Molecular Dynamics on REX-GECO1 Reveal Structural Features Governing Fluorescence

Nikita Rozanov and advisor Dr. Chong Fang

Oregon State University, Department of Chemistry

(Dated: June 4, 2018)

Abstract

Fluorescent proteins have emerged as an essential toolset for bioimaging, creating a demand for engineering proteins with new and improved fluorescent properties. In this thesis, I explore the atomistic structure of REX-GECO1, a newly engineered protein biosensor that has unique optical properties. Since this protein has no available crystal structure, understanding the relationship between its structure and properties is difficult. To overcome this challenge, I use molecular dynamics simulations to predict the protein's structure and use this information to identify structural features that influence fluorescence. Moreover, I use the simulations to obtain thermodynamic information that provides further detail about the protein. These findings will be useful for understanding data obtained from ongoing ultrafast spectroscopic studies.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

Fluorescent proteins (FPs) can respond to incoming light by absorbing and then emitting photons of a different wavelength. Macroscopically, an FP will glow a certain color when placed in light, a property that has made them industrially significant as key components of bioimaging. The REX-GECO1 protein is a recent addition to the FP family, having an unprecedented wavelength difference between the absorption and emission maxima [1]. Its calmodulin (CaM) domain also provides Ca^{2+} sensing capabilities to REX-GECO1, making it a red fluorescent protein biosensor. Since this is a desirable characteristic for bioimaging applications, particularly regarding penetration depth and the avoidance of autofluorescence [1], knowledge of the reaction mechanism and structural features that enable this form of fluorescence will be useful for engineering new FPs.

The structural features of REX-GECO1 can be predicted computationally using molecular dynamics (MD) simulations. These simulations rely on Newtonian models for atoms and molecules to efficiently and accurately describe large molecular systems such as proteins. Simulations at equilibrium, which the system to propagate in time while maintaining constant thermodynamic properties, can reveal a protein's structural behavior in its natural state. In addition, more targeted approaches can obtain detailed thermodynamic information for a specific location within the protein. Free energy perturbation (FEP) is one such approach, which can be applied to REX-GECO1 to predict the protonation state of the photoactive component of the embedded chromophore.

A. Fluorescent Proteins

Fluorescent proteins have a unique ability to emit light when exposed to incident electromagnetic radiation. This optical property has been exploited to make fluorescent proteins a powerful and indispensable tool for in vivo bioimaging [2]. An increase in demand for these proteins has prompted the development of many protein variants with altered and enhanced imaging capabilities. GFP, the first purified and widely used fluorescent protein, has since become the basis for a great variety of new proteins and fluorescent-protein-based biosensors [3, 4]. These expansions allow researchers to choose a protein that has an emission wavelength, Stokes shift, quantum yield, and other properties that are tailored for a specific application. The photoactive chromophore within the protein is not an isolated system, but rather a part of a protein pocket, having an intricate network of surrounding residues. The complexity of the chemistry and photophysics of fluorescent proteins presents a challenge for engineering new ones. Traditionally, new FPs have been engineered through random or coordinated point mutations, but new structural information about their photochemistry, as well as the impact of the local protein environment, has helped introduce informed and targeted changes into FPs.

1. Green Fluorescent Protein

First isolated from the jellyfish *Aequorea victoria* in the early 1960s, GFP has since become widely used in bioimaging, and is the subject of significant scientific research [2]. GFP has a β-barrel structure, where the polypeptide winds eleven strands around a central helical structure [5]. At the center of this cage, the amino acids SER65, TYR66, and GLY67 form the basis for the protein's fluorescence. Specifically, these three residues undergo a post-translational modification to form the SYG chromphore. Following this auto-catalytic cyclization reaction, the two nitrogens adjacent to TYR66 rearrange to form an imidazolinone ring [6]. The end product is shown in Figure 1. The phenol group on the TYR66 forms a singular conjugated structure with nearby imidazolinone ring via a bridging methine group. The extension of the conjugation structure as a result of the post-translational modification is a major prerequisite for fluorescence.

The hydroxyl group on the tyrosyl end of the chromophore is an essential structural feature governing fluorescence. The hydroxyl group is predominantly present in the protonated form. Indeed, the ground state absorbance spectrum of GFP confirms this scenario. It shows two maxima at 395 nm and 475 nm corresponding to the protonated and deprotonated forms, respectively. These peaks are present in 6:1 intensity ratio over a large pH and salt concentration range, indicating that in GFP, the protonated form of the chromophore is predominant largely independent of solvent conditions [5].

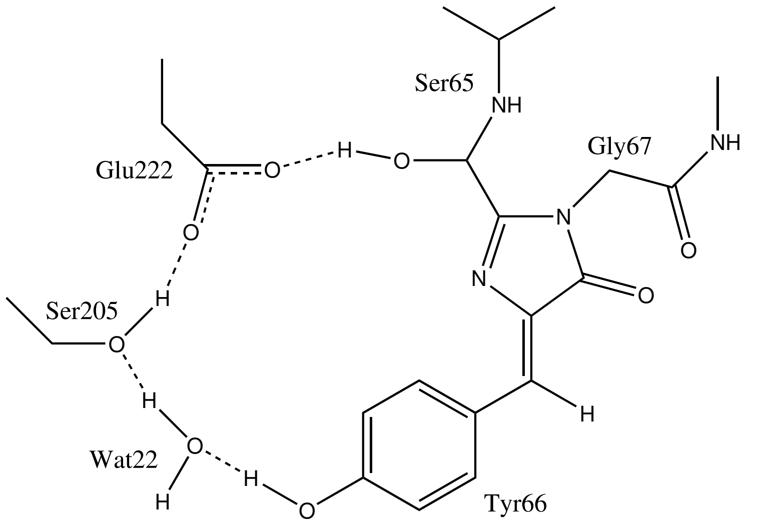

FIG. 1: A representation of proton transfer reaction enabling fluorescence in GFP. The neutral chromophore is able to fluoresce as a result of several protonated nearby residues.

Early studies on the GFP fluorescence mechanism revealed that when the protein is irradiated with 395 nm light, corresponding to the absorption maximum of the protonated chromphore, an excited state proton transfer (ESPT) reaction unfolds. Specifically, the excited state chromophore is much more acidic than its ground state counterpart, which allows swift proton transfer from the hydroxyl to the nearby crystal water. This step triggers concerted proton motions across WAT22, SER205, and GLU222, as illustrated in the Figure 1. After the reaction, GLU222 gains an additional proton, while TYR66 loses one, although this is not the same proton. Rather, a "proton wire" forms between the three residues and the chromophore's tyrosine, and several protons travel across residues during the ESPT reaction.

