Engineering Fluorescent Protein Biosensors to Monitor Glycolytic Metabolism

By

John N. Koberstein

A DISSERTATION

Presented to the Neuroscience Graduate Program,

and Oregon Health & Science University School of Medicine

In partial fulfillment of

The requirements for the degree of:

Doctor of Philosophy

December 2021

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

John N. Koberstein

has been approved

Advisor, Richard H. Goodman

Member and Chair, Brian J. O'Roak

Member, Michael S. Cohen

Member, Philip J. S. Stork

Member, Carsen Schultz

Member, Gary Yellen

Table of Contents

Acknowledgements

I am fortunate to have been granted the opportunity to carry out the research presented in this thesis. I would like to thank my mentor, Richard Goodman. During my time in his lab, I was afforded the time and resources to pursue new ideas, fail repeatedly, and learn from my mistakes. I am thankful to have had his support and guidance throughout this process.

I am tremendously grateful for the assistance and advice provided by my labmates Chadwick Smith and Melissa Stewart. This work would not have been possible without their contributions.

I would also like to thank my dissertation advisory committee, Michael Cohen, Carsten Schultz, Phil Stork, and Brian O'Roak, who have provided me with important feedback. The diverse expertise present in this group has been essential in shaping my graduate studies.

I owe much to my peers across the OHSU graduate programs who have worked tirelessly to improve the conditions for all present and future students. In addition to their brilliance as scientists, they are some of the most kind and thoughtful people I have ever met.

Finally, I would like to thank the many professors and mentors who helped me reach this point: Mike Sardinia, Lee Anne Chaney, Craig Tsuchida, Jason Gerstner, Chris Davis, and Lucia Peixoto. My decision to pursue graduate training in scientific research was preceded by their generous teaching. My life has been thoroughly enriched as a result.

Abstract

Fluorescent protein biosensors have become an important tool for neuroscience research. Highly optimized biosensors for Ca^{2+} are widely used to study neuronal activity in live animals. Underlying the impressive functionality of these fluorescent probes has been a tremendous research and engineering effort to understand their function and engineer new variants with improved performance. While the potential for fluorescent biosensors to enable new directions in biological research is clear, biosensors for other targets have not reached the same widespread usage or impact. Underlying this is a combination of factors including the difficulty of generating new and improved biosensors and the challenge of identifying informative analytes to measure.

To address these issues, I have worked to accelerate biosensor design and optimization using a high-throughput protein engineering approach. In an initial demonstration, I show that hundreds of unique insertions of circularly permuted GFP into maltose binding protein (MBP) can be simultaneously assayed for brightness and dynamic range which resulted in the discovery of new high-dynamic-range variants. Furthermore, to demonstrate the use of massively parallel assays to optimize biosensors, I characterized linker mutations made to an existing pyruvate biosensor. This effort produced a variant with double the dynamic range compared to the starting sequence. In addition, the data generated from this assay was used to train machine learning models to predict dynamic range from linker sequence, providing valuable insights into the relative importance of different biochemical features at each linker position. Taken together this study demonstrates that massively parallel assays provide an efficient method to screen

diverse biosensor libraries for improved variants, while generating crucial sequence-function data.

For studies of metabolism, flux through a pathway is often the parameter of interest. However, flux being a rate is not amenable to measurement by fluorescent biosensors which naturally report concentrations. Recent research suggests that certain pathway intermediates exhibit concentrations that are intrinsically correlated with flux. For glycolysis the intermediate fructose 1,6-bisphosphate (FBP) is proposed to serve as a flux-signaling metabolite. Using the transcription factor CggR, which allows bacteria to sense and respond to changes in flux, I constructed a fluorescent biosensor for FBP. High-throughput assays were used to screen for a suitable site in CggR to insert cpGFP, followed by characterizing linker libraries to optimize dynamic range. The resulting FBP biosensor termed HYlight possess a sensitive ratiometric signal that makes it well suited for use in live cells. The use of HYlight for tracking changes in glycolytic flux with high spatiotemporal resolution was demonstrated primarily in pancreatic beta cells due to unique aspects of glycolytic metabolism in this cell type. The ubiquity and importance of glycolysis combined with the flexibility of fluorescent biosensors suggests HYlight will be broadly useful for studying glycolytic metabolism in many biological systems.

Chapter 1. Introduction

Scientific discovery is often stimulated by the development of tools that enable new ways to observe and measure nature. Metabolism is of fundamental importance to cellular physiology but remains difficult to study with high spatiotemporal resolution in live tissues. Fluorescent protein biosensors enable concentration measurements for specific analytes over time with single cell resolution. These genetically encoded reagents function by combining a ligand-binding domain (LBD) with a fluorescent protein (FP) such that ligand binding results in altered fluorescence intensity (Fig. 1). The site of FP insertion into the LBD and the linkers connecting the two domains are of critical importance for generating functional coupling. Currently, the optimal combination for these parameters cannot be predicted thus necessitating brute force screens for functional sequences. Two key complementary technologies, massively parallel assays and machine learning applied to protein engineering, show promise for overcoming existing bottlenecks. The successful generation of fluorescent biosensors for important metabolites will enable researchers to ask new questions about the regulation of metabolism in live cells.

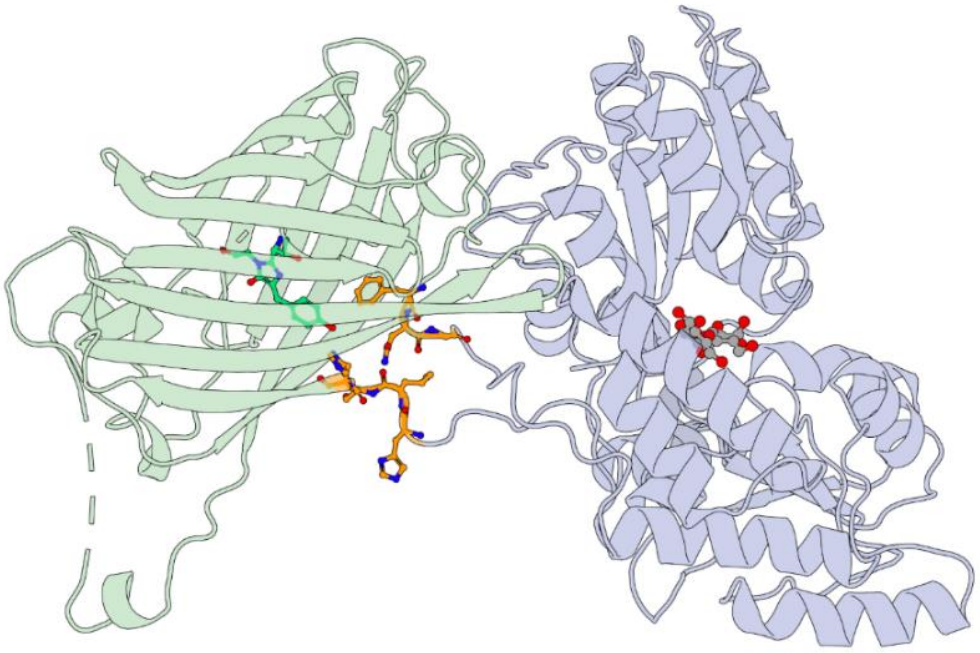

Figure 1. The structure of a prototypical single-FP biosensor (PDB: 3OSQ) consisting of cpGFP (green) inserted into a ligand-binding domain (blue) with linkers (orange) connecting the two domains[1,2].

1.1 Fluorescent proteins and biosensors

Fluorescence

The inner workings of the cell are of fundamental interest to the field of biology. However, it has historically been difficult to measure aspects of function without disturbing cellular integrity. A long-standing solution to this problem has been to probe the cell with light. In fact, the discovery of the cellular basis of life was aided by the technological development of the optical microscope. The utility of light in microscopy is derived from its ability to penetrate the cell with minimal disturbance and be partially absorbed, which generates the contrast needed to visualize cellular structures. The process of absorption results in a passing of energy from the photon to the absorbing molecule. Generally, this energy is dissipated as heat, but for some

molecules, called fluorophores, this energy can be emitted as a lower energy photon of light in a process called fluorescence. The brightness of a fluorophore is determined by how well it absorbs lights, quantified as the extinction coefficient, and how efficiently absorbed photons are converted to emitted photons, quantified as the quantum yield. Fluorescence intensity, which is the product of fluorophore concentration and brightness, can be quantified by measuring the number of emitted photons.

Fluorescence microscopy

Fluorescent molecules alter the energy, and thus the wavelength, of light. This feature presents an opportunity to interrogate the intracellular environment with increased specificity. To image fluorescent molecules inside cells, microscopes have been developed that illuminate the sample with a laser of the specific wavelength required for excitation and then measure filtered light containing only the specific wavelength of fluorescence emission. Fluorescent microscopes can generate images with enhanced contrast owing to the low intrinsic fluorescence exhibited by most cells. Significant research has been devoted to the chemical synthesis of cell permeable fluorescent dyes that interact with intracellular structures. These dyes can be used to “spy on cells” in the words of Roger Tsien[3], effectively sneaking into the cell and reporting back the hidden activities and structures contained within the intracellular environment. While small molecule dyes have found extensive use for fluorescence microscopy, biology has devised its own methods for synthesizing fluorophores which exhibit a variety of unique properties and advantages.

Discovery of a green fluorescent protein

Green fluorescent protein (GFP) has attracted significant research and engineering effort because of its widespread utility in biological research[4,5]. The discovery of this protein in 1962 was prompted by the observable emission of green light from the jellyfish *Aequorea victoria*[6]. Shimomura et al. isolated the pair of proteins responsible for this phenomenon. They showed the chemiluminescent enzyme aequorin oxidates coelenterazine emitting 470 nm light, which in turn excites the fluorescent protein (FP) causing emission of green 508 nm light[7,8]. While both proteins have applications as research tools, the requirement of aequorin, and other chemiluminescent proteins (firefly luciferase), for coelenterazine/luciferin limits their use in organisms that do not possess an endogenous source of the substrate. The finding that transgenic *E. coli* and *C. elegans* expressing the GFP protein from transgenic DNA exhibited fluorescence, indicated that no additional post-transcriptional modifications or cofactors, besides oxygen, were necessary[9]. This discovery suggested that GFP could possibly be utilized as a genetically encoded fluorescent tag in any aerobic biological system.

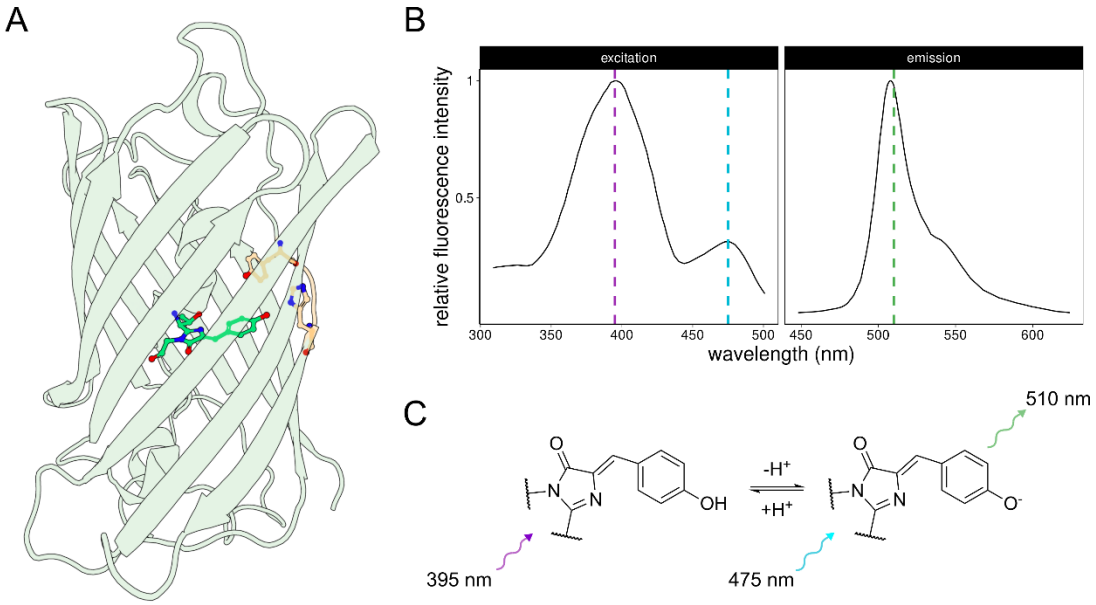

Figure 2. (A) The structure of GFP (PDB: 1EMA)[1,10]. Residues 145-148 (orange) produce are proximal to the fluorophore (bright green). **(B)** The excitation spectrum for wildtype GFP exhibits a major peak at 395 nm and a minor peak at 475 nm, while a single emission peak is present at 510 nm[11]. **(C)** The GFP fluorophore exists in an equilibrium between neutral and anionic states which are responsible for excitation at 395 and 475 nm respectively. Excited state proton transfer (ESPT) occurs with absorption of 395 nm light resulting in emission from a deprotonated state.

GFP structure

The GFP structure consists of an 11-stranded beta-barrel (Fig. 2A). Running through the center of the beta-barrel is an alpha-helix containing the fluorophore[10,12]. The fluorophore is buried near the center of the cylinder which serves to protect it from the surrounding solvent. The GFP amino acids 65-67, which are Ser-Tyr-Gly, form the fluorophore through an intramolecular reaction that proceeds after the protein is folded. First, the imidazolinone ring is formed by nucleophilic attack of the amide of Gly67 on the carbonyl of residue 65 followed by dehydration. Following cyclization, a dehydrogenation of residue Tyr66 results in conjugation between the phenol and imidazolinone moieties[13,14]. Elucidation of the

fluorophore chemical structure was crucial to early engineering efforts to adapt GFP for specific uses.

GFP fluorescence spectrum

The wildtype GFP fluorophore exhibits a major excitation peak at 395 nm and a minor peak at 475 nm (Fig. 2B). Because 375 nm UV light is considered cytotoxic, initial efforts to engineer GFP were focused on improving excitation with 475 nm light. The spectral properties of GFP can be altered by mutating the residues comprising the fluorophore, or by mutating neighboring residues that define the local chemical environment. The fluorophore residues Ser65 and Tyr66 can both tolerate some mutations with interesting effects, while residue Gly67 is strictly conserved in all known GFP mutants that retain fluorescence. An initial observation that the relative intensity of the two peaks is sensitive to pH suggested that the dual excitation wavelengths result from the protonation state of the fluorophore (Fig. 2C). The equilibrium between neutral phenol and anionic phenolate would then determine the relative intensity of the 395 and 475 nm peaks, respectively. The mutation Ser65Thr results in a dramatic increase in the relative intensity of the 475 nm excitation peak. This mutation does not change the conjugated portion of the fluorophore but does alter the side chain resulting in the loss of a key hydrogen bond from Ser65 to Glu222. This change to the hydrogen bond network ultimately prevents ionization of Glu222 which stabilizes the anionic phenolate leading to a dramatic enhancement of 475 nm excitation[15]. The Ser65Thr mutation in combination with a mutation that enhances folding efficiency (Phe64Leu) results in significantly improved brightness and folding at 37° C. This variant commonly

referred to as enhanced GFP (EGFP) became the standard for fluorescence microscopy using widely available 488 nm lasers[16].

Excited State Proton Transfer

While the neutral state predominates in wildtype GFP (wtGFP), excitation by 395 nm light does not directly result in strong fluorescence emission but instead drives the phenol to become more acidic. A transient third state in which the phenol is deprotonated is then responsible for fluorescence emission. This process of excited state proton transfer (ESPT) requires a proton acceptor to facilitate conversion. In wtGFP the proton travels along a chain of hydrogen bonds from the fluorophore to a water molecule, to the side chain hydroxyl of S205, finally terminating in protonation of the Glu222 carboxylate. Mutation of Glu222 to Gly prevents the final step in this pathway and disrupts ESPT. The Ser65Thr mutation also reduces the efficiency of ESPT by shifting the Glu222 sidechain and disrupting the hydrogen bonding pathway[15]. Understanding the complexities of the chemical states and processes underlying GFP fluorescence provide an avenue for generating dynamic changes in fluorescence.

Red fluorescent proteins

Efforts to red-shift the GFP excitation and emission spectra by random mutagenesis while initially successful in producing yellow fluorescent proteins eventually produced diminishing returns. Scientists instead looked for naturally occurring homologous proteins with desired spectral properties. Red fluorescent proteins derived from mushroom coral (*Discosoma sp.*)[17] and bubble-tip anemone (*Entacmaea quadricolor*)[18,19] are spectrally separated from GFP with excitation and emission maxima at greater than 560 and 580 nm respectively. The significant red-

shift is produced by an additional oxidation reaction in the peptide backbone fluorophore residue Gln66 (Ser65 in GFP). RFPs have been iteratively mutated and selected for improved maturation, brightness and monomericity resulting in a unique lineage like that for GFP[20]. The spectral compatibility of GFP and RFP variants has been an important achievement for generating multiplexed fluorescence measurements in single cells.

1.1.1 Fluorescent Proteins as Active Indicators

Sensitivity of GFP fluorescence to pH

While the initial applications of GFP consisted of tagging proteins of interest to detect aspects of spatial distribution and abundance[9], it was quickly discovered that the intensity of emission could serve as an active indicator of cellular parameters. *In vitro* assays of purified proteins revealed that many GFP variants were pH sensitive. The sensitivity of fluorescence intensity at a given excitation wavelength can be attributed to changes in the protonation equilibrium of the fluorophore phenol moiety[13,14]. These sensitized GFP variants can then serve as intracellular reporters to measure the pH of the cytoplasm or specific organelles through appropriate targeting sequences[21]. In this manner, a simple biochemical observation of GFP was utilized to produce one of the first examples of a completely genetically encoded fluorescent biological sensor, which we will refer to as a biosensor.

FRET

An additional motivation underlying the efforts to expand the spectral diversity of fluorescent proteins was an interest in exploiting a phenomenon known as Förster resonance energy transfer (FRET) to produce biosensors. FRET occurs

when two fluorophores are in <100A proximity and the emission spectra of one fluorophore (termed the donor) overlaps with the excitation spectra of the other (termed the acceptor). In such a configuration, excitation of the acceptor can result in energy transfer through space between the two fluorophores causing emission from the donor. The efficiency of this effect depends on the distance and orientation of the two fluorophores, providing a conceptually simple mechanism for generating fluorescent changes in response to biochemical events[22,23]. For example, a simple protease activity biosensor was generated by fusing a blue emitting fluorescent protein (BFP) to EGFP separated by a protease sensitive linker[24]. Initially, FRET efficiency between the BFP donor and GFP acceptor will be high due to the forced proximity, but following cleavage by the protease, FRET efficiency will decrease as the FPs diffuse away from each other. A more generalizable design for dynamic and reversible FRET biosensors consists of a donor and acceptor FP attached to the termini of two interacting proteins. For example, calcium biosensors have been constructed by attaching BFP to M13 and GFP to Calmodulin, which binds M13 in a Ca^{2+} dependent manner[25]. Alternatively, the donor and acceptor can be attached to the N- and C-termini of a single protein that undergoes conformational changes with ligand binding resulting in changes in the distance and orientation of the two fluorophores.

Advantages and disadvantages of FRET

The major advantage of FRET biosensors is the inherently ratiometric readout. Typically, the intensity for both the donor and acceptor are measured and used to calculate a ratio parameter. The donor to acceptor ratio is independent of confounding variables such as differences in protein concentration and movement

artifacts that are often encountered when imaging samples. In addition, the mechanism underlying FRET, changes in distance, is easy to conceptualize and engineer onto existing proteins. While these properties make FRET-based biosensors easier to design, image, and quantify, there are several limitations to their use. By design, two emission channels are occupied by a single FRET biosensor which limits opportunities for measuring multiple biosensors in parallel. In addition, the passage of light through tissue is wavelength dependent with longer wavelengths experiencing decreased scattering, which poses a challenge for quantifying the two emission wavelengths at different tissue depths. Finally, even for highly optimized biosensors, the maximal changes in FRET efficiency elicited by a given biochemical stimulus are often relatively small.

1.1.2 Single Fluorescent Protein Biosensors

Discovery of GFP circular permutation

An increasingly common and useful biosensor design involves dynamic modulation of the intensity derived from a single fluorescent protein. So called single fluorescent protein biosensors (SFPBs) generally rely on a circularly permuted FP (cpFP) in which the original N and C termini have been joined by a flexible linker while new termini have been created by opening an internal site of the protein. The tolerance of GFP to circular permutation stemmed from the discovery that a mutant featuring the replacement of residue Tyr145 in CFP with a hexapeptide surprisingly retained fluorescence[26]. This finding suggested that insertion of larger domains, or circular permutation with the new termini placed at amino acid 145 might also be possible. Indeed, these two modifications were not only tolerated, but of foundational importance for generating improved fluorescent

biosensors. The Tyr145 insertion site resides in a small bulge extruding from the beta barrel that makes space for the fluorophore phenol (Fig. 2A)[10]. By disrupting the beta barrel at residue 145 the key residues surrounding the fluorophore that determine the equilibrium between anionic phenolate and neutral phenol can be replaced. Ideally, these new residues and their interactions with fluorophore will be dynamically rearranged upon binding of Ca^{2+} resulting in changes in the state of the fluorophore and thus emission intensity. Indeed, insertion of Calmodulin between amino acid 145 and 146 of EYFP produced a fusion protein with large changes in 488 nm fluorescence between the calcium-bound and apo states. This construct, created in 1999 and referred to as Camgaroo, was the first engineered single fluorescent protein biosensor[26].

Single-FP biosensors for Ca^{2+}

In 2000, two papers released in quick succession provided the blueprint for what remains the most used series of fluorescent biosensors. Instead of inserting a binding domain directly into an FP, these new biosensors relied on the recently developed circularly permuted GFP (cpGFP) in which the original termini are joined by a short linker and new termini are created at amino acid 145. The two biosensors, pericam[27] and GCaMP[28], each consist of a circularly permuted FP (YFP and GFP respectively), with M13 attached to the N-terminus and Calmodulin to the C-terminus. The residues comprising the junction between the cpGFP termini and Calmodulin/M13 are proximal to the fluorophore and through ligand-dependent conformational changes can alter the local chemical environment resulting in altered fluorescence intensity. Indeed, these two prototypical genetically encoded calcium indicators (GECIs) exhibited relatively large fluorescence responses to

Ca^{2+} when compared to existing FRET biosensors. Additionally, from a theoretical point of view, the single fluorescent protein biosensor design might yield vastly enhanced signal to noise levels with further optimization.

Mechanistic insights into GCaMP function

In the 21 years since its initial publication, the GCaMP biosensor design has been iteratively optimized. Among many properties, one consistent goal is to maximize the relative change in the 488 nm excited fluorescence between the Ca+-bound and apo states, referred to as dynamic range. The brightness of a fluorophore is determined by the concentration, how well the fluorophore absorbs light (the extinction coefficient), and how efficiently absorbed light is translated into emitted light (quantum yield). Therefore, the Ca^{2+} dependent increases in brightness for GCaMP biosensors must ultimately depend on some combination of changes in concentration, extinction coefficient, or quantum yield. For GCaMP6 which has been studied extensively, the major contribution lies in rapid changes to the effective concentration of the fluorophore, in this case specifically the state of the fluorophore which absorbs 488 nm light. Although for GCaMP biosensors the major excitation occurs at 488 nm rather than 400 nm light, the fluorophore still exists in an equilibrium between a neutral phenol and anionic phenolate. In the unbound state, almost all the fluorophore is the protonated neutral form and thus absorbs 405 nm light, but due to inefficient ESPT does not strongly fluoresce. Upon binding Ca^{2+}, a change in the fluorophore pK_a results in a shift to about 50% deprotonated fluorophore. The relative increase in abundance of the anionic state causes the observed increase in fluorescence emission when excited by 488 nm light. The change in pK_a can be largely attributed to the displacement of the

negatively charged Glu148 sidechain away from the fluorophore with binding which alters the local electronic potential[29]. Despite the demonstrated importance of the negatively charged Glu148 in GCaMP2 and GCaMP6m, this residue is often replaced with His, Asp or even nonpolar Ile in other biosensor designs[30]. While knowledge of the mechanism underlying fluorescence switching for existing biosensors is important, the understanding has not yet translated to instructive rules for designing new biosensors.

Quantification of single-FP biosensor signal

The rapid adoption of GCaMP in neuroscience research can be attributed to the unique compatibility between its properties and the signal it measures. GCaMP is excited by a single wavelength and its emission is measured at a single wavelength. The relevant signal is the emission intensity, which is referred to as an intensiometric readout. The intensity of emission is responsive to the Ca^{2+} concentration, but like all intensiometric readouts also critically depends on the concentration of fluorophore in each cell. The biological signal of interest being fast spikes in Ca^{2+} that occur with action potentials is roughly digital. Inference on Ca^{2+} transients thus require only that the relative change in fluorescence intensity within single neurons be reliable recorded, while comparing the intensity between cells is not of interest. As GCaMP has been the prototypical single-FP biosensors, its design has often served as a template although this may not always be appropriate for new analytes.

Fluorescent biosensors applied to cellular metabolism aim to capture an analog signal, the change in analyte concentration. An alternative quantification strategy relies on the dual excitation capacity of GFP. Excitation of the biosensor fluorophore

at two wavelengths can be used to produce a ratio of measured intensity. Importantly, the excitation ratio will not depend on biosensor concentration. Ratiometric biosensors are well suited for comparison of metabolite concentration between cells which is required to investigate aspects of cellular heterogeneity. While ratiometric biosensors offer enhanced utility for metabolism, the specific mechanisms which facilitate ESPT in cpFP-based biosensors remain largely unexplored.

Fluorescence lifetime imaging (FLIM) presents another option for biosensor imaging. Some biosensors exhibit ligand-dependent changes in the average time between photon absorption and emission. Like an excitation ratio, lifetime is independent of fluorophore concentration making it a useful readout for comparing cells with different expression levels. Fluorescence lifetime can be quantified in live cells but requires a specialized microscope setup which has limited its use as a biosensor readout. In contrast to ratiometric readouts, fluorescence lifetime does not depend on instrument settings, making it possible to calibrate biosensor signals across instruments and derive absolute concentrations[31]. Fluorescence lifetime is proportional to quantum yield, indicating that biosensors that function by changes in fluorophore pK_a (notably excitation ratio biosensors, but also GCaMP where the neutral state is non-fluorescent) will not exhibit robust lifetime changes. Understanding the mechanism, whether changes in quantum yield or protonation state, for existing biosensors is an important step to inform the design of new biosensors with optimized readouts.

1.2 Protein engineering

Biosensors represent an interesting and experimentally tractable test case for several broadly important concepts in protein engineering. Specifically, biosensors are allosteric proteins with complex functions dependent on at least two conformational states that exhibit highly epistatic sequence-function landscapes. To solve the problems associated with engineering these proteins, ideas from the broader field of protein science and engineering will be of use.

Semi-rational design

Protein engineering as a discipline seeks to develop an understanding of the sequence-function relationship to inform the design of new proteins with improved properties. Historically the key technique of protein engineering has been the ability to make mutations to the amino acid sequence of a protein and measure the functional impact. Through careful selection of targeted residues and informative substitutions, the physiochemical basis of protein function can be inferred. This methodology can be adapted to engineer proteins by rationally introducing mutations which are predicted to increase function. As demonstrated by the examples regarding GFP and fluorescent biosensors outlined above, this experimental paradigm has proven to be useful for understanding these proteins. In practice however, the iterative improvements in fluorescent biosensor function were not the result of rational design guided by the laws of physics and chemistry. Instead, highly optimized fluorescent biosensors were discovered using a semi-rational approach combining random mutagenesis of important residues with assays for biosensor function. While semi-rational design has been successful in finding improved sequences, this paradigm requires extensive screening incurring

significant time and labor costs and has not necessarily resulted in general insights into biosensor function. Engineering requires predictive models to guide design efforts. The complexity of the protein sequence-function relationship necessitates new experimental methods to elucidate such models. Recent developments in the broader field of protein engineering offer potential solutions to accelerate the design of improved biosensors.

Allostery

The function of single-FP biosensors is to couple ligand-binding with altered fluorescence emission. This phenomenon can be considered an example of allostery in which an effector (ligand-binding) results in a change of activity (fluorescence) at a distal site. Allosteric regulation of protein function is of such importance that the influential French biochemist Jacques Monod famously referred to allostery as the "second secret of life"[32,33]. The dynamic aspects of cellular physiology including regulation of metabolism and gene expression rely on allosteric proteins to sense and respond to stimuli. While allostery has been studied extensively across a wide array of proteins, the mechanism has long been debated. Without a solid understanding of the principles underlying allostery, it will remain difficult to design allosteric regulation into new proteins. Single-FP biosensors present an attractive system for studying allostery due to their simple domain composition and easily detectable output.

Domain arrangement

The initial challenge in constructing a single-FP biosensor is specifying the domain arrangement, which amounts to identifying the site to insert cpGFP into the LBD. A suitable insertion-site will result in the proper folding of both domains

yielding a bright protein still capable of binding the relevant ligand. Ideally, cpFP insertion will also result in ligand dependent changes in fluorescence. In practice however, allosteric coupling is often weak or entirely absent in the initial insertion-variant but can often be generated through mutating the linker residues at the interface of the two domains (refs). Similar findings are present in the broader research on “protein switches”, for examples constructs in which ligand-binding of one domain is coupled to enzymatic activity of a second domain[34]. It is postulated that sites in a protein structure differ in their latent allosteric capacity[35], which would be expressed as differences in the likelihood an inserted domain will produce allosteric coupling of functions. The latent allosteric capacity is likely dictated by structural and dynamic properties but predicting the optimal insertion site to generate a bright high-dynamic biosensor from these properties remains an open problem.

Allosteric coupling and conformational change

A common heuristic for selecting an insertion-site of a FP, assuming the structures are known, is to target regions of the LBD that exhibit conformational changes between the ligand-bound and apo states. This assumption has been thoroughly tested in an experimental effort to generate a single-FP biosensor out of the well-studied maltose binding protein (MBP)[2]. Four different MBP sites with different degrees of conformational change, quantified as the ∆Dihedral between binding states, were chosen for insertion of cpGFP. Hundreds of linker combinations connecting the domains were tested for each site. The site (MBP-317) that exhibits no conformational change upon binding maltose resulted in the lowest mean dynamic range among the tested linker combinations. However, the

site with only moderate conformational change (MBP-165) produced more high-dynamic-range linker variants than the site with the largest ΔDihedral (MBP-175). These results suggest that a minimal degree of ligand induced conformational change is required to generate allosteric coupling. However, the magnitude of conformational change, quantified by ΔDihedral, is not necessarily predictive of the likelihood of allosteric coupling across linkers nor the magnitude of fluorescence change for the top performing linkers. In theory allostery does not require a change in protein conformation and can be generated instead solely through changes in the amplitude and frequency of fluctuations around the mean conformational state[36]. Because of the persistent bias towards exploiting structural factors when designing biosensors, the influence of protein dynamics on biosensor function remains largely unexplored.

Epistasis

The typical protein consists of hundreds of amino acids with only a handful of residues that clearly and directly contribute to function. Amino acids that are distant from the active site can have drastic effects on protein stability and conformational dynamics, making it difficult to predict which residues will be the most informative to mutate. This problem is especially evident for allosterically regulated proteins. Allostery requires that conformational changes induced by ligand-binding be transmitted through the protein to the active site. Many residues lying between the allosteric binding-site and the active site are likely involved in mediating the switch between states, however identifying these residues is difficult even in well-studied proteins. To further complicate matters, the combined effect of two mutations cannot be easily predicted from the effect of each single mutation. For example,

two mutations that increase the stability of a protein when added individually might drastically reduce stability when added in combination. This phenomenon known as epistasis poses a significant challenge to the protein engineer.

Epistasis is evident in linker combinations

The linker amino acids connecting the FP and the LBD are critically important to biosensor function. For semi-rational design, the linkers are often the residues prioritized for saturation mutagenesis when searching for sequences with improved dynamic range. In the case of MBP discussed above, hundreds of linker amino acid sequences were tested at each insertion-site, the majority of which exhibit low dynamic range. However, a few rare combinations exhibit dramatically increased dynamic range. This observed infrequency of high function sequences is a consistent feature present in many other similar screening efforts[37–41], which indicates significant epistasis between linker amino acids. In most cases, only the top variants are sequenced due to cost and labor constraints preventing analysis of the epistatic interactions between linkers. In addition to the linkers, insertion of cpGFP into a ligand-binding domain creates a new interface between the two domains. These interface residues are also critical for biosensor function but are relatively difficult to identify without a solved structure. Furthermore, epistatic interactions between linker and interface residues are probable and need to be accounted for to identify optimized full-length biosensor sequences.

1.2.1 Massively parallel assays

To generate high function biosensors, it is clearly important to test combinations of mutations. However, there is a fundamental mismatch between the throughput of conventional protein assays and the scale of mutational space. For a relatively

small protein consisting of 100 amino acids, the number of possible single mutations alone (100 x 19 = 1,900) exceeds the limits of conventional protein assays, while the number of double mutations (100 x 19 x 49 x 19 = 1,768,900) is orders of magnitude too large. The primary factor limiting throughput of conventional protein function assays is the requirement to test protein variants in isolation. For example, linker mutagenesis screens are typically carried out using medium-throughput assays with 96- or 386-well plates and each well containing a single sequence variant. Isolating and characterizing single variants is resource and labor intensive, which limits the scale of such assays to hundreds of variants. Furthermore, determining the sequence of each tested variant is prohibitively expensive for most laboratories.

Deep Mutational Scanning experimental design

While traditional biochemical assays strictly require physical separation of variants, a new experimental paradigm has emerged that enables protein functional assays to be performed on pooled sequences. These approaches collectively referred to as deep mutational scanning (DMS) or massively parallel assays (MPAs) rely on advances in reading and writing DNA to recast the measurement of protein function into a sequencing problem[42]. These methods operate by expressing many sequences in a population of cells, typically with 1 sequence variant per cell, and then applying a selection such that the relative abundance of the encoding DNA sequence is determined by protein function. In early examples, the selection step consisted of resisting the action of an antibiotic, hence the terminology. However, selection can take other forms such as physical separation of cells based on GFP brightness using a fluorescence activated cell sorter (FACS)[43,44]. Following

selection, the frequency of variants in the population of cells is determined by high-throughput sequencing. Protein function can then be inferred from the sequence read counts using a variety of computational methods[45,46]. These methods can be scaled to characterize hundreds of thousands of sequences with the limits determined by the number of sequences that can be exposed to selection and the number of sequencing reads.

Library design for DMS

This experimental paradigm critically depends on the ability to create DNA libraries with diverse mutations either randomly or programmatically added to a parent sequence. Random mutations are easily achieved through error-prone PCR which relies on polymerases with inherently low fidelity or the use of additives that increase the replication error rate. This approach yields single nucleotide changes per codon generally resulting in conservative mutations and producing a significant fraction of synonymous changes. Despite these limitations, the truly random nature of this approach is useful for discovering which residues are important, especially when the structure of the biosensor is unknown. For semi-rational design, residues can be targeted for mutation using methods such as degenerate primer PCR or oligo pool synthesis. Degenerate primer PCR is a simple and affordable method applicable when mutations are limited to a small region of the protein, for example when specifically targeting linker amino acids. For mutations distributed throughout the entire gene, more complicated methods such as oligo pool synthesis are required. While being more cost and labor intensive, these methods can produce highly specific libraries with defined mutational rates and amino acid substitution frequencies.

Sequencing readout

An additional consideration when designing the library is the mode of sequencing readout. Variants resulting from targeted mutagenesis can be easily identified by directly sequencing the specific location. For gene length mutation libraries, identifying specific variants becomes difficult when the encoding DNA sequence exceeds the sequencing read length. To overcome this limitation, a random DNA barcode can be appended to the library providing a simpler sequencing readout. However, an additional step is then required to generate long reads that span the entire gene length and barcode. This can be accomplished using less common PacBio or Nanopore instruments.

Single variant per cell expression

DMS experiments are generally performed in live cells which maintain a linkage between DNA sequence and protein function, a prerequisite for this type of assay. Alternate approaches have been developed using microfluidics and in vitro transcription/translation systems to isolate DNA-protein pairs. While in vitro methods offer an advantage of controllability, they lose the unique chemical aspects of the intracellular environment which can be crucial for protein function. Numerous examples can be found in the literature where promising biosensor constructs exhibit differences in performance between different cellular contexts, such as a GCaMP2 variant that was bright and highly responsive in *E. coli* lysate but was much dimmer and less responsive to Ca^{2+} when tested in HEK293 lysate[47]. Additionally, it is desirable to introduce only a single variant per cell, which is easily achieved in bacteria or yeast, but becomes more challenging in mammalian cell lines which often represent a more suitable expression context. One solution to this

challenge is the use of engineered cell lines which contain a single unique recombination site placed within the genome. So called “landing pad” cells provide a relevant intracellular environment while enabling targeted recombination of a single variant per cell suitable for high-throughput assays[48].

Massively Parallel Biosensor Assays

The first attempt at applying the principles of DMS to engineering single-FP biosensors came in 2016[49]. In contrast to most DMS efforts, the effect of amino acid mutations was not the focus. Rather, the goal was to evaluate the influence of domain arrangement on biosensor function. A novel transposon-mediated domain-insertion cloning strategy was used to systematically generate cpGFP insertions throughout the amino acid sequence of *E. coli* MBP or the related *Thermococcus litoralis* D-trehalose/D-maltose-binding protein (TMBP). The library was transformed into *E. coli* and evaluated for either brightness or dynamic range. Selecting for variants based on brightness is straightforward using a FACS instrument and sorting for cells above a threshold brightness; brighter variants will be enriched, and dimmer variants will be depleted. The recovered plasmid DNA was then randomly fragmented and sequenced. Reads spanning the junction between MBP/TMBP and cpGFP were used to identify and count unique insertion-variants. The read counts obtained from the library before and after the threshold sort can then be used to infer the relative brightness of each variant.

