

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

John N. Koberstein,* Melissa L. Stewart, Taylor L. Mighell, Chadwick B. Smith, and Michael S. Cohen

Cite This: *ACS Chem. Biol.* 2021, 16, 1709–1720

Read Online

ACCESS |



Metrics & More

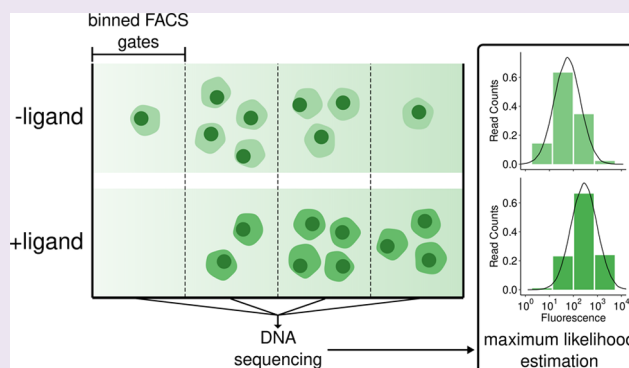


Article Recommendations



Supporting Information

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: the 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 sequences. This high-throughput approach will accelerate the design and optimization of SFPBs, expanding the biosensor toolbox.



INTRODUCTION

Genetically encoded single fluorescent protein biosensors (SFPBs) can unmask aspects of cellular signaling and metabolism that cannot be detected using traditional biochemical approaches.^{1,2} SFPBs can provide crucial information on the subcellular compartmentalization of analytes, resolve changes in concentration over time, and highlight cellular heterogeneity.^{2–7} 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.^{3,8–12} The nature of this allosteric domain coupling is not well understood and has proven difficult to design rationally.¹² 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.¹⁰ 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.¹² 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.^{13–18} In the simplest sort-seq experimental

Received: June 4, 2021

Accepted: August 12, 2021

Published: August 25, 2021



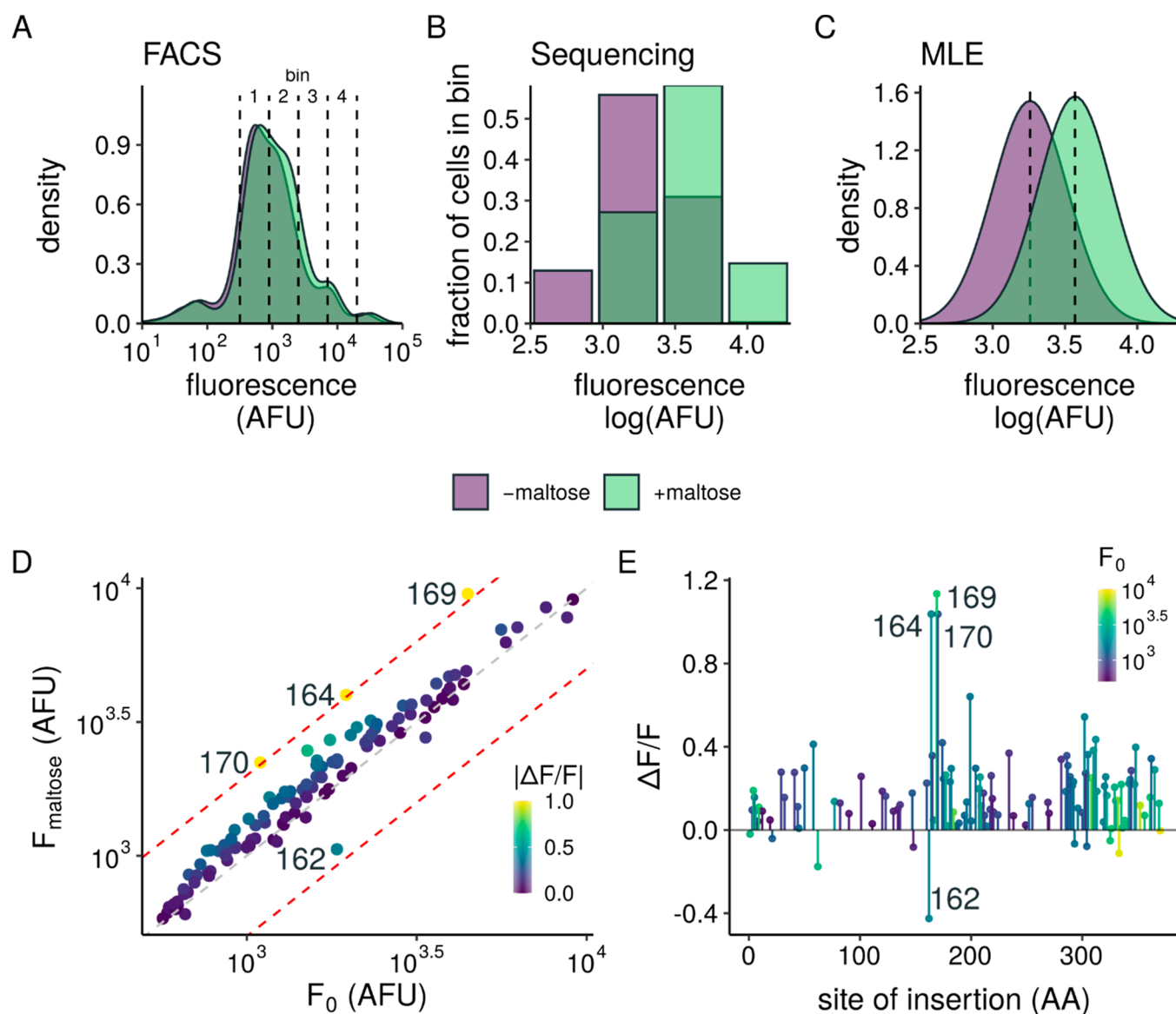


Figure 1. Sort-seq assay of an MBP domain-insertion library. (A) Fluorescence distributions for the MBP domain-insertion library $\pm$ 1 mM maltose. Cells were sorted into the four bins indicated by dashed lines. (B) Sequencing reads are used to estimate the relative fraction of cells containing a given 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 the 113 MBP domain-insertion variants in the presence (F_{maltose}) and absence (F_0) of 1 mM maltose. Colors represent the absolute value of $\Delta F/F$. Red and gray dashed lines indicate $|\Delta F/F| = 1.0$ and 0.0 , respectively. (E) Profile of $\Delta F/F$ estimates for insertions along the MBP primary sequence. Colors indicate the baseline brightness for each insertion site.

