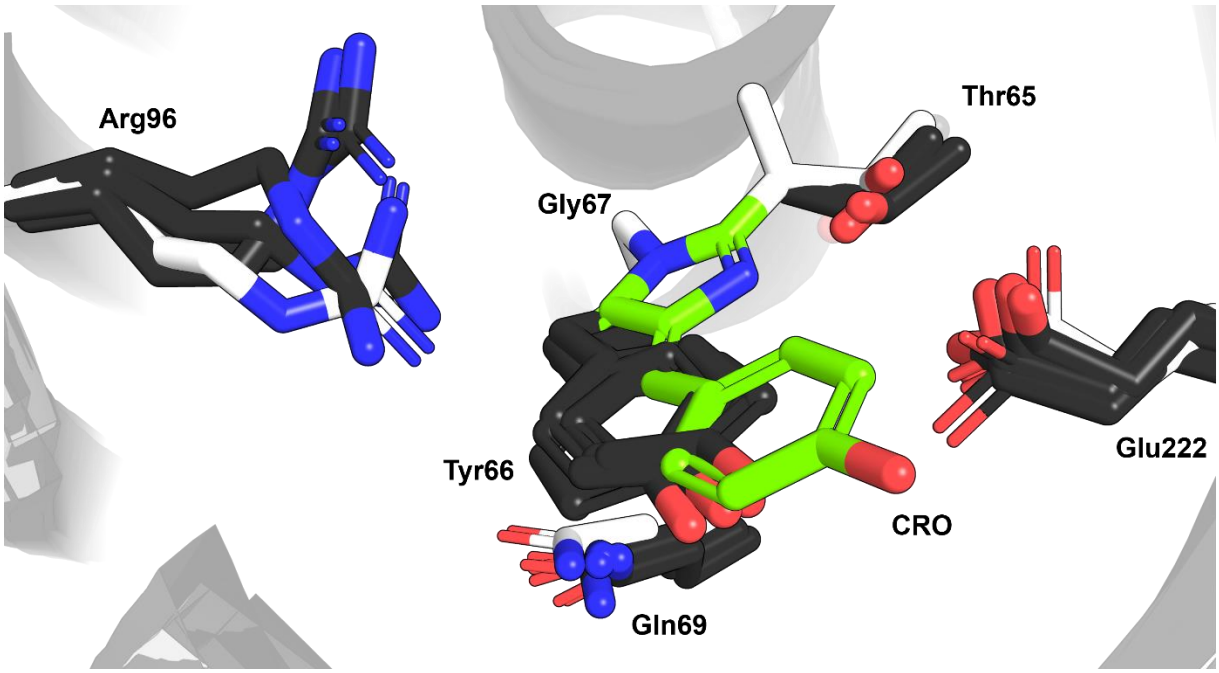



**Figure 15. Overlay of AlphaFold Predictions for Designed Fluorescent Protein Residues Critical for Chromophore Maturation.** AlphaFold2 predicted structures (black) are overlayed with those of the 1EMA[34] crystal structure (white). The mature chromophore of 1EMA (green) is

labelled with its constituent residues. AlphaFold2 cannot predict main-chain cyclization or accurate side-chain rotamer conformations, limiting comparison of the predictions to the crystal structure to general positioning only. Residues critical for chromophore formation or maturation are labelled.

In summary, to design miniature GFPs, we began by pruning residues not within 8 Å of the chromophore from the 1EMA structure, preserving only the critical residues involved in chromophore maturation and fluorescence. We then used RFdiffusion to connect the fragments, designing connecting residues in 3D space while minimizing the number of excessive residues to maintain a compact structure. Next, we employed ProteinMPNN to design amino acid sequences that are predicted to adopt the desired fold, fixing the critical chromophore-forming residues found in the 8 Å spheroid. Finally, we filtered the designs using AF2 predictions, assessing pLDDT scores and visually inspecting the structures to ensure correct positioning of key residues and overall structural feasibility. With this process we designed four GFP variants, G1-G4, collectively termed lilGFPs.

## 2.2 Results and Discussion

### 2.2.1. Fluorescence is Maintained after GFP Miniaturization

GFP has proven historically to be stubbornly difficult to miniaturize while maintaining fluorescence. We show here that it is indeed possible to achieve this aim with a miniaturization of ~25%. As part of this miniaturization, the average number of residues per β-strand for G1, G2, G3, and G4 were reduced by 2.5, 2.3, 2.3, and 2.4 residues, respectively. All four designs tested displayed fluorescence with spectral properties similar to GFP (**Figure 16**), though the fluorescence intensity was orders of magnitude lower. Accurate molar extinction coefficients could not be determined due to sample impurities (**Supplementary Figure 1**), but QY for these

designs ranged from 0.0009 to 0.0019 (**Table 1 and Supplementary Figure 2**). These values are close to the QYs calculated for unfolded, chromophore-containing peptides of GFP (0.0002)[115], and synthetic analogues of the chromophore in solution (0.0005)[57]. The comparable quantum yields suggest that the chromophores in our designs are solvent exposed, consistent with their presumed molten globule structure. Since brightness is the product of QY (Φ) and molar extinction coefficient (ε), even with extinction coefficients similar to GFP, these quantum yields would reduce their brightness by 2-3 orders of magnitude. Additionally, no absorbance peak at the wavelength corresponding to its excitation maximum was detected for each design, indicating that in each protein population only a minor fraction has formed a mature chromophore. While there were no absorbance peaks at the corresponding excitation maxima, QYs were still able to be determined using these spectra. Fluorescence emission was detected more easily than absorbance at these wavelengths due to reduced background, resulting in a lower limit of detection. Since chromophore maturation seems to be stunted, the extinction coefficients are likely much lower than that of GFP, thereby lowering the brightness of these designs even further.

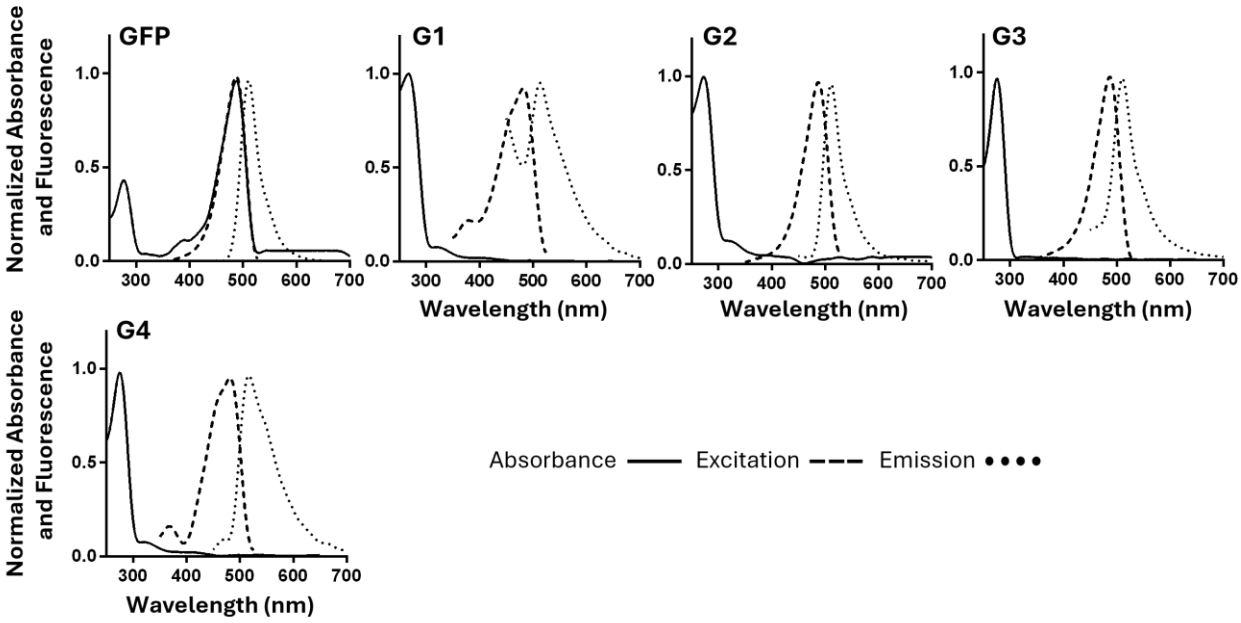



**Figure 16. Fluorescence Spectra of Green Fluorescent Proteins.** Normalized fluorescence spectra of green fluorescent proteins were determined with $\lambda_{ex}$ at 420 nm and $\lambda_{em}$ at 550 nm. Absorbance (−), excitation (--), and emission (··) spectra are shown. Measurements were taken at 25˚C in phosphate saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5) using Black UV-Star 96-Well Microplates (Greiner Bio-One) and a BioTek Synergy H1 Microplate Spectrophotometer (Agilent).

**Table 1. Spectral Properties of Green Fluorescent Proteins.** Fluorescence spectra of green fluorescent proteins were determined with $\lambda_{ex}$ at 420 nm and $\lambda_{em}$ at 550 nm. All designed proteins lacked an absorbance peak at their corresponding $\lambda_{ex}$, with only a peak attributed to tryptophan and tyrosine residues seen at 280 nm. quantum yields (QY) were calculated using the known QY of EGFP as a reference (see *Materials and Methods*).

| Protein | $\lambda_{abs}$ (nm) | $\lambda_{ex}$ (nm) | $\lambda_{em}$ (nm) | Quantum Yield |
|---|---|---|---|---|
| EGFP | 488 | 488 | 509 | 0.6 |
| G1 | N/A | 483 | 450/512 | 0.0009 ± 0.0005 |
| G2 | N/A | 485 | 508 | 0.0019 ± 0.0009 |
| G3 | N/A | 487 | 511 | 0.0010 ± 0.0003 |
| G4 | N/A | 483 | 518 | 0.0010 ± 0.0001 |

Designs G1, G2, G3, and G4 may not be as bright as GFP, but remarkably they exhibit similar spectral characteristics (**Figure 16**). EGFP has absorbance ($\lambda_{abs}$) and excitation ($\lambda_{ex}$) peaks at 488 nm with an emission ($\lambda_{em}$) peak at 509 nm. All four miniaturized designs (G1-G4) show an absorbance peak at 280 nm, corresponding to the absorption of tyrosine and tryptophan residues[116], but lack any observable peak near their $\lambda_{ex}$ , indicating that there is very little chromophore maturation occurring. While this 280 nm peak is also present with EGFP, it's dwarfed by the chromophore's absorption at 488 nm. G1 has an $\lambda_{ex}$ of 483 nm and $\lambda_{em}$ of 512 nm, G2 has an $\lambda_{ex}$ of 485 nm and $\lambda_{em}$ of 508 nm, G3 has an $\lambda_{ex}$ of 487 nm and $\lambda_{em}$ of 511 nm, and G4 has an $\lambda_{ex}$ of 483 nm and $\lambda_{em}$ of 518 nm. Interestingly, G1 has a smaller emission peak at ~450 nm, but this is likely an artifact from its measurement being so close to baseline.

### 2.2.2 Designs are Unstable

All first-round designs suffer from instability issues. This is evidenced by poor expression and solubility, which can be seen with SDS-PAGE gels following immobilized metal affinity chromatography (IMAC) (**Supplementary Figure 1**). Three of the four designs showed a significant amount of insoluble protein trapped in the pellet at their corresponding MW. Of the designs tested, three variants had expression yields under 10 mg/L, and only one design, G3, showed expression levels approaching that of GFP, with a yield of 19.5 mg/L compared to GFP's 22.5 mg/L (**Table 2**).

**Table 2. Expression of Green Fluorescent Proteins.** Expression levels of green fluorescent proteins after being expressed in BL21-Gold (DE3) *E. coli* cells (Agilent) using LB supplemented with 100 µg/mL ampicillin (EGFP) or 50 µg/mL kanamycin (lilGFPs). Cells were grown at 37 °C with shaking at 220 rpm until reaching an optical density at 600 nm of 0.6. Induction of protein expression was triggered with the addition of 1 mM isopropyl β-D-1-

thiogalactopyranoside (Fisher Scientific) and cultures were then incubated for a further 16 hours at 16 °C with shaking at 220 rpm.

| Protein | Expression yield (mg/L) | Expression concentrations (mg/mL) | Expression concentrations (μM) | Final volume (mL) |
|---|---|---|---|---|
| EGFP | 22.5 | 1.5 | 54 | 7.5 |
| G1 | 3.9 | 0.52 | 26 | 7.5 |
| G2 | 8.7 | 1.7 | 84 | 7.5 |
| G3 | 19.5 | 2.6 | 130 | 7.5 |
| G4 | 9.5 | 1.9 | 96 | 7.5 |

All designs showed a tendency to aggregate, which could be related to their insolubility, and was most obvious when performing preparatory size exclusion chromatography (SEC). All designs showed a major elution peak (~10 mL) at a volume lower than GFP and near the column's void volume (~10 mL) (**Figure 17**). Additionally, SDS-PAGE gels of SEC fractions show that most of the protein at the desired MW is eluting in these peaks (**Supplementary Figure 3**). This outcome is unexpected, as GFP, with the larger molecular weight, should have eluted earlier than the lilGFPs. The presence of smaller peaks and shoulders, closer to an appropriate elution volume (Ve), was observed with G2 and G4, but these are still closer to EGFP's Ve, indicating potential dimerization. The species producing these secondary peaks could not be isolated.

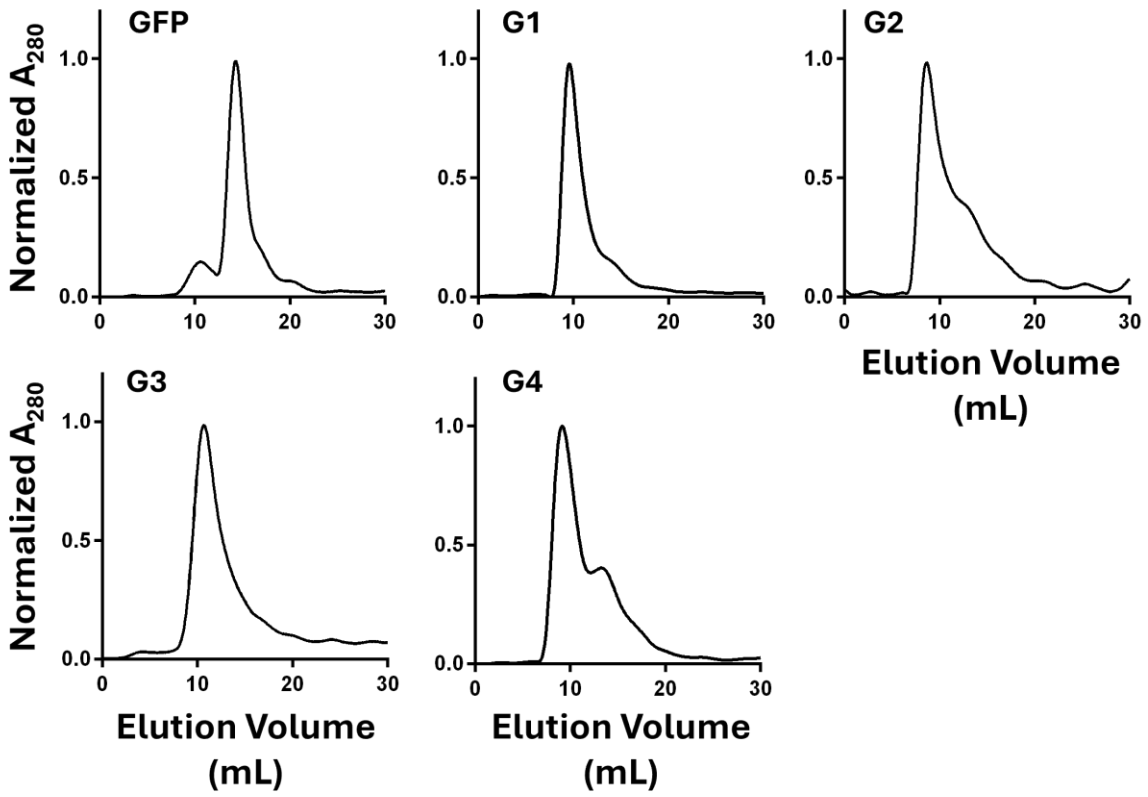



**Figure 17. Chromatograms from Preparatory Size Exclusion Chromatography.** A Bio-Rad BioLogic Duo-Flow FPLC system was used with an ENrich™ SEC 650 10 x 300 column (#7801650). A flow rate of 0.5 mL/min was used with 0.5 mL fractions being collected. Samples were purified into phosphate-saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5). Variants of lilGFP were designed to be monomeric and have a mass ~7 kDa less than GFP, but all elute ~ 10 mL, which is earlier than GFP and near the void volume of the column.

Far-UV circular dichroism (CD) Spectroscopy was used to glean insights into the secondary structure of designs (**Figure 18**). Scans were done at 25˚C and 95˚C. All designs show a mixture of α-helical and β-sheet signal, though none of them exhibited the distinct β-barrel signature typically associated with GFP. It appears that all designs have more α-helical character with more negative signals at 208 and 222 nm, which is to be expected given the designs had shortened β-strands and increased α-helical content (**Figure 13**). At 95 ˚C, only G3 showed a significant negative peak at 200 nm, which is indicative of random coils. This CD analysis utilized three

accumulations per scan, though without replicate samples, so conclusions should be drawn skeptically. This could mean that G1, G2, and G3 are incredibly heat-stable, but when combining the SEC results with those from thermal melting assays, it's more likely that this is a result of instability and/or aggregation.

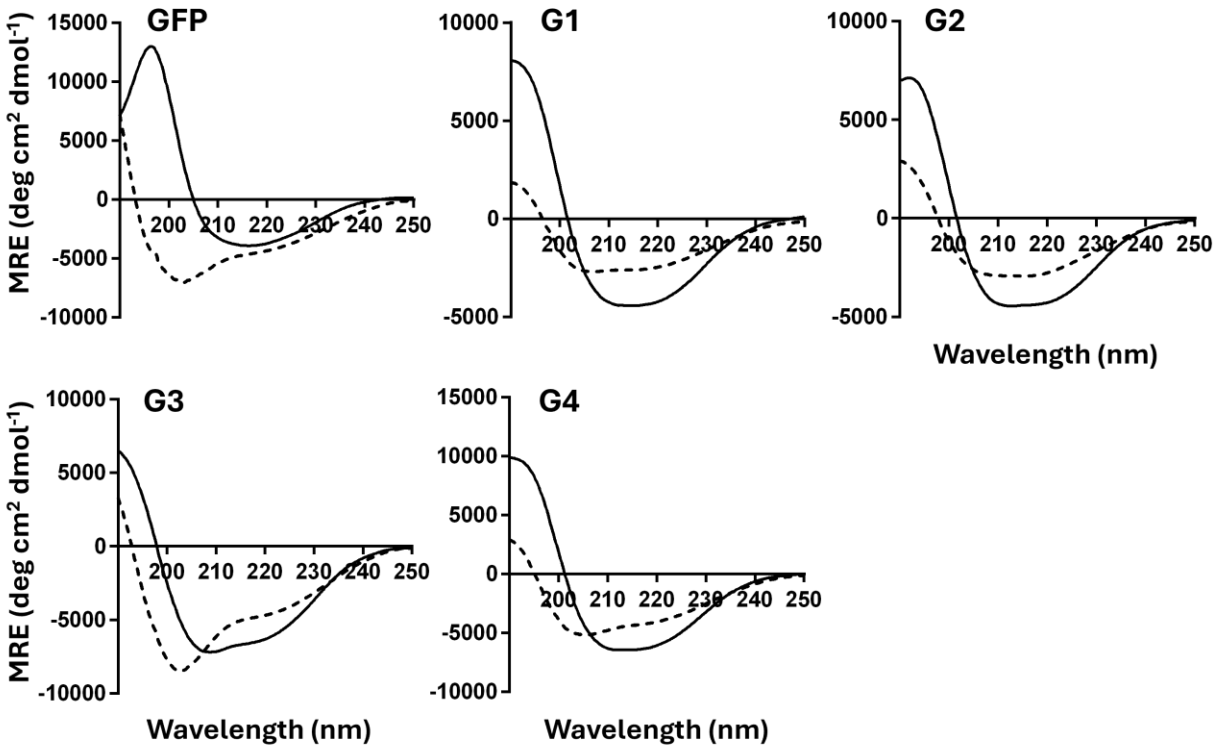



**Figure 18. Circular Dichroism Spectra of First-Round lilGFP Secondary Structure.** Circular Dichroism (CD) spectra were determined using a J-815 CD Spectrometer (Jasco), Peltier Type FDCD Cell Holder (Jasco), and a Hellma Absorption Microcuvette (Millipore Sigma) with a 1 mm pathlength. Purified 10 µM samples were prepared in 20 mM sodium phosphate, pH 7.4. Secondary scans were done with 3 accumulations, a digital integration time (D.I.T.) of 1 second, continuous scanning mode with a scanning speed of 10 nm/min, using a data pitch of 0.2 nm, and band width of 1 nm. Scans were done at 25˚C (−) and 95˚C (--). Using The purity of each sample was determined using ImageJ[117] analysis of SDS-PAGE gels after size exclusion chromatography, yielding the following results: GFP – 98%, G1 – 57%, G2 – 26%, G3 – 89%, and G4 – 72%.

These α-helical/β-sheet spectra could be related to the strange results observed in thermal melt assays (**Figure 19**), potentially indicating the presence of a molten globule conformation. The normalized ellipticity data showed very little cooperative unfolding, which is to be expected for most well-folded proteins where there is only one energy barrier to overcome for the folded-unfolded transition to occur[118]. In the case of G4, there appeared to be some cooperative unfolding, but the unfolded asymptote was never fully reached. Attempts to transform these data to Fraction Unfolded ($F_U$) plots using the Greenfield equations[119] failed for all designs. Even with the equations providing generous fits for the mean residue ellipticity (MRE, mdeg $cm^2$ $dmol^{-1}$) values, there is no, or very gradual, sigmoidal character seen, which lead to unreliable melting temperatures being calculated. The linear character seen could be due to a process of gradual unfolding where several small energy barriers must be overcome during the folded-unfolded transition, and which can be occurring in parallel within the sample population[118]. The $|\Delta_{MRE}|$ during GFP unfolding was 29, whereas values for G1, G2, G3, and G4 were 9.1, 7.5, 14.4, and 12.7, respectively. As these MRE changes range from approximately one-third to one-half of the GFP value, the relative contribution of contaminant proteins to the unfolding data remains unclear. However, the low $|\Delta_{MRE}|$ magnitudes may be indicative of predominantly molten globule-like designs. When coupling this linearity with the fact that the low QY seen with the designs could be caused by a disordered chromophore environment, these melts could indicate that a molten globular form is dominating the folded-unfolded equilibrium of the designs[118].

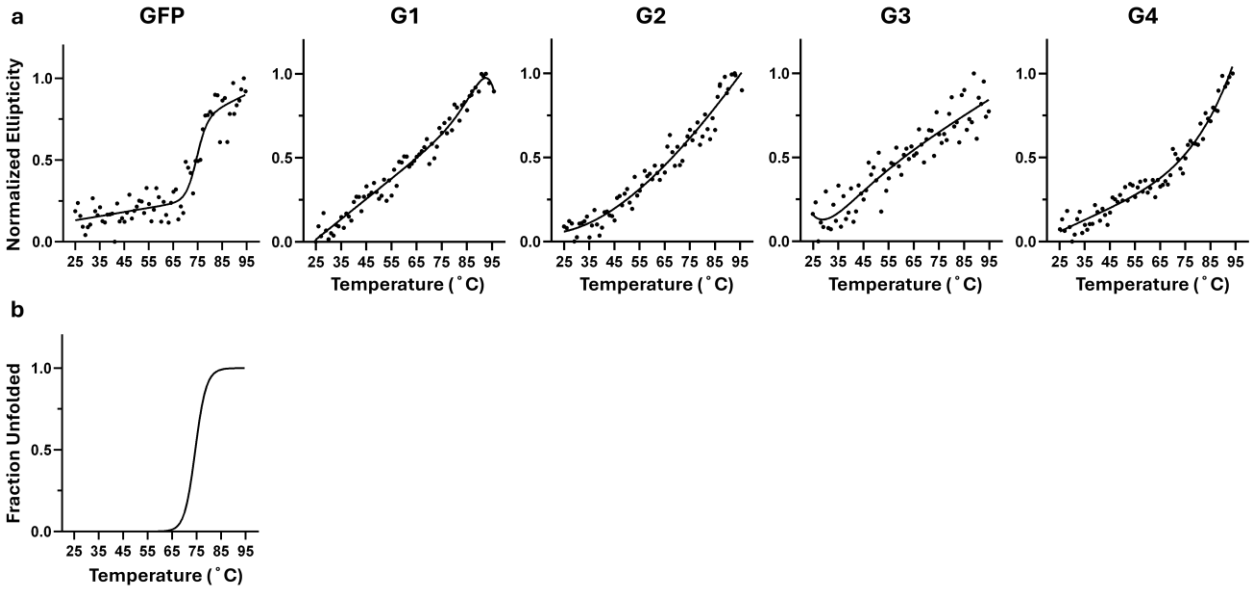



**Figure 19. First-Round Thermal Melt Assays Using Circular Dichroism.** Thermal Melt assays were done using Circular Dichroism with a J-815 CD Spectrometer (Jasco), Peltier Type FDCD Cell Holder (Jasco), and a Hellma Absorption Microcuvette (Millipore Sigma) with a 1 mm pathlength. Purified 10 μM samples were prepared in 20 mM sodium phosphate, pH 7.4. Melting curves were done with a D.I.T. of 8 seconds, standard sensitivity, and a band width of 1 nm. Temperature for the melting curves was increased at a rate of 1˚C/min and sampled every 1˚C after a 10 second wait upon reaching the interval temperature. (a) Normalized ellipticity values (dots) from the thermal melts were used for fitting with the Greenfield method[119] (thin line), resulting in (b) fraction unfolded plots. A $T_M$ of 74.4 ˚C was determined for GFP, while no $T_M$ values were able to be determined for designs. The purity of each sample was determined using ImageJ[117] analysis of SDS-PAGE gels after size exclusion chromatography, yielding the following results: GFP – 98%, G1 – 57%, G2 – 26%, G3 – 89%, and G4 – 72%.

