Large Scale Engineering of Chimeric Histidine Kinases

by

Andrew Sterling Holston

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Chemistry

Dissertation Committee:

Dr. Scott Hansen, Chair

Dr. Calin Plesa, Advisor

Dr. Michael Harms, Core Member

Dr. James Prell, Core Member

Dr. Karen Guillemin, Institutional Representative

University of Oregon

Spring, 2025

DISSERTATION ABSTRACT

Andrew Sterling Holston

Doctor of Philosophy in Chemistry

Title: Large Scale Engineering of Chimeric Histidine Kinases

Sensor histidine kinases (SHKs) represent one of the most abundant and versatile protein families in nature, mediating cellular responses to an extensive range of environmental stimuli. Despite their ubiquity, the majority of SHKs remain functionally uncharacterized due to the limitations of traditional, low-throughput methods. This dissertation addresses the challenge of SHK deorphanization by developing a high-throughput platform for rationally engineering, synthesizing, and profiling the basal activation of chimeric SHKs.

Central to this approach is Degenerate DropSynth, a multiplex gene synthesis technique adapted to assemble multiple variants per gene per droplet in a single emulsion reaction. Chimeric fusion-phase SHK variants were designed via residue insertions or deletions on either side of the fusion junction below the HAMP domain, with variants sampled at one to eight degeneracy levels per parent gene, leading to 21,724 total gene constructs, and this method synthesized these four libraries with a coverage of 75.8% at the amino acid level. Each fusion-phase variant introduces specific angular changes between helical domains to explore effects of register alignment on SHK activity. These libraries were cloned into barcoded plasmids and transformed into *E. coli* BW29655 (Δ*ompR* Δ*envZ*) carrying the response regulator plasmid pSR40.29. This plasmid couples SHK signaling to expression of the superfolderGFP fluorescent reporter, enabling quantification of signaling output.

Following growth in supplemented minimal media, cells were sorted by fluorescence-activated cell sorting into six brightness bins. Barcodes from sorted cells were sequenced via NovaSeq, and per-variant brightness levels were inferred by calculating bin-weighted fluorescence, reported in MEFL units. Across the successfully profiled proteins, fusion phase exerted a dominant effect on basal brightness, with some phase shifts resulting in constitutively active “locked-on” phenotypes and others producing inactive or functional signaling behavior. Analysis revealed that the impact of fusion phase varied across sensor classes, highlighting the structural sensitivity of SHK domain interfaces. A random forest model trained on 13,170

variants was able to predict a fifth of the variance for MEFL brightness based on sequence and phase, offering a framework for computational pre-screening.

This work establishes a scalable, structure-aware design-build-test-learn cycle for SHK engineering, enabling functional mapping at a scale previously inaccessible.

This dissertation includes unpublished co-authored material.

PUBLICATIONS:

Holston, A. S., Hinton, S. R., Lindley, K. A., Kearns, N. C., & Plesa, C. (2023). Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants. *bioRxiv*, 2023-12. In preparation.

Holston, A. S., Jimenez, P., Hinton, S. R., Lippert, L., & Plesa, C. (2024). Precision engineering of fusion phase variants to optimize chimeric histidine kinase functionality. In preparation.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

1. INTRODUCTION

1.1 Overview

The manner in which organisms, especially microscopic organisms, detect information and act on it is a very broad question impactful to most facets of life, piquing the interests of scientists for centuries (Downes & Blunt, 1877; Leeuwenhoek, 1667). Humans use their senses of sight, hearing, smell, taste, and touch to detect changes in the world around them via bringing the information inside through barriers and acting on it. Akin to how a person stepping on a sharp rock sends a signal through their skin to their nervous system, causing them to move, cells use receptors that span their membranes to detect external changes and carry the signal inside of the cell to adjust behavior. For detection of information, sensor histidine kinases (SHKs) are one of the most widely prevalent protein families in nature, where they exist in all domains of life—Archaea, Bacteria, and Eukarya (Alvarez et al., 2016); for eukaryotes, studies have not identified the presence of canonical SHKs in animals, though they are well-represented in other eukaryotic lineages, such as plants and fungi (Kabbara et al., 2019). SHKs sense and respond to an immense array of signals, both chemical and environmental, ranging from small molecules, peptides, and metal ions to pH and light (Ortega et al., 2017). Notably, the majority of SHKs are membrane-embedded and contain an extracellular or periplasmic sensing domain, allowing cells to bring signals from outside of the cellular cytoplasm to inside and act on them; thus, SHKs are an integral part of how many organisms sense and adapt to the conditions of their environments.

These homodimeric SHK proteins consist structurally of discrete domains with distinct functions: a sensing domain to detect activating stimuli, transmembrane segments for localization to the membrane, and internal signal transduction and output kinase domains, which employ a phosphorylation cascade to relay activation (Bhate et al., 2015). When activating stimuli are sensed, conformational changes occur that lead to the transfer of a phosphate from a conserved residue in the SHK to an aspartate on its cognate response regulator (RR). The RR then drives transcriptional changes, allowing the cell to suitably adapt to the environmental conditions. This two-part relay is called a two-component system (TCS), where the SHK and RR serve as the components.

As SHKs act to mediate a variety of cellular response mechanisms upon activation, they present potent targets for engineering, with foreseeable applications in agricultural yields, next-

generation probiotic development, environmental remediation, and increasing the tunability and robustness of genetic circuits (Hatstat et al., 2024; Ishii & Eguchi, 2021; Karan et al., 2009; J. Wang & Childers, 2022). Thus, knowing the activating ligands and interacting partners for SHKs allows for lucrative, impactful deployment and implementation of these in society.

Despite the widespread nature of SHKs, only a minute fraction of them have undergone functional characterization, as the current methods for identifying their activating stimuli are laborious and slow. As these sensors are highly modular, an auspicious approach for "deorphanization" employs chimeras where the sensing domains of uncharacterized SHKs are swapped onto a well-characterized kinase scaffold conserved between all chimeras, thereby facilitating scalable assays with parallel testing of many sensors for determination of activating ligands. However, previous work showed in the field that while functional chimeric SHKs can be created, some chimeric variants may also be constitutively active even in the absence of ligand or non-functional, where no activation occurs under any circumstances, including the presence of their activating ligand (Bi et al., 2016).

Thus, we aim to design and build large gene libraries of chimeric SHKs, express them, and use robust screens to accelerate deorphanization of these integral parts of two-component systems.

1.2 Two-Component Systems

Researchers first began studying two-component systems in the mid-1980s through biochemical characterization of sensory response pathways in bacteria. One such study, conducted by Ninfa and Magasanik, examined the *E. coli* nitrogen regulatory (NR) system that identified a phosphorylation-dependent mechanism linking environmental sensing to gene expression output, laying the foundation for the canonical sensor histidine kinase (SHK)–response regulator (RR) framework (Ninfa & Magasanik, 1986). Later genomic surveys revealed that TCSs are widespread across bacteria, with individual species often encoding dozens to hundreds of these systems to generate responses to varied stimuli such as osmotic pressure, nutrient shifts, and changes in pH (Ortega et al., 2017; Wuichet et al., 2010).

In the 1990s, structural studies elucidated key mechanistic details, including SHK autophosphorylation at a conserved histidine residue and phosphotransfer to an aspartate on the RR (C. Chang & Stewart, 1998). The discovery of non-canonical SHKs, such as hybrid histidine kinases, expanded the framework of the myriad ways in which TCSs may function (Uhl &

Miller, 1996). Though present in all domains of life and ubiquitous in bacteria, they are less common in eukaryotes, and researchers have not identified canonical TCS in the animal kingdom (Stock et al., 2000); however they are present in plants, likely due to lateral gene transfer (Capra & Laub, 2012). The declining cost of genome sequencing highlighted TCS diversity, with systems such as EnvZ/OmpR becoming models for studying signaling specificity and cross-talk avoidance (Ghose et al., 2023). Greater accessibility fueled expanded research, shedding light on their functional roles in cellular pathways.

1.3 Roles of Two-Component Systems in Cells

TCSs regulate pathways and processes critical for cellular function, such as osmoregulation (EnvZ/OmpR), nitrate homeostasis (NarX/NarL), and chemotaxis (CheA) (Cai & Inouye, 2002; Cheung & Hendrickson, 2009; X. Wang et al., 2012). A given bacterium may encode between ten and hundreds of unique TCS systems present (Buschiazzo & Trajtenberg, 2019; Ulrich & Zhulin, 2010). The modular architecture of the typical TCS, which comprises an SHK and its cognate RR, enables flexible environmental sensing when the SHK detects an environmental signal and in turn phosphorylates an aspartate residue on the receiver (Rec) domain of the RR, thereby initiating downstream output activities (Fig. 1).

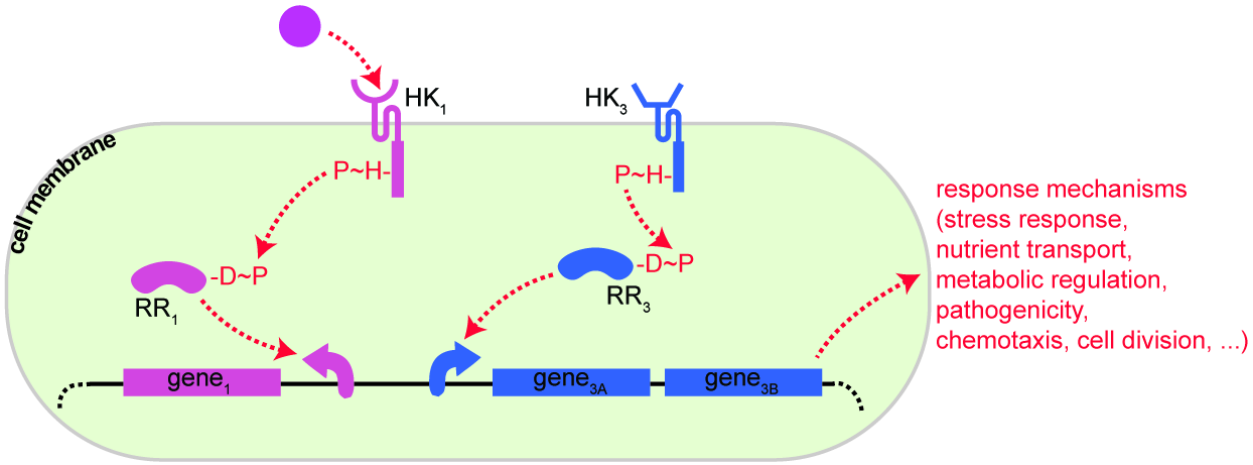


Figure 1 – TCS Signaling Overview: SHK and RR Phosphotransfer. A TCS is comprised of an SHK and an RR, and when the SHK detects stimuli, is phosphorylates its cognate RR, leading to activation of a downstream response

RR effector domains generally fall into five classes based on their output, with these being single-domain, protein-binding, enzymatically active, RNA-binding, and DNA-binding;

the majority of them, comprising nearly 70% of RRs, engage in DNA-binding and likely regulate transcription (Zschiedrich et al., 2016). SHKs, beyond their kinase activity, can also possess other functions, such as methyltransferase activity (Karniol & Vierstra, 2004). SHKs that also have a fused REC domain function as hybrid histidine kinases (HHK) and enable further signal integration, highlighting how TCS structural features position them as key hubs for cellular communication (Brüderlin et al., 2023).

1.4 Structure of a Sensor Histidine Kinase

To grasp how SHKs enable sophisticated signal integration, it is essential to understand how the modular architecture a typical SHK employs underpins this functional versatility. Here we will introduce the canonical architecture of class I SHKs, highlighting the spatial arrangement and conservation of domains through which they process signals into precise molecular responses. By dissecting these structural elements, we can appreciate how SHKs act as molecular switches, coupling external stimuli to intracellular signaling events and ultimately dictating the specificity, efficiency, and directionality of two-component system pathways (Buschiazzo & Trajtenberg, 2019). In cells, SHKs typically form homodimeric structures (Fig. 2).

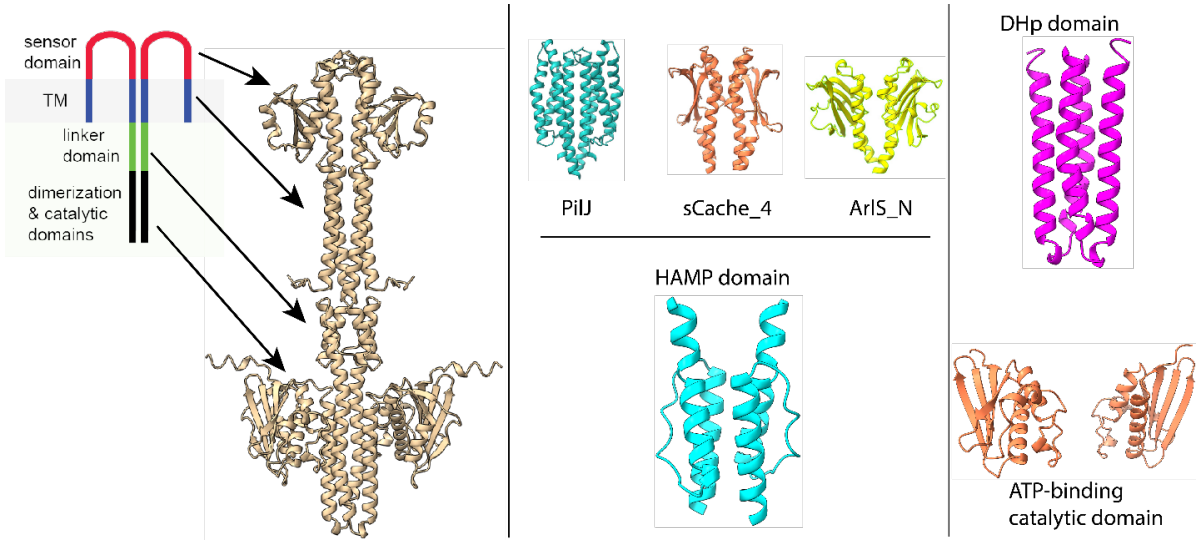


Figure 2 – Domain architecture and representative structures of typical SHKs. Left: Schematic diagram and full-length structure of annotated AlphaFold2-Multimer model of EnvZ, denoting the general domains one may find in a typical SHK. Right: Representative structural models of key SHK domains, including the PilJ, sCache_4, and ArlS_N sensor domains; the HAMP domain; the DHp (Dimerization and Histidine-phosphotransfer) domain; and the ATP-binding catalytic domain.

1.4.1 Sensor Domain

The sensor domain of a canonical SHK serves as the primary module for detecting extracellular or periplasmic stimuli, which then initiates a conformational cascade that propagates through the kinase structure. Located at the N-terminus, this domain typically adopts a variable fold, often α-helical, β-sheet-rich, or mixed α/β architectures, specialized for recognizing specific physicochemical signals, such as ligands, osmotic changes, or redox states (Matilla et al., 2022). In transmembrane SHKs, two transmembrane domains flank the sensor and anchor it to the membrane. Despite their diversity, these sensor domains universally couple perception of their environment to allosteric regulation at the core of the kinase, ensuring a high precision of control over catalytic activities.

1.4.2 Transmembrane Domains

The sensor domain initiates structural rearrangements that propagate through the transmembrane domains. Class I SHKs typically feature two transmembrane (TM) α-helices per monomer, forming an anti-parallel four-helix bundle in the typical homodimeric state. As they flank the sensor domain, the N-terminal helix (TM1) anchors it to the membrane, and the N-terminal helix (TM2) both anchors it to the membrane and also links it to cytoplasmic signaling domains (Bhate et al., 2015). As SHKs are a very large family, exceptions exist, such as DesK, which has five TM helices per monomer, as it also uses these to sense membrane fluidity (Zschiedrich et al., 2016). Mechanisms such as helical rotation, scissor-like motion, or piston-like shifts may mediate signal transduction across the TM bundle, though the type of movement varies across the SHK family, and these mechanical perturbations then pass to the cytoplasmic domains.

1.4.3 Linker Domains

After the transmembrane helices and before the catalytic domains, the presence and types of domains vary, and many of these domains enable signal transmission or cytoplasmic sensing. While not universally present in SHKs, these domains occur widely, with around 30% of SHKs containing HAMP domains, about 40% featuring PAS or GAF domains, and many employing combinations (e.g., HAMP-PAS, tandem HAMP repeats, tandem PAS repeats) to enhance signal processing. The STAC domain is another potential linker domain, though estimates suggest it occurs in only one to five percent of SHKs (Zschiedrich et al., 2016).

1.4.3.1 PAS Domain

The PAS domain takes its name from its initial discoveries in Per, Arnt, and Sim proteins. It acts as a linker domain in some histidine kinases, and it features a conserved three-dimensional fold as a defining feature, despite a limited sequence identity across different proteins. This fold features a single central antiparallel, five-stranded β-sheet, allowing it to serve a role either in signal transduction or as a cytoplasmic sensor, with specific PAS domains detecting intracellular redox potential, oxygen, and small ligands (Taylor & Zhulin, 1999).

1.4.3.2 GAF Domain

The GAF domain derives its name from its identification in cGMP-specific phosphodiesterases, adenylyl cyclases, and FhlA; it shares some structural similarities with the PAS domain, featuring a core of six-stranded antiparallel β-sheets and α-helices. In bacterial SHKs like DosT, GAF domains detect redox signals (such as O_2 via heme iron in DosT) to regulate hypoxia responses (Podust et al., 2008).

1.4.3.3 STAC Domain

The STAC (Solute carrier and Two-component signal transduction-Associated Component) domain acts as a structural module in certain SHKs, such as Pseudomonas CbrA. As a linker domain, STAC adopts a monomeric antiparallel four-helix bundle architecture distinct from the dimeric HAMP domain. The STAC domain may participate in both solute transport and signal transduction, and thus far no known protein contains more than a single STAC domain (Korycinski et al., 2015).

1.4.3.4 HAMP Domain

The HAMP domain derives its name from its presence in Histidine kinases, Adenylate cyclases, Methyl accepting proteins and Phosphatases. It spans approximately fifty residues and typically comprises two amphipathic sequences, AS1 and AS2, which form α-helices. Each helix consists of heptad repeats (a–g), where hydrophobic residues at positions *a* and *d* pack to form a buried hydrophobic core (Airola et al., 2010); an unstructured region connects the helices (Aravind & Ponting, 1999). Together, these helices form a four-helical parallel coiled-coil, enabling the HAMP domain to function as a critical signaling element. Conformational landscape analysis reveals that the domain facilitates signal transduction by enabling coordinated motions, including rotations, shifts, displacements, and tilts (Winski et al., 2024).

There are several competing models concerning the mechanism of signal transduction through the HAMP domain; the dynamic bundle model posits that the HAMP domain toggles between ordered (stable) and disordered (dynamic) states during signal transduction, in which the input signal destabilizes the helical bundle, thereby increasing flexibility and activating downstream signaling. The gearbox mechanism suggests rotational movement of the helices in opposite directions within the domain converts it between active and inactive states; some propose piston- and scissor-like models as well. These mechanistic propositions are not mutually exclusive, and the literature also suggests possible hybrid combinations of these mechanisms (Gushchin et al., 2021), and the precise mechanism may vary between SHKs. Structural studies of the SHK NarQ have shown that it converts the input of piston-like motions to amplified helical rotation at its output helices as a potentially conserved mechanism (Gushchin & Gordeliy, 2018).

Crucially, after the HAMP domain and prior to the next linker or output domain, insertions or deletions cause a disturbance, termed a "stutter," in the heptad repeat, ensuring the coiled-coil structure does not extend into the DHp domain, and this feature may enable bistability in the SHK, allowing it to convert between conformational states of comparable thermodynamic stability (Schmidt et al., 2017).

1.4.4 DHp Domain

SHKs rely on the DHp (Dimerization and Histidine-phosphotransfer) domain as a critical structural component. The DHp domain participates in several key reactions, including autophosphorylating its eponymous conserved histidine residue and subsequently phosphorylating the corresponding response regulator, with the DHp domain of bifunctional SHKs also participating in phosphatase activities through regulating the RR by dephosphorylating it. These reactions drive signal transduction in two-component systems. This domain forms a homodimeric antiparallel four-helix bundle, with a hairpin loop connecting the two α-helices; intriguingly, the length and handedness of the hairpin generally dictate the directionality of autophosphorylation: SHKs with a left-handed four-helix bundle phosphorylate in *cis*, and those with a right-handed four-helix bundle phosphorylate in *trans* (Casino et al., 2014), though this may not stay consistent between different types of DHp domains for SHKs (Trajtenberg et al., 2010).

The domain undergoes structural rearrangements in response to signals that affect the mobility and orientation of the catalytic and ATP-binding domain (CA) and modulates the position of the reactive histidine residue. In the absence of a signal, the DHp domain stabilizes in a kinase-off state, inhibiting autophosphorylation. Signal presence disrupts these structures, allowing phosphorylation to occur (Bhate et al., 2015).

1.4.5 CA Domain

The catalytic and ATP-binding domain (CA) features a canonical mixed α/β sandwich fold, in which three α-helices pack against a five-strand antiparallel β-sheet, forming a compact architecture that supports ATP or ADP binding in a structurally defined pocket. A notable structural element of this domain is the gripper helix, which employs surface-exposed hydrophobic sidechains for the critical role of mediating interactions with the DHp domain (Bhate et al., 2015). This domain is responsible for binding ATP and catalyzing the transfer of the γ-phosphate from ATP to the DHp's catalytic histidine. Importantly, the DHp-CA interface governs the efficiency and specificity of phosphotransfer, minimizing SHK cross-talk with non-cognate RRs (Zschiedrich et al., 2016).

1.5 Structure of a Response Regulator

Response regulators (RRs) are modular proteins that typically contain two primary domains: a conserved receiver (REC) domain on the N-terminus, and a C-terminal effector domain that carries out the output function. This type of architecture enables phosphorylation-dependent regulation of diverse downstream cellular processes.

1.5.1 REC Domain

The REC domain is typically around 120 residues in length and consists of a conserved fold of five parallel β-sheets surrounded by five α-helices. The conserved aspartate residue, which undergoes phosphorylation, resides at the C-terminus of β3 (Gao et al., 2019), and the β1-α1 loop contains a conserved motif that binds Mg^{2+} and stabilizes the phosphorylated Asp residue. Phosphorylation of the Asp residue typically triggers REC-mediated homodimerization via a conserved dimerization interface, though some RRs are also known to form heterodimers (Zschiedrich et al., 2016). An important structural feature, the switch loop (β4-α4), repositions upon RR phosphorylation, enabling activation of the effector domain (Bobay et al., 2012).

1.5.2 Effector Domain

The effector domain, located at the C-terminal end of the RR, connects to the REC domain via a flexible linker. Phosphorylation destabilizes REC-effector interactions, freeing the effector domain to engage in its functional role (Zschiedrich et al., 2016). The effector domain determines the RR's function and exhibits significant structural and functional diversity. Most effector domains participate in DNA binding, with the OmpR/PhoB family serving as a prominent example; this family binds tandem DNA repeats through a winged helix-turn-helix motif (Schmidl et al., 2019). Some effector domains act as enzymatic effectors, functioning as catalytic domains that directly modify proteins or molecules, such as the methylesterase activity of CheB (Galperin, 2006).

However, some RRs exist as single-domain proteins (SDRRs), lacking an effector domain entirely. These SDRRs function exclusively through phosphorylation-mediated conformational changes in their REC domain, often regulating cellular processes via protein-protein interactions (Lori et al., 2018).

1.6 Sensor Domain Diversity

Sensor domains are the most structurally and functionally diverse modules within TCSs, enabling organisms to detect an extraordinary range of chemical and environmental stimuli. These domains serve as the primary input modules for TCSs, determining the specificity and breadth to which a given TCS may respond. Sensor domains have evolved to perceive a wide range of input stimuli; thus, it is no surprise that the structural architecture of these domains also spans a wide spectrum.

1.6.1 Types of Sensor Domains

Histidine kinases exhibit extensive diversity in their sensor domains (Gumerov et al., 2024). A particularly relevant superfamily of sensor domains is the Cache superfamily. One of the principal structural characteristics of Cache domains is the presence of a long N-terminal α-helix that extends directly from the preceding transmembrane segment into the extracellular domain region, with a core comprising three strands that studies theorize form a β-sheet; the Cache superfamily includes members that contain variations of the Cache domain, with some members pertinent to this thesis being the ArlS_N, sCache_3_2, dCache_1, sCache_like, sCache_4, CHASE8, and 2CSK_N domains (Upadhyay et al., 2016). These sensor classes may feature either a single Cache domain (sCache) or a tandem arrangement (dCache).

Another important group is the PAS superfamily, which encompasses domains structurally related to the PAS fold described above; KinB_sensor is a notable example in this family. The 4HB_MCP_1 superfamily includes sensor domains that use four-helical bundles for ligand binding, with examples such as 4HB_MCP_1 and CHASE3 (Ortega et al., 2017). Some sensor classes, like PilJ (composed entirely of α-helices), do not fall into any established superfamily.

1.6.2 Degeneracy of Sensors

While many SHKs are highly specific, some individual sensors respond to inputs spanning chemically and physically diverse stimuli. For example, the SHK PhoQ detects divalent cations such as magnesium and calcium, as well as antimicrobial peptides and pH (Groisman et al., 2021). More broadly, SHKs across the phylogenetic spectrum sense a wide range of environmental cues spanning multiple physicochemical classes.

1.6.3 Orphaned Receptors

Despite many studies on two-component systems, the vast majority of SHKs remain "orphaned"; their sequences can be assigned to known sensor domain families, but their ligands and binding partners are not known and have not been experimentally characterized. Of the roughly 17.7 million unique sensor domains identified, functional mapping only exists for several hundred (Park et al., 2023). This gap is further complicated by the fact that many SHK sensor domains retain overall conservation among paralogs while exhibiting subtle variations in the amino acid residues that form the ligand-binding pocket, thereby potentially altering their specificity.

1.7 Engineering of Histidine Kinases

For engineering histidine kinases, early studies demonstrated that the physical boundaries between the structural domains offer natural junctions where domain swapping or fusion occurs with more minimal disturbance to the global structure. For engineered chimeras, researchers exploited these natural "fusion points" to create SHKs that combine the sensor domain from one protein with the signal transduction and catalytic domains from another, thereby altering the receptor's functional properties via activation of the sensor domain leading to the output response of its fusion partner. Additionally, SHK-chemoreceptor chimeras have been developed that contribute to the knowledge of fusion points; Taz is one such example, which combines the periplasmic, transmembrane, and HAMP domains from the Tar chemoreceptor to the DHp and

CA domains of EnvZ (Utsumi et al., 1989). Thus, engineering SHKs both gives insight into mechanistic function and may provide a way to characterize the ligand specificity of sensor domains of unknown function.

1.7.1 Fusion Points

Researchers have created chimeras with a large variety of fusion junctions (Fig. 3) based on the different domain boundaries (Bi et al., 2016).

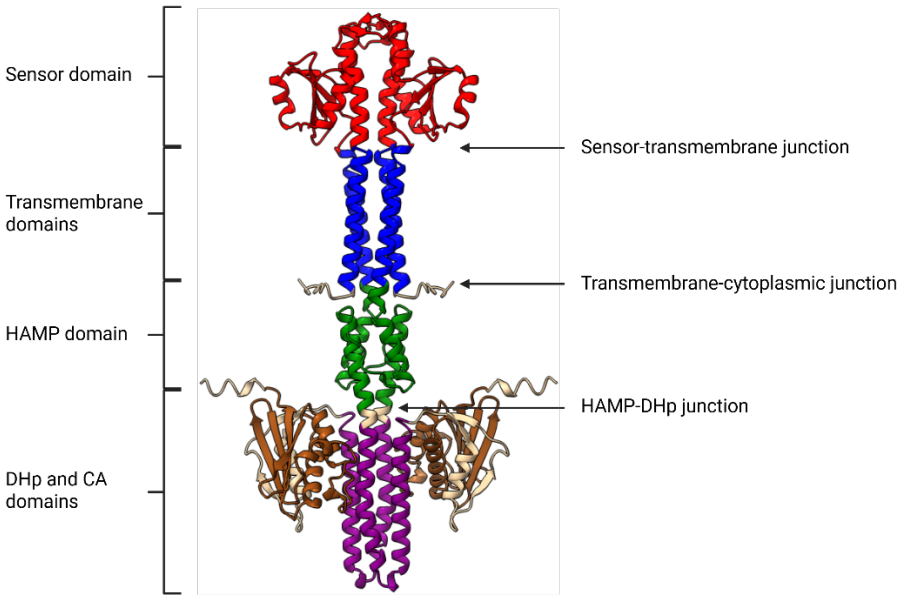


Figure 3 – AlphaFold2-Multimer model of EnvZ histidine kinase from *E. coli*. The left labels correspond to the major functional regions, corresponding to the structure color-coded by domain: the periplasmic sensor domain (red), transmembrane domains (blue), HAMP domain (green), DHp domain (purple), and CA domain (brown). Common fusion junctions between domains are indicated on the right: sensor-transmembrane, transmembrane-cytoplasmic, and HAMP-DHp.

1.7.1.1 Sensor–Transmembrane

The sensor domain connects to transmembrane helices TM1 and TM2 at the sensor–transmembrane boundary in many chimeras. This method alters only the sensor domain and seeks to preserve the continuity of hydrophobic helices in the membrane, with the aim of maintaining proper ligand-induced conformational changes that propagate across the cell envelope, with examples such as in a chimera created that fuses lanthanide-sensing structures in

place of the iron-sensory loops of PmrA (Liang et al., 2013) or one that places the citrate-sensing domain of CitA above the transmembrane helices of the Tar chemoreceptor (Bi et al., 2016).

1.7.1.2 Transmembrane–Cytoplasmic

Fusions at the transmembrane-cytoplasmic junction seek to preserve the continuity of transmembrane helices of TM2 with the structural elements of the linker domains, such as the Tar-EnvZ chimera Tez1 (Zhu & Inouye, 2003). For the Tar and Tap chemoreceptors, Tar–Tap and Tap–Tar chimeras with fusion junctions located just downstream of TM2 demonstrated functionality, though slightly diminished (Weerasuriya et al., 1998).

1.7.1.3 HAMP-DHp

One of the more promising fusion points is the HAMP-DHp junction, and this is the fusion point relevant to this thesis. Fusions here also focus on preserving the continuity of the helical phase, as the HAMP domain connects seamlessly to the N-terminal helix of the DHp domain, forming a continuous, extended helix; if the natural heptad repeat found in this area remains intact, chimeras can properly shift between the kinase-active and phosphatase-active states, with disruptions to the heptad, and thus the coiled-coil structure, altering ligand response (Ferris et al., 2012). Fusions here also necessitate maintaining the rotational phase of the coiled-coil HAMP domain (Ferris et al., 2014). In a study that used systematic fusion mapping to explore a variety of modular fusions, post-HAMP fusion junctions yielded the most successful fusions, with the theorization of signal transduction through the HAMP domain being mechanistically conserved between SHKs (Bi et al., 2016).

1.7.2 Linkers

Some chimeras employ linkers at the fusions of structural junctions, and linker properties influence the efficacy of the fusion. One sensor–transmembrane fusion replaced the native CpxA sensor domain with a designed Zn^{2+}-sensing domain, linking it to TM1 and TM2 using native CpxA-derived linker sequences. In this system, the length of the linker between TM1 and the sensor domain varied basal activation in a pattern-agnostic manner; however, the linker length between the sensor and TM2 correlated with the degree of activation, depending on whether the helical phase of the linker aligned with the phase of the first residue of TM2 (Hatstat et al., 2024). Another study fused the light-sensing YtvA with the heme-binding FixL, with the junction in their common α-helical linker sequence, producing a chimera comprising the light-sensing domain from YtvA, the hybrid α-helical linker, and the DHp and CA domains from

FixL. A variety of deletion and insertion variants were made in the linker region and revealed that the periodicity of the heptad repeat influenced light regulation: all variants differing by seven residues in linker length showed similar light-responsive activity, while a single-residue insertion in the linker region inverted the response (Möglich et al., 2009).

Bifunctional SHKs maintain a balance between their kinase and phosphatase states, but this balance remains susceptible to perturbation. Red-light-responsive photosensory modules, either *Dm*PSM or *Dr*PSM, fused to the α-helical linker, DHp, and CA portions of FixL to create functional chimeras that red light repressed. Additional variants with linker lengths extended or shortened by up to forty residues formed a panel whose activity was subsequently measured. Most functional variants aligned to registers offset by +1 or +3 residues from the typical heptad repeat—positions rarely found in natural light-sensing SHKs; some variants also exhibited inverted, red-light-activated responses, indicating that they inverted the typical kinase-phosphatase balance (Meier et al., 2024). Thus, using linkers at fusion junctions for SHKs can alter both functionality and direction of functionality, leading to sensors that are constitutively active, non-functional, or with an altered balance between kinase-phosphatase states.

1.7.3 Function

The ability of a chimeric SHK to precisely regulate switching between kinase-active and kinase-inactive states, which also corresponds to the phosphatase-active state for bifunctional SHKs, is essential to achieving functional signal transduction. These states are tightly coupled, and disruption of this balance can result in constitutively active ("locked-on") or inactive ("locked-off") conformations. Structural determinants governing this switch are highly sensitive to perturbations, particularly at fusion junctions that disrupt native interdomain contacts or heptad repeats forming the coiled-coil structures (Meier et al., 2024).

Fusion point selection is therefore of crucial importance. Even minor shifts, such as single-residue insertions or deletions, at the fusion junction between domains can shift SHK activity from responsive to unresponsive or inverted, as demonstrated in the light-sensitive SHK YF1, where altering the fusion site in this manner inverted the output response from light-repressed to light-activated (Ohlendorf et al., 2016). These shifts reflect structural mechanisms that alter the ability of the CA and DHp domains to interact, thereby toggling between the kinase-active and phosphatase states.

The phosphatase state has been shown to be particularly sensitive to structural disruption. Chimeras with modified HAMP domains have been observed to adopt alternate conformations of the DHp region that mimic either kinase or phosphatase states, depending on the rotational and packing geometry of the helices (Ferris et al., 2012). Interestingly, variants locked in kinase-active states were much more common than phosphatase-active ones, especially when upstream sensor regulation was removed (Yoshida et al., 2007).

Systematic fusion mapping between the modular domains revealed that even when sensor domains are successfully linked to downstream regions, minor sequence misalignments can abolish signaling or lock receptors into these locked-off or locked-on states, depending on the compatibility of the helices and hydrophilic or hydrophobic properties of residues at the fusion junction. Post-HAMP fusions tend to yield the most consistent results, due to the possibility of the HAMP domain's conserved mechanism across SHKs, and suggest this junction may serve as a valuable site for characterizing ligand specificity in diverse sensor domains (Bi et al., 2016).

Taken together, these findings underscore that SHK function depends not only on compatibility between sensor and effector domains but also on fine-tuned structural and energetic constraints at the fusion interface. Fusion designs must preserve, or purposefully alter, the helical register, coiled-coil rotational phase, and conformational flexibility to avoid failure modes such as locked-on, locked-off, or inverted signaling. These principles are critical for engineering ligand-responsive chimeric SHKs and provide a potential framework for functionally mapping uncharacterized SHKs.

1.7.4 Importance of Orthogonality

In engineering chimeric SHKs, one commonly swaps domains to confer new ligand-binding specificities or to reprogram output signal responses. A central design requirement in these efforts is orthogonality, which refers to an engineered SHK and its corresponding RR interacting exclusively with each other while being insulated from crosstalk with endogenous signaling pathways.

This property is critical for achieving specific, interpretable signal transduction in bacteria, which typically encode dozens to hundreds of SHK-RR pairs, as it ensures that an engineered SHK activates only its intended RR and output pathway. Without this, signals may be confounded by undesired interactions, which may lead to false positives, ambiguous phenotypes, or noise. Thus, orthogonal signaling facilitates clear mapping between input and output, enabling

mechanistic dissection of stimulus perception, autophosphorylation, and phosphotransfer dynamics. Several strategies help achieve orthogonality. One such strategy is swapping the DNA-binding domain (DBD) of the RR to redirect transcriptional output to synthetic promoters not recognized by native regulators, which isolates the engineered system from endogenous pathways (Schmidl et al., 2019). Another strategy is to mutate specificity-determining residues at the protein-protein interface between SHKs and RRs, which thereby insulates them from interactions with native components (McClune et al., 2019). Orthogonality can also be reinforced at the cellular level by deleting native SHKs and RRs that might otherwise interfere with the synthetic pathway.

In our systems, we both employ deletion of native SHKs and RRs along with the use of an RR with a swapped DBD, which activates a synthetic promoter.

1.8 Methods for Studying and Characterizing Histidine Kinases

As histidine kinases are broadly used as environmental sensors, deciphering the detailed molecular events, ranging from identifying activating ligands to conformational transitions and mechanistic catalysis, has driven the application of a wide array of techniques. Here we describe the methodologies that have been implemented to capture static and dynamic states of SHKs, including both experimental and computational.

1.8.1. Experimental Approaches

1.8.1.1 Structure Determination

X-ray crystallography has resolved high-resolution crystal structures of portions of SHKs in various conformational states. For example, structures of the ATP-binding domain or full catalytic cores have provided detailed views about active sites, revealing such things as the positioning of ATP relative to the conserved histidine residue (Trajtenberg et al., 2010). Structures obtained using nonhydrolyzable ATP analogues have captured active-state intermediates by trapping transient species, thereby elucidating the activity-modulating *cis*- and *trans*-phosphorylation mechanisms (Casino et al., 2014). Some sensor domains, both in ligand-free and ligand-bound states, have been resolved with crystal structures (Cheung & Hendrickson, 2009), along with some sensor-TM-HAMP fragments (Gushchin et al., 2020).

Nuclear Magnetic Resonance (NMR) spectroscopy has also been used to resolve some domains, such as the EnvZ kinase domain bound to ATP analogues, thereby illuminating details of the dimerization interface and flexible loops (West & Stock, 2001). Together, crystallography

and NMR have provided atomic-level insights into both static and dynamic aspects of SHK function. However, up to this date, no full-length SHK has been experimentally determined at the atomic level due to the difficulties associated with their length, transmembrane nature, dynamics, flexibility, and transience of intermediate states (Gushchin et al., 2020).

1.8.1.2 Biochemical and Biophysical Techniques

For detection and quantification of ligand binding and catalytic activity, biochemical methods have played a central role. In the past, autophosphorylation, phosphotransfer, and phosphatase functions were typically assayed using radiolabeled [γ-^{32}]-ATP for measuring the kinetics, typically paired with SDS-PAGE and phosphorimaging, autoradiography, or filter-binding methods, with some strategies employing acid-quenching residues to preserve the phosphohistidine group; site-directed mutagenesis to create SHK mutants with defective ATP-binding or phosphorylation (Sankhe et al., 2018; Ueno et al., 2015). As an alternative to radiolabeled ATP, Phos-tag acrylamide gels have also been used (Liu et al., 2017).

Differential scanning fluorimetry has also been used as a label-free method for detecting thermal changes upon binding, providing indirect evidence of inhibitor interactions (Velikova et al., 2016). Förster resonance energy transfer (FRET) has also been widely used to determine domain rearrangement and kinase activity in live cells, such as on ligand addition (Bi et al., 2016). Circular dichroism (CD) spectroscopy is another technique that has seen use in assessing secondary structure in both mutant and wild-type sensors for comparing structural and functional differences (Sankhe et al., 2018).

1.8.2 Computational Approaches

1.8.2.1 Bioinformatics and Coevolutionary Methods

Computational analysis plays a useful role in identifying determinants of dimerization interfaces and ligand specificity. Large-scale multiple sequence alignments (MSAs) and bioinformatic analyses have been employed to identify covarying amino acid pairs within SHKs and partners. For instance, covariation analysis of more than 1200 sequences distinguished residue pairs in the DHp domain and their cognate RRs that coevolve to determine specificity (Skerker et al., 2008). Bioinformatics tools, including PSI-BLAST and clustering algorithms such as AGAPE, have been used to identify and classify conserved sensor domains, kinase domains, and other structural motifs (Zhang & Hendrickson, 2010).

1.8.2.2 Molecular Simulations

Molecular dynamics (MD) simulations provide dynamic insights that complement static crystal structures by capturing the conformational dynamics and ligand-induced motions of SHKs at atomic resolution (Gushchin et al., 2020). MD simulations have elucidated conformational transitions for signal transduction in SHKs, such as for PhoQ, where MD simulations, validated by disulfide cross-linking, modeled and uncovered TM helices using scissoring and rotational motions to enable signal transduction (Lemmin et al., 2013). To probe catalytic mechanisms at higher resolution, hybrid quantum mechanics/molecular mechanics (QM/MM) are employed; this is a multiscale simulation approach where the reactive site is treated with quantum mechanical methods while modeling the remainder of the system with classical mechanics. In one study, QM/MM steered MD with density functional theory showed that WalK catalyzes autophosphorylation via a concerted reaction with a manageable energy barrier (Olivieri et al., 2020).

1.8.2.3 Virtual Docking

Structure-based virtual screening has emerged as a critical strategy for identifying inhibitors that can exploit the conserved ATP-binding pocket of SHKs. *In silico* docking studies are integrated with high-throughput virtual screening to predict binding of small-molecule inhibitors in this pocket (Goswami et al., 2017); these techniques take into account SHK flexibility and are often followed by corroborating MD simulations and *in vitro* kinase assays, thereby using functional data to validate computational predictions (Velikova et al., 2016).

1.8.3 High-Throughput Approaches

Also of interest are high-throughput (HT) approaches, where many SHKs, whether WT or variants, are tested in a scalable manner. One HT screening platform was developed by combining SLAY (surface-localized antimicrobial display) with heterologous TCS expression and sort-seq to measure antimicrobial peptides' (AMPs) interactions with TCS. They screened a library of 117 human AMPs and >3,680 variants against the PhoPQ system, identifying 13 new human AMP activators with diverse structures and charges. A machine learning model trained on this data revealed non-obvious features, such as hydrophobicity periodicity, that predict activation. This platform establishes a scalable method for profiling sensor input specificity and supports rational design of orthogonal signaling modules (Brink et al., 2021).

Another potential HT approach is employing deep mutational scanning (DMS) with a screening approach; in a study, DMS was used to construct a comprehensive library of all 1,140 single-residue variants of the 60-residue DHp domain of EnvZ. These variants were introduced into a fluorescent reporter strain with reporters driven by OmpR-, RstA-, and CpxR-regulated promoters, enabling quantitative assessment of signaling via fluorescence-activated cell sorting (FACS) and deep sequencing (Sort-seq). This allowed them to simultaneously evaluate the kinase activity and signaling specificity of the many EnvZ variants towards three different RRs in vivo. The results revealed that 363 mutations increased signaling to noncognate RRs, confirming that paralog specificity is marginal and vulnerable to disruption by single substitutions (Ghose et al., 2023).

Next-generation sequencing (NGS) platforms have advanced TCS regulon mapping. In a study to determine RR-DNA interactions, all nonessential TCSs in *Staphylococcus aureus* were deleted, then constitutively active RRs were reintroduced one at a time, and applied RNA-seq and mass spectrometry to define the entire regulon for each TCS; this approach allows for the parallel identification of many binding sites from complex mixtures of genomic DNA and generates binding profiles for a suite of RRs (Rapun-Araiz et al., 2020).

Microfluidics may also be employed for HT screening; for a study seeking to elucidate domain coupling, a CpxA library was constructed by microfluidic Golden Gate assembly of four DNA parts, including 81 TM helices, 8 linker variants, and 8 S-helix mutations, enabling the potential assembly of 5,184 possible variants. The library was screened via FACS of a GFP reporter strain followed by Illumina sequencing to quantify enrichment. For the full library, 4,351 variants were successfully assembled and evaluated; this approach uncovered that the S-helix establishes baseline activity, while the TM and linker domains modulate output, often in a context-specific manner, illuminating the modular control of SHK signaling (Clark et al., 2021).

All in all, a large variety of experimental and computational approaches have been used for the determination of SHK mechanisms and ligand specificity, both for single sensors and for many sensors at a time; though much has been discovered, a vast majority of SHKs remain orphaned.

1.9 Importance of Histidine Kinase Deorphanization

SHKs, as pivotal components of TCS, mediate bacterial responses to an array of environmental stimuli, with downstream processes governing essential cellular functions

including regulation of virulence, stress response, and metabolic adaptation. Additionally, they are also absent in mammals, making SHKs attractive targets for diverse applications. Deorphanization of SHKs at scale could fundamentally transform our approach to environmental monitoring, synthetic gene circuit design, and understanding of host–pathogen dynamics by providing a comprehensive catalog of these versatile sensor modules (Capra & Laub, 2012).

1.9.1 Host-Microbe Interactions

In many pathogenic bacteria, SHKs play a central role in sensing host-derived signals such as pH, ion concentrations, and antimicrobial peptides; mapping the ligand specificity of these kinases will clarify how bacteria detect and adapt to host environments during infection (Ishii & Eguchi, 2021). A deeper understanding of these mechanisms will shed light on bacterial strategies for immune evasion and survival under host-imposed stress. This knowledge also creates new opportunities to target virulence pathways with precision. By identifying key SHKs that mediate pathogenic signaling, researchers can develop inhibitors that block bacterial adaptation without harming host cells (Bem et al., 2015). Such agents may function as antivirulence therapies, reducing selective pressure for resistance while limiting progression of disease.

Beyond antimicrobial design, insights into SHK signaling can guide microbiome and probiotic engineering (Rottinghaus et al., 2020). By tuning the sensory capabilities of beneficial microbes, researchers can enhance their ability to detect host cues and respond appropriately, supporting microbiome stability and host health (J. Wang & Childers, 2022). These same principles can be applied in agriculture, where optimized microbial strains could promote plant health and resilience by detecting and responding to soil and root signals (Daudu et al., 2017; F.-F. Wang & Qian, 2019).

1.9.2 Biosensors

SHKs provide a robust foundation for biosensor development due to their modular architecture. Sensor domains respond selectively to defined inputs, making them excellent candidates for signal detection across diverse contexts (Matilla et al., 2022). Their phosphorylation-based signaling enables rapid response times, which are critical for real-time monitoring of output (Bhate et al., 2015). As researchers continue to map sequence-function relationships within individual domains, the ability to engineer SHKs with tailored sensitivity, specificity, and dynamic range will expand significantly (Landry et al., 2018). Identifying

activating ligands opens the door to building multiplexed sensor arrays that distinguish among multiple environmental or clinical cues (Ishii & Eguchi, 2021). When coupled to versatile output modules, such as fluorescent proteins, these sensors can produce easily quantifiable signals suitable for use in field, clinical, or industrial settings (Joshi et al., 2024). For example, researchers adapted two bacterial TCSs into probiotic *E. coli* to function as tetrathionate and thiosulfate sensors in mice (Daeffler et al., 2017), demonstrating the power of *in vivo* applications (Fig. 4).

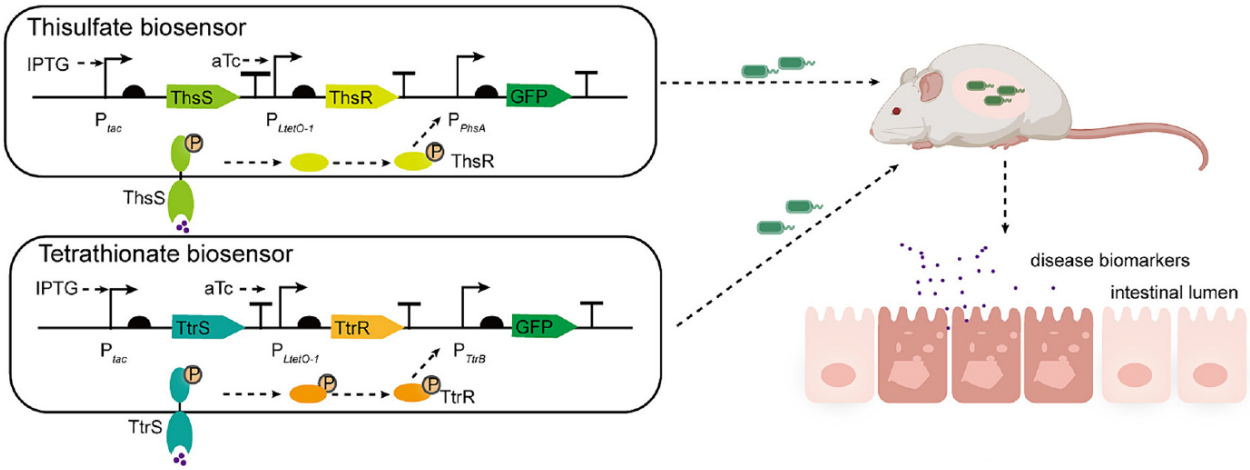


Figure 4 – Engineered SHK Biosensors for Inflammation Detection in Mice . Diagram showing the engineering and adaption of the TtrS/TtrR and ThsS/ThsR TCS biosensors from *Shewanella halifaxensis* for application in sensing gut inflammatory biomarkers in mice (Cao et al., 2024). Figure reproduced with permission from Cao, W., Huang, C., Zhou, X., Zhou, S., & Deng, Y. (2024). Engineering two-component systems for advanced biosensing: From architecture to applications in biotechnology. *Biotechnology Advances, 75*, 108404. Copyright © 2024 Elsevier. Permission obtained from Elsevier via Copyright Clearance Center (License Number: 6023890753385).