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.¹⁹ 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 ($\Delta F/F$)).

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.¹² 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.

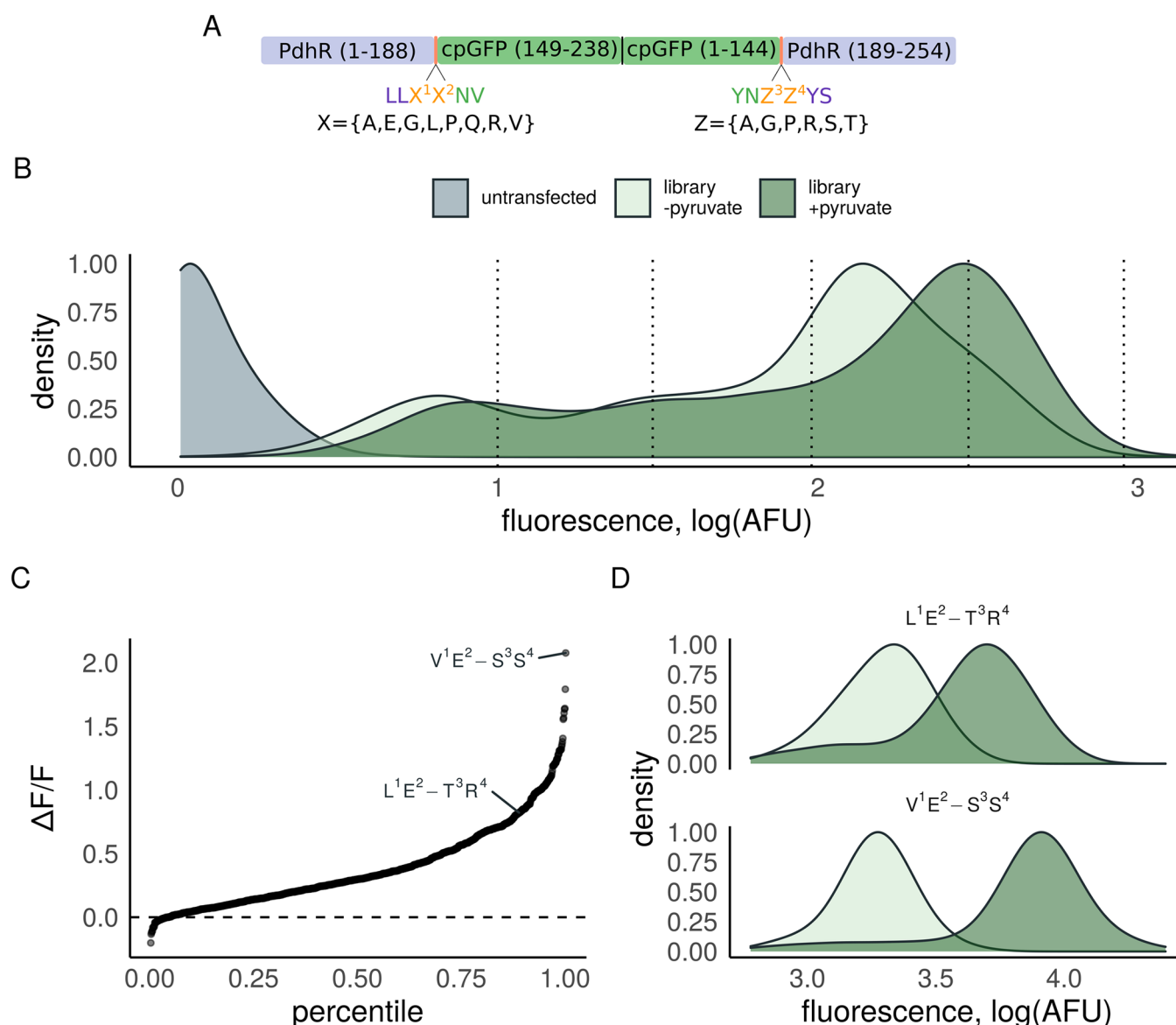


Figure 2. Sort-seq assay of a PyronicSF linker library. (A) Schematic depicting the domain organization of PyronicSF (top) and the designed linker library (bottom). (B) Fluorescence distributions for the PyronicSF linker library expressed in HEK293T Landing Pad cells with (dark green) or without (light green) 10 mM pyruvate. Cells were sorted into the four bins indicated by dashed lines. (C) Dynamic range estimates for 1023 PyronicSF linker variants ordered from lowest to highest $\Delta F/F$. The original linkers ($L^1E^2-T^3R^4$) and top variant ($V^1E^2-S^3S^4$) are labeled. (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) 20 mM pyruvate. Dynamic range is increased from $\Delta F/F = 1.15$ for the original linkers to $\Delta F/F = 2.99$ for variant $V^1E^2-S^3S^4$, validating the estimates from the sort-seq assay.

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.¹⁹ 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 library¹² 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 biosensor²⁰ 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.

RESULTS AND DISCUSSION

Sort-Seq Assay of an MBP Domain-Insertion Library.

A library of circularly permuted GFP (cpGFP) insertions into MBP¹² 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 (Figure 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.). An initial enrichment sort was performed to select for fluorescent variants above a threshold defined by nonfluorescent cells. This sort removes the nonproductive (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, Figure S1B and Data S1), and overall consistency with previous results¹² (Spearman's $\rho = 0.6$, Figure 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.

