High throughput screening and ML-aided engineering of transcription factor based biosensors

By

Nishit Banka

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2024

Date of final oral examination: 08/22/2024

The dissertation is approved by the following members of the Final Oral Committee:

Philip Romero, Assistant Professor, Biochemistry

Andrew Buller, Assistant Professor, Chemistry

Vatsan Raman, Associate Professor, Biochemistry

Jean-Michel Ané, Professor, Bacteriology

Acknowledgements:

My journey throughout PhD has been a rollercoaster ride. I joined UW-Madison in 2019 right before the covid-19 pandemic shut everything down changing the landscape of how we do research. During the pandemic and the aftermath, I faced several challenges including the passing of a dear friend and needed lots of support to come out stronger on the other end. Science is a collaborative endeavor, and I want to acknowledge the various people who helped me reach this point, but it is in no way an exhaustive list.

I want to thank Phil Romero, my advisor, for giving me the platform and tools to explore my scientific and professional curiosities. I was able to indulge in non-academic focused organizations/programs such as WiSolve and the Morgridge Entrepreneurial Bootcamp. It allowed me to learn about my strengths and interests beyond an academic career. Phil being more hands-off made me a lot more independent towards scientific problem solving. I am glad to have joined the lab and will be part of the extended Romero lab alumni network to support future members.

Living seven thousand miles away from family has been one of the tougher parts of PhD but my family has been my anchor these past years. I want to express my heartfelt gratitude to my mom (Namita Banka), my dad (Sanjay Banka) and my sister (Smriti Banka) for the immense support during the tough times and the joyous celebrations during the good. The entire Banka family for their patience and love as I go through more “schooling”. I want to thank the Kedia family for helping me get settled in the US and make my transition as easy as possible. Back in 2020, my amazing roommate and dear friend, Josh Immke, passed away which changed the trajectory of my life in Madison. I have been

blessed to have the Immke family treat me as one of their own while we grieved through this tragic loss.

My committee members, Andrew Buller and Vatsan Raman were a guiding post when I felt lost in my PhD. Andrew helped me gain a good perspective on how to think about science and very importantly, how to communicate it effectively. One of the chapters in this dissertation was a collaboration with the Jean-Michele Ane lab where I worked with a phenomenal post-doc Biswajit Samal, a mentor I wished I had met earlier in my PhD. One of the funnest part of classes was my time spent in the synthetic biology seminar with Scott Coyle, discussing the landscape of what the field has to offer.

I want to mention my community of friends in Madison that made life here possible. My colleagues in the Romero lab became some of my closest friends. I want to specifically acknowledge Mark Mahnke who joined the lab when I did and rode the roller coaster with me, sometimes literally during our visits to six flags. My roommate Jairo Villalona has been a rock throughout graduate school, a needed constant in my life as science would get frustrating. Finally, my projects were made possible by the various core facilities around UW-Madison. The flow core and NGS center played a pivotal role in ensuring my progress. I am happy to have pursued my PhD at UW-Madison.

Table of Contents

Abstract:

Biosensors are biological polymer based technology that detect an input signal, propagate the signal, and allow the quantification of a measurable output. The input can be small molecules, proteins or environmental stress leading to an output such as electrical impulses, luminescence, or fluorescence. Transcription factors (TF) can be an ideal targets for biosensor development as they natively regulate gene expression under an input stimulus. TFs have been used as biosensors in protein and metabolic engineering, point of care testing and autonomous genetic circuits. Transcription factors are abundant amongst pro- and eukaryotic cells that lends well to genomic mining of naturally occurring TF-input interactions. While there has been increased efforts in discovering new TFs, they are not fully characterized which can limit their use as robust biosensors. To overcome challenges faced by native transcription factors, we can employ synthetic biology tools and engineer their functionality for improved sensitivity or specificity. In this dissertation, we combine the use of high throughput screening, NGS data generation and machine learning guided mutation prediction for accelerated engineering of transcription factors. We developed BmoR^{10mut}, a ten amino acid variant with improved activity towards butyrate in a single round of engineering. We want to utilize this biosensor for gut microbiome therapeutics for regulation of butyrate the gut. Using this engineering workflow, we also improved the activity of CymR for the detection of root exudate, p-cumate, by nearly 10-fold in *klebsiella variicola*. We intend to use the improved variants for inducible control of bacterial nitrogen fixation in *kv* and increased delivery of ammonia to plants.

Chapter 1: Transcription factors and their role as biosensors

Introduction:

Transcription Factors as gene regulators:

The development and maintenance of living organisms relies heavily on their ability to control gene expression. This involves responding to external stimuli in an adaptable manner. This ability to adapt to the environment, whether that is a single celled bacteria or a multicellular human, is essential to the survival of the species. Organisms employ several methods to regulate and respond to the environment, a key feature being the use of genetic transcription factors.

Transcription Factors (TFs) are proteins that regulate the expression of genes often under a specific environmental stimuli. Upon receiving a signal, TFs enable the transcription of DNA to messenger RNA (mRNA) for downstream protein synthesis. This stimulus can be a variety of inputs including small molecules[1], protein-protein interactions [2], phosphorylation and dephosphorylation[3] or environmental stress such as heat [4]. To encompass such a wide range of inputs, organisms have developed methods to utilize TFs for efficient control of gene expression. TFs tend to be highly specific towards their input as well as the stretch of DNA they bind to, called the operator sequence.

Bacterial Transcription factors:

Transcription factors are ubiquitously present in all organisms serving this vital role. In this work, I will mostly focus on prokaryotic, specifically, bacterial transcription factors. Most of the current understanding of bacterial TFs began with the introduction of the operon model by Jacob and Monod in the 1960s which postulates that gene expression is controlled by promoters whose activity fixes the transcription of one or more downstream genes [5]. Monod's work on the expression *Escherichia Coli* lactose operon along with Jacob's research on the behavior of bacteriophage lambda's switching between the lytic and lysogenic cycles established a framework for gene transcription dependent on promoters regulated by transcription factors. Through work done following the operon model, it became apparent that it became apparent that process of DNA supercoiling and the abundance of nucleoid-associated proteins that were thought to bind with low sequence specificity to sculpt and compact bacterial chromosomes, also had a direct role on transcription. Despite its widespread acceptance, the operon model painted a more simplistic picture of DNA transcription and that bacTFs operated in a perfect manner[6]. This led to work spanning over five decades of experimentation and generation of a vast amount of data valuable towards understanding fundamentals of gene expression as well as developing key tools for synthetic biology.

The cyclic AMP receptor protein (CRP) is a well characterized transcription factor in its role in regulating the lac operon along with LacI. This operon is required for the transport and metabolism of lactose in several bacterial species. Most bacteria prefer glucose as a source of energy but in no glucose environments, the lac operon offers an

effective way to metabolize lactose using the enzyme Beta-galactosidase. Under low glucose condition, bacteria produce cyclic adenosine monophosphate (cAMP) which binds to CRP forming the CRP-cAMP complex 60bp upstream of the transcription initiation site. The complex promotes RNA polymerase binding to the lac promoters[7]. This activity is interesting to me since the pathway forms a simple network of metabolite crosstalk in a logic gate like manner. If glucose is absent, only then look for lactose. If lactose is present only then make Beta-galactosidase so the host can survive. It gives us a snapshot of how even single cell organisms have evolved to survive in various stressful environmental conditions.

Structural understanding of TFs:

TF domains and their activity:

Unlike siloed graduate students during PhD, transcription factors interact heavily with other proteins and protein-complexes to enable gene expression. This machinery involves DNA interaction with specific σ factors that recruit the highly conserved RNA polymerases (RNAP) along with various activator proteins forming the RNAP holoenzyme. This complex binds and unwinds the coiled promoter DNA forming the open promoter complex (RPo) which allows the RNAP to transcribe DNA into messenger RNA (mRNA)[8]. TFs function as the initiation signal for the formation of the RPo.

TFs are classified based on their interaction with σ factors. TFs that interact with the housekeeping σ70 are called σ70 dependent TFs while σ^{54} interacting TFs are labeled as σ^{54} dependent Transcription factors[9]. There are other σ factors present in bacterial

cells, however σ^{70} and σ^{54} are the two major classes. This provides a fundamental basis for a class of TFs as it also informs their structure. σ^{70} is the most common σ factor utilized by cells and the largest class of TFs in bacteria[10]. On the other hand, σ^{54} TFs tend to be rarer as they are often employed under stress conditions for the host cells[11]. They are not as well studied as the σ^{70} class of TFs but offer a unique mode of transcription initiation. This gives us insight into how cells can have stricter control over gene expression.

The structural understanding of these bacterial TFs and their associations is crucial to the dissemination of their mechanism of action. Transcription factor activity is broadly divided into either activators, which bind DNA upon receiving a signal, or repressors which unbind DNA upon receiving a signal. In both cases, this change in binding leads to either up- or down-regulation of gene expression. Generally, TFs comprise of two domains: an effector binding or protein-protein interaction domain and a DNA binding domain. The former is involved in signal recognition by binding to the effector which in many cases is a metabolite containing information about the extracellular environment[12]. These effector molecules bind at a distal location near the protein surface leading to conformational change in the TF for downstream activity. This coupled interaction amongst the distal region of the protein to alter function is called allostery and the TFs labeled as allosteric transcription factors (aTFs)[13]. The DNA binding domain contains the conserved helix-turn-helix motif that binds directly to the operator region of the DNA[14,15].

In this discussion we focus on proteins that contain both these domains and are singularly involved in both functions of signal propagation. However, bacterial cells also contain two-component systems which comprise of two distinct proteins, one involved in

signal reception and the other nucleotide binding. This system commonly uses phosphotransferase activity by a receptor kinase domain onto a response regulator for DNA binding[16].

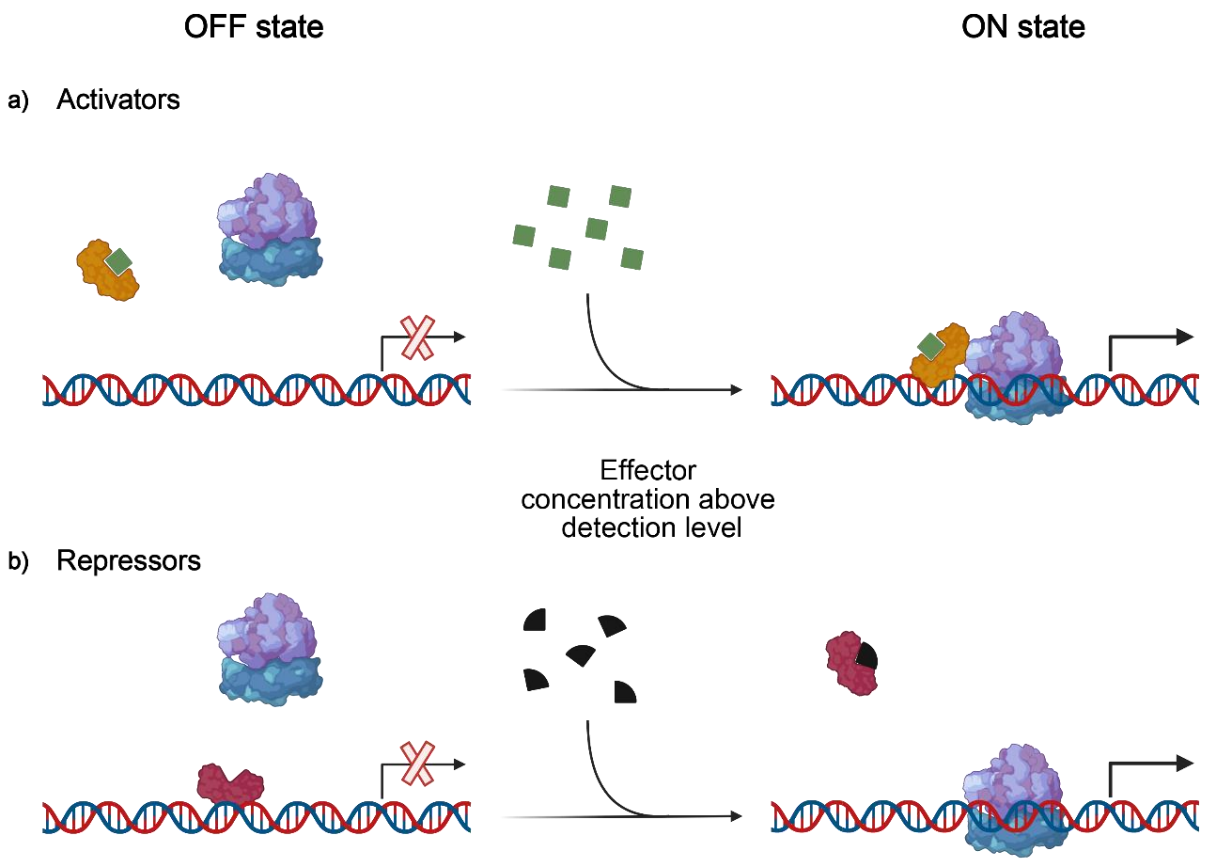


Figure 1: A basic schematic of the types of TF and their mechanism of activity. a) Activators are transcription factors that directly enable downstream gene expression in the presence of detectable effector concentrations. They usually exist in solution and help recruit RNAP complex, bind to their sigma factor or help bend DNA for transcription initiation. Genes under an activator's control tend to be more environmentally sensitive allowing alternate survival methods in the presence of different metabolites. They generally interact with DNA upstream of the promoter region. B) Unlike activators, repressors are transcription factors that natively contact their operator DNA sequence physically blocking the promoter access of the RNAP complex. Upon effector binding, repressors undergo conformation change which leads to DNA unbinding and transcription initiation. A lot of prokaryotic TFs used in synthetic biology are repressors as they are better understood and can be easier to implement in a workflow. Repressors contact DNA either overlapping the promoter sequence or very close to it.

Activators:

Activators are TFs that directly enable the expression of genes by binding to the upstream DNA region (operator) of an open reading frame (ORF) **(Fig. 1a)**. σ^{70}-dependent activators fall into either Class I or Class II. The two classes differ in their DNA binding region as well as their protein-protein interactions. Class I TFs bind DNA in the -61 and -91 region which is quite a bit upstream of the promoter -35 element[17]. The earlier discussed cyclic AMP receptor, CRP, is a well understood example of class I mechanism. Class II activators bind in an overlapping region of the σ factor at the -35 promoter region while contacting the domain 4 of the RNA polymerase σ subunit. The best example of class II TF is the activation of bacteriophage λ PRM promoter by the bacteriophage λCI protein. λCI binds to an intermediate from the closed RNAP holoenzyme to the open complex promoting isomerization and gene activation[18].

The third major type of activators belong to the σ^{54} dependent transcription factors which use ATP hydrolysis to generate enough energy for the isomerization of the closed complex into an open complex. These AAA+ family of proteins contain an additional domain that performs ATPase activity. They bind to an upstream activating sequence (UAS) located at 80-150bp upstream of the ORF similar to eukaroytic enhancer proteins labeling them as bacterial enhancer-binding proteins (bEBPs)[19]. The interaction of UAS-bound bEBP with σ^{54} requires DNA looping, which is often facilitated by DNA bending proteins such as the integrative host factor (IHF)[20]. bEBP family of proteins are not well studied but some of the better understood examples include *E. coli* PspF[21], *Salmonella enterica* NtrC[22], *E. coli* NorR[23]. σ^{54} dependent TFs are typically used by cells under stress

conditions such as phage infection or nitrogen scarcity. BmoR is a σ^{54} dependent transcriptional activator found in *Thaurea Butanovorans* canonically involved in the activation of the butane monooxygenase (Bmo) operon[24]. This operon enables the growth of *T. butanovorans* in short and medium chain alkanes (C2-C8) with BmoR being induced by butanol. Chapter 2 of this dissertation will discuss this protein further along with our engineering efforts.

Repressors:

Repressors are transcription factors that are natively bound to their operator in the resting state to prevent subsequent binding by the RNAP holoenzyme **(Fig. 1b)**. In the presence of the signaling molecule, repressors undergo conformation change causing them to leave the bound DNA allowing the RNAP complex to bind to the promoter and initiate transcription. Repressors are also popularly used as regulatory elements in the synthetic biology field to control plasmid gene expression in host cells[25]. However, there are multiple steps between input of signal and output of gene. While most known repressors are bound to the operator sequence in their resting state, recent studies have shown other methods of repression. TFs in this case bind to the DNA sequence simultaneously with RNAP but inhibit transition of the RPc (closed complex) to RPo (open complex). MerR is one such repressor that prevents the expression of Tn21 mercury-resistance (mer) locus in the absence of mercuric ions (Hg(II)) but binds to the DNA along with the RNAP holoenzyme[26,27].

Promoter clearance is the process of RNA polymerase detaching from the promoter region for extension of the transcribed mRNA along the DNA template. RNAP

can be stalled and gene expression repressed at the +6 to +12 region *in vivo* if the binding to the promoter is very strong[28]. The H-NS protein with the rrnB P1 promoter is an interesting example of clearance based repression method. The global repressor H-NS TF binds to an overlapping promoter region while allowing RNAP binding and RPo complex formation[29]. While initiation does occur, H-NS TF inhibits generation of transcripts larger than two-three nucleotide. H-NS is thought to change the structure of the RNAP open complex which prevents elongation of the transcribed mRNA[30].

One of the most popularly used and studied aTF repressors is LacI serving as a model system for fundamental research in allostery as well as for a broad range of technologies such as synthetic biology systems, biomanufacturing and healthcare. LacI functions as a homotetramer that natively represses the lactose metabolizing operon (lac operon). In the absence of high amounts of lactose, lacI remains bound to its operator preventing expression of lactose metabolizing gene lacZ which encodes lactose metabolizing Beta-galactosidase. In the presence of high lactose concentration, basal activity of Beta-galactosidase converts lactose to allolactose which binds to lacI monomers leading to conformation change, unblocking of the operon promoter region and initiating gene expression of Beta-galactosidase. The conversion of lactose into allolactose serves as a positive feedback loop[31,32].

TF as biosensors:

Whole-cell biosensors:

The above discussed ability of transcription factors lends well to their capability as biosensors, molecules that take in an environmental input to produce a measurable output. Biosensors can be a variety of different macromolecules such as anti-bodies, DNA scaffolds or TFs. Several of these biosensors have been employed for industrial applications and are poised to further increase in their usage[33,34]. Enzyme based biosensors have been used to detect heavy metal ions in water samples to prevent heavy metal toxicity[35,36].

Transcription factors fall into the class of whole-cell biosensors which use biochemical reactions within living host cells to detect and respond to environmental changes. This technology can be effectively targeted towards detecting signals that are valuable for our needs[37,38]. TF based biosensors are popularly used in the metabolic engineering field to improve the yields of key molecules produced by host cells often coupled with engineering efforts to improve those yields[39].

Transcription factors have highly specific switch like response to external inputs which requires organisms to employ a wide variety of TFs for different molecules and interactions. Often, genome mining is used to discover more of these TFs for metabolites of interest and then refactor them for industrial applications. P2TF[40], TF2DNA[41], CollecTF[42] and SigMol[43] are among a few databases that aggregate information about

transcription factors, including their position in fully sequenced genomes, TF binding motifs and quorum sensing molecules as their signaling targets.

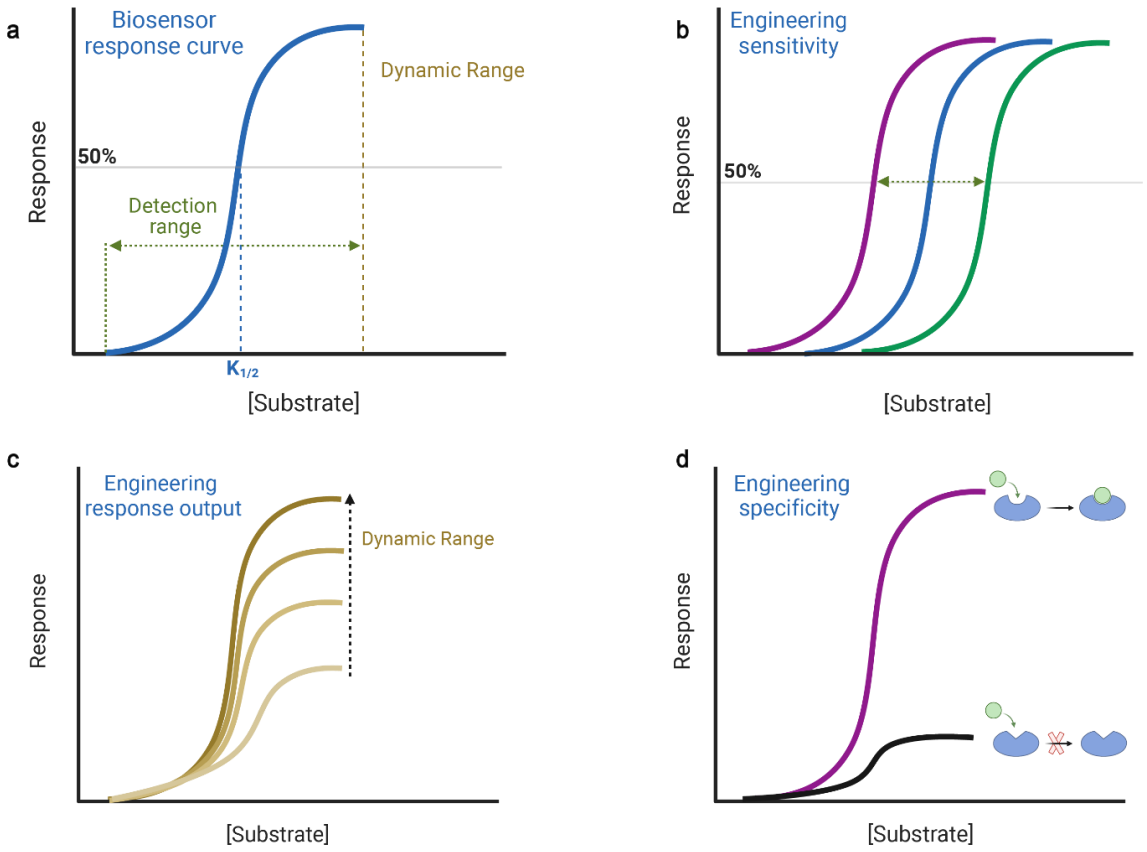


Figure 2: Understanding TF activity and the tunable parameters. a) A typical output of a TF activity for a fluorescent reporter system. The measurements are taken at a range of different substrate concentrations from inactive to maximally active giving us a sigmoidal curve. This concentration range is the detection range of native TFs and can be non-valuable for our needs. The concentration of substrate at half-maximal activity is termed the K1/2. b) The sensitivity of a TF biosensor can be engineered to extend its detection range. This is often done to make the biosensor more sensitive and reduce its K1/2. c) The output difference between the inactive and active state of the switch like response is its dynamic range which heavily impacts the practicality of the TF. For practical use of TFs having a dynamic range of >5-10x is good to avoid a lot of noise in the system. d) Not every known molecule has a regulator that binds to it. Substrate specificity can be engineered to improve binding affinity towards other molecules and/or decrease affinity towards the native substrate.

Efforts towards TF engineering:

Transcription factors natively may not be useful for industrial applications of biosensors due to low dynamic range, non-important detection range or lack of specificity towards target molecule. TF activity is represented as a substrate concentration curve with different facets that can be improved **(Fig 2a)**. The concentration at which we see half-maximal activity of TF is termed $K_{1/2}$ and is the midpoint of detection range. This range can be engineered to make the TF biosensor more effective at concentrations of interest[44] **(Fig 2b)**. The change in output from the "off" state to the "on state" is the dynamic range of the TF and can play a significant role in the practical use of TFs as biosensors. Having a significant dynamic range (>5-10x) can enable effective detection of the signal molecule while low dynamic range can be very noisy[45] **(Fig 2c)**. Finally, the input for the TFs tends to be highly specific with potential basal activity on structurally similar molecules. Engineering TFs can often involve changing or altering specificity to bind to molecules that are highly valuable to us while having no discernible value for nature[46] **(Fig 2d)**.

