Applications of Machine Learning in The Optimization of Genetically Encoded Optogenetic Sensors

Sarah J. Wait

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Andre Berndt, Chair

Douglas Fowler

Frank DiMaio

Program Authorized to Offer Degree:

Molecular Engineering and Sciences

University of Washington

Abstract

Applications of Machine Learning in The Optimization of Genetically Encoded Optogenetic Sensors

Sarah Wait

Chair of the Supervisory Committee:
Andre Berndt
Department of Bioengineering

Naturally occurring proteins provide a wealth of opportunities as tools in research, industry, and medicine. However, native proteins are rarely well suited for usage outside their biological setting. Therefore, the protein's functional ability and stability must be optimized through mutation of its amino acid sequence. This challenge is complicated by the vastness of each protein's mutation space, where mutants containing desired biophysical characteristics are rare and become more difficult to find as more specifications are required. Traditional engineering techniques, such as point-mutation screening, compound this issue by being time- and resource-intensive. Here, we present an alternative approach that harnesses machine learning models to learn from sequence-to-function libraries and screen untested mutants computationally. To showcase this technique, we identified variants of the genetically encoded calcium sensor, GCaMP, that improved the fluorescent response by 5-fold (eGCaMP^{2+}) and increased the decay speed by 3-fold (eGCaMP). To further demonstrate the capabilities of our machine learning platform, we utilized the same approach to engineer the functional capabilities of the red-shifted calcium indicator jRCaMP1b. Our study indicates that machine learning can efficiently learn from complex mutational datasets and harness their predictive power to guide the engineering of functional proteins. This methodology is poised to shift the protein

engineering landscape by providing alternative methods to rapidly engineer proteins for desired characteristics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge and thank my friends and family for their support. **Mom, Dad, and Ben,** thank you for teaching me the importance of perseverance, adaptability, and finding humor in the small things. I wouldn't be the scientist or person I am today without everything you have taught me. **Dennis and Tracy**, thank you for making home feel like it is not so far away.

Andre, thank you for giving me the freedom to be creative and follow my passions, wherever they may lead. I'm grateful for your unwavering support, always lending an open ear, a helping hand, and keeping the lab stocked with snacks to get us through the day. **Dr. Rappleye**, thank you for being my first friend in Seattle. When I started working with you, graduate school finally began to feel like it was starting. **Justin**, your passion for science is contagious, and I feel incredibly fortunate to have learned from you and been in your scientific orbit. **Netta,** thank you for always ensuring I was adequately caffeinated, fed, and home safe at the end of a long day. **Lily, Yuxuan, and Shani**, working with you three has been the greatest joy of my PhD. Thank you for being in my corner and facing the challenges of graduate school alongside me. To the undergrads **Mikayla, Jamison, Lila, and Amanda**, thank you for filling the lab with joy and laughter. I'm so proud of the scientists you've become, and I'm excited to watch your bright futures unfold.

To all my Seattle friends who have celebrated highs and weathered lows and were a wonderful reminder that there is life outside of graduate school. To **Kasey, Tim, and Robert,** thank you for not only sharing a hobby of climbing silly colored rocks but also for the long talks, constant encouragement, and joy of witnessing your lives over the years. To **Ellen, Matt, and Tyler**, who encouraged me to play some mental health-saving hooky and for the long bike rides/runs/skis that reminded me of what's essential. And to **Ryan,** who, with Lily, Yuxuan, and Shani, helped me escape reality, allowed me to treat rules as suggestions, and was the centerpiece of a beautiful chosen family.

Last, but certainly not least, **Sam**—if there was one true lynchpin of this entire journey, it was you. You uprooted your life so I could follow my dreams, and in every challenge I faced, I knew I wasn't facing it alone. Thank you for making home a safe, happy retreat from stress. Every achievement is as much yours as it is mine because it simply wouldn't have been possible without you.

"It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better.

The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again,

because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause;

who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly…"

– Theodore Roosevelt, *The Man In The Arena*

Chapter 1. Genetically Encoded Optogenetic Tools

ABSTRACT

This chapter explores the complexities surrounding pain, emotion, and perception in the mammalian central nervous system, highlighting the challenges researchers face in determining causal linkages between cellular activity and organismal behavior. Optogenetics, a technique pioneered by Karl Deisseroth, has been a groundbreaking innovation in addressing these challenges. Optogenetics enables the monitoring and regulation of neuronal activity using light-responsive genetic tools, allowing researchers to obtain cellular resolution information while preserving the organism's autonomy. This chapter discusses the major subclasses of optogenetic tools, including actuators, such as channelrhodopsin, and sensors, such as GCaMP. It explores the composition of optogenetic actuators and their engineering strategies for optimizing light sensitivity, photocurrent magnitude, kinetics, and activation wavelengths. Similarly, it explores optogenetic sensors' composition and engineering approaches, highlighting the importance of ligand binding domains, peptide linkers, and fluorescent proteins in achieving reliable signals. The chapter concludes by discussing traditional and high-throughput engineering approaches and the potential of machine learning algorithms in advancing protein engineering for optogenetic tools.

1.1 INTRODUCTION

The intricate mechanisms determining pain, emotion, and perception in the mammalian central nervous system are highly complex. Despite decades of research, a complete grasp of its activity's underlying mechanisms remains obscure. This complexity arises from the interplay of networks of neurons that communicate through electrical impulses known as action potentials and chemical messengers like neurotransmitters. Action potentials, characterized by rapid fluctuations in membrane voltage, propagate along neurons until they reach the axon terminal, triggering the release of neurotransmitters and other signaling molecules. These neurotransmitters serve as chemical messengers, facilitating communication between neurons. Depending on their composition and interaction, neurotransmitters can excite, inhibit, or modulate the signaling patterns of downstream cells. Researchers need tools that provide cellular-resolution information to establish causal relationships between neuronal activation and chemical release, while allowing animals to move and make decisions freely.

Historically, deriving the relationship between cellular activity and behavior was limited by several factors. Applying electrodes could modulate cell activity; however, the electrodes indiscriminately stimulate surrounding cells, regardless of cell type. Neurotransmitters and chemical agents can be administered, but they affect large regions of the brain and are limited by diffusion. Although dyes like Fluo4 could visualize certain chemical fluctuations, their effects were temporary and lacked specificity for cell types.

Optogenetics, pioneered by Karl Deisseroth, emerged as a groundbreaking technique that addresses crucial needs in behavioral studies. Optogenetics, or visible genetics, describes genetic tools that respond to light and allow researchers to monitor or even modulate neuronal activity. Optogenetic tools can be classified into two broad categories: actuators and sensors. Actuators

modulate cell activity, prompting neurons to fire in response to brief bursts of light (**Figure 1.1A**). In contrast, sensors are fluorescent tools whose brightness varies with the presence of specific neurotransmitters or chemicals, facilitating the correlation of chemical stimuli with behavior (**Figure 1.1B**). Both tools are genetically encoded, meaning the tool is administered through viral transfection or expressed in transgenic animal models. Using genetic promoters, DIO/Flex vectorization, and targeted viral application, the expression of the tool can be restricted to specific cell types within specific brain regions. Optogenetic tools have become indispensable in neuroscience research due to their ability to be expressed with cell type specificity, be triggered with high temporal accuracy, and allow the cellular environment to remain intact, preserving the organism's autonomy[1].

The proteins that comprise optogenetics tools can be engineered to emphasize desired characteristics. For example, the opsins that comprise optogenetic actuators can be mutated to increase photocurrent amplitude[2], kinetics[3,4], and alter activation wavelength[5]. Sensors can be mutated to increase fluorescent output[6,7], improve ligand sensitivity[8], and alter the fluorescent reporter to alternative color channels[9,10]. Engineering of critical tool features is time and resource-intensive, and the engineering cycles currently do not meet the demand for required tools. Alternative approaches that accelerate the engineering of these protein-based tools are necessary to meet the needs of behavioral studies.

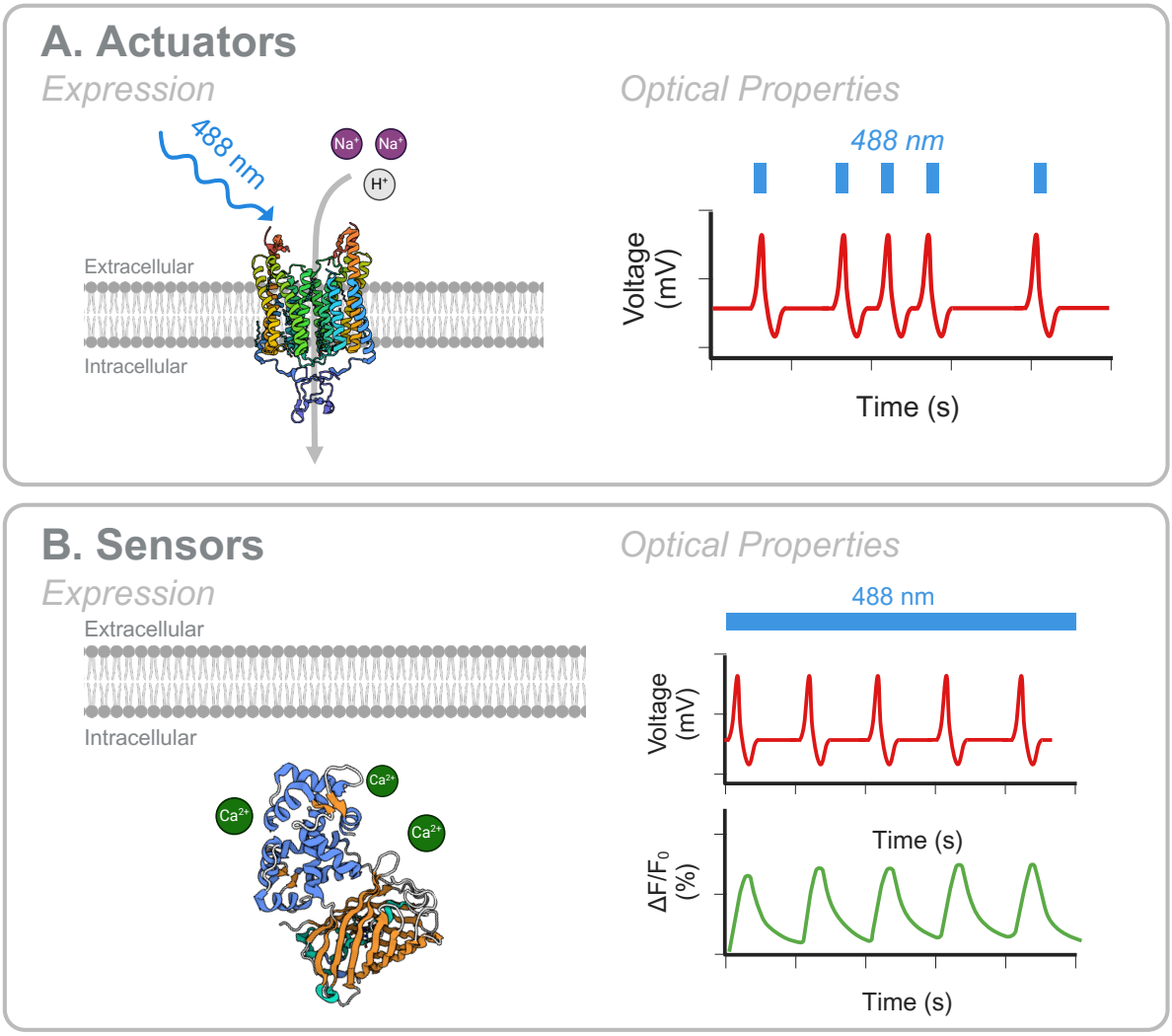


Figure 1.1 Optogenetic Actuators and Sensors

A. Optogenetic actuators are based on fluorescently activated opsins, which are light-sensitive G-protein Coupled Receptor Proteins. In the presence of blue light, the channelrhodopsin C1C2 (PDB: 3UG9), depicted here, will open to allow cations to flow into the cell. This will selectively generate an action potential in the targeted cell under a blue light stimulus.

B. Optogenetic sensors are chimeric proteins that link native sensing domains to fluorescent reporters. These sensors can be expressed in either the cytoplasm or cell membrane. When the native sensing domain interacts with the ligand of interest, conformational changes in the sensing domain will transfer to the fluorescent reporter. These fluorescence fluctuations can be recorded using light microscopy and interpreted to understand cellular activation and neurochemical flux. In the example displayed, GCaMP senses intracellular calcium and, upon calcium binding, produces changes in fluorescent output (PDB: 3WLD).

1.2 DESIGN CONSIDERATIONS AND MECHANISMS OF ACTION OF OPTOGENETIC TOOLS

1.2.1 Optogenetic Actuators: Actuator Composition

In 1982, the bovine rhodopsin was the first opsin to be sequenced, with thousands more being described in the time since[11,12]. Opsins are photoreceptive proteins found in visual systems that, upon light absorption, induce signaling cascades and the opening of ion channels within the receptive cell[13]. Significantly, vertebrate opsins influence ion channels through secondary pathways upon stimulation, where microbial opsins were found to act as unitary light-activated ion pumps or channels[14,15]. In 2005, researchers found that the microbial opsin Channelrhodopsin-2 (ChR2) could be stably expressed in neurons and, upon light administration, could produce action potentials in the host cell[16]. This discovery enabled the precise timing of excitatory events in freely moving animals[17]. Since this discovery, opsins that produce inhibitory events were also developed by expressing halorhodopsin, a light-activated chloride channel, or bacteriorhodopsins, light-activated proton pumps[17–19].

1.2.2 Optogenetic Actuators: Engineering Approaches

Opsins are membrane proteins that contain seven transmembrane helices. Upon binding of retinal, opsins become rhodopsin, indicating that they are functional and photon-sensitive proteins. Retinal covalently interacts with a lysine on the seventh transmembrane helix to form a retinal Schiff base (RSB). Photon absorption causes isomerization of the retinal and acts as the trigger for rhodopsin activity[20]. The residues that interact with the RSB serve as common targets for color and kinetics tuning within rhodopsin engineering.

Within actuator engineering, several fundamental biophysical properties are targeted for optimization, including increased light sensitivity, larger photocurrent magnitude, altered kinetics,

and shifted activation wavelengths. Previous work has shown that mutations to positions within helices 3 and 4 that interact with the RSB yield significant improvements in the light sensitivity of the variants as well as enhancements to peak photocurrent. However, the mutations that improve sensitivity and photocurrent often alter the off-kinetics to the point of creating bi-stable on-states of the rhodopsin to brief pulses of applied light[2,4,21]. To alter the kinetics of rhodopsin, researchers have targeted residues in the retinal binding pocket that stabilize the protonated RSB. For example, the E123T mutation in ChR2 reduced hydrogen bonding between the opsin and retinal, leading to accelerated on- and off-kinetics[3]. Lastly, there has been significant interest in developing actuators that preferentially activate at different wavelengths, such that tools can be co-expressed but independently activated. Multiple tools have been engineered that shift rhodopsin activation towards green and red light stimulus, though currently, activation within the 488nm color band remains, making tools hard to multiplex[5,22,23].

1.2.3 Optogenetic Sensors: Sensor Composition

Fluorescent proteins lay at the heart of optogenetic sensors. The green fluorescent protein (GFP), derived from the *Aequorea Victoria* jellyfish, was the first cofactor-less fluorescent protein to be sequenced and stabilized to be used as a reporter molecule[24–26]. Since the discovery of GFP, much work has been done to generate fluorescent proteins (FPs) with shifted excitation and emission wavelengths to cover the visible light spectra[25,27].

FPs have been broadly adopted as sensing modalities, where, in their basest form, a chimeric protein sensor can be formed by linking the FP gene and a gene of interest and expressing the composite gene in a host organism. By doing so, researchers can observe the production and localization of the protein of interest[28]. With the addition of new wavelengths of FP, researchers could observe how proteins interact with each other and their environment through fluorescence

resonance energy transfer (FRET). FRET describes the energy transfer between two spectrally overlapping FPs, where the emission of the first FP will excite the second FP when positioned within 10 nm distances[29]. Where FRET utilizes two FP channels to study protein conformational changes and protein-protein interactions, *Baird et al.* demonstrated that GFP could be circularly permuted, making GFP conformationally sensitive. *Baird et. al.* additionally showed that when fused to the calcium-sensitive protein Calmodulin (CaM), calcium binding could produce fluorescent changes within the circular permuted GFP (cpGFP)[30]. This discovery allowed researchers to observe the presence of biomolecules within cellular microenvironments and only required using a single fluorophore channel. Throughout all these techniques, the composition of the sensor remains consistent and includes native proteins fused to FPs using peptide linkers.

1.2.4 Optogenetic Sensors: Engineering Approaches

When building single fluorophore genetically encoded fluorescent indicators (GEFIs), engineers must consider the tool's use case, as different tissues will require different ligand selectivity, specificity, kinetic, and fluorophore requirements. To engineer the sensor's biophysical properties, mutations can be introduced to the sensing domain, the peptide linker composition, and the chosen fluorophore.

The ligand binding domain is the first and most crucial component of a GEFI. Most commonly, naturally occurring sensing domains are used for this application; however, some studies have used *de novo* protein design[31]. Multiple naturally occurring sensing proteins can be utilized, as in the case of GCaMP, where CaM and M13 are fused to cpGFP, or single protein binding domains with significant conformational changes can be harnessed, as in the case of G-protein coupled receptor (GPCR) based optogenetic sensors[7,32–35]. The composition of the binding domain determines many of the biophysical characteristics inherent to a GEFI. For example, to

alter the kinetics and sensitivity of GCaMP proteins, typical targets for mutation were the EF-hand domains that directly interact with calcium and the interface between CaM and M13[6,7,33]. GPCR-based sensors have emerged as common binding domains for optogenetic sensors. This advancement is significant, as GPCRs have a conserved 7-transmembrane domain structure[36]. With the discovery that cpGFP inserted into the third intracellular loop could produce fluorescent changes upon ligand binding within the dopamine sensing GPCRs, many similar studies have been published in rapid succession that can use the same design schema on different ligand sensitive GPCRs[34,35,37–39]. These GPCR-based sensors provide a remarkable wealth of opportunity owing to the sheer number of GPCRs expressed within the mammalian nervous system and the selectivity of each for different ligands of interest.

The linkers that fuse the FP to the native sensing domain are crucial for the function of the sensor as they translate conformational changes of the native sensing domain to the FP. As such, many studies have optimized the peptide linkers to significant effect on the overall dynamic range of the sensor[7,33,35,40]. Increasing evidence is also emerging that the native sensing domain plays a role in stabilizing the linkers as mutating residues that display structural proximity can dramatically alter the performance of the sensor[35,37,41]. Recent sensors have also demonstrated that increasing the length of the linkers can improve the affinity of the sensor for the targeted ligand[42]. Linker optimization is a crucial step in all sensor engineering, and optimization efforts should be performed on each new sensor produced.

The most important consideration for FPs is the ability to obtain reliable signals. To that effect, two essential design aspects must be considered: large signal response and on-target activation. Placement of the FP will be the greatest determinant of signal response as conformational changes within the binding domain will determine the solvent accessibility and,

thus, fluorescent readout of the FP. As discussed, standard techniques place the FP between two sensing proteins or within the third intracellular loop of GPCR-based sensors. Common mutation sites within the FP typically lie on the circular permutation linkers and N and C terminus of the FP[7,10,33,42,43]. The N and C terminus of the FP contributes to solvent accessibility and proper folding upon ligand stimulation. Many GEFIs currently use cpGFP. However, there is a need for improved sensors that utilize different FPs for tool multiplexing and the use of beneficial characteristics. For example, lower wavelength fluorophores such as red-shifted FPs exhibit decreased phototoxicity and allow for deeper recording within the tissue. Currently, many sensors that utilize different FPs have been made. However, they are far from as capable as their cpGFP-based counterparts and require further optimization[9,42].

1.3 OPTOGENETIC TOOL ENGINEERING APPROACHES

1.3.1 Traditional Engineering

Isolating protein variants with valuable characteristics is complicated by the vastness of a GEFI's mutational space[44,45]. Traditional engineering of GEFIs has relied on iterative saturation mutagenesis experiments[46]. In this method, a researcher will mutate a given residue position with all 20 amino acids. Variants with favorable responses compared to prior constructs are advanced for future mutations, constituting an evolution-based engineering. In addition to being a time and resource-intensive methodology, this method does not consider how mutations may collaborate or even conflict and limits the sampling of the mutation space to specific residue combinations. Even with structure-guided inference, traditional engineering techniques remain too time- and resource-intensive to meet the growing demand for optogenetic tools. In the coming sections, I will discuss newer techniques that have improved the speed of discovery that are becoming available.

1.3.2 High Throughput Engineering Approaches

Mutation library screening approaches, such as directed evolution (DE) and high-throughput screening (HTS), have been the most effective methods of engineering protein functionality[40,47–51]. These methods create random (DE) or pseudo-random (HTS) mutations on the protein of interest and apply artificial selection pressures to discover variants with desired performance capabilities. The best-performing variants of each screen are advanced as parent constructs for future screens, constituting a mimic of Darwinian evolution. Traditionally, only the mutations that occur in the best-performing variants are identified. These methodologies have been incredibly successful at improving the speed of discovery of new variants, though they still require resource-intensive screening cycles.

1.3.3 Computational Advancements

The relationship between sequence and structure has been well established within protein engineering. Computational predictions of static protein structures or modeling of ligand-protein interactions have been highly successful, as the correct conformation often satisfies the thermodynamic-favorability hypothesis[8,52–54]. However, these successes have not translated well to describing the structure-function relationship of dynamic proteins[47,53]. Current computational models lack the precision required to assess energetic changes induced by single point mutations and do not consider that natural proteins may undergo non-energetically favorable conformational changes[53,55–57]. For this reason, methods of analyzing the sequence-function relationship of proteins require further exploration, which may now be available with machine learning.

Machine learning (ML) algorithms have proven to be powerful data-driven tools in protein engineering. Models have shown the capability to generalize complex biological interactions without previous knowledge of elaborate pathways[58]. These algorithms have already been used to

engineer enzyme functionality, with recent studies using ML to alter fluorescent proteins and optogenetic tools with varying pipeline complexity[8,59–63]. The emergence of promising machine learning studies and developments in high-throughput engineering techniques act as complimentary advancements as they form a symbiotic inference relationship.

1.4 FUTURE DIRECTIONS OF OPTOGENETIC ENGINEERING

There remains a need for continued development of optogenetic tools, including expanding the current tools set as well as optimization of current scaffolds. Within optogenetic actuators, there is a critical need for fast variants that maintain large activation currents, as well as spectrally shifted variants. With additional opsin discovery, there will be more opportunities for actuator development. For optogenetic actuators, even with the large toolset that currently exists, it is vastly outnumbered by the number of biological chemicals that require sensing tools. As such, we still require additional tools as well as improving the biophysical characteristics of existing tools such as sensitivity, kinetics, and specificity. Additionally, within both optogenetic tool types, there is a substantial need for tools with the ability to multiplex. There currently is no microbial opsin-based tool that does not activate within the 488 nm wavelength band, which is the most common channel used by optogenetic sensors. This is compounded by off-target photoactivation of commonly used red-shifted fluorophores such as cp-mApple, which can lead to incorrect analysis when multiplexed with tools like ChR2.

To complete these tasks, the field would still benefit from further advancements in engineering approaches. The high-throughput engineering methods are promising, but within each screen, only the promising variants are sequenced, and thus very little is learned about variants that cause deleterious effects. These high-throughput methodologies present a unique opportunity to generate mutational libraries, should more variants be sequenced within each screen. These

sequence-to-function libraries can serve as inputs to train machine learning algorithms. In this way, high-throughput methods can inform machine learning algorithms, and in turn, machine learning algorithms can inform on possible beneficial mutations and direct downstream mutation library formations. Leveraging machine learning for mutation prediction and optimization could enhance the efficiency and effectiveness of tool development efforts.

Chapter 2. Development of a Machine Learning Pipeline to Guide Optimization of Genetically Encoded Calcium Indicators[1]

ABSTRACT

In this study, we focused on the transformative potential of machine learning in the engineering of genetically encoded fluorescent indicators (GEFIs), protein-based sensing tools that are critical for real-time monitoring of biological activity. GEFIs are complex proteins with multiple dynamic states, rendering optimization by trial-and-error mutagenesis a challenging problem. We applied an alternative approach using machine learning to predict the outcomes of sensor mutagenesis by analyzing established libraries that link sensor sequences to functions. Using the GCaMP calcium indicator as a scaffold, we developed an ensemble of three regression models trained on experimentally derived GCaMP mutation libraries. We used the trained ensemble to perform an in silico functional screen on 1423 novel, uncharacterized GCaMP variants. As a result, we identified the novel ensemble-derived GCaMP (*e*GCaMP) variants, *e*GCaMP and *e*GCaMP^{+}, that achieve both faster kinetics and larger $\Delta F/F_0$ responses upon stimulation than previously published fast variants. Furthermore, we identified a combinatorial mutation with extraordinary dynamic range, *e*GCaMP^{2+}, that outperforms the tested 6th, 7th, and 8th generation GCaMPs. These findings demonstrate the value of machine learning as a tool to facilitate the efficient pre-screening of mutants for functional characteristics. By leveraging the learning capabilities of our ensemble, we were able to accelerate the identification of promising mutations and reduce the experimental burden associated with trial-and-error mutagenesis. Overall, these findings have significant implications for optimizing GEFIs and other protein-based tools, demonstrating the utility of machine learning as a powerful asset in protein engineering.

[1.] This chapter contains text directly from and adapted from: Wait, S.J., Expòsit, M., Lin, S. *et al*. Machine learning-guided engineering of genetically encoded fluorescent calcium indicators. *Nat Comput Sci* **4**, 224–236 (2024). https://doi.org/10.1038/s43588-024-00611-w

2.1 INTRODUCTION

Genetically encoded fluorescent indicators (GEFIs) are protein-based sensors that allosterically link fluorescent proteins to protein domains that bind with specific ligands. Changes in fluorescence intensity upon ligand binding can be recorded spatiotemporally, allowing researchers to monitor ligands such as intracellular second messengers or neuromodulators in freely moving animals[52]. Today, GEFIs are essential tools in neuroscience, with sensors already developed for calcium, dopamine, norepinephrine, endocannabinoids, and opioids, among others[6,7,32–35,40,64–66]. However, to match each sensor's biophysical properties, like dynamic range or kinetics, with specific experimental needs, GEFIs demand extensive mutation-based engineering. Currently used engineering methods, such as trial-and-error mutagenesis, often come with substantial time and resource commitments. As a result, there's a critical need for innovative approaches that can reduce the experimental burden. Here, we use machine learning on a mutational library to investigate the sequence-function relationships of GEFIs and predict the functional characteristics of novel mutants.

Machine learning (ML) algorithms are valuable tools in protein engineering that have the ability to understand complex interactions without prior knowledge of intricate structure-function relationships[58]. These algorithms have been employed to engineer enzymes, fluorescent proteins, and optogenetic tools with varying levels of ML-pipeline complexity[8,59–63]. In this study, we combine the strengths of these previous examples into a novel approach. Firstly, our approach is based on the protein sequence-function relationship, ensuring that the resulting model can be applied to any mutational protein library. This versatility allows for broader adaptability across various protein engineering problems. Secondly, we utilize multiple models in parallel, implementing an ensembling process to inform our final predictions, which increased our model’s

accuracy. Thirdly, our models were trained by testing 554 biophysical amino acid properties. By combining these elements, we developed an ML-based approach with the potential to broadly impact protein engineering. The resulting pipeline not only provides impressive predictive capabilities but remains broadly adaptable to sequence-function libraries, paving the way for more efficient engineering of proteins.

We selected the calcium indicator GCaMP as a protein sensor scaffold to develop this platform. GCaMP is a chimeric protein that consists of circularly permuted GFP (cpGFP) fused to calmodulin (CaM) and calmodulin-binding peptide (CBP). GCaMP sensors have been widely adopted and have seen several generations of improvements to optimize their capabilities[6,7,32,33,64,67,68]. As a result, data from the *in vitro* functional characterizations of >1000 mutants are publicly available[7,33]. Using this data, we developed a stacked ensemble capable of pre-screening the *in vitro* functional characteristics of previously untested GCaMP mutants. This method enabled us to identify mutations that accelerate the off-rate kinetics and increase the $\Delta F/F_0$ response of jGCaMP7s. Our study demonstrates that ML ensembles can effectively learn from complex mutational datasets and that we can harness their predictive power to pre-screen mutation libraries for enhanced biophysical properties. This quality is increasingly important to complement the growing suite of high-throughput protein engineering methodologies, streamlining the analysis process, and further advancing the field.

2.2 RESULTS

2.2.1 Description of Variant Library, Computational Approach, and Predictions on Novel Sequences

To train the ML ensembles, we used published data from two previous publications to form our GCaMP variant libraries, which consisted of 1078 characterized mutants and their *in vitro*

functional characteristics derived from cultured neuron screening[64]. Within the variant library, we focused on the change in fluorescence ($\Delta F/F_0$) to one action potential (AP) stimulation (1AP $\Delta F/F_0$) and decay kinetics of the fluorescent response ($\tau_{1/2}$, decay half-time after 10 APs) (**Figure 2.1A**). When normalized to GCaMP6s as the baseline for target attributes (i.e., 1AP $\Delta F/F_0$ and $\tau_{1/2}$ = 1.0), we can see a broad distribution of variant capabilities and mutation locations within the GCaMP structure (**Figure 2.1B, C; Figure 2.2A**). We found that the sequence similarity is not deterministic for either the $\Delta F/F_0$ or kinetic response, as seen by the variability in mutation performance regardless of GCaMP generation (**Figure 2.2B, D**).

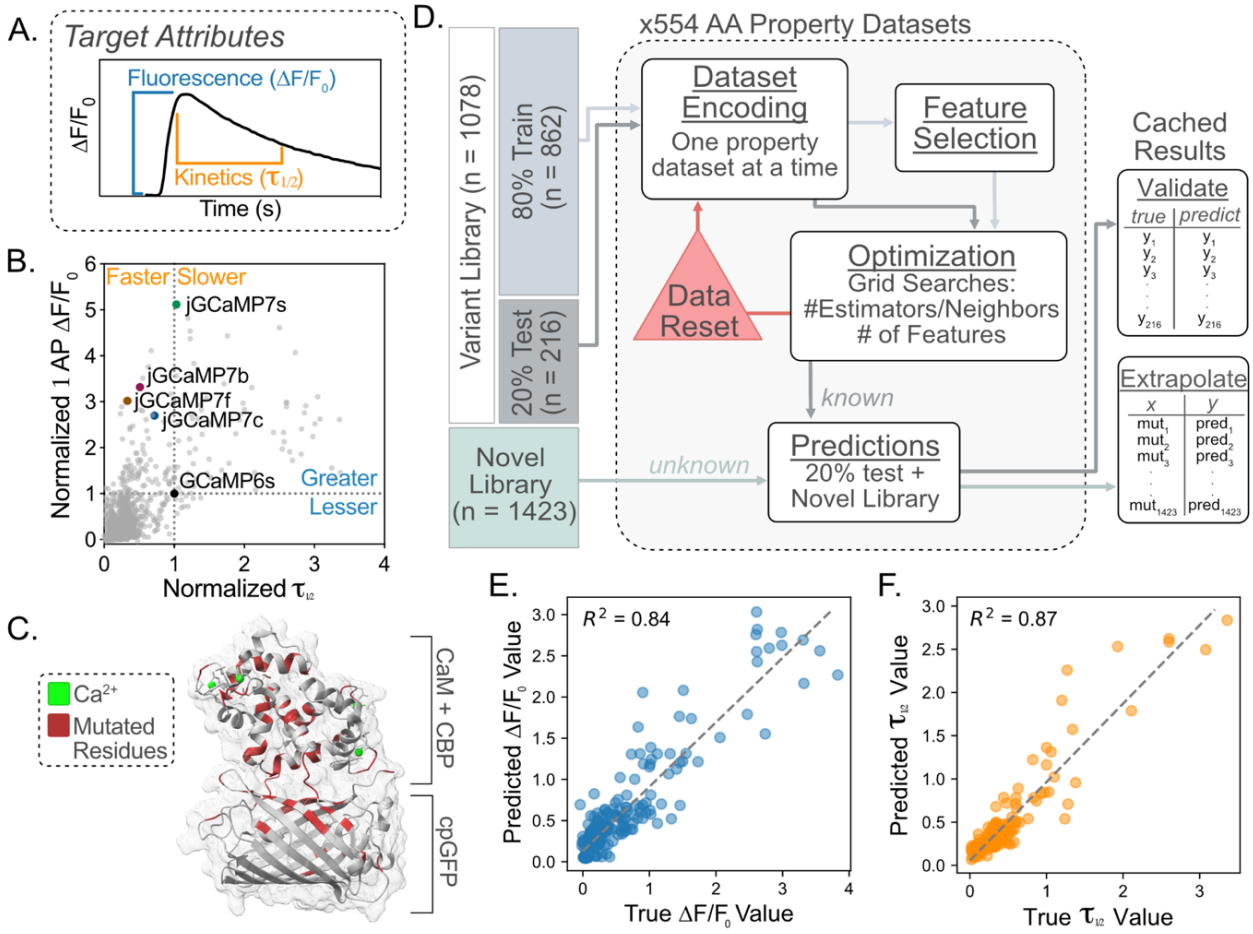


Figure 2.1 Description of Variant Library, Computational Approach, and Ensemble Cross-Validation

A. Description of the biophysical attributes of the GCaMP sensor targeted for engineering. Fluorescence ($\Delta F/F_0$) is the change between the baseline and maximal fluorescence upon

calcium sensing. Kinetics ($\boldsymbol{\tau}_{1/2}$) refers to the decay from maximum $\Delta F/F_0$ to half-maximal $\Delta F/F_0$.

B. Scatter plot depicts the 1AP $\Delta F/F_0$ by the $\boldsymbol{\tau}_{1/2}$ for each of the 1078 variants in the variant library[65]. Each value was normalized to GCaMP6s as 1.0 for 1AP $\Delta F/F_0$ and $\boldsymbol{\tau}_{1/2}$. Published variants are indicated with colored dots and text labels.

C. Crystal structure of GCaMP3-D380Y (RCSB: 3SG3, gray) with 75 residues (red) in which mutation information exists in the variant library[61]. These 75 residues indicate the positions used to form the novel library. Brackets indicate the main GCaMP domains calmodulin (CaM), calmodulin binding peptide (CBP), and circularly permuted GFP (cpGFP).

D. Overview of model training schema. The variant library[65] was split randomly into an 80% training set and a 20% testing set. The data was encoded using the AAINDEX property datasets. The train set underwent feature selection before being optimized using a grid search of key hyperparameters for each model. The optimized model was used to form predictions on the 20% test set and the novel library. The final predictions for both the test set and novel library were cached for downstream analysis.

E. Cross-validation of the fluorescence ensemble. The scatter plot x-axis represents the true $\Delta F/F_0$ value for each variant in the test set, and the y-axis represents the predictions made by the ensemble of the variants in the test set. The dotted line depicts perfect agreement between true values and predicted values. R2 value denotes the coefficient of determination of the scatter data.

F. Cross-validation of the kinetic ensemble. The scatter plot's x-axis represents the true $\boldsymbol{\tau}1/2$ value contained for each variant in the test set and the y-axis represents the predictions made by the ensemble of the variants in the test set. The dotted line depicts perfect agreement between true values and predicted values. R2 value denotes the coefficient of determination of the scatter data.

Before model training, the variants in the library were randomly assigned to training and testing sets at an 80/20 ratio for downstream cross-validation, where the mean values between the train and test sets were not significantly different in either the $\Delta F/F_0$ or kinetics library (**Figure 2.2C, E**). We tried three different methods of encoding our mutation dataset, including one-hot encoding, label encoding, and functional encoding with amino acid property datasets found on AAINDEX[69]. AAINDEX consists of 554 complete matrices that each describe a different AA property, such as size, polarity, or hydrophobicity. Within our hands, we found that the AAINDEX encoding led to the greatest predictive capability of our ensemble (**Figure 2.3C**). We trained and tested our models using each of the 554 AAINDEX property datasets to determine which properties led to the largest R^2 values during cross validation. The top five datasets were isolated

and used to train models that contribute to the final ensemble's predictions (**Figure 2.1D, Figure 2.3A**). Interestingly, the underlying AA properties that led to high R^2 values were associated with hydrophobicity for the $\Delta F/F_0$ library and conformation for the kinetics library (**Figure 2.3B, C, D**; **Table 2.1, 2.2**).

To improve prediction capabilities, we performed a stacked ensemble comprising a random forest regressor (RFR), K-neighbors regressor (KNR), and multi-layer perceptron network regressor (MPNR)[70,71]. The ensemble's predictions for each mutation are the average response from the 15 models (5 AA properties x 3 models). During cross-validation, the ensembles for $\Delta F/F_0$ and kinetics achieved R^2 values greater than 0.80 for predictions made on the test dataset (**Figure 2.1E, F**). The $\Delta F/F_0$ ensemble achieved a higher R^2 value than any of the models contributing to the prediction, which indicates the beneficial collaborative effect of ensembling (**Figure 2.3C**).

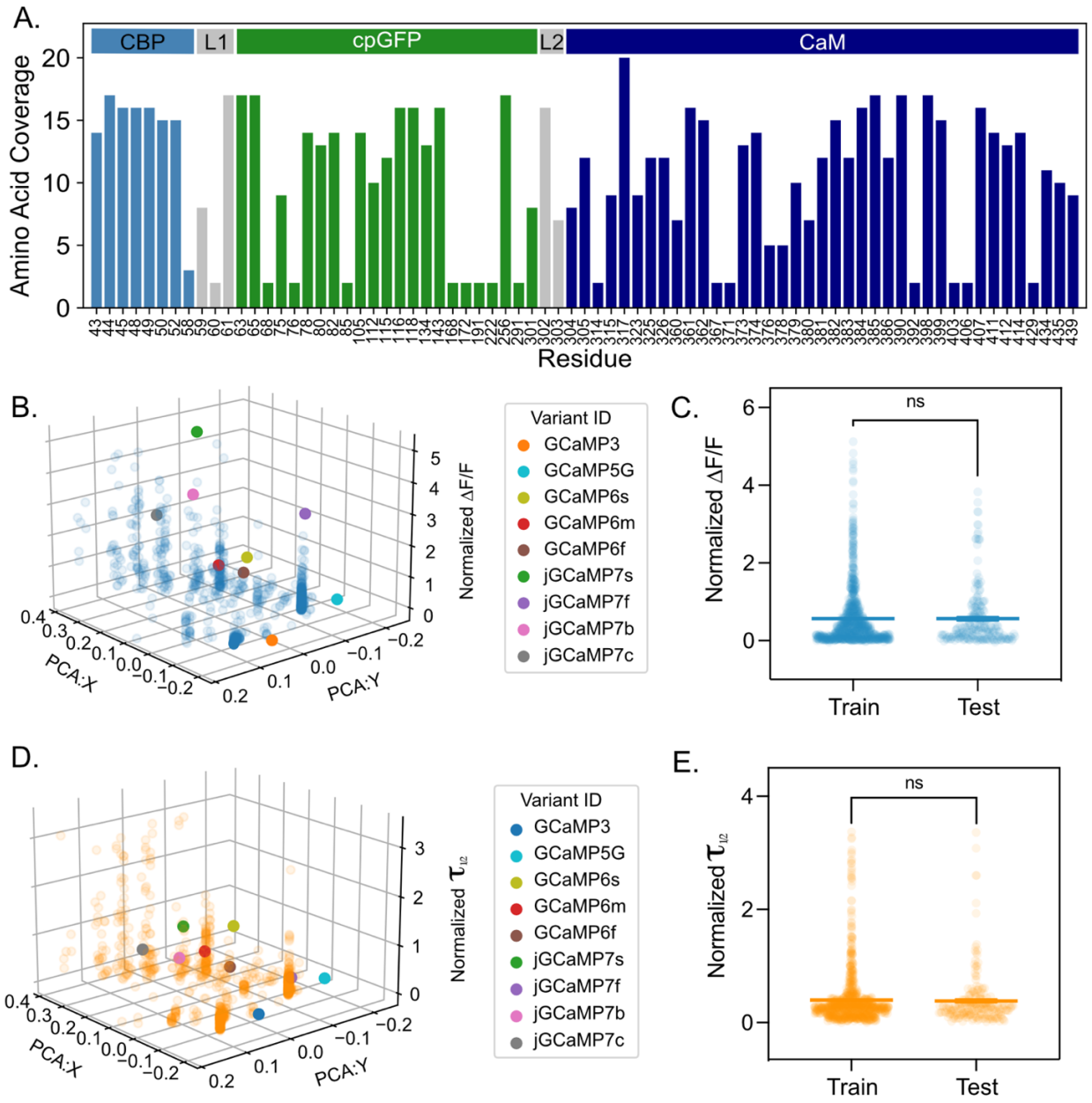


Figure 2.2: Mutation Scope of Chen & Dana dataset & Train/Test Breakdown

A. Bar plot depicts the number of tested amino acids for each residue in the full variant library. Color coding indicates the location within the GCaMP protein, light blue means the residue is in the CBP, gray means the residue is in one of the two linkers, green means the residue is in the cpGFP, and dark blue means the residue is in the CaM (x-axis denotes residue number, bar height indicates # of amino acids).

B. 2D principal component analysis (PCA) of sequences contained in the full variant library with the third dimension displaying the normalized $\Delta F/F_0$ of each variant. Published variants are included as differentially colored dots, indicated in the legend.

C. Span of Normalized $\Delta F/F_0$ values in the train set (n=862) and the test set (n=216). Each dot indicates one variant, where the line designates the mean and error bars SEM. Average values

and distribution did not differ between the two sets. (ns = P-value>0.05, unpaired t-test, two-tailed)

D. 2D PCA of sequences contained in the full variant library with the third dimension displaying the normalized $\boldsymbol{\tau}_{1/2}$ of each variant. Published variants are included as differentially colored dots, indicated in the legend.

E. Span of Normalized $\boldsymbol{\tau}_{1/2}$ values in the train set (n=862) and the test set (n=216). Each dot indicates one variant, where the line designates the mean and error bars SEM. Average values and distribution did not differ between the two sets. (ns = P-value>0.05, unpaired t-test, two-tailed)

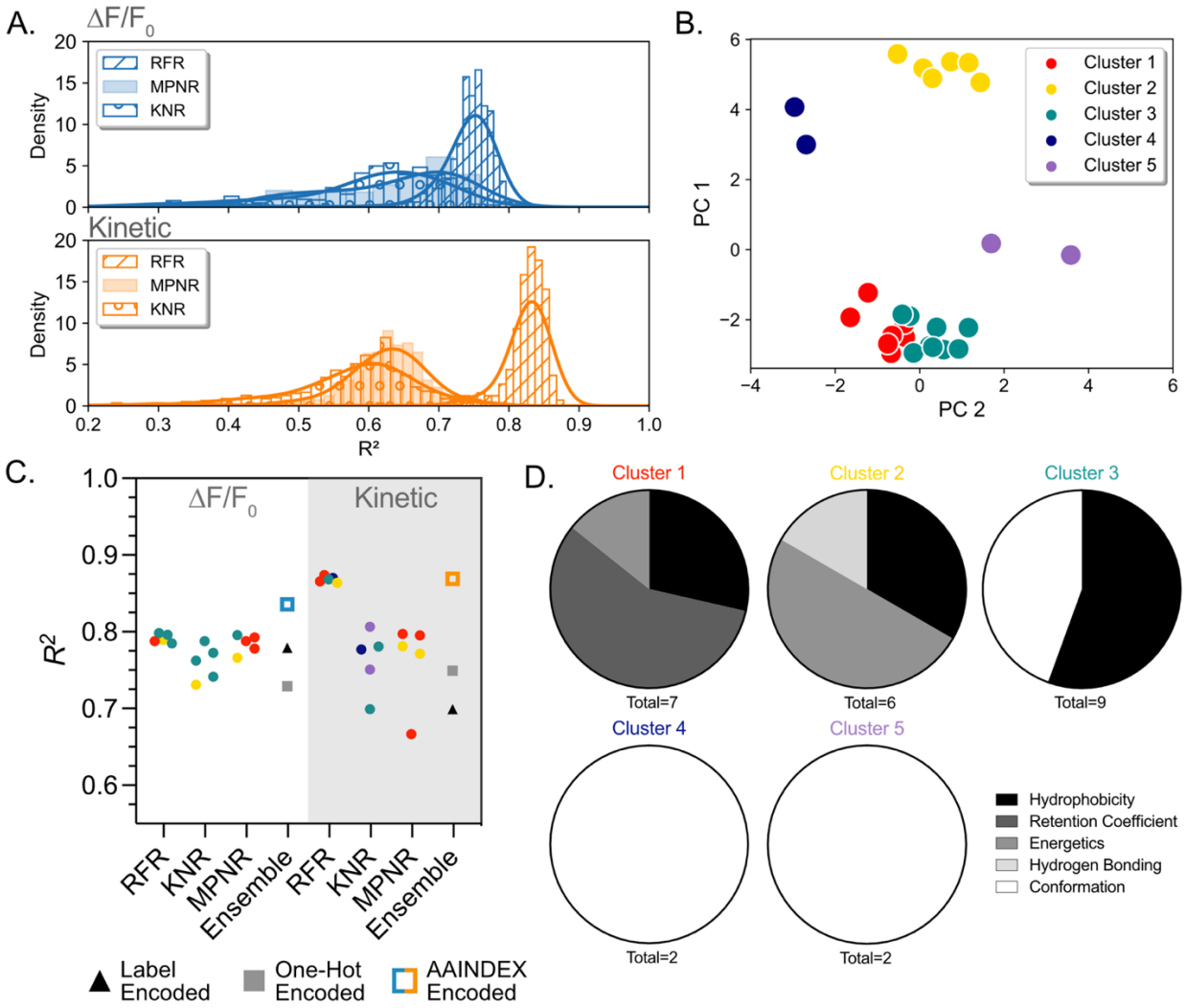


Figure 2.3: ΔF/F₀ and Kinetics Ensembles Display Amino Acid Property Preference

A. Kernel density estimates depict the range of the max R2 values saved for each of the 554 amino acid property datasets optimized for both ensembles. Bar patterning designates regressor type. The top five performing property datasets were advanced for downstream analysis.

B. Principal component analysis of the values in the top datasets from each ensemble (2x15 datasets). (number of components = 4, number of clusters = 5).

C. Scatter plot of R2 values from the top five performing amino acid matrices for each regressor type within each ensemble. Color mapping is indicative of PCA cluster identity (B.). Ensemble R2 indicates the final R2 value of predictions from each contributor model after averaging for the indicated encoding method. Models belonging to the ΔF/F₀ library are plotted on a transparent background, and models belonging to the kinetic library are plotted on an opaque background.

D. Pie-charts depict the amino acid properties that were found in each PCA cluster. Total indicates the number of property matrices within the cluster. Name and name color coordinate with B./C.

Table 2.1: $\Delta F/F_0$ library encoding dataset information.

Fluorescence				
Model	AAINDEX	R^2	Descriptor	Cluster #
RFR	ROSM880102	0.80	Side chain hydropathy, corrected for solvation (Roseman, 1988)	2
	MANP780101	0.797	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)	3
	KANM800104	0.796	Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)	3
	JURD980101	0.795	Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)	3
	MEEJ810102	0.795	Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)	1
MPNR	BROC820101	0.797	Retention coefficient in TFA (Browne et al., 1982)	1
	ZIMJ680105	0.787	RF rank (Zimmerman et al., 1968)	1
	FAUJ830101	0.783	Hydrophobic parameter pi (Fauchere-Pliska, 1983)	1
	CIDH920104	0.768	Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)	3
	BULH740101	0.766	Transfer free energy to surface (Bull-Breese, 1974)	2
KNR	KANM800104	0.78	Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)	3
	LIFS790102	0.77	Conformational preference for parallel beta-strands (Lifson-Sander, 1979)	3
	MANP780101	0.77	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)	3
	BASU050101	0.76	Interactivity scale obtained from the contact matrix (Bastolla et al., 2005)	3
	MIYS990101	0.75	Relative partition energies derived by the Bethe approximation. (Miyazawa-Jernigan, 1999)	2

Table 2.2: Kinetics library encoding dataset information.