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 (Figure 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 (Figure 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 (Figure 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 ($\Delta F/F = (F_{\text{maltose}} - F_0)/F_0$). 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.¹⁹ In all, after filtering for variants with sufficient abundance and expected variance (Figure S2A), estimates of brightness and dynamic range for 113 variants out of a possible 370 were obtained (Figure 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 ($|\Delta F/F| < 0.3$) illustrating the relative rarity of allosterically coupled variants. Internal positive controls MBP-169 ($\Delta F/F = 1.13$) and MBP-170 ($\Delta F/F = 1.04$) 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 ($\Delta F/F = 1.04$). In contrast, insertion at MBP-162 produced a biosensor with the opposite response, exhibiting a decrease in fluorescence with the addition of maltose ($\Delta F/F = -0.43$). 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 (Figure 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 (Figure 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 (Figure S4). Such large disparities in variant abundance are common in many massive mutagenesis libraries. The ability to estimate the 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.^{10,12,21} 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.²⁰ The linker and FP sequences of PyronicSF (Leu¹Glu²-cpGFP-Thr³Arg⁴, referred to as L¹E²-T³R⁴) are derived from the well-characterized SFBP GCaMP3²² and were not optimized for PyronicSF (Figure 2A). We hypothesized that tailoring the linkers to the insertion context would yield improvements to the dynamic range. A library consisting of 2304 unique linker combinations was generated by substituting residues in linkers 1 (L¹E²) and 2 (T³R⁴) with amino acids from the sets {A, E, G, L, P, Q, R, V} and {A, G, P, R, S, T}, respectively (Figure 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 line²³ 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 10 mM pyruvate (Figure 2B). After filtering for variants with sufficient abundance and expected variance (Figure S2B), sort-seq estimates of dynamic range were obtained for 1023 unique linker combinations (Data S3). 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 $\Delta F/F = -0.20$ to 2.08 (Figure 2C). The PyronicSF parent sequence, L¹E²-T³R⁴, served as an internal positive control and was found to have a dynamic range ($\Delta F/F = 0.82$) greater than 88% of tested variants. The variant with the highest estimated dynamic-range, V¹E²-S³S⁴ ($\Delta F/F = 2.08$), was constructed and tested by flow cytometry alongside the original L¹E²-T³R⁴ linkers confirming the increase in dynamic range observed in the sort-seq measurements (Figure 2D). Most single substitutions to the parent sequence (21/24) resulted in a decrease in dynamic range with the substitution of Leu¹ with Val¹ producing the largest increase in function (Figure 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 $V^1E^2S^3S^4$, the measured dynamic range is much greater than the linear sum of the effects of the comprising single substitutions (Figure S5; $\Delta F/F = 2.08$ compared to $\Delta F/F = 0.57$ assuming additivity). Testing substitution effects at each position in serial, as is common in directed evolution experiments,²⁴ 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 a protein sequence can assist in elucidating the biochemical determinants of function and predict additional sequences to test.^{25–30} To this end, the PyronicSF linker sequences were encoded as numerical vectors using the VHSE amino acid descriptor (right principal components score vectors derived from hydrophobic, steric, and electronic properties).³¹ These biochemical encodings of linker sequences were then used as features to train regression models to predict sort-seq derived $\Delta F/F$ values. Random forest^{32–34} was found to outperform other common models when evaluated by 5-fold cross-validation using Spearman's rank correlation coefficient (ρ) to assess accuracy (Figure S6) and was used for all further analysis. The final random forest model was trained on an 80% split of the data set and prediction performance was tested on the held out 20%. Predictions of test set function correlated with measured dynamic range ($\rho = 0.71$, Figure 3, Data S4) suggesting strong predictive power of the model. 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.³² 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 $L^1E^2T^3R^4$ and $V^1E^2S^3S^4$; Figure 4A). Variants containing glutamic acid at this position exhibited the highest mean dynamic range ($\Delta F/F = 0.66 \pm 0.42$; Figure 4B). Interestingly, glutamic acid at this position is featured in many of the GCaMP sensor designs (GCaMP2, GCaMP3, and GCaMP6m) in which detailed structural^{35,36} and biochemical³⁷ studies have found the negatively charged side chain to preferentially stabilize the protonated state of the fluorophore in the ligand-free conformation, which underlies the fluorescence switching mechanism. In contrast, substituting positively charged arginine at this position resulted in the lowest mean dynamic range ($\Delta F/F = 0.20 \pm 0.16$; Figure 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 ($\Delta F/F = 0.12 \pm 0.12$, Figure 4C), from the other amino acids due to its unique cyclic structure. In contrast, variants containing the less constrained small hydrophobic amino acids alanine ($\Delta F/F = 0.53 \pm 0.37$, Figure 4C) and valine ($\Delta F/F = 0.52 \pm 0.39$, Figure 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 ($\Delta F/F = 1.02 \pm 0.39$, Figure 4D) followed by V^1E^2 ($\Delta F/F = 0.95 \pm 0.49$, Figure 4D). While the features of the C-terminal linker were found to be less important for prediction accuracy (Figure 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 (Figure 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 (Figure S8A). This result demonstrates that 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 toward increased function (Figure S8B). By combining sort-seq assays with machine learning guided library design, we can increase both the number of variants tested as well as the likelihood a given variant will exhibit increased function.

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.

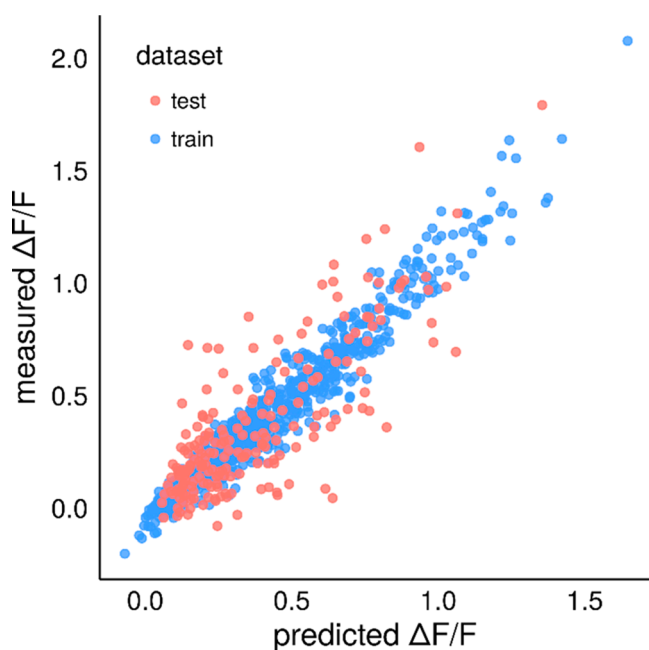


Figure 3. Dynamic range can be predicted from linker sequence. PyronicSF linker variants were numerically encoded using VHSE biochemical descriptors. A random forest model was trained using an 80% split of the encoded linker data (training set, blue) to predict dynamic range. Model performance was evaluated on the held out 20% (test set, red). A strong correlation ($\rho = 0.71$) was observed between model predictions and measured $\Delta F/F$.