Devising a selection scheme for ligand-dependent fluorescence changes is more difficult. To enrich for variants based on dynamic range, the MBP-cpGFP domain-insertion library was subjected to three sequential rounds of FACS. First, cells were sorted for brightness above a threshold in the presence of maltose,

enriching for bright cells in the ligand-bound state. Next, the recovered cells were sorted for fluorescence below a threshold in the absence of maltose, enriching for a dim apo state. Finally, the recovered cells were again sorted for brightness in the presence of maltose. This protocol proved successful in generating an output metric, enrichment, that correlated with ΔF/F across many tested variants. Several high-dynamic range variants were discovered across both MBP and TMBP, but the functional insertion-sites show little overlap between the two related structures. It is likely that aspects of protein dynamics not evident in the static structures underlie these observed differences in the sequence-structure-function relationship.

The systematic generation of domain-insertion variants and characterization of their potential for allostery will be crucial for future efforts to understand this complex relationship. The cloning scheme presented in Nadler et al. is an important advance towards this goal. This method has found use in diverse protein engineering projects including the development of optogenetic ion channels[50] and allosterically regulated Cas9[51]. However, the method for characterizing biosensor libraries has not been replicated outside the initial publication to date. While the logic is sound, guidelines for setting the thresholds for optimal enrichment are not clear. The threshold selected at each step will determine the enrichment of each variant based on brightness by design. However, differences in brightness between variants (up to 100-fold in tested libraries) is much larger than the ligand-induced changes for all but the best biosensors. A more general approach to characterizing biosensors using a high-throughput assay is presented in Chapter 2.

1.2.2 Machine Learning

Highly functional biosensors are rare in the landscape of possible sequences and measuring dynamic range for new variants remains a major bottleneck. In addition to increasing the screening throughput and discovering improved protein variants, massively parallel assays generate datasets which broadly sample the sequence-function landscape. Ideally, this data can then be used to derive a model informed by all tested variants that accurately predicts the function of untested sequences. Several statistical learning approaches, referred to collectively as machine learning (ML) algorithms, aim to automatically learn the mapping between input and output directly from the data with limited assumptions. ML algorithms have proven to be extremely flexible, offering state-of-the-art prediction performance on a wide variety of computational tasks. The most sensational example being AlphaFold, a deep neural network, and its tremendous success on the problem of predicting protein structure from sequence[52]. AlphaFold fundamentally relied on the availability of massive amounts of relevant data, specifically the sequences for millions of known proteins along with thousands of solved protein structures. In general, the performance of machine learning algorithms is dictated by the amount of available training data, of which there is relatively little thus far for fluorescent biosensors. Although deep mutational scanning might produce less accurate measurements than traditional biochemical assays, the increased scale of the data is more important. Pairing DMS with ML, to generate and learn from protein sequence-function data represents a synergistic approach to accomplish data-driven protein engineering.

DMS and ML applied to GFP

Understanding the relationship between GFP structure and function is a crucial aspect of the larger challenge of engineering SFPBs. One of the most impressive DMS assays to date involved fluorescent intensity measurements for 51,715 randomly mutated GFP variants[43]. Fluorescence as a protein function is highly amenable to massively parallel assay owing to the ease of separating cells using a fluorescence activated cell sorter (FACS). Instead of relying on a simple enrichment metric, FACS was used to sort cells into 8 bins spanning the logarithmic scale of brightness. The read counts for each variant in each bin were then used to estimate the per variant log-normal mean fluorescence intensity. The average variant tested contained 3.7 mutations which offers an exciting possibility to explore the combined effects of mutations on GFP function. Most mutations (75%) were deleterious, while up to 30% of the tested variants exhibited significant negative epistatic interactions where the combined effect of slightly deleterious mutations resulted in total non-fluorescence. These results the inherent difficult in engineering fluorescence, let alone dynamically regulated fluorescence. This work demonstrates the importance of generating useful data, as the GFP fitness landscape has become a key resource underlying multiple advances in applying ML to protein engineering[53–57]. Furthermore, the sort-seq assay used to generate this data was adapted for sequencing domain-insertion and linker libraries and applied to characterizing biosensor dynamic range in chapters 2 and 3.

1.3 Metabolism

Metabolism is the collection of chemical reactions that enable cells to produce energy, synthesize and degrade macromolecules, and excrete waste. Metabolic pathways consist of a series of enzyme catalyzed reactions where each product serves as the substrate for the next step in the pathway. The term pathway has connotations of a linear process; however, metabolism is better thought of as a network with certain metabolites acting as connection points between different pathways. Certain metabolites are required across metabolism to facilitate specific types of reactions. For example, adenosine triphosphate (ATP) is the universal phosphate carrier used to transfer chemical energy between reactions. Similarly, NAD^+ is a universal electron carrier that mediates many redox reactions through interconversion with reduced NADH. The ratios of the important cofactors, ATP:ADP and NADH:NAD^+, are broadly informative signals for the overall state of cellular metabolism. Understanding how metabolic pathways for various macromolecules are intertwined with these cofactor pools is of fundamental interest to metabolism research.

Glycolysis

Sugars are an important energetic substrate consumed in the typical mammalian diet. Larger polysaccharides, such as starch, are broken down into monomers. Glucose is the most abundant monosaccharide and serves as the major circulating carbon substrate in mammals. The metabolism of glucose to 2 molecules of pyruvate occurs via 10 enzymatic steps known as glycolysis. This process also yields 2 molecules each of ATP and NADH. Pyruvate has two fates. First, it can be converted to lactate in a reaction referred to as fermentation. The conversion of

pyruvate to lactate by the enzyme lactate dehydrogenase also converts NADH back to NAD^+. The full pathway from glucose to lactate thus results in no net change to the NADH:NAD^+ ratio. Alternatively, pyruvate can be further metabolized in the mitochondria in a process referred to as respiration, whereby pyruvate is imported into the mitochondria where the tricarboxylic acid (TCA) cycle and electron transport chain (ETC) can yield an additional 25-34 molecules of ATP. The ETC requires oxygen as the final electron acceptor, hence the name cellular respiration. Whether the consumption of glucose ultimately results in respiration or fermentation has important implications for ATP yield, redox balance, and the supply of biosynthetic precursors.

Lactate fermentation

It has traditionally been assumed that when oxygen is available, respiration is favored due to the higher ATP yield, and thus lactate production should only occur under anaerobic conditions. Under this view, lactate can be considered a metabolic waste product. In contradiction, it has been observed that a substantial fraction of consumed glucose results in lactate production even in the presence of oxygen. Most prominently, cancer cells ferment glucose to lactate even in abundant oxygen, referred to as the Warburg effect. While this phenotype has pathological connotations, this association is not necessarily warranted given that excessive lactate is produced by most cultured, and non-cancerous, mammalian cells. Given that lactate is not excreted from whole organisms, and most carbon ends up leaving the body as CO_2, it must be assumed that lactate is utilized as fuel[58]. Lactate exists at high concentrations in the blood, can readily enter cells expressing monocarboxylate transporters (MCTs), and the action of LDH is reversible, resulting

in the generation of pyruvate from lactate. Together, these facts suggest that lactate serves as a circulating TCA cycle substrate[58]. In this view, cells can use either circulating glucose to run glycolysis or lactate to fuel the TCA cycle, with the two pathways operating independently. This new view of lactate as an alternative metabolic substrate uncoupled from glucose raises the question of when and where does glycolysis occur.

Glycolytic flux

Glycolysis produces 10 intermediate metabolites from glucose. The individual molecules of each intermediate exist only transiently due to the rapid action of enzymes. As a pool however, the concentration of an intracellular metabolite can be held constant over time when the rate of production is equal to consumption. In this way, the concentrations of glycolytic intermediates are not necessarily informative for evaluating the importance of glycolysis within a given cell type. Instead, the activity of an entire metabolic pathway can be quantified by the parameter flux. Glycolytic flux specifically refers to the rate that glucose-derived carbon atoms flow through the pathway. An increase in glycolytic flux might occur in response to an increase in energy demand or to supply biosynthetic precursors during cellular growth. Regulation of flux is thus of central importance to cellular physiology.

Allosteric regulation of glycolysis

Flux is regulated by the activity of the enzymes involved in the pathway. For multiple glycolytic enzymes, activity is not a static feature but is instead allosterically regulated. For example, phosphofructokinase (PFK) which catalyzes the conversion of fructose 6-phosphate (F6P) to fructose 1,6-bisphosphate (FBP) is

allosterically regulated by multiple factors. PFK1 activity is inhibited by high levels of ATP and activated by AMP such that cellular energy charge partially determines the rate. Downstream metabolites phosphoenolpyruvate (PEP) and citrate, a TCA intermediate, both act as feedback inhibitors of PFK1. Finally, the most potent activator is fructose 2,6-bisphosphate (F-2,6-BP), which like FBP is produced from F6P. These allosteric regulators in combination dictate the rate of FBP production. The lower steps of glycolysis are then regulated both by the supply of FBP, as well as feedforward activation of certain pyruvate kinase isoforms (PKM2 in mammals).

Flux control

While PFK is considered an especially important enzymatic step, its activity is not the sole regulator of glycolytic flux. Instead, multiple enzymes exert partial flux control. Systematic quantification of the extent of flux control by the relevant enzymes and transporters has revealed that the rate of glycolysis is controlled at 4 key steps[59]. The two key enzymatic steps are the phosphorylation of glucose by hexokinase and the phosphorylation of F6P by PFK. Furthermore, two key transport steps, the uptake of glucose into the cell and efflux of lactate out of the cell, exhibit significant flux control. In addition to allosteric regulation of enzyme activity, control over the expression of these enzyme and transporter proteins is critical for regulating glycolytic flux.

Flux-signaling metabolites

Ultimately metabolic flux is of fundamental importance as it constrains the limits of synthetic and energetic aspects of cellular physiology. It is not immediately clear how cells might sense and respond to changes in flux induced by a changing environment or stochastic fluctuations in gene expression. Sensing the availability

of each external substrate and the concentration of each enzyme and transporter individually would pose a significant challenge to the cell. It would be much simpler to sense the concentration of pathway intermediates. While metabolite concentrations do not necessarily change with altered flux, it has been proposed that certain metabolites might exhibit flux dependent changes in concentration[60–62]. So-called flux-signaling metabolites could then serve as an integrated readout of the pathway activity. Allosteric regulation of transcription factor activity would provide a simple means of feeding flux information back onto regulation of gene expression.

FBP is a glycolytic flux signal

FBP has been proposed as a candidate glycolytic flux-signaling metabolite[62]. FBP concentration has been found to correlate with glycolytic flux across varying environmental conditions in a wide range of organisms including bacteria[62], yeast[63], and mammalian cells[59]. Furthermore, in cultured mammalian cells the FBP-flux correlation was maintained across a series of experiments in which the relevant transporters and enzymes were overexpressed, resulting in altered flux[59]. Complementary to this empirical evidence is the existence of FBP-regulated proteins that are poised to take advantage of its flux-signaling capacity. The bacterial Central glycolytic genes Regulator (CggR) is an FBP regulated transcriptional repressor that inhibits the expression of multiple glycolytic genes in the absence of FBP. When bound to FBP, CggR no longer binds DNA and transcription of 5 lower glycolysis enzymes (gapA, pgk, tpi, pgm, eno) is permitted[64]. FBP thus serves as the signal that relates the availability of glucose to the gene expression program thereby enabling its metabolism[60,62]. More recently, similar

functional roles for FBP have been discovered in mammals. For example, the FBP consuming enzyme aldolase, in the absence of its substrate, interacts with the AMP-activated protein kinase (AMPK) in addition to its canonical regulators ATP and AMP[65]. In contrast, when bound to FBP, aldolase activates the mTORC1 complex[66]. In this way, FBP is a signal that mediates the switch between anabolic and catabolic metabolism through the two canonical regulators mTORC1 and AMPK. The accumulating evidence supports that FBP is an informative signal through which glycolytic flux can be inferred.

1.3.1 Technologies for studying metabolism

Metabolomics

Accurate measurement of glycolytic flux can be achieved through analytical chemistry techniques. Specifically, the concentration of intracellular metabolites can be quantified using a combination of chromatography and mass spectrometry to quantify the abundance of many metabolites simultaneously, referred to as metabolomics. To infer flux concentrations, specific labelled metabolites are required which can be distinguished by their heavier mass[67]. Glucose labelled with heavy isotopes will be metabolized and the labelled tracers will be accumulated in intermediate metabolites at a rate that is dependent on flux. While highly accurate, these methods are inherently destructive, requiring metabolites to be extracted from the cell which limits the temporal resolution. Furthermore, disruption of the cell can result in rapid changes in concentration based on the transient lifetime of most metabolites. Metabolism must also be rapidly and completely quenched to accurately capture the relative concentrations present in the cell prior to extraction[68,69].

Extracellular flux assays

Extracellular flux assays are a common alternative to quantitative metabolomics for assessing glycolytic flux[69]. These methods rely on measuring the rate of lactate production and oxygen consumption in the extracellular environment to infer changes in glycolytic flux and the relative balance of fermentation and respiration, respectively. The common Seahorse XF instrument does not measure lactate directly, but instead monitors the extracellular acidification rate (ECAR) as a proxy[70]. Extracellular flux measurements can be performed continuously on live cells which is a significant advantage relative to metabolomics. They also permit the monitoring of dynamic perturbations. A common Seahorse protocol, the glycolytic stress test, calls for the sequential addition of glucose and the drugs oligomycin and 2-deoxyglucose. The addition of glucose to starved cells and resulting increase in ECAR is used as a measure of glycolytic rate. Addition of the mitochondrial inhibitor oligomycin generally results in a further increase in glycolysis in response to the decreased yield in ATP per glucose. Finally, the addition of 2-deoxyglucose, an inhibitor of glycolysis, should decrease ECAR to a level that indicates acidification derived from sources other than lactate. While extracellular flux assays offer an improvement in temporal resolution, they are still limited to measuring the metabolic properties of cultured or primary cells and cannot resolve metabolic heterogeneity.

Fluorescent biosensors for measuring glycolytic flux

Traditional methods for measuring metabolic flux are inherently limited to assaying populations of cells and cannot resolve metabolic heterogeneity between single cells. Fluorescent biosensors provide a possible solution but are not without

challenges. Biosensors specifically enable the measurement of intracellular metabolites abundance which cannot necessarily be equated with flux. Measurements of intracellular pyruvate or lactate can be used to infer relative changes in flux by specifically perturbing either production or consumption and then evaluating the induced rate of change in concentration. For example, the pyruvate biosensor, PyronicSF targeted to the mitochondria, was used to infer flux by inhibiting the mitochondrial pyruvate carrier and measuring the drug-induced rate of fluorescence decrease[71]. Another solution is provided by the existence of flux-signaling metabolites which imprint flux information onto a metabolite concentration that is amenable to biosensor detection. Flux-signaling metabolites, like FBP for glycolysis, can continuously be used to infer flux changes without the need for specific pharmacological perturbations.

1.3.2 Spatial and temporal regulation of glycolytic metabolism

The unique advantages of fluorescent biosensors enable new questions to be asked about the spatiotemporal regulation of metabolism *in vivo*. Rapid metabolic dynamics occurring within single cells can be measured by fluorescent biosensors owing to the enhanced spatiotemporal resolution compared to traditional methods. Fluorescent biosensors can also be used in complex tissues with targeted expression to uncover cell-type specific metabolic phenotypes. While these features are generally useful for studying various tissues, the brain presents a unique confluence of cell-type specific, dynamic, and spatially heterogenous aspects of glycolytic metabolism for which biosensors are especially suited to interrogate.

Brain metabolism

Brain function is energetically expensive. In humans, the brain consumes approximately 20% of the body's total energy supply[72]. In addition to the magnitude of its energetic demands, the brain is also unique in the high variation of energy requirements over time and space. Neurons are specialized to transmit information through the firing of action potentials. This spiking activity of neurons is supported by distinct morphological compartments including the soma, dendrites, axon, and synaptic terminal, each of which exhibits unique patterns of energetic demand. Furthermore, neuronal activity is highly variable over time, generating moment to moment changes in energy consumption.

In contrast to most tissues, the brain appears to obligately consume circulating glucose rather than lactate[73]. Within the brain metabolized glucose is nearly fully oxidized to CO_2 with no net lactate export[74]. While oxidative phosphorylation ultimately produces most brain-wide energy production, this does not necessarily indicate glucose always directly fuels mitochondrial metabolism. Periods of high energy demand in the brain are thought to cause a switch between oxidative phosphorylation and lactate production. Specifically, prolonged neuronal activation has been shown to cause a drastic increase in glucose consumption along with a relatively minor increase in oxygen utilization across the brain[75]. These findings in conjunction with an observed increase in local lactate concentration suggest a decrease in oxidative phosphorylation relative to glycolysis[76,77]. The source of the increase in lactate production relative to consumption remains controversial, with competing lines of evidence suggesting either neurons or astrocytes as the responsible cell type.

Glycolysis in neural cell-types

The temporary redirection of glycolytic metabolism to lactate production has been suggested as evidence for coordinated metabolism between cell types. The astrocyte-neuron lactate shuttle (ANLS) hypothesis posits that glycolysis primarily occurs in astrocytes which then export lactate. The astrocyte-derived lactate is then used by neurons as a high-energy TCA substrate such that neurons do not need to metabolize glucose to maintain high ATP levels by oxidative phosphorylation[78]. Despite significant research, this hypothesis remains controversial, with little direct evidence for the importance of astrocytic glucose consumption to fueling neuronal activity.

An alternative hypothesis suggests that activity induced lactate production results from neuronal glycolysis directly[79–81]. Neurons express the glycolytic enzymes[82,83] and readily consume glucose in culture[84]. Elevated lactate production has been hypothesized to reflect a switch to a fast albeit inefficient resupply of ATP as an alternative to oxidative phosphorylation[85]. Localized anaerobic glycolysis could then function to dynamically match energetic needs at regions of high consumption. In support of this explanation, at the presynaptic terminal in *C. elegans* it has been discovered that glycolytic enzymes can dynamically cluster during periods of high activity[86]. The formation of these so-called glycolytic metabolons implies an increase in local glycolytic flux and lactate production; however, this function has not yet been directly demonstrated.

Application of fluorescent biosensors to neuronal glycolysis

Biosensors for monitoring metabolic properties have provided some key insights into the dynamic regulation of neuronal metabolism. Specifically, the

NADH:NAD^+ ratio biosensor, Peredox, measured in the cytosol of neurons signals a transient increase in NADH in response to neuronal activation[80]. This accumulation of NADH is presumably due to increased glycolytic flux and production of NADH at the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) step. This response was observed in brain slice and *in vivo* and the magnitude of transient change was found to correlate with coexpressed Ca^{2+} biosensor response. The NADH to NAD^+ ratio is also influenced by the ratio of pyruvate to lactate due to the action of LDH which complicates interpretation. To rule out an impact of lactate uptake, an LDH inhibitor was applied and found to further increase the NADH transient resulting from neuronal activity while decreasing the NADH increase from application of extracellular lactate. Furthermore, blocking the plasma membrane transporter for lactate did not diminish the magnitude of the NADH transient. While the creative use of pharmacology to disentangle the interactions between lactate, glycolysis and NADH, produced substantial support for the hypothesis that neurons are the site of stimulated glycolysis, monitoring of FBP could provide a valuable complementary measure. FBP concentration should not be substantially impacted by extracellular lactate uptake but should correlate with increased NADH. Motivated by the potential use for interrogating neuronal glycolysis, we developed a fluorescent biosensor for FBP, detailed in Chapter 3.

Chapter 2. A Sort-Seq Approach to the Development of Single Fluorescent Protein Biosensors

John N. Koberstein[1], Melissa L. Stewart[1], Taylor L. Mighell[2], Chadwick B. Smith[1] & Michael S. Cohen[3]

[1]Vollum Institute, Oregon Health & Science University, Portland, OR 97239, USA.

[2]Department of Molecular and Medical Genetics, OHSU, Portland, OR 97239, USA

[3]Department of Physiology and Pharmacology, OHSU, Portland, OR 97239, USA.

Published in ACS Chemical Biology, August 2021

doi: 10.1021/ACSCHEMBIO.1C00423.

2.1 Abstract

Motivated by the growing importance of single fluorescent protein biosensors (SFPBs) in biological research and the difficulty in rationally engineering these tools, we sought to increase the rate at which SFPB designs can be optimized. SFPBs generally consist of three components: a circularly permuted fluorescent protein, a ligand-binding domain, and linkers connecting the two domains. In the absence of predictive methods for biosensor engineering, most designs combining these three components will fail to produce allosteric coupling between ligand binding and fluorescence emission. While methods to construct diverse libraries with variation in the site of GFP insertion and linker sequences have been developed, the remaining bottleneck is the ability to test these libraries for functional biosensors. We address this challenge by applying a massively parallel assay termed “sort-seq” which combines binned fluorescence-activated cell sorting, next-generation sequencing, and maximum likelihood estimation to quantify the brightness and dynamic range for many biosensor variants in parallel. We applied this method to two common biosensor optimization tasks: choice of insertion site and optimization of linker sequences. The sort-seq assay applied to a maltose-binding protein domain-insertion library not only identified previously described high-dynamic-range variants but also discovered new functional insertion-sites with diverse properties. A sort-seq assay performed on a pyruvate biosensor linker library expressed in mammalian cell culture identified linker variants with substantially improved dynamic range. Machine learning models trained on the resulting data can predict dynamic range from linker sequence. This high-throughput approach will accelerate the design and optimization of SFPBs, expanding the biosensor toolbox.

2.2 Introduction

Genetically encoded single fluorescent protein biosensors (SFPBs) can unmask aspects of cellular signaling and metabolism that cannot be detected using traditional biochemical approaches[87,88]. SPFBs can provide crucial information on the subcellular compartmentalization of analytes, resolve changes in concentration over time, and highlight cellular heterogeneity[88–93]. However, each SFPB specifically measures a single analyte, requiring a unique biosensor for each target. A better understanding of the factors underlying successful designs is necessary to accelerate the development of novel biosensors.

SFPBs can be created by inserting a fluorescent protein (FP) into a ligand-binding domain (LBD) such that ligand binding allosterically regulates fluorescence[2,26,49,89,94,95]. The nature of this allosteric domain coupling is not well understood and has proven difficult to design rationally[49]. In the absence of a predictable relationship between biosensor sequence and function, low-throughput protein-engineering methods are unlikely to discover the optimal sequence because they can explore only a small subset of the possible design space. While the rules for combining domains to produce allostery in biosensors are unclear, the site of insertion and the amino acids connecting the domains are thought to be the two most important parameters[2]. As an alternative to rational design, unbiased library methods have been developed to enable the creation of many domain-insertion and linker sequence variants in parallel[49]. However, the major challenge is efficiently screening these variants.

One potential solution is to apply massively parallel assays that link cellular phenotype with DNA sequencing to characterize sequence-function pairs in a high-throughput fashion. A type of massively parallel assay combining fluorescence-

activated cell sorting (FACS) with sequencing, referred to as sort-seq, is particularly relevant for screening biosensor libraries[96–101]. In the simplest sort-seq experimental design, the library is sorted to collect cells above a threshold intensity. Sequencing then enables determination of the relative abundance of each variant present in the sorted cells compared to the input library, which will correlate with fluorescence intensity[45] . However, in the case of fluorescent biosensors, the goal is not necessarily to improve the brightness, but rather the dynamic range (i.e., the relative change in emission between the ligand-bound and ligand-free (apo) states (ΔF/F)).

A previous study has demonstrated a sort-seq assay relying on three sequential rounds of threshold sorting in the presence and absence of the ligand, that enriches for variants based on dynamic range rather than brightness[49]. While functional biosensor variants were correctly identified, this method exhibits certain biases and limitations. For example, only variants that exhibit a direct relationship between ligand concentration and fluorescence intensity (turn-on) will be enriched, disregarding the possibility of functional variants with the opposite response (turn-off). In addition, the output metric, enrichment, only correlates with dynamic range and does not account for differences in brightness necessitating downstream experiments to quantify these variables for each variant. An ideal method would directly measure the brightness in each state as well as the magnitude and direction of fluorescence change for many variants in parallel to better understand the relationships between these aspects of biosensor function and sequence.

An alternate sort-seq experimental design relies on the sorting of cells into multiple bins spanning the range of fluorescence intensity in order to infer the fluorescence distributions of different variants in an unbiased fashion[45]. The inferred

mean intensity in the ligand-bound and ligand-free states can then be used to calculate the dynamic range for each variant. This binned sort-seq approach should be generally useful for identifying turn-on as well as turn-off sensors across the entire range of observed fluorescence intensity. Here we demonstrate the application of binned sort-seq assays to efficiently screen libraries of biosensor variants. Sort-seq characterization of a previously described maltose-binding protein (MBP) domain-insertion library[49] validates the robustness of this method to identify high-dynamic-range variants independent of baseline brightness and direction of fluorescence change upon ligand binding. Optimization of a pyruvate biosensor[71] through linker mutagenesis screened in mammalian cell culture demonstrates the generalizability of the assay to various contexts and library designs. Machine learning models trained on the resulting data enable prediction of dynamic range from linker sequence. Insights into the biochemical features underlying biosensor function can be drawn from models to aid in further optimization efforts. Together these experiments establish the use of sort-seq assays to efficiently develop high-dynamic-range biosensors.

2.3 Results

Sort-Seq assay of an MBP domain-insertion library.

A library of circularly permuted GFP (cpGFP) insertions into MBP [49] was used as a test-case for the sort-seq screening of biosensor libraries in *E. coli*. The transposon-based cloning strategy used to construct this library simplifies a common first step in biosensor design, the identification of a suitable insertion site in the LBD. This method relies on the random insertion of a modified transposon sequence throughout the LBD which is subsequently replaced with the cpGFP

sequence through Golden Gate assembly (Fig. S1A). This library has previously been characterized by an enrichment assay that identified high-dynamic-range variants (MBP-169 and MBP-170) that serve as internal positive controls (Nadler et al., 2016). An initial enrichment sort was performed to select for fluorescent variants above a threshold defined by non-fluorescent cells. This sort removes the non-productive (out-of-frame and reversed insertions generated by the transposon method) as well as in-frame insertions that result in improper folding of the inserted cpGFP. Enrichment scores calculated from read counts of variant sequences in the naive and sorted libraries revealed enrichment of in-frame variants ($p < 0.001$, Mann-Whitney U test, Data S1B), and overall consistency with previous results[49] (Spearman's $\rho = 0.6$, Fig. S1C-D). These results confirm the usefulness of the transposon cloning strategy for generating many domain-insertion variants in parallel and that a single threshold sort can selectively enrich the desired fluorescent variants.

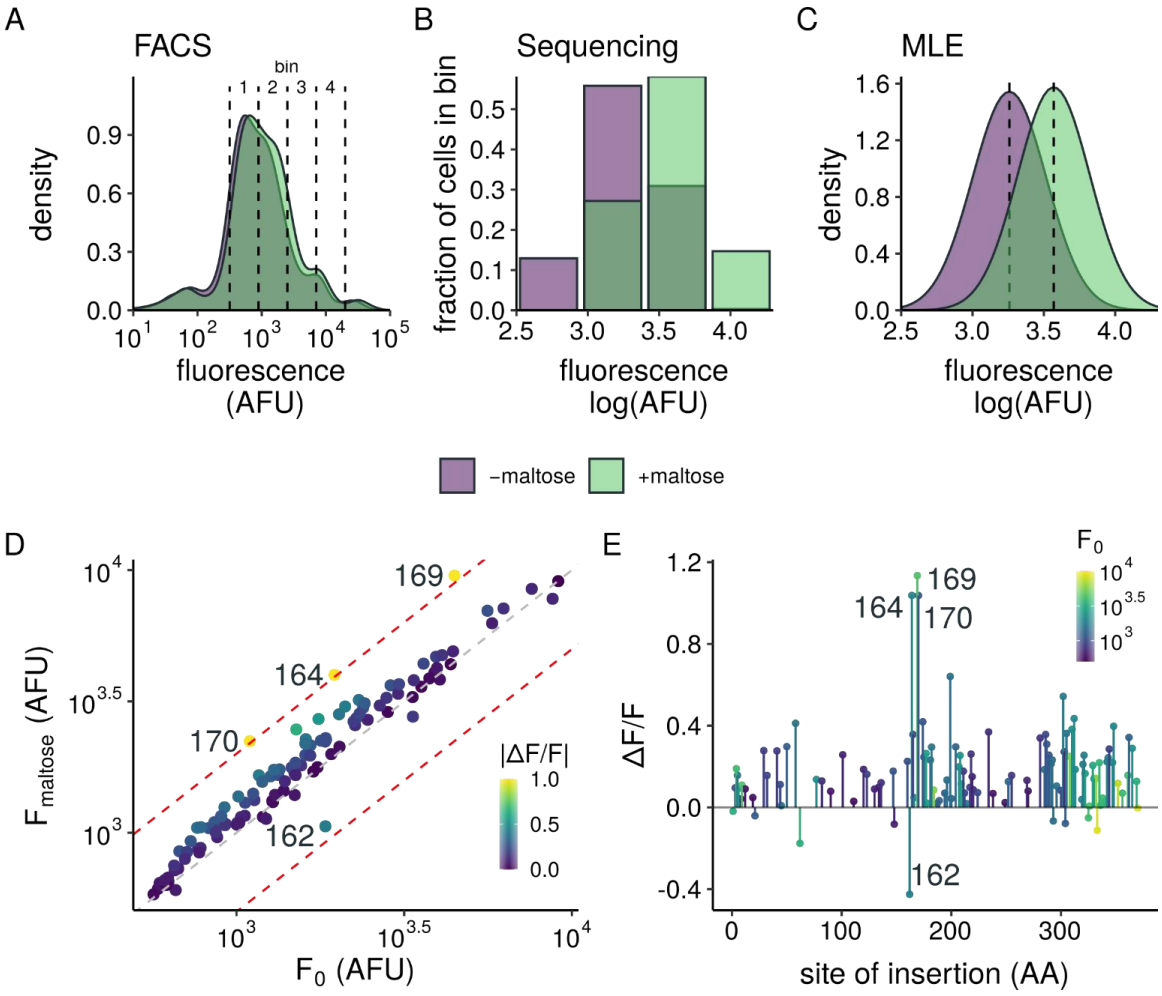

Figure 1. Sort-seq assay of an MBP domain-insertion library. (**A**) Fluorescence distributions for the MBP domain-insertion library ±1mM maltose. Cells were sorted into the 4 bins indicated by dashed lines. (**B**) Sequencing reads are used to estimate the relative fraction of cells containing a given variant sorted into each bin. (**C**) Maximum likelihood estimation is used to infer the log-normal distribution parameters μ and σ represented by dashed line and shaded regions. Dashed lines represent the estimated log-mean (μ) for each condition. Panels B and C represent the sequencing data and MLE distributions for variant MBP-164. (**D**) Estimates of the mean fluorescence intensity for 113 MBP domain-insertion variants in the presence ($F_{maltose}$) and absence (F_0) of 1mM maltose. Colors represent absolute value of ΔF/F. Red and grey dashed lines indicate |ΔF/F|=1.0 and 0.0 respectively. (**E**) Profile of ΔF/F estimates for insertions along the MBP primary sequence. Colors indicate the baseline brightness for each insertion-site.

Estimates of the dynamic range for variants in the MBP library were obtained by binned sort-seq. *E. coli* expressing the enriched library were sorted into bins of equal width on the log-scale spanning the range of fluorescence of the entire library

(Fig. 1A). High-throughput sequencing of the sorted cells generates read counts for each variant in each bin resulting in a course-grained view of the cellular fluorescence distribution analogous to a histogram (Fig. 1B). The parameters describing the input fluorescence distribution, mean and standard deviation, were then inferred from the read counts using a maximum likelihood estimation (MLE) approach assuming a log-normal distribution (Fig. 1C). Cells were sorted in two batches to obtain estimates of the mean fluorescence intensity in the absence (F_0) or presence ($F_{maltose}$) of maltose in order to estimate dynamic range ($\Delta F/F = (F_{maltose} - F_0)/F_0$). The number of bins was limited to 4 in order to minimize downstream sample handling while maintaining a bin width narrower than the width of the fluorescence distribution for a typical variant[45]. In all, after filtering for variants with sufficient abundance and expected variance (Fig. S2A), estimates of brightness and dynamic range for 113 variants out of a possible 370 were obtained (Fig. 1D, Data S2). These data provide a thorough examination of the potential for biosensor function of many insertion sites across the protein.

A majority (93/113 variants) exhibited little to no detectable change in fluorescence (|ΔF/F| < 0.3) illustrating the relative rarity of allosterically coupled variants. Internal positive controls MBP-169 (ΔF/F = 1.13) and MBP-170 (ΔF/F = 1.04) were detected as functional biosensors by sort-seq, exhibiting the largest maltose induced changes. Insertion at amino acid 164 exhibited a response of similar magnitude (ΔF/F = 1.04). In contrast, insertion at MBP-162 produced a biosensor with the opposite response, exhibiting a decrease in fluorescence with the addition of maltose (ΔF/F = -0.43). The proximity of the highest functioning variants in the protein sequence supports the notion that ligand binding is not necessarily allosterically coupled to specific residues but instead broader regions

of the protein (Fig. 1E). MBP-162 and MBP-164 were individually cloned and assayed for function in *E. coli* using a 96-well plate fluorescence reader assay to validate the negative and positive responses to maltose estimated by sort-seq (Fig. S3). These two variants were not detected in the previously published screen likely because of low abundance in the initial library due to intrinsic bias of the transposon reaction, each comprising approximately 1 in 10,000-50,000 sequences (Fig. S4). Such large disparities in variant abundance are common in many massive mutagenesis libraries. The ability to estimate dynamic range of low abundance variants highlights the potential for this assay to expand to even more diverse libraries.

Sort-seq assay of a pyruvate biosensor linker library.

We next asked if the sort-seq assay could be used to optimize the linker regions of an existing SFPB in mammalian cells. Previous studies of various SFPBs show that increases in dynamic range can be achieved by altering the linkers connecting the LBD and circularly permuted fluorescent protein[2,28,49]. We focused on the pyruvate biosensor, PyronicSF, since its dynamic range is easily demonstrated in mammalian cells by the application of exogenous pyruvate which is rapidly transported across the cell membrane. PyronicSF consists of cpGFP inserted into the pyruvate sensitive transcription factor PdhR[71]. The linker and FP sequences of PyronicSF (Leu^{1}Glu^{2}-cpGFP-Thr^{3}Arg^{4}, referred to as L^{1}E^{2}- T^{3}R^{4}) are derived from the well-characterized SFBP GCaMP3[47] and were not optimized for PyronicSF (Fig. 2A). We hypothesized that tailoring the linkers to the insertion context would yield improvements to the dynamic range. A library consisting of 2,304 unique linker combinations was generated by substituting residues in linker 1 (L^{1}E^{2}) and 2 (T^{3}R^{4})

with amino acids from the sets {A, E, G, L, P, Q, R, V} and {A, G, P, R, S, T} respectively (Fig. 2A). This degenerate library encodes the original linker set, which is useful as a positive control, while also sampling from amino acids with diverse biochemical features. The linker library was transfected into the HEK293T Landing Pad cell line[102] and sorted to enrich for recombined cells ensuring genomic integration of a single variant per cell. Recombined cells were sorted into four equal-width bins spanning the range of log fluorescence intensity in the presence or absence of 10mM pyruvate (Fig 2B). After filtering for variants with sufficient abundance and expected variance (Fig. S2B), sort-seq estimates of dynamic range were obtained for 1,023 unique linker combinations (Fig. 2B, Data S3).

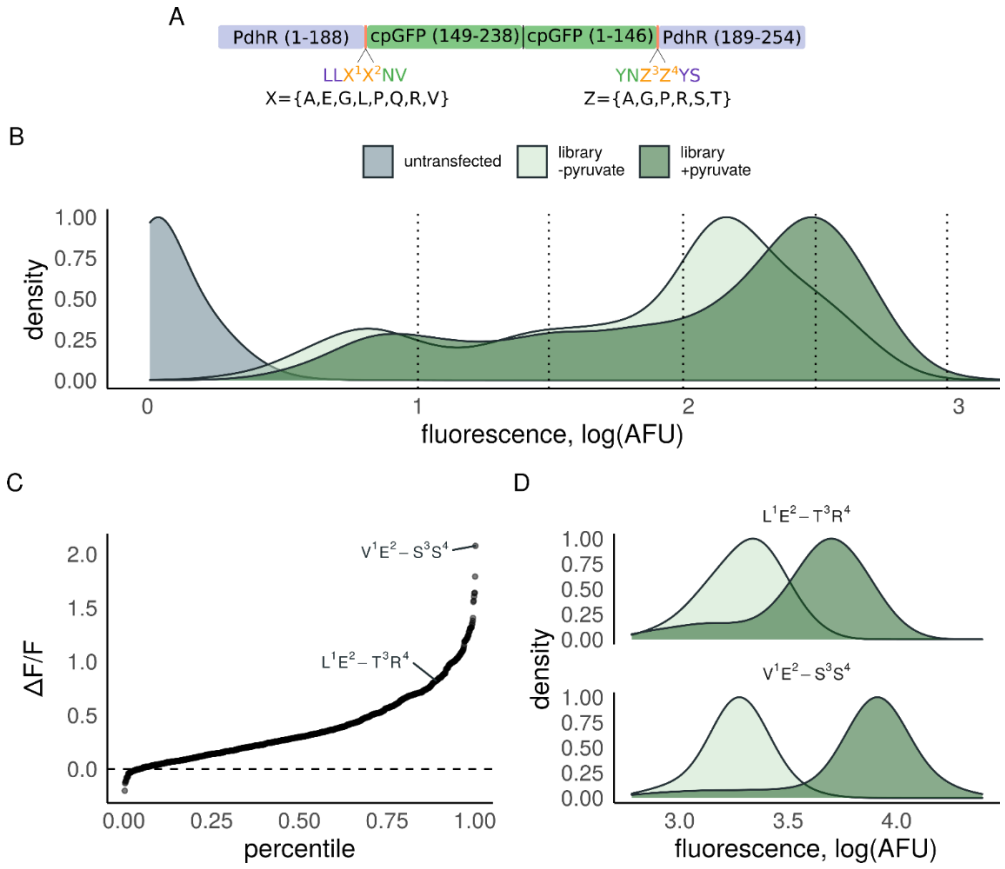

Figure 2. Sort-seq assay of a PyronicSF linker library. (**A**) Schematic depicting the domain organization of PyronicSF (top) and the designed linker library (bottom) (**B**) Fluorescence distributions for the PyronicSF linker library expressed in HEK293T Landing

Pad cells with (dark green) or without (light green) 10mM pyruvate. Cells were sorted into the 4 bins indicated by dashed lines. (**C**) Dynamic range estimates for 1,023 PyronicSF linker variants ordered from lowest to highest ΔF/F. The original linkers (L^1E^2-T^3R^4) and top variant (V^1E^2-S^3S^4) are labelled. (**D**) Flow cytometry fluorescence distributions for HEK293T cells expressing either PyronicSF or the top variant identified from the sort-seq assay with (dark green) or without (light green) 20mM pyruvate. Dynamic-range is increased from ΔF/F = 1.15 for the original linkers to ΔF/F = 2.99 for variant V^1E^2-S^3S^4 validating the estimates from the sort-seq assay.