Despite the widespread use of GFP, a comprehensive picture of its fluorescence mechanism was not known until the late 2000s, primarily due to the difficulties of probing excited-state reactions occurring on femtosecond to picosecond timescales. Using femtosecond stimulated Raman spectroscopy (FSRS) to probe the ultrafast structural motions happening prior to fluorescence, Fang *et al.* [4] show that ESPT occurs through the proton wire described above, where a low-frequency structural mode modulates the conjugated tyrosine and imidazolinone rings due to the modulation of delocalized electrons that gates the ESPT reaction.

Another important mechanism to consider is internal conversion of the chromophore via a *cis-trans* isomerization. In this scenario, the chromophore responds to the energy from an incoming photon by twisting, a reaction that absorbs the energy of the photon without subsequent fluorescence. This type of radiationless pathway is counterproductive for the purpose of designing effective fluorescent proteins because it reduces ratio between the number of photons absorbed to those emitted, otherwise known as the fluorescence quantum yield. Fortunately, the structural constraints of the hydrogen bonding network and the protein β-barrel structure surrounding the chromophore limit the possibility of this unwanted reaction.

2. REX-GECO1

REX-GECO1 is a red-emitting calcium biosensor that consists of a CaM calcium binding protein covalently bound to a GFP derivative via the chicken myosin light chain kinase (M13) region on CaM [1]. Calcium binding to CaM induces a significant change in the absorption spectrum of and fluorescence quantum yield for REX-GECO1. Since the quantum yield is much larger for the Ca^{2+}-bound ($+Ca^{2+}$) form when compared to Ca^{2+}-free ($-Ca^{2+}$), this protein has been used to monitor in vivo calcium ion (Ca^{2+}) concentrations [1]. The absorbance spectrum of the two forms is also different. Specifically, the $+Ca^{2+}$ form shows a broad, singular peak at 480 nm, while the $-Ca^{2+}$ protein has a bimodal absorbance spectrum, with maxima at 440 and 580 nm. The specific structural changes that give rise to these spectral differences remain unknown. In addition, unlike its predecessors, this biosensor has an unprecedented Stokes shift in the $+Ca^{2+}$ form, a quantity that measures the difference between the absorption and emission bands during fluorescence events. Having this property is highly useful for bioimaging because it reduces the interference between the incident and emitted light, meaning that the emission signal can be collected with minimal contamination from incident light. Although the MD simulations that are the topic of this manuscript cannot explain the origin of this phenomenon, the structural features extracted could help correlate atomistic mechanisms and experimental results.

Unlike GFP, REX-GECO1 and its predecessor R-GECO1 have an MYG chromophore, where the serine on the chromophore is mutated into a methionine. This mutation within the chromophore alters its fluorescence properties. Currently, REX-GECO1 has no available crystal structure, presenting a challenge for understanding how its exact structural features influence its photochemistry. The structure of R-GECO1, its parent protein, has, however,

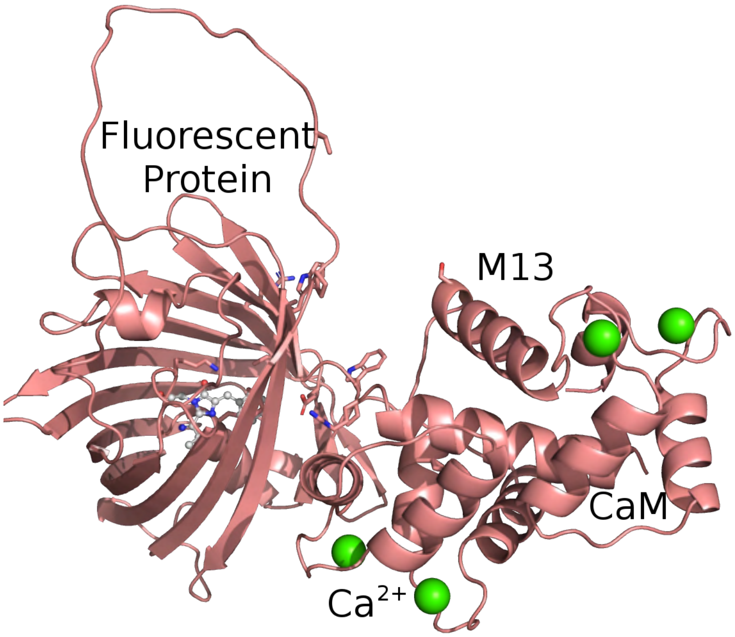

FIG. 2: The REX-GECO1 used for the simulations. Unlike R-GECO1, REX-GECO1 has several point mutations, especially near the chromophore, which is shown as a ball and stick model. In addition, a part of the fluorescent protein region is added since it is not resolved in the crystal structure. The calcium ions are shown as green spheres.

been determined. R-GECO1 has a similar peptide sequence to REX-GECO1, but is lacking several key mutations that give the latter unique fluorescence properties. Figure 2 shows REX-GECO1 with these mutation incorporated. Using this structure as the starting configuration MD simulations will allow the protein to evolve with time. This should allow the structure to adapt to the changes introduced by the mutations, resulting in a prediction for the REX-GECO1 structure.

B. Molecular Dynamics Simulations

MD simulations rely on Newtonian mechanics to describe the forces between atoms in a system, which are specified by a force field. These can be decomposed into the bonded and non-bonded components. There are two common non-bonded interactions: electrostatic interactions between atoms and van der Waals forces, which account for the size of a static electron cloud around atoms, otherwise known as steric effects. Meanwhile, the bonded terms account for the different forces that can result when atoms are physically connected via bonds. A single bond can vibrate like a spring, so a stretching parameter is included. In addition, a group of three atoms can scissor around an equilibrium angle. Finally, groups of four atoms can twist around a common bond, creating a torsion term, or they can be do the same around a central atom, resulting in an improper dihedral. The equation that connects these terms to the potential for the CHARMM force field is given by Equation 1.

$$\begin{aligned} U = &\sum_{\text{bonds}} K_r \left(r - r_0\right)^2 + \sum_{\text{angles}} K_\theta \left(\theta - \theta_0\right) \\ &+ \sum_{\text{dihedrals}} K_\phi \left(1 + \cos(n\phi - \phi_0)\right) + \sum_{\text{impropers}} K_\phi \left(\varphi - \varphi_0\right)^2 \\ &+ \sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}} \end{aligned} \tag{1}$$

This equation incorporates all the bonded and non-bonded forces previously described.

Although researchers generally agree on the main forces that contribute to an accurate Newtonian description of a biological system, different approaches exist for determining their parameters as well as choosing the optimal amount of detail. For example, force fields may rely on different methods to validate their results. Furthermore, some may choose to combine groups of atoms, say a non-polar methyl group, into a coarse grained model, while others explicitly calculate the potential for every atom. For the simulations presented here, the CHARMM force field will be used. It is a popular and well established force field that includes every atom in its calculations.