Kinetics				
Model	AAINDEX	R^2	Descriptor	Cluster #
RFR	ZASB82010	0.88	Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982)	1
	QIAN880130	0.88	Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)	4
	RADA880101	0.88	Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)	1
	NADH010103	0.87	Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)	3
	GUYH850104	0.87	Apparent partition energies calculated from Janin index (Guy, 1985)	2
MPNR	BULH740101	0.79	Transfer free energy to surface (Bull-Breese, 1974)	2
	FAUJ880110	0.79	Number of full nonbonding orbitals (Fauchere et al., 1988)	2
	ZIMJ680105	0.78	RF rank (Zimmerman et al., 1968)	1
	MEEJ800101	0.78	Retention coefficient in HPLC, pH7.4 (Meek, 1980)	1
	ROSM880101	0.77	Side chain hydropathy, uncorrected for solvation (Roseman, 1988)	2
KNR	PALJ810103	0.77	Normalized frequency of beta-sheet from LG (Palau et al., 1981)	3
	PALJ810107	0.77	Normalized frequency of alpha-helix in all-alpha class (Palau et al., 1981)	5
	PALJ810112	0.77	Normalized frequency of beta-sheet in alpha/beta class (Palau et al., 1981)	3
	LEVM780103	0.77	Normalized frequency of reverse turn, with weights (Levitt, 1978)	4
	QIAN880105	0.76	Weights for alpha-helix at the window position of -2 (Qian-Sejnowski, 1988)	5

2.2.2 Identification of Mutations of Interest From Ensemble Predictions

We utilized the trained ensembles to predict a novel library's $\Delta F/F_0$ and kinetics capabilities. This library was created by taking jGCaMP7s and substituting each of the 75 positions previously mutated in the variant library with the remaining 19 amino acids (**Figure 2.1C**). After removing redundant variants, the novel library contained 1423 uncharacterized variants. We calculated the 'Predicted Change From jGCaMP7s' by subtracting the average predicted value of jGCaMP7s from the predicted value for each mutant. We performed an unpaired t-test between the 15 predictions made for each mutant (one from each contributor model) and the 15 predictions made for jGCaMP7s within the same library. These two metrics allowed us to isolate mutations whose predicted value differs significantly from jGCaMP7s (**Figure 2.4Ai**). From these

normalized value predictions, we can ascertain what mutations were predicted to affect the biophysical characteristics of jGCaMP7s in both the $\Delta F/F_0$ and kinetics (**Figure 2.4B, C**). In our model training, the jGCaMP7s sequence was purposely withheld. Nevertheless, the ensemble prediction ranked the base jGCaMP7s sequence within the top 15% of variants for a large $\Delta F/F_0$ response. Consequently, the ensemble predicted most variants to have a decreased $\Delta F/F_0$ capability compared to jGCaMP7s. Variants such as L317E, L317K, L317N, L317D, and L317H were all predicted to have a decreased $\Delta F/F_0$ response (<-2.2 a.u.) compared to jGCaMP7s, while variants such as G392F, G392I, and G392W were all predicted to have an increased (>0.25 a.u.) $\Delta F/F_0$ response (Figure 2B). In the kinetics library, L317E, L317D, L317N, and L317K were all predicted to decay faster (<-0.6 a.u.) than jGCaMP7s, while variants such as A390Y, L302D, and L302C were all predicted to decay slower (>0.3 a.u.) than jGCaMP7s (**Figure 2.4C**). The variants discussed above all fell outside 99.7% (±3σ) of $-\log_{10}$(P-Values), except for large $\Delta F/F_0$ predictions, indicating that the 15 contributing models displayed confidence in the effect of the mutation (±3σ, $\Delta F/F_0$: 0.612, kinetics: 0.242) (**Figure 2.4B, C**).

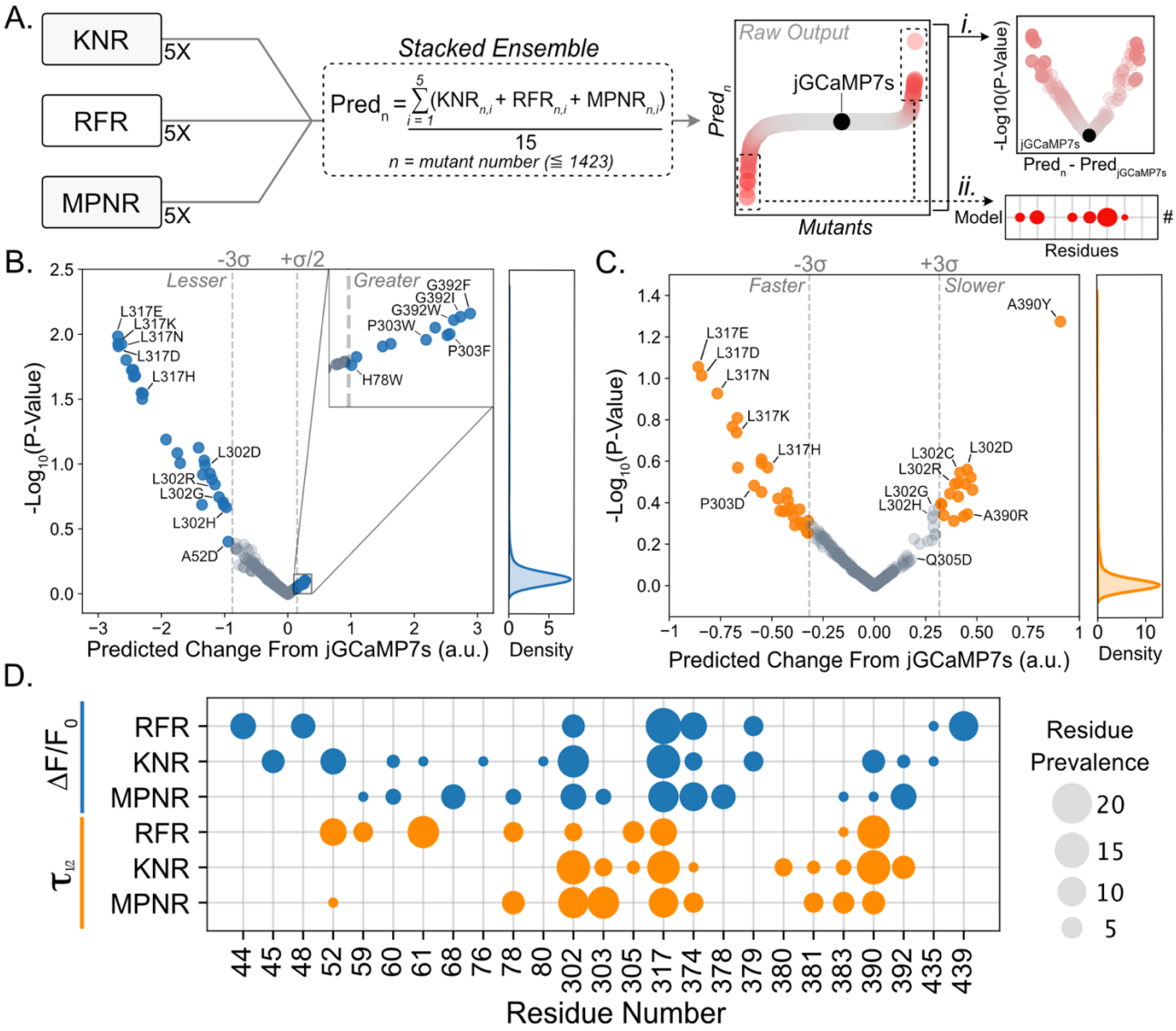


Figure 2.4: Predictions Derived From the Ensembles Led to Mutations of Interest for In Vitro Verification

A. Brief description of prediction analysis. From each model, the predictions from the top five property datasets were combined in the stacked ensemble. The stacked ensemble predictions were formed by averaging the predictions from the 15 contributor models for each variant (Predn) in the novel library. The raw output is thus the prediction (Predn) for each mutant, with a prediction for jGCaMP7s as a benchmark. The volcano plots were formed by subtracting the benchmark jGCaMP7s prediction from the variant prediction (x-axis) and P-values were derived by performing an unpaired t-test between the 15 predictions for variantn and the 15 predictions for jGCaMP7s (i). The bubble plot indicates the prevalence, or number of times a given residue appears, in the top 2.5% and bottom 2.5% of predictions, indicating that the ensemble believed mutations at said residue heavily influence sensor function (ii).

B. Volcano plots depicting the ensemble's prediction for a given mutation change in fluorescent response from jGCaMP7s (x-axis) and the log10(P-value) of the given prediction. P-values were calculated using an unpaired t-test on ensemble predictions (15 models) for jGCaMP7s

and the given mutation. Kernel density estimation (right) depicts the spread of log10(P-values) obtained.

C. Volcano plots depicting the ensemble's prediction for given mutations change kinetic capability from jGCaMP7s (x-axis) and the log10(P-value) of the given prediction. P-values were derived using an unpaired t-test on ensemble prediction (15 models) for jGCaMP7s and given mutation. Kernel density estimation (right) depicts the spread of log10(P-values) obtained.
D. Bubble plot depicting the number of times each residue (x-axis) appeared in the top 2.5% and bottom 2.5% of predicted values for each regressor that comprise each ensemble.

Next, we identified the residues in each library whose mutations had the strongest positive or negative impact on $\Delta F/F_0$ and kinetics. To do so, we isolated the top and bottom 2.5% of the ranked predictions and counted the number of times each residue appears (**Figure 2.4Aii**). We designated these as 'impactful residues', as the ensemble associated mutations at these positions to greatly alter protein function. We found that 22% and 18% of the impactful residues in the $\Delta F/F_0$ and kinetics libraries, respectively, were L317 predictions (**Figure 2.4D**), despite only 1.3% of variants in the novel library harboring an L317 mutation. Similarly, L302 predictions accounted for 14% and 16% of the impactful residues of the $\Delta F/F_0$ and kinetics libraries, respectively (Figure 2.4D*)*. Both L317 and L302 are in key positions of the GCaMP protein, where L317 is located on the interface between CaM and CBP and L302 is in the linker between CaM and cpGFP (**Figure 2.5A, B, C**). In contrast, residue A390 was found to be 4.5 times more impactful in the kinetics predictions than in the $\Delta F/F_0$ predictions. Like L317, A390 is located on the interface between CaM and CBP but on the opposing side (**Figure 2.5D**). Impactful residues for each biophysical property also tended to cluster. For instance, the kinetics library displays 38% prediction prevalence surrounding residue clusters Y380, R381, R383, and L302, P303, Q305. The prevalence of these residues is 2.38x higher in kinetics predictions than the $\Delta F/F_0$ predictions. These residues are located close to each other in 3D space, representing the residue linker and one of the inward loops of CaM (**Figure 2.5E**). Within the $\Delta F/F_0$ predictions, residue clusters N44,

K45, H48, V52, and M374, M378, K379 displayed 31% prediction prevalence, 3.9x higher than the kinetics library. Interestingly, when mapping residues H48, V52, L317, M374, M378, and K379 back onto the crystal structure, we observed that all these residues face inward toward one another, suggesting that they may be involved in interactions essential for the $\Delta F/F_0$ response (**Figure 2.5F**). These observations allow us to concentrate mutation efforts on key residues and identify specific residues or residue interactions that may be most advantageous to target for each biophysical characteristic.

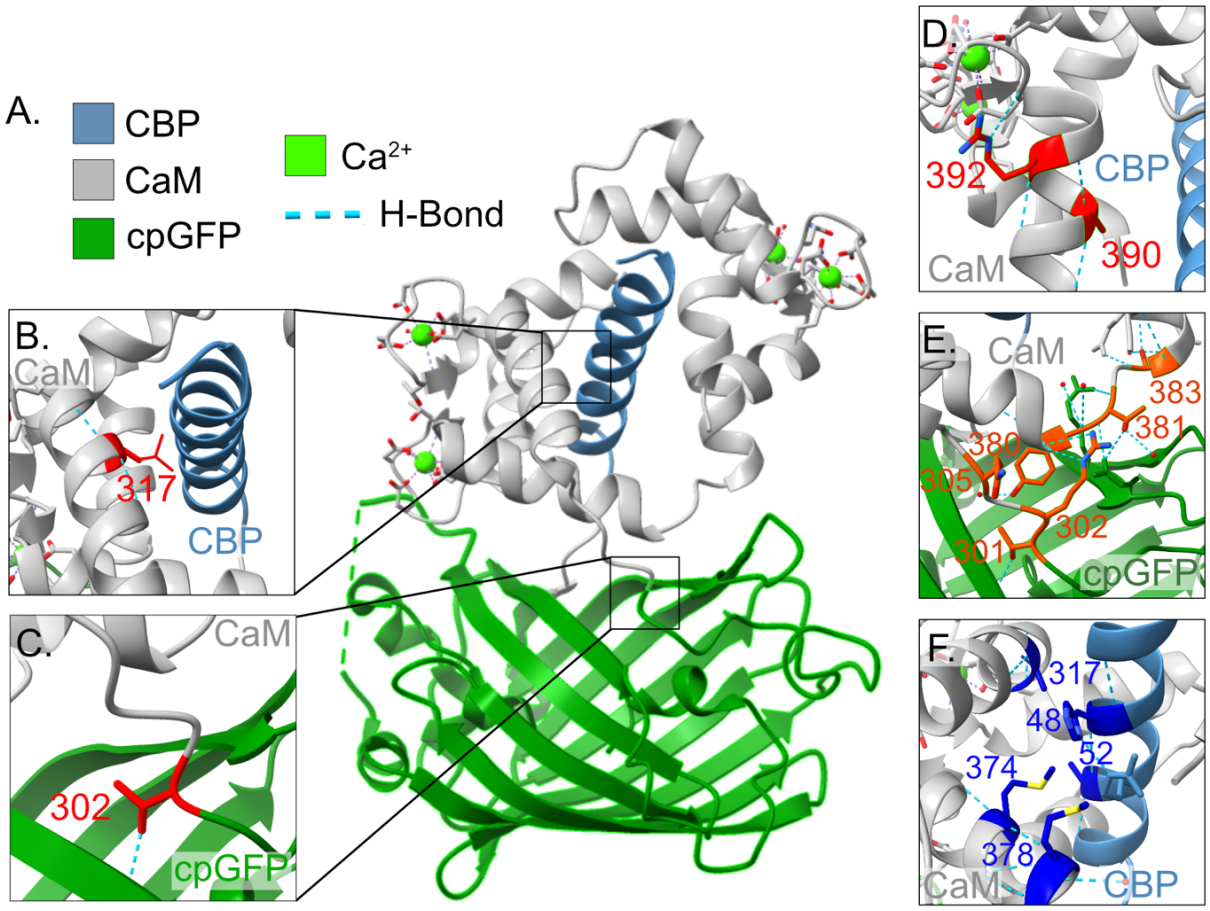


Figure 2.5: In Silico Predictions Indicate Key Residues and Interactions Within the GCaMP Protein

A. Crystal structure of GCaMP3 D380Y (RCSB: 3SG3), with color mapped CaM (gray), CBP (light blue), cpGFP (dark green), Ca2+ (lime green), and hydrogen bonds (light blue dashed lines).

B. Residue A317L rotamer (red) on the interface of CaM (gray) and CBP (light blue).

C. Residue L302 (red), on linker between CaM (gray) and cpGFP (dark green), with hydrogen bonds (light blue dashed lines).
D. Residue A390 (red, left), interfacing with the EF-hand motif, and G392 (red, right) interfacing with CBP (light blue), with color-mapped CaM (gray).
E. Representative image of residues Y380, R381, T383, L302, P303, and Q305 (dark orange) proximity on the crystal structure with color-mapped CaM (gray) and cpGFP (dark green).
F. Representative image of residues K48, V52, L317, M374, and M378 (dark blue) proximity on the crystal structure with color-mapped CaM (gray) and CBP (light blue).

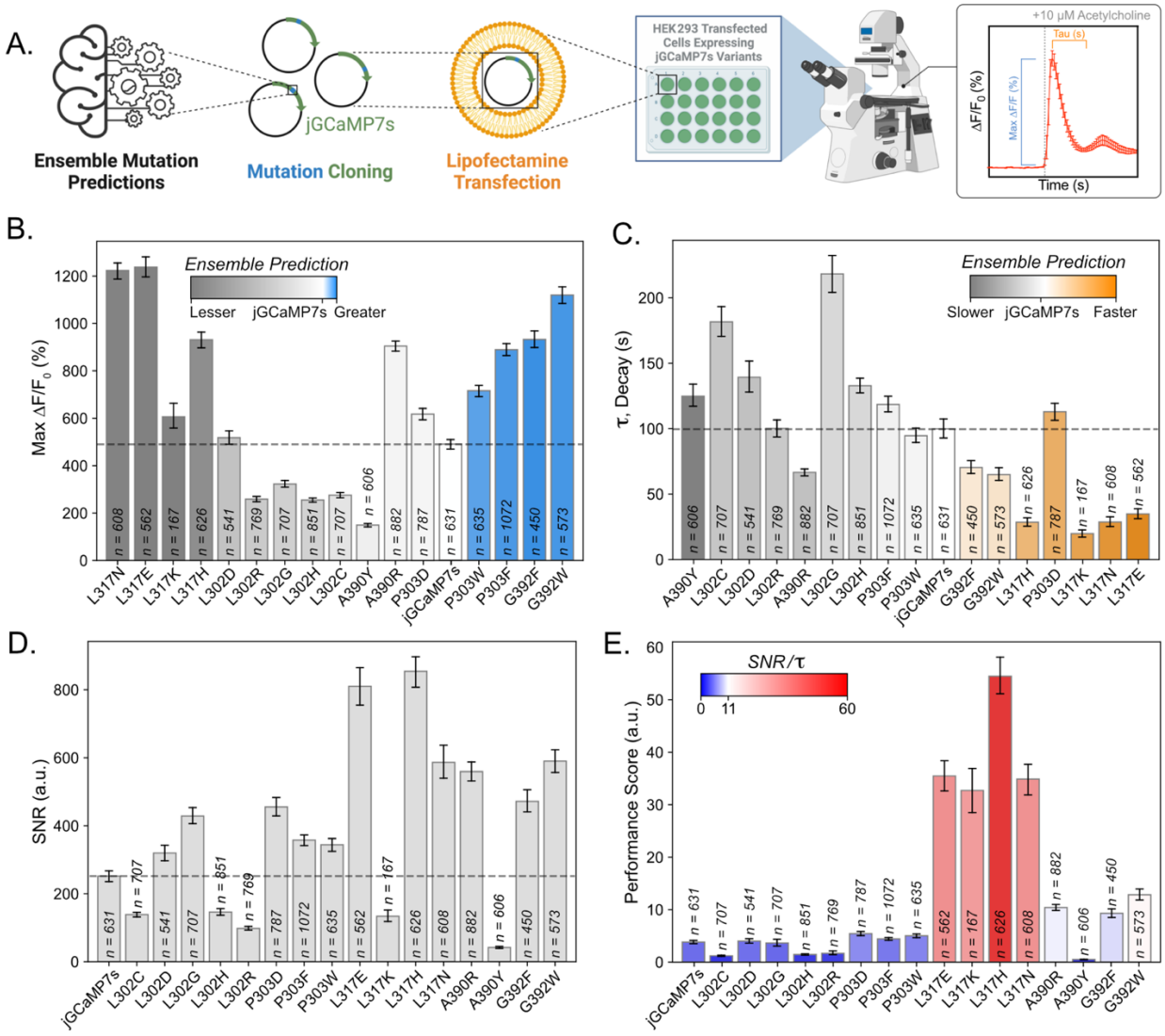


Figure 2.6: Gq/IP3 Assay in HEK293 Cells to Validate Ensemble Predictions

A. Brief description of the methods contained in the figure. Mutation predictions isolated from the ensemble are used as the basis for downstream variant analysis. Variants of interest are cloned into the jGCaMP7s (Addgene, #104463) backbone. These variants are then transfected into HEK293 cells using lipofectamine transfection. Forty-eight hours post-transfection, cells are time-course imaged using an epifluorescent microscope. The stimulation protocol contains a period to collect baseline fluorescence, a bath addition of 10 µM acetylcholine, and a decay period. Visual representations of the qualifications in **B./C.** are found on the representative response trace.

B. Max fluorescent responses (*Eq.1*) that were obtained from each mutant of jGCaMP7s expressed in HEK293 cells and stimulated with ten µM acetylcholine. Heat mapping demonstrates the ensemble's prediction of the given mutation's performance *in vitro*. Mutations are sorted in order of the ensemble's predicted performance. Dotted line depicts mean performance of the base construct, jGCaMP7s.(n = number of cells quantified; bars depict mean + bootstrapped 95% ci).

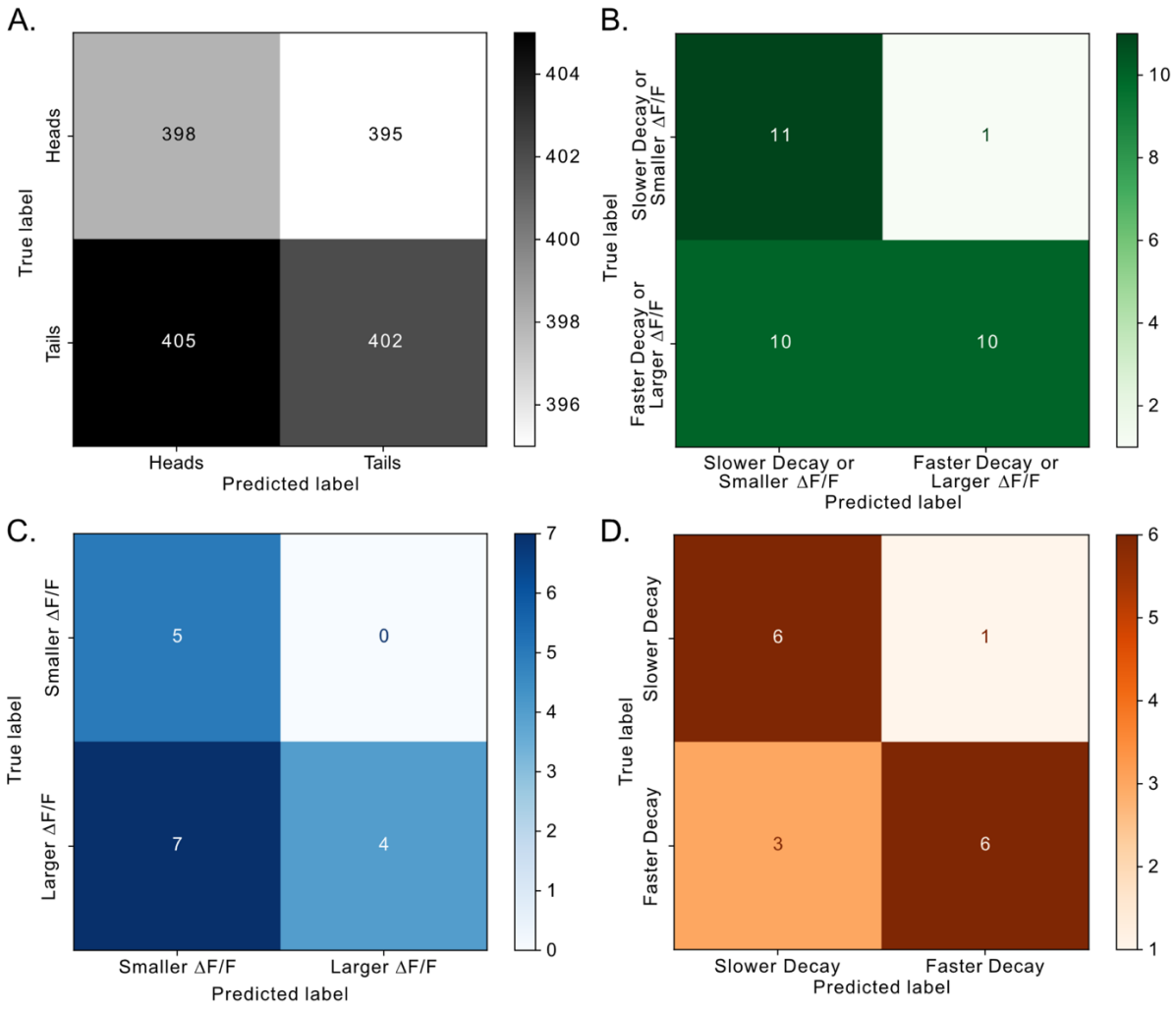
C. Decay values ($\boldsymbol{\tau}$, tau, *Eq.4*) obtained from each mutant of jGCaMP7s expressed in HEK293 cells and stimulated with ten μM acetylcholine. Heat mapping demonstrates the ensemble's prediction of the given mutation's performance *in vitro*. Mutations are sorted in order of the ensemble's predicted performance. Dotted line depicts mean performance of the base construct, jGCaMP7s. (n = number of cells quantified; bars depict mean + bootstrapped 95% ci).
D. Signal-to-noise ratio (SNR, *Eq.2*) of each mutant of jGCaMP7s expressed in HEK293 cells and stimulated with ten μM acetylcholine. Mutations are sorted in ascending order based on residue number and final residue composition. Dotted line depicts mean performance of the base construct, jGCaMP7s. (n = number of cells quantified; bars depict mean + bootstrapped 95% ci).
E. Performance score, consisting of the SNR/$\boldsymbol{\tau}$ (*Eq.2/Eq.4*), obtained from each mutant of jGCaMP7s expressed in HEK293 cells and stimulated with ten μM acetylcholine. Heat mapping highlights the highest-scoring mutants or those with high $\Delta F/F_0$ (%) responses and fast decay speeds. Mutations are sorted in ascending order based on residue number and final residue composition. (n = the number of cells quantified; bars depict mean + 95% bootstrapped ci).

2.2.3 In Vitro Performance of Ensemble Predictions

We tested mutations predicted by the ML-ensemble to enhance biophysical properties by stimulating HEK293 cells with acetylcholine[6,32,72]. This process activates calcium channels in the endoplasmic reticulum (ER) through G_q/IP_3 coupled pathways[73,74] (**Figure 2.6A**). We performed this step as a preliminary screening method to obtain an understanding of variant responses, prior to transitioning promising results to cultured neurons.

Mutations predicted to have a greater $\Delta F/F_0$ than jGCaMP7s (P303F, P303W, G392F, and G392W) all achieved >130% increase in $\Delta F/F_0$ over jGCaMP7s (**Figure 2.6B**, **Table 2.3A**). We also found three variants (L302G, L302H, and L302R) that satisfied their predicted decrease in $\Delta F/F_0$ response, with an average of 1.75x lower $\Delta F/F_0$ response (**Figure 2.6B, Table 2.3A**). The overall accuracy of the $\Delta F/F_0$ model is 0.56 (**Figure 2.7C, E**). The accuracy score is largely affected by the L317 mutants, which were predicted to have a decreased $\Delta F/F_0$ response but displayed the opposite characteristic *in vitro*. For example, all four L317 mutants achieved 2x greater $\Delta F/F_0$ than jGCaMP7s. Within the known variant library used for model training, we found

that all variants characterized that contained the mutations of 317N, 317E, 317K, or 317H saw almost a complete reduction of the $\Delta F/F_0$ capabilities (**Figure 2.8A, B**). Accordingly, we found that the L317H mutation in jGCaMP7f led to the predicted reduction of the $\Delta F/F_0$ (**Figure 2.8C**). This reflects findings in the Dana *et al.* 2019 dataset, in which variant 10.1035 (jGCaMP7f L317H) saw a 95% reduction in $\Delta F/F_0$ to 1AP stimuli compared to 10.9210 (jGCaMP7f)[64]. We speculate that this learned association is why the ensemble predicted mutations at L317 are detrimental to the sensor's $\Delta F/F_0$ response. Regardless, we identified four mutations (P303W, P303F, G392F, G392W) that displayed their predicted increase in $\Delta F/F_0$ as well as five mutations (A390Y, L302C, L302H, L302G, L302R) that displayed the predicted decrease in $\Delta F/F_0$.



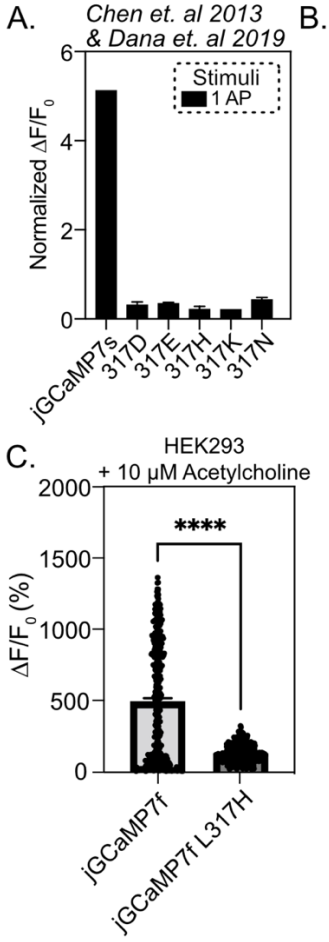
E.

Model	Accuracy	Precision
Random Chance (50/50)	0.5	0.5
Fluorescence + Kinetics	0.66	0.91
Fluorescence	0.56	1
Kinetics	0.75	0.86

Figure 2.7: Estimation of Model Accuracy with Acetylcholine Results
A. Confusion Matrix of 100 simulated coin flip experiments (each containing 16 samples).
B. Confusion matrix of all both models' predictions and acetylcholine performance.
C. Confusion matrix of all fluorescence model's predictions and acetylcholine performance.
D. Confusion matrix of all kinetics model's predictions and acetylcholine performance.
E. Quantification of each confusion matrices accuracy and precision.

The mutations that changed kinetics largely aligned with the ensemble predictions, with an accuracy score of 0.75 (**Figure 2.6C, Figure 2.7D, E**). Variants P303D, L317E, L317H, L317K,

L317N, G392F, and G392W were predicted to accelerate decay kinetics. Of these variants, 85% showed shorter decay times than jGCaMP7s, with L317K displaying a decay time that was 5x faster than jGCaMP7s (**Figure 2.6C, Table 2.3B**). Additionally, 71% of the variants predicted to decrease decay (L302C, L302D, L302G, L302H, L302R, A390R, A390Y) demonstrated the predicted behavior, with L302G exhibiting a decay time 2.18x longer than jGCaMP7s (**Figure 2.6C, Table 2.3B**).



B.

Index	52 AA	317 AA	374 AA	ΔF/F$_0$: 1AP
10.1473	V	L	M	5.12
10.915	A	D	M	0.38
10.1129	A	D	M	0.24
67	A	E	M	0.09
361	A	E	M	0.21
362	A	E	M	0.14
500	A	E	M	0.29
608	A	E	M	0.13
609	A	E	M	0.15
610	A	E	M	0.26
611	A	E	M	0.15
612	A	E	M	0.29
613	A	E	M	0.74
630	A	E	M	0.33
631	A	E	M	0.24
632	A	E	M	0.21
633	A	E	M	0.11
642	A	E	M	0.42
644	A	E	M	0.28
645	A	E	M	0.22
646	A	E	M	0.32
647	A	E	M	0.31
648	A	E	M	0.21
649	A	E	M	0.43
663	A	E	M	0.12
664	A	E	M	0.24
684	A	E	M	0.16
685	A	E	M	0.20
686	A	E	M	0.17
687	A	E	M	0.15
688	A	E	M	0.13
689	A	E	M	0.10
690	A	E	M	0.16
691	A	E	M	0.15
692	A	E	M	0.07
10.1083	A	E	M	0.61
10.1087	A	E	M	0.74

Index	52 AA	317 AA	374 AA	ΔF/F$_0$: 1AP
10.1088	A	E	M	0.65
10.1097	A	E	M	0.27
10.1098	A	E	M	0.64
10.1112	A	E	M	0.28
10.1113	A	E	M	0.76
10.1117	A	E	M	0.43
10.1118	A	E	M	0.43
10.1119	A	E	M	0.42
10.1124	A	E	M	0.49
10.1125	A	E	M	0.29
10.1126	A	E	M	0.34
10.1127	A	E	M	0.90
10.1128	A	E	M	0.32
10.1136	A	E	M	0.60
10.1206	A	E	M	0.26
10.1207	A	E	M	0.28
10.1212	A	E	M	0.16
10.1217	A	E	M	0.59
10.1227	A	E	M	0.74
10.1228	A	E	M	0.21
10.1229	A	E	M	0.51
10.124	A	E	M	0.14
10.1251	A	E	M	0.50
10.1253	A	E	M	0.32
10.1257	A	E	M	0.01
['GCaMP6f', 'GCaMP6f']	A	E	M	0.62
['68', '10.916']	A	E	M	0.20
['10.1175', '10.1182']	A	E	M	0.57
['10.1239', '10.1243']	A	E	M	0.38
['10.108', '10.123']	A	E	M	0.63
['10.1234', '10.1236']	A	E	M	0.55
['10.1241', '10.1258']	A	E	M	0.50
['10.1177', '10.1186']	A	E	M	0.36
10.1035	A	H	M	0.15
10.1259	A	H	M	0.28
10.92	A	K	M	0.21
10.923	A	N	M	0.38
10.1166	A	N	M	0.48

Figure 2.8: Assessment of Ensemble Predictions Compared to In Vitro Behaviors

A. Normalized fluorescent responses to indicated stimuli in cultured neurons, obtained from the Chen *et al. 2013* and Dana *et al. 2019* variant library (bars depict mean + SEM (if applicable).

B. Table of values contained in A. Values are derived from the Chen *et al. 2013* and Dana *et al. 2019* variant library, rows contain the identification number of the cataloged variant (Index), the variants' amino acid identities at residues 52, 317, and 374 (52 AA, 317 AA, 374 AA), and the normalized $\Delta F/F_0$ at 1 AP ($\Delta F/F_0$: 1AP). jGCaMP7s is included in the first row as it's Dana *et al.* 2019 derived identity (10.1473).

C. Max fluorescent responses obtained from listed variants expressed in HEK293 cells and stimulated with acetylcholine. (n = number of cells quantified; bars depict mean + SEM, **** = <0.0001 (unpaired t-test, parametric, two-tailed)). [jGCaMP7f = 497.6 (SEM: 22.15; n=353); jGCaMP7f L317H = 130.7 (SEM: 3.57; n=330)].

For subsequent experiments, we focused on mutations that increased $\Delta F/F_0$ and accelerated decay kinetics, as these biophysical characteristics could improve the detection of fast calcium signaling, such as those found in neurons firing APs. We found that the variants with large $\Delta F/F_0$ responses, including G392W, G392F, P303F, P303W, L317N, L317K, L317E, and L317H, maintained a signal-to-noise ratio (SNR, *Eq. 2*) 1.5x greater than jGCaMP7s (**Figure 2.6D, Table 2.3C**). To highlight variants with large $\Delta F/F_0$ and fast kinetics, we created a performance score by dividing SNR by the tau value (*Eq. 2/Eq. 4*) (**Figure 2.6E**). L317E, L317K, L317H, and L317N achieved performance scores on average 10.28x greater than jGCaMP7s (**Table 2.3D**). Among them, L317H had the highest performance score of 54.49 (a.u.), 14.23x greater than jGCaMP7s. Based on this assessment, we selected the jGCaMP7s L317H variant for further characterization and named it "ensemble-GCaMP" (*e*GCaMP). These *in vitro* results demonstrate that the ensemble could effectively predict sensor functionality, significantly reducing the experimental burden required to identify variants with desired biophysical properties.

Table 2.3: Descriptive Statistics of Ensemble Prediction Screen Results

A.

ΔF/F (%)			
Variant	Mean	95% CI	ROIs Measured (n)
L317N	1224.53	[1188.62, 1257.82]	608
L317E	1239.16	[1195.84, 1282.91]	562
L317K	436.49	[397.03, 477.46]	145
L317H	932.7	[897.68, 967.47]	626
L302D	519.48	[492.15, 549.41]	541
L302R	258.63	[248.38, 271.61]	769
L302G	322.53	[309.5, 335.76]	707
L302H	255.27	[246.09, 264.53]	851
L302C	276.18	[265.04, 287.55]	707
A390Y	149.82	[142.29, 157.56]	606
A390R	906.15	[884.21, 927.9]	882
P303D	620.06	[598.72, 641.78]	787
jGCaMP7s	489.93	[469.53, 510.23]	631
P303W	716.15	[691.97, 740.69]	635
P303F	889.13	[863.36, 914.67]	1072
G392F	932.41	[896.26, 969.28]	450
G392W	1119.84	[1085.01, 1154.64]	573

B.

Tau (s)			
Variant	Mean	95% CI	ROIs Measured (n)
A390Y	124.72	[116.65, 133.5]	606
L302C	181.63	[170.57, 193.2]	707
L302D	139.24	[127.55, 151.71]	541
L302R	100.29	[93.8, 107.2]	769
A390R	66.75	[64.23, 69.38]	882
L302G	218.24	[204.23, 232.75]	707
L302H	132.81	[127.34, 138.61]	851
P303F	118.52	[112.54, 125.0]	1072
P303W	94.61	[89.19, 100.68]	635
jGCaMP7s	99.6	[92.02, 107.83]	631
G392F	70.51	[65.82, 75.64]	450
G392W	65.17	[60.42, 70.23]	573
L317H	28.75	[25.54, 32.31]	626
P303D	113.24	[107.3, 119.51]	787
L317K	4.65	[4.26, 5.08]	145
L317N	28.94	[25.49, 33.0]	608
L317E	35.16	[31.7, 39.14]	562

C.

SNR (a.u.)			
Variant	Mean	95% CI	ROIs Measured (n)
jGCaMP7s	251.74	[234.83, 269.59]	631
L302C	138.67	[131.24, 146.11]	707
L302D	319.8	[297.64, 342.55]	541
L302G	428.67	[405.3, 452.19]	707
L302H	146.09	[136.89, 155.66]	851
L302R	98.1	[92.87, 104.14]	769
P303D	455.22	[429.31, 482.85]	787
P303F	357.71	[341.91, 373.36]	1072
P303W	343.9	[325.42, 363.0]	635
L317E	810.13	[753.73, 866.7]	562
L317K	133.79	[117.73, 151.03]	145
L317H	854.35	[808.94, 899.55]	626
L317N	586.23	[541.01, 636.35]	608
A390R	559.12	[533.85, 585.68]	882
A390Y	41.99	[39.03, 45.06]	606
G392F	471.71	[441.75, 502.39]	450
G392W	590.15	[557.76, 623.86]	573

D.

Performance Score (a.u.)			
Variant	Mean	95% CI	ROIs Measured (n)
jGCaMP7s	3.83	[3.51, 4.15]	631
L302C	1.22	[1.11, 1.34]	707
L302D	4.02	[3.6, 4.47]	541
L302G	3.65	[3.09, 4.45]	707
L302H	1.46	[1.34, 1.59]	851
L302R	1.68	[1.43, 2.06]	769
P303D	5.43	[5.03, 5.85]	787
P303F	4.42	[4.17, 4.69]	1072
P303W	5.02	[4.67, 5.38]	635
L317E	35.48	[32.71, 38.38]	562
L317K	32.71	[28.56, 37.26]	145
L317H	54.49	[51.17, 57.92]	626
L317N	34.89	[32.0, 37.85]	608
A390R	10.4	[9.84, 10.98]	882
A390Y	0.51	[0.46, 0.56]	606
G392F	9.31	[8.5, 10.13]	450
G392W	12.84	[11.83, 13.94]	573

Tables containing the information displayed in Figure 2.6**B,C,D,E**. Tables contain the construct (Variant), the mean response (**A.** $\Delta F/F_0$ (%), **B.** Tau (s), **C.** SNR (a.u), **D.** SNR/Tau (a.u.)), the 95% confidence interval (95% CI), and the number of samples (ROIs Measured (n)).

2.2.4 Combinatorial Mutations and Mutation Transfer Led to the Identification of eGCaMP^{+} and eGCaMP^{2+}

We introduced the 317H mutation into jGCaMP8f[64] to test if the beneficial effects could similarly alter divergent GCaMP iterations. Residue L317 in jGCaMP7s is located in a conserved

region of CaM and is equivalent to A289 in jGCaMP8f (**Figure 2.9A**). The A289H mutation on jGCaMP8f improved the $\Delta F/F_0$ response 4x over jGCaMP8f (**Figure 2.10A**). jGCaMP8f A289H also showed 36% faster decay than jGCaMP8f (**Figure 2.10B**). The fast decay kinetics combined with large $\Delta F/F_0$ responses provide a promising variant that we named "ensemble-GCaMP^{+}" (*e*GCaMP^{+}), which we advanced for further downstream testing.

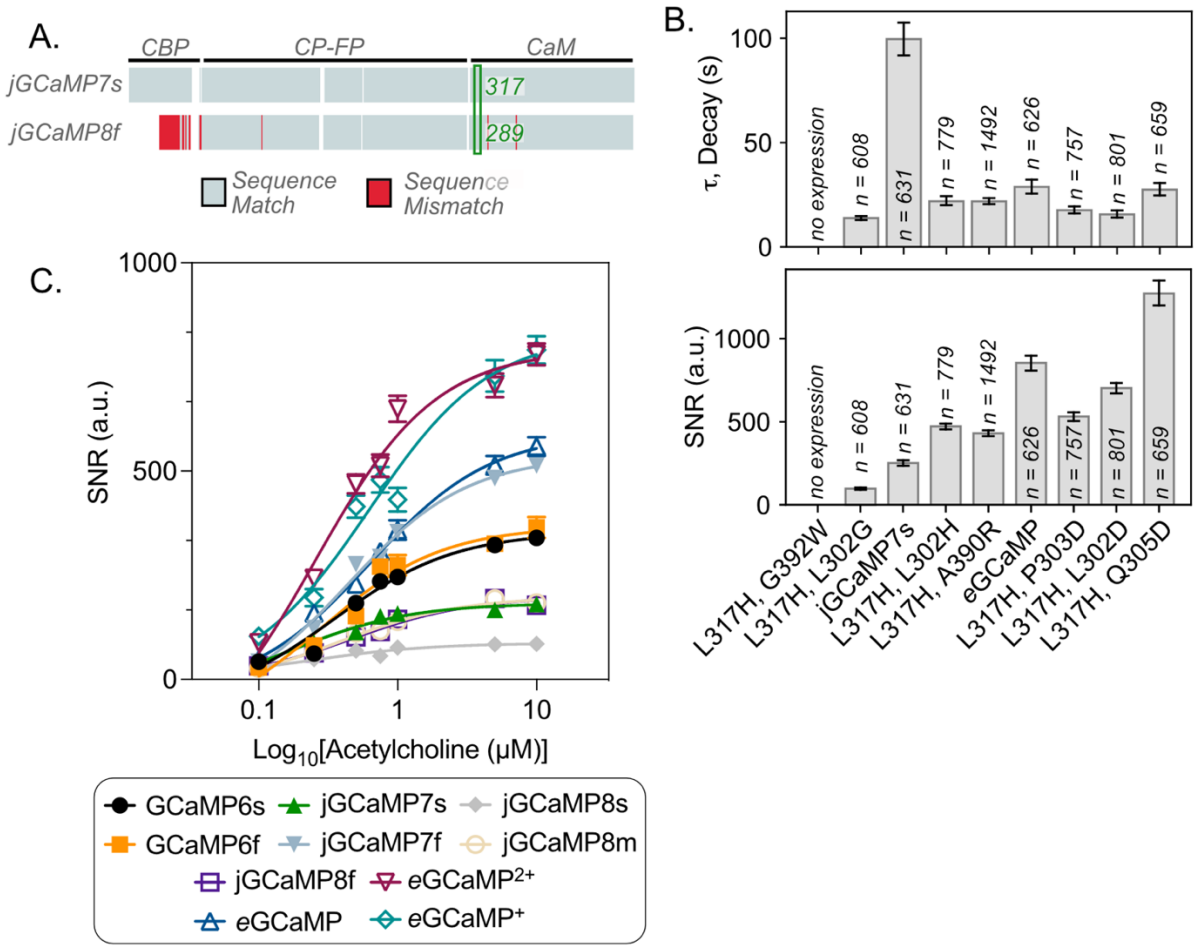


Figure 2.9: Combinatorial Mutation Biophysical Characteristics and Basis for Mutation Transfer

A. Sequence alignment of jGCaMP7s and jGCaMP8f. Light gray indicates identical sequence alignment, and red indicates sequence dissimilarities. Breaks indicate a missing sequence portion caused by additional sequence portions in other constructs. The physical location of the residue L317 in jGCaMP7s is designated with the green box and the residue number of the matching location is included as green text over the sequence. Text along the top of the sequence depicts the physical location in the GCaMP protein: CBP, CaM, or circularly permuted fluorescent protein (cpFP).

B. Decay values ($\boldsymbol{\tau}$, tau, *Eq.4*) obtained from each combinatorial mutant of jGCaMP7s expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Mutations are sorted according to $\Delta F/F_0$ performance (n = the number of cells quantified; bars depict mean + bootstrapped 95% ci). Signal-to-noise ratio (SNR, *Eq.4*) obtained from each combinatorial mutant of jGCaMP7s expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Mutations are sorted according to $\Delta F/F_0$ performance (n = the number of cells quantified; bars depict mean + bootstrapped 95% ci).
C. Signal-to-noise ratio (SNR, *Eq.4*) of indicated variant, expressed in HEK293 cells and stimulated with different acetylcholine concentrations (x-axis). Plotted points indicate the mean SNR response for each variant to indicated stimuli, and error bars display the SEM. The solid line depicts the non-linear fit of scatter data.

Next, we tested a select combination of additional mutations on *e*GCaMP. For example, we chose the jGCaMP7s variants L302D, P303D, A390R, and G392W for their increased $\Delta F/F_0$ *in vitro* (**Figure 2.6B**). Other mutants were selected based on their locations. Namely, L302 and P303 are key functional residues in the linker between cpGFP and CaM[6,75](**Figure 2.10C**). Residue G392 forms a hydrogen bond with residue G398, which lies in one of the EF-hand domains and has been previously observed to influence the Ca^{2+} affinity[6,76] (**Figure 2.5D**), and A390 lies on the interaction face between CaM and CBP (**Figure 2.10D**). We tested Q305 due to its proximity to the linker residues (**Figure 2.10C**), hydrogen bonding interactions with Y380 (**Figure 2.5E**), and prevalence in the impactful residues for kinetics (**Figure 2.4D**).

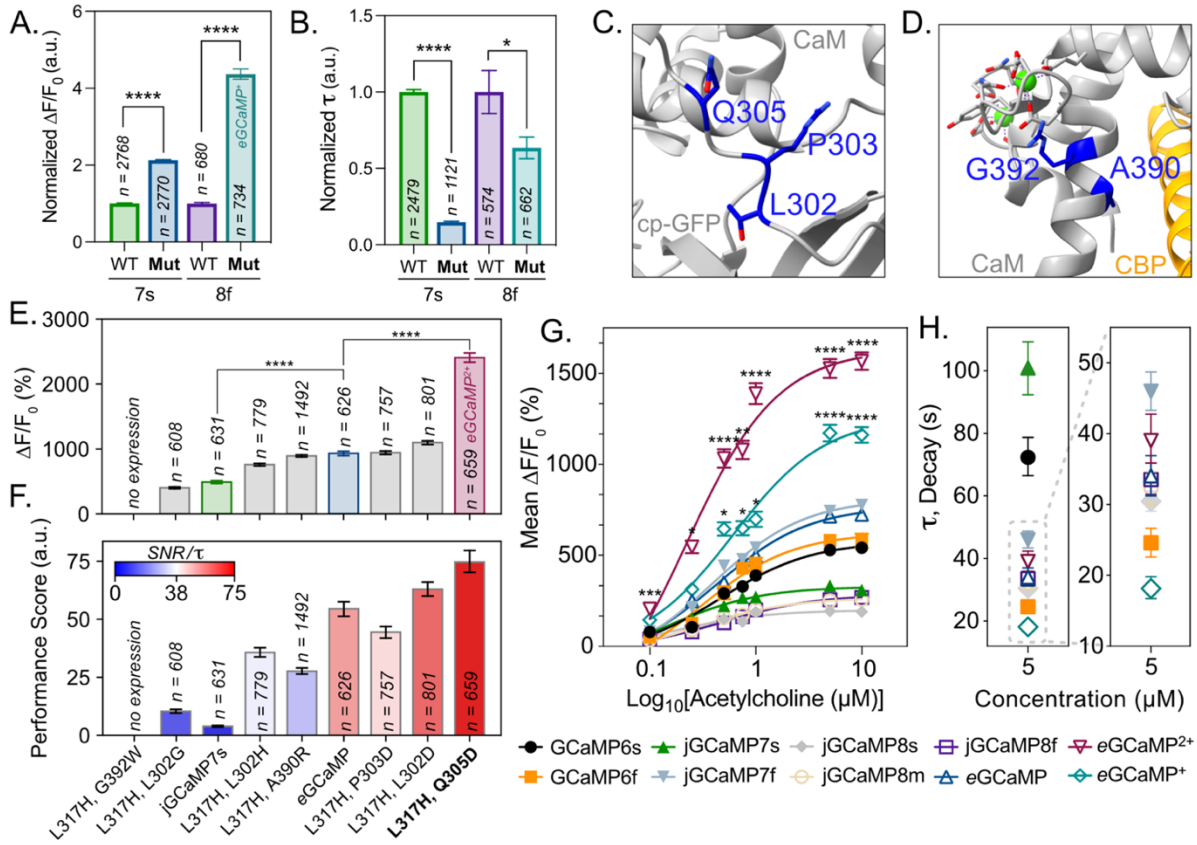


Figure 2.10: Mutation Transfer and Combinatorial Mutation For The Identification of eGCaMP^{+} and eGCaMP^{2+}

A. Max fluorescent responses (*Eq.1*) obtained from each variant indicated on the x-axis, expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Wild Type (WT) indicates the parent construct of either jGCaMP7s (7s) or jGCaMP8f (8f). Mutation (Mut) indicates the parental construct with the addition of L317H in jGCaMP7s and A289H in jGCaMP8f. Each parental/variant pair is normalized to the base construct mean = 1.0 (n = the number of cells quantified; bars depict mean + SEM; **** = <0.0001 (unpaired t-test, two-tailed)). jGCaMP8f A289H is called *e*GCaMP^{+} in Figure 2.10**G.** and 2.10**H.**

B. Decay values (***τ***, tau, *Eq.4*) obtained from each variant indicated on the x-axis, expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Wild Type (WT) indicates the parent construct of either jGCaMP7s (7s) or jGCaMP8f (8f). Mutation (Mut) indicates the parental construct with the addition of L317H in jGCaMP7s and A289H in jGCaMP8f. Each parental/variant pair is normalized to the base construct mean = 1.0 (n = the number of cells quantified; bars depict mean + SEM; **** = <0.0001 (unpaired t-test, two-tailed)). jGCaMP8f A289H is called *e*GCaMP^{+} in Figure 2.10**G.** and 2.10**H.**

C. Crystal structure of GCaMP3-D380Y (RCSB: 3SG3, gray) with Q305 and linker residues P303 and L302 colored in dark blue with sidechains visible. CaM and cpGFP labels are included to orient linker locations.