SHK-based biosensors hold particular promise in environmental applications, where they can detect trace pollutants or metabolic intermediates, and in clinical diagnostics, where they may track biomarkers in real time, facilitating a simpler path to bioremediation. In agriculture, engineered SHKs could support precision monitoring of soil health and pathogen presence, improving crop resilience and sustainability.

1.9.3 Impacts on Synthetic Biology

SHK deorphanization would have a massive impact on the field of synthetic biology, as synthetic biology thrives on modularity and predictability, and well-characterized SHKs provide a valuable set of building blocks for constructing programmable circuits (Lazar & Tabor, 2021). With comprehensive ligand profiles, researchers can design signal transduction systems that respond to specific stimuli and trigger defined genetic outputs. The modularity of SHK domains allows integration with a range of RRs and output elements, enabling new combinations of inputs and behaviors. Through rational engineering and directed evolution, SHKs can be modified to detect non-native ligands or respond to novel stimuli, expanding their functional range (Hatstat et al., 2024). Hybrid constructs with multiple sensing elements can support complex logic operations, allowing cells to process several signals and compute a coordinated response; these engineered circuits can be deployed to control gene expression, manage metabolic flux, or modulate stress responses in industrial bioprocessing (H.-J. Chang et al., 2018). As large-scale ligand discovery proceeds, the number of available SHK modules will grow, offering a broader chemical sensing space that includes ions, metabolites, and signaling molecules. This expanded repertoire will support the development of engineered microbial communities capable of coordinated behavior, distributed logic, and adaptive regulation across varied environments.

Thus, large-scale deorphanization of SHKs would be deeply impactful both for basic biology and also for engineering applications.

1.10 Scope and Aims of this Work

This thesis addresses the central challenge of accelerating the deorphanization and engineering of SHKs by developing and integrating multiplexed gene libraries and a screening platform. The work is motivated by the need to systematically decipher how sequence features, especially domain boundaries and helical fusion phases, govern the basal activation states of SHKs, with the ultimate goal of enabling large-scale deorphanization and rational design of biosensors for a wide range of applications.

The scope of this thesis encompasses three tightly integrated areas:

(1) Multiplexed Gene Library Construction: The development and application of Degenerate DropSynth, a scalable, cost-effective method for the programmable assembly of large libraries of diverse gene variants. This approach enables the systematic creation of

chimeric SHKs with multiple, precisely designed fusion phase variants per parent gene, facilitating exploration of sequence-function relationships at an unprecedented scale.

(2) Genetic Circuit and Plasmid Engineering: The rational design and optimization of genetic circuits, both plasmid-based and genomically integrated, that couple SHK activity to robust fluorescent and life-death selection outputs. These circuits are engineered to maximize signal-to-noise, dynamic range, and selection stringency, enabling efficient discrimination between functional, non-functional, and constitutively active ("locked-on") variants in libraries.

(3) High-Throughput Functional Screening: The establishment of the Locked-On Sorting (LOS) assay, a multiplexed, FACS-based platform for quantifying the basal activation states of thousands of chimeric SHKs in parallel. This assay leverages barcode sequencing and computational analysis to infer variant-specific signaling states and to reveal how helical phase and sensor class modulate SHK activity.

Aims

The specific aims of this work are:

(1) To develop and validate a scalable gene synthesis platform (Degenerate DropSynth) for constructing large, diverse libraries of chimeric SHKs and their local designed mutants. This includes demonstrating the ability to generate up to eight phase variants per gene, analyzing assembly fidelity and coverage, and providing practical guidelines for library design and scale-up

(2) To design and optimize genetic circuits that enable robust selection and screening of SHK libraries, both in plasmid and chromosomal contexts. This includes maximizing dynamic range and minimizing background activity, implementing life-death selection modules, and ensuring modularity and stability for large-scale screening workflows

(3) To systematically characterize the functional consequences of fusion phase and sensor domain diversity in chimeric SHKs using high-throughput LOS assays. This involves quantifying basal activation states across thousands of variants, analyzing the impact of phase alignment and sensor class on signaling output, and training predictive machine learning models to relate sequence features to functional phenotypes

(4) To integrate these platforms into a unified pipeline for the efficient generation, screening, and analysis of chimeric SHK libraries, laying the groundwork for rapid deorphanization and rational engineering of biosensors. The pipeline is designed to be extensible to other protein families and screening modalities.

By achieving these aims, this thesis establishes a comprehensive methodological framework for the scalable engineering and functional screening of modular sensor histidine kinases. The approaches and insights developed in this thesis not only address key points of strain in SHK deorphanization but also provide generalizable strategies for accelerating synthetic biology research across diverse protein systems.

The next chapter, Degenerate DropSynth for Simultaneous Assembly of Diverse Gene Libraries and Local Designed Mutants, focuses on Aim 1 and includes material co-authored with Samuel R Hinton, Kyra A Lindley, Nora C Kearns, and Calin Plesa.

1.10 Introduction Bibliography

Airola, M. V., Watts, K. J., Bilwes, A. M., & Crane, B. R. (2010). Structure of Concatenated HAMP Domains Provides a Mechanism for Signal Transduction. *Structure*, *18*(4), 436–448. https://doi.org/10.1016/j.str.2010.01.013

Alvarez, A. F., Barba-Ostria, C., Silva-Jiménez, H., & Georgellis, D. (2016). Organization and mode of action of two component system signaling circuits from the various kingdoms of life. *Environmental Microbiology*, *18*(10), 3210–3226. https://doi.org/10.1111/1462-2920.13397

Aravind, L., & Ponting, C. P. (1999). The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiology Letters*, *176*(1), 111–116. https://doi.org/10.1111/j.1574-6968.1999.tb13650.x

Bem, A. E., Velikova, N., Pellicer, M. T., Baarlen, P. van, Marina, A., & Wells, J. M. (2015). Bacterial Histidine Kinases as Novel Antibacterial Drug Targets. *ACS Chemical Biology*, *10*(1), 213–224. https://doi.org/10.1021/cb5007135

Bhate, M. P., Molnar, K. S., Goulian, M., & DeGrado, W. F. (2015). Signal Transduction in Histidine Kinases: Insights from New Structures. *Structure*, *23*(6), 981–994. https://doi.org/10.1016/j.str.2015.04.002

Bi, S., Pollard, A. M., Yang, Y., Jin, F., & Sourjik, V. (2016). Engineering Hybrid Chemotaxis Receptors in Bacteria. *ACS Synthetic Biology*, *5*(9), 989–1001. https://doi.org/10.1021/acssynbio.6b00053

Bobay, B. G., Hoch, J. A., & Cavanagh, J. (2012). Dynamics and activation in response regulators: The β4-α4 loop. *BioMolecular Concepts*, *3*(2), 175–182. https://doi.org/10.1515/bmc-2011-0063

Brink, K. R., Mu, A. M., Hoang, K. V., Groszman, K., Gunn, J. S., & Tabor, J. J. (2021). *High-throughput discovery of peptide activators of a bacterial sensor kinase* (p. 2021.06.01.446581). bioRxiv. https://doi.org/10.1101/2021.06.01.446581

Brüderlin, M., Böhm, R., Fadel, F., Hiller, S., Schirmer, T., & Dubey, B. N. (2023). Structural features discriminating hybrid histidine kinase Rec domains from response regulator homologs. *Nature Communications*, *14*(1), 1002. https://doi.org/10.1038/s41467-023-36597-8

Buschiazzo, A., & Trajtenberg, F. (2019). Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level? *Annual Review of Microbiology*, *73*(1), 507–528. https://doi.org/10.1146/annurev-micro-091018-054627

Cai, S. J., & Inouye, M. (2002). EnvZ-OmpR Interaction and Osmoregulation in Escherichia coli. *Journal of Biological Chemistry*, *277*(27), 24155–24161. https://doi.org/10.1074/jbc.M110715200

Cao, W., Huang, C., Zhou, X., Zhou, S., & Deng, Y. (2024). Engineering two-component systems for advanced biosensing: From architecture to applications in biotechnology. *Biotechnology Advances*, *75*, 108404. https://doi.org/10.1016/j.biotechadv.2024.108404

Capra, E. J., & Laub, M. T. (2012). Evolution of Two-Component Signal Transduction Systems. *Annual Review of Microbiology*, *66*(1), 325–347. https://doi.org/10.1146/annurev-micro-092611-150039

Casino, P., Miguel-Romero, L., & Marina, A. (2014). Visualizing autophosphorylation in histidine kinases. *Nature Communications*, *5*(1), 3258. https://doi.org/10.1038/ncomms4258

Chang, C., & Stewart, R. C. (1998). The Two-Component System1: Regulation of Diverse Signaling Pathways in Prokaryotes and Eukaryotes. *Plant Physiology*, *117*(3), 723–731. https://doi.org/10.1104/pp.117.3.723

Chang, H.-J., Mayonove, P., Zavala, A., De Visch, A., Minard, P., Cohen-Gonsaud, M., & Bonnet, J. (2018). A Modular Receptor Platform To Expand the Sensing Repertoire of Bacteria. *ACS Synthetic Biology*, *7*(1), 166–175. https://doi.org/10.1021/acssynbio.7b00266

Cheung, J., & Hendrickson, W. A. (2009). Structural Analysis of Ligand Stimulation of the Histidine Kinase NarX. *Structure*, *17*(2), 190–201. https://doi.org/10.1016/j.str.2008.12.013

Clark, I. C., Mensa, B., Ochs, C. J., Schmidt, N. W., Mravic, M., Quintana, F. J., DeGrado, W. F., & Abate, A. R. (2021). Protein design-scapes generated by microfluidic DNA assembly elucidate domain coupling in the bacterial histidine kinase CpxA. *Proceedings of the National Academy of Sciences*, *118*(12), e2017719118. https://doi.org/10.1073/pnas.2017719118

Daeffler, K. N., Galley, J. D., Sheth, R. U., Ortiz-Velez, L. C., Bibb, C. O., Shroyer, N. F., Britton, R. A., & Tabor, J. J. (2017). Engineering bacterial thiosulfate and tetrathionate sensors for detecting gut inflammation. *Molecular Systems Biology*, *13*(4), 923. https://doi.org/10.15252/msb.20167416

Daudu, D., Allion, E., Liesecke, F., Papon, N., Courdavault, V., Dugé de Bernonville, T., Mélin, C., Oudin, A., Clastre, M., Lanoue, A., Courtois, M., Pichon, O., Giron, D., Carpin, S., Giglioli-Guivarc'h, N., Crèche, J., Besseau, S., & Glévarec, G. (2017). CHASE-Containing Histidine Kinase Receptors in Apple Tree: From a Common Receptor Structure to Divergent Cytokinin Binding Properties and Specific Functions. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.01614

Downes, A., & Blunt, Thos. P. (1877). Researches on the Effect of Light upon Bacteria and other Organisms. *Proceedings of the Royal Society of London*, *26*, 488–500.

Ferris, H. U., Coles, M., Lupas, A. N., & Hartmann, M. D. (2014). Crystallographic snapshot of the *Escherichia coli* EnvZ histidine kinase in an active conformation. *Journal of Structural Biology*, *186*(3), 376–379. https://doi.org/10.1016/j.jsb.2014.03.014

Ferris, H. U., Dunin-Horkawicz, S., Hornig, N., Hulko, M., Martin, J., Schultz, J. E., Zeth, K., Lupas, A. N., & Coles, M. (2012). Mechanism of Regulation of Receptor Histidine Kinases. *Structure*, *20*(1), 56–66. https://doi.org/10.1016/j.str.2011.11.014

Galperin, M. Y. (2006). Structural Classification of Bacterial Response Regulators: Diversity of Output Domains and Domain Combinations. *Journal of Bacteriology*, *188*(12), 4169–4182. https://doi.org/10.1128/jb.01887-05

Gao, R., Bouillet, S., & Stock, A. M. (2019). Structural Basis of Response Regulator Function. *Annual Review of Microbiology*, *73*(Volume 73, 2019), 175–197. https://doi.org/10.1146/annurev-micro-020518-115931

Ghose, D. A., Przydzial, K. E., Mahoney, E. M., Keating, A. E., & Laub, M. T. (2023). Marginal specificity in protein interactions constrains evolution of a paralogous family. *Proceedings of the National Academy of Sciences*, *120*(18), e2221163120. https://doi.org/10.1073/pnas.2221163120

Goswami, M., Wilke, K. E., & Carlson, E. E. (2017). Rational Design of Selective Adenine-Based Scaffolds for Inactivation of Bacterial Histidine Kinases. *Journal of Medicinal Chemistry*, *60*(19), 8170–8182. https://doi.org/10.1021/acs.jmedchem.7b01066

Groisman, E. A., Duprey, A., & Choi, J. (2021). How the PhoP/PhoQ System Controls Virulence and Mg2+ Homeostasis: Lessons in Signal Transduction, Pathogenesis, Physiology, and Evolution. *Microbiology and Molecular Biology Reviews*, *85*(3), 10.1128/mmbr.00176-20. https://doi.org/10.1128/mmbr.00176-20

Gumerov, V. M., Ulrich, L. E., & Zhulin, I. B. (2024). MiST 4.0: A new release of the microbial signal transduction database, now with a metagenomic component. *Nucleic Acids Research*, *52*(D1), D647–D653. https://doi.org/10.1093/nar/gkad847

Gushchin, I., Aleksenko, V. A., Orekhov, P., Goncharov, I. M., Nazarenko, V. V., Semenov, O., Remeeva, A., & Gordeliy, V. (2021). Nitrate- and Nitrite-Sensing Histidine Kinases: Function, Structure, and Natural Diversity. *International Journal of Molecular Sciences*, *22*(11), 5933. https://doi.org/10.3390/ijms22115933

Gushchin, I., & Gordeliy, V. (2018). Transmembrane Signal Transduction in Two-Component Systems: Piston, Scissoring, or Helical Rotation? *BioEssays*, *40*(2), 1700197. https://doi.org/10.1002/bies.201700197

Gushchin, I., Orekhov, P., Melnikov, I., Polovinkin, V., Yuzhakova, A., & Gordeliy, V. (2020). Sensor Histidine Kinase NarQ Activates via Helical Rotation, Diagonal Scissoring, and Eventually Piston-Like Shifts. *International Journal of Molecular Sciences*, *21*(9), 3110. https://doi.org/10.3390/ijms21093110

Hatstat, A. K., Kormos, R., Xu, V., & DeGrado, W. F. (2024). *A designed Zn2+ sensor domain transmits binding information to transmembrane histidine kinases* (p. 2024.10.30.621206). bioRxiv. https://doi.org/10.1101/2024.10.30.621206

Ishii, E., & Eguchi, Y. (2021). Diversity in Sensing and Signaling of Bacterial Sensor Histidine Kinases. *Biomolecules*, *11*(10), Article 10. https://doi.org/10.3390/biom11101524

Joshi, S. H.-N., Jenkins, C., Ulaeto, D., & Gorochowski, T. E. (2024). Accelerating Genetic Sensor Development, Scale-up, and Deployment Using Synthetic Biology. *BioDesign Research*, *6*, 0037. https://doi.org/10.34133/bdr.0037

Kabbara, S., Hérivaux, A., Dugé de Bernonville, T., Courdavault, V., Clastre, M., Gastebois, A., Osman, M., Hamze, M., Cock, J. M., Schaap, P., & Papon, N. (2019). Diversity and Evolution of Sensor Histidine Kinases in Eukaryotes. *Genome Biology and Evolution*, *11*(1), 86–108. https://doi.org/10.1093/gbe/evy213

Karan, R., Singla-Pareek, S. L., & Pareek, A. (2009). Histidine kinase and response regulator genes as they relate to salinity tolerance in rice. *Functional & Integrative Genomics*, *9*(3), 411–417. https://doi.org/10.1007/s10142-009-0119-x

Karniol, B., & Vierstra, R. D. (2004). The HWE Histidine Kinases, a New Family of Bacterial Two-Component Sensor Kinases with Potentially Diverse Roles in Environmental Signaling. *Journal of Bacteriology*, *186*(2), 445–453. https://doi.org/10.1128/jb.186.2.445-453.2004

Korycinski, M., Albrecht, R., Ursinus, A., Hartmann, M. D., Coles, M., Martin, J., Dunin-Horkawicz, S., & Lupas, A. N. (2015). STAC—A New Domain Associated with Transmembrane Solute Transport and Two-Component Signal Transduction Systems. *Journal of Molecular Biology*, *427*(20), 3327–3339. https://doi.org/10.1016/j.jmb.2015.08.017

Landry, B. P., Palanki, R., Dyulgyarov, N., Hartsough, L. A., & Tabor, J. J. (2018). Phosphatase activity tunes two-component system sensor detection threshold. *Nature Communications*, *9*(1), 1433. https://doi.org/10.1038/s41467-018-03929-y

Lazar, J. T., & Tabor, J. J. (2021). Bacterial two-component systems as sensors for synthetic biology applications. *Current Opinion in Systems Biology*, *28*, 100398. https://doi.org/10.1016/j.coisb.2021.100398

Leeuwenhoek, A. V. (1667). Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: Concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Philosophical Transactions of the Royal Society of London*, *12*(133), 821–831. https://doi.org/10.1098/rstl.1677.0003

Lemmin, T., Soto, C. S., Clinthorne, G., DeGrado, W. F., & Peraro, M. D. (2013). Assembly of the Transmembrane Domain of E. coli PhoQ Histidine Kinase: Implications for Signal Transduction from Molecular Simulations. *PLOS Computational Biology*, *9*(1), e1002878. https://doi.org/10.1371/journal.pcbi.1002878

Liang, H., Deng, X., Bosscher, M., Ji, Q., Jensen, M. P., & He, C. (2013). Engineering Bacterial Two-Component System PmrA/PmrB to Sense Lanthanide Ions. *Journal of the American Chemical Society*, *135*(6), 2037–2039. https://doi.org/10.1021/ja312032c

Liu, Y., Rose, J., Huang, S., Hu, Y., Wu, Q., Wang, D., Li, C., Liu, M., Zhou, P., & Jiang, L. (2017). A pH-gated conformational switch regulates the phosphatase activity of bifunctional HisKA-family histidine kinases. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-02310-9

Lori, C., Kaczmarczyk, A., de Jong, I., & Jenal, U. (2018). A Single-Domain Response Regulator Functions as an Integrating Hub To Coordinate General Stress Response and Development in Alphaproteobacteria. *mBio*, *9*(3), e00809-18. https://doi.org/10.1128/mBio.00809-18

Matilla, M. A., Velando, F., Martín-Mora, D., Monteagudo-Cascales, E., & Krell, T. (2022). A catalogue of signal molecules that interact with sensor kinases, chemoreceptors and transcriptional regulators. *FEMS Microbiology Reviews*, *46*(1), fuab043. https://doi.org/10.1093/femsre/fuab043

McClune, C. J., Alvarez-Buylla, A., Voigt, C. A., & Laub, M. T. (2019). Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature*, *574*(7780), 702–706. https://doi.org/10.1038/s41586-019-1639-8

Meier, S. S. M., Multamäki, E., Ranzani, A. T., Takala, H., & Möglich, A. (2024). Leveraging the histidine kinase-phosphatase duality to sculpt two-component signaling. *Nature Communications*, *15*(1), 4876. https://doi.org/10.1038/s41467-024-49251-8

Möglich, A., Ayers, R. A., & Moffat, K. (2009). Design and Signaling Mechanism of Light-Regulated Histidine Kinases. *Journal of Molecular Biology*, *385*(5), 1433–1444. https://doi.org/10.1016/j.jmb.2008.12.017

Ninfa, A. J., & Magasanik, B. (1986). Covalent modification of the glnG product, NRI, by the glnL product, NRII, regulates the transcription of the glnALG operon in Escherichia coli. *Proceedings of the National Academy of Sciences*, *83*(16), 5909–5913. https://doi.org/10.1073/pnas.83.16.5909

Ohlendorf, R., Schumacher, C. H., Richter, F., & Möglich, A. (2016). Library-Aided Probing of Linker Determinants in Hybrid Photoreceptors. *ACS Synthetic Biology*, *5*(10), 1117–1126. https://doi.org/10.1021/acssynbio.6b00028

Olivieri, F. A., Burastero, O., Drusin, S. I., Defelipe, L. A., Wetzler, D. E., Turjanski, A., & Marti, M. (2020). Conformational and Reaction Dynamic Coupling in Histidine Kinases: Insights from Hybrid QM/MM Simulations. *Journal of Chemical Information and Modeling*, *60*(2), 833–842. https://doi.org/10.1021/acs.jcim.9b00806

Ortega, Á., Zhulin, I. B., & Krell, T. (2017). Sensory Repertoire of Bacterial Chemoreceptors. *Microbiology and Molecular Biology Reviews*, *81*(4), e00033-17, e00033-17. https://doi.org/10.1128/MMBR.00033-17

Park, H., Joachimiak, M. P., Jungbluth, S. P., Yang, Z., Riehl, W. J., Canon, R. S., Arkin, A. P., & Dehal, P. S. (2023). A bacterial sensor taxonomy across earth ecosystems for machine learning applications. *mSystems*, *9*(1), e00026-23. https://doi.org/10.1128/msystems.00026-23

Podust, L. M., Ioanoviciu, A., & Ortiz de Montellano, P. R. (2008). 2.3 Å X-ray Structure of the Heme-Bound GAF Domain of Sensory Histidine Kinase DosT of Mycobacterium tuberculosis. *Biochemistry*, *47*(47), 12523–12531. https://doi.org/10.1021/bi8012356

Rapun-Araiz, B., Haag, A. F., De Cesare, V., Gil, C., Dorado-Morales, P., Penades, J. R., & Lasa, I. (2020). Systematic Reconstruction of the Complete Two-Component Sensorial Network in Staphylococcus aureus. *mSystems*, *5*(4), 10.1128/msystems.00511-20. https://doi.org/10.1128/msystems.00511-20

Rottinghaus, A. G., Amrofell, M. B., & Moon, T. S. (2020). Biosensing in Smart Engineered Probiotics. *Biotechnology Journal*, *15*(10), 1900319. https://doi.org/10.1002/biot.201900319

Sankhe, G. D., Dixit, N. M., & Saini, D. K. (2018). Activation of Bacterial Histidine Kinases: Insights into the Kinetics of the cis Autophosphorylation Mechanism. *mSphere*, *3*(3), 10.1128/msphere.00111-18. https://doi.org/10.1128/msphere.00111-18

Schmidl, S. R., Ekness, F., Sofjan, K., Daeffler, K. N.-M., Brink, K. R., Landry, B. P., Gerhardt, K. P., Dyulgyarov, N., Sheth, R. U., & Tabor, J. J. (2019). Rewiring bacterial two-component systems by modular DNA-binding domain swapping. *Nature Chemical Biology*, *15*(7), 690–698. https://doi.org/10.1038/s41589-019-0286-6

Schmidt, N. W., Grigoryan, G., & DeGrado, W. F. (2017). The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: Application to understand signaling in histidine kinases. *Protein Science*, *26*(3), 414–435. https://doi.org/10.1002/pro.3095

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell*, *133*(6), 1043–1054. https://doi.org/10.1016/j.cell.2008.04.040

Stock, A. M., Robinson, V. L., & Goudreau, P. N. (2000). Two-Component Signal Transduction. *Annual Review of Biochemistry*, *69*(Volume 69, 2000), 183–215. https://doi.org/10.1146/annurev.biochem.69.1.183

Taylor, B. L., & Zhulin, I. B. (1999). PAS Domains: Internal Sensors of Oxygen, Redox Potential, and Light. *Microbiology and Molecular Biology Reviews*, *63*(2), 479–506. https://doi.org/10.1128/mmbr.63.2.479-506.1999

Trajtenberg, F., Graña, M., Ruétalo, N., Botti, H., & Buschiazzo, A. (2010). Structural and Enzymatic Insights into the ATP Binding and Autophosphorylation Mechanism of a Sensor Histidine Kinase *. *Journal of Biological Chemistry*, *285*(32), 24892–24903. https://doi.org/10.1074/jbc.M110.147843

Ueno, T. B., Johnson, R. A., & Boon, E. M. (2015). Optimized assay for the quantification of histidine kinase autophosphorylation. *Biochemical and Biophysical Research Communications*, *465*(3), 331–337. https://doi.org/10.1016/j.bbrc.2015.07.121

Uhl, M. A., & Miller, J. F. (1996). Integration of multiple domains in a two-component sensor protein: The Bordetella pertussis BvgAS phosphorelay. *The EMBO Journal*, *15*(5), 1028–1036. https://doi.org/10.1002/j.1460-2075.1996.tb00440.x

Ulrich, L. E., & Zhulin, I. B. (2010). The MiST2 database: A comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Research*, *38*(suppl_1), D401–D407. https://doi.org/10.1093/nar/gkp940

Upadhyay, A. A., Fleetwood, A. D., Adebali, O., Finn, R. D., & Zhulin, I. B. (2016). Cache Domains That are Homologous to, but Different from PAS Domains Comprise the Largest Superfamily of Extracellular Sensors in Prokaryotes. *PLOS Computational Biology*, *12*(4), e1004862. https://doi.org/10.1371/journal.pcbi.1004862

Utsumi, R., Brissette, R. E., Rampersaud, A., Forst, S. A., Oosawa, K., & Inouye, M. (1989). Activation of Bacterial Porin Gene Expression by a Chimeric Signal Transducer in Response to Aspartate. *Science*, *245*(4923), 1246–1249. https://doi.org/10.1126/science.2476847

Velikova, N., Fulle, S., Manso, A. S., Mechkarska, M., Finn, P., Conlon, J. M., Oggioni, M. R., Wells, J. M., & Marina, A. (2016). Putative histidine kinase inhibitors with antibacterial effect against multi-drug resistant clinical isolates identified by in vitro and in silico screens. *Scientific Reports*, *6*(1), 26085. https://doi.org/10.1038/srep26085

Wang, F.-F., & Qian, W. (2019). The roles of histidine kinases in sensing host plant and cell–cell communication signal in a phytopathogenic bacterium. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1767), 20180311. https://doi.org/10.1098/rstb.2018.0311

Wang, J., & Childers, W. S. (2022). The Future Potential of Biosensors to Investigate the Gut-Brain Axis. *Frontiers in Bioengineering and Biotechnology*, *9*. https://doi.org/10.3389/fbioe.2021.826479

Wang, X., Vu, A., Lee, K., & Dahlquist, F. W. (2012). CheA-Receptor Interaction Sites in Bacterial Chemotaxis. *Journal of Molecular Biology*, *422*(2), 282–290. https://doi.org/10.1016/j.jmb.2012.05.023

Weerasuriya, S., Schneider, B. M., & Manson, M. D. (1998). Chimeric Chemoreceptors in Escherichia coli: Signaling Properties of Tar-Tap and Tap-Tar Hybrids. *Journal of Bacteriology*, *180*(4), 914–920. https://doi.org/10.1128/jb.180.4.914-920.1998

West, A. H., & Stock, A. M. (2001). Histidine kinases and response regulator proteins in two-component signaling systems. *Trends in Biochemical Sciences*, *26*(6), 369–376. https://doi.org/10.1016/S0968-0004(01)01852-7

Winski, A., Ludwiczak, J., Orlowska, M., Madaj, R., Kaminski, K., & Dunin-Horkawicz, S. (2024). AlphaFold2 captures the conformational landscape of the HAMP signaling domain. *Protein Science*, *33*(1), e4846. https://doi.org/10.1002/pro.4846

Wuichet, K., Cantwell, B. J., & Zhulin, I. B. (2010). Evolution and phyletic distribution of two-component signal transduction systems. *Current Opinion in Microbiology*, *13*(2), 219–225. https://doi.org/10.1016/j.mib.2009.12.011

Yoshida, T., Phadtare, S., & Inouye, M. (2007). Functional and Structural Characterization of EnvZ, an Osmosensing Histidine Kinase of E. coli. In *Methods in Enzymology* (Vol. 423, pp. 184–202). Elsevier. https://doi.org/10.1016/S0076-6879(07)23008-3

Zhang, Z., & Hendrickson, W. A. (2010). Structural Characterization of the Predominant Family of Histidine Kinase Sensor Domains. *Journal of Molecular Biology*, *400*(3), 335–353. https://doi.org/10.1016/j.jmb.2010.04.049

Zhu, Y., & Inouye, M. (2003). Analysis of the Role of the EnvZ Linker Region in Signal Transduction Using a Chimeric Tar/EnvZ Receptor Protein, Tez1 *. *Journal of Biological Chemistry*, *278*(25), 22812–22819. https://doi.org/10.1074/jbc.M300916200

Zschiedrich, C. P., Keidel, V., & Szurmant, H. (2016). Molecular Mechanisms of Two-Component Signal Transduction. *Journal of Molecular Biology*, *428*(19), 3752–3775. https://doi.org/10.1016/j.jmb.2016.08.003

2. DEGENERATE DROPSYNTH FOR SIMULTANEOUS ASSEMBLY OF DIVERSE GENE LIBRARIES AND LOCAL DESIGNED MUTANTS

2.1 Contributions

I was the primary contributor to the annotation and domain definition of genes, and the selection of genes was a collaborative process between me and Calin Plesa. Calin Plesa was the primary contributor, with contributions from Nora C. Kearns, to designing the DropSynth and degenerate oligo designs. The oligo amplification and processing, the emulsion assembly, and the suppression PCR were done jointly by me and Samuel R. Hinton, with assistance from Kyra A. Lindley; jointly done by myself and Samuel R. Hinton were the Golden Gate assemblies, transformations, and assembly barcode sequencing. Calin Plesa and Samuel R. Hinton were the main contributors to the plasmid design of pHKGG1. Calin Plesa was the main contributor to the error model and post-sequencing analysis. Calin Plesa, myself, and Samuel R. Hinton were contributors to the writing of this chapter, with me and Calin Plesa being the main editors and revisors of the text.

2.2 Introduction

Protein engineering seeks to design proteins with novel properties and functions by determining the sequences corresponding to a particular targeted function or property. Its potential applications are immense and span from drug development and diagnostics to biofuel production and environmental remediation. One important aspect of protein engineering is the creation of chimeric or hybrid proteins (Koide, 2009; Lin & Liu, 2016; Maervoet & and Briers, 2017; Yu et al., 2015). These proteins, especially prevalent in cellular signaling (Gordley et al., 2016), are constructed by combining different modules to create new or enhanced functionalities. This is particularly useful in the development of biomass saccharification (Punt et al., 2011), cellular signaling (Gordley et al., 2016), and complex biosynthesis pathways (Menzella & Reeves, 2007) to name a few. Central to the success of these chimeric proteins are the design and engineering of fusion points and linkers. The choice of fusion point can significantly impact the function of the resulting chimeric protein, as it can influence the spatial arrangement of the protein domains and their ability to interact with each other and with other molecules (Vymětal et al., 2022; Yu et al., 2015). Linkers, on the other hand, are connectors that provide flexibility

and space between the domains, enabling proper folding and independent functioning (Klein et al., 2014; Patel et al., 2022). Their characteristics (length, composition, and conformation) significantly affect the activity, stability, and solubility of the chimeric protein (Nielsen et al., 2016; Patel et al., 2022). Thus, the strategic selection of fusion points and linkers is a key element in the field of protein engineering.

Although most protein engineering efforts in this area have focused on single chimeric proteins, an increasingly desirable approach would be to decipher the rules for engineering entire protein families rather than individual proteins. Towards this end a potential Design-Build-Test-Learn strategy could consist of: (1) designing large amounts of diverse relevant hybrids through metagenomic mining or rational computational approaches, (2) assembling large libraries of specifically designed variants spanning many diverse genes, (3) functionally characterizing the library using a multiplexed functional assay, (4) feeding the resulting data into computational or machine learning (ML) models which can discern the underlying patterns affecting functionality, (5) repeating this process using computational or ML generated variants and feeding the results back into the model until some target threshold for accuracy is achieved. This manuscript tackles the second step in this process and demonstrates how to build large libraries of designed local mutant variants for many diverse genes.

Previous methods to generate fusion or linker variants are unable to generate designed variants for a diverse library of gene fusions. Methods utilizing nuclease mediated truncation such as SHIPREC (Sieber et al., 2001) and ITCHY (Ostermeier et al., 1999) create many fusions at random points. Although PCR primer-based methods such as PATCHY (Ohlendorf et al., 2016) allow control over the fusion points, they are only applicable to single genes and do not work with diverse gene libraries, due to the lack of conserved sequence in the priming region among different library members.

We previously introduced a multiplex gene synthesis method, DropSynth 2.0, and demonstrated that it is capable of assembling 1,536 genes with a median 501 bp in length in a single reaction with 64% coverage and a median 25% percent perfects at the amino acid level (Sidore et al., 2020) (~15% at nucleotide level). The percentage perfects at the DNA or amino acid level is defined as the fraction of total barcodes for each designed sequence (with at least 100 barcodes) that contain no mutations, or synonymous mutations as well, in the case of amino acid perfects. Although this approach increased the assembly scale significantly, in the context of

protein engineering, given the immense number of variants possible, creating a large number of variants for many diverse genes would quickly saturate the 1,536 genes possible in a typical reaction. Scaling would require the use of many separate libraries, which increases reagent and labour cost. As such, we sought to see if multiple variants for the same parent sequence could be synthesized within each droplet. In this approach, DropSynth proceeds exactly as before, Fig. 1a, with oligos processed to expose a 12-nt single stranded barcoded overhang, which is then hybridized and ligated on to corresponding barcoded beads, followed by emulsification into droplets, and emulsion polymerase cycling assembly (ePCA).

The degenerate DropSynth approach leverages the fact that any additional oligo targeted to a barcode which contains the same ePCA overhangs as the gene will also participate in the assembly reaction (Fig. 1b) and create a variant of the gene in a programmable manner. This approach provides several advantages. First, the marginal cost of each additional variant is only the cost of an additional oligo as all other reagents remain unchanged. This results in an estimated cost of $0.20 per 1kbp variant, assuming $0.11 per oligo and a degeneracy level of 8. This represents a 75.9% reduction in per-variant cost compared to assembling these using DropSynth 2.0 when oligos for conserved regions are (redundantly) ordered multiple times, and a 53.3% reduction when variable and conserved regions are split into separate oligo subpools and selectively remixed before assembly. Second, this provides a simple path towards much higher scales without larger sets of barcoded beads. While this study focuses on variants at the end of each gene, modifying only the last oligo, this approach can be readily adapted to create internal defined variants (App. Fig. A1), combinatorial variants (App. Fig. A2), or even entirely different full-length genes using completely orthogonal overlaps. While background-isolation from complex pools is critical to prevent cross-hybridization and proper assembly, small numbers of genes can be successfully assembled together with minimal screening for othogonality (Borovkov et al., 2010).

We apply the degenerate DropSynth approach to the creation of chimeric sensor histidine kinases (SHK) as proof of concept. SHKs are one of the most abundant protein families found in nature with millions of receptors that can sense a wide variety of stimuli including small molecules, light, pH, metal ions, and osmotic pressure (Cai & Inouye, 2002; Chakraborty et al., 2010; Hirose et al., 2008; Matilla et al., 2022; Xu et al., 2020). These homodimeric modular proteins typically contain an extracellular sensing domain, transmembrane domains, and

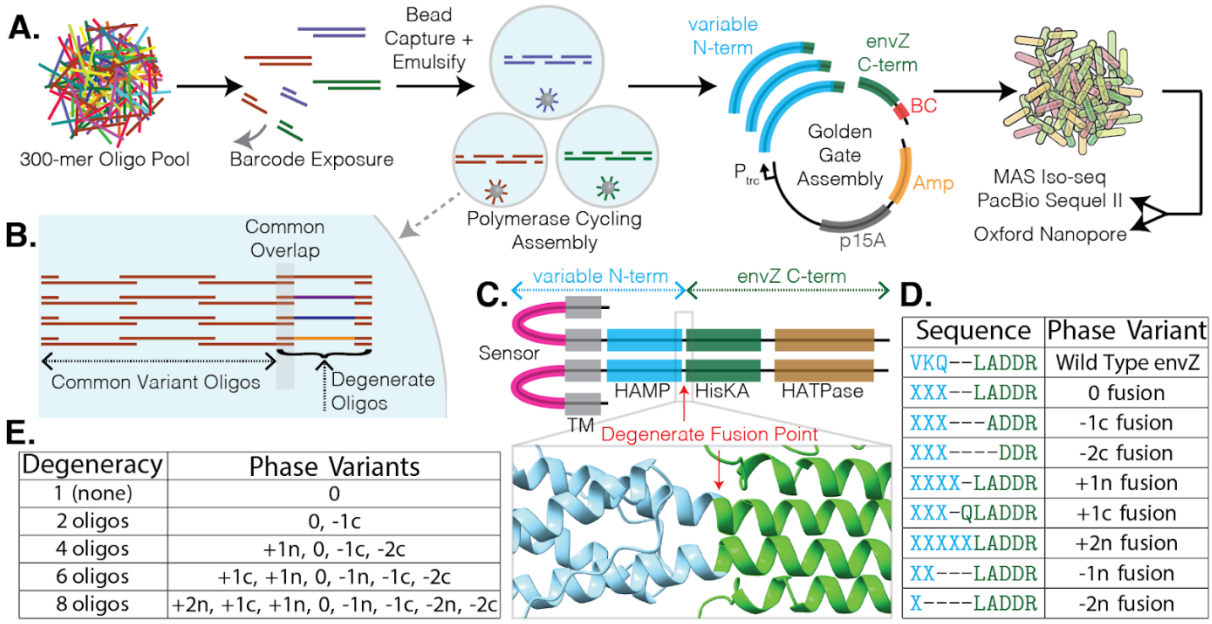
signaling domains which help propagate the activation signal to a kinase domain (Fig. 1c) (Bhate et al., 2015). Phosphorylation of a histidine residue is transferred to the aspartate residue on a specific response regulator protein which can then dimerize and activate transcription of a target promoter (Buschiazzo & Trajtenberg, 2019).

Despite their abundance, activating ligands are only known for a small number of SHKs since individual characterization is a slow and laborious process (Matilla et al., 2022). Large scale characterization and deorphanization of this family could be achieved if the modular sensory domains could be swapped onto a well characterized kinase domain to allow for multiplexed testing of many receptors (Bi et al., 2016; Jung et al., 2018). Many challenges exist in such an approach, as detailed elsewhere (Ganesh et al., 2019; Lazar & Tabor, 2021; Schmidl, 2019; Wang et al., 2013). One issue is the selection of a fusion point due to the diversity and uncertainty in the exact domain boundaries as detected by HMMs or other methods, which makes it difficult to create functional chimeras, even if the signal transduction mechanism of both halves is the same. This complicates placing the upstream (helix) portion into the proper phase orientation with the downstream (helix) portion (Airola et al., 2010; Kaur et al., 2014; Parkinson, 2010; Stewart & Chen, 2010; Wang et al., 2014). In this work we focus on building chimeric fusions in a region just below the HAMP signaling domain, a homodimeric four alpha-helix parallel coiled-coil region. We use degenerate DropSynth to create many phase variants for each sensor domain through the controlled addition or subtraction of amino acid residues on either the C or N terminal fragments of each chimera as shown in Fig. 1d. Any variant can be made as it is programmatically encoded on a corresponding assembly oligo, with the only requirements being that the variant sequence length can fit onto the corresponding oligo and be placed outside the common overlap region.

As a proof of concept, in each library, we tested the assembly of between 1 and 8 different phase variants for each gene, as shown in Fig. 1e. In other words, in the same reaction, some droplets would only assemble one gene with no additional variants (degeneracy of 1), while at the other extreme droplets would assemble 8 different variants of the same gene. This allowed us to control for inter-reaction variability from factors such as the DNA recovery from the emulsions, PCR amplification, and processing yields. We designed two sets of proteins based on their length. One with 1,530 proteins to be assembled with 4x 300-mer oligos, and another with 1,531 proteins to be assembled with 5x 300-mer oligos. We made two different codon

libraries for each set, for a total of four libraries with 6,122 total genes distributed among them. We previously demonstrated that using multiple codon variants enhances the likelihood of successful assembly, likely due to differential effects of oligo synthesis, amplification, and assembly processes on each codon variant. With the extra degenerate variants added, the four libraries encoded for a total of 10,862 proteins and 21,724 genes with distribution of degeneracy levels shown in App. Fig. A3. In the two 4-oligo libraries, we distributed the microbead barcodes relatively uniformly among the different degeneracy levels (median 292 microbead barcodes, s.d. 22), while in the 5-oligo libraries over half of the microbead barcodes had no degeneracy (level 1) (App. Fig. A4), to ensure sufficient statistics for the longer length. The use of 300-mer oligos allowed us to reach assembly lengths of 1 kbp with 5 oligos, effectively doubling the length demonstrated previously with DropSynth 2.0 (Sidore et al., 2020).

Figure 1 (next page). Degenerate DropSynth for rational fusion point engineering. a. Overview of the DropSynth protocol and cloning scheme. b. Within each barcoded microbead droplet, multiple versions of the last oligo are added. These degenerate oligos share a common overlap and can participate in the assembly reaction, with each encoding a different mutant variant. c. Domain architecture of a typical sensor histidine kinase. We assembled SHKs with diverse sensory and HAMP domains each linked to the C-terminal portion of the well characterized histidine kinase EnvZ. Fusion points were generated below the HAMP domain. (bottom insert) AlphaFold2 predicted structure around the fusion site of a NarX-EnvZ chimera. Proper function of these chimeric sensors requires the correct phase orientation between the fused alpha helices. d. By adding or removing residues from the N-terminal or C-terminal fragments, the relative phase orientation between the alpha helices on either side of the fusion point is changed. e. We tested assembly reactions where genes contained between 1 (no degeneracy) and 8 degenerate oligos encoding different phase variants to assess its impact on the gene assembly process.



Degeneracy	Phase Variants
1 (none)	0
2 oligos	0, -1c
4 oligos	+1n, 0, -1c, -2c
6 oligos	+1c, +1n, 0, -1n, -1c, -2c
8 oligos	+2n, +1c, +1n, 0, -1n, -1c, -2n, -2c

Sequence	Phase Variant
VKQ--LADDR	Wild Type envZ
XXX--LADDR	0 fusion
XXX---ADDR	-1c fusion
XXX----DDR	-2c fusion
XXXX-LADDR	+1n fusion
XXX-QLADDR	+1c fusion
XXXXXLADDR	+2n fusion
XX---LADDR	-1n fusion
X----LADDR	-2n fusion

2.3 Materials and Methods

2.3.1 Gene Design

All amino acid sequences (1,127,577 in total) containing a histidine kinase domain were obtained from UniProt release 2021_02. This dataset was then loaded into HMMER (version 3.2.1) (Eddy, 2011) and the domains were annotated for each sequence using the Pfam HMMs (version 33.1). Phobius (version 1.01) (Käll et al., 2004) and TMHMM-py (version 1.3.1) (Krogh et al., 2001) were both used to define transmembrane regions (TMs); only TMs with an overlap consensus of at least nine residues were kept, with the boundaries determined by Phobius being used. Only proteins with 2 TM domains and a HAMP domain were kept, leaving 126,611 proteins. These were further filtered into proteins where the length from the N-terminal to the end of the HAMP domain could fit into a 4x or 5x 300-mer oligo DropSynth assembly. Sequences are provided in Supplementary Data file 1.

2.3.2 DropSynth Oligo Design

The DropSynth oligos were designed using a series of custom scripts available at (https://github.com/PlesaLab/DropSynth_code_2023). These scripts were significantly optimized compared to older versions of the design scripts. Briefly, some of the changes include a switch to a “recipe”-based workflow with all parameters in a single file and the use of Lattice-

Automation's seqfold python library for minimum free energy structure calculations, instead of unafold (hybrid-ss-min; see App. Fig. A16 for a comparison). We implemented a programmable database for handling all restriction enzyme sites required, the ability to split as many genes as necessary in the first step with subsequent (384x, 1536x) library allocation, and a virtual assembly and translation to verify oligo designs. In order to offset barcoded microbead effects, we provide the option to do microbead barcode reversal between libraries. We improved codon optimization with lower split failures through the use of several hardcoded rules and added the ability to require certain sequences (controls) in each library. We added support for single oligo processing for very small genes and DNA (non-protein) constructs. We improved oligo junction length handling, with genes that fail due to length placed into a special file for input into higher oligo splits. As before, each for the four subpools was given a set of unique 15-mer subpool amplification primers, as shown in App. Table A2.

2.3.3 Degenerate Oligo Design

We created an R script to create all necessary degenerate oligos. Briefly, we initially designed DropSynth oligos using the +2 N fusion variants, since all other variants are as long or shorter. This ensured that all variants could fit on the last oligo. All DropSynth oligos were loaded and the payload sequence between the BtsI sites was determined. The overlap sequence between the last and second-to-last gene fragments was determined. We then determined if the degenerate mutation could be made based on the distance between the end cloning site (GACGTGAGACC) and the end of the overlap. Since the overlap sequence could not be modified, degenerate oligos were designed only if there was sufficient length after the overlap to implement the desired mutation. If sufficient, the script made degenerate oligos by selectively removing or mutating codons for each phase variant and level of degeneracy (Fig. 1d,e). If codons were removed, the overall length of the oligo was maintained by adding random bases into the padding region between the end cloning site and assembly amplification reverse primer site (skpp504R), checking to make sure no illegal restriction sites were introduced. If codons were changed, the padding was left unchanged, but the new sequence was still screened for illegal restriction sites. All successful degenerate oligo designs were combined together with the full DropSynth oligo set. All oligo sequences are provided in Supplementary Data file 2.

2.3.4 Oligo Amplification and Processing

Oligo designs were ordered as part of a pool of 58,500 300-mer oligos from Twist Bioscience. Processing followed the same protocol as detailed previously. Briefly, the OLS pool was resuspended to a concentration of 19 ng/µL. Subpools were amplified using 18 - 20 cycles (as first determined using qPCR) with Kapa HiFi. Bulk amplification of each subpool was then carried out using 8 - 11 cycles (determined by qPCR) with 0.5 to 1 ng of template. Between 7 - 9 ug of bulk amplified DNA was put into the nt.BspQI nick processing, with a yield ranging from 4.2 - 6.2 µg (corresponding to a range of molar yields of 52 - 75%).

2.3.5 Emulsion Assembly and Suppression PCR

Briefly, the DropSynth reaction was carried out using 1.3 µg of processed DNA. After emulsion breaking with chloroform, the correct length assemblies were (blind) size-selected with an agarose gel. DNA was extracted from gel slices with an NEB Monarch DNA Gel Extraction Kit and eluted in 30 µL. Of this, 1 µL was used as template in a suppression PCR reaction (first on a qPCR) using 25 - 28 cycles. These reactions were cleaned up, quantified by Qubit, and 0.36 pmol was subsequently used in each Golden Gate reaction as described below.

2.3.6 Plasmid Design of pHKGG1

Plasmid pSR348 containing the complete NarX SHK under the LacI-=inducible promoter was received from Addgene (#124713) and sequence verified using full plasmid Oxford Nanopore sequencing. Golden Gate Assembly was used to change the antibiotic selection maker from spectinomycin resistance to carbenicillin creating the plasmid pSR348_Carb. Several clones were isolated, and their plasmids were extracted and sequence verified in the same manner as pSR348. After swapping the selection marker, we replaced the NarX CDS with the wild-type EnvZ SHK (amplified from E. *coli* MG1655) to generate the plasmid EnvZ_pSR348_Carb. Finally, site-directed mutagenesis was used to remove a KpnI restriction site for backup cloning purposes creating the final plasmid renamed pHKGG1. A complete plasmid map and all primers used to generate pHKGG1 are found in the supporting information (App. Fig. A17) and App. Table A1. All transformations after each cloning step were done in 10-Beta cells (NEB) unless otherwise noted.