In order to be useful for many different biological systems, force fields only include the basic building blocks used to reproduce the systems, such as the individual amino acids, ions, and small molecules. It would not be practical or efficient to determine a complete set of new parameters for thousands of atoms when simulating a typical protein system. Choosing a set of parameters for these building blocks that are both accurate and transferable to a variety of systems is challenging. In the context of fluorescent proteins simulations, the chromophore is a non-standard amino acid, meaning that it is not included as part of a standard force field distribution. Parameters for certain chromophores, however, may be available in literature for a particular force field so researchers can directly use them. Another approach is to use software that automates the determination of parameters for a particular force field. The availability of a force field, or a method to obtain one, for a chromophore is an important consideration when designing the simulations for fluorescent proteins.

Another consideration specific to fluorescent protein simulations is that MD can only capture the ground state mechanics of the system based on thermodynamic considerations. Furthermore, bonds cannon be created or broken with a traditional MD simulation, meaning that non-equilibrium reactive pathways cannot be explored. Quantum mechanical effects are not explicitly accounted for in this approximation, but rather implicitly integrated into the non-bonded and some of the bonded terms. For fluorescent protein systems, this may not always be adequate because the chromophore can undergo photochemical reactions such as proton transfer. Because the creation and destruction of chemical bonds is not part of the classical MD framework, it is important to restrict these simulations to an non-reactive ground state.

1. Free Energy Perturbation (FEP) Theory

FEP is a computational method to calculate the differences in free energy between two configurations. Since free energy quantifies a system's thermodynamic stability, FEP can predict whether changes to a protein's structure are favorable. FEP harnesses the ability to modify the system's potential during the simulation to construct a thermodynamic pathway. During a simulation, the potential energy of the system and its individual components is fully specified, which allows for evaluating thermodynamic properties, including the free energy difference. Under constant pressure and temperature (NPT) conditions, the Gibbs free energy G is defined as

$$G = -\frac{1}{\beta}\ln Z$$

where $\beta = \frac{1}{k_B T}$. The partition function Z is defined in terms of all available energy states E_i through

$$Z = \sum_i e^{-\beta E_i}.$$

Assuming the kinetic energy between the states remains constant, the free energy difference between two states can be expressed as a function of the ensemble averages of potential energies in each state. This relationship is useful because the potential energy is easily measurable during a molecular dynamics simulation.

In practice, however, several states along the thermodynamic path between the endpoints are simulated. Specifically, the path is partitioned into several windows, each containing a potential that is a mixture of the two end states. In each window, the system is allowed to equilibrate, and the equilibrium potential energy is measured. The aggregate free energy difference for N windows is defined by Equation 2 [7].

$$\Delta G = -\frac{1}{\beta} \ln \left(\frac{1}{N} \sum_{i=1}^{N} \exp[-\beta \Delta U(\Gamma_i)] \right) \tag{2}$$

In Equation 2, Γ_i represents a microstate along the discretized reaction coordinate, which is denoted by λ.

2. $\mathrm{pK_a}$ *Estimation*

The FEP approach described previously can be applied to quantitatively understand a variety of protein properties. Estimating the $\mathrm{pK_a}$ of a protonation site within a protein is an important application of the FEP approach. This is especially useful within the context of understanding fluorescent proteins, where knowledge of the protonation state of the chromophore can help identify the fluorescence mechanism.

The free energy measured through molecular dynamics can be related to the acid dissociation constant K_a by

$$\Delta G = -N_A k_B T \ln K_a,$$

where K_a is the equilibrium constant for acid-base reaction defined by Equation 3.

$$\mathrm{HA} + \mathrm{B} \rightleftharpoons \mathrm{BH}^+ + \mathrm{A}^- \tag{3}$$

Here the proton donor HA, otherwise known as a Lewis acid, donates a proton to an acceptor B. From there, K_a is also directly related to the $\mathrm{pK_a}$ through

$$\mathrm{pK_a} = -\log_{10} K_a.$$

Combining the two equations yields Equation 4, which relates the $\mathrm{pK_a}$ to the Gibbs free energy difference, where $R = N_A k_B$ is the macroscopic ideal gas constant.

$$\mathrm{pK_a} = \frac{\Delta G}{\ln(10) N_A k_B T} \approx \frac{\Delta G}{2.303RT}. \tag{4}$$

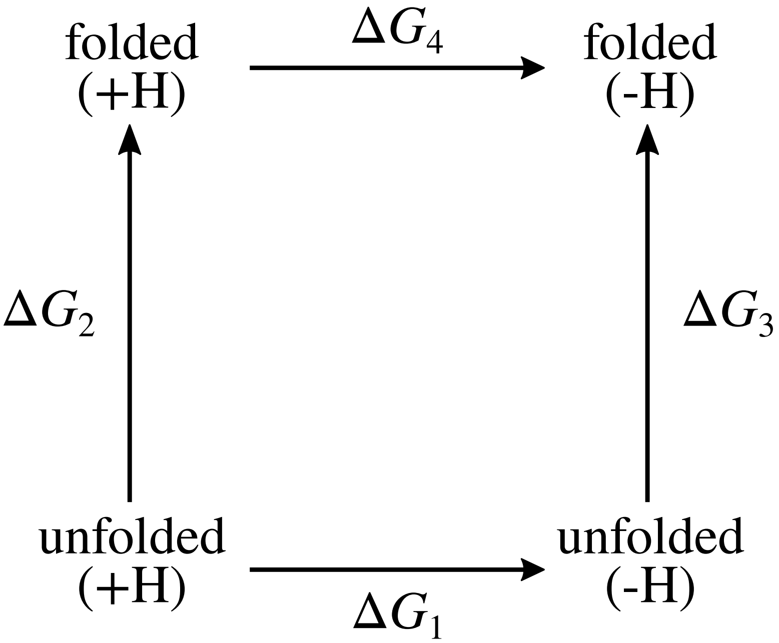

FIG. 3: The thermodynamic cycle used for determining $\mathrm{pK_a}$.