Within an increasingly disordered microenvironment, a chromophore can undergo a greater range of torsional motions in the excited state with reduced energetic barriers[120]. These extra degrees of freedom would promote nonradiative decay through mechanisms such as TICT and *cis-trans* isomerization of the chromophore's phenolate moiety[121], releasing energy as heat and kinetic energy instead of as photons[122], thereby reducing QY. The results in this chapter suggest that the proteins may be forming molten globules, which could contribute to the types of nonradiative

decay described above. This indicates the presence of structural defects that are impairing function and that need to be addressed. Chapter 3 will outline the attempts made to remedy these issues.

## Chapter 3: Improving Brightness in Designed lilGFPs

### 3.1 Iterative Computational Design of a Miniature GFP

Given the low QY and stability of first-round designs, a second round was undertaken to optimize the process based on lessons learned from the initial designs and recent developments in the field of MLAPD. After designing and characterizing G1-G4, several new technical developments and insights emerged that influenced this second design strategy. Firstly, it was found that comparing the inner helix RMSD of the designs to the 1EMA crystal structure could serve as a predictor for chromophore maturation[123]. Secondly, a new version of ProteinMPNN, called LigandMPNN, was published, which allows for the incorporation of ligands and noncanonical amino acids[92]. It was also observed that pLDDT values in the mid to high 90s were increasingly necessary for accurate structural predictions[100,124], a threshold that had not been reached with earlier designs. Finally, it became apparent that refining designs through iterative and cyclical rounds of sequence optimization using ProteinMPNN/LigandMPNN and AF2 was beneficial for achieving successful outcomes[100]. Specifically, using an AF2 model from a previous iteration as input for ProteinMPNN instead of the initial RFdiffusion model was show to produce more stable and better-folded designs[100].

With the newfound insight that GFP-like proteins with an average inner helix RMSD of 0.8 Å, when compared to 1EMA, could form a chromophore[123], we finally had a metric for validating designs based on function, rather than relying exclusively on structural accuracy. We aimed to synergize this information with the recent ability to incorporate a mature chromophore

into the sequence design workflow, which promised to enhance the precision of our predictions and further refine our design process.

Our second round of design used G1, G2, and G3 sequences and AF2 models as the starting templates for three parallel streams of the same workflow. A G4 stream was not undertaken since G3 and G4 share a backbone and only differ in sequence. For each template, the chromophore from 1EMA was inserted into the first round AF2 model and the existing residues 65-67 were removed (**Figure 20**). The chromophore containing template was then standardized using Triad (Protabit), which is a flexible-backbone minimization that optimizes bond lengths, dihedral angles, and rotamer configurations. Cleaned chromophore-containing templates were then used as inputs for LigandMPNN, which produced arrays of one hundred sequences for G1 and G2, and one thousand sequences for G3. In this round of design, only the residues within 4 Å of the chromophore (as seen in 1EMA) were fixed. Each sequence was then placed on its respective cleaned chromophore-containing backbone in Triad, and the structure was cleaned, which is a fixed-backbone repacking and minimization using Triad's *cleanSequences.py* app to remove any clashes introduced by the new residues. This workflow allowed for the ranking of structures using Triad's Phoenix energy score[125], as well as measurement of the inner helix RMSD between the AF2 models, 1EMA, and the Triad-generated models. Since AF2 still couldn't incorporate a mature chromophore, we swapped back Thr65, Tyr66, and Gly67 into the sequences for structure prediction, ensuring that the key residues for chromophore formation were represented in the model.

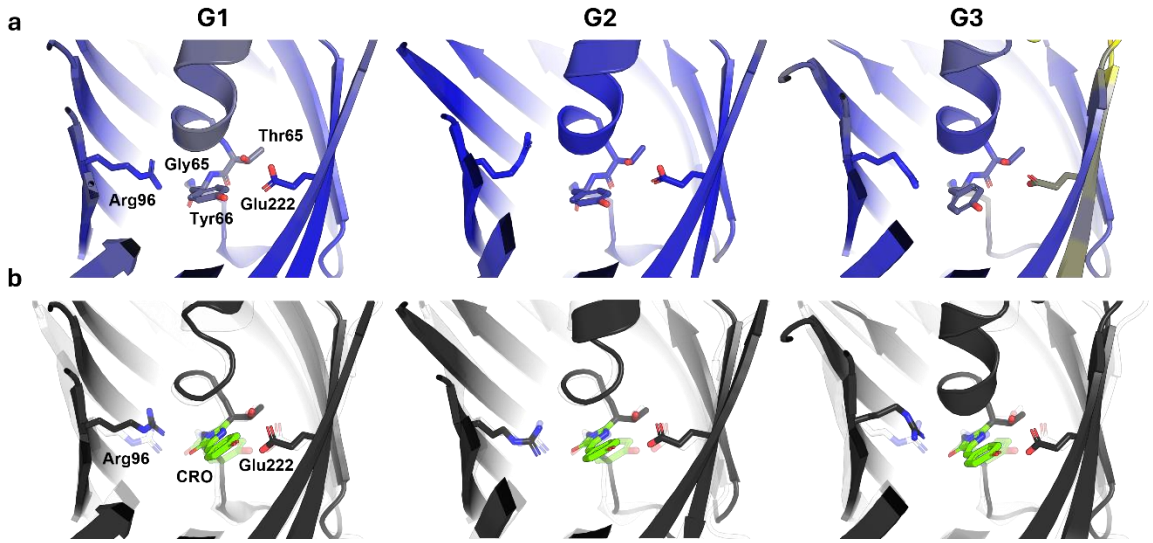



**Figure 20. Structures of G1 – G3 AlphaFold2 Predictions After Chromophore Insertion and Flexible-Backbone Minimization in Triad.** (a) G1, G2, and G3 AlphaFold2 predictions (red-yellow-blue) were used as the staring templates for a second round of lilGFP design. Analogous residues to Thr65, Tyr66, and Gly67 were removed and (b) the chromophore (green) from a GFP crystal structure (white)(PDB: 1EMA[34]) was inserted into the model. Rigid backbone repacking was done on the chromophore-containing designs in Triad (Protabit LLC) using the *cleanSequences.py* application.

With this additional workflow (**Figure 21**) we could now rank designs based on the likelihood of their folds being adoptable using pLDDT values, their relative stabilities using the Triad Phoenix energy score, and their potential to form a chromophore using Inner Helix RMSD. This provided a comprehensive evaluation of our candidates in terms of both their structural and functional potential.

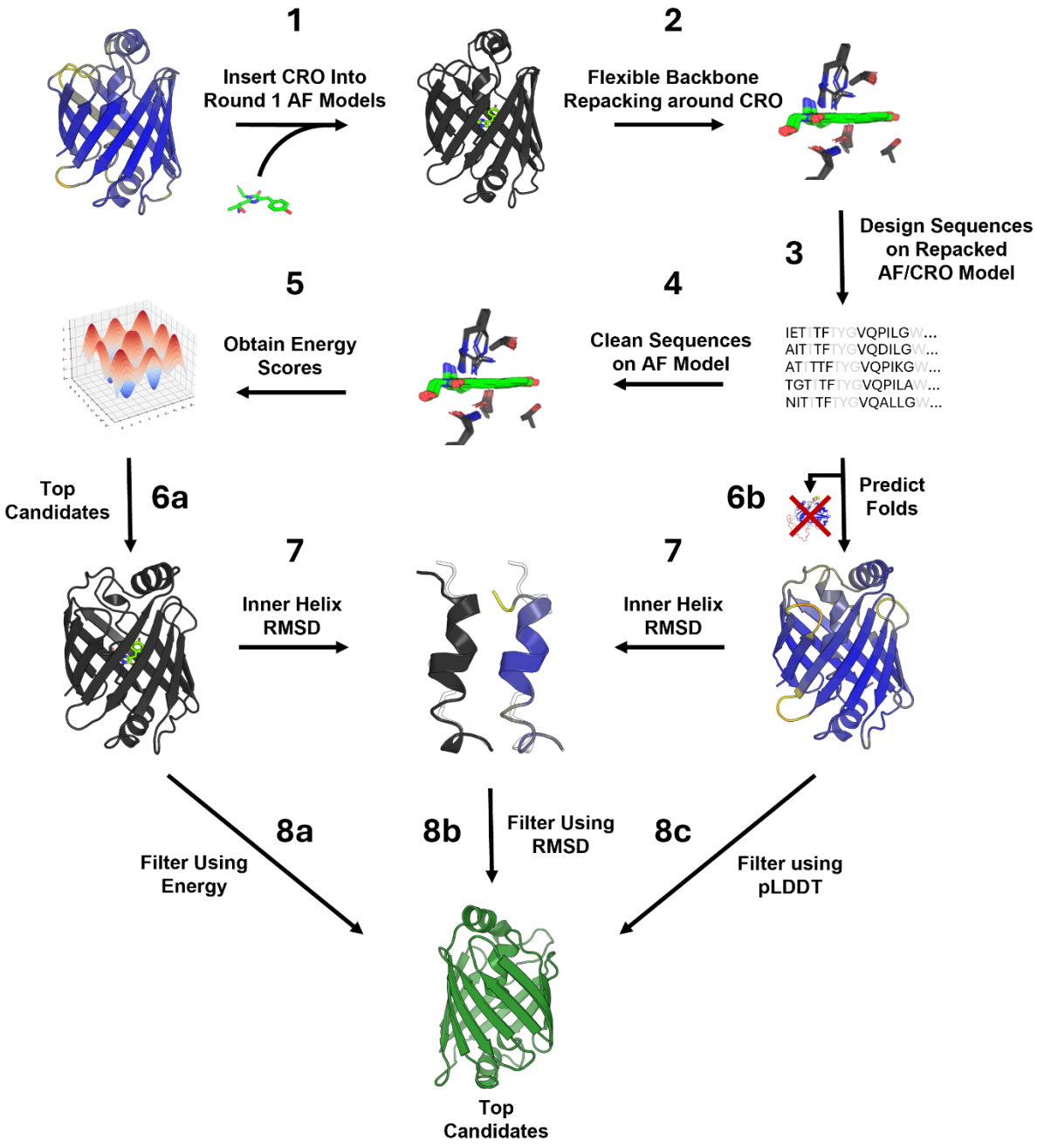



**Figure 21. Enhanced Design Process Used for Second Round lilGFPs.** AlphaFold2 (AF2) models of first round lilGFPs G1, G2, and G3 were used as the starting template for round two designs. (1) The chromophore from a GFP crystal structure (PDB: 1EMA[34]) was inserted into each AF2 model, replacing Thr65,Tyr66, and Gly67. (2) Flexible backbone repacking around the chromophore was done using the Triad software (Protabit LLC). (3) Sequences for the repacked, chromophore-containing AF2 models were designed using LigandMPNN[92]. (4) Sequences from LigandMPNN were cleaned on their respective repacked, chromophore-containing AF2 backbones. (5) Triad energy scores were obtained from previous cleaning step. (6a) Structures with negative energy values were selected. (6b) LigandMPNN sequences from (3) were

submitted to AlphaFold2 for structure prediction. (7) The inner helices of both cleaned and AF2 structures underwent RMSD comparisons with that of PDB: 1EMA. Final candidates for experimental characterization were chosen after (8a) cleaned structures were filtered based on the lowest Triad energy scores, (8b) cleaned and AF2 structures were filtered based on the lowest inner helix RMSD compared to PDB: 1EMA, and (8c) AF2 structures were filtered based on the highest pLDDT values.

The following filtering criteria were applied to the second round of designs: 1) Inner Helix $RMSD_{AF\text{-}1EMA} \leq 0.8$ Å, 2) Inner Helix $RMSD_{AF\text{-}Triad} \leq 0.5$ Å, 3) Inner Helix $RMSD_{Triad\text{-}1EMA} \leq 0.5$ Å, 4) Triad energy ≤ -500, and 5) the highest possible pLDDT (**Table 3**). After filtering, we settled on twelve designs to experimentally characterize (**Figure 22**). Although these criteria could not be applied in the initial round of design, we performed the analysis on G1, G2, and G3, as they served as templates for the second round. Notably, G1 and G2 would have passed a similar filtering step, with the only disqualifying metric for G3 being the Triad energy score. Since all first-round designs exhibit an Inner Helix $RMSD_{AF\text{-}1EMA}$ value lower than the reported 0.8 Å average for chromophore-forming proteins[123], it is unsurprising that they exhibited fluorescence. Although most first-round designs passed the second-round filtering criteria, this does not necessarily invalidate the second-round selection process. All second-round designs exhibited higher pLDDT scores and lower Triad energy values. We hypothesized that once a certain threshold of inner helix $RMSD_{AF\text{-}1EMA}$ was reached, further reduction in RMSD no longer correlated with increased fluorescence. Taken together, these results suggest that the second-round designs were more promising and likely to outperform those from the first round. In total, twenty-two second-round designs passed filtering, with the final twelve being chosen based on the criteria above.

**Table 3. Filtering Parameters for Designed Green Fluorescent Proteins.** First round, second round and G1-Enhanced designs were subject to different filtering criteria, and the results are shown here. First round and G1-Enhanced designs were filtered using length, molecular weight, average pLDDT scores and visual inspection. G1, G2, and G3 were later evaluated according to the same criteria as the second-round designs but had already been experimentally characterized. Second-round designs were filtered the same as pervious rounds, but with the addition of four criteria: 1) inner helix RMSD between AlphaFold2 (AF2) predictions and that of a crystal structure (PDB: 1EMA[34]), 2) inner helix RMSD between AF2 predictions and the inner helix of corresponding chromophore-containing models after rigid backbone repacking in Triad, 3) inner helix RMSD between the aforementioned Triad models and that of 1EMA, and 4) Triad energy scores after fixed backbone repacking after chromophore insertion.

| Protein | Amino Acids | Molecular Weight (Da) | pLDDT | Inner Helix $RMSD_{AF^1\text{-}1EMA^2}$ (Å) | Inner Helix $RMSD_{AF\text{-}Triad^3}$ (Å) | Inner Helix $RMSD_{Triad\text{-}1EMA}$ (Å) | Triad Energy |
|---|---|---|---|---|---|---|---|
| **Round 1** | | | | | | | |
| G1 | 170 | 18,798 | 91.1 | 0.44 | 0.31 | 0.48 | -533 |
| G2 | 171 | 19,018 | 93.3 | 0.42 | 0.28 | 0.46 | -509 |
| G3 | 167 | 18,953 | 91.1 | 0.36 | 0.34 | 0.40 | -428 |
| G4 | 167 | 18,673 | 91.3 | N/A | N/A | N/A | N/A |
| **G1-Enhanced** | | | | | | | |
| G1-Emerald | 170 | 19,148 | 90.7 | N/A | N/A | N/A | N/A |
| G1-GreenLantern | 170 | 18,941 | 90.4 | N/A | N/A | N/A | N/A |
| G1-Superfolder | 170 | 18,952 | 92.7 | N/A | N/A | N/A | N/A |
| **Round 2** | | | | | | | |
| G1A50 | 170 | 18,644 | 93.6 | 0.64 | 0.45 | 0.46 | -543 |
| G2A78 | 171 | 18,695 | 94.6 | 0.46 | 0.38 | 0.46 | -536 |
| G3CRO1k-92 | 167 | 18,435 | 94.1 | 0.52 | 0.38 | 0.40 | -534 |
| G3CRO1k-219 | 167 | 18,089 | 94.4 | 0.42 | 0.45 | 0.39 | -567 |
| G3CRO1k-281 | 167 | 18,199 | 93.9 | 0.68 | 0.41 | 0.44 | -542 |
| G3CRO1k-368 | 167 | 18,263 | 94.2 | 0.48 | 0.47 | 0.38 | -530 |
| G3CRO1k-385 | 167 | 18,411 | 94.3 | 0.47 | 0.30 | 0.40 | -520 |
| G3CRO1k-570 | 167 | 18,074 | 94.2 | 0.73 | 0.49 | 0.43 | -580 |
| G3CRO1k-698 | 167 | 18,269 | 93.7 | 0.51 | 0.27 | 0.44 | -537 |
| G3CRO1k-799 | 167 | 18,424 | 94.0 | 0.68 | 0.42 | 0.47 | -553 |
| G3CRO1k-941 | 167 | 18,698 | 94.4 | 0.46 | 0.34 | 0.39 | -508 |
| G3CRO1k-964 | 167 | 18,313 | 94.0 | 0.60 | 0.38 | 0.43 | -516 |

Round 1 designs had 6x H-tags and 3 amino acid linkers. Round 2 designs had 6x H-tags and no linkers. All second-round designs but G1A50 had a tryptophan residue added for purification. Amino acid lengths and molecular weights reported do not include H-tag or linker residues, only starting methionine. Backbone atoms in residues 60-74 of 1EMA and analogues residues in designs were used for RMSD calculations.
[1]AlphaFold2 model, [2] PDB: 1EMA,[3] Chromophore-containing model repacked in Triad

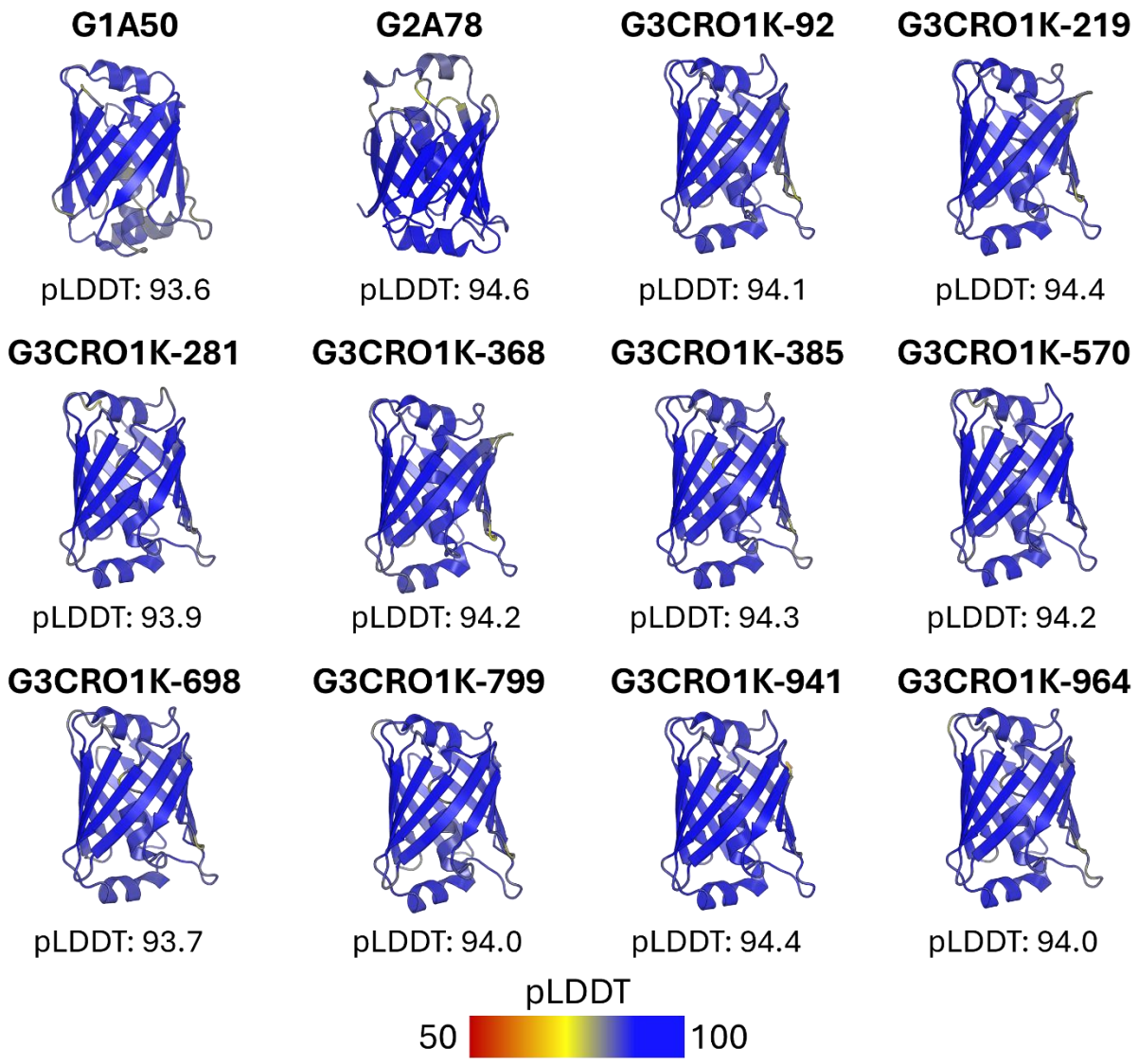



**Figure 22. AlphaFold2 Predicted Structures of Second-Round lilGFP Designs.** The second round of computational design of lilGFPs yielded twelve candidates, seen coloured according to their AlphaFold2 residue-wise pLDDT scores (red-yellow-blue). All second-round designs had higher average pLDDT scores than the first round designs. The final twelve candidates were filtered from a total of 1400 designs.

### 3.2 Simulating Alternative Starting Templates for GFP Miniaturization

Using EGFP as a starting template for this design process was a logical choice due to its well characterized structure and simplified excitation spectrum. However, we couldn't help but wonder if another enhanced form of GFP would have produced better results. We were particularly

interested in what could have been had we used mEmerald, mGreenLantern, or Superfolder GFP as a starting template for miniaturization. Since these variants must share structural homology by necessity, we hypothesized that we could bypass the initial pruning step of the design process and simply insert substitutions into G1 (**Supplementary Table 1**). This resulted in the G1-Enhanced variants G1-Emerald with 28 substitutions, G1-GreenLantern with 39 substitutions, and G1-Superfolder with 27 substitutions (**Figure 23**). In the cases of G1-Emerald and G1-GreenLantern, this resulted in lower pLDDT scores than G1's 91.1, but G1-Superfolder surpassed G1 with a score of 92.7.

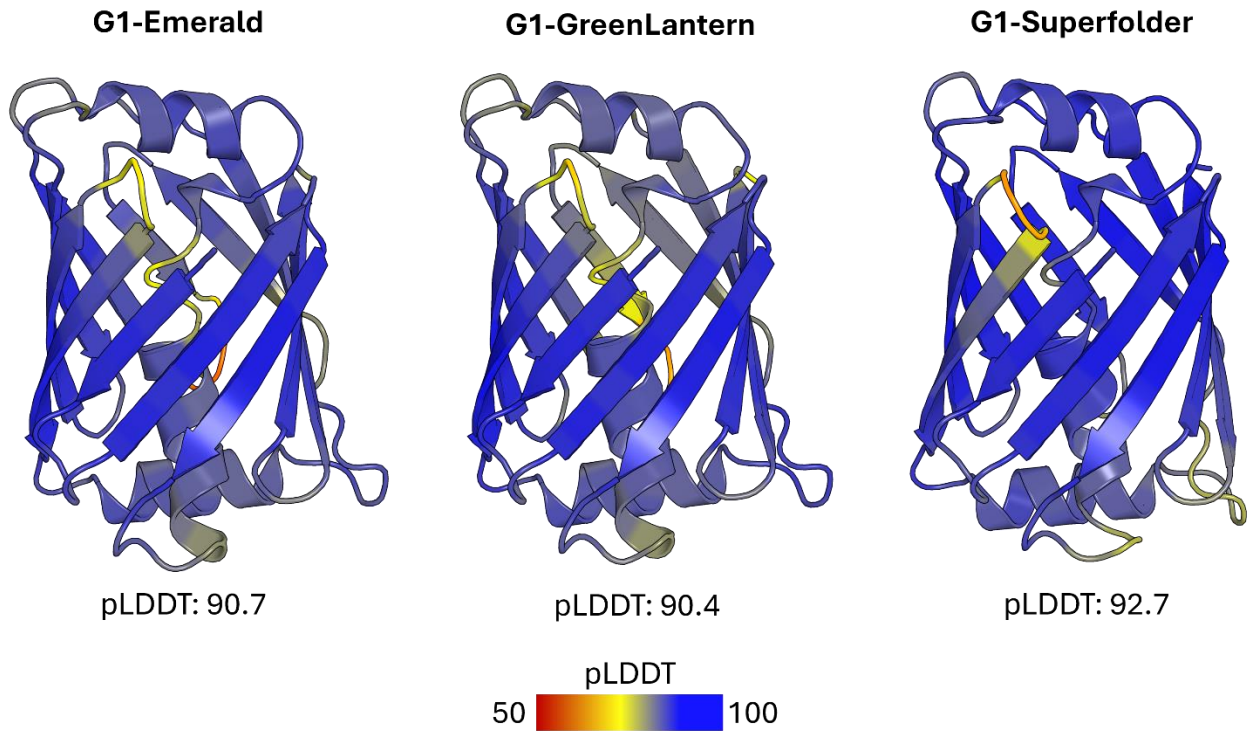



**Figure 23. AlphaFold2 Predicted Structures of G1-Enhanced Designs.** Any mutations found in mEmerald, mGreenLantern, or Superfolder GFP that would have been preserved in the 8 Å spheroid of fragments in the first design round were substituted into G1. This created the G1-Enhanced variants seen here, coloured according to their AlphaFold2 residue-wise pLDDT scores (red-yellow-blue).