Similar to the domain insertion library, variation in the linker regions produced changes in brightness spanning over an order of magnitude. The estimated dynamic range was also found to be highly variable with values ranging from ΔF/F = -0.20 to 2.08 (Fig. 2C). The PyronicSF parent sequence, L^1E^2-T^3R^4, served as an internal positive control and was found to have a dynamic range (ΔF/F = 0.82) greater than 88% of tested variants. The variant with the highest estimated dynamic-range, V^1E^2-S^3S^4 (ΔF/F = 2.08) was constructed and tested by flow cytometry alongside the original L^1E^2-T^3R^4 linkers confirming the increase in dynamic-range observed in the sort-seq measurements (Fig. 2D). Most single substitutions to the parent sequence (21/24) resulted in a decrease in dynamic range with the substitution of Leu^1 with Val^1 producing the largest increase in function (Fig. S5). As would be expected given the proximity of the amino acid within a given linker as well as the presumed physical proximity of the two linkers, the effect of multiple substitutions results in epistatic interactions. In the case of V^1E^2-S^3S^4, the measured dynamic range is much greater than the linear sum of the effects of the comprising single substitutions (Fig. S5; ΔF/F = 2.08 compared to ΔF/F = 0.57 assuming additivity). Testing substitution effects at each position in serial, as is common in directed evolution experiments[103], would not arrive at this combination evidencing the utility of testing combinatorial libraries.

In addition to discovering highly functional variants, another benefit of this approach is the opportunity to learn from the numerous suboptimal variants. Machine learning algorithms trained to predict functional activity from protein sequence can assist in elucidating the biochemical determinants of function and predict additional sequences to test[56,104–108]. To this end, the PyronicSF linker sequences were encoded as numerical vectors using the VHSE amino acid descriptor (8 principal components score vectors derived from hydrophobic, steric, and electronic properties)[109]. These biochemical encodings of linker sequences were then used as features to train regression models to predict sort-seq derived ΔF/F values. Random forest[110–112] was found to outperform other common models when evaluated by 5-fold cross-validation using Spearman's rank correlation coefficient (ρ) to assess accuracy (Fig. S6) and was used for all further analysis. The final random forest model was trained on an 80% split of the dataset and prediction performance was tested on the held out 20%. Predictions of test set function correlated with measured dynamic range (ρ = 0.71, Fig. 3, Data S4) suggesting strong predictive power of the model.

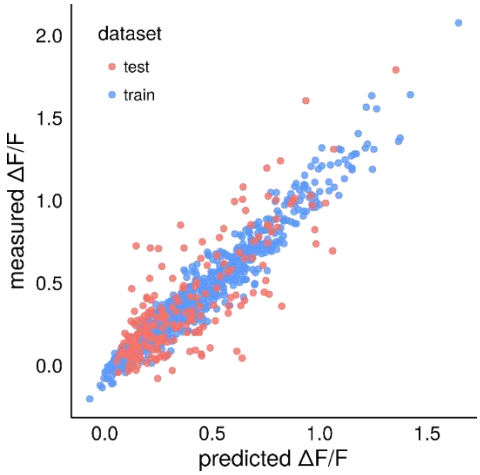

Figure 3. Dynamic range can be predicted from linker sequence. PyronicSF linker variants were numerically encoded using VHSE biochemical descriptors. A random forest

model was trained using an 80% split of the encoded linker data (training set, blue) to predict dynamic range. Model performance was evaluated on the held out 20% (test set, red). A strong correlation ($\rho = 0.71$) was observed between model predictions and measured $\Delta F/F$.

Insights into the sequence determinants of biosensor function can be drawn from interpreting how useful the various biochemical features are for predicting function. The importance of a given feature can be estimated from the change in model accuracy when predictions are made using a permutation of that feature vector[110]. For example, the most important feature according to this method describes electronic properties (VHSE5) of the second amino acid in the N-terminal linker (E^2 in L^1E^2-T^3R^4 and V^1E^2-S^3S^4; Fig. 4A). Variants containing glutamic acid at this position exhibited the highest mean dynamic range ($\Delta F/F = 0.66 \pm 0.42$; Fig. 4B). Interestingly, glutamic acid at this position is featured in many of the GCaMP sensor designs (GCaMP2, GCaMP3, and GCaMP6M) in which detailed structural[113,114] and biochemical[29] study have found the negatively charged side chain to preferentially stabilize the protonated state of the fluorophore in the ligand-bound conformation which underlies the fluorescence switching mechanism. In contrast, substituting positively charged arginine at this position resulted in the lowest mean dynamic range ($\Delta F/F = 0.20 \pm 0.16$; Fig. 4B). The second most important feature corresponds to backbone properties (VHSE8) of the first amino acid. This feature segregates proline, the most deleterious amino acid at this position ($\Delta F/F = 0.12 \pm 0.12$, Fig. 4C), from the other amino acids due to its unique cyclic structure. In contrast, variants containing the less constrained small hydrophobic amino acids alanine ($\Delta F/F = 0.53 \pm 0.37$, Fig. 4C) and valine ($\Delta F/F = 0.52 \pm 0.39$, Fig. 4C) exhibited the highest mean dynamic range. Despite the presumed importance of glutamic acid at amino acid 2, the N-terminal linker combination with the highest mean dynamic range was A^1A^2 ($\Delta F/F = 1.02 \pm 0.39$,

Fig. 4D) followed by V^1E^2 ($\Delta F/F = 0.95 \pm 0.49$, Fig. 4D). While the features of the C-terminal linker were found to be less important for prediction accuracy (Fig. 4A), this is not to say that these positions do not contribute to biosensor function. Rather, C-terminal substitution effects can be substantial and seemingly more dependent on interactions with the other linker amino acids (Fig. S7).

Training the model with variable training set sizes reveals sharp decreases in the mean absolute prediction error when using up to 20% of the data and diminishing returns upon further increases (Fig. S8A). This result demonstrates a machine learning model trained on a subset of a complex library is useful for efficiently prioritizing untested sequence variants to evaluate experimentally. To test this hypothesis in silico, the random forest model trained on 20% of the data (241 variants) was used to predict the dynamic range for all 2304 possible combinations encoded by the degenerate bases. A second library that could be encoded by degenerate bases was then designed by selecting amino acids at each position that appeared in more than 10% of the top 100 predicted variants. Comparing the distribution of dynamic range across all variants in the original library to the predicted high-dynamic-range subset reveals a shift towards increased function (Fig. S8B). By combining sort-seq assays with machine learning guided library design, we can both increase the number of variants tested as well as the likelihood a given variant will exhibit increased function.

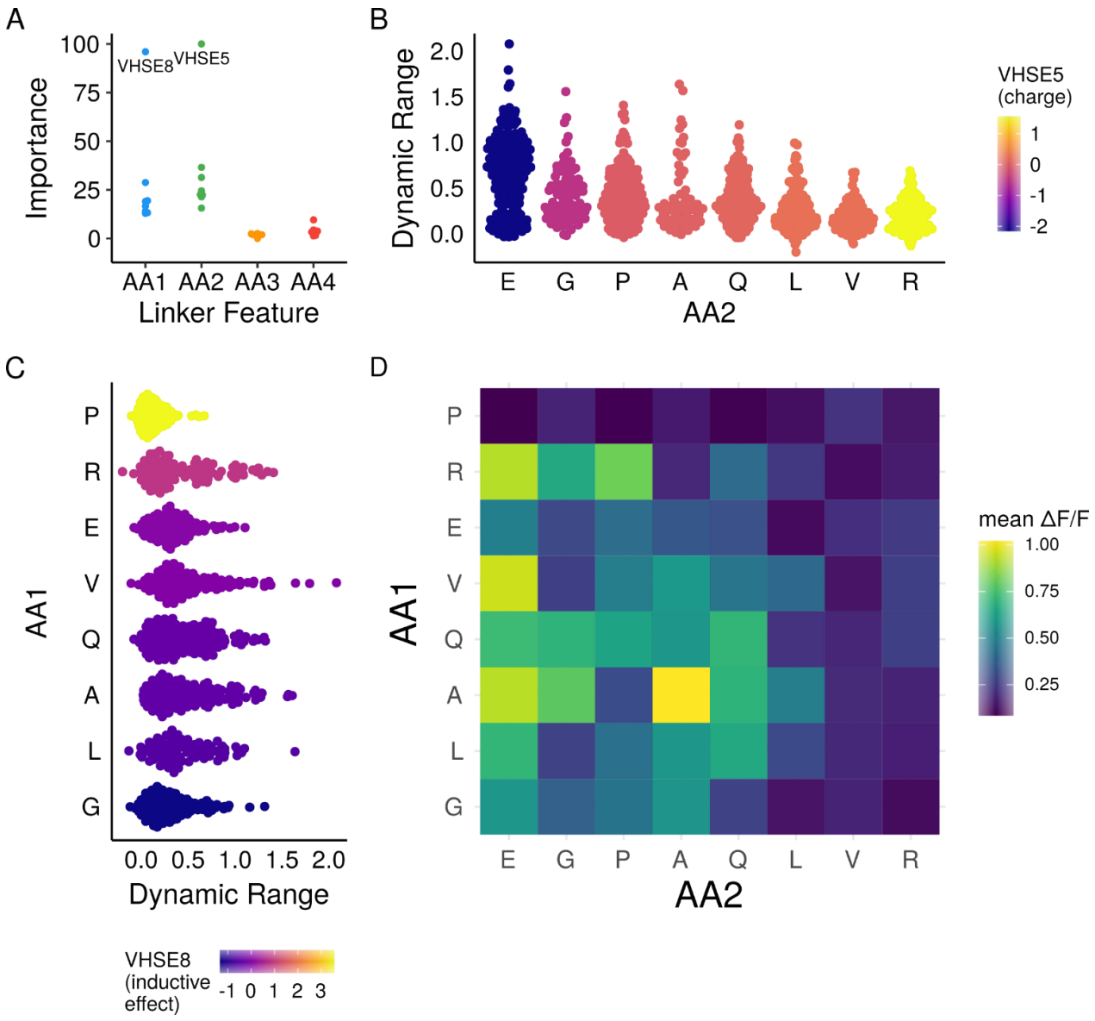

Figure 4. Analysis of the biochemical basis of linker function. (**A**) The importance of each feature can be estimated from the impact on model importance. Two features of the N-terminal linker amino acids, VHSE8 of the first and VHSE5 of the second amino acid, are the most important. (**B**) Distributions of variant dynamic-range estimates sorted by VHSE5 of second N-terminal linker amino acid. Negatively charged glutamic acid exhibits the highest mean ΔF/F, while positively charged arginine exhibits the lowest. (C**)** Distributions of variant dynamic-range estimates sorted by VHSE8 of the first N-terminal linker amino acid. VHSE8 separates proline, which is distinctly deleterious, from the remaining amino acids. (**D**) Mean dynamic-range for N-terminal linker pairs highlights certain combinations that on average exhibit increased function such as A^1A^2 and V^1E^2.

2.4 Discussion

Technologies to construct protein variants in parallel have outpaced the ability to assay these constructs for function. Here we show that binned sort-seq assays enable estimation of biosensor dynamic range by directly quantifying the

fluorescence intensity of ligand-bound and unbound states for many variants in parallel. The application of this method to domain-insertion as well as linker variant libraries presents a high-throughput approach to the development of novel biosensor constructs as well as optimization of existing SFPBs.

Characterization of an MBP domain-insertion library using the sort-seq assay both confirms previous findings and expands the number of high-dynamic-range variants detected. Identifying functional biosensors in a pooled library is made especially difficult given that high-dynamic-range variants can exhibit differences in brightness and direction of fluorescence change. In the case of the MBP domain-insertion library, the three variants with the largest dynamic range (164, 169, 170) differ significantly in ligand-free brightness and another variant (162) exhibits a considerable dynamic range but with an inverse response to maltose. Despite these differences, these variants were all correctly identified as high-dynamic-range biosensors, highlighting the advantages of the binned sort-seq approach for discovering functional biosensors independent of brightness and response direction.

The ability to screen biosensor libraries in both live *E. coli* and HEK293T cells provides a significant advantage in regard to screening conditions. Although protein function can be sensitive to the context of expression, SFPBs are commonly developed using assays of purified protein or expressed in *E. coli*. Differences in protein translation, trafficking, and analyte abundance can lead to reduced performance when changing the expression context. Functional screens performed in mammalian cell culture using the HEK293T Landing Pad line can reduce these differences when developing sensors for in vivo mammalian studies[47]. The assay conditions can be even further refined by converting a given cultured mammalian

cell of interest into a Landing Pad line using a simple lentiviral toolkit[102]. Despite the potential advantages, live cell FACS assays may not be suitable for all biosensor/analyte pairs as it requires that the measured signal can be manipulated and sustained for the duration of the sort. For a ligand like pyruvate which readily enters HEK293 cells[71], the combination of Landing Pad cell lines and sort-seq simplifies the engineering process by screening for biosensors in an appropriate context.

Biosensor engineering is an iterative process. Each round of mutagenesis and screening ideally will result in improvements in biosensor function as well as insights to guide future rounds of optimization. Physiologically relevant differences in concentration, such as the estimated 12μM difference between astrocyte cytosolic and mitochondrial pyruvate concentrations, are relatively small compared to the sensitive range of the biosensor (approximately 0.1mM to 10mM)[71]. The 160% improvement in dynamic range resulting from our assay of PyronicSF linker variants is an important step towards more accurate detection of such small but physiologically important changes in pyruvate concentration. Additional improvements to PyronicSF, most importantly increasing the binding affinity of PyronicSF (K_d=480μM) to better match the intracellular pyruvate concentration (33μM)[71] will be necessary to make this tool generally useful.

In addition to the improved variants identified by this assay, the resulting dataset as a whole is a valuable resource to guide future experiments. In most biosensor engineering experiments, only the most functional variants are sequenced. However, obtaining sequence-function pairs for variants with improved function as well as mutations that decrease function is useful to better understand the sequence-function landscape. In the case of PyronicSF, mutational paths from the

original linker to an improved sequence mostly traveled through intermediates with decreased function. Understanding these epistatic interactions between amino acids is essential for identifying optimized linker combinations. Machine learning algorithms applied to this problem show promise for accurately predicting dynamic range from linker sequence, identifying biochemical features contributing to biosensor function and designing new libraries biased towards high-dynamic-range variants.

In this study, variation was limited to either the site of insertion or the composition of the linkers. For a fixed insertion site such as PyronicSF, some linker variants produce strong coupling while others exhibit no coupling of ligand binding with fluorescence. This suggests that the arbitrary linkers imposed by the transposon cloning method might mask insertion sites that would be functional with different linkers. In addition, analysis of model performance revealed that only a fraction of the collected linker data for this insertion-site was required to generate robust predictions. This result suggests experimental resources could be better spent on a shallow survey of a more complex library than a comprehensive examination of all possible combinations at a few positions. Highly complex libraries containing variable linkers at each insertion site can be constructed using oligo library synthesis methods[115] and should be amenable to characterization by sort-seq. Studying hundreds of linker combinations at many insertion sites will help elucidate the interactions between these features. Furthermore, the increase in sequence diversity generated by combinatorial linker and insertion-site libraries might uncover highly functional variants that would be difficult to discover by testing each parameter separately. In the short term, binned sort-seq assays provide an efficient method for systematically generating and improving SFPBs for scientific

research. After many such experiments, the accumulated data might yield general insights into the principles governing biosensor design.

2.5 Methods

Domain-insertion library cloning.

The DNA coding sequence for amino acids 1-370 of *E. coli* MBP (malE) was obtained as a gBlock (IDT DNA, Table S1). BsaI sites were added to ends by PCR with complementary overhangs for golden gate cloning into pATT-Dest (Addgene plasmid #79770) using primers MBP-BsaI-GG-F and MBP-BsaI-GG-R (the sequences for all primers used for cloning can be found in Table S1). Mu-BsaI transposon was digested from pUC-KanR-Mu-BsaI (Addgene plasmid # 79769) with BglII and HindIII in Buffer 3.1 (NEB) at 37°C overnight and purified by gel extraction (NucleoSpin Gel and PCR Clean-up). Transposition was performed using 100 ng of purified MuA-BsaI transposon, pATT-MBP plasmid DNA at a 1:2 molar ratio relative to transposon, 4 µl of 5x MuA reaction buffer, and 1 µl of 0.22 µg/µl MuA transposase (Thermo Fisher) in a total volume of 20µL. Reactions were incubated at 30˚C for 18 hours, followed by 75˚C for 10 minutes. Reactions were cleaned up using the DNA Clean & Concentrator-5 Kit (Zymo Research Corp.) and eluted in 6µL of water. Transformation was performed using 2µL of reaction in 25µl of E. Cloni 10G ELITE cells (Lucigen) in 1.0-mm Bio-Rad cuvettes using a Gene Pulser Xcell Electroporation System (settings: 10 µF, 600 Ω, 1.8 kV). Cells were immediately resuspended in 975µL Recovery Media and shaken at 250rpm for 1 hour at 37°C. A 10µL aliquot of transformed cells were plated on carbenicillin (100 µg/ml) and chloramphenicol (25 µg/ml) to select for the presence of pATT plasmid backbone and transposon insertion to assess library coverage. The remaining

transformed cells were pelleted and resuspended in 50mL LB with 100 µg/ml carbenicillin and 25 µg/ml chloramphenicol. Cultures were grown at 250rpm, 37°C overnight followed by plasmid DNA purification using a HiSpeed Plasmid Midi Kit (Qiagen).

Transposed pATT-MBP plasmid was digested with Esp3I and the band corresponding to MBP plus the transposon sequence was isolated by gel extraction (NucleoSpin Gel and PCR Clean-up). The isolated transposed ORF was cloned into the expression vector pTKEI-Dest (Addgene Plasmid #79784) by golden gate cloning using 40 fmoles pTKEI-Dest, 40 fmoles of purified MBP-transposon, 10 units Esp3I (NEB), 800 units T4 DNA ligase (NEB) and 1X T4 DNA Ligase Reaction Buffer (NEB) in a total volume of 20µL. The reaction was incubated 2 minutes at 37°C, 5 minutes at 16 °C for 50 cycles followed by 20 minutes at 60°C and 20 minutes at 80 °C. Reactions were purified using a DNA Clean & Concentrator-5 Kit and eluted with 6 µl water. Transformation was performed using 2µL of reaction in 25µl of 10G ELITE *E. coli* (Lucigen) in 1.0-mm Biorad cuvettes using a Gene Pulser Xcell Electroporation System (settings: 10 µF, 600 Ω, 1.8 kV). Cells were immediately resuspended in 975µL Recovery Media and shaken at 250rpm for 1 hour at 37°C. A 10µL aliquot of transformed cells were plated on LB agar plates containing 50 µg/ml kanamycin and 25 µg/ml chloramphenicol to select for pTKEI-Dest backbone and transposon insertions in order to assess library coverage. The remaining transformed cells were pelleted and resuspended in 6mL LB with 1% glucose, 50 µg/ml kanamycin, and 25 µg/ml chloramphenicol. Cultures were grown at 250rpm, 37°C overnight followed by plasmid DNA purification using a QIAprep Spin Miniprep Kit (Qiagen).

BsaI sites and compatible overhangs were added by PCR amplification of cpGFP from pTKEI-Mal-B2 (Addgene Plasmid #79756) using primers cpGFP-BsaI-GG-F and cpGFP-BsaI-GG-R. Inserted transposons were replaced with cpGFP by BsaI-mediated Golden Gate cloning using 40 fmoles of pTKEI-MBP-transposon, 40 fmoles of purified cpGFP, 10 units Esp3I (NEB), 800 units T4 DNA ligase (NEB) and 1X T4 DNA Ligase Reaction Buffer (NEB) in a total volume of 20µL. The reaction was incubated 2 minutes at 37°C, 5 minutes at 16 °C for 50 cycles followed by 20 minutes at 60 °C and 20 minutes at 80 °C. Reactions were purified using a DNA Clean & Concentrator-5 Kit and eluted with 6 µl water. Transformation was performed using 1µL of reaction in 40µl of Tuner DE3 cells (Novagen) by heat shock at 42°C for 30s. Cells were immediately resuspended in 975µL Recovery Media and shaken at 250rpm for 1 hour at 37°C. A 10µL aliquot of transformed cells was plated on LB agar plates containing 50 µg/ml kanamycin and 25 µg/ml chloramphenicol to select for pTKEI-Dest backbone and transposon insertions to assess remaining transposon. The remaining culture was pelleted and resuspended in 6mL LB with 1% glucose and 25µg/mL kanamycin and grown overnight. The following day, FACS samples were inoculated from this culture and a glycerol stock was prepared for storage. Plasmid DNA was extracted from the remaining culture using a QIAprep Spin Miniprep Kit.

MBP domain-insertion library FACS.

Approximately 7 hours prior to sorting, a 100µL aliquot of the library culture to be sorted as well as FACS controls (pTKEI-MBP and pTKEI-cpGFP) were added to 5mL of MOPS EZ Rich Defined Medium (Teknova) containing 60 µg/mL kanamycin and 0.4% glycerol. Cultures were shaken at 250rpm, 37°C for

approximately 2 hours until the OD600 was between 0.6 and 0.8. Expression was induced by adding 0.5mM IPTG followed by incubation for 4 hours. Samples were transported on ice to the core facility for sorting and incubated an additional 30 minutes at 37 °C, 250 rpm. Immediately before starting sort, the samples were diluted 1:100 in phosphate-buffered saline (PBS) and kept on ice.

Sorts were performed using a BD FACSAria III instrument equipped with a 488nm laser for excitation and a 530/30nm emission filter for GFP measurements and 561nm laser with 610/20 filter for mCherry measurements. To minimize the sorting of aggregated cells, *E. coli* expressing either GFP or mCherry were mixed evenly and the sort rate was adjusted until less than 1% of cells were double positive for green and red fluorescence. Cells expressing MBP alone were used to adjust instrument voltages and establish a baseline for cellular autofluorescence. Cells expressing the MBP-cpGFP domain-insertion library were sorted to collect cells above a threshold set at the upper range of autofluorescence. Cells were sorted into a 15mL conical tube containing 5mL LB supplemented with 1% glucose. For the initial enrichment sort, a total of 4.4×10^6 cells were collected. Immediately after sorting, cells were incubated for 1 hour at 37°C, 250rpm. The recovered cultures were diluted to 15mL LB with addition of 1% glucose. An aliquot was diluted 1:50 and spread on LB agar plates containing 50 µg/mL kanamycin to assess cell survival. The remaining culture was incubated with 50µg/mL kanamycin overnight at 30°C, 250rpm. The next day, cultures for FACS were inoculated following the steps above and a glycerol stock was prepared to store the fluorescence enriched library. Plasmid DNA for sequencing was extracted from the remaining culture using a QIAprep Spin Miniprep Kit.

Two samples were prepared for the second round of FACS in order to perform the binned sort with and without the addition of maltose. The two samples were prepared as described above with a 1.5 hour offset to account for time taken to sort the first sample. After transporting the first induced sample to the core, 1mM maltose was added before incubating an additional 30 minutes. Immediately before starting sort, the samples were diluted 1:100 in phosphate- buffered saline (PBS) or 1:100 in PBS with added 1mM maltose and kept on ice. Cells expressing MBP or cpGFP were used to determine the range of fluorescence to be sorted. The lower bound was set at the upper range of autofluorescence indicated by MBP expressing cells. The upper bound was set to include approximately 50% of the cells expressing cpGFP. Four equal width gates on the log scale were set to span this range. The +/- maltose samples were sorted using the same gates for 1.5 hours each. The cells for each bin were collected in 5mL tubes containing 1mL LB with 1% glucose. A range of 4,200 to 444,000 cells were collected for each bin approximately proportional to the relative density of cells in each bin. Cells were recovered as described above and grown overnight. The following day, plasmid DNA for each bin was extracted using a QIAprep Spin Miniprep Kit.

MBP domain-insertion library sequencing.

The ORF to be sequenced was PCR amplified from pTKEI plasmid using primers pTKEI-seqamp-F and pTKEI-seqamp-R (the sequences for all primers used for DNA sequencing can be found in Table S1). 50μL reactions were prepared with a final concentration of 0.2ng/μL plasmid DNA, 0.25μM forward and reverse primer, 1X SeqAmp PCR buffer, 1X SeqAmp Polymerase (Clontech), 1X SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions:

[94°C 60s, (98°C 10s, 55°C 15s, 68°C 60s, plate read) x 29 cycles]. The number of cycles was determined such that reactions were in the exponential phase of amplification upon completion of the program. Reactions were cleaned with AMPure XP beads and eluted in 40μL elution buffer (5 mM Tris/HCl, pH 8.5).

Amplicons were fragmented and tagged (tagmented) in a 25μL reaction containing 1ng/μL amplicon, 1X TD buffer and 0.5μL TDE1 enzyme from the Nextera DNA Sample Prep Kit (Illumina). Tagmentation reactions were cleaned up using a Nucleospin column and eluted in 15μL elution buffer. Tagmented DNA was amplified with primer i5-Nex2p and a unique indexed primer per sample (i7-TXX-NEX2p). 25μL reactions were prepared containing 1μL tagmented DNA, 0.5μM forward and reverse primer, 1X KAPA HiFi Hotstart Readymix (KHF), and 1X SYBR Green. Amplification was monitored by qPCR with cycling conditions: [72°C 3 minutes, 95°C 20 seconds, (98°C 20 seconds, 52°C 15 seconds, 72°C 30 seconds, plate read, 72°C 8 seconds) x 20 cycles]. Reactions were removed during the exponential phase of amplification. PCR products were run on a 1.5% agarose gel to visualize distribution of tagmented DNA size and to estimate relative concentrations using FIJI gel analysis. Indexed samples were pooled normalizing for relative concentration. Pooled products were run on a 1.5% agarose gel, cutting out a band at approximately 500bp which was then purified using the NucleoSpin Gel and PCR Clean-up column. The concentration of the pooled library was quantified using a Qubit fluorometer and size distribution was assessed using a HS DNA chip on the Bioanalyzer 2100 instrument (Agilent). The library was sequenced using 2x75bp paired-end reads on an Illumina MiSeq (v3 Reagent kit).

MBP domain-insertion library enrichment analysis.

Paired end reads were merged using BBMerge[116]. MBP-cpGFP insertion sites were counted using the dipseq analysis pipeline developed by the Savage lab available at (https://github.com/SavageLab/dipseq)[49]. This python package identifies reads that contain sequences originating from both MBP and cpGFP. Junction reads are then trimmed of the cpGFP sequence plus transposon scar before aligning the remaining sequence to MBP to identify the site of insertion. The output is a file containing counts for each insertion site (including out-of-frame and reverse insertions) in each sample which was used for enrichment and sort-seq analysis.

Read counts deriving from the sorted and naive libraries were converted to fractional counts: $f_i^j = {r_i^j}/{t^j}$ where *r* is the number of read counts for insertion at position *i* in sample *j* and *t* is the total number of reads in sample *j*. Variants with a count of 0 in any sample were filtered out and not used in further analysis. Enrichment scores were calculated as $E_j = \log_2({f_i^{sort}}/{f_i^{naive}})$.

Maltose biosensor plate reader assay.

MBP-162 and MBP-164 were cloned by Gibson assembly and tested for function individually (sequences for biosensor constructs can be found in the Supplementary Text). The backbone was opened up by PCR amplification of pTKEI-MBP using primers MBP-162-GA-F1/R1 and MBP-164-GA-F1/R1. The cpGFP insert was amplified from pTKEI-cpGFP using primer sets MBP-162-GA-F2/R2 and MBP-164-GA-F2/R2. 20µL Gibson assembly reactions were prepared with 50ng backbone, 5:1 molar ratio of insert to backbone and 1X Gibson Assembly

Master Mix (NEB) and incubated at 50°C for 1 hour before transformation into One Shot TOP10 Chemically Competent *E. coli* (Invitrogen). Single colonies were picked from LB agar plates with 50µg/mL kanamycin the next day and Sanger sequenced to confirm correct assembly. Expression plasmids with the correct sequence were transformed into Tuner DE3 cells for testing. Single colonies of each biosensor variant to be tested were picked into a deep 96-well plate with each well containing 200 µL of LB supplemented with 10% glycerol, 1% glucose and 60 µg/ml kanamycin. Six wells were inoculated for each variant to accommodate 3 replicates for 2 conditions (+/- ligand). Plates were incubated overnight at 30°C, 300rpm. The following day, 5µL of culture was used to inoculate 200µl MOPS EZ Rich Defined Medium (Teknova) containing 60 µg/mL kanamycin and 0.4% glycerol in a clear bottom black 96-well plate. The plate was incubated at 37°C, 300rpm in a Molecular Devices i3 plate reader while measuring OD600 and fluorescence (485/5nm excitation and 515/5 nm emission) every 600s for 2.5 hours. Upon OD600 exceeding 0.6 for all wells, expression was induced by adding 0.5mM IPTG. Following 2 hours of incubation to allow for adequate protein expression, either 1mM maltose in PBS or equal volume of PBS alone was added to each well. After adding ligand, wells were incubated and monitored for another 1 hour. Dynamic range was calculated as $\Delta F/F_0=(F_t-F_0)/ F_0$ where F_t is the fluorescence at time t and F_0 is the fluorescence at the start of experiment. The difference between the calculated $\Delta F/F_0$ for each variant with and without maltose was calculated to account for changes in fluorescence due to increased protein expression over time.

PyronicSF linker library cloning.

DNA encoding PyronicSF was obtained from Addgene (Plasmid #124812) in the expression vector pcDNA3.1(-). PyronicSF was cloned into the holding plasmid HC_Kan_RFP-p7 (Addgene Plasmid #100615) by Golden Gate Assembly. PCR was used to amplify the coding region in 4 parts to add BsaI recognition sites and compatible overhangs to ends while also removing internal BsaI/Esp3I sites that would interfere with future cloning steps using primers PyronicSF-GG-comp-F1/2/3/4 and PyronicSF-GG-comp-R1/2/3/4. Assembly was performed in a 10µL reaction containing 13 pmol of each PCR fragment and holding plasmid, 10 units BsaI-HFv2 (NEB), 100 units T4 DNA ligase (NEB), and 1X T4 DNA ligase buffer (NEB). The reaction was incubated at 37°C for 5 minutes, 16°C for 10 minutes for 40 cycles, followed by 16°C for 20 minutes, 60°C for 30 minutes and 75°C for 6 minutes before transformation into One Shot TOP10 Chemically Competent *E. coli*.

Mutations to the PyronicSF linkers were generated using degenerate primer PCR. The 5' linker was mutated using two SNA codons and the 3' linker using two VST codons. The backbone sequence including PdhR was opened by amplification with primers PyronicSF-VSTx2-GA-F1 and PyronicSF-SNAx2-GA-R1. cpGFP was amplified using primers PyronicSF-VSTx2-GA-F2 and PyronicSF-SNAx2-GA-R2. Overlapping homologous sequences were annealed by Gibson assembly. The reaction was cleaned using a DNA Clean & Concentrator-5 Kit (Zymo Research Corp.) and eluted in 6µL of water. The library was transformed by electroporation using 1µL of cleaned reaction in 25µL E. cloni 10G ELITE cells. A 5µL aliquot of recovered transformation was plated on LB agar containing 50µg/mL kanamycin. The remaining recovered culture was diluted into 50mL LB containing 50µg/mL

kanamycin and grown overnight at 37°C, 250rpm. Transformation produced well over 20,000 unique transformants, enough to cover 2,304 variants in library 10X.

A golden gate compatible recombination plasmid, referred to as EMMA-attB-Dest, containing a Bxb1 attB site was cloned using parts from the Extensible Mammalian Modular Assembly (EMMA) Toolkit[117]. The PyronicSF linker library was cloned into EMMA-attB-Dest using an Esp3I mediated Golden Gate reaction. In a 10µL reaction containing 13pmol of the PyronicSF linker library in the holding plasmid, 13pmol of EMMA-attB-Dest, 5 units Esp3I (NEB), 100 units T4 DNA ligase (NEB), and 1X T4 DNA ligase buffer (NEB). The reaction was incubated at 37°C for 5 minutes, 16°C for 10 minutes for 40 cycles, followed by 16°C for 20 minutes, 60°C for 30 minutes and 75°C for 6 minutes. The Golden Gate reaction was cleaned using a DNA Clean & Concentrator-5 Kit, eluted in 6µL of water and transformed by electroporation using 1µL reaction added to 25µL of E. cloni 10G ELITE cells. Recovered cells were diluted into 50mL LB with 100µg/ml carbenicillin and grown overnight at 37°C, 250rpm. Plasmid DNA for transfection was purified using a Qiagen Plasmid Maxi Kit.

HEK293T Landing Pad transfection.

HEK293T Landing Pad (TetBxb1BFP) cells obtained from the Fowler lab[102] were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 25mM glucose and 4mM L-glutamine (Gibco 11965092) supplemented with 10% fetal bovine serum (FBS) and 25mM HEPES. Cultures were maintained with 2µg/ml doxycycline (Sigma-Aldrich) which was removed 1 day prior to transfection. For each transfection, a total of 3µg of plasmid DNA and 6µL FuGENE 6 (Promega) were combined in 300µL Opti-MEM (Gibco). A 50:50 mixture by mass of pCAG-

NLS-HA-Bxb1 (Addgene Plasmid #51271), for Bxb1 recombinase expression, and a recombination plasmid containing the attB sequence were used. The DNA transfection reagent mixture was added to a 6-well plate containing $1x10^6$ freshly seeded cells per well. Cells were incubated with transfection reagents for two days before expanding each well to a 15cm plate. After 1 day of growth in the 15cm plate, 2µg/ml doxycycline was added to induce expression. Cells were grown for an additional 7 days after induction before any cytometric analysis or FACS was performed.

PyronicSF linker library FACS.

Cells at 50% confluency in a 15cm plate were detached by trypsinization, pelleted and resuspended in 2mL DMEM supplemented with 50µg/mL gentamicin. Sorts were performed using a BD Influx instrument equipped with a 488nm laser for excitation and a 530/40nm emission filter for GFP measurements and 405nm laser with 460/50 filter for mTagBFP measurements. Cells transfected with Bxb1 recombinase but no attB plasmid were used to adjust instrument voltages and establish a baseline of autofluorescence in the green channel and presence of mTagBFP expression. Cells transfected with an attB plasmid but without expression of Bxb1 recombinase were used to determine the level of green fluorescence derived from plasmid expression as opposed to genomic integration, doxycycline induced expression. Cells transfected with the PyronicSF linker library and Bxb1 recombinase were sorted to collect cells positive for GFP and negative for mTagBFP fluorescence to enrich for recombined cells. A total of 300,000 GFP+/BFP- cells were collected in a 15cm tube containing 5mL DMEM supplemented with 10% FBS and 50µg/mL gentamicin. Sorted cells were pelleted,

resuspended in 800µL media and plated in a 12-well plate. Cells were expanded over about 7 days to reach 50% confluence in two 15cm plates before the second round of sorting.

The two samples for the second round of FACS were detached from 15cm plates by trypsinization, pelleted and resuspended in 2mL DMEM supplemented with 50µg/mL gentamicin and either 10mM pyruvate or an equal volume of water. Cells transfected with Bxb1 recombinase but no attB plasmid were used to adjust instrument voltages and establish a baseline of autofluorescence in the green channel. Four equal width gates on the log scale were set to span the range of log(AFU) from 1.0 to 3.0. The +/- pyruvate samples were sorted using the same gates for a duration of 1.5 hours each. The cells for each bin were collected in 5mL tubes containing 1mL DMEM supplemented with 10% FBS and 50µg/mL gentamicin. A range of 200,000 to 1,315,000 cells were collected for each bin approximately proportional to the relative density of cells in each bin. Sorted cells for each bin were individually pelleted, resuspended and plated in either a 24-well (<500,000 cells), 12-well (>500,000 and <900,000 cells) or 6-well plate (>900,000 cells). All samples were expanded to 50% in a 10cm plate before harvesting and storing at -20°C after pelleting and washing with PBS.

PyronicSF linker library sequencing.

Genomic DNA was extracted from approximately 5M cells using a Qiagen DNeasy Blood & Tissue Kit. The linker regions flanking cpGFP were PCR amplified from the genome using primers PyronicSF-LLseq-F and PyronicSF-LLseq-R.

Four replicate 50µL reactions were prepared for each sample with a final concentration of 5ng/µL genomic DNA, 0.25µM forward and reverse primer, 1X

SeqAmp PCR buffer, 1X SeqAmp Polymerase (Clontech), 1X SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions: [94°C 60s, (98°C 10s, 55°C 15s, 68°C 60s, plate read) x 26 cycles]. The number of cycles was determined such that reactions were in the exponential phase of amplification upon completion of the program. Replicate reactions were pooled and cleaned with a Nucleospin column and eluted in 15µL elution buffer (5 mM Tris/HCl, pH 8.5).