At first glance, it may appear that calculating ΔG for a proton annihilation reaction for the chromophore in the protein pocket is sufficient. However, this approach will not take into account the energy changes associated with a bond forming between the proton and its acceptor, as shown in Equation 3. Furthermore, the chromophore may have several plausible acceptors within the protein matrix, making it difficult to model the thermodynamics of this bond formation. Instead, these problems can be resolved by calculating free energy differences relative to a reference reaction. In other words, comparing ΔG for a proton annihilation reaction in a model compound to the same reaction happening in the protein pocket can show how the protein environment shifts ΔG. Simonson *et al.* [8] defined the thermodynamic cycle shown in Figure 3 to calculate the $\mathrm{pK_a}$ of an amino acid within the protein using the concepts described previously. Specifically, a model compound with a known $\mathrm{pK_a}$ is introduced to model the unfolded state of the protein, which acts as a reference state for thermodynamic calculations. In practice, this compound can be a tripeptide with the central residue being the amino acid of interest. In this case, the tripeptide includes the chromophore surrounded by two adjacent amino acids.

The free energy differences for removing a proton for both the protein system and model compound are calculated using MD simulations. From there, their difference forms a double

Gibbs free energy difference, which is used to calculate the pK_a shift between the two systems. Finally, since the pK_a of the model compound is known, the absolute pK_a for the amino acid in the protein matrix emerges from the pK_a shift.

II. METHODS

A. Equilibrium Molecular Dynamics

Equilibrium molecular dynamics simulations are conducted for the protonated forms of both the Ca^{2+} free and bound forms of the REX-GECO1 protein to understand the chromophore's local environment. Since REX-GECO1 has no available crystal structure, starting coordinates from the parent R-GECO1 structure are used. An equilibrium molecular dynamics simulation allows the starting protein structure to adapt to the structural differences introduced by the point mutations present in REX-GECO1.

The R-GECO1 protein in the $+Ca^{2+}$ state (PDB: 4I2Y)[9] is prepared by manually adding a missing residue chain using PyMOL [10], which is not resolve in the published crystal structure because of its high mobility. The protein is further modified by creating the mutations present in REX-GECO1, also using PyMOL. Furthermore, the nitrogen atom on the methionine part of the chromophore is corrected to an oxygen to remain consistent with prior characterizations of GFP derived chromophores.

The July 2017 version of the modified CHARMM force field is used [11] to model the forces present in amino acids and ions in the system; the TIP3P water model is used. Force field parameters for the protonated chromophore are taken, from Mironov *et al.* [12]. The deprotonated chromophore is modeled similarly, using parameters from Reuter *et al.* [13]. The systems are subsequently solvated in a water box that has a 14 Å padding distance between the protein and its edges. The solvent is neutralized and set to a $0.01\,\mathrm{mol\,L^{-1}}$ ionic concentration with sodium and chloride ions. The resulting system is shown in Figure 4.

Simulations are conducted using the NAMD 2.12 [14] integrator. The time step is set to 2 fs, along with a rigid constraint on all bonds using the SHAKE algorithm. In addition, electrostatics are approximated using a Particle Mesh Ewald (PME) with a Fourier grid spacing of 1.0 Å. Meanwhile, van der Waals parameters are described using a switching function with a 10.0 Å switching distance, a 12.0 Å cutoff. A 14.0 Å pair list with a 10 step

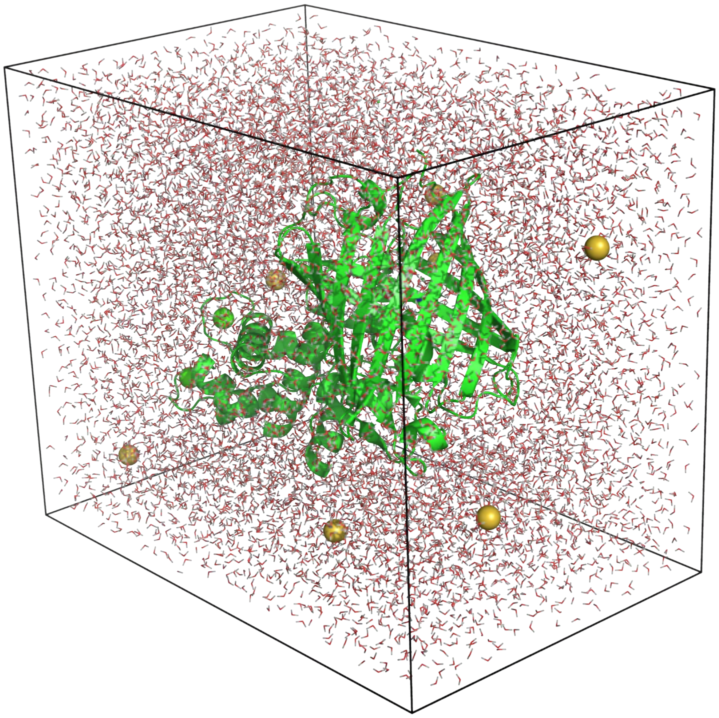

FIG. 4: An example of one of the systems that has been simulated. The REX-GECO1 protein is solvated in a water box and ionized to a small ionic concentration in order to better approximate physiological conditions. Chloride counterions are shown as yellow spheres in the solvent.

cycle is used to identify atoms that are within the cutoff distance. Van der Waals interactions are calculated at every step, whereas full electrostatics are computed every other step. In addition, atoms that are connected by two bonds or less are excluded from non-bonded calculations, and scaling parameters are applied to atoms connected by three bonds.

Temperature during all simulations is maintained at 300 K (room temperature), and is controlled via a Langevin bath applied to all atoms except for hydrogens ($\tau = 5.0\,\mathrm{ps}^{-1}$). Where appropriate, constant pressure control is applied though a Langevin piston set to a 1.013 25 bar target, which has a 100 fs oscillation period and a 50 fs damping constant. The walls of the system boundary are applied to expand isotropically in all directions to correct for deviation in pressure. Periodic boundary conditions are applied at system's boundaries.

The two systems are initially minimized for 10 000 steps with the conjugate gradient algorithm to remove any unfavorable contacts. Following, the system is equilibrated under constant pressure and temperature (NPT) conditions. During this simulation, a harmonic constraint with a 10 kcal mol^{-1} Å^{-2} force constant is applied to the protein backbone, excluding the M13 region. Subsequently, a 1 ns constant volume (NVT) simulation is conducted. During this process, the constraints are slowly released from 10 to 5 to 1 kcal mol^{-1} Å^{-2} in 250 ps steps. For the last 250 ps, the system equilibrates without any restraints. Finally, both systems undergo 20 ns unconstrained simulations. To maintain a constant cell size, NVT conditions are chosen for the production simulations.

B. Free Energy Perturbation (FEP)

While equilibrium molecular dynamics can yield important structural insights, they do not disclose how thermodynamic variables change for different chemical states. Knowledge of the free energy difference between two states is particularly useful because it can determine which state is thermodynamically favorable, and by what magnitude. In the context of current simulations, the free energy difference between the protonated and deprotonated forms is estimated (for both Ca^{2+} bound and free forms). After determining several reference constants, the free energy is converted into a pK_a, which is a more commonplace metric in biophysics.