2.3.7 Golden Gate Assembly

A three-fragment Golden Gate assembly was used to clone our DropSynth-generated libraries. The first fragment (A) of the assembly consisted of a variable segment composed of

SHK sensory domains produced from the degenerate DropSynth protocol. The overhang CATA was used at the start, where the final adenine base is the start of the ATG codon of the gene. The overhang GACG was used at the end, where GAC encodes residue D232 in the EnvZ protein. The second fragment (B) contained a conserved portion of the EnvZ gene corresponding to amino acid residues D232-G450 and a 24 basepair quasi-randomer barcode region (NNBBDDBBVVHHDDBBVVHNDDNN) downstream of the stop codon. The final fragment (C) was derived from pHKGG1 and contained the p15A Ori, Amp selection marker, lac repressor, and P_{trc} inducible promoter. The fragment (B) containing the conserved portion of EnvZ and the 24 bp barcode along with the pHKGG1 backbone were both generated via PCR, and their respective primers are tabulated in App. Table A1. The sequences of the three Golden Gate fragments are provided in App. Table A4.

For each Degenerate DropSynth library, four Golden Gate reactions were run using the following temperature program: 1) Incubate at 37°C for 20 hours 2) Heat inactivation of both enzymes at 80°C for 20 mins 3) Final hold at 12°C. The molar ratios used were 0.18 pmol for FrgC (backbone), 0.36 pmol for FrgB (conserved region of EnvZ and barcode), and 0.36 pmol for FrgA (degenerate DropSynth synthesized library) for a total of 470 ng of DNA in the reaction. The reagent volumes were 2.5 µL of 10x T4 DNA Ligase Buffer (NEB), 2.5 µL of 10 mM ATP (NEB), 0.25 µL T4 DNA Ligase (NEB), 0.75 µL BsaI-HFv2 (NEB), and water to complete the remaining volume to a total of 25 µL. Assemblies were then pooled and cleaned using the Monarch PCR and DNA Cleanup Kit (NEB) and then drop dialyzed on a 0.05 um membrane filter (Sigma Millipore) for a minimum of 30 minutes. Purified DNA was used as the input for two electroporations (Bio-Rad MicroPulser), which were then combined and plated. Serial dilutions were used to calculate the total number of CFUs and are listed in App. Table A3.

2.3.8 Assembly Barcode Sequencing (MAS ISO-seq) and Analysis

Each library was PCR amplified with one of the HK_PB_0#_FWD+REV primer pairs using Q5 DNA polymerase and10 ng of template (miniprepped plasmid) in a 50 µL reaction for 11-12 cycles. To submit samples for MAS ISO-seq, the four libraries were mixed into a single 30µL sample containing 1.25ng/µL per library (5ng/µL total) and submitted to the UOregon GC3F core for library preparation and sequencing. Briefly, PacBio MAS ISO-Seq for 10x Single Cell 3' kit (102-659-600) was used to generate arrays for sequencing on a Sequel II instrument producing 120.20 Gbases of unique molecular data in 6.14 million raw reads. These samples

were also sequenced by Plasmidsaurus using Oxford Nanopore sequencing with R10.4.1 flowcells, v14 chemistry, and basecalled with Guppy 6.5.7. This produced between 10,592,215 and 12,245,627 reads for each sample. All sequencing data are available from the NIH sequencing read archive (SRA) under the BioProject PRJNA1049019.

For the MAS ISO-seq data, skera (0.1.0) was used to split the MAS arrays within 1,268,200 CCS reads, producing 15,907,686 split segments. Lima was then used to demultiplex the split segments, resulting in 2,740,434 to 4,826,815 reads per library. For both MAS ISO-seq and Oxford Nanopore data, a custom Python script was used to first identify the constant regions flanking the barcode (GTCGCTGCCGAACAGC-24N-AGGAGAAGAGCGCACG), allowing up to 3 mismatches. Then each read was scanned for the presence of the NdeI (CATATG) site at the start codon and the conserved region in the vector immediately flanking the cloning site (GACGACCGCACGCTGCTG, which corresponds to residues 232-237 (DDRTLL) of envZ). The script outputs this variable region and each associated barcode, as well as the barcodes counts. Barcode counts were input into Starcode (1.4) for collapse with a distance of 1 using the sphere algorithm. A consensus call was made for each barcode using a simple majority call. Genes were aligned to their closest parent sequence using minimap2, with k-mers set to 10. To assign reads to degenerate variants, the last 16 bp of each variable region was taken. Perfect matches to designs were taken as is, for the remainder, we calculated the Levenshtein distance between each 16 bp end sequence and all degenerate variants for that parent sequence. The read was assigned based on the smallest Levenshtein distance. In case of a tie (rare), a random assignment was made. All subsequent analysis and plots were carried out in R (4.3.1).

2.3.9 Error Model

The percentage of perfect oligos from oligo synthesis are given by

$$P_{olisynth} = \left(1 - E_{synth}\right)^{L_{oli} * N_{oli}} \quad (1)$$

where L_{oli} is the length of the oligos (300 nt), N_{oli} is the number of oligos used in the assembly (4 or 5), and E_{synth} is the estimated error rate of the synthesis process, discussed in text. The errors introduced by PCR amplification can be split into the amplification of the oligos and the suppression PCR of the assembled genes. The percentage of perfects after both PCRs should be given by

$$P_{PCR} = P_{oliPCR} * P_{supPCR} = (1 - E_{PCR})^{L_{oli} * C_{oli} * N_{oli}} * (1 - E_{PCR})^{L_{gene} * C_{gene}} \quad (2)$$

where L_{gene} is the length of the gene, C_{oli} is the number of PCR cycles used in oligos amplification, C_{gene} is the number of suppression PCR cycles used in gene amplification, and E_{PCR} is the error rate of the polymerase, discussed in text. We model the ePCA process using

$$P_{ePCA} = P_{endOligos} * P_{intOligos} = (1 - E_{ePCA})^{\left(\frac{L_{oli}}{N_{oli}} - \frac{L_{overlap}}{2}\right)*2} (1 - E_{ePCA})^{\left(\frac{L_{oli}}{N_{oli}} - L_{overlap}\right)(N_{oli}-2)} \quad (3)$$

where the first part accounts for the two oligos on the end while the latter part accounts for internal oligos. $L_{overlap}$ is the typical length of the overlap between fragments and E_{ePCA} is the error rate of the assembly process. The total estimated number of percentage perfects is then given by

$$P_{Total} = P_{ePCA} * P_{PCR} * P_{PCR} * P_{olisynth} \quad (4)$$

The only unknown parameter which is allowed to vary during fits is E_{ePCA}.

2.4 Results and Discussion

Increased levels of degenerate oligos have a minimal impact on the percentage perfects of assembled genes when different levels of degeneracy are uniformly represented. The four libraries were successfully assembled (Fig. 2ab) using the standard DropSynth protocol, Golden Gate cloned into a randomly barcoded expression vector (Schmidl, 2019), and long-read sequenced using both MAS ISO-seq (PacBio Sequel II) and Oxford Nanopore (Al'Khafaji et al., 2024). Among genes with at least 100 barcodes, we found a median of 16.6% and 17.3% DNA perfects for the 4 oligo library codon 1 and codon 2, respectively (Fig. 2c-top). We explored the impact of adding increasing numbers of degenerate oligos on the rates of perfects observed. Comparing the rates of genes from the no degeneracy set to those with 2, 4, 6, or 8 we found no statistically significant differences except for Codon 2 degeneracy level 6, which showed a weakly significant decrease (p=0.014). These results suggest that the presence of additional assembly oligos has no impact on the percentage of perfects.

For the five oligo assemblies, we see a median of 10.2% and 13.0% DNA perfects codon 1 and codon 2, respectively (Fig. 2c-bottom). For the codon 1 library, we see significant decreases for degeneracy levels of 4 (8.2%, p=0.02), 6 (7.6%, p=0.02), and 8 (6.4%, p=9E-6) relative to the no degeneracy case. A similar pattern is observed in the codon 2 library with significant decreases for degeneracy levels of 4 (11.2%, p=0.0004), 6 (11.6%, p=0.0004), and 8 (11.6%, p=0.03). These decreases are attributed to the imbalance in the final representation

between the variants from different degeneracy levels, which is exponentially increased by the suppression PCR amplification, as discussed later, and can be observed when plotting the perfects rate against the number of barcodes observed (App. Fig. A5). This effect is larger for the 5 oligo libraries due to their higher inequality among degeneracy levels (App. Fig. A4). Comparing the rates seen at the DNA level to those at the amino acid level, where synonymous mutations have been collapsed onto the parent sequence, we see a consistent 4.0% (s.d. 0.2%) higher rate at the protein level for 4 oligo assemblies and a 2.7% (s.d. 0.8%) difference for 5 oligo assemblies, as shown in App. Fig. A6. As expected, this suggests that the fraction of synonymous mutations decreases as length increases.

Deletions are the dominant errors found in assemblies. To quantify the relative frequencies of various error types per kilobase pair (kbp), we analyzed CIGAR alignment strings. Our analysis revealed comparable frequencies for insertions (median 1.47, s.d. 0.22) and single base deletions (median 1.44, s.d. 0.25), as shown in App. Fig. A7. Interestingly, mismatch frequencies were similar in the 4-oligo libraries (median 1.47) but significantly higher in the 5-oligo libraries (median 4.56). Across all libraries, the occurrence of multi-base deletions was notably high (median 5.75, s.d. 2.08). Delving deeper into deletion lengths, we observed that single base deletions were most common (App. Fig. A8a). However, the presence of a considerable number of longer deletions indicates a higher probability that any given deletion is part of a longer multi-base deletion rather than an isolated single-base deletion (App. Fig. A8b). The higher rates of deletions are consistent with previous reports using microarray derived oligos (Kosuri & Church, 2014). Further investigation is required to determine whether these long multi-base deletions primarily originate from oligo synthesis or the ePCA process. Additionally, the elevated mismatch rates in the 5-oligo libraries could potentially be attributed to the ePCA process itself, given that other parameters like the number of cycles are comparable to those in the 4-oligo libraries, and mismatch rates should not be highly dependent on sequencing depth. We note that these numbers were derived with minimap2 (Li, 2018), a general purpose aligner, as opposed to an exhaustive global alignment like Needleman-Wunsch (Needleman & Wunsch, 1970). Minimap2 is designed for speed and handles large-scale sequence data efficiently by using heuristics to quickly identify approximate matches, while Needleman-Wunsch is focused on accuracy, providing an exact alignment but at a much higher computational cost.

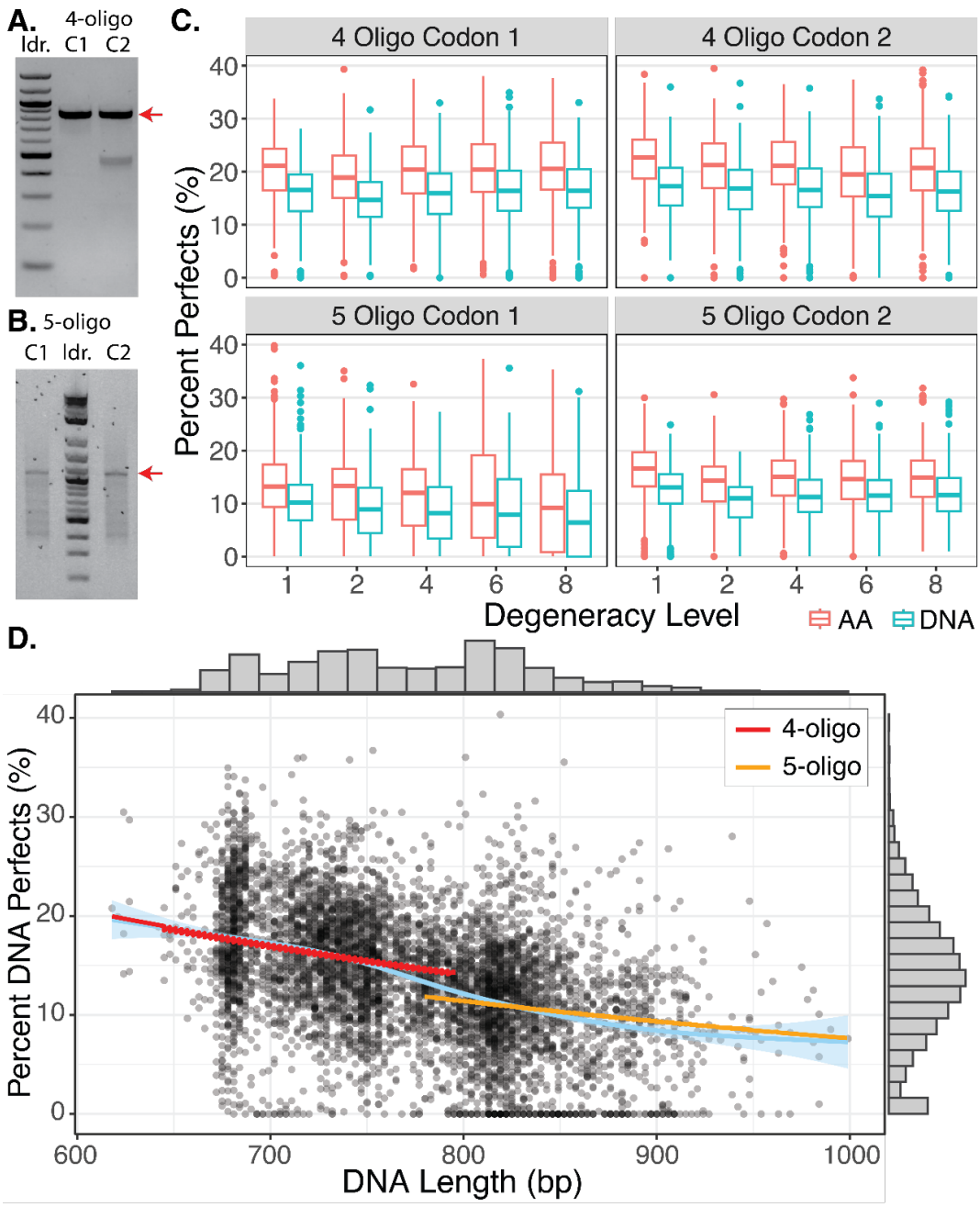


Figure 2. 4x and 5x 300-mer oligo assemblies. a.-b. Assembly of the two 4-oligo and 5-oligo libraries respectively. c. The percentage perfects observed for variants with at least 100 barcodes shown both at the AA level (collapsed on synonymous mutations) and DNA level. d. The percentage of DNA perfects as a function of length showing a consistent reduction from ~20% at 600 bp down to ~8% by 1 kbp. The blue line is a smoothed general additive model (GAM) fit with the confidence interval around the line displayed, while the colored lines are fits for 4 and 5 oligos with the model described in the text.

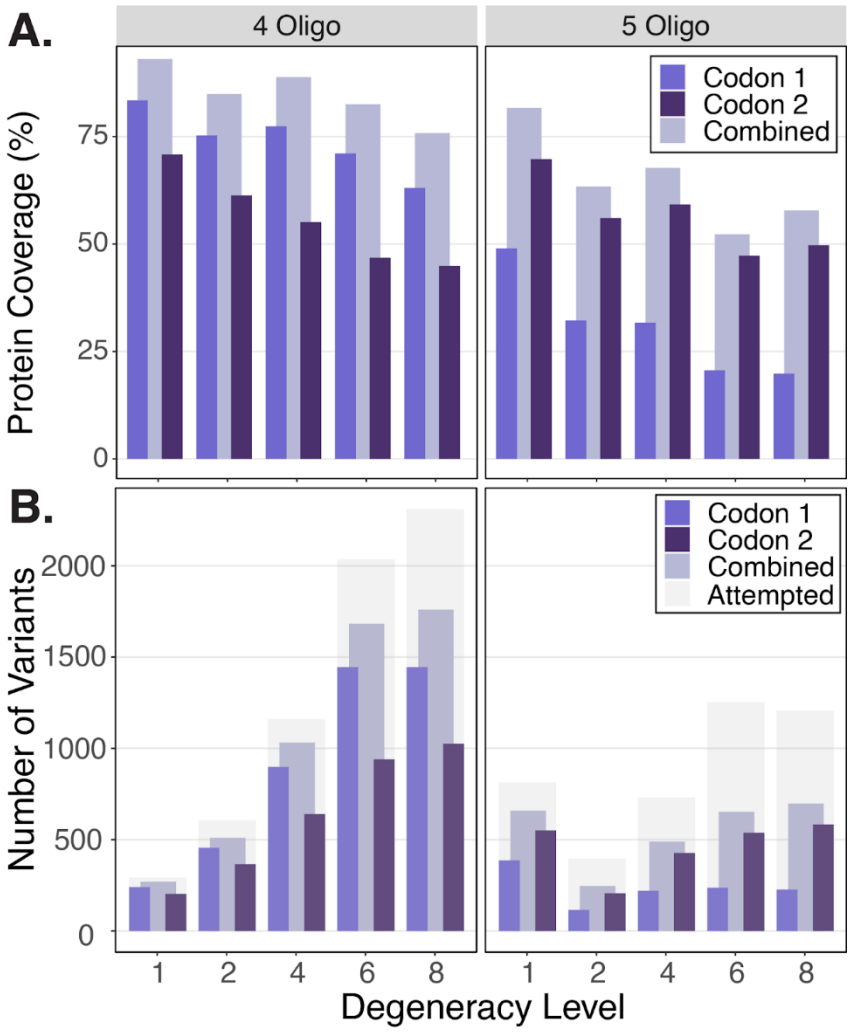
The rate of perfect assemblies correlates inversely with gene length, a trend we attribute to the propagation and combination of errors throughout the oligo synthesis, PCR amplification, and emulsion PCR (ePCA) assembly processes. This relationship is illustrated in Fig. 2d, where we plot the percentage of perfect assemblies for genes across all degeneracy levels as a function of their length. The 17 genes shorter than 650 bp show a median perfect assembly rate of 18.2% (s.d. 4.7%), while the 22 genes above 950 bp exhibit a rate of 8.2% (s.d. 5.2%). To understand this trend, we developed a simple model incorporating the error rates from oligo synthesis, PCR amplification, and ePCA assembly. With an error rate of 5.52E-6 errors per base per cycle (Hestand et al., 2016), Kapa HiFi amplification of the oligos and assembled genes has a relatively modest impact on the percentage of perfects. For example, Kapa HiFi based PCR amplification, performed over 27-28 cycles for oligos and 29 cycles for assembled genes, would theoretically allow 68% of 1 kbp genes to remain error-free post-PCR. For oligo synthesis with a 1 in 3000 bp error rate (based on vendor provided rates), we expect a 61% probability that all 5 oligos (300-mers) are perfect, which would reduce to 47% if the error rate increases to 1 in 2000 bp. This calculation, admittedly conservative, does not account for the selective pressure against propagation of errors in critical regions such as primers, restriction sites, and overlaps during the assembly process. By combining these factors, our model suggests an overall estimated perfect assembly rate of 41% for 1 kbp genes, far higher than the actual observed rate of 8.2%.

To further discern between different sources of error, we Nanopore sequenced a different 300-mer DropSynth library after oligo processing and bead capture, but immediately prior to the DropSynth reaction. We observe a median 25.9% perfect oligo sequences (266 bp length) as shown in App. Fig. A9, which represents a lower bound due to the error rate of nanopore sequencing. This is far lower than the 84.8% perfects predicted by our model for the 52 cycles of PCR amplification between oligo synthesis and Nanopore sequencing. This suggests that the actual error rates for PCR and oligo synthesis are higher than those used in our calculations.

Decreases in observed coverage are attributed to reduced representation at higher degeneracy levels due to suppression PCR amplification bias. We determined the coverage for all libraries and degeneracy levels, where coverage is defined by the number of variants for which at least one perfect protein sequence is observed. To improve coverage compared to DropSynth 2.0, we aimed to diminish PCR bias by lowering the total number of amplification cycles, thereby reducing its impact on library uniformity. We combined both PacBio and

Nanopore data to maximize the sequencing depth of our data, as a perfect sequence in a Nanopore read is likely to be a true perfect. For the 4 oligo libraries, we observe 84% and 71% coverage for a degeneracy level of 1, which reduces to 63% and 45% by degeneracy of 8, as shown in Fig. 3a. However, when the results from the two different codon version libraries are combined over their common protein sequences this improves to 93% for degeneracy of 1 and 76% by degeneracy level 8. Despite an average 0.69-fold decrease in percentage coverage, this reduction is modest compared to the 8-fold increase in scale brought about by higher degeneracy. In absolute numbers, the total count of observed genes rose from 244 and 207 at degeneracy level of 1 to 1448 and 1026 by degeneracy level of 8, due to the increased scale for the 4 oligo libraries, which have a roughly uniform distribution of microbead barcodes (median 292, s.d. 22) among different degeneracy levels (App. Fig. A4). For the 5 oligo libraries, we observe 49% and 70% coverage for a degeneracy of 1, dropping to 20% and 50% by degeneracy of 8. Combining over the two codon versions again raised coverage levels to 82% for a degeneracy of 1 and 58% for a degeneracy level of 8.

Figure 3 (next page). Protein coverage. The number of designed protein variants for which we see at least one perfect amino acid sequence. Coverage data are presented for each codon library, both individually and combined over the same proteins. a. The percentage observed relative to the number of variants designed. b. The absolute numbers of variants observed, contrasted against the total designed variants, which are depicted in light gray. Notably, the percent coverage decreases less than the increase in degeneracy scale, resulting in a net increase in the total number of successfully assembled protein variants.



We looked for trends which could explain this drop in coverage. In examining the fraction of barcodes observed versus the corresponding fraction of overall designs as a function of degeneracy level (App. Fig. A10), we observe an exponential decrease with increasing degeneracy. We still see this decay when these numbers are scaled by the degeneracy level to account for lower amounts of DNA (App. Fig. A11). We hypothesize that this trend results from the suppression PCR amplification post-assembly. Each barcoded bead in the DropSynth assembly has a limited DNA loading capacity, leading to a situation where the amount of DNA for each variant is inversely proportional to the degeneracy level. Consequently, a variant in a droplet with a degeneracy level of 8 will, on average, contain 8 times less assembled DNA than a variant with a degeneracy level of 1. This disparity is further exacerbated during suppression PCR, which significantly amplifies these initial differences, leading to variants in droplets with a

degeneracy level of 0 using more of the total sequencing reads, lowering the chances of observing some of the variants assembled with a degeneracy level of 8. We hypothesize this amplification bias results from differences in PCR efficiency due to the initial abundances and competition between the sequences. Suppression PCR amplification has a lower overall PCR efficiency than regular PCR due to the inverted terminal repeats in the priming region. A simple PCR amplification model fits the log transformed relationship between barcodes observed per variant versus expected variant concentration relationship quite well, with R^2 values ranging from 0.960 to 0.993 (App. Fig. A12). Here, we calculate the expected variant concentration based on the proportion of barcoded microbeads with a given degeneracy level to the total number of variants at that level.

In order to determine the acceptable range of degeneracy combinations, we analyzed the relationship between the fold-change in output versus the fold-change in input DNA (App. Fig. A12). A linear fit on a log-log scale yielded a slope of 1.59, indicating that to maintain less than a 10-fold difference in output DNA amounts, the input levels should not differ by more than 4-fold. For a 2-fold difference in input, the expected output difference would be limited to approximately 3-fold. This provides some practical guidance for designing degenerate DropSynth reactions and underscores the importance of maintaining a consistent degeneracy level across all variants when possible.

Looking ahead, if a full 1536x library is created with a uniform degeneracy level of 8, encoding 12,288 variants, we conservatively estimate that protein coverage would be at least ~8,000 for a single library and about 9,000 when combined over two different codon versions for 4 oligo assemblies, and roughly 6,000 for a single library and about 7,000 when combined over two codon versions for 5 oligos. These estimates are based on sufficient sequencing depth and are likely conservative, considering the potential improvements achievable with a consistent degeneracy across all library barcodes.

We determined the uniformity in the representation of the variants. The Gini coefficient, which quantifies distribution equality on a scale from 0 (perfect equality) to 1 (perfect inequality), was calculated for each degeneracy level in all four libraries (App. Fig. A14). This ranges from 0.71 to 0.88 (median 0.81, s.d. 0.06) for the 4 oligo libraries, with a slightly increasing trend as degeneracy increases. In contrast, the 5 oligo libraries exhibited higher Gini coefficients, ranging from 0.83 to 0.93 (median 0.87, s.d. 0.04), indicating less uniformity in

variant representation. These values are in line with previous observations (ranging from 0.69 to 0.94), suggesting the increased degeneracy has a minor effect relative to other factors such as PCR bias.

Barcode mapping of constructs over 500 bp requires long read sequencing, beyond what can be achieved with Illumina sequencing. We investigated the use of Oxford Nanopore sequencing as an alternative to PacBio MAS-Iso-seq. Despite Oxford Nanopore's inherently higher error rate in raw reads, previous studies have demonstrated that this can be significantly mitigated by collapsing reads onto barcodes (Karst et al., 2021; Zurek et al., 2020). To determine the consensus sequence for each barcode, we employed a read-based majority call approach. When we compared the percentage perfects obtained from each sequencing method, PacBio consistently outperformed Oxford Nanopore, with a median of 7.8% (s.d. 4.2%) higher percentage perfects, as depicted in App. Fig. A15. This comparison underlines the notable difference in error rates between the two sequencing platforms. It also implies that achieving comparable accuracy with Oxford Nanopore might necessitate more sophisticated consensus calling strategies, along with higher sequencing depths, to match the performance seen with PacBio.

This study underscores the efficacy of Degenerate DropSynth as a robust, scalable tool for protein engineering and synthetic biology, especially in the construction of large gene libraries of chimeric proteins. We demonstrate the capability to assemble up to eight distinct variants for each gene by isolating multiple overlap-compatible fragments into emulsion PCR (ePCA) droplets, utilizing barcoded microbeads. Beyond linker optimization and fusion proteins, this method is valuable in any scenario requiring the generation of variants within a specific region or across multiple regions of diverse genes. Applications include introducing diversity into regions of interest, such as catalytic sites across various homologs, engineering distant higher-order mutants for functional genomics, and the combinatorial creation of genetic circuits for synthetic biology, among others. In addition to functionally characterizing the sensor histidine kinases assembled in this study, future gene synthesis work will explore the scalability of this approach, which, we believe, has yet to reach its full potential. Potential strategies include increasing the number of variants per barcode beyond 8 while maintaining uniform degeneracy across the entire library, employing combinatorial assembly with multiple variable fragment positions, and integrating larger bead sets (exceeding 1536x) with degeneracy. Such strategies

could enable assembly of libraries with tens to hundreds of thousands of gene variants. Considering the observed 8% rate of perfect assemblies for genes of 1 kbp length, the use of longer lengths may require either error-correction or substantial oversampling in subsequent functional assays. In sum, Degenerate DropSynth significantly enhances the scope, length, and cost-efficiency of gene library construction, thereby facilitating the exploration and understanding of protein families through large-scale, programmable assembly.

2.5 Data Availability

The oligo and protein data underlying this chapter (Supplementary files) are available on figshare, at https://dx.doi.org/10.6084/m9.figshare.25091615 while sequencing data is available from the NIH sequencing read archive (SRA) under the BioProject PRJNA1049019.

2.6 Supplementary Data

Supplementary Data available at https://doi.org/10.6084/m9.figshare.25091615 . Referenced appendix tables and figures are found in the Appendix.

2.7 Funding

Burroughs Wellcome Fund [grant number 1018211 to C.P.]; National Institutes of Health T32 Molecular Biology and Biophysics Training Program [grant number 5T32GM007759 to A.S.H.].

2.8 Bridge

This chapter presents some of the considerations and initial bioinformatic methods for the design of the fusion phase variants for the chimeric SHKs. The next chapter, Plasmid Designs, focuses on the design and testing of genetic circuits for the characterization of these chimeras.

We thank the UOregon GC3F core staff, Doug Turnbull and Tina Arredondo for productive discussion on the use of MAS ISO-seq. Plasmid pSR348 was a gift from Jeffrey Tabor (Addgene plasmid # 124713; http://n2t.net/addgene:124713 ; RRID:Addgene_124713).

2.9 Degenerate DropSynth Bibliography

Airola, M. V., Watts, K. J., Bilwes, A. M., & Crane, B. R. (2010). Structure of Concatenated HAMP Domains Provides a Mechanism for Signal Transduction. *Structure*, *18*(4), 436–448. https://doi.org/10.1016/j.str.2010.01.013

Al'Khafaji, A. M., Smith, J. T., Garimella, K. V., Babadi, M., Popic, V., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M. A., Blaum, E. M., Day, A., Costello, M., Bowers, T., Gabriel, S., Banks, E., Philippakis, A. A., Boland, G. M., Blainey, P. C., &

Hacohen, N. (2024). High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, *42*(4), 582–586. https://doi.org/10.1038/s41587-023-01815-7

Bhate, M. P., Molnar, K. S., Goulian, M., & DeGrado, W. F. (2015). Signal Transduction in Histidine Kinases: Insights from New Structures. *Structure*, *23*(6), 981–994. https://doi.org/10.1016/j.str.2015.04.002

Bi, S., Pollard, A. M., Yang, Y., Jin, F., & Sourjik, V. (2016). Engineering Hybrid Chemotaxis Receptors in Bacteria. *ACS Synthetic Biology*, *5*(9), 989–1001. https://doi.org/10.1021/acssynbio.6b00053

Borovkov, A. Y., Loskutov, A. V., Robida, M. D., Day, K. M., Cano, J. A., Le Olson, T., Patel, H., Brown, K., Hunter, P. D., & Sykes, K. F. (2010). High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Research*, *38*(19), e180. https://doi.org/10.1093/nar/gkq677

Buschiazzo, A., & Trajtenberg, F. (2019). Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level? *Annual Review of Microbiology*, *73*(1), 507–528. https://doi.org/10.1146/annurev-micro-091018-054627

Cai, S. J., & Inouye, M. (2002). EnvZ-OmpR Interaction and Osmoregulation in Escherichia coli. *Journal of Biological Chemistry*, *277*(27), 24155–24161. https://doi.org/10.1074/jbc.M110715200

Chakraborty, S., Li, M., Chatterjee, C., Sivaraman, J., Leung, K. Y., & Mok, Y.-K. (2010). Temperature and Mg2+ Sensing by a Novel PhoP-PhoQ Two-component System for Regulation of Virulence in Edwardsiella tarda*. *Journal of Biological Chemistry*, *285*(50), 38876–38888. https://doi.org/10.1074/jbc.M110.179150

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Ganesh, I., Kim, T. W., Na, J.-G., Eom, G. T., & Hong, S. H. (2019). Engineering Escherichia coli to Sense Non-native Environmental Stimuli: Synthetic Chimera Two-component

Systems. *Biotechnology and Bioprocess Engineering*, *24*(1), 12–22. https://doi.org/10.1007/s12257-018-0252-2

Gordley, R. M., Bugaj, L. J., & Lim, W. A. (2016). Modular engineering of cellular signaling proteins and networks. *Current Opinion in Structural Biology*, *39*, 106–114. https://doi.org/10.1016/j.sbi.2016.06.012

Hestand, M. S., Houdt, J. V., Cristofoli, F., & Vermeesch, J. R. (2016). Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *784–785*, 39–45. https://doi.org/10.1016/j.mrfmmm.2016.01.003

Hirose, Y., Shimada, T., Narikawa, R., Katayama, M., & Ikeuchi, M. (2008). Cyanobacteriochrome CcaS is the green light receptor that induces the expression of phycobilisome linker protein. *Proceedings of the National Academy of Sciences*, *105*(28), 9528–9533. https://doi.org/10.1073/pnas.0801826105

Jung, K., Fabiani, F., Hoyer, E., & Lassak, J. (2018). Bacterial transmembrane signalling systems and their engineering for biosensing. *Open Biology*, *8*(4), 180023. https://doi.org/10.1098/rsob.180023

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, *338*(5), 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*, *18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

Kaur, H., Singh, S., Rathore, Y. S., Sharma, A., Furukawa, K., Hohmann, S., Gang, A., & Mondal, A. K. (2014). Differential Role of HAMP-like Linkers in Regulating the Functionality of the Group III Histidine Kinase DhNik1p *. *Journal of Biological Chemistry*, *289*(29), 20245–20258. https://doi.org/10.1074/jbc.M114.554303

Klein, J. S., Jiang, S., Galimidi, R. P., Keeffe, J. R., & Bjorkman, P. J. (2014). Design and characterization of structured protein linkers with differing flexibilities. *Protein Engineering, Design and Selection*, *27*(10), 325–330. https://doi.org/10.1093/protein/gzu043

Koide, S. (2009). Generation of new protein functions by nonhomologous combinations and rearrangements of domains and modules. *Current Opinion in Biotechnology*, *20*(4), 398–404. https://doi.org/10.1016/j.copbio.2009.07.007

Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: Technologies and applications. *Nature Methods*, *11*(5), 499–507. https://doi.org/10.1038/nmeth.2918

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Lazar, J. T., & Tabor, J. J. (2021). Bacterial two-component systems as sensors for synthetic biology applications. *Current Opinion in Systems Biology*, *28*, 100398. https://doi.org/10.1016/j.coisb.2021.100398

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lin, C.-Y., & Liu, J. C. (2016). Modular protein domains: An engineering approach toward functional biomaterials. *Current Opinion in Biotechnology*, *40*, 56–63. https://doi.org/10.1016/j.copbio.2016.02.011

Maervoet, V. E. T., & and Briers, Y. (2017). Synthetic biology of modular proteins. *Bioengineered*, *8*(3), 196–202. https://doi.org/10.1080/21655979.2016.1222993

Matilla, M. A., Velando, F., Martín-Mora, D., Monteagudo-Cascales, E., & Krell, T. (2022). A catalogue of signal molecules that interact with sensor kinases, chemoreceptors and transcriptional regulators. *FEMS Microbiology Reviews*, *46*(1), fuab043. https://doi.org/10.1093/femsre/fuab043

Menzella, H. G., & Reeves, C. D. (2007). Combinatorial biosynthesis for drug development. *Current Opinion in Microbiology*, *10*(3), 238–245. https://doi.org/10.1016/j.mib.2007.05.005

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nielsen, M. L., Isbrandt, T., Petersen, L. M., Mortensen, U. H., Andersen, M. R., Hoof, J. B., & Larsen, T. O. (2016). Linker Flexibility Facilitates Module Exchange in Fungal Hybrid PKS-NRPS Engineering. *PLOS ONE*, *11*(8), e0161199. https://doi.org/10.1371/journal.pone.0161199

Ohlendorf, R., Schumacher, C. H., Richter, F., & Möglich, A. (2016). Library-Aided Probing of Linker Determinants in Hybrid Photoreceptors. *ACS Synthetic Biology*, *5*(10), 1117–1126. https://doi.org/10.1021/acssynbio.6b00028

Ostermeier, M., Shim, J. H., & Benkovic, S. J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotechnology*, *17*(12), 1205–1209. https://doi.org/10.1038/70754

Parkinson, J. S. (2010). Signaling Mechanisms of HAMP Domains in Chemoreceptors and Sensor Kinases. *Annual Review of Microbiology*, *64*(Volume 64, 2010), 101–122. https://doi.org/10.1146/annurev.micro.112408.134215

Patel, D. K., Menon, D. V., Patel, D. H., & Dave, G. (2022). Linkers: A synergistic way for the synthesis of chimeric proteins. *Protein Expression and Purification*, *191*, 106012. https://doi.org/10.1016/j.pep.2021.106012

Punt, P. J., Levasseur, A., Visser, H., Wery, J., & Record, E. (2011). Fungal Protein Production: Design and Production of Chimeric Proteins. *Annual Review of Microbiology*, *65*(Volume 65, 2011), 57–69. https://doi.org/10.1146/annurev.micro.112408.134009

Schmidl, S. R. (2019). Rewiring bacterial two-component systems by modular DNA-binding domain swapping. *Nature Chemical Biology*, *15*, 14.

Sidore, A. M., Plesa, C., Samson, J. A., Lubock, N. B., & Kosuri, S. (2020). DropSynth 2.0: High-fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Research*, *48*(16), e95. https://doi.org/10.1093/nar/gkaa600

Sieber, V., Martinez, C. A., & Arnold, F. H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nature Biotechnology*, *19*(5), 456–460. https://doi.org/10.1038/88129

Stewart, V., & Chen, L.-L. (2010). The S Helix Mediates Signal Transmission as a HAMP Domain Coiled-Coil Extension in the NarX Nitrate Sensor from *Escherichia coli* K-12. *Journal of Bacteriology*, *192*(3), 734–745. https://doi.org/10.1128/JB.00172-09

Vymětal, J., Mertová, K., Boušová, K., Šulc, J., Tripsianes, K., & Vondrasek, J. (2022). Fusion of two unrelated protein domains in a chimera protein and its 3D prediction: Justification of the x-ray reference structures as a prediction benchmark. *Proteins: Structure, Function, and Bioinformatics*, *90*(12), 2067–2079. https://doi.org/10.1002/prot.26398

Wang, B., Barahona, M., Buck, M., & Schumacher, J. (2013). Rewiring cell signalling through chimaeric regulatory protein engineering. *Biochemical Society Transactions*, *41*(5), 1195–1200. https://doi.org/10.1042/BST20130138

Wang, B., Zhao, A., Novick, R. P., & Muir, T. W. (2014). Activation and Inhibition of the Receptor Histidine Kinase AgrC Occurs through Opposite Helical Transduction Motions. *Molecular Cell*, *53*(6), 929–940. https://doi.org/10.1016/j.molcel.2014.02.029

Xu, Y., Zhao, Z., Tong, W., Ding, Y., Liu, B., Shi, Y., Wang, J., Sun, S., Liu, M., Wang, Y., Qi, Q., Xian, M., & Zhao, G. (2020). An acid-tolerance response system protecting exponentially growing Escherichia coli. *Nature Communications*, *11*(1), 1496. https://doi.org/10.1038/s41467-020-15350-5

Yu, K., Liu, C., Kim, B.-G., & Lee, D.-Y. (2015). Synthetic fusion protein design and applications. *Biotechnology Advances*, *33*(1), 155–164. https://doi.org/10.1016/j.biotechadv.2014.11.005

Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A., & Hollfelder, F. (2020). UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nature Communications*, *11*(1), 6023. https://doi.org/10.1038/s41467-020-19687-9

3. PLASMID DESIGNS

3.1 Introduction

3.1.1 Response Regulator Circuit Goals

Sensor histidine kinases (SHKs) are one of the most widespread protein families in nature, recognized for their ability to detect and respond to a vast range of environmental signals—from small molecules and metal ions to pH and osmotic stress. Structurally, these homodimeric proteins contain discrete domains that perform distinct functions: an extracellular sensing domain to bind external stimuli, transmembrane segments for membrane localization, and internal signaling and kinase domains that relay activation through a phosphorylation cascade. When an external ligand binds, the conformational changes trigger the transfer of a phosphate from a conserved histidine residue in the SHK to an aspartate on a cognate response regulator (RR).

This, in turn, drives transcriptional changes, allowing the cell to adapt appropriately. Despite their pervasiveness, only a fraction of SHKs have been functionally characterized, as current methods for identifying their activating stimuli are slow and painstaking. A promising strategy for large-scale "deorphanization" of these receptors is to exploit their modular design: by swapping their extracellular sensing domain onto a single, well-characterized kinase scaffold, many receptors could be tested in parallel for their specific ligands.

With this aim in mind, we constructed a large library of chimeric SHKs (described in chapter 2) by engineering fusion boundaries just below a region called the HAMP signaling domain—a four-helix bundle involved in transmitting conformational changes from the sensing module to the kinase module. By designing these hybrid proteins, we hoped to generate diverse receptor variants that would activate transcription only upon exposure to their respective ligands. The chimeric SHKs resulting from our library can generally be sorted into three distinct categories:

(1) Non-functional receptors, which do not activate under any conditions, even when their cognate ligand is present.
(2) "Locked-on" receptors, which are always active regardless of ligand presence; and
(3) Functional receptors, which behave in the desired manner and only respond when their specific ligand is introduced.

To maximize the odds of finding functional SHKs, we needed a robust screening pipeline. A key challenge is to eliminate "locked-on" variants, which would otherwise confound downstream assays by always remaining active. Our strategy, therefore, was to build genetic circuits that could differentiate truly ligand-dependent activity from constitutive activation. We decided upon two complementary screening systems, with the goal of eventually integrating these into the genome of *E. coli*. Although they share similar core designs—using both a life-death selection and a backup fluorescence readout—each system serves a distinct purpose. One system focuses on culling the "locked-on" variants by killing any cells whose SHK derivatives are active without ligand stimulation. The other is dedicated to a more comprehensive characterization of ligand response, removing variants that show no response when ligand is introduced. In both cases, surviving cells should, in principle, harbor chimeric SHKs that function appropriately: off in the absence of ligand and activated on cue once a ligand is present.

By using a stringent life-death selection with fluorescence-based sorting as an alternative, these circuits allow for high-throughput selection and sorting of thousands of chimeric SHK candidates simultaneously. This chapter details the rational and technical steps behind constructing these dual selection circuits, as well as the ways in which they enable both the removal of problematic "locked-on" variants and the direct measurement of ligand-responsive activity.

Effectively, a variant going through this pipeline will end up in one of three boxes: from the locked-on assay, it will either end up eliminated due to being locked-on or activated by something in the minimal media or pass onto the next assay; from the characterization assay, the variant will either be characterized with its ligand being determined, or it will not be characterized, in which case either its activating ligand is not present in the chemical panel, or the variant is catalytically dead (Fig. 1).

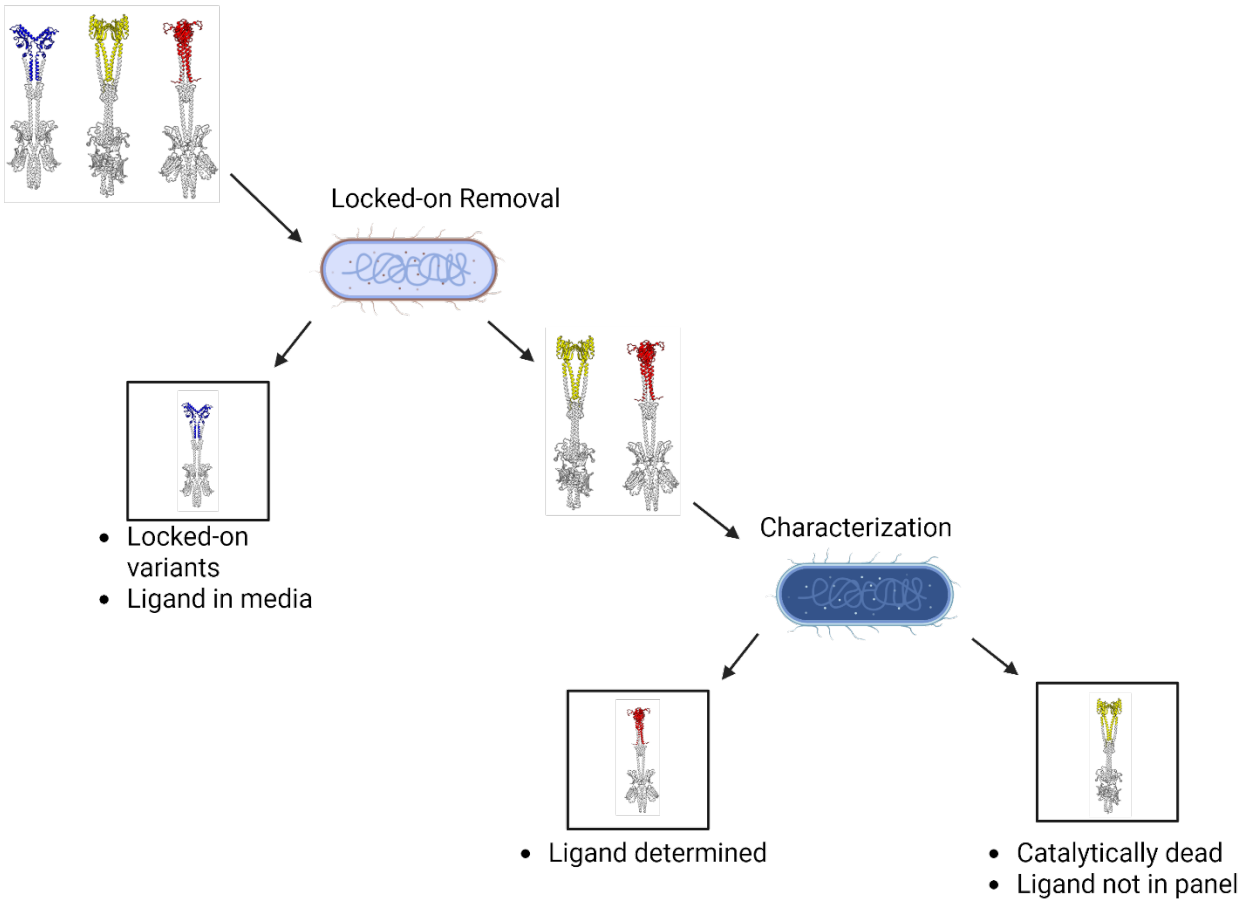


Figure 1 – Outcomes from Locked-On and Characterization Sorting Pipelines. The three possible outcomes from this full pipeline: sorted out for being locked-on or activated by something in the minimal media, characterized, or uncharacterized, with the uncharacterized variants either being catalytically dead or their activating ligand not being present in the screened chemicals.

3.1.2 Design Goals

3.1.2.1 Maximize Signal-to-Noise

A central requirement in constructing robust genetic circuits is to achieve a high signal-to-noise ratio (SNR). In the context of sensing and reporting on chimeric SHK function, signal refers to the output generated upon legitimate ligand-dependent activation, whereas noise includes any unwanted basal activity (Fig. 2)—either background expression when no ligand is present, cross talk with other SHKs, or nonspecific responses to irrelevant stimuli. High baseline noise makes it difficult to distinguish true positives from false signals. Minimizing noise (e.g., by using minimal media, optimizing promoter strength, ribosome binding site efficiency, and

expression levels) is crucial for creating a clear on/off distinction and improving the reliability of downstream screening.

3.1.2.2 Minimize Selection Ratio and Mutant Escape Frequency

In parallel with maintaining a clear on/off response, the selection system must reliably remove undesired variants. Two common pitfalls are:

- False Positives: Variants that display activity even without ligand (e.g., “locked-on”) or that respond nonspecifically.
- Escape Frequency: The proportion of cells harboring non-functional or spurious variants that still survive selection.

A strong selection system enforces a stringent kill-on-failure mechanism, ensuring only the most promising, ligand-responsive variants remain. Dual-antibiotic selection can bolster stringency and reduce escape frequency by requiring that cells pass two independent “tests” to survive.

Furthermore, selection markers can be titratable, allowing the strength of selection to be fine-tuned. By adjusting antibiotic concentrations, one can calibrate the stringency to match the specific screening goals—whether that’s a more permissive screen to preserve borderline variants or a very strict screen to enrich only the best performers. It is also critical to ensure these selection markers are absent from the base strain and any assay plasmids carrying the chimeric SHK, as overlapping antibiotic markers or unintentional resistance cassettes would undermine the specificity of the selection process.

3.1.2.3 Maximize Dynamic Range

An additional goal is to maximize the circuit’s dynamic range, this being defined as the fold-change between baseline output (no ligand) and induced output (with ligand). A large dynamic range enables quantitative evaluation of each chimeric SHK’s sensitivity and specificity. This is especially valuable when screening large libraries, as researchers can more easily differentiate between weak, moderate, and strong activation levels. Achieving a wide dynamic range often entails carefully balancing promoter strength, regulatory elements, and feedback loops so that the maximum induced signal is as high as possible while keeping the off state near background levels.

3.1.2.3 Additional Considerations

Beyond these primary goals, other elements are key for a successful screening platform:

- Genomic Integration: Placing the circuit in the chromosome rather than on a plasmid can mitigate variability due to copy number differences and improve stability of the selection system.
- Modularity: Building the kill genes, fluorescent markers, and promoters so that each component can be swapped or fine-tuned enables rapid iteration and customization.
- Metabolic Burden: Minimizing toxicity from high-level expression of circuit components is crucial; otherwise, growth or viability issues could skew the selection outcomes.

By carefully optimizing signal-to-noise, applying stringent selection markers (including titratable antibiotics) to reduce both false positives and escape frequency, and maximizing dynamic range, these circuits will more effectively isolate and characterize functional chimeric histidine kinases.