The methods to make such improvements seemed arduous and difficult till directed evolution using random mutagenesis was showcased by the seminal work of Frances Arnold's with Subtilisin engineering[47,48]. In 1999, Zhou and Arnold used directed evolution to evolve *Bacillus subtilis* subtilisin E towards improved activity at higher temperature. They engineered the WT sub E to be as heat tolerant as its thermophilic homolog thermitase from *Thermoactinomyces vulgaris* with an increase in its optimal working temperature by 17°C[49]. This method does not involve extensive prior knowledge of the

system, nor does it need a well-defined crystal structure, major limitation of rational design methods used previously. Similarly, as the discovery of new TFs with new targets becomes commonplace, engineering methods to improve their functionality follows.

Machado, Currin and Dixon published their work on engineering PcaV, a MarR family repressor, for the detection of hydroxyl-substituted benzoic acids. PcaV is natively involved in the regulation of catabolic genes for protocatechuic acid (PCA) in *Streptomyces coelicolor* but not much is known about this family of TFs[50,51]. With a need to detect aromatic aldehydes, the researchers looked at PcaV that canonically binds to an aromatic aldehyde and sought to widen its specificity towards other structurally similar molecules including vanillin. Since there wasn't much information about PcaV, they utilized directed evolution methods to achieve their goals. They made the variant Van2, a three amino acid substitution from starting PcaV sequence, with 8-fold enhanced selectivity towards vanillin and 3,4-dihydroxybenzaldehyde with no activity towards its native substrate, PCA[52].

When enzymatic activity cannot be linked with a fluorescent like output, the screening methods for enzyme engineering can be limited to analytical instruments such as LCMS. This limits the library's functional exploration space to a small fraction. TF based biosensors have found a valuable home here. They can be coupled with enzymatic reactions to produce a screenable output such as GFP production to enable high throughput screens. Metabolic engineering methods work by targeting enzymes in a metabolic pathway to improve the overall flux of metabolites and increase yield of a desired products such as biofuel valuable alcohols or building blocks such as

terpenoids[53,54]. Li et. al. coupled the production of intracellular malonyl-coenzyme (CoA) with β-Galactosidase activity for plate screening of high malonyl-CoA producing strains. This relied on the production of malonyl-CoA by a knockout library of upstream enzymes using transposon addition. Malonyl-CoA gets transformed into triacetic acid lactone (TAL) via an exogenous 2-pyrone synthase. TAL is detected by AraC producing β-Galactosidase which acts upon the X-Gal present in the plate agar. More galactosidase activity leads to bluer colonies with increased production of malonyl Coa. Without this coupled method, malonyl-CoA detection required the use of LC-MS a lower throughput, time and resource intensive method. They identified a significant role played by Iron in malonyl-CoA production with colonies producing 4x more TAL than WT[55].

Within the past decade, there has been an immense leap in the use of computational methods in engineering proteins. Machine learning approaches towards designing better variants or de novo proteins have advanced significantly[56]. Recent work done by d'Oelsnitz et. al. focuses on combining the use of biosensors and machine learning to evolve an Amaryllidaceae enzyme[57]. The authors, using directed evolution, initially engineered RamR to be highly sensitive towards 4′-O-methylnorbelladine, a key branch point for Amaryllidaceae alkaloid. Along with the engineered biosensor, they utilized a structure based neural network (mutCompute X) to generate highly active variants of the Amaryllidaceae enzyme, rapidly screened using the previously developed biosensor. They found variants with 60% improved titers, 2 fold increase in catalytic activity and 3-fold lower off-product regioisomer formation. This is a very neat setup to show the feasibility of combining different synthetic biology tools to achieve our engineering goals.

However, not a lot of work has been done to use machine learning to evolve transcription factors themselves. Advent of AlphaFold 3 set a new benchmark in protein engineering as now we have access to high quality structural information for most proteins[58]. This provides a new avenue for TF engineering along with other methods to accelerate the process of new variant development. In later chapters, I will talk about two projects which involve transcription factor engineering using directed evolution combined with machine learning models to generate enhanced variants.

Challenges in TF engineering:

Transcription factors provide a strong fundamental technology for biosensor applications however, they can be difficult to implement. In this section, I wanted to highlight some challenges I faced in my efforts towards TF engineering.

A major roadblock can be the lack of understanding of the transcription factor. While we can engineer proteins without the need to fully understand them, having insufficient information regarding the underlying biology can lead to prolonged project times and lots of troubleshooting. This manifests as problems in assay development especially with reproducibility between run and biological samples. Systems developed by other researchers may not always work as described, which can be due to differences in researcher skill, equipment usage or even just the physical environment of the scientists. The vastness of available prokaryotic TFs is undermined by our lack of understanding of the different kinds of TFs and usually a few get selectively optimized. Taking time to understand your plasmid system and its underlying biological mechanism, given the distinct functions of the transcription factor, can be very valuable. Engineering

TFs can often involve parallel screening to ensure specificity. Mutations that impact effector binding can lead to promiscuous activity. In many cases this can be a trivial problem as the application for biosensor removes the presences off-targets present but this can be hurdle when used in therapeutic applications. This problem is often addressed by conducting those parallel screens and reproducible output to ensure desired specificity.

Another challenge with TF engineering that often gets overlooked (I am guilty of this as well) is the variability of activity due to the phenotypic variants generated by a population of bacteria. Since TFs facilitate gene expression, bacteria adapting to different environments can lead to variable expression of the TFs and their outputs. I observed this in my work with TFs and aimed to mitigate it with multiple passages of colonies on plates and picking colonies with optimal TF activity. Another not so addressed issue is the variability when using TFs in different host cells. While TF activity is optimized for either DNA binding or effector binding, their function is usually engineered in model organisms such as *Escherichia coli* or *saccharomyces cerevisiae.* While the easier answer may sound like using the intended host during the engineering process, not every bacteria or organism is usable in laboratory settings. Sometimes it's feasible to ignore this problem since the engineered variant gets used in the same or other model organisms. However, it is prudent to transfer the biosensor in the host organism of interest to ensure activity remains desirable.

The above represent an overarching theme of ensuring reproducibility in biological ventures. Synthetic chemical methods became popular for drug development and natural product synthesis due to ease of access, faster time scales and reproducibility. In our

advancement in synthetic biology to promote green chemistry and improve upon our industrial methodology we can overlook fundamental gaps in utilizing this technology. As these challenges stand in front of every TF based biosensor, it is possible to mitigate them to a good extent. For me the process involved lots of trial and error as well as sharing my engineering experience with colleagues to avoid similar mistakes.

Future of biosensors:

The environmental demands of our society are increasing as the need to protect our planet grows rapidly. Addressing this issue while maintaining the current state of the world would require the shift in the current chemical manufacturing, replacing organic solvents with water, minimizing use of precious metal catalysts with enzymes in living organisms and increasing the use of less-toxic reagents. TF based biosensors can play a key role to enable this biomanufacturing model for chemical production. Several known TFs have been developed to detect important amino acids [59], antibiotics[60], sugars[61] and plant metabolites[62]. They have been used extensively in metabolic engineering for improved chemical synthesis. A good potential target for TF biosensor based engineering would be in developing minimal cells which are host organisms with basic survival functions intact along with exogenous over expressed pathways. This allows minimal cells to act as chemical factories without requiring a host of different resources that are needed by native organisms. Using TF biosensors, we can engineer metabolic pathways to improve their flux along with learning about non-essential biochemical pathways.

As I look towards industrial applications of TF based biosensors, it holds immense value in academic settings. I say this especially in regard to the effect of endogenous

metabolites present in the gut. Our understanding of the spatial organization of the gut is poor to say the least. Lots of amazing work is being done to learn about this large community of bacteria and fungi, adding gut metabolite sensing TFs can give us a glimpse into the network of this community. Using tissue penetrating reporter systems such anaerobic florescent proteins[63] or bioluminescence[64], it is possible to apply these biosensors in live mouse models to monitor real time spatial arrangement of metabolite in the gut.

In my opinion, there is an unfulfilled need to develop a library of accessible biosensors for academic and industrial applications. Some very talented colleagues built an online database (GroovDB) for known TF-effector combinations[65] with over a 100 regulators and 150 effectors. Tools such as these will be vital in information sharing and parallel testing for researchers that want to use TFs but do not have to engineer them. A step forward would be to make a physical library of lab strains and plasmids containing these biosensors for “off-the-shelf” use. While this approach of building a biosensor library wasn’t my focus during graduate school, I think there is room for developing this technology down the line.

References:

(1) Pennypacker, K. R. Pharmacological Regulation of Transcription Factor Binding. *Karger Publishers*, 1995, *51*, 1–12. https://doi.org/10.1159/000139311.

(2) Göös, H.; Kinnunen, M.; Salokas, K.; Tan, Z.; Liu, X.; Yadav, L.; Zhang, Q.; Wei, G.-H.; Varjosalo, M. Human Transcription Factor Protein Interaction Networks. *Nature Portfolio*, 2022, *13*. https://doi.org/10.1038/s41467-022-28341-5.

(3) Whitmarsh, A. J.; Davis, R. J. Transcription Factor AP-1 Regulation by Mitogen-Activated Protein Kinase Signal Transduction Pathways. *Springer Science+Business Media*, 1996, *74*, 589–607. https://doi.org/10.1007/s001090050063.

(4) Wang, X.; Tan, N.; Chung, F. Y.; Yamaguchi, N.; Gan, E.-S.; Ito, T. Transcriptional Regulators of Plant Adaptation to Heat Stress. *Multidisciplinary Digital Publishing Institute*, 2023, *24*, 13297–13297. https://doi.org/10.3390/ijms241713297.

(5) Jacob, F.; Monod, J. On the Regulation of Gene Activity. *Cold Spring Harbor Laboratory Press*, 1961, *26*, 193–211. https://doi.org/10.1101/sqb.1961.026.01.024.

(6) Dillon, S. C.; Dorman, C. J. Bacterial Nucleoid-Associated Proteins, Nucleoid Structure and Gene Expression. *Nature Portfolio*, 2010, *8*, 185–195. https://doi.org/10.1038/nrmicro2261.

(7) Malan, T. P.; Kolb, A.; Buc, H.; McClure, W. R. Mechanism of CRP-cAMP Activation of Lac Operon Transcription Initiation Activation of the P1 Promoter. *Elsevier BV*, 1984, *180*, 881–909. https://doi.org/10.1016/0022-2836(84)90262-6.

(8) Murakami, K.; Masuda, S.; Darst, S. A. Structural Basis of Transcription Initiation: RNA Polymerase Holoenzyme at 4 Å Resolution. *American Association for the Advancement of Science*, 2002, *296*, 1280–1284. https://doi.org/10.1126/science.1069594.

(9) Chandrangsu, P.; Helmann, J. D. Sigma Factors in Gene Expression, 2014. https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000854.pub3.

(10) Paget, M. S.; Helmann, J. D. The Σ70family of Sigma Factors. *Genome Biology* 2003, *4* (1), 203. https://doi.org/10.1186/gb-2003-4-1-203.

(11) Danson, A. E.; Jovanovic, M.; Buck, M.; Zhang, X. Mechanisms of Σ54-Dependent Transcription Initiation and Regulation. *Journal of Molecular Biology* 2019, *431* (20), 3960–3974. https://doi.org/10.1016/j.jmb.2019.04.022.

(12) Martínez-Antonio, A.; Janga, S. C.; Salgado, H.; Collado-Vides, J. Internal-Sensing Machinery Directs the Activity of the Regulatory Network in Escherichia Coli. *Elsevier BV*, 2006, *14*, 22–27. https://doi.org/10.1016/j.tim.2005.11.002.

(13) Monod, J.; Changeux, J.-P.; Jacob, F. Allosteric Proteins and Cellular Control Systems. *Elsevier BV*, 1963, *6*, 306–329. https://doi.org/10.1016/s0022-2836(63)80091-1.

(14) Seshasayee, A. S. N.; Bertone, P.; Fraser, G. M.; Luscombe, N. M. Transcriptional Regulatory Networks in Bacteria: From Input Signals to Output Responses. *Elsevier BV*, 2006, *9*, 511–519. https://doi.org/10.1016/j.mib.2006.08.007.

(15) Babu, M. M. Evolution of Transcription Factors and the Gene Regulatory Network in Escherichia Coli. *Oxford University Press*, 2003, *31*, 1234–1244. https://doi.org/10.1093/nar/gkg210.

(16) Stock, A.; Robinson, V.; Goudreau, P. N. Two-Component Signal Transduction. *Annual Reviews*, 2000, *69*, 183–215. https://doi.org/10.1146/annurev.biochem.69.1.183.

(17) Browning, D. F.; Busby, S. The Regulation of Bacterial Transcription Initiation. *Nature Portfolio*, 2004, *2*, 57–65. https://doi.org/10.1038/nrmicro787.

(18) Nickels, B. E.; Dove, S. L.; Murakami, K.; Darst, S. A.; Hochschild, A. Protein–Protein and Protein–DNA Interactions of Σ70 Region 4 Involved in Transcription Activation by λcI. *Elsevier BV*, 2002, *324*, 17–34. https://doi.org/10.1016/s0022-2836(02)01043-4.

(19) Kim, T. K.; Ebright, R. H.; Reinberg, D. Mechanism of ATP-Dependent Promoter Melting by Transcription Factor IIH. *American Association for the Advancement of Science*, 2000, *288*, 1418–1421. https://doi.org/10.1126/science.288.5470.1418.

(20) Yoshua, S.; Watson, G. D.; Howard, J.; Velasco-Berrelleza, V.; Leake, M. C.; Noy, A. Integration Host Factor Bends and Bridges DNA in a Multiplicity of Binding Modes with Varying Specificity. *Oxford University Press*, 2021, *49*, 8684–8698. https://doi.org/10.1093/nar/gkab641.

(21) Jovanović, G.; Weiner, L.; Model, P. Identification, Nucleotide Sequence, and Characterization of PspF, the Transcriptional Activator of the Escherichia Coli Stress-Induced Psp Operon. *American Society for Microbiology*, 1996, *178*, 1936–1945. https://doi.org/10.1128/jb.178.7.1936-1945.1996.

(22) Carlo, S. D.; Chen, B.; Hoover, T. R.; Kondrashkina, E.; Nogales, E.; Nixon, B. T. The Structural Basis for Regulated Assembly and Function of the Transcriptional Activator NtrC. *Cold Spring Harbor Laboratory Press*, 2006, *20*, 1485–1495. https://doi.org/10.1101/gad.1418306.

(23) Hutchings, M. I.; Mandhana, N.; Spiro, S. The NorR Protein of Escherichia Coli Activates Expression of the Flavorubredoxin Gene norV in Response to Reactive Nitrogen Species, 2002. https://journals.asm.org/doi/10.1128/JB.184.16.4640-4643.2002.

(24) Dubbels, B. L.; Sayavedra-Soto, L. A.; Arp, D. J. Butane Monooxygenase of 'Pseudomonas Butanovora': Purification and Biochemical Characterization of a Terminal-Alkane Hydroxylating Diiron Monooxygenase. *Microbiology Society*, 2007, *153*, 1808–1816. https://doi.org/10.1099/mic.0.2006/004960-0.

(25) Rantasalo, A.; Kuivanen, J.; Penttilä, M.; Jäntti, J.; Mojžita, D. Synthetic Toolkit for Complex Genetic Circuit Engineering in Saccharomyces Cerevisiae. *American Chemical Society*, 2018, *7*, 1573–1587. https://doi.org/10.1021/acssynbio.8b00076.

(26) Brown, N. L.; Stoyanov, J.; Kidd, S. P.; Hobman, J. L. The MerR Family of Transcriptional Regulators. *Oxford University Press*, 2003, *27*, 145–163. https://doi.org/10.1016/s0168-6445(03)00051-2.

(27) Chengli Fang, Y. Z. undefined. Bacterial MerR Family Transcription Regulators: Activationby Distortion, 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9909328.

(28) Ellinger, T.; Behnke, D.; Bujard, H.; Gralla, J. D. Stalling of Escherichia Coli RNA Polymerase in the +6 to +12 Region in Vivo Is Associated with Tight Binding to Consensus Promoter Elements. *Elsevier BV*, 1994, *239*, 455–465. https://doi.org/10.1006/jmbi.1994.1388.

(29) Rodgers, D.; Le, C.; Pimentel, C.; Tuttobene, M. R.; Subils, T.; Escalante, J.; Nishimura, B.; Véscovi, E. G.; Sieira, R.; Bonomo, R. A.; Tolmasky, M. E.; Ramírez, M. S. Histone-like Nucleoid-Structuring Protein (H-NS) Regulatory Role in Antibiotic Resistance in Acinetobacter Baumannii. *Nature Portfolio*, 2021, *11*. https://doi.org/10.1038/s41598-021-98101-w.

(30) Schröder, O.; Wagner, R. The Bacterial DNA-Binding Protein H-NS Represses Ribosomal RNA Transcription by Trapping RNA Polymerase in the Initiation Complex. *Elsevier BV*, 2000, *298*, 737–748. https://doi.org/10.1006/jmbi.2000.3708.

(31) Becker, N. A.; Peters, J. P.; Lionberger, T. A.; Maher, J. Mechanism of Promoter Repression by Lac Repressor–DNA Loops. *Oxford University Press*, 2012, *41*, 156–166. https://doi.org/10.1093/nar/gks1011.

(32) Hirsh, J.; Schleif, R. In Vivo Experiments on the Mechanism of Action of L-Arabinose C Gene Activator and Lactose Repressor. *Elsevier BV*, 1973, *80*, 433–444. https://doi.org/10.1016/0022-2836(73)90414-2.

(33) Carpenter, A. C.; Paulsen, I. T.; Williams, T. Blueprints for Biosensors: Design, Limitations, and Applications. *Multidisciplinary Digital Publishing Institute*, 2018, *9*, 375–375. https://doi.org/10.3390/genes9080375.

(34) Cheng, F.; Tang, X.; Kardashliev, T. Transcription Factor-Based Biosensors in High-Throughput Screening: Advances and Applications. *Wiley-Blackwell* 2018, *13* (7). https://doi.org/10.1002/biot.201700648.

(35) Odobašić, A.; Šestan, I.; Begić, S. Biosensors for Determination of Heavy Metals in Waters. *IntechOpen*, 2019. https://doi.org/10.5772/intechopen.84139.

(36) Alfadaly, R. A.; Elsayed, A.; Hassan, R. Y. A.; Noureldeen, A.; Darwish, H.; Gebreil, A. Microbial Sensing and Removal of Heavy Metals: Bioelectrochemical Detection and Removal of Chromium(VI) and Cadmium(II). *Multidisciplinary Digital Publishing Institute*, 2021, *26*, 2549–2549. https://doi.org/10.3390/molecules26092549.

(37) Shang, Y.; Song, X.; Bowen, J.; Corstanje, R.; Gao, Y.; Gaertig, J.; Gorovsky, M. A. A Robust Inducible-Repressible Promoter Greatly Facilitates Gene Knockouts, Conditional Expression, and Overexpression of Homologous and Heterologous Genes in Tetrahymena Thermophila. *National Academy of Sciences*, 2002, *99*, 3734–3739. https://doi.org/10.1073/pnas.052016199.

(38) Cerminati, S.; Soncini, F. C.; Checa, S. K. A Sensitive Whole-Cell Biosensor for the Simultaneous Detection of a Broad-Spectrum of Toxic Heavy Metal Ions. *Royal Society of Chemistry*, 2015, *51*, 5917–5920. https://doi.org/10.1039/c5cc00981b.

(39) Teng, Y.; Zhang, J.; Jiang, T.; Zou, Y.; Gong, X.; Yan, Y. Biosensor-Enabled Pathway Optimization in Metabolic Engineering. *Elsevier BV*, 2022, *75*, 102696–102696. https://doi.org/10.1016/j.copbio.2022.102696.

(40) Ortet, P.; Luca, G. D.; Whitworth, D. E.; Barakat, M. P2TF: A Comprehensive Resource for Analysis of Prokaryotic Transcription Factors. *BioMed Central*, 2012, *13*. https://doi.org/10.1186/1471-2164-13-628.

(41) Pujato, M.; Kieken, F.; Skiles, A. A.; Tapinos, N.; Fiser, A. Prediction of DNA Binding Motifs from 3D Models of Transcription Factors; Identifying TLX3 Regulated Genes. *Oxford University Press*, 2014, *42*, 13500–13512. https://doi.org/10.1093/nar/gku1228.

(42) Kılıç, S.; White, E. R.; Sagitova, D. M.; Cornish, J. P.; Erill, I. CollecTF: A Database of Experimentally Validated Transcription Factor-Binding Sites in Bacteria. *Oxford University Press*, 2013, *42*, D156–D160. https://doi.org/10.1093/nar/gkt1123.

(43) Rajput, A.; Kaur, K.; Kumar, M. SigMol: Repertoire of Quorum Sensing Signaling Molecules in Prokaryotes. *Oxford University Press*, 2015, *44*, D634–D639. https://doi.org/10.1093/nar/gkv1076.

(44) Tillotson, B. J.; Goulatis, L. I.; Parenti, I.; Duxbury, E.; Shusta, E. V. Engineering an Anti-Transferrin Receptor ScFv for pH-Sensitive Binding Leads to Increased Intracellular Accumulation. *Public Library of Science*, 2015, *10*, e0145820–e0145820. https://doi.org/10.1371/journal.pone.0145820.

(45) D'Ambrosio, V.; Jensen, M. K. Lighting up Yeast Cell Factories by Transcription Factor-Based Biosensors. *Oxford University Press*, 2017, *17*. https://doi.org/10.1093/femsyr/fox076.

(46) Taylor, N. D.; Garruss, A. S.; Moretti, R.; Chan, S.; Arbing, M. A.; Cascio, D.; Rogers, J. K.; Isaacs, F. J.; Kosuri, S.; Baker, D.; Fields, S.; Church, G. M.; Raman, S. Engineering an Allosteric Transcription Factor to Respond to New Ligands. *Nat Methods* 2016, *13* (2), 177–183. https://doi.org/10.1038/nmeth.3696.

(47) Chen, K.; Arnold, F. H. Enzyme Engineering for Nonaqueous Solvents: Random Mutagenesis to Enhance Activity of Subtilisin E in Polar Organic Media. *Springer Nature*, 1991, *9*, 1073–1077. https://doi.org/10.1038/nbt1191-1073.

(48) Chen, K.; Arnold, F. H. Tuning the Activity of an Enzyme for Unusual Environments: Sequential Random Mutagenesis of Subtilisin E for Catalysis in Dimethylformamide. *National Academy of Sciences*, 1993, *90*, 5618–5622. https://doi.org/10.1073/pnas.90.12.5618.

(49) Zhao, H.; Arnold, F. H. Directed Evolution Converts Subtilisin E into a Functional Equivalent of Thermitase. *Oxford University Press*, 1999, *12*, 47–53. https://doi.org/10.1093/protein/12.1.47.

(50) Davis, J. R.; Sello, J. K. Regulation of Genes in Streptomyces Bacteria Required for Catabolism of Lignin-Derived Aromatic Compounds. *Springer Science+Business Media*, 2009, *86*, 921–929. https://doi.org/10.1007/s00253-009-2358-0.

(51) Davis, J. R.; Brown, B. L.; Page, R.; Sello, J. K. Study of PcaV from Streptomyces Coelicolor Yields New Insights into Ligand-Responsive MarR Family Transcription Factors. *Oxford University Press*, 2013, *41*, 3888–3900. https://doi.org/10.1093/nar/gkt009.