The chromophore is prepared using the dual topology paradigm, where both forms of the mutated chromophore are present during the simulations. FEP simulations are conducted for 10 ns with 100 windows. During each window, the first 20 ps are reserved for equilibration and no energy data is collected. Soft core potential are employed, with a van der Waals radius-shifting coefficient of 5. Van der Waals forces are used throughout the entire simulation while electrostatics are only present for $\lambda > 0.5$. First, the proton annihilation reaction ($\lambda : 0 \rightarrow 1$) is finished, followed by the reverse reaction ($\lambda : 1 \rightarrow 0$). This approaches helps improve the reliability and error analysis of the resulting data. The simulation results are analyzed using the Bennett Acceptance Ratio (BAR) [15] to determine the total free energy change.

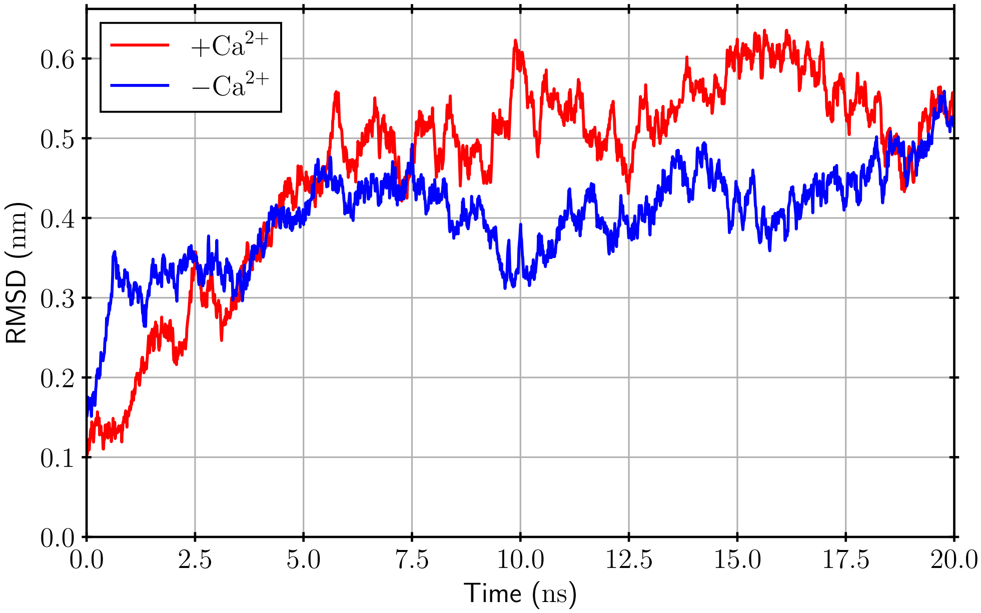

FIG. 5: The root mean square deviation (RMSD) of the α carbons of REX-GECO1 relative to the R-GECO1 crystal structure during the equilibrium simulations is shown for the Ca^{2+} bound and free forms. Atoms that are not resolved in the crystal structure are not included in the calculation, and neither are the mutated residues. Coordinates are sampled at a 10 ps rate.

III. RESULTS

A. Structural Features

The 20 ns equilibrium MD simulations on the $+Ca^{2+}$ and $-Ca^{2+}$ forms are intended to allow the REX-GECO1 systems to respond to the structural changes introduced by point mutations. The root mean square deviation (RMSD) is a standard method for quantifying the difference between two protein structures. The RMSD for n atoms is defined as

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}},$$

where $\hat{y}_i$ and y_i are the atomic coordinates from the two proteins. Typically, only the alpha carbons are used in these calculations because these atoms, as opposed to ones on the side chain, better reflect the structure of the protein. At the start of both simulations, there

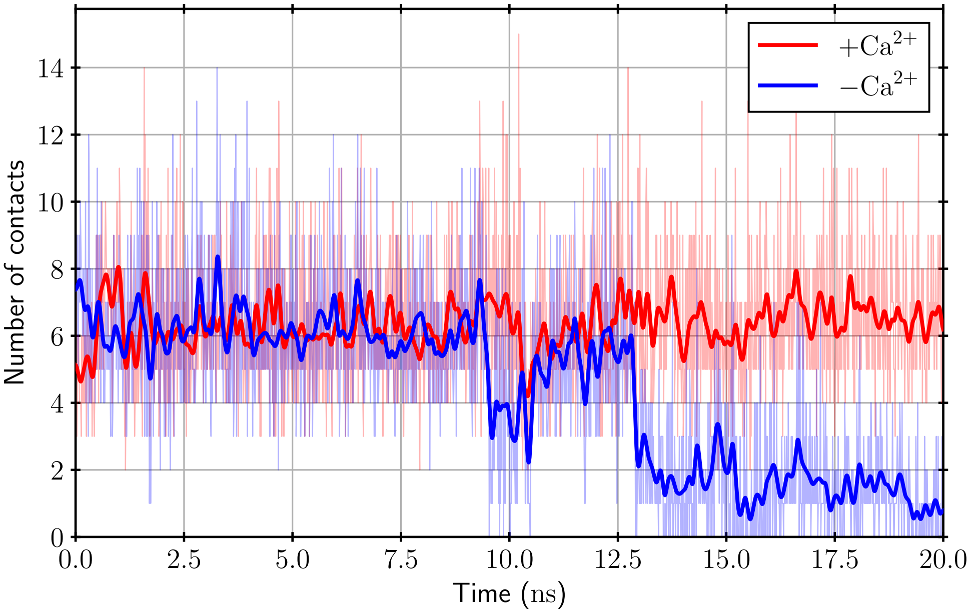

FIG. 6: The number of contacts within 3.5 Å of the hydroxyl group on the chromophore as a function of time for simulations on both protein forms. Coordinate data are collected every 10 ps during the simulation, and are shown as the faint, thin lines. Since the data contain a large amount of high frequency noise, a 10 point smooth is applied, which is shown as the thicker, solid lines.

is a rise in the RMSD, which is expected because the starting structure is from R-GECO1, and serves as an initial guess for REX-GECO1.