Pooled first round PCR products were amplified a second time with primer i5-IPE2p and a unique indexed primer per sample (i7-iPE2p-XX). 25µL reactions were prepared containing 1µL round 1 DNA, 0.5µM forward and reverse primer, 1X KAPA HiFi Hotstart Readymix (KHF), and 1X SYBR Green. Amplification was monitored by qPCR with cycling conditions: [95°C 3 minutes, (98°C 20 seconds, 60°C 15 seconds, 72°C 30 seconds, plate read, 72°C 8 seconds) x 8 cycles]. Reactions were removed during the exponential phase of amplification. PCR products were run on a 1.5% agarose gel to ensure only a single band had been produced and to estimate relative concentrations using FIJI gel analysis. Indexed samples were pooled normalizing for relative concentration. Pooled products were run on a 1.5% agarose gel, cutting out a band at 933bp which was then purified using the NucleoSpin Gel and PCR Clean-up column. The concentration of the pooled library was quantified using a Qubit fluorometer and size distribution was assessed using a HS DNA chip on the Bioanalyzer 2100 instrument (Agilent). The library was sequenced using 2x75bp paired-end reads on an Illumina MiSeq (v3 Reagent kit) loaded at a final concentration of 14pM with 15% PhiX spiked-in. Sequencing reads were processed using an R script available at:

https://github.com/jnkoberstein/biosensor-sort-seq. Briefly, reads were trimmed to remove all but the linker sequence using Cutadapt (Martin, 2011) and

allowed degenerate codons were counted using the ShortRead R/Bioconductor package[118]. The resulting table of read counts for each linker sequence in each sample was used as input for sort-seq analysis. A relatively high amount (28.6% of reads in the naive library) of the parent sequence (including bases not allowed using the degenerate scheme) was detected likely as a result of carryover from the plasmid used as a source for cloning the linker library. These reads were included for downstream MLE fitting but excluded from all further machine learning analyses in favor of the parent amino acid sequence encoded by allowed bases.

Sort-seq data analysis.

Sort-seq data analysis was performed as thoroughly described by Peterman et al.[45,99] using functions written in R available at: https://github.com/jnkoberstein/biosensor-sort-seq.The sorting experiment involves m gates of width *w* spanning a range of logarithmic fluorescence. Sorting gate *j* is defined by its upper and lower boundaries, L_j and U_j. Given read counts r_{ij} the number of read counts of variant i in bin *j*, the mean fluorescence assuming a log-normal distribution can be estimated using a maximum likelihood approach. Raw sequencing data was processed to obtain read counts of each variant in each bin using the dipseq python package for the MBP library or a custom R script for the PyronicSF library. A proportionality constant d_j relating read counts to the total number of sorted cells in each bin b_j is set as

$$d_j = \left(\sum_i b_{ij}\right) \Big/ \left(\sum_i r_{ij}\right)$$

Given the count data and gate parameters, the MLE is computed as the values $\hat{\mu}$ and $\hat{\sigma}$ that maximize the log-likelihood function

$$\log L(\mu, \sigma | r) = \sum_{j=1}^{m} r_j \log\left(F_{\mu,\sigma}(U_j) - F_{\mu,\sigma}(L_j)\right).$$

The values $\hat{\mu}$ and $\hat{\sigma}$ for each variant *i* were obtained by minimizing $-\log L(\mu, \sigma | r_{ij})$ over μ and σ using the Nelder-Mead algorithm while keeping the read counts and experimental parameters fixed. The log-normal mean fluorescence is then calculated as $F = 10^{(\hat{\mu}+\hat{\sigma}^2/2)}$. Mean fluorescence was calculated for both samples and then used to calculate dynamic range as $\frac{\Delta F}{F_0} = (F_l - F_0)/F_0$ where F_l is the fluorescence intensity in the sample with added ligand and F_0 is the fluorescence intensity in the absence of ligand. The MBP library dataset was filtered to keep only variants in which more than 100 cells were collected between the two samples and the variance in each sample was between 0.1 and 0.3. The PyronicSF library was filtered to keep only variants in which an estimate of more than 500 cells were collected for each sample and variance in each sample was between 0.1 and 0.4.

PyronicSF-VE-SS flow cytometry.

PyronicSF-VE-SS was cloned by Gibson assembly. The backbone was opened up by PCR amplification of HC_Kan_PyronicSF-p7 using primers PyronicSF-VE-SS-GA-F1/R1. The cpGFP insert was amplified from HC_Kan_PyronicSF-p7 using primer sets PyronicSF-VE-SS-GA-F2/R2. 20µL Gibson assembly reactions were prepared with 50ng backbone, 5:1 molar ratio of insert to backbone and 1X Gibson Assembly Master Mix (NEB) and incubated at 50°C for 1 hour before transformation into One Shot TOP10 Chemically Competent *E. coli* (Invitrogen). Single colonies were picked from LB agar plates with 50µg/mL kanamycin the next day and Sanger sequenced to confirm correct assembly. PyronicSF-VE-SS was cloned into EMMA-attB-Dest using an Esp3I mediated Golden Gate reaction as described for library cloning. The resulting plasmid, EMMA-attB-PyronicSF-VE-SS was transfected and

recombined into Landing Pad cells. Similarly, a construct containing the original linkers, EMMA-attB-PyronicSF-LE-TR was recombined into Landing Pad cells. Recombined cells grown to approximately 50% confluency in a 6-well plate were detached by trypsinization, pelleted and resuspended in 400µL DMEM. Cytometric evaluation was performed on a BD LSR II equipped with a 488nm laser for excitation and a 530/30nm emission filter for GFP measurements and 405nm laser with 440/40 filter for mTagBFP measurements. Events were gated for live cells using FSC-A and SSC-A and single cells using FSC-A and FSC-H. Recombined cells were analyzed by gating for loss of BFP fluorescence. GFP intensity for recombined cells was analyzed in the presence and absence of 20mM pyruvate.

Regression Models.

PyronicSF linker sequences were converted to 32 length numerical vectors by encoding each of the 4 amino acids as the 8 representative VHSE biochemical features. Model training was performed using the R packages caret[119] with the random forest model implemented in ranger[120] . An 80% split (819 variants) was used for model training and the remaining 20% (204 variants) was used for testing the final model. Fivefold cross-validation was used for hyperparameter tuning. Grid search was performed to find the optimal values for the following hyperparameters: mtry=5, 7, 13; splitrule=variance, extratrees, maxstat; min.node.size=1, 5, 7. Feature importance was calculated using the permutation method. The learning curve was generated by training the model with subsets consisting of 2.5% to 80% of the data while evaluating mean absolute error on the 20% test set and on the held out cross-validation resampling set with hyper-parameters set to mtry=7, splitrule=extratrees, and min.node.size=5. Random forest prediction performance

was compared to K-nearest neighbors (KNN), ridge regression and gaussian process (GP) models by 5-fold cross-validation. Hyperparameters for each model were tuned separately by 5-fold cross-validation with the final top model used for comparison. Tested hyperparameters were: k=1-10 for KNN, lambda=0, $1x10^{-4}$, $1x10^{-3}$, 0.1 and 1 for ridge regression, and sigma = 0, $1x10^{-4}$, $1x10^{-3}$, $1x10^{-2}$, 0.1, 0.5, and 1 for gaussian process using the radial basis function kernel.

Data Availability.

Raw sequencing reads associated with both the MBP and PyronicSF sort-seq experiments can accessed from the NCBI Sequence Read Archive (SRA) with the accession code PRJNA732942.

Author Contributions

Author Contributions J.N.K. performed sort-seq experiments, analyzed data, characterized biosensor performance by flow cytometry and plate reader assays, prepared figures and wrote the manuscript. M.L.S. and C.B.S. designed and cloned biosensor libraries and edited the manuscript. T.L.M. assisted with sort-seq experimental design and data analysis. M.S.C. supervised research and edited the manuscript.

Acknowledgements

We thank the OHSU Flow Cytometry Core for assistance with fluorescence activated cell sorting and cytometric analysis; Y. Jia and the OHSU Molecular Technologies Core for sequencing services; and D. Fowler and K. Matrayek for sharing the HEK293T Landing Pad cell line.

2.6 Supplemental Material

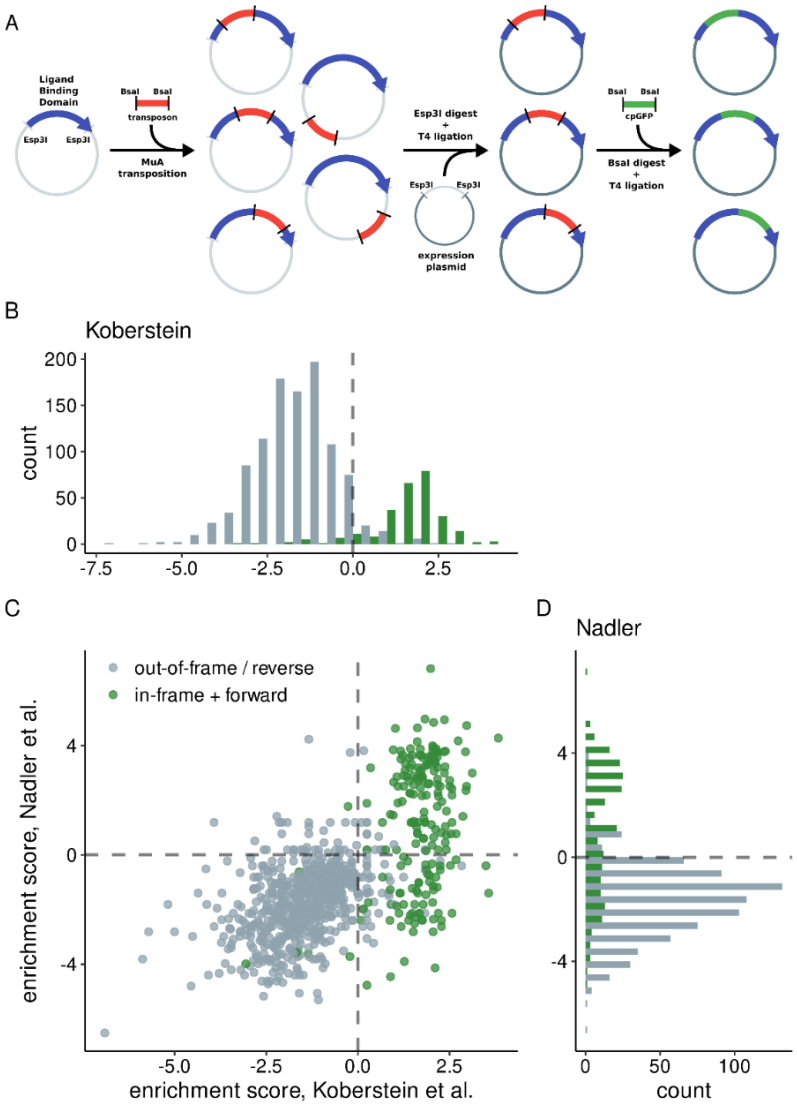

Figure S1. Enrichment of productive cpGFP insertions into MBP. (A) Overview of the transposon mediated domain-insertion cloning method. An engineered transposon sequence is inserted throughout the sequence of a target ligand-binding domain. Subsequent Golden Gate cloning steps move the transposed ORF into an expression plasmid and finally replace the transposon sequence with the coding sequence of cpGFP. Adapted with permission from Nadler, D. C.; Morgan, S.-A.; Flamholz, A.; Kortright, K. E.; Savage, D. F. Rapid Construction of Metabolite Biosensors Using Domain-Insertion Profiling. *Nature Communications* **2016**, 7, 12266. Copyright 2016 Springer Nature. **(B)** Distribution of enrichment values from this study for MBP domain-insertion variants following a single enrichment sort for GFP+ cells. **(C)** Comparison of enrichment values for MBP domain-insertion variants across studies reveals consistent enrichment of variants on the basis of fluorescence intensity generated by in-frame and forward insertion of GFP (Spearman's ρ = 0.6). **(D)** Distribution of enrichment values from Nadler et al. for MBP domain-insertion variants following a single enrichment sort for GFP+ cells.

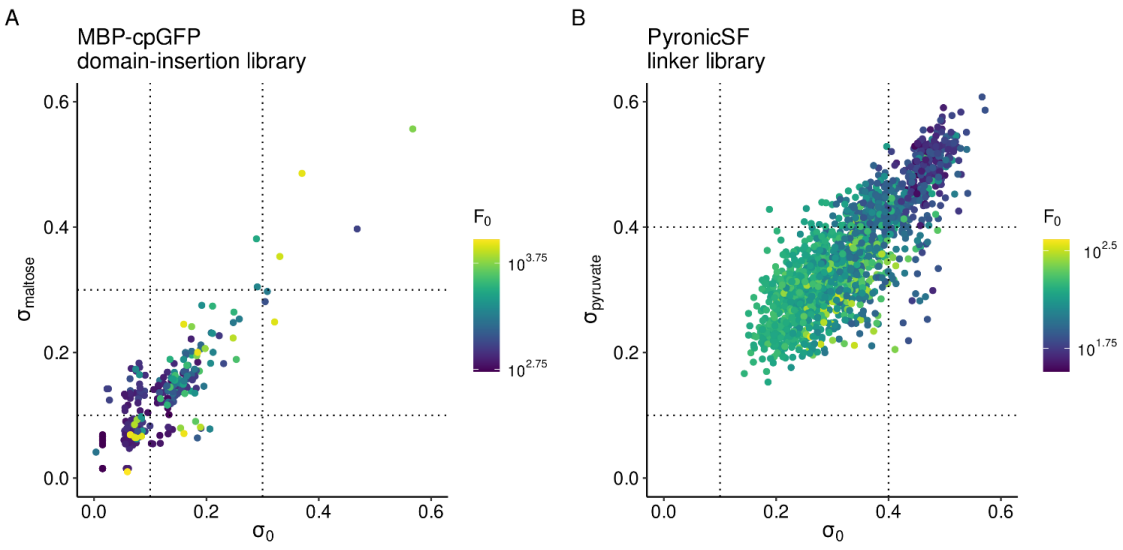

Figure S2. Sort-seq estimates of variance are correlated for individual variants across conditions. (A) Sort-seq derived estimates of log-normal variance (σ) compared between conditions for all productive MBP-cpGFP domain-insertion variants. **(B)** Sort-seq derived estimates of log-normal variance (σ) compared between conditions for all PyronicSF linker variants. Dashed lines indicate the lower and upper bounds used to filter out variants with lower and higher variance than expected, respectively.

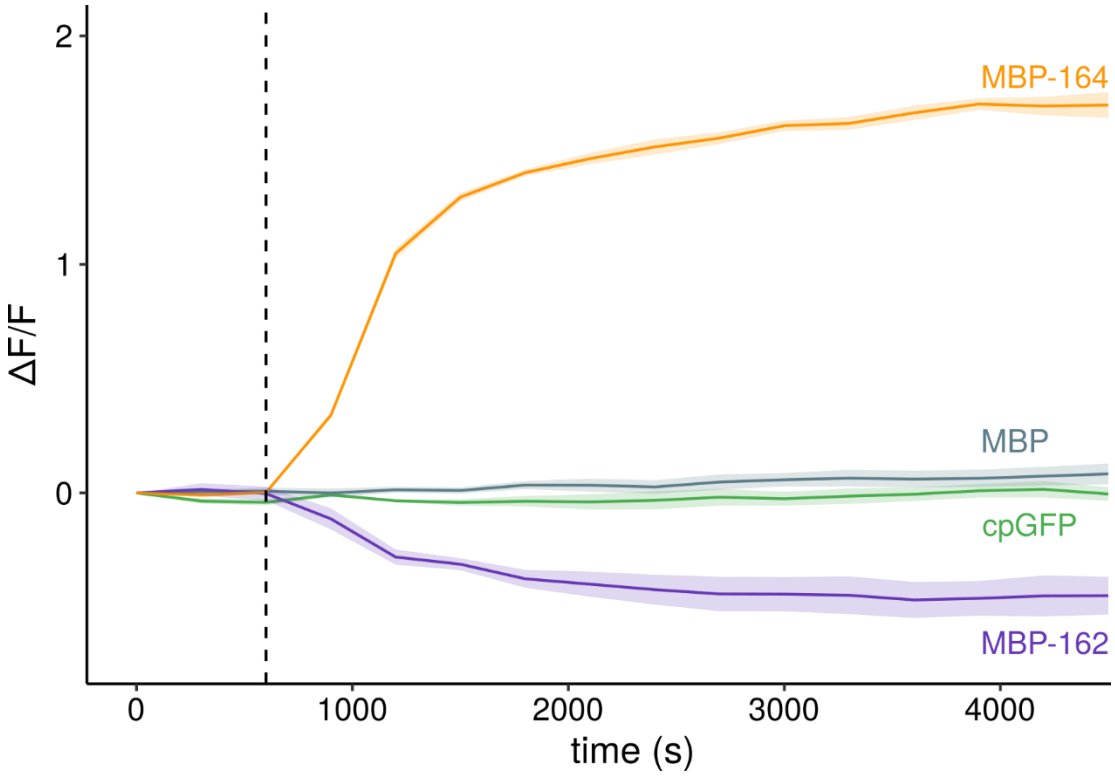

Figure S3. MBP-162 and MBP-164 function as biosensors with opposite responses to increased maltose concentration. The fluorescence of MBP-162 and MBP-164 were monitored while expressed in *E. coli* following the addition of 1mM maltose (dashed line) to the media. Controls expressing either MBP or cpGFP were monitored for non-specific changes with addition of maltose. Fluorescence was measured for wells with either maltose or an equal volume of PBS added and the difference between conditions calculated to account for increases in intensity due to changes in protein expression over time. Data are mean ± s.d. for three replicates for each condition.

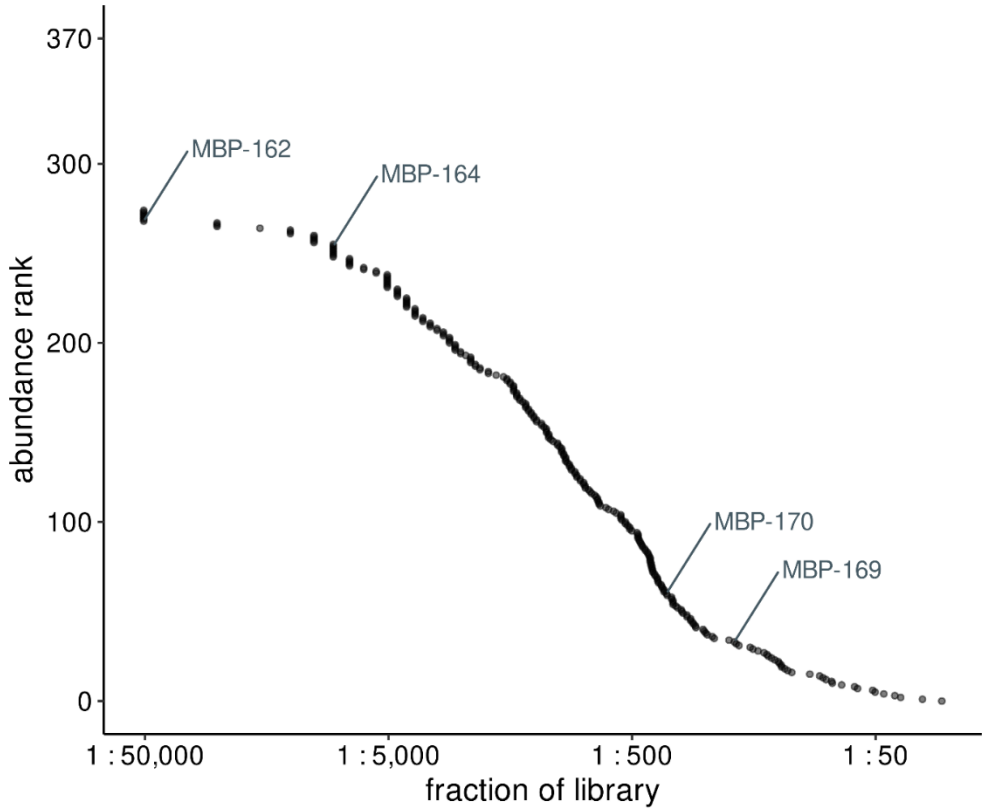

Figure S4. Domain-insertion variant abundance following the fluorescence enrichment sort. The fraction of reads containing a given variant versus the ranked abundance, with 1 being the most abundant variant and 370 being the least. Out of the 370 possible domain-insertion variants, 275 appear at least once in the high-throughput sequencing reads. Abundance is highly variable with some variants appearing only once in approximately 50,000 sequencing reads while others appear as often as once in every 50 reads.

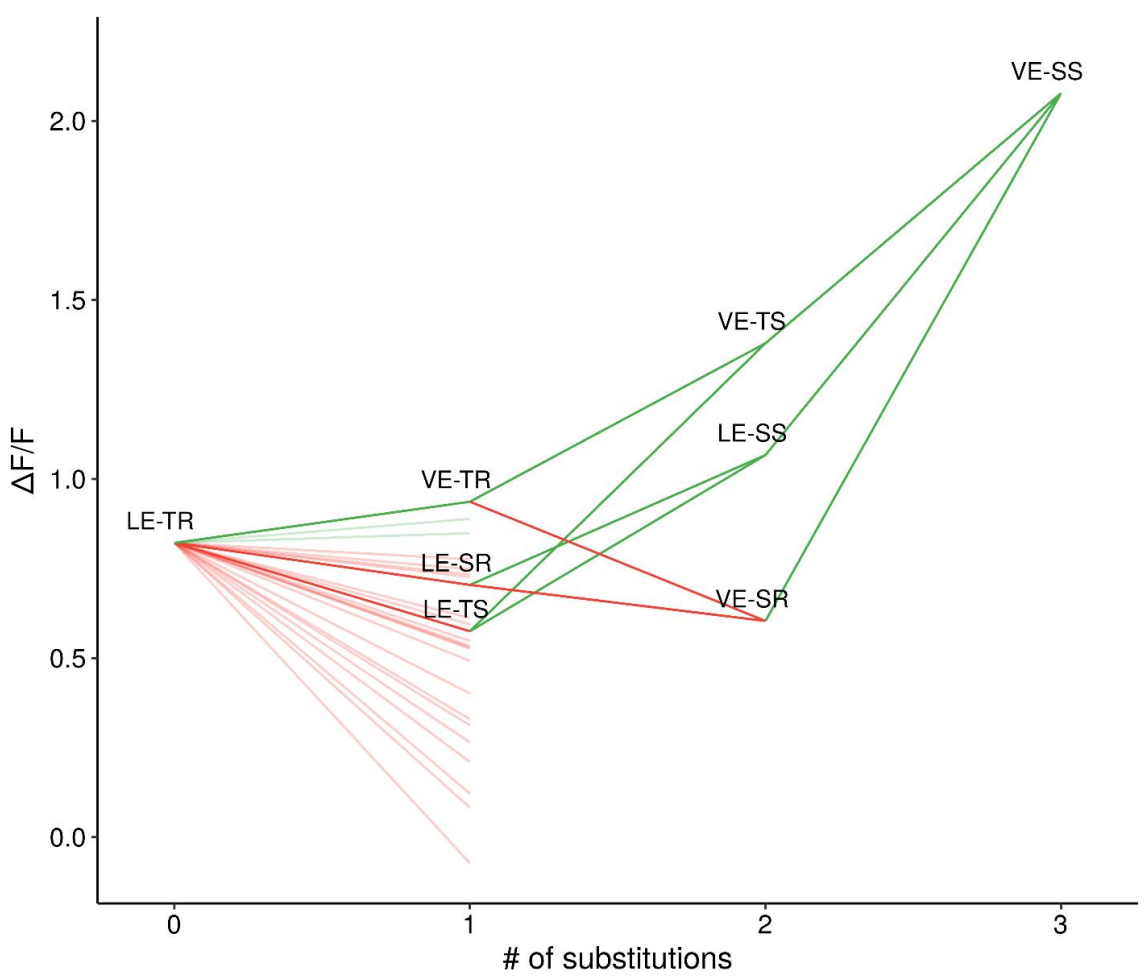

Figure S5. Dynamic range for linker variants versus the number of substitutions to the parent sequence. Five out of the six paths from the parent sequence (LE-TR) to the optimized sequence (VE-SS) travel through intermediates with decreased function (dark lines; green/red indicates increase/decrease in function relative to parent sequence). A majority (21/24) of the tested single substitutions to the parent sequence (additional shaded lines) result in a decrease in function. The total increase in dynamic range of VE-SS compared to the parent sequence is much greater (ΔF/F=2.08) than the sum of the effects of the comprising single substitutions (ΔF/F = 0.57).

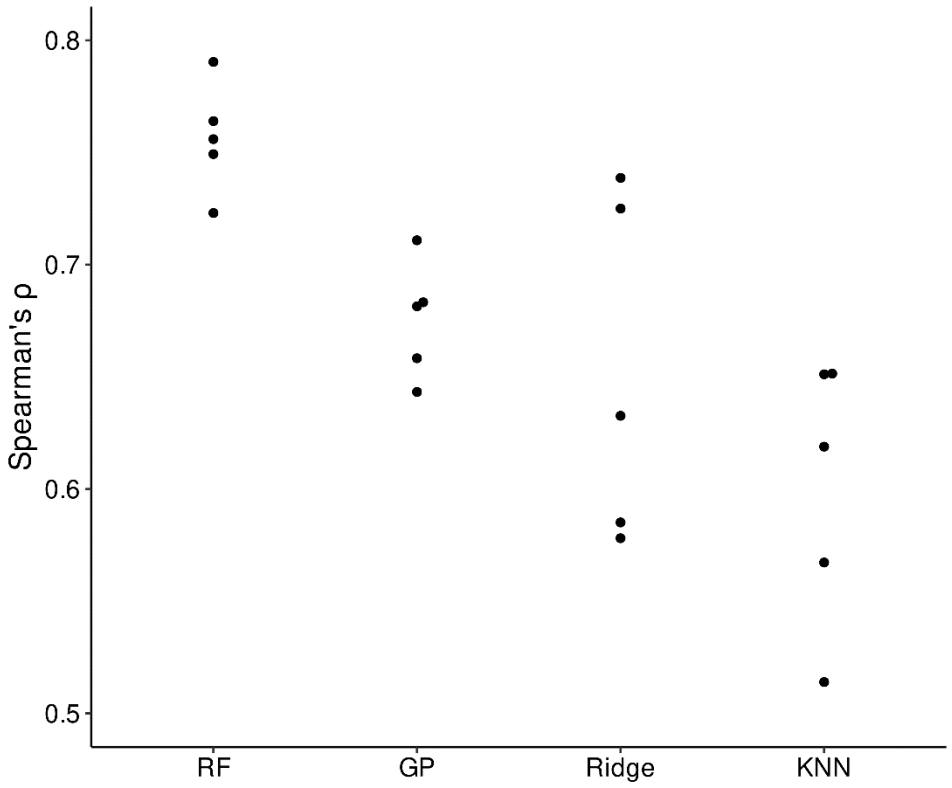

Figure S6. Comparison of different regression models for predicting dynamic range from sequence. Each model was assessed by 5-fold cross-validation using the Spearman rank correlation coefficient calculated between the model prediction and measured dynamic range for the held-out fold. Random forest exhibits the highest mean correlation (ρ=0.76) across the tested models followed by Gaussian Process (GP, ρ=0.68), ridge regression (Ridge, ρ=0.65) and K-nearest neighbors (KNN, ρ=0.60).

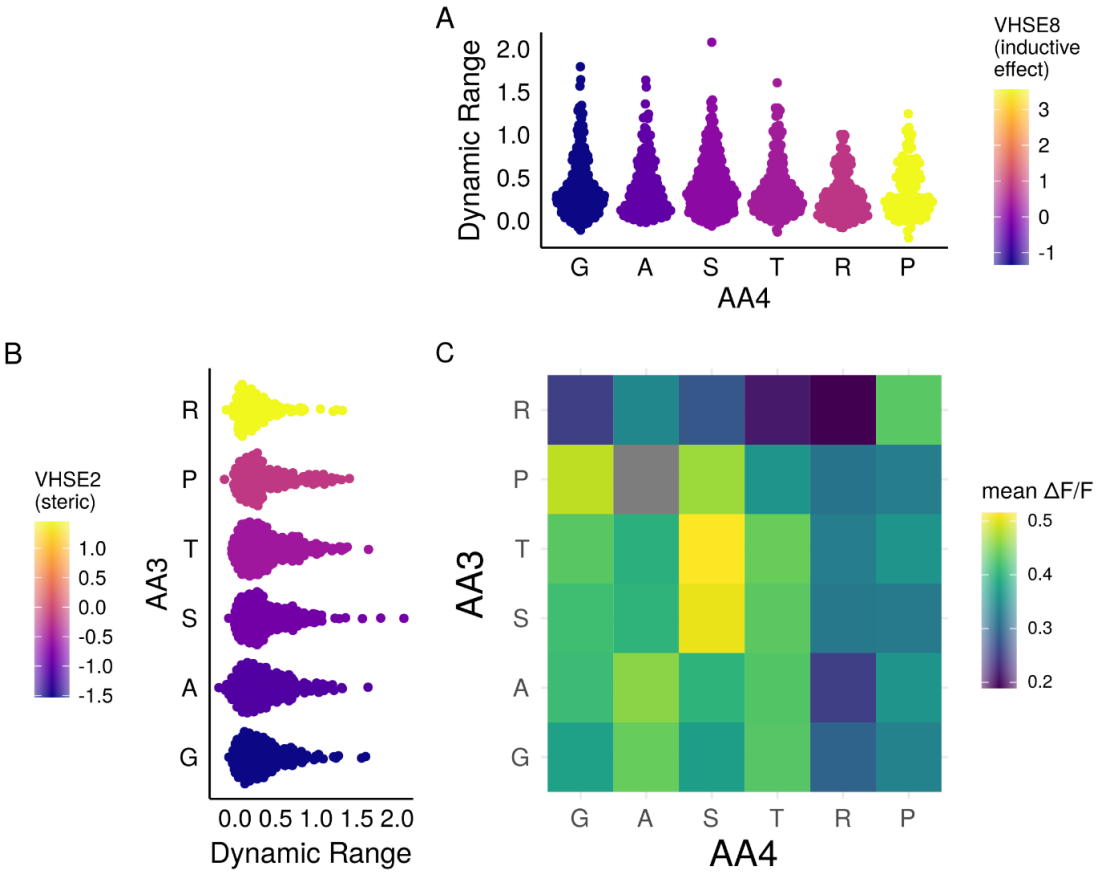

Figure S7. Analysis of the biochemical basis of C-terminal linker function. (A) Distributions of variant dynamic-range estimates sorted by VHSE8 of second C-terminal linker amino acid. **(B)** Distributions of variant dynamic-range estimates sorted by VHSE2 of the first C-terminal linker amino acid. **(C)** Mean dynamic-range for C-terminal linker pairs. The combination P^3A^4 is excluded (grey) as no variants with the linker combination passed filtering.

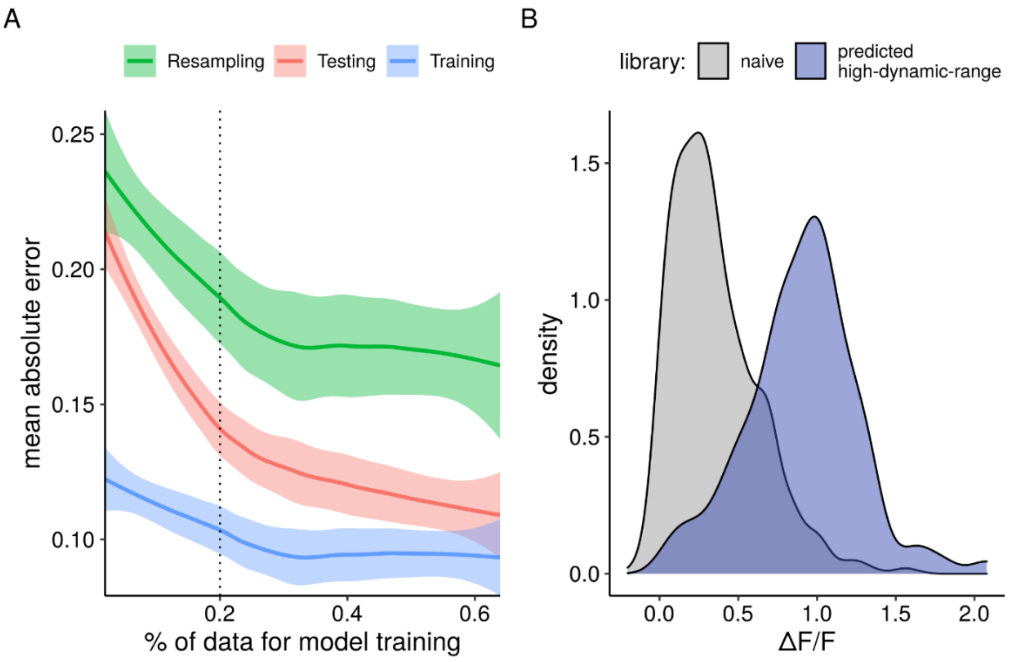

Figure S8. A random forest model trained on a subset of the collected data can be used to design a library with an increase in dynamic range. (A) Mean absolute error as a function of the percent of data used to train the model evaluated on predictions of dynamic range for the test, training, and resampled data. Lines represent a locally weighted smoothing (LOESS) fit and shaded regions represent a 95% confidence interval. The vertical dashed line represents the 20% of data used to train the model in panel B. **(B)** Distribution of dynamic range estimates for variants in the naive library compared to the high-dynamic-range variants predicted by a model trained on 20% of the data (192 variants).

Supplementary Text. Sequences of the constructs used in this study.