(52) Machado, L. F. M.; Currin, A.; Dixon, N. Directed Evolution of the PcaV Allosteric Transcription Factor to Generate a Biosensor for Aromatic Aldehydes. *BioMed Central*, 2019, *13*. https://doi.org/10.1186/s13036-019-0214-z.

(53) Martin, V. J. J.; Pitera, D. J.; Withers, S. T.; Newman, J. D.; Keasling, J. D. Engineering a Mevalonate Pathway in Escherichia Coli for Production of Terpenoids. *Nature Portfolio*, 2003, *21*, 796–802. https://doi.org/10.1038/nbt833.

(54) Peralta-Yahya, P.; Zhang, F.; Cardayré, S. B. del; Keasling, J. D. Microbial Engineering for the Production of Advanced Biofuels. *Nature Portfolio*, 2012, *488*, 320–328. https://doi.org/10.1038/nature11478.

(55) Li, H.; Chen, W.; Jin, R.; Jin, J.; Táng, S. Biosensor-Aided High-Throughput Screening of Hyper-Producing Cells for Malonyl-CoA-Derived Products. *BioMed Central*, 2017, *16*. https://doi.org/10.1186/s12934-017-0794-6.

(56) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nature Methods* 2019, *16* (8), 687–694. https://doi.org/10.1038/s41592-019-0496-6.

(57) d'Oelsnitz, S.; Diaz, D. J.; Acosta, D. J.; Schechter, M. W.; Minus, M. B.; Howard, J. R.; Loy, J. M.; Do, H.; Alper, H. S.; Ellington, A. D. Synthetic Microbial Sensing and Biosynthesis of Amaryllidaceae Alkaloids. *Cold Spring Harbor Laboratory*, 2023. https://doi.org/10.1101/2023.04.05.535710.

(58) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* 2024, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(59) Verma, N.; Singh, A. K.; Singh, A. K. L-Arginine Biosensors: A Comprehensive Review. *Elsevier BV*, 2017, *12*, 228–239. https://doi.org/10.1016/j.bbrep.2017.10.006.

(60) Wang, Y.; Li, S.; Xue, N.; Wang, L.; Zhang, X.; Zhao, L.; Guo, Y.; Zhang, Y.; Wang, M. Modulating Sensitivity of an Erythromycin Biosensor for Precise High-Throughput

Screening of Strains with Different Characteristics. *American Chemical Society*, 2023, *12*, 1761–1771. https://doi.org/10.1021/acssynbio.3c00059.

(61) Yoo, E. H.; Lee, S.-Y. Glucose Biosensors: An Overview of Use in Clinical Practice. *Multidisciplinary Digital Publishing Institute*, 2010, *10*, 4558–4576. https://doi.org/10.3390/s100504558.

(62) Wang, R.; Cress, B. F.; Yang, Z.; Hordines, J.; Zhao, S.; Jung, G. Y.; Wang, Z.; Koffas, M. Design and Characterization of Biosensors for the Screening of Modular Assembled Naringenin Biosynthetic Library in Saccharomyces Cerevisiae. *American Chemical Society*, 2019, *8*, 2121–2130. https://doi.org/10.1021/acssynbio.9b00212.

(63) Streett, H.; Charubin, K.; Papoutsakis, E. T. Anaerobic Fluorescent Reporters for Cell Identification, Microbial Cell Biology and High-Throughput Screening of Microbiota and Genomic Libraries. *Elsevier BV*, 2021, *71*, 151–163. https://doi.org/10.1016/j.copbio.2021.07.005.

(64) Buckley, S. M. K.; Delhove, J.; Perocheau, D.; Karda, R.; Rahim, A. A.; Howe, S. J.; Ward, N. J.; Birrell, M. A.; Belvisi, M. G.; Arbuthnot, P.; Johnson, M. R.; Waddington, S. N.; McKay, T. R. In Vivo Bioimaging with Tissue-Specific Transcription Factor Activated Luciferase Reporters. *Nature Portfolio*, 2015, *5*. https://doi.org/10.1038/srep11842.

(65) d'Oelsnitz, S.; Love, J. D.; Diaz, D. J.; Ellington, A. D. GroovDB: A Database of Ligand-Inducible Transcription Factors. *American Chemical Society*, 2022, *11*, 3534–3537. https://doi.org/10.1021/acssynbio.2c00382.

Chapter 2: Machine learning aided engineering of BmoR as a biosensor for butyrate

Abstract:

Butyrate is a short chain fatty acid produced in the gut with several beneficial function towards the host. Characterizing and regulating butyrate in the gut can be a valuable therapeutic avenue which requires the ability to detect butyrate and initiate downstream activity. Here, we engineered a σ^{54} transcription factor, BmoR, using NGS based data generation and machine learning assisted mutation prediction. Our engineered variant, $BmoR^{10mut}$, contains 10 mutations at once compared to WT with a 2-fold increase in activity. $BmoR^{10mut}$ also exhibits 3-fold improvement in activity in spent media assay of a simple butyrate producing gut microbiome community.

Introduction:

Cardiovascular diseases (CVD) account for 31% of the deaths worldwide, making them the leading cause of death around the globe[1]. Recently, studies focused on the gut microbiome have demonstrated a close link between gut health and cardiovascular health[2,3]. Short-chain fatty acids (SCFAs) are major metabolites in the gut produced via fermentation of the dietary fibers and resistant starch. Butyrate, an SCFA along with acetate and propionate, serves not only as the primary source of energy for the colonocytes, but also regulates many cellular functions within the gut and host organs[4,5]. Butyrate has been shown to have significant impact outside the gut being involved with

blood pressure regulation, improved glucose and lipid profiles, improved cardiovascular health and promoting sleep[6]. Gut microbes such as butyrate producing firmicutes develop a microbiome community by cross feeding metabolites to sustain growth. Low-fiber diets can lead to a lack of metabolites exchange and low production of butyrate causing bowel inflammation[7,8].

The vast collection of microbes in the gut are thought to exist in specific spatial niches throughout the G.I. tract such as the colon, intestinal lumen, mucus layer separating the lumen and gut epithelium and even the microscopic folds in the gut epithelium (crypts) producing a variety of different metabolites that impact the host[9]. It is important to understand this metabolite composition with accurate measurement of their levels in the various parts of the gut[10]. Specifically, the amount of gut butyrate is variable throughout the gut without a strong consensus on precise values with some reporting a range of 10-40mM[11,12]. There is an unmet a need to better understand butyrate levels in the gut as well as the capability to maintain a healthy concentration of butyrate enabling novel live microbial therapeutics[7,13]. One crucial aspect in this process involves the ability to detect butyrate in the gut.

In this work, we utilized an NGS based directed evolution workflow to engineer a transcription factor (TF), BmoR, with improved activity towards butyrate. We utilized error-prone pcr to generate the mutant library that was screened using florescence assisted cell sorting (FACS). Over the last decade, we have seen a significant increase in the use of machine learning to aid engineering efforts[14] often using data rich input[15]. We used an

enrichment-based ML model, Positive Unlabeled learning (PU-learning), to analyze the screened dataset and obtain favorable mutations.

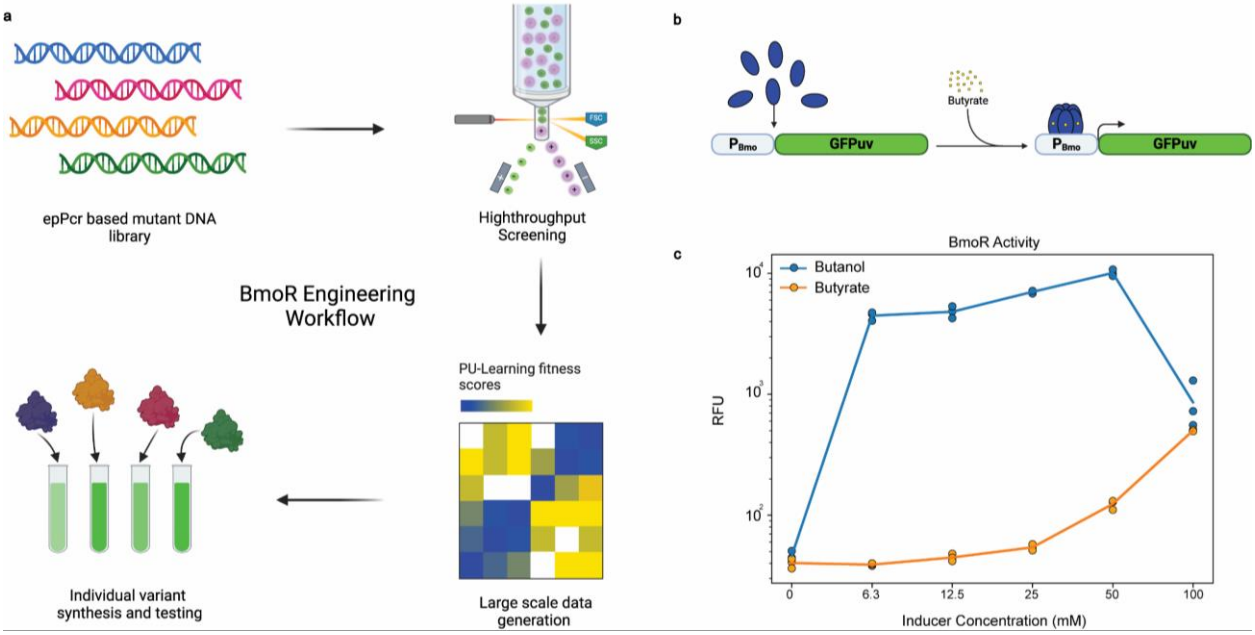


Figure 3: Engineering workflow and initial activity of BmoR. (a) Error-prone pcr generates a random mutagenic library that can be transformed in cells of interest. This method efficiently uses ultra-highthrouput screening for data generation for short read NGS. Data analysis of obtained data can be used to predict mutations for selection as well as generate a DMS like dataset. This workflow can be repeated for further engineering. **(b)** Our BmoR engineering system utilizes a GFPuv reporter (ex: 405nM, em: 510nM) under BmoR inducible promoter and operator. The BmoR protein itself is on a constitutive P_{bmoR} promoter obtained from *Thauera Butanivorans* and acts as an activator when induced by an effector, in this case butyrate/butanol. **(c)** WT BmoR shows activity with a 10-fold increase at 100mM butyrate (orange) albeit much lower than butanol (blue) with nearly 100-fold activity at 6.3mM. BmoR activity doesn't saturate at 100mM butyrate, but higher concentrations start becoming toxic to the cells. Samples were tested using flow cytometry and analyzed with FlowJo. Dots represent individual samples while line values represent the mean (n = 3). Note: We see a drop in the FI. value of 100mM butanol due to severe toxicity and cell death.

Results:

BmoR as a putative biosensor for butyrate:

Bacterial transcription factors (TFs) have been used more often to develop a whole range of biosensors for high throughput screening of small molecules and potential therapeutic applications[16]. They canonically bind to specific molecules which can be furthered engineered to enhance their activity either towards that specific molecule or potentially a different target[17,18]. These TFs, upon binding their molecule of interest, induce downstream gene expression making them an effective tool in gene circuits. TFs have been previously utilized in metabolic engineering as well as whole-cell strain engineering to improve titers of industrially relevant molecules[19]. In past work, E. coli Lrp operon was shown to natively bind butyrate but the levels of detection are lower than physiological levels with a two-component system[20].

To minimize multiple proteins involved in our system, we selected BmoR, a σ^{54} TF natively found in *Thauera butanivorans* that binds to small-medium chain alcohols (C4-C6) particularly butanol[21]. BmoR contains the three domains, the ligand binding domain, central AAA+ domain, and helix-turn-helix DNA-binding domains common to members of the σ^{54} bacterial enhancer binding protein (bEBP) family [22]. It was initially used as a biosensor to improve butanol titers in engineered strains[23,24]. Further work has been done on BmoR to improve its activity towards other small chain aliphatic and branched alcohols[25,26]. With structural similarity between butyrate and butanol we hypothesized that BmoR could bind to butyrate and induce fluorescent protein expression. We aimed to use a traditional directed evolution workflow generating a mutagenic library using error-prone

pcr transformed into *E. Coli* cells. With an *in vivo* GFPuv reporter system, we used a FACS based screening method to comb through large libraries followed by NGS sequencing and machine learning based data analysis to obtain the improved mutations for BmoR **(fig 3a)**.

The *in vivo* BmoR engineering circuit was obtained from the Keasling lab[23] **(fig. 3b)**. BmoR acts as an activator that initiates gene expression via a hexameric central pore which interacts with the substrate. However, the mechanism for substrate activity and TF interaction with ATP is not well known for BmoR and most other σ^{54} TFs[27]. Working with BmoR, we found its activity to be very variable between experiments even with heavy optimization attempts. We refactored this system to obtain more robust and reproducible results. The refactoring involved codon optimizing the BmoR sequence for E. Coli as well as introduced the CA(129,130)TC mutation in the BmoR upstream activating sequence (UAS)[32]. Initial screening of BmoR showed activity towards butyrate at higher mM ranges with 10-fold increased activity at 100mM from 0mM, although nearly 15x lower than butanol **(Fig 3c)**. BmoR doesn't saturate at 100mM butyrate which constitutes 10% of media volume after which higher concentrations of butyrate lead to increased toxicity in cells. We already observe this level of toxicity at 100mM butanol where the sharp decrease in fluorescence was directly due to increased cell death.

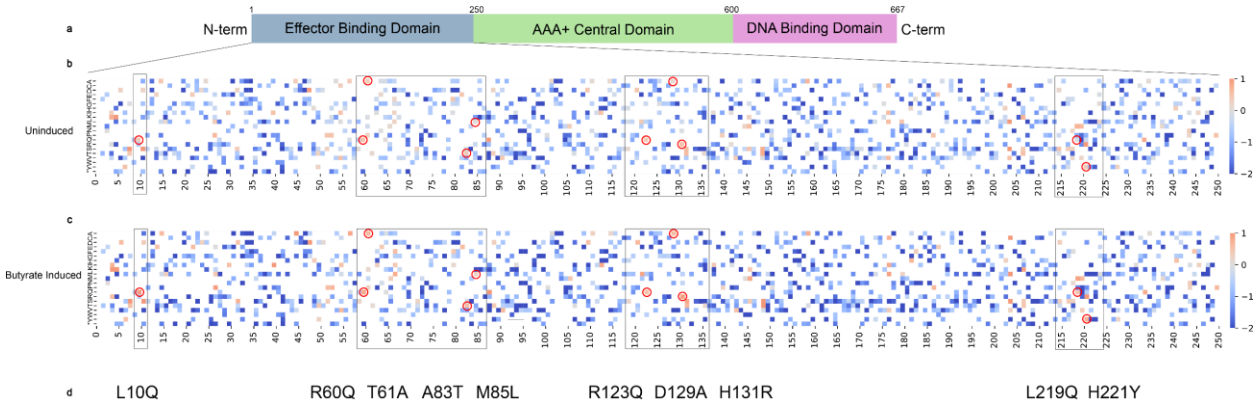


Figure 4: BmoR FACS screening and NGS data analysis. (a) BmoR has three distinct domains common to bacterial σ^{54} transcription factors: Effector binding domain, AAA+ ATPase central domain and the Helix-turn-Helix (HTH) motif containing DNA binding domain. To minimize background noise in our library, we restricted our engineering efforts to the effector binding domain consisting of the top 250 amino acids. **(b)&(c)** We analyzed the NGS dataset using PU-learning to obtain quantitative effects of observed mutations. The heatmap shown here compares the coefficient values of all observed mutations for each position in either the uninduced library (top) or the butyrate induced library (bottom). Red indicates activating mutations while blue is inactivating ones. We see mutations mostly being inactive while active mutations scattered throughout the region. To improve sensitivity, we looked for mutations with negative/low coefficients for the uninduced population while being positive for the induced populations. **(d)** The list shows all the 10 mutations we picked, making a single mutant sequence for each mutation as well as a sequence containing all 10 mutations ($BmoR^{10mut}$). The grey boxes in the heatmaps highlight regions containing the picked mutations with the red circles highlighting the amino acid picked.

DMS of BmoR effector binding region:

BmoR is proposed to form a hexameric complex upon effector binding and induce gene expression from a distance by looping DNA, similar to eukaryotic enhancer proteins[24]. Due to a gap in the understanding of the ligand binding region of σ^{54} based transcription factors along with minimal information known about BmoR, we decided to engineer BmoR using random mutagenesis. However, this multi-domain, 667 amino acid protein with distinct functions has a massive sequence space to explore which can be quite a challenge when generating random libraries. We limited our exploration space to target the first 250 amino acids containing the effector binding domain **(fig 4a)**. We used error-prone pcr, an easy and very effective method[28], for the generation of a random mutagenized library. We found by varying the amount of Mn added, we could control the mutational rate of the mutagenic region in a near linear fashion **(S1)** giving us better control on the sequence-space exploration. We utilized golden gate cloning, a highly effective ligation method to insert the mutagenized region into the our backbone[29] to obtain a large library size. Given the scarless nature of golden gate cloning, we used a CcdB based selection method to avoid false positives upon ligation.

We screened the transformed library at 0mM (uninduced) and 100mM (induced) butyrate to find BmoR variants with improved sensitivity towards butyrate. We observe that the induced library has an increase in its activity as compared to the uninduced population **(S2)** while most of the library was inactive, as expected. We collected the top 25% of the GFP fluorescence population of the samples to compare the mutational pattern amongst them. High throughput NGS can lead to high background sequencing in regions of no interest of the plasmid. Using restriction enzyme digestion, we reduced

background sequencing by about 4.5 fold and increased read efficiency. We obtained 5-10 million reads for each of the samples giving us ~10x oversampling to better capture the entirety of the mutational space. We aligned the reads using Bowtie, an efficient short read aligner[30], to generate a multiple sequence alignment (MSA) off the reference sequence, WT BmoR. We then strip the aligned reads to obtain count for each mutation observed in the pre-sorted and post-sorted populations followed by recreating mutagenic sequences using the WT while substituting the observed mutations at each residue based on their read count.

This set of sequences acts as the input for our Positive-Unlabeled learning (PU-learning) algorithm, an enrichment based regression model, meant to give relative quantitative data for each observed mutation[31]. For each condition, the model outputs a set of coefficients **(S3)**, positive values indicating enriched mutations in the post-sorted conditions while negative being de-enriched. Positive and negative values can be interpreted as activating or deleterious mutations respectively, for that specific condition. Our replicates show a good correlation between each other with the pearson correlation, r>0.90 **(S4)**. A heatmap of the PU-learning coefficients highlights the mutational pattern at various residues showcasing a non-exhaustive DMS library of the BmoR N-terminus effector binding region for uninduced populations (fig 4b) and the butyrate induced population **(fig 4c)**. Most mutations are deleterious while we see activating mutations scattered throughout the region. We set filtering parameters focused on low background activity selecting mutations with low uninduced coefficients rather than high butyrate induced coefficients. The grey boxes show the regions of interest in the BmoR effector binding region and the red circles indicate our selected residues. Using a 99% confidence

level ($p<0.01$) and filtering the coefficients, we obtained a list of 10 mutations for further testing **(fig 4d)**. We generated a set of 11 sequences, one for each of the individual mutations as well as one sequence containing all the 10 mutations which we call $BmoR^{10mut}$.

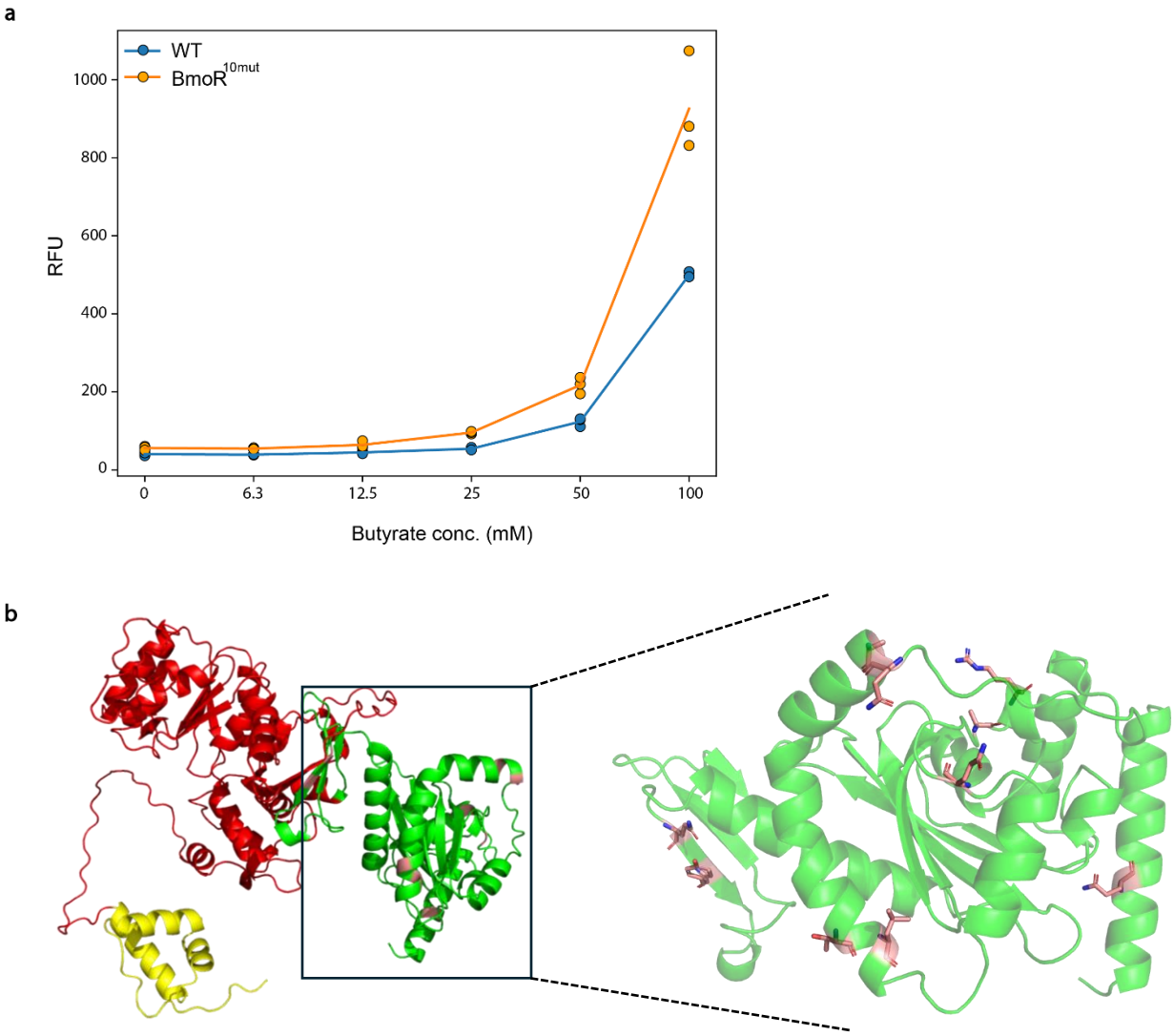


Figure 5: Testing and visualizing the variants. (a) After we codon optimized BmoR for *E. Coli* as well as added CA->TC mutations in the BmoR operator region, we observe improved robustness of our system. We do see a near 2-fold increase in activity of BmoR^{10mut} over WT in a more robust manner. Furthermore, there is 1.2-1.5 fold increase in activity of BmoR^{10mut} in the 10-50mM physiological range more relevant for the gut microbiome. Dots represent individual samples while line values represent the mean (n = 3). **(c)** We used AlphaFold 3 to predict the structure of BmoR^{10mut} with high confidence. Left: BmoR^{10mut} containing the three distinct domains of BmoR joined by loop like linking sequences. Green: Effector binding domain, Red: Central ATPase domain and Yellow: DNA binding domain. Right: BmoR^{10mut} with the 10 picked mutations in the effector binding region. We see them mostly on the surface of the protein with hydrophilic amino acids potentially interacting with the solvent or other BmoR monomers for improved stability.