### 3.3 Improving Green Fluorescence with Mutagenesis

To improve the fluorescence of our computationally designed lilGFPs, we turned to *in vitro* mutagenesis. Mutagenesis, specifically in the context of Directed Evolution, has become the method of choice for improving protein function[126]. Directed Evolution has proven to be particularly effective when starting with a protein that already demonstrates some degree of activity and pairing it with a high-throughput selection method[126]. Since our designs exhibited marginal fluorescence, Directed Evolution appeared to be the most viable path forward.

We chose fluorescence-activated cell sorting (FACS) as our selection method due to its efficiency and ease of use, as it is already set up to sort cells based on green fluorescence. For the creation of the mutagenic libraries, we used a variation of error-prone PCR (epPCR)[127] (see *Chapter 4.4 Random Mutagenesis*), with the primary distinction being the absence of $MnCl_2$, which is typically included in traditional epPCR protocols to increase the mutation rate. This variation was used to introduce a manageable level of mutations that didn't saturate the libraries.

When evolving a protein, higher stability is crucial because it can promote evolvability[128,129]. This is because mutations that are beneficial for activity, and most mutations in general, tend to destabilize proteins[126]. Therefore, for a protein to "survive" this trade-off and continue evolving, it needs to be stable enough to withstand the mutations required to improve its activity. For this reason, we chose G3 as our candidate for Directed Evolution, since it had the best expression levels of all designs, nearly at that of GFP (see *Chapter 2.2 Results and Discussion)*. As mentioned above, we used epPCR to generate mutant libraries of G3. We performed multiple iterative rounds of mutagenesis with two distinct mutagenic trajectories running in parallel. This resulted in a mutagenic library of 300,000 G3 variants, which we then sorted with FACS.

For sorting, cells were stained with Propidium Iodide (PI, $\lambda_{em}$ = 617 nm), which serves as a viability dye by only binding the DNA of dead cells[130]. During sorting runs, a PI gate was set up

on the Y-axis with a bandpass of 617/30 nm and a lowpass of 600 nm, this functioned as the "Dead Cell" gate by excluding cells with a fluorescence signal between 600 nm and 647 nm. On the X-axis was the gate for green fluorescence with a bandpass of 525/50 nm, which was to collect improved G3 variants displaying fluorescence between 475 nm and 575 nm. However, despite considerable effort, zero variants showed detectable improvement in fluorescence, meaning none were successfully selected for a second round of mutagenesis. That said, analysis of the G3 mutant library by flow cytometry suggested the presence of potentially beneficial mutations; however, the resulting fluorescence increases were insufficient to meet the selection threshold during FACS.

## 3.4 Results and Discussion

### 3.4.1 Iterative Rounds of Sequence Design Have no Impact on Fluorescence

Overall, the second round of designs displayed similar results to those of the first round. QY ranged from 0.00004 to 0.002 **(Figure 24 and Table 4)**, though it must be noted that no biological replicate measurements were made for second-round designs. Similar to the first round, all second-round designs exhibited impurity issues, as evidenced by SDS-PAGE analysis following IMAC purification. Most of the protein at the desired molecular weight was found to be localized in the insoluble pellet fractions, with minimal protein detected in the final elution fractions (**Supplementary Figure 4**). This prevented accurate extinction coefficients from being determined, since contaminants, like the presumed *ArnA* enzyme[131] (MW: 74 kDa)[132], were disproportionally purified. As with the first-round designs, the only notable absorbance peak can be attributed to tyrosine and tryptophan at 280 nm, which again indicates that only a minor percentage of each sample is undergoing complete chromophore maturation. An interesting trend observed among all the second-round designs is a shift in their excitation peaks, which now range between 442 and 460 nm. Additionally, all emission peaks were red-shifted slightly between 510

and 516 nm. These changes may be reflecting subtle modifications in the chromophore environment caused by relaxing the cutoff for residue fixing from 8 Å to 4 Å during sequence design.

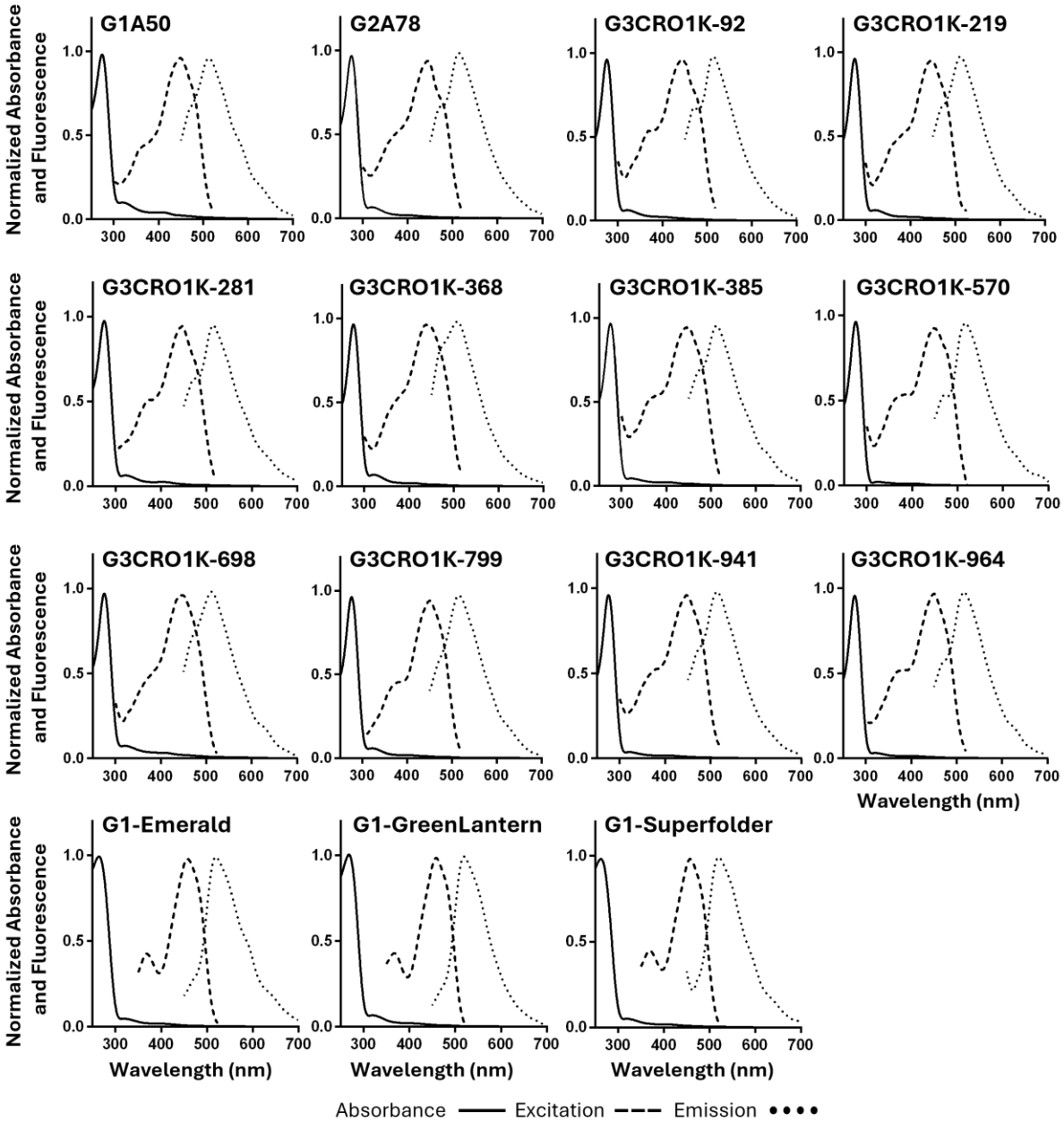



**Figure 24. Fluorescence Spectra of Second-Round Green Fluorescent Proteins.** Normalized fluorescence spectra of green fluorescent proteins were determined with $\lambda_{ex}$ at 420 nm and $\lambda_{em}$ at 550 nm. Absorbance (−), excitation (--), and emission (··) spectra are shown. Measurements were

taken at 25˚C in phosphate saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5) using Black UV-Star 96-Well Microplates (Greiner Bio-One) and a BioTek Synergy H1 Microplate Spectrophotometer (Agilent).

**Table 4. Spectral Properties of Green Fluorescent Proteins.** Fluorescence spectra of green fluorescent proteins were determined with $\lambda_{ex}$ at 420 nm and $\lambda_{em}$ at 550 nm. All designed proteins lacked an absorbance peak at their corresponding $\lambda_{ex}$, with only a peak attributed to tryptophan and tyrosine residues seen at 280 nm. quantum yields (QY) were calculated using the known QY of EGFP as a reference (see *Materials and Methods*).

| Protein | $\lambda_{abs}$ (nm) | $\lambda_{ex}$ (nm) | $\lambda_{em}$ (nm) | Quantum Yield |
|---|---|---|---|---|
| EGFP | 488 | 488 | 509 | 0.6 |
| G1A50 | N/A | 450 | 514 | 0.0004 |
| G2A78 | N/A | 442 | 512 | 0.0006 |
| G3CRO1k-92 | N/A | 447 | 512 | 0.0004 |
| G3CRO1k-219 | N/A | 445 | 511 | 0.001 |
| G3CRO1k-281 | N/A | 451 | 512 | N/A |
| G3CRO1k-368 | N/A | 443 | 511 | 0.002 |
| G3CRO1k-385 | N/A | 451 | 510 | 0.002 |
| G3CRO1k-570 | N/A | 460 | 515 | 0.0005 |
| G3CRO1k-698 | N/A | 443 | 515 | 0.0005 |
| G3CRO1k-799 | N/A | 458 | 512 | 0.001 |
| G3CRO1k-941 | N/A | 451 | 516 | 0.00004 |
| G3CRO1k-964 | N/A | 454 | 514 | 0.001 |
| G1-Emerald | N/A | 458 | 521 | 0.001 |
| G1-GreenLantern | N/A | 460 | 522 | 0.002 |
| G1-Superfolder | N/A | 457 | 520 | 0.0006 |

### 3.4.2 Enhanced GFP Insertions have Negligible Effects

For first round designs, EGFP was used as a starting template. To explore whether beginning with another enhanced variant of GFP could have yielded better results, we used G1 as a secondary template by making substitutions based on mEmerald, mGreenLantern, and Superfolder GFP. For any residues found in these variants that would have also been preserved in the initial pruning of EGFP (**Figure 10**) the appropriate substitutions were carried out on G1, thereby creating G1-Emerald, G1-GreenLantern, and G1-Superfolder variants.

No biological replicates measurements of these variants were taken, but there was no drastic improvement in brightness (**Table 4**). The quantum yields of these variants ranged from 0.0006 to 0.002, which is still far too low to be useful. Absorbance peaks other than those at 280 nm were absent (**Figure 24**). Excitation peaks were blue-shifted similar to second-round design with a range between 457 and 460 nm. These G1-Enhanced variants had similar impurity issues as previous rounds of design (**Supplementary Figure 5).** Furthermore, emission peaks were drastically red-shifted, ranging from 520-522 nm. Arguably, these G1-Enhanced variants were worse GFPs than G1.

### 3.4.3 Designs Remain Unstable

As mentioned earlier, the second-round designs exhibited similar issues to the first round, potentially to a greater extent. Of the twelve second-round designs expressed, nine displayed a significant amount of insoluble protein, which was retained in the pellet at their expected MW following IMAC purification (**Supplementary Figure 4**). Furthermore, all designs, except for G3CRO1K-570, showed minimal soluble protein at their corresponding molecular weights. The G1-Enhanced variants showed no improvement. Given the low quantum yield, insolubility, and poor expression levels of the second-round designs, only two variants, G3CRO1K-92 and

G3CRO1K-570, were selected for further characterization. Moreover, the incongruous chromatograms from preparative SEC (**Figure 25 and Supplementary Figure 3**), which mirrored the issues observed in the first round (**Figure 17**), suggest that these designs are also likely undergoing misfolding and/or aggregation.

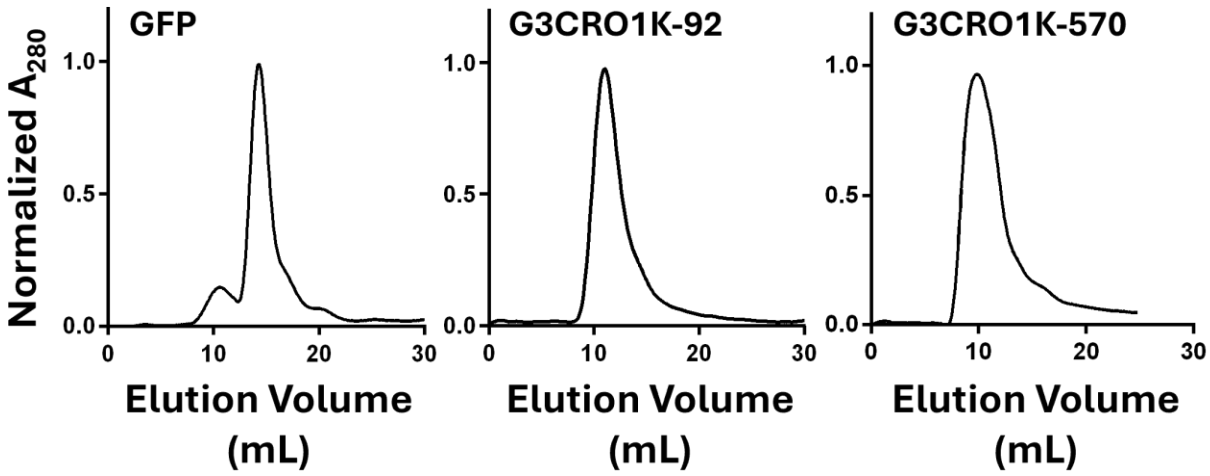



**Figure 25. Chromatograms from Second-Round Preparatory Size Exclusion Chromatography.** A Bio-Rad BioLogic Duo-Flow FPLC system was used with an ENrich™ SEC 650 10 x 300 column (#7801650). A flow rate of 0.5 mL/min was used with 0.5 mL fractions being collected. Samples were purified into phosphate-saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5). Variants of lilGFP were designed to be monomeric and have a mass ~7 kDa less than GFP, but all elute ~ 10 mL, which is earlier than GFP and near the void volume of the column.

Secondary structure analysis using CD revealed that these designs exhibited semi-folded and non-GFP-like structures (**Figure 26**). Both designs displayed negative signals at 208 nm, 218 nm, and 222 nm, suggesting the presence of both α-helical and β-sheet character. Additionally, each design showed a negative signal at 200 nm, which, while not a distinct peak, could indicate the presence of random coil structures. At 95˚C, the negative peak at 200 nm became more pronounced, suggesting some degree of unfolding at the elevated temperature. However, since

there was minimal loss of signal in the 215-225 nm range, most notably for G3CRO1K-570, it appears that any unfolding observed is likely only partial.

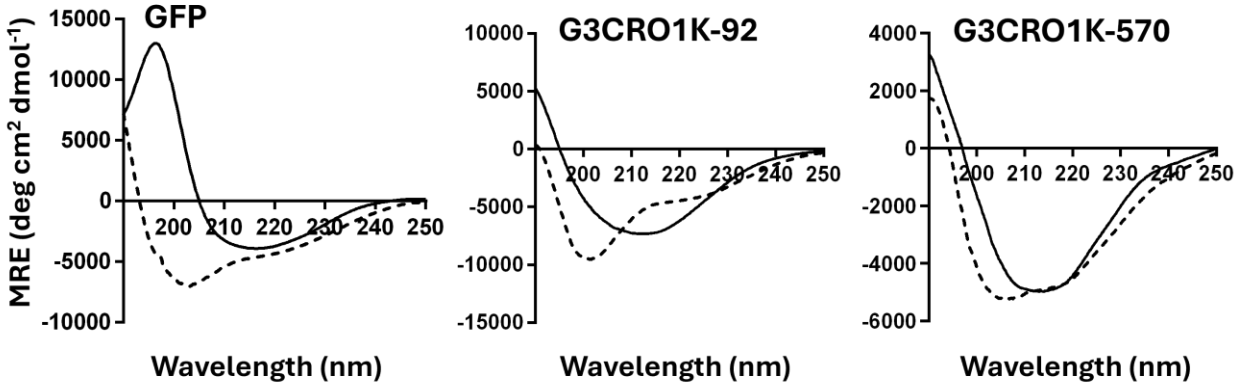



**Figure 26. Circular Dichroism Spectra of Second-Round lilGFP Secondary Structure.** Circular Dichroism (CD) spectra were determined using a J-815 CD Spectrometer (Jasco), Peltier Type FDCD Cell Holder (Jasco), and a Hellma Absorption Microcuvette (Millipore Sigma) with a 1 mm pathlength. Purified 10 μM samples were prepared in 20 mM sodium phosphate, pH 7.4. Secondary scans were done with 3 accumulations, a digital integration time (D.I.T.) of 1 second, continuous scanning mode with a scanning speed of 10 nm/min, using a data pitch of 0.2 nm, and band width of 1 nm. Scans were done at 25˚C (−) and 95˚C (--). The purity of each sample was determined using ImageJ[117] analysis of SDS-PAGE gels after size exclusion chromatography, yielding the following results: GFP – 98%, G3CRO1K-92 – 86%, G3CRO1K-570 – 96%.

CD was also used to perform thermal melting assays on G3CRO1K-92 and G3CRO1K-570 (**Figure 27**). Both designs appeared to show some melting as their ellipticity values began returning to baseline with increasing temperature, but also lacked the sigmoidal character typical of a well-folded protein undergoing cooperative unfolding[133]. No upper asymptote, typical of the unfolded state in a folded-unfolded two-state model[119], was seen for either design, hinting that either complete melts were not achieved, or that gradual and non-cooperative unfolding is occurring. The $|\Delta_{MRE}|$ during unfolding for G3CRO1K-92 and G3CRO1K-570 was 0.1 and 7.25, respectively. Compared to the $|\Delta_{MRE}|$ of 29 observed for GFP unfolding, these markedly lower values suggest that the second-round designs may populate a significant fraction of molten

globule-like conformations. Due to these factors, fitting of the Mean Residue Ellipticity (deg $cm^2$ $dmol^{-1}$) values using the Greenfield method[119] did not result in reliable $T_M$ or $\Delta H_{unfolding}$ values, meaning that no Fraction Unfolded plots could be obtained.

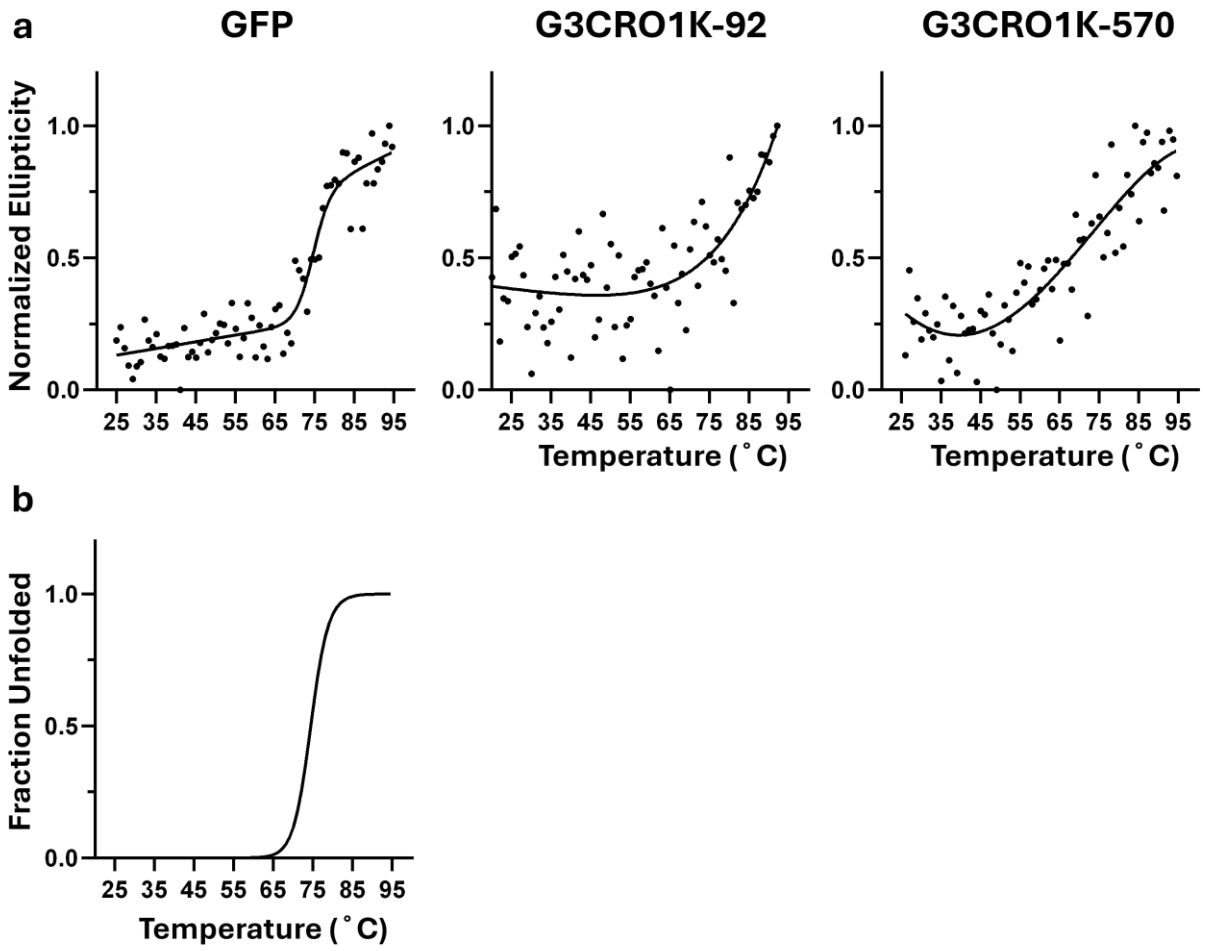



**Figure 27. Second-Round Thermal Melt Assays using Circular Dichroism.** Thermal Melt assays were done using Circular Dichroism with a J-815 CD Spectrometer (Jasco), Peltier Type FDCD Cell Holder (Jasco), and a Hellma Absorption Microcuvette (Millipore Sigma) with a 1 mm pathlength. Purified 10 μM samples were prepared in 20 mM sodium phosphate, pH 7.4. Melting curves were done with a D.I.T. of 8 seconds, standard sensitivity, and a band width of 1 nm. Temperature for the melting curves was increased at a rate of 1˚C/min and sampled every 1˚C after a 10 second wait upon reaching the interval temperature. (a) Normalized ellipticity values (dots) from the thermal melts were used for fitting with the Greenfield equations[119] (thin line), resulting in (b) fraction unfolded plots. A $T_M$ of 74.4 ˚C was determined for GFP, while no $T_M$ values were able to be determined for designs. The purity of each sample was determined

using ImageJ[117] analysis of SDS-PAGE gels after size exclusion chromatography, yielding the following results: GFP – 98%, G3CRO1K-92 – 86%, G3CRO1K-570 – 96%.