Since the structural changes happening in the vicinity of the chromophore have the greatest impact on the resulting fluorescence properties, a more targeted method for assessing structural features is needed. Measuring the number of nearby contacts to the chromophore, or more specifically the hydroxyl group on its tyrosine ring, which directly affects the ESPT pathway, is a simple yet informative method to better understand the protein pocket. Recall that the proton on this hydroxyl group is commonly involved in ESPT in fluorescent protein systems, so its environment in the REX-GECO1 system is particularly relevant. The number of contacts within 3.5 Å of this hydroxyl group is shown in Figure 6 for simulations on both forms of the protein. The number of nearby contacts shows that the structures of the two protein forms diverge by the end of the simulation. Specifically, the $-Ca^{2+}$ form shows a sharp decrease in contacts at 13 ns, while for $+Ca^{2+}$ they remain mostly con-

+Ca^{2+} final			**−Ca^{2+} mid**			**−Ca^{2+} final**		
residue	atom	% time	residue	atom	% time	residue	atom	% time
GLU78	CD	99.60	GLU78	CD	99.60	TRP300	HB3	32.40
GLU78	OE2	94.00	TRP300	HB3	92.40	LEU114	HB2	20.60
TRP300	HB3	91.80	TRP300	HB2	84.20	HOH	H2	12.80
TRP300	HB2	89.00	GLU78	OE1	83.60	TRP300	HD1	12.80
GLU78	OE1	60.60	GLU78	OE2	63.00	HOH	O	12.80

TABLE I: The five most frequent atomic contacts within 3.5 Å of the hydroxyl end of the chromophore are tabulated for various time windows for both the +Ca^{2+} and −Ca^{2+} forms. Each atom's residue, name, as well as the percentage of time it is within the cutoff is tabulated. The data designated as final corresponds to a 15-20 ns time window, whereas mid contacts are between 2.5 and 7.5 ns. The −Ca^{2+}-mid form is grayed out because it is not representative of a protein configuration that is expected in a physical system, and it is mainly included as a comparison.

stant throughout the simulation. This indicates that a lack of Ca^{2+} binding to the CaM region of REX-GECO1 causes nearby residues to move away from the hydroxyl end of the chromophore.

The simulation data can be further analyzed by inspecting the residues that are most often in contact with the hydroxyl group. Table I shows this information for select time segments of the simulations. It is necessary to choose specific time windows for this analysis because, as witnessed in Figure 6, the structure of the protein pocket changes throughout the simulation. The aforementioned table shows contacts for the 15-20 ns range, which should be representative of the protein pocket at the end of the simulation. In addition, contacts for the 2.5-7.5 ns time range are tabulated for the −Ca^{2+} form as a comparison.

Before the structural rearrangement observed near 12.5 ns in the −Ca^{2+} form MD simulations, both forms of the protein have a similar pattern of nearby contacts. This is seen by comparing the +Ca^{2+} final and −Ca^{2+} mid contacts in Table I, which is dominated by the same atoms from GLU78 and TRP300 in both forms. Towards the end of the −Ca^{2+} simulation, however, the percentage of time atoms are in contact decreases significantly. This structural change is important for explaining the different fluorescent properties of both protein forms, and will be elaborated in the Discussion section and a future publication.

B. Energetics

Energy data gathered from the MD simulations provide a thermodynamic comparison between the two protein forms. For example, changes in the chromophore's energetics in the equilibrium simulations may correlate to structural changes and suggest whether they are energetically favorable. The non-bonded interactions, consisting of van der Waals and electrostatic forces, are most relevant for this investigation because they describe how the protein pocket affects the chromophore. Figure 7 shows the non-bonded energies between the chromophore's two rings and the rest of the system for the equilibrium simulations.

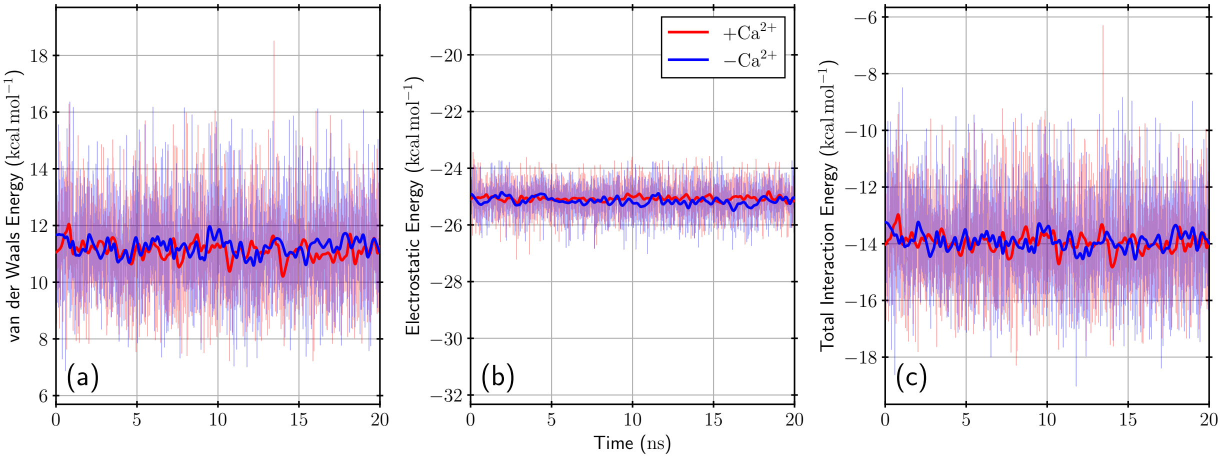

FIG. 7: Non-bonded energetics of the chromophore during the equilibrium simulations are shown. Specifically, the energy between the two rings on the chromophore and the rest of the system are used. The van der Waals (a), electrostatic (b), and total (c) non-bonded interaction energies are displayed.

Despite the dramatic changes to the protein pocket witnessed in the $-$Ca^{2+} form, equilibrium non-bonded energies do not show any time-dependent changes. For both simulations, the van der Waals, electrostatic, and total non-bonded energies of the chromophore remain largely constant.

The energy data from the FEP simulations, on the other hand, provide a more definitive result about the differences between the two protein forms. Specifically, these simulations quantitatively estimate whether the protonated chromophore is thermodynamically favorable using the cycle shown in Figure 3.