Bmor^{10mut} design and test:

We assessed the individual variants with titrating butyrate from 0mM to 100mM showing an improvement in the fold-change of our variant sequences with the BmoR^{10mut} performing the best **(S5)**. Post-refactoring, we see a robust response of GFPuv expression and a 2-fold increase in activity of the BmoR^{10mut} with a minimal increase in background **(Fig 5a)**. Note: We did not refactor each individual variant as the BmoR^{10mut} was consistently shown to be the best variant.

Alphafold3 predicted structure of BmoR^{10mut} shows three distinct domains linked by loop-like stretches of DNA **(Fig 5b, left)**. The Central domain (red) is highly conserved among σ^{54} TFs. It contains significantly conserved motifs such as the highly GAFTGx (407-412aa) motif that contacts AAA+ loop σ^{54} factor during ATP hydrolysis and the GxxxxGK (360-366 AA) motif interacting with the phosphate in ATP[33]. The C-terminal DNA binding domain (yellow) contains the Helix-Turn-Helix motif that binds to the UAS to facilitate DNA looping and gene transcription. The effector binding domain (green) is highly variable in this class of TFs and not well understood. We mapped the 10 mutations onto the predicted structure to visualize the residues in 3D space **(Fig 5b, right)**. We see the distribution of mutations to be mostly on the surface of the protein with electronegative atoms (N and O) jetting out. Most mutations appear prominently on one side of BmoR which could indicate stabilizing protein-protein interactions as BmoR is hypothesized to oligomerize upon ligand binding[24].

Biosensor activity in a simple gut community:

We also wanted to evaluate the capability of our transcription factor to assess butyrate levels produced by butyrate producing gut communities. We looked at a simple community consisting of *Bifidobacterium longum subs. infantis* (BL) and *Anaerostipes caccae* (AC)[34]. AC is a butyrate producer while BL is a not a butyrate producer. However, *Bifidobacterium longum* is shown to enhance butyrate production by improving metabolite cross-feeding[35] **(Fig 6a)**. We assessed the ability of the BmoR containing cells to detect differences in the butyrate conc. between spent media samples of monocultures of AC, BL and a co-culture of AC-BL (pair). Analytically determined, the samples had 0, 18 mM and 35 mM butyrate, respectively. We observe a shift in GFPuv fl. in butyrate containing populations when tested with both WT and BmoR^{10mut} with the latter having a wider separation between the samples **(Fig 6b)**. Compared to the WT, we see a 2- and 3-fold improvement in the engineered variant's activity in the AC and Pair butyrate containing samples, respectively **(Fig 6c)**.

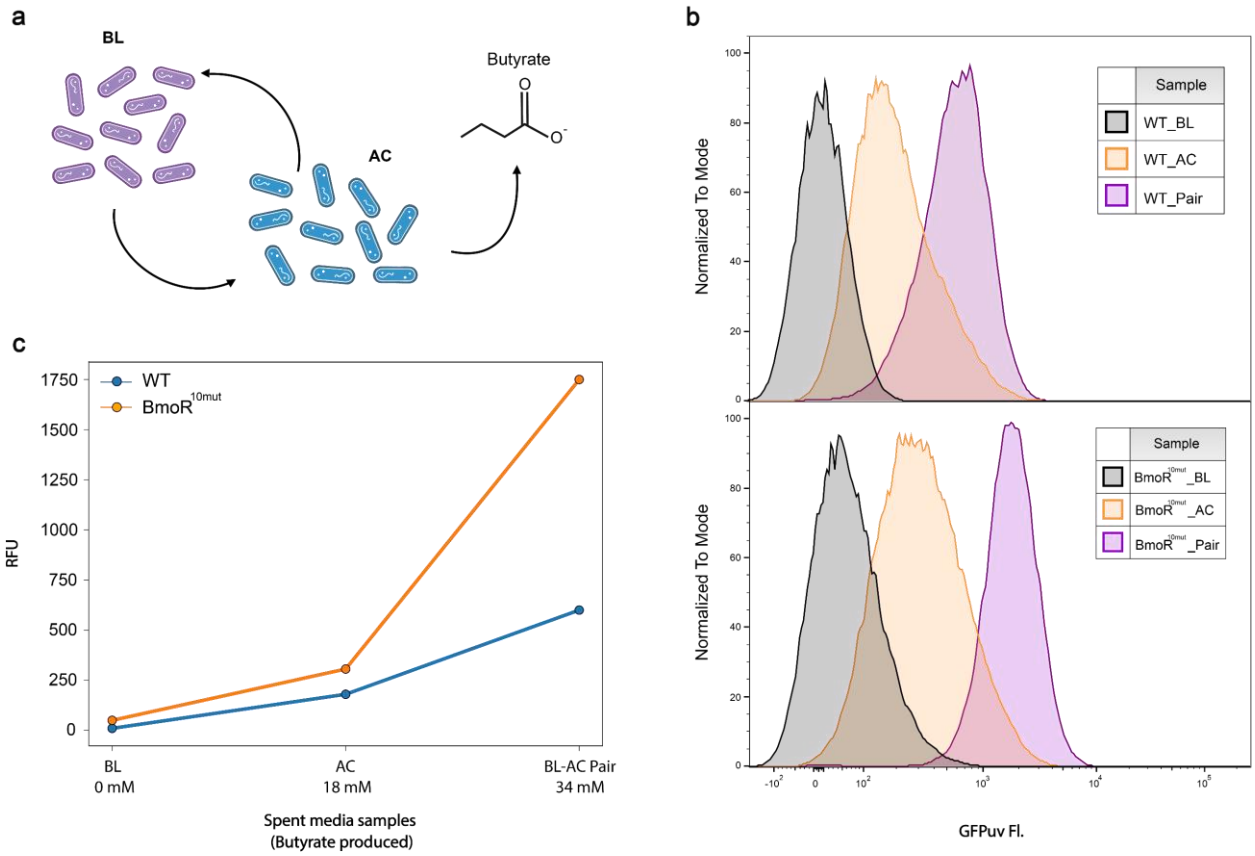


Figure 6: Spent media assay. (a) Here is a schematic of the simple butyrate producing community we tested. The two-member synthetic gut community comprises of *Bifidobacterium longum subs. infantis* (BL) and *Anaerostipes caccae* (AC). AC is a butyrate producer while BL is an enhancer and doesn't produce butyrate. The two species were cultured separately BL, AC, and together (Pair) yielding a low (0mM), medium (18mM) and high (34mM) butyrate production, respectively. This was measured analytically using LCMS. **(b)** Histogram of the activity of WT BmoR (top) and BmoR^{10mut} (bottom) in the various spent media samples. We see a rightward shift in the peaks of BmoR^{10mut} in the each of the samples. **(c)** The activity of WT and BmoR^{10mut} represented by line plot. We see BmoR^{10mut} have a steeper increase in fluorescence with increasing concentration of butyrate with a near 2-fold increase in AC spent media and a near 3-fold increase in the BL-AC Pair spent media samples as compared to WT. The values represent median activity of samples measured using flow cytometry.

Discussion:

Transcription factor engineering is a popular method for in vivo detection of small molecules leading to downstream activity which has valuable applications in building gene circuits[17,36]. Previous work was done on BmoR was focused on improving its activity towards its native activator, butanol, and other alcohols for improved titer production in metabolic engineering applications[26]. With our goal to build a biosensor for butyrate, a structurally similar molecule to butanol, BmoR provided a good target for protein engineering. However, BmoR is an activator, with various distinct functions, belonging to the σ^{54} dependent class of TFs making it a tricky protein to engineer[22] . To minimize variance among replicates of BmoR expression assay, we codon optimized the BmoR sequence and added two mutations TC→CA in the BmoR operator region. This helped improve the robustness of our system.

Initial work we did with BmoR showed some activity 50-100mM butyrate with nearly 10x dynamic range although nearly 15-fold lower than its activity towards butanol. This initial activity is a good first step to improve the sensitivity of BmoR's towards butyrate. The high throughput screening assay paired with NGS enabled a machine learning workflow for our engineering efforts. Using PU-learning on NGS reads we created a novel DMS dataset of the N-terminus activator binding region for BmoR. We filtered the PU-learning dataset to obtain 10 mutations proposed to have increased BmoR activity and low background activity. The combination of the 10 mutations, $BmoR^{10mut}$, showed a near 2-fold increase in its activity compared to WT with low background increase. Here, we showed an example of 10 substitutions introduced at once to improve activity when

collective introduction of mutations in a single sequence can be deleterious[37]. The mutations being distributed throughout the N-terminus region indicate more indirect interactions with the activator potentially showcasing epistatic effect of the mutational set.

The PU-learning model applied onto the obtained NGS reads allowed us to generate a better delineation between active and inactive mutations. This method can consider positive and negative selection by comparing data obtained under those varying condition enabling wide parallel screening from a single starting point. We generated a heatmap of the BmoR N-terminus region comparing mutations between the induced and uninduced populations. As expected, most mutations are deleterious or neutral towards BmoR activity with various pockets of activating mutations scattered throughout the protein. While there aren't any obvious trends that stand out, we focused on mutations that have increased activity in the butyrate induced population with low activity in the uninduced population. Obtaining the ten mutations was not difficult but getting a functioning protein when combining all ten mutations was very interesting. This method of utilizing NGS datasets for machine learning prediction is well suited for an accelerated engineering process. Moreover, it enables us to combine mutations that would not be easily accessible in a traditional directed evolution output. In future work, we want to create combinatorial libraries of picked mutations to find the optimal combinations.

Finally, we tested our biosensors in butyrate containing spent media of a simple butyrate producing community. This was an effective test as we got a good range of butyrate concentrations produced (0, 18mM and 34mM) to test our biosensor. We observed a 3-fold improvement in activity of $BmoR^{10mut}$ in this spent media assay as

compared to WT. Given that the various butyrate levels in the samples are in physiological range, BmoR^{10mut} is a valuable start as a biosensor for butyrate detection. Depending on need, work can be done to further improve BmoR's capabilities.

The use of BmoR^{10mut} as a biosensor for the gut still requires further testing. We would eventually like to put BmoR^{10mut} in mouse models upon verifying its activity *in vivo* with an expanded gut microbiome community. This would involve using alternative reporter systems better suited for anaerobic conditions such as Flavin-based fluorescent proteins (FbFPs)[38,39]. Developing biosensors such as BmoR^{10mut} is promising for use in a variety of applications such as *in vivo* protein activity characterization, downstream protein of interest expression, gene circuit design, enzyme engineering and therapeutic needs[40–42].

Materials and methods:

Strains, Media, and Materials:

The original WT BmoR plasmid, BmoR_GAF, was obtained from the Keasling Lab. Several modifications were made to this plasmid over time to implement golden gate cloning, reduce WT re-ligation and improve BmoR expression assay. The list of plasmids is in supplemental Table 1. The Luria Broth (Miller) media (DSL24400-2000), Sodium butyrate (303410-100G) and 1-butanol (B7906-500ML) was sourced from Sigma Aldrich. The Carbenicillin (Disodium) obtained from Goldbio (C-103-5). The LB (miller) Agar (30620042-4) used for making plates was obtained from bioworld. The consumables used in the project were sourced from various vendors. The 15mL round bottom culture tubes (352059) were corning Falcon tubes. We used VWR 15mL conical bottom tubes (525-0636) and 50mL conical bottom tubes (525-0610). The black with clear bottom 96-well microtiter plates were obtained from Thermo scientific (165305). The restriction enzymes and mastermixes for molecular biology were sourced from New England Biolabs.

Error-prone pcr:

We amplified the 1-250aa of the effector binding domain of BmoR using error-prone PCR to introduce random mutations. We reaction mixture comprised of (1M) Betaine, Standard Taq buffer (1x), $MgCl_2$ (5.5mM), DNA template (~200ng), 1:1:2.5:2.5 ratio of A:G:T:C (0.2mM A,G and 0.5mM T,C), $MnCl_2$ (0.05mM), forward and reverse primers (0.4mM each) and Taq DNA polymerase (5 units/100uL reaction volume). The pcr reaction was typically made in 400uL total volume and divided into 8 tubes of 50uL run in 8-pcr strip tubes (Thermo scientific, AB2000). The reaction was cycled with initial melting

at 95°C for 30sec, followed by 16 cycles of 95°C, 20s melt, 63°C,30 sec extension and 68°C,1 min elongation. The final extension was done at 68°C for 5 min. The pcr was run for 16 cycles to minimize amplification bias in a Bio-Rad thermocycler (T100). We digested the amplified library with DpnI (NEB R0176S) to remove DNA template followed by purification via Zymo DNA clean and concentration kit (cat# D4013). The insert amplification was verified using a 1% agarose gel. We wanted to keep the mutational rate in the library low (~1-2aa/gene) to allow generating a healthy functional fraction of BmoR variants while maximizing the size of the obtained library.

The backbone was stored in Thermo scientific One shot ccdB survival cells (A10460) as these cells contain the anti-toxin ccdA. The cells were grown overnight in 5mL LB cultures and plasmid prepped the following day using QIAprep Spin Miniprep Kit (Qiagen, 27104) following the kit's protocol to obtain backbone plasmid.

Electrocomp cell preparation:

We used an E Coli. MG1655 derivative (ΔadhE) for this study which required generation of electrocompetent cells. The cells in this study containing no plasmid, were grown overnight in LB culture as seed culture. The following day, we diluted the seed culture 1:100 in 200mL of SOB media (RPI, S25000-1000) grown to ~0.8-1 OD and cooled to 4°C. The 200mL culture was divided in 4, 50mL conical tubes and centrifuged at 2500xg, to 4°C for 7 minutes on a Thermo scientific Sorvall ST 16 tabletop centrifuge. The supernatant was discarded and each of the tube washed twice with 50mL, 10% glycerol containing MilliQ water. The cells from each tube resuspended in 1mL of 10% glycerol water collected in a 15mL conical tubes with a final centrifugation step. The cells were concentrated in 500uL of total volume and ready for transformation. Cells not used

immediately are aliquoted 75uL each in 1.5mL Eppendorf tubes, flash frozen with liquid nitrogen and stored as stocks in -80°C.

Golden gate cloning and transformation:

A 50 uL golden gate mastermix was made by combining 1M Betaine, 1x T4 DNA ligase buffer, 1000units T4 DNA Ligase (NEB, M0202S), and 50units BsaI-HFv2 (NEB, R3733S) restriction enzyme. The insert and backbone were added in a >2:1 molar ratio respectively and water to reach the total volume of 50uL in a single pcr tube. The reaction was cycled for 1min @37°C and 1min @16°C for ~50-60 cycles. The samples were cleaned using the Zymo DNA clean and concentration kit.

We transformed the ligated plasmid library into the freshly prepared electrocompetent cells via electroporation in 1mm cuvettes (Fisher, FB101) at 1.80Kv using the Bio-rad MicroPulser. The transformed cells were recovered for 1 hour @37°C and diluted into 50mL, carbenicillin containing LB media. After ~12-14 hr growth @30°C, we made 5, 1mL aliquots of the culture in 15% Glycerol to store as redundant stocks for future use.

Post 1 hr recovery, dilution plates with 500x, 5,000x and 50,000x dilution were inoculated from the 50mL culture and grown overnight @37°C. We obtained a library size of ~1.2*10^6 transformants, analyzed by counting colonies. Individual random 10 colonies were picked for colony pcr followed by Sanger sequencing (Functional Biosciences).

Flow Cytometry and FACS:

WT BmoR, Empty and any variants were grown overnight in 5mL LB cultures with carbenicillin for 250rpm, 20-22h @30°C. We then diluted the saturated cultures 1:100 (50uL in 5000uL) in fresh 5mL LB media with antibiotics. The expression cultures are grown for 250rpm, 16 h @37°C with varying concentrations of butyrate: 0, 1.5, 3.13 6.3, 12.5, 25, 50 and 100mM added after inoculation. The following day, expression cultures were washed 2 times with 1mL phosphate-buffered saline, PBS (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4 and 2 mM KH_2PO_4) with centrifugation for 10 min @3000xg. We resuspended the cells in 1mL PBS and then diluted the resuspension (1:50) in 1mL fresh PBS in 5mL polypropylene tubes (Fisher cat# 352058). For Flow cytometry, we ran the samples in a BD LSRFortessa X-20. The samples were excited @405nM and detected @510nM.

For FACS, the samples above are ready for sorting. To collect the sorted cells, we made a collection tube containing 1mL of LB Media with antibiotics for each sample of sorted cells. The samples were run in a BD FACSAria III at the same excitation and emission. The collected cells were grown overnight to an OD ~1-2 to avoid significant amount of cell death. The samples were analyzed using FlowJo. We duplicated the screening to get a second replicate as a means of verifying integrity of the obtained NGS data in the next step.

Next-generation sequencing:

The pre-sorted and post-sorted cultures from the FACS screening were miniprepped using the QIAprep Spin Miniprep Kit to obtain the library of plasmids in each

sample. We digested the post-screening plasmid library samples using PpuMI and EcoNI to obtain a 1200bp region for illumina sequencing. The digests were run on a 1% agarose gel and extracted using QIAquick Gel Extraction Kit (Qiagen, 28704) following the kit's protocol. The samples were submitted to UW-Madison Next generation Sequencing core to prepare NGS libraries using the Celero™ DNA-Seq Library Prep for the Illumina Nova-seq. We collected $10^7 \pm 2*10^6$ reads for pre-sorted samples and $5*10^6 \pm 10^6$ for the post-sorted populations.

Spent media assay:

Strain culturing for monoculture and coculture supernatants were performed according to methods in Clark 2021, scaled to a 30 mL culture volumes in 50 mL sterile falcon tubes to provide sufficient material for biosensor assays. The target inoculation density for each strain was 0.005, totaling 0.01 for the coculture. Optical density was measured using 200 uL of sample in a Tecan F200 plate reader in standard clear, flat bottom 96-well microplates (Grenier). Inoculation volumes were calculated as Volume(inoc) = Volume(well)*0.01 OD / (Preculture OD). Strains cultures were incubated anaerobically at 37°C for approximately 48 hours. Supernatant was harvested by decanting after centrifugation at 4000 rpm for 20 minutes in Sorvall ST 16R centrifuge (Thermo Scientific). WT and $BmoR^{10mut}$ were grown overnight in 5mL LB cultures with carbenicillin for 250rpm, 20-22h @30°C. We diluted the saturated cultures 1:100 (50uL in 5000uL) in each of the obtained supernatants, AC, BL, and AC-BL coculture (Pair). The samples were prepared, and fluorescence measured according to the Flow setup described earlier. The data was analyzed using FlowJo.

Organic acids were quantified using high performance liquid chromatography according to Clark 2021. Supernatant samples were thawed in a room temperature water bath before addition of 2 µL of H2SO4 to 180 uL of supernatant to precipitate any components that might be incompatible with the running buffer. The samples were then centrifuged at 2400 × g for 10 min and then 150 µL of each sample was filtered through a 0.2 µm filter using a vacuum manifold before transferring 70 µL of each sample to an HPLC vial. HPLC analysis was performed using a Shimadzu HPLC system equipped with a SPD-20AV UV detector (210 nm). Compounds were separated on a 250 × 4.6 mm Rezex© ROA-Organic acid LC column (Phenomenex Torrance, CA) run with a flow rate of 0.2 mL min−1 and at a column temperature of 50 °C. The samples were held at 4 °C prior to injection. Separation was isocratic with a mobile phase of HPLC grade water acidified with 0.015 N H2SO4 (415 µL L−1). At least two standard sets were run along with each sample set. Standards were 100, 20, and 4 mM concentrations of butyrate. The injection volume for both sample and standard was 20 µL. The resultant data was analyzed using the Shimadzu LabSolutions software package.

Data Analysis:

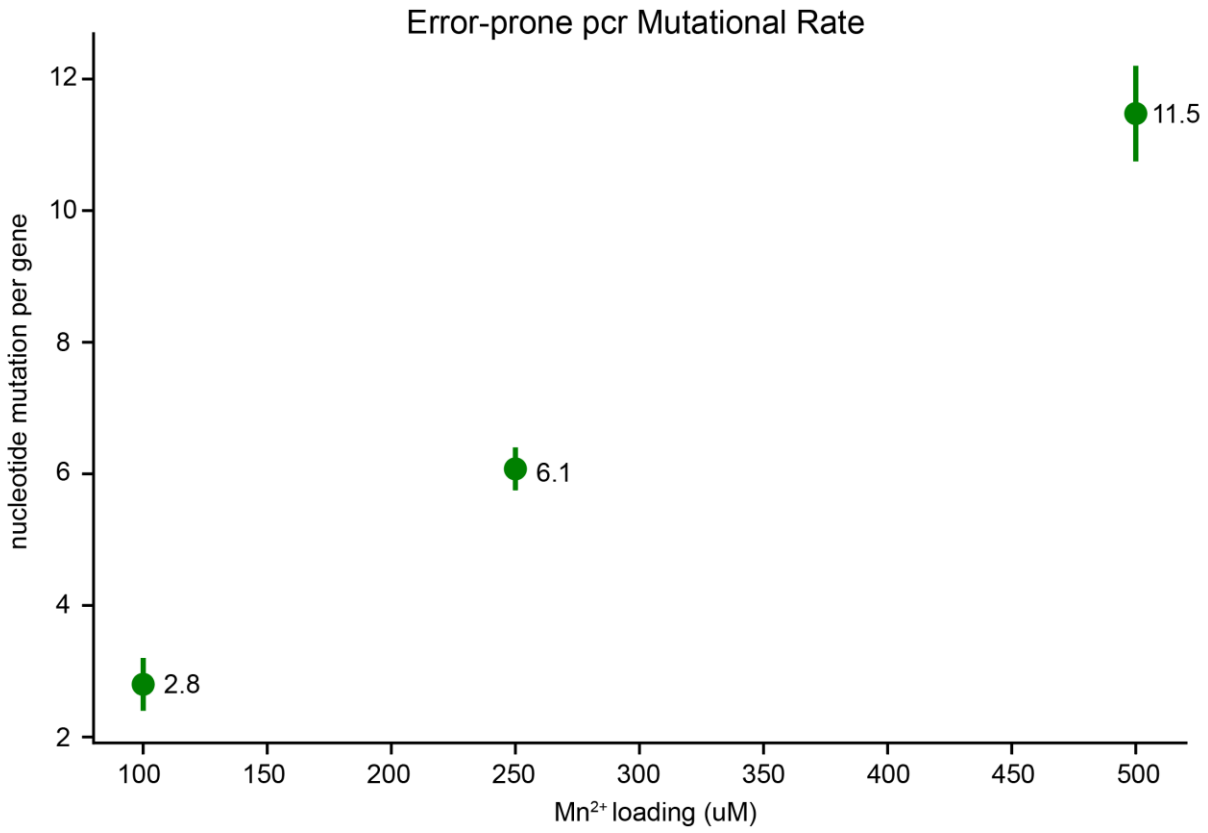
The data analysis and machine learning for the project was done using python. The code is in the romerolab Github repo.