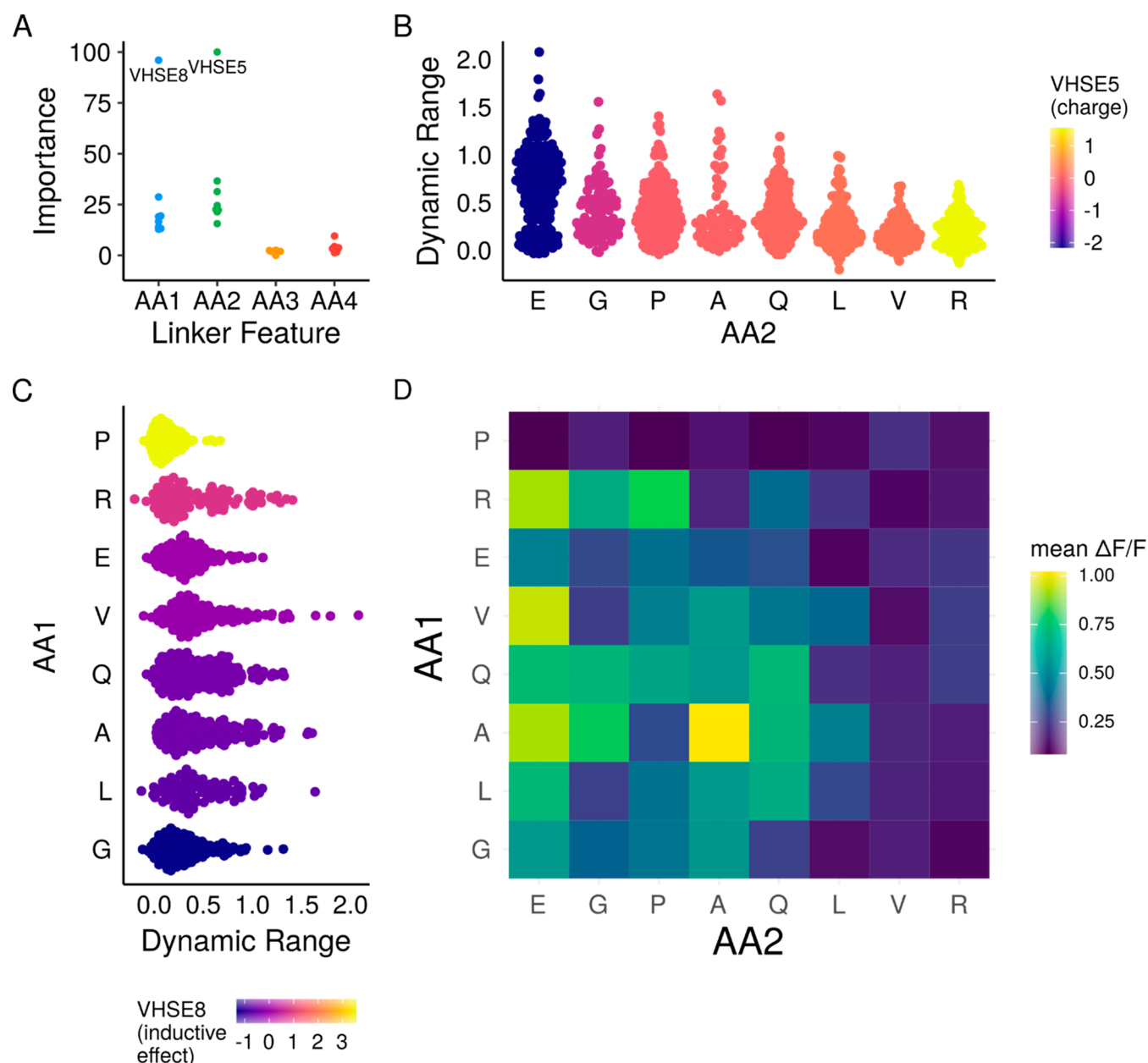


Figure 4. Analysis of the biochemical basis of linker function. (A) The importance of each feature can be estimated from the impact on model importance. Two features of the N-terminal linker amino acids, VHSE8 of the first and VHSE5 of the second amino acid, are the most important. (B) Distributions of variant dynamic-range estimates sorted by VHSE5 of second N-terminal linker amino acid. Negatively charged glutamic acid exhibits the highest mean $\Delta F/F$, 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 .

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.²² 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.²³ Despite the potential advantages, live cell FACS assays may not be suitable for all biosensor/analyte pairs as they require the ability of the measured signal to be manipulated and sustained for the duration of the sort. For a ligand like pyruvate which readily enters HEK293 cells,²⁰ 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 12 μM difference between astrocyte cytosolic and mitochondrial pyruvate concentrations, are relatively small compared to the sensitive range of the biosensor (approximately 0.1 mM to 10 mM).²⁰ The 160% improvement in dynamic range resulting from our assay of PyronicSF linker variants is an important step toward 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 ($K_d = 480 \mu\text{M}$) to better match the intracellular pyruvate concentration (33 μM)²⁰ will be necessary to make this tool generally useful.

In addition to the improved variants identified by this assay, the resulting data set 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 a linker sequence, identifying biochemical features contributing to biosensor function and designing new libraries biased toward 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 methods³⁸ 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.