### 3.4.4 Miniaturized GFPs can be Enhanced through Mutagenesis

To remedy the issues mentioned above we attempted to create an improved variant using Directed Evolution, with G3 acting as a starting template. To generate mutant libraries epPCR was used (see *Chapter 4.4 Random Mutagenesis*) leveraging imbalanced deoxynucleotide triphosphates (dNTPs), increased $MgCl_2$ concentration, and the inherent error-rate of *Taq* Polymerase due to its lack of proofreading ability[134]. We obtained a final mutant library of 300,000 variants, which resulted from two parallel trajectories of iterative rounds of mutagenesis (**Table 5**). mutant trajectory 1 (MT1) averaged 2 mutations per gene and round over four rounds, while mutant trajectory 2 (MT2) averaged 1 mutation per gene and round over seven rounds. Our final mutant library (MT1 + MT2) contained variants with mutations ranging from 0 to 10 mutations per gene, as confirmed by Sanger and Long-Read sequencing. Sanger sequencing was used to identify mutation of the genes, while Long-Read sequencing confirmed that no mutations were occurring in promoter regions (**Supplementary Figure 6**).

**Table 5. Mutagenic Libraries of Green Fluorescent Proteins.** G3 was used as the template for two streams of iterative rounds of mutagenesis (MT1 and MT2, respectively). MT1 averaged 2 mutations per gene and round, while MT2 averaged 1 mutation per gene and round. Mutagenesis was done using epPCR (see *Materials and Methods)* for a total of twelve rounds. A mutagenic library containing 300,000 G3 variants was created, with confirmed mutations ranging from 0 to 10 per gene. To determine the average number of mutations per gene and mutagenesis round, 5 isolated colonies from each transformation of mutant libraries were sequenced using the Sanger protocol.

| Round | Library Size | Average Mutations gene$^{-1}$ |
|---|---|---|
| **MT1** | | |
| NH1 | ≤ 20,980 | 1 |
| NH2 | ≤ 29,560 | 4 |
| NH3 | ≤ 59,400 | 5 |
| NH4 | ≤ 49,000 | 7 |
| **MT2** | | |
| OB1 | ≤ 33,200 | 1 |
| OB2 | ≤ 29,800 | 3 |
| OB3 | ≤ 27,000 | 2 |
| OB4 | ≤10,000 | 3 |
| OB5/6 | ≤ 12,600 | 2 |
| OB7 | ≤ 10,000 | 5 |
| OB8 | ≤ 15,000 | 4 |

For Directed Evolution to take place, improved variants must be selected and further diversified. For High-Throughput selection we employed fluorescence-activated cell sorting (FACS). Using the Sony MA900 (Sony Biotechnology) cell sorter, unmutated G3 was unable to be differentiated from background noise (**Figure 28**). We hypothesized that upon the introduction of beneficial mutations, a population of improved variants would be able to be separated from noise and therefore selected. However, none of the mutants met the selection criteria (**Figure 28**).

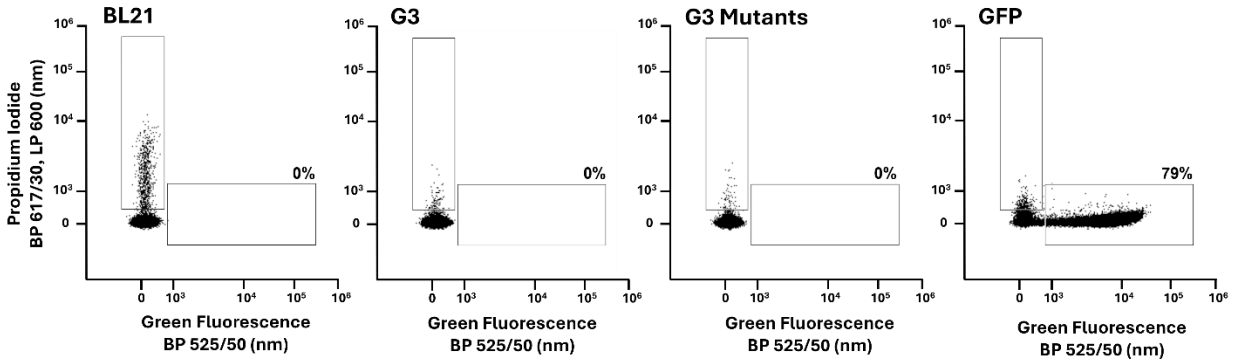



**Figure 28. Cytograms of Green Fluorescent Proteins Generated During Fluorescence Activated Cell Sorting (FACS).** Populations of BL21-Gold (DE3) cells expressing a) empty pET-29b(+), b) G3, c) MT1/MT2 mutants, and d) EGFP were sorted by FACS. Cells were expressed and then stored in the dark at 4˚C for 48 hours to allow for chromophore maturation. Populations were concentrated to $10^6$ cells $mL^{-1}$ and stained with propidium iodide (PI) to assess viability. A PI gate was set up on the Y-axis and a GFP gate was set up on the X-axis. Cells were sorted at 4000 events $second^{-1}$. For analysis of each population, 100,000 cells were used.

Since the starting point (G3) exhibited low fluorescence that was indistinguishable with FACS, we used a flow cytometer with a lower detection limit to ensure that no improved variants were overlooked. The Sony MA900 that was used for FACS had a relevant fluorescence sensitivity of FITC $\leq$ 94 MESF[135], meaning that it can detect 94 or fewer Molecules of Equivalent Soluble Fluorophore (MESF) using Fluorescein Isothiocyanate (FITC) as a standard green fluorophore. On the other hand, the CytoFLEX S Cytometer (Beckman Coulter) has a fluorescence sensitivity of FITC $\leq$ 30 MESF[136]. This enhanced sensitivity allowed us to identify improvements in fluorescence that had previously gone undetected.

When looking at histograms of the Flow Cytometry results (**Figure 29**) one can see that the average green fluorescence of G3 variants has been reduced compared to G3. This is to be expected since random mutations are known to be largely deleterious, with approximately 1/3 having severely deleterious effects on protein function[126,137]. However, a closer look at the upper limit of green fluorescence in both the G3 and mutant variants (**Figure 29b,c**) revealed a shoulder

region indicating that a small population of G3 variants exhibited enhanced fluorescence above that of G3. The small size of this population is not surprising given that only 0.01-0.5% of random mutations will be beneficial[126]. Based on these data (n=10,000), 0.32% of mutants showed green fluorescence above 3500 FITC-H, compared to 0.01% for G3. These results suggest that beneficial mutations were successfully introduced into the G3 template; however, it remains unclear whether the observed improvement in green fluorescence arises from enhanced photophysical properties of individual proteins or from indirect effects such as increased expression levels or solubility.

Having demonstrated that G3 can be mutagenized to enhance the fluorescence of expressing cells, and that these improvements are detectable using instrumentation with a lower detection limit, evolving these designs should be achievable with the proper selection method. Even if the observed fluorescence increase is solely attributable to improved expression or solubility, this would still represent a successful outcome, as these enhancements confer advantageous properties that could render G3 variants more amenable to acquiring additional, potentially beneficial mutations. The CytoFLEX SRT Benchtop Cell Sorter (Beckman Coulter) has the same fluorescence sensitivity as the CytoFLEX S Cytometer, and would therefore be a suitable instrument for the task.

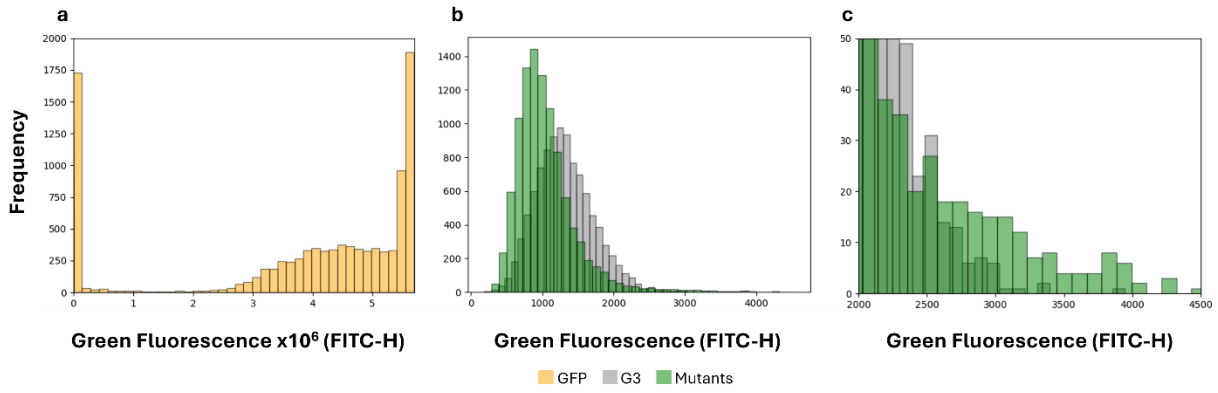



**Figure 29. Histograms of Flow Cytometry Data for Green Fluorescent Proteins.** Flow Cytometry was carried out on GFP, G3, and G3 mutant populations using a CytoFLEX S Cytometer (Beckman Coulter). Cells were expressed and then stored in the dark at 4˚C for 48 hours to allow for chromophore maturation. Populations were concentrated to $10^6$ cells $mL^{-1}$ and stained with propidium iodide (PI) to assess viability. For each population of GFP (orange), G3 (grey), and G3 mutant samples (green), 10,000 randomized cells were analyzed.

## Chapter 4. Materials and Methods

### 4.1 Initial Computational Design of Miniature Green Fluorescent Proteins

Wildtype $GFP^{S65T}$'s crystal structure (PDB: 1EMA)[34] was used as the starting template for GFP miniaturization. Using PyMOL 2.5.4, residues 12-19, 27-34, 39-45, 56-64, 68-72, 83-98, 105-113, 117-124, 144-152, 161-167, 179-185, 199-206 and 217-224 were conserved, while all others were deleted. Using PyMOL's *Mutagenesis* function and backbone dependent rotamers, a C70S substitution was made to avoid the formation of undesirable disulfide bridges upon protein expression. A .pdb file was created from these remaining atomic coordinates, titled *cleavedGFP*. To connect the fragments left by the absence of 1EMA's chromophore (PDB ID: CRO), three residues were diffused between *cleavedGFP*'s F64 and V68 using RFdiffusion[89]. PyMOL's *Mutagenesis* function and backbone dependent rotamers were then used to recapitulate the wildtype sequence identity (T65-Y66-G67) on this diffused backbone, creating the *preGFP* structure. To monomerize the design, connecting structures with lengths between 1 and 22 residues

were diffused, using RFdiffusion, between *preGFP* fragments according to their N-C connectivity in 1EMA. Diffused structures were then visually inspected for the presence of a β-can and capping α-helices. Structures were rediffused on a fragment-by-fragment basis as necessary. Structures which passed visual inspection had their sequences redesigned using ProteinMPNN[91], except for analogous residues of those found within 8 Å of the 1EMA chromophore. Redesigned sequences from ProteinMPNN were submitted to AlphaFold 2.3.2[93] and filtered using visual inspection and pLDDT scoring.

### 4.2 Computational Redesign of Initial Miniature Green Fluorescent Proteins

AlphaFold models of G1, G2, and G3 were used as templates for the second round of miniature green fluorescent protein design. Atoms analogous to those of residues 65, 66, and 67 in the 1EMA PDB were deleted and replaced by the chromophore found in the 1EMA crystal structure. The chromophore-containing templates were then standardized in Triad (Protabit), which optimized bond lengths and side-chain dihedral angles. Second round sequences were then obtained by running LigandMPNN's ligand_mpnn model on the standardized chromophore-containing templates. Analogous residues of those found within 4 Å of the 1EMA chromophore were fixed. Parameters used were; ligand_mpnn_checkpoint = ligandmpnn_v_32_005_25.pt, checkpoint_path_sc = ligandmpnn_sc_v_32_002_16.pt, ten batches of size one hundred, a temperature of 0.1, pack_side_chains = 1, two packs per design, pack_with_ligand_context = 1, and repack_everything = 0. 100 G1 sequences, 100 G2 sequences, and 1400 G3 sequences were submitted to AlphaFold 2.3.2[93]. Triad was used to clean second-round sequences on their respective standardized, chromophore-containing templates, and scored using the Phoenix energy function[125]. Second round designs were filtered using AlphaFold pLDDT's, Triad energy scores,

and the RMSD between inner helices of AlphaFold models, cleaned Triad models, and the 1EMA crystal structure.

### 4.3 Green Fluorescent Protein Genes

All amino acid sequences for GFPs within this document can be found in **Supplementary Table 2**. Stephen L. Mayo generously gifted the wildtype $GFP^{S65T}$ gene encoded on a pET-11a vector[138]. Miniature GFPs were N-terminally his-tagged and their codon-optimized genes were cloned by Twist Bioscience into pET-29b(+) vectors (Novagen) using *NdeI* and *XhoI* (New England BioLabs) restriction sites. All first round designs had a three amino acid linker for their his-tags. All second-round designs except G1A50 had a tryptophan added to their his-tag to aid with protein quantification. All plasmids from Twist Bioscience were transformed into *E. coli* BL21-Gold (DE3) cells (Agilent) for protein expression. Mutants of miniature GFPs were prepared as described in the *Random Mutagenesis* section.

### 4.4 Random Mutagenesis

Random mutagenesis was performed using G3 as the starting template, using a modified version of the Cirino *et al.* protocol[127]. In short, 25 rounds of error-prone polymerase chain reaction catalyzed by 5 U *Taq* DNA Polymerase in 1X Standard *Taq* Buffer pH 8.3 (New England Biolabs) supplemented with a mixture of deoxynucleotides (0.2 mM dATP, 0.2 mM dGTP, 0.6 mM dTTP, and 0.6 mM dCTP, Thermo Fisher Scientific) and 2.5 mM $MgCl_2$ were used for the introduction of mutations. 5 ng of template DNA was used with primers at a concentration of 1 μM each. After following the Cirino *et al.* protocol, 25 rounds of conventional *Taq* DNA Polymerase (New England Biolabs) amplification were performed on 10 ng of the mutant amplicons, according to the manufacturer's published protocol. Each mutant library was cloned into a pET-29b(+) vector

(Novagen) using *NdeI* and *XhoI* (New England BioLabs) restriction enzymes. Ligated plasmids were transformed into *E. cloni* 10G Elite electrocompetent *E. coli* cells (Lucigen) and plated on LB Top Agar (Teknova) supplemented with 50 μg/mL kanamycin. After 12-16 hours of incubation at 37 °C, all colonies were collected using 5 mL of 50 μg/mL kanamycin supplemented LB. Collected colonies were then amplified overnight at 37 °C with 220 rpm shaking and had their plasmids extracted using the E.Z.N.A. Plasmid DNA Mini Kit II (Omega Biotek). Collected and purified DNA was transformed into BL21(DE3) Electrocompetent *E. coli* cells (Sigma-Aldrich) for protein expression and Fluorescence-Activated Cell Sorting. Mutations were confirmed by DNA sequencing.

**4.5 Fluorescence-Activated Cell Sorting**

Electrocompetent *E. coli* cells (Sigma-Aldrich) containing random mutant libraries described in the *Random Mutagenesis* section were used as inoculant for 5 mL of LB supplemented with 50 μg/mL kanamycin. Cells were grown at 37 °C with shaking at 225 rpm until reaching an optical density at 600 nm of 0.6. Induction of protein expression was triggered with the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (Fisher Scientific) and cultures were then incubated for a further 16 hours at 16 °C with shaking at 225 rpm. After incubation, cells were washed twice with sterile phosphate buffer (PBS) (20 mM phosphate and 50 mM sodium chloride, pH 7.4), pelleted by centrifugation, and stored in darkness at 4 °C for 48 hours to allow for chromophore maturation. Cell pellets were then resuspended in 5 mL of sterile PBS, pH 7.4. Cells were stained for viability by adding 5 μL of Propidium Iodide (Invitrogen) to 100 μL of cells, incubating in darkness for 30 minutes, and then adding 900 μL of sterile PBS, pH 7.4, to a final concentration of $10^6$ cells/mL. A MA900 Multi-Application Cell Sorter (Sony Biotechnology) was used for sorting of cells

expressing green fluorescent protein based on their fluorescence intensity ($\lambda_{ex}$ = 488 nm, $\lambda_{em}$ = BP 525/50 nm).

### 4.6 Flow Cytometry

BL21(DE3) Electrocompetent E. coli cells (Sigma-Aldrich) containing random mutant libraries described in the *Random Mutagenesis* section were used as inoculant for 5 mL of LB supplemented with 50 μg/mL kanamycin. Cells were grown at 37 °C with shaking at 225 rpm until reaching an optical density at 600 nm of 0.6. Induction of protein expression was triggered with the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (Fisher Scientific) and cultures were then incubated for a further 16 hours at 16 °C with shaking at 225 rpm. After incubation, cells were washed twice with sterile phosphate buffer (20 mM phosphate and 50 mM sodium chloride, pH 7.4), pelleted by centrifugation, and stored in darkness at 4 °C for 48 hours to allow for longer chromophore maturation. Cell pellets were then resuspended in 5 mL of sterile PBS, pH 7.4. Cells were stained for viability by adding 5 μL of Propidium Iodide (Invitrogen) to 100 μL of cells, incubating in darkness for 30 minutes, and then adding 900 μL of sterile PBS, pH 7.4, to a final concentration of $10^6$ cells/mL. A CytoFLEX S Cytometer (Beckman Coulter) was used for flow cytometry experiments with a flow rate of 10 μL/min.

### 4.7 Protein Expression and Purification in Large Batches

GFP variants were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using LB supplemented with 100 μg/mL ampicillin (pET-11a) or 50 μg/mL kanamycin (pET-29b(+)). Cells were grown at 37 °C with shaking at 220 rpm until reaching an optical density at 600 nm of 0.6. Induction of protein expression was triggered with the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (Fisher Scientific) and cultures were then incubated for a further 16 hours

at 16 °C with shaking at 220 rpm. After incubation, cells were washed twice with sterile phosphate buffer (20 mM phosphate and 50 mM sodium chloride, pH 7.5), pelleted by centrifugation, and stored in darkness at 4 °C for 48 hours to allow for longer chromophore maturation. Cell pellets were resuspended in 7.5 mL of lysis buffer (100 mM potassium phosphate buffer, 15 mM imidazole, 1 mg/mL lysozyme, and 50 U benzonase nuclease [Novagen], pH 8). Cells were lysed with an EmulsiFlex-B15 cell disruptor (Avestin), cellular debris was removed via centrifugation, and lysates were filtered with 0.45 µm Acrodisk syringe filters (Pall Corporation). Proteins were purified by immobilized metal affinity chromatography using Ni-NTA (Qiagen) resin and Econo-Pac chromatography columns (Bio-Rad) according to the manufacturer's protocol. Elution fractions were desalted using Econo-Pac 10DG Desalting Columns (Bio-Rad) into 100 mM phosphate buffer, pH 7.5. An additional gel purification step was carried out for circular dichroism experiments into 20 mM sodium phosphate, using a BioLogic DuoFlow fast protein liquid chromatography system (Bio-Rad) and an ENrich™ SEC 650 10 x 300 column (Bio-Rad).

### 4.8 Quantification of Protein

Absorbance measurements of 10X diluted samples were made at 280 nm using a NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Fisher Scientific) and its *Other Protein (ε + MW)* function. Based on the *Beer-Lambert law* (**Equation 7**) and published absorption coefficients[139]. The sample's concentration ($c$) is determined in mg/mL, using the absorbance at 280 nm for $A$, the protein's extinction coefficient ($M^{-1}$ $cm^{-1}$) as calculated by the Expasy ProtParam online tool (Swiss Institute of Bioinformatics) as $\varepsilon$, and the path length (0.1 cm) as $l$. The protein's percent extinction coefficient (ε1%) is calculated from user-input values (**Equation 8**), and then converted to a concentration with units of mg/mL after multiplying the equation by a factor of 10

(**Equation 9**). Reported values were converted to molarity when necessary, using dimensional analysis.

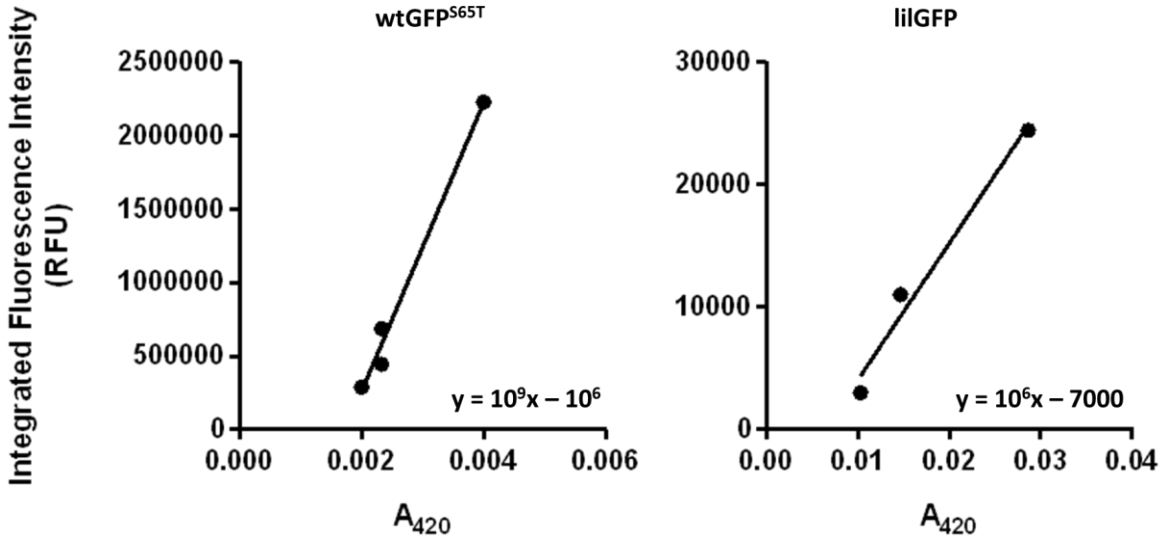
$$\varepsilon 1\% \left(\frac{g}{100\ mL{*}\ cm}\right) = \frac{\varepsilon_{molar}{*}10}{MW} \quad \text{eq. [8]}$$

$$c\ \left(\frac{mg}{mL}\right) = \frac{A{*}10}{\varepsilon 1\%} \quad \text{eq. [9]}$$

### 4.9 Characterization of Spectral Properties

Spectra for chromophore absorption, excitation and emission were measured at 25 °C in phosphate-saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5) using Black UV-Star 96-Well Microplates (Greiner Bio-One) and a BioTek Synergy H1 Microplate Spectrophotometer (Agilent).

Quantum yields (Φ) of GFP variants were calculated by comparing their integrated fluorescence intensity with that of an equally absorbing sample of wtGFP$^{S65T}$ (Φ = 0.66[22]) with excitation at 420 nm. This can be seen plotted below and with **Equation 10**.



$$\Phi_{\text{lilGFP}} = \Phi_{\text{wtGFP}^{\text{S65T}}} \frac{\text{m}_{\text{lilGFP}}}{\text{m}_{\text{wtGFP}^{\text{S65T}}}} \quad \text{eq. [10]}$$

$$\Phi_{\text{lilGFP}} = (0.66)\frac{10^6}{10^9}$$

$$\Phi_{\text{lilGFP}} = 0.0007$$

Quantum yield measurements are not impacted by non-fluorescent contaminants in a sample due to only green chromophore-containing molecules contributing to the absorbance and emission values (420 nm and 435-700 nm, respectively) used in their associated calculations.

Extinction coefficients (ε) of chromophores were determined using the *Beer-Lambert law* (**Equation 7**). For each sample, serial dilutions were made in phosphate-saline buffer (100 mM potassium phosphate and 100 mM sodium chloride, pH 7.5) using UV-Star 96-Well Microplates (Greiner Bio-One) and subject to complete absorbance scans with a BioTek Synergy H1 Microplate Spectrophotometer (Agilent). Pathlengths *(l)* were determined per well using absorbance reads of 900 nm and 977 nm, and **Equation 11**[140,141] shown below, where *k* is a known constant. Concentrations (*c*) were determined using absorbance readings at 280 nm as described in *Quantification of Protein*. Final chromophore extinction coefficient values were determined by calculating the slope of the absorbance reading at the wavelength of the chromophore's maximal absorbance ($A_{\text{max}}$), divided by the pathlength vs *c* (**Equation 12**).

$$l = \frac{(A_{977} - A_{900})_{sample}}{(A_{977} - A_{900})_{1.0\ cm\ water}} = \frac{(A_{977} - A_{900})_{sample}}{k} = \frac{(A_{977} - A_{900})_{sample}}{0.18} \quad \text{eq. [11]}$$

$$A = \varepsilon * l * c \rightarrow \frac{A}{l} = \varepsilon * c \equiv y = m * x + b \quad \text{eq. [12]}$$

### 4.10 Circular Dichroism

Circular Dichroism spectra were obtained using a J-815 CD Spectrometer (Jasco), Peltier Type FDCD Cell Holder (Jasco), and a Hellma Absorption Microcuvette (Millipore Sigma) with a 1 mm pathlength. Purified samples were prepared in 20 mM sodium phosphate, pH 7.4. Secondary scans were done with 3 accumulations, a digital integration time (D.I.T.) of 1 second, continuous scanning mode with a scanning speed of 10 nm/min, using a data pitch of 0.2 nm, and band width of 1 nm. Melting curves were done with a D.I.T. of 8 seconds, standard sensitivity, and a band width of 1 nm. Temperature for the melting curves was increased at a rate of 1 ˚C/min and sampled every 1 ˚C after a 10 second wait upon reaching the interval temperature.