Supplementary Figures:

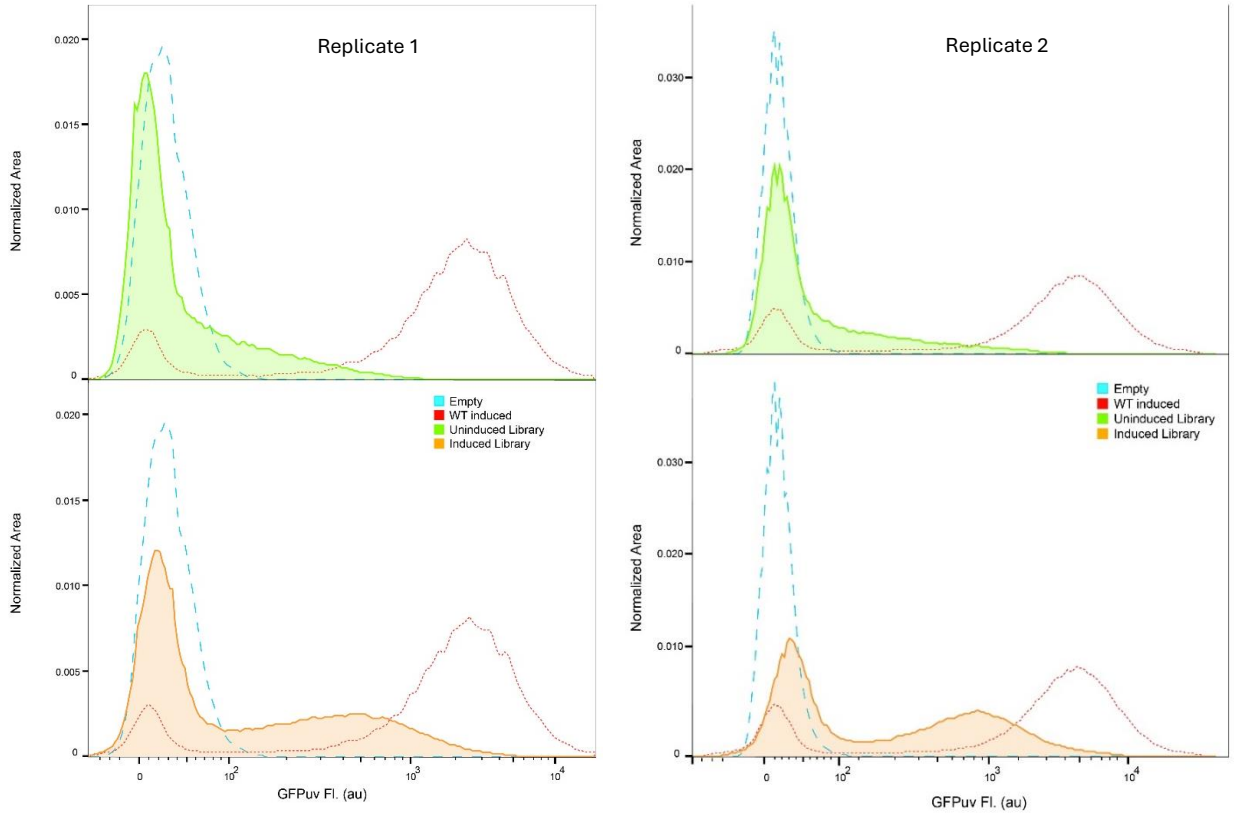
Table 1:

Plasmid	Composition	Description	Source
BmoR_GAF	P_{bmoR}-bmoR; P_{bmo}-gfpuv; colE1; amp^r	Original BmoR plasmid obtained from literature	Dietrich et. al. (2013)
BmoR_GG	P_{bmoR}-bmoR; P_{bmo}-gfpuv; colE1; amp^r	WT BmoR plasmid with golden gate compatibility	This study
BmoR_10mut_GG	P_{bmoR}-$bmoR^{10mut}$; P_{bmo}-gfpuv; colE1; amp^r	$BmoR^{10mut}$ plasmid with golden gate compatibility	This study
BmoR_opt_GG	P_{bmoR}-bmoR; P_{bmo}-gfpuv; colE1; amp^r	The refactored version of the WT BmoR.	This study
BmoR_opt_10mut_GG	P_{bmoR}-$bmoR^{10mut}$; P_{bmo}-gfpuv; colE1; amp^r	The refactored version of the $Bmor^{10mut}$.	This study
BmoR_GG_BB	P_{bmoR}-bmoR-ccdB; P_{bmo}-gfpuv; colE1; amp^r	ccdB containing backbone for	This study
BmoR_opt_GG_BB	P_{bmoR}-bmoR-CcdB; P_{bmo}-gfpuv; colE1; amp^r	The refactored version of the backbone plasmid.	This study
pJ7_pET22hc_ccdB	LacI ; T7- ccdB; F1 ori; amp^r	Plasmid used to clone ccdB into the above plasmids	Romero Lab stock

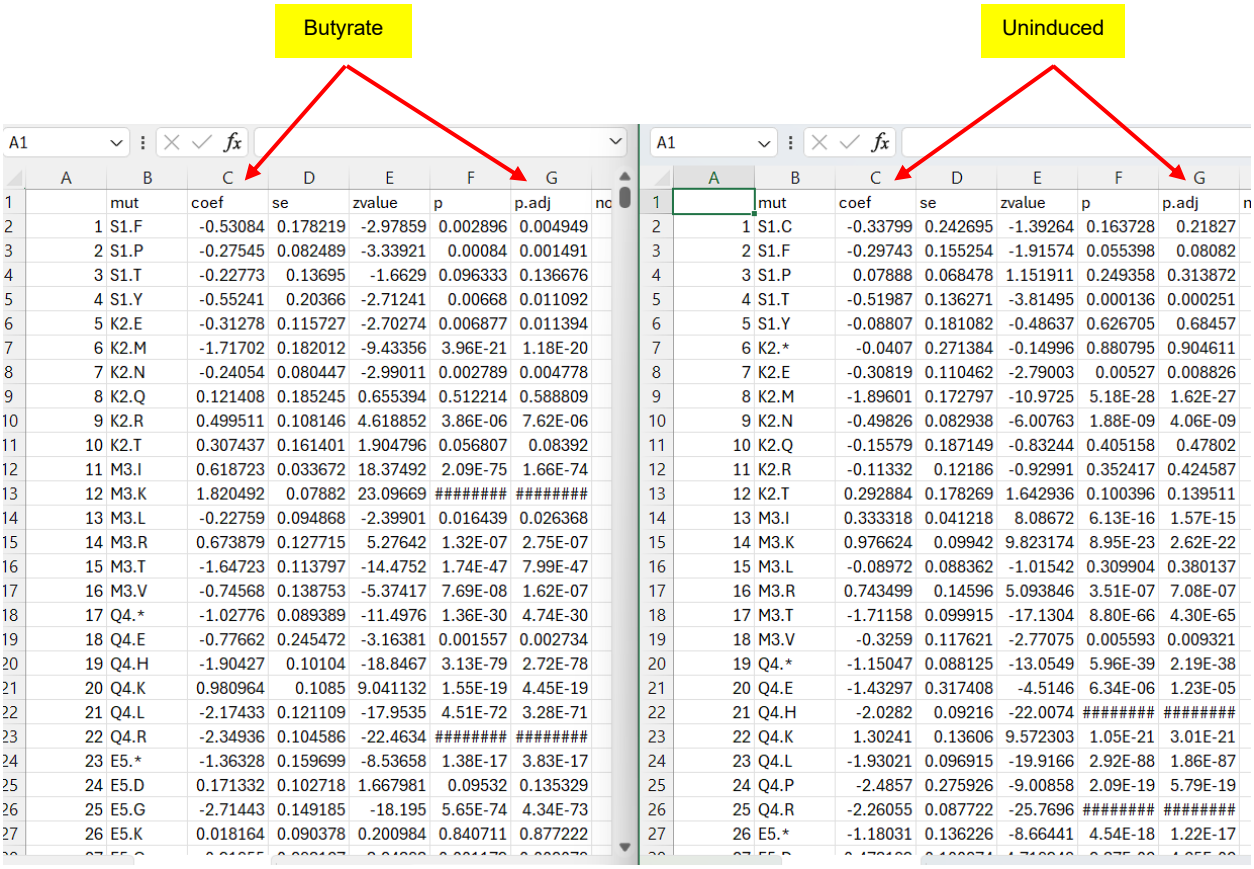
Supp Table 1: The plasmids used in this study



S1: We compared the effect of Mn^{2+} loading in generating library diversity. We observe a near linear relationship of Mn^{2+} and nucleotide mutation/gene. Our per gene count is 750aa. Increased mutations per gene gives a highly diverse library while potentially compromising the percent active library.



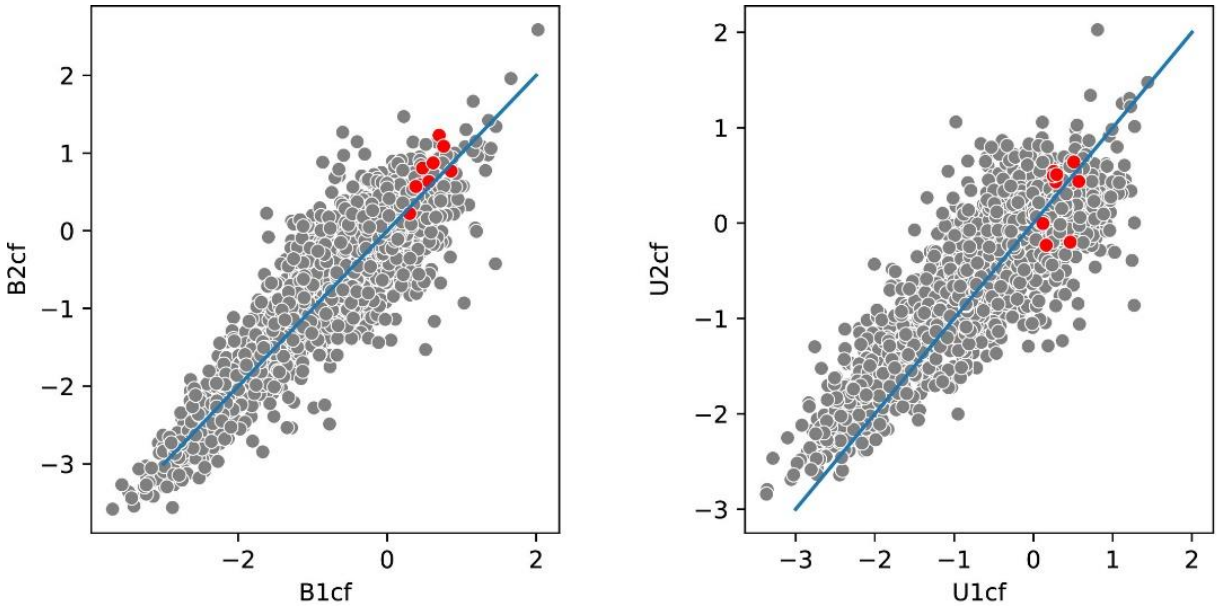
S2: We screened the error-prone pcr generated mutagenic library using FACS at either 0mM or 100mM butyrate. We set a threshold gate at 25% of the top performers to obtain and compare the mutational pattern between the two conditions. We replicated the FACS screen for data integrity validation for downstream data analysis.



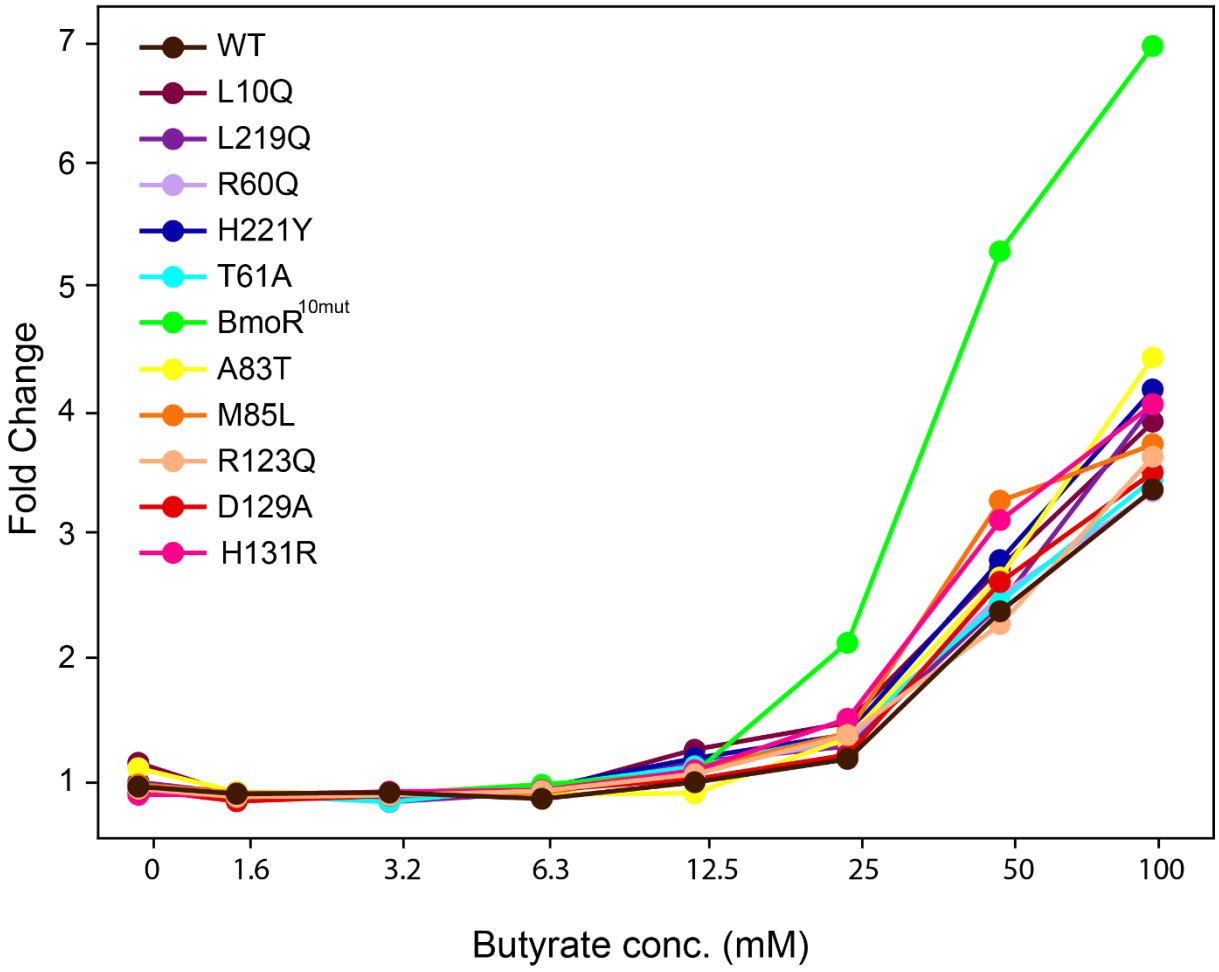
		mut	coef	se	zvalue	p	p.adj	nc
1		mut	coef	se	zvalue	p	p.adj	
2	1	S1.F	-0.53084	0.178219	-2.97859	0.002896	0.004949	
3	2	S1.P	-0.27545	0.082489	-3.33921	0.00084	0.001491	
4	3	S1.T	-0.22773	0.13695	-1.6629	0.096333	0.136676	
5	4	S1.Y	-0.55241	0.20366	-2.71241	0.00668	0.011092	
6	5	K2.E	-0.31278	0.115727	-2.70274	0.006877	0.011394	
7	6	K2.M	-1.71702	0.182012	-9.43356	3.96E-21	1.18E-20	
8	7	K2.N	-0.24054	0.080447	-2.99011	0.002789	0.004778	
9	8	K2.Q	0.121408	0.185245	0.655394	0.512214	0.588809	
10	9	K2.R	0.499511	0.108146	4.618852	3.86E-06	7.62E-06	
11	10	K2.T	0.307437	0.161401	1.904796	0.056807	0.08392	
12	11	M3.I	0.618723	0.033672	18.37492	2.09E-75	1.66E-74	
13	12	M3.K	1.820492	0.07882	23.09669	########	########	
14	13	M3.L	-0.22759	0.094868	-2.39901	0.016439	0.026368	
15	14	M3.R	0.673879	0.127715	5.27642	1.32E-07	2.75E-07	
16	15	M3.T	-1.64723	0.113797	-14.4752	1.74E-47	7.99E-47	
17	16	M3.V	-0.74568	0.138753	-5.37417	7.69E-08	1.62E-07	
18	17	Q4.*	-1.02776	0.089389	-11.4976	1.36E-30	4.74E-30	
19	18	Q4.E	-0.77662	0.245472	-3.16381	0.001557	0.002734	
20	19	Q4.H	-1.90427	0.10104	-18.8467	3.13E-79	2.72E-78	
21	20	Q4.K	0.980964	0.1085	9.041132	1.55E-19	4.45E-19	
22	21	Q4.L	-2.17433	0.121109	-17.9535	4.51E-72	3.28E-71	
23	22	Q4.R	-2.34936	0.104586	-22.4634	########	########	
24	23	E5.*	-1.36328	0.159699	-8.53658	1.38E-17	3.83E-17	
25	24	E5.D	0.171332	0.102718	1.667981	0.09532	0.135329	
26	25	E5.G	-2.71443	0.149185	-18.195	5.65E-74	4.34E-73	
27	26	E5.K	0.018164	0.090378	0.200984	0.840711	0.877222	

		mut	coef	se	zvalue	p	p.adj	n
1		mut	coef	se	zvalue	p	p.adj	
2	1	S1.C	-0.33799	0.242695	-1.39264	0.163728	0.21827	
3	2	S1.F	-0.29743	0.155254	-1.91574	0.055398	0.08082	
4	3	S1.P	0.07888	0.068478	1.151911	0.249358	0.313872	
5	4	S1.T	-0.51987	0.136271	-3.81495	0.000136	0.000251	
6	5	S1.Y	-0.08807	0.181082	-0.48637	0.626705	0.68457	
7	6	K2.*	-0.0407	0.271384	-0.14996	0.880795	0.904611	
8	7	K2.E	-0.30819	0.110462	-2.79003	0.00527	0.008826	
9	8	K2.M	-1.89601	0.172797	-10.9725	5.18E-28	1.62E-27	
10	9	K2.N	-0.49826	0.082938	-6.00763	1.88E-09	4.06E-09	
11	10	K2.Q	-0.15579	0.187149	-0.83244	0.405158	0.47802	
12	11	K2.R	-0.11332	0.12186	-0.92991	0.352417	0.424587	
13	12	K2.T	0.292884	0.178269	1.642936	0.100396	0.139511	
14	13	M3.I	0.333318	0.041218	8.08672	6.13E-16	1.57E-15	
15	14	M3.K	0.976624	0.09942	9.823174	8.95E-23	2.62E-22	
16	15	M3.L	-0.08972	0.088362	-1.01542	0.309904	0.380137	
17	16	M3.R	0.743499	0.14596	5.093846	3.51E-07	7.08E-07	
18	17	M3.T	-1.71158	0.099915	-17.1304	8.80E-66	4.30E-65	
19	18	M3.V	-0.3259	0.117621	-2.77075	0.005593	0.009321	
20	19	Q4.*	-1.15047	0.088125	-13.0549	5.96E-39	2.19E-38	
21	20	Q4.E	-1.43297	0.317408	-4.5146	6.34E-06	1.23E-05	
22	21	Q4.H	-2.0282	0.09216	-22.0074	########	########	
23	22	Q4.K	1.30241	0.13606	9.572303	1.05E-21	3.01E-21	
24	23	Q4.L	-1.93021	0.096915	-19.9166	2.92E-88	1.86E-87	
25	24	Q4.P	-2.4857	0.275926	-9.00858	2.09E-19	5.79E-19	
26	25	Q4.R	-2.26055	0.087722	-25.7696	########	########	
27	26	E5.*	-1.18031	0.136226	-8.66441	4.54E-18	1.22E-17	

S3: The PU-learning outputs an excel file with various parameters for each mutation. For our needs, we focus on two, the coefficients value and the P_{adj} values that show the model's confidence in its output. These parameters then act as filters for our dataset when deciding on mutations.



S4: Correlation scatter plot of the replicates. Left: Butyrate coefficients, right: Uninduced coefficients. The pearson r >0.90 indicates decent correlation amongst the replicates. The grey dots represent all mutations while the red dots indicate the chosen mutations.



S5: Activity of WT BmoR, BmoR^{10mut} and the single mutation variants. We observe BmoR10mut to have 2-fold improved activity as compared to WT and the other single variants. This outcome, while promising was very qualitative and had to be optimized for robust activity. We needed to re-factor the system by introducing codon-optimization and BmoR operator region mutations (TC->CA) along with all our optimization results to obtain more reproducible results. Since we observed BmoR10mut to be a better variant, we focused on refactoring only the WT and the BmoR10mut.

References:

1. Roth, G. A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *J. Am. Coll. Cardiol.* (2020) doi:10.1016/j.jacc.2020.11.010.

2. Wang, L. *et al.* The role of the gut microbiota in health and cardiovascular diseases. *Springer Nat.* 3, (2022).

3. The gut microbiota as a novel regulator of cardiovascular function and disease. (2018).

4. Li, Q. *et al.* Butyrate Suppresses the Proliferation of Colorectal Cancer Cells via Targeting Pyruvate Kinase M2 and Metabolic Reprogramming. *Elsevier BV* 17, 1531–1545 (2018).

5. Canani, R. B., Costanzo, M. D. & Leone, L. The epigenetic effects of butyrate: potential therapeutic implications for clinical practice. *BioMed Cent.* 4, (2012).

6. Guilloteau, P. *et al.* From the gut to the peripheral tissues: the multiple effects of butyrate. *Nutr. Res. Rev.* (2010) doi:10.1017/s0954422410000247.

7. Bridgeman, S. C. *et al.* Butyrate generated by gut microbiota and its therapeutic role in metabolic syndrome. *Pharmacol. Res. Print* (2020) doi:10.1016/j.phrs.2020.105174.

8. Besten, G. den *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Elsevier BV* 54, 2325–2340 (2013).

9. Nguyen, J., Pepin, D. M. & Tropini, C. Cause or effect? The spatial organization of pathogens and the gut microbiota in disease. *Microbes Infect.* 23, 104815 (2021).

10. Liu, J. *et al.* Functions of Gut Microbiota Metabolites, Current Status and Future Perspectives. *Aging Dis.* 13, 1106–1126 (2022).

11. Martin-Gallausiaux, C., Marinelli, L., Blottière, H. M., Larraufie, P. & Lapaque, N. SCFA: mechanisms and functional importance in the gut. *Proc. Nutr. Soc.* 80, 37–49 (2021).

12. Bach Knudsen, K. E. *et al.* Impact of Diet-Modulated Butyrate Production on Intestinal Barrier Function and Inflammation. *Nutrients* 10, 1499 (2018).

13. Du, K., Bereswill, S. & Heimesaat, M. M. A literature survey on antimicrobial and immune-modulatory effects of butyrate revealing non-antibiotic approaches to tackle bacterial infections. (2021).

14. Greenhalgh, J. C., Fahlberg, S. A., Pfleger, B. F. & Romero, P. A. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nat. Portf.* 12, (2021).

15. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* 16, 687–694 (2019).

16. Vaaben, T. H., Vazquez-Uribe, R. & Sommer, M. O. A. Characterization of Eight Bacterial Biosensors for Microbial Diagnostic and Therapeutic Applications. *ACS Synth. Biol.* (2022) doi:10.1021/acssynbio.2c00491.

17. Mahr, R. & Frunzke, J. Transcription factor-based biosensors in biotechnology: current state and future prospects. *Appl. Microbiol. Biotechnol.* 100, 79–90 (2016).

18. Cheng, F., Tang, X. & Kardashliev, T. Transcription Factor-Based Biosensors in High-Throughput Screening: Advances and Applications. *Wiley-Blackwell* 13, (2018).

19. Liu, Y., Liu, Y. & Wang, M. Design, Optimization and Application of Small Molecule Biosensor in Metabolic Engineering. (2017).