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). The 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 5 $\times$ MuA reaction buffer, and 1 μL of 0.22 $\mu\text{g}/\mu\text{L}$ 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. coli* 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 $\mu\text{g}/\text{mL}$) and chloramphenicol (25 $\mu\text{g}/\text{mL}$) 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 $\mu\text{g}/\text{mL}$ carbenicillin and 25 $\mu\text{g}/\text{mL}$ chloramphenicol. Cultures were grown at 250 rpm, at 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 fmol pTKEI-Dest, 40 fmol of purified MBP-transposon, 10 units of Esp3I (NEB), 800 units of T4 DNA ligase (NEB), and 1 $\times$ 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. Reactions were purified using a DNA Clean & Concentrator-5 Kit and eluted with 6 μL of water. Transformation was performed using 2 μL of reaction in 25 μL 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 975 μL of Recovery Media and shaken at 250 rpm for 1 h at 37 °C. A 10 μL aliquot of transformed cells was plated on LB agar plates containing 50 $\mu\text{g}/\text{mL}$ kanamycin and 25 $\mu\text{g}/\text{mL}$ 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 6 mL of LB with 1% glucose, 50 $\mu\text{g}/\text{mL}$ of kanamycin, and 25 $\mu\text{g}/\text{mL}$ of chloramphenicol. Cultures were grown at 250 rpm at 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 fmol of pTKEI-MBP-transposon, 40 fmol of purified cpGFP, 10 units of Esp3I (NEB), 800 units of T4

DNA ligase (NEB), and 1 × T4 DNA Ligase Reaction Buffer (NEB) in a total volume of 20 μ L. 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. Reactions were purified using a DNA Clean & Concentrator-5 Kit and eluted with 6 μ L of water. Transformation was performed using 1 μ L of reaction in 40 μ L of Tuner DE3 cells (Novagen) by heat shock at 42 °C for 30 s. Cells were immediately resuspended in 975 μ L of Recovery Media and shaken at 250 rpm for 1 h at 37 °C. A 10 μ L aliquot of transformed cells was plated on LB agar plates containing 50 μ g/mL kanamycin and 25 μ g/mL chloramphenicol to select for pTKEI-Dest backbone and transposon insertions to assess the remaining transposon. The remaining culture was pelleted and resuspended in 6 mL of LB with 1% glucose and 25 μ g/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 h prior to sorting, a 100 μ L aliquot of the library culture to be sorted as well as FACS controls (pTKEI-MBP and pTKEI-cpGFP) were added to 5 mL of MOPS EZ Rich Defined Medium (Teknova) containing 60 μ g/mL kanamycin and 0.4% glycerol. Cultures were shaken at 250 rpm at 37 °C for approximately 2 h until the OD₆₀₀ was between 0.6 and 0.8. Expression was induced by adding 0.5 mM IPTG followed by incubation for 4 h. Samples were transported on ice to the core facility for sorting and incubated an additional 30 min at 37 °C at 250 rpm. Immediately before starting the 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 488 nm laser for excitation and a 530/30 nm emission filter for GFP measurements and a 561 nm laser with a 610/20 filter for mCherry measurements. To minimize the sorting of aggregated cells, *E. coli* expressing either GFP or mCherry was 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 15 mL conical tube containing 5 mL 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 h at 37 °C at 250 rpm. The recovered cultures were diluted to 15 mL of LB with the addition of 1% glucose. An aliquot was diluted 1:50 and spread on LB agar plates containing 50 μ g/mL of kanamycin to assess cell survival. The remaining culture was incubated with 50 μ g/mL of kanamycin overnight at 30 °C at 250 rpm. 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 h offset to account for time taken to sort the first sample. After transporting the first induced sample to the core, 1 mM maltose was added before incubating an additional 30 min. Immediately before starting the sort, the samples were diluted 1:100 in phosphate-buffered saline (PBS) or 1:100 in PBS with added 1 mM 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 $\pm$ maltose samples were sorted using the same gates for 1.5 h each. The cells for each bin were collected in 5 mL tubes containing 1 mL of LB with 1% glucose. A range of 4200 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). Then, 50 μ L reactions were prepared with a final concentration of 0.2 ng/ μ L of plasmid DNA, 0.25 μ M 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 AMPure XP beads and eluted in 40 μ L of elution buffer (5 mM Tris/HCl, pH 8.5).

Amplicons were fragmented and tagged (tagmented) in a 25 μ L reaction containing 1 ng/ μ L amplicon, 1 × TD buffer, and 0.5 μ L TDE1 enzyme from the Nextera DNA Sample Prep Kit (Illumina). Tagmentation reactions were cleaned up using a Nucleospin column and eluted in 15 μ L of elution buffer. Tagmented DNA was amplified with primer i5-Nex2p and a unique indexed primer per sample (i7-TXX-NEX2p). The 25 μ L reactions were prepared containing 1 μ L of tagmented DNA, 0.5 μ M forward and reverse primer, 1 × KAPA HiFi Hotstart Readymix (KHF), and 1 × SYBR Green. Amplification was monitored by qPCR with cycling conditions: [72 °C 3 min, 95 °C 20 s, (98 °C 20 s, 52 °C 15 s, 72 °C 30 s, plate read, 72 °C 8 s) × 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 2 × 75bp paired-end reads on an Illumina MiSeq (v3 Reagent kit).

MBP Domain-Insertion Library Enrichment Analysis. Paired end reads were merged using BBMerge.³⁹ MBP-cpGFP insertion sites were counted using the dipseq analysis pipeline developed by the Savage lab available at (<https://github.com/SavageLab/dipseq>).¹² 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 derived 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 $E_j = \log_2(f_i^{\text{sort}}/f_i^{\text{naive}})$.