## Chapter 5. Summary and Outlook

Substantial miniaturization of GFP, while maintaining autocatalytic chromophore maturation and fluorescence, is indeed achievable, as demonstrated by our designs. Our designs shared between 35 and 82 percent sequence identity amongst each other (**Supplementary Table 3**), meaning this wasn't the result of oversampling a small region of sequence space. However, achieving fluorescence levels close to that of GFP remains elusive with current computational methods. This limitation likely arose from our approach, which did not explicitly design chromophore cyclization and maturation. Instead, we relied on state-of-the-art MLAPD, which, while powerful, may not yet be able to predict the complex dynamics needed to promote efficient chromophore maturation.

Given the rapid pace with which these methods are being developed, it is possible that software now exists which was unavailable during our GFP miniaturization efforts, and that they could potentially be used to design enhanced fluorescence while using the same pipeline. As mentioned earlier, RFdiffusion All-Atom and RFdiffusion2 can model ligands, which was not possible when

we used RFdiffusion to nucleate backbone around the 8 Å spheroid of GFP fragments. At that time, we had to diffuse uncyclized analogues corresponding to Thr65, Tyr66, and Gly67 instead. In doing so, we may have designed backbones which contain chromophore pockets that are less compatible with a mature chromophore. While we used the soluble_mpnn model of ProteinMPNN in the first round of sequence design, we employed the more recent LigandMPNN in the second round, specifically implementing the ligand_mpnn model. Given the continued aggregation of our designs and the limited representation of GFP's chromophore, a rare post-translational modification, in the PDB used to train both models[123], it may, in retrospect, have been beneficial to explore the soluble_mpnn variant of LigandMPNN. For validation of our designs we had used AlphaFold2, which, while proven to be accurate, is now outperformed by AlphaFold3 and Chai-1[142]. Incorporating these newer models as part of a final filtering stage in a third round of miniaturization, or even simply cross-validating our previous designs, would be worth exploring.

Barring a third design round, our results suggest that evolving these lilGFPs may be a viable path forward, even though we didn't succeed in doing so here. We did however demonstrate that mutagenesis can potentially enrich a population of the designs for improved green fluorescence, indicating that when coupled with an appropriate selection method, Directed Evolution should be effective. If simple random mutagenesis can improve fluorescence, the accumulation of synergistic beneficial mutation, such as in epistatic networks[143], should compound this effect.

To discuss why these designs are believed to be evolvable, we must first consider the factors that hinder their fluorescence. Since no structures for the designs have been determined, we can only hypothesize about the underlying causes. The most compelling evidence suggesting that these designs need improvement, and are likely amenable to such improvement, is the absence of a significant absorbance peak at their excitation maxima (**Figures 14 and 22**), despite the

presence of some fluorescence. The lack of a significant absorbance peak indicates that 4-(p-hydroxybenzylidene)imidazolidin-5-one chromophore is not maturing effectively[64]. This could be the result of poor folding, cyclization and/or maturation.

It is difficult to assert which process is acting as the rate-limiting or obstructing step in the maturation of these miniaturized GFP variants. However, folding of the β-barrel is crucial for the cyclization of the critical residues Thr65/Tyr66/Gly67, which are involved in chromophore formation. While evolving improved folding efficiency may not directly increase fluorescence, it would increase the proportion of proteins available to undergo cyclization and maturation[36]. If during evolution the fold were to be changed slightly in such a way that led to less solvent-induced quenching or increased chromophore rigidity, then fluorescence would be improved via better quantum yields. The SEC chromatograms showed undesired higher-order oligomers that elute near the void volume (**Figures 15 and 23**), thermal melts with little to no cooperative unfolding (**Figures 17 and 25**), and the likely presence of inclusion bodies (**Supplementary Figures 1, 4, and 5**) point to there being a folding issue occurring. This would not be surprising, since with regular GFP variants, the first step in the folding equilibrium is the formation of a molten globule, which is a partially folded state prone to aggregation[144]. If GFPs normal folding pathway was perturbed by the significant sequence alteration, it would come as no surprise if this aggregation was exacerbated. In future studies, hydrophobic fluorescent probes such as 1-anilino-8-naphthalenesulfonic acid could be employed to assess the prevalence of molten globule populations in these designs, as it is known to interact with hydrophobic surface patches exposed in the molten globule state.[145]. These folding issues could theoretically all be improved via Directed Evolution as larger regions of sequence space are sampled.

If the proper β-barrel fold is being adopted, there is no guarantee that the internal architecture is conducive to efficient cyclization of the Thr65/Tyr66/Gly67 tripeptide. In GFP and its fluorescent variants, the kinked central helix is imperative for cyclization to occur. As mentioned earlier, this kink (tight-turn) aligns Gly67's lone electron pair with the $\pi^*$ orbital of Thr65's carbonyl carbon, priming it for attack (**Figure 3**). Additionally, this kink also precludes the formation of 9 out of 12 possible hydrogen bonds on the backbone of the central helix, thereby reducing the number of hydrogen bonds needing to be broken for cyclization, lowering its enthalpic barrier[57]. If the central helix in our designs is misaligned upon folding, it would be disastrous for cyclization and therefore fluorescence. Another complicating factor is the potential misalignment of key residues, such as those analogous to Arg96 and Glu222. In GFP, the loss of positive charge at position 96 changes the rate-limiting step in chromophore maturation from oxidation to cyclization[57]. If the analogous arginines in our designs are displaced, it could lead to a similar disruption in maturation. As mentioned earlier, Glu222 mediates its effects through a network of structured water molecules. Without a structure, there is no telling what havoc may be caused to this network by unintended and unpredicted changes to the internal chromophore environment. Even if the positioning of GFP residues that we preserved in our 8 Å spheroid of fragments was recapitulated to a sub-angstrom RMSD, the presence of novel capping residues on both sides of the β-barrel could have any number of untold consequences on maturation dynamics. This could be caused by an inward pressure disrupting normal side-chain behaviour, or alternatively a lack of outward resistance which would allow for a more loosely packed chromophore environment.

Given that our designs do indeed exhibit some fluorescence, one aspect that can be more accurately diagnosed is the low quantum yields observed in these variants. If we are to assume that

at least a small proportion of our expressed designs are adopting the β-barrel fold, based on there being fluorescence with the correct spectra, then there are two very likely causes of the low quantum yield which originate from the same source, the complete structural remodeling of the capping residues on both ends of the β-barrel. As discussed in *Chapter 1.3 The Chromophore and its Role in Fluorescence*, a rigid chromophore is essential for minimizing energy lost to processes of non-radiative decay, such as TICT. With the extensive changes made to the capping architecture around the chromophore's local environment, the newly introduced residues (or absence of former residues) may be interfering with the critical residues analogous to GFPs. The internal rotameric environment of GFP is thought to be influenced by the rigidity of the surrounding β-barrel backbone, which leaves sidechain packing largely dependent on shape and charge complementarity[146]. By drastically remapping this internal landscape, sidechains critical for maintaining a rigid chromophore environment may now be allowed to sample non-productive conformations. A potential example of this would be Tyr66, which when in a photoexcited state is normally prevented from adopting the rotamer associated with the lowest-energy non-radiative decay pathway by the steric hindrance of two other sidechains in the microenvironment[146]. Disruption of either of these residues could promote nonradiative decay and result in the observed lower quantum yield.

The second most likely source of low quantum yields connected to the new capping environments is solvent- and oxygen-induced quenching. At its core, GFP maturation relies on a buried network of solvent-mediated hydrogen bonds[146]. Gln69, Gln94, Arg96, and Gln183 are of particular importance and form hydrogen bonds with the chromophore either directly or indirectly via water molecules[146]. Molecular oxygen availability is also essential (**Scheme 1 and 2**). While water and oxygen are clearly critical for fluorescence, exposure of the chromophore to an excess

of either can promote fluorescence quenching[144]. In the case of water, this occurs when a photoexcited molecule transfers an electron to water instead of emitting that energy as a photon. Oxygen quenching in solution results mainly from diffusive encounters of oxygen with the excited fluorophore[147], and is non-destructive since no complex is formed between the two[148]. Water's seemingly paradoxical effect can be explained by the fact that a single water molecule cannot quench fluorescence, but a cluster of water molecules can, as they are able to orient themselves to accept electrons without requiring additional energy[149]. In wild-type GFP, there is a known water pore lined by residues 146, 147, 167, 205, and 206[69], using 1EMA as a reference. This pore is theorized to provide the water molecules necessary to build up the hydrogen bond network implicated with Glu222 earlier in *Chapter 1.3 The Chromophore and its Role in Fluorescence*[69]. Slight enlargement of the pore in the engineered TurboGFP variant has been attributed to increased water diffusion and improved chromophore maturation[69]. However, with our drastic disruption of GFPs capping environment, including this pore (**Figure 30**), excess bulk solvent may be able to penetrate the interior of our designs and quench fluorescence by being allowed to form clusters capable of accepting electrons. This influx of water would also allow for more diffusion of fluorescence-quenching oxygen. Furthermore, the introduction of water can alter the polarity of the region in question, which may affect the mutual disposition of the chromophore's singlet and triplet states[149]. If the triplet state is lower than that of the singlet state, it's possible for the singlet state to transition to the triplet, with corresponding quenching of fluorescence[149] (**Figure 5**).

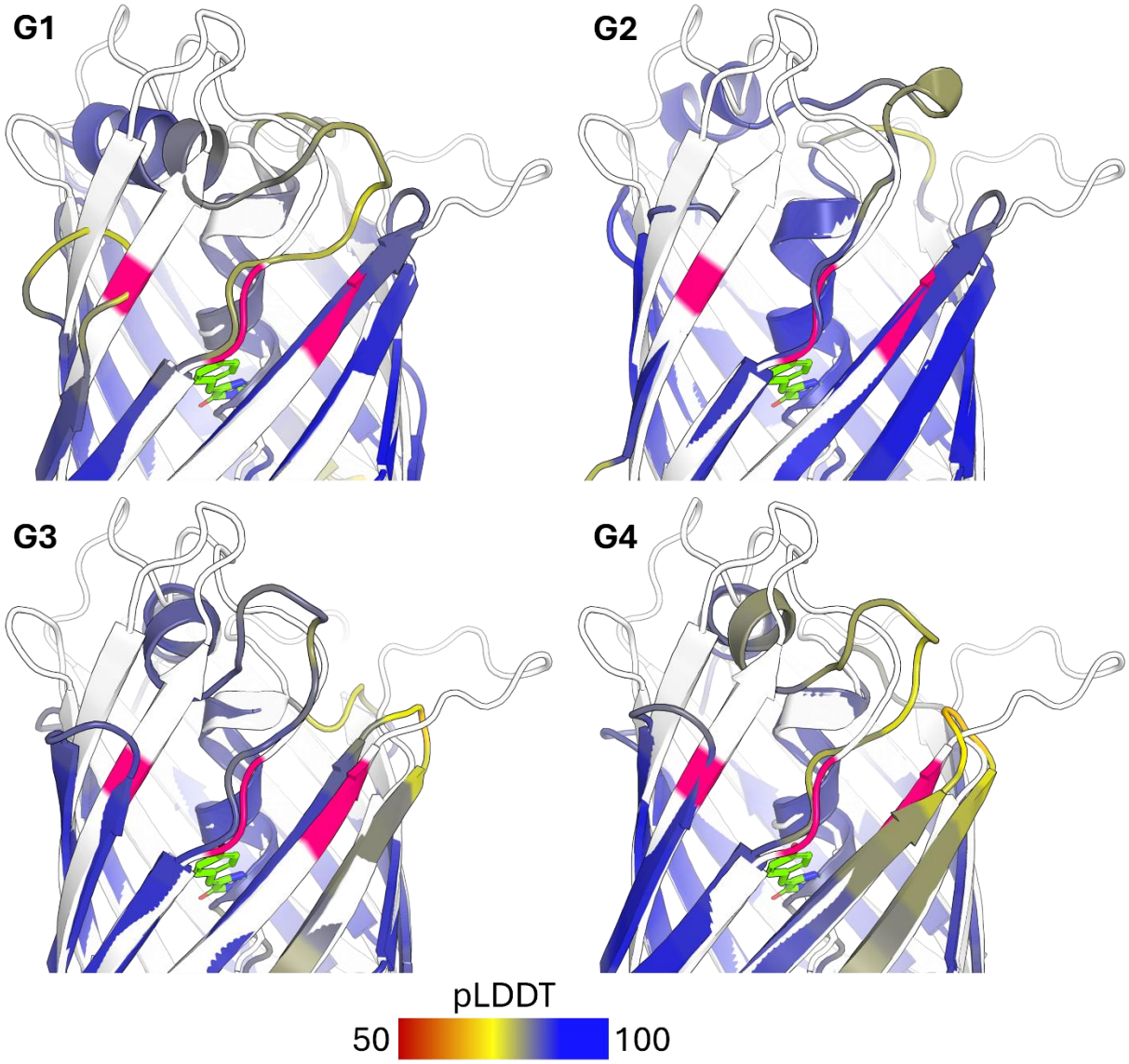



**Figure 30. Water Pore Comparison Between 1EMA and G1-G4 AlphaFold2 Predictions.** Residues 146, 147, 167, 205, and 206 (pink) in 1EMA line a pore known to allow limited water entry to GFP's interior[69]. AlphaFold2 models of G1-G4 are coloured by their per-residue pLDDT score (red-yellow-blue, increasing). The crystal structure of 1EMA (white) is shown as reference. G1 and G4 have low pLDDT values in this region, indicating caution must be shown interpreting these structures.

It's conceivable that these issues, if they are indeed occurring, could be overcome with Directed Evolution. There is some evidence suggesting that mutagenesis alone can improve the

fluorescence of our lilGFPs (**Figure 29**), and iterative rounds of selection should amplify these improvements. Directed Evolution has been shown to improve protein stability[126], quantum yields of FPs[150], and chromophore maturation[151]. Taken together, these successes provide a clear motivation to pursue a new Directed Evolution campaign with our designs. Our previous attempts to evolve G3 were stymied by instrument sensitivity limitations. However, having now determined that we have access to instrumentation with the required resolution to sort our mutant library, the only remaining question is when to begin this next phase of evolution.

## Supplementary Information

**Supplementary Table 1. Substitutions Made to G1 to Create G1-Enhanced Variants.**
Substitutions were only made if there were differing residues with G1 within the 8 Å spheroid of fragments initially preserved in the design process.

| Variant | Substitutions from G1 |
|---|---|
| G1-Emerald | M6L, G8F, G19Y, L32F, E34K, F42L, S50A, I54V, D56E, P73F, I75Y, I92L, F94A, D96E, Y98W, Q99K, L102E, N119K, Y122I, S123M, I131T, L146T, S147K, D148E, G149N, A156K, I157L |
| G1-GreenLantern | M6L, G8F, S12R, G19N, L32F, E34D, F42L, T43G, Q47A, S50A, T51D, I54V, D56E, P73S, I75Y, E91V, I92L, F94A, N97D, Y98W, Q99E, L102R, D103E, E111D, Y115F, N119K, Y122I, S123M, V127A, I131T, L146T, D148E, G149S, T153H, A156K, I157L, L163K, F165R |
| G1-Superfolder | M6L, G8Y, S12R, G19N, L32F, F42L, D56E, P73S, I75Y, I92L, F94A, D96E, Q99E, D103E, E111P, T114S, Y115F, Y122I, S123T, N124D, V127A, L146T, A156V, I157L, D158K, G159D |

**Supplementary Table 2. Amino acid Sequences of Green Fluorescent Proteins.**

| Protein | Sequence |
|---|---|
| EGFP | MHHHHHHSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGK<br>LPVPWPTLVTTFTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRA<br>EVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIR<br>HNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTA<br>AGITHGMDELYK |
| EGFP Fragments Used for RFdiffusion Template | PILVELDG—SVSGEGEG—GKLTLKF—WPTLVTTFTYGVQSFS—FKSAMPEG<br>YVQERTIF—YKTRAEVKF—TLVNRIEL—YNSHNVYIM—KVNFKIR—DHYQQ<br>NT—YLSTQSAL—MVLLEFVT |
| G1 | MHHHHHHMKTPILVE**M**DGGSVSGEGEGEGGKLTLKFDSDLSLEEWPTLVTTFTYGVQ**S**<br>FSTGKIGDDFKSAMPEGYVQERTI**P**G**I**KTRAEVKFGTLVNRIE**I**EFEDNYQVALDGAR<br>AALKEDGTYNSHNVY**YS**NNKVNFKIRGDHYQQNTGKNLKLSDGYLSTQSA**I**DGMVLLE<br>FVTDK |
| G2 | MHHHHHHMKT**K**ILVE**MKT**PY**D**V**K**GEGEGDGKLTLKFDLEGGIWPTLVTTFTYGVQ**S**FS<br>SDPAARAAF**QG**AMP**G**GYVQER**R**I**S**GG**FEV**RAEVK**A**G**P**LVNRIEL**K**IPEEALKKFETKP<br>QEFKYNSHNV**SYS**DA**E**V**T**FKIRGLDH**E**Q**R**NTRVDGTPVTFSSDV**T**LSTQS**YI**KDM**D**LL<br>EFV**G**PV |
| G3 | MHHHHHHMVE**K**ILVEL**EM**EE**K**VSGEGEGKDG**F**LTLKFDARAWPTLVTTFTYGVQ**S**FSD<br>DKSPASQVFKSAMPEGYVQERTIFG**I**KTRAEVK**E**E**P**LVNRIE**I**DAKLEELLERDPLLV<br>NEDGTYNSHNVYI**K**NG**E**VNFKIRGIDHYQQNTGKNLKPEET**T**LSTQSALDMVLLEFV**K**<br>ES |
| G4 | MHHHHHHMKT**K**ILV**K**L**KA**EK**EYE**GEG**V**GEDG**F**LTLKFDAEPWPTLVTTFTYGVQ**S**FSD<br>SSSPISQAFKSAMPEGYVQERTI**E**G**IEV**RAEVK**P**D**P**LVNRIELNAKLEELLARDPLLV<br>NDDGTYNSHNV**NYE**NGKVNFKIRGI**N**HYQQNTGKNLNLKPA**T**LSTQS**YG**SMVLLEFV**K**<br>PA |
| G1-Emerald | MHHHHHHPILVELD**F**GSVSGEGEGE**Y**GKLTLKFDSDLSFEKWPTLVTT**L**TYGVQ**S**F**A**T<br>GKVGEDFKSAMPEGYVQERTIFGYKTRAEVKFGTLVNRIELEAEENWKVAEDGARAAL<br>KEDGTYNSH**K**VYIMNNKVNFK**T**RGDHYQQNTGKNLKTKENYLSTQS**K**LDGMVLLEFVT<br>DS |

| | |
|---|---|
| G1-GreenLantern | MHHHHHHPILVELDFGSVRGEGEGENGKLTLKFDSDLSFEDWPTLVTTLGYGVASFAD<br>GKVGEDFKSAMPEGYVQERTISGYKTRAEVKFGTLVNRIVLEAEDDWEVAREGARAAL<br>KDDGTFNSHKVYIMNNKANFKTRGDHYQQNTGKNLKTSESYLSHQSKLDGMVLKERVT<br>DS |
| G1-Superfolder | MHHHHHHPILVELDYGSVRGEGEGENGKLTLKFDSDLSFEEWPTLVTTLTYGVQSFST<br>GKIGEDFKSAMPEGYVQERTISGYKTRAEVKFGTLVNRIELEAEENYEVALEGARAAL<br>KPDGSFNSHNVYITDNKANFKIRGDHYQQNTGKNLKTSDGYLSTQSVLKDMVLLEFVT<br>DV |
| G1A50 | MHHHHHHKIVVEVKGANITGEGEGKDGKIDLKFKTDLSSEEFPAFVTTFTYGVQCFAT<br>GEIGEAFKSAFPEGYTQTRNMPGITTKATVTEGPIKNKIDITYNDDYDKALSGAKAAL<br>HSDGTYNSHDVNYKNNKVTFNIGGNHEMLFEGKGLKIPEGKLHTKSKLDGTHLKEEVK<br>WV |
| G2A78 | MHHHHHHWKIIVKLKNPYDITGEGTGDGVLNLTFNLTGGIFPYLVTTFTYGVQCFSAD<br>ADLRAAFQGAFPEGYTQTRDISGGLKVTATVTAGPVTNEIELKIPEETIKELYTKPED<br>YKYNSHNVTFTNTETTFTINGRKHTLKNIKKDGTPVTLSKPVTAYTKSTLEGNKLTEE<br>VGPV |
| G3CRO1k-92 | MHHHHHHWKIEVILEMDKTITGSGTGENGKIVLKLDSKVIPPLVTTFTYGVQCFT<br>EPDTPVYKIYTSALPEGYVQKREIGGIKSEGVVTADPVVNRIKIDAKMEELEKLAPL<br>FVDENGNYYDHDVHVKNGKVTFTINGVEVKQEVTGKNLLPLEATIHTSLTLDDVLTEV<br>IKLK |
| G3CRO1k-219 | MHHHHHHWDVQVVLKMDKTITGSGTGADGLITLTLDSKVIPGLVTTFTYGVQCLADPG<br>TPVARMYTAALPEGYEQEREIGGIKSKGKVTADPLVNQIDIDAKMEEIEKLNPLFIDE<br>DGNYYPHLVHVSNGKVTFEINGVKVEQTVTGPNLLPLEATIETELTLSDVLTERIRLA |
| G3CRO1k-281 | MHHHHHHWAIEVNLTMDKTITGSGTGADGLITLTLDSKVIPGLVTTFTYGVQCFAEPD<br>TPVARMFNAAFPEGYVQKRKIGGIESTGKVTRDPVVNDIKIDAKLEEIEKLNPLFVNE<br>DGTYYPHLVHVENGKVTFTINGVKVEQEVTGPNLLPLTATIFTEKTLSDVLTERIRLA |

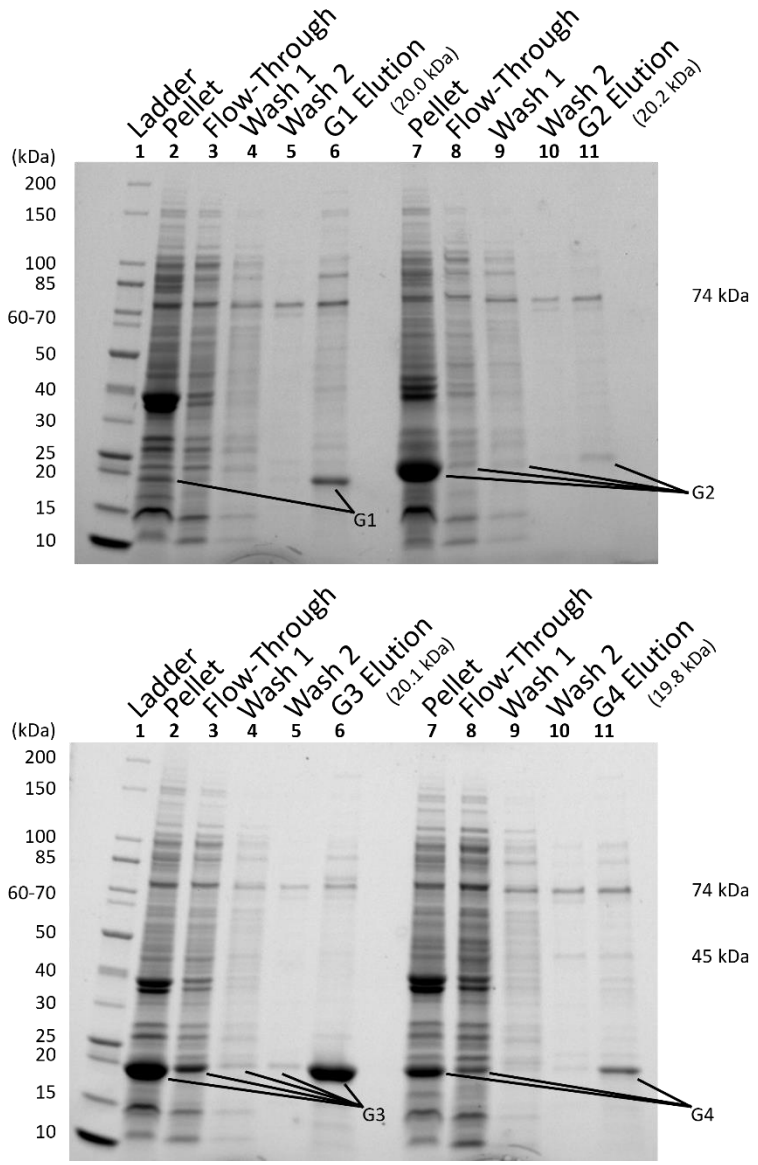
| | |
|---|---|
| G3CRO1k-368 | MHHHHHHWDILVRLKMDKTITGSGEGRDGLIVLTLDSKVLPPLVTTFTYGVQCCADPS<br>TPVHRMYSAALPEGYTQTREIGGIKSTGRVTADPLVNDIDIDAKLEELEKLAPLFVDE<br>KGNYYPHLVHVANGKVTFEINGVKVEQRVTGPRLLPLEATIRTSLTLDDVLTERIEAA |
| G3CRO1k-385 | MHHHHHHWKIEVNLKMDKTITGSGEGKDGKITLTLDSEVVPPLVTTFTYGVQCFAKPD<br>TPVARMYRAALPEGYEQTRTIFGIEAKKRVTADPVRNDIDIDAKMEELLEKAPLLVDD<br>DGSYYPHLVHVANGEVTFTINGVDVKQRVTGPRLLPLEDTIFTSLTLGKTLTEEIRAA |
| G3CRO1k-570 | MHHHHHHWDILVRLKMDKTITGSGTGEDGLITLTLDSKVEPGLVTTFTYGVQM<br>FAEPGTPVHKIFTAALPEGYTQTREIGGIKSEGRVTADPVVNDITIDAEMAQILELNP<br>RFVDADGNYYPHLVHVANGEVTFTINGVDVKQTVTGPNLLPLEATIETELTLSDTLTE<br>RIRLA |
| G3CRO1k-698 | MHHHHHHWDILVNLKMDKTITGSGTGKDGLITLTLDSKVIPPLVTTFTYGVQCFTDPS<br>SPVHKIFTSALPEGYEQTREIFGIKSKKKVTADPLTNDIDIDAKLEELLEKAPLLVNE<br>DGSYYDHLVHVSNGKVTFEINGVKVEQTVTGPNLLPLTATIYTSLTLSDTLTERIRLA |
| G3CRO1k-799 | MHHHHHHWPLLVELHMDKTITGSGTGEDGRIVLTLDSEVIPGLVTTFTYGVQCFADPS<br>TPVRRMYTSALPEGYTQKREIGGIKSEKEVTADPVVNRIKIDAKLEEIEKLNPLFVDA<br>DGNYYPHDVHVENGEVTFTINGVDVKQRVTGPRLLPLKDTIHTSLTLSDTLTEVIRAK |
| G3CRO1k-941 | MHHHHHHWKIIVELNMDKTITGEGEGKDGLITLKLDTEVVPPLVTTFTYGVQCFAEPD<br>TPVYKMYTAAMPEGYEQTREIGGIKSKGKVTKDPVVNKIDIDAKLEELLRLAPRFVDA<br>DGRYYPHEVHVKDGEVTFEINGVKVKQKVTGKNLLPLEDTIFTSHTLSDVLTERIERL |
| G3CRO1k-964 | MHHHHHHWKIIVNLKMDKTITGSGEGKDGKIVITLDSEVEPGLVTTFTYGVQCCTDPS<br>TPVHKIYTAALPEGYEQTRTIGGIESKGKVTADPLVNDIEIDAKMEEILKLNPLFVNE<br>DGTYYPHEVHIENGKETFTINGVKVEQKVTGPKLLPLKDTIYTSLTLSDTLTEVIRAK |

Mutations from EGFP are highlighted in bold and underlined with a solid line. Residues designed on *de novo* fragments are underlined with a dotted line. Fragment breaks are denoted by dashes (—). A C70S mutation was carried out for G1,G2,G3,G4, and G1 enhanced designs. G2A78 and all G3CRO1k designs had a tryptophan placed at the C-terminal end of their polyhistidine-tag to aid in purification and quantification.