D. Crystal structure of GCaMP3-D380Y (RCSB: 3SG3, gray) with A390 and G392 colored dark blue with sidechains visible. Bound Ca^{2+} (green spheres) in the EF-Hand motifs and the CBP (orange) are included.

E. Max fluorescent responses (*Eq.1*) obtained from each combinatorial variant of jGCaMP7s expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Mutations are sorted in order of $\Delta F/F_0$ performance and identified on the x-axis of 2.10**D**. (n = the number of cells quantified; bars depict mean + bootstrapped 95% ci[77]; **** = <0.0001 (unpaired t-test)).
F. Performance score, consisting of the SNR/$\boldsymbol{\tau}$ (*Eq.2/Eq.4*), obtained from each combinatorial variant of jGCaMP7s expressed in HEK293 cells and stimulated with 10 μM acetylcholine. Mutations are sorted in order of $\Delta F/F_0$ performance. (n = number of cells quantified; bars depict mean + bootstrapped 95% ci) jGCaMP7s L317H Q305D is called *e*GCaMP^{2+} in Figure 2.10**G.** and 2.10**H.**
G. Fluorescent responses ($\Delta F/F_0$, *Eq. 1*) of indicated variant, expressed in HEK293 cells and stimulated with different acetylcholine concentrations (x-axis). Plotted points indicate the mean $\Delta F/F_0$ response for each variant to indicated stimuli, and error bars display the SEM. The solid line depicts the non-linear fit of scatter data. Additional information on plotted points is included in Supplementary Table 3. (* = <0.05; ** < 0.01; *** < 0.001; **** < 0.0001 (unpaired t-test between indicated variant and jGCaMP7f)).
H. Kinetic decay ($\boldsymbol{\tau}$, tau, *Eq.4*) of the indicated variant, expressed in HEK293 cells and stimulated with 5 μM acetylcholine. Plotted points indicate the mean $\boldsymbol{\tau}$ for each variant to the indicated stimuli, and error bars display the SEM. Additional information on plotted points is included in Supplementary Table 3.

All combinations, except for L317H/G392W, led to functional proteins (**Figure 2.10E, F; Figure 2.9B**). On average, all variants exhibited decay times 5.0x faster than jGCaMP7s (**Figure 2.9B**; **Table 2.4B**). Within the tested variants, 50% displayed equal or improved $\Delta F/F_0$ response to that of *e*GCaMP (**Figure 2.10E**; **Table 2.4A**). We observed the largest $\Delta F/F_0$ in the L317H/Q305D mutation, with an almost 2.5-fold increase in $\Delta F/F_0$ over *e*GCaMP and a 5-fold increase over jGCaMP7s (**Figure 2.10E**; **Table 2.4A**). The variant also achieved the highest performance score (i.e., large SNR, fast decay) of all variants, a 1.36x fold increase over *e*GCaMP (**Figure 2.10F**; **Figure 2.9B**; **Table 2.4D**). We chose the jGCaMP7s L317H/Q305D for further characterization and named it "ensemble-GCaMP^{2+}" (*e*GCaMP^{2+}).

Table 2.4: Descriptive Statistics of Combinatorial Mutation Screen Results

A.

ΔF/F (%)				
Variant	Acetylcholine Stimuli (μM)	Mean	95% CI	ROIs Measured (n)
L317H, L302G	10	401.8	[387.48, 416.08]	608
jGCaMP7s	10	489.93	[469.11, 510.71]	631
L317H, L302H	10	758.06	[738.7, 777.22]	779
L317H, A390R	10	892.11	[875.99, 908.27]	1492
e GCaMP (L317H)	10	932.7	[897.84, 967.47]	626
L317H, P303D	10	942.58	[916.89, 968.52]	757
L317H, L302D	10	1098.63	[1070.71, 1126.54]	801
L317H, Q305D	10	2407	[2335.38, 2478.47]	659

B.

Tau (s)				
Variant	Acetylcholine Stimuli (μM)	Mean	95% CI	ROIs Measured (n)
L317H, L302G	10	13.82	[12.91, 14.78]	608
jGCaMP7s	10	99.6	[92.15, 107.49]	631
L317H, L302H	10	21.86	[19.91, 24.12]	779
L317H, A390R	10	21.83	[20.57, 23.24]	1492
e GCaMP (L317H)	10	28.75	[25.57, 32.35]	626
L317H, P303D	10	17.61	[16.02, 19.49]	757
L317H, L302D	10	15.59	[14.03, 17.49]	801
L317H, Q305D	10	27.39	[24.62, 30.51]	659

C.

SNR (a.u.)				
Variant	Acetylcholine Stimuli (μM)	Mean	95% CI	ROIs Measured (n)
L317H, L302G	10	97.7	[91.52, 104.22]	608
jGCaMP7s	10	251.74	[234.96, 269.24]	631
L317H, L302H	10	471.92	[454.95, 489.27]	779
L317H, A390R	10	430.55	[413.14, 448.84]	1492
e GCaMP (L317H)	10	854.35	[809.76, 901.75]	626
L317H, P303D	10	531.58	[504.94, 558.54]	757
L317H, L302D	10	702.85	[672.35, 733.8]	801
L317H, Q305D	10	1271.48	[1200.07, 1343.99]	659

D.

SNR/Tau (a.u.)				
Variant	Acetylcholine Stimuli (μM)	Mean	95% CI	ROIs Measured (n)
L317H, L302G	10	10.35	[9.56, 11.17]	608
jGCaMP7s	10	3.83	[3.51, 4.15]	631
L317H, L302H	10	35.66	[33.73, 37.61]	779
L317H, A390R	10	27.58	[26.26, 29.05]	1492
e GCaMP (L317H)	10	54.49	[51.21, 57.96]	626
L317H, P303D	10	44.43	[41.78, 47.16]	757
L317H, L302D	10	62.97	[59.92, 66.04]	801
L317H, Q305D	10	74.62	[69.97, 79.53]	659

Tables containing the information displayed in Figures 2.10**E,F** and Figure 2.9**B,C**. Tables contain the construct (Variant), the concentration of the acetylcholine stimulus (Acetylcholine Stimuli (μM)), the mean response ((**A.** $\Delta F/F_0$ (%), **B.** Tau (s), **C.** SNR (a.u), **D.** SNR/Tau (a.u.)), the 95% confidence interval (95% CI), and the number of samples (ROIs Measured (n)).

Table 2.5: Descriptive Statistics of Acetylcholine Concentration Curve Results

A.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
10	GCaMP6s	542.06	16.3	1218
10	GCaMP6f	580.06	19.75	776
10	jGCaMP7s	309.67	8.69	1401
10	jGCaMP7f	776.22	8.83	1883
10	jGCaMP8s	189.90	4.34	1129
10	jGCaMP8m	257.06	4.63	1514
10	jGCaMP8f	268.50	5.21	1288
10	eGCaMP	724.49	14.36	1618
10	eGCaMP2+	1567.52	24.54	1924
10	eGCaMP+	1162.09	21.91	1467

B.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
5	GCaMP6s	529.34	14.26	1275
5	GCaMP6f	582.71	14.46	1355
5	jGCaMP7s	332.85	10.67	910
5	jGCaMP7f	743.66	7.57	1867
5	jGCaMP8s	194.31	4.17	1143
5	jGCaMP8m	250.46	5.08	1203
5	jGCaMP8f	256.31	4.69	1358
5	eGCaMP	715.41	14.6	1527
5	eGCaMP2+	1526.99	26.47	1448
5	eGCaMP+	1171.96	23.31	1161

C.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
1	GCaMP6s	389.18	18.02	785
1	GCaMP6f	447.79	19.66	584
1	jGCaMP7s	271.94	11.82	686
1	jGCaMP7f	542.35	9.69	1524
1	jGCaMP8s	188.81	5.06	749
1	jGCaMP8m	203.67	6.75	760
1	jGCaMP8f	200.35	5.39	1065
1	eGCaMP	497.50	15.62	1000
1	eGCaMP2+	1389.23	29.65	1125
1	eGCaMP+	698.07	21.61	996

D.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
0.75	GCaMP6s	329.41	15.49	804
0.75	GCaMP6f	429.50	14.13	1016
0.75	jGCaMP7s	268.25	9.28	1006
0.75	jGCaMP7f	477.93	9.81	1361
0.75	jGCaMP8s	135.62	4.28	805
0.75	jGCaMP8m	173.83	5.27	983
0.75	jGCaMP8f	164.20	3.92	1671
0.75	eGCaMP	474.77	11.93	1666
0.75	eGCaMP2+	1079.67	26	1278
0.75	eGCaMP+	648.48	19.95	1034

E.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
0.5	GCaMP6s	290.84	14.57	826
0.5	GCaMP6f	293.70	18.37	657
0.5	jGCaMP7s	223.09	11.05	846
0.5	jGCaMP7f	442.49	9.13	1574
0.5	jGCaMP8s	151.18	6.43	462
0.5	jGCaMP8m	157.39	5.01	1035
0.5	jGCaMP8f	128.58	4.69	1100
0.5	eGCaMP	350.12	12.21	1175
0.5	eGCaMP2+	1033.06	25.58	1353
0.5	eGCaMP+	644.64	19.25	1121

F.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
0.25	GCaMP6s	105.54	15.68	302
0.25	GCaMP6f	128.22	12.6	491
0.25	jGCaMP7s	156.55	11.77	453
0.25	jGCaMP7f	214.02	7.26	1577
0.25	jGCaMP8s	102.76	5.92	477
0.25	jGCaMP8m	91.13	4.29	935
0.25	jGCaMP8f	79.28	4.02	934
0.25	eGCaMP	249.34	13.68	784
0.25	eGCaMP2+	546.50	18.21	1471
0.25	eGCaMP+	313.19	15.23	976

G.

Concentration (μM)	Variant	Mean $\Delta F/F_0$ (%)	SEM	Samples (n)
0.1	GCaMP6s	78.84	14.76	474
0.1	GCaMP6f	50.79	13.63	356
0.1	jGCaMP7s	77.28	9.16	381
0.1	jGCaMP7f	81.91	4.79	1385
0.1	jGCaMP8s	58.18	5.5	300
0.1	jGCaMP8m	35.94	3.16	626
0.1	jGCaMP8f	40.50	3.1	937
0.1	eGCaMP	71.75	6.5	896
0.1	eGCaMP2+	205.69	14.34	807
0.1	eGCaMP+	145.32	10.61	808

H.

Concentration (μM)	Variant	τ, Decay (s)	SEM	Samples (n)
5	GCaMP6s	72.31	3.06	940
5	GCaMP6f	24.58	1.03	1357
5	jGCaMP7s	100.9	4.27	676
5	jGCaMP7f	45.97	1.42	2088
5	jGCaMP8s	30.44	0.69	1379
5	jGCaMP8m	32.28	1.19	1325
5	jGCaMP8f	33.52	1.13	1457
5	eGCaMP	34.02	1.49	1564
5	eGCaMP2+	39.02	1.74	1509
5	eGCaMP+	18.13	0.79	1297

Tables containing the information displayed in Figures 2.10**G,H**. (**A.,B.,C.,D.,E.,F.,G.**) Tables contain the concentration of the acetylcholine stimulus (Concentration (μM)), the construct (Variant), the mean response (Mean $\Delta F/F_0$ (%)), the standard error of the mean (SEM), and the number of samples (Samples (n)). (**H.**) The table contains the concentration of the acetylcholine stimulus (Concentration (μM)), the construct (Variant), the speed of off-decay (Tau Off (s)), the standard error of the mean (SEM), and the number of samples (Samples (n)).

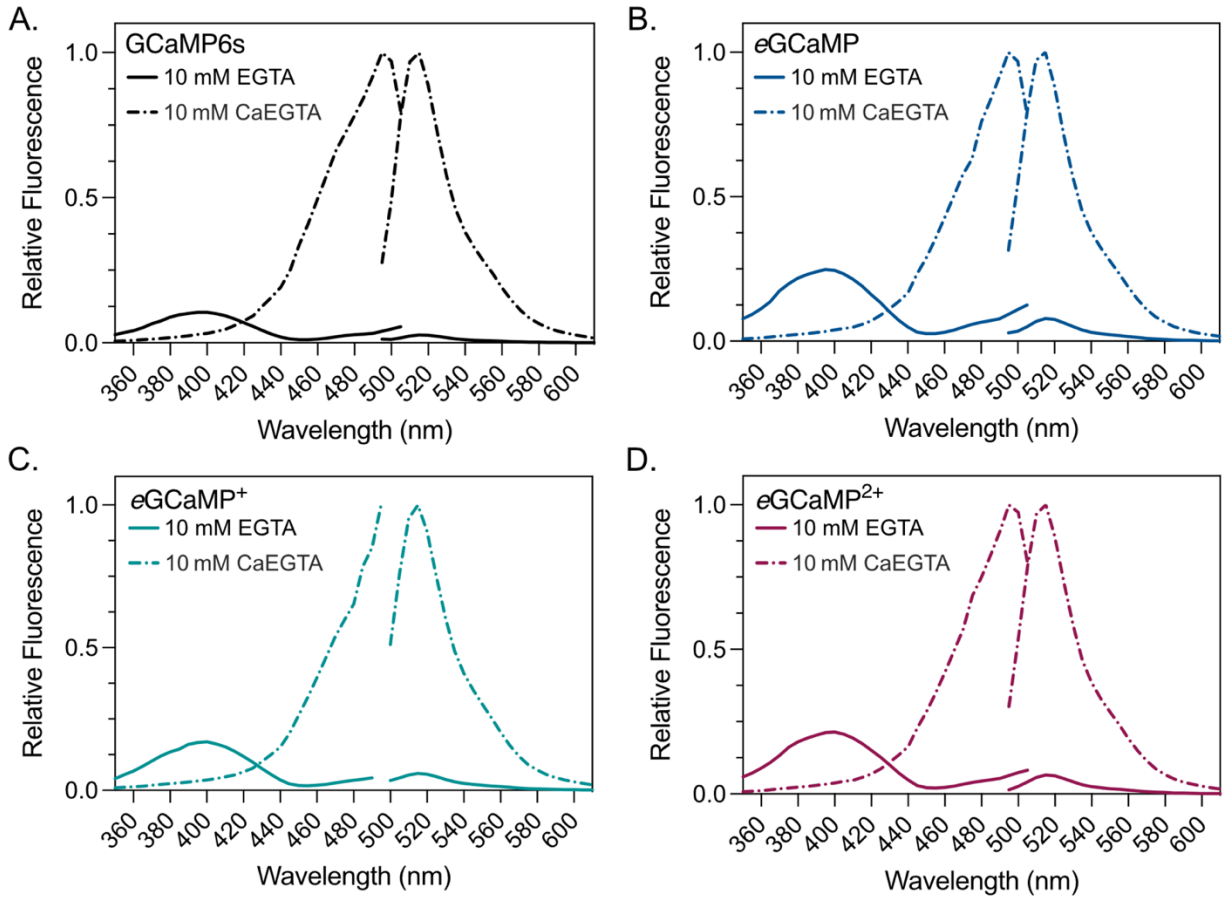


Figure 2.11: Excitation and Emission Spectra of eGCaMP Sensors

A. Purified GCaMP6s protein diluted into buffer containing either 10 mM EGTA or 10 mM CaEGTA. Emission spectra were calculated using a fixed excitation at 450 nm and excitation spectra were calculated using a fixed emission at 520 nm.
B. Purified *e*GCaMP protein diluted into buffer containing either 10 mM EGTA or 10 mM CaEGTA. Emission spectra were calculated using a fixed excitation at 450 nm and excitation spectra were calculated using a fixed emission at 520 nm.
C. Purified *e*GCaMP^{+} protein diluted into buffer containing either 10 mM EGTA or 10 mM CaEGTA. Emission spectra were calculated using a fixed excitation at 450 nm and excitation spectra were calculated using a fixed emission at 520 nm.
D. Purified *e*GCaMP^{2+} protein diluted into buffer containing either 10 mM EGTA or 10 mM CaEGTA. Emission spectra were calculated using a fixed excitation at 450 nm and excitation spectra were calculated using a fixed emission at 520 nm.

We benchmarked the biophysical and photophysical properties of *e*GCaMP, *e*GCaMP^{2+}, and *e*GCaMP^{+} against published variants including widely used constructs such as GCaMP6s, GCaMP6f, jGCaMP7s, jGCaMP7f, jGCaMP8s, jGCaMP8m, and jGCaMP8f[7,33,64]. The excitation and emission spectra of the *e*GCaMP variants remains unchanged from the previously published

GCaMPs, with excitation peaks at ~495 nm and emission peaks at ~515 nm (**Figure 2.11A-D**). Using c-terminally red fluorescent protein (RFP) tagged constructs, we found that *e*GCaMP, *e*GCaMP^{+}, and *e*GCaMP^{2+} maintained higher dynamic ranges and SNRs but have lower baseline fluorescence than GCaMP6s, jGCaMP7s, and jGCaMP8f (**Figure 2.10F**; **Figure 2.12A-D**). In the acetylcholine concentration curve, we found that the three ensemble variants demonstrated both impressive $\Delta F/F_0$ responses and signal to noise ratios compared to previously published GCaMPs (**Figure 2.10G**; **Figure 2.9C**). At every tested concentration, *e*GCaMP^{+} and *e*GCaMP^{2+} maintained larger $\Delta F/F_0$s than all previously published variants (**Figure 2.10G**; **Table 2.5**). For example, eGCaMP^{2+} achieved 2.5x greater $\Delta F/F_0$s at 0.1μM acetylcholine than the highest performing published variant, with decay times comparable to jGCaMP7f (**Figure 2.10G, H**; **Table 2.5**). Additionally, the decay time of *e*GCaMP^{+} was the fastest of all tested variants (46% faster than jGCaMP8f, its parental construct) while the maximum $\Delta F/F_0$, was second only to *e*GCaMP^{2+} (**Figure 2.10G, H**; **Table 2.5**). *e*GCaMP achieved a $\Delta F/F_0$ close to jGCaMP7f but with a 26% faster decay (**Figure 2.10G, H**; **Table 2.5**). Using purified proteins, we found that the *e*GCaMP and *e*GCaMP^{2+} variants achieved similar K_d's to those published for jGCaMP8f[78] (**Table 2.6**). *e*GCaMP^{+} displayed a K_d shifted to the micromolar range, which is consistent with previously published studies finding a tradeoff between sensitivity and kinetics[7,33,78] (**Table 2.6**). The *e*GCaMPs also demonstrated slightly diminished saturated extinction coefficients compared to GCaMP6f, but all displayed larger saturated quantum yields (**Table 2.6**).

Table 2.6: Photophysical Properties of Purified eGCaMP Proteins

Sensor	Kd (nM)	Hill Coefficient	$\varepsilon_{Saturated}$ (x1000) (M^{-1} cm^{-1})	$\varphi_{Saturated}$
GCaMP6s	120.8 [110.6, 132.2]	2.014 [1.716, 2.387]	N/A	N/A
GCaMP6f	291.3 [256.4, 333.1]	1.857 [1.544, 2.261]	65.276	0.6
jGCaMP7s	46.2 [39.3, 53.7]	2.138 [1.596, 1.918]	N/A	N/A
eGCaMP	354.8 [262.8, 516.4]	1.761 [1.087,3.339]	62.726	0.68
eGCaMP2+	358.7 [310.4, 418.8]	1.925 [1.540, 2.461]	60.070	0.72
eGCaMP+	1885 [1.082, 34.02]	0.9976 [0.4875, 1.871]	58.988	0.63

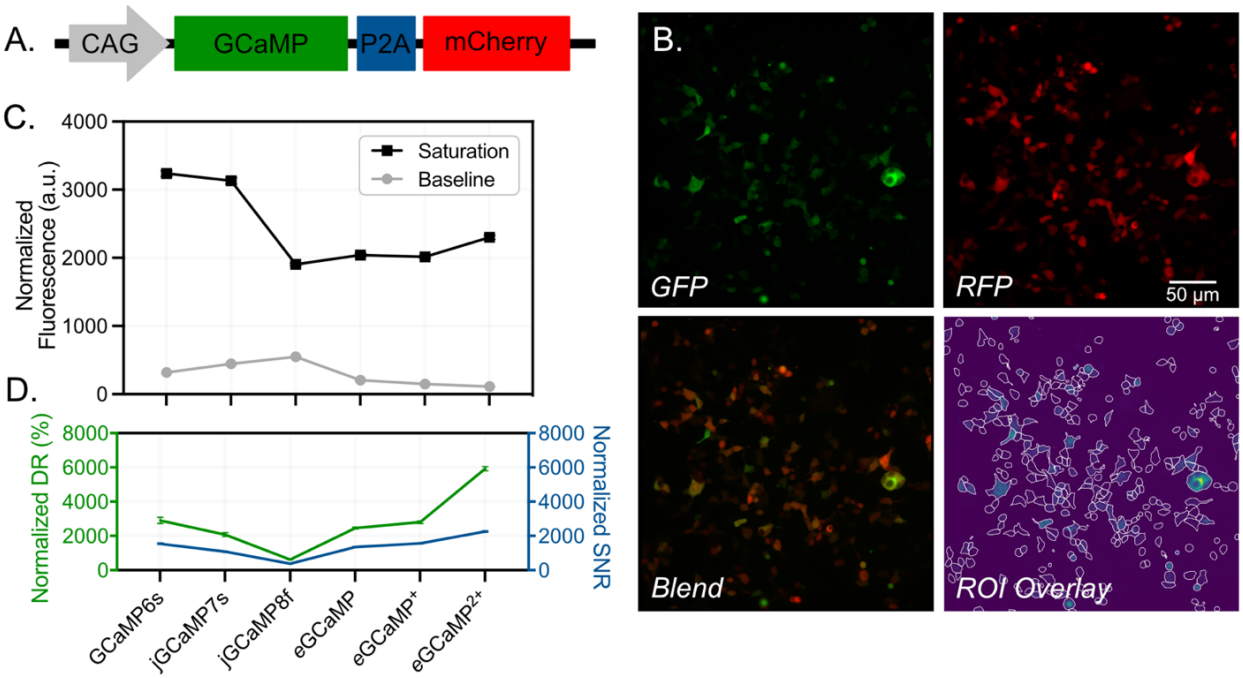


Figure 2.12: Ratiometric analysis of baseline fluorescence for eGCaMP, eGCaMP^{+}, and eGCaMP^{2+}

A. GCaMP variants GCaMP6s, jGCaMP7s, jGCaMP8f, *e*GCaMP, *e*GCaMP^{+}, *e*GCaMP^{2+} were transformed into a CAG driven vector, with a mCherry control fluorophore added after a self-cleaving motif (P2A) within the reading frame.

B. Representative images taken with 488 nm wavelength excitation (GFP), 585nm wavelength excitation (RFP), an overlap of the two channels depicting the ratio of GFP intensity to RFP intensity (Blend), and the Cellpose derived ROIs used for analysis (ROI overlay). The scale bar depicts 50µm.

C. Normalized fluorescence intensity of each ratiometric variant (x-axis shared with D.) at baseline (gray) and saturation conditions (black). Scatter points depict mean value and error bars depict SEM.

D. Normalized dynamic range (DR, green, left y-axis) and signal-to-noise ratio (SNR, blue, right y-axis) for each ratiometric variant (x-axis). Scatter points depict mean value and error bars depict SEM.

2.2.5 eGCaMP, eGCaMP^{+}, and eGCaMP^{2+} Performance in Primary Neurons

Next, we tested the *e*GCaMP variants in cultured primary rat cortical neurons stimulated using extracellular electrical fields[7,33,79]. We included these previously published variants to benchmark the responses from our sensors under identical experimental conditions. *e*GCaMP^{2+} displayed a $\Delta F/F_0$ of 10.1% in response to 1 AP stimuli, similar to amplitudes obtained by jGCaMP8f (**Figure 2.13A**; **Table 2.7A**). *e*GCaMP^{2+}*'s* impressive response amplitudes became more apparent with increasing numbers of elicited action potentials. At 10 AP, *e*GCaMP^{2+} achieved 2.34x greater $\Delta F/F_0$ response than jGCaMP7s (the next closest variant), and at 80 AP stimuli, *e*GCaMP^{2+} achieved 1.82x greater $\Delta F/F_0$ response than GCaMP6s (the next closest variant) (**Figure 2.13B, C**; **Table 2.7B, C**). These results were recapitulated in saturation responses, where the average $\Delta F/F_0$ response to 40 mM KCl was 1938% for cells expressing *e*GCaMP^{2+} (**Figure 2.13E**; **Table 2.7E**). This $\Delta F/F_0$ is 2x greater than those observed in GCaMP6s, the sensor that saw the second-greatest responses to 40 mM KCl (**Figure 2.13E**; **Table 2.7E**). While the KCl saturation responses were quantified using the cell body, the proximal projections in *e*GCaMP^{2+} similarly maintained >1000% $\Delta F/F_0$ increases (**Figure 2.13F**). At 80 AP trains, both *e*GCaMP and *e*GCaMP^{+} achieved higher $\Delta F/F_0$ response amplitudes than the previously published fast variants GCaMP6f and jGCaMP8f (**Figure 2.13C**; **Table 2.7C**). These results are compounded by both *e*GCaMP and *e*GCaMP^{+} achieving 10 AP half decay times ($\tau_{1/2}$) of 1.17s and 0.74s for each variant, respectively. These decay times are faster than jGCaMP8f's, whose 10 AP half decay time was 1.49s (**Figure 2.13D**; **Table 2.7D**). Furthermore, *e*GCaMP

decayed 8x faster than jGCaMP7s, highlighting the ability of the ensemble to correctly predict the single point mutation's functional effect (**Figure 2.13D**; **Table 7**).

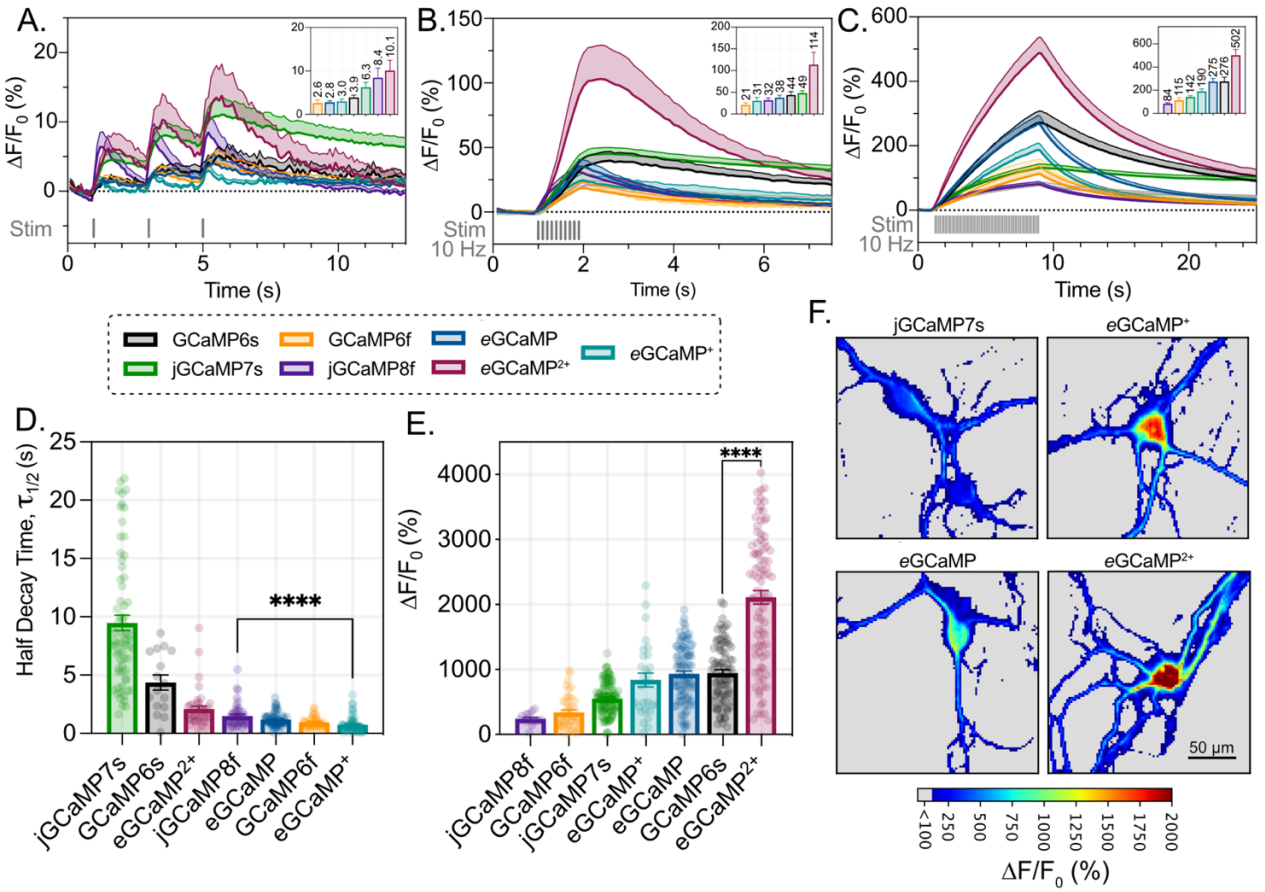


Figure 2.13: eGCaMP, eGCaMP^{+}, and eGCaMP^{2+} Fluorescence and Kinetics Characteristics in Primary Neurons

A. $\Delta F/F_0$ (%) recordings of each variant to 1 AP stimuli applied at 0.5 Hz over 6 seconds (lines depict mean, shading depicts SEM). The applied stimulus is shown in gray. Graph inset displays max $\Delta F/F_0$ (%) of each variant to first applied AP (mean + SEM, above-bar annotations = mean).

B. $\Delta F/F_0$ (%) recordings of each variant to 10 AP stimuli applied at 10 Hz over 1 second (lines depict mean, shading depicts SEM). The applied stimulus is shown in gray. Graph inset displays max $\Delta F/F_0$ (%) of each variant to 10 AP stimulus (mean + SEM, above-bar annotations = mean).

C. $\Delta F/F_0$ (%) recordings of each variant to 80 AP stimuli applied at 10 Hz over 8 seconds (lines depict mean, shading depicts SEM). The applied stimulus is shown in gray. Graph inset displays max $\Delta F/F_0$ (%) of each variant to 80 AP stimulus (mean + SEM, above-bar annotations = mean).

D. Half decay time values after 10 AP stimuli, scatter depicts neurons quantified. (bars depict mean + SEM; * = 0.045 (Unpaired t-test, Two-tailed)).

E. Maximum $\Delta F/F_0$ (%) achieved after stimulation with 40 mM KCl. (bars depict mean + SEM; **** = < 0.0001 (Unpaired t-test, Two-tailed)).

F. Representative images of maximal fluorescent response to 40 mM KCl stimulation variant indicated above image. Heat Mapping displays $\Delta F/F_0$ (%) achieved by each pixel. (Scale bar = 50 μm).

Table 2.7: Descriptive Statistics of Primary Neuron Recording

A.

1 AP							
Construct	GCaMP6s	GCaMP6f	jGCaMP7s	jGCaMP8f	*e*GCaMP	*e*GCaMP^{+}	*e*GCaMP^{2+}
Number of values	33	48	102	72	56	49	47
Mean	3.911	2.626	6.252	8.438	2.783	3.026	10.1
Std. Deviation	3.197	5.021	11.51	18.88	2.845	3.702	16.25
Std. Error of Mean	0.5564	0.7247	1.139	2.225	0.3802	0.5289	2.371

B.

10 AP							
Construct	GCaMP6s	GCaMP6f	jGCaMP7s	jGCaMP8f	*e*GCaMP	*e*GCaMP^{+}	*e*GCaMP^{2+}
Number of values	49	55	134	50	82	77	53
Mean	43.77	20.99	48.53	31.94	37.81	30.54	113.8
Std. Deviation	53.41	34.86	57.84	35.49	50.59	64.68	207.4
Std. Error of Mean	7.63	4.701	4.996	5.019	5.587	7.371	28.49

C.

80 AP							
Construct	GCaMP6s	GCaMP6f	jGCaMP7s	jGCaMP8f	*e*GCaMP	*e*GCaMP^{+}	*e*GCaMP^{2+}
Number of values	49	63	111	72	88	86	54
Mean	276.1	114.6	142.4	84.28	275.3	190.4	502.5
Std. Deviation	252.1	173.6	126.7	64.44	243.7	193.1	370.1
Std. Error of Mean	36.01	21.87	12.03	7.594	25.97	20.82	50.37

D.

10 AP Decay							
Construct	GCaMP6s	GCaMP6f	jGCaMP7s	jGCaMP8f	*e*GCaMP	*e*GCaMP^{+}	*e*GCaMP^{2+}
Number of values	16	44	70	47	75	62	43
Mean	4.37	0.9547	9.479	1.49	1.168	0.7355	2.099
Std. Deviation	2.583	0.4496	5.487	0.9786	0.5577	0.6476	1.595
Std. Error of Mean	0.6457	0.06778	0.6558	0.1427	0.0644	0.08225	0.2433

E.

40 mM KCl							
Construct	GCaMP6s	GCaMP6f	jGCaMP7s	jGCaMP8f	*e*GCaMP	*e*GCaMP^{+}	*e*GCaMP^{2+}
Number of values	82	44	121	15	96	29	92
Mean	940.5	333.3	546.4	233.2	928.8	833	2110
Std. Deviation	465.3	253.7	239.1	89.34	445.1	572.1	1009
Std. Error of Mean	51.39	38.24	21.74	23.07	45.43	106.2	105.2

Tables containing the information displayed in Figures 2.13**A,B,C,D**. Tables **A.**, **B.**, and **C.** contain the construct (Construct), number of samples (Number of Values), the mean $\Delta F/F_0$ (%) response (Mean), the standard deviation (Std. Deviation), and the standard error of the mean (Std. Error of Mean) at 1 AP, 10 APs and 80 APs, respectively. Table **D.** contains the construct (Construct), number of samples (Number of Values), the mean half decay time (s) (Mean), the standard deviation (Std. Deviation), and the standard error of the mean (Std. Error of Mean) after 10 AP stimuli. Table **E.** contains the construct (Construct), number of samples (Number of Values), the

mean $\Delta F/F_0$ (%) response (Mean), the standard deviation (Std. Deviation), and the standard error of the mean (Std. Error of Mean) after 40 mM KCl Stimulus.

2.2.6 eGCaMP$^+$ and eGCaMP^{2+} Performance in vivo

Based on our cultured primary rat cortical neuron screening, we chose to test the performance of both eGCaMP^{2+} and eGCaMP$^+$ *in vivo*, compared to GCaMP6f. We injected each variant of Cre-dependent GCaMP virus in the medial prefrontal cortex (mPFC), and a retrograde Cre virus in the nucleus accumbens (NAc; **Figure 2.14A**). This labeled a relatively sparse population of mPFC to NAc projections neurons with the GCaMP sensor. An optical fiber was implanted above the mPFC to measure the GCaMP fluorescence signal in response to brief foot shocks, which has been previously shown to elicit responses in these neurons[80]. Histology images showed qualitatively similar GCaMP expression in mPFC cell bodies and axons in NAc across all groups of mice (**Figure 2.14B**). All three GCaMP variants exhibited a time-locked increase in fluorescence during the foot shock, followed by a slow decay in the sensor fluorescence (**Figure 2.14C**). We calculated the mean response to the foot shock for each sensor and found that eGCaMP^{2+} exhibited a larger change in response compared to GCaMP6f and eGCaMP$^+$ (**Figure 2.14D**), similar our results in culture. We also calculated the mean sensor decay response 3 to 4 s after the foot shock (which is dependent on the off kinetics of the sensor; a faster sensor will return to baseline after the shock, while a slower sensor will still maintain a higher signal after the shock). We found that the eGCaMP$^+$ decay response was smaller than that of eGCaMP^{2+}, supporting our previous findings the eGCaMP$^+$ has the fastest off kinetics among our sensors (**Figure 2.14E**).

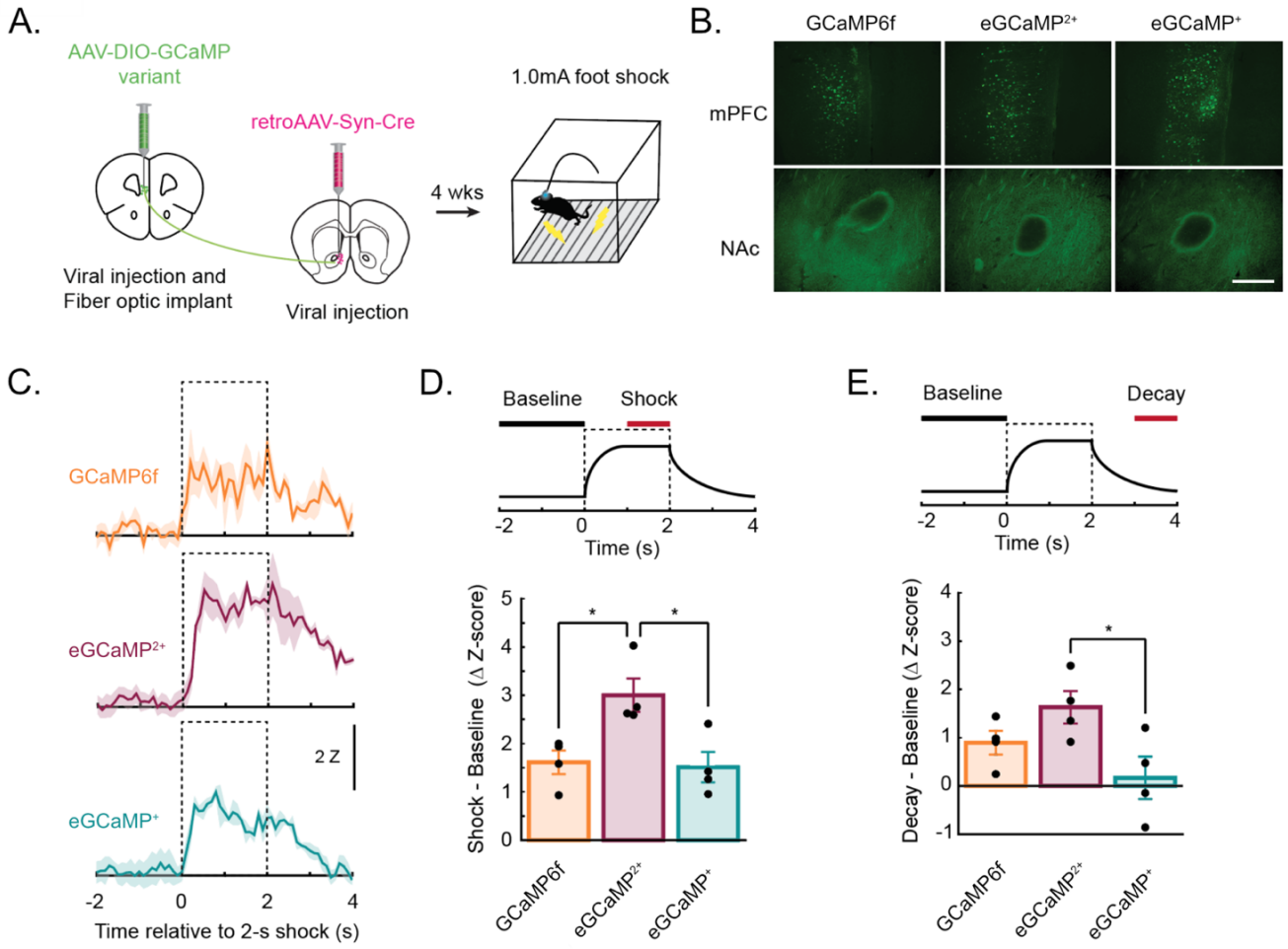


Figure 2.14: In Vivo Performance of eGCaMP+ and eGCaMP^{2+} expressed in mPFC

A. Experimental timeline. Mice were injected with an AAV-Cre dependent-GCaMP variant in the mPFC and an retroAAV-Syn-Cre was injected in NAc. An optic fiber was implanted above the mPFC to allow for light delivery and fluorescence recording.

B. Representative fluorescence images of GCaMP expression in mPFC and NAc (stained with anti-GFP-Alexafluor488). Scale bar, 130 µm.

C. Mean Z-scored fluorescence changes in response to a foot shock (n=4 total shock trials, collected from 2 mice for each GCaMP variant).

D. Comparison of the mean shock response between the three GCaMP variants. Top: schematic of how the shock response was calculated (see methods). Bottom: Mean change in Z-scored fluorescence response to shock (n=4 total shock trials, collected from 2 mice for each GCaMP version). P-values were calculated using a One-way ANOVA followed by Tukey's multiple comparisons in panels (D) and (E): *P<0.05. All data show mean +/- SEM.

E. Comparison of the mean decay to shock between the three GCaMP variants. Top: schematic of how the decay to shock was calculated (see methods). Bottom: Mean change in Z-scored fluorescence decay to shock (n=4 total shock trials, collected from 2 mice for each GCaMP version). P-values were calculated using a One-way ANOVA followed by Tukey's multiple comparisons in panels (D) and (E): *P<0.05. All data show mean +/- SEM.

2.3 DISCUSSION

Incorporating machine learning into our engineering pipeline enabled us to efficiently identify new GCaMP variants with enhanced $\Delta F/F_0$ responses and decay kinetics. We achieved impressive predictive performance in the cross-validation phase by using an ensemble of three regressor models, encoding our dataset with amino acid characteristics, and focusing solely on sequence inputs for learning. These predictive capabilities translated to the *in vitro* space, where many *in silico* predicted characteristics accurately reflected the mutant's true performance with an accuracy score of 0.66 (**Figure 2.7B,E**). As a result of these engineering efforts, we identified three new variants, *e*GCaMP, *e*GCaMP^{+}, and *e*GCaMP^{2+}.

With the functional predictions gathered from the model, we were able to not only gather mutations that directed sensor engineering but also able to observe the learning and predictive patterns to better understand the protein function. For example, when we mapped the residues the ensemble predicted would be influential back onto the GCaMP crystal structure, we found that the highlighted residues were in structurally significant parts of the GCaMP protein and often faced inward toward one other (**Figure 2.5**). This phenomenon may indicate that the ensemble is learning which residue interactions are important for protein function and govern the given biophysical property. As such, these residue interactions constitute a promising basis for further mutational studies and may even be used to influence future mutation library generation.

While the constructs presented here have not been previously described, clues from the literature may explain the impact of these mutations. For example, Residue L317 is known to be involved in extensive hydrophobic interactions between CaM and CBP[76]. Each mutation at L317 that the ensemble proposed is capable of forming hydrogen bonds, which may destabilize the CaM and CBP interactions, accelerate kinetics, and alter $\Delta F/F_0$ responses. Within the known variant

library, the GCaMP variants that contained a 317E/H/K/N mutation had decreased $\Delta F/F_0$ capability compared to jGCaMP7s, an association in which the ensemble learned (**Figure 2.8A**). However, each previously characterized variant that contained a mutation at residue 317 also contained an Alanine at residue 52 (**Figure 2.8B**). When we tested the L317H variant in jGCaMP7f, which contains A52, we observed the loss of $\Delta F/F_0$ capabilities that the model predicted and mirrored previous findings from the Dana *et al.* 2019 study (**Figure 2.8C**). These interactions had a substantial effect on protein function and may constitute a promising target for further mutation library studies.

The impressive dynamic range of the Q305D mutation in *e*GCaMP^{2+} may result from intraprotein interactions within CaM. One possible explanation is that the decreased R-group length in the Q305D mutation requires a more substantial conformational change to form the hydrogen bond with residue Y380 (**Figure 2.5E**). The resulting conformational change may have downstream effects on both the cpGFP/CaM linker (**Figure 2.10A**) and on residue R381, which faces inward toward the chromophore (**Figure 2.5E**). Hence, the dramatic effects of this mutation on the $\Delta F/F_0$ suggest a collaborative role between the cpGFP/CaM linker and inward loop of CaM in stabilizing the phenol/phenolate transition of the chromophore[81,82,83].

We made several critical design decisions while forming this methodology, such as our encoding method, chosen models, ensemble, devotion to sequence-only inputs, and limitation to single point mutation exploration. Dataset encoding is a crucial step in model training as it determines the underlying patterns on which the generalizations are formed[58]. For this reason, we encoded the sequence with biophysical properties underlying the amino acids in each position to form meaningful learning patterns. We derived our AA property datasets from the online repository AAINDEX[69]; however, other similar online databases exist[84]. Encoding with the

property matrices improved the cross-validation R^2 value by an average of 20% over one-hot encoded or label-encoded libraries (**Figure 2.3C**). Analysis of the top performing datasets within each model additionally provides insight into how the model was able to learn and served as a lens into the interprotein interactions. For instance, we found that AA property datasets that described hydrophobicity were commonly associated with higher-performing predictive capabilities in the $\Delta F/F_0$ model (**Figure 2.3B-D; Table 2.1**), meaning that some of the modifications in protein behavior may be due in part to key hydrophobic interactions. In comparison, AA property datasets associated with protein folding and energetics were common amongst the higher-performing predictive capabilities in the kinetics model (**Figure2.3B-D; Table 2.2**).

The reason we chose to include five property datasets in the final ensemble prediction was twofold. The first was that, during our training, we found that the top-performing datasets often achieved R^2 values that were remarkably similar (**Table 2.1&2.2; Figure 2.3**). Given the marginal superiority of the top-performing dataset over its counterparts, a strategic choice was made to include additional matrices. The selection of five datasets was made semi-arbitrarily, as it afforded the desired additional insights without dramatically impacting computational demands, processing time, and storage requirements. The second reason we chose to include more was due to the type of ensemble we were performing. Within the stacked ensemble, each final ensemble prediction was determined through unweighted averaging. This method is not free from outlier corruption, meaning that if one model's prediction is vastly different from the others, it will influence how that prediction is considered in the final ensemble predictions. The addition of more datasets/models enables some buffering to happen and for a large sample size to determine our ensemble's mean predictions.

Ensembling ML models (i.e., considering the input from multiple models) is preferable over single model predictions, as no singular model is perfectly optimized to perform all tasks[85]. We consider inputs from a random forest regressor (RFR), a K-neighbors regressor (KNR), and a multi-layer perceptron network regressor (MPNR). Decision tree learning methods, such as RFRs, are computationally efficient models well suited for small training libraries, such as the variant library, making them a strong foundation within our ensemble's learning[58]. KNRs are computationally demanding but simple[86], where KNR's similarity metric can capture the degree of variability between the performances of nearly identical sequences. The similarity metric highlights residues whose mutation led to large differences in the targeted biophysical property. MPNRs are deep-learning models capable of extracting high-level features from the data, making them useful for identifying key residues or properties that lead to the observed biophysical response[58]. The three selected models have diverse learning strategies and make different assumptions about the data, which is important when ensembling. When the predictions from each model are ensembled, the cross-validation predictive accuracy matched or improved the sole contributor's performance (**Figure 2.3C**).

While structural insights guided the engineering of previously published GCaMPs, we developed the ensemble pipeline to be structure-agnostic. This design consideration was crucial, as we aim to engineer subsequent GEFIs using this pipeline without relying on molecular structures. Due to the exclusion of structure information, extrapolation outside of the observed sequence space may be difficult. This tool is best suited for data generalization and exploration within a sequence space with only minor variations from the training dataset, such as point mutations at tested residues. However, one could incorporate spatial information from crystal structures or structure prediction tools in the ensemble's learning to aid extrapolation in the future.