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. The 20 μ L Gibson assembly reactions were prepared with 50 ng of backbone, a 5:1 molar ratio of insert to backbone, and 1 × Gibson Assembly Master Mix (NEB) and incubated at 50 °C for 1 h before transformation into One Shot TOP10 Chemically Competent *E. coli* (Invitrogen). Single colonies were picked from LB agar plates with 50 μ g/mL of 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 μ L of LB supplemented with 10% glycerol, 1% glucose, and 60 μ g/mL kanamycin. Six wells were inoculated for each variant to accommodate three replicates for two conditions ($\pm$ ligand). Plates were incubated overnight at 30 $^{\circ}$ C at 300 rpm. The following day, 5 μ L of culture was used to inoculate 200 μ L of MOPS EZ Rich Defined Medium (Teknova) containing 60 μ g/mL of kanamycin and 0.4% glycerol in a clear-bottom, black, 96-well plate. The plate was incubated at 37 $^{\circ}$ C and 300 rpm in a Molecular Devices i3 plate reader while measuring OD600 and fluorescence (485/5 nm excitation and 515/5 nm emission) every 600s for 2.5 h. Upon OD600 exceeding 0.6 for all wells, expression was induced by adding 0.5 mM IPTG. Following 2 h of incubation to allow for adequate protein expression, either 1 mM maltose in PBS or an equal volume of PBS alone was added to each well. After adding the ligand, wells were incubated and monitored for another 1 h. 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 the 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 four 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 10 μ L reaction containing 13 pmol of each PCR fragment and holding plasmid, 10 units of BsaI-HFv2 (NEB), 100 units of T4 DNA ligase (NEB), and 1 $\times$ T4 DNA ligase buffer (NEB). The reaction was incubated at 37 $^{\circ}$ C for 5 min and 16 $^{\circ}$ C for 10 min for 40 cycles, followed by 16 $^{\circ}$ C for 20 min, 60 $^{\circ}$ C for 30 min, and 75 $^{\circ}$ C for 6 min 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 6 μ L of water. The library was transformed by electroporation using 1 μ L of cleaned reaction in 25 μ L of *E. coli* 10G ELITE cells. A 5 μ L aliquot of recovered transformation was plated on LB agar containing 50 μ g/mL kanamycin. The remaining recovered culture was diluted into 50 mL of LB containing 50 μ g/mL kanamycin and grown overnight at 37 $^{\circ}$ C at 250 rpm. Transformation produced well over 20 000 unique transformants, enough to cover 2304 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.⁴⁰ The PyronicSF linker library was cloned into EMMA-attB-Dest using an Esp3I mediated Golden Gate reaction, a 10 μ L reaction containing 13 pmol of the PyronicSF linker library in the holding plasmid, 13 pmol of EMMA-attB-Dest, 5 units of Esp3I (NEB), 100 units of T4 DNA ligase (NEB), and 1 $\times$ T4 DNA ligase buffer (NEB). The reaction was incubated at 37 $^{\circ}$ C for 5 min and 16 $^{\circ}$ C for 10 min for 40 cycles, followed by 16 $^{\circ}$ C for 20 min, 60 $^{\circ}$ C for 30 min, and 75 $^{\circ}$ C for 6 min. The Golden Gate reaction was cleaned using a DNA Clean & Concentrator-5 Kit, eluted in 6 μ L of water and transformed by electroporation using 1 μ L of the reaction added to 25 μ L of *E. coli* 10G ELITE cells. Recovered cells were diluted into 50 mL of LB with 100 μ g/mL carbenicillin and grown overnight at 37 $^{\circ}$ C

at 250 rpm. 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 lab²³ 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) and 25 mM HEPES. Cultures were maintained with 2 μ g/mL doxycycline (Sigma-Aldrich), which was removed 1 day prior to transfection. For each transfection, a total of 3 μ g of plasmid DNA and 6 μ L of FuGENE 6 (Promega) were combined in 300 μ L of 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 six-well plate containing 1×10^6 freshly seeded cells per well. Cells were incubated with transfection reagents for 2 days before expanding each well to a 15 cm plate. After 1 day of growth in the 15 cm plate, 2 μ g/mL of 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 15 cm plate were detached by trypsinization, pelleted, and resuspended in 2 mL of DMEM supplemented with 50 μ g/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 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 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 15 cm tube containing 5 mL of DMEM supplemented with 10% FBS and 50 μ g/mL gentamicin. Sorted cells were pelleted, resuspended in 800 μ L of media, and plated in a 12-well plate. Cells were expanded over about 7 days to reach 50% confluence in two 15 cm plates before the second round of sorting.

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 μ g/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) from 1.0 to 3.0. The $\pm$ pyruvate 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 μ g/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 six-well plate (>900 000 cells). All samples were expanded to 50% in a 10 cm plate before harvesting and storing at -20° C after pelleting and washing with PBS.

PyronicSF Linker Library Sequencing. Genomic DNA was extracted from approximately 5 M 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 50 μ L reactions were prepared for each sample with a final concentration of 5 ng/ μ L genomic DNA, 0.25 μ M forward and reverse primer, 1 $\times$ SeqAmp PCR buffer, 1 $\times$ SeqAmp Polymerase (Clontech), and 1 $\times$ SYBR Green (Invitrogen). Amplification was monitored by qPCR with cycling conditions: [94 $^{\circ}$ C 60s, (98 $^{\circ}$ C 10s,

55 °C 15s, 68 °C 60s, plate read) $\times$ 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 15 μ L of 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). Then, 25 μ L reactions were prepared containing 1 μ L of round 1 DNA, 0.5 μ M forward and reverse primer, 1 $\times$ KAPA HiFi Hotstart Readymix (KHF), and 1 $\times$ SYBR Green. Amplification was monitored by qPCR with cycling conditions: [95 °C 3 min, (98 °C 20 s, 60 °C 15 s, 72 °C 30 s, plate read, 72 °C 8 s) $\times$ 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 933 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 $\times$ 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. 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), and allowed degenerate codons were counted using the ShortRead R/Bioconductor package.⁴¹ The resulting table of read counts for each linker sequence in each sample was used as an 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.^{16,19} 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 were 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_i relating read counts to the total number of sorted cells in each bin b_j is set as

$$d_i = \left(\sum_j b_j \right) / \left(\sum_i r_{ij} \right)$$

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

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

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 $F = 10^{(\hat{\mu} + \hat{\sigma}_i/2)}$. Mean fluorescence was calculated for both samples and then used to calculate the dynamic range as $\frac{\Delta F}{F_0} = (F_i - F_0)/F_0$ where F_i 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 data set 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 primer 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. Then, 20 μ L of Gibson assembly reactions were prepared with 50 ng of backbone, a 5:1 molar ratio of insert to backbone, and 1 $\times$ Gibson Assembly Master Mix (NEB) and incubated at 50 °C for 1 h before transformation into One Shot TOP10 Chemically Competent *E. coli* (Invitrogen). Single colonies were picked from LB agar plates with 50 μ g/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 linker, EMMA-attB-PyronicSF-LE-TR, was recombined into Landing Pad cells. Recombined cells grown to approximately 50% confluency in a six-well plate were detached by trypsinization, pelleted, and resuspended in 400 μ L of DMEM. Cytometric evaluation was performed on a BD LSR II equipped with a 488 nm laser for excitation and a 530/30 nm emission filter for GFP measurements and 405 nm laser with a 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 the loss of BFP fluorescence. GFP intensity for recombined cells was analyzed in the presence and absence of 20 mM pyruvate.