PyronicSF-LE-TR

PdhR 1-188

Linkers

cpGFP

PdhR 189-254

Amino Acid:

MGSAYSKIRQPKLSDVIEQQLEFLILEGTLRPGEKLPPERELAKQFDVSRPSLREAIQRLEAKGLLLRR
QGGGTFVQSSLWQSFSDPLVELLSDHPESQYDLLETRHALEGIAAYYAALRSTDEDKERIRELHHAIE
LAQQSGDLDAESNAVLQYQIAVTEAAHNVVLLHLLRCMEPMLAQNVRQNFELLLENVYIKADKQKN
GIKANFKIRHNIEDGGVQLAYHYQQNTPIGDGPVLLPDNHYLSVQSKLSKDPNEKRDHMVLLEFVTA
AGITLGMDELYKGGTGGSMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYIQERTIFFKDDGNYKTRAEVKFE
GDTLVNRIELKGIDFKEDGNILGHKLEYNTRYSRREMLPLVSSHRTRIFEAIMAGKPEEAREASHRHL
AFIEEILLDRSREESRRERSLRRLEQRKNSG*

DNA:

ATGGGCAGCGCATATAGCAAAATTCGGCAGCCCAAACTGAGCGATGTGATTGAGCAGCAGCTGGAG
TTTCTGATTCTGGAAGGCACCCTGAGGCCTGGAGAGAAACTGCCCCCTGAGCGCGAACTCGCCAAGC
AGTTCGACGTGAGTCGACCATCACTGAGGGAGGCTATCCAGAGGCTGGAAGCAAAGGGACTGCTCC
TGAGGAGACAGGGAGGAGGGACTTTCGTGCAGAGCTCCCTGTGGCAGAGCTTCAGCGACCCCCTGG
TCGAGCTCCTGTCTGACCACCCAGAAAGTCAGTACGATCTCCTGGAGACAAGACATGCTCTGGAAGG
CATCGCCGCTTACTATGCAGCCCTGCGGTCCACTGACGAGGATAAGGAACGCATCCGAGAGCTGCAC
CATGCCATTGAACTCGCTCAGCAGTCAGGAGATCTGGATGCAGAGAGCAACGCCGTGCTGCAGTACC
AGATTGCAGTCACCGAGGCTGCACACAATGTGGTCCTCCTGCATCTCCTGAGGTGCATGGAGCCAAT
GCTGGCCCAGAACGTGAGACAGAATTTTGAGCTCCTGCTCGAAAACGTCTATATCAAGGCCGACAA
GCAGAAAAACGGCATTAAGGCTAACTTCAAGATCAGACACAACATCGAGGATGGTGGCGTGCAGCT
GGCCTACCATTATCAGCAGAACACACCAATCGGAGATGGACCAGTGCTGCTCCCAGATAATCACTAC
CTGAGCGTCCAGTCCAAGCTGTCTAAAGACCCTAACGAGAAGCGGGATCATATGGTGCTGCTCGAA
TTTGTCACAGCCGCTGGGATCACTCTGGGTATGGACGAGCTCTATAAAGGAGGGACCGGTGGCAGT
ATGGTGTCAAAGGGCGAGGAACTGTTCACAGGAGTGGTCCCCATTCTGGTGGAGCTCGACGGCGAT
GTCAATGGACACAAATTTTCCGTGTCTGGCGAGGGCGAAGGAGATGCTACCTACGGGAAGCTGACA
CTCAAATTCATCTGCACCACAGGCAAGCTGCCAGTGCCCTGGCCTACTCTGGTCACTACCCTCACCT
ACGGGGTGCAGTGTTTCTCCAGATATCCCGACCACATGAAGCAGCATGATTTCTTTAAATCTGCTAT
GCCTGAGGGGTACATCCAGGAACGGACAATTTTCTTTAAGGACGATGGTAACTACAAAACACGCGC
AGAGGTGAAGTTCGAAGGCGACACTCTGGTCAATCGAATCGAGCTGAAGGGAATTGACTTTAAAGA
AGATGGGAACATCCTGGGTCACAAGCTGGAGTACAATACTAGGTATTCTCGGCGCGAAATGCTGCC
ACTCGTGTCTAGTCACAGGACCAGAATCTTTGAGGCAATTATGGCCGGAAAGCCCGAGGAAGCTAG
AGAAGCAAGTCACCGGCATCTGGCCTTCATCGAGGAAATTCTGCTCGACCGGAGCCGCGAGGAATCC
CGAAGGGAGCGCAGCCTGAGGCGACTCGAACAGCGAAAGAACTCAGGCTAA

PyronicSF-VE-SS

PdhR 1-188

Linkers

cpGFP

PdhR 189-254

Amino Acid:

MGSAYSKIRQPKLSDVIEQQLEFLILEGTLRPGEKLPPERELAKQFDVSRPSLREAIQRLEAKGLLLRR
QGGGTFVQSSLWQSFSDPLVELLSDHPESQYDLLETRHALEGIAAYYAALRSTDEDKERIRELHHAIE
LAQQSGDLDAESNAVLQYQIAVTEAAHNVVLLHLLRCMEPMLAQNVRQNFELLVENVYIKADKQKN
GIKANFKIRHNIEDGGVQLAYHYQQNTPIGDGPVLLPDNHYLSVQSKLSKDPNEKRDHMVLLEFVTA
AGITLGMDELYKGGTGGSMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYIQERTIFFKDDGNYKSSAEVKFEG
DTLVNRIELKGIDFKEDGNILGHKLEYNTRYSRREMLPLVSSHRTRIFEAIMAGKPEEAREASHRHLA
FIEEILLDRSREESRRERSLRRLEQRKNSG*

DNA:

ATGGGCAGCGCATATAGCAAAATTCGGCAGCCCAAACTGAGCGATGTGATTGAGCAGCAGCTGGAG
TTTCTGATTCTGGAAGGCACCCTGAGGCCTGGAGAGAAACTGCCCCCTGAGCGCGAACTCGCCAAGC
AGTTCGACGTGAGTCGACCATCACTGAGGGAGGCTATCCAGAGGCTGGAAGCAAAGGGACTGCTCC
TGAGGAGACAGGGAGGAGGGACTTTCGTGCAGAGCTCCCTGTGGCAGAGCTTCAGCGACCCCCTGG
TCGAGCTCCTGTCTGACCACCCAGAAAGTCAGTACGATCTCCTGGAGACAAGACATGCTCTGGAAGG
CATCGCCGCTTACTATGCAGCCCTGCGGTCCACTGACGAGGATAAGGAACGCATCCGAGAGCTGCAC
CATGCCATTGAACTCGCTCAGCAGTCAGGAGATCTGGATGCAGAGAGCAACGCCGTGCTGCAGTACC
AGATTGCAGTCACCGAGGCTGCACACAATGTGGTCCTCCTGCATCTCCTGAGGTGCATGGAGCCAAT
GCTGGCCCAGAACGTGAGACAGAATTTTGAGCTCCTGGTAGAAAACGTCTATATCAAGGCCGACAA
GCAGAAAAACGGCATTAAGGCTAACTTCAAGATCAGACACAACATCGAGGATGGTGGCGTGCAGCT
GGCCTACCATTATCAGCAGAACACACCAATCGGAGATGGACCAGTGCTGCTCCCAGATAATCACTAC
CTGAGCGTCCAGTCCAAGCTGTCTAAAGACCCTAACGAGAAGCGGGATCATATGGTGCTGCTCGAA
TTTGTCACAGCCGCTGGGATCACTCTGGGTATGGACGAGCTCTATAAAGGAGGGACCGGTGGCAGT
ATGGTGTCAAAGGGCGAGGAACTGTTCACAGGAGTGGTCCCCATTCTGGTGGAGCTCGACGGCGAT
GTCAATGGACACAAATTTTCCGTGTCTGGCGAGGGCGAAGGAGATGCTACCTACGGGAAGCTGACA
CTCAAATTCATCTGCACCACAGGCAAGCTGCCAGTGCCCTGGCCTACTCTGGTCACTACCCTCACCT
ACGGGGTGCAGTGTTTCTCCAGATATCCCGACCACATGAAGCAGCATGATTTCTTTAAATCTGCTAT
GCCTGAGGGGTACATCCAGGAACGGACAATTTTCTTTAAGGACGATGGTAACTACAAAACACGCGC
AGAGGTGAAGTTCGAAGGCGACACTCTGGTCAATCGAATCGAGCTGAAGGGAATTGACTTTAAAGA
AGATGGGAACATCCTGGGTCACAAGCTGGAGTACAATAGTAGTTATTCTCGGCGCGAAATGCTGCC
ACTCGTGTCTAGTCACAGGACCAGAATCTTTGAGGCAATTATGGCCGGAAAGCCCGAGGAAGCTAG
AGAAGCAAGTCACCGGCATCTGGCCTTCATCGAGGAAATTCTGCTCGACCGGAGCCGCGAGGAATCC
CGAAGGGAGCGCAGCCTGAGGCGACTCGAACAGCGAAAGAACTCAGGCTAA

PyronicSF-SNAx2/VSTx2

PdhR 1-188

Linkers

X={A,G,E,L,Q,R,P,V}

Z={A,G,P,R,S,T}

cpGFP

PdhR 189-254

Amino Acid:

MGSAYSKIRQPKLSDVIEQQLEFLILEGTLRPGEKLPPERELAKQFDVSRPSLREAIQRLEAKGLLLRR
QGGGTFVQSSLWQSFSDPLVELLSDHPESQYDLLETRHALEGIAAYYAALRSTDEDKERIRELHHAIE
LAQQSGDLDAESNAVLQYQIAVTEAAHNVVLLHLLRCMEPMLAQNVRQNFELLXXNVYIKADKQKN
GIKANFKIRHNIEDGGVQLAYHYQQNTPIGDGPVLLPDNHYLSVQSKLSKDPNEKRDHMVLLEFVTA
AGITLGMDELYKGGTGGSMVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYIQERTIFFKDDGNYKZZAEVKFE
GDTLVNRIELKGIDFKEDGNILGHKLEYNTRYSRREMLPLVSSHRTRIFEAIMAGKPEEAREASHRHL
AFIEEILLDRSREESRRERSLRRLEQRKNSG*

DNA:

ATGGGCAGCGCATATAGCAAAATTCGGCAGCCCAAACTGAGCGATGTGATTGAGCAGCAGCTGGAG
TTTCTGATTCTGGAAGGCACCCTGAGGCCTGGAGAGAAACTGCCCCCTGAGCGCGAACTCGCCAAGC
AGTTCGACGTGAGTCGACCATCACTGAGGGAGGCTATCCAGAGGCTGGAAGCAAAGGGACTGCTCC
TGAGGAGACAGGGAGGAGGGACTTTCGTGCAGAGCTCCCTGTGGCAGAGCTTCAGCGACCCCCTGG
TCGAGCTCCTGTCTGACCACCCAGAAAGTCAGTACGATCTCCTGGAGACAAGACATGCTCTGGAAGG
CATCGCCGCTTACTATGCAGCCCTGCGGTCCACTGACGAGGATAAGGAACGCATCCGAGAGCTGCAC
CATGCCATTGAACTCGCTCAGCAGTCAGGAGATCTGGATGCAGAGAGCAACGCCGTGCTGCAGTACC
AGATTGCAGTCACCGAGGCTGCACACAATGTGGTCCTCCTGCATCTCCTGAGGTGCATGGAGCCAAT
GCTGGCCCAGAACGTGAGACAGAATTTTGAGCTCCTGSNASNAAACGTCTATATCAAGGCCGACAA
GCAGAAAAACGGCATTAAGGCTAACTTCAAGATCAGACACAACATCGAGGATGGTGGCGTGCAGCT
GGCCTACCATTATCAGCAGAACACACCAATCGGAGATGGACCAGTGCTGCTCCCAGATAATCACTAC
CTGAGCGTCCAGTCCAAGCTGTCTAAAGACCCTAACGAGAAGCGGGATCATATGGTGCTGCTCGAA
TTTGTCACAGCCGCTGGGATCACTCTGGGTATGGACGAGCTCTATAAAGGAGGGACCGGTGGCAGT
ATGGTGTCAAAGGGCGAGGAACTGTTCACAGGAGTGGTCCCCATTCTGGTGGAGCTCGACGGCGAT
GTCAATGGACACAAATTTTCCGTGTCTGGCGAGGGCGAAGGAGATGCTACCTACGGGAAGCTGACA
CTCAAATTCATCTGCACCACAGGCAAGCTGCCAGTGCCCTGGCCTACTCTGGTCACTACCCTCACCT
ACGGGGTGCAGTGTTTCTCCAGATATCCCGACCACATGAAGCAGCATGATTTCTTTAAATCTGCTAT
GCCTGAGGGGTACATCCAGGAACGGACAATTTTCTTTAAGGACGATGGTAACTACAAAACACGCGC
AGAGGTGAAGTTCGAAGGCGACACTCTGGTCAATCGAATCGAGCTGAAGGGAATTGACTTTAAAGA
AGATGGGAACATCCTGGGTCACAAGCTGGAGTACAATVSTVSTTATTCTCGGCGCGAAATGCTGCCA
CTCGTGTCTAGTCACAGGACCAGAATCTTTGAGGCAATTATGGCCGGAAAGCCCGAGGAAGCTAGA
GAAGCAAGTCACCGGCATCTGGCCTTCATCGAGGAAATTCTGCTCGACCGGAGCCGCGAGGAATCCC
GAAGGGAGCGCAGCCTGAGGCGACTCGAACAGCGAAAGAACTCAGGCTAA

MBP-162

MBP 1-162

Linkers

cpGFP

MBP 161-370

Amino Acid:

MSKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHD
RFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIP
ALDKELKAKGKSALMFNLQEPYFTWPLIAASYNVFIMADKQKNGIKANFKIRHNIEDGGVQLAYHYQ
QNTPIGDGPVLLPDNHYLSVQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYKGGTGGSMVSKG
EELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQCFS
RYPDHMKQHDFFKSAMPEGYIQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGH
KLEYNFNASIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFN
KGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLL
TDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAA
SGRQTVDEALKDAQTRITKSGHHHHHH*

DNA:

ATGTCCAAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGACTC
GCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAA
CTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACG
ACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGA
CAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCT
GTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG
ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAA
CCGTACTTCACCTGGCCGCTGATTGCTGCATCTTATAACGTCTTTATCATGGCCGACAAGCAGAAGA
ACGGCATCAAGGCGAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCGCCTATCA
CTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCGTG
CAGTCCAAACTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACC
GCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGCGGTACCGGAGGGAGCATGGTGAGC
AAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGC
CACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCA
TCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAG
TGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCT
ACATTCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTATAAGACACGCGCTGAGGTTAAGT
TCGAGGGCGACACTCTGGTTAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACA
TCCTGGGCCATAAGCTTGAATATAACTTCAACGCGTCAATTGCTGCTGACGGGGGTTATGCGTTCA
AGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTC
TGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCAG
AAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATCG
ACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCGT
TCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTCC
TCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCCG
TAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAACG
CCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTGC
GGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTCGTATC
ACCAAGAGCGGTCACCATCACCATCACCATTAA

MBP-164

MBP 1-164

Linkers

cpGFP

MBP 163-370

Amino Acid:

MSKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAHD
RFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIP
ALDKELKAKGKSALMFNLQEPYFTWPLIAADASYNVFIMADKQKNGIKANFKIRHNIEDGGVQLAYH
YQQNTPIGDGPVLLPDNHYLSVQSKLSKDPNEKRDHMVLLEFVTAAGITLGMDELYKGGTGGSMVS
KGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLTYGVQC
FSRYPDHMKQHDFFKSAMPEGYIQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNIL
GHKLEYNFNASADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAF
NKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYL
LTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINA
ASGRQTVDEALKDAQTRITKSGHHHHHH*

DNA:

ATGTCCAAAATCGAAGAAGGTAAACTGGTAATCTGGATTAACGGCGATAAAGGCTATAACGGACTC
GCTGAAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCACCGTTGAGCATCCGGATAAA
CTGGAAGAGAAATTCCCACAGGTTGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACACG
ACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAATCACCCCGGACAAAGCGTTCCAGGA
CAAGCTGTATCCGTTTACCTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCGATCGCT
GTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGCTGCCGAACCCGCCAAAAACCTGGGAAGAG
ATCCCGGCGCTGGATAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAACCTGCAAGAA
CCGTACTTCACCTGGCCGCTGATTGCTGCTGATGCATCTTATAACGTCTTTATCATGGCCGACAAGC
AGAAGAACGGCATCAAGGCGAACTTCAAGATCCGCCACAACATCGAGGACGGCGGCGTGCAGCTCG
CCTATCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCT
GAGCGTGCAGTCCAAACTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTT
CGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGCGGTACCGGAGGGAGCAT
GGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGT
AAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTG
AAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACG
GCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCC
CGAAGGCTACATTCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTATAAGACACGCGCTGA
GGTTAAGTTCGAGGGCGACACTCTGGTTAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGA
CGGCAACATCCTGGGCCATAAGCTTGAATATAACTTCAACGCGTCAGCTGACGGGGGTTATGCGTT
CAAGTATGAAAACGGCAAGTACGACATTAAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGG
TCTGACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAGACACCGATTACTCCATCGCA
GAAGCTGCCTTTAATAAAGGCGAAACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATC
GACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCTTCAAGGGTCAACCATCCAAACCG
TTCGTTGGCGTGCTGAGCGCAGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGAGTTC
CTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCGGTTAATAAAGACAAACCGCTGGGTGCC
GTAGCGCTGAAGTCTTACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCATGGAAAAC
GCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAGATGTCCGCTTTCTGGTATGCCGTGCGTACTG
CGGTGATCAACGCCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGCGCAGACTCGTAT
CACCAAGAGCGGTCACCATCACCATCACCATTAA

Supplementary Table 1

Type	Name	Sequence
cloning primers	MBP-BsaI-GG-F	cacaccaggtctcaGTCCAAAATCGAAGAAGGTAAACTGGTAATCTGG
	MBP-BsaI-GG-R	cacaccaggtctcaCGCTCTTGGTGATACGATCTGCGC
	cpGFP-BsaI-GG-F	caatgcggtctcgcatcTATAACGTCTTTATCATGGC
	cpGFP-BsaI-GG-R	caatgcggtctcgacgcGTTGAAGTTATATTCAAGCT
	MBP-162-GA-F1	AATATAACTTCAACGCGTCAATTGCTGCTGACGGGGGTTA
	MBP-162-GA-R1	TAAAGACGTTATAAGATGCAGCAATCAGCGGCCAGGTGAA
	MBP-162-GA-F2	TTCACCTGGCCGCTGATTGCTGCATCTTATAACGTCTTTA
	MBP-162-GA-R2	TAACCCCCGTCAGCAGCAATTGACGCGTTGAAGTTATATT
	MBP-164-GA-F1	AATATAACTTCAACGCGTCAGCTGACGGGGGTTATGCGTT
	MBP-164-GA-R1	TAAAGACGTTATAAGATGCATCAGCAGCAATCAGCGGCCA
	MBP-164-GA-F2	TGGCCGCTGATTGCTGCTGATGCATCTTATAACGTCTTTA
	MBP-164-GA-R2	AACGCATAACCCCCGTCAGCTGACGCGTTGAAGTTATATT
	PyronicSF-VSTx2-GA-F1	AGCTGGAGTACAATVSTVSTTATTCTCGGCGCGAAATGCT
	PyronicSF-SNAx2-GA-R1	TCGGCCTTGATATAGACGTTTNSTNSCAGGAGCTCAAAAT
	PyronicSF-SNAx2-GA-F2	ATTTTGAGCTCCTGSNASNAAACGTCTATATCAAGGCCGA
	PyronicSF-VSTx2-GA-R2	AGCATTTCGCGCCGAGAATAASBASBATTGTACTCCAGCT
	PyronicSF-VE-SS-GA-F1	AGCTGGAGTACAATAGTAGTTATTCTCGGCGCGAAATGCT
	PyronicSF-VE-SS-GA-R1	TTGATATAGACGTTTTCTACCAGGAGCTCAAAATTCTGTCT
	PyronicSF-VE-SS-GA-F2	ATTTTGAGCTCCTGGTAGAAAACGTCTATATCAAGGCCGA

	PyronicSF-VE-SS-GA-R2	TCGCGCCGAGAATAACTACTATTGTACTCCAGCTTGTGAC
sequencing primers	PyronicSF-LLseq-F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNACGTGAGACAGAATTTTGAGCT
	PyronicSF-LLseq-R	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNTAGACACGAGTGGCAGCATTTC
	i5-IPE2p	AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGAC
	i7-TXX-iPE2p	CAAGCAGAAGACGGCATACGAGAT[i7index]GTGACTGGAGTTCAGACGTGTGC
	i7-T71-IPE2p	CAAGCAGAAGACGGCATACGAGATaactaggcgcGTGACTGGAGTTCAGACGTGTGC
	i7-T72-IPE2p	CAAGCAGAAGACGGCATACGAGATtcgctaagcaGTGACTGGAGTTCAGACGTGTGC
	i7-T73-IPE2p	CAAGCAGAAGACGGCATACGAGATtatatactaaGTGACTGGAGTTCAGACGTGTGC
	i7-T74-IPE2p	CAAGCAGAAGACGGCATACGAGATacttgctagaGTGACTGGAGTTCAGACGTGTGC
	i7-T75-IPE2p	CAAGCAGAAGACGGCATACGAGATaaccattggaGTGACTGGAGTTCAGACGTGTGC
	i7-T76-IPE2p	CAAGCAGAAGACGGCATACGAGATtcgcggttggGTGACTGGAGTTCAGACGTGTGC
	i7-T77-IPE2p	CAAGCAGAAGACGGCATACGAGATcgtagttaccGTGACTGGAGTTCAGACGTGTGC
	i7-T78-IPE2p	CAAGCAGAAGACGGCATACGAGATtccaatcatcGTGACTGGAGTTCAGACGTGTGC
	i7-T79-IPE2p	CAAGCAGAAGACGGCATACGAGATaatcgataatGTGACTGGAGTTCAGACGTGTGC
	i7-T80-IPE2p	CAAGCAGAAGACGGCATACGAGATccattatctaGTGACTGGAGTTCAGACGTGTGC
	i7-T81-IPE2p	CAAGCAGAAGACGGCATACGAGATtcaacgtaagGTGACTGGAGTTCAGACGTGTGC
	i7-T82-IPE2p	CAAGCAGAAGACGGCATACGAGATtctaatagtaGTGACTGGAGTTCAGACGTGTGC
	i7-T83-IPE2p	CAAGCAGAAGACGGCATACGAGATaaccgctggtGTGACTGGAGTTCAGACGTGTGC
	i7-T84-IPE2p	CAAGCAGAAGACGGCATACGAGATgatcgcttctGTGACTGGAGTTCAGACGTGTGC
	i7-T85-IPE2p	CAAGCAGAAGACGGCATACGAGATctaactagatGTGACTGGAGTTCAGACGTGTGC
	i7-T86-IPE2p	CAAGCAGAAGACGGCATACGAGATgctggaacttGTGACTGGAGTTCAGACGTGTGC

sequencing primers	i5-NEX2p	AATGATACGGCGACCACCGAGATCTACACGTCTCGTGGGCTCG GAGATG
	i7-TXX-NEX2p	CAAGCAGAAGACGGCATACGAGAT[i7 index]GTCTCGTGGGCTCGGAGATG
	i7-T71-NEX2p	CAAGCAGAAGACGGCATACGAGATaactaggcgcGTCTCGTGGGC TCGGAGATG
	i7-T72-NEX2p	CAAGCAGAAGACGGCATACGAGATtcgctaagcaGTCTCGTGGGC TCGGAGATG
	i7-T73-NEX2p	CAAGCAGAAGACGGCATACGAGATtatatactaaGTCTCGTGGGC TCGGAGATG
	i7-T74-NEX2p	CAAGCAGAAGACGGCATACGAGATacttgctagaGTCTCGTGGGC TCGGAGATG
	i7-T75-NEX2p	CAAGCAGAAGACGGCATACGAGATaaccattggaGTCTCGTGGGC TCGGAGATG
	i7-T76-NEX2p	CAAGCAGAAGACGGCATACGAGATtcgcggttggGTCTCGTGGGC TCGGAGATG
	i7-T77-NEX2p	CAAGCAGAAGACGGCATACGAGATcgtagttaccGTCTCGTGGGC TCGGAGATG
	i7-T78-NEX2p	CAAGCAGAAGACGGCATACGAGATtccaatcatcGTCTCGTGGGC TCGGAGATG
	i7-T79-NEX2p	CAAGCAGAAGACGGCATACGAGATaatcgataatGTCTCGTGGGC TCGGAGATG
	i7-T80-NEX2p	CAAGCAGAAGACGGCATACGAGATccattatctaGTCTCGTGGGCT CGGAGATG

gBlock	MBP 1-370	gcaggctccaccatggggcatatgtccAAAATCGAAGAAGGTAAACTGG TAATCTGGATTAACGGCGATAAAGGCTATAACGGACTCGCTG AAGTCGGTAAGAAATTCGAGAAAGATACCGGAATTAAAGTCA CCGTTGAGCATCCGGATAAACTGGAAGAGAAATTCCCACAGGT TGCGGCAACTGGCGATGGCCCTGACATTATCTTCTGGGCACAC GACCGCTTTGGTGGCTACGCTCAATCTGGCCTGTTGGCTGAAA TCACCCCGGACAAAGCGTTCCAGGACAAGCTGTATCCGTTTAC CTGGGATGCCGTACGTTACAACGGCAAGCTGATTGCTTACCCG ATCGCTGTTGAAGCGTTATCGCTGATTTATAACAAAGATCTGC TGCCGAACCCGCCAAAAACCTGGGAAGAGATCCCGGCGCTGGA TAAAGAACTGAAAGCGAAAGGTAAGAGCGCGCTGATGTTCAA CCTGCAAGAACCGTACTTCACCTGGCCGCTGATTGCTGCTGAC GGGGGTTATGCGTTCAAGTATGAAAACGGCAAGTACGACATT AAAGACGTGGGCGTGGATAACGCTGGCGCGAAAGCGGGTCTG ACCTTCCTGGTTGACCTGATTAAAAACAAACACATGAATGCAG ACACCGATTACTCCATCGCAGAAGCTGCCTTTAATAAAGGCGA AACAGCGATGACCATCAACGGCCCGTGGGCATGGTCCAACATC GACACCAGCAAAGTGAATTATGGTGTAACGGTACTGCCGACCT TCAAGGGTCAACCATCCAAACCGTTCGTTGGCGTGCTGAGCGC AGGTATTAACGCCGCCAGTCCGAACAAAGAGCTGGCGAAAGA GTTCCTCGAAAACTATCTGCTGACTGATGAAGGTCTGGAAGCG GTTAATAAAGACAAACCGCTGGGTGCCGTAGCGCTGAAGTCTT ACGAGGAAGAGTTGGCGAAAGATCCACGTATTGCCGCCACCAT GGAAAACGCCCAGAAAGGTGAAATCATGCCGAACATCCCGCAG ATGTCCGCTTTCTGGTATGCCGTGCGTACTGCGGTGATCAACG CCGCCAGCGGTCGTCAGACTGTCGATGAAGCCCTGAAAGACGC GCAGACTCGTATCACCaagagcggtcaccatcaccatcaccattaagcggcc gcactcgagatatcta

Chapter 3. Monitoring Glycolytic Dynamics in Single Cells Using a Fluorescent Biosensor for Fructose 1,6-Bisphosphate

John N. Koberstein[1,4], Melissa L. Stewart[1,4], Chadwick Smith[1], Andrei I. Tarasov[2], Frances M. Ashcroft[3], Philip Stork[1], and Richard H. Goodman[1*]

[1]Vollum Institute, Oregon Health & Science University, Portland, OR 97239, USA

[2]School of Biomedical Sciences, Ulster University, Coleraine, UK

[3]Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, UK

[4]These authors contributed equally

3.1 Abstract

Cellular metabolism is regulated over space and time to ensure that energy production is efficiently matched with consumption. Fluorescent biosensors are useful tools for studying metabolism as they enable real-time detection of metabolite abundance with single-cell resolution. For monitoring glycolysis, the intermediate fructose 1,6-bisphosphate (FBP) is a particularly informative signal as its concentration is strongly correlated with flux through the whole pathway. Using GFP insertion into the ligand-binding domain of the *B. subtilis* transcriptional regulator CggR, we developed a fluorescent biosensor for FBP termed HYlight. We demonstrate HYlight can reliably report the real-time dynamics of glycolysis in living cells and tissues, driven by various metabolic or pharmacological perturbations, alone or in combination with other physiologically relevant signals. This biosensor therefore provides a new paradigm for studying glycolytic dynamics capable of uncovering cell-to-cell heterogeneity, subcellular localization, and temporal dynamics in cultured cells, primary tissues, and potentially multicellular *in vivo* systems.

3.1 Introduction

Glycolysis is an ancient metabolic pathway used by nearly all living organisms for the production of energy and biosynthetic precursors. Given the central role of this pathway, it is not surprising that the ten enzymatic steps responsible for glycolysis, elucidated by Embden, Meyerhof, and Parnas in the 1920s, comprise some of the best studied reactions in all of biochemistry. Perhaps equally important was the earlier demonstration in 1910 by Harden and Young that orthophosphoric acid and a component later determined to be NAD^+ were required for glycolysis. The former is needed for production of fructose 1,6-bisphosphate (FBP), the first intermediate of the glycolytic pathway to be identified[121], which in turn, has an emerged as an indicator of glycolytic flux overall[59,60,62,63,122].

FBP, produced by phosphofructokinase (PFK) from fructose-6-phosphate, sits at an important juncture in the glycolytic pathway. The enzymatic activity of PFK is highly regulated, being inhibited by the downstream products phosphoenolpyruvate (PEP), citrate, and ATP, and activated by fructose 2,6-bisphosphate (F2,6BP) and AMP. The final step of glycolysis, conversion of PEP to pyruvate, is catalyzed by pyruvate kinase, one isoform of which is activated by FBP. The kinetics and allosteric regulation of glycolytic enzymes collectively result in a correlation between the concentration of FBP and the overall flux through the glycolytic pathway [60,62,63]. This correlation has been detected across diverse lifeforms, including bacteria[62], yeast[63], and mammals[59], suggesting that FBP serves as a flux-signaling metabolite that is utilized by cells to sense and respond to their internal metabolic dynamics[61].

The role of FBP in signaling glycolytic flux has been demonstrated in the bacterium, *Bacillus subtilis*, where the transcription factor CggR (Central glycolytic

gene Repressor) is regulated by FBP binding. In the absence of FBP, CggR represses the expression of multiple glycolytic enzymes[60,62,123]. FBP thus serves as an internal readout of glycolytic flux induced by glucose availability. Yeast have been engineered that utilize CggR to regulate transcription of a fluorescent protein, thus enabling evaluation of glycolytic flux in single cells by fluorescence microscopy[63]. This method is limited, however, by slow readout kinetics and a lack of subcellular resolution. An alternative biosensing paradigm used for other ligands employs insertion of a circularly permuted fluorescent protein (cpFP) directly into a ligand-binding domain, such that ligand binding results in altered fluorescence intensity[2,26–28,49,124]. Such single fluorescent protein biosensors (SFPBs), while difficult to engineer, can provide fast kinetics and a spatially resolved measure of the concentration of specific analytes. Here, we describe the development and utilization of an SFPB, based on CggR, that can monitor real-time changes intracellular FBP levels in living cells. This biosensor was named HYlight after Harden and Young, who discovered FBP was a glycolytic intermediate.

We chose pancreatic beta cells as a model system to evaluate the utility of HYlight because of the essential role of glycolytic metabolism in the coupling of glucose stimulus to insulin secretion by these cells. In beta cells, the elevation of plasma glucose leads to the enhanced metabolism of the sugar, via glycolysis and oxidative phosphorylation. This resulting elevation in ATP subsequently closes ATP-sensitive K+ (K_{ATP}) channels, triggering plasma membrane electrical activity, Ca^{2+} influx into the cytosol, and insulin secretion[125–127]. Notably, beta cell glucose handling is adjusted to sense the substrate availability rather than respond to the energy demand, thereby enabling the glucose sensor role for the cell. In particular, glucose uptake is not rate-limiting for glycolytic flux, and glucose phosphorylation

is mediated by glucokinase (hexokinase IV), which exhibits a higher K_m than other hexokinases and lacks product inhibition. These adaptations result in glycolytic flux remaining sensitive to physiologically relevant concentrations of plasma glucose in the millimolar range. Additionally, the low levels of lactate production and transport produce a tight relationship between extracellular glucose and the rates of glycolysis and oxidative metabolism in beta cells[128]. We show here that beta cell FBP levels, monitored by HYlight, increase with glucose concentration in a dose-dependent manner and decrease upon the inhibition of glycolysis, providing compelling evidence that FBP can serve as an indicator of glycolytic flux in these cells. Beta cell lines exhibit metabolic and ionic oscillations, which were detected by HYlight coexpressed with R-GECO1 in single cells, uncovering the temporal relationship between FBP and Ca^{2+} dynamics. Finally, imaged simultaneously in hundreds of beta cells within islets of Langerhans, glucose-stimulated FBP changes exhibited a clear per-cell heterogeneity.

3.3 Results

Development of the FBP biosensor HYlight

When designing an SFPB, the most important decisions are the choices of ligand-binding domain (LBD) and fluorescent protein (FP), the site of FP insertion into the LBD, and the composition of the linkers connecting the two domains. For the FBP-binding domain, we used a truncation of *B. subtilis* CggR (residues 89-340, Fig. 1A) that excludes the N-terminal DNA-binding domain (residues 1-88) to prevent unnecessary DNA-binding by biosensor constructs[123]. We used a circularly permuted GFP (cpGFP) variant derived from a series of maltose biosensors[2,49]. High-throughput assays were used to characterize the brightness and dynamic

range for libraries containing random variation in the site of GFP insertion and composition of the linker amino acids (Fig. 1B).

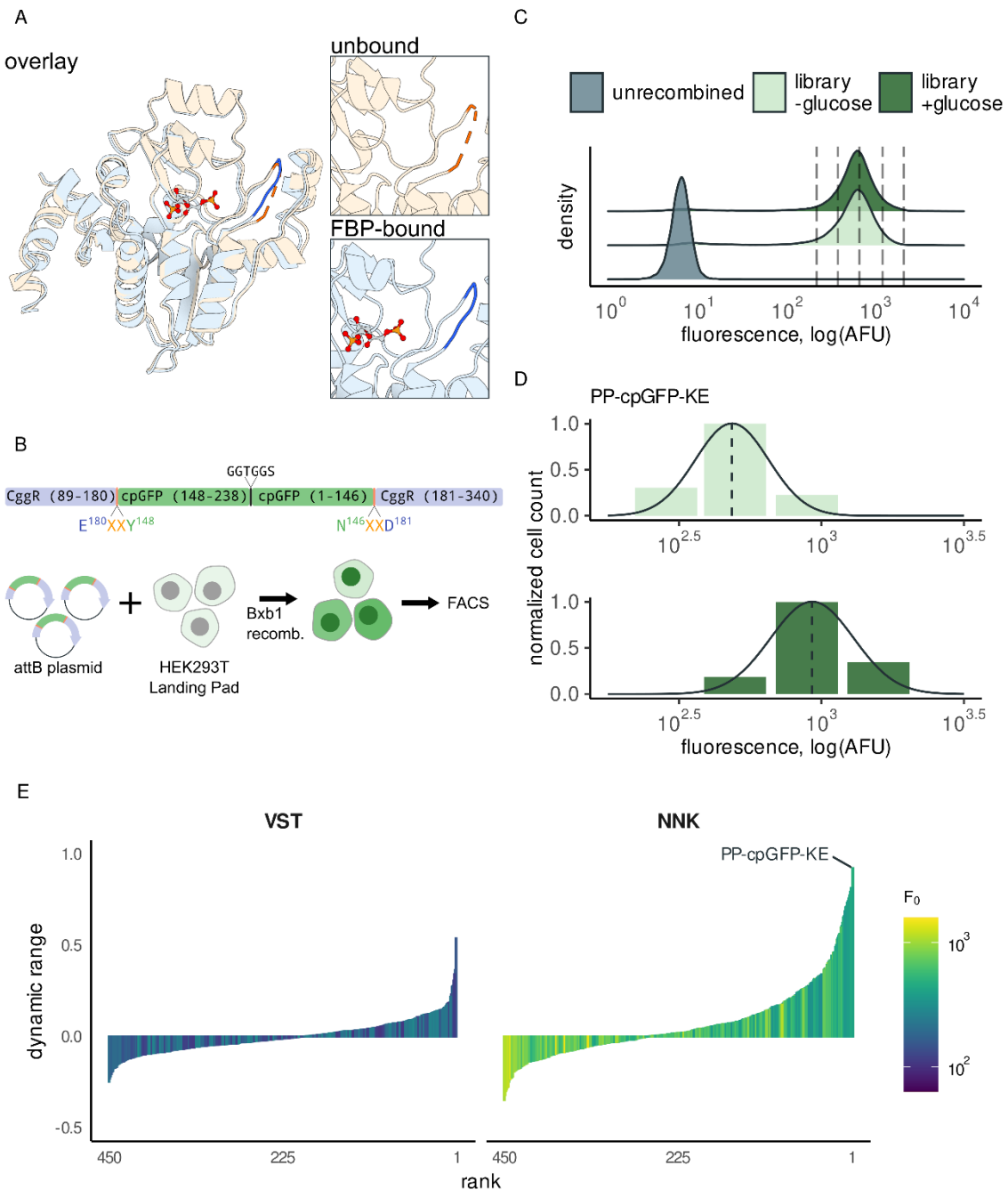

Figure 1. Discovery of a high-dynamic-range FBP biosensor by sort-seq assay. (A) Structural comparison of CggR in apo state (PDB 2OKG, orange) vs FBP bound (PDB 3BXF, blue) reveals a loop (residues 177-183) that undergoes a disorder-to-order transition upon binding FBP. **(B)** A library of linker variants was assayed for function in HEK293T Landing Pad cells. Circularly permuted GFP was inserted into CggR at residue 180 with two flanking linker amino acids on either side. Linker amino acids were encoded by the degenerate codon VST, which translates to a limited set of 8 amino acids, or by the fully degenerate codon NNK, which includes all 20 amino acids. The library was placed into an

attB plasmid that enables genomic recombination into the HEK293T Landing Pad genome by the Bxb1 recombinase. **(C)** Fluorescence distributions of the CggR-180-NNK library expressed in HEK293T cells exposed to 0mM or 25mM glucose. Dotted lines indicate the bins used for sort-seq. **(D)** The number of cells sorted into each bin indicated by bar height along with the maximum likelihood density estimates for the variant with the highest dynamic range, CggR-180-PP-cpGFP-KE. **(E)** Dynamic range (ΔF/F) estimates for all variants after screening libraries with linker residues substituted with amino acids encoded by a limited set using VST codons (left) or fully degenerate NNK codons (right)

To measure the FBP-induced changes in fluorescence for hundreds of sequence variants in parallel, we utilized sort-seq, a high-throughput functional assay combining fluorescence-activated cell sorting (FACS) and DNA sequencing[45,96,129]. To ensure that expression was limited to a single sequence variant per cell (a requirement of this assay) we employed the HEK293T Landing Pad cell line which enables genomic integration of the DNA library into a single Bxb1 recombination site (Fig. 1B)[48,130]. To characterize FBP induced changes in fluorescence, we performed sort-seq assays in the presence and absence of 25mM glucose, which produces high and low intracellular FBP concentrations, respectively. Recombined cells were sorted into four bins spanning the range of the observed library fluorescence intensity (Fig. 1C). Read counts for variants in each bin generated by DNA sequencing provided a view of the log-normal fluorescence distribution analogous to a histogram, from which the mean was inferred using a maximum likelihood estimator (Fig. 1D). Biosensor dynamic range ($\Delta F/F = (F_{ligand} - F_{min}) / F_{min}$) was then calculated as the relative difference between the estimated mean fluorescence in the glucose fed (F_{ligand}) and starved states (F_{min}) for each variant.

A key challenge of biosensor design is identifying a site in which cpGFP insertion is tolerated, allowing both domains to fold, while also potentially permitting allosteric coupling between ligand binding and fluorescence. Initially, multiple CggR

positions (156 out of 251 possible) were tested for permissibility of cpGFP insertion using a transposon-mediated domain-insertion cloning strategy (Fig. S1A). None of the insertion-site variants responded to changes in glucose ($|\Delta F/F| < 0.3$ for all tested insertion-site variants, Fig. S1C). Nonetheless, insertions into the loop comprising residues 177-183 (numbered according to the full-length protein) were relatively bright compared to the other tested sites ($F_{min} = 2.1 \pm 0.04$ log(AFU), $p < 0.005$ Mann–Whitney U test). Structures of the CggR ligand binding domain in the apo- and FBP-bound states indicated that this loop undergoes an FBP-dependent disordered-to-ordered transition (Fig. 1A). To minimize disruption of the FBP-binding site, we chose the variant CggR-180, which places cpGFP at the apex of the loop, for further optimization efforts.

Only a fraction of the possible linkers will produce strong allosteric coupling between a given cpFP and LBD. We designed a library containing cpGFP inserted between residues 180-181 of CggR with short linkers on either side consisting of two amino acids each (Fig. 1B). The linker amino acids were encoded by either the degenerate codons VST that translate to the amino acids A, G, P, R, S, T or NNK which included all 20 amino acids. In the case of the NNK library, the number of combinations (20^4 = 160,000) was much greater than what could readily be characterized. To reduce the number of combinations to an experimentally tractable number, 10,000 cells from the top 10% by brightness were sorted in the presence of glucose, expanded, and used as input for subsequent assays using this library.

After filtering out low abundance variants and those outside the expected range, we estimated the brightness and dynamic range for 900 linker variants between the two libraries (Fig. 1E). Although most variants did not display glucose-dependent alterations in fluorescence, both libraries contained rare variants that showed large

fluorescence changes. The fully degenerate NNK library produced more high-dynamic-range variants than the more limited VST library (Fig. 1E). Notably, biosensors with fluorescence increases (turn-on) and decreases (turn-off) upon addition of glucose were both identified. The variant with the largest dynamic range (ΔF/F = 0.92) contained the linker pairs Pro-Pro and Lys-Glu on the N and C-terminal ends of cpGFP, respectively. This variant, named HYlight, was used to explore FBP dynamics *in vitro* and in live pancreatic beta cells.

In vitro characterization of HYlight

When excited by 488 nm light, purified HYlight protein exhibited an increase in emission at 510 nm with increasing [FBP] (ΔF/F = 1.5, Fig. 2 A-C). An additional major excitation peak was discovered at 400 nm whose emission was slightly red-shifted (Fig. 2B). This latter signal decreased with increasing [FBP] (Fig. 2 A-C). The ratio of the emission values from the two excitation wavelengths provided a readout with increased sensitivity (ΔR/R = 3.0, Fig. 2D) and reduced variability resulting from differences in biosensor concentration (Fig. S2A) and pH (Fig. 2E and Fig. S2B). While HYlight utilized as an intensiometric biosensor with 488 nm excitation alone had a moderately high dynamic range, the unexpected discovery of a second major excitation peak measured at 405 nm substantially improved its ability to respond to changes in FBP specifically and sensitively.