One of the major hurdles of protein engineering is the susceptibility of proteins to experience epistasis, in which combinations of mutations non-additively influence the phenotypic characteristics of a protein[87]. Though the mutation library we worked with had >1000 well-characterized variants, the large number of mutated residues renders the dimensionality incredibly large. For a library such as this, there are 1.18e+91 possible combinations of residues over 70 mutated residue positions, meaning that the variant library is only a small sampling of the theoretical mutation space. We felt that the risk of epistasis upon combinatorial mutation was too great and that the relatively limited size of the library in comparison to its dimensionality rendered this application better suited to single-point mutation testing. Though investigation of a combinatorial library was not used in this study, others have shown promise using machine learning to engineer protein combinatorial libraries[63].

In previous studies, the authors employed a volume approach, in which they tested over a thousand variants iteratively and chose to fully characterize those determined to have optimal kinetic and maximal fluorescence capabilities. Because of the sheer number of experiments, they could split their variants into kinetic regimes and determine the best possible variant within each regime based on multiple biophysical properties[80]. The approach we employ here allowed much of the screening to occur *in silico*, which reduced the experimental burden while achieving similar outcomes. We trained and selected variants for downstream testing based on their predicted performance for one biophysical property: $\Delta F/F_0$ or off-kinetics. As a result, the selected variants display compensation within favorable biophysical characteristics, such as a lower baseline fluorescence, as it was never a criterion for their predictions (**Figure 2.12**). However, the lower baseline did not impact the performance of eGCaMPs in neuron cultures or *in vivo* fiber photometry (**Figure 2.14**). Hence, it would be an acceptable tradeoff in many use scenarios.

However, as a consideration for future studies, metrics for other favorable biophysical characteristics could be included in either ensemble training or variant analysis to preserve them within the final variants.

The machine learning ensemble used in this study has demonstrated an impressive capacity to guide fluorescent biosensor engineering. The ensemble's predictions helped identify variants with large $\Delta F/F_0$s and fast decay kinetics, while highlighting clusters of impactful residues for each biophysical property, which may be further exploited by mutation library-based high-throughput screening. These findings illustrate the ensemble's ability to guide engineering efforts and improve experimental efficiency. Moreover, since our model's learning is based solely on the sequence-function relationship and all contributor model optimization is unbiased, the final ensemble platform can be broadly applied to any genotype-to-phenotype mutation library. Applying this ML platform to mutation studies of proteins with quantifiable output characteristics, including other protein sensors, has the potential to accelerate the engineering of these proteins.

2.4 METHODS

2.4.1 Data Preprocessing

The Chen and Dana studies provide a functional characterization of >1000 GCaMP variants that span the GCaMP6 and jGCaMP7 iterations[84]. The experimental conditions from each study were standardized across experiments, allowing a direct comparison of the GCaMP mutation's properties[88]. Each study normalized the results to base constructs for data such as the $\Delta F/F_0$ response (*Eq.1*) to stimuli of 1 AP, 3 AP, 10 AP, 160 AP, and decay half-time after 10 AP. To cross-compare mutation libraries, we re-normalized the *Chen et al.* 2013 dataset such that GCaMP6s was 1.0 for all metrics. The authors linked the functional ability of each variant to a primary key identifier and the identities of the mutations within each variant. The list of mutations

was relative to either GCaMP3 or GCaMP6s for *Chen et al. 2013* and *Dana et al. 2019,* respectively. To generate a dataset compatible with ML algorithms, we replaced the list of mutations with a Pandas DataFrame containing one column per residue. The resultant data structure comprised 453 columns: one column containing the primary key identifiers present in the parent datasets, 451 columns corresponding to the sequence of the GCaMP variant, and the final column containing each variant's empirically derived performance. The mutations that occurred in each variant were reflected in their respective sequence positions within the DataFrame. Any duplicated variants that were present were isolated, and their responses were averaged before compiling them back into the variant library. This duplicate data consideration ensures that each variant only occurs once in the final variant library and ameliorates instances of data leakage between train and test data.

$$\Delta F/F_0 = \frac{(F - F_0)}{F_0} * 100 \qquad \textit{(Eq.1)}$$

The resultant dataset is the basis for our dependent and independent variables used to train our ML algorithms. Within the variant library used for model training, the independent variable consists of the sequence of each mutation. The dependent variable is the $\Delta F/F_0$ response (1AP $\Delta F/F_0$) or kinetics capability ($\tau_{1/2}$). However, because the sequence is a series of string-type values, the complexities of the identities of each amino cannot be understood by the algorithms. The sequences need to be encoded with quantitative values. Dataset encoding can be performed in several capacities: label encoding, one-hot encoding, or by adding functional information. Within our label encoding, we randomly assigned an integer value to each amino acid and replaced each residue label in the GCaMP sequence with the dummy label. For one-hot encoding, the full extent of possible residues at each position is considered in a Boolean manner (20 amino acids x 450 residue positions). The start codon, methionine, is considered a one in column 1-M, where every

other 1-x contains a zero. Finally, to perform encoding with functional data, we developed a dictionary of amino acid properties by web scraping the AAINDEX database[69,89]. AAINDEX consists of matrices that each describe a different AA property (e.g., Size (Dawson, 1972), Polarity (Grantham, 1974), Hydrophobicity (Jones, 1975)). The general shape and composition of each one of the property datasets is a list of 20 float type values, in which the order is linked to the amino acid and the float type value is a quantitative value that is dependent on the property in question. We used the 554 complete property datasets to formulate an unbiased model training paradigm in two steps. To perform the encoding, we replace each amino acid in the sequence with the corresponding value from the property dataset, i.e., the float type value that exists for that amino acid's position in the property dataset list. The final variant library used in model training consisted of the fully encoded GCaMP sequence and the variants empirically derived performance capability.

2.4.2 Generation of the novel variant library:

To generate a library of unknown sequences, we performed a single-point saturation of the jGCaMP7s sequence at 75 residue locations. These 75 residues correspond to the 75 residues that contain mutagenesis information in the variant library. The outcome was a novel point-saturation-mutation library that contained 1500 sequences. To ensure each variant was a previously untested sequence, we removed variants that had sequences redundant to any that occurred in the variant library, including any redundancies with jGCaMP7s, such that the final point-saturation-mutation library contained 1423 variants. Specifically, 75 variants were redundant with the base jGCaMP7s sequence, and two variants (jGCaMP7s L317A and jGCaMP7s H78K) were redundant with previously characterized variants. The $\Delta F/F_0$ and kinetics ensembles generated predictions of the

functional capabilities of the 1423 novel variants in the novel library with jGCaMP7s included as a control. These final predictions serve as the basis for mutations considered for *in vitro* testing.

2.4.3 Ensemble Training

The learning capabilities of any model are limited when tasked to predict outcomes where the factors underlying response have innumerable contributing factors. Under this assumption, we trained and optimized three regressors that would each contribute to the mutation predictions we tested *in vitro*. Our goal was to ensemble these weak learners and focus our downstream efforts on mutually agreed upon mutations. The models that we developed were from the pip installable package Scikit Learn in Python 3.8.5 to develop a Random Forest Regressor (RFR), K-Neighbors Regressor (KNR), and a Multi-layer Perceptron Network Regressor (MPNR). The models were trained on the encoded sequence of each variant linked to their empirically derived performance capability. The performance capabilities correspond to their $\Delta F/F_0$ response to one AP or half decay time after 10 AP. The data was split into train/test sets at a ratio of 80:20 with a random seed of 42 for downstream optimization efforts. Due to the inherent complexity of the 451-residue feature space of the GCaMP sequence, we performed the 'SelectKBest' feature selection function found in Scikit Learn to rank the importance of each input feature before model training. This feature selection was critical to reduce the dimensionality of the data and, ultimately, decreasing the required runtime. Optimization of the model was done by grid-search hyperparameter tuning. We used the coefficient of determination (R^2) and mean squared error (MSE) to track the fitting of each model. Additionally, we optimized each model using the key considerations that govern model performance, such as the number of neighbors in KNR and the number of estimators in RFR. Conditions that lead to the highest R^2 of the test set predictions were compared between each AA property dataset used for encoding to individually optimize and associate predictive

capabilities with the underlying amino acid property. This process was repeated over each of the 554 datasets for the three models (~1662x). Each model's top five performing property datasets were advanced to generate predictions on the novel variant library. Each contributor model (5 AA property x 3 Regressor models) forms predictions independently, and the final predictions are the average response from each contributor model for each target attribute ($\Delta F/F_0$ response 1AP $\Delta F/F_0$ or kinetics capability $\tau_{1/2}$). The predicted values returned by the ensemble are numeric values originating from a normalized library, making the predictions unitless. For example, smaller numeric values in the $\Delta F/F_0$ library would correspond to a predicted decreased $\Delta F/F_0$ response, and smaller numeric values in the kinetics library would correspond to a predicted faster decay speed.

2.4.4 PCA Clustering

Each feature within the data was first scaled using Sklearn's StandardScaler. We passed the scaled data into Sklearn's PCA function with no defined number of components. We chose the optimum number of components by finding where the explained variance of the PCA of the data passed 0.8. We reinitialized the PCA with the determined number of principal components and fit the function with the standardized data. We then used the principal component space coordinates to find the ideal number of clusters for K-Means clustering. We determined the ideal number of clusters by using the 'elbow method' on the Within Cluster Sum of Square. After finding the clusters, we labeled each input to their K-means-defined cluster.

2.4.5 Molecular Cloning

Predicted mutations were reflected into the CMV-jGCaMP7s backbone (Addgene ID: 104463) using point-mutation primers ordered from Integrated DNA Technologies (IDT) and PCR amplification with either Q5-polymerase (New England Biolabs; M0492L) or Superfi-II

polymerase (Invitrogen; 12368010). Amplification of the DNA fragment was verified with agarose gel electrophoresis. Blunt-end DNA circularization was achieved with Kinase, Ligase, and DpnI enzyme (KLD) treatment (New England Biolabs: E0554S). Circularized DNA was transformed into competent *E.Coli* cells (DH5α or TOP10) and grown on agar plates that contain either ampicillin or kanamycin selection antibiotic (50 µg/mL). Upon colony formation, single colonies were picked and grown in 5mL cultures containing LB Broth (Fisher BioReagents; BP9723-2) and selection antibiotic (ampicillin/kanamycin; 50 µg/mL) overnight (37°C, 230 RPM). DNA was isolated using Machery Nagel DNA prep kits (Machery Nagel; 740490.250). Sanger sequencing (Genewiz; Seattle, WA) of the isolated plasmid DNA was used to confirm the presence of the intended mutation.

Genes encoding the GCaMP variants were cloned into a CAG-driven backbone, pCAG-Archon1-KGC-EGFP-ER2-WPRE (Addgene; #108423), using Gibson assembly (New England Biolabs; E2621L). All subsequences were verified with Sanger sequencing (Genewiz; Seattle, Wa).

2.4.6 Acetylcholine Assays

Human Embryonic Kidney (HEK293; ATCC Ref: CRL-1573) cells were cultured in Dulbecco's Modified Eagle Medium + GlutaMAX (Gibco; 10569-010) supplemented with 10% fetal bovine serum (Biowest; S1620). When cultures reached 85% confluency, the cultures were seeded at 100,000 cells per well or 50,000 cells per well in 24-well and 48-well plates, respectively. 24 hours after cell seeding, the cells were transfected using Lipofectamine3000 (Invitrogen; L3000015) at 1000 ng of DNA per well of a 24-well plate, according to the manufacturer's instructions.

48 hours post-transfection, the plates were prepared for imaging by washing and then replacing culturing media volume with imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM

KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES (triple supplemented with 1% Glutamax (Gibco; 35050-1), 1% Sodium Pyruvate (GIBCO; 11360-070), and 1% MEM Non-Essential Amino Acids (Gibco; 11140-050)). Crystalline power Acetylcholine Chloride (Alfa Aesar; L02168.14) was resuspended into imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES) into 2x the desired final concentration. During imaging, 1:1 volumes of the acetylcholine-tyrodes imaging solution were hand-pipetted into the bath volume to bring the final acetylcholine concentration to the desired concentration. Imaging was performed on a sCMOS camera (Photometrics Prime95B) on an epifluorescent microscope (Leica DMI8) using a 20X objective (Leica HCX PL FLUOTAR L 20x/0.40 NA CORR). A Lumencor Light Engine LED and Semrock Filters (Excitation: FF01-474-27; Emission: FF01-520/35) were used for fluorescence imaging.

2.4.7 Analysis of Fluorescent Assays

Analysis of HEK293 cell fluorescence imaging data was done by FluorAREA, a custom cloud-based semi-automated time series fluorescence data analysis platform written in Python. First, the cell segmentation quality of the selected Cellpose[90] model was manually verified. For the segmentation of cells expressing cytosolic fluorescent indicators, model 'cyto' was selected as our base model. If the selected Cellpose model was low-performing, we further trained the Cellpose model using the Cellpose 2.0 human-in-the-loop system[91]. Using an "optimized" segmentation model, fluorescence time-series data is extracted for each region of interest. This allows for unbiased extraction of change in cellular fluorescence information for a complete set of experimental samples. Using the raw fluorescence data, % fluorescence change from the baseline ($\Delta F/F_0$) over time was calculated using *Eq.1*. The signal-to-noise ratio (SNR) was calculated using *Eq.2.*

$$SNR = \frac{(F_{max} - F_0)}{standard\ deviation\ (F_0)} \qquad (Eq.2)$$

The exponential decay constant ($\boldsymbol{\lambda}$) was calculated using *Eq.3*, where F(t) is the change in fluorescence at a time (t) after the max fluorescence (F_0) was achieved. Importantly, F_0 was normalized to 1.0, such that F(t) depicts the change in fluorescence over time, t.

$$F(t) = F_0 e^{-\lambda t} \qquad (Eq.3)$$

The exponential time constant ($\boldsymbol{\tau}$) was isolated by using the known reciprocal relationship of $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ (Eq.4).

$$\tau = \frac{1}{\lambda} \qquad (Eq.4)$$

The dynamic range (DR) was defined as the ratio of the max fluorescence intensity to the baseline fluorescence intensity (*Eq.5*). All $\Delta F/F_0$, SNR, $\boldsymbol{\tau}$, and DR values were quantified using a custom python script.

$$DR = \frac{F_{max}}{F_0} \qquad (Eq.5)$$

2.4.8 Optical Properties of Purified Proteins

Proteins were purified by large-scale protein purification and SEC purification, as previously described[31,92]. Purified protein isolates were diluted to 10μM in 30 mM MOPS, 100 mM KCl, pH 7.2 with either 10 mM CaEGTA (high Ca^{2+}) or 10 mM EGTA buffers (low Ca^{2+}) (Invitrogen; C3008MP). Protein absorbance spectra were recorded for each condition using a UV-vis spectrophotometer (NanoDrop 2000/2000c Spectrophotometers; Thermo Scientific). Fluorescence emission and excitation spectra for each condition were measured with a spectrum-capable plate reader (SpectraMax M5; Molecular Devices).

For calcium titrations, GCaMP protein was first diluted (0.5 μM protein) in triplicate in high Ca^{2+} or low Ca^{2+} buffers. These two solutions were mixed in various ratios to give 11 different

free calcium concentrations (Invitrogen; C3008MP). GCaMP fluorescence (excitation 485 nm; emission 535 nm) was measured using a SpectraMax M5 (Molecular Devices). Calcium titration curves were fit (Prism; GraphPad) to sigmoidal binding functions, and the Hill coefficient and Kd for Ca^{2+} binding for the GCaMP variants were extracted.

The absorbance under saturating conditions was measured using 2 μM protein diluted into high Ca^{2+} buffer at 500 nm (DU800 spectrophotometer; Beckman Coulter). The chromophore concentration was measured from the absorbance (447 nM) of protein denatured by 1 M NaOH (extinction coefficient 44,000 M−1 cm−1). The extinction coefficient was calculated using Beers' law, where the absorbance was at that of the saturated protein at 500 nm, and the concentration was extracted using the absorbance of the denatured protein.

Quantum yield measurements were measured at 460 nm light using an integrating-sphere spectrometer (Hamamatsu) for 0.3 μM protein diluted in +Ca buffer.

2.4.9 Isolation of Cortical Neurons

Primary cortical neurons were prepared as previously described[93,94]. Briefly, 24-well tissue culture plates were coated with matrigel (mixed 1:20 in cold-PBS, Corning; 356231) solution and incubated at 4°C overnight prior to use. Sterile dissection tools were used to isolate cortical brain tissue from P0 rat pups (male and female). Tissue was minced until 1mm pieces remained, then lysed in equilibrated (37°C, 5% CO_2) enzyme (20 U/mL Papain (Worthington Biochemical Corp; LK003176) in 5mL of EBSS (Sigma; E3024)) solution for 30 minutes at 37°C, 5% CO_2 humidified incubator. Lysed cells were centrifuged at 200xg for 5 minutes at room temperature, and the supernatant was removed before cells were resuspended in 3 mLs of EBSS (Sigma; E3024). Cells were triturated 24x with a pulled Pasteur pipette in EBSS until homogenous. EBSS was added until the sample volume reached 10 mLs prior to spinning at 0.7 rcf for 5 minutes at room

temperature. Supernatant was removed, and enzymatic dissociation was stopped by resuspending cells in 5 mLs EBSS (Sigma; E3024) + final concentration of 10 mM HEPES Buffer (Fisher; BP299-100) + trypsin inhibitor soybean (1 mg/ml in EBSS at a final concentration of 0.2%; Sigma, T9253) + 60 µl of fetal bovine serum (Biowest; S1620) + 30 µl 100 U/mL DNase1 (Sigma;11284932001). Cells were washed 2x by spinning at 0.7 rcf for 5 minutes at room temperature and removing supernatant + resuspending in 10 mLs of Neuronal Basal Media (Invitrogen; 10888022) supplemented with B27 (Invitrogen; 17504044) and glutamine (Invitrogen; 35050061) (NBA++). After final wash spin and supernatant removal, cells were resuspended in 10 mLs of NBA++ prior to counting. Just before neurons were plated, matrigel was aspirated from the wells. Neurons were plated on the prepared culture plates at desired seeding density. Twenty-four hours after plating, 1µM AraC (Sigma; C6645) was added to the NBA++ growth media to prevent the growth of glial cells. Plates were incubated at 37°C and 5% CO_2 and maintained by exchanging half of the media volume for each well with fresh, warmed Neuronal Basal Media (Invitrogen; 10888022) supplemented with B27 (Invitrogen; 17504044) and glutamine (Invitrogen; 35050061) every three days.

2.4.10 Calcium Phosphate Transfection of Primary Cortical Neurons

Isolated primary cortical neurons were transfected using the calcium phosphate transfection kit from Sigma Aldrich (Sigma-Aldrich; CAPHOS-1KT). Half of the neuron media was changed 24 hours before transfection, saving the removed conditioned media to add to the neurons after transfection. Reagents were mixed in a ratio of 3 µl $CaCl_2$: 24.5 µl H_2O: 1000 ng DNA before being added dropwise to bubbled 2x HEPES Buffered Saline (30 µl). The final solution was vortexed for 4 seconds and left undisturbed for 20 minutes. The solution was added

dropwise to each well of neurons in a 24-well plate and shaken to distribute equally. Neurons were left to incubate for 1 hr at 37^0C with 5% CO_2. The cells were rinsed twice with HBSS before adding the conditioned media removed from the day prior and mixed with half-fresh media.

2.4.11 Electrical Field Stimulation

On the day of imaging, ~24-36 hours post-transfection, cells were washed once with imaging solution and then transferred to E-Stim Tyrode's (pH = 7.33; 150 mM NaCl, 4 mM KCl, 3 mM $CaCl_2$, 1 mM $MgCl_2$, 10 mM Dextrose, 10 mM HEPES)[79]. A custom wire holding piece was designed to fit into 48-well plates with silver wires 10 mm apart. 100 mA pulses, with a 3 ms pulse width, were administered at either 0.5Hz or 10Hz frequency using a pulse generator (Warner Instruments; SIU-102B), triggered with Sutter Instruments Integrated Patch Amplifier with Patch Panel, time-locked using Igor Pro 8. Imaging was performed with a digital camera (Hamamatsu ORCA-Flash4.0; C11440) at 100ms exposure attached to an epifluorescent microscope (Leica DM IL). The light was generated using a SOLA Light Engine (Lumencor; SOLA SE 5-LCR-SB) with a 488 nm wavelength filter lens. Bulk fluorescence traces were acquired using FIJI imaging software with background subtraction (rolling = 50 stack) and hand-drawn ROIs. The baseline was defined as the first 50 measurements before the event trigger. Max $\Delta F/F_0$ and decay values were obtained using a custom Python script. Final traces were plotted in Prism9.

2.4.12 Potassium Chloride Assays

On the day of imaging, ~24-36 hours post-transfection, cells were washed once with imaging solution, then replaced with imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES (triple supplemented with 1% Glutamax (Gibco; 35050-1), 1% Sodium Pyruvate (Gibco; 11360-070), and 1% MEM Non-Essential Amino Acids (Gibco; 11140-050)). Powdered Potassium Chloride (Sigma; P9541-500G)

was diluted in ddH_2O to a concentration of 2M. This solution was then diluted to 80mM in imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES). During imaging, 1:1 volumes of KCl solution were hand-pipetted into the bath to bring the final KCl concentration to the desired concentration. Imaging was performed on a sCMOS camera (Photometrics Prime95B) on an epifluorescent microscope (Leica DMI8) using a 20X objective (Leica HCX PL FLUOTAR L 20x/0.40 NA CORR). A Lumencor Light Engine LED, and Semrock Filters (Excitation: FF01-474-27; Emission: FF01-520/35) were used for fluorescence imaging. Bulk fluorescence traces were acquired using FIJI imaging software with background subtraction (rolling = 50 stack) and hand-drawn ROIS. The baseline was defined as the first 30 measurements before KCl addition. Max $\Delta F/F_0$ values were obtained using a custom Python script. Final traces were plotted in Prism9.

2.4.13 Animals

6–7-week-old male and female C57BL/6J mice were obtained from Jackson Laboratories and maintained on a 12h reverse light-dark cycle (lights on at 9PM) at 22ºC, group-housed with same-sex cage mates and given ad-lithium access to food and water. Mice were left undisturbed for 1-week following arrival before the start of testing. All experiments occurred in the dark cycle. All experiments were conducted in accordance with UC Davis's Institutional Animal Care and Use Committee.

2.4.14 Stereotaxic Surgery

Mice were anesthetized under 1.5-2% isoflurane and placed in a stereotaxic apparatus (RWD) on a heat pad. Three different AAV cre-dependent GCaMP variants were tested: AAV5-Syn-FLEX-GCaMP6f (Addgene #100834; final titer 1.1E+13 GC/ml vg/mL) AAV1-EF1a-DIO-eGCaMP+ (Fred Hutch Virus Core; final titer 1.25E+12 IU/ml); or AAV1-EF1a-DIO-eGCaMP2+

(Fred Hutch Virus Core; final titer 6.80E+11 IU/ml). 1µL of the AAV cre-dependent GCaMP variant was infused into the medial prefrontal cortex (mPFC; M/L: -0.35, A/P: 1.98, D/V: -2.25 mm relative to bregma), and 500nL of retroAAV-Syn-Cre (Addgene #105553; final titer 9.50E+12 GC/ml) was infused into the nucleus accumbens (NAc; M/L: -0.35, A/P: +1.25, D/V: -4.6 mm). Injections were performed at a rate of 150nL/min using a Hamilton syringe controlled by an injection pump (World Precision Instruments). The virus was allowed to diffuse for 5 mins before withdrawing the needle. Chronically implantable fibers (RWD; 400µm core, 0.37NA, 1.25mm ceramic ferrule) were implanted above the mPFC injection site (M/L: -0.35, A/P: 1.98, D/V: -1.5 mm) to allow for blue light delivery and fluorescence signal recording. Recordings began 4 weeks after surgery to allow sufficient time for viral expression.

2.4.15 Fiber photometry recording

Fiber photometry recordings were performed using RWD's Tricolor Multi Channel Fiber Photometry System. Briefly, 470nm and 410nm light pulses were alternately delivered through a 400µm patchcord (0.57NA; Doric Lenses) connected to an optical fiber implanted above the PFC. Fluorescence was recorded with a cMOS sensor using RWD software at a frequency of 20Hz. The 410nm trace was linearly scaled to the 470nm trace and subtracted for each recording. The corrected 470nm trace was then z-scored for further analysis.

2.4.16 Shock delivery

During the fiber photometry recording, mice were given a 2 s, 1.0-mA foot shock 2 times, separated by at least 60 seconds. Shocks were delivered using a behavior box with a built-in shock floor (Med Associates Inc.). The time of shock delivery was synchronized to the fiber photometry recording using TTL time stamps.

2.4.17 Fiber photometry analysis

Data analysis was performed using MATLAB (Mathworks v 2020b). The 410nm trace was linearly scaled to the 470nm trace and subtracted for each recording. The corrected 470nm trace was then z-scored for further analysis. To calculate the mean shock response, the mean trace from t=1 to 2s after the shock onset was calculated, and then the mean baseline trace from t=-2 to 0s before the shock was subtracted from that. To calculate the mean decay after the shock, the mean trace from t=3 to 4s after the shock onset was calculated, and then the mean baseline trace from t=-2 to 0s before the shock was subtracted from that.

2.4.18 Histology

Mice were anesthetized under 5% isofluorane and perfused with 20mL cold phosphate-buffered saline (PBS), followed by 20mL of cold 4% paraformaldehyde (PFA). Brains were extracted and post-fixed overnight in PFA before being transferred to PBS. Brains were sliced on a vibratome (Leica) to a thickness of 60µm. For immunostaining, brain slices were first washed in PBS with 0.3 % Triton-X then blocked for 60 min in PBS with 0.3% Triton-X and 5% normal donkey serum. Slices were stained overnight with anti-GFP-AlexaFluor488 antibody (1:1000 in blocking solution, Life Technologies cat#A-21311) at 4ºC. Histology images were captured using a Keyence BZ-X180 fluorescence microscope, with an 80W halide lamp and PlanApo 10x 0.45 NA air objective. GFP fluorescence was visualized using the commercially provided GFP set excitation/emission filters. Images were processed using ImageJ (Fiji).

Chapter 3. Machine Learning Directed Engineering of Red-Shifted Genetically Encoded Calcium Indicators

ABSTRACT

Genetically encoded calcium indicators (GECIs) are pivotal optogenetic sensing tools that enable the tracking of calcium transients, particularly action potentials, in freely moving animals. The extensively engineered family of GCaMP has revolutionized the field of optogenetics, though much of the utility is limited to blue-light activation wavelengths. The abundance of tools solely reliant on cpGFP limits multiplexing capabilities and necessitates a shift toward spectra outside blue light activation for GECIs and other GEFIs. This chapter explores methodologies for potentiating the development of red-shifted GECIs. Rational engineering of XCaMP-R, though promising, yielded inferior results compared to its machine learning counterparts. Machine learning-guided optimization of jRCaMP1b, based on mRuby3, identified mutations enhancing fluorescence responses and decay kinetics. Furthermore, the ML insights were used to facilitate the design of mutation libraries targeting specific biophysical properties. This study showcases a symbiotic relationship between machine learning algorithms and high-throughput screening, promising accelerated discovery and efficacy of optogenetic tool engineering.

3.1 INTRODUCTION

Genetically encoded calcium indicators (GECIs) are broadly adopted optogenetic sensing tools that provide insight into calcium transients and enable researchers to track action potentials within freely moving animals[9,10,43]. The family of GCaMP sensors, cpGFP-based GECIs, has seen several decades of engineering and improvements, and outstanding tools currently exist[87]. Their success overshadows the development of red-shifted GECIs. This phenomenon is partly due to the abundance of bright fluorescent proteins (FPs) available within the blue light activation and the diminished number at every other excitation wavelength (**Figure 3.1A**). While cpGFP will remain a prominent FP in sensor engineering, the abundance of tools that contain cpGFP makes co-expression of multiple optogenetic tools an impossibility. For this reason, a critical need exists for GECIs and GEFIs that are shifted into spectra outside of the blue light activation.

Previous studies have introduced red-shifted fluorophores into the GCaMP scaffold and produced functional sensors using the FPs mRuby and mApple (**Figure 3.1B**). However, the capabilities of these sensors are nowhere near their cpGFP-based counterparts[9,10,43]. Currently, there is a tradeoff within commonly used sensors. jRGECO1a is highly sensitive and fast but contains a photo artifact when stimulated with 488nm light (**Figure 3.1C**). In comparison, jRCaMP1b is stable at 488 nm light but has a diminished affinity for Ca^{2+} and a smaller dynamic range than jRGECO1a. To multiplex red-shifted GECIs with optogenetic actuators and other tools, we require a sensitive variant that displays large $\Delta F/F_0$s, a variant with faster decay kinetics, and we need these tools to have stable artifact-free fluorescent read-outs.

We propose several methodologies to help potentiate the development of red-shifted GECIs. First and foremost, we attempted the rational engineering of XCaMP-R by transferring the promising mutations from our eGCaMP variants onto the XCaMP-R backbone, in an attempt to

improve the kinetics and $\Delta F/F_0$. We additionally utilized a library of published library of jRCaMP1b variants and performed machine learning-based optimization. Preliminarily, we have identified promising variants that improve the change in fluorescence and the speed of decay of jRCaMP1b. We further demonstrated that, with the predictions from the machine learning screening, we could identify key residues that play a role in interprotein interactions that determine sensor functionality. We used the model inferences as targets for mutation libraries that we tested using high-throughput opto-MASS screening[40]. From our high-throughput screening, we identified promising variants that display large $\Delta F/F_0$'s and fast decay kinetics. With the approaches we present, we can create a symbiotic relationship between machine learning algorithms and high-throughput screening that can dramatically improve the speed of discovery and efficacy of optogenetic tool engineering.

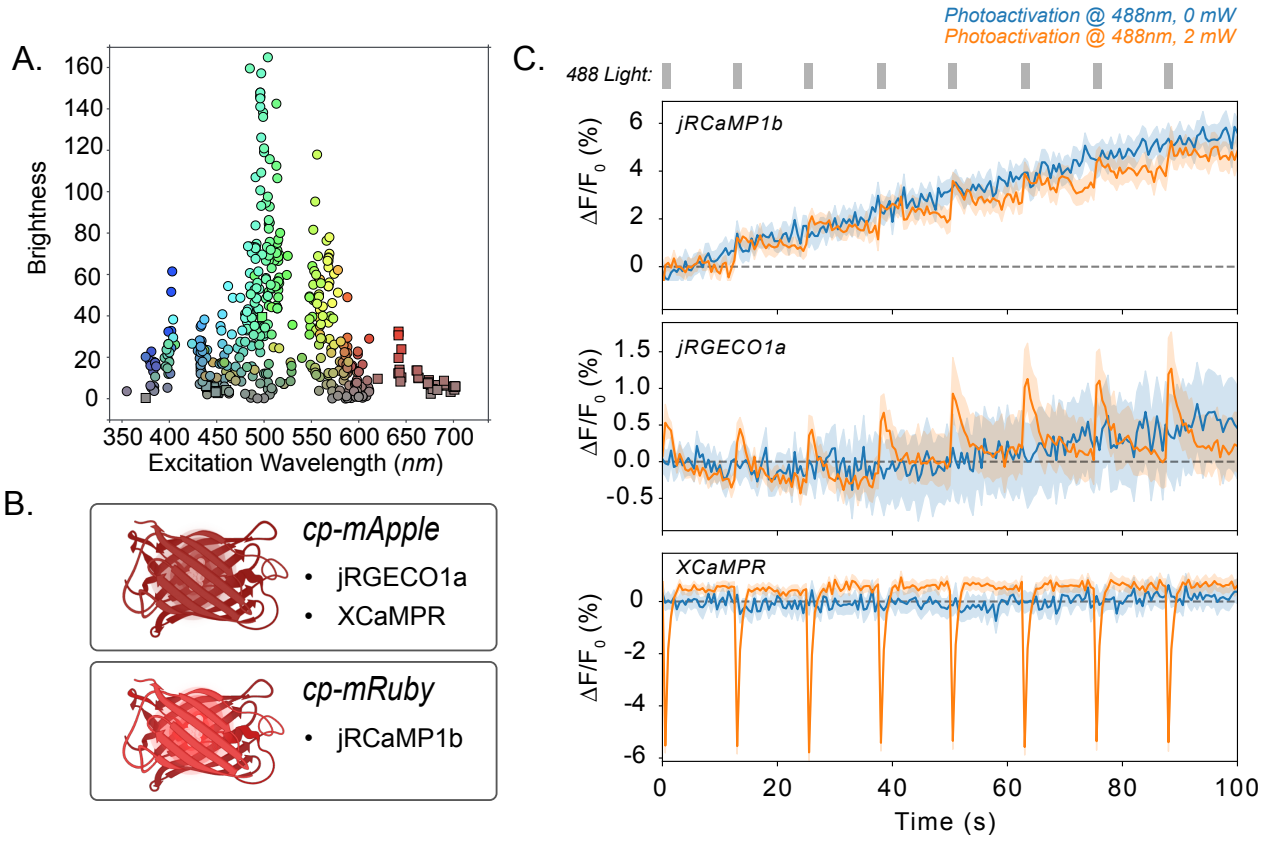


Figure 3.1 Currently Available Red-Shifted Calcium Indicators

A. Scatter plot displays the currently available fluorophores plotted by their excitation wavelength by the FP's brightness, information derived from FPBase[95].

B. Figure depicting the red-shifted fluorophores used in previously published red-calcium indicators jRGECO1a[96], XCaMP-R[9], and jRCaMP1b[96].

C. Plots depict the $\Delta F/F_0$ (%) of each indicated construct when exposed to 2 mW flashes of 488 nm light (orange) or when no light stimulus is added (blue). Lines depict the mean; shading represents the 95% confidence interval. The applied stimulus is shown in gray.

3.2 RESULTS

3.2.1 Rational Engineering of XCaMP-R

As a canonical starting point, we decided to transfer the mutations from our high-performing eGCaMP and eGCaMP^{2+} variants onto the XCaMP-R backbone[9,41]. We chose XCaMP-R as the basis for this engineering as it displayed reliable single-action potential responses *in vivo and* robust folding and expression[9]. The mutations within the eGCaMP variants lie on

conserved calmodulin and linker regions between eGCaMP, eGCaMP^{2+}, and XCaMP-R (**Figure 3.2A,B**). The first mutation, A323H that led to eGCaMP, when transferred to XCaMP-R, could not produce similar improvements in fluorescence response as we observed jGCaMP7s (**Figure 3.2B,C, Table 3.1**). However, this XCaMP-R A323H did lead to a slightly improved speed of decay, though not a significant improvement (**Figure 3.2D, Table 3.2).** This observation is consistent with the prediction that mutations at this location would increase the decay speed of GCaMP7s[10,86]. The addition of the second mutation, Q311D did lead to a roughly 25% improvement in saturation $\Delta F/F_0$ (**Figure 3.2B,C, Table 3.1**). Within both variants, there was a diminished response at 5μM acetylcholine compared to XCaMP-R, which may be indicative of reduced sensitivities of these variants (**Figure 3.2C, Table 3.1)**. When expressed in primary cortical neurons, we observed robust expression of all three constructs, as well as a robust response to 40mM KCl. However, there was not a significant improvement over the parental XCaMP-R (**Figure 3E,F)**.

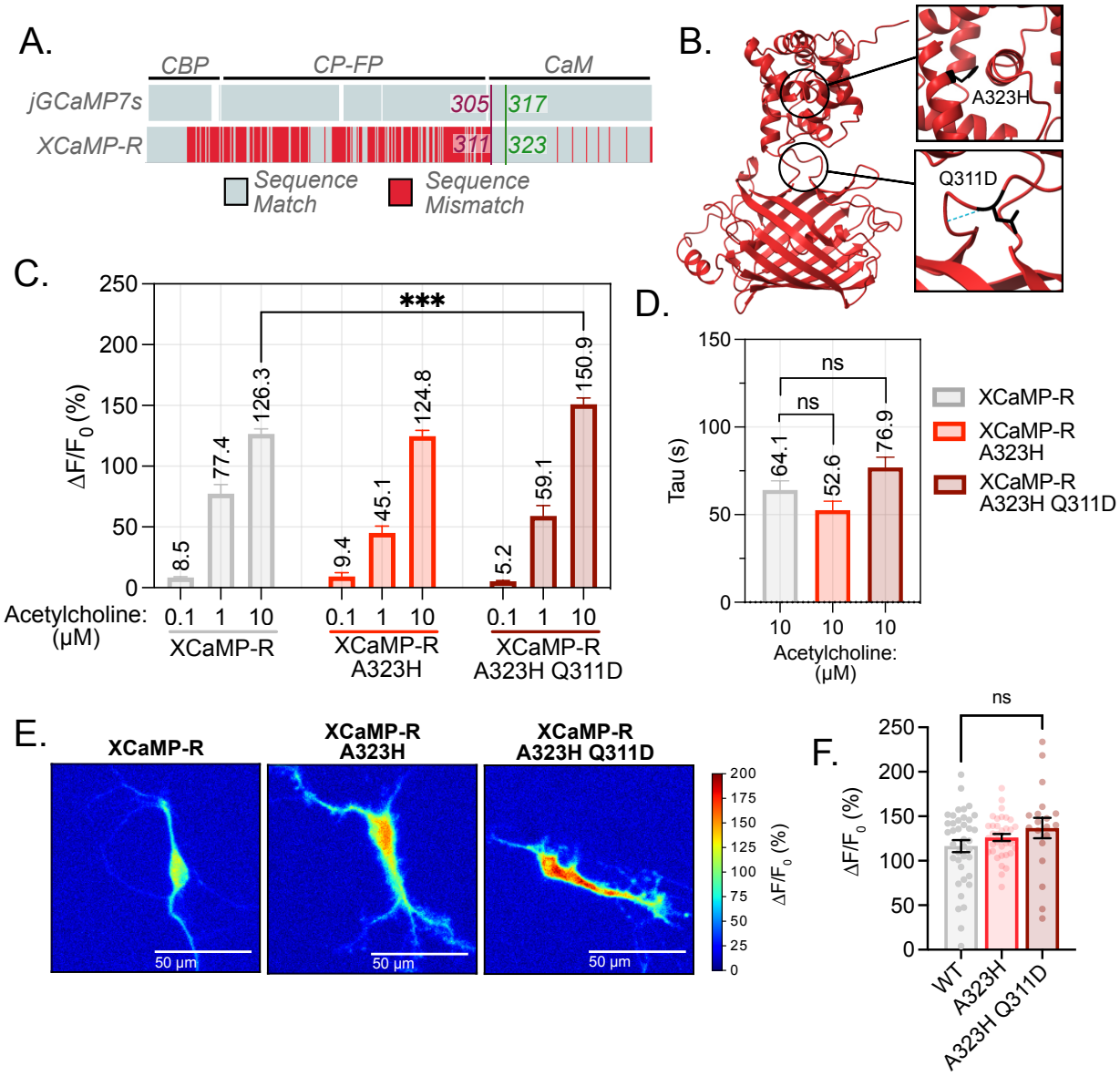


Figure 3.2 Rational Engineering of XCaMP-R Using Previous Identified Mutation Targets

A. Sequence alignment of jGCaMP7s and XCaMP-R. Light gray indicates identical sequence alignment, and red indicates sequence dissimilarities. Breaks indicate a missing sequence portion caused by additional sequence portions in other constructs. The physical location of the residue L317 in jGCaMP7s is designated with the green box and the residue number of the matching location is included as green text over the sequence. Text along the top of the sequence depicts the physical location in the GCaMP protein: CBP, CaM, or circularly permuted fluorescent protein (cpFP).

B. Alphafold Predicted Crystal structure of XCaMP-R (Red) with A323H and Q311D colored in black with sidechains visible.

C. Max fluorescent responses (*Eq.1*) obtained from each variant indicated on the x-axis, expressed in HEK293 cells and stimulated with indicated concentration of acetylcholine (Bars depict mean + SEM; **** = <0.0001 (unpaired t-test, two-tailed)).

D. Decay values ($\boldsymbol{\tau}$, tau, *Eq.4*) obtained from each variant indicated on the x-axis, expressed in HEK293 cells and stimulated with 10 μM acetylcholine (bars depict mean + SEM; ns = >0.05 (unpaired t-test, two-tailed)).
E. Representative images of maximal fluorescent response to 40 mM KCl stimulation variant indicated above image. Heat Mapping displays $\Delta F/F_0$ (%) achieved by each pixel. (Scale bar = 50 μm).
F. Maximum $\Delta F/F_0$ (%) achieved after stimulation with 40 mM KCl. (bars depict mean + SEM; ns = >0.05 (Unpaired t-test, Two-tailed)).

Table 3.1 Descriptive Statistics of $\Delta F/F_0$ Responses of XCaMP-R Variants in Acetylcholine Assay

Variant	XCaMP–R			XCaMP–R A323H			XCaMP–R A323H Q311D		
Acetylcholine (μM)	0.1	1	10	0.1	1	10	0.1	1	10
Number of values	337	123	265	169	127	154	218	78	138
Mean	8.497	77.39	126.3	9.366	45.08	124.8	5.218	59.06	150.9
Std. Deviation	11.26	83.47	67.68	39.61	63.44	59.43	9.361	74.68	61.66
Std. Error of Mean	0.6131	7.527	4.158	3.047	5.63	4.789	0.634	8.455	5.249

Table 3.2 Descriptive Statistics of Decay Time of XCaMP-R Variants in Response to 10μM Aceylcholine

Variant	XCaMP–R	XCaMP–R A323H	XCaMP–R A323H Q311D
Acetylcholine (μM)	10	10	10
Number of values	265	154	138
Mean	64.1	52.59	76.89
Std. Deviation	85.22	63.73	69.06
Std. Error of Mean	5.235	5.135	5.879

Within the XCaMP-R engineering, we saw improvements that were far inferior to the responses we observed in GCaMP. There are many possible reasons for this outcome, including the rigidity of the FP, making mutations challenging to transfer between scaffolds[9]. In addition to poor improvement over the parental construct, we found that other red-shifted GECIs outperformed our XCaMP-R variants (**Figure 3.3A**). Additionally, XCaMP-R contains the FP m-

Apple, which displays photoactivation under 488nm light (**Figure 3.3B)**. Without significant improvement to photoactivation properties, this variant would remain useless for multiplexing use cases. For this reason, we chose to shift our engineering efforts towards different GECIs.

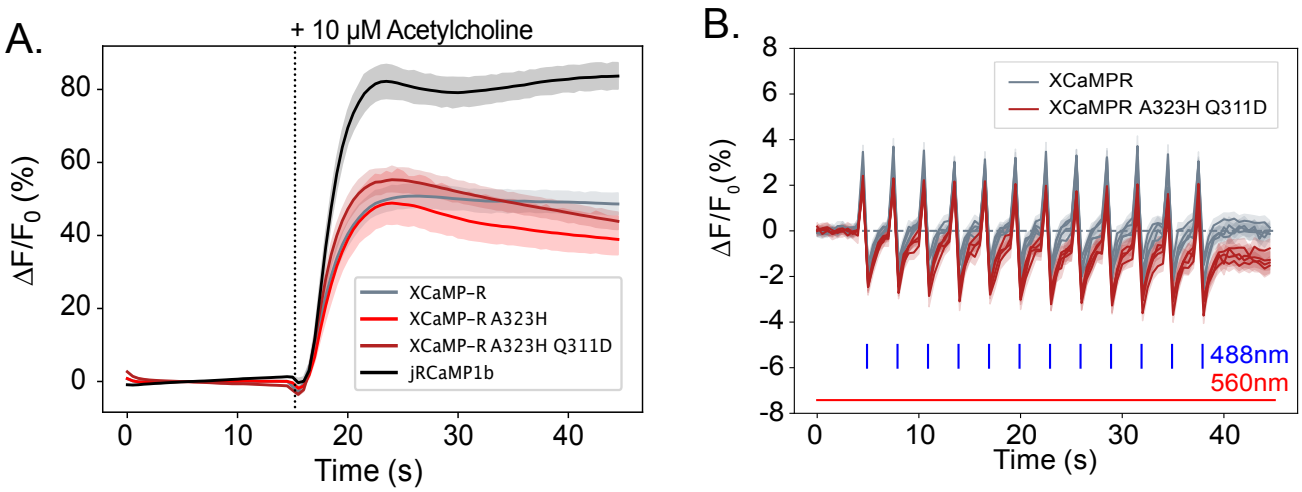


Figure 3.3 Complications with Engineering XCaMP-R Variants.
A. $\Delta F/F_0$ (%) recordings of each variant to 10 µM Acetylcholine stimuli applied after 15 seconds second (lines depict mean, shading depicts SEM). Stimulus applied at the dotted line.
B. $\Delta F/F_0$ (%) recordings of each variant to 500ms of 488 nm stimulus. Stimulus was applied every 2.5 seconds (lines depict mean, shading depicts SEM).

3.2.2 ML Guided Optimization of Biophysical Properties of jRCaMP1b

jRCaMP1b was an advantageous next target for engineering for several reasons. The first is that jRCaMP1b is based on the red-shifted FP, mRuby3, which does not display photoactivation to 488nm light, making it a reliable target for multiplexing[71,72] (**Figure 3.1C**). Users of the tool tended to favor other mApple-based red-shifted tools due to greater dynamic responses and better Ca^{2+} affinities. Fortunately, jRCaMP1b was engineered using the same neuron culture screening approach as our GCaMP variants and similarly contained published variant libraries that contained >1000 variants[10,88] (**Figure 3.4A**). These mutation locations were broadly distributed throughout the protein and included regions in the CaM, CBP, and cpFP (**Figure 3.4B**).

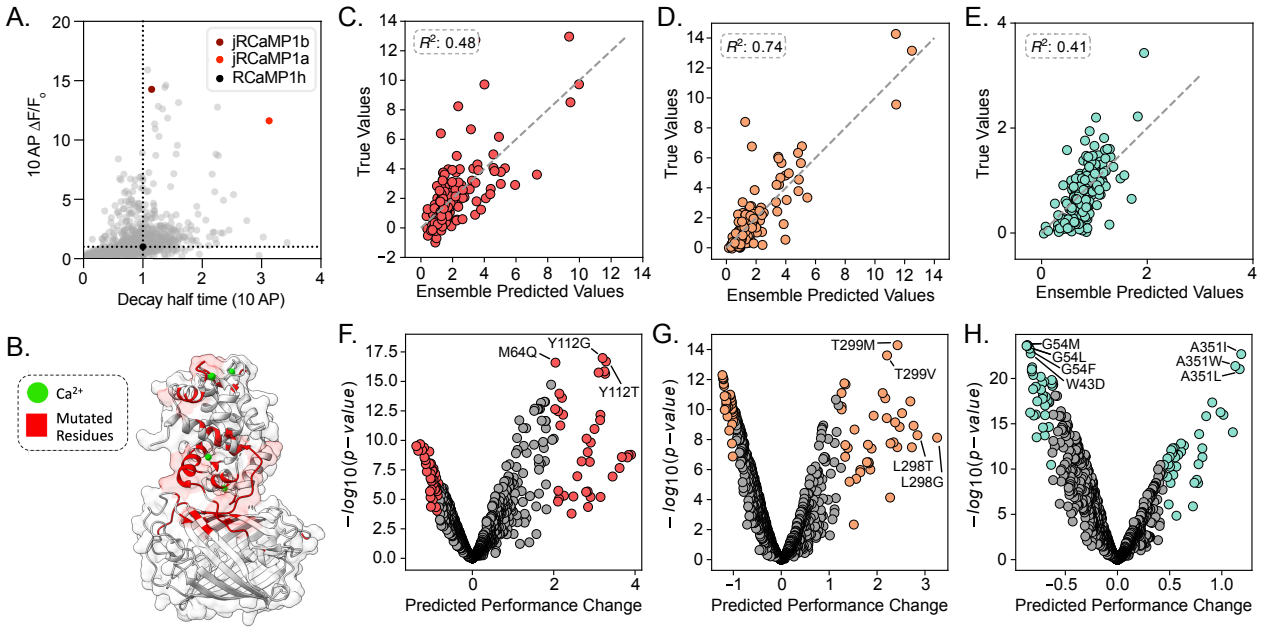


Figure 3.4 Ensemble Predictions on jRCaMP1b Variant Libraries

A. Scatter plot depicts the 10AP $\Delta F/F_0$ and Decay half time after 10 AP for each of the 1067 variants in the jRCaMP1b variant library[92]. Published variants are indicated with colored dots.