Regression Models. PyronicSF linker sequences were converted to 32-length numerical vectors by encoding each of the four amino acids as the eight representative VHSE biochemical features. Model training was performed using the R package caret⁴² with the random forest model implemented in ranger.⁴³ An 80% split (819 variants) was used for model training, and the remaining 20% (204 variants) was used for testing the final model. A 5-fold 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 hyperparameters 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; $\lambda = 0, 1 \times 10^{-4}, 1 \times 10^{-3}, 0.1$, and 1 for ridge regression; and $\sigma = 0, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 0.1, 0.5$, and 1 for Gaussian process using the radial basis function kernel.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acschembio.1c00423>.

Oligo/DNA sequences used for cloning, Sequences for biosensor constructs (PDF)

Sort-seq data sets (XLSX)

Accession Codes

Raw sequencing reads associated with both the MBP and PyronicSF sort-seq experiments can be accessed from the NCBI

Sequence Read Archive (SRA) with the accession code PRJNA732942.

AUTHOR INFORMATION

Corresponding Author

John N. Koberstein – Vollum Institute, Oregon Health & Science University, Portland, Oregon 97239, United States; orcid.org/0000-0003-0622-5289; Email: koberste@ohsu.edu

Authors

Melissa L. Stewart – Vollum Institute, Oregon Health & Science University, Portland, Oregon 97239, United States
Taylor L. Mighell – Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, Oregon 97239, United States; Present Address: Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain
Chadwick B. Smith – Vollum Institute, Oregon Health & Science University, Portland, Oregon 97239, United States
Michael S. Cohen – Department of Physiology and Pharmacology, Oregon Health & Science University, Portland, Oregon 97239, United States; orcid.org/0000-0002-7636-4156

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acscchembio.1c00423>

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.

Funding

This work was supported by NIH Grant 1R01AG055431-01. J.N.K. received support through a training grant supported by the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health under award number T32NS007466. J.N.K. is an ARCS scholar (Achievement Rewards for College Scientists Foundation, Inc., Oregon Chapter).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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.

REFERENCES

(1) Greenwald, E. C.; Mehta, S.; Zhang, J. Genetically Encoded Fluorescent Biosensors Illuminate the Spatiotemporal Regulation of Signaling Networks. *Chem. Rev.* **2018**, *118* (24), 11707–11794.
(2) Okumoto, S.; Jones, A.; Frommer, W. B. Quantitative Imaging with Fluorescent Biosensors. *Annu. Rev. Plant Biol.* **2012**, *63* (1), 663–706.
(3) Arce-Molina, R.; Cortés-Molina, F.; Sandoval, P.; Galaz, A.; Alegría, K.; Schirmeier, S.; Barros, L.; Martín, A. S. A Highly Responsive Pyruvate Sensor Reveals Pathway-Regulatory Role of the Mitochondrial Pyruvate Carrier MPC. *bioRxiv* **2019**, 611806.

(4) Hou, B. H.; Takanaga, H.; Grossmann, G.; Chen, L. Q.; Qu, X. Q.; Jones, A. M.; Lalonde, S.; Schweissgut, O.; Wiechert, W.; Frommer, W. B. Optical Sensors for Monitoring Dynamic Changes of Intracellular Metabolite Levels in Mammalian Cells. *Nat. Protoc.* **2011**, *6* (11), 1818–1833.
(5) Miyawaki, A.; Niino, Y. Molecular Spies for Bioimaging-Fluorescent Protein-Based Probes. *Mol. Cell* **2015**, *58*, 632–643.
(6) Ni, Q.; Mehta, S.; Zhang, J. Live-Cell Imaging of Cell Signaling Using Genetically Encoded Fluorescent Reporters. *FEBS J.* **2018**, *285* (2), 203–219.
(7) Pendin, D.; Greotti, E.; Lefkimiatis, K.; Pozzan, T. Exploring Cells with Targeted Biosensors. *J. Gen. Physiol.* **2017**, *149* (1), 1–36.
(8) Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. Circular Permutation and Receptor Insertion within Green Fluorescent Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (20), 11241–11246.
(9) Doi, N.; Yanagawa, H. Design of Generic Biosensors Based on Green Fluorescent Proteins with Allosteric Sites by Directed Evolution. *FEBS Lett.* **1999**, *453* (3), 305–307.
(10) Marvin, J. S.; Schreiter, E. R.; Echevarría, I. M.; Looger, L. L. A Genetically Encoded, High-Signal-to-Noise Maltose Sensor. *Proteins: Struct., Funct., Genet.* **2011**, *79* (11), 3025–3036.
(11) Marvin, J. S.; Borghuis, B. G.; Tian, L.; Cichon, J.; Harnett, M. T.; Akerboom, J.; Gordus, A.; Renninger, S. L.; Chen, T.-W.; Bargmann, C. I.; Orger, M. B.; Schreiter, E. R.; Demb, J. B.; Gan, W.-B.; Hires, S. A.; Looger, L. L. An Optimized Fluorescent Probe for Visualizing Glutamate Neurotransmission. *Nat. Methods* **2013**, *10* (2), 162–170.
(12) Nadler, D. C.; Morgan, S.-A.; Flamholz, A.; Kortright, K. E.; Savage, D. F. Rapid Construction of Metabolite Biosensors Using Domain-Insertion Profiling. *Nat. Commun.* **2016**, *7*, 12266.
(13) Kinney, J. B.; Murugan, A.; Callan, C. G.; Cox, E. C. Using Deep Sequencing to Characterize the Biophysical Mechanism of a Transcriptional Regulatory Sequence. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (20), 9158–9163.
(14) Kosuri, S.; Goodman, D. B.; Cambray, G.; Mutalik, V. K.; Gao, Y.; Arkin, A. P.; Endy, D.; Church, G. M. Composability of Regulatory Sequences Controlling Transcription and Translation in *Escherichia Coli*. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (34), 14024–14029.
(15) Noderer, W. L.; Flockhart, R. J.; Bhaduri, A.; Diaz De Arce, A. J.; Zhang, J.; Khavari, P. A.; Wang, C. L. Quantitative Analysis of Mammalian Translation Initiation Sites by FACS-Seq. *Mol. Syst. Biol.* **2014**, *10*, 748.
(16) Peterman, N.; Lavi-Itzkovitz, A.; Levine, E. Large-Scale Mapping of Sequence-Function Relations in Small Regulatory RNAs Reveals Plasticity and Modularity. *Nucleic Acids Res.* **2014**, *42* (19), 12177–12188.
(17) Sharon, E.; Kalma, Y.; Sharp, A.; Raveh-Sadka, T.; Levo, M.; Zeevi, D.; Keren, L.; Yakhini, Z.; Weinberger, A.; Segal, E. Inferring Gene Regulatory Logic from High-Throughput Measurements of Thousands of Systematically Designed Promoters. *Nat. Biotechnol.* **2012**, *30* (6), 521–530.
(18) Sharon, E.; van Dijk, D.; Kalma, Y.; Keren, L.; Manor, O.; Yakhini, Z.; Segal, E. Probing the Effect of Promoters on Noise in Gene Expression Using Thousands of Designed Sequences. *Genome Res.* **2014**, *24* (10), 1698–1706.
(19) Peterman, N.; Levine, E. Sort-Seq under the Hood: Implications of Design Choices on Large-Scale Characterization of Sequence-Function Relations. *BMC Genomics* **2016**, *17* (1), 206.
(20) Arce-Molina, R.; Cortes-Molina, F.; Sandoval, P. Y.; Galaz, A.; Alegría, K.; Schirmeier, S.; Barros, L. F.; San Martín, A. A Highly Responsive Pyruvate Sensor Reveals Pathway-Regulatory Role of the Mitochondrial Pyruvate Carrier MPC. *eLife* **2020**, *9*, e53917.
(21) Nakai, J.; Ohkura, M.; Imoto, K. A High Signal-to-Noise Ca²⁺ Probe Composed of a Single Green Fluorescent Protein. *Nat. Biotechnol.* **2001**, *19* (2), 137–141.
(22) Tian, L.; Hires, S. A.; Mao, T.; Huber, D.; Chiappe, M. E.; Chalasani, S. H.; Petreanu, L.; Akerboom, J.; McKinney, S. A.; Schreiter, E. R.; Bargmann, C. I.; Jayaraman, V.; Svoboda, K.; Looger, L. L. Imaging Neural Activity in Worms, Flies and Mice with