While many dual excitation ratiometric biosensors have been developed, in most cases the second excitation peak was an unexpected outcome of attempts to optimize other biosensor properties through mutagenesis. Excitation at 400 nm is a characteristic of wtGFP caused by excited state proton transfer (ESPT) from the neutral state of the fluorophore to nearby residues [131]. The specific mutation

introduced into EGFP (S65T) that stabilizes the anionic state of the fluorophore responsible for 488 nm excitation[15] is included in the cpGFP present in HYlight. A second mutation, H148Y, is relatively uncommon amongst biosensors, present only in HYlight and the disaccharide biosensors from which the cpGFP sequence used in HYlight was obtained[30]. Due to its proximity to the fluorophore, we reasoned that Tyr148 might promote ESPT and 400 nm excitation by directly acting as the proton acceptor. Reverting this mutation back to the wildtype His148 indeed produced a biosensor with significantly diminished 400 nm excitation (Fig. S2D).

The apparent affinity of HYlight for FBP (11μM, Fig. 2D) is comparable to previously reported measurements by isothermal calorimetry using only the ligand-binding domain of CggR [123]. This binding is tighter than that reported using the full-length protein in thermal shift assays, however [60,63]. While binding to FBP has been well characterized, it has not been determined whether CggR also binds fructose 2,6-bisphosphate (F2,6BP), which is not produced in *B. subtilis* but is present in eukaryotic cells at much lower concentrations than FBP. F2,6BP affected HYlight fluorescence similarly to FBP but with slightly weaker affinity (Fig. S2C, Kd = 7.2μM for F2,6BP and 3.8 μM for FBP). Given the considerably lower concentration of F2,6BP, it is unlikely that it influences the observed fluorescence in cells, however.

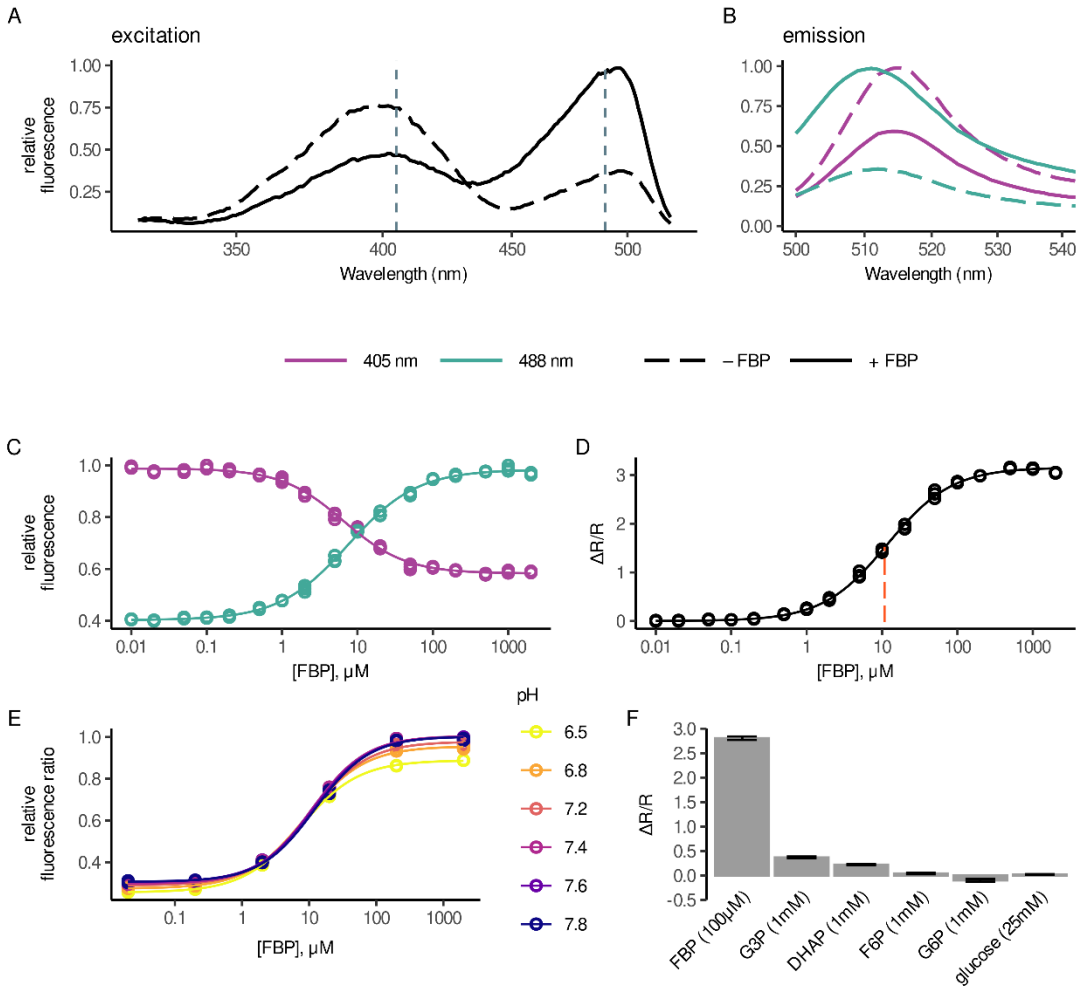

Figure 2. In vitro characterization of HYlight. (A) Excitation spectra from HYlight in the presence (solid) and absence (dashed) of 1mM FBP. **(B)** Emission spectra from HYlight in the presence (solid) and absence (dashed) of 1mM FBP. Purple lines indicate excitation set at 405 nm while cyan lines indicate excitation set at 488 nm. **(C)** Normalized HYlight emission induced by excitation at 405 nm (purple) or 488 nm (cyan) as a function of FBP concentration. **(D)** Relative change of the fluorescence ratio (ΔR/R) resulting from 488 nm and 405 nm excitation as a function of [FBP]. **(E)** Fluorescence ratio as a function of pH across FBP concentrations. **(F)** Relative HYlight fluorescence ratio for FBP compared to other glycolytic metabolites. G3P = glyceraldehyde 3-phosphate, DHAP = dihydroxyacetone phosphate, F6P = fructose 6-phosphate, G6P = glucose 6-phosphate. See also Figure S2.

To examine the specificity of HYlight, we tested other glycolytic metabolites, namely dihydroxyacetone phosphate (DHAP), glyceraldehyde 3-phosphate (G3P), fructose 6-phosphate (F6P), glucose 6-phosphate (G6P), and glucose. Minimal

changes in the HYlight fluorescence ratio were observed in the presence of supraphysiological concentrations of these other metabolites (Figure 2E). The absence of a fluorescence response to other ligands does not necessarily preclude binding, however. While F6P alone does not produce changes in HYlight fluorescence, it is capable of displacing FBP from the binding site with an apparent affinity of 321 µM (Figure S2E). In the presence of 300 µM F6P, the FBP dose-response curve for HYlight was modestly shifted to a higher concentration (K_d^{app} = 20 µM, Figure S2F) consistent with the determined affinity. The reported intracellular concentrations [132,133] are not greater than the affinity for F6P suggesting that competition will not dramatically alter the affinity of HYlight for FBP when expressed in cells.

While the fluorescence ratio of HYlight is largely insensitive to pH, this results from similar changes in fluorescence intensity for both 405 and 488 nm excitation across the pH range tested. In cases where dual excitation is not feasible, single wavelength measurements should be complemented by experimental controls to determine the extent of pH-dependent changes. To address this issue, we sought to generate a variant of HYlight that is incapable of binding FBP. We hypothesized that replacing the CggR residue Thr152 with Glu would result in the charged side chain of Glu occupying the region of the binding site normally occupied by the 6-phosphate. This T152E variant exhibited no changes in fluorescence ratio across FBP concentrations (Fig. S2B). In addition, the T152E variant exhibited similar sensitivity to pH changes as HYlight. This “binding dead” version of HYlight was used as an experimental control to detect possible artifactual changes in fluorescence caused by factors other than FBP.

HYlight can detect cell type specific relationships between glucose and FBP

We probed the relationship between the extracellular glucose and intracellular FBP levels by measuring HYlight fluorescence in living HEK293 cells and insulin-secreting MIN6 beta cells, using flow cytometry. Fluorescence was monitored using 405 and 488 nm excitation and the relative change ($\Delta R/R_0$) in fluorescence ratio ($R=F_{488}/F_{405}$) was calculated relative to the no glucose condition (R_0). In both cell types, the median $\Delta R/R_0$ exhibited a glucose-dependent increase, but the concentration of glucose that elicited a half-maximal $\Delta R/R_0$ ($R_{0.5}$) differed between cell types (Fig. 3 A,B). In MIN6 cells, $R_{0.5}$ was achieved at 6.4 ± 1.6 mM glucose (Fig. 3B) which is close to the glucose concentration reported for half-maximal activation of glucokinase [134]. In contrast, the $R_{0.5}$ occurred at much lower glucose concentration in HEK293 cells (1.0 ± 0.2 mM glucose, $p < 0.05$ vs MIN6). This is consistent with a key role for hexokinase (HKI $K_{0.5}$ = 41 µM, HKII $K_{0.5}$ = 340 mM) and a lack of glucokinase in HEK293 cells [135,136]. The glucokinase activator dorzagliatin (10 µM) significantly lowered the concentration eliciting half-maximal activation ($R_{0.5}$ = 1.7 ± 0.4 mM glucose, $p < 0.05$ vs untreated cells) in MIN6 cells.

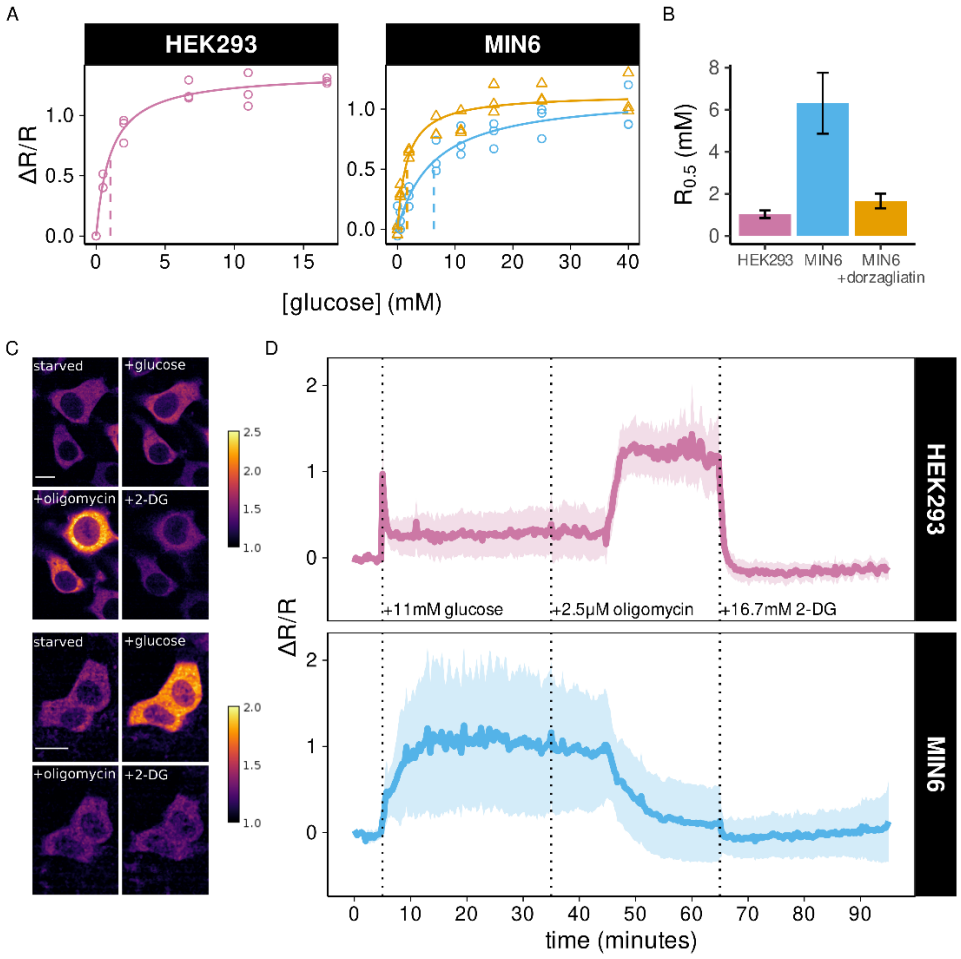

Figure 3. HYlight imaging reveals differences in metabolic phenotype between HEK293 and MIN6 cells. (A) HYlight fluorescence ratio as a function of glucose concentration follows Michaelis-Menten kinetics. Differences in the magnitude of maximal change and EC50 are observed between HEK293 (purple), MIN6 (blue) and MIN6 cells treated with dorzagliatin (orange). Each point represents the median fluorescence ratio across cells measured by flow cytometry relative to cells incubated in 0 mM glucose. Lines indicate fitted Michaelis-Menten equation with R0.5 shown as a vertical dashed line. **(B)** Comparison of R0.5 estimates ± standard error for HEK293 (1.0 ± 0.2 mM), MIN6 (6.7 ± 1.6 mM), and MIN6 cells treated with 10 μM dorzagliatin (1.7 ± 0.4 mM). **(C)** Example of fluorescence ratio images in HEK293 cells (top) and MIN6 cells (bottom) after 1 hour of glucose starvation and following addition of 11 mM glucose, 2.5 μM oligomycin, and 16.7 mM 2-deoxyglucose (2-DG). Scale bar, 10 μm. **(D)** Quantification of the change in fluorescence ratio (ΔR/R) for HEK293 cells (top, n=67 cells over 3 separate experiments) and MIN6 cells (bottom, n=222 cells over 3 separate experiments) following the metabolic perturbations shown in (C). ΔR/R was normalized to the glucose starved state at the beginning of each experiment. Solid line represents the mean while shaded ribbon represents the mean ± standard deviation.

HYlight can differentiate between cell types based on response to metabolic perturbations

We next examined the use of HYlight for detecting cell type-specific differences in FBP dynamics in HEK293 and MIN6 cells monitored by time-lapse confocal microscopy. Cells were starved of glucose for 1 hour prior to imaging, followed by the sequential addition of 11mM glucose, 2.5 µM oligomycin and 16.7 mM 2-deoxyglucose (2-DG) to mimic a common Seahorse XF protocol referred to as the Glycolytic Stress Test. Fluorescence was monitored using 405 and 488 nm excitation and the change in fluorescence ratio ($\Delta R/R_0$) was calculated relative to the ratio at the start of experiment (R_0). In response to 11mM glucose, HEK293 cells exhibited a transient spike in fluorescence ratio, which rapidly declined to a steady plateau ($\Delta R/R = 0.41 \pm 0.20$, Fig. 3C-D), possibly due to ATP feedback inhibition of PFK1. The glucose-induced increase in fluorescence ratio exhibited a slower increase but plateaued at a higher level, in MIN6 cells, ($\Delta R/R = 0.83 \pm 0.19$, $p<0.05$ vs HEK293 cells, Fig. 3E-F). Oligomycin, an inhibitor of the mitochondrial ATP-synthase, produced a further, sustained increase in fluorescence ratio in HEK293 cells, presumably reflecting an increased anaerobic glycolytic capacity of HEK293 cells when oxidative phosphorylation is inhibited. In contrast, the glucose-induced increase in fluorescence ratio was largely reversed by oligomycin in MIN6 cells. This likely reflects the inability of beta cells to metabolize glucose to lactate[137,138], which results in insufficient glycolytic ATP for FBP production by PFK. Finally, 2-deoxyglucose (2-DG), reduced glycolysis in HEK293 cells, as evidenced by a rapid decline in fluorescence ratio back to the baseline (Fig. 3E-F), while only minimally decreasing the already low FBP levels in MIN6 cells.

Multiplexed imaging with HYlight resolves temporal relationship between oscillating signals

In addition to a sustained increase in FBP upon exposure to 11 mM glucose, large amplitude oscillations in FBP were observed in MIN6 cells, consistent with reports of a glycolytic oscillator in beta cell lines[139]. The dynamics of beta cell glycolysis under constant external conditions presents an opportunity to investigate cell-intrinsic aspects of glycolytic regulation using HYlight. Beta cell lines are also known to exhibit oscillations in plasma membrane electrical potential and cytosolic Ca^{2+}, but the temporal relationship between ionic and glycolytic oscillations is unclear. To demonstrate the use of HYlight for multiplexed imaging, we next examined the relationship between FBP and Ca^{2+} dynamics in individual MIN6 cells by co-expressing HYlight and a spectrally compatible Ca^{2+} sensor, R-GECO1[140]. As expected, the initial increase in FBP stimulated by glucose preceded the increase in Ca^{2+} (Fig. 4A-B). The glucose-induced oscillations of FBP and Ca^{2+} in individual cells were in phase, with a periodicity of approximately 60 seconds (Fig. 4C). Cross-correlation analysis of the two signals averaged over all cells indicated the two signals were maximally correlated with little (15s) or no time offset (Fig. 4D).

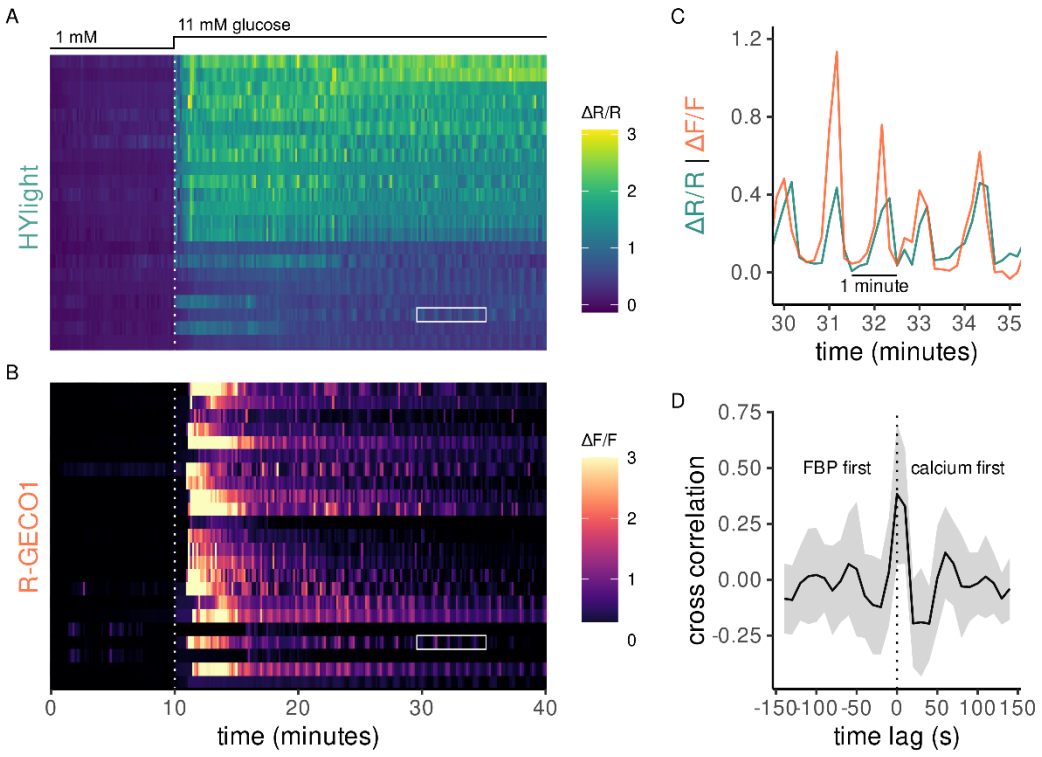

Figure 4. Multiplexed fluorescence imaging of HYlight and R-GECO1 uncovers temporal relationships between FBP and Ca^{2+} in MIN6 cells. (A) HYlight fluorescence ratio measured in MIN6 cells following an increase from 1 to 11 mM glucose. **(B)** R-GECO1 fluorescence measured in the same cells as (A) revealed a delayed increase in Ca^{2+} relative to FBP. **(C)** The HYlight fluorescence ratio and R-GECO1 fluorescence for a single cell indicated by white box in (A) and (B). Oscillations in both signals occur synchronously. **(D)** The cross correlation of R-GECO1 and HYlight signal following glucose stimulation (t=20-40 minutes) averaged over all cells with shaded ribbon representing the mean ± the standard deviation. The maximum cross correlation occurs in simultaneous frames or at a 1 frame (15s) lag with the R-GECO1 signal preceding HYlight.

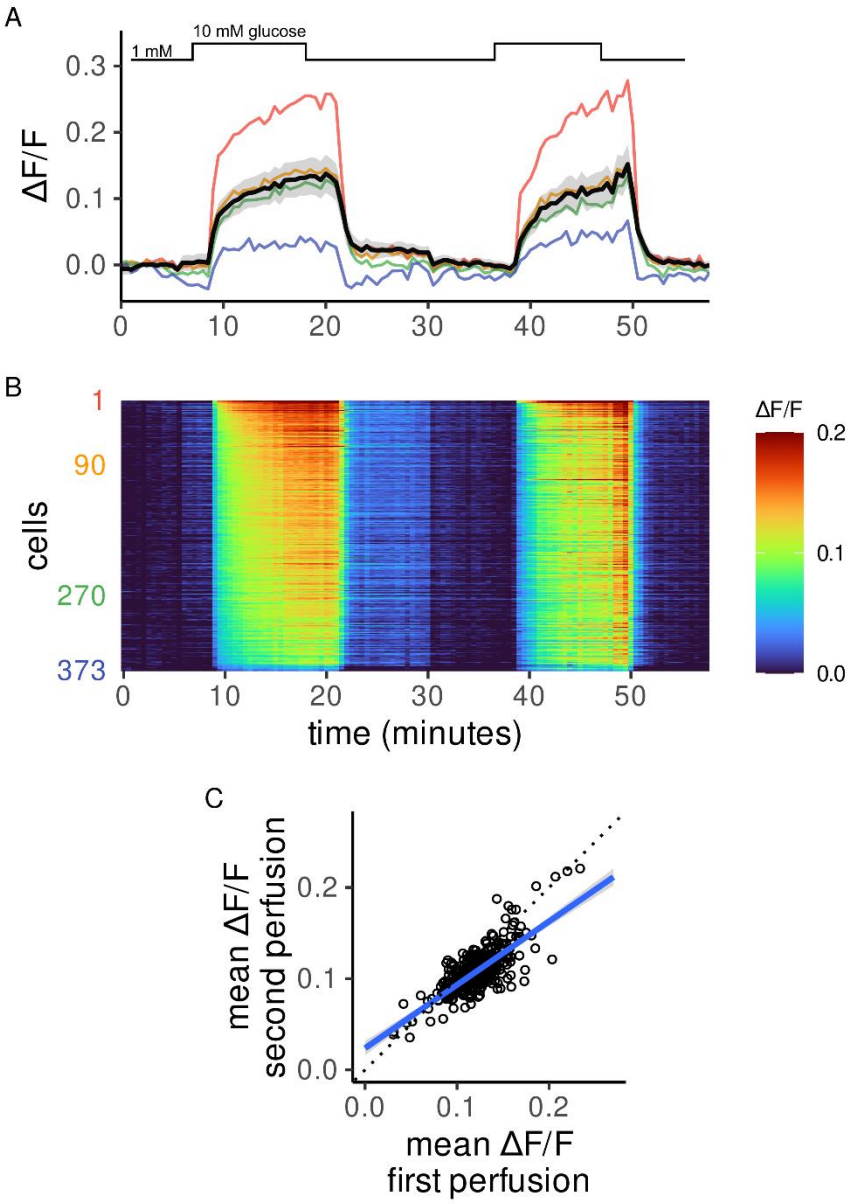

Figure 5. HYlight uncovers heterogeneity among primary beta cells in response to glucose. (A) Beta cells imaged with repeated cycling between 1 mM and 10 mM glucose. The solid black line represents the mean and the grey shaded region represents the mean ± standard deviation. Four traces from example cells are shown as colored lines. **(B)** Heatmap of fluorescence changes where each row represents a single cell quantified over time. Cells are ordered from highest (top) to lowest (bottom) mean response. Example cells from (A) are indicated by numerical ranks on Y-axis. **(C)** The mean change in fluorescence ratio during the first and second exposure to 10 mM glucose is heterogeneous across cells but consistent within individual cells over time (R=0.75). The blue line represents a linear regression fit, while dotted black indicates the unity line.

HYlight reveals differences in glycolytic dynamics at the single cell level in intact islets

A key advantage of fluorescent biosensor imaging is the high spatial resolution which permits evaluation of the metabolic properties of single cells. In MIN6 beta cells the FBP response to elevated glucose was highly heterogeneous (Fig. 3F). Glucose-induced increases in FBP were also readily detected in individual beta cells within isolated mouse islets expressing HYlight, delivered via a beta cell specific AAV vector. As in MIN6 cells, the individual FBP responses of primary islet beta cells were heterogeneous (Fig. 5A-B). To determine whether the variability in FBP response in individual cells persisted over multiple stimulations we treated cells with cycles of low (1 mM) and high (11 mM) glucose. The per-cell magnitude of HYlight increase was highly correlated between sequential rounds of glucose stimulation (R = 0.75, Fig. 5C).

3.4 Discussion

We developed HYlight to monitor FBP dynamics with cellular resolution, an effort motivated by a growing appreciation for the role of this metabolite in reflecting the state of the glycolytic pathway. CggR, a transcription factor that enables bacteria to sense and respond to changes in glycolytic flux, was utilized to engineer an SFPB via high-throughput functional assays in live cells. Inserting cpGFP into a loop of CggR that undergoes a structural transition with ligand binding produced a relatively bright construct, but one that lacked responsiveness to FBP. High-throughput characterization of the linker amino acids connecting cpGFP to CggR revealed multiple high-dynamic-range FBP biosensors. Given that only 900 of the 160,000 possible linker amino acid combinations were tested, it is likely that linker

variants with improved performance remain undiscovered. The many pairs of biosensor sequence and associated dynamic range collected in this study might prove useful as training data for emerging machine learning methods to predict the function for untested sequences.

The ratiometric signal and high-dynamic-range of HYlight make it an especially useful tool for monitoring metabolism in live cells. The high-performance of HYlight combined with the large changes in FBP[141] that track with alterations of flux make this sensor well suited for assaying glycolytic metabolism. HYlight thus complements existing fluorescent biosensors relevant to glycolysis covering the major pathway inputs (glucose[142]), outputs (pyruvate[89]/lactate[39,143]) and cofactors (NADH:NAD^{+}[144] and ATP:ADP[124]). In addition to its important role within glycolysis, interactions between FBP and cellular signaling pathways that broadly regulate cellular metabolism are becoming increasingly apparent in eukaryotes. For example, in high-glucose conditions aldolase bound to FBP contributes to the activation of mTORC1[66], and blocks activation of AMPK [65]. These interactions between glycolytic metabolism and cellular signaling pathways mediated by FBP present additional cases where HYlight may be useful.

Pancreatic beta cells whose glycolytic rate depends on glucose supply provided an ideal system to test HYlight and an opportunity to detect FBP dynamics in relation to normal physiology. HYlight detected changes in FBP across a range of glucose concentrations in MIN6 cells and isolated islets. Additionally, perturbations thought to increase glycolytic flux, either by increasing glucokinase activity in MIN6 cells or inhibiting mitochondrial function in HEK293 cells were associated with an increased HYlight fluorescence ratio providing further evidence that FBP levels reflect glycolytic flux. Different beta cells within the same islet exhibited

heterogeneous responses to glucose that were maintained over time, indicating a cell intrinsic source of variability. Despite the high variability in FBP levels among individual HEK293 and MIN6 cells, the median HYlight fluorescence ratio in response to different concentrations of glucose, revealed by flow cytometry, conformed to the known properties of hexokinase and glucokinase. The detection of fast (~1 minute period) oscillations in FBP in MIN6 cells, in synchrony with those of Ca^{2+} likely reflects the hierarchy of metabolic and electrical signaling in the beta cell. The concurrence of FBP and Ca^{2+} peaks that we observed contrasts with prior studies using PKAR, a FRET sensor for PKM2 tetramerization, which showed, also in MIN6 cells, that FBP oscillated out of phase with Ca^{2+} oscillations[145]. Based on the regulation of PKM2 tetramerization by FBP binding, this FRET change was proposed to reflect FBP dynamics, a prediction worth reconsidering in light of our findings using HYlight. It is noteworthy that the FBP oscillations were not evident in beta cells within intact islets, where these cells are electrically coupled and Ca^{2+} waves have been shown to propagate[146].

Beyond the example of beta cells presented in this work, the ubiquity and importance of glycolysis provides additional potential uses for HYlight across a range of cell types, tissues, and organisms. Precise measurement of glycolytic flux requires metabolic labeling followed by destructive analytical chemistry techniques [69]. To monitor metabolic flux in live cells, instruments that measure the rate of oxygen consumption and extracellular acidification have been developed, most notably the Seahorse XF Analyzer[70]. However, these methods are limited to measuring bulk properties of cell populations, thereby obscuring the heterogeneity among individual cells. Additionally, Seahorse assays do not lend themselves to perfusion experiments, making it difficult to assess effects of removing an agent.

While HYlight measurement of FBP does not provide absolute quantification of flux, our qualitative findings from flow cytometry and fluorescence microscopy reflect the known metabolic properties elucidated through bulk assays, with enhanced spatiotemporal resolution. These advantages, combined with the relative affordability and flexibility of a fluorescent biosensor, suggest that HYlight could be broadly useful for metabolic assays in a wide variety of biological contexts.

In cancer biology, a field where measures of glycolytic metabolism are commonly performed, it has become apparent that cellular heterogeneity is a major contributor to disease progression and therapeutic response. The preference of cancer cells for aerobic fermentation (Warburg effect) varies from cell to cell in solid tumors, depending on the level of hypoxia and interactions with the tumor microenvironment[147,148]. Glycolytic heterogeneity is also relevant in immunology, where acute elevations in glycolysis can characterize cell type-specific responses of leukocytes to specific challenges[149,150]. Imaging of HYlight-expressing cells using a fluorescence microscope can detect metabolic characteristics of individual cells. Additionally, subcellular compartmentalization of glycolysis, well-known in trypanosomes, where glycolytic enzymes are segregated into discrete organelles, and yeast, where glycolytic enzymes are enriched in liquid condensates known as G bodies, is also emerging as an important response of neurons to activity-dependent energy demands[86,151]. We anticipate that HYlight will be useful in characterizing glycolytic regulation at the subcellular level in addition to the examples provided in the current work.

3.5 Methods

Biosensor library cloning

pHO_pTEFmut7_CggR_R250A_ble was a gift from Matthias Heinemann (Addgene plasmid # 124585 ; http://n2t.net/addgene:124585; RRID:Addgene_124585). The coding sequence for CggR (89-340) was amplified from pHO_pTEFmut7_CggR_R250A_ble using primers CggR-p7-GG-F and CggR-p7-GG-R to add BsaI restriction sites and compatible overhangs for Golden Gate assembly into the plasmid HC_Kan_RFP-p7. All Golden Gate reactions were carried out using 40 fmol of vector (HC_Kan_RFP-p7), 40 fmol of purified insert (CggR amplicon), 10 units of BsaI (NEB), 800 units of T4 DNA ligase (NEB), and 1 × T4 DNA Ligase Reaction Buffer (NEB) in a total volume of 20 µL. The reaction was incubated 2 min at 37°C and 5 min at 16°C for 50 cycles and followed by 20 min at 60°C and 20 min at 80°C. The R250A mutation in the resulting plasmid HC_Kan_CggR-R250A-p7 was reverted to wildtype using the primers CggR-A250R-Q5-F and CggR-A250R-Q5-R and Q5 site-directed mutagenesis.

A holding plasmid for cloning CggR-180 linker libraries was generated by Gibson Assembly. HC_Kan_CggR-EBD-WT was linearized at amino acid 180 by PCR using primers CggR-180-Dest-GA-F2 and CggR-180-Dest-GA-R2 which added homologous ends to a lacZ cassette amplified from pATT-Dest with primers CggR-180-Dest-GA-F1 and CggR-180-Dest-GA-R1. pATT-Dest was a gift from David Savage (Addgene plasmid # 79770 ; http://n2t.net/addgene:79770 ; RRID:Addgene_79770). The resulting plasmid HC_Kan_CggR-180-Dest was used to clone linker libraries by Golden Gate Assembly using the BsaI sites added with lacZ cassette. Golden Gate compatible cpGFP was amplified from pTKEI-Mal-B2

with primers CggR-cpGFP-VSTx4-GG-F and CggR-cpGFP-VSTx4-GG-R for the VST library or CggR-cpGFP-NNKx4-GG-F and CggR-cpGFP-NNKx4-GG-R for the NNK library. pTKEI-Mal-B2 was a gift from David Savage (Addgene plasmid # 79756 ; http://n2t.net/addgene:79756 ; RRID:Addgene_79756).

Transposon-mediated domain-insertion profiling (DIP) was used to construct a library of cpGFP insertion into CggR [49]. The Mu-BsaI transposon was digested from pUCKanR-Mu-BsaI (Addgene plasmid # 79769) with BglII and HindIII in Buffer 3.1 (NEB) at 37°C overnight and purified by gel extraction (NucleoSpin Gel and PCR Clean-up). pUCKanR-Mu-BsaI was a gift from David Savage (Addgene plasmid # 79769 ; http://n2t.net/addgene:79769 ; RRID:Addgene_79769). Transposition was performed using 100 ng of purified MuA-BsaI transposon, pATT-CggR plasmid DNA at a 1:2 molar ratio relative to transposon, 4 µL of 5× MuA reaction buffer, and 1 µL of 0.22 µg/µL MuA transposase (Thermo Fisher) in a total volume of 20 µL. Reactions were incubated at 30°C for 18 h, followed by 75°C for 10 min. Reactions were cleaned up using the DNA Clean & Concentrator-5 Kit (Zymo Research Corp.) and eluted in 6 µL of water. Transformation was performed using 2 µL of reaction in 25 µL of E. Cloni 10G ELITE cells (Lucigen) in 1.0 mm Bio-Rad cuvettes using a Gene Pulser Xcell Electroporation System (settings: 10 µF, 600 Ω, 1.8 kV). Cells were immediately resuspended in 975 µL Recovery Media and shaken at 250 rpm for 1 h at 37°C. A 10 µL aliquot of transformed cells was plated on carbenicillin (100 µg/mL) and chloramphenicol (25 µg/ mL) to select for the presence of pATT plasmid backbone and transposon insertion to assess library coverage. The remaining transformed cells were pelleted and resuspended in 50 mL of LB with 100 µg/mL carbenicillin and 25 µg/mL chloramphenicol. Cultures were grown at 250 rpm, at

37°C overnight, followed by plasmid DNA purification using a HiSpeed Plasmid Midi Kit (Qiagen).

The CggR-DIP and CggR-180 linker libraries were moved to a recombination plasmid, EMMA-attB-Dest, by Golden Gate Assembly as described previously using Esp3I in place of BsaI. The domain-insertion library was amplified using primers CggR-BsmbI-GG-F and CggR-BsmbI-GG-R to add compatible overhangs. The PCR product was purified by gel extraction (NucleoSpin Gel and PCR Clean-up) and used as the insert for the Golden Gate Assembly reaction. Reactions were cleaned up using the DNA Clean & Concentrator-5 Kit and eluted in 6 µL of water. Bacterial transformation was performed using 25 µL of E. Cloni 10G ELITE Chemically Competent cells combined with 2 µL of purified Golden Gate reaction and heat shock for 45 seconds in a 42°C water bath. A small aliquot (5 µL) of the outgrowth media was plated on LB agar with carbenicillin to assess transformation efficiency. The remaining outgrowth media was diluted into 100 mL LB supplemented with 50 µg/mL carbenicillin and grown overnight at 37°C. Plasmid DNA for transfection was purified using a Qiagen Plasmid Maxi Kit.

HEK293T Landing Pad Transfection

The HEK293T Lentiviral Landing Pad (LLP-iCasp9-Blast) cell line [48,130] obtained from Kenneth Matreyek and Doug Fowler was used to express libraries for sort-seq experiments. Cells were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 25 mM glucose and 4 mM L-glutamine (Gibco 11965092) supplemented with 10% fetal bovine serum (FBS). BFP expression was induced with 2 µg/mL doxycycline (Sigma-Aldrich), which was removed 1-2 days prior to transfection. Recombination was achieved by transfecting cells with a total of 3 µg of plasmid

DNA (1.5 µg each of attB recombination plasmid and pCAG-NLS-HA-Bxb1) and 6 µL of FuGENE 6 (Promega) in 300 µL of Opti-MEM (Gibco). pCAG-NLS-HA-Bxb1 was a gift from Pawel Pelczar (Addgene plasmid # 51271 ; http://n2t.net/addgene:51271 ; RRID:Addgene_51271). The DNA transfection reagent mixture was incubated for 15 minutes at room temperature before adding to a six-well plate containing 1×10^6 freshly seeded cells per well. Cells were incubated with the transfection mixture for 24-48 hours before expanding each well to a 10 cm plate. After 1 day of growth to reach approximately 90% confluency in the 10 cm plate, 2 µg/mL of doxycycline was added to induce expression. Selection for recombined cells was accomplished by adding 1 nM AP1903 approximately 24 hours after adding doxycycline. Recombined cells were grown for an additional 7 days after induction before FACS was performed.

Fluorescence activated cell sorting

Three sort-seq experiments were conducted, characterizing a single DIP library and two linker libraries, with minor experimental differences. Cells at 50% confluency in a 15 cm plate were detached by trypsinization, pelleted, and resuspended in 2 mL of DMEM supplemented with 50 µg/mL of gentamicin. Sorts were performed using a BD Influx instrument equipped with a 488 nm laser for excitation and a 530/40 nm emission filter for GFP measurements and a 405 nm laser with 460/50 filter for mTagBFP measurements. Cells transfected with Bxb1 recombinase but no attB plasmid were used to adjust instrument voltages and establish a baseline of autofluorescence in the green channel and the presence of mTagBFP expression. Cells transfected with an attB plasmid but without expression of Bxb1 recombinase were used to determine the level of green

fluorescence derived from plasmid expression as opposed to genomic integration and doxycycline induced expression. Cells transfected with the biosensor library plasmid (CggR-DIP or CggR-180-VST/NNNK) and Bxb1 recombinase were sorted to collect cells positive for GFP and negative for mTagBFP fluorescence to enrich for recombined cells.

The CggR-DIP library was initially sorted for fluorescence above the background to enrich for productive insertions with a total of 70,000 cells collected. Cells were collected in a 15 cm tube containing 5 mL of DMEM supplemented with 10% FBS and 50 µg/ mL gentamicin. Sorted cells were pelleted, resuspended in 800 µL of media, and plated in a 12-well plate with 2 µg/mL doxycycline. Cells were expanded over 8 days to reach 50% confluence in two 15 cm plates before the second round of sorting. The CggR-180-NNK was initially sorted to bottleneck the number of unique combinations, collecting a total of 2,000 cells above the 95th percentile brightness. Cells were collected in a single well of a 96-well plate containing 100 µL of DMEM supplemented with 10% FBS, 50 µg/ mL gentamicin and 2 µg/mL doxycycline. Cells were expanded over 14 days to reach 50% confluence in two 15 cm plates before the second round of sorting.