20. Bai, Y. & Mansell, T. J. Production and Sensing of Butyrate in a Probiotic E. coli Strain. *Int. J. Mol. Sci.* 21, 3615 (2020).

21. Kurth, E. G., Doughty, D. M., Bottomley, P. J., Arp, D. J. & Sayavedra-Soto, L. A. Involvement of BmoR and BmoG in n-alkane metabolism in ‘Pseudomonas butanovora’. *Microbiology* 154, 139–147 (2008).

22. Buck, M., Gallegos, M.-T., Studholme, D. J., Guo, Y. & Gralla, J. D. The Bacterial Enhancer-Dependent ς54(ςN) Transcription Factor. *J. Bacteriol.* (2000) doi:10.1128/jb.182.15.4129-4136.2000.

23. Dietrich, J. A., Shis, D. L., Alikhani, A. & Keasling, J. D. Transcription Factor-Based Screens and Synthetic Selections for Microbial Small-Molecule Biosynthesis. *ACS Synth. Biol.* 2, 47–58 (2013).

24. Yu, H. *et al.* Engineering transcription factor BmoR for screening butanol overproducers. *Metab. Eng.* 56, 28–38 (2019).

25. Yu, H. *et al.* Establishment of BmoR-based biosensor to screen isobutanol overproducer. *Microb. Cell Factories* 18, 30 (2019).

26. Wu, T., Chen, Z., Guo, S., Zhang, C. & Huo, Y.-X. Engineering Transcription Factor BmoR Mutants for Constructing Multifunctional Alcohol Biosensors. *ACS Synth. Biol.* 11, 1251–1260 (2022).

27. Danson, A. E., Jovanovic, M., Buck, M. & Zhang, X. Mechanisms of σ54-Dependent Transcription Initiation and Regulation. *J. Mol. Biol.* 431, 3960–3974 (2019).

28. Lin-Goerke, J., Robbins, D. J. & Burczak, J. D. PCR-Based Random Mutagenesis Using Manganese and Reduced dNTP Concentration. *Biotechniques/BioTechniques* 23, 409–412 (1997).

29. Engler, C., Kandzia, R. & Marillonnet, S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PloS One* 3, e3647–e3647 (2008).

30. Langmead, B. Aligning Short Sequencing Reads with Bowtie. (2010).

31. Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Syst.* 12, 92-101.e8 (2021).

32. Kim, N. M., Sinnott, R. W., Rothschild, L. N. & Sandoval, N. R. Elucidation of Sequence–Function Relationships for an Improved Biobutanol In Vivo Biosensor in E. coli. *Front. Bioeng. Biotechnol.* 10, (2022).

33. Bush, M. & Dixon, R. The Role of Bacterial Enhancer Binding Proteins as Specialized Activators of σ54-Dependent Transcription. *Microbiol. Mol. Biol. Rev.* 76, 497–529 (2012).

34. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and function. *bioRxiv* 2020.08.19.241315 (2020) doi:10.1101/2020.08.19.241315.

35. Belenguer, Á. *et al.* Two Routes of Metabolic Cross-Feeding between Bifidobacterium adolescentis and Butyrate-Producing Anaerobes from the Human Gut. *Am. Soc. Microbiol.* 72, 3593–3599 (2006).

36. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. (2015).

37. Loewe, L. & Hill, W. G. The population genetics of mutations: good, bad and indifferent. 1153–1167 (2010) doi:https://doi.org/10.1098/rstb.2009.0317.

38. Lobo, L. A., Smith, C. J. & Rocha, E. R. Flavin mononucleotide (FMN)-based fluorescent protein (FbFP) as reporter for gene expression in the anaerobe Bacteroides fragilis. *Oxf. Univ. Press* 317, 67–74 (2011).

39. Drepper, T. *et al.* Reporter proteins for in vivo fluorescence without oxygen. *Nat. Portf.* 25, 443–445 (2007).

40. Design and Development of Biosensors for the Detection of Heavy Metal Toxicity. (2011) doi:10.4061/2011/343125.

41. Ibraheem, A. & Campbell, R. E. Designs and applications of fluorescent protein-based biosensors. *Elsevier BV* 14, 30–36 (2010).

42. Palmer, A. E., Qin, Y., Park, J. G. & McCombs, J. E. Design and application of genetically encoded biosensors. *Elsevier BV* 29, 144–152 (2011).

Chapter 3: Engineering CymR for nanomolar level detection of *p*-cumate

Abstract:

Cereal crop production heavily relies on the use of inorganic nitrogen fertilizers to meet global food demands but polluting the environment in the process. To create a sustainable alternative to fertilizers, we propose enhancing the natural nitrogen fixation ability of soil bacteria. Constitutive expression of nitrogenase enzymes leads to toxicity to the host cells; We hypothesized the use of plant-root released cumate (*p*-isopropyl benzoate) to put nitrogen fixation under inducible control. CymR is a regulator that natively binds to cumate at levels >1uM but we need it to be sensitive in the nanomolar range. In this work, we performed one round directed evolution using high throughput screening and ML-guided mutation prediction to create an enhanced variant of CymR. Our best variants showed nearly 10x enhanced activity compared to WT with detection in the 100-1000nM range.

Introduction:

Nitrogen is an essential element needed by all organisms as a building block to make DNA and protein macro molecules. Despite being the most abundant element in the atmosphere, gaseous Nitrogen (N_2) is not easily assimilated by most organism in its atmospheric state. The process of converting N_2 into its more bioavailable form ammonia (NH_3) is called biological nitrogen fixation (BNF), a very energy demanding process. It is estimated that for 1 mol of N_2 consumption, the host cell consumes 16 mol of ATP[1]. Plants

and other autotrophs do not have capability of BNF and instead rely on other organisms called diazotrophs to fulfill their nitrogen requirements. Diazotrophs are bacteria or archaea that possess the capability of BNF by catalyzing the conversion of N_2 into NH_3 nitrogenase enzyme complex[2,3]. This symbiotic relationship provides a resource rich environment for the microbes while eliminating the metabolic load of BNF onto the plants.

The availability of fixed nitrogen poses a massive limitation on our production of cereal crops such as corn, rice and wheat which provide 50% of the global calories[4]. To offset the lack of key minerals for agriculture, we rely heavily on the use of nitrogen based chemical fertilizers. With overuse of fertilizers to unsustainable levels, we have created a flurry of environmental problems from greenhouse gas emission[5,6] increase to dead zones in coastal oceans[7]. Moreover, half of the available nitrogen gets lost to the environment rather than being absorbed by plant roots[8]. So, not only do we end up polluting the land and water with chemical fertilizers, we do not even get the maximum benefit. There has been a strong need to address the use of fertilizers while maintaining or even increasing the crop yields to serve societal needs especially as the global hunger stress grows rapidly[9–11].

A potential approach to eliminate this reliability on fertilizers involved engineering plants to contain the BNF activity themselves. First-generation of self-fertilizing cereal crops (Maize and barley) focused on improving their association with the rhizobacteria by secreting symbiosis-promoting secondary metabolites called rhizopines[12–14]. Second-generation of self-fertilizing cereal crops (rice) aim to improve the uptake of nitrogen into plant roots via the soil environment by expressing NF receptors[15] and key regulators of

nodule organogenesis[16] to further enhance crop association to the soil microbiome. The third generation of self-fertilizing crop production aims to eliminate the need for bacteria by introducing N-fixation directly into plants. Researchers thought to achieve this by introducing nitrogenase subunits in mitochondria and chloroplasts due to the oxygen sensitive nature of N-fixation[17–19]. While engineered crops provide a sustainable path forward, the task to do so is an extremely intensive undertaking in synthetic biology. It needs strong collaborative and multi-disciplinary expertise to engineer these crops for an eco-friendly agricultural solution. The complexity of plant metabolism, weak portability of nitrogenase gene clusters from bacteria and slow growth of the plants for rapid phenotyping prevent crop modification from being a viable short-term approach for the growing food insecurity. Apart from the scientific barriers, society at large is currently unable to accept “genetically modified plants” as an option for crop production and consumption[20].

An alternate approach to increasing nitrogen availability in plants is right under its roots. We can harness the capability of diazotrophs that reside in the plant rhizosphere by engineering strains of soil bacteria that natively fix nitrogen to do the process better as they can already provide up to 20-25% of cereal crop nitrogen requirements[21]. This would provide a viable short-term alternative to chemical fertilizers while we learn to make and accept modified crops for a sustainable and real future. BNF in proteobacteria is very tightly regulated by the PII signal transduction proteins, GlnB and GlnK, that integrate several metabolic signals (glutamine, ADP, ATP and 2-oxoglutarate) for the expression of nitrogenase enzyme NifA called the GS-GOGAT metabolic pathway[22]. Glutamine (Glu) and 2-oxoglutarate (2-OG) serve as the switch between nitrogen fixation and nitrogen

assimilation as they represent the nitrogen:carbon ratio in the cell. Under low ammonia conditions represented by high amounts of 2-OG, PII enzymes are uridynalyted by the uridylyltransferase (UTase) activity of GlnD which prevents the interaction of the PII complex to nifL-nifA regulatory complex which controls expression of downstream nitrogen fixation genes[23]. Brewin et. al. tested several methods of enhancing ammonia production in *Azotobacter vinelandii*, a common soil proteobacteria. Among various proteins involved in the production of ammonia, they found only mutations that deactivate the regulatory NifLA complex led to accumulation of up to 35mM (1000x higher than normal) of fixed nitrogen. This significant increase in ammonia can be attributed to the deregulation of the very tightly controlled BNF[24]. However, this level of nitrogenase activity puts a massive metabolic load on the bacterial cells making this a toxic system. Secondly, they also saw higher levels of intracellular ammonium than in the extracellular medium which brings into question how much fixed nitrogen is being delivered to plants. Using a constitutively active nitrogen fixation system is not a viable approach to engineering rhizosphere bacteria.

A good alternative to constitutive expression would be to put the symbiotic bacteria under inducible control of plant root exudates, primary and secondary metabolites that are released by plants via the roots into the soil. Root exudates help prevent growth of harmful organisms in the plant soil microbiome while promoting the growth of helpful symbiotic microorganisms. They can account for nearly 10% of fixed carbon made by plants and released into the soil for the rhizosphere[25,26]. Root exudates are also involved in shaping the environment such as the soil pH levels, nutrient availability and chelating toxic compounds. Measuring the amount of root exudates is a non-trivial task that can

involve complex root extraction and analysis methods. Kawasaki et. al. describe a sterile hydroponic system to characterize root exudates in wheat and barley showcasing the first example of aluminum-activated malate exudation in major wheat roots[27]. There aren't any known studies that directly characterize the concentration of *p*-cumate as root exudates or in plant roots, however, unpublished work done by my collaborators indicate nanomolar concentration of *p*-cumate using analytical methods.

To study released *p*-cumate in the soil and put BNF under inducible control, we hypothesized the use of transcription factor based biosensor that can detect low levels of *p*-cumate *in vivo*. In this work we aimed to engineer parent CymR, a bacterial transcription factor, for sensing cumate in the 100-1000nM range. We combined with our lab's expertise on ML modeling for mutation prediction with traditional D.E. methods to discover variants with nearly 10x improved sensitivity towards CA. This can be integrated in the genome for enhanced nitrogenase activity without heavily increasing the metabolic load on the host cells.

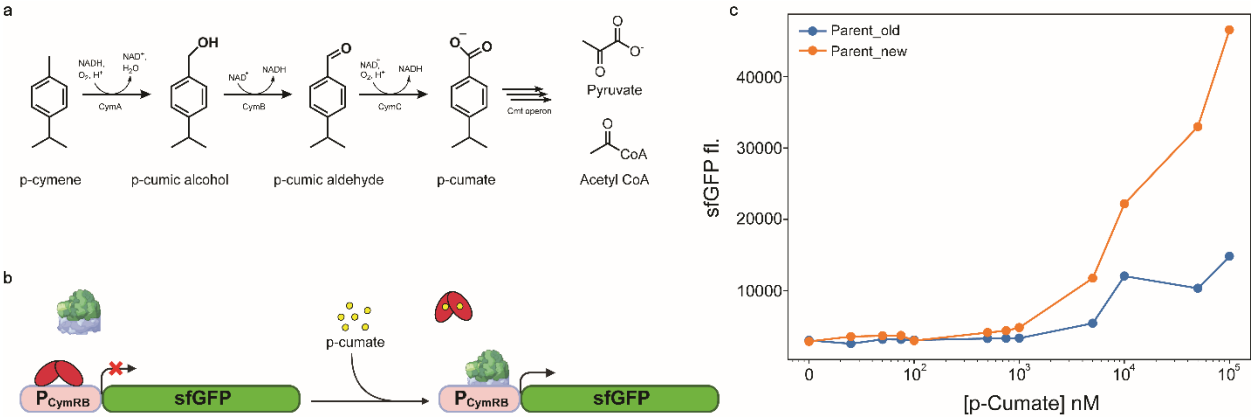


Figure 7: Elucidating *Pseudomonas Putida* derived CymR activity. (a) *p*-cumate is key intermediate made by soil bacteria serving as a valuable source of carbon. Enzymes in the Cym operon oxidize *p*-cymene into *p*-cumate which gets metabolized into TCA relevant molecules (pyruvate and acetyl CoA). **(b)** Directed evolution requires a robust highthroughput reporter assay to quickly assess the fitness landscape. In our system, the CymR homodimer is natively bound to the operator region physically blocking the access of RNA polymerase to the promoter. In the presence of enough *p*-cumate, CymR binds to the signal, undergoes conformation change and enables RNAP-mediated gene expression of sfGFP. The fluorescent output can be detected using flow cytometry. **(c)** Parent CymR shows activity towards *p*-cumate at concentrations >1uM, saturating around 1000uM. Initially, our plasmid system was oriented such that the sfGFP was upstream of the CymR expression, Parent_old (blue). For easier genomic integration, we reoriented the plasmid to have CymR upstream of sfGFP, Parent_new (orange). We noticed this also improved the dynamic range of CymR by 3.5-fold. Fl. activity was assessed using a plate reader.

Results:

While we obtain a lot of alkyl substituted aromatic carbons through the burning of fossil fuels, a terpene *p*-cymene (*p*-isopropyltoluene) is commonly found as a volatile oil in several plant species with several beneficial effects to humans[28]. In plants, *p*-cymene serves as a building block to make valuable metabolites such as pyruvate, isobutyrate and acetyl coenzyme. This chemical transformation requires conversion of *p*-cymene into *p*-cumate (*p*-isopropylbenzoate) as an intermediate **(Fig 7a)**. The 11Kb Cym operon in *Pseudomonas putida* F1 facilitates this transformation by oxidation of the methyl-substituted aromatic compound[29] via in a manner similar to other known aromatic compounds including xylene, methylnaphthalenes, toluene and *p*-cresol. The Cym operon constitutes 6 genes CymAa, CymAb CymB, CymC, CymD, CymE encoding *p*-cymene monooxygenase (hydroxylase, reductase), *p*-cumic alcohol dehydrogenase, membrane protein and acetyl coenzyme A synthetase, respectively[30]. The activity of this operon is controlled by the upstream transcriptional regulator CymR.

CymR, a TetR-based transregulator, for *p*-cumate detection:

CymR[1] is a putative regulatory protein that represses the expression of the Cym operon, as well as the downstream Cmt operon. It was found that the effector for CymR is *p*-cumate and not *p*-cymene leading to a feedback loop for the production of *p*-cumate that will be further catabolize *p*-cumate via the CymR induced Cmt operon[30]. CymR is a TetR-type regulator that has been utilized as an efficient inducible system in various

[1] In this work, we used an engineered version of WT CymR, $CymR^{AM}$ developed by the Voight lab[43] to have >100x dynamic range for effective inducible control. Throughout this chapter, I will be referring to $CymR^{AM}$ as parent or parent CymR.

organisms implying a wide range of utility starting from a model strain[31]. A structural characteristic of this family of regulators is their winged helix-turn-helix motif (wHTH) present in the N-terminus region. CymR exists as a dimer with the N-term region for each monomer interacting with DNA on opposite ends of the oligomer. They look like tiny wings on either ends giving them the winged HTH nomenclature.

We implemented a system where sfGFP is put under cumate responsive inducible control **(Fig 7b)**. So, in the presence of >10mM cumate, we see a 100-fold increase in activity from its inactive state. The system was originally oriented such that the CymR regulator gene was directly downstream of the sfGFP encoding gene (S6). We learned that this leads to an increase in noise as well as a decrease in the maximal activity. To mitigate this issue and make it easier for genome integration we swapped the position of the regulator to be directly upstream of sfGFP **(S6)**. This showed a 4-fold improvement in the dynamic range of the parent CymR **(Fig 7c)**. CymR is a strong engineering candidate for increased sensitivity towards *p*-cumate. We followed a directed evolution workflow but incorporated NGS data generation and machine learning guided mutation prediction for an accelerated engineering process. We obtained ~$7.5*10^6$ mutants of CymR using error-prone pcr mutagenesis

We observed activity of the parent in both *E. Coli* and *K. variicola* with a $K_{1/2}$ of ~10uM with potential for nanomolar sensitivity. We subjected the mutagenized library to 0, 100, 500 and 1000nM cumate for parallel screening with FACS with the 0nM sample being our background control. We saw 40% of our library active at 1uM with 15% of the library having increased activity compared to the parent. Within the individual concentrations, we see 8.7%, 9.3%, 10.2% and 13.6% of total cells at 0, 100, 500 and

1000nM, respectively, with activity higher than WT. We gated the top 5% of the 1000nM with each population falling in the 3-5% of GFP high cells collected amounting to ~400,000 to ~600,000 cells. Our one round of engineering comprised of 5 samples to compare mutations that are enriched within each concentration using PU-Learning[32].

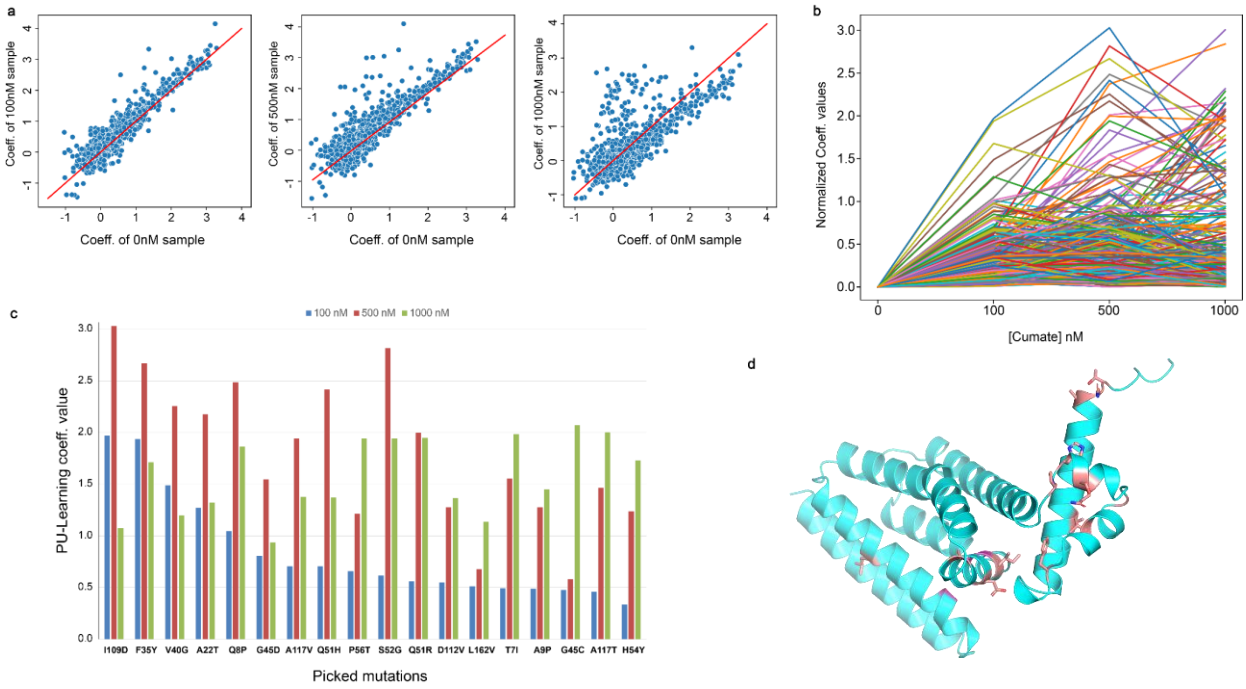


Figure 8: CymR DMS dataset and mutation selection. (a) Correlation scatterplot of the 100, 500 and 1000nM to the 0nM sample from left to right, respectively. We observe a spike in unique mutations with increased coeffs. in each of the induced datasets over the uninduced set. This trend is more pronounced as the concentration increases from 100 to 500 to 1000nM. **(b)** Comparing the normalized coeffs. (induced – uninduced) revealed that the top mutations have a peak coeff. value at 500nM rather than 1000nM. We observe a decrease in the coeff. value at 1000nM potentially due to a larger number of activating mutations in the post-sorted sequencing pool. **(c)** We selected the top 18 mutations present in the 85th percentile of each induced dataset. The mutations were selected with a focus on the 100nM sample set. **(d)** Using Alphafold 3, we generated a predicted structure of CymR with high confidence. The pink areas highlight the residues to be mutated. Most of the mutations are localized to the N-terminus region involved in DNA binding. The central 60-100 aa stretch sees no change in residue possibly due to the disruption of protein-protein interaction of the CymR homodimer

Mutational pattern of CymR DMS dataset:

PU-Learning compares the enrichment of mutations in the 0, 100, 500 and 1000nM samples to the entire library pre-selection giving us quantitative coefficients for each observed mutation. The dataset contained 1263 unique mutations averaging 6 mutations at each residue position. The ROC curves for the individual samples show a corrected area under curve value of about 0.7 indicating good output of the model. Given our screening methods, we intended to see an increase in active mutations in the increasing concentration samples. We observe exactly that; Comparing individual mutations between samples we see a higher number of unique mutations present in the 1000nM sample than the 500nM followed by 100nM samples that are active over the uninduced samples **(fig 8a)**. This was a useful check as parent CymR shows some activity at 1000nM indicating more mutations can be tolerated at this concentration level.

We hypothesized that activating mutations at 100nM should be present in increased abundance at higher concentrations and created a set of mutations that follow these parameters. However, to test this hypothesis we ranked mutations in each sample normalized to their uninduced coefficients **(fig 8b)**. Surprisingly, we noticed that the best mutations peaked at 500nM with a decreasing value at 1000nM. A possible explanation for this is the relative amount of a single mutation in the pool of the positive mutations. PU-Learning is an enrichment based model that infers information from the count of a mutation in the pre-sorted population to that of their post-sorted population. In 1000nM, we observe a larger number of unique mutations in the activating region relative to 500nM and 100nM. This leads to a decrease in the quantified coeff. value of a particular mutation at 1000nM as compared to 500nM.

We shifted our focus from creating a set of mutations with ascending coeff. values to selecting a set of the top based on their rank in each population. The top 18 mutations present in the 87th percentile of each set were selected as our test mutations from round one **(fig 8c)**. Most of the selected mutations were concentrated along the N-terminus which, in CymR, functions as the DNA binding region **(fig 8d)**. The central stretch of CymR which is involved in dimerization was largely untouched. We created a combinatorial library (~156,000 variants) of the selected mutations with a mean distribution of 8 mutations per gene.