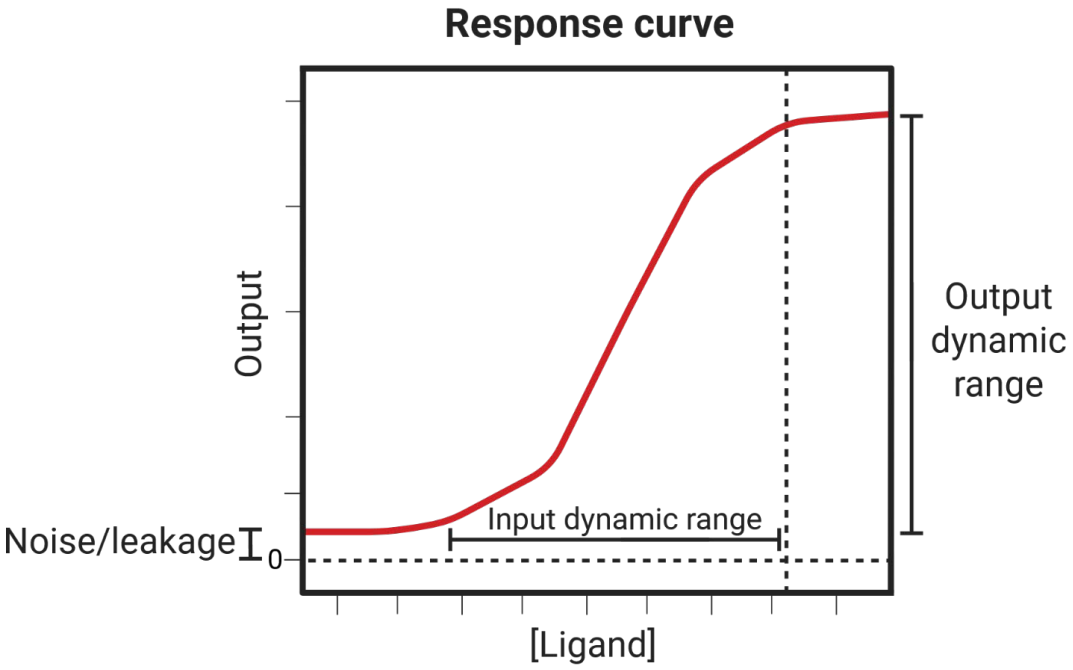


Figure 2 – Schematic of Signal-to-Noise and Dynamic Range Concepts. An example response curve of input versus output, highlighting the basal leakage/noise of a system when no ligand is present, the dynamic range (level of basal noise as the floor and maximum activation as the ceiling) of output, and the dynamic range for input (starting when signal increases above noise and ending when output signal is at a maximum).

3.1.3 Initial Response Regulator Plasmid

To enable fluorescence-based detection of EnvZ (or EnvZ-chimeric) activity, we utilized a plasmid construct originally developed in the lab of Jeffrey Tabor at Rice University, designated pSR40.29 (Schmidl et al., 2019). This plasmid (Fig. 3) is designed for measuring SHK signaling output by coupling EnvZ-dependent phosphorylation events to a fluorescent reporter gene. Because EnvZ is an osmosensing HK naturally found in E. coli, pSR40.29 has proven useful for assaying EnvZ function—and, by extension, the function of EnvZ-based chimeras—under controlled lab conditions.

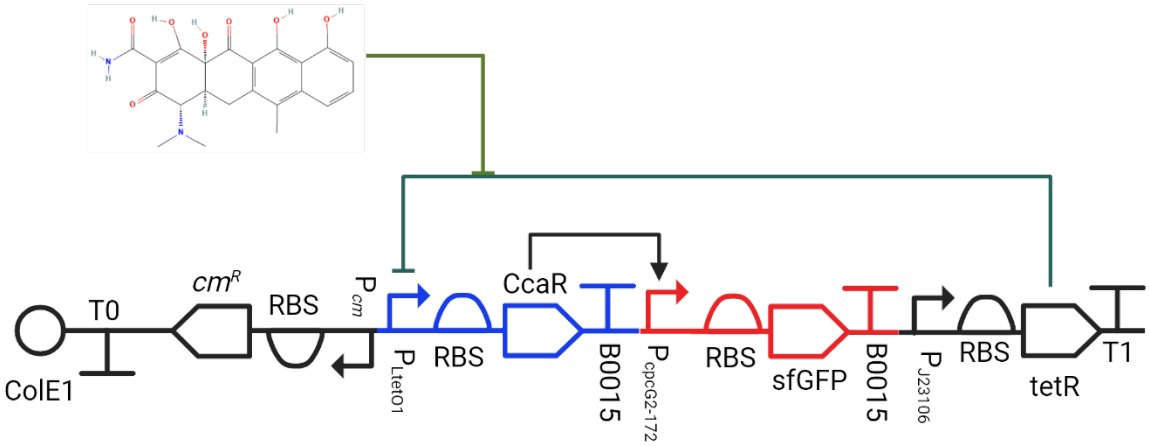


Figure 3 – Synthetic Biology Open Language (SBOL) Diagram of pSR40.29. This plasmid encodes a three-module transcriptional circuit regulating sfGFP expression. OmpR/CcaR, a chimeric response regulator, is under the control of the TetR-repressible promoter P_{LtetO1} and activates the output promoter $P_{cpcG2\text{-}172}$ upstream of sfGFP. TetR is constitutively expressed and represses P_{LtetO1}, forming a negative feedback loop. Anhydrotetracycline (aTc) relieves TetR repression, enabling sfGFP expression. SBOL glyphs represent genetic parts: arrows for promoters, semicircles for ribosome binding sites (RBSs), pointed rectangles for coding sequences, and T-bars for terminators. Colors are employed to indicate separate transcriptional units; blue: *ccaR*; red: *sfGFP*; black: *tetR* and backbone elements.

Built on a pBR322 derivative, pSR40.29 carries the pMB1 origin (part of the ColE1 origin family) and generally maintains 15–20 copies per cell. For antibiotic selection, it encodes

CmR (chloramphenicol resistance) under the CmR promoter. Critically, TetR is constitutively produced (driven by the J23106 promoter), providing inducible control of other plasmid elements via pLtetO-1.

The plasmid's core functionality arises from a chimeric response regulator (RR):

- The receiver domain is derived from OmpR, which becomes phosphorylated by EnvZ (or EnvZ-chimeras).
- The effector domain is taken from CcaR, a response regulator orthogonal to *E. coli* regulators.

This fusion RR is placed under the pLtetO-1 promoter, which is repressed by TetR unless the inducer anhydrotetracycline (aTc) is added. Once synthesized and phosphorylated, the CcaR effector domain can then drive transcription from its cognate promoter. In the default configuration, this RR-inducible output promoter (cpcG2-172) controls the expression of sfGFP (Superfolder GFP), enabling a straightforward, fluorescence-based readout of SHK activity.

3.1.3.1 Establishing Baseline Signal and Noise

We first assessed the signal-to-noise ratio of pSR40.29 by co-transforming it with a plasmid expressing the wild-type EnvZ into the *E. coli* strain BW29655 (a K-12 BW25113 derivative lacking both the ompR and envZ genes). This strain ensures that any observed output is solely attributed to the plasmid-borne RR and SHK. Because EnvZ is responsive to changes in osmolarity, we tested various sucrose concentrations as the osmolyte.

For all plate reader fluorescence assays, we used supplemented M9 minimal medium (1× M9 salts, 2 mM MgSO4, 0.1 mM CaCl2) containing 0.4% (wt/vol) glucose and 0.2% (wt/vol) casamino acids, along with appropriate inducers (such as aTc) and antibiotics (chloramphenicol for plasmid maintenance and any additional antibiotics if required). By monitoring sfGFP fluorescence under these defined conditions, we quantified both the maximum inducible signal (at high osmotic strength) and the basal (background) fluorescence in the absence of activating stimuli. These measurements allowed us to benchmark the plasmid's dynamic range, signal-to-noise ratio, and growth characteristics, setting a clear baseline for subsequent modifications or the testing of chimeric EnvZ variants (Fig. 4; Fig. 5).

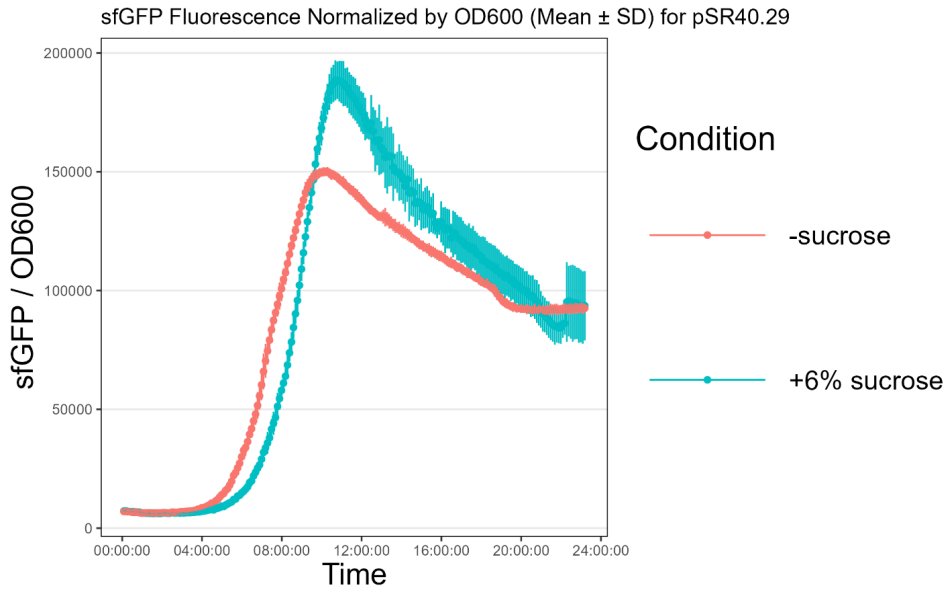


Figure 4 – Plate reader sfGFP fluorescence data of pSR40.29 in absence of sucrose or presence of 6% w/v sucrose. The max value for the system with ligand present is 150,000 AU, and the timeframe for maximum signal-to-noise ratio starts about ten hours into the experiment and continues for around four hours, with a peak dynamic range of 41,734 AU at eleven hours.

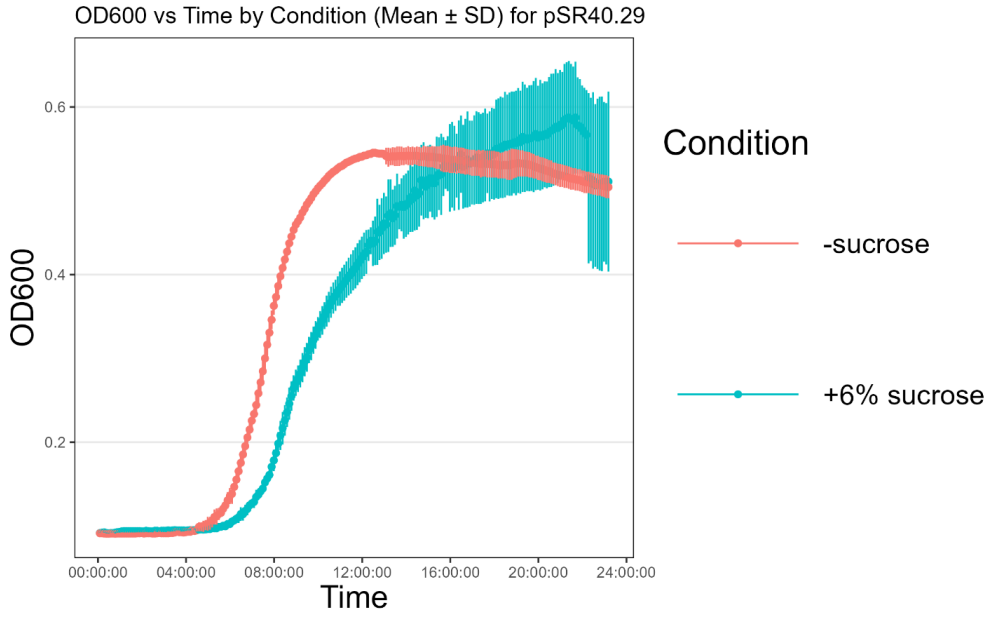


Figure 5 – Plate reader OD600 data of pSR40.29 in absence of sucrose or presence of 6% w/v sucrose. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly seven to eight hours before wells started to reach an OD600 of 0.2.

After confirming that pSR40.29 provides robust and measurable output in response to EnvZ activity, we turned our attention to designing and implementing our own circuit. The insights gleaned from this initial plasmid—especially regarding maximal output and basal leakage—guided our optimization steps for integrating additional genetic elements, implementing dual-antibiotic selection, and ultimately ensuring the high-fidelity detection of ligand-dependent histidine kinase signaling events.

3.2 Locked-On Sorting

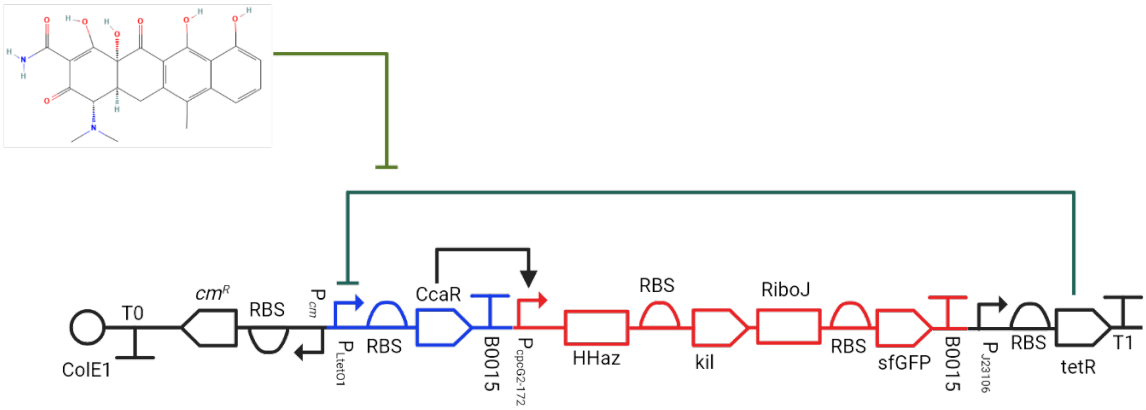
3.2.1 Locked-On Sorting – First General Design

3.2.1.1 pSR40.29-kil

One major challenge when screening chimeric histidine kinases (HKs) is dealing with "locked-on" variants—those that remain constitutively active regardless of ligand presence. To overcome this, we designed a circuit that couples SHK-mediated output to a life-death selection: cells expressing the SHK that is constitutively active (and thus driving the reporter promoter) are selectively eliminated under specific conditions.

Our initial circuit design was built from pSR40.29, adapting it into a new plasmid called pSR40.29-kil (Fig. 6). The key modification was to equip the plasmid with a kill gene (kil) under the same response regulator-inducible promoter (cpcG2-172) that controls sfGFP, thereby linking HK-dependent gene expression to a life-death mechanism.

Figure 6 (next page) – SBOL Diagram of pSR40-kil. This plasmid encodes a regulated system linking cell death (via *kil*) and sfGFP expression to the aTc-titratable activation of the CcaR response regulator. CcaR is under the control of a TetR-repressible promoter P_{LtetO1} and activates the output promoter $P_{cpcG2\text{-}172}$ upstream of an engineered transcript encoding a self-cleaving hammerhead ribozyme (HHaz), the toxin gene *kil*, and sfGFP. The *RiboJ* insulator separates *kil* and *sfGFP* to reduce crosstalk. TetR, expressed by the constitutive promoter P_{J23106}, represses P_{LtetO1}, and its repression is inhibited by aTc, enabling circuit activation. Colors are employed to indicate distinct transcriptional units; blue: *ccaR*; red: toxin and *sfGFP*; black: *tetR* and backbone elements.



3.2.1.1.1 Aptazyme Integration

We inserted a theophylline-responsive aptazyme sequence upstream of the kil gene. Aptazymes are engineered riboswitches that modulate gene expression through ligand-induced self-cleavage (Wieland & Hartig, 2008).

Crucially, in our design this aptazyme masks the ribosome binding site (RBS) for kil in the absence of theophylline, preventing basal translation of kil. Even low-level expression of kil can be detrimental to cell viability, so this mechanism was added to strictly prevent kil leakage in uninduced conditions. When theophylline is present, the aptazyme structure changes, exposing the RBS and allowing kil translation. This ensures kil is only expressed (and cells killed) if two conditions are met: (a) the cpcG2-172 promoter is activated by the SHK/RR system, and (b) theophylline is provided in the medium.

3.2.1.1.2 RiboJ Insulator

A RiboJ insulator sequence (Clifton et al., 2018) was placed between the kil gene and sfGFP. This prevents translational coupling and any unintended impact of kil expression on the fluorescent reporter, preserving a clean and independent readout.

3.2.1.1.3 Kil Protein

The Kil protein inhibits the essential cell division factor FtsZ (W. Chen et al., 2019). Without functional FtsZ, the bacterium cannot complete cell division, leading to elongation and eventual lysis. By placing kil under HK-dependent control, any cells harboring “locked-on” HKs (and thus perpetually activating the promoter) will express lethal kil—provided theophylline is added. The life-death selection involving the Kil protein was tested separately prior to its inclusion in this system by quantifying the number of colony-forming units in non-selective and selective conditions for three dilution series replicates, leading to a calculated escape frequency

for the Kil protein of 1:993,686. Thus, around one out of every one million cells surviving selection would contain escaped variants.

With pSR40.29-kil, we can thus perform:

- Life-death selection (in the presence of theophylline): only cells whose SHK is not locked-on will survive, since locked-on mutants driving continuous kil expression will die.
- Fluorescence-based sorting (in the absence of theophylline): aptazyme-mediated repression of kil prevents kill activity, allowing us to rely solely on sfGFP output for selection.

We validated this design by transforming pSR40.29-kil into E. coli BW29655 (lacking ompR and envZ) and measuring the circuit's output without theophylline, showing that sfGFP has clear separation between nine and thirteen hours (Fig. 7; Fig. 8), but with a low signal-to-noise ratio, with expectations of these parameters improving upon integration, as there would be less variability in copy number.

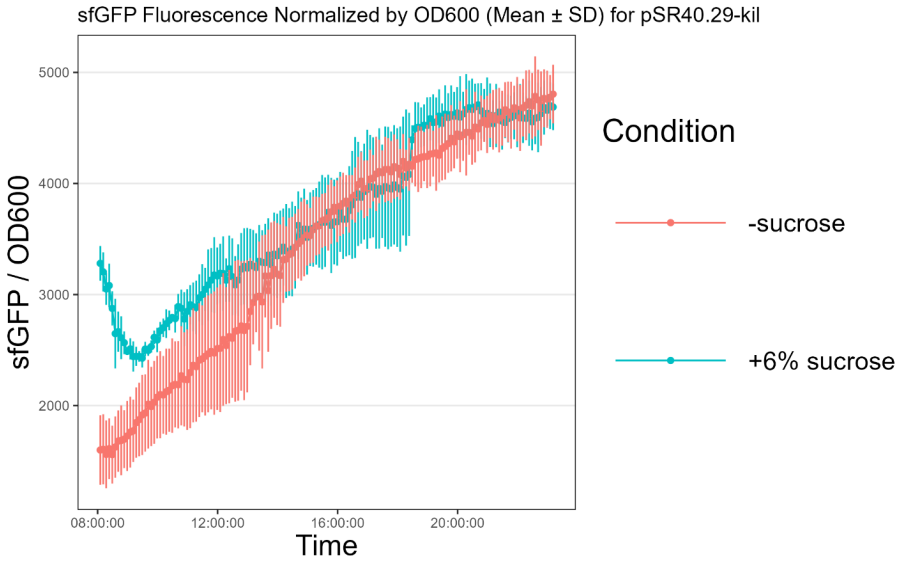


Figure 7 – Plate reader sfGFP fluorescence data of pSR40.29-kil in absence of sucrose or presence of 6% w/v sucrose with no theophylline present in either condition. The timeframe for maximum signal-to-noise ratio starts about nine hours into the experiment and continues for around four hours, with a peak dynamic range of 749 AU at nine hours, a max value for the system with ligand present during this timeframe around 3,000 AU, and a signal-to-noise ratio of 1.43.

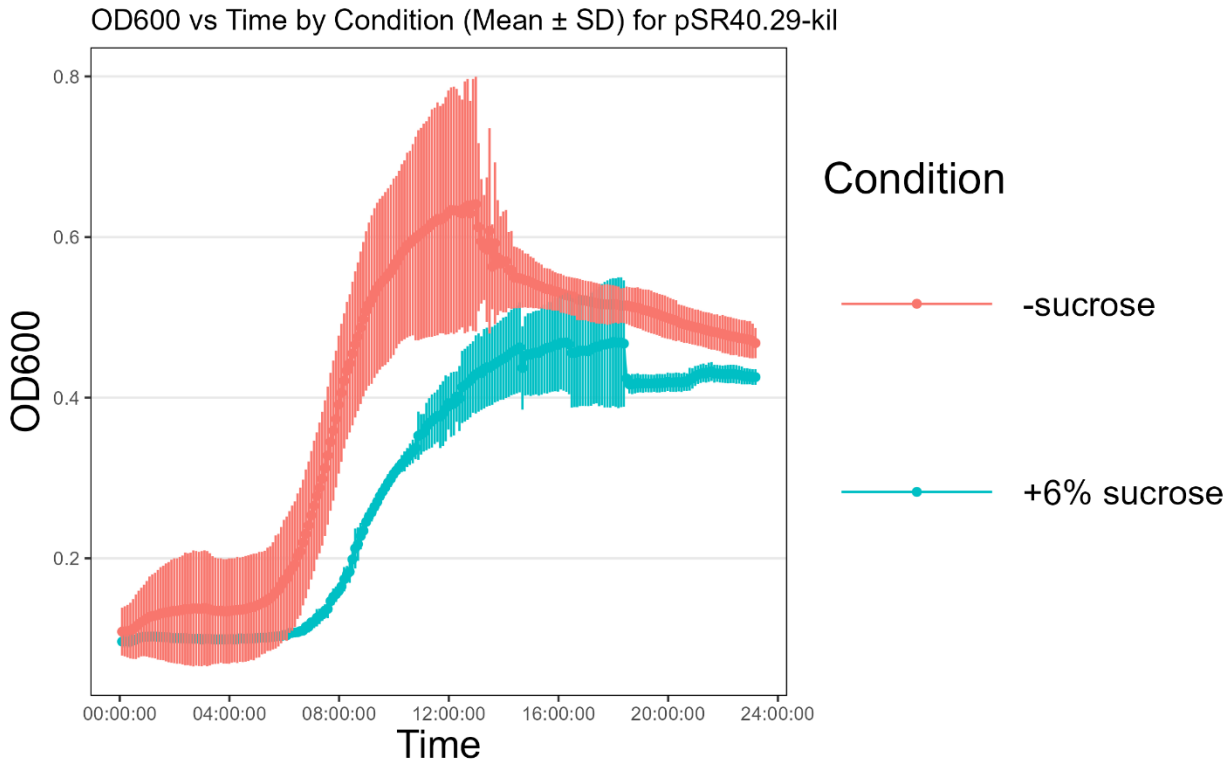


Figure 8 – Plate reader OD600 data of pSR40.29-kil in absence of sucrose or presence of 6% w/v sucrose with no theophylline present in either condition. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly seven to eight hours before wells started to reach an OD600 of 0.2.

3.2.1.2 pSpin INTEGRATE One-Plasmid System

Although plasmid-based circuits are convenient for rapid construction and testing, chromosomal integration offers more stable copy numbers and reduces variability due to plasmid copy fluctuations, with potential to improve signal-to-noise ratios. To achieve genomic integration, we employed the INTEGRATE system (pSL1521, or pSpin) developed by the Sternberg lab at Columbia University (Vo et al., 2021). This system uses a Tn7-like transposon from Vibrio cholerae guided by a Type I-F CRISPR–Cas mechanism for programmable, RNA-guided transposition.

3.2.1.2.1 Targeting the nth-ydgr Locus

We chose the nth-ydgr locus in E. coli as our integration site. Because it lies farther from the origin of replication, copy number variation is minimized during growth. Previous work had used it successfully (Urtecho et al., 2018).

A specific crRNA spacer was designed for this locus. After ordering and annealing single-stranded oligonucleotides, we employed Golden Gate assembly to insert the spacer into the pSL1521 vector, creating pSpin-nth (Fig. 9).

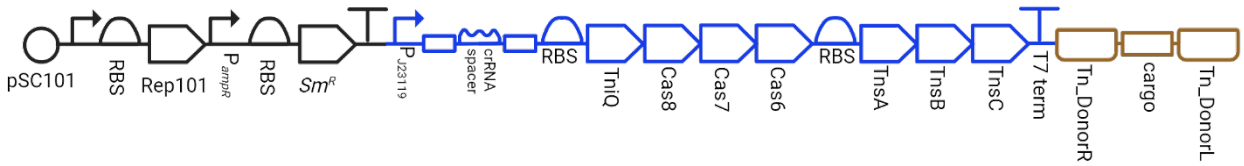


Figure 9 – SBOL Diagram of pSpin/pSL1521. This plasmid encodes the Type I-F CRISPR-Cas system from *Vibrio cholerae*, including CRISPR array, Cascade (TniQ–Cas8–Cas7–Cas6), and transposition machinery (TnsA–TnsB–TnsC). Expression is driven by the constitutive promoter P_{J23119}. The crRNA spacer guides sequence-specific targeting. A synthetic mini-transposon flanked by Tn_{DonorL} and Tn_{DonorR} sites enables site-specific integration of desired cargo. The plasmid also includes a pSC101 origin, streptomycin resistance, and *Rep101* for replication control. Components are color-coded; black: plasmid backbone and maintenance elements; blue: CRISPR-Cas machinery; brown: transposon donor cassette.

By co-delivering pSpin-nth alongside the circuit of interest (e.g., pSR40.29-kil or its derivatives) after insertion of mini-transposon donor sites or placing the circuit elements onto pSpin-nth itself, the CRISPR–Cas transposition system integrates the entire construct into the host chromosome. This reduces metabolic burden from plasmid maintenance and ensures stable inheritance of the life-death selection circuit.

3.2.1.3 pSpin-40-kil

Following our initial success with the plasmid-based kill circuit (pSR40.29-kil), we aimed to stabilize the system by integrating it into the *E. coli* BW29655 genome. To achieve this, we constructed pSpin-40-kil, using pSpin (pSL1521)—a vector that supports CRISPR-guided transposition into the nth-ydgr locus—as the backbone. We removed the original antibiotic resistance portions from pSR40.29-kil to avoid redundancy and inserted the remaining kill circuit

into pSpin using Golden Gate assembly, thus creating a single plasmid that could both target nth-ydgr and carry the kil-based life-death selection module.

Transformation of pSpin-40-kil into BW29655 and subsequent selection yielded multiple colonies. To confirm successful integration, we performed PCR amplification of the nth-ydgr locus followed by sequencing. The results verified that our construct was precisely inserted into the intended genomic location, indicating that CRISPR-guided transposition had occurred as designed.

Next, we evaluated the functionality of the integrated circuit using a plate reader. In particular, we measured sfGFP fluorescence under induced and uninduced conditions. Although we did detect sfGFP from the chromosome, there was no discernible difference between the two conditions, suggesting that the fluorescence readings were driven by baseline expression (noise) rather than true ligand-dependent induction (Fig. 10); this may be due to the ratio of SHK:RR being off, as previous work has shown dynamic range is strongly affected by the relative ratio of each (Landry et al., 2018). We also assessed how the presence or absence of theophylline (which should toggle the kil system on or off) influenced growth rates (Fig. 11). Here too, there was essentially no difference beyond normal variability, implying that kil expression was not robust enough to exert a lethal effect on the cells.

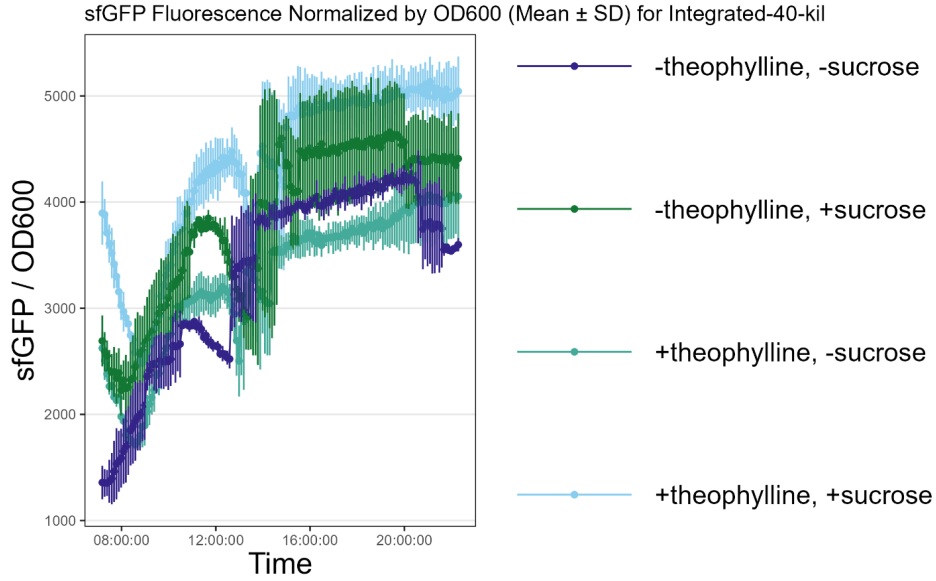


Figure 10 – Plate reader sfGFP fluorescence data of the integrated 40-kil circuit in BW29655. The trends shown here indicate significant noise in the system, with no clear, usable timeframe of clean separation.

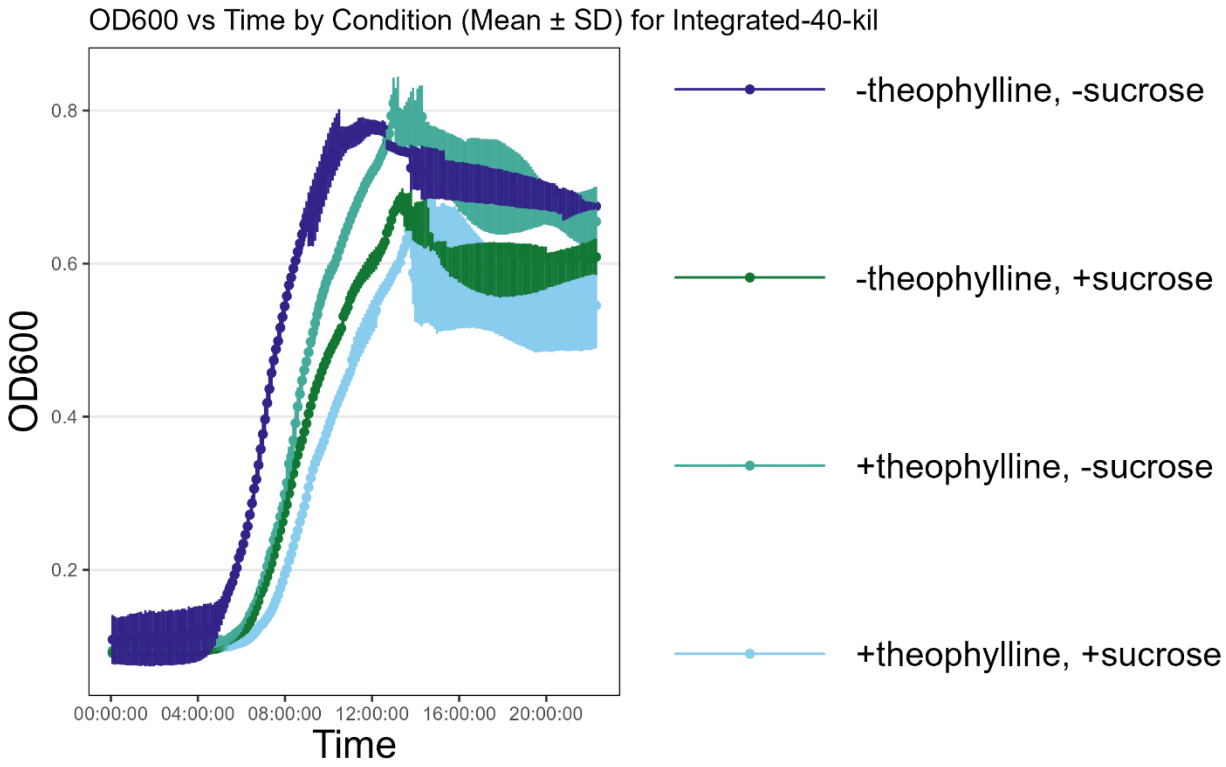


Figure 11 – Plate reader OD600 data of the integrated 40-kil circuit in BW29655. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly six to eight hours before wells started to reach an OD600 of 0.2. The Kil protein seems to have a minimal effect on the growth rates.

Given these observations, we suspect that chromosomal context and/or reduced copy number hindered the circuit's performance, preventing it from generating a sufficiently large on/off dynamic range. It is also possible that local regulatory effects in the nth-ydgr region diminished the circuit's responsiveness. Consequently, we decided to alter the orientation of the integrated circuit in an effort to improve functionality, with the hope that reconfiguring its genomic arrangement would restore the clear induction response and kill-based selection we observed in the plasmid-based format.

3.2.1.4 pSR40.29-kilFlip

Because the original life-death selection circuit did not exhibit the desired efficacy—particularly regarding the kil-mediated kill switch—and sfGFP expression appeared dominated by background noise, we created a reoriented version of the design while preserving most of the

functional elements. In this variant, named pSR40.29-kilFlip, we changed the order of key genetic parts so that sfGFP would be expressed first, followed by the RiboJ insulator, then the ribosome binding site (RBS), then the theophylline-responsive aptazyme, and finally the kil gene (Fig. 12). Our goal was to reduce potential interference between the kil module and the fluorescent reporter, as well as to minimize leaky kil expression before sfGFP could be reliably detected.

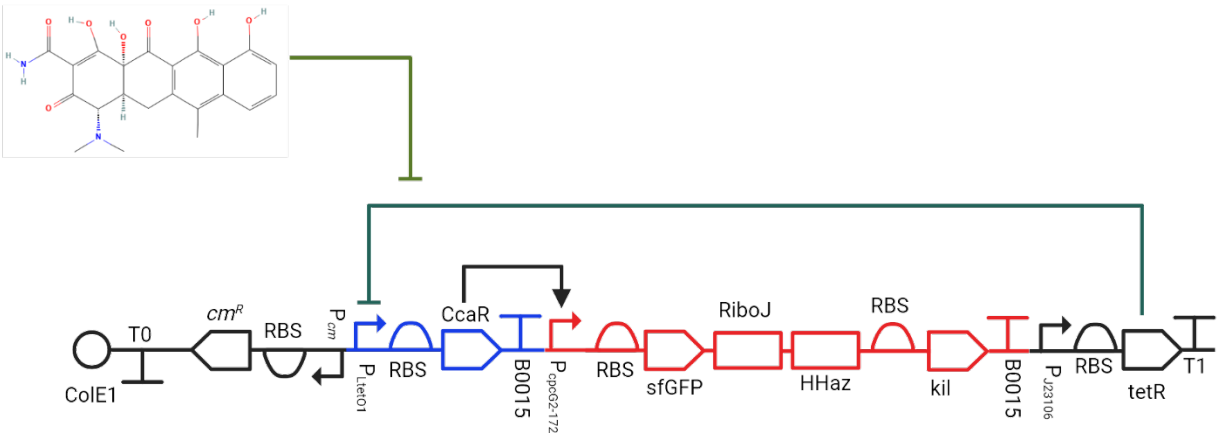


Figure 12 – SBOL Diagram of pSR40.29-kilFlip. This plasmid encodes the same circuit as pSR40.29-kil, but with the order of *sfGFP* and *kil* reversed. Upon induction by aTc, CcaR is expressed, and once phosphorylated by an SHK, activates the $P_{cpcG2\text{-}172}$ promoter, leading to co-expression of sfGFP and the toxin gene *kil*. A RiboJ insulator and HHaz ribozyme are included to buffer expression and modulate transcript stability. sfGFP now precedes *kil*, in contrast to the arrangement in pSR40.29-kil. As before, TetR represses P_{LtetO1}, forming a feedback loop that can be disrupted by aTc. Colors are employed to indicate distinct transcriptional units; blue: *ccaR*; red: toxin and *sfGFP*; black: *tetR* and backbone elements.

In the original configuration, even minor basal translation of kil could reduce the viability of cells before sufficient sfGFP expression occurred, thus skewing fluorescence measurements. Flipping these elements places sfGFP (along with its RBS) upstream, providing a clearer readout of promoter activity before the kill system is triggered. The aptazyme and the kil gene remain downstream, thus helping ensure that kil is produced only under tightly controlled, theophylline-inducible conditions. We implemented the reorientation through PCR amplification of the relevant segments from the previous plasmid. This modular approach allowed us to preserve all

core functional parts—such as the cpcG2-172 promoter, the aptazyme sequence, and the kil gene—while rearranging their positions to reduce undesired crosstalk.

3.2.1.4.1 Initial Testing of pSR40.29-kilFlip

We co-transformed pSR40.29-kilFlip alongside a sensor histidine kinase (SHK) plasmid into *E. coli* BW29655 and subjected the cells to plate reader assays to assess both fluorescence output and growth behavior.

Under inducing conditions, pSR40.29-kilFlip displayed an 8- to 10-fold increase in sfGFP signal relative to the uninduced state (Fig. 13). By comparison, the original orientation typically yielded only 2- to 3-fold induction. This suggests the reversed arrangement reduced leaky expression and improved the dynamic range of fluorescence detection.

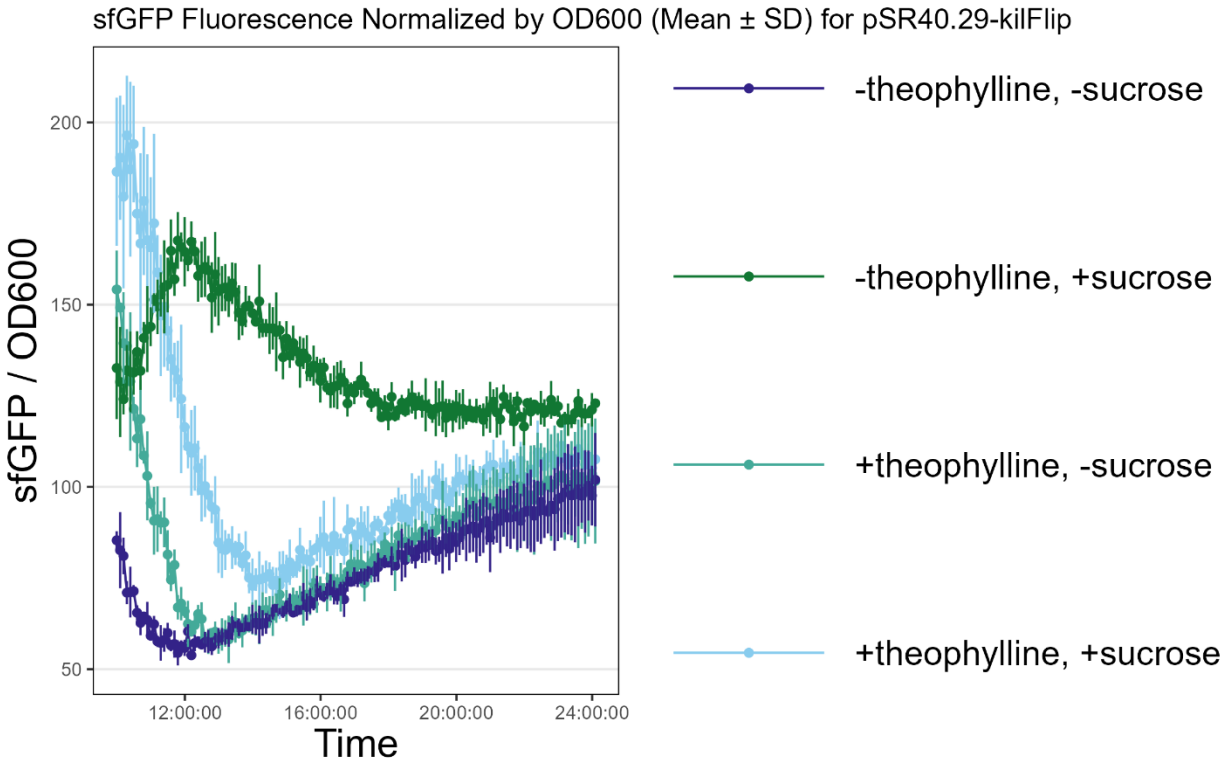


Figure 13 – Plate reader sfGFP fluorescence data of the 40-kilFlip circuit. The timeframe for maximum signal-to-noise ratio starts about ten hours into the experiment and continues for around six hours, with a peak dynamic range of 113 AU at twelve hours, a max value for the system with ligand present and without theophylline during this timeframe around 170 AU, and a max signal-to-noise ratio of 3.11.

When the theophylline aptazyme was activated, the kil gene substantially slowed population growth, indicating effective kill-switch activity (Fig. 14). We observed a twofold increase in the doubling time compared to cells grown under non-inducing conditions. Even though some basal toxicity was still detectable, it was significantly less severe than in the previous configuration, allowing a clearer on/off distinction for kil-mediated killing.

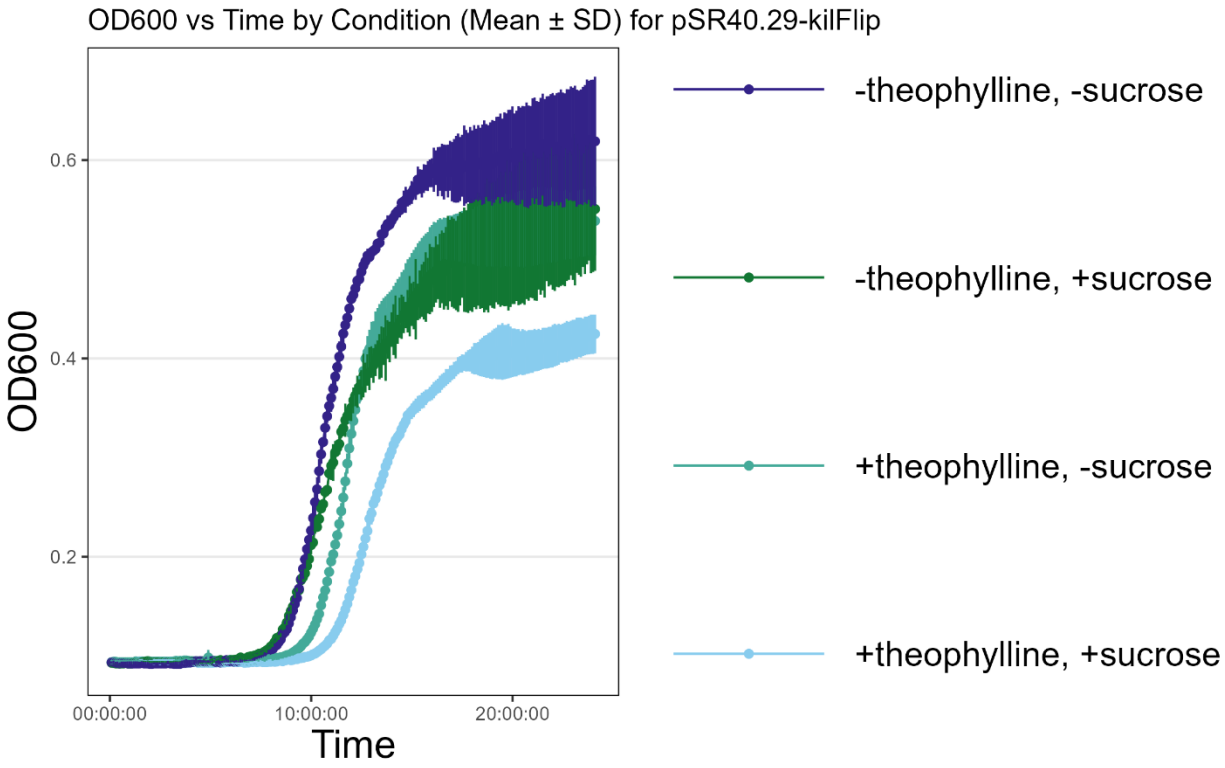


Figure 14 – Plate reader OD600 data of the 40-kilFlip circuit. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly ten to thirteen hours before wells started to reach an OD600 of 0.2, with the Kil protein seeming to contribute to slower growth for some conditions.

Overall, these preliminary results supported the notion that flipping the order of elements provided a more robust platform for distinguishing true signal from noise. sfGFP expression was more clearly tied to genuine promoter activation, and the kil module exhibited tighter control, yielding a measurable impact on cell viability only when intended. Encouraged by the improved dynamic range and kill effect observed with pSR40.29-kilFlip in plasmid form, we decided to integrate this new construct into the chromosome.

3.2.1.5 pDonor-40-kilFlip

Despite multiple attempts at integrating the pSR40.29-kilFlip cargo into the pSpin backbone via various cloning methods, we encountered persistent difficulties. Consequently, we adopted a two-plasmid system (Vo et al., 2021), comprising pSL1777 (pEffector), which provides the CRISPR–Cas transposition machinery, and pSL1119 (pDonor), a vector designed to house the DNA payload destined for integration. We cloned our kilFlip circuit into the pDonor backbone, generating pDonor-40-kilFlip, which was then co-transformed with pEffector into *E. coli* BW29655. Successful genomic insertion was confirmed by PCR amplification of the integration site and subsequent Sanger sequencing.

To remove the donor and effector plasmids post-integration, we employed a modified version of pFree_cm (Lauritsen et al., 2017), which was engineered to cure plasmids with a CDF origin in addition to its original curing targets. After transforming pFree into the newly engineered LOS (Locked-On Sorting) strain, both pDonor-40-kilFlip and pEffector were effectively eliminated, and we then cured pFree by heat treatment. At this stage, no extraneous plasmids carrying antibiotic resistance or transposition systems remained in the strain, leaving the integrated circuit in the genome.

We next introduced a plasmid expressing a Tar-EnvZ chimeric sensor histidine kinase (SHK) into this strain and performed plate reader assays to quantify both kill-system efficiency and sfGFP output. While the presence of the kil gene did appear to slow growth rates marginally—on the order of a 25% reduction compared to the no-theophylline control (Fig. 15)—the effect remained modest, suggesting that the kill system was operating at a much lower efficiency than intended. Moreover, sfGFP levels under induced conditions were again indistinguishable from baseline noise (Fig. 16), indicating insufficient dynamic range for reliable detection, with a maximum signal-to-noise ratio of 1.1 and a maximum dynamic range of 18 AU. These outcomes prompted us to collect quantitative data on expression of mRNA to compare the plasmid-based circuit versus the integrated versions, aiming to pinpoint possible transcriptional or translational bottlenecks.

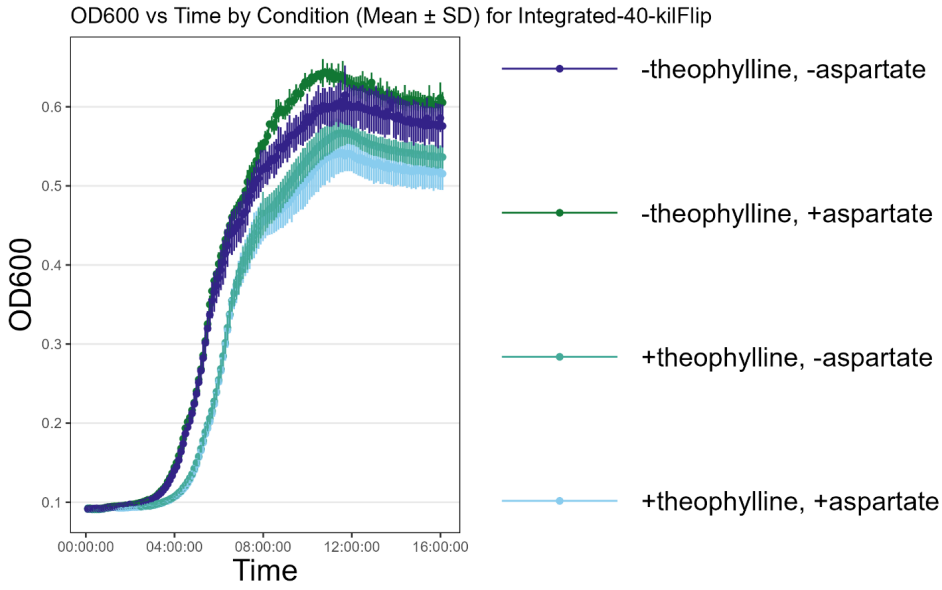


Figure 15 – Plate reader OD600 data of the integrated 40-kilFlip circuit. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly four to six hours before wells started to reach an OD600 of 0.2, with the expression of the Kil protein noticeably slowing growth.