For all three sort-seq experiments, the two samples for the second round of FACS were detached from 15 cm plates by trypsinization, pelleted, and resuspended in 2 mL of DMEM supplemented with 50 µg/mL gentamicin and either 10 mM pyruvate or an equal volume of water. Cells transfected with Bxb1 recombinase but no attB plasmid were used to adjust instrument voltages and establish a baseline of autofluorescence in the green channel. Four equal width gates on the log scale were set to span the range of log(AFU) covered by the distribution of each library. The ± glucose samples were sorted using the same

gates for a duration of 1.5 h each. The cells for each bin were collected in 5 mL tubes containing 1 mL of DMEM supplemented with 10% FBS and 50 µg/mL gentamicin. The -glucose samples for the CggR-DIP and CggR-180-VST libraries were supplemented with 10 mM pyruvate assuming an additional carbon source would be beneficial, while CggR-180-NNK did not include added pyruvate to the -glucose sort-seq sample. Cells were sorted into each bin at approximately proportional amounts to the relative density of cells in each bin (Table 1). The collected cells for each bin were individually pelleted, resuspended, and plated in either a 24-well (<200,000 cells), 12-well (>200,000 and <500,000 cells), six-well plate (>400,000 and <2M cells) or 10 cm dish (>2M cells). All samples were expanded to 50% confluent in a 10 cm plate before harvesting by trypsinization and centrifugation. Cell pellets were washed with PBS and stored at −20°C. Genomic DNA was extracted from cell pellets containing approximately 5 M cells using a Qiagen DNeasy Blood & Tissue Kit.

Domain-insertion library DNA Sequencing

The ORF to be sequenced was PCR amplified from the Landing Pad genomic integration site using primer specific to the genome LP-Ptet-F and one specific to the integrated plasmid sequence LP-STOP-R (the sequences for all primers used for DNA sequencing can be found in Table S1). Each 50 µL reactions was prepared with a final concentration of 10 ng/µL of genomic DNA, 0.25 µM forward and reverse primer, 1 × SeqAmp PCR buffer, 1 × SeqAmp Polymerase (Clontech), and 1 × SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions: [94°C 60s, (98°C 10s, 55°C 15s, 68°C 60s, plate read) × 29 cycles]. The number of cycles was determined such that reactions were in the exponential phase

of amplification upon completion of the program. Reactions were cleaned with a NucleoSpin Gel and PCR Clean-up kit and eluted in 15 µL of water.

Amplicons were fragmented and tagged (tagmented) in a 20 µL reaction containing 2.5 ng/µL amplicon, 1 × TD buffer, and 0.5 µL TDE1 enzyme from the Nextera DNA Sample Prep Kit (Illumina). Tagmentation reactions were cleaned up using a NucleoSpin column and eluted in 15 µL of water. Tagmented DNA was amplified with primer i5-Nex2p and a unique indexed primer per sample (i7- TXX-NEX2p). The 25 µL reactions were prepared containing 1 µL of tagmented DNA, 0.5 µM forward and reverse primer, 1 × KAPA HiFi Hotstart ReadyMix, and 1 × SYBR Green. Amplification was monitored by qPCR with cycling conditions: [72°C 3 min, 95°C 20 s, (98°C 20s, 52°C 15 s, 72°C 30 s, plate read, 72°C 8s) × 13 cycles]. Reactions were removed during the exponential phase of amplification.

PCR products were run on a 1.5% agarose gel to visualize distribution of tagmented DNA size and to estimate relative concentrations using FIJI gel analysis. Indexed samples were pooled while normalizing for relative concentration. Pooled products were run on a 1.5% agarose gel, cutting out a band at approximately 500bp, which was then purified using the NucleoSpin Gel and PCR Clean-up column. The concentration of the pooled library was quantified using a Qubit fluorometer, and size distribution was assessed using a HS DNA chip on the Bioanalyzer 2100 instrument (Agilent). The library was sequenced using 2 × 75bp paired-end reads on an Illumina MiSeq (v3 Reagent kit).

Linker library DNA Sequencing

The entire ORF was initially PCR amplified from the Landing Pad genomic integration site using a forward primer specific to the genome, LP-Ptet-F, and a

reverse primer specific to the integrated plasmid sequence, LP-STOP-R, to limit amplification of any residual plasmid DNA. Each 50 µL reactions was prepared with a final concentration of 20 ng/µL of genomic DNA, 0.25 µM forward and reverse primer, 1 × SeqAmp PCR buffer, 1 × SeqAmp Polymerase (Clontech), and 1 × SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions: [94°C 60s, (98°C 10s, 55°C 15s, 68°C 60s, plate read) × 14 cycles]. Reactions were cleaned with a NucleoSpin Gel and PCR Clean-up kit and eluted in 15 µL of water.

The linker regions flanking cpGFP were PCR amplified in a second round using primers CggR-180-seq-F2 and CggR-180-seq-R2. Two replicate 50 µL reactions were prepared for each sample with a final concentration of 5 ng/µL genomic DNA, 0.25 µM forward and reverse primer, 1 × SeqAmp PCR buffer, 1 × SeqAmp Polymerase (Clontech), and 1 × SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions: [94°C 60s, (98°C 10s, 55°C 15s, 68°C 60s, plate read) × 10 cycles]. The number of cycles was determined such that reactions were in the exponential phase of amplification upon completion of the program. Replicate reactions were pooled and cleaned with a NucleoSpin column and eluted in 15 µL of elution buffer (5 mM Tris/HCl, pH 8.5).

Second-round PCR products were amplified a third time with primer i5-IPE2p and a unique indexed primer per sample (i7- iPE2p-XX). Then, 25 µL reactions were prepared containing 1 µL of round 2 DNA, 0.5 µM forward and reverse primer, 1 × KAPA HiFi Hotstart Readymix (KHF), and 1 × SYBR Green. Amplification was monitored by qPCR with cycling conditions: [95°C 3 min, (98°C20 s, 60°C15s, 72°C 30 s, plate read, 72°C8s) × 8 cycles]. Reactions were removed during the exponential phase of amplification. PCR products were run on a 1.5% agarose gel

to ensure only a single band had been produced and to estimate relative concentrations using FIJI gel analysis. Indexed samples were pooled, normalizing for relative concentration. Pooled products were run on a 1.5% agarose gel, cutting out a band at 941 bp, which was then purified using the NucleoSpin Gel and PCR Clean-up column. The concentration of the pooled library was quantified using a Qubit fluorometer, and size distribution was assessed using a HS DNA chip on the Bioanalyzer 2100 instrument (Agilent). The library was sequenced using 2 × 75 bp paired-end reads on an Illumina MiSeq (v3 Reagent kit) loaded at a final concentration of 14 pM with 15% PhiX spiked in.

Sort-seq data analysis

Paired end reads were merged using BBMerge [116]. CggR-cpGFP insertion sites were counted using the dipseq analysis pipeline developed by the Savage lab available at (https://github.com/SavageLab/dipseq). This Python package identifies reads that contain sequences originating from both CggR and cpGFP. Junction reads are then trimmed to remove the cpGFP sequence plus transposon scar before aligning the remaining sequence to CggR to identify the site of insertion. The output is a file containing counts for each insertion site (including out-of-frame and reverse insertions) in each sample, which was used for enrichment and sort-seq analysis.

Sort-seq data analysis was performed as previously described[45,129] using functions written in R available at: https://github.com/jnkoberstein/biosensor-sort-seq. Raw sequencing data were processed to obtain read counts of each variant in each bin using the dipseq python package for the CggR-DIP library or a custom R script for the CggR-180 library. Mean fluorescence was estimated for both samples

using a maximum likelihood estimator and then used to calculate the dynamic range as $\Delta F/F = (F_I - F_0) / F_0$ where F_I is the fluorescence intensity in the sample with added glucose and F_0 is the fluorescence intensity in the absence of glucose. The CggR-DIP and CggR-180 libraries were filtered to keep only variants in which an estimate of more than 500 cells were collected for each sample and variance in each sample was less than 0.3.

Protein Purification

A protein expression destination vector was constructed by moving the CggR-180-Dest sequence from HC_Kan_CggR-180-Dest into the plasmid pSMT3. pSMT3 was a gift from Arthur Glasfeld that consists of a pET-28b vector modified to include an N-terminal 6xHIS fused to *S. cerevisiae* Smt3 (yeast homolog of SUMO). Briefly, the CggR-180-Dest sequence was amplified using primers pSMT3-CggR-EBD-GA-F2 and pSMT3-CggR-EBD-GA-R2 while the pSMT3 vector was amplified and linearized using pSMT3-CggR-EBD-GA-F1 and pSMT3-CggR-EBD-GA-R1. The two PCR fragments were purified and combined in a 20 µL Gibson Assembly reaction using 50 ng total of backbone amplicon and a 3-to-1 molar backbone:insert ratio. HYlight (CggR-180-PPKE) was cloned into pSMT3-CggR-180-Dest by amplifying cpGFP from pTKEI-Mal-B2 using primers CggR-180-PP-F CggR-180-KE-R to add the linker sequences, complementary overhangs and BsaI restriction sites for Golden Gate Cloning.

pSMT3-CggR-180-PPKE was transformed into E. cloni EXPRESS BL21(DE3) Competent Cells and plated on LB agar with kanamycin. A single colony was picked into 40 mL LB with kanamycin and grown overnight at 37°C while shaken at 250rpm. The following day 2 mL of the overnight culture was used to inoculate 400

mL LB with kanamycin which was incubated at 37°C for approximately 2 hours until reaching an OD600 between 0.5 and 0.6. Protein expression was then induced by adding IPTG to a concentration of 750 µM. Induced cultures were grown for an additional 3-4 hours at 30°C. Cultures were pelleted by centrifugation at 5000g for 20 minutes.

Cell pellets were resuspended in 8 mL lysis buffer (50 mM Na_2HPO_4, 350 mM NaCl and 20% sucrose, pH 8) in a 15 mL Falcon tube and a dash of lysozyme was added. Cells were lysed by sonication using 10s short bursts followed by 50s cooldown for 10 cycles. Lysates were centrifuged at 15,000g for 20 minutes to pellet the insoluble fraction. Cleared lysate was added to a column containing TALON Metal Affinity Resin and incubated while nutating at 4°C for at least 1 hour. The flow through was drained and the beads bound to HIS-tagged protein were washed 2 times with Wash Buffer (50 mM Na_2HPO_4 and 350 mM NaCl, pH 8) before eluting in Wash Buffer with 200 mM imidazole. The SMT3 domain was removed by incubation with the Ulp1 protease. The protein containing solution was run over a column containing TALON Metal Affinity Resin to separate the 6xHIS-tagged SMT3 from the now untagged protein which was captured in the flow through. The concentration of collected protein was measured by Bradford assay prior to further characterization of biosensor function.

HYlight (and various binding and fluorescence mutations) were tested for affinity and dynamic range in a 96-well plate format measured by a BMG CLARIOstar Plus Microplate Reader. Purified protein diluted to approximately 10 µM in PBS in a 96-well plate at a volume of 100 µL per well. Excitation scans were collected at 550 nm emission with a 20 nm bandwidth, and emission scans were performed with 405- and 488-nm excitation. FBP solutions were prepared at 10X

final concentration in PBS. Tenfold dilutions of 5, 2, and 1 mM FBP were tested in triplicate to generate FBP binding curves.

Cell culture and Islet Preparation

HEK293T cells were maintained in DMEM supplemented with 10% FBS and transfected using lipofectamine 2000 as per the manufacturer's protocol. MIN6 cells were maintained in DMEM containing 15% FBS, 1% Pen/Strep and 70 µM β-mercaptoethanol. MIN6 cells were transduced with an insulin promoter driven, β-cell specific adeno-associated virus [152] encoding HYlight for 2 hours and then returned to MIN6 culture media. Primary islets were freshly isolated from 8 weeks old male mice (Jackson Laboratories) and isolated as previously described [153]. Islets were cultured in RPMI-1640 supplemented with 10% FBS and 1% Pen/Strep. Immediately following isolation, islets were infected with a β-cell specific AAV encoding HYlight for 2 hours in 5% CO_2 at 37˚C and then returned to islet culture media. Experiments were conducted 2 days after infection.

Imaging and Quantification

MIN6 or HEK293T cells were plated onto glass bottom 35 mm pie dishes at $2x10^5$ or $1x10^5$ cells per quadrant, respectively and either transfected or transduced as described above. One hour prior to imaging, culture media was changed to 500 µL of 0 mM or low glucose in media containing 145 mM NaCl, 5 mM KCl, 1.2 mM $MgCl_2$, 2.6 mM $CaCl_2$, and 10 mM HEPES, pH 7.4. Live cell imaging was performed on a Nikon Eclipse TiE inverted microscope encased in a Okolab H101 stage-top incubator with temperature, humidity and CO_2 control with a Yokogawa CSU-W1 spinning disk confocal unit. Images were taken using a 60X oil-based objective (NA 1.4, Plan Apo VC OFN 25) and using the perfect focus system (PFS) for automatic

correction of drift and to aid in stability during long term time-lapse imaging. Cells were maintained in 5% CO_2 at 37˚C for the duration of the experiment. Cells were excited at 488 nm and 405 nm with emission wheel 525/25 nm. Images were acquired every 10-15 seconds. To test sensor responses, glucose or drugs were applied directly to the media during the imaging session. For the glycolytic stress test cells were glucose starved for 1 hour prior to imaging. Starved cells were imaged for 5 minutes, followed by 11 mM glucose for 30 minutes, 2.5 μM Oligomycin for 30 minutes and lastly 2-deoxyglucose for 30 minutes. For the imaging of whole islets, a peristaltic pump perfusion system was used to deliver media which was alternated between low and high glucose. Images were processed using CellPose [154] to segment individual cells and generate ROIs, which were quantified using FIJI. Background fluorescence intensity was calculated from a region containing no cells and this value was subtracted from all pixel values. Mean intensity values for both the 488 and 405 channels were measured for each ROI and used to calculate the excitation ratio ($R = F_{488}/F_{405}$). The change in excitation ratio was calculated as $\Delta R/R = (R_t - R_{t=0})/R_{t=0}$ where $R_{t=0}$ is the excitation ratio as the start of experiment.

Flow cytometry

MIN6 cells were seeded in 6 well plates (10^6 per well). The following day MIN6 cells were transduced with a β-cell specific AAV encoding HYlight. Cells were incubated with 5% CO_2 at 37˚C for two days. HEK293T cells were plate in 6 well plates ($7.5x10^6$ per well). The following day HEK293T cells were transfected using Lipofectamine 2000 per the manufacturer's protocol. Cells were incubated with 5% CO_2 at 37˚C for 24 hours. One hour prior to harvest, MIN6 or HEK293 culture media

was exchanged for media containing 145 mM NaCl, 5 mM KCl, 1.2 mM $MgCl_2$, 2.6 mM $CaCl_2$, 10 mM HEPES and glucose ranging from 0-25 mM. For experiments in MIN6 cells using the glucokinase activator dorzagliatin (Selleckchem, Cat. #S6921), cells were treated with 10 μM dorzagliatin for 1 hour in media containing 145 mM NaCl, 5 mM KCl, 1.2 mM $MgCl_2$, 2.6 mM $CaCl_2$, 10 mM HEPES and glucose ranging from 0-25 mM. Following the 1-hour incubation, cells were trypsinized for 2 minutes, spun down at 2,000rpm for 3 minutes and resuspended in 500ul of media containing the same concentration of glucose prior to harvest, with or without Dorzagliatin. Cells were kept at room temperature for the remainder of the experiment. Data was collected on a BD Symphony Flow Cytometer using lasers 488-1 (Ex. 488 nm, Em. 530/30 nm) and 405-2 (Ex. 405 nm, Em. 525/50 nm). Cells were gated to exclude dead cells and debris (using forward and side scatter area) followed by a standard doublet-exclusion (using forward scatter area and width). Fluorescent cells were gated for on the diagonal of a plot of 405-2 vs 488-1 excluding all non-transfected cells. At least 10^4 fluorescent cells were evaluated per sample and three samples (independent wells) were prepared per glucose concentration. The excitation ratio ($\mathrm{R} = \mathrm{F}_{488} - \mathrm{F}_{405}$) was calculated per cell on log(AFU) values. Michaelis-Menten curves were fit to the median ratio calculated per sample.

Author Contributions

Conceptualization, J.N.K., P.J.S.S., and R.H.G.; Methodology, J.N.K., C.B.S., M.L.S., F.M.A., and A.I.T.; Investigation, J.N.K., C.B.S., M.L.S. and A.I.T.; Writing-Original draft, J.N.K., P.J.S.S., and R.H.G.; Writing-Review and editing, J.N.K., M.L.S., A.I.T., F.M.A., P.J.S.S., and R.H.G.; Funding acquisition, P.J.S.S., A.I.T.,

and R.H.G.; Resources, J.N.K. and R.H.G.; Supervision, P.J.S.S, F.M.A., and R.H.G.

Acknowledgements

We thank Markus Grompe and Sunghee Chai for the beta cell specific AAV vector, Xander Viray and James Frank for help with calcium measurements, Matthew Schleisman for assistance with FACS, Yibing Jia for help with DNA sequencing, Stephanie Kaech for her expertise in microscopy, and Carsten Shultz and Gary Yellen for helpful discussions. Portions of this work were carried out in the OHSU Imaging Core, Flow Cytometry Core, Molecular Technologies Core and by R01 AG055431 (to R.H.G.).

3.6 Supplemental Materials

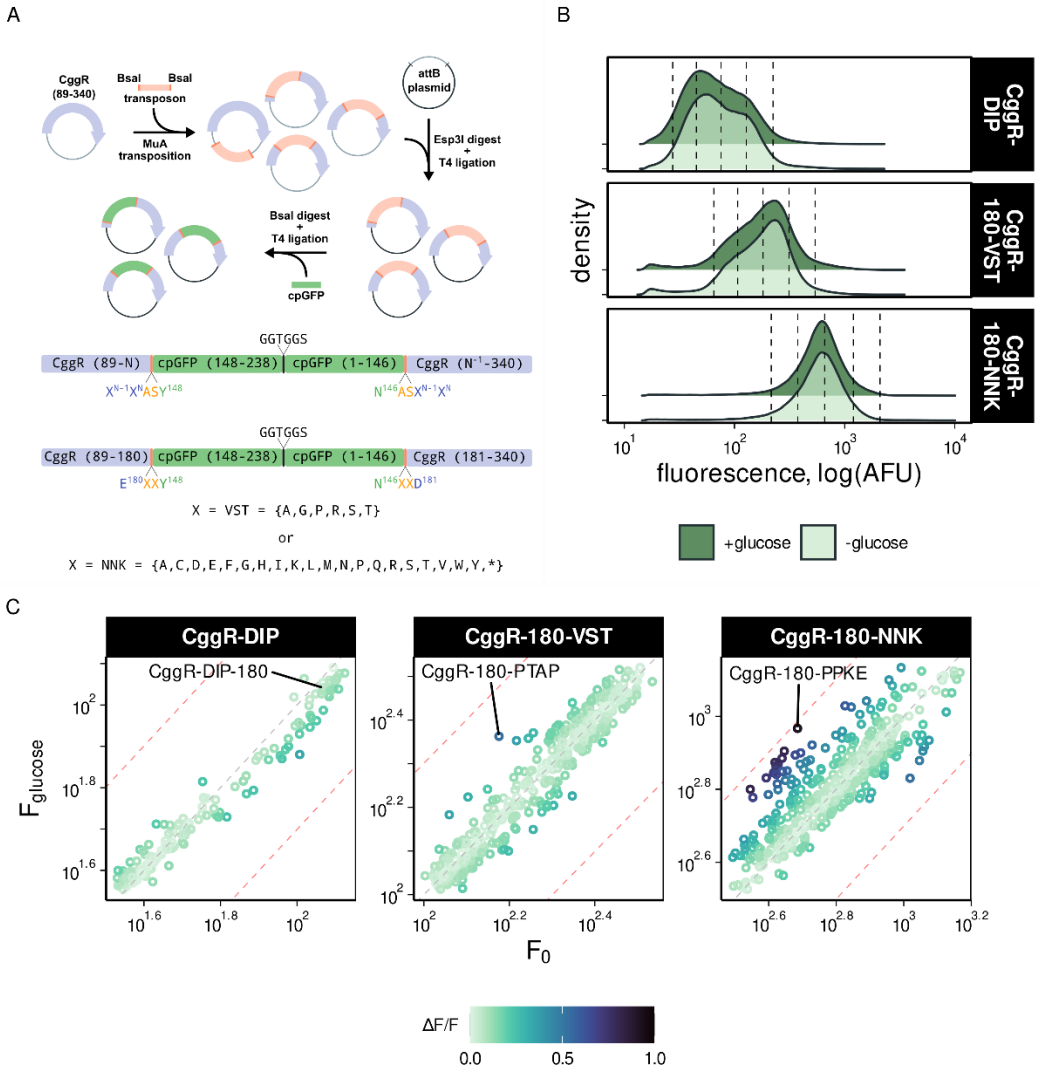

Figure S1. Discovery of a high-dynamic-range FBP biosensor by sort-seq assay. (A) A transposon based cloning strategy was used to construct a domain-insertion profiling (DIP) library with variable insertion of cpGFP into CggR. The BsaI restriction sites flanking the transposon sequence produce short Ala-Ser linkers on either side of cpGFP, along with a duplication of first two amino acids preceding cpGFP (N^{-1} and N). Two additional libraries consisting of cpGFP insertion at CggR residue 180 with variable linkers were constructed using PCR with primers containing degenerate nucleotides. These libraries are referred to as CggR-180-VST and CggR-180-NNK. **(B)** Cellular fluorescence distributions and FACS gate configurations for the three separate libraries evaluated by sort-seq. The CggR-180-NNK library was initially sorted for a limited number of bright cells to bottleneck the number of unique variants present while also enriching especially bright variants. CggR-DIP was initially sorted for variants with fluorescence intensity above un-recombined cells, while the naive CggR-180-VST library was used as input for sort-seq. **(C)** Sort-seq estimates of the mean fluorescence for each variant in glucose starved (F_0) and fed ($F_{glucose}$) conditions.

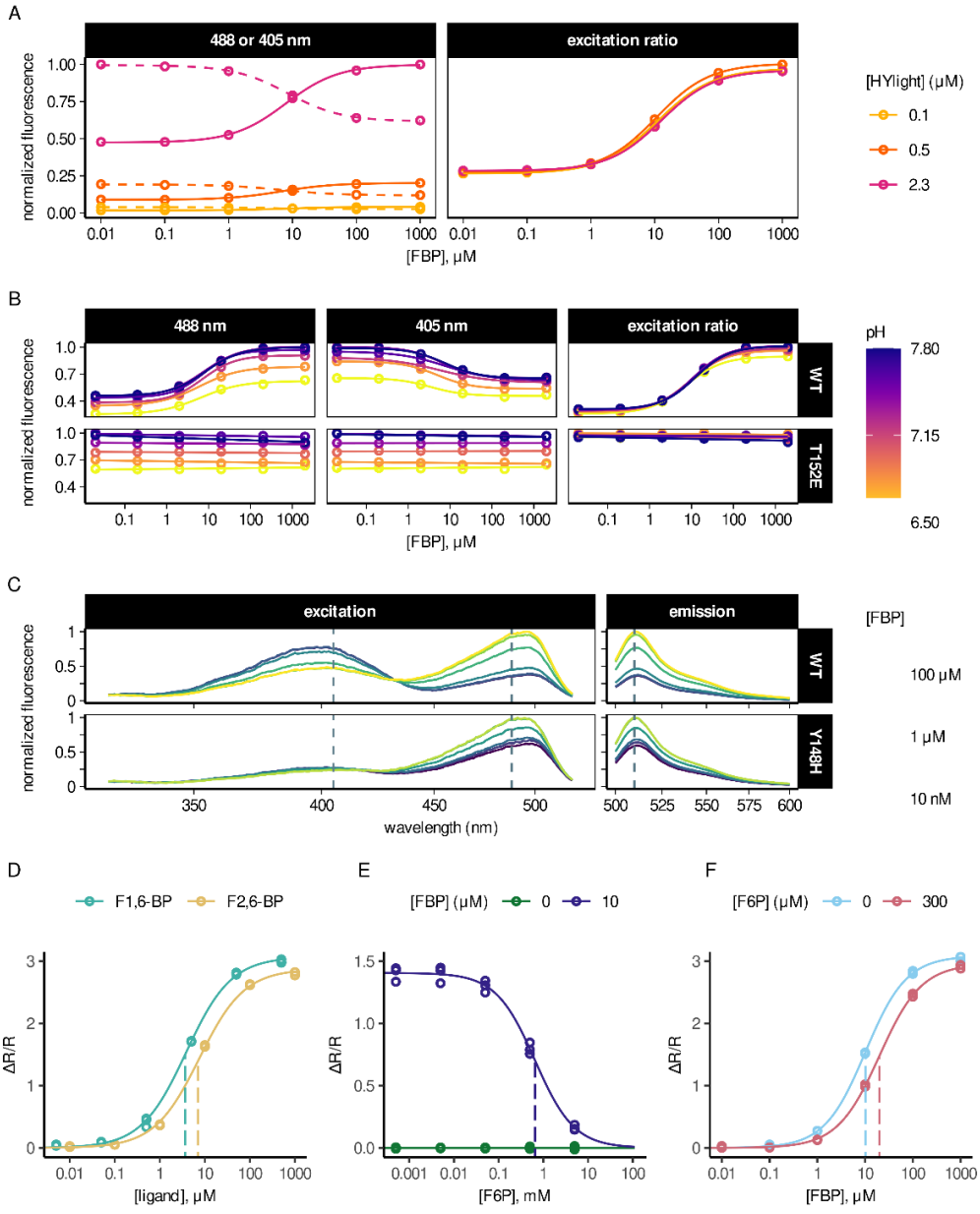

Figure S2. In vitro characterization of HYlight (A) HYlight excitation ratio consistently reports [FBP] across varied protein concentrations. Dashed lines in left panel correspond to 405 nm and solid line to 488 nm excitation. **(B)** HYlight exhibits pH dependent changes for both 488 nm and 405 nm excitation, however the ratio of these two parameters is largely pH insensitive. Mutation of Thr^{152} to Glu (numbered according to CggR) results in a loss of FBP sensitivity while retaining similar pH sensitivity across both 488 and 405 nm excitation. **(C)** The prominent neutral fluorophore excitation peak (405 nm) and red-shifting of the anionic fluorophore excitation peak (maximum at 495 nm) observed for HYlight can be attributed to two specific mutations. Substitution of Tyr148 back to wildtype His (numbered according to wtGFP) results in a loss of 405 nm excitation. Other biosensors including ratiometric pericam (Nagai et al., 2001), Perceval (Berg et al., 2009), FGBP (Hu et al., 2018), GEX-GECO1 (Zhao et al., 2011) and ExRai-AKAR (Mehta et al., 2018) exhibit similar dual

excitation capabilities. None of these other ratiometric biosensors contain H148Y. However, ratiometric-pericam, Perceval, and FGBP contain the mutation H148D, which may function similarly. **(D)** Relative fluorescence ratio as a function of fructose 1,6-bisphosphate (F1,6-BP) and fructose 2,6-bisphosphate (F2,6-BP) concentration exhibits similar maximal changes in fluorescence ratio ($\Delta R/R = 3.0$) with slightly different estimated K_d (3.8 µM for F1,6-BP vs. 7.2 µM for F2,6-BP). **(E)** Fructose 6-phosphate (F6P) can compete with FBP for the HYlight binding site. In the presence of 10 µM FBP (purple) addition of F6P results in a decrease in $\Delta R/R$ with an apparent affinity of 321 µM. In the absence of FBP (green) addition of F6P results in no change to the fluorescence ratio. **(F)** F6P shifts the affinity of HYlight for FBP from 10.2 µM (no F6P) to 20.0 µM (300 µM F6P) consistent with the affinity for F6P determined in (E).

Chapter 4. Summary, Conclusions and Future Directions

Fluorescent biosensors have the potential to revolutionize the study of *in vivo* metabolism, but significant challenges limit their current use. A primary obstacle is the difficulty in developing new biosensors and improving existing designs. In chapter two, we present the application of a high-throughput cell-sorting and sequencing assay to the problem of quantifying biosensor dynamic range. This work expands upon prior literature by discovering new functional domain-insertion variants in a previously characterized library[49] and further extends this approach to the characterization of linker mutations. In chapter three, we apply these techniques to the engineering of an FBP biosensor using the ligand-binding domain CggR. We show this this new biosensor, HYlight, can be used to interrogate glycolytic metabolism with high spatiotemporal resolution. The flexibility of HYlight presents many new opportunities to investigate glycolytic heterogeneity, temporal dynamics, and subcellular localization.

4.1 Methods for high-throughput biosensor characterization

Domain-insertion and linker variation

In chapter two, we generated and characterized domain-insertion and linker libraries. Both libraries contained functional bright and responsive variants that were successfully detected by sort-seq. In chapter three, a domain-insertion library consisting of cpGFP insertions into CggR did not yield any high-dynamic-range variants. Mutating the linkers at a single site, CggR-180, resulted in several functional variants with both direct and inverse fluorescent responses to increased FBP. It is likely that promising CggR insertion-sites were left undiscovered due to suboptimal linkers. The demonstration here, along with previous evidence that the

linkers are of critical importance to generate function[2], suggests that an improved approach will be required to identify the optimal insertion-site.

One simple solution is to change the linkers imposed by the transposon scar, AS-AS, to linkers that have shown to function across multiple sites. However, identifying a pair of generally useful linkers is a non-trivial problem and altering this sequence will impact transposition efficiency. A more complicated alternative is to generate libraries in which many linkers at many insertion sites are simultaneously generated. Oligo-pool synthesis has been previously used to generate domain-insertion libraries[115], and the choice of junction sequence between domains is not constrained as it is for transposition-based methods. Devising efficient cloning and downstream sequencing strategies for domain-insertion libraries with variable linkers will be critical for discovering highly optimized biosensors.

Towards data-driven biosensor design

The application of massively parallel assays to biosensor libraries was motivated by two complementary aspects of this approach. The primary goal was to increase testing throughput to efficiently discover the rare functional sequences. We were successful in characterizing approximately 100-200 domain-insertion variants and 400-1000 linker variants per library. These efforts are on par with the more extensive medium-throughput screening efforts performed to date[37,38,155,156]. While the top performing variants discovered in these screens do not exhibit comparable dynamic range to the best existing biosensors, this likely reflects inefficiencies in library design rather than screening approach.

A secondary benefit is the generation of datasets necessary to enable the application of machine learning to biosensor design. The collection of existing

biosensor designs is currently of limited use for this purpose, given that only a handful of sequenced variants are presented per publication[2,37,38,71,157], with notable exceptions[155,156,158], and this data is scattered across many such publications. In addition to discovering the variant termed HYlight, we produced brightness and dynamic range measurements for 900 additional CggR-180 linker variants. It is likely that HYlight is not the optimal sequence among the 160,000 possible linker combinations in terms of dynamic range. HYlight is certainly not the global optimum when considering the entire biosensor sequence. Tailoring ML algorithms to the task of predicting biosensor brightness and dynamic range from sequence will be important to make full use of this data. Gaussian Process (GP) models are a promising method for efficiently searching the space of possible linkers for the optimal combination. This class of model produces an output prediction as well as an estimate of prediction uncertainty. Using the confidence-interval generated by GP models, a new library could be designed that consists of linker combinations that are predicted to be bright and responsive along with variants from unexplored regions of sequence space for which predictions are relatively uncertain[159–161]. ML feedback onto library design will be useful for decreasing the number of non-functional sequences tested and increasing the likelihood that variants with improved properties are present in the pool during each round of screening.

Directed evolution of fluorescent biosensors

Directed evolution has emerged as a promising technique to discover highly functional fluorescent biosensors. Traditionally directed evolution is aided by high-throughput selection techniques which are applicable to protein functions such as binding and catalysis but has proven difficult to apply to biosensors. Recently, high-

dynamic-range biosensors for citrate[41] and lactate[39] have been generated using upwards of ten rounds of error-prone PCR screened by *in vitro* characterization of dynamic range. The top variants identified by this approach are highly sensitive ($\Delta F/F > 10$) approaching performance comparable to early iterations of GCaMP. The use of low-throughput selection techniques is however not ideal.

A system for physically picking variants based on dynamic range has been created using microscopy and robotics that offers high-throughput selection[162]. To evolve improved voltage biosensors, cultured HEK293 cells expressing a library of variants were imaged while manipulating membrane potential. Cells containing variants exhibiting a combination of appropriate trafficking to membrane, high brightness and voltage sensitivity were then physically picked from the plate using a robotically controlled suction pipette. A similar system relying on tagging cells with light using a photoactivatable fluorescent protein, such that selected cells can later be recovered by FACS[163], has been develop but has not yet been applied to directed evolution of proteins. While promising, these high-throughput selection techniques require specialized microscope setups which may limit their adoption.

A key challenge for ML-based biosensor engineering will be to outperform these emerging directed evolution approaches. The outcome of traditional directed evolution depends on the starting sequence and the shape of the fitness landscape. By focusing on iterative improvements in function at each step, it is possible to get stuck at a local optimum in which all mutations decrease function. This makes it especially difficult to cross large “valleys” of diminished function separating fitness peaks. A hybrid method, ML-directed evolution provides desirable aspects of each approach[164,165]. Instead of relying on random mutagenesis, each round consists of screening ML-designed libraries specifically targeting regions of sequence space

predicted to contain highly functional variants. The resulting data can then be used to update the ML model to inform the next library design. By leveraging the extent of known sequence-function information, it becomes possible to make large jumps in sequence space while minimizing the amount of low fitness sequences that are tested[105]. ML-directed evolution thus represents a possible path to efficiently generate biosensors with GCaMP-level functionality with less experimental effort.

Rational design

The development of GFP and GCaMP occurred through cycles of random mutagenesis, functional screening, and mechanistic interrogation. The approach envisioned in this thesis, combining massively parallel assays and machine learning, is not a deviation from this paradigm, but rather just a change in scale. The complexity of allosteric proteins necessitates complex models, a problem well suited for machine learning algorithms. Biophysics informed ML models[166–168] will be of use to bridge the gap between predictive performance and mechanistic insight into biosensor function. A truly rational biosensor design process might consist of human-interpretable ML models capable of *in silico* selection for optimal performance along multiple dimensions (brightness, dynamic range, binding affinity, and kinetics). The collection of sequence-function data presented here is just an initial step towards this larger goal.

4.2 Assessment of the spatiotemporal regulation of glycolytic flux

Limitations of the FBP-flux relationship

In this work, we did not directly evaluate the relationship between glycolytic flux, FBP concentration, and HYlight fluorescence ratio. Instead, we relied on the

numerous examples of the flux-FBP correlation present in the literature combined with the results of our experiments using HYlight. Across the cell types, conditions and perturbations tested, the observed HYlight fluorescence ratio exhibited relative changes consistent with the expected changes in flux. Together this empirical foundation generated in our lab and others supports the use of HYlight for the inference of flux changes. It would however be beneficial to perform a systematic evaluation of glycolytic flux, FBP concentration and HYlight fluorescence ratio. Performing paired metabolomics and fluorescence measurements for cell samples across multiple conditions would be useful to determine the exact degree of correlation between these variables. Additionally, the range of FBP concentrations that HYlight detects in the intracellular environment could be defined from the resulting data. The competitive binding of HYlight to glycolytic metabolites other than FBP currently makes such an assessment difficult when relying on assays of purified protein *in vitro*. It might also be counterintuitively useful to identify edge-cases in which FBP-flux correlation is not maintained. Elucidating such conditions would provide insights into the limits of flux detection by HYlight and valuable guidance on contexts where using this fluorescent biosensor is appropriate.

Multiplexed detection of metabolic properties

As an initial demonstration of multiplexed imaging, we used a red Ca^{2+} biosensor, R-GECO1, in combination with HYlight to detect synchronized electrical and metabolic oscillations in beta cells. The interaction of FBP with other aspects of metabolism was of interest, but the simultaneous detection of multiple metabolic parameters is currently limited by spectral overlap. Highly functional fluorescent biosensors for citrate (Citron/Citroff, cpGFP), pyruvate (PyronicSF, cpGFP),

NADH:NAD^+ (Peredox, cp-T-Sapphire) and ATP:ADP (Perceval, cp-mVenus) have all been developed using variants of GFP with significant spectral overlap. Our initial attempts at developing an FBP biosensor based on cp-mRuby (Appendix I) provides a starting point towards the goal of multiplexed metabolic measurements. However, the inconsistent performance of the red FBP biosensors when evaluated by microscopy limits their use and warrants further evaluation.

Simultaneous detection of glycolytic flux and TCA flux would be useful for disentangling the glucose consumption terminates in pyruvate or lactate production. An increase in TCA flux without a corresponding increase in glycolytic flux could be taken as evidence for increased lactate uptake from the extracellular environment. Citrate has been proposed to act as a flux-signaling metabolite for the TCA based on its high variance across condition and numerous regulatory interactions in bacteria. Whether this role is maintained in eukaryotes where the TCA cycle instead occurs in mitochondria is unclear. Citrate concentrations have been found to lack correlation with glycolytic flux in cultured iBMK cells[59] demonstrating that TCA and glycolytic flux are not necessarily coupled. In contrast, the existing citrate biosensors exhibit a strong response to glucose stimulation when expressed in the INS1 pancreatic beta cells[41]. This finding is to be expected if citrate acts a flux-signaling metabolite given the exhibit strong coupling of glycolytic and TCA flux driven by glucose concentration in this cell type. Citrate also acts as an allosteric inhibitor of PFK1, which suggests that increases in cytosolic citrate could result in decreases in FBP. Multiplexed monitoring of citrate and FBP is an intriguing target, should a red fluorescent biosensor for either be developed.