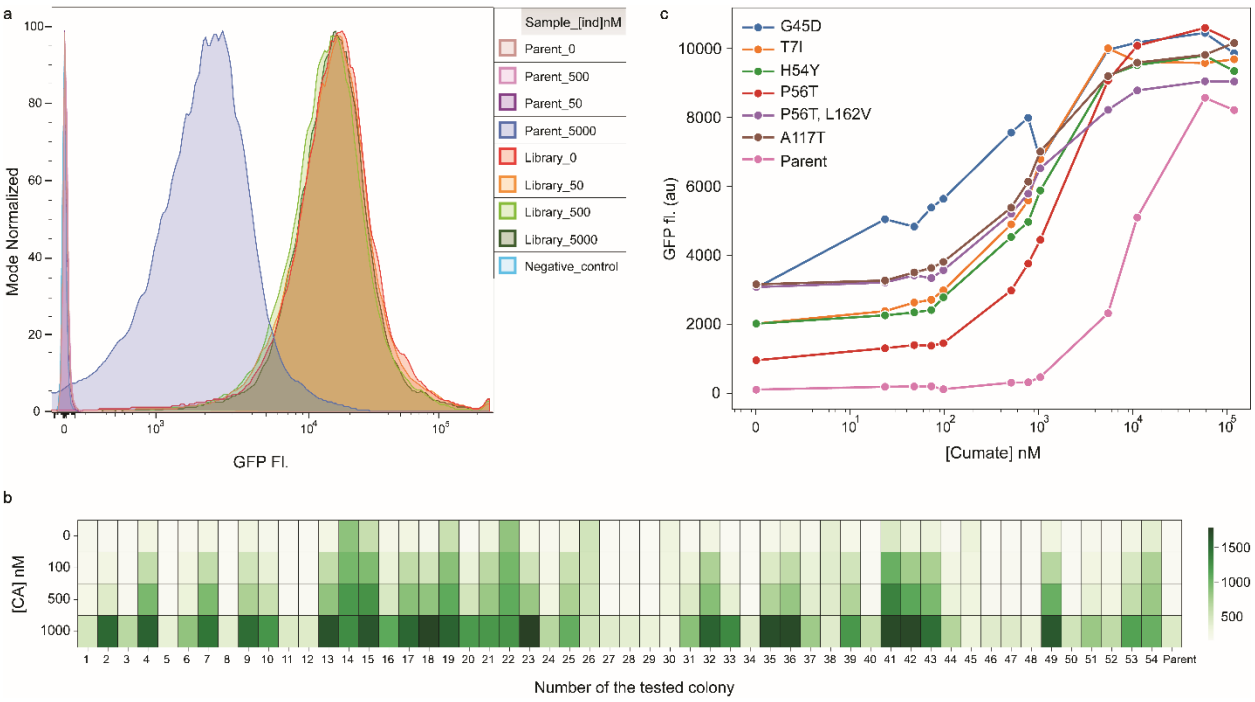


Figure 9: Combing through the combinatorial library. (a) The activity of parent CymR and the combinatorial library at 0, 50, 500 and 5000nM *p*-cumate. The combinatorial library sees no shift in the histogram in a concentration gradient while the parent shows no activity below 1000nM and nearly 1000x activity at 5000nM. This data was collected using BDFortessa Flow cytometer. **(b)** We conducted negative-positive screening to isolate variants with increased sensitivity on agar plates. The heatmap shows activity of 54 selected colonies and the parent in liquid media at 0, 100, 500 and 1000nM, top to bottom. We found several colonies with higher sensitivity than WT while decreasing the dynamic range. 32 colonies were sent for sanger sequencing. **(c)** We transformed the top 7 variants in *klebsiella variicola* and measured their activity using Flow cytometry. We observe a 10-fold increase in sensitivity in all our variants with varying degree of increase in background activity. Our best variant, $CymR^{P56T}$, showed activity in the 100-1000nM range with >2-fold decreased background compared to other variants.

Finding the needle in a combinatorial library haystack:

The combinatorial library, not surprisingly, was mostly dead. Given a big set of mutations introduced at the same time, it is possible to have broken the biosensor leading to extremely high uninduced activity. We observed the combinatorial library being 4x more active than induced WT but no difference among the different concentrations (fig 9a). We performed two rounds of negative selection by screening the library at 0nM cumate and selecting variants with low GFP fluorescence. At this point we were able to see a shift in population among the 100, 500 and 1000nM samples **(S7)**. To obtain the best active variants from this targeted combinatorial library, we set a selection threshold of the top 1% of active mutations in the 1000nM population set.

We subjected the sorted cells to a solid phase induction assay to pick the best colonies **(S8)**. Out of ~300 colonies screened on plates; we selected 50 colonies for subsequent liquid culture assay. We saw quite a few active mutations at with the best ones showing a 3-fold increased activity at 500nM. This was a promising result to learn specific mutations involved in the increased sensitivity towards cumate **(fig 9b)**. We hypothesized that there is a strong possibility for the screened colonies to be non-unique. We chose 32 variants for sequencing which included the highly active variants as well as the parent as a positive control. Our hypothesis was confirmed as we observed 12 unique mutation sets with most of the variants being single mutants. Testing our individually picked colonies showed measurable detection of cumate in the 100-1000nM range but with ~10x increase in background activity. G45D was used as a test case for learning about the potential increase in background noise while P56T, T7I and H54Y gave the most promising output (K_d = ~800-1200nM) compared to the WT (K_d = ~8000nM). Out of

the 12 unique variants, we selected 7 for testing within *Klebsiella variicola*. While the dynamic range of the parent and variants was lowered in *kv*, it did reflect the activity observed in the initial *E. Coli* host **(fig 9c)**. This confirmed that 2-3x increased activity at 500nM of the variants was in fact real giving us positive variants from a single round of engineering.

Characterizing bacterial localization around plant roots:

We do not fully understand the localization of soil bacteria in plant roots. They are thought to be presence in root hair, root tips and potentially even the shoot. Secondly, it is important to know where root exudates are being released and in what quantities to understand plant stress response. We aimed to test the high performing variants T7I, H54Y and P56T in soil bacteria localization studies which would require integration of the GUS (β-glucuronidase) reporter system instead of the GFP. GUS is a luminesce based assay that uses β-glucuronidase to convert the colorless X-gluc into a bright blue diX-indigo.

This is a popular reporter system used in plant rhizosphere assay but can be tricky due to high background activity. We used the parent CymR controlled GUS expression and a constitutively expressing GUS strain to compare whether bacterial localization was even visible. We see a darker blue in the constitutive sample, as expected, along with lower GUS activity in the parent sample **(fig 10)**. This is one of the first examples showing direct spatial arrangement of soil bacteria (*kv*) in response to a plant released metabolite (cumate).

NOTE: This section contains some of the most recent data collected before my dissertation defense. The result section is incomplete and will be fully presented in the publication of this work.

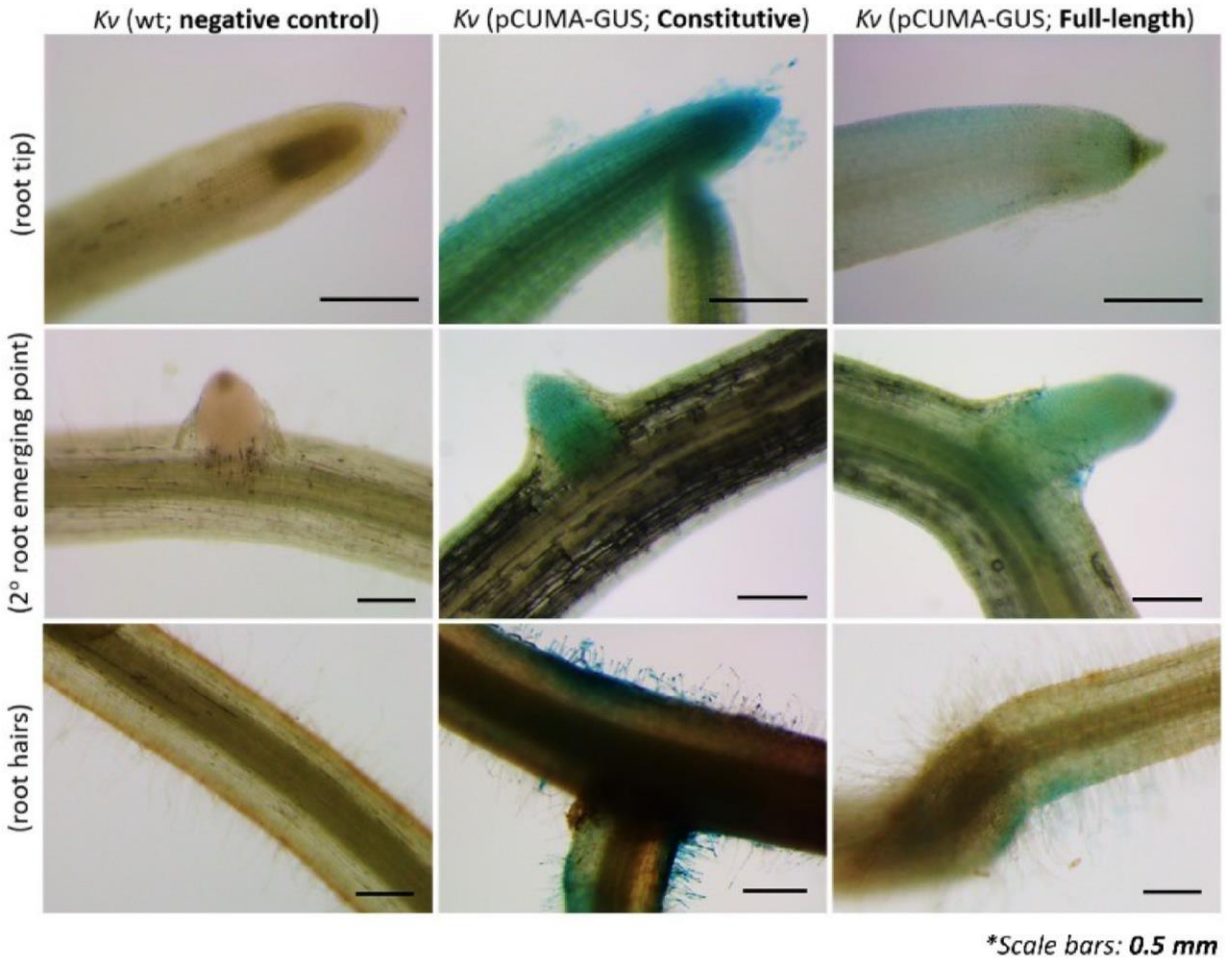


Figure 10: *In Planta* assessment of CymR. This is one of the first examples of characterizing the localization of *klebsiella variicola* in sorghum plant roots. The left column is the negative sample is WT *kv* with no GUS system showing no color in any of root cross-sections. The middle column is GUS under constitutive expression acting as positive control for luminescence. We see strong blue color in the root tips (top) and root hair (bottom) with some activity in the root emergence points. This indicates soil bacteria can penetrate root tissues potentially aid in root emergence. The right column is *kv* with GUS expression under *p*-cumate control. The regulator is parent CymR, and we observe some blue color in the root cross-sections.

Discussion:

Efficient nitrogen fixation is a requirement to meet the rising food demand across the globe. Inorganic fertilizers are an effective tool in our arsenal for providing fixed nitrogen to crops but not without its limitations. Inorganic fertilizers are massive pollutants to the environment known to contain heavy metal contamination thus creating a bigger problem in the long term[33]. Current technology relying on increase in yields of cereal crops is reaching a plateau and is in desperate need for innovation. Moreover, the consumer mentality to modifying plants and crops is very hostile due to past outbreaks and lack of understanding[34]. Until this mentality shifts and innovation in plant biology is within our reach, there is an unmet short term need for effective fixed nitrogen. We propose the engineering of bacteria in the plant microbiome that natively fix oxygen to increase their production in a non-toxic manner. Root exudates are metabolites secreted by plants in the soil via their roots, often as a means of communication or defense [35,36]. Cumate is a root exudate with anti-fungal activities[37,38] produced by several cereal crops such as sorghum and corn. We intended to use cumate as an inducer to regulate nitrogen fixation in *klebsiella variicola* for increased NH_4^+ delivery to sorghum crop without compromising the fitness of the bacteria.

Engineering BNF in *kv* requires detecting cumate using the σ70 transcription factor CymR derived from *Pseudomonas putida*. CymR natively detects cumate at >1000nM concentration. In non-cumin plants the amount of cumate released is unknown but potentially thought to be in the 100-1000nM range which lends well to the directed evolution of CymR for increased sensitivity. Traditional DE methods would need engineering at 1000nM with multiple rounds of engineering for a stepwise increase in

sensitivity. However, we used a machine learning guided approach to accelerate this process. This enables us to screen via FACS at 0nM, 100nM, 500nM and 1000nM and generate a mutational dataset containing ~1300 unique mutations for a 200 amino acid protein. The goal of simultaneous screening was to find mutations that are enriched in each set giving us increased confidence in the activity of variants at 100 and 500nM cumate.

The selected 18 mutations were enriched in each set with 500nM being the peak. This was interesting since we anticipated an increase in PU-learning coeff with increasing concentration, but we saw increased enrichment at 500nM with a decrease in value at 1000nM. I attribute this trend to the increased pool of activating mutations at 1000nM as compared to a smaller pool in the 500nM sample. This lowered the enrichment value of our highly active mutations in the 1000nM dataset. This wasn't really a big problem since our goal was to focus on the lower concentration, so selecting the top 18 mutations from the 87th percentile made sense. Although, we see the top mutations clustered in the N-terminus region of CymR rather than the in the central stretch of CymR (60-105 aa) which we think is involved in protein-protein interaction to form a homodimer. A possible explanation here is that making mutations in homodimer interacting region can be tricky since its effect gets doubled as the change in residue impacts each monomer creating a "double mutant" for the homodimer[39]. PU-Learning doesn't directly infer epistatic interactions among the observed mutations, so we created a combinatorial library to account for the epistasis and find viable combinations for parent CymR variants.

Our combinatorial library was a set of mostly broken sequences given several mutations present in each variant. While this wasn't a great outcome it was easily

salvageable. Screening this library for low fl. output when induced followed by a positive screen induced at 100, 500 and 1000nM gave a set of highly sensitive variants with P56T being nearly 8x more sensitive with the least increase in background. The alphafold3 predicted structure of CymR shows P56T mutation in the loop region right beside the helix-turn-helix motif potentially stabilizing the conformer upon ligand binding. We see that G45D mutation increases background activity ~5-fold over the P56T variant and is located within the HTH motif. We hypothesize that the presence of the acidic aspartate residue prevents interaction with the phosphates of the DNA given both molecules are negatively charged. Another interesting observation we had was in the P56T, L162V variant. It retains the sensitivity of the P56T mutation but has nearly 2-3x increased background. Further work can be done to learn the epistatic interaction between the two mutations by creating a site saturated library at the two residues, but this is outside the scope of this work. I was mildly disappointed that most of our variants were single mutants instead of a combination of multiple mutations, a capability of our ML-guided mutational approach. However, this method did yield variants with desirable activity in the 100-1000nM range make our engineering efforts worthwhile.

Future directions:

We see desirable activity of our best variant, CymR^{P56T}, in *klebsiella variicola* as well. The next steps to this project involve testing the capability of CymR to effectively regulate nitrogen fixation under cumate control. This will involve genome integration of the CymR variants upstream of nitrogenase enzymes in *k. variicola.* We plan on using the acetylene reduction assay (ARA), a popular nitrogenase activity assay, for *in vitro* assessment of our variants. Finally, we aim to create this technology for increased BNF and delivery of ammonia to plants. This will require *in planta* testing of the engineered *k. variicola*. Apart from the ability of the soil microbe to deliver fixed nitrogen, we will also gain insights into the spatial arrangement of soil bacteria from the intake of nitrogen in plants.

In this work, we developed nanomolar sensitive *p*-cumate detecting biosensors using highthroughput screening and ML-guided mutation prediction for inducible control of nitrogen fixation. This workflow can be adopted to detect other root exudates or plant metabolites for academic and industrial applications. Cinnamic acid[40], naringenin[41] and flavonoids[42] are strong initial targets for engineering biosensors. Root exudate induced nitrogen fixation can help supplement if not completely replace inorganic nitrogen based fertilizers.

Materials and methods:

Strains, Media, and Materials:

The CymR plasmid, CymR_old, was made as used in the literature[43]. One major modification was made to this plasmid to reposition the regulator upstream of the reporter, CymR_new. The Luria Broth (Miller) media (DSL24400-2000) and Cuminic acid (268402-5G) were sourced from Sigma Aldrich. The Kanamycin (monosulfate) obtained from Goldbio (K-120-5). The LB (miller) Agar (30620042-4) used for making plates was obtained from bioworld. The consumables used in the project were sourced from various vendors. The 15mL round bottom culture tubes (352059) were corning Falcon tubes. We used VWR 15mL conical bottom tubes (525-0636) and 50mL conical bottom tubes (525-0610). The black with clear bottom 96-well microtiter plates were obtained from Thermo scientific (165305). The restriction enzymes and mastermixes for molecular biology were sourced from New England Biolabs.

Error-prone pcr:

We amplified the entire CymR gene (202 aa) using error-prone PCR to introduce random mutations. We reaction mixture comprised of (1M) Betaine, Standard Taq buffer (1x), $MgCl_2$ (5.5mM), DNA template (~200ng), 1:1:2.5:2.5 ratio of A:G:T:C (0.2mM A,G and 0.5mM T,C), $MnCl_2$ (0.05mM), forward and reverse primers (0.4mM each) and Taq DNA polymerase (5 units/100uL reaction volume). The pcr reaction was made in 200uL total volume and divided into 8 tubes of 50uL run in 8-pcr strip tubes (Thermo scientific, AB2000). The reaction was cycled with initial melting at 95°C for 30sec, followed by 16 cycles of 95°C, 20s melt, 56°C, 60 sec extension and 68°C,1 min elongation. The final

extension was done at 68°C for 5 min. The pcr was run for 16 cycles to minimize amplification bias in a Bio-Rad thermocycler (T100). We digested the amplified library with DpnI (NEB R0176S) to remove DNA template followed by purification via Zymo DNA clean and concentration kit (cat# D4013). The insert amplification was verified using a 1% agarose gel. We wanted to keep the mutational rate in the library low (~1-2aa/gene) to allow generating a healthy functional fraction of CymR variants while maximizing the size of the obtained library.

Gibson ligation and transformation:

A 15 uL Gibson mastermix was made by combining 1M Betaine, 5x isothermal reaction buffer, 40U/uL Taq DNA (NEB, M0208L), 1U/uL T5 exonuclease (NEB, M0663L), 2U/uL Phusion High-Fidelity DNA Polymerase (NEB, M0530L). The insert and backbone were added in a >2:1 molar ratio respectively and water to reach the total volume of 20uL in a single pcr tube. The reaction was run isothermally at 50°C for an hour. The samples were cleaned using the Zymo DNA clean and concentration kit (Zymo, D4003).

We transformed the ligated plasmid library into commercial chemically competent E. cloni cells via electroporation in 1mm cuvettes (Fisher, FB101) at 1.80Kv using the Bio-rad MicroPulser (1652100). The transformed cells were recovered for 1 hour @37°C and diluted into 50mL, kanamycin containing LB media to be grown at 30°C. Simultaneously, plates with 500x, 5,000x and 50,000x dilution were inoculated from the 50mL culture and grown overnight @37°C. After ~12-14 hr growth at 30°C, we made 5, 1mL aliquots of the culture in 15% Glycerol to store as redundant stocks for future use.

We obtained a library size of ~7.5*10^6 transformants, analyzed by counting colonies. Individual random 10 colonies were picked for colony pcr followed by Sanger sequencing (Functional Biosciences).

Plate reader assay:

Seed cultures of parent CymR, Negative control and any variants were inoculated from colony plates or glycerol stocks in 5mL LB media containing Kanamycin for 20-22h at 30°C, 250rpm. The overnight saturated cultures were diluted 1:100 fresh 250uL LB media with antibiotics. The expression cultures were grown in corning 96-well plates (3340) for 16 h at 37°C, 250rpm with varying concentrations (0-100uM) of cuminic acid added after inoculation. The fluorescence of the samples was measured using a Biotek Agilent plate reader with excitation at 488nM, emission at 510nM, analyzed using excel and python.

Flow Cytometry and FACS:

WT CymR, Negative control and any variants were grown overnight in 5mL LB cultures with kanamycin for 250rpm, 20-22h @30°C. The overnight saturated cultures were diluted 1:100 (50uL in 5000uL) in fresh 5mL LB media with antibiotics. The expression cultures were grown for 16 h @37°C, 250rpm with varying concentrations (0-100uM) of cuminic acid added after inoculation. The following day, expression cultures were washed 2 times with 1mL phosphate-buffered saline, PBS (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4 and 2 mM KH_2PO_4) with centrifugation for 10 min @3000xg. We resuspended the cells in 1mL PBS and then diluted the resuspension (1:50) in 1mL fresh PBS in 5mL polypropylene tubes (Fisher cat# 352058). For Flow cytometry, we ran the samples in a BD LSRFortessa X-20. The samples were excited @488nM and detected @510nM.

For FACS, the assay and sample prep follows the above described method. 1mL LB collection tubes with kanamycin were made to collect the sorted cells for each sample. The samples were run in a BD FACSAria III at the same excitation and emission. 20-21 million cells were observed by the machine with the collection threshold gate at 5% for the 1000nM sample. The same threshold gate was used for each of the samples. The collected cells were grown overnight to an OD ~1-2 to avoid significant amount of cell death and stored in 15% glycerol as stocks for later use. The FACS data was analyzed using FlowJo.

Next-generation sequencing:

The pre-sorted and post-sorted cultures from the FACS screening were miniprepped using the QIAprep Spin Miniprep Kit to obtain the library of plasmids in each sample. The whole plasmid for each sample was submitted to UW-Madison Next generation Sequencing core to prepare NGS libraries using the Celero DNA-Seq Library Prep for the Illumina Nova-seq. We collected $2.2*10^8$ reads for pre-sorted samples and $5*10^7$ for the post-sorted populations.

Solid Phase induction:

LB agar plates were made containing kanamycin and varying amounts of *p*-cuminic acid (0, 100, 500, 1000 and 5000nM). Individual colonies from the 100nM active variant plate were chosen and spotted in the same geographical location for each agar plate. 5 sets of 5 plates were created to spot 250 colonies. The plates were incubated overnight at 37°C, 250 rpm. The following day, the plates were analyzed under blue light (~480nM).