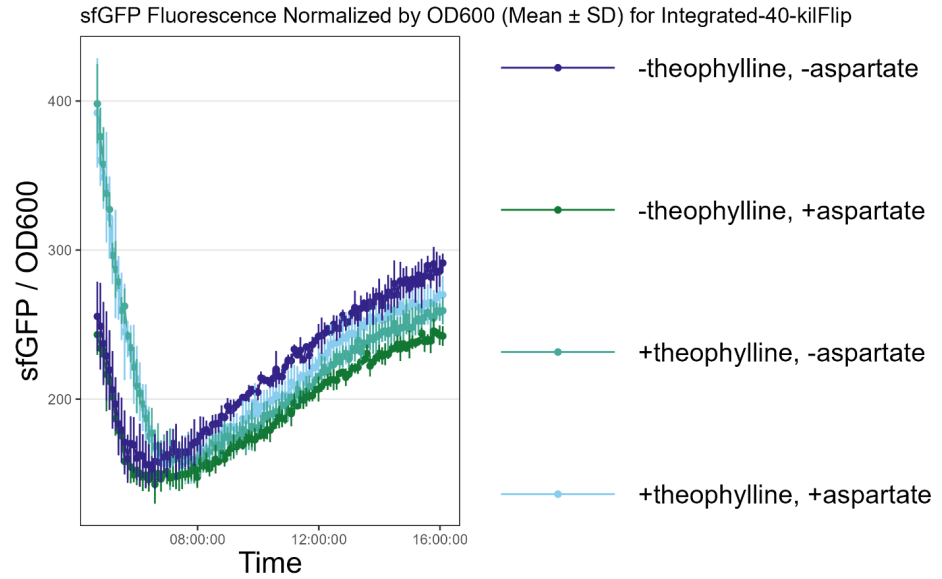


Figure 16 – Plate reader sfGFP fluorescence data of the integrated 40-kilFlip circuit. The reliable timeframe for maximum signal-to-noise ratio starts about eight hours into the experiment and continues for around four hours, with a peak dynamic range of around 18 AU, a max value for the system with ligand present and without theophylline during this timeframe around 250 AU, and a max signal-to-noise ratio of 1.1.

In light of these challenges, we decided to design a second-generation locked-on sorting system to improve both kill-switch potency and fluorescent readout. Our immediate priorities include rearranging genetic elements, optimizing promoter and RBS strengths, and investigating alternative integration sites or approaches.

3.2.1.6 Transcript Expression Testing

3.2.1.6.1 Plasmid and Integrated Construct

To understand why our integrated constructs showed lower signal-to-noise ratios than plasmid-based systems, we performed reverse-transcription quantitative PCR (RT-qPCR) to compare mRNA levels in the two setups. Specifically, we tested a combination of the Tar-EnvZ sensor histidine kinase (SHK) plus pSR40.29-kilFlip (the two-plasmid system) versus the same Tar-EnvZ SHK with the integrated LOS strain. We also examined whether ligand presence or absence (in this case, aspartate for Tar-EnvZ) influenced transcript abundance, creating four experimental conditions in total (plasmid-based or integrated, each with or without ligand).

For each gene target, we designed probes labeled with FAM dye and Zen–Iowa Black quenchers. Our key targets included kil, sfGFP (the fluorescent reporter), a conserved region of EnvZ (to detect transcripts common to both Tar-EnvZ and the full-length EnvZ), and the OmpR portion of the OmpR–CcaR fusion. Additionally, we used idnT, hcaT, and cysG as reference genes because they consistently exhibit stable expression in our conditions. Total RNA was isolated with the Monarch® Total RNA Miniprep Kit (New England Biolabs, NEB), ensuring minimal genomic DNA contamination and consistent yields. We then employed the Luna® Universal Probe One-Step RT-qPCR Kit (also NEB) for cDNA synthesis and real-time amplification, combining reverse transcription and PCR into a single streamlined protocol. Each sample was prepared and run in triplicate to reinforce the statistical validity of our measurements. After background correction ($Q_{corr} = Q_{sample} - Q_{noRT}$), any negative values were set to zero ($Q_{corr,\ floored}$). To prevent infinite fold-changes when the baseline $Q_{corr,\ floored}$ was zero, we then added a small pseudocount (ε)—chosen as the smallest non-zero $Q_{corr,\ floored}$ observed across all genes and conditions ($\varepsilon = 0.57854$). Fold-changes were computed as $(Q_{corr,\ floored} + \varepsilon)/(Q_{baseline,\ floored} + \varepsilon)$. This strategy ensures that true zero values yield a neutral ratio of 1 and that larger signals are only minimally biased by the pseudocount.

The results revealed only a modest (~0.5-1.5-fold) difference in transcript levels between ligand-induced and uninduced samples, both for the plasmid-based system and the integrated

system (Table 1). This modest ratio paralleled our plate reader data, which similarly indicated weak separation between the on and off states. However, when we compared absolute mRNA abundances, the two-plasmid system showed far higher expression—on the order of 5- to 280-fold—for kil, the OmpR–CcaR fusion, and sfGFP in the integrated system. This substantial difference highlights the influence of copy number: with a single integrated copy in the chromosome, the system struggles to generate the transcript levels necessary for a strong on/off discrimination, while plasmid-based constructs profit from higher multiplicities that amplify the signal.

Target gene	Condition	Cq ref	Cq ref NRT	Cq tgt	Cq tgt NRT	Q, sample	Q, noRT	Q, corr floored	Ratio vs baseline, with pseudocount	epsilon used
sfGFP	Integrated -Asp	30.14	33.49	31.02	34.33	0.54	0.56	0.00	1.00	0.57854
sfGFP	Integrated +Asp	31.94	32.89	31.72	33.67	1.16	0.58	0.58	2.00	0.57854
sfGFP	Two-plasmid -Asp	29.80	33.79	22.54	26.35	153.18	173.36	0.00	1.00	0.57854
sfGFP	Two-plasmid +Asp	29.44	34.67	22.16	28.07	155.91	96.86	59.05	103.07	0.57854
kil	Integrated -Asp	31.31	33.34	29.28	33.87	4.09	0.69	3.40	1.00	0.57854
kil	Integrated +Asp	31.71	32.22	31.43	33.67	1.21	0.36	0.85	0.36	0.57854
kil	Two-plasmid -Asp	30.90	33.09	21.13	25.73	871.77	164.03	707.74	178.25	0.57854
kil	Two-plasmid +Asp	30.55	33.88	20.33	27.48	1192.86	84.31	1108.55	279.12	0.57854
EnvZ	Integrated -Asp	29.52	31.66	21.54	28.30	252.44	10.28	242.16	1.00	0.57854
EnvZ	Integrated +Asp	30.37	30.91	23.95	27.57	85.17	10.06	75.10	0.31	0.57854
EnvZ	Two-plasmid -Asp	30.34	32.38	20.04	26.60	1258.76	54.96	1203.79	4.96	0.57854
EnvZ	Two-plasmid +Asp	29.99	32.80	19.90	28.22	1092.34	23.83	1068.51	4.40	0.57854
OmpR	Integrated -Asp	30.14	33.49	24.69	32.93	43.73	1.48	42.25	1.00	0.57854
OmpR	Integrated +Asp	31.94	32.89	27.89	33.15	16.62	0.83	15.79	0.38	0.57854
OmpR	Two-plasmid -Asp	29.80	33.79	17.20	25.58	6228.63	297.25	5931.38	138.51	0.57854
OmpR	Two-plasmid +Asp	29.44	34.67	16.60	26.80	7320.49	232.86	7087.63	165.51	0.57854

Table 1 – RT-qPCR–Derived Expression Metrics and Pseudocount-Adjusted Foldchanges. RT-qPCR–derived expression metrics and pseudocount-adjusted fold-changes for circuit (sfGFP, kil, EnvZ, RR1, RR4) under four conditions, these being the genomically integrated and two-plasmid systems, each with and without the presence of asparate. For each gene–condition pair, raw relative quantity ($Q_{sample} = 2^{(Cq_{ref} - Cq_{tgt})}$) and background ($Q_{noRT} = 2^{(Cq_{ref,\ noRT} - Cq_{tgt,\ noRT})}$) were computed, then background-corrected ($Q_{corr} = Q_{sample} - Q_{noRT}$) and floored at zero ($Q_{corr,\ floored}$). To avoid division by zero when the baseline $Q_{corr,\ floored}$ was zero, a pseudocount ε equal to the smallest non-zero $Q_{corr,\ floored}$ observed (0.57854) was added to both numerator and denominator. Pseudocount-adjusted fold-change was then calculated as: Ratio: $(Q_{corr,\ floored} + \varepsilon)/(Q_{baseline,\ floored} + \varepsilon)$

These findings help explain why the integrated construct underperformed. The single-copy arrangement constrains transcript output, resulting in a small gap between baseline (noise) and induced (signal), thereby producing a poor dynamic range for screening. Compounding the issue, the promoter regulated by OmpR–CcaR could not be readily replaced without sacrificing specificity or orthogonality. Since the promoter strength could not simply be boosted in the integrated context, and the copy number was fixed at one per cell, our results suggested that a genomic approach was unlikely to provide the robust performance needed for locked-on sorting under the current design parameters.

Consequently, after confirming through RT-qPCR that the integrated circuit's lower expression levels were a key bottleneck, we opted to revert to a two-plasmid configuration and redesign the response regulator plasmid. By exploiting higher copy numbers, we anticipate generating more pronounced differences between induced and uninduced states. This experience outlined how transcript-level measurements can pinpoint bottlenecks in synthetic circuit performance and guide subsequent improvements in design.

3.2.2 Locked-On Sorting – Second General Design

3.2.2.1 pSR40.29-LRv2

In seeking to overcome the limitations of our first locked-on sorting constructs, we designed and built a new plasmid called pSR40.29-LRv2 (Fig. 17). Our primary goals were to (1) enhance the dynamic range of the fluorescent readout, (2) improve the versatility and tunability of the life-death selection, and (3) address issues related to plasmid stability and unwanted recombination events. Below is an overview of the most significant alterations and the reasoning behind them.

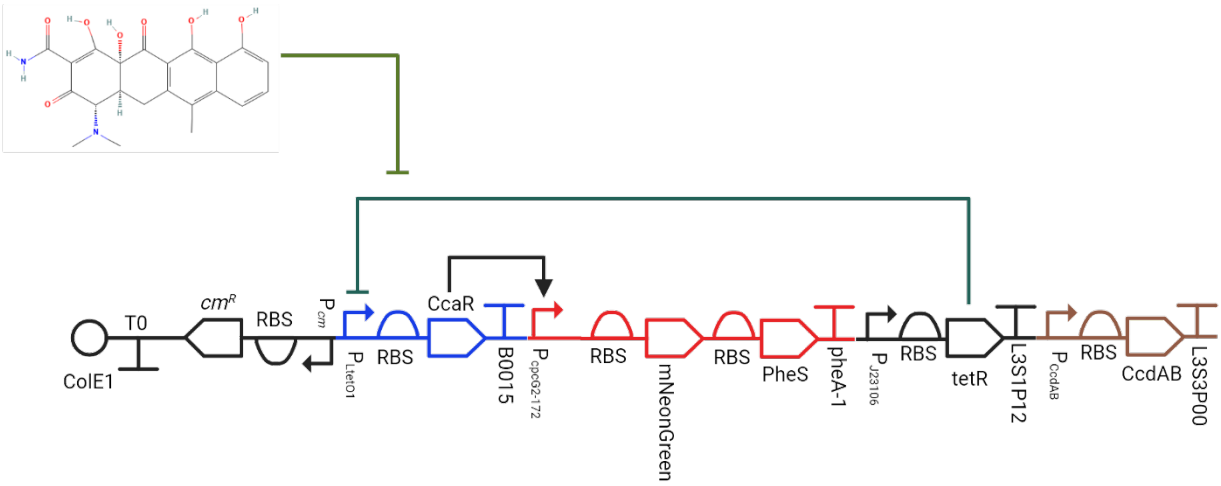


Figure 17 – SBOL Diagram of pSR40.29-LRv2. This plasmid implements an inducible system that co-expresses the fluorescent protein mNeonGreen and a toxic PheS variant, enabling both fluorescent reporting and negative selection. CcaR is under control of the TetR-repressed promoter P_{LtetO1} and activates $P_{cpcG2\text{-}172}$ to drive the expression of the output module. TetR, constitutively expressed from, represses CcaR, forming a feedback loop tunable by aTc. A third module expresses the *ccdAB* toxin-antitoxin operon from the weak constitutive promoter P_{CcdAB}, providing an orthogonal post-locked-on-sorting selection against this plasmid. Colors are employed to indicate distinct transcriptional units; blue: *ccaR*; red: *mNeonGreen*, *pheS*; *tetR* and backbone elements; brown: *ccdAB* system.

1. Enhanced Fluorescent Reporter

One major change was replacing sfGFP with the mNeonGreen fluorescent protein. Compared to sfGFP, mNeonGreen offers a higher quantum yield, greater brightness, and a shorter maturation time—attributes that together produce a more sensitive and rapid fluorescence signal. We also increased the strength of its ribosome binding site (RBS) by two orders of magnitude, from a translation initiation rate (TIR) of 1692 AU to 911,556 AU, with these values predicted by the RBS Calculator by De Novo DNA (Reis & Salis, 2020). These modifications should provide a larger, more robust fluorescent signal upon promoter activation, thereby boosting the signal-to-noise ratio.

2. Switching from Kil to PheS for Life-Death Selection

Instead of the Kil protein, which can sometimes suffer from detrimental background leakage, we introduced PheS (A249G), a modified phenylalanyl-tRNA synthetase. This mutated enzyme incorporates 4-chloro-DL-phenylalanine (fenclonine) into proteins, which is cytotoxic (Maranhao & Ellington, 2017; Thyer et al., 2013). Unlike Kil, whose expression can quickly kill cells even with slight transcriptional leakage, PheS toxicity is tunable by adjusting fenclonine concentration in the growth medium. This titratable approach allows finer control over stringency: modest fenclonine levels can remove constitutively high expressors without drastically reducing overall cell viability, while stronger induction can sharply discriminate true locked-on mutants.

3. Plasmid Removal Features and ccdB System

A practical obstacle in screening large sensor histidine kinase (SHK) libraries is preventing contamination or carryover of the locked-on sorting plasmid during subsequent manipulations (e.g., miniprepping libraries). To mitigate this, we introduced multiple BsaI restriction sites into pSR40.29-LRv2. Because BsaI is a “forbidden” restriction site absent from the SHK plasmids, any intact locked-on sorting plasmid that appears during library preparation can be selectively digested. Furthermore, we embedded the CcdB toxin from the CcdB/CcdA toxin-antitoxin system (Vandervelde et al., 2017). If any residual pSR40.29-LRv2 plasmids evade digestion and transform into cells lacking its antitoxin (CcdA), the toxicity of CcdB will be lethal, ensuring that only the desired SHK plasmid persists.

4. Updated Terminators to Reduce Recombination

We addressed concerns about repeat regions and recombination hot spots by modifying several terminators. The original pSR40.29 relied on repeating certain terminators multiple times—a setup prone to homologous recombination, especially in strains such as ours that retain a functional RecA. To circumvent this, we replaced these repeats with strong, distinct terminators. Specifically, the transcript for mNeonGreen and PheS ends with the pheA-1 terminator, which is both strong and naturally occurring. Meanwhile, the terminators for TetR and the ccdAB operon were switched to synthetic terminators, L3S1P12 and L3S3P00, respectively (Y.-J. Chen et al., 2013). By employing these diverse, well-characterized terminators, we aimed to limit unwanted recombination while maintaining robust termination efficiency.

3.2.2.1.1 Modification of Genomic GyrA

To successfully use pSR40.29-LRv2, we initially planned to introduce the R462C mutation into the gyrA gene of the E. coli BW29655 genome. This single-amino-acid substitution (achieved via a single C-to-T transversion in the DNA sequence) confers resistance to CcdB—the toxin used in our locked-on sorting plasmid—by preventing it from binding to the GyrA subunit of DNA gyrase. With the R462C variant, cells would be able to tolerate the presence of CcdB, enabling straightforward selection for correct recombinants when ccdB is expressed.

3.2.2.1.1.1 First Approach: Recombineering with pTac-ccdB and Oligonucleotide Mutagenesis

Our initial strategy centered on recombineering using a pTac-ccdB plasmid, which allows inducible expression of CcdB via IPTG (Besmer et al., 2006). We reasoned that inducing CcdB would kill all cells lacking the gyrA R462C mutation, enriching for potential mutants. To introduce the point mutation, we treated the strain with RecA and co-transformed a phosphorothioated oligo carrying the desired C-to-T transversion in gyrA along with the pTac-ccdB plasmid (Murphy & Marinus, 2010; van Loenhout et al., 2009; Wannier et al., 2021). The phosphorothioation protected the oligo from intracellular nucleases, theoretically enhancing mutation efficiency.

Despite carefully optimizing transformation and selection protocols, we only recovered false positives—colonies that appeared resistant but carried no actual R462C substitution. This likely stemmed from leaky or insufficient CcdB expression under our conditions or spontaneous mutations elsewhere in the genome that partially rescued growth. In any event, further attempts to validate these colonies via sequencing consistently showed no sign of the intended mutation.

3.2.2.1.1.2 Second Approach: λ-Red Recombineering with pTKRED

After the initial failures, we turned to a more robust, λ-Red-mediated recombineering using pTKRED (Kuhlman & Cox, 2010). In this system, IPTG induces λ-Red recombination functions (gam, bet, exo), promoting homologous recombination with donor DNA. We built a donor construct (Fig. 18) containing the mutated gyrA (R462C) allele linked to a kanamycin-resistance marker (kanR) flanked by FRT sites (Schlake & Bode, 2002). This donor was then electroporated into BW29655 carrying pTKRED, and transformants were plated on selective media (kanamycin) and weakly selective media.

Once again, we observed false positives—colonies that grew on selective plates but did not actually contain the desired gyrA mutation when we checked them by PCR and sequencing. Possible explanations include partial recombination that inserted only the kanR marker without the specific point mutation or secondary mutations providing nonspecific CcdB tolerance.

3.2.2.1.2 Deciding Against CRISPR-Based Approaches and Removing ccdAB

Although CRISPR/Cas9 methods could have been employed for more precise genome editing, we ultimately opted to remove the ccdAB operon from pSR40.29-LRv2 altogether rather than continue pursuing the gyrA route. Given that pSR40.29-LRv2 already harbored multiple other design features—such as BsaI cut sites and an improved fluorescent reporter—that provided robust selection and screening capabilities, the additional complexity of engineering gyrA to tolerate CcdB no longer seemed essential. We concluded that these built-in safeguards (BsaI restriction for plasmid removal and the new life-death mechanism with PheS) were sufficient for our library workflows without requiring the gyrA R462C mutation.

In summary, while the notion of making BW29655 intrinsically resistant to CcdB via the R462C substitution in gyrA was conceptually sound, practical limitations—including high false-positive rates and multiple unsuccessful attempts to introduce the single-base change—drove us toward a simpler solution. By abandoning ccdAB-based selection and focusing on alternative kill-switch mechanisms, we streamlined the system and avoided potentially confounding genome modifications.

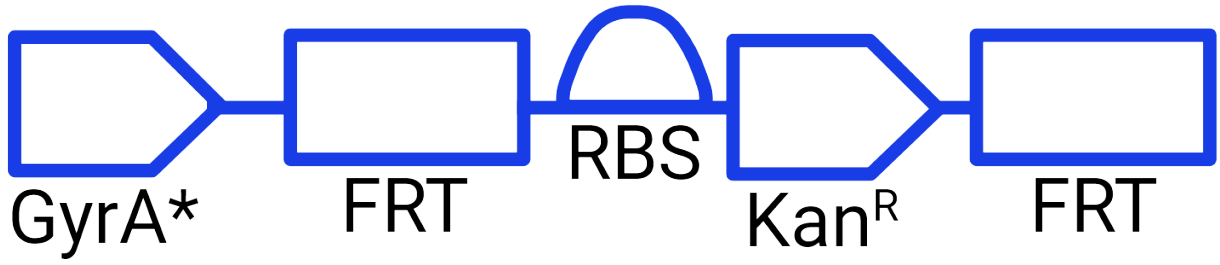


Figure 18 – SBOL Diagram of GyrA-FRT Construct. This construct contains the mutant *gyrA* (R462C), followed by a kanamycin resistance gene *kan*^{R} under its own RBS, flanked by Flp recognition target (FRT) sites for future excision. This cassette was used as donor DNA for λ-Red recombineering into *E. coli* strain BW29655 carrying the inducible recombination plasmid pTKRED.

3.2.3 Locked-On Sorting – Third General Design

3.2.3.1 pJH998

3.2.3.1.1 CcdAB Operon Removal

For the purpose of moving past the practical limitations presented by the inclusion of the selection based on the ccdAB operon, and as alternative kill-switch mechanisms are present, we sought to make the third locked-on sorting construct via removal of the ccdAB operon. We implemented this via PCR amplification of all desired parts. However, due to difficulties that arise when PCR primers are designed to anneal directly in terminator regions of sequences, the parts removed were the terminator directly after the TetR gene and the majority of the ccdAB operon, leaving an 80 basepair scar after TetR's stop codon and retaining the L3S3P00 synthetic terminator, previously the terminator for the ccdAB operon, as the terminator for the TetR gene. This new version of the locked-on sorting plasmid, which is thus a derivative of pSR40.29-LRv2 with the CcdAB operon removed, was named pJH998 (Fig. 19).

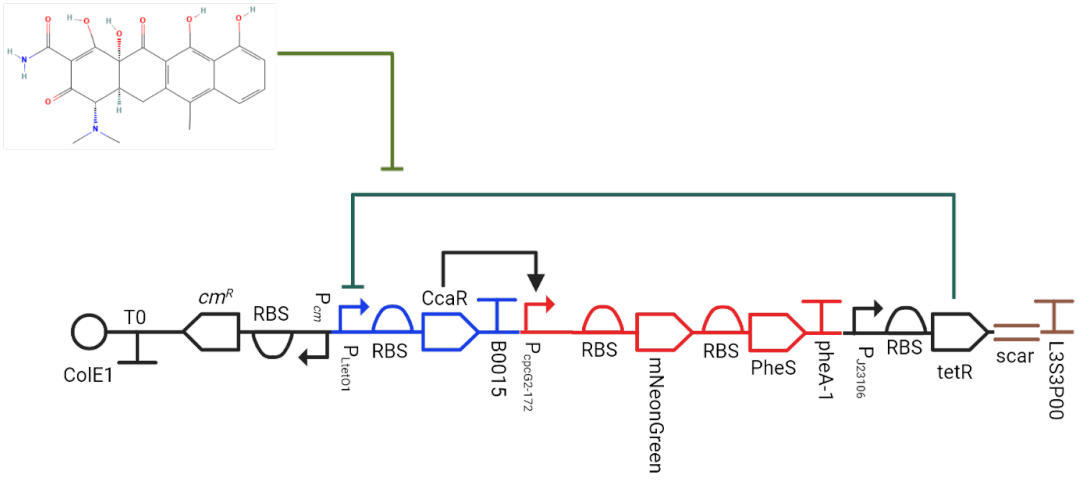


Figure 19 – SBOL Diagram of pJH998. This plasmid retains the same core logic as pSR40.29-LRv2: this plasmid implements an inducible system that co-expresses the fluorescent protein mNeonGreen and a toxic PheS variant, enabling both fluorescent reporting and negative selection. CcaR is under control of the TetR-repressed promoter P_{LtetO1} and activates $P_{cpcG2\text{-}172}$ to drive the expression of the output module. TetR, constitutively expressed from, represses CcaR, forming a feedback loop tunable by aTc. Unlike pSR40.29-LRv2, the downstream *ccdAB* operon has been removed, leaving a residual transcriptional scar followed by a terminator (L3S3P00). Colors are employed to indicate distinct transcriptional units; blue: *ccaR*; red: *mNeonGreen*, *pheS*; black: *tetR* and backbone elements; brown: scar and terminator.

3.2.3.1.2 pJH991

For assays with a fluorescence readout, it is very useful to have a positive control plasmid containing the fluorescent protein of interest in a manner responsive to induction. mNeonGreen is used as the fluorescent protein in pJH998, so we used PCR amplification to remove a portion of pJH998, resulting in mNeonGreen now being under the control of the pLtetO-1 promoter, which is repressed by TetR unless the inducer aTc is added, and this new positive control plasmid was named pJH991 (Fig. 20).

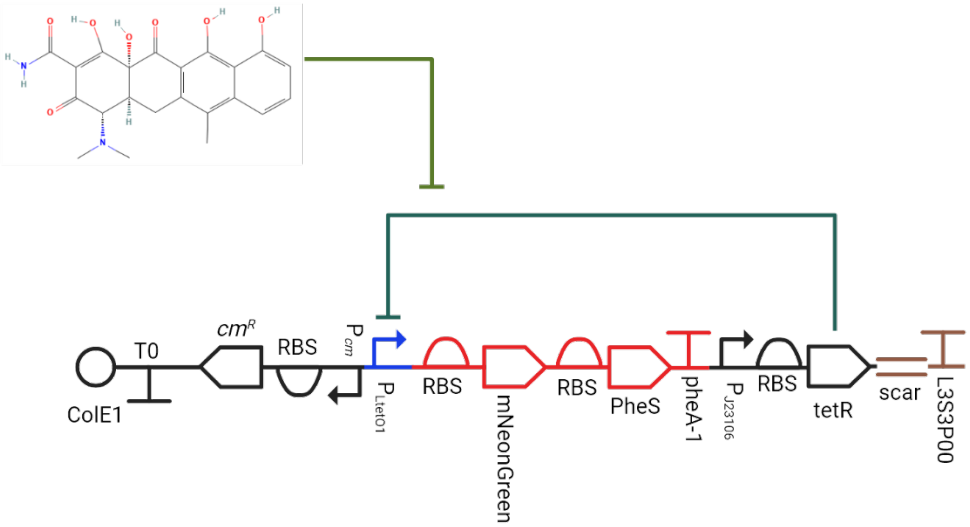


Figure 20 – SBOL Diagram of pJH991. This plasmid eliminates the *ccaR* response regulator and associated $P_{cpcG2\text{-}172}$ promoter, resulting in direct expression of the mNeonGreen–PheS* operon from the TetR-repressible promoter P_{LtetO1}. In the absence of aTc, TetR represses expression; upon addition of aTc, repression is relieved, allowing induction of mNeonGreen and PheS. This minimal version removes signal sensing in favor of direct control of fluorescent induction via aTc. Colors are employed to indicate distinct transcriptional units; blue: *ccaR*; red: *mNeonGreen*, *pheS*; *tetR* and backbone elements; brown: scar and terminator.

We co-transformed sequencing-verified pJH998 alongside a sensor histidine kinase (SHK) plasmid into *E. coli* BW29655. We also transformed pJH991 into the same strain of *E. coli*. From glycerol stocks of these transformed strains, we then inoculated supplemented M9 media (as described above) for the purpose of growing overnight cultures for the purpose of subjecting the cells to plate reader assays to assess both fluorescence output and growth behavior. However, when it came time the next day to collect the cells and start the assay, there

was no measured change in growth, based on the OD600. The components to make up the supplemented M9 media were all remade, and to determine if there was an issue with the glycerol stocks, with which we inoculated the overnight cultures. These grew overnight, but to a lower OD600 than anticipated.

When these cultures were subjected to a plate reader assay, they grew much more slowly than normal, even considering pJH991 had extra cells spiked in so as to still be useful as a positive control, with pJH998 reaching an OD600 of 0.2 in eighteen hours after wells were started with an initial OD600 of 0.005 (Fig. 21), compared to other circuits typically reaching an OD600 of 0.2 in six to nine hours with starting OD600s of 0.0005, an order of magnitude smaller. The measured mNeonGreen expression seemed not much better than noise for this system, which likely was affected by the very slow growth experienced (Fig. 22).

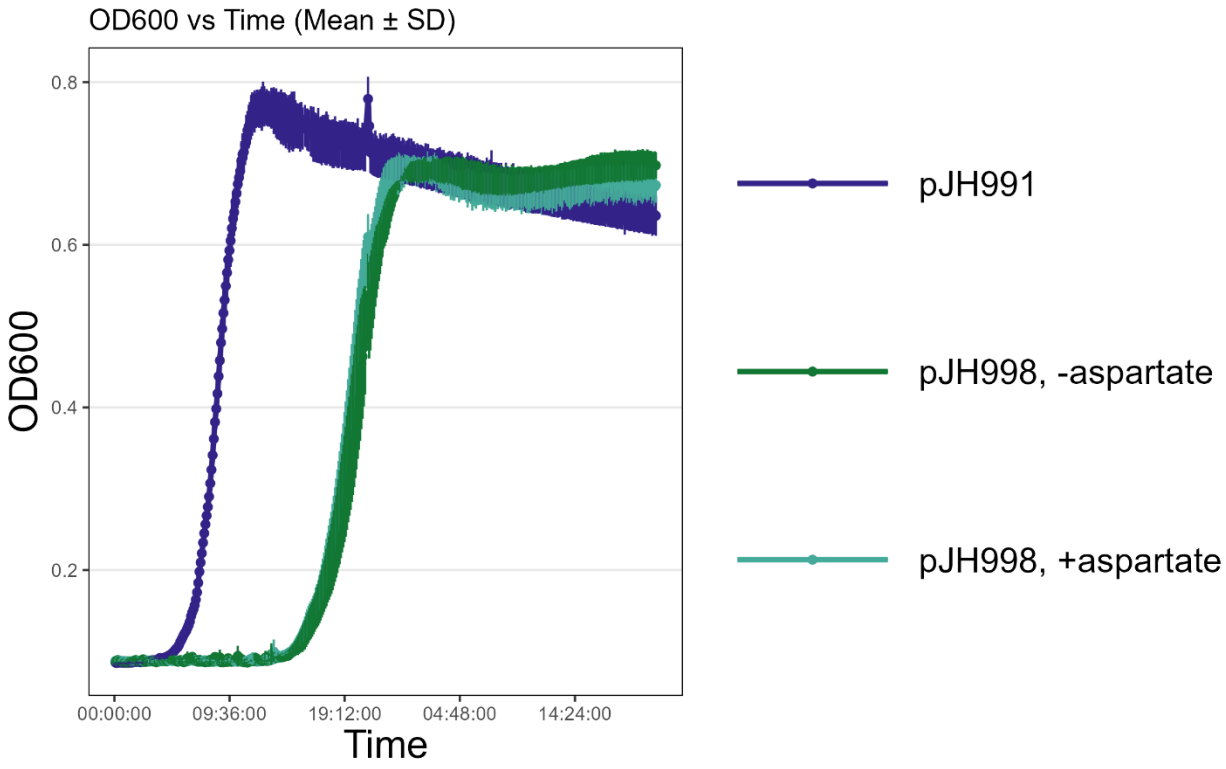


Figure 21 – Plate reader OD600 data of pJH998 and pJH991. With a starting OD600 of 0.005 for all conditions in this experiment, it took roughly eight to eighteen hours before wells started to reach an OD600 of 0.2, indicating a severe issue with growth rate.

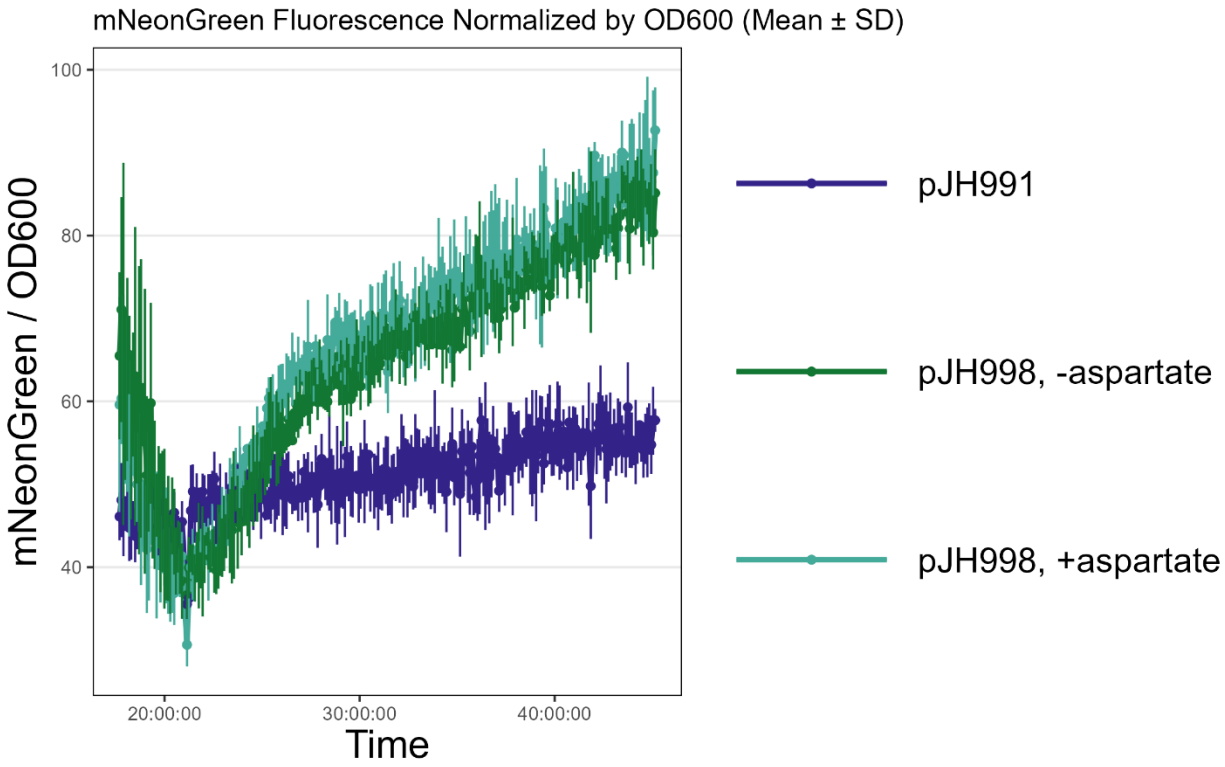


Figure 22 – Plate reader mNeonGreen fluorescence data of pJH998 and pJH991. The issues with growth rate seemingly caused the expression of fluorescence to be mainly attributable to noise in the system, with max fluorescence values below 100AU, and no discernible dynamic range to report.

As can be seen in the above graphs, pJH991 had an almost-flat OD600 curve, and pJH998 grew rather slowly. Based on its slow growth rate, and thus a lower fluorescence output, and on analysis of the growth of these cultures in LB, it was decided to make other variants to determine which part was causing growth issues in supplemented M9 media. As both the full circuit and the shorter positive control circuit were experiencing growth issues, we aimed to alter parts shared between them to observe effects on growth.

3.2.3.2 Initial Growth Troubleshooting

For the goal of troubleshooting slow growth in liquid media, a variety of plasmids were made for the purpose of determining the root cause.

3.2.3.2.1 pJH991 and pJH998 Derivatives

As slowed growth was seen in both pJH991 and pJH998, we made plasmid derivatives of both for each set of changes made for growth analysis.

3.2.3.2.1.1 pJH992 and pJH999

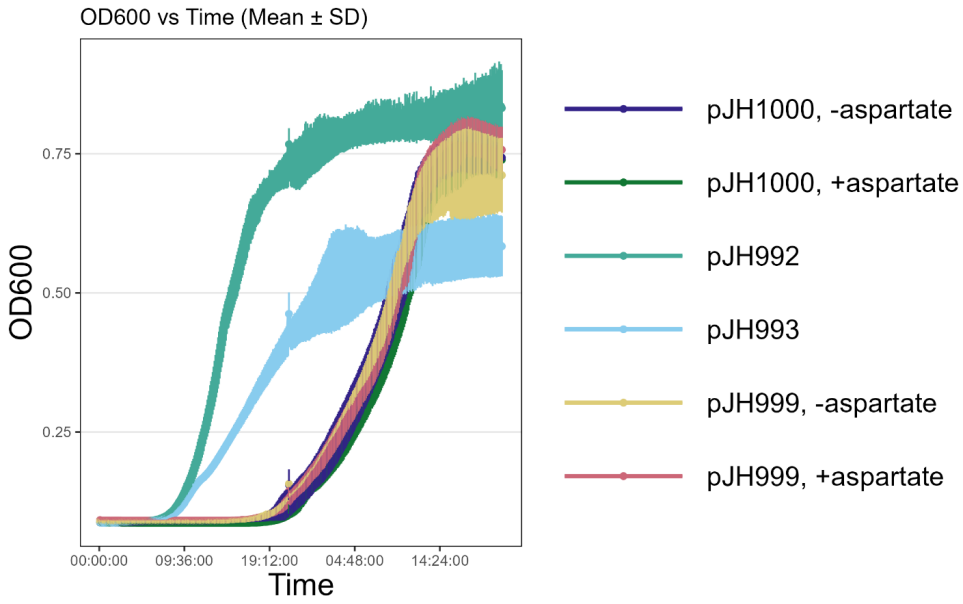
The first set of derivatives made, with pJH992 being the positive control and pJH999 being the locked-on sorting circuit, we kept the strong RBS but switched the fluorescent protein output from mNeonGreen to sfGFP, resulting in context-dependent strength of the RBS lowering from a predicted translation initiation rate of 911,556 AU to 36,544 AU, thus still being a strong RBS. This would help determine if growth problems were caused by the combination of a strong RBS and mNeonGreen in this context.

3.2.3.2.1.2 pJH993 and pJH1000

The next set of designed troubleshooting plasmids were pJH993 and pJH1000, with pJH993 being the positive fluorescent control and pJH1000 being the locked-on sorting circuit. The changes made here were swapping the strong RBS and mNeonGreen for the RBS and sfGFP from pSR40.29, thus seeing if the issue was elsewhere on the plasmid.

We transformed the pJH991 derivatives into *E. coli* BW29655, and the pJH998 derivatives were co-transformed alongside an SHK plasmid into the same strain, and these were then subjected to plate reader assays to assess growth behavior. It was observed they had very slow growth rates in supplemented M9 (Fig. 23), with pJH992 reaching an OD600 of 0.2 in eleven hours after wells were started with an initial OD600 of 0.005, pJH993 taking thirteen hours, and pJH999 and pJH1000 taking around twenty-five to twenty-seven hours.

Figure 23 (next page) – Plate reader OD600 data of pJH992, pJH993, pJH999, and pJH1000. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly nine to twenty-eight hours before wells started to reach an OD600 of 0.2.

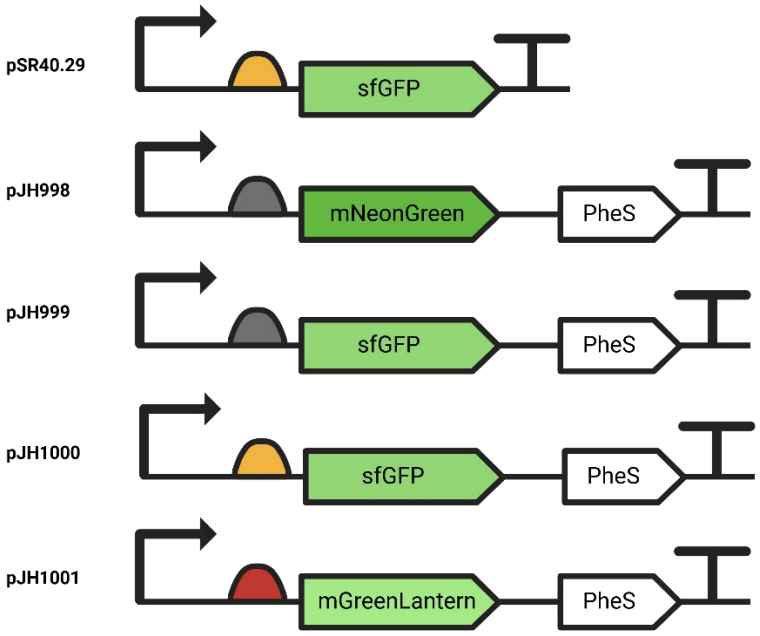


3.2.3.2.1.3 pJH994 and pJH1001

Also designed, assembled, and sequence-verified were pJH994 and pJH1001, with pJH994 being the positive control derivative. The design changes for these altered the existing RBS for a new RBS with a predicted translation initiation rate of 50,030 (AU), and with the fluorescent protein switched from mNeonGreen to mGreenLantern.

The comparison of these plasmids designed for troubleshooting growth issues can be seen in a color-coordinated manner in Fig. 24.

Figure 24 (next page) – Simplified SBOL Diagrams Comparing Plasmid RBS and Reporter Variants. These simplified SBOL diagrams are comparing pSR40.29, pJH998, pJH999, pJH1000, and pJH1001. Each construct features the $P_{cpcG2\text{-}172}$ promoter driving expression of a fluorescent reporter (sfGFP, mNeonGreen, or mGreenLantern) paired with RBSs of varying strength. RBS strength is color-coded by origin of RBS: orange (from pJH998), gray (from pSR40.29), and red (as in pJH1001). This series was used to evaluate how RBS strength and fluorescent reporter identity affected expression output and growth.



However, before pJH994 and pJH1001 were subjected to a plate reader assay, it was already determined and decided to move to the next phase of growth troubleshooting, as the interplay between RBS and fluorescent protein did not seem to be responsible for the growth issues observed.

3.2.3.3 Secondary Growth Troubleshooting

For the second phase of growth troubleshooting, we wanted to see what other parts present on the locked-on sorting circuit may be responsible for these issues in growth. For this purpose, the plasmids we made to test these issues were all derived from pJH1000, as pJH1000 is closest in design to pSR40.29, and pSR40.29 did not show evidence of growth issues in supplemented M9 media or in LB media.

3.2.3.3.1 pJH1000 Derivatives

For the purposes of growth troubleshooting, we decided to alter certain parts of the circuit to ascertain if they were a major cause for the growth issues in the context of this version of the locked-on removal circuit. The parts we chose to alter for the purposes of generating derivatives of pJH1000 were the chimeric response regulator (RR), the presence of a fluorescent protein reporter, and the PheS* selection toxin system.

3.2.3.3.1.1 pJH1002, pJH1003, pJH1004

We implemented the creation of these troubleshooting derivatives by PCR amplifying all but the desired deletion from pJH1000. For pJH1002, everything except the chimeric response regulator was amplified; for pJH1003, the RBS and gene for PheS were removed, though a 51 base pair scar remained after the stop codon of sfGFP, thus having sfGFP be the only gene under the control of the RR-inducible output promoter cpcG2-172, similar to how it is in pSR40.29; for pJH1004, the RBS and sfGFP were removed scarlessly, leaving the gene for the PheS toxin as the only output of the cpcG2-172 promoter. A simplified comparison of these derived circuits is shown in Fig. 25.

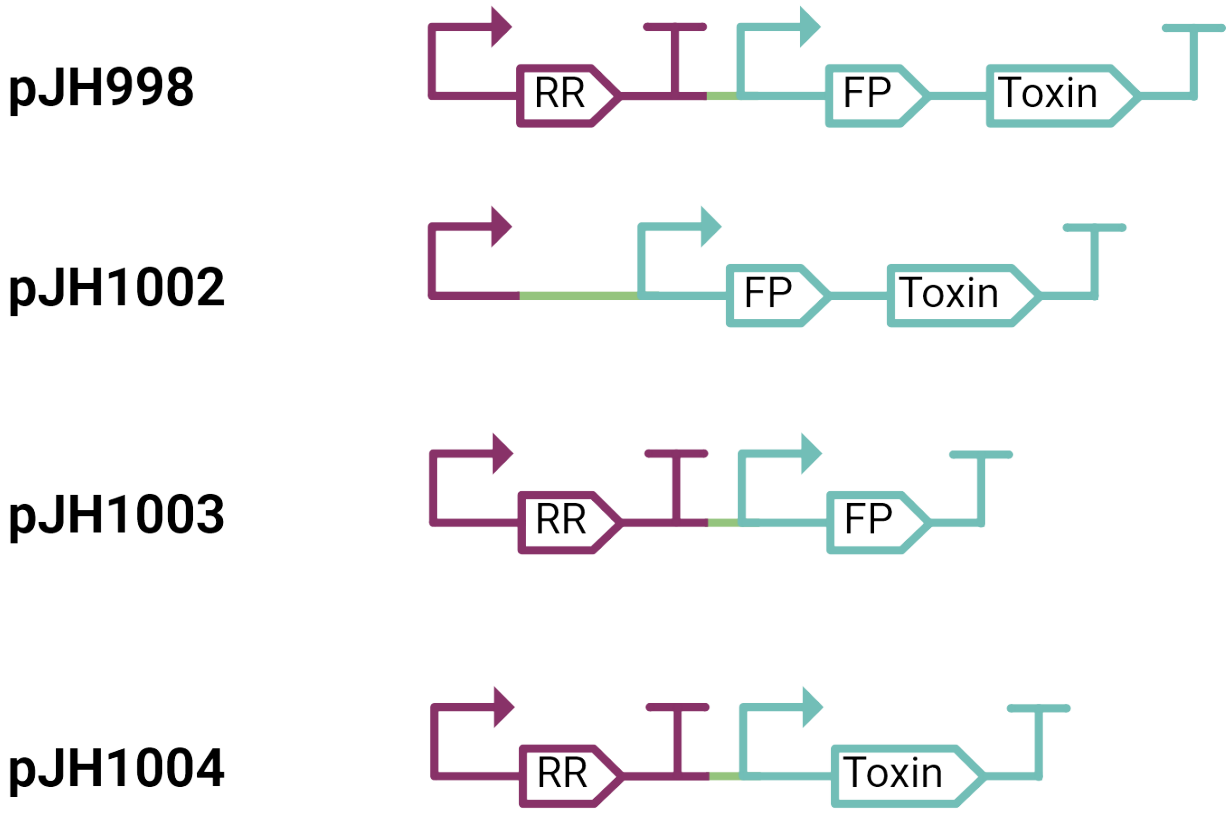


Figure 25 – Simplified SBOL Diagram Showing Alterations for pJH1002, pJH1003, and pJH1004. Each version has one part removed when compared to pJH998. pJH1002 has OmpR/CcaR removed, pJH1003 has PheS removed, and pJH1004 has sfGFP removed. This series was used to evaluate how the presence or absence of certain parts affected growth rate.

These were all assembled, sequence-verified, and co-transformed alongside an SHK plasmid into *E. coli* BW29655. These transformed strains all grew at an acceptable level in LB media, allowing glycerol stocks to be generated. Overnight supplemented M9 media overnight cultures were inoculated for the purpose of employing a plate reader assay the following day to analyze their growth rates; however, in the morning after the standard 16-18 hours of overnight growth, the OD600 values for these cultures were very low, indicating growth issues in supplemented M9 media. Thus no plate reader assay was run, as this indicated the experienced growth issues were caused by portions elsewhere in the plasmid.

3.2.3.3.1.2 Decision Against pJH998 Derivatives

In summary, as pJH998 nor any of its derivatives grew well in supplemented M9 media, and pSR40.29-LRv2 was never tested in supplemented M9 in the context of the *E. coli* BW29655 strain, thus meaning it was unknown whether it incurred growth issues, the decision was made to move to a more simplified circuit very closely derived from pSR40.29.

3.2.4 Locked-On Sorting – Fourth General Design

As all derivatives of pJH998 had growth issues in supplemented M9 media, we sought to design new versions of the locked-on sorting circuit based on the pSR40.29 plasmid with the main goal of enhancing the utility of the fluorescent readout while minimally impacting the growth rate in *E. coli* BW29655. We opted to not include a stringent life-death selection system on these versions of the locked-on sorting circuit, instead planning to rely on restriction enzyme cutsites and alternative fluorescent proteins for downstream removal of the locked-on sorting plasmid.

3.2.4.1 Derivatives of pSR40.29 with Small Modifications

Reminiscent of earlier plasmids designed for troubleshooting, to enhance the utility of the fluorescent readout, we made three separate variants that differed from the original pSR40.29 by switching out the RBS, the fluorescent protein, or both, along with potential additions of BsaI cutsites; these variants were pSH429, pSH430, and pSH433 (Fig. 26).