Beta cells and neurons

To validate the performance of HYlight in live cells, we chose pancreatic beta cells due to the ease of manipulating glycolytic flux in this cell type. The observation of oscillations in FBP and provided a clear example of the capacity of HYlight to uncover dynamic changes in glycolytic metabolism and resolve the temporal relationships with other cellular signals. While the primary motivation for engineering an FBP biosensor was specifically for use in neurons, the dynamic aspects of beta cells are reminiscent of neuronal physiology[169,170]. The overlap between these two cell-types provides some suggestion that HYlight might be useful to interrogate glycolytic dynamics driven by activity in neurons. In particular, we envision monitoring FBP using HYlight will provide a unique lens to view glycolytic regulation during neuronal activation.

Sub-cellular regulation of glycolysis

Biosensors also critically enable subcellular measurements, which we have not yet explored using HYlight. Clusters of glycolytic enzymes, called glycolytic metabolons, have recently been discovered in *C. elegans* neurons[86]. These metabolons are dynamically formed near synapses in response to hypoxia and optogenetic stimulation. PFK, the FBP producing enzyme, is implicated in the formation of these metabolons which display properties consistent with liquid-phase-separated droplets[171]. The functional consequences of glycolytic metabolon formation are difficult to measure *in vivo*. The development of HYlight presents a unique opportunity to evaluate whether metabolons exhibit elevated FBP and support enhanced glycolytic flux relative to the surrounding cytoplasm. It is likely that targeting of HYlight to the phase-separated droplet will be required which

complicates interpretation given the unique properties of condensates. It is also unclear if the FBP-flux correlation holds at subcellular scales or if it is an emergent property of cells. Despite these complications, glycolytic metabolons represent a fascinating new system to understand *in vivo* glycolytic regulation at subcellular scales.

References

1. Tomasello, G., Armenia, I. & Molla, G. The Protein Imager: A full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36**, 2909–2911 (2020).
2. Marvin, J. S., Schreiter, E. R., Echevarría, I. M. & Looger, L. L. A genetically encoded, high-signal-to-noise maltose sensor. *Proteins* **79**, 3025–36 (2011).
3. Tsien, R. Y. Building and breeding molecules to spy on cells and tumors. in *FEBS Letters* vol. 579 927–932 (John Wiley & Sons, Ltd, 2005).
4. Schultz, C. Fluorescent Revelations. *Chemistry and Biology* vol. 16 107–111 (2009).
5. Tsien, R. Y. The green fluorescent protein. *Annual Review of Biochemistry* vol. 67 509–544 (1998).
6. SHIMOMURA, O., JOHNSON, F. H. & SAIGA, Y. Extraction, purification and properties of aequorin, a bioluminescent. *Journal of cellular and comparative physiology* **59**, 223–239 (1962).
7. Shimomura, O. & Johnson, F. H. Properties of the Bioluminescent Protein Aequorin. *Biochemistry* **8**, 3991–3997 (1969).
8. Shimomura, O. Structure of the chromophore of Aequorea green fluorescent protein. *FEBS Letters* **104**, 220–222 (1979).
9. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).
10. Ormö, M. *et al.* Crystal structure of the Aequorea victoria green fluorescent protein. *Science* **273**, 1392–1395 (1996).
11. Lambert, T. J. FPbase: a community-editable fluorescent protein database. *Nature Methods* vol. 16 277–278 (2019).
12. Yang, F., Moss, L. G. & Phillips, G. N. The Molecular Structure of Green Fluorescent Protein. *Nature Biotechnology* **14**, 1246–1251 (1996).
13. Heim, R., Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 12501–12504 (1994).
14. Cubitt, A. B. *et al.* Understanding, improving and using green fluorescent proteins. *Trends in Biochemical Sciences* **20**, 448–455 (1995).
15. Brejc, K. *et al.* Structural basis for dual excitation and photoisomerization of the Aequorea victoria green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2306–2311 (1997).
16. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). in *Gene* vol. 173 33–38 (Elsevier, 1996).
17. Matz, M. v. *et al.* Fluorescent proteins from nonbioluminescent Anthozoa species. *Nature Biotechnology* **17**, 969–973 (1999).

18. Wiedenmann, J. *et al.* A far-red fluorescent protein with fast maturation and reduced oligomerization tendency from Entacmaea quadricolor (Anthozoa, Actinaria). *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11646–11651 (2002).
19. Merzlyak, E. M. *et al.* Bright monomeric red fluorescent protein with an extended fluorescence lifetime. *Nature Methods* **4**, 555–557 (2007).
20. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature Biotechnology* **22**, 1567–1572 (2004).
21. Llopis, J., McCaffery, J. M., Miyawaki, A., Farquhar, M. G. & Tsien, R. Y. Measurement of cytosolic, mitochondrial, and Golgi pH in single living cells with green fluorescent proteins. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6803–6808 (1998).
22. Stryer, L. Fluorescence energy transfer as a spectroscopic ruler. *Annual review of biochemistry* vol. 47 819–846 (1978).
23. Tsien, R. Y., Bacskai, B. J. & Adams, S. R. FRET for studying intracellular signalling. *Trends in Cell Biology* **3**, 242–245 (1993).
24. Mitra, R. D., Silva, C. M. & Youvan, D. C. Fluorescence resonance energy transfer between blue-emitting and red-shifted excitation derivatives of the green fluorescent protein. in *Gene* vol. 173 13–17 (Elsevier, 1996).
25. Romoser, V. A., Hinkle, P. M. & Persechini, A. Detection in living cells of Ca2+-dependent changes in the fluorescence emission of an indicator composed of two green fluorescent protein variants linked by a calmodulin-binding sequence. A new class of fluorescent indicators. *Journal of Biological Chemistry* **272**, 13270–13274 (1997).
26. Baird, G. S., Zacharias, D. A. & Tsien, R. Y. Circular permutation and receptor insertion within green fluorescent proteins. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 11241–11246 (1999).
27. Nagai, T., Sawano, A., Park, E. S. & Miyawaki, A. Circularly permuted green fluorescent proteins engineered to sense Ca2+. *Proceedings of the National Academy of Sciences* **98**, 3197–3202 (2001).
28. Nakai, J., Ohkura, M. & Imoto, K. A high signal-to-noise ca2+ probe composed of a single green fluorescent protein. *Nature Biotechnology* **19**, 137–141 (2001).
29. 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).
30. Nasu, Y., Shen, Y., Kramer, L. & Campbell, R. E. Structure- and mechanism-guided design of single fluorescent protein-based biosensors. *Nature Chemical Biology* 1–10 (2021) doi:10.1038/s41589-020-00718-x.

31. Koveal, D., Díaz-García, C. M. & Yellen, G. Fluorescent Biosensors for Neuronal Metabolism and the Challenges of Quantitation. *Current Opinion in Neurobiology* vol. 63 111–121 (2020).
32. Morris, B., Monod, J. & Wainhouse, A. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. *Technology and Culture* vol. 13 (1972).
33. Monod, J., Changeux, J. P. & Jacob, F. Allosteric proteins and cellular control systems. *Journal of Molecular Biology* **6**, 306–329 (1963).
34. Choi, J. H., Laurent, A. H., Hilser, V. J. & Ostermeier, M. Design of protein switches based on an ensemble model of allostery. *Nature Communications* **6**, 6968 (2015).
35. Coyote-Maestas, W., He, Y., Myers, C. L. & Schmidt, D. Domain Insertion Permissibility is a Measure of Engineerable Allostery in Ion Channels. *bioRxiv* 334672 (2018) doi:10.1101/334672.
36. Cooper, A. & Dryden, D. T. F. Allostery without conformational change - A plausible model. *European Biophysics Journal* **11**, 103–109 (1984).
37. Zhang, J. F. *et al.* An ultrasensitive biosensor for high-resolution kinase activity imaging in awake mice. *Nature Chemical Biology* 1–8 (2020) doi:10.1038/s41589-020-00660-y.
38. Patriarchi, T. *et al.* Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. *Science* **360**, eaat4422 (2018).
39. Nasu, Y. *et al.* A genetically encoded fluorescent biosensor for extracellular l-lactate. *Nature Communications* **12**, 2021.03.05.434048 (2021).
40. Dong, C. *et al.* Psychedelic-inspired drug discovery using an engineered biosensor. *Cell* **184**, 2779-2792.e18 (2021).
41. Zhao, Y., Shen, Y., Wen, Y. & Campbell, R. E. High-Performance Intensiometric Direct- And Inverse-Response Genetically Encoded Biosensors for Citrate. *ACS Central Science* **6**, 1441–1450 (2020).
42. Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science. *Nature Methods* vol. 11 801–807 (2014).
43. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
44. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50**, 874–882 (2018).
45. Peterman, N. & Levine, E. Sort-seq under the hood: Implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
46. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).
47. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* **6**, 875–881 (2009).

48. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Research* **45**, e102–e102 (2017).
49. Nadler, D. C., Morgan, S.-A., Flamholz, A., Kortright, K. E. & Savage, D. F. Rapid construction of metabolite biosensors using domain-insertion profiling. *Nature Communications* **7**, 12266 (2016).
50. Coyote-Maestas, W., He, Y., Myers, C. L. & Schmidt, D. Domain insertion permissibility-guided engineering of allostery in ion channels. *Nature Communications* **10**, 290 (2019).
51. Oakes, B. L. *et al.* CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification. *Cell* vol. 176 254-267.e16 (2019).
52. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
53. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Systems* **11**, 461-477.e9 (2020).
54. Sinai, S. *et al.* AdaLead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv* (2020).
55. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *bioRxiv* 2020.01.23.917682 (2020) doi:10.1101/2020.01.23.917682.
56. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322 (2019).
57. Rao, R. *et al.* Evaluating protein transfer learning with TAPE. in *Advances in Neural Information Processing Systems* vol. 32 (Neural information processing systems foundation, 2019).
58. Rabinowitz, J. D. & Enerbäck, S. Lactate: the ugly duckling of energy metabolism. *Nature Metabolism* **2**, 566–571 (2020).
59. Tanner, L. B. *et al.* Four Key Steps Control Glycolytic Flux in Mammalian Cells. *Cell Systems* **7**, 49-62.e8 (2018).
60. Bley Folly, B. *et al.* Assessment of the interaction between the flux-signaling metabolite fructose-1,6-bisphosphate and the bacterial transcription factors CggR and Cra. *Molecular Microbiology* **109**, 278–290 (2018).
61. Litsios, A., Ortega, Á. D., Wit, E. C. & Heinemann, M. Metabolic-flux dependent regulation of microbial physiology. *Current Opinion in Microbiology* vol. 42 71–78 (2018).
62. Kochanowski, K. *et al.* Functioning of a metabolic flux sensor in Escherichia coli. *Proceedings of the National Academy of Sciences* **110**, 1130–1135 (2013).
63. Monteiro, F. *et al.* Measuring glycolytic flux in single yeast cells with an orthogonal synthetic biosensor. *Molecular Systems Biology* **15**, (2019).

64. Ludwig, H. *et al.* Transcription of glycolytic genes and operons in Bacillus subtilis: Evidence for the presence of multiple levels of control of the gapA operon. *Molecular Microbiology* **41**, 409–422 (2001).
65. Zhang, C. S. *et al.* Fructose-1,6-bisphosphate and aldolase mediate glucose sensing by AMPK. *Nature* **548**, 112–116 (2017).
66. Li, M. *et al.* Aldolase is a sensor for both low and high glucose, linking to AMPK and mTORC1. *Cell Research* vol. 31 478–481 (2021).
67. Jang, C., Chen, L. & Rabinowitz, J. D. Metabolomics and Isotope Tracing. *Cell* vol. 173 822–837 (2018).
68. Lu, W. *et al.* Metabolite measurement: Pitfalls to avoid and practices to follow. *Annual Review of Biochemistry* **86**, 277–304 (2017).
69. Teslaa, T. & Teitell, M. A. Techniques to monitor glycolysis. in *Methods in Enzymology* vol. 542 91–114 (Academic Press, 2014).
70. Ferrick, D. A., Neilson, A. & Beeson, C. Advances in measuring cellular bioenergetics using extracellular flux. *Drug Discovery Today* vol. 13 268–274 (2008).
71. Arce-Molina, R. *et al.* A highly responsive pyruvate sensor reveals pathway-regulatory role of the mitochondrial pyruvate carrier MPC. *eLife* **9**, (2020).
72. Rolfe, D. F. S. & Brown, G. C. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews* **77**, 731–758 (1997).
73. Hui, S. *et al.* Glucose feeds the TCA cycle via circulating lactate. *Nature* **551**, 115–118 (2017).
74. Hyder, F., Rothman, D. L. & Bennett, M. R. Cortical energy demands of signaling and nonsignaling components in brain are conserved across mammalian species and activity levels. *Proceedings of the National Academy of Sciences of the United States of America* vol. 110 3549–3554 (2013).
75. Madsen, P. L., Cruz, N. F., Sokoloff, L. & Dienel, G. A. Cerebral oxygen/glucose ratio is low during sensory stimulation and rises above normal during recovery: Excess glucose consumption during stimulation is not accounted for by lactate efflux from or accumulation in brain tissue. *Journal of Cerebral Blood Flow and Metabolism* **19**, 393–400 (1999).
76. Hu, Y. & Wilson, G. S. A temporary local energy pool coupled to neuronal activity: Fluctuations of extracellular lactate levels in rat brain monitored with rapid-response enzyme-based sensor. *Journal of Neurochemistry* **69**, 1484–1490 (1997).
77. Mazuel, L. *et al.* A neuronal MCT2 knockdown in the rat somatosensory cortex reduces both the NMR lactate signal and the BOLD response during whisker stimulation. *PLoS ONE* **12**, e0174990 (2017).
78. Magistretti, P. J., Sorg, O., Yu, N., Martin, J. L. & Pellerin, L. Neurotransmitters regulate energy metabolism in astrocytes: Implications for

the metabolic trafficking between neural cells. *Developmental Neuroscience* **15**, 306–312 (1993).
79. Díaz-García, C. M. *et al.* Neuronal Stimulation Triggers Neuronal Glycolysis and Not Lactate Uptake. *Cell Metabolism* **26**, 361-374.e4 (2017).
80. Díaz-García, C. M. & Yellen, G. Neurons rely on glucose rather than astrocytic lactate during stimulation. *Journal of Neuroscience Research* vol. 97 883–889 (2019).
81. Yellen, G. Fueling thought: Management of glycolysis and oxidative phosphorylation in neuronal metabolism. *Journal of Cell Biology* **217**, 2235–2246 (2018).
82. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *Journal of Neuroscience* **34**, 11929–11947 (2014).
83. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nature Neuroscience* **18**, 1819–1831 (2015).
84. Bak, L. K. *et al.* Neuronal glucose but not lactate utilization is positively correlated with NMDA-induced neurotransmission and fluctuations in cytosolic Ca2+ levels. in *Journal of Neurochemistry* vol. 109 87–93 (John Wiley & Sons, Ltd, 2009).
85. Sahlin, K., Tonkonogi, M. & Söderlund, K. Energy supply and muscle fatigue in humans. in *Acta Physiologica Scandinavica* vol. 162 261–266 (John Wiley & Sons, Ltd, 1998).
86. Jang, S. R. *et al.* Glycolytic Enzymes Localize to Synapses under Energy Stress to Support Synaptic Function. *Neuron* **90**, 278–291 (2016).
87. Greenwald, E. C., Mehta, S. & Zhang, J. Genetically encoded fluorescent biosensors illuminate the spatiotemporal regulation of signaling networks. *Chemical Reviews* **118**, 11707–11794 (2018).
88. Okumoto, S., Jones, A. & Frommer, W. B. Quantitative Imaging with Fluorescent Biosensors. *Annual Review of Plant Biology* **63**, 663–706 (2012).
89. Arce-Molina, R. *et al.* A highly responsive pyruvate sensor reveals pathway-regulatory role of the mitochondrial pyruvate carrier MPC. *bioRxiv* 611806 (2019) doi:10.1101/611806.
90. Hou, B. H. *et al.* Optical sensors for monitoring dynamic changes of intracellular metabolite levels in mammalian cells. *Nature Protocols* **6**, 1818–1833 (2011).
91. Miyawaki, A. & Niino, Y. Molecular Spies for Bioimaging-Fluorescent Protein-Based Probes. *Molecular Cell* vol. 58 632–643 (2015).
92. Ni, Q., Mehta, S. & Zhang, J. Live-cell imaging of cell signaling using genetically encoded fluorescent reporters. *The FEBS Journal* **285**, 203–219 (2018).
93. Pendin, D., Greotti, E., Lefkimmiatis, K. & Pozzan, T. Exploring cells with targeted biosensors. *The Journal of General Physiology* **149**, 1–36 (2017).

94. Doi, N. & Yanagawa, H. Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Letters* **453**, 305–307 (1999).
95. Marvin, J. S. *et al.* An optimized fluorescent probe for visualizing glutamate neurotransmission. *Nature Methods* **10**, 162–170 (2013).
96. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences* **107**, 9158–9163 (2010).
97. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 14024–14029 (2013).
98. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS -seq . *Molecular Systems Biology* **10**, 748 (2014).
99. Peterman, N., Lavi-Itzkovitz, A. & Levine, E. Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. *Nucleic Acids Research* **42**, 12177–12188 (2014).
100. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology* **30**, 521–530 (2012).
101. Sharon, E. *et al.* Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research* **24**, 1698–1706 (2014).
102. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Research* **45**, (2017).
103. Reetz, M. T. & Carballeira, J. D. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature Protocols* **2**, 891–903 (2007).
104. Bedbrook, C. N. *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature Methods* **16**, 1176–1184 (2019).
105. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *Nature Methods* **18**, 389–396 (2021).
106. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences* **116**, 8852–8858 (2019).
107. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *Journal of Chemical Information and Modeling* **60**, 2773–2790 (2020).

108. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648 (2018).
109. Mei, H., Liao, Z. H., Zhou, Y. & Li, S. Z. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers - Peptide Science Section* **80**, 775–786 (2005).
110. Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
111. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* vol. 9 (2019).
112. Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947–1958 (2003).
113. Akerboom, J. *et al.* Crystal structures of the GCaMP calcium sensor reveal the mechanism of fluorescence signal change and aid rational design. *Journal of Biological Chemistry* **284**, 6455–6464 (2009).
114. Chen, Y. *et al.* Structural insight into enhanced calcium indicator GCaMP3 and GCaMPJ to promote further improvement. *Protein and Cell* **4**, 299–309 (2013).
115. Coyote-Maestas, W., Nedrud, D., Okorafor, S., He, Y. & Schmidt, D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. *Nucleic Acids Research* **48**, 11 (2020).
116. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE* **12**, e0185056 (2017).
117. Martella, A., Matjusaitis, M., Auxillos, J., Pollard, S. M. & Cai, Y. EMMA: An Extensible Mammalian Modular Assembly Toolkit for the Rapid Design and Production of Diverse Expression Vectors. *ACS Synthetic Biology* **6**, 1380–1392 (2017).
118. Morgan, M. *et al.* ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
119. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1–26 (2008).
120. Wright, M. N. & Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (2017).
121. Young, W. J. & Harden, A. The alcoholic ferment of yeast-juice. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character* **77**, 405–420 (1906).
122. Kochanowski, K. *et al.* Functioning of a metabolic flux sensor in Escherichia coli. *Proceedings of the National Academy of Sciences* **110**, 1130–1135 (2013).
123. Řezáčová, P. *et al.* Crystal structures of the effector-binding domain of repressor Central glycolytic gene Regulator from Bacillus subtilis reveal

ligand-induced structural changes upon binding of several glycolytic intermediates. *Molecular Microbiology* **69**, 895–910 (2008).
124. Berg, J., Hung, Y. P. & Yellen, G. A genetically encoded fluorescent reporter of ATP:ADP ratio. *Nature Methods* **6**, 161–166 (2009).
125. Ashcroft, F. M., Harrison, D. E. & Ashcroft, S. J. H. Glucose induces closure of single potassium channels in isolated rat pancreatic β-cells. *Nature* **312**, 446–448 (1984).
126. Nicholls, D. G. The pancreatic β-cell: A bioenergetic perspective. *Physiological Reviews* **96**, 1385–1447 (2016).
127. Rorsman, P. & Ashcroft, F. M. Pancreatic β-cell electrical activity and insulin secretion: Of mice and men. *Physiological Reviews* **98**, 117–214 (2018).
128. Matschinsky, F. M. & Wilson, D. F. The central role of glucokinase in glucose homeostasis: A perspective 50 years after demonstrating the presence of the enzyme in islets of Langerhans. *Frontiers in Physiology* **10**, 148 (2019).
129. Koberstein, J. N., Stewart, M. L., Mighell, T. L., Smith, C. B. & Cohen, M. S. A Sort-Seq Approach to the Development of Single Fluorescent Protein Biosensors. *ACS Chemical Biology* **16**, 1709–1720 (2021).
130. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz910.
131. Lossau, H. *et al.* Time-resolved spectroscopy of wild-type and mutant Green Fluorescent Proteins reveals excited state deprotonation consistent with fluorophore-protein interactions. *Chemical Physics* **213**, 1–16 (1996).
132. Park, J. O. *et al.* Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nature Chemical Biology* **12**, 482–489 (2016).
133. Berman, H. K. & Newgard, C. B. Fundamental Metabolic Differences between Hepatocytes and Islet β-cells Revealed by Glucokinase Overexpression†. *Biochemistry* **37**, 4543–4552 (1998).
134. Matschinsky, F. M. Regulation of pancreatic β-cell glucokinase: From basics to therapeutics. in *Diabetes* vol. 51 S394–S404 (American Diabetes Association, 2002).
135. Aleshin, A. E. *et al.* Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes revelant to allosteric regulation. *Journal of Molecular Biology* **296**, 1001–1015 (2000).
136. Ardehali, H. *et al.* Functional organization of mammalian hexokinase II: Retention of catalytic and regulatory functions in both the NH2- and COOH-terminal halves. *Journal of Biological Chemistry* **271**, 1849–1852 (1996).
137. Sekine, N. *et al.* Low lactate dehydrogenase and high mitochondrial glycerol phosphate dehydrogenase in pancreatic β-cells. Potential role in nutrient sensing. *Journal of Biological Chemistry* **269**, 4895–4902 (1994).
138. Zhao, C., Wilson, M. C., Schuit, F., Halestrap, A. P. & Rutter, G. A. Expression and distribution of lactate/monocarboxylate transporter isoforms in pancreatic islets and the exocrine pancreas. *Diabetes* **50**, 361–366 (2001).

139. Civelek, V. N., Deeney, J. T., Fusonie, G. E., Corkey, B. E. & Tornheim, K. Oscillations in oxygen consumption by permeabilized clonal pancreatic β-cells (HIT) incubated in an oscillatory glycolyzing muscle extract roles of free Ca2+, substrates, and the ATP/ADP ratio. *Diabetes* **46**, 51–56 (1997).
140. Zhao, Y. *et al.* An expanded palette of genetically encoded Ca2+ indicators. *Science* **333**, 1888–1891 (2011).
141. Lorenz, M. A., el Azzouny, M. A., Kennedy, R. T. & Burant, C. F. Metabolome response to glucose in the β-cell line INS-1832/13. *Journal of Biological Chemistry* **288**, 10923–10935 (2013).
142. Keller, J. P. *et al.* In vivo glucose imaging in multiple model organisms with an engineered single-wavelength sensor. *Cell Reports* **35**, 109284 (2021).
143. Galaz, A. *et al.* Highly responsive single-fluorophore indicator to explore lactate dynamics in high calcium environments. *bioRxiv* 2020.10.01.322404 (2020) doi:10.1101/2020.10.01.322404.
144. Hung, Y. P., Albeck, J. G., Tantama, M. & Yellen, G. Imaging cytosolic NADH-NAD + redox state with a genetically encoded fluorescent biosensor. *Cell Metabolism* **14**, 545–554 (2011).
145. Merrins, M. J., van Dyke, A. R., Mapp, A. K., Rizzo, M. A. & Satin, L. S. Direct measurements of oscillatory glycolysis in pancreatic islet β-cells using novel fluorescence resonance energy transfer (FRET) biosensors for pyruvate kinase M2 activity. *Journal of Biological Chemistry* **288**, 33312–33322 (2013).
146. Benninger, R. K. P., Head, W. S., Zhang, M., Satin, L. S. & Piston, D. W. Gap junctions and other mechanisms of cell-cell communication regulate basal insulin secretion in the pancreatic islet. *Journal of Physiology* **589**, 5453–5466 (2011).
147. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology* vol. 20 1349–1360 (2018).
148. Kondo, H. *et al.* Single-cell resolved imaging reveals intra-tumor heterogeneity in glycolysis, transitions between metabolic states, and their regulatory mechanisms. *Cell Reports* **34**, 108750 (2021).
149. Norata, G. D. *et al.* The Cellular and Molecular Basis of Translational Immunometabolism. *Immunity* vol. 43 421–434 (2015).
150. Pearce, E. L. & Pearce, E. J. Metabolic pathways in immune cell activation and quiescence. *Immunity* vol. 38 633–643 (2013).
151. Rangaraju, V., Calloway, N. & Ryan, T. A. Activity-driven local ATP synthesis is required for synaptic function. *Cell* **156**, 825–835 (2014).
152. Pekrun, K. *et al.* Using a barcoded AAV capsid library to select for clinically relevant gene therapy vectors. *JCI Insight* **4**, 131610 (2019).
153. Tarasov, A. I. *et al.* The mitochondrial Ca2+ uniporter MCU is essential for glucose-induced atp increases in pancreatic β-cells. *PLoS ONE* **7**, e39722 (2012).

154. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. doi:10.1038/s41592-020-01018-x.
155. Dana, H. *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods* **16**, 649–657 (2019).
156. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
157. Cambronne, X. A. *et al.* Biosensor reveals multiple sources for mitochondrial NAD^+. *Science* **352**, 1474–7 (2016).
158. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *Journal of Neuroscience* **32**, 13819–13840 (2012).
159. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E193–E201 (2013).
160. Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *arXiv* (2021) doi:10.1016/j.sbi.2021.11.002.
161. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature Methods* **16**, 687–694 (2019).
162. Piatkevich, K. D. *et al.* A robotic multidimensional directed evolution approach applied to fluorescent voltage reporters. *Nature Chemical Biology* **14**, 352–360 (2018).
163. Hasle, N. *et al.* High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Molecular Systems Biology* **16**, e9442 (2020).
164. Wittmann, B. J., Yue, Y. & Arnold, F. H. Machine learning-assisted directed evolution navigates a combinatorial epistatic fitness landscape with minimal screening burden. *bioRxiv* 1–15 (2020) doi:10.1101/2020.12.04.408955.
165. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences of the United States of America* (2019) doi:10.1073/pnas.1901979116.
166. Otwinowski, J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Molecular Biology and Evolution* **35**, 2345–2354 (2018).
167. Faure, A. J. *et al.* Global mapping of the energetic and allosteric landscapes of protein binding domains. *bioRxiv* 2021.09.14.460249 (2021) doi:10.1101/2021.09.14.460249.
168. Morrison, A. J., Wonderlick, D. R. & Harms, M. J. Ensemble epistasis: Thermodynamic origins of nonadditivity between mutations. *Genetics* **219**, (2021).
169. Arntfield, M. E. & van der Kooy, D. β-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. *BioEssays* **33**, 582–587 (2011).

170. Eberhard, D. Neuron and beta-cell evolution: Learning about neurons is learning about beta-cells. *BioEssays* vol. 35 584 (2013).
171. Jang, S. R. *et al.* Phosphofructokinase relocalizes into subcellular compartments with liquid-like properties in vivo. *Biophysical Journal* **120**, 1170–1186 (2021).
172. Shcherbakova, D. M., Subach, O. M. & Verkhusha, V. v. Red fluorescent proteins: Advanced imaging applications and future design. *Angewandte Chemie - International Edition* vol. 51 10724–10738 (2012).
173. Akerboom, J. *et al.* Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. *Frontiers in Molecular Neuroscience* **6**, 2 (2013).
174. Dana, H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *eLife* **5**, (2016).

Appendix I. Development of a red FBP biosensor

Simultaneous imaging of multiple biosensors in single cells can reveal the spatial and temporal relationships between metabolic and cell signaling pathways. Multiplexed imaging requires spectral separation of the two signals. Most existing biosensors are based on GFP and its many derivatives which exhibit significant spectral overlap. In addition, most optogenetic reagents requires excitation by blue light. These factors limit the opportunities for experiments requiring multiple biosensors and optogenetic actuators. In contrast, red FP based biosensors are spectrally separated from GFP presenting new opportunities for multiplexed imaging. In addition, longer wavelength light results in reduced autofluorescence, light-scattering and phototoxicity when imaging tissues[172]. Despite these desirable properties, red FP-based biosensors have proven difficult to engineer. Only a few high-performance red biosensors currently exist, notably the calcium biosensors R-GECO and RCaMP, from which to base new designs.

Insertion of cpGFP at CggR-180 followed by high-throughput linker screening was utilized to generate a green FBP biosensor termed HYlight in Chapter 3. This same approach was applied to generate an FBP biosensor using a red FP. Among the red FPs, cp-mApple is the most common in biosensor designs[30,140], however it exhibits blue-light photoactivation that limit its use with other biosensors and optogenetic actuators[173]. Instead, we used cpmRuby derived from jRCaMP1a[174] which does not exhibit photoactivation. The linkers connecting CggR to cpmRuby were encoded by NNK codons resulting in combinations of all 20 amino acids across the 4 positions.

The library was recombined into the HEK293T Landing Pad genome[48,130]. Library expressing cells were initially sorted to bottleneck the number of variants.

Specifically, 2,000 cells each from the 10%, 50% and 90% were selected to capture a range of brightness. The sorted and expanded cells were incubated in either 0- or 25-mM glucose and sorted into 4 variable-width bins each containing approximately 25% of the distribution (Fig. 1A). Following the bin-sort, cells were separately expanded, and genomic DNA was extracted for high-throughput sequencing.

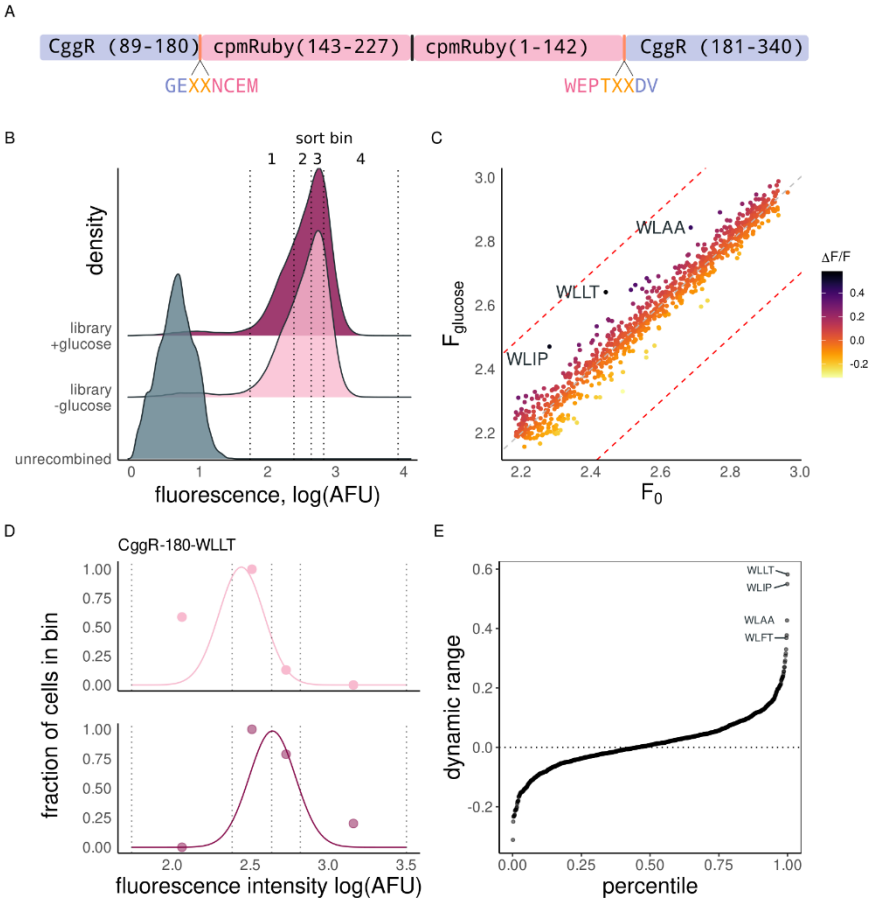

Figure 1. Development of an FBP biosensor using the red fluorescent protein cp-mRuby. (A) Circularly permuted mRuby was inserted into CggR at residue 180 with two flanking linker amino acids on either side. Linker amino acids were encoded by the degenerate codon NNK, which includes all 20 amino acids. **(B)** Fluorescence distributions of the CggR-180-cp-mRuby-NNK library expressed in HEK293T cells exposed to 0mM (light pink) or 25mM glucose (dark pink). Dotted lines indicate the bins used for sort-seq. **(C)** Estimates of brightness in the glucose starved (F_0) and glucose-fed ($F_{glucose}$) samples were used to calculate dynamic range ($\Delta F/F$) for 906 linker variants. **(D)** The number of cells sorted into each bin (points) along with the maximum likelihood density estimates (lines) for the variant with the highest dynamic range, CggR-180-WLLT. **(E)** Dynamic range ($\Delta F/F$)

estimates for all variants ordered by rank with the variants containing the amino acid pair WL as the N-terminal linker labelled.

A maximum likelihood approach was used to estimate mean the fluorescence from the sequencing read counts (Fig. 2C)[45,129]. Dynamic range ($\Delta F/F = (F_{glucose} - F_0) / F_0$) was calculated from the relative change in brightness between the starved (F_0) and high ($F_{glucose}$) samples (Fig. 1B). Estimates for 906 variants were produced in total, with most variants exhibiting no change in fluorescence. The largest glucose-induced increases ($\Delta F/F > 0.4$) were observed for 3 variants, WLIP, WLLT, and WLAA, each containing the same N-terminal linker pair.

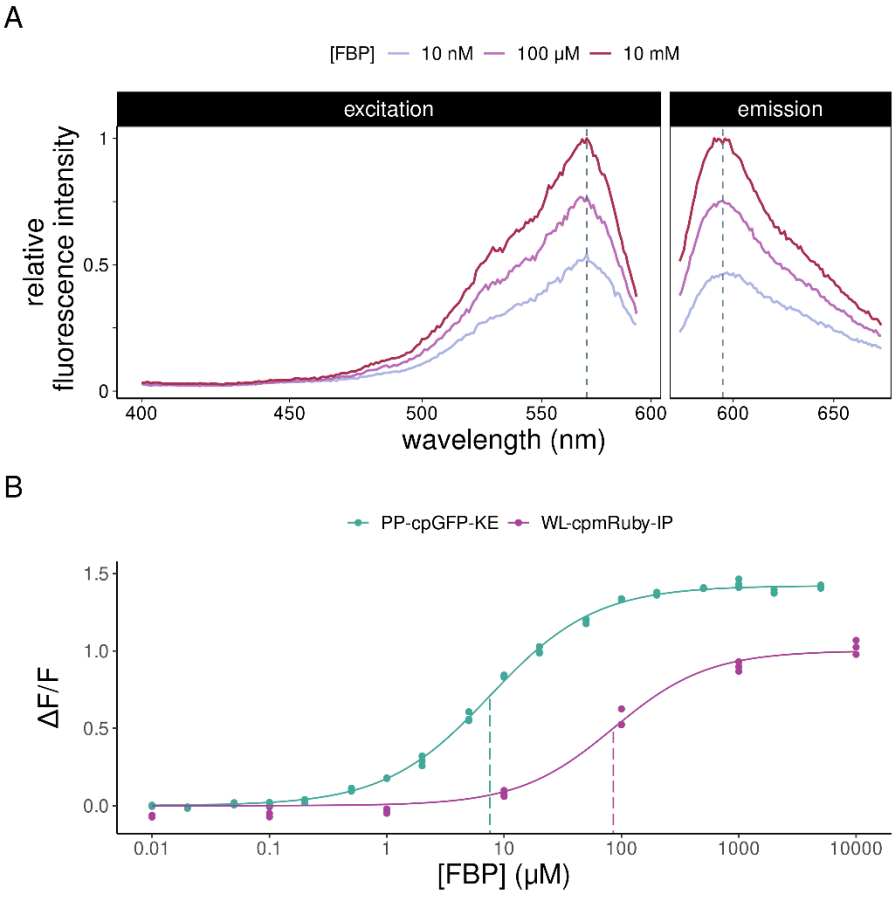

Figure 2. In vitro characterization of CggR-180-WL-cpmRuby-IP. (A) Excitation and emission spectra reveal peaks at 570 and 595 nm respectively. **(B)** The relative change in fluorescence intensity in response to increasing concentration of FBP for CggR-180-WL-cpmRuby-IP (purple, $\Delta F/F$ = 1.0, 560 nm excitation) and CggR-180-PP-cpGFP-KE (teal, $\Delta F/F$ = 1.5, 488 nm excitation). The affinity is significantly reduced for the cpmRuby-based biosensor compared to the cpGFP-based biosensor (K_d = 80 µM compared to 10 µM respectively, indicated by dashed lines).

The variant CggR-180-WL-cpmRuby-IP was purified and characterized *in vitro*. The excitation and emission spectra exhibited maxima at 570 and 595 nm respectively. As observed for RCaMP1a, the excitation and emission spectra are slightly red-shifted relative to mRuby (558/590 nm ex/em)[173]. The *in vitro* the dynamic range ($\Delta F/F$ = 1.0) in response to saturating FBP was higher than the glucose induced changes measured by sort-seq ($\Delta F/F$ = 0.55). The apparent affinity for FBP (K_d = 80 µM) was substantially reduced compared to the previously developed green FBP biosensor HYlight (K_d = 10 µM). The decrease in binding strength might be a general feature caused by cpmRuby insertion at this site or could be specifically caused by the linkers used in this variant. It is not clear how this change in affinity will affect function given that the intracellular concentration of FBP has not been conclusively determined. When expressed in HEK293 cells and evaluated by microscopy, this variant occasionally exhibited decreases in intensity over time (data not shown). This observed non-specific decrease might reflect bleaching of the fluorophore or cellular acidification due to toxicity from biosensor overexpression. These new red FBP biosensors represent a promising starting point with considerable dynamic range *in vitro* but require further validation before they can be confidently used in cells.