B. Crystal structure of RCaMP1h (RCSB: 3U0K, gray) with mutated residues (red) in that exist in the variant library. These residues indicate the positions used to form the novel library.

C. Cross-validation of the 1AP response ensemble. The scatter plot x-axis represents the true $\Delta F/F_0$ value for each variant in the test set, and the y-axis represents the predictions made by the ensemble of the variants in the test set. The dotted line depicts perfect agreement between true values and predicted values. R^2 value denotes the coefficient of determination of the scatter data.

D. Cross-validation of the 10AP response ensemble. The scatter plot x-axis represents the true $\Delta F/F_0$ value for each variant in the test set, and the y-axis represents the predictions made by the ensemble of the variants in the test set. The dotted line depicts perfect agreement between true values and predicted values. R^2 value denotes the coefficient of determination of the scatter data.

E. Cross-validation of the kinetic ensemble. The scatter plot's x-axis represents the true $\boldsymbol{\tau}_{1/2}$ value contained for each variant in the test set, and the y-axis represents the predictions made by the ensemble of the variants in the test set. The dotted line depicts perfect agreement between true values and predicted values. R^2 value denotes the coefficient of determination of the scatter data.

F. Volcano plots depicting the ensemble's prediction for a given mutation change in fluorescent response to 1AP from jRCaM1b (x-axis) and the log10(P-value) of the given prediction. P-values were calculated using an unpaired t-test on ensemble predictions (15 models) for jRCaMP1b and the given mutation.

G. Volcano plots depicting the ensemble's prediction for a given mutation change in fluorescent response to 10AP from jRCaM1b (x-axis) and the log10(P-value) of the given prediction. P-values were calculated using an unpaired t-test on ensemble predictions (15 models) for jRCaMP1b and the given mutation.

H. Volcano plots depicting the ensemble's prediction for a given mutation change in decay speed after 10AP from jRCaM1b (x-axis) and the log10(P-value) of the given prediction. P-values were calculated using an unpaired t-test on ensemble predictions (15 models) for jRCaMP1b and the given mutation.

Using the machine learning algorithm ProteiML we developed and discussed in Chapter 2, we trained our models on the jRCaMP1b data. We used ProteiML to form predictions on a novel mutant library that contained 1484 variants. We trained five models corresponding to the following biophysical characteristics: 1AP $\Delta F/F_0$, 10AP $\Delta F/F_0$, 160AP $\Delta F/F_0$, Half Decay Time to 10AP, and Baseline Fluorescence. Within the cross-validation phase, each model obtained R^2 as follows: 1AP ensemble ($R^2 = 0.48$, **Figure 3.4C**), 10AP ensemble ($R^2 = 0.74$, **Figure3.4D**), 160AP ensemble ($R^2 =0.55$), Decay ensemble ($R^2 =0.41$, **Figure3.4E**), and Baseline fluorescence ($R^2 =0.23$). We observed a broad distribution in our model's ability to predict, and we suspect that this is a result of the composition of the library. For instance, the authors actively targeted the fluorescent response, and we observed a broad distribution in fluorescence capabilities that correlates with a larger R^2. The range of kinetic capabilities was much more limited, which correlates with a lower R^2. Regardless, we were able to extract predicted locations within each ensemble that can be used for downstream *in vitro* testing (**Figure 3.4F, G, H**).

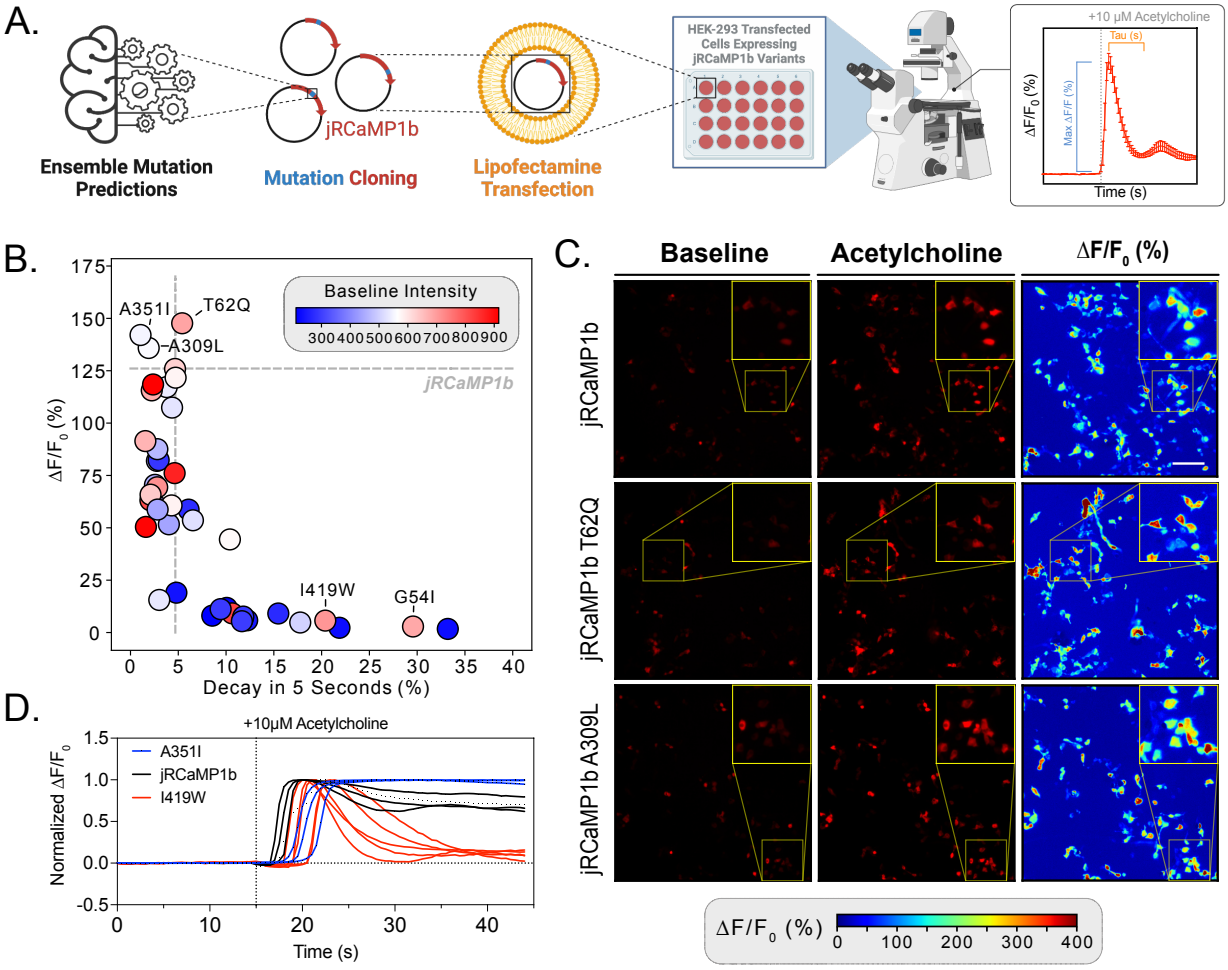


Figure 3.5 Acetylcholine Assay in HEK293 Cells to Validate Ensemble Red-Calcium Indicator Predictions

A. Mutation predictions isolated from the ensemble are used as the basis for downstream variant analysis. Variants of interest are cloned into the jRCaMP1b backbone, then transfected into HEK293 cells using lipofectamine transfection. Forty-eight hours post-transfection, cells are time-course imaged using an epifluorescent microscope. Visual representations of the quantifications in **B.** are found on the representative response trace. Figure made using Biorender.

B. The scatter plot depicts the average fluorescent responses (*Eq.1*) and percent decay after 5 seconds (*Eq.3*) from each variant of jRCaMP1b expressed in HEK293 cells and stimulated with ten μM acetylcholine. Heat mapping indicates the baseline fluorescence intensity of the variant. Dashed lines depict the mean performance of the base construct, jRCaMP1b. Constructs with large $\Delta F/F_0$ and fast decays are indicated with lines and text labels.

C. The image panel depicts the responses of parental (jRCaMP1b) and two high $\Delta F/F_0$ variants (T62Q and A309L) before (Baseline) and after (+10μM Acetylcholine) stimulus. $\Delta F/F_0$ of the image was calculated pixel-by-pixel, and the values of the heat mapping are indicated in the color bar beneath the image panel. The scale bar depicts 200 μm.

D. The time course depicts representative traces of the decay kinetics of the parental construct (jRCaMP1b), a slow variant (A351I), and a fast variant (I419W). The $\Delta F/F_0$ responses were normalized so that the maximum response was 100%. Each trace indicates a single cell.

Mutations that were predicted to affect sensor responses were cloned into the parental jRCaMP1b, and the functionality of each variant was tested using a stimulus of acetylcholine, which activates calcium channels in the endoplasmic reticulum[73,74] (**Figure 3.5A**). We tested 40 variants predicted to alter the sensor's biophysical characteristics and quantified each construct's baseline fluorescence, max $\Delta F/F_0$, and percent decay within 5 seconds after max $\Delta F/F_0$ (**Figure 3.5, Table 3.3,3.4,3.5**).

We found multiple promising candidates for each biophysical property, including T62Q and A351I, which displayed a ~ 22% and ~17% increase, respectively, in $\Delta F/F_0$ as compared to the jRCaMP1b (**Figure 3.5B,C; Table 3.3**). We additionally observed variants that dramatically shift the speed of decay of jRCaMP1b. For example, the signals from variants G54I, I419W, and G54M decayed 3x faster than those of jRCaMP1b (**Figure 3.5B,D; Table 3.5**). These responses also match the predictions made by the kinetics ensemble. Conversely, signals from variants such as A351I, L298T, and W43D decayed nearly 2.8x times slower than jRCaMP1b (**Figure 3.5B,D; Table 3.5**). The predictions made for A351I and L298T matched what was observed *in vitro*; however, the observed response from W43D was the opposite of what was predicted by the model **(Figure 3.4H)**. We additionally observe variants that display impressive baseline fluorescence, including A59N, which was the correct prediction from a baseline fluorescence model (data not shown, **Table 3.4**). Within this screen, we have found many promising variants for the biophysical properties we are hoping to target within our jRCaMP1b construct, including improved $\Delta F/F_0$ as

well as altered speeds of decay. However, we do not observe any variants that display both fast speeds of decay as well as improved dynamic ranges (**Figure 3.5B**).

Table 3.3 Descriptive Statistics of the $\Delta F/F_0$ of Each jRCaMP1b Variant in the Acetylcholine Assay

Variant	Mean ΔF/F (%)	95% Confidence Interval	Sample Number
T62Q	147.66	['140.79,154.36']	185
A351I	142.06	['134.07,150.09']	236
A309L	135.87	['127.45,144.34']	180
jRCaMP1b	125.86	['122.04,129.7']	398
A52K	121.68	['117.72,125.7']	264
A59N	118.44	['112.06,124.6']	151
S310D	117.19	['111.87,122.5']	229
A351W	115.62	['110.46,120.69']	263
G368H	107.41	['101.48,113.45']	195
L298T	91.41	['87.74,94.84']	156
D318I	87.55	['83.45,91.72']	294
A309I	82.33	['75.89,88.51']	134
T299M	81.97	['79.19,84.69']	368
H48A	76.1	['73.62,78.45']	359
T299Y	70.52	['65.67,75.24']	235
D318F	69.36	['60.49,78.03']	89
Y112T	65.83	['62.61,69.16']	275
A59G	63.15	['59.24,67.09']	181
A52E	60.69	['56.5,64.95']	222
D318W	59.9	['56.07,63.58']	192
G368E	53.59	['49.05,58.21']	131
A309H	51.68	['44.57,59.11']	108
W43D	50.44	['49.26,51.62']	481
S58V	44.46	['39.28,49.79']	299
A309F	19.02	['18.15,19.93']	523
V329E	15.56	['14.68,16.44']	241
G368S	11.76	['10.57,13.01']	102
N60S	11.18	['10.12,12.28']	357
G54F	9.16	['8.59,9.76']	384
M64Q	9.09	['8.1,10.08']	246
D318A	8.05	['7.34,8.79']	129
Y112G	7.61	['6.94,8.3']	176
N60G	6.25	['5.34,7.2']	524
M326H	5.85	['5.4,6.29']	174
G54M	5.69	['5.1,6.31']	191
T299V	5.41	['5.02,5.81']	343
I419W	4.79	['4.33,5.25']	311
G54I	2.88	['2.48,3.29']	223
G368Q	2.27	['1.93,2.63']	133
M326I	1.72	['1.63,1.82']	256

Table containing the information displayed in Figure 3.5. Table contains the construct (Variant), the mean response (Mean $\Delta F/F_0$ (%)), the 95% confidence interval (95% CI), and the number of samples (Sample Number)).

Table 3.4 Descriptive Statistics of the Baseline Fluorescence of Each jRCaMP1b Variant in the Acetylcholine Assay

Variant	Mean Baseline	95% Confidence Interval	Sample Number
A59N	915.81	['854.19,981.15']	151
W43D	909.11	['880.35,938.75']	481
H48A	867.13	['841.6,893.56']	359
G54F	822.65	['798.72,847.47']	384
A59G	809.73	['778.22,841.8']	181
D318I	715.51	['664.58,769.66']	89
G54M	707.53	['666.27,751.02']	191
T62Q	690.09	['659.87,721.83']	185
G54I	682.87	['655.85,710.35']	223
A351W	679.02	['647.98,710.86']	263
L298T	666.45	['638.9,695.48']	156
Y112T	636.64	['613.32,660.18']	275
jRCaMP1b	628.16	['608.92,647.62']	398
A52E	582.35	['557.98,607.98']	222
A52K	578.03	['558.43,598.23']	264
S58V	571.03	['552.77,590.78']	299
A309L	555.47	['523.91,587.25']	180
A351I	549.55	['523.83,575.75']	236
S310D	537.51	['510.58,565.36']	229
V329E	532.34	['514.19,551.08']	241
G368H	525.29	['500.99,550.94']	195
G368E	521.17	['483.54,561.59']	131
I419W	500.4	['487.88,513.33']	311
D318W	465.13	['451.24,479.42']	294
T299Y	454.97	['438.6,471.62']	235
T299M	454.65	['443.86,466.0']	368
N60G	449.36	['439.49,459.1']	524
A309H	438.82	['405.62,474.28']	108
D318A	427.51	['408.35,447.73']	192
T299V	296.38	['293.89,298.86']	343
N60S	291.59	['288.7,294.45']	357
Y112G	289.32	['285.56,293.18']	176
A309I	286.64	['277.93,295.47']	134
M64Q	269.78	['266.41,273.1']	246
M326H	245.49	['242.57,248.38']	174
A309F	236.59	['235.1,238.09']	523
G368Q	228.25	['225.11,231.58']	133
M326I	228.1	['226.56,229.65']	256
D318F	212.78	['211.23,214.34']	129
G368S	211.45	['209.83,213.14']	102

Table containing the information displayed in Figure 3.5. Table contains the construct (Variant), the mean response (Mean Baseline Fluorescence), the 95% confidence interval (95% CI), and the number of samples (Sample Number)).

Table 3.5 Descriptive Statistics of the Decay Speeds of Each jRCaMP1b Variant in the Acetylcholine Assay

Variant	Mean Decay Within 5 Seconds (%)	95% Confidence Interval	Sample Number
M326I	33.2	['31.14,35.25']	133
G54I	29.56	['27.88,31.33']	154
G368Q	21.87	['19.03,24.78']	60
G54M	20.38	['19.61,21.16']	169
I419W	17.76	['17.09,18.5']	163
M64Q	15.47	['14.69,16.28']	162
M326H	12.21	['11.15,13.37']	120
Y112G	11.81	['11.23,12.43']	133
N60G	11.73	['10.08,13.56']	140
T299V	11.66	['11.11,12.24']	217
G54F	10.64	['10.33,10.96']	286
S58V	10.4	['8.99,11.91']	138
G368S	10.07	['9.08,11.22']	76
N60S	9.43	['8.38,10.61']	165
D318A	8.56	['7.78,9.41']	93
G368E	6.53	['6.12,6.93']	67
T62Q	5.41	['5.22,5.59']	115
A309F	4.8	['4.67,4.93']	348
A52K	4.72	['4.61,4.82']	163
jRCaMP1b	4.66	['4.5,4.83']	213
H48A	4.62	['4.37,4.89']	200
G368H	4.35	['4.1,4.61']	96
A52E	4.3	['4.07,4.53']	94
A309H	4.01	['3.73,4.34']	51
S310D	3.79	['3.63,3.95']	112
V329E	3.01	['2.89,3.13']	160
A309I	2.95	['2.76,3.14']	95
D318W	2.85	['2.74,2.98']	106
D318I	2.81	['2.71,2.9']	129
D318F	2.79	['2.54,3.06']	40
T299M	2.76	['2.68,2.85']	221
T299Y	2.56	['2.41,2.73']	113
A59N	2.36	['2.24,2.48']	80
A351W	2.23	['2.09,2.36']	99
Y112T	2.1	['1.79,2.46']	132
A59G	2.08	['1.97,2.2']	87
A309L	1.9	['1.75,2.06']	63
W43D	1.61	['1.57,1.66']	286
L298T	1.56	['1.41,1.72']	89
A351I	1.05	['0.96,1.15']	92

Table containing the information displayed in Figure 3.5. Table contains the construct (Variant), the mean response (Mean Decay Within 5s (%)), the 95% confidence interval (95% CI), and the number of samples (Sample Number)).

While the results from the acetylcholine assay provide insight into dynamic ranges and kinetic responses of the jRCaMP1b variants, we are unable to understand critical biophysical

properties such as sensitivity. Without this information, we cannot validate predictions made on the 1AP and 10AP ensembles within our HEK293 assay. To validate these predictions, we instead transfected these constructs into primary cortical neurons, where we employed a dual optogenetic approach to validate our variants. We multiplexed the light-gated ion channel, Channelrhodopsin[93] (ChR2), alongside our jRCaMP1b variants, such that brief pulses of 488 nm light would induce Ca^{2+} dependent changes in the fluorescence of our jRCaMP1b variants (**Figure 3.6A,B**). We added increasing light stimulus of 1 flash, 10 flashes, and 80 flashes to understand our variants' sensitivity before adding 40mM potassium chloride (KCl) to fully saturate the sensor (**Figure 3.6B**).

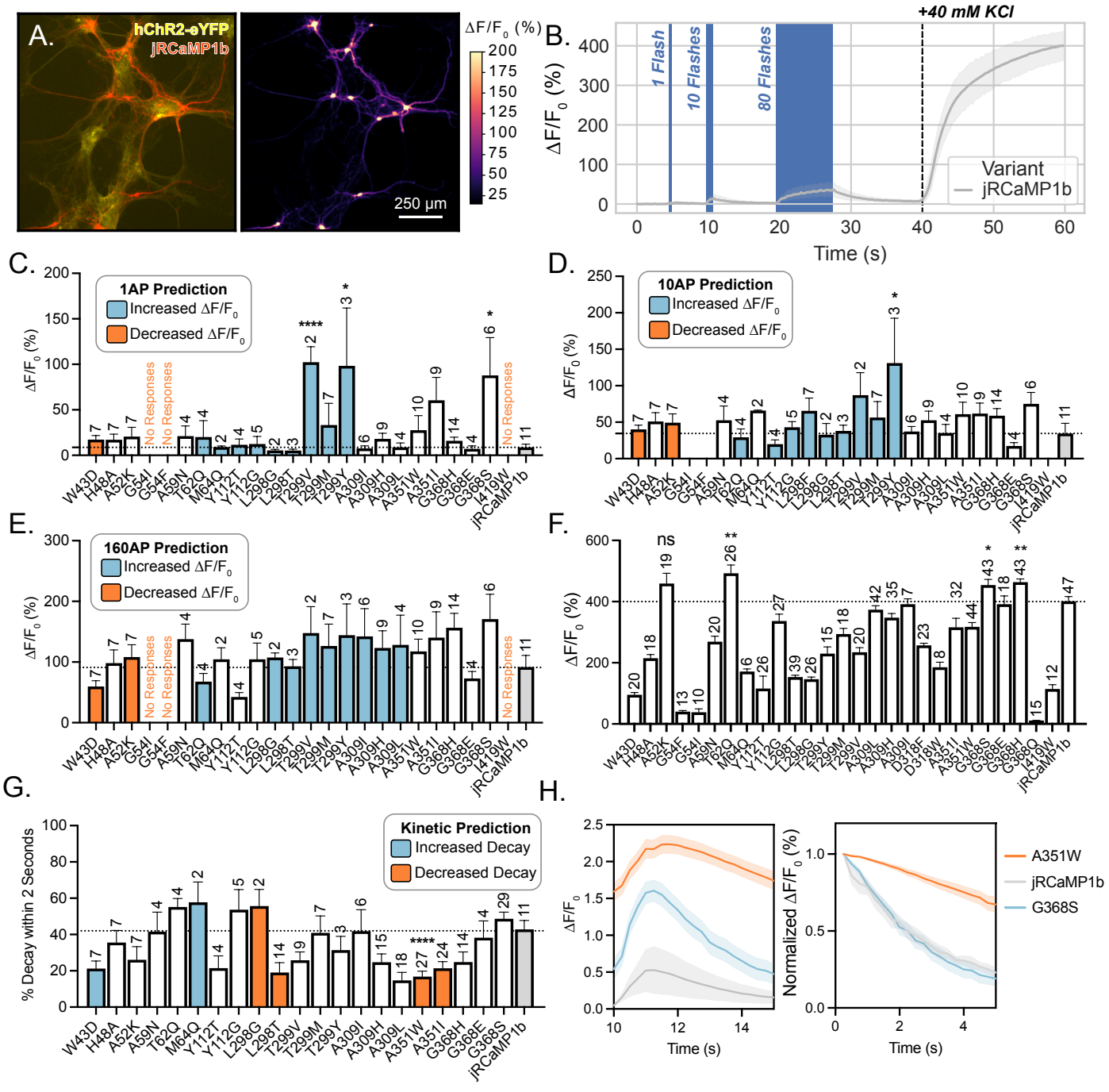


Figure 3.6 Validation of Ensemble Predictions Using Multiplexed Optogenetics in Primary Cortical Neurons

A. The left image panel displays the co-expression of hChR2-YFP (yellow) and jRCaMP1b (Red) in primary cortical neurons. The right panel depicts the change in fluorescence ($\Delta F/F_0$) of the neurons displayed in the left panel. Scale bar = 250 µm.

B. The time course displays the stimulus paradigm of the primary cortical neurons. Briefly, five ms light pulses were administered at 5 seconds, 10 seconds, and 20 seconds at a rate of 10 Hz,

where applicable. After 40 seconds, 40 mM KCl was added to fully saturate the sensor. Line depicts the mean responses, and shading depicts the SEM.

C. The bar diagram demonstrates the average ΔF/F responses to one flash of 488 nm light, where the bar height corresponds to the average response, and the error bars depict the SEM. The text above each bar indicates the sample size and the stars indicate the significance. **** = P<0.0001, * = P<0.05. Colored bars demonstrate ensemble predictions, where the coloring is described within the legend. The dotted line indicates the average response from the parental construct jRCaMP1b.
D. The bar diagram demonstrates the average $\Delta F/F_0$ responses to ten flashes of 488 nm light, where the bar height corresponds to the average response, and the error bars depict the SEM. The text above each bar indicates the sample size. Colored bars demonstrate ensemble predictions, where the coloring is described within the legend. The dotted line indicates the average response from the parental construct jRCaMP1b.
E. The bar diagram demonstrates the average $\Delta F/F_0$ responses to eighty flashes of 488 nm light, where the bar height corresponds to the average response, and the error bars depict the SEM. The text above each bar indicates the sample size. Colored bars demonstrate ensemble predictions, where the coloring is described within the legend. The dotted line indicates the average response from the parental construct jRCaMP1b.
F. The bar diagram demonstrates the average $\Delta F/F_0$ responses to 40 mM KCl, where the bar height corresponds to the average response, and the error bars depict the SEM. The text above each bar indicates the sample size and the stars indicate the significance. **** = P<0.0001, * = P<0.001. Colored bars demonstrate ensemble predictions, where the coloring is described within the legend. The dotted line indicates the average response from the parental construct jRCaMP1b.
G. The bar diagram demonstrates the % decay of the $\Delta F/F_0$ after ten flash stimuli, where the bar height corresponds to the average response, and the error bars depict the SEM. The text above each bar indicates the sample size and the stars indicate the significance. **** = P<0.0001. Colored bars demonstrate ensemble predictions, where the coloring is described within the legend. The dotted line indicates the average response from the parental construct jRCaMP1b.
H. The left time course demonstrates the $\Delta F/F_0$ response to 10 flash stimuli, and the right panel is the same data with the maximum responses normalized to a max of 1.0. The color of each line is indicative of the variant tested (found on the right side of the time courses). Line depicts the mean responses, and shading depicts the SEM.

At one flash stimulus, we found that variants T299V, T299Y, and G368S displayed 10x larger average responses than those of jRCaMP1b (**Figure 3.6C, Table 3.6**). These results match the prediction from the 1AP model, whereas variants such as G54F, G54I, and I419W, predicted to decrease $\Delta F/F_0$, showed no responses to 1 flash stimulus. The variant W43D again displayed the opposite response to the model's prediction, outperforming jRCaMP1b to 1AP stimulus. This mirrors what was observed in the HEK293 assay with the slow decay speeds of W43D but may

indicate an improvement in the sensitivity of this variant. At 10 and 80 flashes of 488 nm light, T299V, T299Y, and G368S continue to outperform jRCaMP1b (**Figure 3.6D,E, Table 3.7**), though to a lesser extent than what was observed at one flash stimulus. This phenomenon indicates a higher affinity for these variants to calcium, as we see that the overall dynamic range of these variants is comparable to, if not lower, than those of jRCaMP1b (**Figure 3.6F**). At the 80 flashes stimulus, we found that 11/14 variants responses matched the 160AP model predictions, including W43D, which is likely fully saturated at this stimulus, achieving lower $\Delta F/F_0$s than jRCaMP1b (**Figure 3.6F, Table 3.8, Table 3.9**).

Table 3.6 Descriptive Statistics of $\Delta F/F_0$ Responses of jRCaMP1b Variants in Cultured Neurons to Single Flash Stimulus

Variant	Sample Size	Mean	Standard Deviation	SEM
T299V	2	102.3	24.04	17
T299Y	3	98.4	110.1	63.57
G368S	6	87.86	101.6	41.47
A351I	9	60.46	75.59	25.2
T299M	7	33.29	63.1	23.85
A351W	10	27.54	50.98	16.12
A59N	4	21.06	22.58	11.29
A52K	7	20.68	26.56	10.04
T62Q	4	20.14	35.93	17.96
A309H	9	18.12	18.3	6.101
W43D	7	17.34	11.88	4.491
H48A	7	17.1	16.28	6.155
G368H	14	15.86	15.31	4.091
Y112G	5	12.35	18.92	8.462
Y112T	4	11.67	12.41	6.205
M64Q	2	9.131	1.625	1.149
A309L	4	8.775	9.601	4.801
jRCaMP1b	11	8.683	11.97	3.609
A309I	6	7.539	7.024	2.868
G368E	4	7.161	6.62	3.31
L298G	2	5.437	1.708	1.208
L298T	3	5.314	2.464	1.423

Table containing the information displayed in Figure 3.6. The table includes the construct (Variant), the mean response, the sample size, the mean response to 1 flash of 488 nm light, the standard deviation, and the standard error of the mean.

Table 3.7 Descriptive Statistics of $\Delta F/F_0$ Responses of jRCaMP1b Variants in Cultured Neurons to Ten Flash Stimulus

Variant	Sample Size	Mean	Standard Deviation	SEM
T299Y	3	131.1	106.3	61.36
T299V	2	86.95	43.71	30.91
G368S	6	75.25	38.25	15.62
M64Q	2	66	0.8399	0.5939
L298F	7	65.58	46.3	17.5
A351I	9	61.86	43.17	14.39
A351W	10	60.95	51.95	16.43
G368H	14	58.98	36.54	9.767
T299M	7	56.33	58.26	22.02
A59N	4	52.68	39.18	19.59
A309H	9	52.62	37.86	12.62
H48A	7	50.93	32.31	12.21
A52K	7	49.26	31.86	12.04
Y112G	5	43.07	17.09	7.644
W43D	7	40.23	15.27	5.772
L298T	3	38.07	13.51	7.802
A309I	6	37.31	16.5	6.736
A309L	4	35.04	23.77	11.88
jRCaMP1b	11	34.58	46.03	13.88
L298G	2	32.92	21.77	15.39
T62Q	4	29.28	22.81	11.41
Y112T	4	19.82	12.01	6.004
G368E	4	17.4	9.102	4.551

Table containing the information displayed in Figure 3.6. The table includes the construct (Variant), the mean response, the sample size, the mean response to 10 flashes of 488 nm light, the standard deviation, and the standard error of the mean.

Table 3.8 Descriptive Statistics of $\Delta F/F_0$ Responses of jRCaMP1b Variants in Cultured Neurons to Eighty Flash Stimulus

Variant	Sample Size	Mean	Standard Deviation	SEM
G368S	6	170.8	100.5	41.02
L298F	7	166	99.53	37.62
G368H	14	156.4	89.33	23.87
T299V	2	147.7	61.64	43.59
T299Y	3	144.3	89.3	51.56
A309I	6	142.3	112.2	45.81
A351I	9	139.9	127.6	42.53
A59N	4	137.7	49.76	24.88
A309L	4	128.3	98.07	49.03
T299M	7	126.4	95.82	36.22
A309H	9	123.2	85	28.33
A351W	10	117.5	64.18	20.3
A52K	7	108.2	53.83	20.34
L298G	2	107.4	10.66	7.536
Y112G	5	104.5	59.75	26.72
M64Q	2	104.4	26.83	18.97
H48A	7	98.23	57.61	21.78
L298T	3	93.05	19.52	11.27
jRCaMP1b	11	91.71	63.47	19.14
G368E	4	73.05	22.68	11.34
T62Q	4	67.72	26.91	13.45
W43D	7	59.56	25.39	9.598
Y112T	4	42.46	14.21	7.106

Table containing the information displayed in Figure 3.6. The table includes the construct (Variant), the mean response, the sample size, the mean response to 80 flashes of 488 nm light, the standard deviation, and the standard error of the mean.

Table 3.9 Descriptive Statistics of $\Delta F/F_0$ Responses of jRCaMP1b Variants in Cultured Neurons to 40 mM KCl

Variant	Sample Size	Mean	Standard Deviation	SEM
T62Q	26	492.5	139.7	27.4
G368H	43	463.5	69.97	10.67
A52K	19	458.9	147.4	33.82
G368S	43	454.3	120.2	18.33
jRCaMP1b	47	400.9	107.4	15.66
G368E	18	391.7	113.5	26.76
A309I	7	391.4	47.01	17.77
A309L	42	373.4	85.92	13.26
A309H	35	347.5	81.74	13.82
Y112G	27	336.6	117.1	22.54
A351W	44	317.6	91.56	13.8
A351I	32	316	166	29.34
T299M	18	294.2	77.5	18.27
A59N	20	268.9	81.28	18.17
D318F	23	257.1	32.79	6.838
T299V	20	234.2	66.93	14.97
T299Y	15	230.2	83.48	21.56
H48A	18	214.7	52.2	12.3
D318W	8	185.5	47.13	16.66
M64Q	6	171.6	20.57	8.396
L298T	39	153.3	36.92	5.912
L298G	26	146.4	35.32	6.927
Y112T	26	116	205.2	40.24
I419W	12	114.3	48.61	14.03
W43D	20	95.54	31.62	7.071
G54F	13	40.51	11.51	3.191
G54I	10	38.32	34.25	10.83
G368Q	15	11.24	2.945	0.7604

Table containing the information displayed in Figure 3.6. The table includes the construct (Variant), the mean response, the sample size, the mean response to 40 mM KCl, the standard deviation, and the standard error of the mean.

To quantify the decay kinetics of our variants, we normalized each signal such that the maximal response to 10 flashes is 1. We then calculated the difference between the maximum and that of the signal two seconds after the max to obtain the decay within two seconds. Using this metric, we found that many of the variants predicted to slow the decay of jRCaMP1b, such as A351I, A351W, and L298T, all decreased the speed of decay (**Figure 3.6G, Table 3.10**). Conversely, many of the constructs predicted to speed the rate of decay, such as G54I, G54F, and I419W, did not show any responses to 10 flash stimuli and, therefore, could not be included in this

analysis. Regardless, we still found two promising variants, M64Q and G368S, that demonstrate improved decay kinetics compared to jRCaMP1b, though not significantly faster (**Figure 3.6G, Table 3.10**). Of these two, G368S is maintains larger $\Delta F/F_0$ responses to 10 flash stimuli while maintaining similar decay kinetics to those of jRCaMP1b (**Figure 3.6G,H, Table 3.10**).

From our cultured neuron screening, we were able to validate our ensemble predictions and outline multiple promising variants such as T299Y and T229V, which display significantly larger $\Delta F/F_0$s to 1 flash stimulus, A351I and A351W that display slow decay kinetics, and G368S, which has large $\Delta F/F_0$ responses to each stimulus while maintaining fast decay kinetics. We additionally prove the multiplexing capability of these variants alongside blue light-activated Channelrhodopsins. Downstream, these variants will be further characterized both *in vitro* and in purified protein.

Table 3.10 Descriptive Statistics of Decay Speeds of jRCaMP1b Variants in Cultured Neurons

Variant	Sample Size	Mean	Standard Deviation	SEM
M64Q	2	57.78	15.77	11.15
L298G	2	55.68	13.01	9.201
T62Q	4	55.16	9.322	4.661
Y112G	5	53.63	24.88	11.12
G368S	29	48.66	19.9	3.695
jRCaMP1b	11	42.85	16.23	4.892
A309I	6	41.86	29.01	11.84
A59N	4	41.58	21.57	10.79
T299M	7	40.81	24.72	9.343
G368E	4	38.27	18.53	9.263
D318W	5	37.33	26.63	11.91
H48A	7	35.45	17.87	6.754
T299Y	3	31.36	13.2	7.622
A52K	7	26.09	19.22	7.265
T299V	9	25.84	13.68	4.559
G368H	14	24.8	21.09	5.638
A309H	15	24.69	17.98	4.642
Y112T	4	21.47	13.54	6.772
A351I	24	21.37	18.01	3.675
W43D	7	21.19	11.04	4.174
L298T	14	19.04	20.28	5.421
D318F	23	18.76	11.87	2.476
A351W	27	16.78	15.53	2.989
A309L	18	14.72	18.72	4.413

Table containing the information displayed in Figure 3.6. The table includes the construct (Variant), the mean response, the sample size, the mean decay after two seconds to 10 flash stimuli, the standard deviation, and the standard error of the mean.

3.2.3 The Effect of Combinatorial Mutagenesis on jRCaMP1b Performance

During our eGCaMP engineering, we found that combining Q305D with L317H led to impressive capabilities that we dubbed our eGCaMP^{2+} variant. Toward this similar goal, we combined multiple mutations, both from positive results from our HEK293 cell screen and through some rational engineering, with insights from our eGCaMP study. We focused on residue locations such as A52, T62, Q297, A309, M345, A351, and G368. We chose the mutations A52V, A309H, and Q297D not only for their essential locations on the jRCaMP1b protein but also because these are homologous mutations for improvements to the GCaMP sensor (**Figure 3.7A**). A52V was one of the mutations that led to improved sensitivity of the jGCaMP7s construct[95], and the A309H and

Q297D mutations are found on eGCaMP and eGCaMP^{2+}[97]. We chose the mutation T62Q due to its large dynamic range in both our HEK293 assays as well as our primary cortical neuron assays (**Figure 3.5, 3.6**). We chose variants G368Q and A351W for their kinetic capabilities in the HEK293 assay (**Figure 3.5**). M345L was selected due to its impressive abilities that we found in our high throughput screening assays (Chapter 3.2.4).

In our preliminary HEK293 screening assay, we found that T62Q A351W had an $\Delta F/F_0$ increase of 194%, nearly doubling the response from jRCaMP1b, which achieved 85.9% increase (**Figure 3.7B**). A higher $\Delta F/F_0$ ability was also found within T62Q G368Q and T62Q M345L, which achieved 130% and 120% respectively (**Figure 3.7B**). Notably, the A52V mutation had little effect on the $\Delta F/F_0$ abilities of our variants, and we do not see an improvement in $\Delta F/F_0$ within the A52V A309H Q297D variant that is the hallmark of eGCaMP^{2+}. We also do not observe a decreased baseline fluorescence, which is common with our eGCaMP variants (**Figure 3.7C**). To that effect, each tested combinatorial variant displayed baseline fluorescence comparable to or greater than that of jRCaMP1b (**Figure 3.7C**). None of the combinatorial variants were able to achieve the speed of decay that exceeds that of jRCaMP1b, T62Q A351W mimics the speed of decay of the A351W solo mutation, which is significantly slower than that of jRCaMP1b (**Figure 3.7D**).

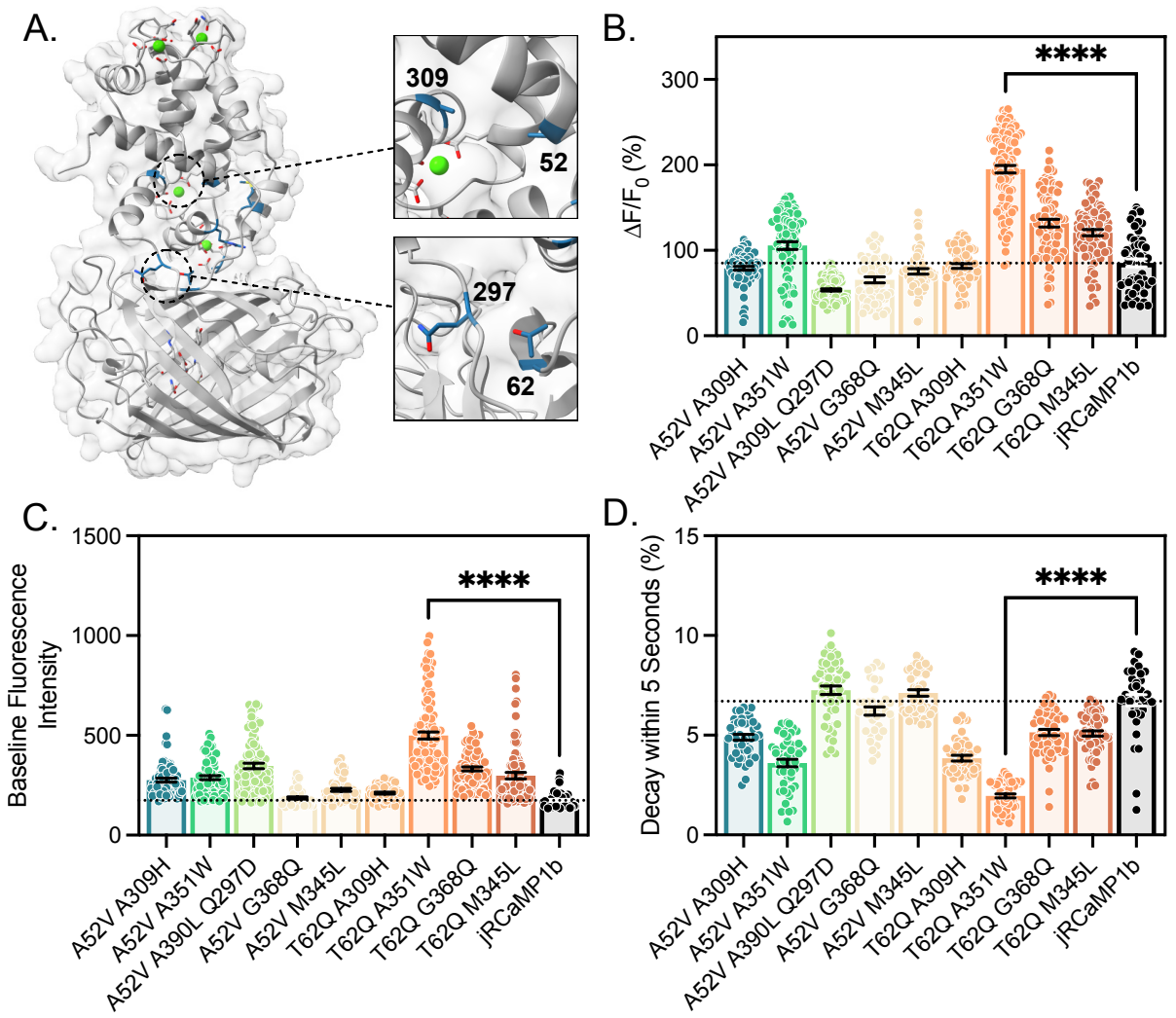


Figure 3.7 Effects of Combinatorial Mutations on jRCaMP1b

A. Crystal structure of RCaMP1h (RCSB: 3U0K, gray) with the residues targeted for combinatorial mutation highlighted (blue). The insets display the proximity between amino acids, and text labels indicate the residue number.

B. The bar diagram demonstrates each variant's average $\Delta F/F_0$ response to 10μM acetylcholine addition, where the bar height corresponds to the average response and the error bars depict the SEM. The scatter plots depict each of the analyzed cells. The dotted line indicates the average response from the parental construct jRCaMP1b. **** = P<0.0001.

C. The bar diagram demonstrates each variant's baseline fluorescence, where the bar height corresponds to the average baseline, and the error bars depict the SEM. The scatter plots depict each of the analyzed cells. The dotted line indicates the average response from the parental construct jRCaMP1b. **** = P<0.0001.

D. The bar diagram demonstrates each variant's average decay within 5 seconds after the maximum, where the bar height corresponds to the average decay, and the error bars depict the SEM. The scatter plots depict each of the analyzed cells. The dotted line indicates the average response from the parental construct jRCaMP1b. **** = P<0.0001.

We then transitioned these variants into primary cultured neurons. Similar to our solitary mutation screening, we underwent a multiplexing optogenetic approach, where we co-expressed each variant with ChR2 to drive action potentials within our cultures. The cells were exposed to the same stimulation protocol, including 5ms flashes of 488 nm light, administered at one flash, ten flashes (10 Hz), 80 flashes (10 Hz), and 40 mM KCl(**Figure 3.8A**). After single flash administration, we found that many of our combinatorial variants achieved $\Delta F/F_0$ responses greater than those of jRCaMP1b. This includes variants A52V A309H, A52V M345L, and T62Q G368Q, which achieved average responses of 74%, 34%, and 30%, respectively, compared to jRCaMP1b that achieved 8.6% (**Figure 3.8B, Table 3.11**). These responses were consistent at the 10-flash stimulus, where A52V A309H continued to display the greatest $\Delta F/F_0$ response, nearly 4x greater than that of jRCaMP1b (**Figure 3.8C, Table 3.11**). At the 80-flash and 40 mM KCl stimulus, the maximum responses were achieved by T62Q A351W (**Figure 3.8D,E, Table 3.11, Table 3.12**). We hypothesize that the dynamic range of the A52V A309H variant is not greater than that of jRCaMP1b, but that the sensitivity is greater, allowing the transition to occur at lower stimulus ranges. Conversely, T62Q A351W displays a large dyanamic range, but a similar activation profile to jRCaMP1b.

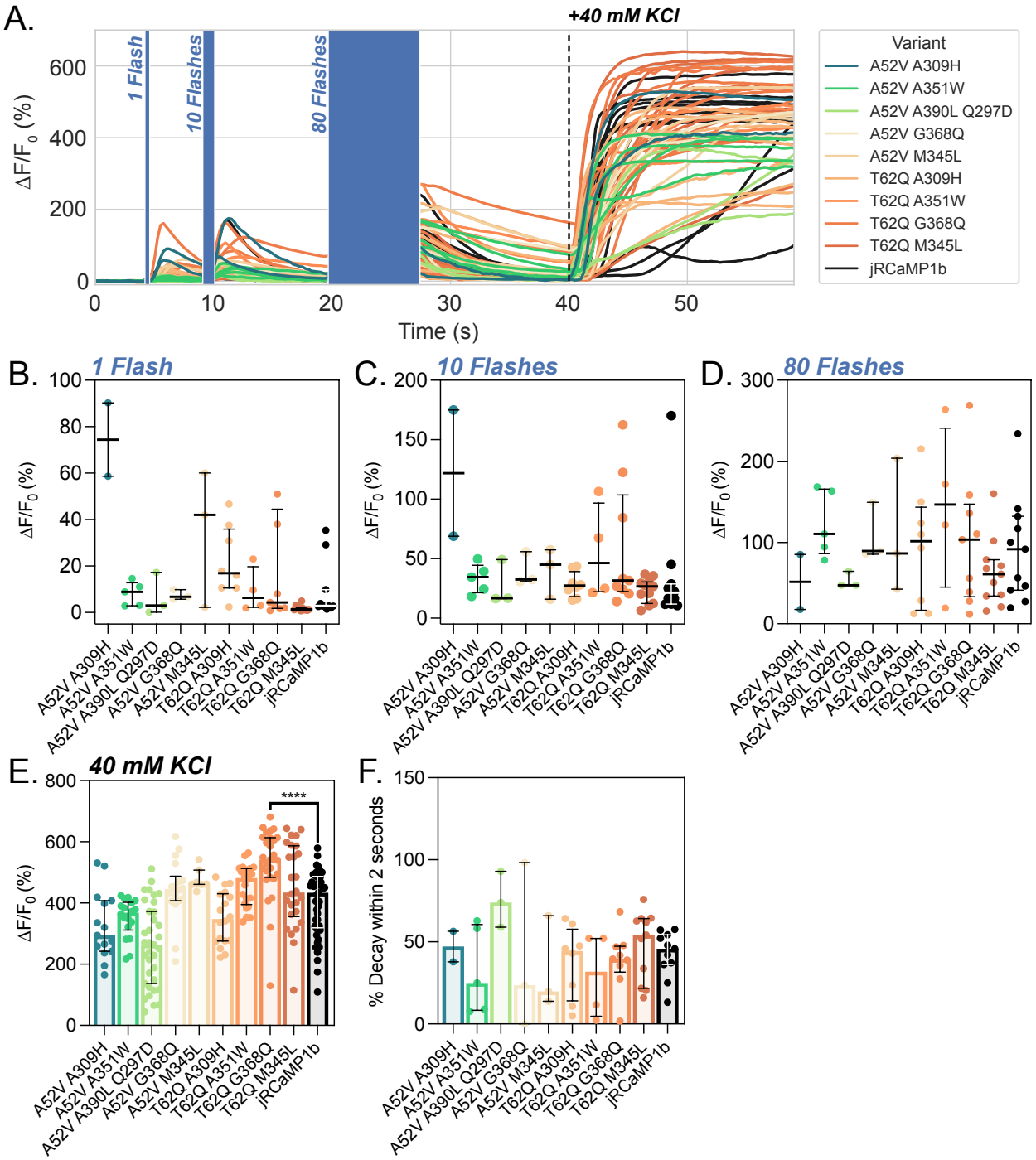


Figure 3.8 Effects of Combinatorial Mutations on jRCaMP1b

A. The time course displays the stimulus paradigm of the primary cortical neurons. Briefly, five ms light pulses were administered at 5 seconds, 10 seconds, and 20 seconds at a rate of 10 Hz, where applicable. After 40 seconds, 40 mM KCl was added to fully saturate the sensor. Each line depicts the responses to a single cell, and the corresponding variant color key is located on the righthand side of the graph.

B. The scatter plot demonstrates the average $\Delta F/F_0$ responses to one flash of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response.
C. The scatter plot demonstrates the average $\Delta F/F_0$ responses to 10 flashes of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response.
D. The scatter plot demonstrates the average $\Delta F/F_0$ responses to eighty flashes of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response.
E. The bar plot demonstrates the average $\Delta F/F_0$ responses to 40 mM KCl, where the bar height indicates the mean and the error bars indicate the SEM. Each dot indicates a single cell's response. **** = $P<0.0001$.
F. The bar plot demonstrates the average decay within 2 seconds after 10 flash stimuli, where the bar height indicates the mean and the error bars indicate the SEM. Each dot indicates a single cell's response.

In the decay calculations, none of the tested variants display decay profiles faster than jRCaMP1b. However, we do see that T62Q A351W and T62Q G368Q tended to have slower decay speeds than jRCaMP1b, which, paired with their improved $\Delta F/F_0$ at 1-flash and 10-flash stimulus, is consistent with the increased sensitivity of these variants (**Figure 3.8F**, **Table 3.11**). The results from the cultured neurons approximate the decay kinetics that we observed within our acetylcholine screen. Therefore, it is not surprising that none of these variants outperformed jRCaMP1b, though we do see that A52V A309H maintained similar decay kinetics while simultaneously achieving >4x $\Delta F/F_0$ responses to 1-flash and 10-flash stimulus (**Figure 3.8B,C,F**, **Table 3.11**).