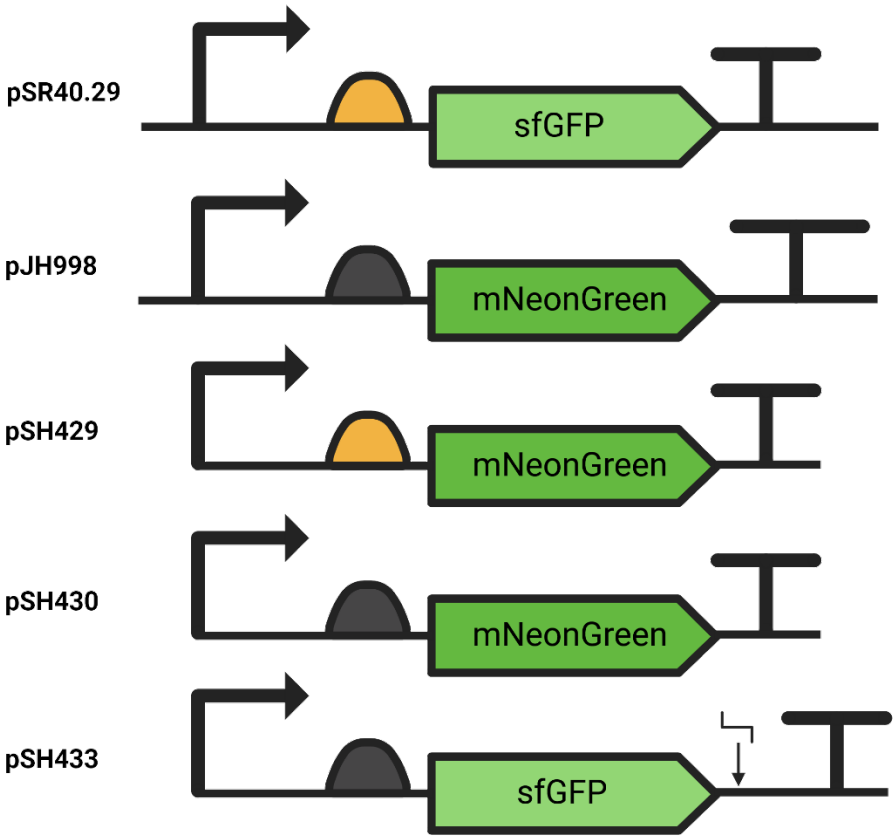


Figure 26 – Simplified SBOL diagrams showing major alterations for pSH429, pSH430, and pSH433, and comparing these to pSR40.29 and pJH998. Each construct features the $P_{cpcG2\text{-}172}$ promoter driving expression of a fluorescent reporter (sfGFP or mNeonGreen) paired with RBSs of varying strength. RBS strength is color-coded by origin of RBS: orange (strongest, from pJH998), or gray (weakest, from pSR40.29). This series was used to evaluate how RBS strength and fluorescent reporter identity affected expression output in absence of other parts from pSR40.29-LRv2 being present.

3.2.4.1.1 pSH429

The first derivative design was pSH429. The only change made from pSR40.29 was swapping mNeonGreen in to replace sfGFP, with the RBS present in pSR40.29 being kept. The fluorescent properties of mNeonGreen are preferable to those of sfGFP for our purposes, as mNeonGreen offers a higher quantum yield, greater brightness, and a shorter maturation time. After assembly and sequence verification, this was co-transformed alongside an SHK plasmid into *E. coli* BW29655.

3.2.4.1.2 pSH430

The next derivative design was that of pSH430, where the changes from pSR40.29 included swapping out both sfGFP and its RBS for mNeonGreen and the RBS used with it in pJH998. After assembly and sequence-verifying, pSH420 was co-transformed along with an SHK plasmid into *E. coli* BW29655.

Both pSH429 and pSH430 were then subjected to characterization via a plate reader assay (Fig. 27). While there was separation between the absence and presence of ligand conditions for both pSH429 and pSH430, with an adequate signal-to-noise ratio, the raw fluorescence signal, both prior to and after normalization, was so low as to raise concern about the maximum dynamic range available for these systems.

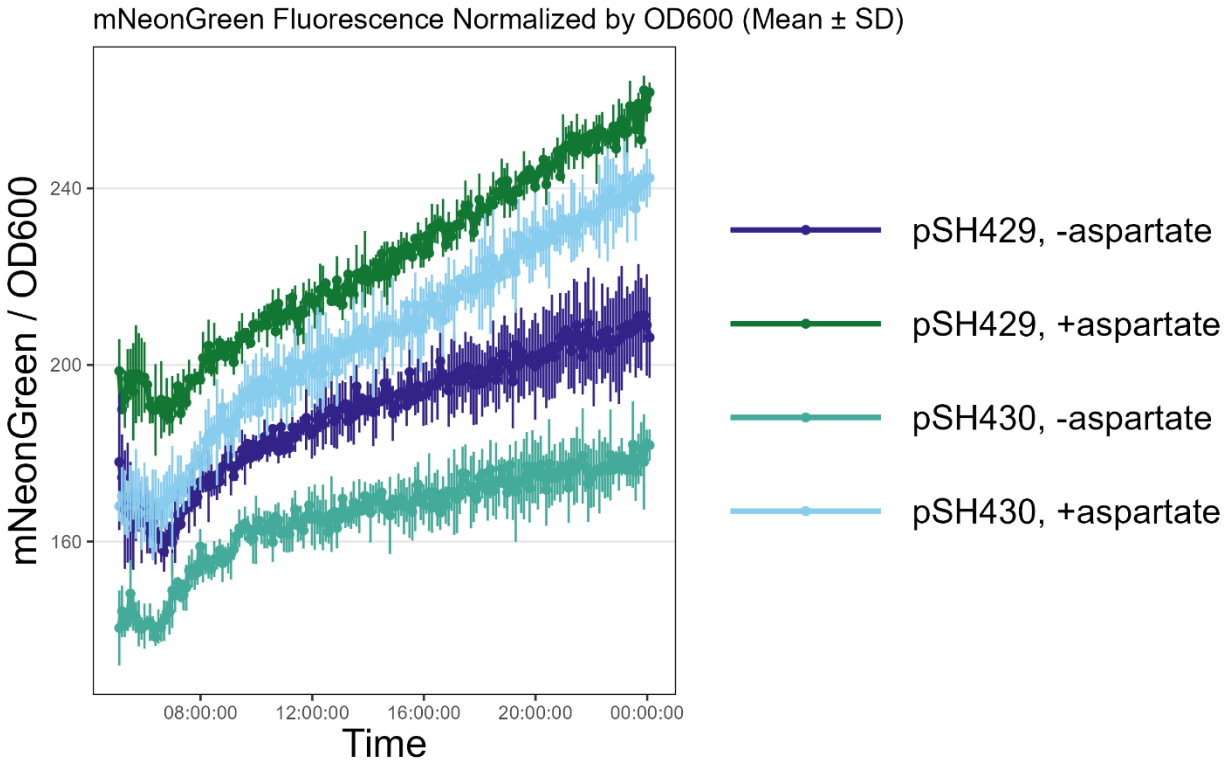


Figure 27 – Plate reader data of mNeonGreen fluorescence data of pSH429 and pSH430. The timeframe for maximum signal-to-noise ratio starts about seven hours into the experiment and continues for several hours, with a peak dynamic range of 60 AU at nine hours for both systems, a max value for these systems with ligand present during this timeframe in the 170-250 AU range, and a signal-to-noise ratio of 1.27 for pSH429 and 1.26 for pSH430.

3.2.4.1.3 pSH433

The third variation in this round was pSH433, which included the alterations of switching in the strong RBS used for mNeonGreen in pJH998 in place of the one pSR40.29 used for sfGFP, while keeping sfGFP as the fluorescent protein, and adding an additional BsaI cutsite after the end of sfGFP and before its terminator. After sequence verification, this was then co-transformed alongside an SHK plasmid into *E. coli* BW29655, and as they both used sfGFP as their fluorescent reporter, a plate reader assay was run to directly compare pSH433 to pSR40.29. While there was separation between the presence and absence of ligand for pSH433, its signal-to-noise ratio was worse than that of pSR40.29 (Fig. 28), in addition to its much lower fluorescence values.

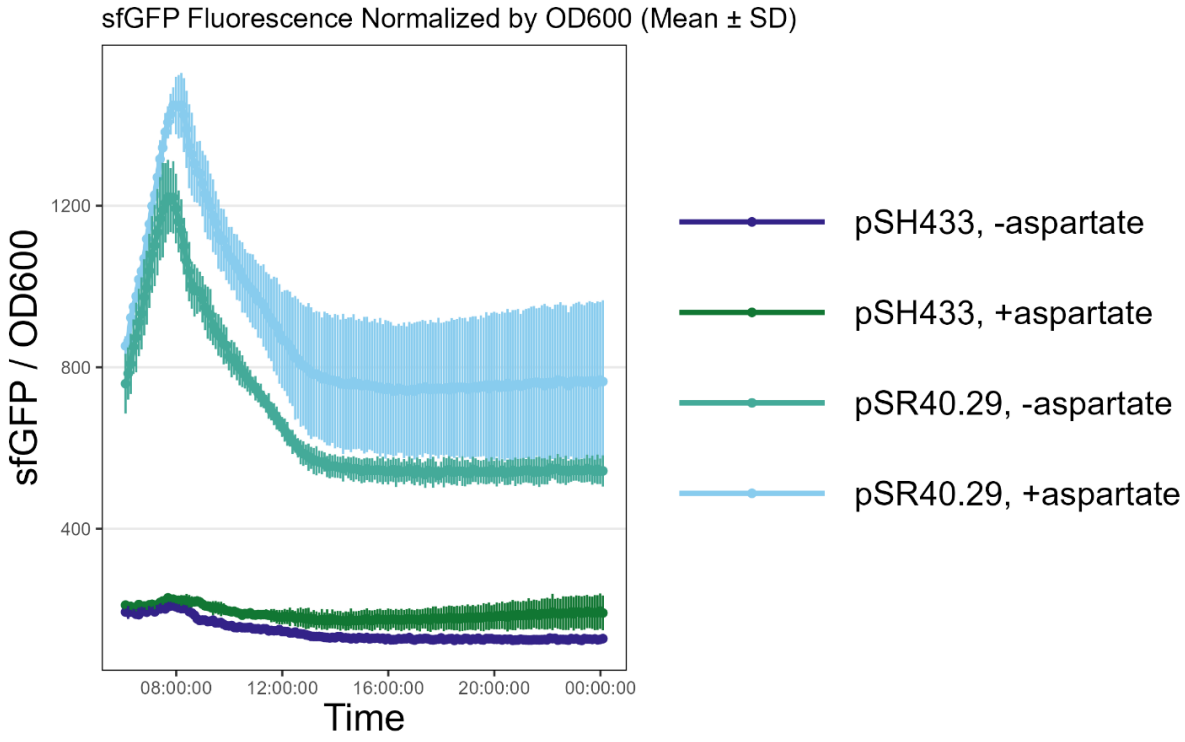


Figure 28 – Fluorescence Comparison of pSR40.29 and pSH433. Plate reader data of sfGFP fluorescence data for comparing pSR40.29 and pSH433; the timeframe for maximum signal-to-noise ratio starts about eight to nine hours into the experiment and continues for several hours, with a peak dynamic range of 330 AU at eight hours for pSR40.29 and a peak dynamic range of 60AU for pSH433, a max brightness value with ligand present during this timeframe of 1,440 AU for pSR40.29 and 220 AU for pSH433, and a signal-to-noise ratio of 1.35 for pSH429 and 1.29 for pSH433.

Thus, for pSH429, pSH430, and pSH433, though they successfully did not cause issues in the growth rate for our strain of *E. coli*, the desired enhancement in fluorescence was not only absent but worse than the fluorescence characteristics in pSR40.29. Thus, the decision was made to move forward with pSR40.29's combination of RBS and sfGFP. With this in mind, some level of alteration was desired so as to facilitate removal of the locked-on sorting plasmid for downstream experiments.

3.2.5 Locked-On Sorting – Fifth General Design

For the next iteration of designs, our objective was simple: add more BsaI restriction sites so the plasmid may be digested and not remain in the populations after a fluorescence-based sorting approach is used to separate locked-on variants from the rest of the libraries.

3.2.5.1 pBJ23 and pBJ232

The two designed variants are each only a single alteration away from pSR40.29, as they both are pSR40.29 but with another BsaI restriction site added to the plasmid; pSR40.29 already contains a single BsaI restriction site between the CmR promoter and the pLtetO-1 promoter. For pBJ23, a second restriction site was added between the origin and the T0 terminator, and for pBJ232, the second restriction site was added between the B0015 terminator for sfGFP and the J23106 promoter for TetR (Fig. 29).

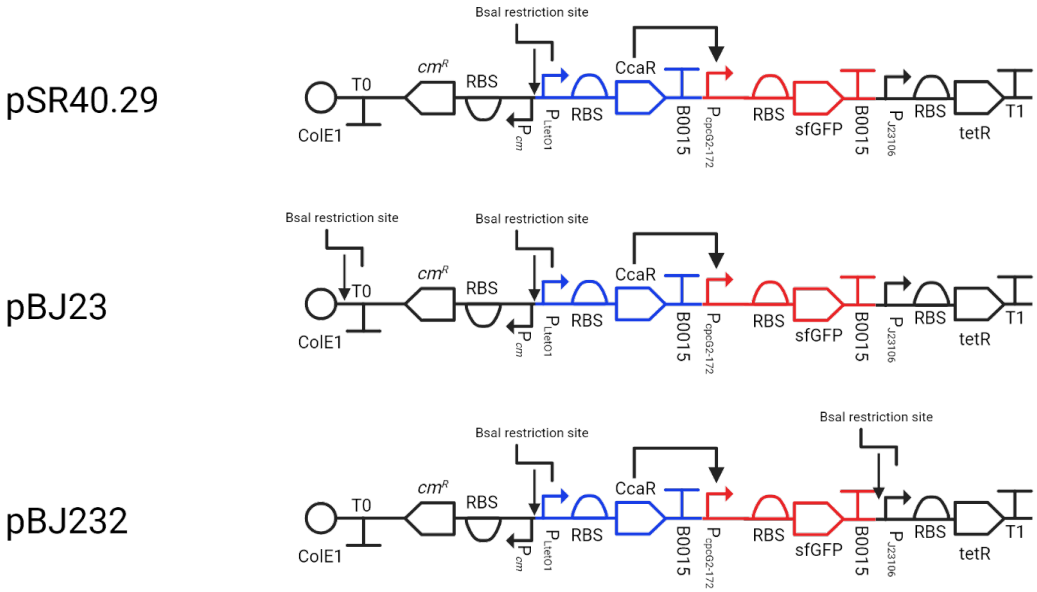


Figure 29 – SBOL Diagram Comparing pSR40.29, pBJ23, and pBJ232. These all have the same parts as pSR40.29, but pBJ23 and pBJ232 each contain an additional BsaI restriction site when compared to pSR40.29.

Both pBJ23 and pBJ232 were assembled, sequence-verified, and co-transformed alongside an SHK plasmid into *E. coli* BW29655. After this, these strains were then employed in plate reader assays that were run to characterize the signal-to-noise ratio for pBJ23 and pBJ232 (Fig. 30).

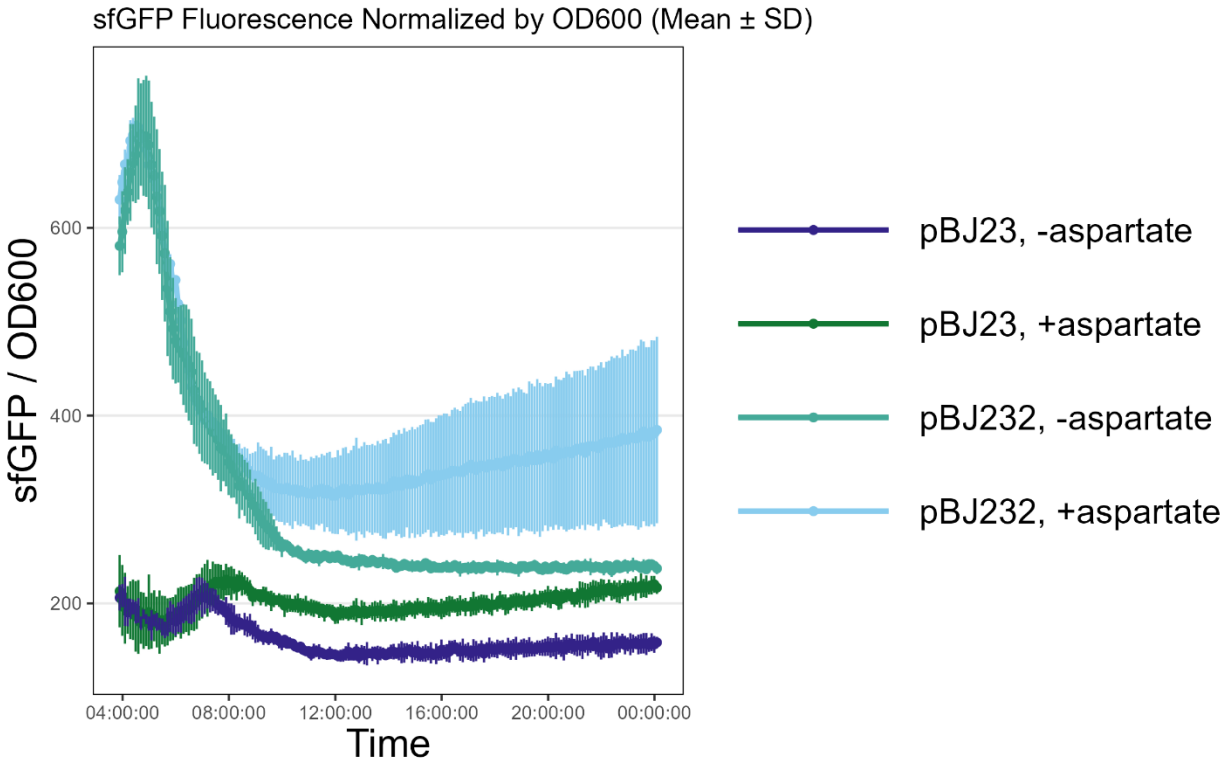


Figure 30 – Plate reader sfGFP fluorescence data of pBJ23 and pBJ232. The timeframe for maximum signal-to-noise ratio starts about nine to ten hours into the experiment and continues for several hours, with a peak dynamic range of 70 AU at fourteen hours for pBJ23 and a peak dynamic range of 98 AU for pBJ232, a max brightness value with ligand present during this timeframe of 300 AU for pBJ23 and 380 AU for pBJ232, and a signal-to-noise ratio of 1.31 for pBJ23 and 1.32 for pBJ232.

As pBJ232 had decent separation between the conditions of absence versus presence of ligand and thus a decent signal-to-noise ratio while having decent fluorescence values, another plate reader assay was run to compare it directly to pSR40.29 (Fig. 31). It was observed that pSR40.29 had both a better signal-to-noise ratio, which persisted for a longer period of time, and also a stronger signal in general than pBJ232.

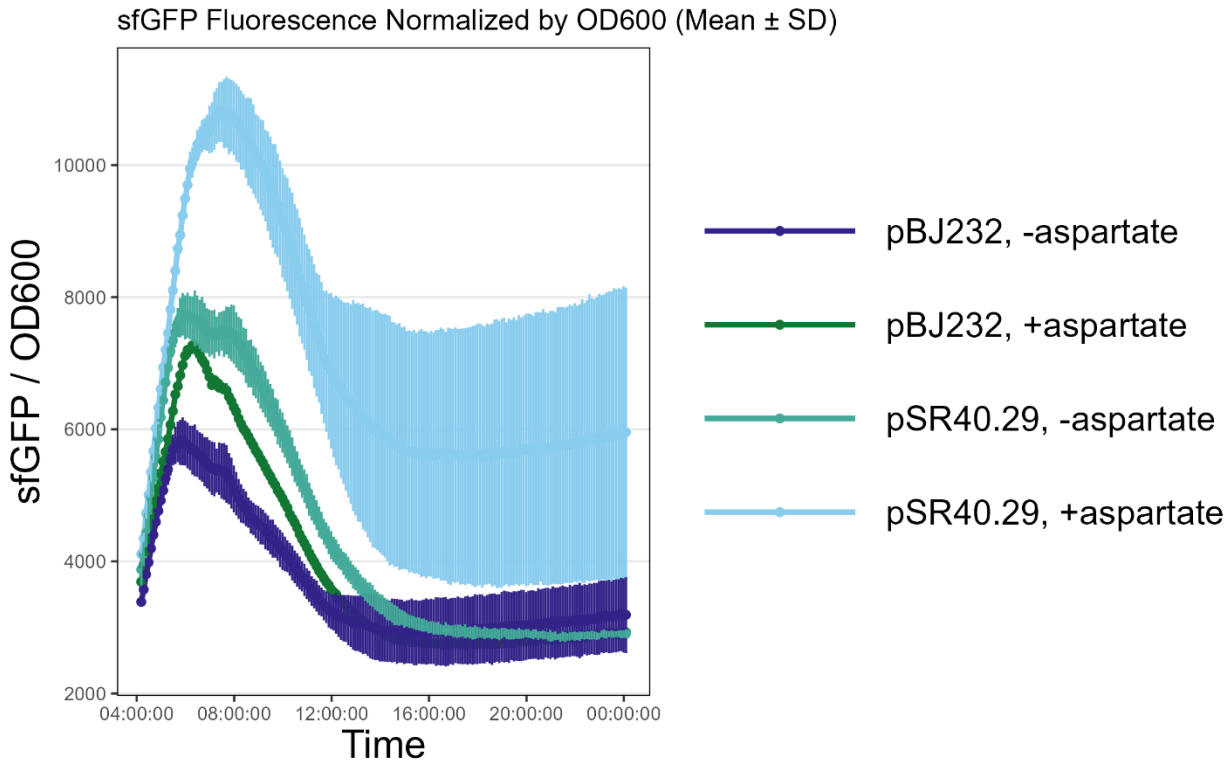


Figure 31 – Fluorescence Comparison of pSR40.29 and pBJ232. Plate reader data of sfGFP for comparing pSR40.29 and pBJ232; the timeframe for maximum signal-to-noise ratio starts about six hours into the experiment and continues for several hours, with a peak dynamic range of 1,350 AU at seven hours for pBJ232 and a peak dynamic range of 3,364 AU for pSR40.29, a max brightness value with ligand present during this timeframe of 6,745 AU for pBJ232 and 10,803 AU for pSR40.29, and a signal-to-noise ratio of 1.25 for pBJ232 and 1.83 for pSR40.29.

3.2.5.2 Locked-On Sorting – Final Plasmid Choice

In summary, as pSR40.29 had a more robust fluorescent output, both in terms of a larger, more persistent signal and a better signal-to-noise ratio, than pBJ232, it was decided to move forward with pSR40.29 as the plasmid to use for sorting variants into different bins to determine their level of activation in the absence of ligand, and thus which variants are locked-on.

3.3 Chemical Screening

After locked-on variants are sorted out from the library, the library would then be screened against a variety of chemicals for a more comprehensive characterization of ligand response, and this would entail removing any variants that show no response upon ligand

introduction; thus, a separate genetic system would need to be designed for this purpose. The goal of this system, then, is to remove cells containing variants that are inactive in the presence of ligand and retain cells whose variants are activated upon ligand introduction. Ideally, this entails the combination of both stringent life-death selection and an alternative fluorescence-based readout as a backup system.

By incorporating a dual-antibiotic selection strategy for life-death selection, cells would need to pass two different sets of checks to survive, thus lowering the likelihood of non-functional variants escaping selection.

3.3.1 pSR40.29-dualAB

For our system built to characterize variants regarding their response to ligands, we initially built our circuit design from pSR40.29 by altering it into a new plasmid called pSR40.29-dualAB (Fig. 32). This plasmid was designed with the goal of genomic integration in mind.

3.3.1.1 Life-death selection

For life-death selection, we designed a dual-antibiotic system using the genes conferring resistance to spectinomycin and chloramphenicol, as these are both titratable antibiotics that allow us to fine-tune the stringency of selection.

3.3.1.2 RiboJ Insulator

A RiboJ insulator sequence (Clifton et al., 2018) was placed between the spectinomycin resistance gene and sfGFP. This avoids translational coupling and negates any unintended impact of $spec^R$ expression on the fluorescent reporter, preserving a clean and independent readout.

3.3.1.3 Circuit Orientation

As cells often like to mutate circuits so as to escape antibiotic selection, the life-death selection circuit necessitated a more complicated design. Thus, on one strand, the RR-inducible output promoter was used to control the expression of spectinomycin resistance and sfGFP, and the single terminator present after sfGFP in pSR40.29 was swapped for a double terminator, with the strong, natural terminator ECK120029600 and a strong, synthetic terminator of L3S1P12. After this, but now reversed and on the opposing strand, a second operon was designed that also employed the RR-inducible cpcG2-172 promoter to drive the expression of a chloramphenicol resistance gene, and using two synthetic terminators, these being L3S1P56 and L3S3P00 (Y.-J. Chen et al., 2013).

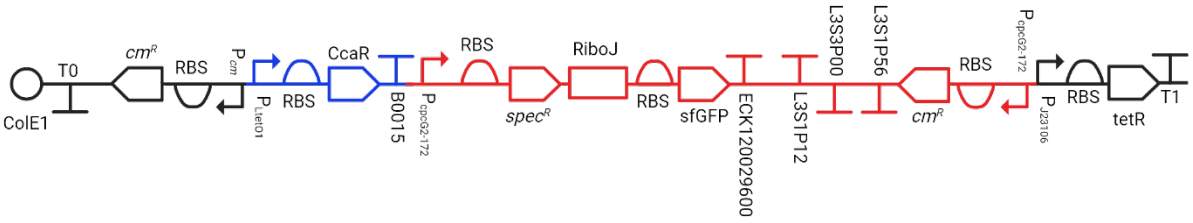


Figure 32 – SBOL Diagram of pSR40.29-dualAB. This plasmid contains a $P_{cpcG2\text{-}172}$-driven operon regulated by the response regulator CcaR, which is itself repressed by TetR. In the presence of aTc, upon phosphorylation by an SHK, CcaR activates expression of an output operon containing a spectinomycin resistance gene *spec*^{R} and sfGFP, separated by a RiboJ insulator. The operon also includes a second *cm*^{R} cassette driven by a separate $P_{cpcG2\text{-}172}$ promoter for dual-antibiotic selection. Terminator variants (L3S3P00, L3S1P56, L3S1P12, etc.) are included between operon components to reduce readthrough and tune expression. TetR, expressed by the constitutive promoter P_{J23106}, represses P_{LtetO1}, and its repression is inhibited by aTc, enabling circuit activation. Colors indicate separate transcriptional units; blue: *ccaR*; red: output operons; black: *tetR* and backbone elements.

Following assembly and sequence verification, pSR40.29-dualAB was co-transformed alongside an SHK plasmid into BW29655, and a plate reader assay was for a basal characterization of the system, measuring both sfGFP expression and the strength of life-death selection in the presence and absence of stimuli and spectinomycin (Fig. 33), as chloramphenicol resistance is also present on the backbone; measurable separation was observed between the no stimuli and stimuli present conditions.

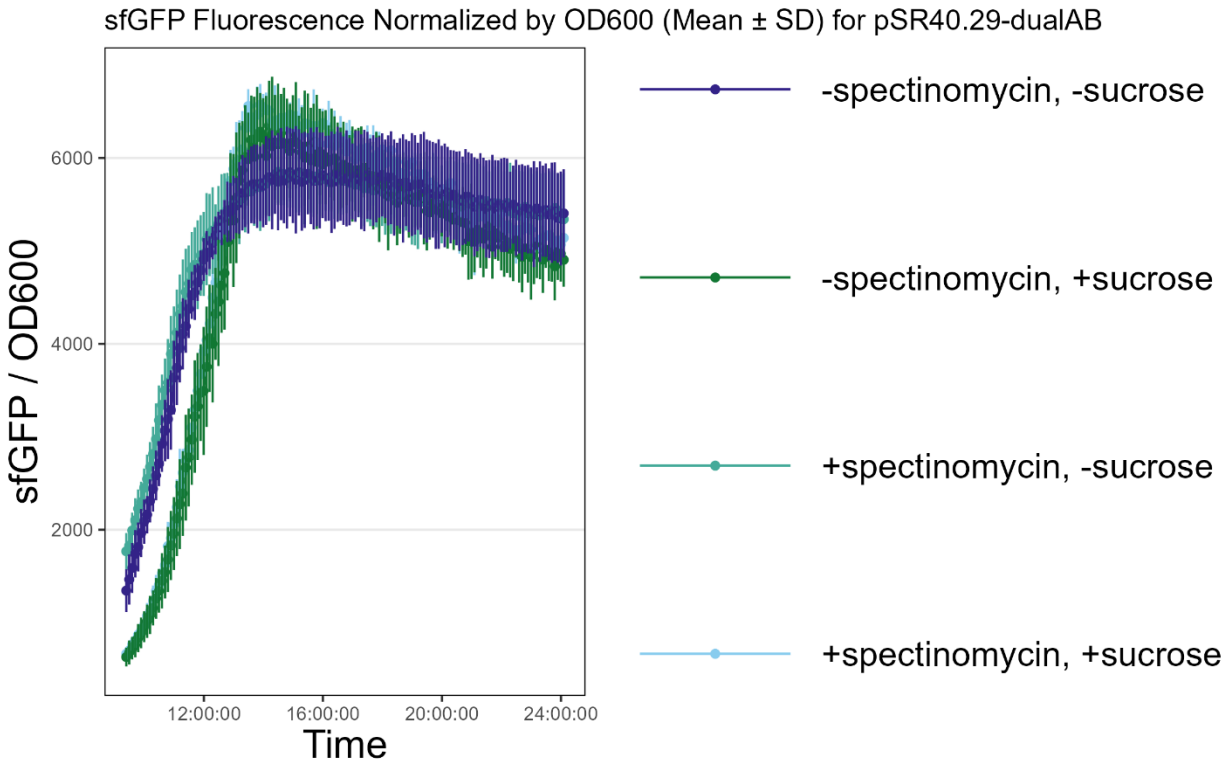


Figure 33 – Plate reader sfGFP fluorescence data comparing presence and absence of pSR40.29-dualAB for different ligand and spectinomycin antibiotic selection conditions. The timeframe for maximum signal-to-noise ratio starts about thirteen hours into the experiment and continues for several hours, with a peak dynamic range of 886 AU, a max brightness value with ligand present during this timeframe of 5,850-6,570 AU for these conditions, and a signal-to-noise ratio of 1.16.

However, for the same assay, when analyzing the OD600, the presence of stimuli, this being sucrose, seemed to have a larger effect on the growth rate of the cells than did the presence of spectinomycin (Fig. 34).

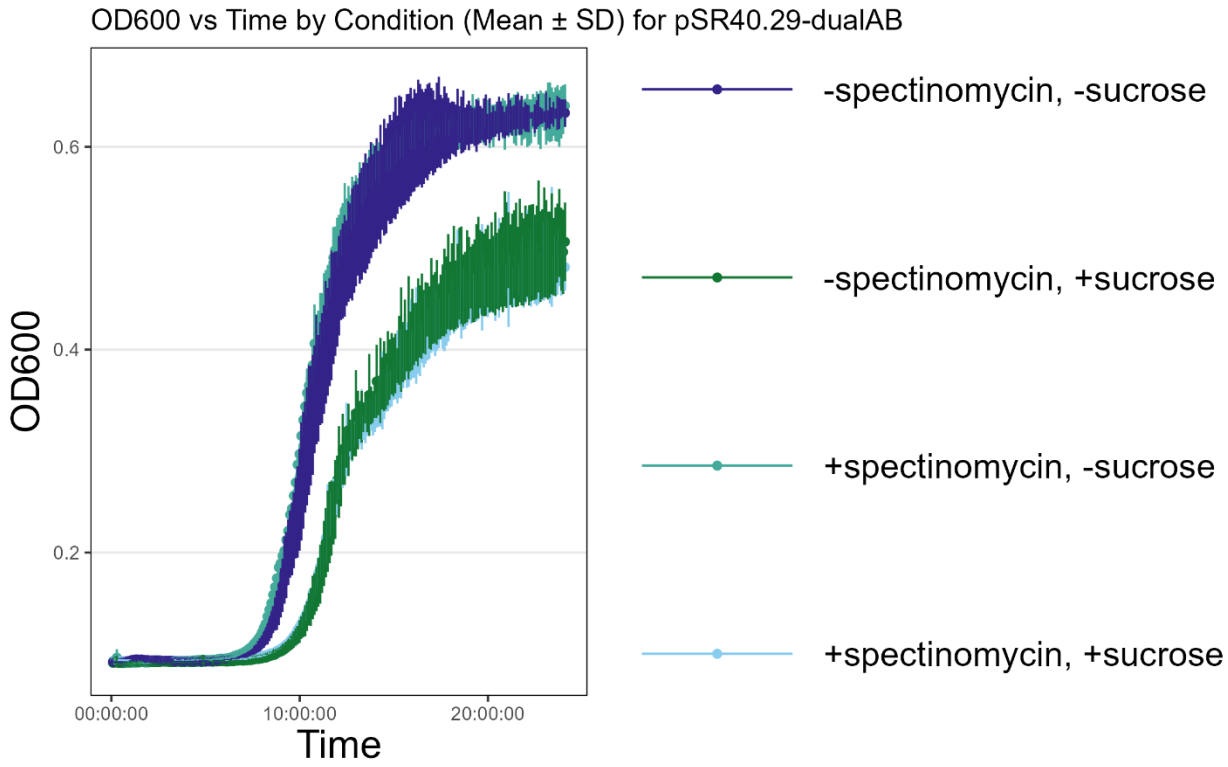


Figure 34 – Plate reader OD600 data comparing presence and absence of pSR40.29-dualAB for different ligand and spectinomycin antibiotic selection conditions. With a starting OD600 of 0.0005 for all conditions in this experiment, it took roughly nine to eleven hours before wells started to reach an OD600 of 0.2.

Though steps were made towards the integration of this construct into the genome of BW29655, which would thus allow a more robust measurement of the efficacy of dual versus single antibiotic selection, however, before this was reached, the low expression levels of sfGFP were noticed for integrated 40-kilFlip, with potential consequences that would apply to this system as well.

3.4 Plasmid Designs Summary

This chapter described the rationale, design, and evaluation of synthetic circuits for high-throughput screening of chimeric sensor histidine kinases (SHKs), with a focus on distinguishing ligand-dependent signaling from constitutive activation or complete inactivity. The development and use of genetic circuits capable of reliably distinguishing locked-on variants from catalytically dead or functional variants is central to accelerating large-scale SHK

deorphanization. Our approach prioritized three key design goals: (1) maximizing signal-to-noise ratio to ensure clear differentiation between ligand-induced and background activity, (2) minimizing escape frequency to reduce the persistence of spurious or locked-on variants, and (3) maximizing dynamic range between basal and ligand-induced outputs to enable quantitative discrimination between weak and strong SHK activation. Secondary considerations included circuit modularity, genomic integration, and minimization of metabolic burden.

Through the course of this work, we engineered and characterized multiple iterations of genetic circuit architecture, each aiming to improve functional screening fidelity relative to previous designs. These included constructs employing kill-switches for life-death selection, alternative ribosome binding sites and fluorescent protein combinations for tuning expression, and features enabling plasmid removal. Although pSR40.29 lacks a life-death selection module or plasmid removal mechanism, more stringent selection schemes were explored and consistently underperformed. Common issues included leaky expression, metabolic instability, and undesirable growth defects, with no designs matching the reliability of the fluorescence-only-based circuit of the pSR40.29 plasmid.

Built on a pBR322-derived backbone with moderate copy number and chloramphenicol resistance, pSR40.29 couples EnvZ- and chimera-EnvZ-dependent phosphorylation to sfGFP expression via a chimeric OmpR-CcaR response regulator. This architecture enabled consistent, quantifiable readout of SHK activity through fluorescence alone. When co-expressed with chimeric SHKs in the ΔompR ΔenvZ BW29655 strain of *E. coli*, pSR40.29 provided a usable dynamic range under supplemented M9 minimal media conditions. Ultimately, although other designs incorporated more complex selection mechanisms, pSR40.29 remained the most effective system for high-throughput screening.

While this fluorescence-based strategy lacks the stringency of survival-based assays, it is sufficient for characterizing locked-on variants via fluorescence-activated cell sorting (FACS) and served as the foundation for downstream genotype–phenotype analysis. Although more stringent circuits remain a goal for future integration and iterative refinement, pSR40.29 provided the necessary performance for initial SHK deorphanization efforts due to it offering the best balance of sensitivity, reliability, and scalability. As such, it served as the primary platform for functional characterization of the chimeric SHK library described in the subsequent chapter.

3.5 Plasmid Designs Bibliography

Besmer, E., Market ,Eleonora, & and Papavasiliou, F. N. (2006). The Transcription Elongation Complex Directs Activation-Induced Cytidine Deaminase-Mediated DNA Deamination. *Molecular and Cellular Biology*, *26*(11), 4378–4385. https://doi.org/10.1128/MCB.02375-05

Chen, W., Li, Y., Wu, G., Zhao, L., Lu, L., Wang, P., Zhou, J., Cao, C., & Li, S. (2019). Simple and efficient genome recombineering using kil counter-selection in Escherichia coli. *Journal of Biotechnology*, *294*, 58–66. https://doi.org/10.1016/j.jbiotec.2019.01.024

Chen, Y.-J., Liu, P., Nielsen, A. A. K., Brophy, J. A. N., Clancy, K., Peterson, T., & Voigt, C. A. (2013). Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods*, *10*(7), 659–664. https://doi.org/10.1038/nmeth.2515

Clifton, K. P., Jones, E. M., Paudel, S., Marken, J. P., Monette, C. E., Halleran, A. D., Epp, L., & Saha, M. S. (2018). The genetic insulator RiboJ increases expression of insulated genes. *Journal of Biological Engineering*, *12*(1), 23. https://doi.org/10.1186/s13036-018-0115-6

Kuhlman, T. E., & Cox, E. C. (2010). Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research*, *38*(6), e92. https://doi.org/10.1093/nar/gkp1193

Landry, B. P., Palanki, R., Dyulgyarov, N., Hartsough, L. A., & Tabor, J. J. (2018). Phosphatase activity tunes two-component system sensor detection threshold. *Nature Communications*, *9*(1), 1433. https://doi.org/10.1038/s41467-018-03929-y

Lauritsen, I., Porse, A., Sommer, M. O. A., & Nørholm, M. H. H. (2017). A versatile one-step CRISPR-Cas9 based approach to plasmid-curing. *Microbial Cell Factories*, *16*(1), 135. https://doi.org/10.1186/s12934-017-0748-z

Maranhao, A. C., & Ellington, A. D. (2017). Evolving Orthogonal Suppressor tRNAs To Incorporate Modified Amino Acids. *ACS Synthetic Biology*, *6*(1), 108–119. https://doi.org/10.1021/acssynbio.6b00145

Murphy, K. C., & Marinus, M. G. (2010). RecA-independent single-stranded DNA oligonucleotide-mediated mutagenesis. *F1000 Biology Reports*, *2*, 56. https://doi.org/10.3410/B2-56

Reis, A. C., & Salis, H. M. (2020). An automated model test system for systematic development and improvement of gene expression models. ACS synthetic biology, 9(11), 3145-3156. https://doi.org/10.1021/acssynbio.0c00394

Schlake, T., & Bode, J. (2002, May 1). Use of Mutated FLP Recognition Target (FRT) Sites for the Exchange of Expression Cassettes at Defined Chromosomal Loci (world) [Research-article]. ACS Publications; American Chemical Society. https://doi.org/10.1021/bi00209a003

Schmidl, S. R., Ekness, F., Sofjan, K., Daeffler, K. N.-M., Brink, K. R., Landry, B. P., Gerhardt, K. P., Dyulgyarov, N., Sheth, R. U., & Tabor, J. J. (2019). Rewiring bacterial two-component systems by modular DNA-binding domain swapping. Nature Chemical Biology, 15(7), 690–698. https://doi.org/10.1038/s41589-019-0286-6

Thyer, R., Filipovska, A., & Rackham, O. (2013). Engineered rRNA Enhances the Efficiency of Selenocysteine Incorporation during Translation. Journal of the American Chemical Society, 135(1), 2–5. https://doi.org/10.1021/ja3069177

van Loenhout, M. T. J., van der Heijden, T., Kanaar, R., Wyman, C., & Dekker, C. (2009). Dynamics of RecA filaments on single-stranded DNA. Nucleic Acids Research, 37(12), 4089–4099. https://doi.org/10.1093/nar/gkp326

Vandervelde, A., Drobnak, I., Hadži, S., Sterckx, Y. G.-J., Welte, T., De Greve, H., Charlier, D., Efremov, R., Loris, R., & Lah, J. (2017). Molecular mechanism governing ratio-dependent transcription regulation in the ccdAB operon. Nucleic Acids Research, 45(6), 2937–2950. https://doi.org/10.1093/nar/gkx108

Vo, P. L. H., Ronda, C., Klompe, S. E., Chen, E. E., Acree, C., Wang, H. H., & Sternberg, S. H. (2021). CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. Nature Biotechnology, 39(4), Article 4. https://doi.org/10.1038/s41587-020-00745-y

Wannier, T. M., Ciaccia, P. N., Ellington, A. D., Filsinger, G. T., Isaacs, F. J., Javanmardi, K., Jones, M. A., Kunjapur, A. M., Nyerges, A., Pal, C., Schubert, M. G., & Church, G. M. (2021). Recombineering and MAGE. Nature Reviews Methods Primers, 1(1), 1–24. https://doi.org/10.1038/s43586-020-00006-x

Wieland, M., & Hartig, J. S. (2008). Improved Aptazyme Design and In Vivo Screening Enable Riboswitching in Bacteria. Angewandte Chemie International Edition, 47(14), 2604–2607. https://doi.org/10.1002/anie.200703700

4. LOCKED-ON SORTING

4.1 Introduction

While sensor histidine kinase (SHK) architecture and domain structures are well-characterized (Berntsson et al., 2017), fundamental gaps remain in understanding how certain sequence features and perturbations in structure influence basal signaling states, particularly in engineered SHKs where chimeras are generated via fusing N-terminal domains to heterologous kinase modules (Lehning et al., 2017). If structural determinants governing signal transduction are perturbed, chimeras can end up in constitutively active ("locked-on") or inactive ("locked-off") conformations; if locked-on variants are not removed from libraries, they can confound downstream assays by always remaining active, generating false positives.

This chapter demonstrates our locked-on sorting (LOS), a high-throughput assay for profiling the basal activation states of chimeric SHKs through fluorescence-activated cell sorting (FACS) and computational analysis. Traditional approaches to characterizing SHKs face critical limitations, such as structural conformation studies lacking throughput for library-scale analysis, and previous chimeric designs often producing non-functional designs due to improper phasing at fusion junctions (Meier et al., 2024). Our strategy addresses these challenges through four key innovations:

(1) Phase-aware library design incorporating up to eight helical phase variants per chimera, compensating for α-helical periodicity that governs signal transduction.
(2) Multiplexed FACS sorting coupled with barcode-based sequencing enabling quantification of inferred brightness per variant.
(3) Representing a large variety of sensor classes and organisms, providing a robust array of sensors for testing.
(4) Machine learning models for predicting SHK brightness from sequence features, allowing for functional enrichment in chimeric designs.

This chapter is structured in a manner to first establish the mechanistic basis for phase-dependent signaling in SHKs, following by descriptions of our analysis pipeline developed in-house. The following sections discuss how the interplay of helical phase and sensor class modulate basal activation states before culminating in a random-forest machine learning model. This chapter aims to provide a practical toolkit for the initial steps to accelerating SHK

deorphanization and new insights into phase-dependent signaling of SHKs through the lens of four libraries of chimeric SHKs.

4.1.1 Locked-On Sorting Goals

To accelerate the de-orphanization of sensor histidine kinases (SHKs), we established a multiplexed "locked-on" sorting workflow that quantifies basal activity for thousands of chimeric variants in parallel. By systematically varying both sensor class (Fig. 1) and the rotational fusion phase, we created a data set suitable for training predictive sequence-to-activity models that can pre-enrich functional designs before library construction. For this purpose, we transformed our pooled library of barcoded plasmids encoding different chimeric SHKs into the ΔompR ΔenvZ BW29655 *E. coli* host strain containing the plasmid pSR40.29, which is an RR plasmid designed for measuring SHK signaling output by coupling SHK phosphorylation events to expression of a sfGFP reporter (Table 1, Fig. 2). We grew the transformed cells in supplemented M9 minimal media with antibiotics, aTc, and IPTG to maintain plasmids and for circuit induction. Variants that exhibit differential levels of sfGFP expression, from basal activity to high fluorescence, are sorted via FACS into multiple bins with discrete brightness gates, and the distribution of each barcode across these gates provides an estimate of its intrinsic, ligand-independent activity. These cells are then grown on selective media and collected, with a portion set aside for DNA extraction. We PCR-amplified the barcode regions of the SHK plasmids and sent the amplicons for NovaSeq sequencing for determination of each barcode's bin distribution. This approach enables high-throughput quantification of SHK signaling states and identification of variants exhibiting locked-on phenotypes (Fig. 3).

Figure 1 (next page) – AlphaFold2-Multimer models for three chimeras representing the HisK_sensor, 2CSK_N, and PilJ sensor classes.

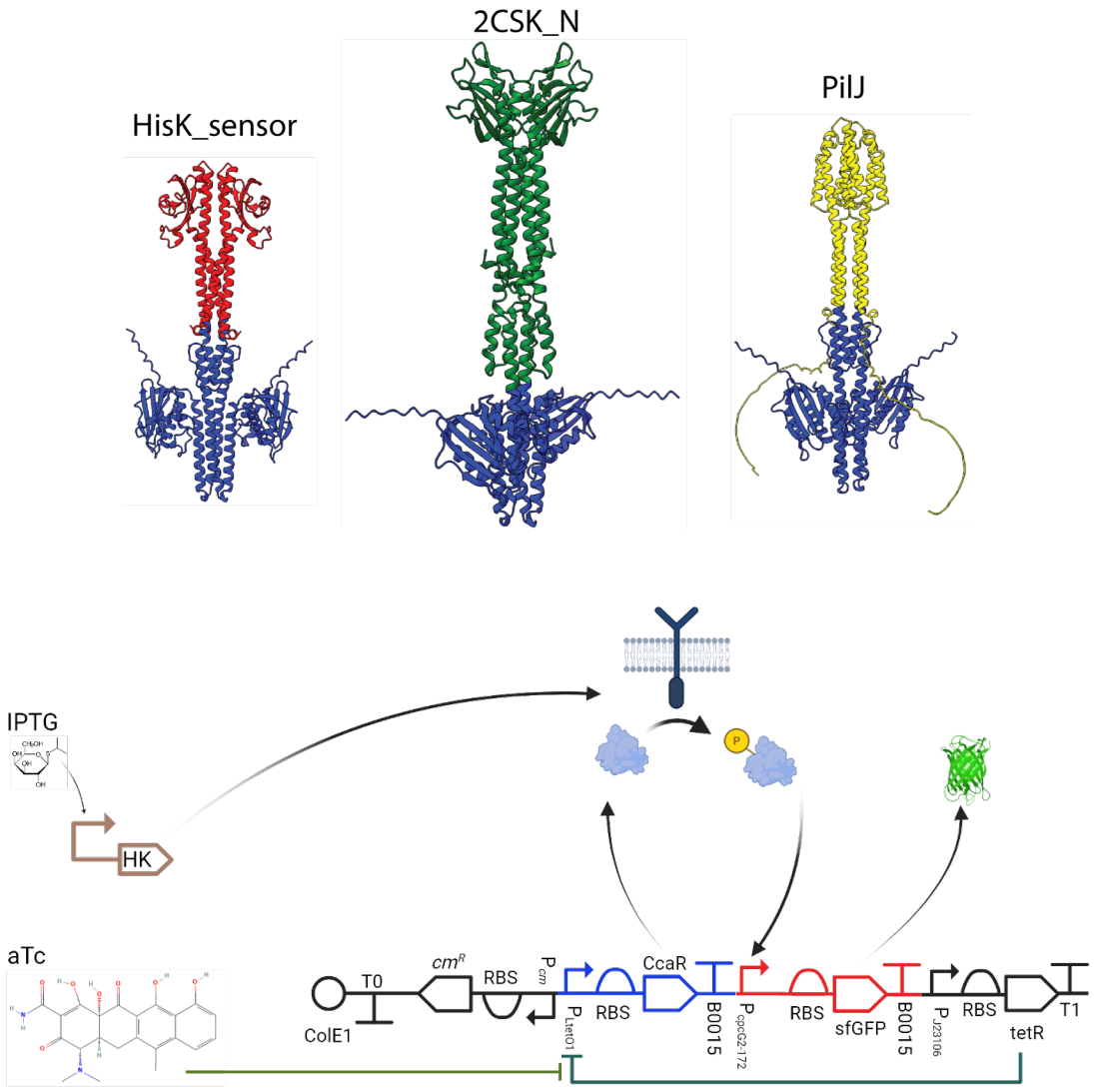


Figure 2 – Simplified molecular schematic of the LOS assay for testing ligand-independent activity of a chimeric SHK. IPTG induces expression of the SHK; if activated in the absence of ligand, the SHK will drive phosphorylation of the chimeric response regulator CcaR produced by pSR40.29. Phosphorylated CcaR drives sfGFP expression from the output promoter PcpcG2-172. The Synthetic Biology Open Language diagram at the bottom shows the genetic circuit of pSR40.29; it encodes three modules: (1) CcaR under TetR-repressible PLtetO1 control, (2) sfGFP under PcpcG2-172, and (3) constitutive tetR expression forming a negative feedback loop. Anhydrotetracycline (aTc) relieves TetR repression, enabling CcaR and sfGFP expression. SBOL glyphs denote genetic parts, with colors indicating separate transcriptional units: blue (ccaR), red (sfGFP), and black (tetR and backbone). This assay enables identification of SHK variants that are active in the absence of ligand.