Improved GCaMP Calcium Indicators. *Nat. Methods* **2009**, *6* (12), 875–881.

(23) Matreyek, K. A.; Stephany, J. J.; Fowler, D. M. A Platform for Functional Assessment of Large Variant Libraries in Mammalian Cells. *Nucleic Acids Res.* **2017**, *45* (11), e102.

(24) Reetz, M. T.; Carballeira, J. D. Iterative Saturation Mutagenesis (ISM) for Rapid Directed Evolution of Functional Enzymes. *Nat. Protoc.* **2007**, *2* (4), 891–903.

(25) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315–1322.

(26) Bedbrook, C. N.; Yang, K. K.; Robinson, J. E.; Mackey, E. D.; Gradinaru, V.; Arnold, F. H. Machine Learning-Guided Channelrhodopsin Engineering Enables Minimally Invasive Optogenetics. *Nat. Methods* **2019**, *16* (11), 1176–1184.

(27) Biswas, S.; Khimulya, G.; Alley, E. C.; Esvelt, K. M.; Church, G. M. Low- N Protein Engineering with Data-Efficient Deep Learning. *Nat. Methods* **2021**, *18* (4), 389–396.

(28) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine Learning-Assisted Directed Protein Evolution with Combinatorial Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (18), 8852–8858.

(29) Xu, Y.; Verma, D.; Sheridan, R. P.; Liaw, A.; Ma, J.; Marshall, N. M.; McIntosh, J.; Sherer, E. C.; Svetnik, V.; Johnston, J. M. Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773.

(30) Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H. Learned Protein Embeddings for Machine Learning. *Bioinformatics* **2018**, *34* (15), 2642–2648.

(31) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs. *Biopolymers* **2005**, *80* (6), 775–786.

(32) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*:1–45, 5–32.

(33) Probst, P.; Wright, M.; Boulesteix, A.-L. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining Knowl. Discov.* **2019**, *9* (3), e1301.

(34) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958.

(35) Akerboom, J.; Rivera, J. D. V.; Rodríguez Guilbe, M. M.; Malavé, E. C. A.; Hernandez, H. H.; Tian, L.; Hires, S. A.; Marvin, J. S.; Looger, L. L.; Schreiter, E. R. Crystal Structures of the GCaMP Calcium Sensor Reveal the Mechanism of Fluorescence Signal Change and Aid Rational Design. *J. Biol. Chem.* **2009**, *284* (10), 6455–6464.

(36) Chen, Y.; Song, X.; Ye, S.; Miao, L.; Zhu, Y.; Zhang, R. G.; Ji, G. Structural Insight into Enhanced Calcium Indicator GCaMP3 and GCaMPJ to Promote Further Improvement. *Protein Cell* **2013**, *4* (4), 299–309.

(37) Barnett, L. M.; Hughes, T. E.; Drobizhev, M. Deciphering the Molecular Mechanism Responsible for GCaMP6m's Ca²⁺-Dependent Change in Fluorescence. *PLoS One* **2017**, *12* (2), e0170934.

(38) Coyote-Maestas, W.; Nedrud, D.; Okorafor, S.; He, Y.; Schmidt, D. Targeted Insertional Mutagenesis Libraries for Deep Domain Insertion Profiling. *Nucleic Acids Res.* **2020**, *48* (2), e11.

(39) Bushnell, B.; Rood, J.; Singer, E. BBMerge – Accurate Paired Shotgun Read Merging via Overlap. *PLoS One* **2017**, *12* (10), e0185056.

(40) 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 Synth. Biol.* **2017**, *6* (7), 1380–1392.

(41) Morgan, M.; Anders, S.; Lawrence, M.; Aboyoun, P.; Pagès, H.; Gentleman, R. ShortRead: A Bioconductor Package for Input, Quality Assessment and Exploration of High-Throughput Sequence Data. *Bioinformatics* **2009**, *25* (19), 2607–2608.

(42) Kuhn, M. Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software* **2008**, *28* (1), 1–26.

(43) Wright, M. N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **2017**, *77* (1), 1–17.