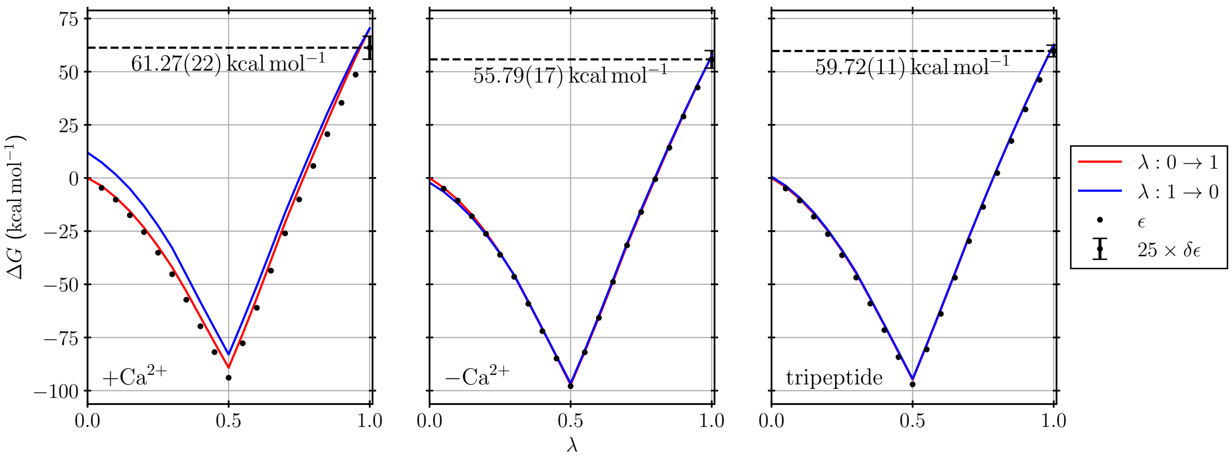

FIG. 8: Gibbs free energy differences ΔG are shown with respect to the FEP reaction coordinate for the $+$Ca^{2+}, $-$Ca^{2+}, and tripeptide systems. The forward reaction (proton annihilation, $\lambda : 0 \rightarrow 1$) is shown in red, while the reverse is in blue ($\lambda : 1 \rightarrow 0$). In addition, the BAR is used to estimate the combined free energy change, which is shown as a dashed line. In addition, the individual ΔG estimates are shown as dots along with their standard error (magnified by 25 for clarity).

The Gibbs free energy change ΔG for proton annihilation in the two protein forms (ΔG_4) as well as for the model tripeptide (ΔG_1) are shown as a function of the reaction coordinate λ in Figure 8. Qualitatively, $\Delta G_4 > \Delta G_1$ for the $+$Ca^{2+} form, meaning that there is a positive shift in free energy associated with the deprotonation. Consequently, this implies that deprotonation is less favorable for the $+$Ca^{2+} form when compared to the model compound. On the other hand, the $-$Ca^{2+} form shows the opposite trend: the deprotonated form is more favorable under the same conditions.

Although the pK$_a$ for the model compound has not been determined experimentally, it is assumed that it is about the similar to that for the GFP chromophore in solution. This model compound, known as HBDI, has pK$_a$ = 8.3 [16]. Under this assumption, the pKa's are calculated to be 9.4 and 5.4 for the $+$Ca^{2+} and $-$Ca^{2+} forms, respectively.

IV. DISCUSSION

The MD simulations were able to identify major structural differences between the two forms of the REX-GECO1 protein. Specifically, the $+\mathrm{Ca}^{2+}$ state has a more compact protein pocket, where several nearby residues are in proximity to the chromophore's hydroxyl end. In contrast, the $-\mathrm{Ca}^{2+}$ form has these same residues farther away, making them less likely to significantly interact with the chromophore. In context of the protein's photochemistry, the $-\mathrm{Ca}^{2+}$ form is not expected to have a significant ESPT pathway since there are no nearby proton acceptors. Meanwhile, GLU78 appears as a clear candidate for a proton acceptor in an ESPT reaction in the $+\mathrm{Ca}^{2+}$ form. To better visualize the protein pocket in both cases, the average structure from both equilibrium simulations is taken from the 15 to 20 ns time range and shown in Figure 9.

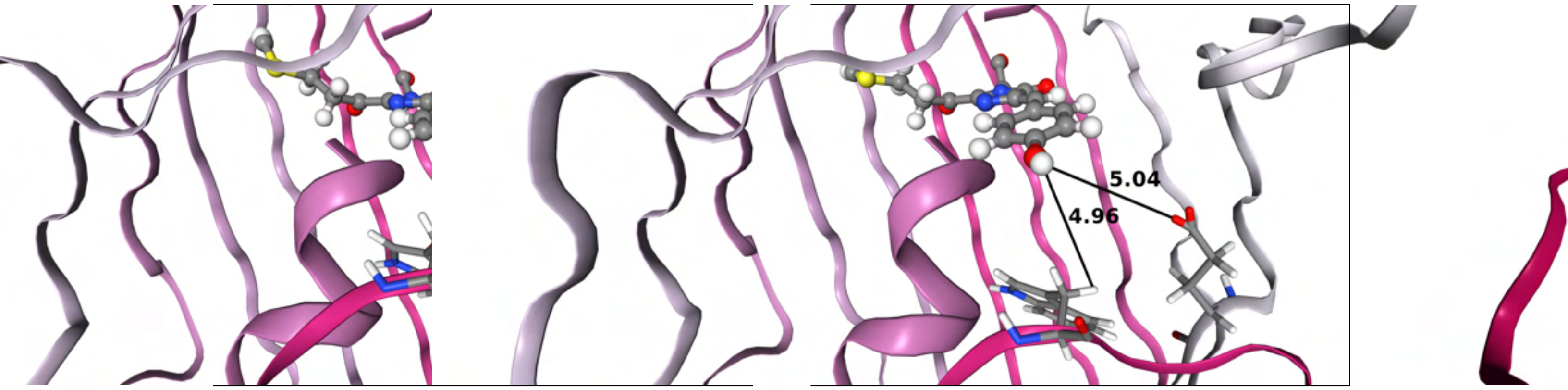

FIG. 9: The chromophore and its environment predicted by the MD simulations for the protonated chromophore in the $+\mathrm{Ca}^{2+}$ and $-\mathrm{Ca}^{2+}$ protein forms. This prediction averages the atomic coordinates from 15 to 20 ns in both simulations. Residues in proximity to the chromophore are highlighted, and select atomic distances are reported in Angstroms. Specifically, the residue right of the chromophore is GLU78, and the other is TRP300. Both of these are most frequently in contact with the hydroxyl group, as witnessed in Table I.

This representation adds further clarity to the data shown in Table I. In particular, the O···H distance between GLU78 and the chromophore is significantly different between the two simulations: 1.9 Å for $+\mathrm{Ca}^{2+}$ and 5.0 Å for $-\mathrm{Ca}^{2+}$. The former is comparable to the O···O distance in GFP, which is 2.7 Å [5]. These data suggest that ESPT is feasible for the $+\mathrm{Ca}^{2+}$ form because of a short interatomic distance between the chromophore's hydroxyl group and the oxygens on GLU78, while for the $-\mathrm{Ca}^{2+}$ form, this distance is too large for efficient proton transfer.