Colony Counts and Bottleneck Values By Library - Post-Transformation into BW29655 Containing pSR40.29				
Library	Dilution Plate	Colonies Present	$\frac{CFU}{mL}$	Library Bottleneck (millions)
4oligo codon1	10^{-4}	360	$3.60 * 10^7$	10.8
4oligo codon2	10^{-4}	540	$5.40 * 10^7$	16.2
5oligo codon1	10^{-3}	1928	$1.93 * 10^7$	5.78
5oligo codon 2	10^{-3}	541	$5.41 * 10^6$	1.62

Table 1 – Colony counts and bottlenecks after transforming libraries into BW2655 strain which already contained pSR40.29

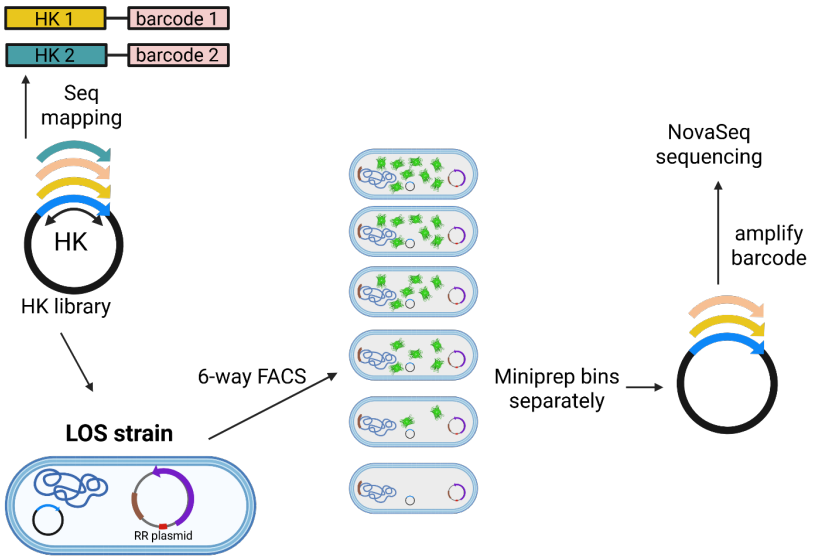


Figure 3 – Schematic overview of the locked-on sorting (LOS) assay for profiling basal activation of barcoded chimeric SHKs. A pooled library of plasmids encoding different chimeric SHKs, each linked to a unique DNA barcode, is transformed into the ΔompR ΔenvZ BW29655 *E. coli* host strain harboring the response regulator (RR) plasmid pSR40.29. Following transformation, cells are cultured in supplemented M9 minimal media with antibiotics for plasmid maintenance and aTc and IPTG for circuit induction. Variants that exhibit differential sfGFP expression levels—ranging from low basal activity to high fluorescence—are sorted via FACS into multiple bins. These cells are then grown and collected, with a portion set aside for DNA extraction. The barcode regions of the SHK plasmids are then amplified and sent for NovaSeq sequencing to determine the bin distribution of each variant. This approach enables high-throughput quantification of SHK signaling states and identification of variants exhibiting locked-on phenotypes.

4.2 Methods

4.2.1 Initial Library Design Considerations

Several considerations were taken into account when designing the library, chief among these being (1) drawing from a diverse set of organisms, (2) drawing from several classes of sensor domains, and (3) building fusion phase variants to raise the chances of building functional chimeras.

4.2.1.1 Diversity of Source Organisms

Histidine kinases are one of the most widespread protein families in nature, so we deliberately sampled across the tree of life when constructing our libraries. The four libraries collectively represent 1789 species, 79 classes, and six major kingdoms (*Pseudomonadati*, *Bacillati*, *Methanobacteriati*, *Thermotogati*, *Thermoproteati*, and *Fungi*), in addition to many strains and subspecies (Table 2).

Presorting Taxonomic Diversity - By Library					
Taxon	4oligo codon1	4oligo codon2	5oligo codon1	5oligo codon2	All Libraries Combined
Domain	3	2	3	3	3
Kingdom	6	4	5	5	6
Phylum	46	40	29	31	56
Class	61	53	43	48	79
Order	105	90	66	79	131
Family	193	169	104	125	250
Genus	403	342	180	250	615
Species	846	697	527	803	1789
Subspecies	11	11	12	18	32
Strain	122	101	98	155	301

Table 2 – Post-assembly, pre-sorting, taxonomic breakdown of source SHK organisms, illustrating representation across all domains of life and broad phylogenetic breadth at the kingdom-, class-, and species-level.

4.2.1.2 Sensor Class Diversity

To capture structural breadth, we included 17 distinct sensory domain classes in the libraries, ranging from predominantly α-helical architectures to mixed secondary-structure domains. However, a defining feature of all chosen sensors are class I HKs, containing only two

transmembrane domains and with their only linker domain being the HAMP domain. In addition to well-studied sensor classes, we also included two different sensor classes that are domains of unknown function (DUFs), DUF1189 and DUF3365, for which no member protein has yet been functionally annotated (Table 3).

Pre-Sorting Sensor Class Library Diversity					
SensorClass	4oligo codon1	4oligo codon2	5oligo codon1	5oligo codon2	All Libraries Combined
Sensor_TM1	0	0	27	73	88
sCache_like	77	67	3	8	92
sCache_4	127	101	22	37	180
sCache_3_2	71	70	13	19	111
RisS_PPD	62	46	9	13	96
PilJ	165	110	11	20	216
KinB_sensor	66	50	3	4	89
HK_sensor	72	49	2	4	84
HisK_sensor	91	74	0	0	99
DUF3365	10	10	79	123	151
DUF1189	54	48	0	0	60
dCache_1	71	60	4	9	94
CHASE8	71	61	9	17	98
CHASE3	59	41	138	261	371
ArlS_N	80	60	236	282	419
4HB_MCP_1	73	59	103	137	252
2CSK_N	65	51	0	1	79

Table 3 - Pre-sorting designs were made with proteins representing seventeen different sensor classes, with the number of different sensors present varying both via design and library assembly efficiency.

4.2.1.3 Fusion Phase Variants

A persistent challenge in chimeric SHK design is the choice of fusion point, which can bias constructs toward “locked-on” or “locked-off” states, as mentioned in introduction section 1.7.3. Many fusion points lead to locked-on or locked-off variants, with functional variants being rarer. As SHKs are thought to use different types of helical movement for their interdomain signal transduction, disturbing the helical phase could lead to disruptions in overall structure. For protein α-helices with ~3.6 residues per helical turn, each residue shift rotates the downstream

helix by ~100°. If a fusion point is chosen where the helical phases of the N-terminal and C-terminal fused portions are not aligned, then it is theorized the sensor has a higher likelihood of being artificially manipulated into a different basal signaling. In addition to the α-helical phase, the rotational phase of the coiled-coil dimer interface must also be considered, because each SHK functions as a parallel homodimer, with a disturbance in the heptad repeats between the HAMP and DHp domains, termed a "stutter," ensuring the coiled-coil structure is not continuous between these domains, enabling conversion between conformational signaling states (Schmidt et al., 2017). Mis-aligning this coiled-coil heptad repeat, or the stutter, can further twist the kinase core and is explicitly accounted for in our downstream analyses. For our designs, when choosing fusion points based on computationally defined domain boundaries, there is a small amount of uncertainty introduced on top of the intrinsic ignorance of the precise helical phase of the protein. Thus, for our designs, we incorporated between one and eight different phase variants of each chimera, and this was accomplished by including fusion alterations where residues were programmatically added or subtracted from either side of the fusion point below the HAMP domain (Table 4); this also served to sample sequence space around the stutter, where hydrophobic residues are found at the first and fourth positions of the heptad repeat, as well as the first position of the stutter (Table 5).

Sequence	Phase Variant
VKQ--LADDR	Wild Type envZ
XXX--LADDR	0 fusion
XXX---ADDR	-1c fusion
XXX----DDR	-2c fusion
XXXX-LADDR	+1n fusion
XXX-QLADDR	+1c fusion
XXXXXLADDR	+2n fusion
XX---LADDR	-1n fusion
X----LADDR	-2n fusion

Table 4 – Strategy for generating fusion-phase variants. For each chimera, residues were computationally inserted or deleted on one or both sides of the fusion junction.

Variant	Alignment																			
WT-EnvZ	*a*	*b*	*c*	*d*	*e*	*f*	*g*	-	-	a/d	e	f	g	a	b	c	d	e	f	g
0	*a*	*b*	*c*	*d*	*e*	*f*	*g*	-	-	a/d	e	f	g	a	b	c	d	e	f	g
-1c	*a*	*b*	*c*	*d*	*e*	*f*	*g*	-	-	-	e	f	g	a	b	c	d	e	f	g
-2c	*a*	*b*	*c*	*d*	*e*	*f*	*g*	-	-	-	-	f	g	a	b	c	d	e	f	g
+1n	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*a/d*	-	a/d	e	f	g	a	b	c	d	e	f	g
+1c	*a*	*b*	*c*	*d*	*e*	*f*	*g*	-	g	a/d	e	f	g	a	b	c	d	e	f	g
+2n	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*a/d*	*e*	a/d	e	f	g	a	b	c	d	e	f	g
-1n	*a*	*b*	*c*	*d*	*e*	*f*	-	-	-	a/d	e	f	g	a	b	c	d	e	f	g
-2n	*a*	*b*	*c*	*d*	*e*	-	-	-	-	a/d	e	f	g	a	b	c	d	e	f	g

Table 5 – Heptad repeat alignment of chimeric phase variants relative to the wild-type (WT) coiled-coil register of EnvZ. Letters *a* through *g* denote positions within the canonical heptad repeat, with positions *a* and *d* (highlighted in yellow) forming the hydrophobic core of the coiled-coil. Variants incorporate ±1 or ±2 residue insertions or deletions at the junction between the HAMP and DHp domains to shift fusion phase. Each shift perturbs the downstream register, with phase misalignments disrupting core *a/d* positions. The natural WT stutter between the HAMP and DHp domains is preserved in the phase 0 variant but is variably displaced in the other variants. Italicized and underlined positions are contributed by the N-terminal fusion, with the remaining positions being contributed by C-terminal EnvZ.

Each single-residue offset changes the relative orientation of the catalytic and dimerization domains (Fig. 4). AlphaFold2-Multimer models of representative variants illustrate how these phase shifts re-angle the cytoplasmic kinase core, providing a structural rationale for their divergent basal activities.

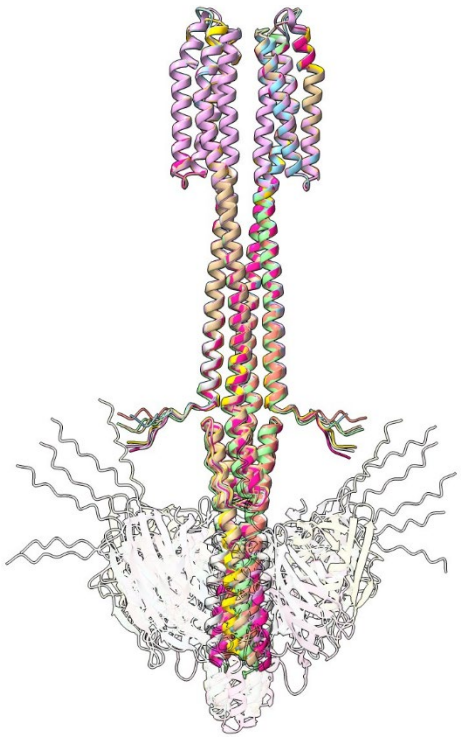


Figure 4 – AlphaFold2-Multimer predicted models aligned by sensor and transmembrane domains, showing how phase can alter angles in the kinase domains.

4.2.2 Library Sorting

To quantify basal activity in the absence of exogenous ligands, we analyzed each chimeric library by fluorescence-activated cell sorting (FACS). For these experiments, libraries were inoculated to an initial OD600 of 0.005 and grown in supplemented M9 minimal medium (1× M9 salts, 2 mM MgSO4, 0.1 mM CaCl2) containing 0.4% (wt/vol) glucose and 0.2% (wt/vol) casamino acids, with chloramphenicol and carbenicillin for plasmid maintenance and 100ng/mL aTc and 1mM IPTG for system induction for eight hours. As controls, the same strain was grown (i) without any plasmid (baseline autofluorescence) and (ii) with an aTc-inducible sfGFP plasmid (maximal fluorescence). The library and the controls were all then harvested by spinning down 10mL of culture at 3500rcf for 10min at 4C. The media was then decanted, and the cells resuspended in 10mL of 1x phosphate-buffered saline (PBS). The cells in PBS were then strained through a 25µM filter to remove any cell clumps present, and the library and controls were diluted to 70million/mL in PBS. For the purposes of converting arbitrary fluorescent units to MEFL later, Spherotech RCP-30-5A lot #3285117 rainbow calibration beads

were prepared by vortexing two drops in 2mL PBS and measured on the cell sorter. After measuring the controls, the libraries were then run on a Cytek Aurora Cell Sorter and sorted into six different bins based on strength of fluorescence, as shown in Figs. 5-8. Sorted populations were recovered on selective agar and grown overnight to enable downstream sequencing.

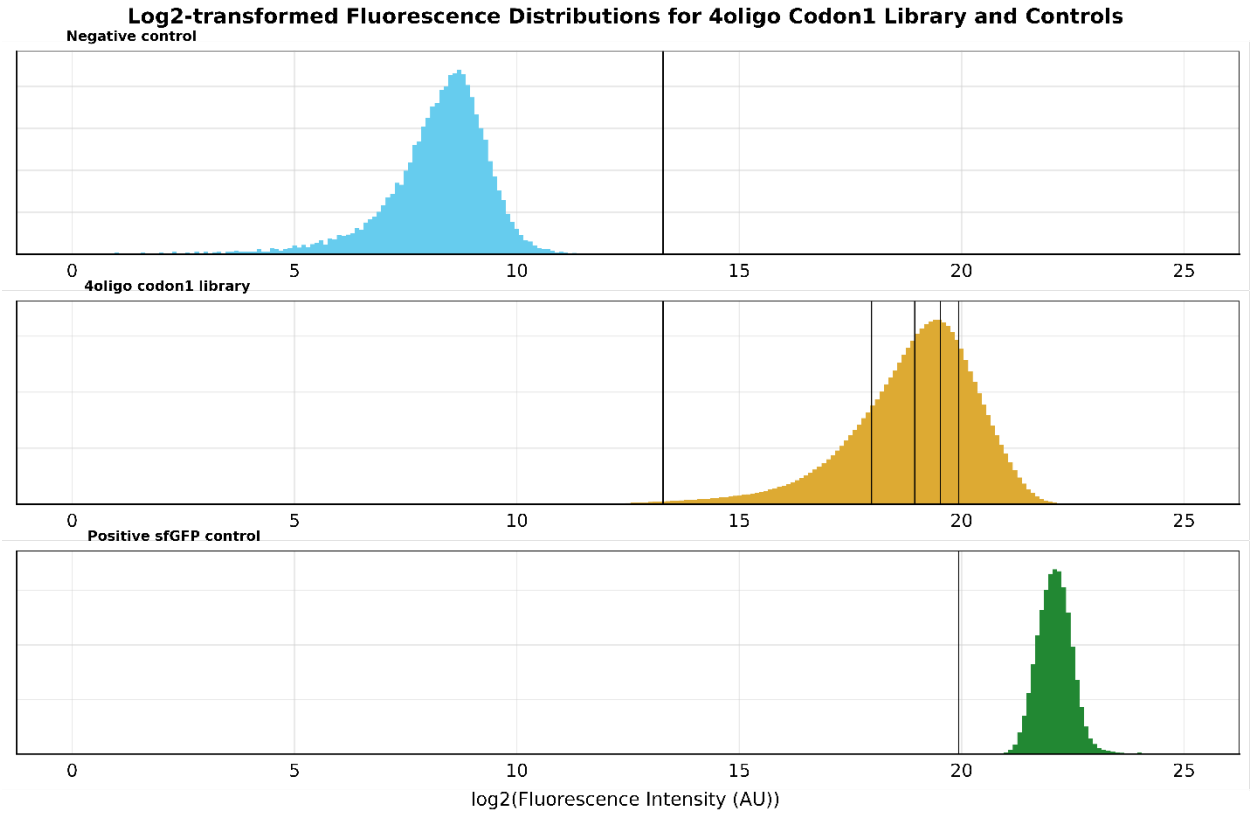


Figure 5 – Log2 transformed FACS distribution for the 4oligo codon1 library, along with BW29655 cells with no plasmid as a negative control, and BW29655 containing aTc-inducible sfGFP expression as a positive control.

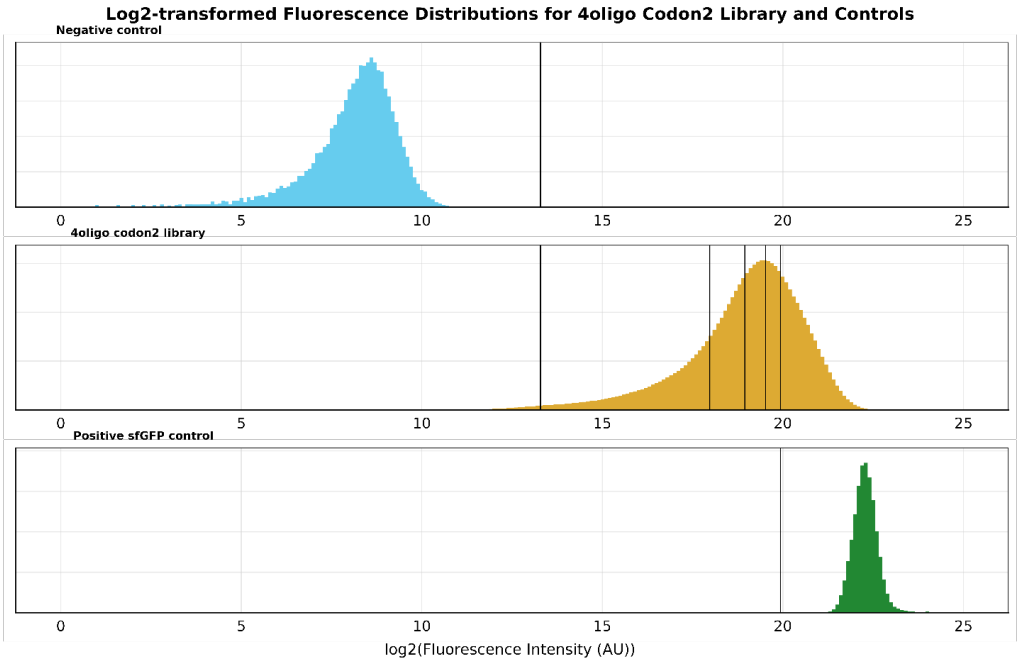


Figure 6 – Log2 transformed FACS distribution for the 4oligo codon2 library, along with BW29655 cells with no plasmid as a negative control and BW29655 containing aTc-inducible sfGFP expression as a positive control.

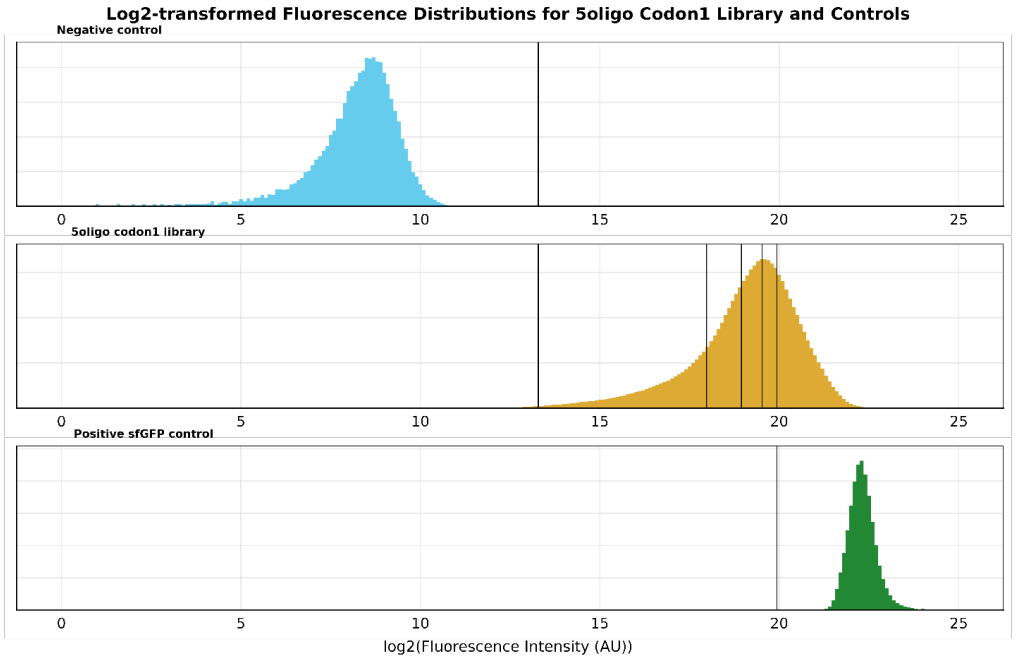


Figure 7 – Log2 transformed FACS distribution for the 5oligo codon1 library, along with BW29655 cells with no plasmid as a negative control and BW29655 containing aTc-inducible sfGFP expression as a positive control.

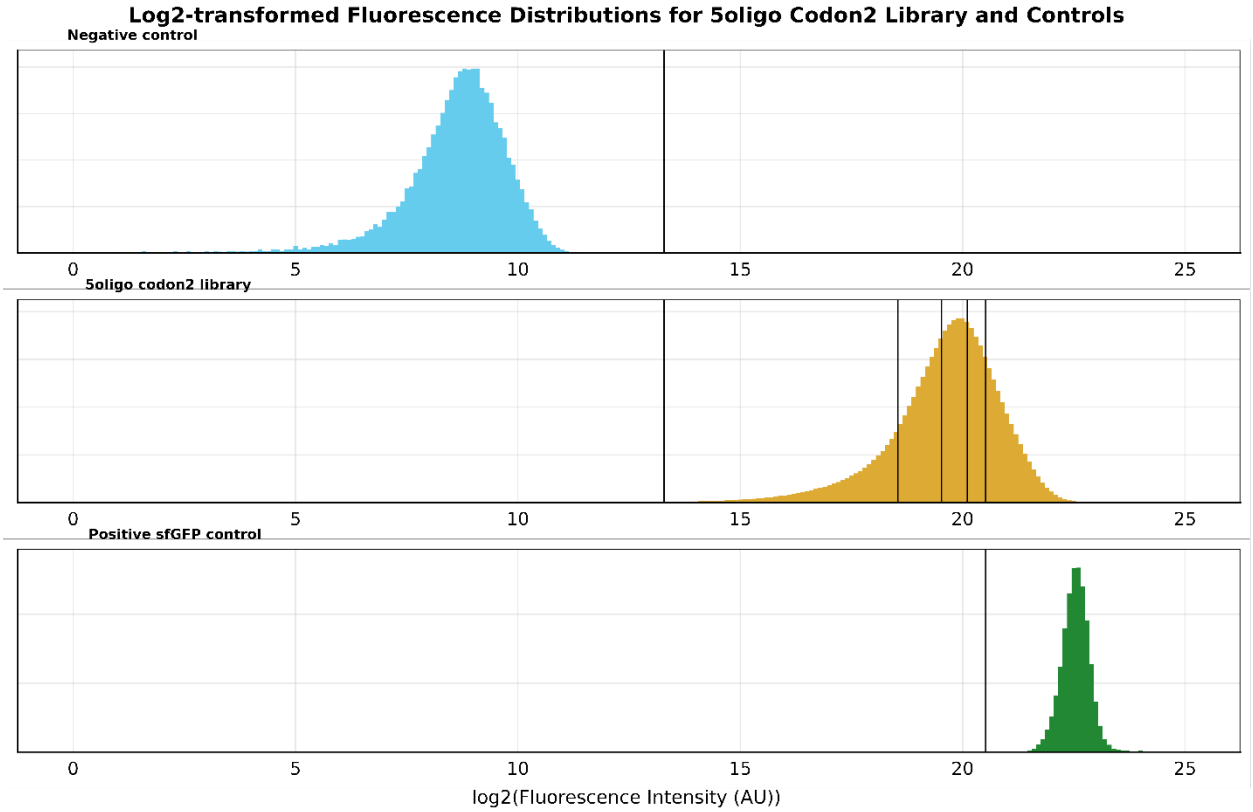


Figure 8 – Log2 transformed FACS distribution for the 5oligo codon2 library, along with BW29655 cells with no plasmid as a negative control and BW29655 containing aTc-inducible sfGFP expression as a positive control.

4.2.3 Preparing Sorted Samples for Sequencing

The day after the FACS experiments, a minimal amount of LB media was poured onto the selective plates, and the plates, corresponding to the different sorted bins, were carefully scraped to remove all colonies present. This scrape was then split, with some being miniprepped and the majority going to glycerol stocks. From the miniprepped cells per bin, the barcode region of the SHK plasmid was amplified using a mixture of five sets of shotgun primers for our first PCR, and then this amplification product was again amplified with Stubby TruSeq UDI primers, with these second sets of primers being specific to each bin. Amplicons from all fluorescence bins and libraries were quantified, pooled at equimolar ratios, and sequenced on an Illumina NovaSeq 6000 at the University of Oregon Genomics & Cell Characterization Core Facility.

4.2.4 Sequencing Analysis Pipeline

Demultiplexed FASTQ files were processed with an in-house workflow. BBMerge from BBTools was used to merge overlapping paired reads into a single read. Next, hts_primers from HTStream was used to flip the sequences into the correct orientation and trim the primer sequence from the ends of the sequences. The quality data was removed to turn the FASTQ files into FASTA format, and the first 24 bases (the length of the barcode on the SHK plasmids) were cut out, with the number of times each barcode appeared counted. Starcode, a DNA sequence clustering software, was then used to cluster and collapse all barcodes and their counts at a Levenshtein distance of one for the final output of this initial pipeline. Processing was performed independently for every library-bin combination, yielding barcode-count matrices that fed subsequent activity-inference analyses.

4.2.5 Inferring Brightness

Mapping data from barcode–coding-region sequencing (Oxford Nanopore and MAS ISO-seq PacBio HiFi) were imported into R for curation. The sequences were trimmed down to directly after their first stop, if applicable, and then the sequences were length-filtered so that only sequences no shorter than ten amino acids less than the shortest designed sequence remained, though ones containing a premature stop codon were allowed to be 30 amino acids shorter than the shortest design. To resolve collisions, in which two distinct stutter variants encoded identical amino acid sequences and could not be uniquely linked to a single barcode set, one variant was randomly removed; the remaining variant was then assigned all associated barcodes. Each variant was annotated as missense, insertion, deletion, or nonsense. Read counts per barcode were totaled and converted to fractional abundances; the resulting table was exported for downstream brightness calculations.

4.2.5.1 Determining Median Fluorescence Per Barcode

To calculate an inferred brightness, a workflow was adapted from a similar study (Biswas et al., 2021). The barcode count files from the output of the NovaSeq analysis pipeline were brought into a script and made into a count table, C. A relative abundance table, R, was made by dividing the columns of the count table by their sums. The output file made from the mapping data was brought in, and each column of the relative abundance table was divided element-wise by this to obtain a fold-change table, F. Then, each row of the fold-change table was divided by its sum to generate a table of adjusted abundances, A. This is then used for defining a discrete

probability mass function used for determining the fractional median bin number where a variant would appear. Total reads per barcode, along with how many and which bins each barcode appeared in, were also calculated. Now, given the measured fluorescence of the Spherotech rainbow calibration beads from the FACS experiment, the intensity of the fluorescence peaks being specific to the bead lot #3285117, after which FlowCal (Castillo-Hair et al., 2016) was used to analyze the rainbow calibration beads to determine the slope of the line used in the equation for conversion of arbitrary fluorescent units to molecules of equivalent fluorescein (MEFL). Bin boundaries and mid-points were converted to MEFL, allowing each barcode's median-bin index to be expressed as an inferred brightness in MEFL.

4.2.5.2 MEFL by Phase Variant

As this analysis was done using all mapped variants at the barcode level, an inferred brightness could be determined for each barcode. Using this to determine the inferred brightness at the amino acid level, the barcodes were grouped by protein sequence, and the brightness at the barcode level was multiplied by the total reads for that barcode divided by the total reads for all barcodes for a protein sequence to determine its fractional contribution, and then these were summed to obtain the weighted inferred brightness. If all barcodes for a given protein sequence were only found to be in a single bin via sequencing, these were flagged so as to be easily filtered from the dataset for downstream analysis when desired, as the inferred brightness was unable to be properly calculated, as it was impossible to determine where in the bin each would reside. The final dataset joins each protein's weighted MEFL with its sensor domain, class, phase variant, and taxonomic origin, enabling downstream analyses of how fusion phase and sensor class modulate basal activity.

4.3 Analysis of Inferred Brightness for Phase Variants

After the data was thus processed and filtered, it was then ready for analysis of which trends were present in the data.

4.3.1 Phase Variant Affects Brightness

To assess the impact of rotational register on basal activity, brightness values were grouped by the combination of sensor class and phase, after which the median MEFL for each group was plotted (Fig. 9). Two general patterns became immediately apparent. First, the 0 and the +1c phase, which correspond to either no residue shift or a one–residue addition on the C-terminal side of the fusion, were almost invariably the dimmest variants across the majority of

sensor classes. Second, the –1n phase, representing a one–residue deletion on the N-terminal side, was repeatedly the brightest. The 2n and –2c registers did not display a uniform tendency in either direction, an observation that could imply class-specific structural constraints rather than a universal phase effect.

In the absence of activating ligands, basal activity across phase variants also reflects how each fusion perturbs both the heptad repeat and the naturally occurring stutter at the HAMP-DHp fusion junction. The -1n phase, which subtly shifts the N-terminal fusion and removes the stutter and reestablishes quasi-continuous heptad spacing, consistently yields high activity, likely by biasing the system towards a conformation that mimics an activated state. In contrast, the 0 phase retains wild-type residue continuity, including the stutter, and shows consistently low activity, likely because it neither disrupts nor promotes activation in the chimeric context, yielding a structurally neutral baseline. Interestingly, the -1c variant also shows elevated activity despite disrupting the heptad core: its deletion removes a hydrophobic *a*/*d* residue, breaking the canonical stutter between the coiled-coil interfaces; this suggests its high activity arises not from preserved structure but from destabilization of the DHp interface in a manner that pushes the system toward activation. For the +1c phase variant, although it preserves coiled-coil formation within DHp, the additional residue misaligns its heptad repeat with the HAMP stutter, disrupting hydrophobic register at the HAMP-DHp junction, but not in the manner used in the natural stutter. Rather than leading to a locked-on state, this misalignment may weaken interdomain coupling and prevent proper signal propagation, resulting in consistently low ligand-independent activity despite intact helices. Other variants, such as -2n, -2c, and +2n, more drastically alter both the heptad and stutter, leading to sensor class-specific outcomes that may reflect differences in tolerance for structural mismatch. These results indicate that phase-dependent basal activity arises not just from helical or heptad alignment but also from whether the fusion retains or disrupts the crucial stutter feature that regulates HAMP-to-DHp communication.

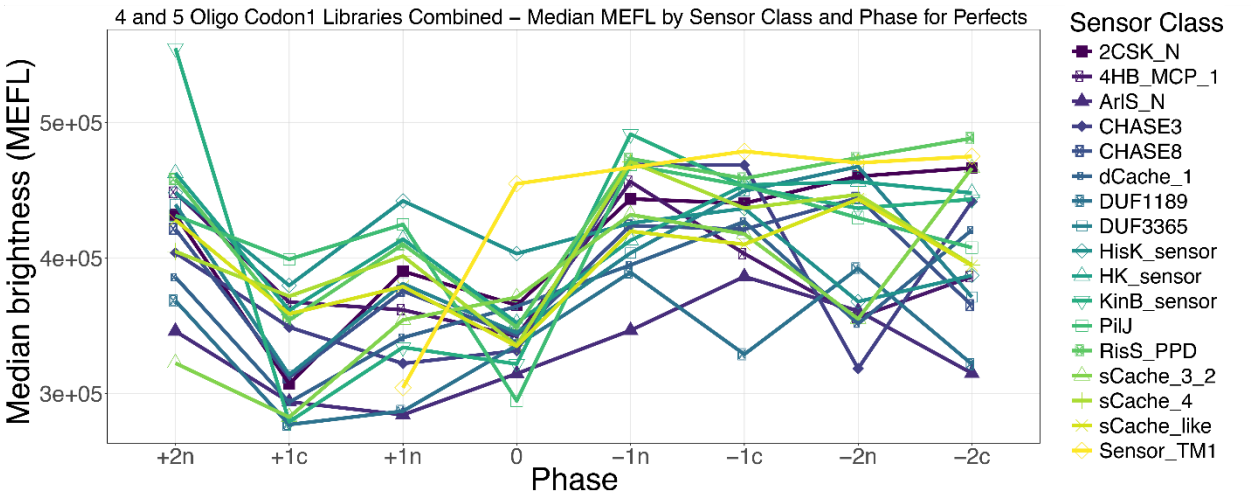


Figure 9 – Median inferred brightness (MEFL) on the y-axis with phase on the x-axis for all sensor classes for the combined codon1 4oligo and 5oligo libraries, with values only representing perfect constructs at the amino acid level.

Because a single summary statistic can obscure the underlying distributions, the two sensor classes that were most abundant in the 4oligo codon1 library were examined independently in greater detail. For sCache_4 sensors, the distribution echoed the global trend: variants in the –1n phase formed a distribution skewed toward high brightness, whereas those in the 0 and +1c phases were strongly biased toward low brightness (Fig. 10). PilJ sensors exhibited a comparable ordering across phases (Fig. 11). When non-designed missense variants were included alongside perfect constructs, the brightness distributions broadened in both classes, yet the relative ordering of the phases remained intact, indicating that phase effects dominate over the modest noise introduced by single-amino-acid substitutions (Fig. 12).

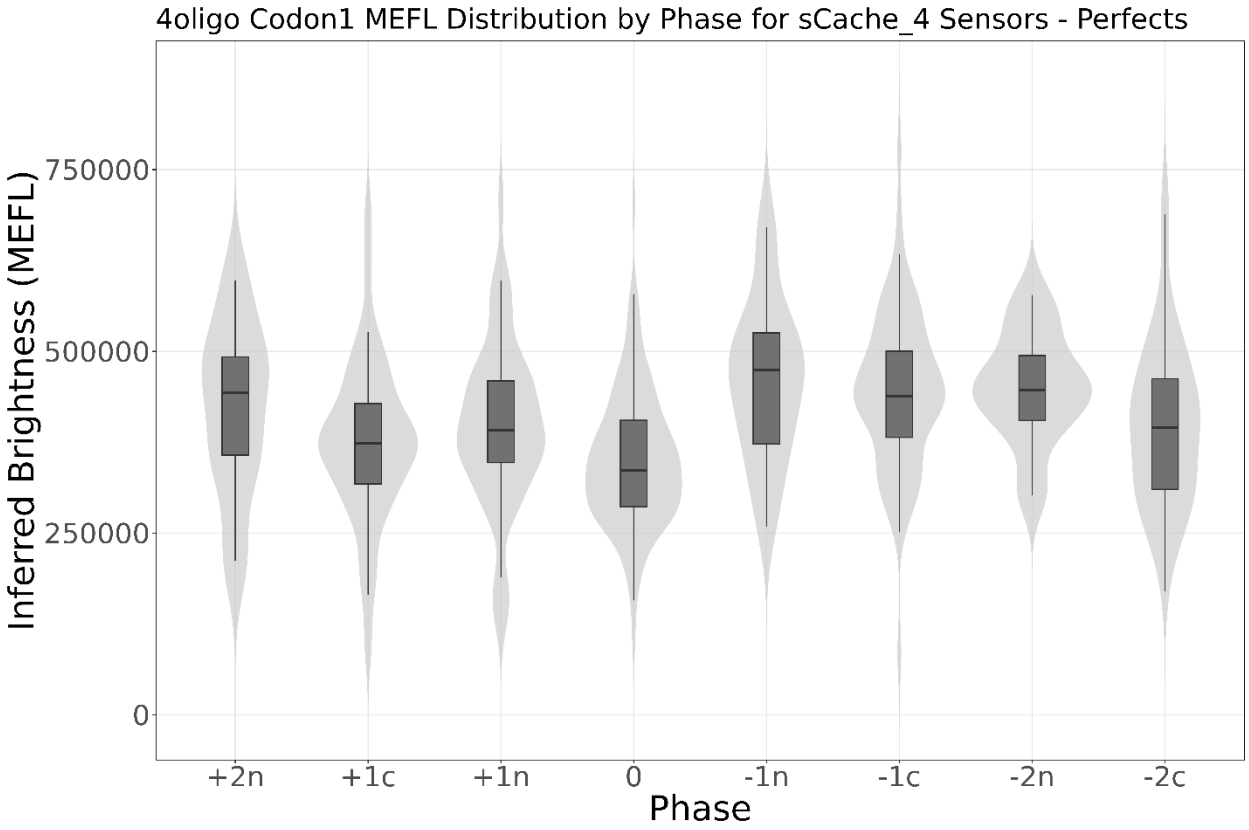


Figure 10 – Violin boxplots illustrating the distribution of inferred brightness (MEFL) for 4oligo codon1 sCache_4 sensors across different phases, only representing perfect constructs. Each violin represents the distribution of brightness values for a specific phase, with the embedded boxplots indicating the interquartile range and median. The x-axis denotes the phase, while the y-axis displays the inferred brightness in MEFL. This highlights phase-dependent variation in brightness, enabling comparison of both the whole distribution and its averages for different phases.

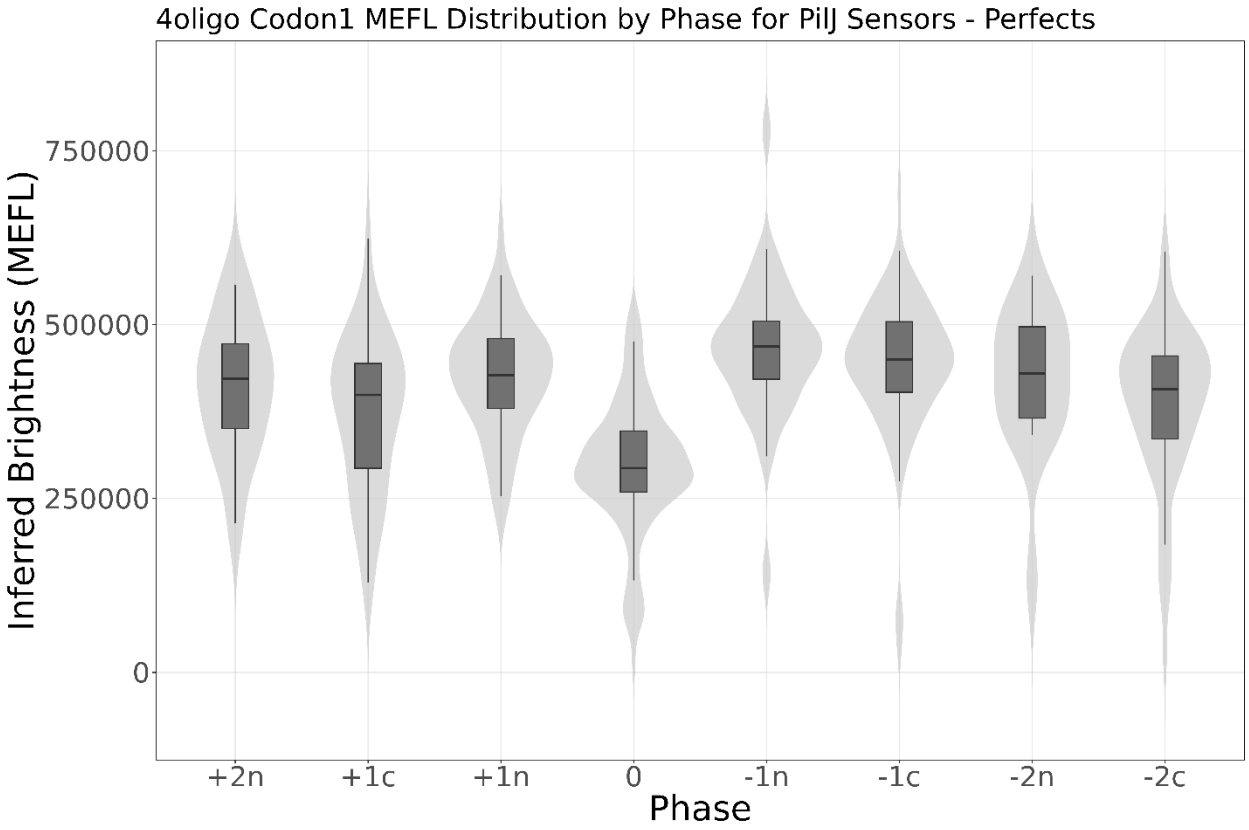


Figure 11 – Violin boxplots illustrating the distribution of inferred brightness (MEFL) for 4oligo codon1 PilJ sensors across different phases, only representing perfect constructs. Each violin represents the distribution of brightness values for a specific phase, with the embedded boxplots indicating the interquartile range and median. The x-axis denotes the phase, while the y-axis displays the inferred brightness in MEFL. This highlights phase-dependent variation in brightness, enabling comparison of both the whole distribution and its averages for different phases.

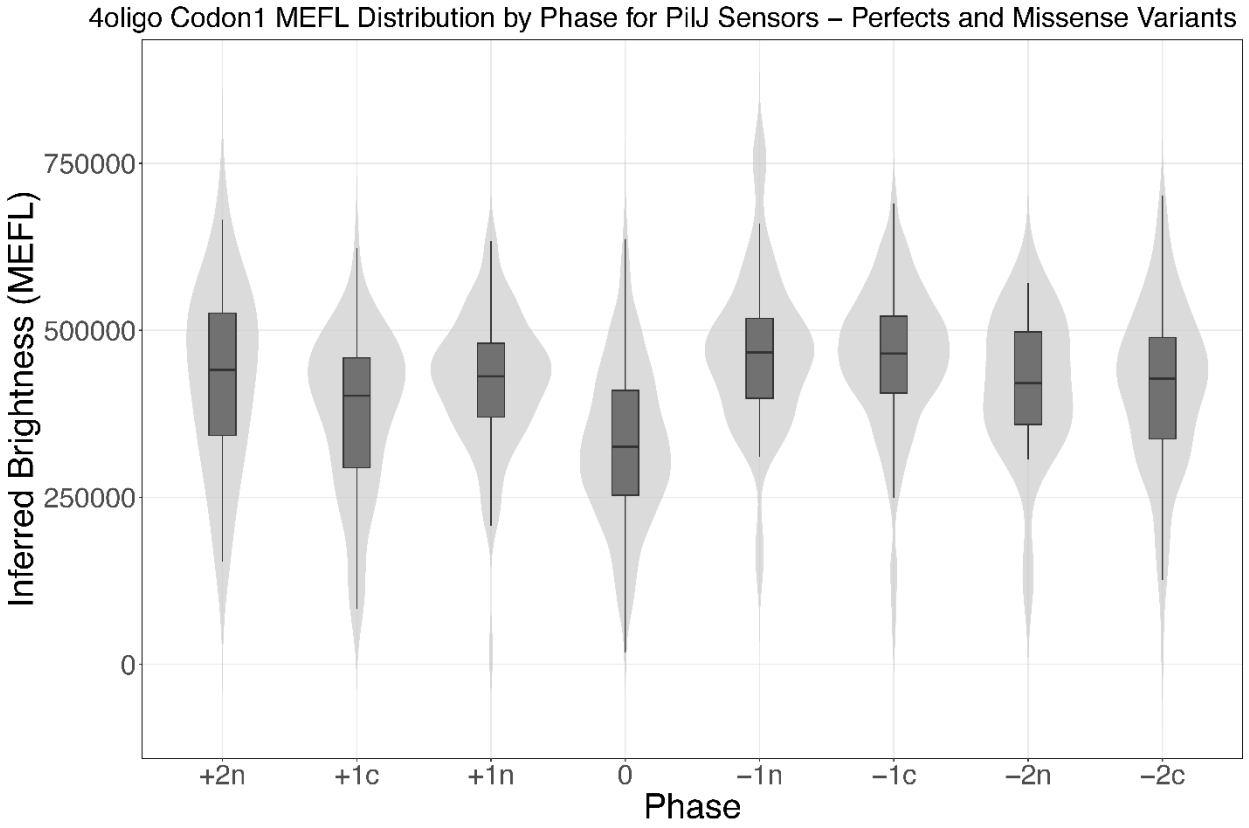


Figure 12 – Violin boxplots illustrating the distribution of inferred brightness (MEFL) for 4oligo codon1 PilJ sensors across different phases, only representing perfect and missense constructs. Each violin represents the distribution of brightness values for a specific phase, with the embedded boxplots indicating the interquartile range and median. The x-axis denotes the phase, while the y-axis displays the inferred brightness in MEFL. This highlights phase-dependent variation in brightness, enabling comparison of both the whole distribution and its averages for different phases.

4.3.2 Effects of Sensor Class on Brightness Distribution

By measuring output characteristics, such as minimum and maximum brightness, within groups of variants with the same protein ID, distributions of these characteristics can be made for sensor classes. These, when split up by sensor domain, help inform functional differences between sensor architectures. Minimum brightness speaks to the potential leakiness or basal activation of these sensors, while maximum brightness indicates how strongly these sensors could potentially be activated. Taken together, these minima and maxima provide an

approximate dynamic-range envelope for each architecture and show how effectively a given sensor domain can modulate kinase output (Fig. 13).

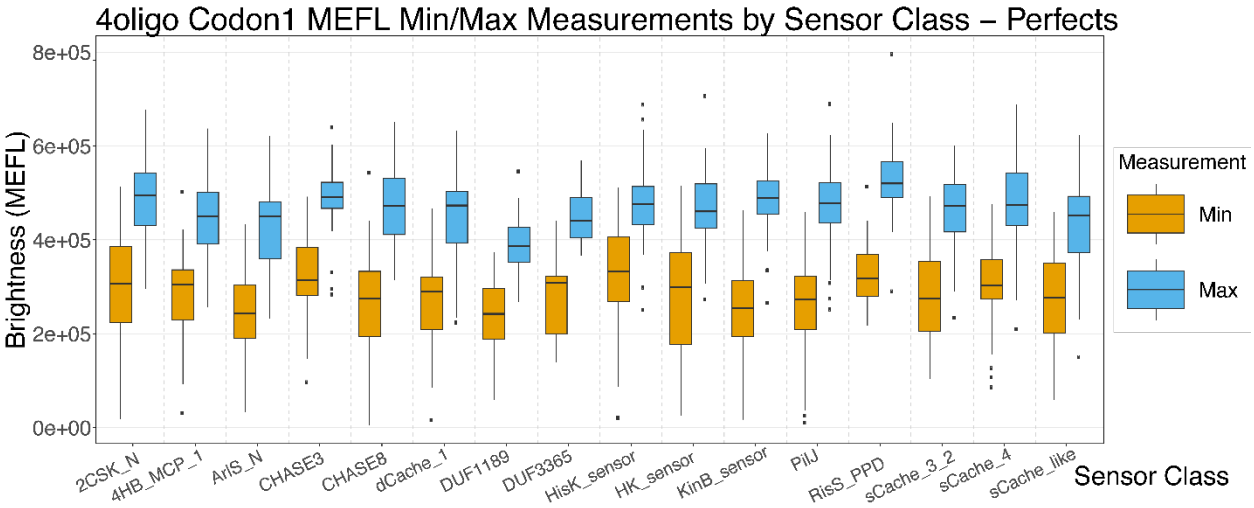


Figure 13 – Inferred brightness (MEFL) plotted against sensor class. For each protein ID with at least two perfect sequences, the brightest and dimmest variants were extracted; these values form the maximum- and minimum-brightness distributions shown. The spread between the two distributions illustrates the dynamic range characteristic of each sensor class.

4.3.3 Phase Variant Effects on Angle of Domains Below Fusion

Because an ideal α-helix completes one 360-degree turn in roughly 3.6 residues, each residue that is inserted or deleted at the fusion junction should rotate the downstream helix by about 100 degrees. On this geometric basis, the library's phase variants can be assigned ideal angular offsets, which are summarized in Table 6.

Ideal Angle Change by Phase Variant		
Phase	Angle 0 Reference (°)	Angle 2n Reference (°)
-2c	-200°	0°
-2n	-200°	0°
-1c	-100°	100°
-1n	-100°	100°
0	0°	200°
+1n	100°	300°
+1c	100°	300°
+2n	200°	40°

Table 6 - Theoretical ideal rotational offsets predicted for each phase variant, calculated from a helical pitch of 3.6 residues per turn.