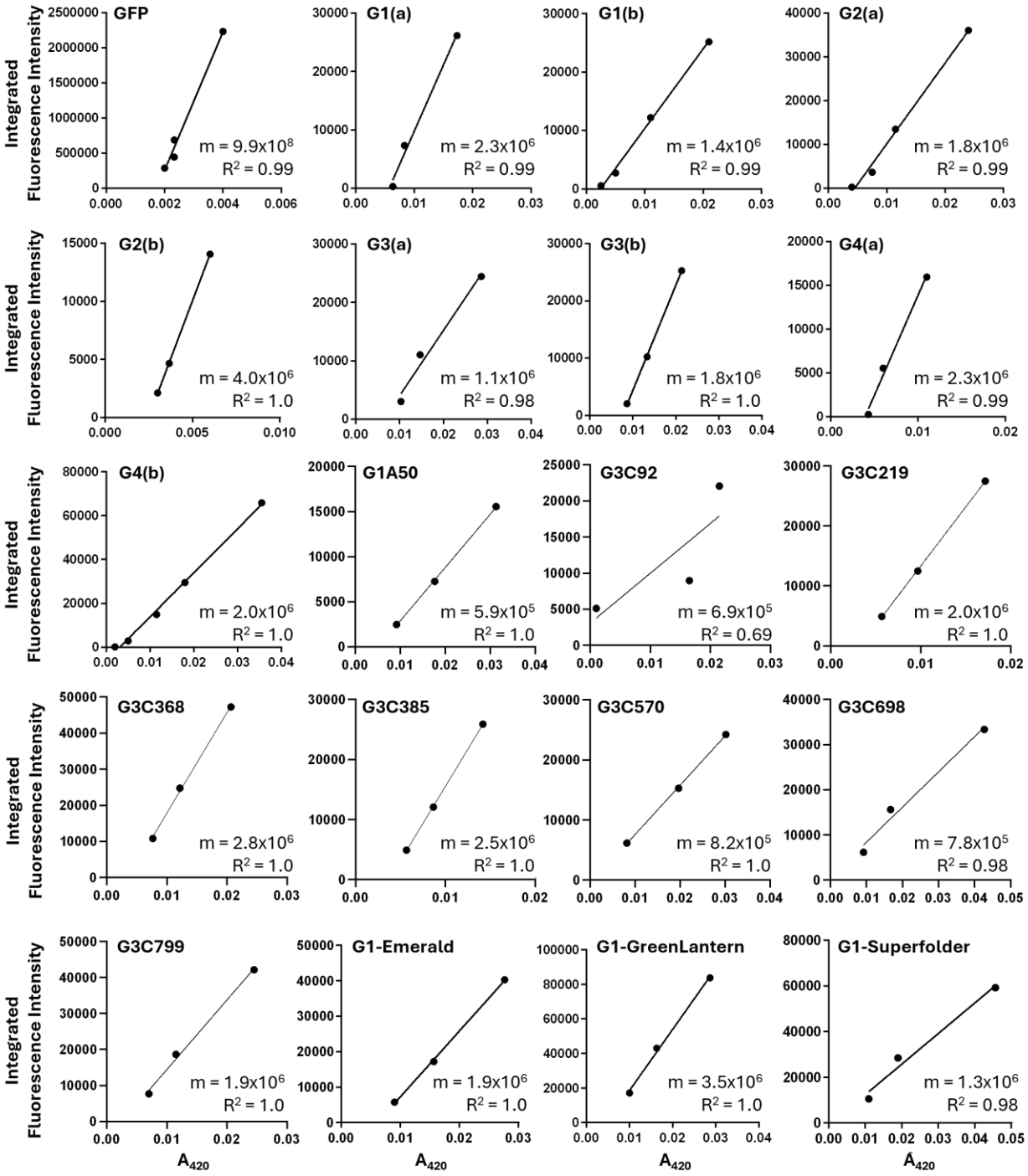
**Supplementary Table 3. Sequence Percent Identity Matrix of First and Second Round lilGFP Designs.** The Percent Identity Matrix was created using Clustal2.1 (EMBL-EBI) after performing a sequence alignment with Clustal Omega1.2.4 (EMBL-EBI).

| Design | G2 | G2A78 | 941 | 92 | 799 | 385 | 964 | 698 | 570 | 368 | 219 | 281 | G1A50 | G1 | G3 | G4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G2 | 100 | 59 | 38 | 36 | 37 | 37 | 36 | 36 | 35 | 37 | 35 | 35 | 39 | 57 | 58 | 58 |
| G2A78 | 59 | 100 | 42 | 40 | 42 | 41 | 43 | 42 | 43 | 40 | 40 | 42 | 44 | 38 | 43 | 44 |
| 941 | 38 | 42 | 100 | 73 | 70 | 73 | 71 | 71 | 72 | 73 | 72 | 71 | 44 | 40 | 52 | 46 |
| 92 | 36 | 40 | 73 | 100 | 78 | 69 | 71 | 72 | 73 | 75 | 73 | 74 | 41 | 41 | 50 | 47 |
| 799 | 37 | 42 | 70 | 78 | 100 | 73 | 76 | 71 | 77 | 77 | 76 | 75 | 38 | 39 | 48 | 45 |
| 385 | 37 | 41 | 73 | 69 | 73 | 100 | 74 | 76 | 73 | 75 | 71 | 71 | 43 | 37 | 49 | 46 |
| 964 | 36 | 43 | 71 | 71 | 76 | 74 | 100 | 77 | 72 | 75 | 75 | 74 | 40 | 38 | 48 | 47 |
| 698 | 36 | 42 | 71 | 72 | 71 | 76 | 77 | 100 | 77 | 78 | 80 | 75 | 41 | 40 | 53 | 50 |
| 570 | 35 | 43 | 72 | 73 | 77 | 73 | 72 | 77 | 100 | 77 | 80 | 76 | 38 | 38 | 46 | 43 |
| 368 | 37 | 40 | 73 | 75 | 77 | 75 | 75 | 78 | 77 | 100 | 82 | 76 | 40 | 37 | 47 | 45 |
| 219 | 35 | 40 | 72 | 73 | 76 | 71 | 75 | 80 | 80 | 82 | 100 | 82 | 37 | 37 | 46 | 44 |
| 281 | 35 | 42 | 71 | 74 | 75 | 71 | 74 | 75 | 76 | 76 | 82 | 100 | 38 | 38 | 48 | 48 |
| G1A50 | 39 | 44 | 44 | 41 | 38 | 43 | 40 | 41 | 38 | 40 | 37 | 38 | 100 | 55 | 46 | 45 |
| G1 | 57 | 38 | 40 | 41 | 39 | 37 | 38 | 40 | 38 | 37 | 37 | 38 | 55 | 100 | 68 | 64 |
| G3 | 58 | 43 | 52 | 50 | 48 | 49 | 48 | 53 | 46 | 47 | 46 | 48 | 46 | 68 | 100 | 77 |
| G4 | 58 | 44 | 46 | 47 | 45 | 46 | 47 | 50 | 43 | 45 | 44 | 48 | 45 | 64 | 77 | 100 |

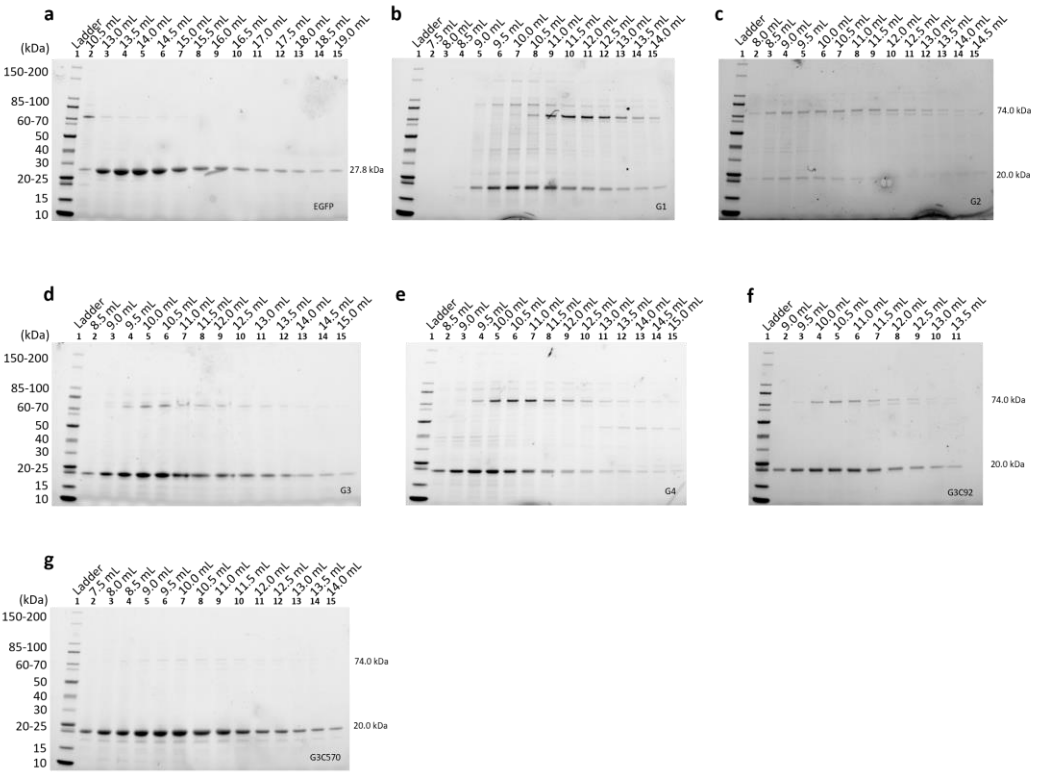
Designs that don't begin with the letter (G) normally have the (G3CRO1K-) prefix.



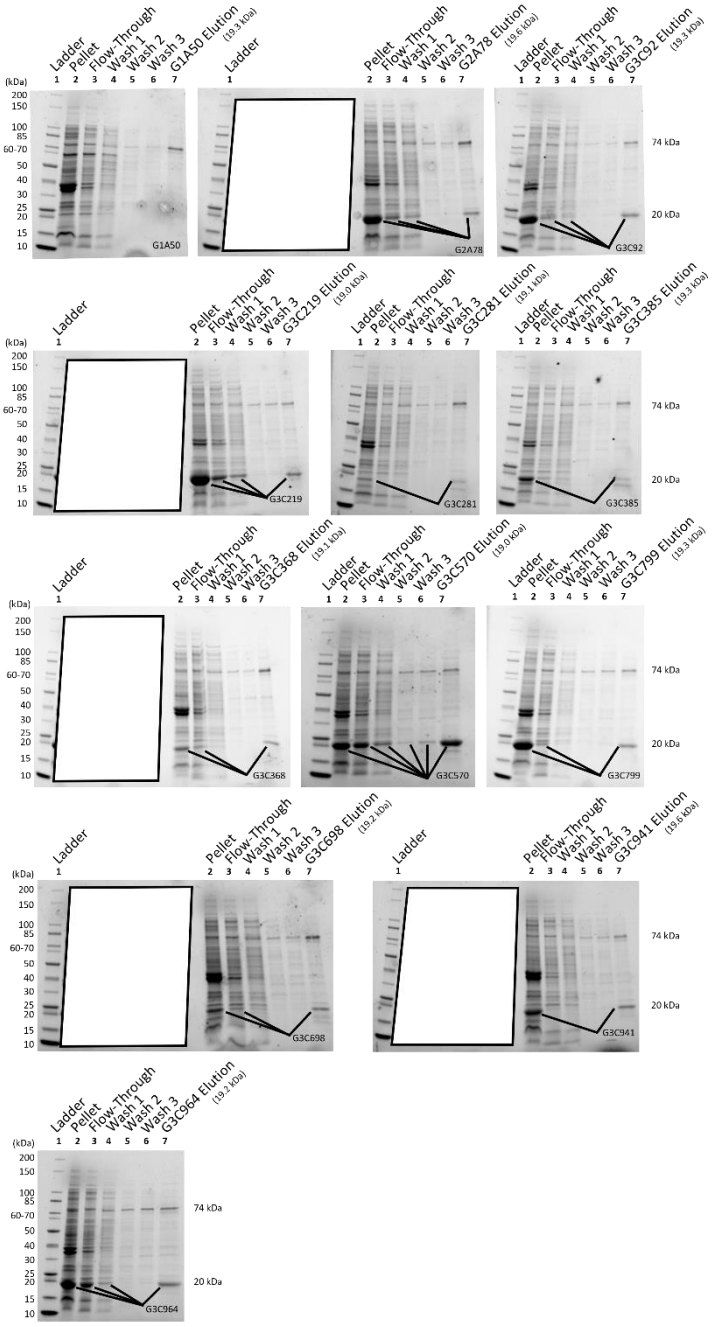

**Supplementary Figure 1. Representative SDS-PAGE of IMAC Purification of G1 – G4.** Purification fractions were loaded on 4–15% Mini-PROTEAN® TGX Stain-Free™ Gels (Bio-Rad). Lanes correspond to: 1) Unstained Protein Standard Broad Range (10-200 kDa) ladder (P7717S, NEB), 2) pellet, 3) Flow-Through, 4) 20 mM imidazole wash, 5), 40 mM imidazole wash and 6) 250 mM imidazole elution. ArnA (74 kDa) is an *E. coli* protein known to bind Ni-NTA with high affinity[131], and is likely seen here.



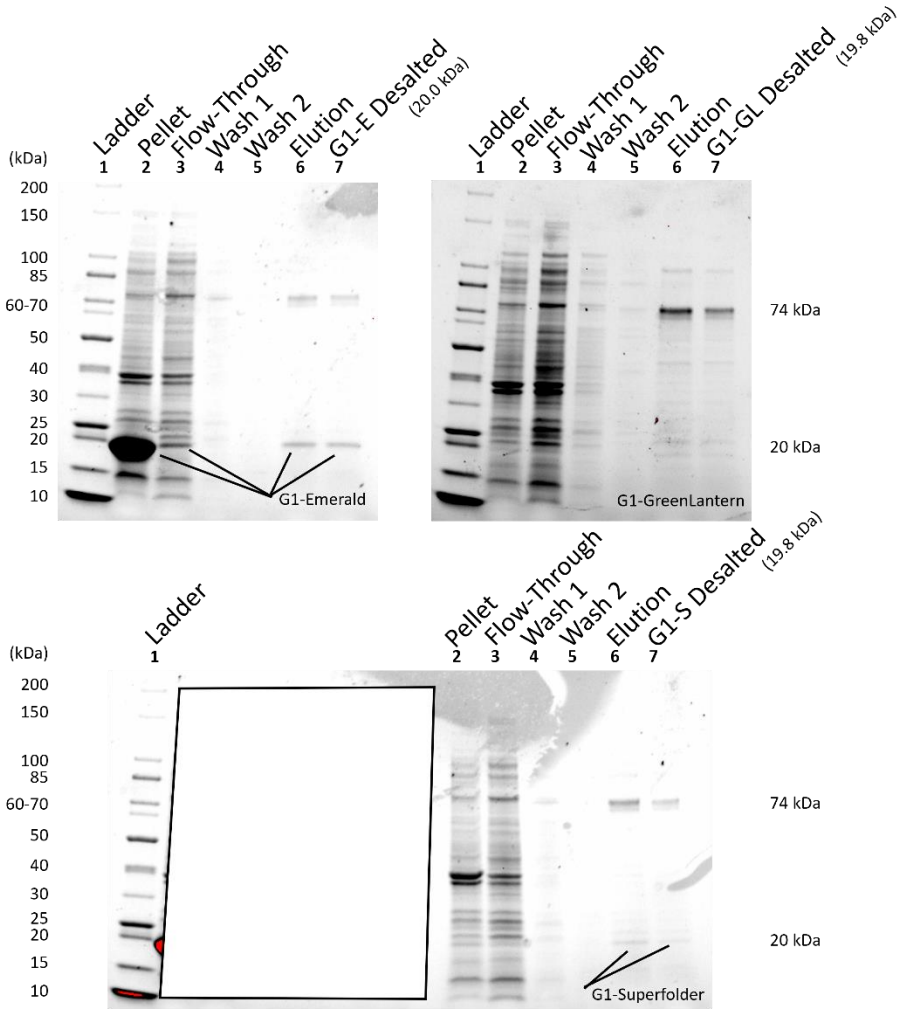

**Supplementary Figure 2. Plots Used to Determine Quantum Yields of Green Fluorescent Proteins.** Integrated Fluorescence Intensity values are plotted against absorbance at the wavelength of chromophore excitation (420 nm). Experimentally obtained quantum yields (QY) can be determined by comparing the resulting slope to that of another fluorescent protein with a known quantum yield. EGFP (QY: 0.6) was used as a reference for calculations.



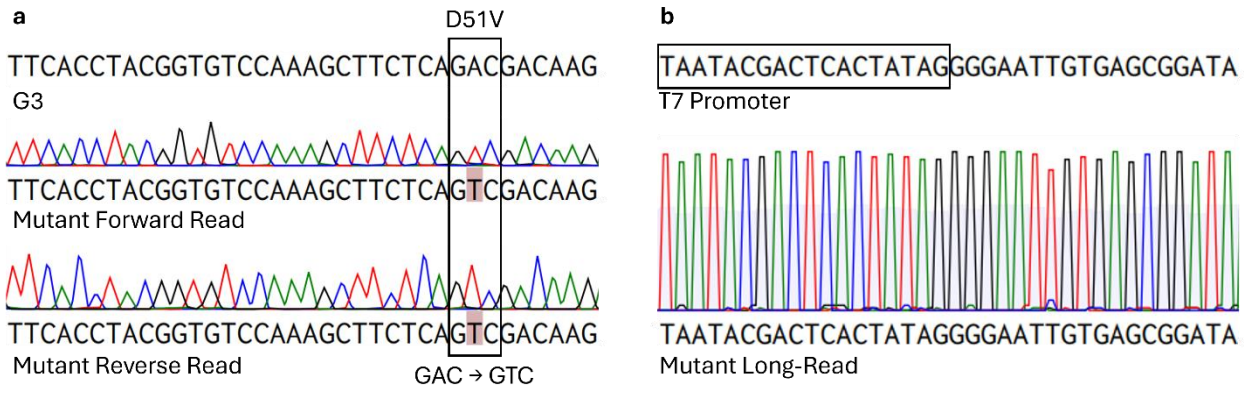

**Supplementary Figure 3. Representative SDS-PAGE Gels Following Size Exclusion Chromatography.** Fractions (0.5 mL) from Size Exclusion Chromatography were loaded on 4–15% Mini-PROTEAN® TGX Stain-Free™ Gels (Bio-Rad). Unstained Protein Standard Broad Range (10-200 kDa) ladder (P7717S, NEB) was used (Lanes 1). Gels correspond to: a) GFP (27.8 kDA), b) G1 (20.0 kDa), c) G2 (20.2 kDa), d) G3 (20.1 kDa), e) G4 (19.8 kDa), f) G3CRO1K-92 (19.3 kDa), and g) G3CRO1K-570 (19.0 kDa). ArnA (74 kDa) is an *E. coli* protein known to bind Ni-NTA with high affinity[131], and is likely seen here.



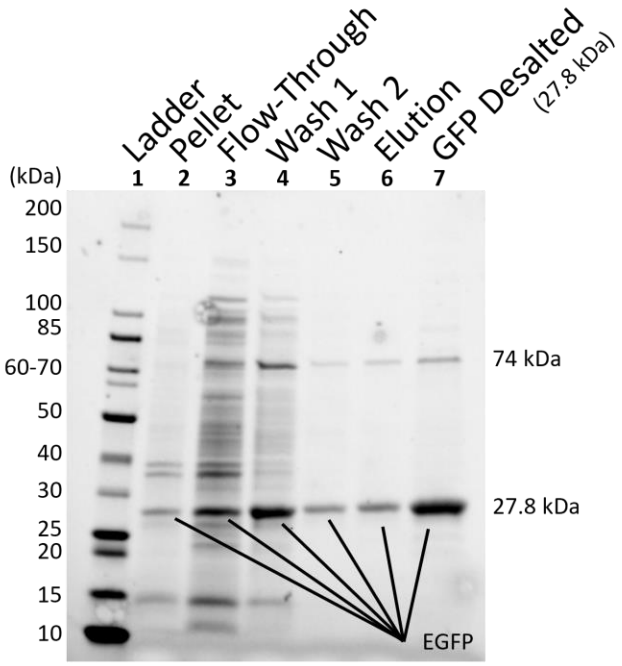

**Supplementary Figure 4. Representative SDS-PAGE of IMAC Purification of Second-Round Designs.** Purification fractions were loaded on 4–15% Mini-PROTEAN® TGX Stain-Free™ Gels (Bio-Rad). Lanes correspond to 1), Unstained Protein Standard Broad Range (10-200 kDa) ladder (P7717S, NEB) 2) pellet, 3) Flow-Through, 4) 20 mM imidazole wash, 5) 40 mM imidazole wash, 6) 60 mM imidazole wash, and 7) 250 mM imidazole elution. ArnA (74 kDa) is an *E. coli* protein known to bind Ni-NTA with high affinity[131], and is likely seen here.