In planta GUS staining protocol for corn/sorghum seedling roots:

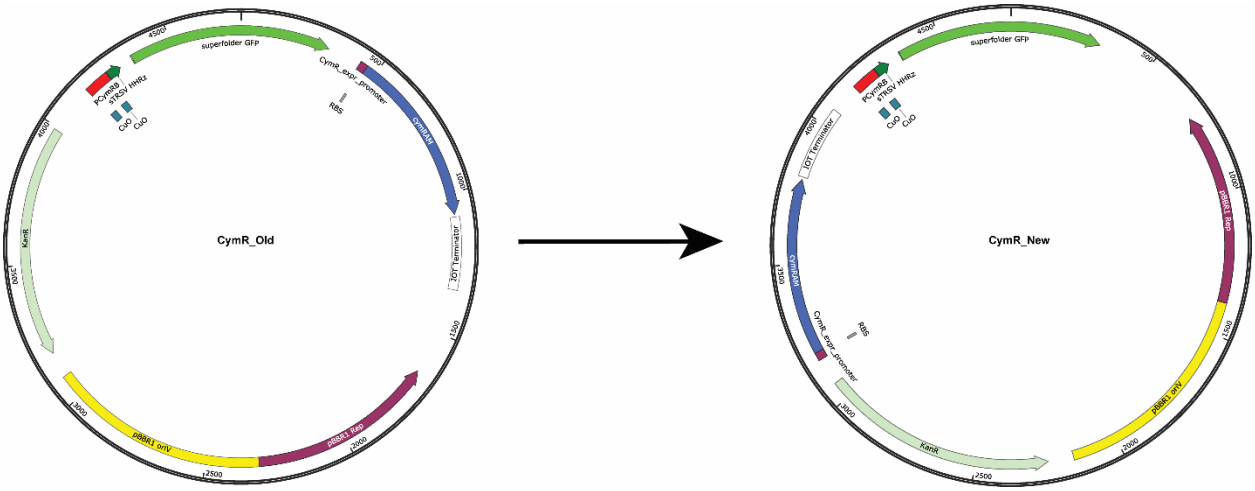
The 7 days old corn/sorghum seedlings (3-4 seedlings in each growth pouches under aseptic condition) were inoculated with 10 ml of the GUS reporter strains (OD_{600} = 0.1) and incubated for another 7 days in the plant growth chamber. On day 7 after bacterial inoculation, the whole seedlings were harvested out of the pouches gently and washed three times with sterile MQ water. Then for each combination, depending on the sample volume, the whole seedlings were transferred into the GUS assay buffer in separate 50 ml falcons/glass beakers in such a way that the shoot part remains out of the tube/beaker and only the roots remain submerged into the GUS buffer. All the seedlings inoculated with each of the bacterial strain(s) were subjected to vacuum infiltration for 2hrs to facilitate X-Gluc penetration into the infiltrated seedling roots at room temperature. The infiltration cups containing the infiltrated seedlings were incubated at 37°C overnight (16-20 hr). The following day, the seedlings were removed out of the GUS buffer, washed gently using fresh MQ water (at least three times), and transferred into fresh beakers/falcon tubes. Subsequently, chlorophyll was completely removed from the stained seedlings by incubating in 70-95% ethanol for 96-120 hr at 37°C (in dark). Finally, the roots were observed under white light by using a bright field stereomicroscope for blue colored stained regions.

*Note: For the GUS staining purpose, 1 mM of X-Gluc (5-bromo-4-chloro-3-indolyl-β-D-glucuronide) in GUS assay buffer [50 mM sodium di-hydrogen phosphate (pH 7.0), 10 mM EDTA, 0.1% sodium lauryl sarcosine, 0.1% Triton X-100 and 10 mM β-mercaptoethanol] was used

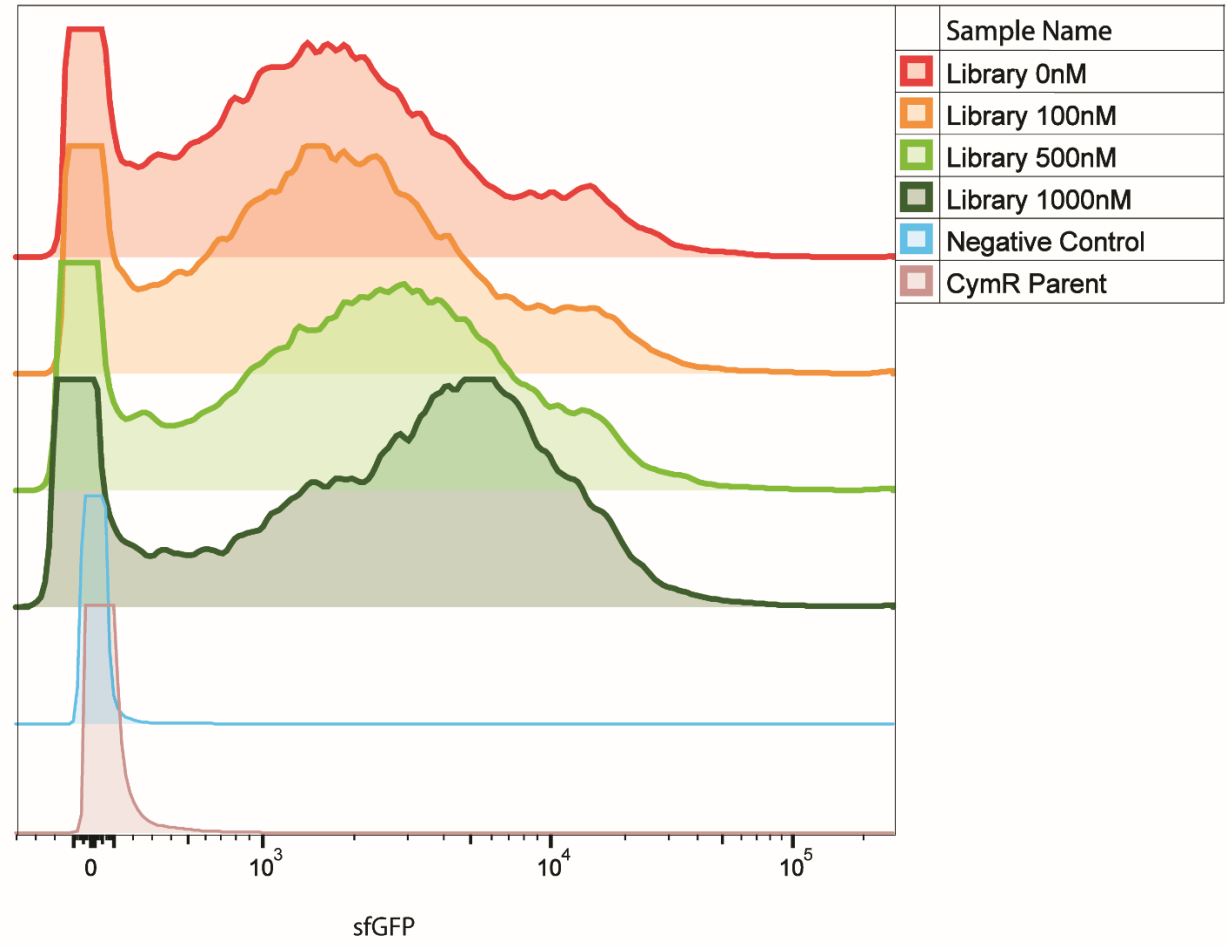
Data Analysis:

The data analysis and machine learning for the project was done using python. The code is in a Github repo.

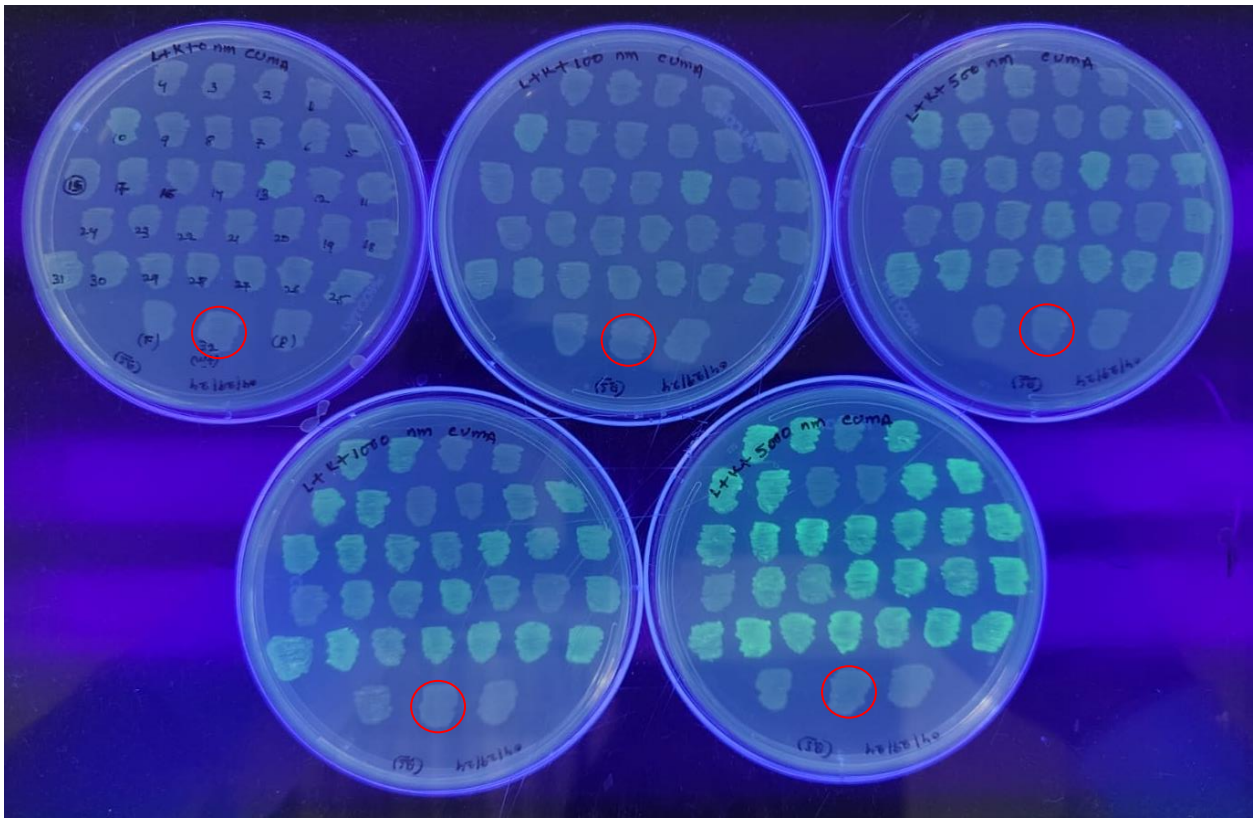
Supplementary figures:



S6: Changes in our plasmid system. Our plasmid system originally had the CymR gene downstream of sfGFP (left). For easier genome integration, we realized it would be better to have the CymR and subsequent variants upstream of the reporter gene (right). This change also improved the dynamic range of our system, given us a second benefit.



S7: Salvaging our broken combinatorial library. Our combinatorial library showed elevated fluorescence activity at all concentrations of *p*-cumate. To find variants with increased sensitivity without completely losing the dynamic range, we ran two negative screens (collecting cells that are inactive under 0nM inducer condition) followed by a positive screen (shown above). We observe a shift in Fl. activity with increasing concentration. We sorted the top 1% of cells in the above populations to find variants with desirable activity towards *p*-cumate.



S8: Solid Phase induction assay of active variants. We plated the sorted cells from the 100nM sample of our combinatorial library salvage screen on agar plates containing 0, 100, 500, 1000, 5000nM of p-cumate (top left to bottom right, respectively) alongside the parent (circled in red). We illuminated the plates with blue light and found several colonies with activity in the 100-1000nM.

References:

(1) Sohm, J. A.; Webb, E. A.; Capone, D. G. Emerging Patterns of Marine Nitrogen Fixation. *Nat Rev Microbiol* 2011, *9* (7), 499–508. https://doi.org/10.1038/nrmicro2594.

(2) Hoffman, B. M.; Lukoyanov, D.; Yang, Z.-Y.; Dean, D. R.; Seefeldt, L. C. Mechanism of Nitrogen Fixation by Nitrogenase: The Next Stage. *Chem. Rev.* 2014, *114* (8), 4041–4062. https://doi.org/10.1021/cr400641x.

(3) Seefeldt, L. C.; Hoffman, B. M.; Dean, D. R. Mechanism of Mo-Dependent Nitrogenase. *Annual Review of Biochemistry* 2009, *78* (Volume 78, 2009), 701–722. https://doi.org/10.1146/annurev.biochem.78.070907.103812.

(4) Ludemann, C. I.; Gruere, A.; Heffer, P.; Dobermann, A. Global Data on Fertilizer Use by Crop and by Country. *Sci Data* 2022, *9* (1), 501. https://doi.org/10.1038/s41597-022-01592-z.

(5) Erisman, J. W.; Galloway, J.; Seitzinger, S.; Bleeker, A.; Butterbach-Bahl, K. Reactive Nitrogen in the Environment and Its Effect on Climate Change. *Current Opinion in Environmental Sustainability* 2011, *3* (5), 281–290. https://doi.org/10.1016/j.cosust.2011.08.012.

(6) Walling, E.; Vaneeckhaute, C. Greenhouse Gas Emissions from Inorganic and Organic Fertilizer Production and Use: A Review of Emission Factors and Their Variability. *Journal of Environmental Management* 2020, *276*, 111211. https://doi.org/10.1016/j.jenvman.2020.111211.

(7) Diaz, R. J.; Rosenberg, R. Spreading Dead Zones and Consequences for Marine Ecosystems. *Science* 2008, *321* (5891), 926–929. https://doi.org/10.1126/science.1156401.

(8) Wang, L.; Huang, D. Nitrogen and Phosphorus Losses by Surface Runoff and Soil Microbial Communities in a Paddy Field with Different Irrigation and Fertilization Managements. *PLoS One* 2021, *16* (7), e0254227. https://doi.org/10.1371/journal.pone.0254227.

(9) Peng, S.; Tang, Q.; Zou, Y. Current Status and Challenges of Rice Production in China. *Plant Production Science* 2009, *12* (1), 3–8. https://doi.org/10.1626/pps.12.3.

(10) *122 million more people pushed into hunger since 2019 due to multiple crises, reveals UN report*. https://www.who.int/news/item/12-07-2023-122-million-more-people-pushed-into-hunger-since-2019-due-to-multiple-crises--reveals-un-report (accessed 2024-07-27).

(11) *Hunger*. Food and Agriculture Organization of the United Nations. http://www.fao.org/hunger/en/ (accessed 2024-07-27).

(12) Haskett, T. L.; Paramasivan, P.; Mendes, M. D.; Green, P.; Geddes, B. A.; Knights, H. E.; Jorrin, B.; Ryu, M.-H.; Brett, P.; Voigt, C. A.; Oldroyd, G. E. D.; Poole, P. S. Engineered Plant Control of Associative Nitrogen Fixation. *Proceedings of the National Academy of Sciences* 2022, *119* (16), e2117465119. https://doi.org/10.1073/pnas.2117465119.

(13) Bishop, P. E.; Joerger, R. D. Genetics and Molecular Biology of Alternative Nitrogen Fixation Systems. *Annual Review of Plant Biology* 1990, *41* (Volume 41, 1990), 109–125. https://doi.org/10.1146/annurev.pp.41.060190.000545.

(14) Geddes, B. A.; Paramasivan, P.; Joffrin, A.; Thompson, A. L.; Christensen, K.; Jorrin, B.; Brett, P.; Conway, S. J.; Oldroyd, G. E. D.; Poole, P. S. Engineering Transkingdom Signalling in Plants to Control Gene Expression in Rhizosphere Bacteria. *Nat Commun* 2019, *10* (1), 3430. https://doi.org/10.1038/s41467-019-10882-x.

(15) *Functional analysis of chimeric lysin motif domain receptors mediating Nod factor-induced defense signaling in Arabidopsis thaliana and chitin-induced nodulation signaling in Lotus japonicus - Wang - 2014 - The Plant Journal - Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1111/tpj.12450 (accessed 2024-07-27).

(16) Soyano, T.; Kouchi, H.; Hirota, A.; Hayashi, M. NODULE INCEPTION Directly Targets NF-Y Subunit Genes to Regulate Essential Processes of Root Nodule Development in Lotus Japonicus. *PLOS Genetics* 2013, *9* (3), e1003352. https://doi.org/10.1371/journal.pgen.1003352.

(17) Soto, G.; Fox, A. R.; Ayub, N. D. Exploring the Intrinsic Limits of Nitrogenase Transfer from Bacteria to Eukaryotes. *J Mol Evol* 2013, *77* (1), 3–7. https://doi.org/10.1007/s00239-013-9578-8.

(18) Burén, S.; Young, E. M.; Sweeny, E. A.; Lopez-Torrejón, G.; Veldhuizen, M.; Voigt, C. A.; Rubio, L. M. Formation of Nitrogenase NifDK Tetramers in the Mitochondria of Saccharomyces Cerevisiae. *ACS Synth. Biol.* 2017, *6* (6), 1043–1055. https://doi.org/10.1021/acssynbio.6b00371.

(19) Oldroyd, G. E.; Dixon, R. Biotechnological Solutions to the Nitrogen Problem. *Current Opinion in Biotechnology* 2014, *26*, 19–24. https://doi.org/10.1016/j.copbio.2013.08.006.

(20) Funk, C. *About half of U.S. adults are wary of health effects of genetically modified foods, but many also see advantages*. Pew Research Center. https://www.pewresearch.org/short-reads/2020/03/18/about-half-of-u-s-adults-are-wary-of-health-effects-of-genetically-modified-foods-but-many-also-see-advantages/ (accessed 2024-07-08).

(21) Montañez, A.; Blanco, A. R.; Barlocco, C.; Beracochea, M.; Sicardi, M. Characterization of Cultivable Putative Endophytic Plant Growth Promoting Bacteria Associated with Maize Cultivars (*Zea Mays* L.) and Their Inoculation Effects *in Vitro*. *Applied Soil Ecology* 2012, *58*, 21–28. https://doi.org/10.1016/j.apsoil.2012.02.009.

(22) *Nitrogen control in bacteria*. https://doi.org/10.1128/mr.59.4.604-622.1995.

(23) Zeng, Y.; Guo, L.; Gao, Y.; Cui, L.; Wang, M.; Huang, L.; Jiang, M.; Liu, Y.; Zhu, Y.; Xiang, H.; Li, D.-F.; Zheng, Y. Formation of NifA-PII Complex Represses Ammonium-Sensitive Nitrogen Fixation in Diazotrophic Proteobacteria Lacking NifL. *Cell Reports* 2024, *43* (7). https://doi.org/10.1016/j.celrep.2024.114476.

(24) Brewin, B.; Woodley, P.; Drummond, M. The Basis of Ammonium Release in nifLMutants of Azotobacter Vinelandii. *Journal of Bacteriology* 1999, *181* (23), 7356–7362. https://doi.org/10.1128/jb.181.23.7356-7362.1999.

(25) Bais, H. P.; Weir, T. L.; Perry, L. G.; Gilroy, S.; Vivanco, J. M. THE ROLE OF ROOT EXUDATES IN RHIZOSPHERE INTERACTIONS WITH PLANTS AND OTHER

ORGANISMS. *Annual Review of Plant Biology* 2006, *57* (Volume 57, 2006), 233–266. https://doi.org/10.1146/annurev.arplant.57.032905.105159.

(26) Jones, D. L.; Nguyen, C.; Finlay, R. D. Carbon Flow in the Rhizosphere: Carbon Trading at the Soil–Root Interface. *Plant Soil* 2009, *321* (1), 5–33. https://doi.org/10.1007/s11104-009-9925-0.

(27) Kawasaki, A.; Okada, S.; Zhang, C.; Delhaize, E.; Mathesius, U.; Richardson, A. E.; Watt, M.; Gilliham, M.; Ryan, P. R. A Sterile Hydroponic System for Characterising Root Exudates from Specific Root Types and Whole-Root Systems of Large Crop Plants. *Plant Methods* 2018, *14* (1), 114. https://doi.org/10.1186/s13007-018-0380-x.

(28) Balahbib, A.; El Omari, N.; Hachlafi, N. E.; Lakhdar, F.; El Menyiy, N.; Salhi, N.; Mrabti, H. N.; Bakrim, S.; Zengin, G.; Bouyahya, A. Health Beneficial and Pharmacological Properties of *P*-Cymene. *Food Chem Toxicol* 2021, *153*, 112259. https://doi.org/10.1016/j.fct.2021.112259.

(29) Davis, J. B.; Raymond, R. L. Oxidation of Alkyl-Substituted Cyclic Hydrocarbons by a Nocardia during Growth on n-Alkanes. *Applied Microbiology* 1961, *9* (5), 383–388. https://doi.org/10.1128/am.9.5.383-388.1961.

(30) Eaton, R. W. *P*-Cymene Catabolic Pathway in Pseudomonas Putida F1: Cloning and Characterization of DNA Encoding Conversion of *p*-Cymene to *p*-Cumate. *Journal of Bacteriology* 1997, *179* (10), 3171–3180. https://doi.org/10.1128/jb.179.10.3171-3180.1997.

(31) Mullick, A.; Xu, Y.; Warren, R.; Koutroumanis, M.; Guilbault, C.; Broussau, S.; Malenfant, F.; Bourget, L.; Lamoureux, L.; Lo, R.; Caron, A. W.; Pilotte, A.; Massie, B. The

Cumate Gene-Switch: A System for Regulated Expression in Mammalian Cells. *BMC Biotechnology* 2006, *6* (1), 43. https://doi.org/10.1186/1472-6750-6-43.

(32) Song, H.; Bremer, B. J.; Hinds, E. C.; Raskutti, G.; Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Systems* 2021, *12* (1), 92-101.e8. https://doi.org/10.1016/j.cels.2020.10.007.

(33) Mortvedt, J. J. Heavy Metal Contaminants in Inorganic and Organic Fertilizers. *Fertilizer Research* 1995, *43* (1), 55–61. https://doi.org/10.1007/BF00747683.

(34) Marris, C. Public Views on GMOs: Deconstructing the Myths. *EMBO Rep* 2001, *2* (7), 545–548. https://doi.org/10.1093/embo-reports/kve142.

(35) Baetz, U.; Martinoia, E. Root Exudates: The Hidden Part of Plant Defense. *Trends in Plant Science* 2014, *19* (2), 90–98. https://doi.org/10.1016/j.tplants.2013.11.006.

(36) Vranova, V.; Rejsek, K.; Skene, K. R.; Janous, D.; Formanek, P. Methods of Collection of Plant Root Exudates in Relation to Plant Metabolism and Purpose: A Review. *Journal of Plant Nutrition and Soil Science* 2013, *176* (2), 175–199. https://doi.org/10.1002/jpln.201000360.

(37) Sun, Y.; Wang, Y.; Han, L. R.; Zhang, X.; Feng, J. T. Antifungal Activity and Action Mode of Cuminic Acid from the Seeds of Cuminum Cyminum L. against Fusarium Oxysporum f. Sp. Niveum (FON) Causing Fusarium Wilt on Watermelon. *Molecules* 2017, *22* (12), 2053. https://doi.org/10.3390/molecules22122053.

(38) Wang, Y.; Zhang, J.; Sun, Y.; Feng, J.; Zhang, X. Evaluating the Potential Value of Natural Product Cuminic Acid against Plant Pathogenic Fungi in Cucumber. *Molecules* 2017, *22* (11), 1914. https://doi.org/10.3390/molecules22111914.

(39) Ispolatov, I.; Yuryev, A.; Mazo, I.; Maslov, S. Binding Properties and Evolution of Homodimers in Protein–Protein Interaction Networks. *Nucleic Acids Research* 2005, *33* (11), 3629–3635. https://doi.org/10.1093/nar/gki678.

(40) Mehmood, A.; Hussain, A.; Irshad, M.; Hamayun, M.; Iqbal, A.; Rahman, H.; Tawab, A.; Ahmad, A.; Ayaz, S. Cinnamic Acid as an Inhibitor of Growth, Flavonoids Exudation and Endophytic Fungus Colonization in Maize Root. *Plant Physiology and Biochemistry* 2019, *135*, 61–68. https://doi.org/10.1016/j.plaphy.2018.11.029.

(41) Szoboszlay, M.; White-Monsant, A.; Moe, L. A. The Effect of Root Exudate 7,4′-Dihydroxyflavone and Naringenin on Soil Bacterial Community Structure. *PLoS One* 2016, *11* (1), e0146555. https://doi.org/10.1371/journal.pone.0146555.

(42) *Multifaceted roles of flavonoids mediating plant-microbe interactions* | *Microbiome* | *Full Text*. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-022-01420-x (accessed 2024-07-28).

(43) Meyer, A. J.; Segall-Shapiro, T. H.; Glassey, E.; Zhang, J.; Voigt, C. A. Escherichia Coli “Marionette” Strains with 12 Highly Optimized Small-Molecule Sensors. *Nat Chem Biol* 2019, *15* (2), 196–204. https://doi.org/10.1038/s41589-018-0168-3.