With these variants, A52V A309H should be verified for its sensitivity comparable to our T299 variants. Other promising variants that require further validation include T62Q A351W and T62Q G368Q, which display large dynamic ranges. We will perform *in vivo* testing and characterization in purified protein with these variants.

Table 3.11 Descriptive Statistics of $\Delta F/F_0$ Responses of jRCaMP1b Variants in Cultured Neurons

		1 AP		
Variant	Sample Size	Mean	Std. Deviation	Std. Error of Mean
A52V A309H	2	74.42	22.32	15.78
A52V A351W	5	8.079	5.118	2.289
A52V A390L Q297D	3	6.792	9.217	5.321
A52V G368Q	3	7.425	2.12	1.224
A52V M345L	3	34.77	29.57	17.07
T62Q A309H	8	21.57	15.26	5.394
T62Q A351W	4	9.418	9.707	4.854
T62Q G368Q	9	30.03	52.47	17.49
T62Q M345L	11	1.852	1.222	0.3684
jRCaMP1b	11	8.683	11.97	3.609

		10 AP		
Variant	Sample Size	Mean	Std. Deviation	Std. Error of Mean
A52V A309H	2	121.9	75	53.04
A52V A351W	5	33.25	12.31	5.507
A52V A390L Q297D	3	27.46	18.81	10.86
A52V G368Q	3	39.66	14.03	8.098
A52V M345L	3	39.47	21.26	12.28
T62Q A309H	8	28.2	10.5	3.711
T62Q A351W	4	55.1	40.17	20.09
T62Q G368Q	9	57.79	53.09	17.7
T62Q M345L	11	23.38	10.15	3.059
jRCaMP1b	11	34.58	46.03	13.88

		80 AP		
Variant	Sample Size	Mean	Std. Deviation	Std. Error of Mean
A52V A309H	2	51.59	47.88	33.85
A52V A351W	5	123.1	40.77	18.23
A52V A390L Q297D	3	53.13	10	5.776
A52V G368Q	3	108.3	35.92	20.74
A52V M345L	3	111	83.34	48.12
T62Q A309H	8	93.32	72.29	25.56
T62Q A351W	4	144.3	102	50.98
T62Q G368Q	9	101.4	80.64	26.88
T62Q M345L	11	63.53	41.33	12.46
jRCaMP1b	11	91.71	63.47	19.14

		Decay		
Variant	Sample Size	Mean	Std. Deviation	Std. Error of Mean
A52V A309H	2	47.14	13.21	9.34
A52V A351W	5	32.49	26.51	11.86
A52V A390L Q297D	3	75.3	17.05	9.845
A52V G368Q	3	40.63	51.37	29.66
A52V M345L	3	33.18	28.56	16.49
T62Q A309H	8	37.62	22.16	7.834
T62Q A351W	4	29.52	26.25	13.12
T62Q G368Q	9	38.69	17.81	5.938
T62Q M345L	11	47.9	21.1	6.361
jRCaMP1b	11	42.61	13.96	4.208

Tables containing the information displayed in Figure 3.8. The table includes the construct (Variant), the sample size, the mean response to the indicated stimulus, the standard deviation, and the standard error of the mean.

Table 3.12 Descriptive Statistics of KCL Responses of jRCaMP1b Variants in Cultured Neurons

		KCl		
Variant	Sample Size	Mean	Std. Deviation	Std. Error of Mean
A52V A309H	15	324.5	110.3	28.47
A52V A351W	18	351.3	63.19	14.89
A52V A390L Q297D	35	262.6	132.9	22.47
A52V G368Q	18	442.2	99.65	23.49
A52V M345L	7	477	35.32	13.35
T62Q A309H	18	351	86.08	20.29
T62Q A351W	22	457.4	67.85	14.47
T62Q G368Q	27	530.5	115.7	22.26
T62Q M345L	26	453.9	136.6	26.8
jRCaMP1b	47	400.9	107.4	15.66

Tables containing the information displayed in Figure 3.8. The table includes the construct (Variant), the sample size, the mean response to the indicated stimulus, the standard deviation, and the standard error of the mean.

3.2.4 Ensemble Directed High-Throughput Screening of Mutation Libraries

Within our previous study, we observed that residues most commonly predicted by the model to be influential to sensor performance often displayed proximity when mapped back onto the crystal structure and in a biophysical property-dependent manner (**Figure 2.5**). We hypothesized that these locations indicate critical intraprotein interactions that govern the given biophysical characteristic and may serve as advantageous positions for further mutation library formation. The current study has observed the same trend of ensemble predictions displaying proximity when mapped back onto the crystal structure.

Using the predictions on the novel variant library from our trained ensembles, we isolated the top and bottom 2.5% of model predictions and counted the number of times each residue appeared. Using this metric, we can observe patterns in which the ensemble commonly predicts mutations at specific residues to be highly influential over the sensor's functional ability (**Figure 3.9A**). We observed patterns of residues that remain exclusive within a given biophysical characteristic (i.e., residues 351 and 357 in kinetics predictions) (**Figure 3.9A**). Interestingly, residue locations such as 43, 54, 60, 61, 112, and 298 were shared between 1AP and 10 AP predictions. This highlights residues such as 421, which is exclusively represented within the 1AP predictions and remarkably close to the calcium-sensing EF-hand domain, and indicates that his region is essential for the sensitivity of the variant (**Figure 3.9A,B**). Similarly, residues such as 43 and 54 are shared between the 1AP, 10AP, and kinetics predictions, which may be indicative of the inverse correlation often observed between sensitivity and rate of decay[96] (**Figure 3.9A**).

Similar to our previous observations, we found that many of these influential residue sites, when mapped back onto the protein's crystal structure, displayed remarkable proximity in 3D space (**Figure 3.9B,C,D**). From these results, we have determined several mutation libraries to be

formed: a sensitivity library that saturates sites 112, 140, and 298 (**Figure 3.9A,B**), a brightness library that saturates sites 56, 62, 299, and 309 (**Figure 3.9A,**C), and a decay library with saturated sites at 93, 298, 351, and 367 (**Figure 3.9A,D**). Sites were determined not only on cloning feasibility (sites needed to be able to be Gibson assembled) but additionally on exclusive representation within the biophysical characteristic's predictions.

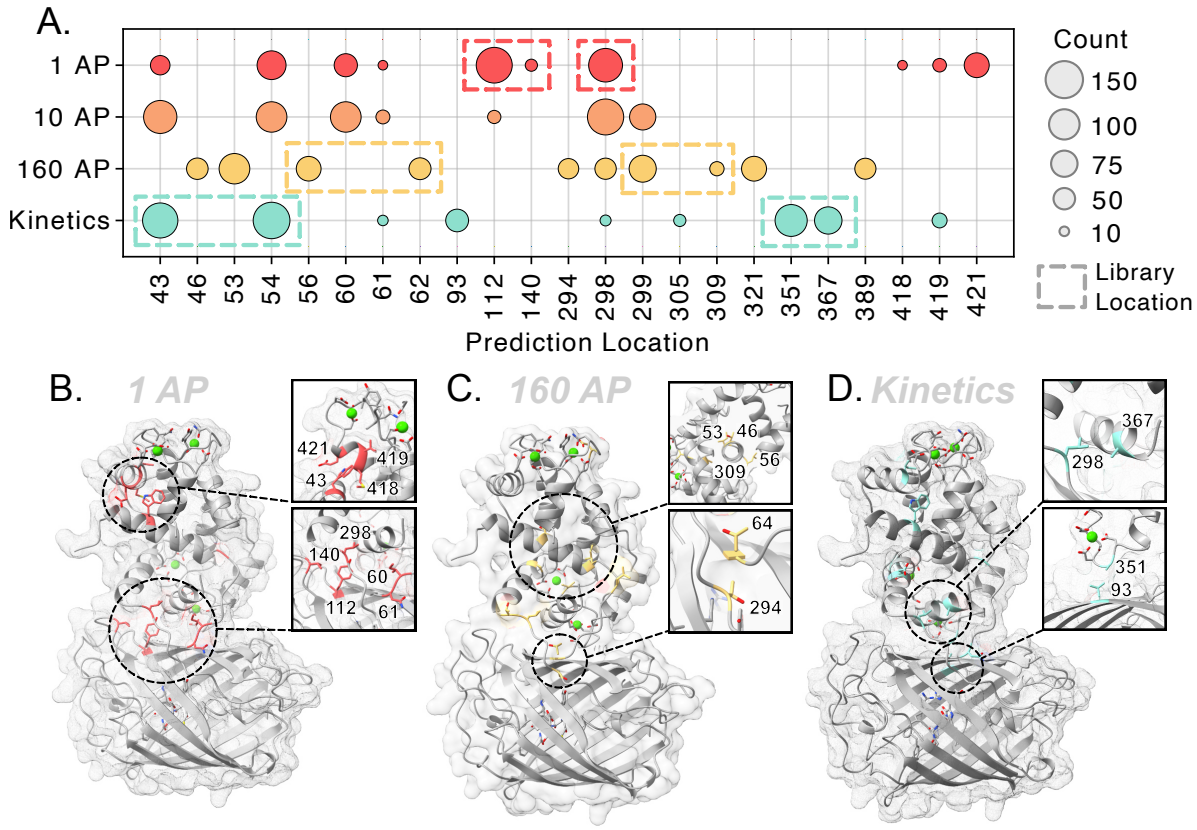


Figure 3.9 Structural Insights Derived from Ensemble Predictions

A. Bubble plot depicting the number of times each residue (x-axis) was referenced within the top and bottom 2.5% of the ensemble's predictions for each biophysical property (y-axis). Dashed boxes indicate sites chosen for mutation library formation.

B. Crystal structure of RCaMP1h (RCSB: 3U0K, gray) with the influential residues from the 1AP predictions (red) highlighted. The inset is used to display the proximity between amino acids, and text labels indicate the residue number.

C. Crystal structure of RCaMP1h (RCSB: 3U0K, gray) with the influential residues from the 160AP predictions (yellow) highlighted. The inset is used to display the proximity between amino acids, and text labels indicate the residue number.

D. Crystal structure of RCaMP1h (RCSB: 3U0K, gray) with the influential residues from the kinetics predictions (teal) highlighted. The inset is used to display the proximity between amino acids, and text labels indicate the residue number.

We leveraged these insights from our ensemble predictions to guide our high-throughput screening efforts. With the influential residues from each biophysical property, we synthesized mutation libraries using degenerate codon PCR and Gibson assembly cloning in a backbone compatible with the landing pad cell line (**Figure 3.10A**). The landing pad cell line contains a singular Tet-inducible Bxb1 recombination site that will irreversibly undergo one recombination event[98]. This enables us to express a library of genetically distinct variants within the same culture of HEK293T cells. Alongside our jRCaMP1b variants, we additionally recombined a calcium-selective opsin called CapChR2[99](**Figure 3.10B**). By incorporating an opsin alongside our jRCaMP1b variants, we can optically introduce calcium into the cell nuancedly. This enables us to gather information such as the sensitivity of each variant during our high-throughput screening. Importantly, because HEK293T are not excitable cells, traditional opsins permeable to multiple cations would not produce a calcium signal to the same effect, which is why we opted for CapChR2.

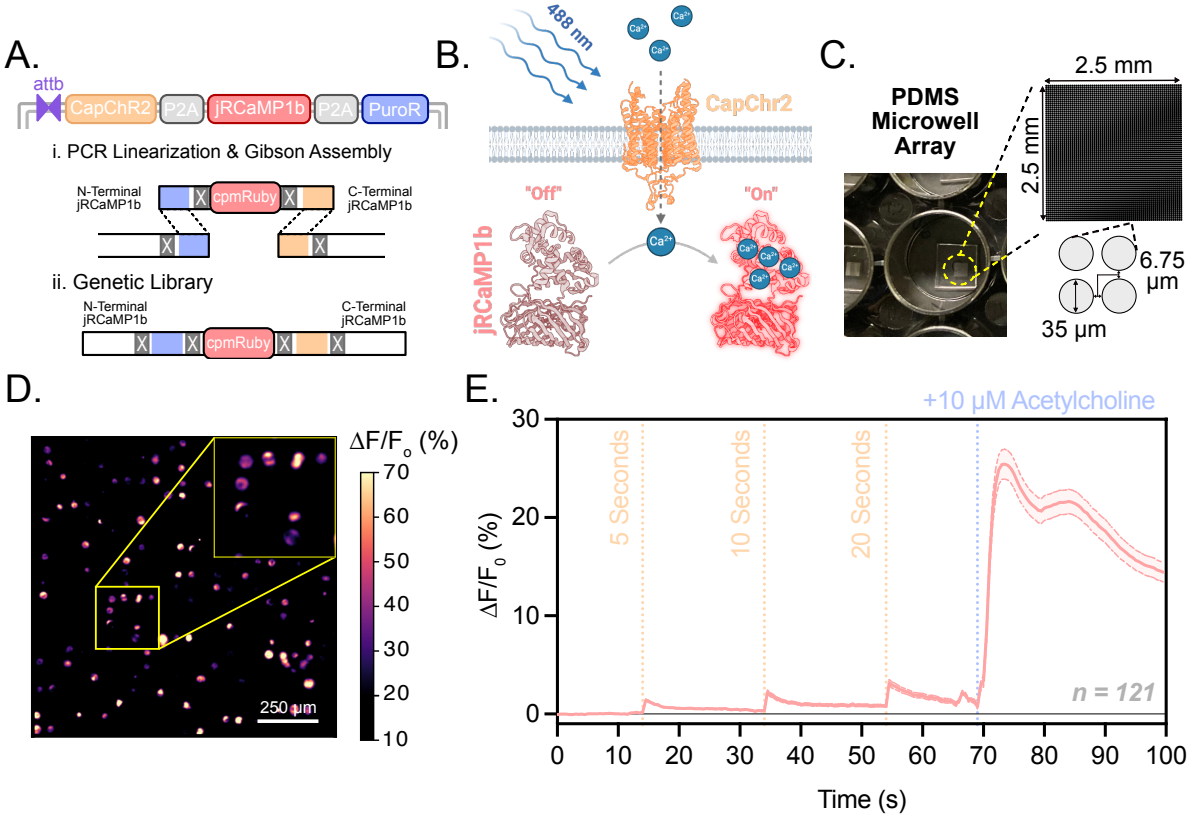


Figure 3.10 Formation of a High-Throughput Screening Paradigm for jRCaMP1b

A. The genetic cassette demonstrates what will be recombined into landing pad cells[98] , and the steps below illustrate the mutagenesis steps that were performed to make genetic variant libraries.

B. The diagram depicts the mechanism of action of CapChR2[99], a calcium-selective channel rhodopsin variant expressed alongside each jRCaMP1b variant. Briefly, pulses of 488 nm light induce calcium permeability across the membrane via CapChR2, which can be sensed by the expressed jRCaMP1b variant.

C. Image displays PDMS microwell array in which recombined library cells are seeded for Opto-MASS screening (figure adapted from Rappleye et al. 2023[40])

D. $\Delta F/F_0$ responses of jRCaMP1b (isotype) recombined in landing-pad cells to 10 µM acetylcholine stimulus after seeding on PDMS microwell arrays. $\Delta F/F_0$ (%) heat-mapped values are indicated on the right-hand side of the image, and the scale bar = 250 µm.

E. Time course displays the parental response of jRCaMP1b (isotype) recombined in landing-pad cells to both CapChR2 stimulus and ten µM acetylcholine. Dotted lines indicate added stimulus, where CapChR2 stimulus is in orange and acetylcholine stimulus is in lavender. (n indicates the number of cells analyzed)

We formed our mutation libraries as previously discussed at the chosen sites and verified their randomization using nanopore sequencing (**Figure 3.11A-C**). These plasmid libraries were then transfected into landing pad cells, where they underwent recombination and selection. With our libraries expressed in landing pad cells, we seed the cells into polydimethylsiloxane (PDMS) microwell arrays so each cell is physically separated and differential responses can be observed (**Figure 3.10C,D**). Within this configuration, we developed an imaging paradigm in which increasing pulse lengths of 488nm light are applied before a saturating amount of acetylcholine is added to the cells (**Figure 3.10E**). We observe a dose-dependent increase in fluorescence to each 488 nm stimulus, indicating the sensitivity or level of activation to low doses of calcium influx. In contrast, the acetylcholine stimulus enables us to see the sensor's dynamic range and kinetics (**Figure 3.10E**).

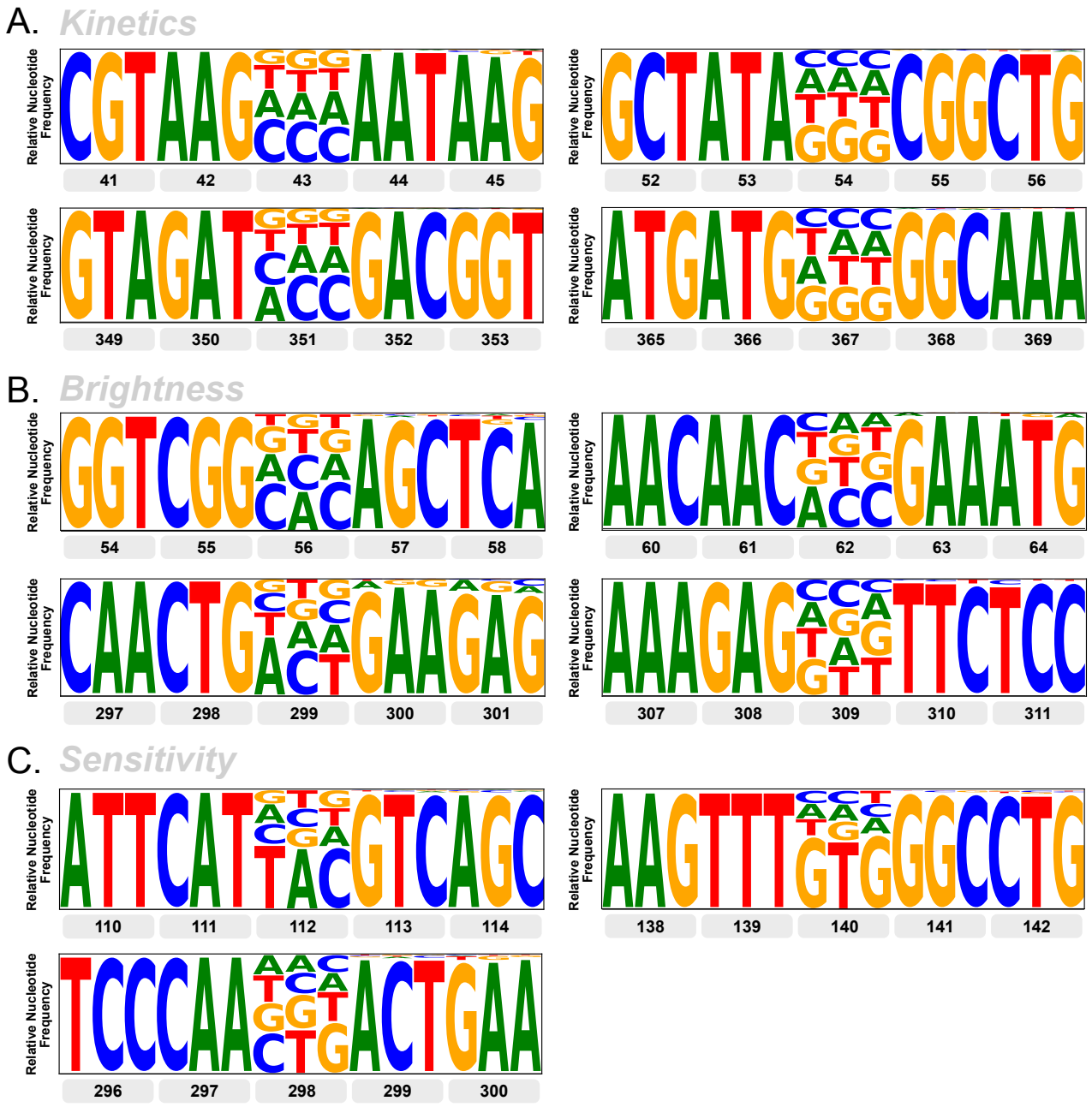


Figure 3.11 Validation of the Randomization Achieved in Each Library

A. Logo maps indicate the relative nucleotide frequency over mutagenesis regions 43, 54, 351, and 367 chosen for the kinetics library.

B. Logo maps indicate the relative nucleotide frequency over mutagenesis regions 56, 62, 299, and 309 chosen for the brightness library (160AP).

C. Logo maps indicate the relative nucleotide frequency over mutagenesis regions 112, 140, and 298 chosen for the sensitivity library (1AP).

While screening our libraries, we observed a broad distribution of $\Delta F/F_0$ capabilities within each library (**Figure 3.12A**). These responses, however, were heavily biased toward reducing the sensor's capabilities compared to jRCaMP1b (**Figure 3.12A**). This demonstrates how rare beneficial mutations are within our sensor's mutational space. Within these screens, we performed two screening paradigms, one in which we incorporated the use of the CapChR2 opsin and one in which we solely tested the acetylcholine stimulus (**Figure 3.12B, C**). We focused on the amplitude of $\Delta F/F_0$ response to CapChr2 stimulation, $\Delta F/F_0$ response to acetylcholine stimulation, and the decay speed after acetylcholine stimulation. We found many variants in both stimulus paradigms that display promising qualities, though we could not identify the genetic differences from every pick (**Figure 3.12B,C)**. Variants B1 and S1 displayed large $\Delta F/F_0$s after acetylcholine stimulus (**Figure 3.12B).** Variants K1 and K2 displayed fast decay speeds after acetylcholine stimulus (**Figure 3.12B).** Confirming the known correlation between sensitivity and kinetics, pick K3 shows the greatest increase after the blue light stimulus while displaying slow decay kinetics after the acetylcholine stimulus (**Figure 3.12B).** Similarly, within our acetylcholine screening paradigm, we observe that S2 (a sensitivity library pick) displays fast decay kinetics (**Figure 3.12C**). Also, within the acetylcholine screening paradigm, we retrieved B2 and B3, which display large $\Delta F/F_0$ responses and slow decay kinetics (**Figure 3.12C**).

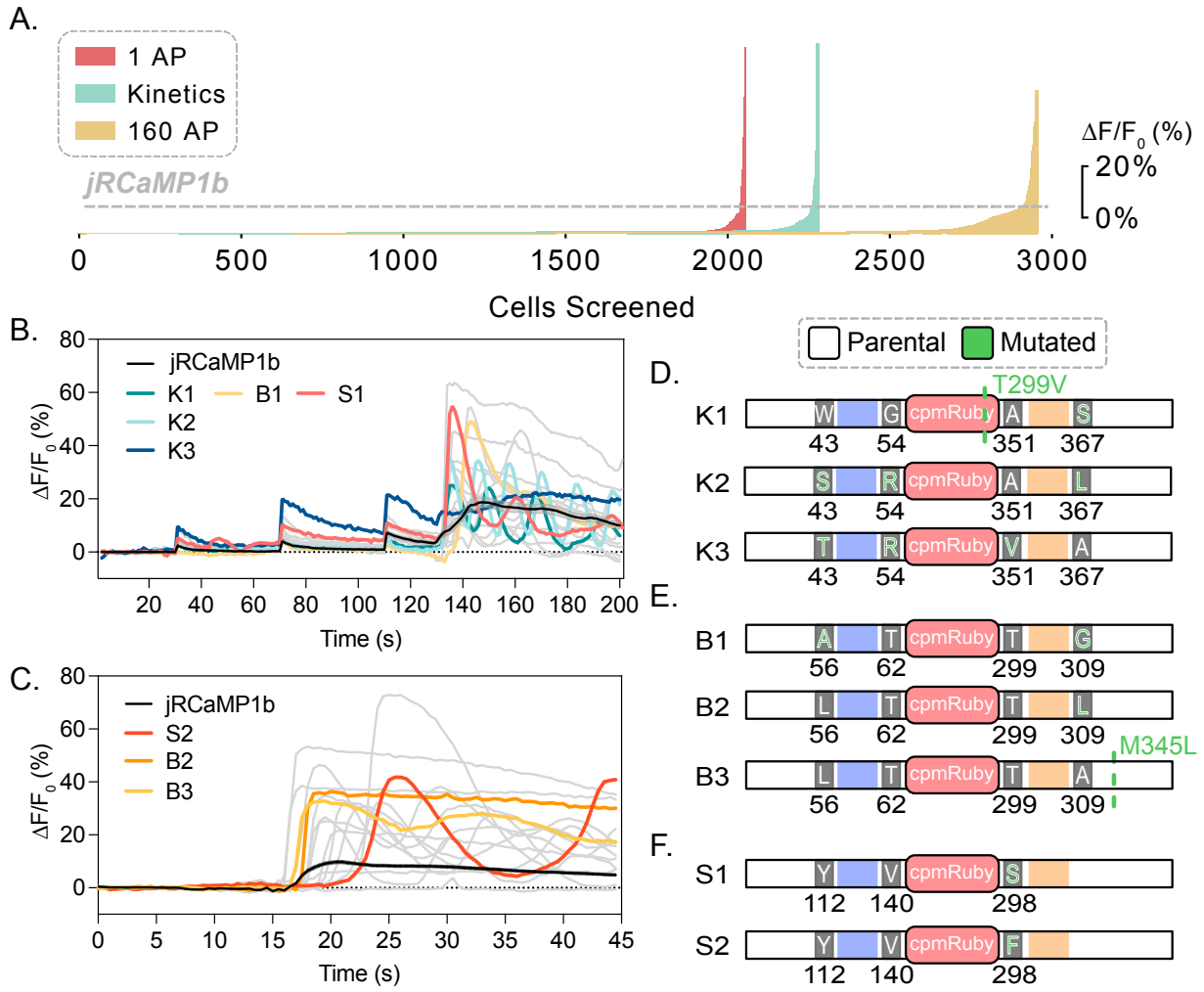


Figure 3.12 Opto-MASS Screen of Sensitivity, Kinetics, and Brightness jRCaMP1b Libraries

A. The waterfall plot demonstrates the ranked $\Delta F/F_0$ responses observed within the color-mapped jRCaMP1b libraries. Each bar represents the response from one cell.

B. The time course demonstrates cell responses that showed improvements for one of the three target qualities. Each gray line depicts the response from a cell pick in which the genetic identity could not be recovered. The colored lines depict recovered genetic identities alongside the parental response (black). Recovered picks are color-mapped and labeled depending on the origin library.

C. The time course demonstrates cell responses that improved one of two target qualities. Each gray line depicts the response from a cell pick in which the genetic identity could not be recovered. The colored lines depict recovered genetic identities alongside the parental response (black). Recovered picks are color-mapped and labeled depending on the origin library.

D. The genetic cassette demonstrates the recovered genetic identity of each mutated site within the labeled pick (left-hand side). White lettering indicates no mutation compared to jRCaMP1b, and green lettering indicates a mutated residue. Dotted lines indicate off-target mutations that were recovered alongside the mutation's identity.

E. The genetic cassette demonstrates the recovered genetic identity of each mutated site within the labeled pick (left-hand side). White lettering indicates no mutation compared to jRCaMP1b, and green lettering indicates a mutated residue. Dotted lines indicate off-target mutations that were recovered alongside the mutation's identity.
F. The genetic cassette demonstrates the recovered genetic identity of each mutated site within the labeled pick (left-hand side). White lettering indicates no mutation compared to jRCaMP1b, and green lettering indicates a mutated residue.

After retrieving these cells from the microwell array, we identified their genetic identities using rt-PCR (**Figure 3.12D, E, F**). Within the kinetics picks, K1 and K2, which displayed fast kinetics but low $\Delta F/F_0$ to CapChr2 stimulus, both maintained the genetic identity of A351, whereas the sensitive variant K3 harbored a mutation of A351V (**Figure 3.12D)**. This recapitulates what we had observed previously in the single mutation screening in which mutations to A351V improved the 1AP $\Delta F/F_0$ responses of jRCaMP1b (**Figure 3.6C, Table 3.6**). Interestingly, K1 picked up an off-target mutation of T299V in addition to an A367S mutation. This mutation had been tested previously and significantly improved the 1AP responses of jRCaMP1b. However, we currently do not understand the epistasis effects that may arise when combined with A367S. Notably, mutations to G368Q almost entirely reduced the sensor's functionality, and similar effects may occur here (**Figure 3.6F, Table 3.6**). Within the brightness library picks, we find that all three picks maintained the identity of T62 and T299, and both B2 and B3 maintained the identity of L56 (**Figure 3.12E**). Pick B2, which harbors an A309L mutation, has already been previously tested in our single mutant testing. Within the acetylcholine screen, we observe a similar response in which, within HEK293 cells, A309L outperforms jRCaMP1b. However, this response was not maintained within primary cultured neurons (**Figure 3.5, 3.6**). Variant B3 maintained the parental identity at each of the four chosen mutation residues and instead incorporated an off-target mutation of M345L. M345 was previously an untested residue, though when mapped back onto the crystal structure, this residue lies at the critical location at the insertion site between CBP and

CaM. The mutation of methionine to a small aliphatic leucine possibly reduces steric hindrances and promotes more favorable interactions between the two domains. Within the sensitivity variants, we found that both picks maintained Y112 and V140 and focused mutations at L298 (**Figure 3.12F)**. Mutation L298S is particularly important, as it displayed significant $\Delta F/F_0$ responses at the blue light stimulus and fast decay speeds after acetylcholine stimulus.

We moved forward with three of these variants to assess the responses of the picks within cultured neurons (**Figure 3.13A**). We found that M345L and L298F displayed large $\Delta F/F_0$ responses to single-flash stimuli, which was maintained throughout both 10-flash and 80-flash stimuli, though they do not achieve greater dynamic ranges in response to KCl (**Figure 3.13 B, C, D, E**). We do observe the same decay kinetics within M345L and A309L that we had observed within the Opto-MASS screen, M345L behaved similarly to jRCaMP1b, and A309L reduced the speed of decay (**Figure 3.13 F**). The responses for L298F were not similar to those we observed in our screening. In particular, the decay kinetics of the L298F pick were fast; however, in cultured neurons, we observed a complete reduction in the decay speed (**Figure 3.13 F**). There may be several reasons for this. Firstly, an undiscovered off-target mutation within the S2 pick may not be reflected within the DNA used for neuron screening. Another possibility is that the L298F mutation was a contaminating DNA that arose during rt-PCR amplification. With these results, we still require further testing of these libraries to identify promising variants. However, we must improve the gene recovery process, as we missed a lot of information on promising variants, simply because we could not amplify their DNA during rt-PCR.

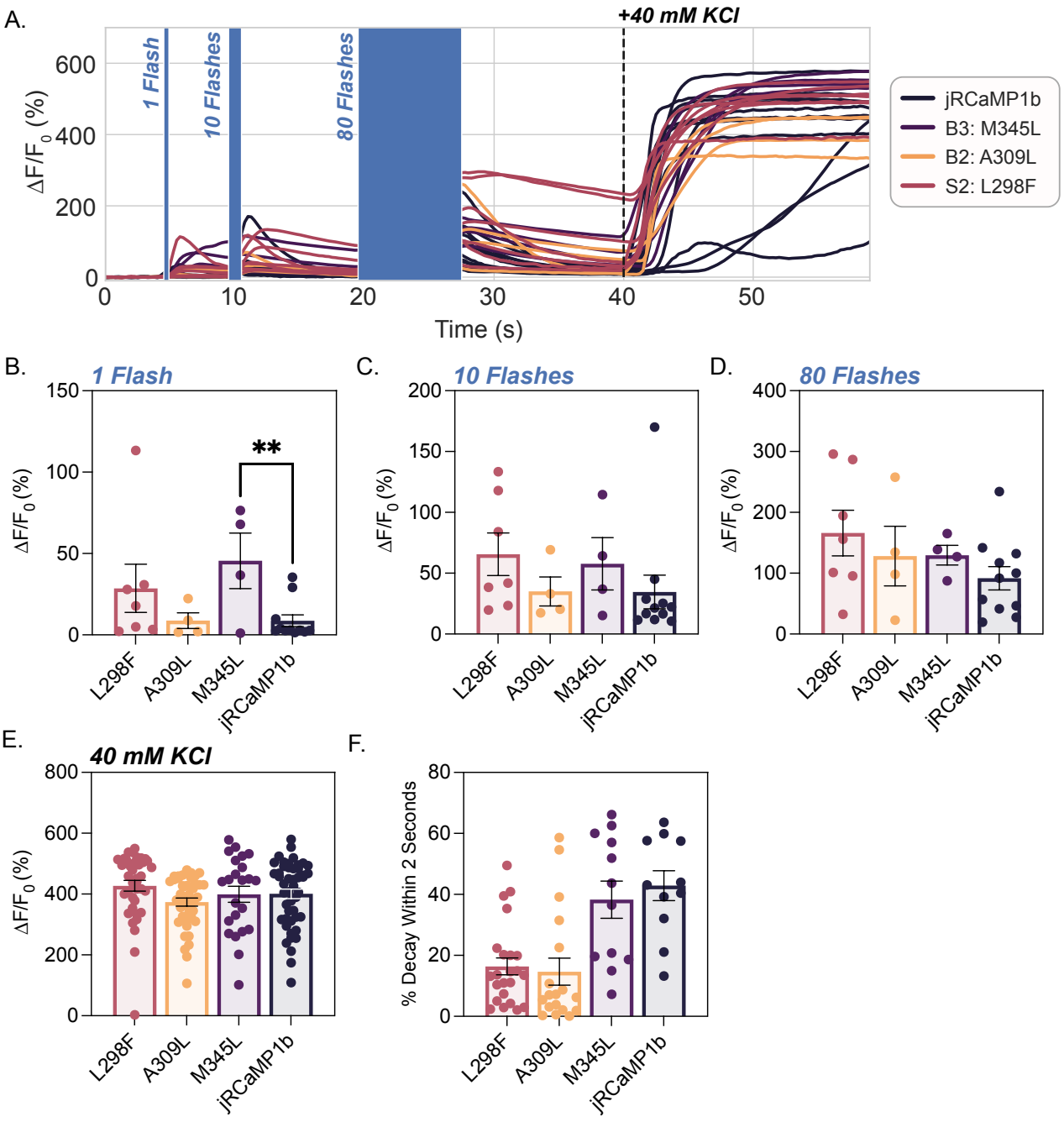


Figure 3.13 Opto-MASS Screen of Sensitivity, Kinetics, and Brightness jRCaMP1b Libraries

A. The time course displays the stimulus paradigm of the primary cortical neurons. Briefly, five ms light pulses were administered at 5 seconds, 10 seconds, and 20 seconds at a rate of 10 Hz, where applicable. After 40 seconds, 40 mM KCl was added to fully saturate the sensor. Each line depicts the responses to a single cell, and the corresponding variant color key is located on the righthand side of the graph.

B. The scatter plot demonstrates the average $\Delta F/F_0$ responses to one flash of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response. ** = P<0.001.
C. The scatter plot demonstrates the average $\Delta F/F_0$ responses to 10 flashes of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response.
D. The scatter plot demonstrates the average $\Delta F/F_0$ responses to eighty flashes of 488 nm light, where the bar indicates the mean and the error bars indicate the 95% confidence interval. Each dot in the scatter plot indicates a single cell's response.
E. The bar plot demonstrates the average $\Delta F/F_0$ responses to 40 mM KCl, where the bar height indicates the mean and the error bars indicate the SEM. Each dot indicates a single cell's response.
F. The bar plot demonstrates the average decay within 2 seconds after 10 flash stimuli, where the bar height indicates the mean and the error bars indicate the SEM. Each dot indicates a single cell's response.

3.3 DISCUSSION

In this chapter, we further implement our machine learning pipeline onto a new protein of interest, jRCaMP1b. We demonstrated the ensemble's efficiency in identifying variants that display improved sensitivities and differential kinetics to the parental jRCaMP1b. Within this version, we altered our models slightly to include a random forest regressor, a k-neighbors regressor, and a Gaussian process regressor. We chose to implement a Gaussian process regressor due to its proven capabilities in the engineering of enzymes[61,100–103] and the algorithm's ability to capture uncertainty within small datasets[58]. In our previous implementation, we observed that some contributing models would make predictions across a much wider range of values, which impacted the average predictions due to outliers. To circumvent these effects, we extended dataset scaling and feature selection to each model. With this ensemble configuration, we achieved R^2 values of 0.48, 0.74, and 0.41 for predictions made on the 1AP $\Delta F/F_0$, 10AP $\Delta F/F_0$, and 10AP decay, respectively (**Figure 3.4 C,D,E**). These R^2 values were considerably lower than what we achieved with the GCaMP model, and we predict that this phenomenon is due to a limited range of responses recorded within the training dataset. For example, due to the previous engineering

being performed on two combined datasets, many predicted variants behaved poorly compared to the template jRCaMP1b, which may introduce unintentional bias. This iteration uses data from a single origin, and with this format of data, we observe a more significant variance between the model predictions and the actual empirical value. However, the model predicted compared to the true value follows a linear correlation, and the model's predictions display impressive agreement with what was observed *in vitro* (**Figure 3.5, Figure 3.6**)**.** As a result of these engineering efforts, we have identified many mutations, such as T299V, T299Y, T62Q, A351I, A351W, and G368S, that each alter the biophysical characteristics of jRCaMP1b.

All four mutated residues occur in structurally significant portions of the jRCaMP1b protein, particularly on the interacting face between the sensing and fluorescent proteins. Residues T62, G368, and T299 occur on or near the linker residues. The linker residues translate the conformational changes of the sensing protein to the fluorescent reporter, and it has been shown that modifications in both composition and length can dramatically affect the sensor performance [40,104,105]. Residue 368 resides on one of the inner loops of the calmodulin and contains a Vander Waal overlap of >-0.04 Å with residues V349 and S58. V349 is directly adjacent to one of the EF-hand domains, and S58 occurs in the linker region between the CBP and cp-mRuby. This optimal positioning between these two crucial protein components may lend G368 to assist in translating calcium binding to the formation of the Calmodulin/CBP complex. Residue A351 occurs directly within one of the EF-hand domains, which lends credence to why mutations at this residue dramatically affect the sensitivity of our sensor. For example, the tryptophan in the A351W may form favorable interactions with G361 to promote the binding and retention of calcium.

We found that the ensemble's predictions captured an incredible capacity for nuance. For example, residue A309, highlighted within the 10 AP and 160 AP predictions to improve $\Delta F/F_0$,

did not show improvements to 1 AP stimuli but did show improvements at the 10 AP and 80 AP stimuli. This residue is also homologous to the site L317 in GCaMP, which was mutated to improve the $\Delta F/F_0$ and decay kinetics. Interestingly, mutations to this residue were not predicted to affect the kinetics and did not display improvements to the decay speed achieved with eGCaMP. The ensemble's predictions also detail a known phenomenon in which there is a tradeoff between sensitivity and decay kinetics[96]. We can observe this particularly clearly with mutations such as G54I, G54F, and I419W, which all displayed fast decay kinetics when screened in HEK293 cells and did not produce any fluorescence change to AP stimulus in cultured neurons. The ensemble predicted all three mutations to decrease $\Delta F/F_0$ and increase decay speeds. While not predicted as clearly, we observe the same trend with A351I and A351W, which achieved impressive responses to 1AP stimulus and had slow decay kinetics.

In addition to the single mutations that we derived from the ensemble predictions, we were also able to analyze the interprotein interactions that govern each biophysical property by mapping the ensemble predictions back onto the protein's structure. We found similar behavior to what we had observed previously, in which residues commonly mutated by the ensemble for each biophysical property showed distinctive proximity within 3D space. Capitalizing on this discovery, we chose to form mutation libraries at these clustered residues. We screened these mutation libraries in a high-throughput manner to discover the effects of combinatorial mutations that may have otherwise been too experimentally burdensome to investigate. Within this screening, we identified multiple mutations that display impressively fast kinetics and sensitivities. While the mutations we gathered still necessitate validation, the approach validates that ensemble predictions can be paired with high-throughput screening in a symbiotic manner.

The direct benefit of these engineering efforts will be the development of red-shifted GECIs that are sensitive, fast, and lack a photo-activation artifact. Toward that goal, we have already identified several promising jRCaMP1b variants using our developed machine-learning pipeline. As a juxtaposition, we performed rational engineering on XCaMP-R using known beneficial mutations from eGCaMP. These engineering efforts were primarily in vain and were far less successful than our machine learning-derived variants.

In conclusion, by harnessing the insights gleaned from the ML pipeline, we have already made significant strides in identifying promising variants for red-shifted GECIs, surpassing the outcomes of traditional rational engineering approaches. This underscores the potential of ML-driven engineering to accelerate the development of sensors with enhanced sensitivity and performance characteristics. Moving forward, continued refinement and optimization of our ML pipeline are vital to unlocking further advancements in this field, paving the way for rapid and efficient innovation in protein engineering.

3.4 METHODS

3.4.1 Data Preprocessing Before ProteiML Predictions

The Dana study provided a functional characterization of >1000 RCaMP variants[96]. To generate a dataset compatible with ProteiML, we generated a Pandas Dataframe with the following structure: one column indicating the mutation primary identifier, 442 columns that correspond to each residue in the jRCaMP1b sequence, and the last column that contains empirically derived performance of the variant. Each row contains a unique variant with a unique primary key identifier and sequence.

$$\Delta F/F_0 = \frac{(F - F_0)}{F_0} * 100 \qquad \textit{(Eq.1)}$$

The resultant dataset is the basis for our dependent and independent variables needed to train ProteiML. We generated four datasets, the sequence data linked to each biophysical characteristic such as 1AP $\Delta F/F_0$, 10AP $\Delta F/F_0$, 160AP $\Delta F/F_0$, and kinetics capability ($\tau_{1/2}$).

3.4.2 Molecular Cloning

Predicted mutations were reflected into the pGP-CMV-NES-jRCaMP1b backbone (Addgene ID: 63136) using point-mutation primers ordered from Integrated DNA Technologies (IDT) and PCR amplification with either Q5-polymerase (New England Biolabs; M0492L) or Superfi-II polymerase (Invitrogen; 12368010). Amplification of the DNA fragment was verified with agarose gel electrophoresis. Blunt-end DNA circularization was achieved with Kinase, Ligase, and DpnI enzyme (KLD) treatment (New England Biolabs: E0554S). Circularized DNA was transformed into competent *E.Coli* cells (DH5α or TOP10) and grown on agar plates that contain either ampicillin or kanamycin selection antibiotic (50 µg/mL). Upon colony formation, single colonies were picked and grown in 5mL cultures containing LB Broth (Fisher BioReagents; BP9723-2) and selection antibiotic (ampicillin/kanamycin; 50 µg/mL) overnight (37°C, 230 RPM). DNA was isolated using Machery Nagel DNA prep kits (Machery Nagel; 740490.250). Sanger sequencing (Azenta; Seattle, WA) of the isolated plasmid DNA was used to confirm the presence of the intended mutation.

Genes encoding the jRCaMP1b variants were cloned into a CAG-driven backbone, pCAG-Archon1-KGC-EGFP-ER2-WPRE (Addgene; #108423), using Gibson assembly (New England Biolabs; E2621L). All subsequences were verified with Sanger sequencing (Genewiz; Seattle, Wa).

3.4.3 Acetylcholine Assays

Human Embryonic Kidney (HEK293; ATCC Ref: CRL-1573) cells were cultured in Dulbecco's Modified Eagle Medium + GlutaMAX (Gibco; 10569-010) supplemented with 10% fetal bovine serum (Biowest; S1620). When cultures reached 85% confluency, the cultures were seeded at 100,000 cells per well or 50,000 cells per well in 24-well and 48-well plates, respectively. 24 hours after cell seeding, the cells were transfected using Lipofectamine3000 (Invitrogen; L3000015) at 1000 ng of DNA per well of a 24-well plate, according to the manufacturer's instructions.

48 hours post-transfection, the plates were prepared for imaging by washing and then replacing culturing media volume with imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES (triple supplemented with 1% Glutamax (Gibco; 35050-1), 1% Sodium Pyruvate (GIBCO; 11360-070), and 1% MEM Non-Essential Amino Acids (Gibco; 11140-050)). Crystalline power Acetylcholine Chloride (Alfa Aesar; L02168.14) was resuspended into imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES) into 2x the desired final concentration. During imaging, 1:1 volumes of the acetylcholine-tyrodes imaging solution were hand-pipetted into the bath volume to bring the final acetylcholine concentration to the desired concentration. Imaging was performed on a sCMOS camera (Photometrics Prime95B) on an epifluorescent microscope (Leica DMI8) using a 20X objective (Leica HCX PL FLUOTAR L 20x/0.40 NA CORR). A Lumencor Light Engine LED and Semrock Filters were used for fluorescence imaging.

3.4.4 Analysis of Fluorescent Assays

Analysis of HEK293 cell fluorescence imaging data was done using FUSE, a custom cloud-based semi-automated time series fluorescence data analysis platform written in Python. First, the cell segmentation quality of the selected Cellpose[90] model was manually verified. For the segmentation of cells expressing cytosolic fluorescent indicators, model 'cyto' was selected as our base model. If the selected Cellpose model was low-performing, we further trained the Cellpose model using the Cellpose 2.0 human-in-the-loop system[91]. Using an "optimized" segmentation model, fluorescence time-series data is extracted for each region of interest. This allows for unbiased extraction of change in cellular fluorescence information for a complete set of experimental samples. Using the raw fluorescence data, % fluorescence change from the baseline ($\Delta F/F_0$) over time was calculated using *Eq.1*.

The rate of decay was approximated using *Eq.3*, where F(t) is the change in fluorescence at a time (t) after the max fluorescence (F_0) was achieved. Importantly, F_0 was normalized to 1.0, such that F(t) depicts the change in fluorescence over time, t.

3.3.5 Isolation of Cortical Neurons

Primary cortical neurons were prepared as previously described[93,94]. Briefly, 24-well tissue culture plates were coated with matrigel (mixed 1:20 in cold-PBS, Corning; 356231) solution and incubated at 4°C overnight prior to use. Sterile dissection tools were used to isolate cortical brain tissue from P0 rat pups (male and female). Tissue was minced until 1mm pieces remained, then lysed in equilibrated (37°C, 5% CO_2) enzyme (20 U/mL Papain (Worthington Biochemical Corp; LK003176) in 5mL of EBSS (Sigma; E3024)) solution for 30 minutes at 37°C, 5% CO_2 humidified incubator. Lysed cells were centrifuged at 200xg for 5 minutes at room temperature, and the supernatant was removed before cells were resuspended in 3 mLs of EBSS (Sigma; E3024). Cells

were triturated 24x with a pulled Pasteur pipette in EBSS until homogenous. EBSS was added until the sample volume reached 10 mLs prior to spinning at 0.7 rcf for 5 minutes at room temperature. Supernatant was removed, and enzymatic dissociation was stopped by resuspending cells in 5 mLs EBSS (Sigma; E3024) + final concentration of 10 mM HEPES Buffer (Fisher; BP299-100) + trypsin inhibitor soybean (1 mg/ml in EBSS at a final concentration of 0.2%; Sigma, T9253) + 60 µl of fetal bovine serum (Biowest; S1620) + 30 µl 100 U/mL DNase1 (Sigma;11284932001). Cells were washed 2x by spinning at 0.7 rcf for 5 minutes at room temperature and removing supernatant + resuspending in 10 mLs of Neuronal Basal Media (Invitrogen; 10888022) supplemented with B27 (Invitrogen; 17504044) and glutamine (Invitrogen; 35050061) (NBA++). After final wash spin and supernatant removal, cells were resuspended in 10 mLs of NBA++ prior to counting. Just before neurons were plated, matrigel was aspirated from the wells. Neurons were plated on the prepared culture plates at desired seeding density. Twenty-four hours after plating, 1µM AraC (Sigma; C6645) was added to the NBA++ growth media to prevent the growth of glial cells. Plates were incubated at 37°C and 5% CO_2 and maintained by exchanging half of the media volume for each well with fresh, warmed Neuronal Basal Media (Invitrogen; 10888022) supplemented with B27 (Invitrogen; 17504044) and glutamine (Invitrogen; 35050061) every three days.

3.4.6 Calcium Phosphate Transfection of Primary Cortical Neurons

Isolated primary cortical neurons were transfected using the calcium phosphate transfection kit from Sigma Aldrich (Sigma-Aldrich; CAPHOS-1KT). Half of the neuron media was changed 24 hours before transfection, saving the removed conditioned media to add to the neurons after transfection. Reagents were mixed in a ratio of 3 µl $CaCl_2$: 24.5 µl H_2O: 1000 ng

DNA before being added dropwise to bubbled 2x HEPES Buffered Saline (30 µl). The final solution was vortexed for 4 seconds and left undisturbed for 20 minutes. The solution was added dropwise to each well of neurons in a 24-well plate and shaken to distribute equally. Neurons were left to incubate for 1 hr at 37^0C with 5% CO_2. The cells were rinsed twice with HBSS before adding the conditioned media removed from the day prior and mixed with half-fresh media.