In addition, the FEP simulations help explain the difference in the absorption spectrum between the two protein forms. The $+\mathrm{Ca}^{2+}$ chromophore is calculated to be moderately basic ($\mathrm{pK_a} = 9.4$), implying that the protonated form dominates. This result is in agreement with previous results, which predict that a nearby residue enables ESPT from the chromophore. In contrast, the $-\mathrm{Ca}^{2+}$ form is primarily deprotonated according to the FEP simulations. While it is observed that it is more acidic than the $+\mathrm{Ca}^{2+}$ protein, as expected, the degree of acidity implies that the deprotonated form is largely dominant. However, two significant absorbance peaks are present, meaning that the protonated population should not be negligible. Still, it is important to understand that the MD model has inherent limitations based on the approximations that are made. As a result, it is reasonable to expect an error associated with numerical results. Overall, the FEP results lend some support that the protonation state is plausible reason for the absorbance lineshape. Still, other factors need to be considered before a determination can be made regarding the origin of the bimodal $-\mathrm{Ca}^{2+}$ absorbance spectrum.

V. CONCLUSION

MD simulations on the REX-GECO1 FP reveal structural features driving the photochemistry of the system. In the fluorescent $+\mathrm{Ca}^{2+}$ form, residues are compactly arranged around the chromophore, creating a more supportive environment for ESPT. Specifically, GLU78 is identified as the likely proton acceptor from the hydroxyl end of the chromophore, driving the fluorescence mechanism. Meanwhile, in the $-\mathrm{Ca}^{2+}$ form, the chromophore has significantly fewer nearby contacts, which presents a significant barrier to proton transfer reactions. This helps explain the difference in fluorescence quantum yield between the two forms of the protein. In addition, the protein is expected to be acidic in the $+\mathrm{Ca}^{2+}$ state, further supporting the possibility of ESPT. Meanwhile, the $-\mathrm{Ca}^{2+}$ protein is predicted to be largely deprotonated. This suggests that a low chromophore $\mathrm{pK_a}$ in combination with the lack of nearby contacts allows for non-radiative pathways to compete with fluorescence in the $-\mathrm{Ca}^{2+}$ form. Understanding the interplay between structure and fluorescence properties is especially useful for REX-GECO1, which has an unprecedented Stokes shift for the

+Ca^{2+} form. These findings provide structural data to begin understanding the protein's fluorescence mechanism.

Future work on this project will focus on correlating the structural features observed in the simulations to experimental spectroscopy data. These protein structures will be reconciled with ground state Raman spectra of the chromophore aided by quantum calculations. This type of comparison will test how well the simulations agree with experiment. Furthermore, Raman spectra can be obtained computationally for the chromophore using either QM/MM on the full system or QM only on relevant atoms. These methods can not only confirm experimental data, but can also help assign complex structural modes with a higher degree of accuracy.

[1] Jiahui Wu, Ahmed S. Abdelfattah, Loïs S. Miraucourt, Elena Kutsarova, Araya Ruangkittisakul, Hang Zhou, Klaus Ballanyi, Geoffrey Wicks, Mikhail Drobizhev, Aleksander Rebane, Edward S. Ruthazer, and Robert E. Campbell, "A long Stokes shift red fluorescent Ca^{2+} indicator protein for two-photon and ratiometric imaging," Nat. Commun. **5**, 5262 (2014).

[2] Gregor Jung, ed., *Fluorescent Proteins I*, Springer Series on Fluorescence, Vol. 11 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012) pp. 115–132.

[3] Roger Y. Tsien, "The Green Fluorescent Protein," Annu. Rev. Biochem. **67**, 509–544 (1998).

[4] Chong Fang, Renee R. Frontiera, Rosalie Tran, and Richard A. Mathies, "Mapping GFP structure evolution during proton transfer with femtosecond Raman spectroscopy," Nature **462**, 200–204 (2009).

[5] S. James Remington, "Green fluorescent protein: A perspective," Protein Sci. **20**, 1509–1519 (2011).

[6] Yingying Ma, Qiao Sun, Hong Zhang, Liang Peng, Jian Guo Yu, and Sean C. Smith, "The mechanism of cyclization in chromophore maturation of green fluorescent protein: A theoretical study," J. Phys. Chem. B **114**, 9698–9705 (2010).

[7] Andrew Pohorille, Christopher Jarzynski, and Christophe Chipot, "Good Practices in Free-Energy Calculations," , 10235–10253 (2010), arXiv:arXiv:1011.1669v3.

[8] Thomas Simonson, Jens Carlsson, and David A. Case, "Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models," J. Am. Chem. Soc. **126**, 4167–4180 (2004).

[9] Jasper Akerboom, Nicole Carreras Calderón, Lin Tian, Sebastian Wabnig, Matthias Prigge, Johan Tolö, Andrew Gordus, Michael B. Orger, Kristen E. Severi, John J. Macklin, Ronak Patel, Stefan R. Pulver, Trevor J. Wardill, Elisabeth Fischer, Christina Schüler, Tsai-Wen Chen, Karen S. Sarkisyan, Jonathan S. Marvin, Cornelia I. Bargmann, Douglas S. Kim, Sebastian Kügler, Leon Lagnado, Peter Hegemann, Alexander Gottschalk, Eric R. Schreiter, and Loren L. Looger, "Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics," Front. Mol. Neurosci. **6**, 2 (2013), arXiv:NIHMS150003.

[10] Schrödinger LLC, "The PyMOL Molecular Graphics System, Version 1.7," (2015).

[11] Jing Huang and Alexander D. MacKerell, "CHARMM36 all-atom additive protein force field:

Validation based on comparison to NMR data," J. Comput. Chem. **34**, 2135–2145 (2013).

[12] Vladimir A. Mironov, Maria G. Khrenova, Bella L. Grigorenko, Alexander P. Savitsky, and Alexander V. Nemukhin, "Thermal isomerization of the chromoprotein asFP595 and its kindling mutant A143G: QM/MM molecular dynamics simulations," J. Phys. Chem. B **117**, 13507–13514 (2013).

[13] Nathalie Reuter, Hai Lin, and Walter Thiel, "Green fluorescent proteins: Empirical force field for the neutral and deprotonated forms of the chromophore. Molecular dynamics simulations of the wild type and S65T mutant," J. Phys. Chem. B **106**, 6310–6321 (2002).

[14] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten, "Scalable molecular dynamics with NAMD," J. Comput. Chem. **26**, 1781–1802 (2005), arXiv:NIHMS150003.

[15] Charles H Bennett, "Efficient estimation of free energy differences from Monte Carlo data," J. Comput. Phys. **22**, 245–268 (1976).

[16] Christina Scharnagl and Robert a. Raupp-Kossmann, "Solution pKa Values of the Green Fluorescent Protein Chromophore from Hybrid Quantum-Classical Calculations," J. Phys. Chem. B **108**, 477–489 (2004).