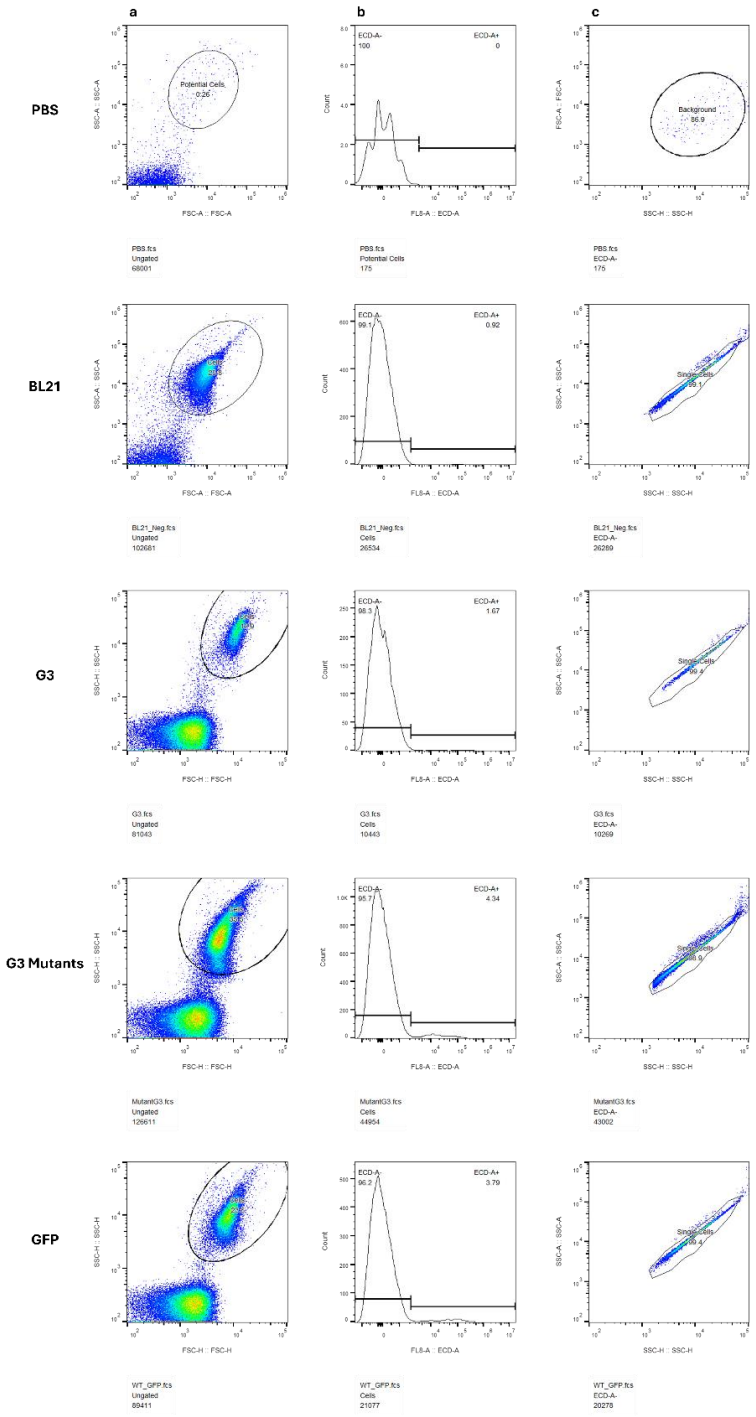

**Supplementary Figure 5. Representative SDS-PAGE of IMAC Purification of G1 – G4 Designs.** Purification fractions were loaded on 4–15% Mini-PROTEAN® TGX Stain-Free™ Gels (Bio-Rad). Lanes correspond to: 1) Unstained Protein Standard Broad Range (10-200 kDa) ladder (P7717S, NEB), 2) pellet, 3) Flow-Through, 4) 20 mM imidazole wash, 5), 40 mM imidazole wash, 6) 250 mM imidazole elution, and 7) desalted fraction. ArnA (74 kDa) is an *E. coli* protein known to bind Ni-NTA with high affinity[131], and is likely seen here.




**Supplementary Figure 6. Representative Sanger and Long-Read Sequencing Results for G3 Mutants.** After each round of mutagenesis, plasmid DNA was sent to Genome Quebec for sequencing. (a) Sanger sequencing was used to verify the presence of mutations using both forward and reverse strands for validation. Plasmid DNA was also sent to Plasmidsaurus for (b) Long-Read sequencing to confirm that mutations were confined to reading frames and not occurring in promoter regions. Chromatograms displaying the relative abundance of nucleotides A (green), T (red), G (black), C (blue) were used in conjunction with sequence alignments to determine the presence of mutations. For sample preparation, isolated colonies were cultured, and their plasmids were extracted and purified following the manufacturer's protocol (Omega Bio-Tek).




**Supplementary Figure 7. Representative SDS-PAGE of IMAC Purification of GFP.** Purification fractions were loaded on 4–15% Mini-PROTEAN® TGX Stain-Free™ Gels (Bio-Rad). Lanes correspond to: 1) Unstained Protein Standard Broad Range (10-200 kDa) ladder (P7717S, NEB), 2) pellet, 3) Flow-Through, 4) 20 mM imidazole wash, 5), 40 mM imidazole wash, 6) 250 mM imidazole elution, and 7) desalted fraction. ArnA (74 kDa) is an *E. coli* protein known to bind Ni-NTA with high affinity[131], and is likely seen here.




**Supplementary Figure 8. Gating Strategy Used for Analysis of Green Fluorescent Proteins Using Flow Cytometry.** Gating strategy used to isolate (a) cells, (b) living cells, and (c) singlet cells. Samples of PBS, BL21-Gold(DE3) cells expressing empty pET-29b(+), G3, G3 mutants (MT1 + MT2), and GFP were analyzed. Samples were concentrated to $10^6$ cells $mL^{-1}$ in sterile PBS and analyzed at a flow rate of 10 µL $min^{-1}$. Analysis was done with a CytoFlex S Cytometer (Beckman Coulter).

## References


(1) Bergendahl, L. T.; Gerasimavicius, L.; Miles, J.; Macdonald, L.; Wells, J. N.; Welburn, J. P. I.; Marsh, J. A. The Role of Protein Complexes in Human Genetic Disease. *Protein Science* **2019**, *28* (8), 1400–1411. https://doi.org/10.1002/pro.3667.

(2) Karamanou, M.; Panayiotakopoulos, G.; Tsoucalas, G.; Kousoulis, A. A.; Androutsos, G. From Miasmas to Germs: A Historical Approach to Theories of Infectious Disease Transmission. *Le Infezioni in Medicina* **2012**, *20* (1), 52–56.

(3) Smith, K. A. Louis Pasteur, the Father of Immunology? *Front. Immun.* **2012**, *3*, 68–68. https://doi.org/10.3389/fimmu.2012.00068.

(4) Kozloff, E. N. Marine Invertebrates of the Pacific Northwest. *Journal of the Marine Biological Association of the United Kingdom* **1997**, *77* (1), 286–286. https://doi.org/10.1017/S0025315400034081.

(5) Tian, F.; Xu, G.; Zhou, S.; Chen, S.; He, D. Principles and Applications of Green Fluorescent Protein-Based Biosensors: A Mini-Review. *Analyst* **2023**, *148* (13), 2882–2891. https://doi.org/10.1039/D3AN00320E.

(6) Hoffman, R. M. Application of GFP Imaging in Cancer. *Laboratory Investigation* **2015**, *95* (4), 432–452. https://doi.org/10.1038/labinvest.2014.154.

(7) Koveal, D.; Rosen, P. C.; Meyer, D. J.; Díaz-García, C. M.; Wang, Y.; Cai, L.; Chou, P. J.; Weitz, D. A.; Yellen, G. A High-Throughput Multiparameter Screen for Accelerated Development and Optimization of Soluble Genetically Encoded Fluorescent Biosensors. *Nat Commun* **2022**, *13* (1), 2919. https://doi.org/10.1038/s41467-022-30685-x.

(8) *Fluorescent Proteins: Methods and Protocols*; Sharma, M., Ed.; Methods in Molecular Biology; Springer US: New York, NY, **2023**; Vol. 2564. https://doi.org/10.1007/978-1-0716-2667-2.

(9) Palmer, A. E.; Qin, Y.; Park, J. G.; McCombs, J. E. Design and Application of Genetically Encoded Biosensors. *Trends in Biotechnology* **2011**, *29* (3), 144–152. https://doi.org/10.1016/j.tibtech.2010.12.004.

(10) Soboleski, M. R.; Oaks, J.; Halford, W. P. Green Fluorescent Protein Is a Quantitative Reporter of Gene Expression in Individual Eukaryotic Cells. *FASEB j.* **2005**, *19* (3), 1–20. https://doi.org/10.1096/fj.04-3180fje.

(11) Misteli, T.; Spector, D. L. Applications of the Green Fluorescent Protein in Cell Biology and Biotechnology. *Nat Biotechnol* **1997**, *15* (10), 961–964. https://doi.org/10.1038/nbt1097-961.

(12) Eason, M. G.; Pandelieva, A. T.; Mayer, M. M.; Khan, S. T.; Garcia, H. G.; Chica, R. A. Genetically Encoded Fluorescent Biosensor for Rapid Detection of Protein Expression. *ACS Synth. Biol.* **2020**, *9* (11), 2955–2963. https://doi.org/10.1021/acssynbio.0c00407.

(13) Remington, S. J. Green Fluorescent Protein: A Perspective. *Protein Science* **2011**, *20* (9), 1509–1519. https://doi.org/10.1002/pro.684.

(14) Ramesh, B.; Frei, C. S.; Cirino, P. C.; Varadarajan, N. Functional Enrichment by Direct Plasmid Recovery After Fluorescence Activated Cell Sorting. *BioTechniques* **2015**, *59* (3), 157–161. https://doi.org/10.2144/000114329.

(15) Kovačević, G.; Ostafe, R.; Balaž, A. M.; Fischer, R.; Prodanović, R. Development of GFP-Based High-Throughput Screening System for Directed Evolution of Glucose Oxidase. *Journal of Bioscience and Bioengineering* **2019**, *127* (1), 30–37. https://doi.org/10.1016/j.jbiosc.2018.07.002.
(16) Prasher, D. C.; Eckenrode, V. K.; Ward, W. W.; Prendergast, F. G.; Cormier, M. J. Primary Structure of the Aequorea Victoria Green-Fluorescent Protein. *Gene* **1992**, *111* (2), 229–233. https://doi.org/10.1016/0378-1119(92)90691-H.
(17) Chalfie, M.; Yuan, T.; Euskirchen, G.; Ward, W. W.; Prasher, D. C. Green Fluorescent Protein as a Marker for Gene Expression. *Science* **1994**, *263* (5148), 802–805. https://doi.org/10.1126/science.8303295.
(18) Heim, R.; Prasher, D. C.; Tsien, R. Y. Wavelength Mutations and Posttranslational Autoxidation of Green Fluorescent Protein. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91* (26), 12501–12504. https://doi.org/10.1073/pnas.91.26.12501.
(19) Wachter, R. M.; King, B. A.; Heim, R.; Kallio, K.; Tsien, R. Y.; Boxer, S. G.; Remington, S. J. Crystal Structure and Photodynamic Behavior of the Blue Emission Variant Y66H/Y145F of Green Fluorescent Protein. *Biochemistry* **1997**, *36* (32), 9759–9765. https://doi.org/10.1021/bi970563w.
(20) Heim, R.; Cubitt, A.; Tsien, R. Improved Green Fluorescence. *Nature* **1995**, *373* (6516), 663–663. https://doi.org/10.1038/373663a0.
(21) Zylka, M. J.; Schnapp, B. J. Optimized Filter Set and Viewing Conditions for the S65T Mutant of GFP in Living Cells. *BioTechniques* **1996**, *21* (2), 220–226. https://doi.org/10.2144/96212bm11.
(22) Heim, R.; Tsien, R. Y. Engineering Green Fluorescent Protein for Improved Brightness, Longer Wavelengths and Fluorescence Resonance Energy Transfer. *Current Biology* **1996**, *6* (2), 178–182. https://doi.org/10.1016/S0960-9822(02)00450-5.
(23) Zacharias, D. A.; Violin, J. D.; Newton, A. C.; Tsien, R. Y. Partitioning of Lipid-Modified Monomeric GFPs into Membrane Microdomains of Live Cells. *Science* **2002**, *296* (5569), 913–916. https://doi.org/10.1126/science.1068539.
(24) Valbuena, F.; Fizgerald, I.; Strack, R. L.; Andruska, N.; Smith, L.; Glick, B. S. A Photostable Monomeric Superfolder GFP. *Traffic* **2020**, *21* (8), 534–544. https://doi.org/10.1111/tra.12737.
(25) Shinoda, H.; Ma, Y.; Nakashima, R.; Sakurai, K.; Matsuda, T.; Nagai, T. Acid-Tolerant Monomeric GFP from Olindias Formosa. *Cell Chemical Biology* **2018**, *25* (3), 330–338. https://doi.org/10.1016/j.chembiol.2017.12.005.
(26) Ivorra-Molla, E.; Akhuli, D.; McAndrew, M. B. L.; Scott, W.; Kumar, L.; Palani, S.; Mishima, M.; Crow, A.; Balasubramanian, M. K. A Monomeric StayGold Fluorescent Protein. *Nat Biotechnol* **2024**, *42* (9), 1368–1371. https://doi.org/10.1038/s41587-023-02018-w.
(27) Zhang, H.; Lesnov, G. D.; Subach, O. M.; Zhang, W.; Kuzmicheva, T. P.; Vlaskina, A. V.; Samygina, V. R.; Chen, L.; Ye, X.; Nikolaeva, A. Yu.; Gabdulkhakov, A.; Papadaki, S.; Qin, W.; Borshchevskiy, V.; Perfilov, M. M.; Gavrikov, A. S.; Drobizhev, M.; Mishin, A. S.; Piatkevich, K. D.; Subach, F. V. Bright and Stable Monomeric Green Fluorescent Protein Derived from StayGold. *Nat Methods* **2024**, *21* (4), 657–665. https://doi.org/10.1038/s41592-024-02203-y.

(28) Close, D. W.; Paul, C. D.; Langan, P. S.; Wilce, M. C. J.; Traore, D. A. K.; Halfmann, R.; Rocha, R. C.; Waldo, G. S.; Payne, R. J.; Rucker, J. B.; Prescott, M.; Bradbury, A. R. M. Thermal Green Protein, an Extremely Stable, Nonaggregating Fluorescent Protein Created by Structure-guided Surface Engineering. *Proteins* **2015**, *83* (7), 1225–1237. https://doi.org/10.1002/prot.24699.
(29) Balleza, E.; Kim, J. M.; Cluzel, P. Systematic Characterization of Maturation Time of Fluorescent Proteins in Living Cells. *Nat Methods* **2018**, *15* (1), 47–51. https://doi.org/10.1038/nmeth.4509.
(30) Dáder, B.; Burckbuchler, M.; Macia, J.-L.; Alcon, C.; Curie, C.; Gargani, D.; Zhou, J. S.; Ng, J. C. K.; Brault, V.; Drucker, M. Split Green Fluorescent Protein as a Tool to Study Infection with a Plant Pathogen, Cauliflower Mosaic Virus. *PLoS ONE* **2019**, *14* (3), e0213087–e0213087. https://doi.org/10.1371/journal.pone.0213087.
(31) Paulk, A. M.; Williams, R. L.; Liu, C. C. Rapidly Inducible Yeast Surface Display for Antibody Evolution with OrthoRep. *ACS Synth. Biol.* **2024**, *13* (8), 2629–2634. https://doi.org/10.1021/acssynbio.4c00370.
(32) Cao, P.; Shi, H.; Zhang, S.; Chen, J.; Wang, R.; Liu, P.; Zhu, Y.; An, Y.; Zhang, M. A Robust High-throughput Functional Screening Assay for Plant Pathogen Effectors Using the TMV-GFP Vector. *The Plant Journal* **2024**, *119* (1), 617–631. https://doi.org/10.1111/tpj.16774.
(33) Pham, T. D.; Poletti, C.; Tientcheu, T. M. N.; Cuccioloni, M.; Spurio, R.; Fabbretti, A.; Milon, P.; Giuliodori, A. M. FAST, a Method Based on Split-GFP for the Detection in Solution of Proteins Synthesized in Cell-Free Expression Systems. *Sci Rep* **2024**, *14* (1), 8042–8042. https://doi.org/10.1038/s41598-024-58588-5.
(34) Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. Crystal Structure of the Aequorea Victoria Green Fluorescent Protein. *Science* **1996**, *273* (5280), 1392–1395. https://doi.org/10.1126/science.273.5280.1392.
(35) Myatt, D. P.; Hatter, L.; Rogers, S. E.; Terry, A. E.; Clifton, L. A. Monomeric Green Fluorescent Protein as a Protein Standard for Small Angle Scattering. *BSI* **2017**, *6* (3–4), 123–134. https://doi.org/10.3233/BSI-170167.
(36) Tsien, R. Y. The Green Fluorescent Protein. *Annu. Rev. Biochem.* **1998**, *67* (1), 509–544. https://doi.org/10.1146/annurev.biochem.67.1.509.
(37) Cubitt, A. B.; Heim, R.; Boyd, A. E.; Adams, S. R.; Gross, L. A.; Tsien, R. Y. Understanding, Improving and Using Green Fluorescent Proteins. *Trends in Biochemical Sciences* **1995**, *20* (11), 448–455. https://doi.org/10.1016/s0968-0004(00)89099-4.
(38) Wood, T. I.; Barondeau, D. P.; Hitomi, C.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Defining the Role of Arginine 96 in Green Fluorescent Protein Fluorophore Biosynthesis. *Biochemistry* **2005**, *44* (49), 16211–16220. https://doi.org/10.1021/bi051388j.
(39) Kremers, G.-J.; Gilbert, S. G.; Cranfill, P. J.; Davidson, M. W.; Piston, D. W. Fluorescent Proteins at a Glance. *Journal of Cell Science* **2011**, *124* (15), 2676–2676. https://doi.org/10.1242/jcs.095059.

(40) Phillips, G. N. Structure and Dynamics of Green Fluorescent Protein. *Current Opinion in Structural Biology* **1997**, *7* (6), 821–827. https://doi.org/10.1016/S0959-440X(97)80153-4.
(41) Jach, G.; Binot, E.; Frings, S.; Luxa, K.; Schell, J. Use of Red Fluorescent Protein from Discosoma Sp. (dsRED) as a Reporter for Plant Gene Expression. *The Plant Journal* **2001**, *28* (4), 483–491. https://doi.org/10.1046/j.1365-313X.2001.01153.x.
(42) Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent Proteins from Nonbioluminescent Anthozoa Species. *Nat Biotechnol* **1999**, *17* (10), 969–973. https://doi.org/10.1038/13657.
(43) Wall, M. A.; Socolich, M.; Ranganathan, R. The Structural Basis for Red Fluorescence in the Tetrameric GFP Homolog DsRed. *Nature Structural Biology* **2000**, *7* (12), 1133–1138. https://doi.org/10.1038/81992.
(44) Salih, A.; Larkum, A.; Cox, G.; Kühl, M.; Hoegh-Guldberg, O. Fluorescent Pigments in Corals Are Photoprotective. *Nature* **2000**, *408* (6814), 850–853. https://doi.org/10.1038/35048564.
(45) Baubet, V.; Le Mouellic, H.; Campbell, A. K.; Lucas-Meunier, E.; Fossier, P.; Brûlet, P. Chimeric Green Fluorescent Protein-Aequorin as Bioluminescent $Ca^{2+}$ Reporters at the Single-Cell Level. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (13), 7260–7265. https://doi.org/10.1073/pnas.97.13.7260.
(46) Teranishi, K. Luminescence of Imidazo[1,2-a]Pyrazin-3(7H)-One Compounds. *Bioorganic Chemistry* **2007**, *35* (1), 82–111. https://doi.org/10.1016/j.bioorg.2006.08.003.
(47) Yang, F.; Moss, L. G.; Phillips, G. N. The Molecular Structure of Green Fluorescent Protein. *Nature Biotechnology* **1996**, *14* (10), 1246–1251. https://doi.org/10.1038/nbt1096-1246.
(48) Costantini, L. M.; Fossati, M.; Francolini, M.; Snapp, E. L. Assessing the Tendency of Fluorescent Proteins to Oligomerize Under Physiologic Conditions. *Traffic* **2012**, *13* (5), 643–649. https://doi.org/10.1111/j.1600-0854.2012.01336.x.
(49) Yarbrough, D.; Wachter, R. M.; Kallio, K.; Matz, M. V.; Remington, S. J. Refined Crystal Structure of DsRed, a Red Fluorescent Protein from Coral, at 2.0-Å Resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (2), 462–467. https://doi.org/10.1073/pnas.98.2.462.
(50) Ward, W. W.; Prentice, H. J.; Roth, A. F.; Cody, C. W.; Reeves, S. C. Spectral Perturbations of the Aequorea Green FLuorescent Protein. *Photochem & Photobiology* **1982**, *35* (6), 803–808. https://doi.org/10.1111/j.1751-1097.1982.tb02651.x.
(51) Shinoda, H.; Lu, K.; Nakashima, R.; Wazawa, T.; Noguchi, K.; Matsuda, T.; Nagai, T. Acid-Tolerant Reversibly Switchable Green Fluorescent Protein for Super-Resolution Imaging under Acidic Conditions. *Cell Chemical Biology* **2019**, *26* (10), 1469–1479. https://doi.org/10.1016/j.chembiol.2019.07.012.
(52) Costantini, L. M.; Snapp, E. L. Fluorescent Proteins in Cellular Organelles: Serious Pitfalls and Some Solutions. *DNA and Cell Biology* **2013**, *32* (11), 622–627. https://doi.org/10.1089/dna.2013.2172.
(53) Kremers, G.-J.; Gilbert, S. G.; Cranfill, P. J.; Davidson, M. W.; Piston, D. W. Fluorescent Proteins at a Glance. *Journal of Cell Science* **2011**, *124* (15), 2676–2676. https://doi.org/10.1242/jcs.095059.

(54) Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. Circular Permutation and Receptor Insertion within Green Fluorescent Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (20), 11241–11246. https://doi.org/10.1073/pnas.96.20.11241.
(55) Li, X.; Zhang, G.; Ngo, N.; Zhao, X.; Kain, S. R.; Huang, C.-C. Deletions of the Aequorea Victoria Green Fluorescent Protein Define the Minimal Domain Required for Fluorescence. *Journal of Biological Chemistry* **1997**, *272* (45), 28545–28549. https://doi.org/10.1074/jbc.272.45.28545.
(56) Dopf, J.; Horiagon, T. M. Deletion Mapping of the Aequorea Victoria Green Fluorescent Protein. *Gene* **1996**, *173* (1), 39–44. https://doi.org/10.1016/0378-1119(95)00692-3.
(57) Craggs, T. D. Green Fluorescent Protein: Structure, Folding and Chromophore Maturation. *Chem. Soc. Rev.* **2009**, *38* (10), 2865–2875. https://doi.org/10.1039/b903641p.
(58) Heilemann, M. Light at the End of the Tunnel. *Angew Chem Int Ed* **2009**, *48* (22), 3908–3910. https://doi.org/10.1002/anie.200900696.
(59) Fu, J. L.; Kanno, T.; Liang, S.-C.; Matzke, A. J. M.; Matzke, M. GFP Loss-of-Function Mutations in *Arabidopsis Thaliana*. *G3 Genes|Genomes|Genetics* **2015**, *5* (9), 1849–1855. https://doi.org/10.1534/g3.115.019604.
(60) Barondeau, D. P.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Understanding GFP Chromophore Biosynthesis: Controlling Backbone Cyclization and Modifying Post-Translational Chemistry †. *Biochemistry* **2005**, *44* (6), 1960–1970. https://doi.org/10.1021/bi0479205.
(61) Rosenow, M. A.; Huffman, H. A.; Phail, M. E.; Wachter, R. M. The Crystal Structure of the Y66L Variant of Green Fluorescent Protein Supports a Cyclization−Oxidation−Dehydration Mechanism for Chromophore Maturation. *Biochemistry* **2004**, *43* (15), 4464–4472. https://doi.org/10.1021/bi0361315.
(62) Barondeau, D. P.; Putnam, C. D.; Kassmann, C. J.; Tainer, J. A.; Getzoff, E. D. Mechanism and Energetics of Green Fluorescent Protein Chromophore Synthesis Revealed by Trapped Intermediate Structures. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (21), 12111–12116. https://doi.org/10.1073/pnas.2133463100.
(63) Ma, Y.; Zhang, H.; Sun, Q.; Smith, S. C. New Insights on the Mechanism of Cyclization in Chromophore Maturation of Wild-Type Green Fluorescence Protein: A Computational Study. *J. Phys. Chem. B* **2016**, *120* (24), 5386–5394. https://doi.org/10.1021/acs.jpcb.6b04406.
(64) Bartkiewicz, M.; Kazazić, S.; Krasowska, J.; Clark, P. L.; Wielgus-Kutrowska, B.; Bzowska, A. Non–Fluorescent Mutant of Green Fluorescent Protein Sheds Light on the Mechanism of Chromophore Formation. *FEBS Letters* **2018**, *592* (9), 1516–1523. https://doi.org/10.1002/1873-3468.13051.
(65) Pletneva, N. V.; Pletnev, V. Z.; Lukyanov, K. A.; Gurskaya, N. G.; Goryacheva, E. A.; Martynov, V. I.; Wlodawer, A.; Dauter, Z.; Pletnev, S. Structural Evidence for a Dehydrated Intermediate in Green Fluorescent Protein Chromophore Biosynthesis. *Journal of Biological Chemistry* **2010**, *285* (21), 15978–15984. https://doi.org/10.1074/jbc.M109.092320.