3.3.7 Cortical Neuron Stimulus Protocol

The cultured neurons were transduced with pAAV-CaMKIIa-hChR2(H134R)-EYFP (Addgene; 26969-AAV1) three days post dissection and given 5 days to express. On the day of imaging, ~24-36 hours post-transfection, cells were washed once with imaging solution and then transferred to 1xTyrode's (pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES). Flashes of 488 nm light were triggered using Sutter Instruments Integrated Patch Amplifier with Patch Panel, time-locked using Igor Pro 8. Five millisecond light pulses were administered at 10 hz. Powdered Potassium Chloride (Sigma; P9541-500G) was diluted in ddH_2O to a concentration of 2M. This solution was then diluted to 80mM in imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES). During imaging, 1:1 volumes of KCl solution were hand-pipetted into the bath to bring the final KCl concentration to the desired concentration. Imaging was performed on a sCMOS camera (Photometrics Prime95B) on an epifluorescent microscope (Leica DMI8) using a 20X objective (Leica HCX PL FLUOTAR L 20x/0.40 NA CORR). A Lumencor Light Engine LED and Semrock Filters were used for fluorescence imaging. Bulk fluorescence traces were acquired using FIJI imaging software with background subtraction (rolling = 50 stack) and hand-drawn ROIS. The baseline was defined as the first 15 measurements before light addition. Max $\Delta F/F_0$ values were obtained using a custom Python script.

3.4.8 Generation of Randomized Variant Libraries

Randomized mutation libraries were generated using Gibson assembly cloning. Inserts and backbones were amplified using PCR (Superfi-II polymerase (Invitrogen; 12368010)). Amplification of the DNA fragments was verified with agarose gel electrophoresis, and circularization was achieved using Gibson assembly (New England Biolabs; E2621L). The assembly was purified using the Machery Nagel PCR cleanup kit (Macher Nagel; 740611.250) before being transformed into 33 µL of electrocompetent cells (NEB cat no. C3020K) (2000 V, τ = 5 ms) in ice-cold cuvettes (1 mm gap) with one µL of the elution. The electroporated cells were rescued in 967 µl of SOC for 1 hour at 37 °C. The rescue media was added to 100 mL of Luria Broth (LB) with supplemented with ampicillin and grown overnight at 37 °C and 240 rpm. Library DNA was isolated using the Machery–Nagel NucleoBond Xtra Midi EF kit (Machery Nagel; 740420.50). The mutagenesis of the library was verified using nano-pore sequencing (Plasmidsaurus).

3.3.9 Landing Pad Transfections

HEK293T landing pad cells that contain a recombination cassette, developed by the fowler lab, were stably recombined with the library prep using a double transfection protocol. The landing pad cells were standard cultured in 1–2 µg/mL doxycycline. One day before transfection, the landing pad cells were removed from doxycycline. On the day of transfection, the landing pad cells were seeded at 250,000 cells per well into 6-well dishes. The Fugene transfection reagents (per well) were prepared as followed: 3 µg of plasmid DNA encoding pCAG-Bxb1 recombinase; 6 µL of Fugene6 (Promega cat. # E2693); 300 µL of Opti-MEM. The reagents were left to complex undisturbed for 15 min and then added to the cell suspension. After 24 hours, the fugene transfection is repeated with the addition of the library plasmid in place of the pCAG-Bxb1

recombinase DNA. The transfected cells were transitioned to 4 μg/mL doxycycline media after 24 hours and 4 μg/mL doxycycline + 1 μg/mL doxycycline media after 48 hours.

3.4.10 Polydimethylsiloxane Microwell Array Formation and Seeding

The microwell arrays were fabricated, as discussed previously[40]. Briefly, an etched silicon wafer acts as a negative mold for the PDMS (Sylgard 184, Corning). The PDMS was formulated by mixing an equal ratio of 10:1 (w/w) and vigorously mixing. This mixture was poured onto the negative mold and placed into a desiccator for 10 minutes to remove air bubbles. The PDMS-containing wafer was then placed into a 70°C incubator overnight to cure. The cured PDMS was removed from the mold and plasma treated before being stamped with bovine serum albumin (BSA) (FischerSci; Cat #BP1600). The PDMS arrays were then cut to size using a scalpel and placed into a 24-well dish.

3.4.11 Cell Seeding Onto Microwell Arrays

With the PDMS arrays placed in 24-well dishes, the entire plate was plasma-treated before quickly adding standard growth media to each well. The plate with media in the wells was then placed under vacuum to remove microbubbles from inside the wells. The landing pad cells were then removed from their standard growth dishes using 0.05% Trypsin-EDTA and before being counted. The cell suspension was normalized to a concentration of 500,000 cells per mL. 40 μL of this cell suspension was then carefully pipetted above each microwell array. The cells were returned to the incubator for 10 min to allow for the cells to seed into the cells through gravity. After 10 min, the 24-well plates were then placed in a centrifuge and spun down at 100 RCF for 5 min, to further seed the cells deeper within the wells. The media within each well was changed to remove any cells that did not occupy the microwells and to replace the array cells with doxycycline media.

3.4.12 Reverse Transcription PCR

Reverse transcription PCR was performed as discussed previously, and the current section contains excerpts from the manuscript[40]. Single-cell recovery tubes were prepared by diluting 5 µL of 0.1 M DTT into 200 µL of TE buffer; 5 µL of the TE/DTT solution was added to each PCR tube. After library screening, the single cell recovery tubes containing 5 µL 0.1 M DTT/ TE and the library pick were removed from the dry ice and placed on wet ice. Each tube was processed with reagents from the SuperScript IV First-Strand synthesis kit (Invitrogen cat. # 18091050). To each tube, 0.5 µL of 0.1 M DTT and 0.5 µL of RNase inhibitor were added. The samples were placed on dry ice for 5 min and then moved back to wet ice. Next, 0.5 µL of the following was added to each tube, DI H20, a 10 mM dNTP mix, and 2 µM primer.

Next, the primers were annealed to the mRNA by incubating the samples at 95 °C for 30 s, 4 °C for 1 min, and 65 °C for 5 min. The samples were then returned to the wet ice. Next, 2 µL of SSIV RT 5× Master Mix was added to each sample. The samples were pipetted up and down thoroughly. Finally, 0.5 µL of the SSIV RT Enzyme was added to each tube, and the samples were thoroughly pipetted up and down. To perform the reverse transcriptase reaction, the samples were incubated at 53 °C for 10 min and then at 80 °C for 10 min. Next, 0.5 µL of RNaseH was added to each tube, and the samples were incubated for 20 min at 37 °C to remove any mRNA from the cDNA. Samples were stored at −20 °C before PCR amplification of cDNA with Q5 (New England Biolabs) or SuperFiII (ThermoFischer) using recombination-specific primers.

Chapter 4. Optimization and Exploration of High-Throughput Screening Using the Red-Shifted Dopamine Indicator, $GRAB_{rDA2m}$

ABSTRACT

The current glaring limitation of the machine learning engineering approach is the necessity of large mutational datasets for model training. The curation of such datasets can lead to very long lead times and financial burdens on the engineering lab. In an effort to form mutation datasets within the Berndt lab, we investigated two approaches for mutant testing. The first is through implementing the Berndt lab-developed Optogenetic Microwell Array Screening System (Opto-MASS)[40]. Opto-MASS dramatically increases the throughput of variant screening; however, we encountered challenges in accurately recovering the genetic identities of the promising variants. This chapter discusses the limitations of the current pipeline and presents efforts to optimize the cell recovery method. The second approach addresses the challenge of combinatorial mutations, which often lead to misfolded proteins or non-functional sensors. To reduce the resources spent screening non-viable mutants, we investigated Fluorescence-Activated Cell Sorting (FACS) as a tool to enrich variant libraries with viable mutations. As a target for this engineering, I have aimed to optimize $GRAB_{rDA2m}$[105], a genetically encoded red-shifted dopamine indicator. The outcome of the effort expended in this chapter is not only an engineered $GRAB_{rDA2m}$ sensor but a more productive identification of mutant identities.

4.1 INTRODUCTION

While we, among others, have demonstrated the capabilities of machine learning in protein engineering, the scope and scalability of this approach heavily depend on the availability of mutational libraries[8,58,59,61,62,97]. The form and function of every protein are unique. Therefore, mutation libraries would need to be curated for each protein that requires functional engineering. The generation of these mutational libraries is hugely time- and labor-intensive. For example, the curation of the datasets used in Chapters 2 and 3 extended over six years. During this time, each of the more than 2,000 mutants tested was selected based on structural insights, cloned into the chosen construct, packaged into viruses, and tested on primary hippocampal neurons[7,10,33]. In comparison, the training and engineering of sensors based on these libraries was highly efficient, highlighting the immense need to expedite the formation of these mutation libraries.

Within the Berndt lab, we are uniquely equipped to generate large mutation libraries using our optogenetic microwell array screening system, Opto-MASS[40]. Within this platform, we clone a mutation library and stably transfect the mutants into HEK293 landing pad cells[98]. These cells are seeded onto PDMS microwell arrays and screened for functional abilities. Variants that exhibit improved capabilities compared to the parental construct are isolated from the arrays using glass micropipettes, and the identity of each variant is recovered via RT-PCR. Using this platform, we have successfully improved the dopamine sensor dLight1.1 and opioid sensor mLight[35]. However, we have encountered issues with a low recovery rate of the genes of interest. We hypothesized that this can be attributed to the method of cell recovery from the PDMS microwell array. We are investigating the optimization of the glass micropipette used for picking, as well as modifying the sequencing method used on the amplified DNA.

We also found that in our opto-MASS screening, most of the tested mutants were non-functional. This is due to epistasis within our mutation library, where the impact of combinatorial mutations can non-linearly affect the folding and functionality of our sensors[87]. As a result, most mutants tested in our Opto-MASS screening show no effect or a worsening of sensor capabilities. In this chapter, we explore the use of Fluorescence-Activated Cell Sorting (FACS) to enrich our libraries with promising functional variants.

A promising candidate for these engineering efforts is the red dopamine sensor $GRAB_{rDA2m}$. The current suite of dopamine sensors combines dopamine-sensing G-protein coupled receptors (GPCRs), DRD1 or DRD2, with a fluorescent reporter[35,65,104,105]. The development of these tools has enabled researchers to monitor dopamine dynamics in behavioral studies of reward, learning, and movement[106–109]. Recent advancements have been made to expand the color palette of dopamine indicators to include red and yellow fluorescent proteins, facilitating multiplexing[104,105]. However, red-shifted dopamine indicators have been based on cpmApple, which displays photoactivation when exposed to blue light. This effect has already caused a misinterpretation of rdLight1.1 signals when multiplexed with the blue-light-activated opsin ChR2[110,111]. Interestingly, despite containing mApple, the $GRAB_{rDA2m}$ and $GRAB_{rDA2h}$ variants do not display photoactivation to blue light, although they exhibit smaller $\Delta F/F_0$s compared to their photoactive counterparts, $GRAB_{rDA3m}$ and $GRAB_{rDA3h}$. We aim to improve the $\Delta F/F_0$ of the non-photo-switching variant, $GRAB_{rDA2m}$, due to its suitability for multiplexing experiments and its slow decay kinetics, which are advantageous for FACS.

This chapter aims to optimize our ability to generate mutation libraries through two main mechanisms. First, by refining our identification of promising variant DNA in high-throughput screening, and second, by enriching our libraries with beneficial variants using FACS. Through

these efforts, this study seeks not only to improve $GRAB_{rDA2m}$ but also to contribute to more efficient mutant testing and machine learning dataset generation in protein engineering

4.2 RESULTS

4.2.1 Biophysical Properties of $GRAB_{rDA2m}$ Prior to Sensor Engineering

$GRAB_{rDA2m}$ is a GPCR-based sensor that links the red fire ant D_2R receptor to the cp-mApple fluorescent protein to form a red-shifted dopamine indicator (**Figure 4.1A**)[105]. $GRAB_{rDA2m}$ maintains a robust $\Delta F/F_0$ response to dopamine addition with impressive trafficking of the sensor to the cell membrane (**Figure 4.1B, C**). Of the current red-shifted dopamine indicators available, $GRAB_{rDA2m}$ displays decreased sensitivities, which can even be observed in slow onset kinetics of the sensor (**Figure 4.1D**). Importantly, $GRAB_{rDA2m}$, despite utilizing the cpmApple fluorescent protein, does not display photoactivation in response to blue light stimulation, making it one of the only genetically encoded, red-shifted dopamine indicator free from this artifact (**Figure 4.1E**). We chose $GRAB_{rDA2m}$ as a target for optimization of our gene recovery efforts primarily due to the extended activation of the sensor after dopamine addition. As FACS was a proposed tool for screening efforts, we determined that the off-kinetics provided extended periods for screening between tube exchanges. Conversely, the fast off-kinetics of the jRCaMP1b sensors, that decays within a manner of seconds, severely limits applicability of FACS as a tool for calcium sensor engineering. The secondary benefit of engineering $GRAB_{rDA2m}$ is the improvement of a red-shifted dopamine indicator capable of multiplexing.

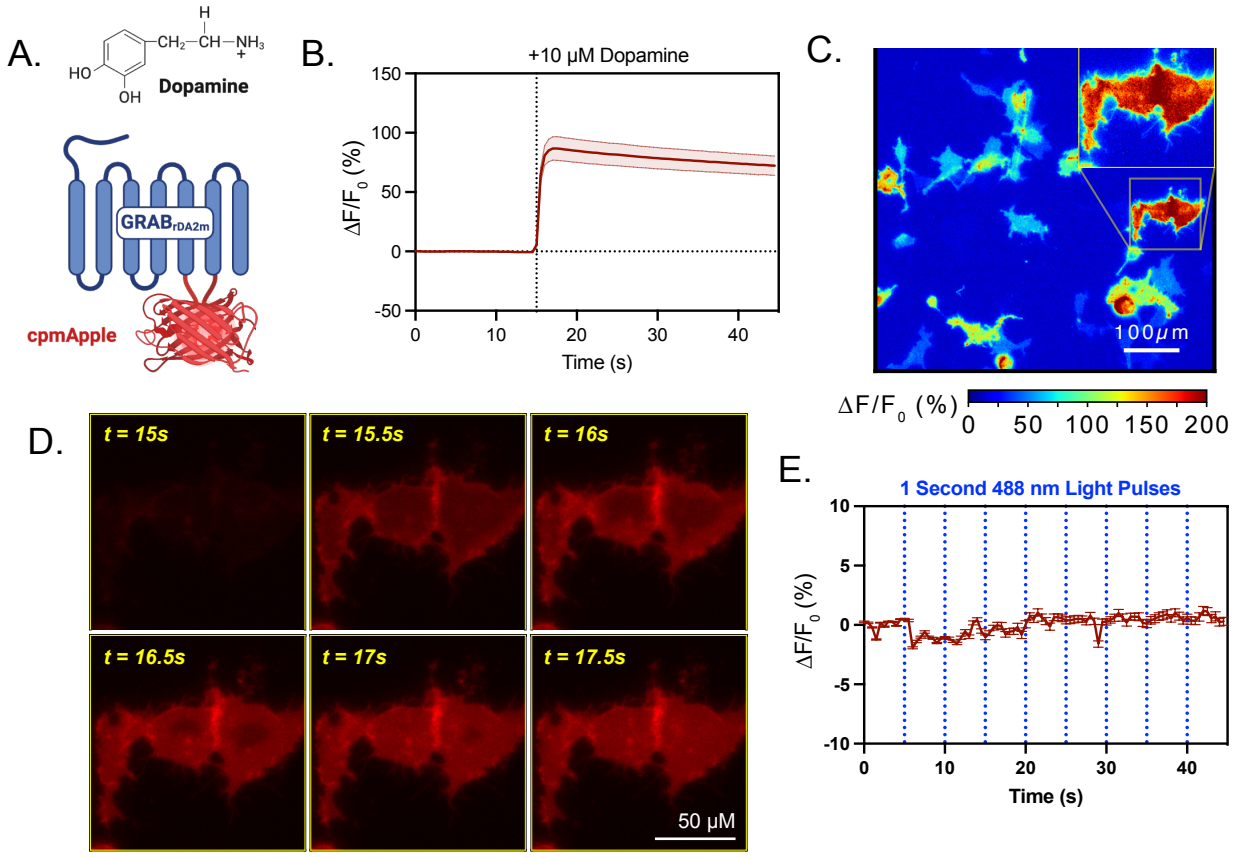


Figure 4.1 Biophysical characteristics of $GRAB_{rDA2m}$

A. Schematic of sensor topology alongside Dopamine, the sensed molecule.
B. The time course displays the $\Delta F/F_0$ response of pDisplay-$GRAB_{rDA2m}$ to 10µM Dopamine.
C. Image displays the $\Delta F/F_0$ response of pDisplay-$GRAB_{rDA2m}$ to 10µM Dopamine. Each pixel is heat mapped with the corresponding max $\Delta F/F_0$ achieved. Scale bar = 100 µm.
D. The image set displays the onset of the $\Delta F/F_0$ response of pDisplay-$GRAB_{rDA2m}$ to 10µM Dopamine. Each image is 0.5s apart, where the yellow text indicates the time. Scale bar = 50 µm.
E. The time course displays the photoactivation profile of $GRAB_{rDA2m}$ to 1s pulses of 488 nm light.

4.2.2 Generation of a Site-Saturated Mutation Library of the GRAB$_{rDA2m}$ Linkers

Within the GRAB$_{rDA2m}$ sensor, we determined six sites within the linker residues to fully saturate (**Figure 4.2A**). To form our mutation libraries, we cloned the GRAB$_{rDA2m}$ sensor into the landing pad compatible plasmid, which contains an attb recombination cassette and a P2A-Puro that will enable selection resistance (**Figure 4.2B**). To mutate the linker residues, we performed PCR with degenerate codons, ensuring that each nucleotide was equally represented at the six chosen residue locations (**Figure 4.2C**). Alongside the library, we additionally cloned the wild-type GRAB$_{rDA2m}$ sensor to act as a control, where we saw responses consistent with the non-landing pad vector when seeded onto the PDMS micro-well arrays (**Figure 4.1B, Figure 4.2D, E**).

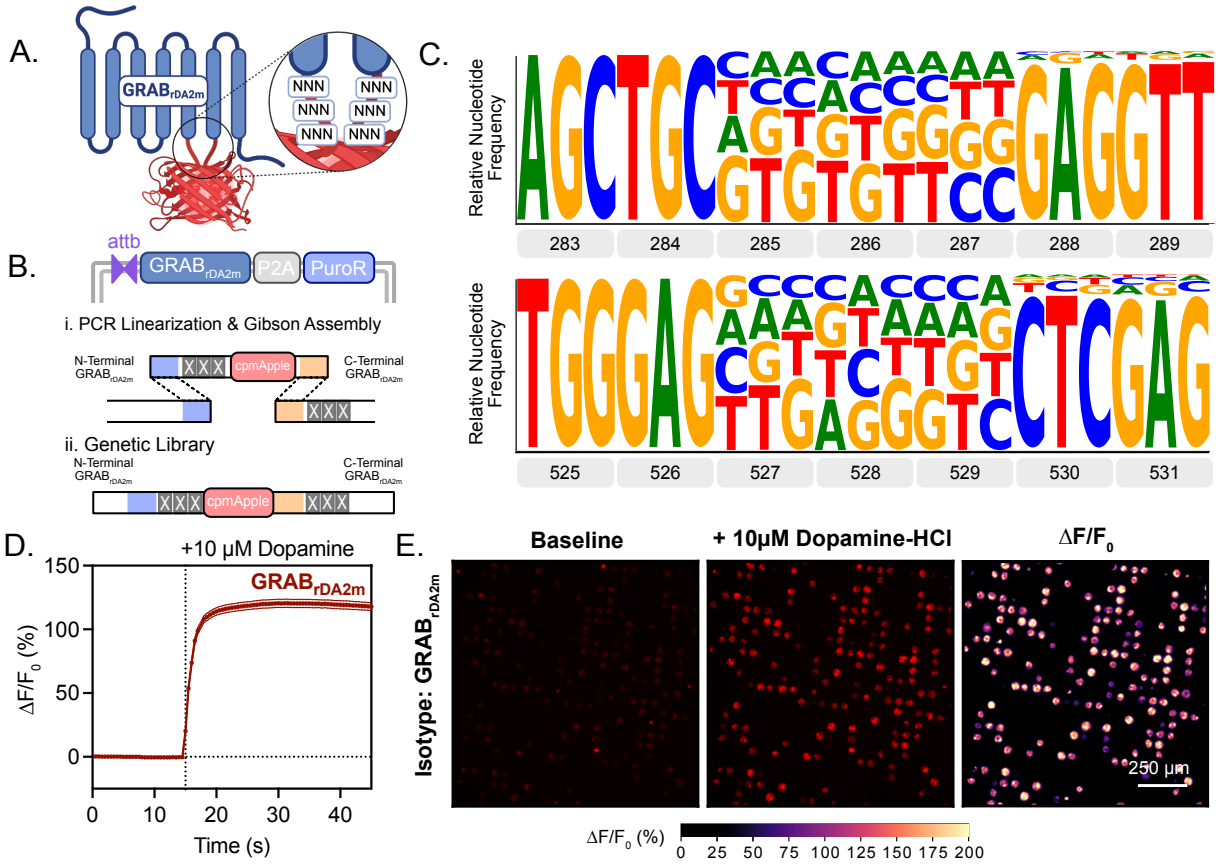


Figure 4.2 Library Construction of a GRAB$_{rDA2m}$ Linker Library for Opto-MASS Screening

A. Schematic of the residues targeted for generating a mutational dopamine sensor library from GRAB$_{rDA2m}$ (denoted by X). The six residues are located in the linker between the cpmApple and GPCR domains and were targeted by site-saturated randomized mutagenesis.
B. Schematic of the dopamine sensor library constructed using Gibson Assembly. Colored regions are indicative of Gibson Assembly overlaps that are included during cloning.
C. The logo maps display the relative nucleotide frequency achieved at the indicated residues.
D. The time course displays the $\Delta F/F_0$ response of the stably transfect landing pad-GRAB$_{rDA2m}$ cells to 10μM Dopamine.
E. The image set displays the fluorescence of landing pad-GRAB$_{rDA2m}$ before and after adding 10μM Dopamine and the $\Delta F/F_0$ achieved by each cell. Scale bar = 250 μm.

4.2.3 Opto-MASS Screen of the GRAB$_{rDA2m}$ Linker Library with Optimized Pipettes and Sequencing Analysis

In the currently published version of Opto-MASS, we struggled to recover the genes of interest. This manifested in no observable DNA amplification after rt-PCR. As such, we found ourselves only recovering the genetic identity of one pick out of every ten. Because of this, I

proposed several methods to try to increase the efficiency of our Opto-MASS picking protocol. The first was to alter the glass micropipettes' width to isolate our cells of interest. Second, we changed the method of sequencing used to analyze our cell picks.

In the development of the Opto-MASS pipeline, we used glass micropipettes, commonly used in whole-cell patch clamp electrophysiology, to physically isolate the promising variant cells from the PDMS arrays. The standard patch pipette has a very narrow opening at the tip that allows for the pipette to contact a small portion of the cell membrane (**Figure 4.3A**). With these pipettes, we would pierce the cell to remove it from the PDMS array. During experimentation, we found that the surface tension of the imaging solution caused the cell of interest to dislodge, resulting in a low recovery rate. To circumvent this issue, we modulated the temperatures used to pull the glass micropipettes. We found that by decreasing the temperature of the second step, we were able to achieve a slightly wider opening at the end of the micropipette. The resultant pipette had an opening at the tip that was much wider and capable of entirely fitting the cell within the end of the tip (**Figure 4.3B&C**). With this pipette, after aligning the end of the pipette with a cell of interest, we were able to apply small amounts of negative pressure, which kept the cell stuck to the end of the pipette and reduced the effect of surface tension on cell recovery.

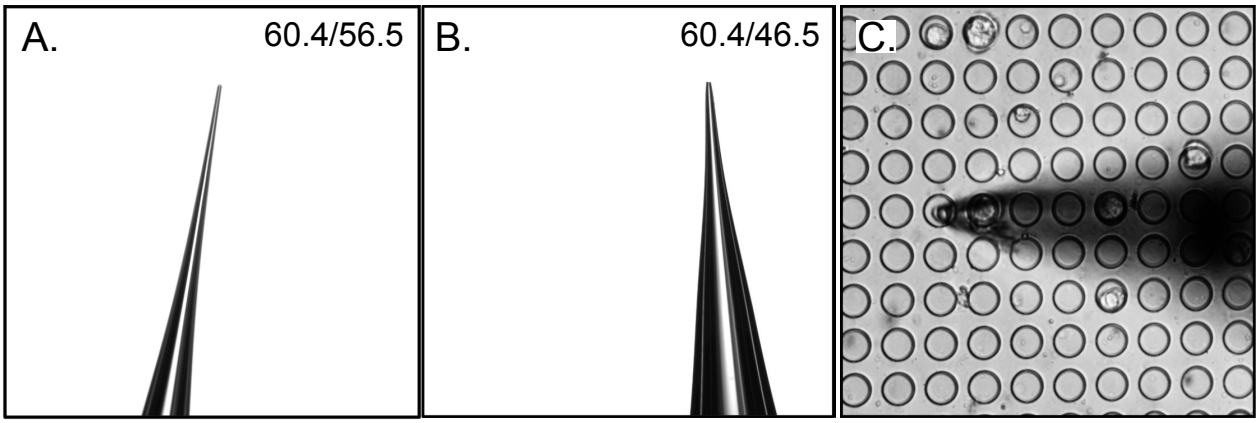


Figure 4.3 Optimized Micropipettes Used for Cell Picking

A. The image demonstrates a glass micropipette commonly used in patch clamp electrophysiology and used during the first iteration of Opto-MASS picking. The numbers in the top right corner indicate the temperatures used during pipette pulling.
B. The image demonstrates the optimized glass micropipette. The numbers in the top right corner indicate the temperatures used during pipette pulling.
C. The image demonstrates picking a cell of interest using the optimized glass micropipettes.

In addition to modulating the picking pipettes, we transitioned from sequencing the pick DNA using Sanger sequencing to nanopore sequencing. We found that during our Sanger sequencing of recovered cells, we often got straightforward peaks corresponding to the variant DNA (**Figure 4.4A**). However, occasionally contaminating DNA would lead to multiple overlapping peaks in the Sanger sequencing results, which made us unable to determine the identity of the mutated residue (**Figure 4.4B**). With nanopore sequencing, we receive raw, unique reads, allowing us to determine which residue was the most commonly found and, therefore, more likely to be our cell of interest. Using the raw reads obtained from nanopore sequencing, we can delineate the identities of each contaminating DNA to recover the variant of interest's identity (**Figure 4.4C**).

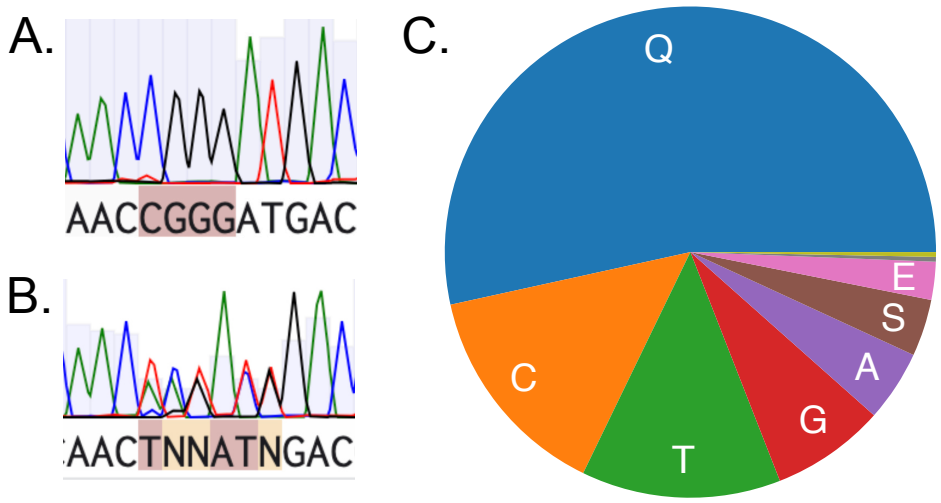


Figure 4.4 Identification of the Variant DNA Using Nanopore Sequencing
A. The image demonstrates the sequencing results obtained using Sanger sequencing on a variant pick that contains a single variant.
B. The image demonstrates the sequencing results obtained using Sanger sequencing on a variant pick that contains contaminating DNA.
C. The image demonstrates the sequencing results obtained using nanopore sequencing on a variant pick that contains contaminating DNA.

With these two modifications to the Opto-MASS protocol, we continued screening the $GRAB_{rDA2m}$ linker library. We screened the $\Delta F/F_0$ capabilities of approximately 14,000 cells at dopamine concentrations of 10µM and 1µM (**Figure 4.5A**). We predominantly focused on the pick results from the 10µM dopamine screen, in which we found multiple promising variants (**Figure 4.5B**). For the variants that did not outright surpass the parental $GRAB_{rDA2m}$ capabilities, we picked the cell if the capability vastly outperformed the capabilities of any of the other cells on the array. For example, ROI number 181 achieved 3x the $\Delta F/F_0$ of the second-ranked variant (**Figure 4.6A, B,&C**). We can also observe interesting differential kinetics between some of our variants. For example, pick 12:181 decays faster than pick 18:90 throughout the imagine paradigm (**Figure 4.5B).**

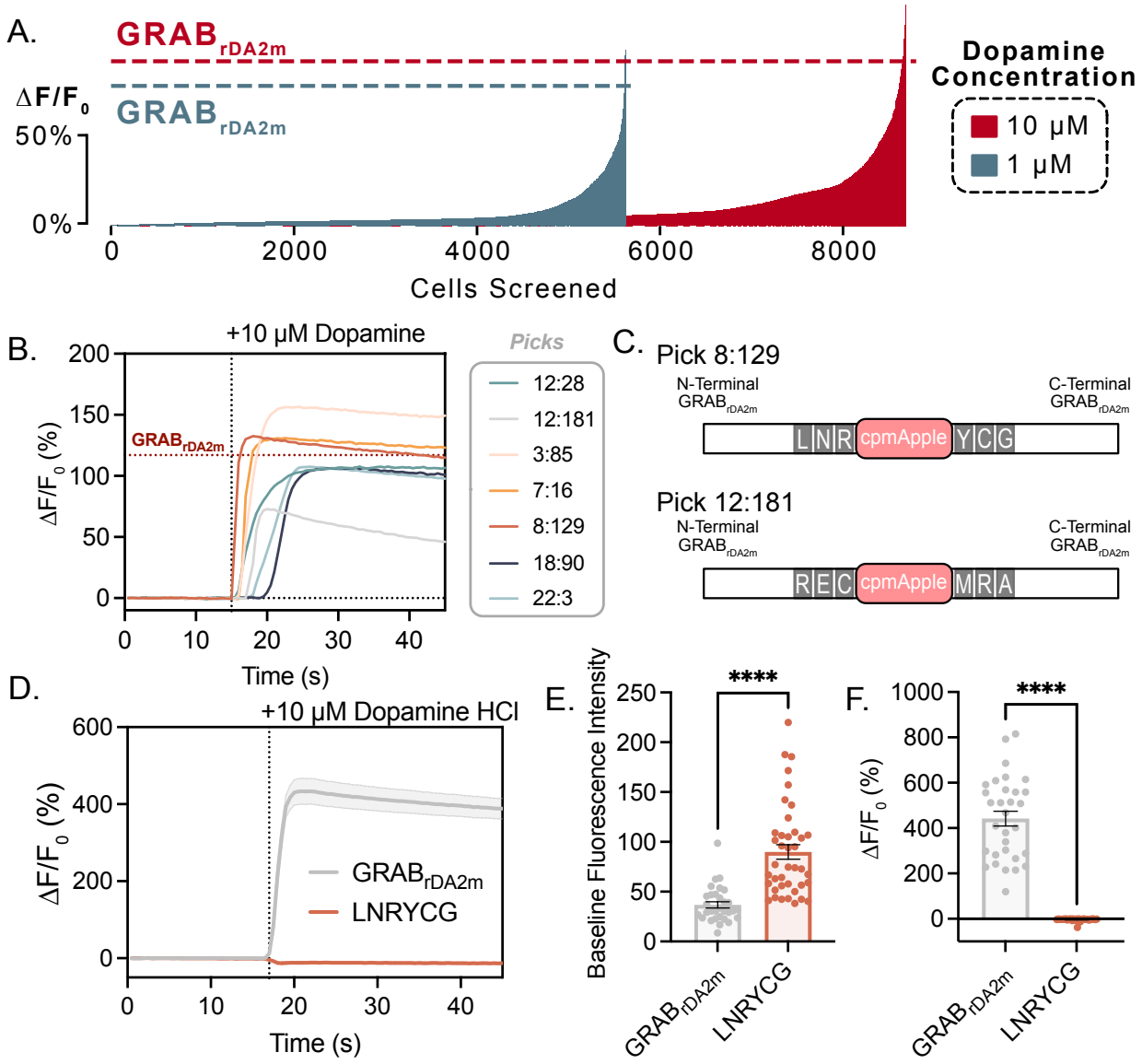


Figure 4.5 High-throughput Screening of The GRAB$_{rDA2m}$ Linker Library

A. The waterfall plot shows the ΔF/F$_0$ responses from the GRAB$_{rDA2m}$ linker library screening at the indicated dopamine concentration. The dotted lines depict the responses from the parental construct at the indicated dopamine concentration.

B. The time course demonstrates variants that displayed promising ΔF/F$_0$s. Each line depicts the response from a single cell pick. The horizontal dotted line represents the average response from the parental GRAB$_{rDA2m}$. The vertical dotted line represents the time dopamine was added dropwise.

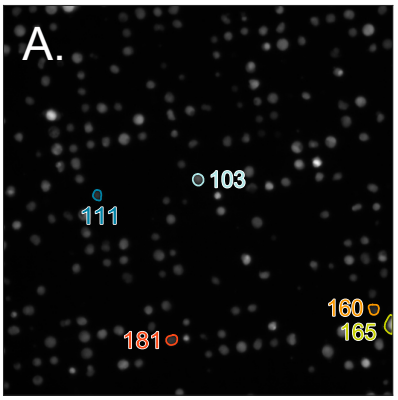
C. The genetic cassette demonstrates the recovered genetic identity of two of the GRAB$_{rDA2m}$ picks.

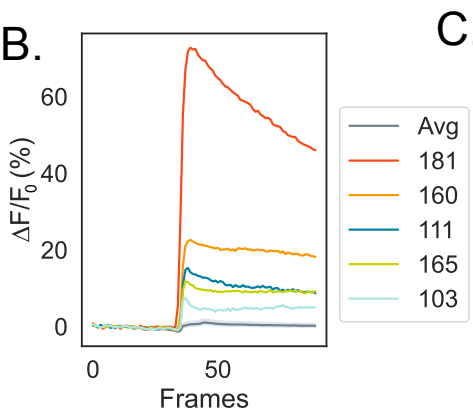
D. The bar plot displays the average baseline fluorescence for each indicated variant. Each scatter dot indicates the baseline of an analyzed cell. The bar depicts the mean and standard error of the mean. **** = P<0.0001.

E. The bar plot displays the average $\Delta F/F_0$ for each indicated variant in response to 10μM Dopamine. Each scatter dot indicates the max $\Delta F/F_0$ of an analyzed cell. The bar depicts the mean and standard error of the mean. **** = P<0.0001.

Of the picks we recovered from the array, we could only obtain the genetic identity of two of the picks. However, this was due to the availability of rt-PCR primers as opposed to picking changes. In picks 8:129 and 12:181, we recovered linker identities of LNRYCG and RECMRA, respectively (**Figure 4.5C**). Both picks displayed large quantities of contaminating DNA, which we suspect may be a factor of new picking pipettes. During the picking process, we observed that the picking pipette with the larger opening caused the imaging solution equalize within the pipette, due to hydrostatic pressure. This enables the RNA within the bath solution to be amplified during the rt-PCR.

We cloned the 8:129 variant into the CAG-GRAB$_{rDA2m}$ vector to validate its functionality. Interestingly, we observed no responses from the variant to 10 μM dopamine stimulus (**Figure 4.5D&F)**. However, this variant interestingly had a greater baseline fluorescence (**Figure 4.5E**). We suspect that this variant could have been one of the contaminating RNAs that amplified, rather than the actual variant we observed in the Opto-MASS screening. Again, we suspect this may be due to the increased width of the picking pipette and the need for further optimization.





C.

Rank	ROI	$\Delta F/F_0$
1	181	73
2	160	23
3	111	15
4	165	12
5	103	8

Figure 4.6 Example Output of Opto-MASS Screening Results

A. The image depicts $GRAB_{rDA2m}$ library cells seeded onto PDMS microwell arrays.
B. The time course demonstrates top five variants of the screened well alongside the well's average $\Delta F/F_0$ response. Each line depicts the response from a single cell pick.
C. The table indicates the ranked response of each of the cells analyzed in panel A, linked to their ROI identity, and $\Delta F/F_0$ achieved.

4.2.3 Fluorescence Activated Cell Sorting-Based Investigation of The $GRAB_{rDA2m}$ Linker Library

Within the Opto-MASS screening, we observed many variants display low functionality and no fluorescence (**Figure 4.5A**). This linker library, additionally, is very large, having an observable mutation space of 64,000,000 variants. Using the traditional Opto-MASS pipeline, we can screen approximately 10,000 variants daily. At this rate, we would have to screen for 100 days to screen 1/64 of the library. To speed up the discovery rate of promising variants, we have investigated the incorporation of FACS to pre-screen the library and remove any non-functional or non-viable mutations.

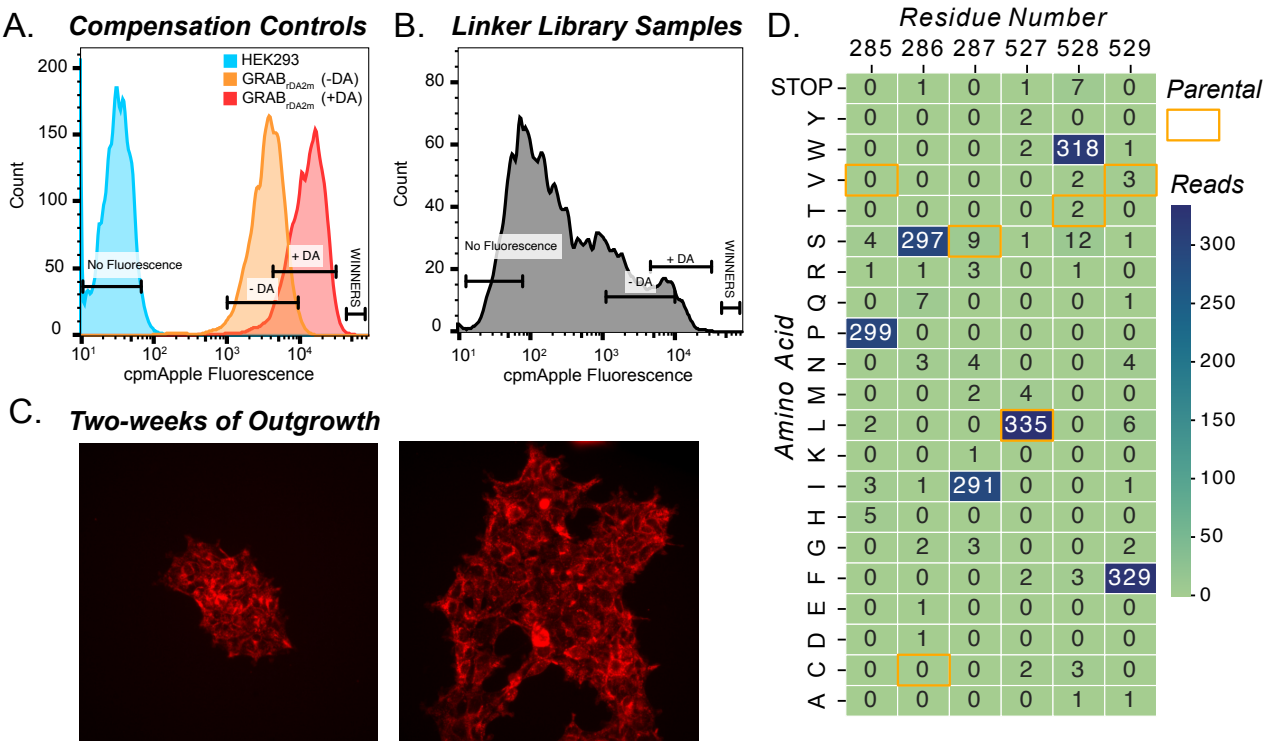


Figure 4.7 Fluorescence Activated Cell Sorting of The GRAB$_{rDA2m}$ Linker Library

A. The histogram plot displays the compensation controls and determination of gating used during sorting. No fluorescence indicates cp-mApple fluorescence of non-expressing HEK293 cells, GRAB$_{rDA2m}$ (-DA) indicates parental construct before dopamine addition, and GRAB$_{rDA2m}$ (+DA) is parental after dopamine addition. Brackets indicated gated populations, where the "winners" bracket consisted of cells sorted after dopamine addition.
B. The histogram plot displays the sorted linker library population.
C. The images demonstrate the growth of the sorted cells after two weeks of outgrowth.
D. The logo map displays the relative nucleotide frequency achieved at the indicated residues.

During FACS, we first sorted non-fluorescent HEK293 and parental GRAB$_{rDA2m}$ populations with and without dopamine (**Figure 4.7A**). We used the fluorescent responses of each population to determine gates where there is no fluorescence, baseline fluorescence of GRAB$_{rDA2m}$, stimulated fluorescence of GRAB$_{rDA2m}$, and a gate above the +DA GRAB$_{rDA2m}$ that we planned to sort. We observed a broad range of responses when we screened the variant library, though very few fell within the "winner" category (**Figure 4.7B**). We screened 3 million cells, and of this population, only 300 fell within the "winner" category. We allowed these cells to recover in tissue culture for two weeks before performing rt-PCR (**Figure 4.7C**). Within our rt-PCR results, we

found that the vast majority of the cells we recovered contained the mutation identity of PSILWF instead of the parental identity of VCSLTV (**Figure 4.7D**). Interestingly, the leucine at position 527 was conserved and may be important for sensor function.

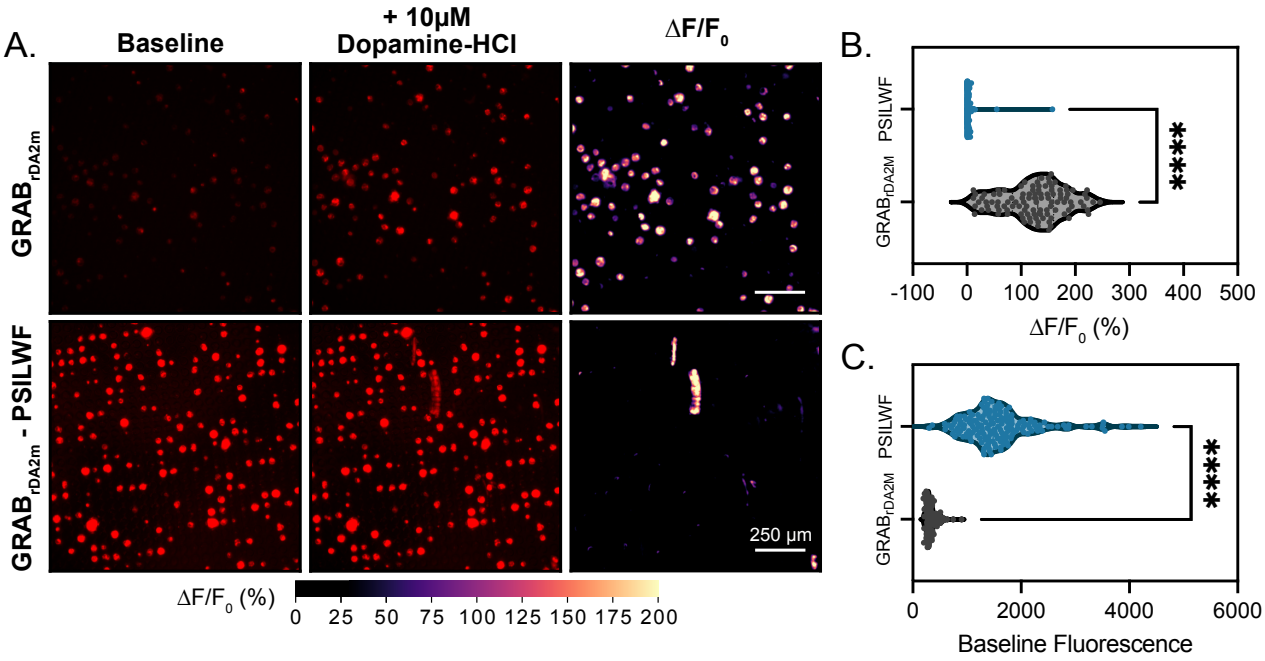


Figure 4.8 *In Vitro* Responses of FACS GRAB$_{rDA2m}$ Variant

A. The image set displays the fluorescence of landing pad-GRAB$_{rDA2m}$ and variant PSILWF before and after adding 10μM Dopamine and the $\Delta F/F_0$ achieved by each cell. Scale bar = 250 μm.
B. The violin plot displays the average $\Delta F/F_0$ for each indicated variant in response to 10μM dopamine. Each scatter dot indicates the max $\Delta F/F_0$ of an analyzed cell. **** = P<0.0001.
C. The violin plot displays the average baseline fluorescence for each indicated variant. Each scatter dot indicates the baseline of an analyzed cell. **** = P<0.0001.

After allowing the sorted cells to outgrow until confluency, we seeded the enriched library cells into PDMS microwell arrays and applied 10μM dopamine (**Figure 4.8A**). We found very little variance in the population's capabilities, including baseline fluorescence and in $\Delta F/F_0$, which is consistent with the sequencing results displaying striking agreement towards a singular variant (**Figure 4.8B, C**). The PSILWF variant displayed a very bright baseline fluorescence with no

change in fluorescence after dopamine addition. This indicates that we preferentially sorted cells that were very bright but do not display dynamic fluorescence.

4.3 DISCUSSION

$GRAB_{rDA2m}$ is a strong, red-shifted dopamine sensor with impressive membrane localization and photostability, though it has smaller $\Delta F/F_0$s than its photoactive counterparts. Within our optimization, we aimed to improve $GRAB_{rDA2m}$, not only due to its promising capabilities as a multiplexable dopamine indicator, but also due to the slow decay kinetics of the sensor. This was advantageous for our engineering efforts as it enabled us to have longer sorting cycles between stimulus addition.

We started our optimization by generating a site-saturated linker library of the $GRAB_{rDA2m}$ sensor. We chose three residues on either side of the cp-mApple, which has linker lengths of five residues and three residues for the 5' and 3' ends, respectively. We limited the number of sites to six due to the theoretical mutation space of 64 million, which would take a considerable amount of screening to properly sample a library of this size. With the six sites we chose, we balanced the mutated sites equally around the fluorescent protein. We chose the linkers of this protein because the residue linkers translate the conformational changes of the GPCR to the fluorescent protein. This makes the residue linkers a straightforward starting point for engineering.

To optimize the gene recovery process of our Opto-MASS screening, we altered two aspects of our current process. We began by altering the pipettes that are used for picking. We pulled pipettes that had a wider opening at the end of the picking pipette, which enabled us to provide negative pressure when picking the cell from the array. Objectively, this process worked well, allowing us to hold the cell at the tip of the pipette and reduce the loss of the cells due to surface tension. However, the slight opening at the pipette's end and the negative pressure

application allowed bath media inside the pipette. This introduced contaminating DNA to the gene recovery process, making discerning the promising variant's gene difficult after sequencing. This led to potentially incorporating the wrong variant in 8:129 and no variant functionality during validation. We additionally altered the sequencing process from Sanger sequencing to nanopore sequencing, which was much more interpretable than the results from Sanger sequencing. Moving forward, we can still improve the recovery process. One such alteration that could be made is to attempt cellular outgrowth with the optimized picking pipette that holds the cells. By allowing the cells to have several days to replicate, we can improve the recovery rate of the gene of interest while also diluting any possible contaminating DNA that may have been introduced during picking.

We investigated FACS as an avenue to quickly and crudely prescreen the library to remove nonviable variants and enrich the library for promising candidates. Toward this endeavor, we formed sorting gates based on non-fluorescence controls and the parental construct in both the apo and saturated states. With these controls, we designated four gates: no fluorescence, WT -DA, WT + DA, and winners. While sorting the linker library sample, we observed a massive shift of the average fluorescence of the library towards the no fluorescence gate. This highlights our observation *in vitro,* where many of the mutations were non-viable. After sorting the few cells that populated the “winner” gate and letting them outgrow, we were able to test their ability against the parental construct. We found that the variant was much brighter but lost its dynamic nature. This bias was introduced during our screening process but is something that could be mitigated using sophisticated gating strategies. For example, we could employ an initial sort and gate for cells with a proper baseline fluorescence before adding dopamine and sorting for cells displaying a promising fluorescence shift.

In conclusion, our optimization efforts to improve the $GRAB_{rDA2m}$ sensor have provided valuable insights into both the strengths and limitations of the current approaches. While we successfully identified potential avenues for improvement, our efforts were hampered by challenges in the gene recovery process and the introduction of bias during the FACS. These obstacles underscore the need for further refinement of our screening methods, particularly in improving the efficiency and accuracy of cell picking and reducing contamination during the gene recovery process. Ultimately, while the current approach has not yet resulted in a significantly improved sensor, it is a crucial process toward better understanding the steps necessary for forming sequence-to-function libraries. The strategies outlined here will provide a solid foundation for future endeavors to overcome the current limitations.

4.4 METHODS

4.4.1 Molecular Cloning

Predicted mutations were reflected into the pDisplay-GRAB-rDA2m-IRES-EGFP-CAAX backbone (Addgene; 208694) using point-mutation primers ordered from Integrated DNA Technologies (IDT) and PCR amplification with either Q5-polymerase (New England Biolabs; M0492L) or Superfi-II polymerase (Invitrogen; 12368010). Amplification of the DNA fragment was verified with agarose gel electrophoresis. Blunt-end DNA circularization was achieved with Kinase, Ligase, and DpnI enzyme (KLD) treatment (New England Biolabs: E0554S). Circularized DNA was transformed into competent *E.Coli* cells (DH5α or TOP10) and grown on agar plates that contain either ampicillin or kanamycin selection antibiotic (50 μg/mL). Upon colony formation, single colonies were picked and grown in 5mL cultures containing LB Broth (Fisher BioReagents; BP9723-2) and selection antibiotic (ampicillin/kanamycin; 50 μg/mL) overnight (37°C, 230 RPM). DNA was isolated using Machery Nagel DNA prep kits (Machery Nagel;

740490.250). Sanger sequencing (Azenta; Seattle, WA) of the isolated plasmid DNA was used to confirm the presence of the intended mutation.

Genes encoding the $GRAB_{rDA2m}$ variants were cloned into a CAG-driven backbone, pCAG-Archon1-KGC-EGFP-ER2-WPRE (Addgene; #108423), using Gibson assembly (New England Biolabs; E2621L). All subsequences were verified with Sanger sequencing (Genewiz; Seattle, WA).

4.4.2 Dopamine-HCL Assays

Human Embryonic Kidney (HEK293; ATCC Ref: CRL-1573) cells were cultured in Dulbecco's Modified Eagle Medium + GlutaMAX (Gibco; 10569-010) supplemented with 10% fetal bovine serum (Biowest; S1620). When cultures reached 85% confluency, the cultures were seeded at 100,000 cells per well or 50,000 cells per well in 24-well and 48-well plates, respectively. 24 hours after cell seeding, the cells were transfected using Lipofectamine3000 (Invitrogen; L3000015) at 1000 ng of DNA per well of a 24-well plate, according to the manufacturer's instructions.

48 hours post-transfection, the plates were prepared for imaging by washing and then replacing culturing media volume with imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES (triple supplemented with 1% Glutamax (Gibco; 35050-1), 1% Sodium Pyruvate (GIBCO; 11360-070), and 1% MEM Non-Essential Amino Acids (Gibco; 11140-050)). Crystalline power Dopamine Hydrochloride (Sigma Aldrich; H8502-25G) was resuspended into imaging solution (Tyrode's pH = 7.33; 125mM NaCl, 2mM KCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 30 mM Dextrose, 25 mM HEPES) into 2x the desired final concentration. During imaging, 1:1 volumes of the acetylcholine-tyrodes imaging solution were hand-pipetted into the bath volume to bring the final acetylcholine concentration to the desired

concentration. Imaging was performed on a sCMOS camera (Photometrics Prime95B) on an epifluorescent microscope (Leica DMI8) using a 20X objective (Leica HCX PL FLUOTAR L 20x/0.40 NA CORR). A Lumencor Light Engine LED and Semrock were used for fluorescence imaging.

4.4.3 Analysis of Fluorescent Assays

Analysis of HEK293 cell fluorescence imaging data was done using FUSE, a custom cloud-based semi-automated time series fluorescence data analysis platform written in Python. First, the cell segmentation quality of the selected Cellpose[90] model was manually verified. For the segmentation of cells expressing cytosolic fluorescent indicators, model 'cyto' was selected as our base model. If the selected Cellpose model was low-performing, we further trained the Cellpose model using the Cellpose 2.0 human-in-the-loop system[91]. Using an "optimized" segmentation model, fluorescence time-series data is extracted for each region of interest. This allows for unbiased extraction of change in cellular fluorescence information for a complete set of experimental samples. Using the raw fluorescence data, % fluorescence change from the baseline ($\Delta F/F_0$) over time was calculated using *Eq.1*.

4.4.4 Generation of Randomized Variant Libraries

Randomized mutation libraries were generated using Gibson assembly cloning. Inserts and backbones were amplified using PCR (Superfi-II polymerase (Invitrogen; 12368010)). Amplification of the DNA fragments was verified with agarose gel electrophoresis, and circularization was achieved using Gibson assembly (New England Biolabs; E2621L). The assembly was purified using the Machery Nagel PCR cleanup kit (Macher Nagel; 740611.250) before being transformed into 33 μL of electrocompetent cells (NEB cat no. C3020K) (2000 V, τ = 5 ms) in ice-cold cuvettes (1 mm gap) with one μL of the elution. The electroporated cells were

rescued in 967 µl of SOC for 1 hour at 37 °C. The rescue media was added to 100 mL of Luria Broth (LB) with supplemented with ampicillin and grown overnight at 37 °C and 240 rpm. Library DNA was isolated using the Machery–Nagel NucleoBond Xtra Midi EF kit (Machery Nagel; 740420.50). The mutagenesis of the library was verified using nano-pore sequencing (Plasmidsaurus).

4.4.5 Landing Pad Transfections

HEK293T landing pad cells that contain a recombination cassette, developed by the fowler lab, were stably recombined with the library prep using a double transfection protocol. The landing pad cells were standard cultured in 1–2 µg/mL doxycycline. One day before transfection, the landing pad cells were removed from doxycycline. On the day of transfection, the landing pad cells were seeded at 250,000 cells per well into 6-well dishes. The Fugene transfection reagents (per well) were prepared as followed: 3 µg of plasmid DNA encoding pCAG-Bxb1 recombinase; 6 µL of Fugene6 (Promega cat. # E2693); 300 µL of Opti-MEM. The reagents were left to complex undisturbed for 15 min and then added to the cell suspension. After 24 hours, the fugene transfection is repeated with the addition of the library plasmid in place of the pCAG-Bxb1 recombinase DNA. The transfected cells were transitioned to 4 µg/mL doxycycline media after 24 hours and 4 µg/mL doxycycline + 1 µg/mL doxycycline media after 48 hours.

4.4.6 Polydimethylsiloxane Microwell Array Formation and Seeding

The microwell arrays were fabricated, as discussed previously[40]. Briefly, an etched silicon wafer acts as a negative mold for the PDMS (Sylgard 184, Corning). The PDMS was formulated by mixing an equal ratio of 10:1 (w/w) and vigorously mixing. This mixture was poured onto the negative mold and placed into a desiccator for 10 minutes to remove air bubbles. The PDMS-containing wafer was then placed into a 70°C incubator overnight to cure. The cured PDMS was

removed from the mold and plasma treated before being stamped with bovine serum albumin (BSA) (FischerSci; Cat #BP1600). The PDMS arrays were then cut to size using a scalpel and placed into a 24-well dish.

4.4.7 Cell Seeding Onto Microwell Arrays

With the PDMS arrays placed in 24-well dishes, the entire plate was plasma-treated before quickly adding standard growth media to each well. The plate with media in the wells was then placed under vacuum to remove microbubbles from inside the wells. The landing pad cells were then removed from their standard growth dishes using 0.05% Trypsin-EDTA and before being counted. The cell suspension was normalized to a concentration of 500,000 cells per mL. 40 µL of this cell suspension was then carefully pipetted above each microwell array. The cells were returned to the incubator for 10 min to allow for the cells to seed into the cells through gravity. After 10 min, the 24-well plates were then placed in a centrifuge and spun down at 100 RCF for 5 min, to further seed the cells deeper within the wells. The media within each well was changed to remove any cells that did not occupy the microwells and to replace the array cells with doxycycline media.

4.4.8 Reverse Transcription PCR

Reverse transcription PCR was performed as discussed previously, and the current section contains excerpts from the manuscript[40]. Single-cell recovery tubes were prepared by diluting 5 µL of 0.1 M DTT into 200 µL of TE buffer; 5 µL of the TE/DTT solution was added to each PCR tube. After library screening, the single cell recovery tubes containing 5 µL 0.1 M DTT/ TE and the library pick were removed from the dry ice and placed on wet ice. Each tube was processed with reagents from the SuperScript IV First-Strand synthesis kit (Invitrogen cat. # 18091050). To each tube, 0.5 µL of 0.1 M DTT and 0.5 µL of RNase inhibitor were added. The samples were

placed on dry ice for 5 min and then moved back to wet ice. Next, 0.5 μL of the following was added to each tube, DI H20, a 10 mM dNTP mix, and 2 μM primer.

Next, the primers were annealed to the mRNA by incubating the samples at 95 °C for 30 s, 4 °C for 1 min, and 65 °C for 5 min. The samples were then returned to the wet ice. Next, 2 μL of SSIV RT 5× Master Mix was added to each sample. The samples were pipetted up and down thoroughly. Finally, 0.5 μL of the SSIV RT Enzyme was added to each tube, and the samples were thoroughly pipetted up and down. To perform the reverse transcriptase reaction, the samples were incubated at 53 °C for 10 min and then at 80 °C for 10 min. Next, 0.5 μL of RNaseH was added to each tube, and the samples were incubated for 20 min at 37 °C to remove any mRNA from the cDNA. Samples were stored at −20 °C before PCR amplification of cDNA with Q5 (New England Biolabs) or SuperFiII (ThermoFischer) using recombination-specific primers.

3.3.12 Fluorescence Activated Cell Sorting

To prepare cells for FACS, the cells were trypsinized for 5 minutes at 37 °C and rescued using a standard growth medium to generate a single-cell suspension. The cell population was counted using a hemocytometer and the remaining cell population was transferred to a sterile 15mL tube and pelleted at 0.5 rcf for 5 minutes. After spinning, the supernatant was discarded, and flow sorting buffer (98 mL HBSS (Gibco; 14025-092), 2 mL FBS (Biowest;91620), 1 mL HEPES (1M) (Gibco; 15630-080) was added to the pellet to achieve a final cell concentration of 2 million cells per 1 mL of sorting buffer. After sorting buffer was added, the cells were kept on ice throughout the flow sorting steps. The cells were sorted on a BD FACS ARIA III, and just prior to being loaded onto the machine, the cells were passed through cap of a 5 mL Polystyrene Round-Bottom Tube with Cell-Strainer Cap (Falcon; 352235). Cells were sorted into sterile 15 mL tubes that contained recovery media (10 mL of DMEM + 10% FBS + 200 μL P/S). After

sorting, the recovered cells were spun at 0.5 rcf for 5 minutes, and the supernatant was removed using a 1000 μL pipette. Fresh recovery media was used to resuspend the cell pellet and the cells were transferred to a sterile 24-well plate to allow for outgrowth.

Chapter 5. Proposed Advancements in Sequence-to-Function Library Generation for Supervised Model Training

ABSTRACT

This chapter explores advancements that will help develop sequence-to-function libraries for regressor and classifier machine-learning implementations. We begin by discussing the formation of mutation datasets and proper data-handling techniques. Key considerations for improving future datasets are outlined, including defining sequence space, ensuring balanced mutation representation, and identifying both 'good' and 'bad' mutations for proper model training. We then discuss an accelerated approach that can be employed to develop sequence-to-function libraries for regressor training. The approach we outline includes a collaboration with the Fowler lab to perform *in situ* sequencing to link empirically derived sensor performance to a DNA barcode. We end this chapter by discussing fluorescence-activated cell sorting as a mechanism to develop libraries suited for classifier training. In this method, we would employ an iterative sorting technique to capture the level of fluorescence change after ligand addition that will be used to sort the library into separate dates. Combining the gated performance with the identity of the mutants returned during rt-PCR can be used as input for classifier training. Implementing these methodologies promises to accelerate the generation of mutation libraries for supervised model training, enhancing the applicability of machine learning in future protein engineering studies.

5.1 PROPOSED IMPROVEMENTS

5.1.1 Considerations for Improved Sequence-to-Function Libraries

We would like to acknowledge that the dataset used to train the ensembles in both the GCaMP and RCaMP studies were biased toward influential residues, as the mutated residues were chosen through crystal structure analysis and previous empirical insight. Whether this is truly a limitation is difficult to ascertain. We observed that highly mutated positions tend to come to the forefront of final predictions; however, this does not mean that the residues are not influential or that the mutations that the ensemble suggests cannot be exploited further. Likewise, even with biases in the mutation library, it did not preclude less explored residues from being chosen as influential in sensor performance. We ultimately benefitted more from the well-characterized mutation library that was already published than what was detracted via biased predictions. We would like to point out that the mutational dataset was not intentionally formed with machine learning in mind, yet the information found within was still invaluable and capable of training machine learning models.

As machine learning studies become more prevalent, several considerations for data acquisition may help generate datasets better suited for machine learning extrapolation. First, sequence space and dimensionality must be well-defined. Smaller dimensionality offers more in-depth analysis and comprehension of combinatorial mutations. At the same time, larger numbers of residue positions will span a much greater sequence space but limit the study to small iterations from the starting sequence. Data acquisition should have equal numbers of mutations per residue in their characterization to avoid any potential biases that may arise due to unbalanced prevalence. Furthermore, identifying 'bad' mutations is as vital to proper training as 'good' mutations. The use case of iterative model training, in which the user is informed by machine learning and then retrains

the model with additional information, is an ideal application of this technology. However, testing only promising variants should be avoided, as this may introduce bias into the dataset during retraining. Testing mutations at sites where the ensemble shows significant prediction variability can increase understanding.

5.1.2 In-Situ Sequencing to Generate Mutation Libraries for Regressor Training

After developing our machine learning pipeline, a significant goal was the in-house development of sequence-to-function libraries that can be used to train our regressor models. Importantly, the output for regressor models are continuous numerical values, which provides us the level of nuance between variants that we were able to leverage in chapters two and three[112]. While the data we have utilized thus far has been invaluable, the data was acquired over seven years[7,113], a rate-limiting step within our machine learning pipeline. In collaboration with the Fowler Lab, we propose an accelerated approach that couples several projects between our two labs.

Using the landing pad cells developed in the Fowler Lab, we will express variant libraries based on machine learning insights. One modification to the Opto-MASS protocol within this instance is the inclusion of short, unique DNA barcodes. These cells will then be sparsely seeded onto 96-well plates, where phenotype will be recorded using traditional stimulation techniques, such as applying acetylcholine or stimulating with 488nm light to observe photoactivation tendencies. After obtaining phenotypic capabilities for each cell, the Fowler Lab will perform SENSE, an all-optical *in situ* sequencing method based on FISSEQ, which is being developed within the Fowler lab[114]. In this method, the stimulated well is fixed using PFA to preserve the DNA barcode, and cDNA is formed by *in situ* reverse transcription. The resultant cDNA undergoes amplification and sequencing using reversibly terminated fluorescent deoxyribonucleotides. The

resultant data is an image stack that contains fluorescent label information for each position in the DNA Barcode that can then be used to link the identity of the barcode to the mutant contained. Using FUSE, a deep-learning-based cell analysis pipeline developed in the Berndt Lab, we will assign ROIs to cells during their stimulus protocol, which can then be transferred to their *in situ* sequencing results (**Figure 5.1**).

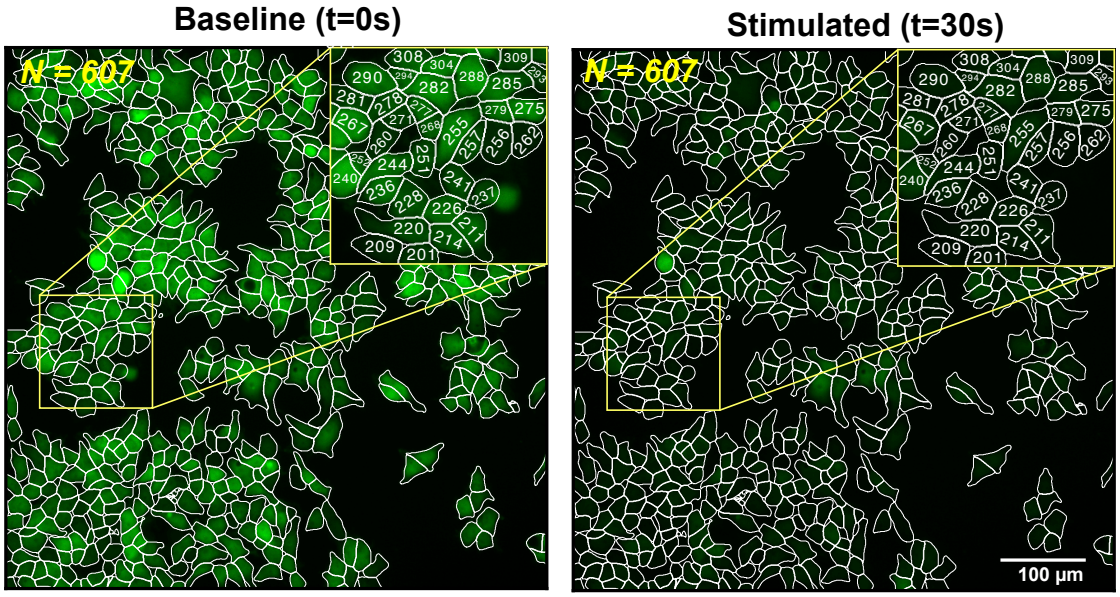


Figure 5.1 Demonstration of FUSE, a Deep-Learning Cell Segmentation Pipeline
Images depict cell segmentations before and after acetylcholine stimulus in CaMPARI expressing cells. ROI labels are formed using FUSE and are used to generate $\Delta F/F_0$ quantifications. (N=607 cells segmented, scale bar depicts 100 µM, images are separated by 30s/60 Frames).

With the successful application of this *in situ* sequencing technique, we gain the ability to improve the speed of mutation library generation dramatically. This also provides the ability for reciprocal learning, in which the mutation screen is informed by the ML pipeline, which is, in turn, further trained by screening results. This methodology forms a symbiotic relationship between ML algorithms and high-throughput screening. It has the potential to significantly reduce the time required to form sequence-to-function libraries needed for model training.

5.1.3 Fluorescence Activated Cell Sorting to Generate Mutation Libraries for Classifier Training

In Chapter 4, we investigated the use of fluorescence-activated cell sorting (FACS) to enrich our mutation libraries with functional variants. The successful application of this methodology potentiates the ability to create sequence-to-function libraries through strategic gating that would be compatible with classifiers. Notably, within classifiers, we would obtain a qualitative output of the sensor performance (i.e., small $\Delta F/F_0$, medium $\Delta F/F_0$, large $\Delta F/F_0$) that would also be returned to us during predictions on novel sequences, as opposed to the continuous values that we have obtained in our regressor training used thus far. While we do lose the nuance that we have been afforded with the training of regressors, this methodology seeks to reduce time necessary to develop sequence-to-function libraries.

Within this approach, we propose first to screen our compensation controls to define gates that capture the wild type's change in fluorescence (**Figure 5.2A, D**). After determining the gates, we will then pre-sort the linker library samples prior to adding the stimulating ligand. We will focus on sorting the non-fluorescent samples, 1, and samples that show baseline fluorescence similar to our wild type, 2* (**Figure 5.2B, D**). With the 2* population, we will then add the stimulating ligand and re-sort the library, where we will sort the final 2, 3, and 4 populations (**Figure 5.2C, D**). With these populations, we can perform rt-PCR to determine which mutants comprise each population.

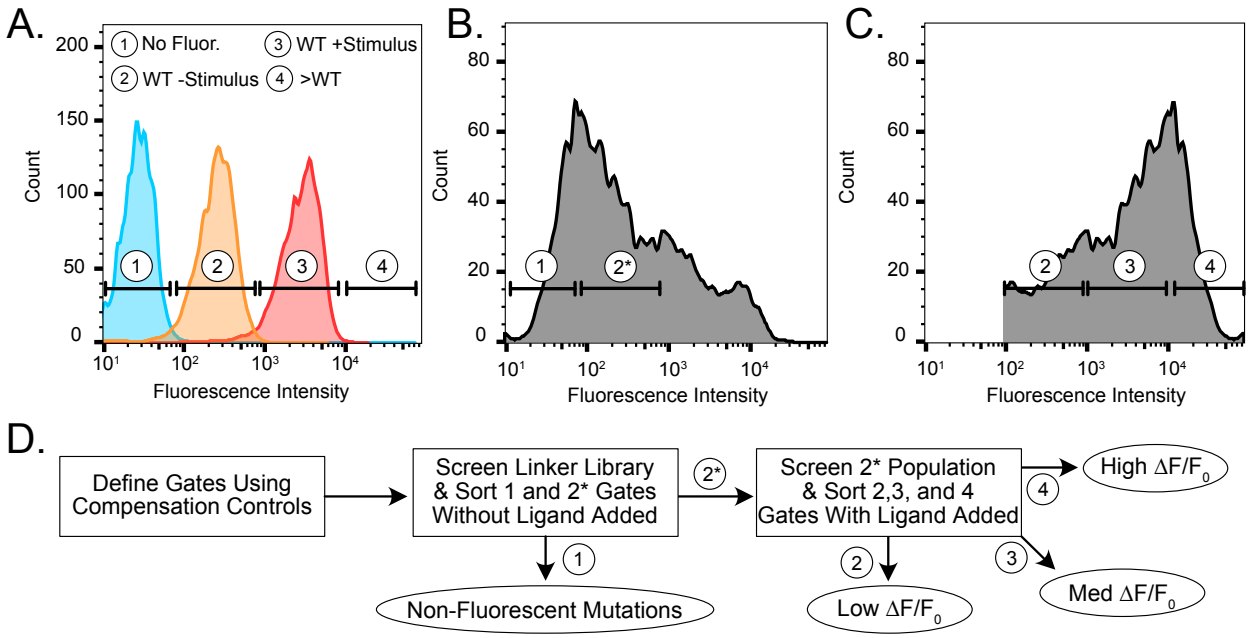


Figure 5.2 Potential Sorting Strategy to Generate Mutation Libraries

a. Hypothetical histogram plot of compensation controls and defined gates from defined populations.
b. Hypothetical histogram plot of linker library sample prior to ligand addition and gates used for sorting.
c. Hypothetical histogram plot of 2* population after ligand addition and gates used for sorting.
d. The flow diagram depicts steps taken to sort populations 1,2,3, and 4 for rt-PCR.

With the sorted cell populations, we will perform rt-PCR to amplify the DNA of variants within each sorted population. After nanopore sequencing, we can link the variant sequence to the performance rate (1: non-fluorescent, 2: no change, 3: medium change, 4: large change), which can be used as input to a machine-learning classifier. This would entail altering the current methodology that we have been employing to replace the regressor models with classifiers. However, the current model architecture can remain intact to capture similar capabilities.

Chapter 6. Concluding Remarks

Optogenetics has enabled unprecedented understanding of the inner workings of the mammalian nervous system. It harnesses naturally occurring light-sensitive proteins to alter or monitor how cells interact within their native environment. Optogenetics is widely applicable in cellular biology; however, each application requires tools with specific biophysical characteristics that match experimental needs. To obtain these tools, researchers have historically undertaken long, resource-intensive engineering cycles to find improved proteins with the desired biophysical characteristics. Each engineering cycle is based on hypotheses that necessitate a deep understanding of microscopic intraprotein interactions, requiring low angstrom crystal structures and intellectual burden. These trial-and-error-based mutagenesis studies limit the throughput of protein engineering and expose a critical need for improved mutational insights.

Toward this goal, I first demonstrated that machine learning can guide our engineering of optogenetic sensors. To do so, I developed a machine learning ensemble that comprised a random forest regressor, K-neighbors regressor, and multilayer perceptron network regressor. This ensemble was trained using a sequence-to-function library of the protein GCaMP. Within cross-validation, we observed impressive learning with R^2 values above 0.8. We used this method to predict the functional capabilities of 1426 novel sequences and identified mutations to test *in vitro* based on our models' predictions. We found that many ensemble predictions matched the observed outcome *in vitro*. We successfully implemented a computational tool capable of learning from empirically derived sequence-to-function libraries and suggesting mutations that led to desired biophysical characteristics. Using this technique, we identified three new GCaMP variants that we termed ensemble-GCaMPs (eGCaMPs) that display not only large $\Delta F/F_0$s but also fast decay kinetics.

We demonstrated that this same methodology these results were not specific to the GCaMP study and that this same methodology can be used to engineer other sensors, including the red-shifted calcium indicator jRCaMP1b. We followed the same computational approach with a sequence-to-function library for jRCaMP1b and sought to optimize the kinetics, sensitivity, and dynamic range. We took care to directly link the model predictions with the observed performance in primary cortical neurons. We found that the model's suggestions impressively matched what we observed *in vitro*, with many promising candidate variants that show impressive sensitivity and kinetics. We also took this study one step further to demonstrate that the insights derived from the model predictions can be used as targets for combinatorial mutagenesis through high-throughput screening. Within this study, we lend credence to the validity of the results derived in the engineering of GCaMP and illustrate that the model predictions can be used in non-traditional ways that complement high-throughput techniques.

The rate-limiting step of the computational approach outlined in this thesis is the availability of sequence-to-function libraries that can be used to train models. We conclude our study by attempting to assuage common issues with the current high-throughput screening implementation and investigate new practices that can be used to develop mutation libraries for both regressor and classier supervised learning methods. With the successful execution of these techniques, we can create a symbiotic relationship between high-throughput screening and machine learning to dramatically accelerate the engineering of functional proteins.

In conclusion, this thesis presents a novel computational approach to optogenetic sensor engineering by leveraging machine learning to overcome the inherent limitations of traditional trial-and-error-based mutagenesis. By reducing the resource and intellectual burden of protein engineering, this methodology provides an efficient and data-driven alternative that can

significantly streamline the development of optogenetic tools with tailored biophysical characteristics. The success of this approach was observed across multiple sensors, including GCaMP and RCaMP, demonstrating the versatility of this framework to different optogenetic sensors. This work establishes a promising new paradigm in protein engineering, positioning machine learning as an indispensable tool for accelerating the creation of optogenetic tools and beyond.

Bibliography

1. Deisseroth, K. Optogenetics. *Nat. Methods* **8**, 26–29 (2011).
2. Gradinaru, V., Thompson, K. R. & Deisseroth, K. eNpHR: a Natronomonas halorhodopsin enhanced for optogenetic applications. *Brain Cell Biol.* **36**, 129–139 (2008).
3. Gunaydin, L. A. *et al.* Ultrafast optogenetic control. *Nat. Neurosci.* **13**, 387–392 (2010).
4. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12**, 229–234 (2009).
5. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
6. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
7. Dana, H. *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nat. Methods* **16**, 649–657 (2019).
8. Unger, E. K. *et al.* Directed Evolution of a Selective and Sensitive Serotonin Sensor via Machine Learning. *Cell* **183**, 1986-2002.e26 (2020).
9. Inoue, M. *et al.* Rational Engineering of XCaMPs, a Multicolor GECI Suite for In Vivo Imaging of Complex Brain Circuit Dynamics. *Cell* **177**, 1346-1360.e24 (2019).
10. Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *Elife* **5**, (2016).
11. Ovchinnikov, Y. A. Rhodopsin and bacteriorhodopsin: structure—function relationships. *FEBS Lett.* **148**, 179–191 (1982).
12. Nathans, J. & Hogness, D. S. Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. *Cell* **34**, 807–814 (1983).
13. Shichida, Y. & Matsuyama, T. Evolution of opsins and phototransduction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2881–2895 (2009).
14. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
15. Kim, C. K., Adhikari, A. & Deisseroth, K. Integration of optogenetics with complementary methodologies in systems neuroscience. *Nat. Rev. Neurosci.* **18**, 222–235 (2017).
16. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
17. Gradinaru, V. *et al.* Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141**, 154–165 (2010).
18. Zhang, F. *et al.* Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633–639 (2007).
19. Lozier, R. H., Bogomolni, R. A. & Stoeckenius, W. Bacteriorhodopsin: a light-driven proton pump in Halobacterium Halobium. *Biophys. J.* **15**, 955–962 (1975).
20. Yizhar, O., Fenno, L., Zhang, F., Hegemann, P. & Diesseroth, K. Microbial opsins: A family of single-component tools for optical control of neural activity. *Cold Spring Harb. Protoc.* **2011**, top102 (2011).
21. Kleinlogel, S. *et al.* Ultra light-sensitive and fast neuronal activation with the Ca2+-permeable channelrhodopsin CatCh. *Nat. Neurosci.* **14**, 513–518 (2011).
22. Zhang, F. *et al.* Red-shifted optogenetic excitation: a tool for fast neural control derived from Volvox carteri. *Nat. Neurosci.* **11**, 631–633 (2008).

23. Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).
24. 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* **111**, 229–233 (1992).
25. Heim, R. & Tsien, R. Y. Engineering green fluorescent protein for improved brightness, longer wavelengths and fluorescence resonance energy transfer. *Curr. Biol.* **6**, 178–182 (1996).
26. Heim, R., Cubitt, A. B. & Tsien, R. Y. Improved green fluorescence. *Nature* **373**, 663–664 (1995).
27. Heim, R., Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12501–12504 (1994).
28. Tsien, R. Y. Building and breeding molecules to spy on cells and tumors. *FEBS Lett.* **579**, 927–932 (2005).
29. Liu, L., He, F., Yu, Y. & Wang, Y. Application of FRET Biosensors in Mechanobiology and Mechanopharmacological Screening. *Front. Bioeng. Biotechnol.* **8**, (2020).
30. 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.* **96**, 11241–11246 (1999).
31. Klima, J. C. *et al.* Incorporation of sensing modalities into de novo designed fluorescence-activating proteins. *Nat. Commun.* **12**, 856 (2021).
32. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nat. Methods* **6**, 875–881 (2009).
33. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
34. Feng, J. *et al.* A Genetically Encoded Fluorescent Sensor for Rapid and Specific In Vivo Detection of Norepinephrine. *Neuron* **102**, 745-761.e8 (2019).
35. Patriarchi, T. *et al.* Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science* **360**, (2018).
36. Pierce, K. L., Premont, R. T. & Lefkowitz, R. J. Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.* **3**, 639–650 (2002).
37. Duffet, L. *et al.* A genetically encoded sensor for in vivo imaging of orexin neuropeptides. *Nat. Methods* **19**, 231–241 (2022).
38. Sun, F. *et al.* A genetically encoded fluorescent sensor enables rapid and specific detection of dopamine in flies, fish, and mice. *Cell* **174**, 481-496.e19 (2018).
39. Wan, J. *et al.* A genetically encoded sensor for measuring serotonin dynamics. *Nat. Neurosci.* **24**, 746–752 (2021).
40. Rappleye, M. *et al.* Optogenetic microwell array screening system: A high-throughput engineering platform for genetically encoded fluorescent indicators. *ACS Sens.* (2023) doi:10.1021/acssensors.3c01573.
41. Wait, S. J., Rappleye, M., Lee, J. D., Smith, N. & Berndt, A. Machine Learning Ensemble Directed Engineering of Genetically Encoded Fluorescent Calcium Indicators. *bioRxiv* 2023.04.13.536801 (2023) doi:10.1101/2023.04.13.536801.
42. Zhuo, Y. *et al.* Improved dual-color GRAB sensors for monitoring dopaminergic activity*in vivo*. *bioRxiv* 2023.08.24.554559 (2023) doi:10.1101/2023.08.24.554559.

43. Akerboom, J. *et al.* Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. *Front. Mol. Neurosci.* **6**, (2013).
44. Pierce, N. A. & Winfree, E. Protein design is NP-hard. *Protein Eng.* **15**, 779–782 (2002).
45. Mandecki, W. The game of chess and searches in protein sequence space. *Trends Biotechnol.* **16**, 200–202 (1998).
46. Acevedo-Rocha, C. G., Hoebenreich, S. & Reetz, M. T. Iterative saturation mutagenesis: a powerful approach to engineer proteins by systematically simulating Darwinian evolution. *Methods Mol. Biol.* **1179**, 103–128 (2014).
47. Arnold, F. H. Design by Directed Evolution. http://www.cheme.caltech.edu/groups/fha/old_website_2011_06_08/Arnold_ACR_1998.pdf (1998).
48. Farinas, E. T., Schwaneberg, U. & Glieder, A. Directed evolution of a cytochrome P450 monooxygenase for alkane oxidation. *Synthesis & Catalysis* doi:10.1002/1615-4169(200108)343:6/7<601::AID-ADSC601>3.0.CO;2-9.
49. Koveal, D. *et al.* A high-throughput multiparameter screen for accelerated development and optimization of soluble genetically encoded fluorescent biosensors. *Nat. Commun.* **13**, 2919 (2022).
50. Piatkevich, K. D. *et al.* A robotic multidimensional directed evolution approach applied to fluorescent voltage reporters. *Nat. Chem. Biol.* **14**, 352–360 (2018).
51. Nadler, D. C., Morgan, S.-A., Flamholz, A., Kortright, K. E. & Savage, D. F. Rapid construction of metabolite biosensors using domain-insertion profiling. *Nat. Commun.* **7**, 12266 (2016).
52. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
53. Baker, D. What has de novo protein design taught us about protein folding and biophysics? *Protein Sci.* **28**, 678–683 (2019).
54. Okwei, E. *et al.* Rosetta's Predictive Ability for Low-Affinity Ligand Binding in Fragment-Based Drug Discovery. *Biochemistry* (2023) doi:10.1021/acs.biochem.2c00649.
55. Katritch, V., Cherezov, V. & Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.* **33**, 17–27 (2012).
56. Deupi, X. & Kobilka, B. K. Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology* **25**, 293–303 (2010).
57. Dou, J. *et al.* Sampling and energy evaluation challenges in ligand binding protein design. *Protein Sci.* **26**, 2426–2437 (2017).
58. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
59. Saito, Y. *et al.* Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration. *bioRxiv* 2021.08.13.456323 (2021) doi:10.1101/2021.08.13.456323.
60. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
61. Saito, Y. *et al.* Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth. Biol.* **7**, 2014–2022 (2018).
62. Bedbrook, C. N. *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176–1184 (2019).

63. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8852–8858 (2019).
64. Zhang, Y. *et al.* Fast and sensitive GCaMP calcium indicators for imaging neural populations. *bioRxiv* 2021.11.08.467793 (2021) doi:10.1101/2021.11.08.467793.
65. Sun, F. *et al.* Next-generation GRAB sensors for monitoring dopaminergic activity in vivo. *Nat. Methods* **17**, 1156–1166 (2020).
66. Dong, A. *et al.* A fluorescent sensor for spatiotemporally resolved imaging of endocannabinoid dynamics in vivo. *Nat. Biotechnol.* **40**, 787–798 (2022).
67. Tian, L., Akerboom, J., Schreiter, E. R. & Looger, L. L. Chapter 5 - Neural activity imaging with genetically encoded calcium indicators. in *Progress in Brain Research* (eds. Knöpfel, T. & Boyden, E. S.) vol. 196 79–94 (Elsevier, 2012).
68. Nakai, J., Ohkura, M. & Imoto, K. A high signal-to-noise Ca(2+) probe composed of a single green fluorescent protein. *Nat. Biotechnol.* **19**, 137–141 (2001).
69. Kawashima, S. *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202-5 (2008).
70. Dong, X., Yu, Z., Cao, W., Shi, Y. & Ma, Q. A survey on ensemble learning. *Frontiers of Computer Science* **14**, 241–258 (2020).
71. Zhou, Z.-H. Ensemble Learning. in *Machine Learning* (ed. Zhou, Z.-H.) 181–210 (Springer Singapore, Singapore, 2021).
72. Yang, Y. *et al.* Improved calcium sensor GCaMP-X overcomes the calcium channel perturbations induced by the calmodulin in GCaMP. *Nat. Commun.* **9**, 1504 (2018).
73. Song, Z., Wang, Y., Zhang, F., Yao, F. & Sun, C. Calcium Signaling Pathways: Key Pathways in the Regulation of Obesity. *Int. J. Mol. Sci.* **20**, (2019).
74. Nausch, B., Heppner, T. J. & Nelson, M. T. Nerve-released acetylcholine contracts urinary bladder smooth muscle by inducing action potentials independently of IP3-mediated calcium release. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **299**, R878-88 (2010).
75. Souslova, E. A. *et al.* Single fluorescent protein-based Ca2+ sensors with increased dynamic range. *BMC Biotechnol.* **7**, 37 (2007).
76. Ding, J., Luo, A. F., Hu, L., Wang, D. & Shao, F. Structural basis of the ultrasensitive calcium indicator GCaMP6. *Sci. China Life Sci.* **57**, 269–274 (2014).
77. Dragicevic, P. Fair Statistical Communication in HCI. in *Modern Statistical Methods for HCI* (eds. Robertson, J. & Kaptein, M.) 291–330 (Springer International Publishing, Cham, 2016).
78. Zhang, Y. *et al.* Fast and sensitive GCaMP calcium indicators for imaging neural populations. *Nature* **615**, 884–891 (2023).
79. Fenno, L. E. *et al.* Comprehensive Dual- and Triple-Feature Intersectional Single-Vector Delivery of Diverse Functional Payloads to Cells of Behaving Mammals. *Neuron* **107**, 836-853.e11 (2020).
80. Kim, C. K. *et al.* Molecular and circuit-dynamical identification of top-down neural mechanisms for restraint of reward seeking. *Cell* **170**, 1013-1027.e14 (2017).
81. Akerboom, J. *et al.* Crystal structures of the GCaMP calcium sensor reveal the mechanism of fluorescence signal change and aid rational design. *J. Biol. Chem.* **284**, 6455–6464 (2009).

82. Barnett, L. M., Hughes, T. E. & Drobizhev, M. Deciphering the molecular mechanism responsible for GCaMP6m's Ca2+-dependent change in fluorescence. *PLoS One* **12**, e0170934 (2017).
83. Nasu, Y., Shen, Y., Kramer, L. & Campbell, R. E. Structure- and mechanism-guided design of single fluorescent protein-based biosensors. *Nat. Chem. Biol.* **17**, 509–518 (2021).
84. Ofer, D. & Linial, M. ProFET: Feature engineering captures high-level protein functions. *Bioinformatics* **31**, 3429–3436 (2015).
85. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
86. Yao, Z. & Ruzzo, W. L. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* **7 Suppl 1**, S11 (2006).
87. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
88. Wardill, T. J. *et al.* A neuron-based screening platform for optimizing genetically-encoded calcium indicators. *PLoS One* **8**, e77728 (2013).
89. AAindex. https://www.genome.jp/aaindex/?fbclid=IwAR3qnzYQsc3iI2Env6iGQ2K2JkPunC_f7Uv0vSzxCw8tMCItO5T3hZFKPxI.
90. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
91. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Methods* (2022) doi:10.1038/s41592-022-01663-4.
92. Klima, J. C. *et al.* Bacterial expression and protein purification of mini-fluorescence-activating proteins. *Research Square* (2021) doi:10.21203/rs.3.pex-1077/v1.
93. Catapano, L. A., Arnold, M. W., Perez, F. A. & Macklis, J. D. Specific neurotrophic factors support the survival of cortical projection neurons at distinct stages of development. *J. Neurosci.* **21**, 8863–8872 (2001).
94. Martin, D. L. Synthesis and release of neuroactive substances by glial cells. *Glia* **5**, 81–94 (1992).
95. Lambert, T. FPbase fluorescent protein property visualization. *FPbase* https://www.fpbase.org/chart/.
96. Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *Elife* **5**, (2016).
97. Wait, S. J. *et al.* Machine learning-guided engineering of genetically encoded fluorescent calcium indicators. *Nat. Comput. Sci.* **4**, 224–236 (2024).
98. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).
99. Fernandez Lahore, R. G. *et al.* Calcium-permeable channelrhodopsins for the photocontrol of calcium signalling. *Nat. Commun.* **13**, 7844 (2022).
100. Mellor, J., Grigoras, I., Carbonell, P. & Faulon, J.-L. Semisupervised Gaussian process for automated enzyme search. *ACS Synth. Biol.* **5**, 518–528 (2016).
101. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335–342 (2014).
102. Jokinen, E., Heinonen, M. & Lähdesmäki, H. mGPfusion: predicting protein stability changes with Gaussian process kernel learning and data fusion. *Bioinformatics* **34**, i274–i283 (2018).

103. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E193-201 (2013).
104. Patriarchi, T. *et al.* An expanded palette of dopamine sensors for multiplex imaging in vivo. *Nat. Methods* **17**, 1147–1155 (2020).
105. Zhuo, Y. *et al.* Improved green and red GRAB sensors for monitoring dopaminergic activity in vivo. *Nat. Methods* **21**, 680–691 (2024).
106. Hamid, A. A., Frank, M. J. & Moore, C. I. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733-2749.e16 (2021).
107. Lee, S. J. *et al.* Cell-type-specific asynchronous modulation of PKA by dopamine in learning. *Nature* **590**, 451–456 (2021).
108. Kim, H. R. *et al.* A unified framework for dopamine signals across timescales. *Cell* **183**, 1600-1616.e25 (2020).
109. Liu, C. *et al.* An action potential initiation mechanism in distal axons for the control of dopamine release. *Science* **375**, 1378–1385 (2022).
110. Mohebi, A., Collins, V. L. & Berke, J. D. Accumbens cholinergic interneurons dynamically promote dopamine release and enable motivation. *Elife* **12**, (2023).
111. Taniguchi, J. *et al.* Comment on "Accumbens cholinergic interneurons dynamically promote dopamine release and enable motivation." *eLife* vol. 13 (2024).
112. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160 (2021).
113. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
114. Feldman, D. *et al.* Optical pooled screens in human cells. *Cell* **179**, 787-799.e17 (2019).

Vita

PUBLICATIONS

Mollaoglu G, Jones A, **Wait SJ**, Mukhopadhyay A, Jeong S, Arya R, Camolotto SA, Mosbruger TL, Stubben CJ, Conley CJ, Bhutkar A, Vahrenkamp JM, Berrett KC, Cessna MH, Lane TE, Witt BL, Salama ME, Gertz J, Jones KB, Snyder EL, Oliver TG. "The Lineage-Defining Transcription Factors SOX2 and NKX2-1 Determine Lung Cancer Cell Fate and Shape the Tumor Immune Microenvironment". *Immunity*. 2018 Oct 16;49(4):764-779.e9. DOI: 10.1016/j.immuni.2018.09.020. PubMed PMID: 30332632; PubMed Central PMCID: PMC6197489.

Chalishazar MD, **Wait SJ**, Huang F, Ireland AS, Mukhopadhyay A, Lee Y, Schuman S, Guthrie MR, Berrett K, Vahrenkamp J, Hu Z, Kudla M, Modzelewska K, Wang G, Ingolia NT, Gertz J, Lum DH, Cosulich SC, Bomalaski JS, DeBerardinis RJ & Oliver TG. "MYC-driven small cell lung cancer is metabolically distinct and vulnerable to arginine depletion." *Clinical Cancer Research*. 2019 Aug 15; 25(16): 5107-5121. DOI: 10.1158/1078-0432.CCR-18-4140. PubMed PMID: 31164374; PubMed Central PMCID: PMC6697617.

Ireland AS, Micinski AM, Kastner DW, Guo B, **Wait SJ**, Spainhower KB, Conley CC, Chen OS, Guthrie MR, Soltero D, Qiao Y, Huang X, Tarapcsak S, Devarakonda S, Chalishazar MD, Gertz J, Moser JC, Marth G, Puri S, Witt BL, Spike BT, Oliver TG. "MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate." *Cancer Cell.* 2020 July 13; 38(1):60-78. DOI: 10.1016/j.ccell.2020.05.001

Poirier JT, George J, Owonikoko TK, Berns A, Brambilla E, Byers LA, Carbone D, Chen HJ, Christensen CL, Dive C, Farago AF, Govindan R, Hann C, Hellmann MD, Horn L, Johnson JE, Ju YS, Kang S, Krasnow M, Lee J, Lee S, Lehman J, Lok B, Lovly C, MacPherson D, McFadden D, Minna J, Oser M, Park K, Park K, Pommier Y, Quaranta V, Ready N, Sage J, Scagliotti G, Sos ML, Sutherland KD, Travis WD, Vakoc CR, **Wait SJ**, Wistuba I, Wong KK, Zhang H, Daigneault J, Wiens J, Rudin CM, Oliver TG. "New Approaches to SCLC Therapy: From the Laboratory to the Clinic." *Journal of Thoracic Oncology*. 2020 April 01; 15(4): 520-540. DOI: 10.1016/j.jtho.2020.01.016.

Olsen RR, Kastner DW, Ireland AS, Groves SM, Pozo K, Whitney CP, Guthrie MR, **Wait SJ**, Soltero D, Witt BL, Quaranta V, Johnson JE, Oliver TG. "ASCL1 represses a latent osteogenic program in small cell lung cancer in multiple cells of origin." *Genes & Development.* 2021 June 01; 35(11-12): 847-869. DOI: https://doi.org/10.1101/gad.348295.121

Demetra P. Kelenis, Kathia E. Rodarte, Rahul K. Kollipara, Karine Pozo, Shreoshi Pal Choudhuri, Kyle B. Spainhower, **Sarah J. Wait**, Victor Stastny, Trudy G. Oliver, and Jane E. Johnson. "Inhibition of Karyopherin β1-mediated nuclear import disrupts oncogenic lineage-defining transcription factor activity in small cell lung cancer." *Cancer Res* (2022) 82 (17): 3058–3073. Doi: https://doi.org/10.1158/0008-5472.CAN-21-3713

Rappleye, Michael, **Sarah J. Wait**, Justin Daho Lee, Jamison C. Siebart, Lily Torp, Netta Smith, Jeanot Muster, Kenneth A. Matreyek, Douglas M. Fowler, and Andre Berndt. 2023.

"Optogenetic Microwell Array Screening System: A High-Throughput Engineering Platform for Genetically Encoded Fluorescent Indicators." *ACS Sensors*, November. https://doi.org/10.1021/acssensors.3c01573.

Wait, Sarah J., Marc Expòsit, Sophia Lin, Michael Rappleye, Justin Daho Lee, Samuel A. Colby, Lily Torp, et al. 2024. "Machine Learning-Guided Engineering of Genetically Encoded Fluorescent Calcium Indicators." ***Nature Computational Science*** 4 (3): 224–36.

Lee, Justin Daho, Amanda Nguyen, Zheyu Ruby Jin, Aida Moghadasi, Chelsea E. Gibbs, **Sarah J. Wait**, Kira M. Evitts, et al. 2024. "Far-Red and Sensitive Sensor for Monitoring Real Time H_2O_2Dynamics with Subcellular Resolution and in Multi-Parametric Imaging Applications." *Accepted Nature Chemical Biology.* 2024.

SELECTED CONFERENCE PUBLICATIONS & TALKS

Wait S., Gargantiel M., Torp L., Wang Y., Lee J., Lin S., Kim C., Berndt A., "Machine Learning Directed Engineering of Genetically Encoded Calcium Indicators." Invited talk: 2024 Institute for Stem Cell Regenerative Medicine Symposium, Seattle, WA, USA – Received 1st Abstract Award

Wait S., Torp L., Lee J., Berndt A., "Protein sensors for real-time monitoring of reactive oxygen species, calcium, and opioids." Invited talk: 2024 Society for Biomaterials, Seattle, WA, USA – Received 1st Place Flash Talk

Wait S., Rappleye M., Lee J., Berndt A., "Machine Learning Directed Engineering of Genetically Encoded Calcium Indicators." *Optics and the Brain 2023*, Vancouver, British Columbia, Canada. 24-27 April, 2023. ISBN: 978-1-957171-21-0. Invited talk: 2023 Optics and the Brain, Vancouver, BC, CA

Wait SJ, Chalishazar MD, Huang F, Bomalaski JS, DeBerardinis RJ & Oliver TG. "MYC-driven SCLC has unique metabolic vulnerabilities." Journal of Thoracic Oncology. 2020 February; 15(2): S1-S40. Invited Talk: 2020 AACR-IASLC International Joint Conference: Lung Cancer Translational Science: From the Bench to the Clinic, San Diego, CA, USA