(66) Ma, Y.; Sun, Q.; Smith, S. C. The Mechanism of Oxidation in Chromophore Maturation of Wild-Type Green Fluorescent Protein: A Theoretical Study. *Phys. Chem. Chem. Phys.* **2017**, *19* (20), 12942–12952. https://doi.org/10.1039/C6CP07983K.
(67) Sniegowski, J. A.; Lappe, J. W.; Patel, H. N.; Huffman, H. A.; Wachter, R. M. Base Catalysis of Chromophore Formation in Arg96 and Glu222 Variants of Green Fluorescent Protein. *Journal of Biological Chemistry* **2005**, *280* (28), 26248–26255. https://doi.org/10.1074/jbc.M412327200.
(68) Lemay, N. P.; Morgan, A. L.; Archer, E. J.; Dickson, L. A.; Megley, C. M.; Zimmer, M. The Role of the Tight-Turn, Broken Hydrogen Bonding, Glu222 and Arg96 in the Post-Translational Green Fluorescent Protein Chromophore Formation. *Chemical Physics* **2008**, *348* (1–3), 152–160. https://doi.org/10.1016/j.chemphys.2008.02.055.
(69) Li, B.; Shahid, R.; Peshkepija, P.; Zimmer, M. Water Diffusion in and out of the β-Barrel of GFP and the Fast Maturing Fluorescent Protein, TurboGFP. *Chemical Physics* **2012**, *392* (1), 143–148. https://doi.org/10.1016/j.chemphys.2011.11.001.
(70) Ganini, D.; Leinisch, F.; Kumar, A.; Jiang, J.; Tokar, E. J.; Malone, C. C.; Petrovich, R. M.; Mason, R. P. Fluorescent Proteins Such as eGFP Lead to Catalytic Oxidative Stress in Cells. *Redox Biology* **2017**, *12*, 462–468. https://doi.org/10.1016/j.redox.2017.03.002.
(71) Kohata, S.; Nakanotani, H.; Hosokai, T.; Yasuda, T.; Tsuchiya, Y.; Adachi, C. Anti-Stokes Emission Utilizing Reverse Intersystem Crossing. *Angew Chem Int Ed* **2025**, *64* (7), e202419323. https://doi.org/10.1002/anie.202419323.
(72) *Reviews in Fluorescence 2016*; Geddes, C. D., Ed.; Reviews in Fluorescence; Springer International Publishing: Cham, Switzerland, **2017**. https://doi.org/10.1007/978-3-319-48260-6.
(73) Swanson, M. S. *A Concise Introduction to Quantum Mechanics:*; Morgan & Claypool Publishers: Temple Circus, Temple Way, Bristol, UK, **2018**. https://doi.org/10.1088/978-1-6817-4716-3.
(74) *FRET - Förster Resonance Energy Transfer: From Theory to Applications*, 1st ed.; Medintz, I., Hildebrandt, N., Eds.; Wiley-VCH Verlag GmbH: Weinheim, Germany, **2014**. https://doi.org/10.1002/9783527656028.
(75) Shrestha, D.; Jenei, A.; Nagy, P.; Vereb, G.; Szöllősi, J. Understanding FRET as a Research Tool for Cellular Studies. *IJMS* **2015**, *16* (4), 6718–6756. https://doi.org/10.3390/ijms16046718.
(76) Baffour-Awuah, N. Y. A.; Zimmer, M. Hula-Twisting in Green Fluorescent Protein. *Chemical Physics* **2004**, *303* (1–2), 7–11. https://doi.org/10.1016/j.chemphys.2004.04.022.
(77) Drobizhev, M.; Molina, R. S.; Callis, P. R.; Scott, J. N.; Lambert, G. G.; Salih, A.; Shaner, N. C.; Hughes, T. E. Local Electric Field Controls Fluorescence Quantum Yield of Red and Far-Red Fluorescent Proteins. *Front. Mol. Biosci.* **2021**, *8*, 633217–633217. https://doi.org/10.3389/fmolb.2021.633217.
(78) Gowri, M. R.; Ramanathan, G. Planarity Is One of the Essential Requirements for Fluorescence in Red Fluorescent Protein Chromophore Analogs. *Journal of Molecular Structure* **2023**, *1274*, 134481–134481. https://doi.org/10.1016/j.molstruc.2022.134481.

(79) Wenzel, T. Molecular and Atomic Spectroscopy, **2025**. https://chem.libretexts.org/@go/page/111548?pdf (accessed 2025-02-07).
(80) Mayerhöfer, T. G.; Mutschke, H.; Popp, J. Employing Theories Far beyond Their Limits—The Case of the (Boguer-) Beer–Lambert Law. *ChemPhysChem* **2016**, *17* (13), 1948–1955. https://doi.org/10.1002/cphc.201600114.
(81) Bulcha, J. T.; Wang, Y.; Ma, H.; Tai, P. W. L.; Gao, G. Viral Vector Platforms within the Gene Therapy Landscape. *Sig Transduct Target Ther* **2021**, *6* (1), 53–76. https://doi.org/10.1038/s41392-021-00487-6.
(82) Costantini, L. M.; Snapp, E. L. Going Viral with Fluorescent Proteins. *J Virol* **2015**, *89* (19), 9706–9708. https://doi.org/10.1128/JVI.03489-13.
(83) Romei, M. G.; Boxer, S. G. Split Green Fluorescent Proteins: Scope, Limitations, and Outlook. *Annu. Rev. Biophys.* **2019**, *48* (1), 19–44. https://doi.org/10.1146/annurev-biophys-051013-022846.
(84) Liang, G.-T.; Lai, C.; Yue, Z.; Zhang, H.; Li, D.; Chen, Z.; Lu, X.; Tao, L.; Subach, F. V.; Piatkevich, K. D. Enhanced Small Green Fluorescent Proteins as a Multisensing Platform for Biosensor Development. *Front. Bioeng. Biotechnol.* **2022**, *10*, 1039317–1039317. https://doi.org/10.3389/fbioe.2022.1039317.
(85) Devkota, K.; Shonai, D.; Mao, J.; Soderling, S.; Singh, R. Miniaturizing, Modifying, and Augmenting Nature's Proteins with Raygun. bioRxiv August 16, **2024**. https://doi.org/10.1101/2024.08.13.607858.
(86) Notin, P.; Rollins, N.; Gal, Y.; Sander, C.; Marks, D. Machine Learning for Functional Protein Design. *Nat Biotechnol* **2024**, *42* (2), 216–228. https://doi.org/10.1038/s41587-024-02127-0.
(87) Wang, C.; Alamdari, S.; Domingo-Enrich, C.; Amini, A. P.; Yang, K. K. Toward Deep Learning Sequence–Structure Co-Generation for Protein Design. *Current Opinion in Structural Biology* **2025**, *91*, 103018. https://doi.org/10.1016/j.sbi.2025.103018.
(88) Romero-Romero, S.; Kordes, S.; Michel, F.; Höcker, B. Evolution, Folding, and Design of TIM Barrels and Related Proteins. *Current Opinion in Structural Biology* **2021**, *68*, 94–104. https://doi.org/10.1016/j.sbi.2020.12.007.
(89) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.
(90) Chu, A. E.; Kim, J.; Cheng, L.; El Nesr, G.; Xu, M.; Shuai, R. W.; Huang, P.-S. An All-Atom Protein Generative Model. *Proc. Natl. Acad. Sci. U.S.A.* **2024**, *121* (27), e2311500121. https://doi.org/10.1073/pnas.2311500121.
(91) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust Deep Learning–Based Protein Sequence Design Using

ProteinMPNN. *Science* **2022**, *378* (6615), 49–56. https://doi.org/10.1126/science.add2187.
(92) Dauparas, J.; Lee, G. R.; Pecoraro, R.; An, L.; Anishchenko, I.; Glasscock, C.; Baker, D. Atomic Context-Conditioned Protein Sequence Design Using LigandMPNN. bioRxiv December 23, **2023**. https://doi.org/10.1101/2023.12.22.573103.
(93) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.
(94) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.
(95) Lauko, A.; Pellock, S. J.; Sumida, K. H.; Anishchenko, I.; Juergens, D.; Ahern, W.; Jeung, J.; Shida, A.; Hunt, A.; Kalvet, I.; Norn, C.; Humphreys, I. R.; Jamieson, C.; Krishna, R.; Kipnis, Y.; Kang, A.; Brackenbrough, E.; Bera, A. K.; Sankaran, B.; Houk, K. N.; Baker, D. Computational Design of Serine Hydrolases. *Science* **2025**, eadu2454. https://doi.org/10.1126/science.adu2454.
(96) Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y. A.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; Rives, A. Simulating 500 Million Years of Evolution with a Language Model.
(97) Huguet, G.; Vuckovic, J.; Fatras, K.; Thibodeau-Laufer, E.; Lemos, P.; Islam, R.; Liu, C.-H.; Rector-Brooks, J.; Akhound-Sadegh, T.; Bronstein, M.; Tong, A.; Bose, A. J. Sequence-Augmented SE(3)-Flow Matching For Conditional Protein Backbone Generation. arXiv December 11, **2024**. https://doi.org/10.48550/arXiv.2405.20313.
(98) Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Kalvet, I.; Lee, G. R.; Morey-Burrows, F. S.; Humphreys, I. R.; McHugh, R.; Vafeados, D.; Li, X.; Sutherland, A.; Hitchcock, A.; Hunter, C. N.; Baek, M.; DiMaio, F. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. bioRxiv September 10, **2023**. https://doi.org/10.1101/2023.10.09.561603.
(99) Ahern, W.; Yim, J.; Tischer, D.; Salike, S.; Woodbury, S. M.; Kim, D.; Kalvet, I.; Kipnis, Y.; Coventry, B.; Altae-Tran, H. R.; Bauer, M.; Barzilay, R.; Jaakkola, T. S.; Krishna, R.;

Baker, D. Atom Level Enzyme Active Site Scaffolding Using RFdiffusion2. bioRxiv October 4, **2025**. https://doi.org/10.1101/2025.04.09.648075.
(100) Braun, M.; Tripp, A.; Chakatok, M.; Kaltenbrunner, S.; Totaro, M. G.; Stoll, D.; Bijelic, A.; Elaily, W.; Hoch, S. Y. Y.; Aleotti, M.; Hall, M.; Oberdorfer, G. Computational Design of Highly Active de Novo Enzymes. bioRxiv August 3, **2024**. https://doi.org/10.1101/2024.08.02.606416.
(101) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat Biotechnol* **2023**, *41* (8), 1099–1106. https://doi.org/10.1038/s41587-022-01618-2.
(102) Mignon, D.; Druart, K.; Michael, E.; Opuu, V.; Polydorides, S.; Villa, F.; Gaillard, T.; Panel, N.; Archontis, G.; Simonson, T. Physics-Based Computational Protein Design: An Update. *J. Phys. Chem. A* **2020**, *124* (51), 10637–10648. https://doi.org/10.1021/acs.jpca.0c07605.
(103) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. **2008**.
(104) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res* **2020**, *49*, D480–D489.
(105) Mitchell, A. L.; Almeida, A.; Beracochea, M.; Bolan, M.; Burgand, J.; Cochrane, G.; Crusoe, M. R. MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Research* **2020**, *48*, D570-78. https://doi.org/10.1093/nar/gkz1035.
(106) Elofsson, A. Progress at Protein Structure Prediction, as Seen in CASP15. *Current Opinion in Structural Biology* **2023**, *80*, 102594–102594. https://doi.org/10.1016/j.sbi.2023.102594.
(107) Qiu, X.; Li, H.; Ver Steeg, G.; Godzik, A. Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development. *Biomolecules* **2024**, *14* (3), 339. https://doi.org/10.3390/biom14030339.
(108) Stepanenko, O.; Verkhusha, V.; Kuznetsova, I.; Uversky, V.; Turoverov, K. Fluorescent Proteins as Biomarkers and Biosensors: Throwing Color Lights on Molecular and Cellular Processes. *CPPS* **2008**, *9* (4), 338–369. https://doi.org/10.2174/138920308785132668.
(109) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379* (6637), 1123–1130. https://doi.org/10.1126/science.ade2574.
(110) Qiao, Z.; Nie, W.; Vahdat, A.; Miller, T. F.; Anandkumar, A. State-Specific Protein–Ligand Complex Structure Prediction with a Multiscale Deep Generative Model. *Nat Mach Intell* **2024**, *6* (2), 195–208. https://doi.org/10.1038/s42256-024-00792-z.
(111) AlphaFold and Beyond. *Nat Methods* **2023**, *20* (2), 163–163. https://doi.org/10.1038/s41592-023-01790-6.
(112) Kufareva, I.; Abagyan, R. Methods of Protein Structure Comparison. In *Homology Modeling*; Orry, A. J. W., Abagyan, R., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, **2011**; Vol. 857, pp 231–257. https://doi.org/10.1007/978-1-61779-588-6_10.

(113) Reva, B. A.; Finkelstein, A. V.; Skolnick, J. What Is the Probability of a Chance Prediction of a Protein Structure with an RMSD of 6 Å? *Folding and Design* **1998**, *3* (2), 141–147. https://doi.org/10.1016/S1359-0278(98)00019-4.
(114) Zhao, H.; Zhang, H.; She, Z.; Gao, Z.; Wang, Q.; Geng, Z.; Dong, Y. Exploring AlphaFold2's Performance on Predicting Amino Acid Side-Chain Conformations and Its Utility in Crystal Structure Determination of B318L Protein. *IJMS* **2023**, *24* (3), 2740–2740. https://doi.org/10.3390/ijms24032740.
(115) Rajbongshi, B. K.; Rather, S. R.; Bhowmik, S.; Sen, P. Ultrafast Excited State Relaxation of a Model Green Fluorescent Protein Chromophore: Femtosecond Fluorescence and Transient Absorption Study. *Journal of Molecular Structure* **2023**, *1275*, 134538. https://doi.org/10.1016/j.molstruc.2022.134538.
(116) Reinmuth-Selzle, K.; Tchipilov, T.; Backes, A. T.; Tscheuschner, G.; Tang, K.; Ziegler, K.; Lucas, K.; Pöschl, U.; Fröhlich-Nowoisky, J.; Weller, M. G. Determination of the Protein Content of Complex Samples by Aromatic Amino Acid Analysis, Liquid Chromatography-UV Absorbance, and Colorimetry. *Anal Bioanal Chem* **2022**, *414* (15), 4457–4470. https://doi.org/10.1007/s00216-022-03910-1.
(117) Rueden, C. T.; Schindelin, J.; Hiner, M. C.; DeZonia, B. E.; Walter, A. E.; Arena, E. T.; Eliceiri, K. W. ImageJ2: ImageJ for the next Generation of Scientific Image Data. *BMC Bioinformatics* **2017**, *18* (1), 529. https://doi.org/10.1186/s12859-017-1934-z.
(118) Malhotra, P.; Udgaonkar, J. B. How Cooperative Are Protein Folding and Unfolding Transitions? *Protein Science* **2016**, *25* (11), 1924–1941. https://doi.org/10.1002/pro.3015.
(119) Greenfield, N. J. Determination of the Folding of Proteins as a Function of Denaturants, Osmolytes or Ligands Using Circular Dichroism. *Nat Protoc* **2006**, *1* (6), 2733–2741. https://doi.org/10.1038/nprot.2006.229.
(120) Tsai, M.; Tsai, S.; Huang, Y.; Wang, C.; Sun, S.; Yang, J. Hydrogen Bonding-Induced H-Aggregation for Fluorescence Turn-On of the GFP Chromophore: Supramolecular Structural Rigidity. *Chemistry A European J* **2020**, *26* (27), 5942–5945. https://doi.org/10.1002/chem.202000358.
(121) Fang, C.; Frontiera, R. R.; Tran, R.; Mathies, R. A. Mapping GFP Structure Evolution during Proton Transfer with Femtosecond Raman Spectroscopy. *Nature* **2009**, *462* (7270), 200–204. https://doi.org/10.1038/nature08527.
(122) Xu, C.; Ye, R.; Shen, H.; Lam, J. W. Y.; Zhao, Z.; Zhong Tang, B. Molecular Motion and Nonradiative Decay: Towards Efficient Photothermal and Photoacoustic Systems. *Angew Chem Int Ed* **2022**, *61* (30), e202204604. https://doi.org/10.1002/anie.202204604.
(123) Hartley, S. M.; Tiernan, K. A.; Ahmetaj, G.; Cretu, A.; Zhuang, Y.; Zimmer, M. AlphaFold2 and RoseTTAFold Predict Posttranslational Modifications. Chromophore Formation in GFP-like Proteins. *PLoS ONE* **2022**, *17* (6), e0267560–e0267560. https://doi.org/10.1371/journal.pone.0267560.
(124) Abbas, U.; Chen, J.; Shao, Q. Assessing Fairness of AlphaFold2 Prediction of Protein 3D Structures. In *Proceedings of the 14th ACM International Conference on Bioinformatics*; New York, NY, USA, **2023**; pp 1–10. https://doi.org/10.1145/3584371.3612943.

(125) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative Approach to Computational Enzyme Design. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (10), 3790–3795. https://doi.org/10.1073/pnas.1118082108.
(126) Bloom, J. D.; Arnold, F. H. In the Light of Directed Evolution: Pathways of Adaptive Protein Evolution. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (Supplement 1), 9995–10000. https://doi.org/10.1073/pnas.0901522106.
(127) *Directed Evolution Library Creation: Methods and Protocols*; Arnold, F. H., Georgiou, G., Eds.; Methods in molecular biology; Humana Press: Totowa, N.J, 2003.
(128) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (15), 5869–5874. https://doi.org/10.1073/pnas.0510098103.
(129) Wang, X.; Minasov, G.; Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs. *Journal of Molecular Biology* **2002**, *320* (1), 85–95. https://doi.org/10.1016/S0022-2836(02)00400-X.
(130) Crowley, L. C.; Scott, A. P.; Marfell, B. J.; Boughaba, J. A.; Chojnowski, G.; Waterhouse, N. J. Measuring Cell Death by Propidium Iodide Uptake and Flow Cytometry. *Cold Spring Harb Protoc* **2016**, *2016* (7), 647–651. https://doi.org/10.1101/pdb.prot087163.
(131) Andersen, K. R.; Leksa, N. C.; Schwartz, T. U. Optimized E. Coli Expression Strain LOBSTR Eliminates Common Contaminants from His-tag Purification. *Proteins* **2013**, *81* (11), 1857–1861. https://doi.org/10.1002/prot.24364.
(132) Gatzeva-Topalova, P. Z.; May, A. P.; Sousa, M. C. Crystal Structure and Mechanism of the Escherichia Coli ArnA (PmrI) Transformylase Domain. An Enzyme for Lipid A Modification with 4-Amino-4-Deoxy-l-Arabinose and Polymyxin Resistance. *Biochemistry* **2005**, *44* (14), 5328–5338. https://doi.org/10.1021/bi047384g.
(133) Seelig, J.; Seelig, A. Chemical Protein Unfolding – A Simple Cooperative Model. *J. Phys. Chem. B* **2023**, *127* (39), 8296–8304. https://doi.org/10.1021/acs.jpcb.3c03558.
(134) Reetz, M. T. *Directed Evolution of Selective Enzymes: Catalysts for Organic Chemistry and Biotechnology*, 1st ed.; Wiley-VCH: Weinheim, Germany, **2016**. https://doi.org/10.1002/9783527655465.
(135) Sony Biotechnology. Sony MA900 Multi-Application Cell Sorter Brochure, **2024**.
(136) Beckman Coulter. CytoFLEX, CytoFLEX S, and CytoFLEX LX Flow Cytometers, **2023**.
(137) Camps, M.; Herman, A.; Loh, E.; Loeb, L. A. Genetic Constraints on Protein Evolution. *Critical Reviews in Biochemistry and Molecular Biology* **2007**, *42* (5), 313–326. https://doi.org/10.1080/10409230701597642.
(138) Treynor, T. P.; Vizcarra, C. L.; Nedelcu, D.; Mayo, S. L. Computationally Designed Libraries of Fluorescent Proteins Evaluated by Preservation and Diversity of Function. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (1), 48–53. https://doi.org/10.1073/pnas.0609647103.
(139) Thermo Scientific. NanoDrop One User Guide, **2016**. https://tools.thermofisher.com/content/sfs/manuals/3091-NanoDrop-One-Help-UG-en.pdf (accessed 2025-02-20).

(140) Held, P. Nucleic Acid Quantitation Using BioTek's Scanning Microplate Spectrophotometer, **2021**. https://www.agilent.com/cs/library/applications/quantitation-of-nucleic-acids-5994-2700EN-agilent.pdf (accessed 2025-02-20).
(141) Lampinen, J.; Raitio, M.; Perälä, A.; Oranen, H.; Harinen, R. Microplate Based Pathlength Correction Method for Photometric DNA Quantifi Cation Assay, **2012**. https://static.thermoscientific.com/images/D20827~.pdf (accessed 2025-02-20).
(142) Chai Discovery; Boitreaud, J.; Dent, J.; McPartlon, M.; Meier, J.; Reis, V.; Rogozhnikov, A.; Wu, K. Chai-1: Decoding the Molecular Interactions of Life. bioRxiv October 11, **2024**. https://doi.org/10.1101/2024.10.10.615955.
(143) Gros, P.-A.; Le Nagard, H.; Tenaillon, O. The Evolution of Epistasis and Its Links With Genetic Robustness, Complexity and Drift in a Phenotypic Model of Adaptation. *Genetics* **2009**, *182* (1), 277–293. https://doi.org/10.1534/genetics.108.099127.
(144) Wielgus-Kutrowska, B.; Narczyk, M.; Buszko, A.; Bzowska, A.; Clark, P. L. Folding and Unfolding of a Non-Fluorescent Mutant of Green Fluorescent Protein. *J. Phys.: Condens. Matter* **2007**, *19* (28), 285223–285223. https://doi.org/10.1088/0953-8984/19/28/285223.
(145) Uversky, V. N.; Winter, S.; Löber, G. Use of Fluorescence Decay Times of 8-ANS-Protein Complexes to Study the Conformational Transitions in Proteins Which Unfold through the Molten Globule State. *Biophysical Chemistry* **1996**, *60* (3), 79–88. https://doi.org/10.1016/0301-4622(96)00009-9.
(146) Banerjee, S.; Schenkelberg, C. D.; Jordan, T. B.; Reimertz, J. M.; Crone, E. E.; Crone, D. E.; Bystroff, C. Mispacking and the Fitness Landscape of the Green Fluorescent Protein Chromophore Milieu. *Biochemistry* **2017**, *56* (5), 736–747. https://doi.org/10.1021/acs.biochem.6b00800.
(147) Arık, M.; Çelebi, N.; Onganer, Y. Fluorescence Quenching of Fluorescein with Molecular Oxygen in Solution. *Journal of Photochemistry and Photobiology A: Chemistry* **2005**, *170* (2), 105–111. https://doi.org/10.1016/j.jphotochem.2004.07.004.
(148) Lakowicz, J. R.; Weber, G. Quenching of Fluorescence by Oxygen. Probe for Structural Fluctuations in Macromolecules. *Biochemistry* **1973**, *12* (21), 4161–4170. https://doi.org/10.1021/bi00745a020.
(149) Dobretsov, G. E.; Syrejschikova, T. I.; Smolina, N. V. On Mechanisms of Fluorescence Quenching by Water. *BIOPHYSICS* **2014**, *59* (2), 183–188. https://doi.org/10.1134/S0006350914020079.
(150) Manna, P.; Hung, S.-T.; Mukherjee, S.; Friis, P.; Simpson, D. M.; Lo, M. N.; Palmer, A. E.; Jimenez, R. Directed Evolution of Excited State Lifetime and Brightness in FusionRed Using a Microfluidic Sorter. *Integr. Biol.* **2018**, *10* (9), 516–526. https://doi.org/10.1039/C8IB00103K.
(151) Subach, O. M.; Gundorov, I. S.; Yoshimura, M.; Subach, F. V.; Zhang, J.; Grüenwald, D.; Souslova, E. A.; Chudakov, D. M.; Verkhusha, V. V. Conversion of Red Fluorescent Protein into a Bright Blue Probe. *Chemistry & Biology* **2008**, *15* (10), 1116–1124. https://doi.org/10.1016/j.chembiol.2008.08.006.