We can then look at the correlation between fusion phase and inferred brightness (Fig. 14); the strongest correlations occur between phase variants with identical or closely aligned angle changes, such as -2n and +2n (40° difference in angle), -1c and -1n (0° difference), and +1n and +1c (0° difference). This suggests that variants with matching or close-to-matching rotational offsets tend to produce similar functional outputs. Moderate correlations between variants like +1n and +2n, along with -2c and +2n, further support the idea that small angular differences remain similar in function, while larger or mismatched shifts reduce similarity in function.

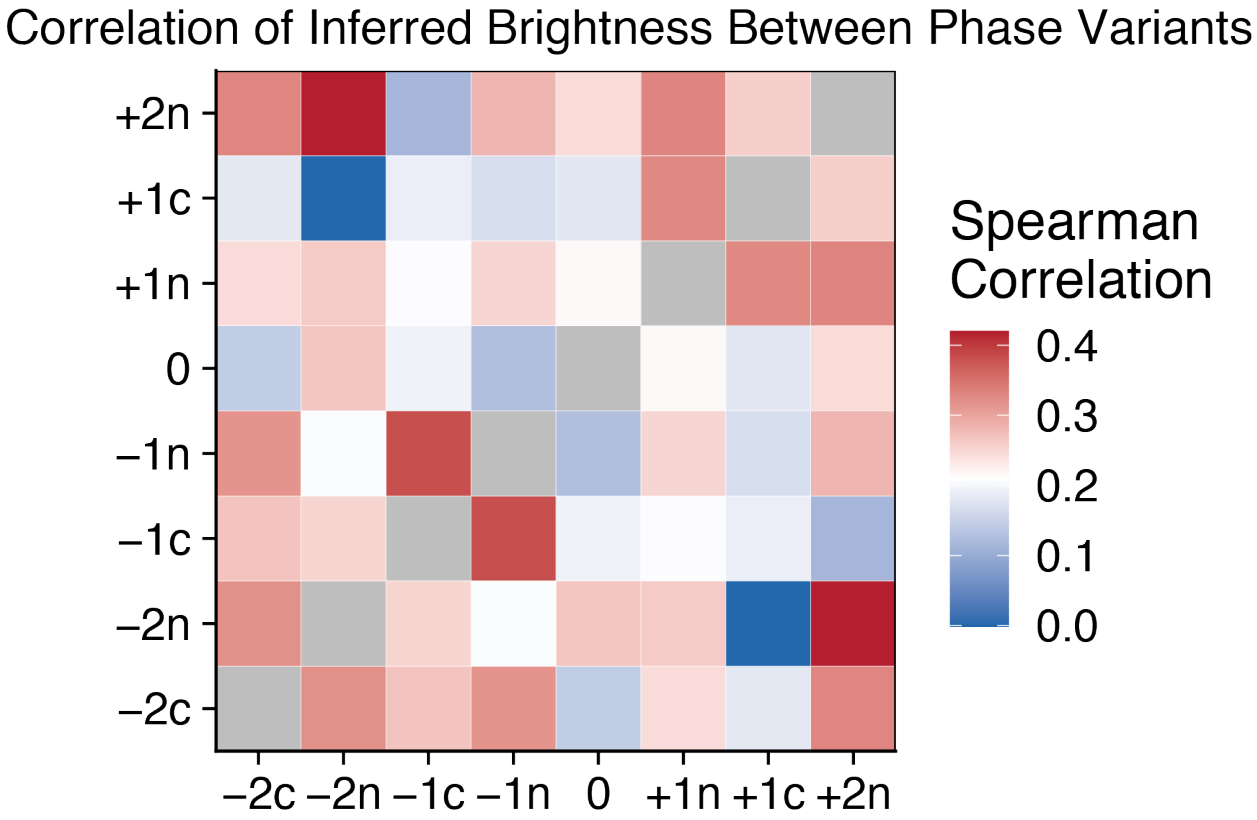


Figure 14 - Heatmap showing Spearman correlation for inferred brightness between phase variants.

A subset of proteins for which all phase variants were successfully designed was further analyzed structurally. Homodimeric models were generated with AlphaFold2-Multimer, and the actual inter-domain rotation below the fusion site was measured for each model. For variants that

also possessed experimentally determined brightness, the measured rotation was correlated with MEFL values (Fig. 15). Across 123 such constructs the overall correlation is weak, a result that may reflect both modelling inaccuracies and the limited sample size. Focusing on a single phase, for example the –1n register, produces a modest improvement in correlation but remains constrained by the small number of data points available.

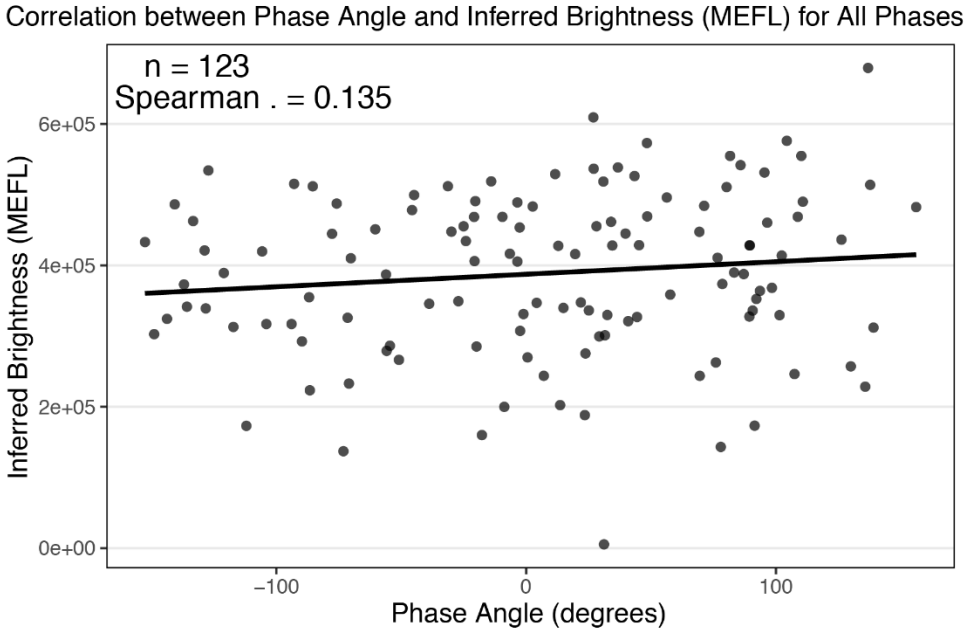


Figure 15 – Scatter-plot of inferred brightness versus AlphaFold2-Multimer-derived rotational angle for 123 fully modelled variants spanning all phases.

4.3.4 Random Forest Model

We evaluated whether the primary sensor sequence encodes enough information to estimate basal activity. The 4oligo and 5oligo codon1 libraries data was fed into a random forest machine learning model using ESM-2 generated embeddings to predict chimeric SHK brightness using the primary N-terminal sequence data, excluding the conserved EnvZ kinase region below the fusion point. Perfect and missense variants were fed into this after being filtered so no single-bin variants were present, and every variant had at least three barcodes; this resulted in a library with 16,463 variants, which was divided into 80% training, 10% validation, and 10% test sets. Sequence representations were generated using 480-dimensional embeddings from the 35M parameter ESM-2 protein language model, as larger ESM-2 models did not improve results. Embedding vectors were concatenated with a one-hot encoded phase indicator (adding sensor

class did not increase accuracy) using scikit-learn 1.0.2 (Python 3.7.12) and then used to train a random-forest regressor. Each training sample was assigned a weight calculated as the log of (number of reads × number of barcodes + 1), normalized to the mean. Random forest hyperparameters were tuned via 3-fold randomized search, yielding a final model with 80 trees, a maximum depth of 14, and a minimum sample split of 2. The resulting model explained 16 % of the variance in the validation set ($R^2 = 0.163$; Spearman $\rho = 0.435$) and 21 % on the held-out test set ($R^2 = 0.210$; Spearman $\rho = 0.462$), with RMS errors of 1.23×10^5 MEFL. A scatter plot of predicted versus measured brightness (test set) and the histogram of log-transformed weights were generated (Fig. 16). Training an XGBoost regressor on the same features yielded comparable metrics, indicating that the ESM-derived, phase-aware feature set captures the majority of the predictive signal.

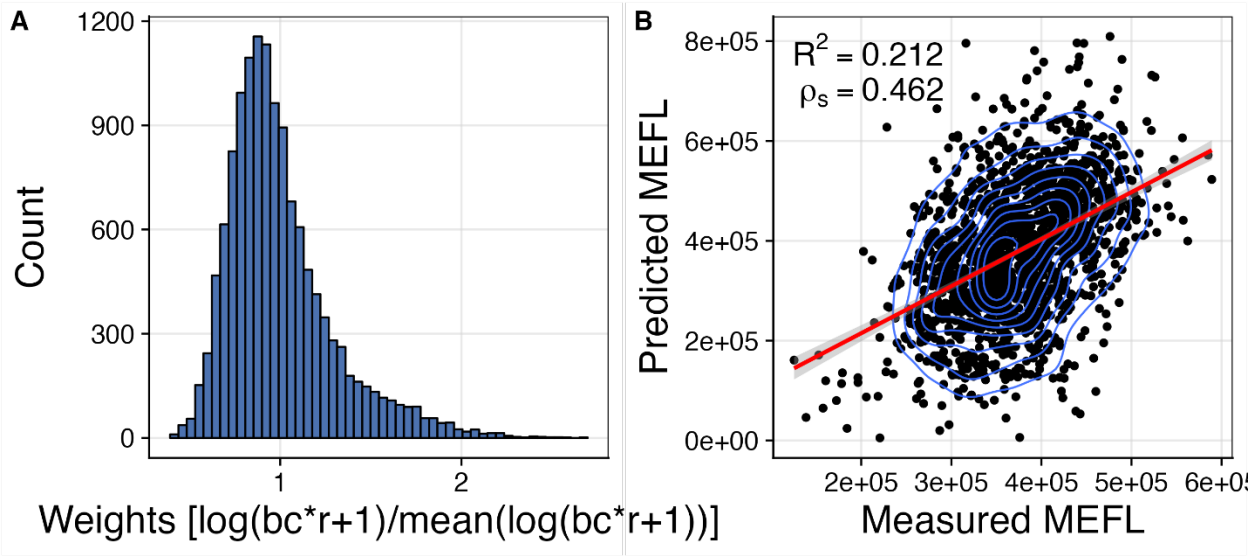


Figure 16 – Histogram of weights and scatter plot for random forest model. A. Histogram of log-transformed weights for random forest model. B. Scatter plot of predicted versus measured brightness in the test set for the random forest model. The red line is a smoothed linear model (lm) fit with the standard error around the line displayed in grey.

4.4 Locked-On Sorting Summary

This chapter presents a comprehensive high-throughput workflow for locked-on sorting to determine how fusion phase variants affect the basal activity of chimeric sensor histidine kinases

(SHKs). We systematically varied the fusion junctions in eight phase registers, accounting for both α-helical and coiled-coil rotational phases, across a diverse range of SHKs from a large set of taxa. Our findings demonstrate that specific phase variants, notably the -1n variant, tend to consistently yield elevated basal activity, likely due to restoration of quasi-continuous heptad repeats and disruption of the natural stutter, mimicking an activated state. In contrast, the 0 and +1c variants typically exhibit low activity, either through preserving a neutral signaling state or misaligning the natural stutter in a way that inhibits signal propagation. These trends are consistent across multiple sensor classes and support the model that both helical rotation and disruption or retention of structural motifs like the stutter are central to phase-dependent SHK function.

Importantly, this establishes that the fusion phase not only alters brightness distributions but also correlates with predicted rotational offsets. Phase variants with similar theoretical angular rotations (e.g., -1c and -1n) show stronger correlations in functional output, reinforcing the role of geometric alignment in signal transduction. The development of a machine learning model using ESM-2 embeddings further underscores the predictive potential of primary sequence and phase for inferring basal activity. Overall, this work provides foundational insight into the structural logic governing SHK fusion design, highlighting phase variation as a key axis for enriching functional chimeras in engineered designs of chimeric SHKs.

4.5 Locked-On Sorting Bibliography

Berntsson, O., Diensthuber, R. P., Panman, M. R., Björling, A., Gustavsson, E., Hoernke, M., Hughes, A. J., Henry, L., Niebling, S., Takala, H., Ihalainen, J. A., Newby, G., Kerruth, S., Heberle, J., Liebi, M., Menzel, A., Henning, R., Kosheleva, I., Möglich, A., & Westenhoff, S. (2017). Sequential conformational transitions and α-helical supercoiling regulate a sensor histidine kinase. *Nature Communications*, *8*, 284. https://doi.org/10.1038/s41467-017-00300-5

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., & Church, G. M. (2021). Low-N protein engineering with data-efficient deep learning. *Nature Methods*, *18*(4), 389–396. https://doi.org/10.1038/s41592-021-01100-y

Castillo-Hair, S. M., Sexton, J. T., Landry, B. P., Olson, E. J., Igoshin, O. A., & Tabor, J. J. (2016). FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. *ACS Synthetic Biology*, *5*(7), 774–780. https://doi.org/10.1021/acssynbio.5b00284

Lehning, C. E., Heidelberger, J. B., Reinhard, J., Nørholm, M. H. H., & Draheim, R. R. (2017). A Modular High-Throughput In Vivo Screening Platform Based on Chimeric Bacterial Receptors. *ACS Synthetic Biology*, *6*(7), 1315–1326. https://doi.org/10.1021/acssynbio.6b00288

Meier, S. S. M., Multamäki, E., Ranzani, A. T., Takala, H., & Möglich, A. (2024). Leveraging the histidine kinase-phosphatase duality to sculpt two-component signaling. *Nature Communications*, *15*(1), 4876. https://doi.org/10.1038/s41467-024-49251-8

Schmidt, N. W., Grigoryan, G., & DeGrado, W. F. (2017). The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: Application to understand signaling in histidine kinases. *Protein Science*, *26*(3), 414–435. https://doi.org/10.1002/pro.3095

5. CONCLUSION

The modularity of sensor histidine kinases (SHKs) presents a compelling opportunity to decode their signaling logic and lay the groundwork for developing plug-and-play biosensors. This dissertation demonstrates a comprehensive strategy for characterization of the basal activity of chimeric SHK variants by building, expressing, and analyzing thousands of sensors in a high-throughput fluorescence-based screening platform. From multiplexed gene synthesis to functional sorting, each part of this work was designed to address key challenges in SHK deorphanization: variable domain architecture, fusion phase incompatibility, and the prevalence of locked-on conformations. By coupling systematic precision engineering with scalable sorting strategies, this platform enables large-scale evaluation of SHK variant libraries in a format suitable for quantitative analysis and future predictive modeling.

At the synthesis level, we extended the capabilities of DropSynth by developing a degenerate oligo design strategy that allows construction of phase-shifted fusion variants within a single emulsion droplet, thereby reducing synthesis costs while maintaining high diversity. By targeting the junction immediately downstream of the HAMP domain, we created between one and eight phase variants per sensor that sampled discrete helical orientations at the fusion site. The final four gene library designs represented 10,862 unique proteins encoded across 21,724 gene constructs, each designed with oligo compatibility, junction phase variation, and codon usage in mind. This design captured sensor diversity from an extensive array of organisms across three domains of life, a long list of genera, and seventeen distinct sensor classes, including domains of unknown function. The resulting constructs retained consistent domain logic, ensuring compatibility with the signal transduction scaffold of *E. coli* EnvZ.

To support high-throughput screening of basal activity, we developed and validated a series of transcriptional reporter circuits. These circuits, centered around the pSR40.29 plasmid and its derivatives, provided both fluorescence-based and life-death selection modalities. Early circuit iterations emphasized modularity, insulation, and adjustable selection stringency. The use of an orthogonal response regulator, inducible promoters, and a tightly regulated output minimized noise and improved signal clarity in our assay conditions.

We then implemented locked-on sorting (LOS), a multiplexed fluorescence-activated cell sorting (FACS) and sequencing approach, to quantify basal activity across the chimeric SHK

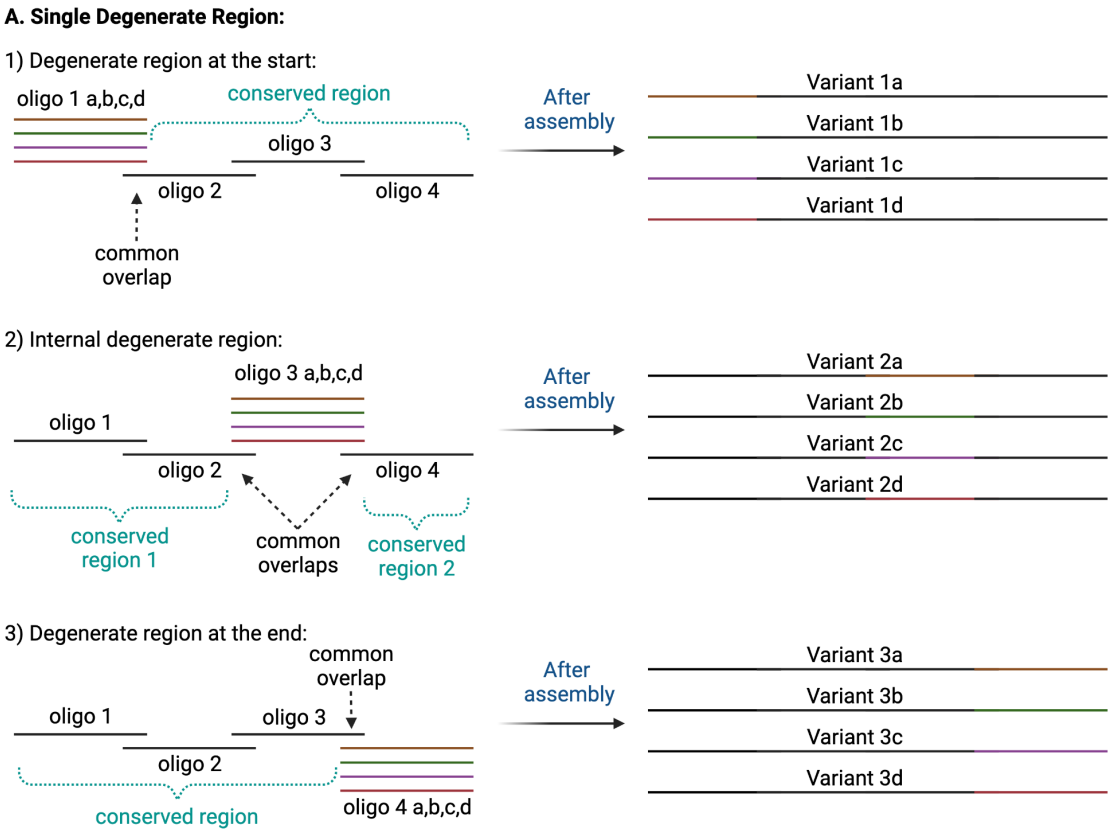
libraries. By analyzing barcode distributions across fluorescence bins, we inferred ligand-independent activity at single-variant resolution. This pipeline not only identified constitutively active variants but also revealed broad trends in signaling output driven by sensor class and fusion phase. Variants aligned with the preserved heptad register at the HAMP-DHp junction generally exhibited lower basal activity, while misaligned variants showed higher activation levels, consistent with known mechanisms of coiled-coil signal propagation.

Together, these methods form an integrated pipeline for engineering and analyzing modular signaling proteins at an unprecedented scale. The degenerate DropSynth approach enabled the systematic construction of variant libraries, the plasmid circuit ensured orthogonal measurement of SHK output, and the locked-on sorting pipeline coupled experimental selection with a sequencing readout to produce a quantitative map of basal signaling across thousands of engineered sensors. In addition to generating a valuable dataset for machine learning-based sequence-function prediction, this work provides a generalizable framework for studying modular signal processing and synthetic biosensor design.

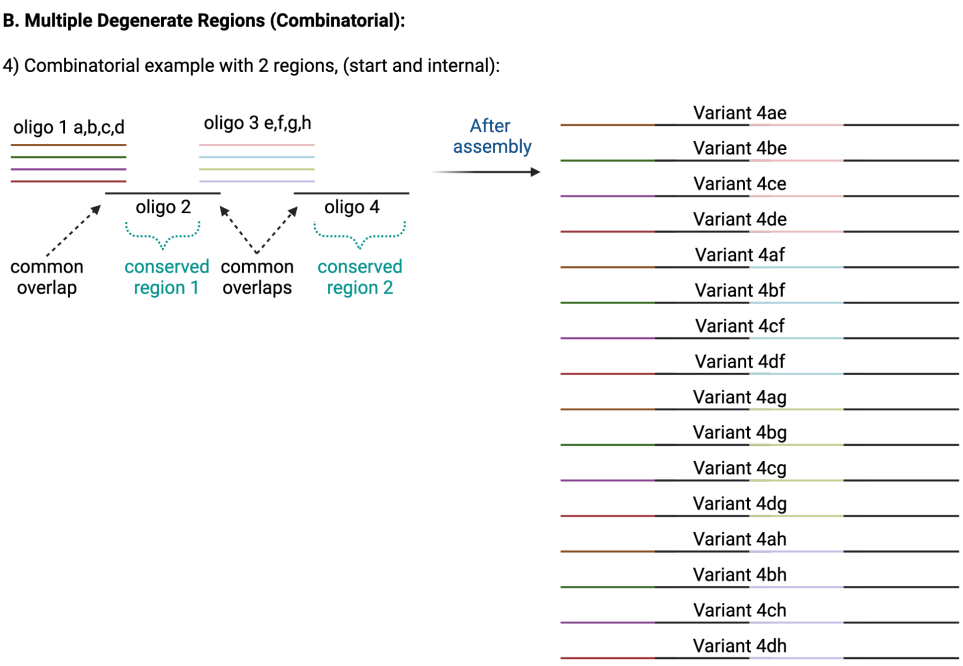
The next steps to move this forward would be screening these engineered libraries of chimeric SHKs against large chemical libraries. By knowing which variants are locked-on and to what degree, testing and knowing which variants are genuinely activated in the presence of various chemicals is then possible. Through iteration of designs, expression, and screening of chimeric variants, sequence-function relationships can be built from the generated data, enabling the future design of *de novo* biosensors for ligands of interest.

By combining design, synthesis, and functional selection in a scalable workflow, this dissertation offers a path towards comprehensive functional annotation of the millions of SHKs in metagenomic space and the rapid assembly of bespoke sensing circuits for synthetic biology.

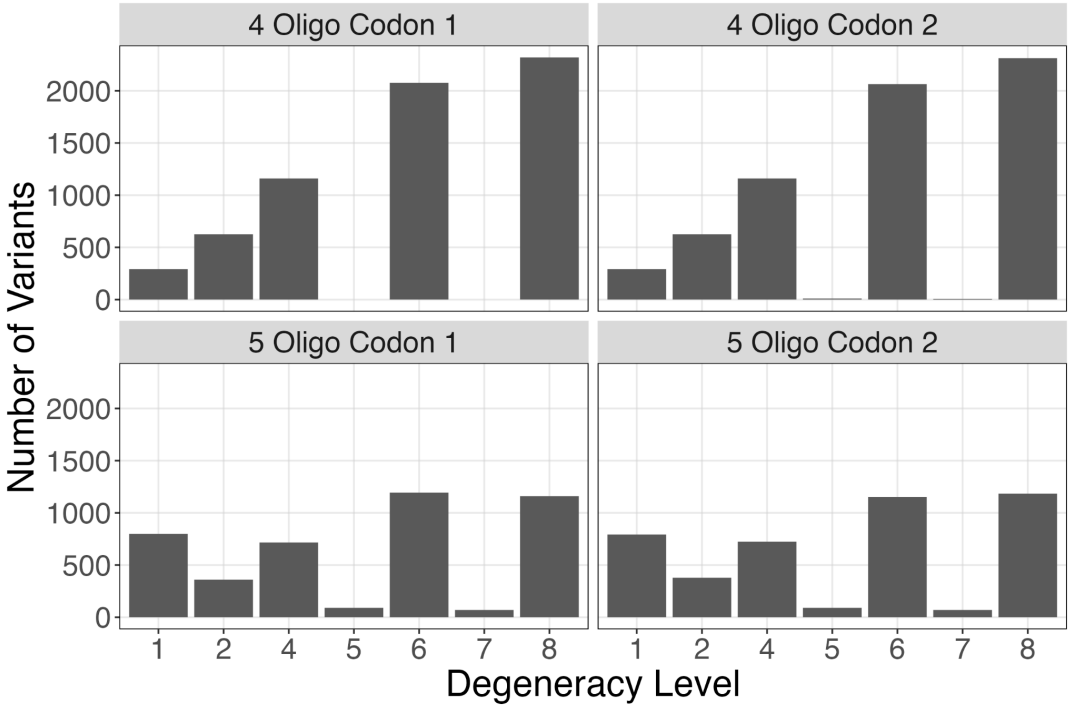
APPENDIX: SUPPLEMENTAL FIGURES AND TABLES FOR DEGENERATE DROPSYNTH



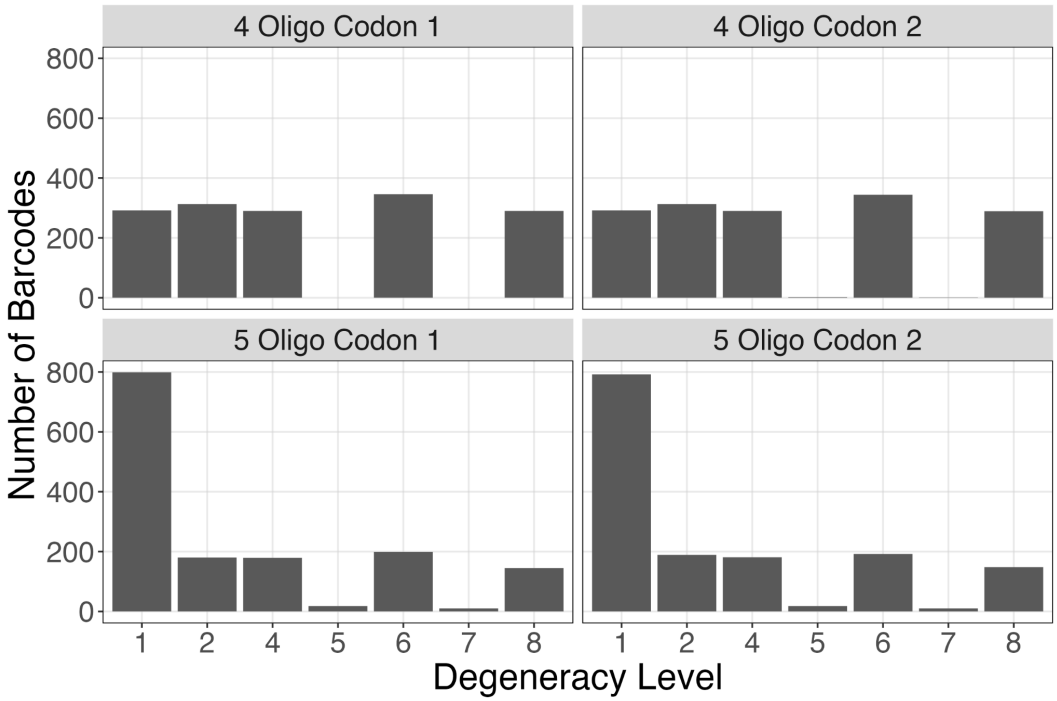
Appendix Figure A1 - Strategies for Introducing Oligo-Level Degeneracy. Degenerate DropSynth could be used to introduce variation in many parts of an assembled gene by creating multiple oligos for gene fragments at 1) the start, 2) internally, or 3) at the end.



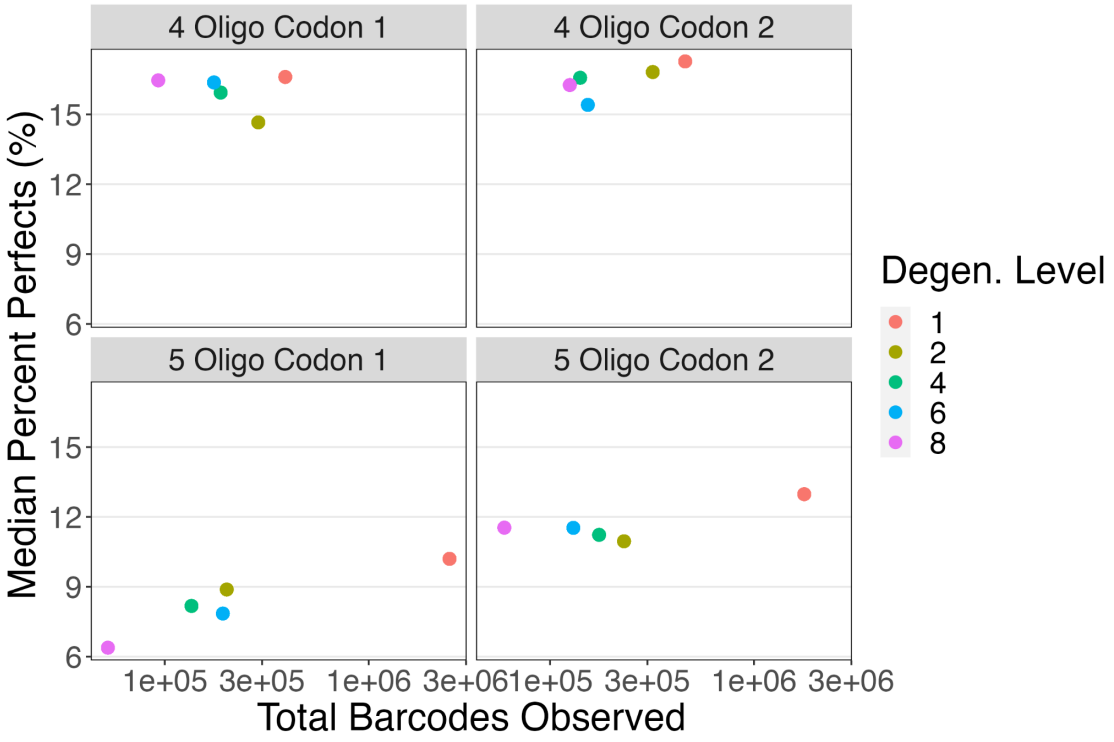
Appendix Figure A2 - Combinatorial Assembly from Multiple Degenerate Fragments. Introducing degeneracy in multiple fragments leads to combinatorial assemblies.



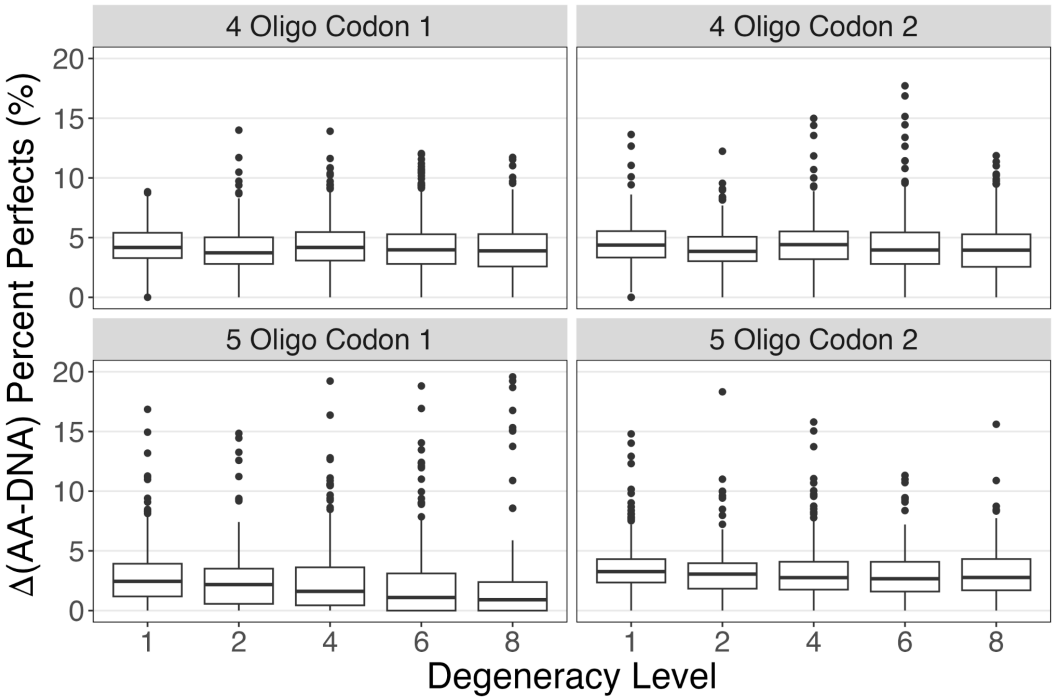
Appendix Figure A3 - Variant Counts by Degeneracy Level Across Libraries. The absolute number of variants designed at each degeneracy level for each of the four libraries tested.



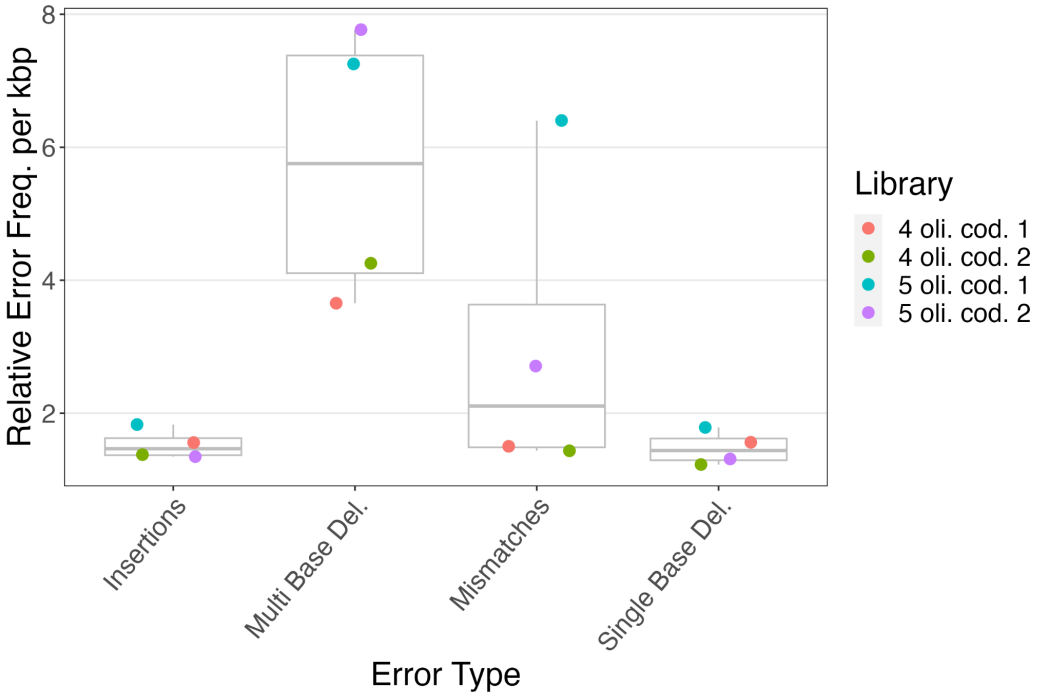
Appendix Figure A4 - Barcode Utilization by Degeneracy and Library Type. The number of barcodes used (out of 1536) for each degeneracy level in each of the four libraries tested. The 4 oligo libraries had a far more uniform distribution.



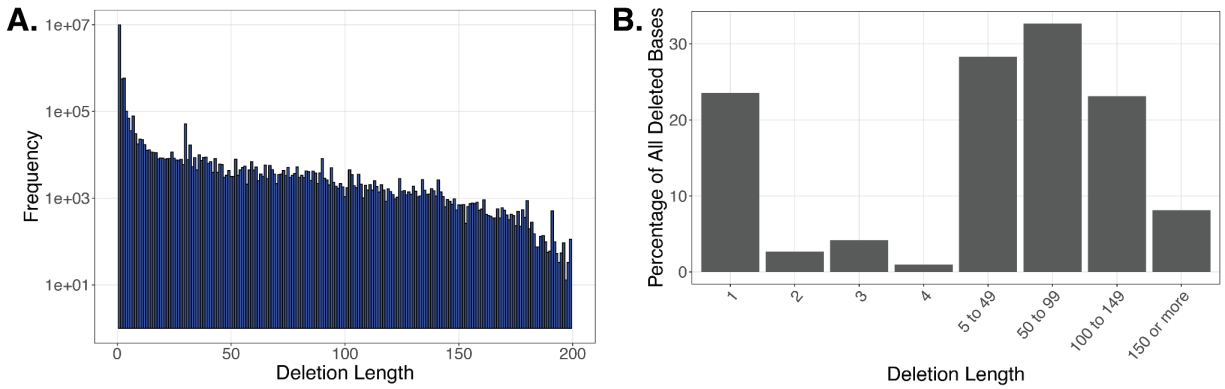
Appendix Figure A5 - Correlation Between Barcode Count and Degeneracy Level for Perfects by Library Type. When plotting the median percentage of perfects, we note a correlation with the total number of barcodes observed and the degeneracy levels which is much stronger with the 5 oligo libraries (0.88 and 0.96 Pearson) (bottom-row) compared to the 4 oligo libraries (-0.38 and 0.67 Pearson) (top-row).



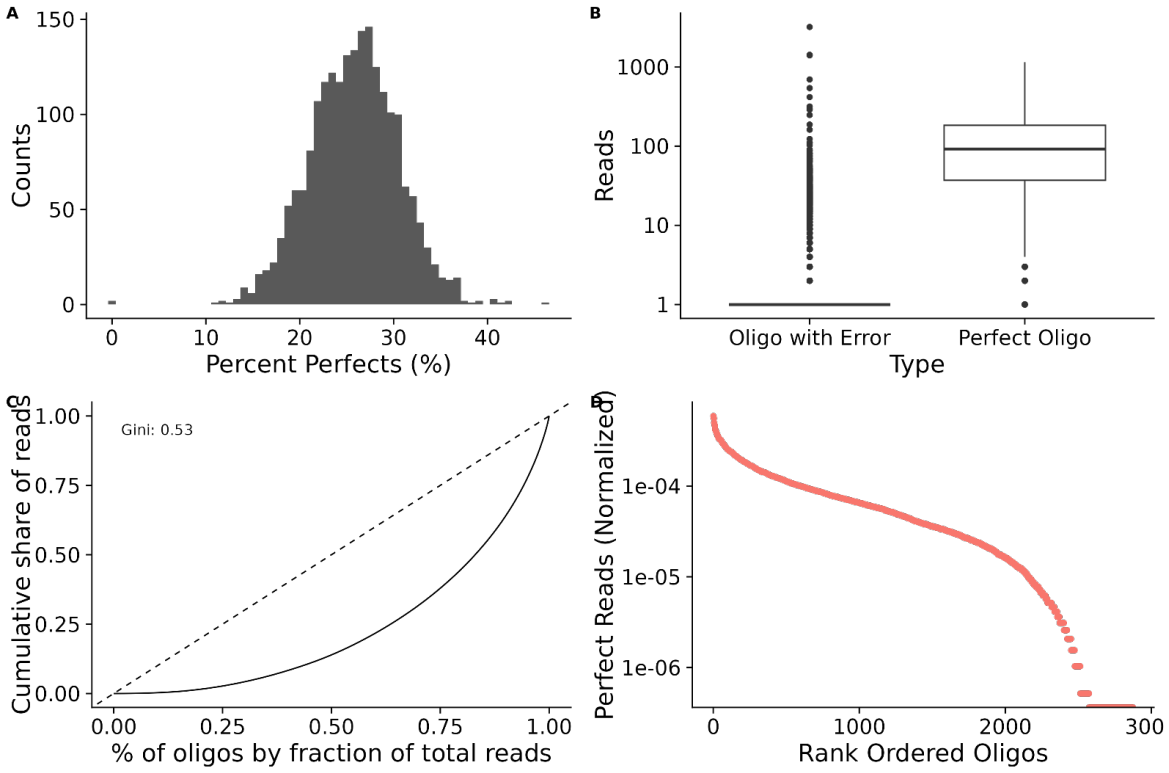
Appendix Figure A6 - Comparison of Percentage Perfects at DNA vs Protein Level. The difference in percentage perfects observed at the protein amino acid level (synonymous mutations collapsed) and the DNA level. We see a relatively consistent difference of 4.0% (s.d. 0.2%) with 4 oligos and 2.7% (s.d. 0.8%) with 5 oligos.



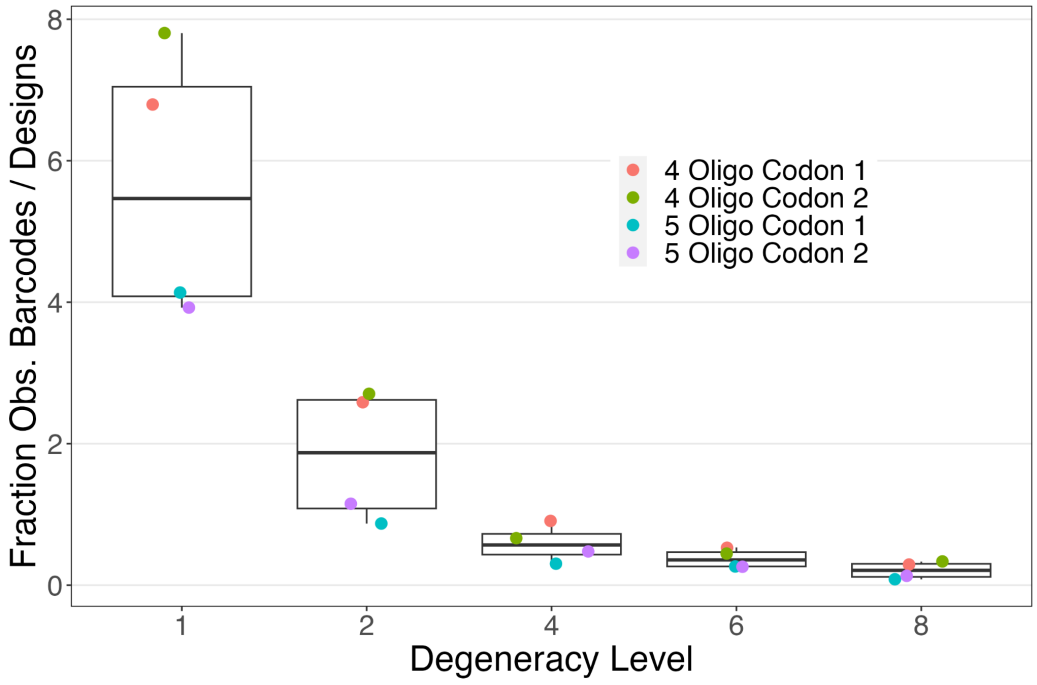
Appendix Figure A7 - Per-kb Error Rates by Type and Library Configuration. Analysis of the CIGAR alignment strings produced by minimap2 reveals the relative error frequencies per kbp for different types of errors. We find equivalent rates of insertions and single base deletions among the four libraries and much higher rates of multi base deletions. In the 4-oligo libraries we see mismatch rates comparable to insertions and single base deletions, while in the 5-oligo libraries mismatches are substantially higher.



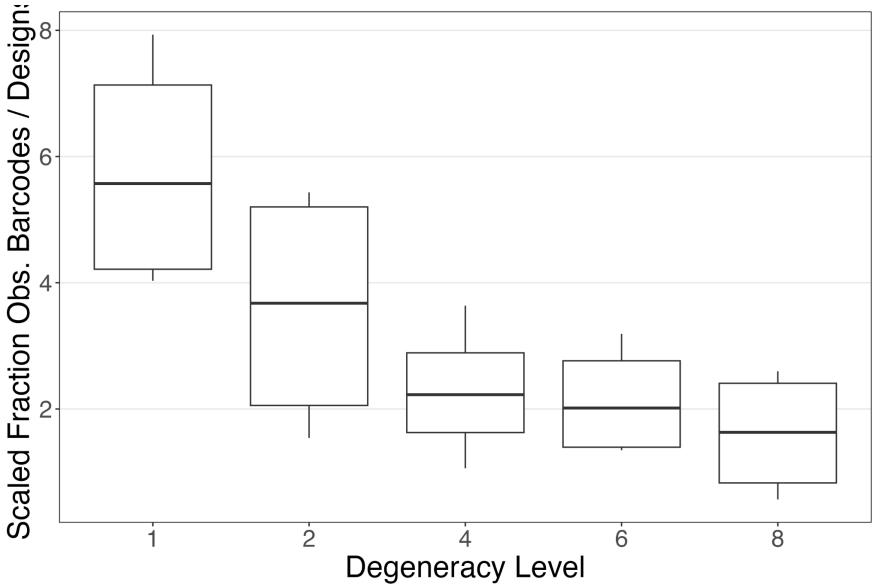
Appendix Figure A8 - Distribution of Deletion Lengths Across Libraries. A. A histogram of the length of deletions observed across all 4 libraries. While single-base deletions are dominant, a substantial amount of long deletions are observed. B. Many multi-base deletions are long. At any given base in a deletion, only 23.6% are from single-deletions, while 28.3% are deletions of 5-49 bp in length, 32.7% are deletions of 50 to 99 bp in length, 23.1% are deletions of 100 to 149 bp in length, and 8.1% are 150 bp or longer.



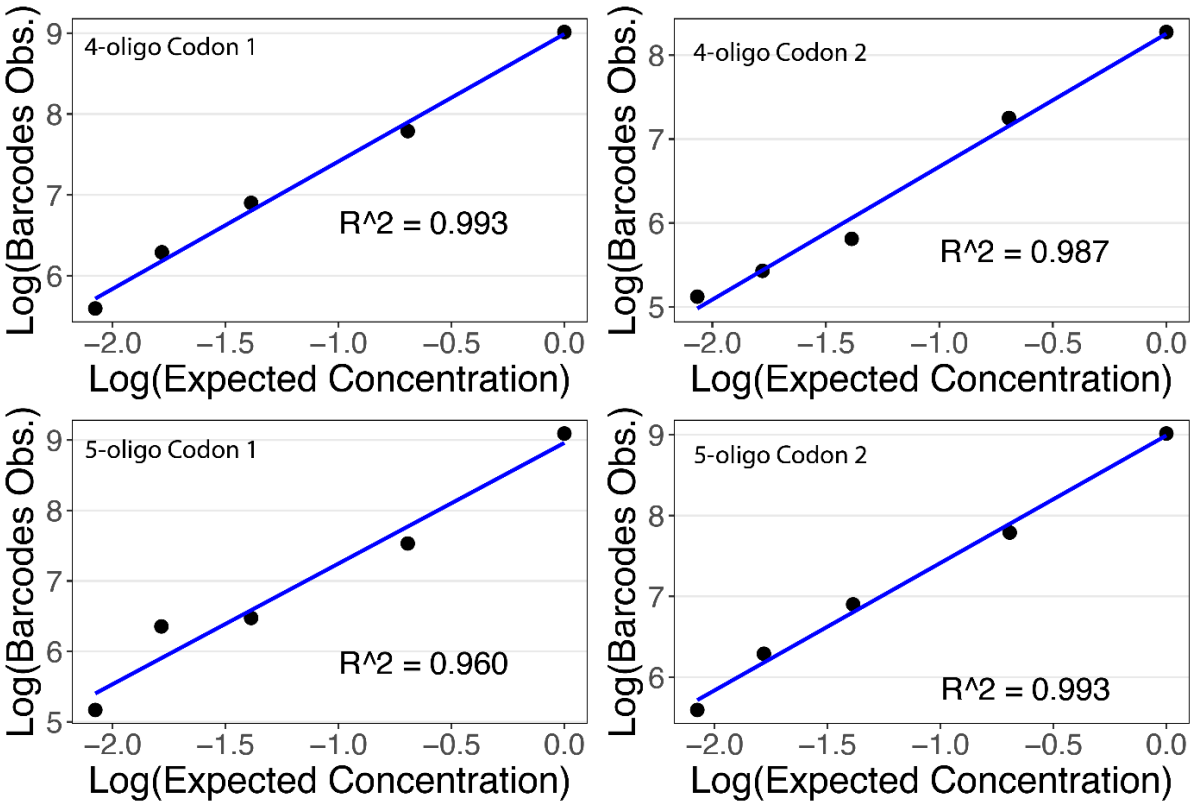
Appendix Figure A9 – Quality Metrics of Oligos Prior to Assembly. A 300-mer DropSynth library was nanopore sequenced after the nick processing and bead capture, but immediately prior to the DropSynth reaction. A total of 52 PCR cycles occurred between oligo synthesis and NGS sequencing. The reads for the 266-bp region of interest were aligned to the designed oligos, and the following statistics were generated. A. The distribution of percent perfects for 2,875 oligos showed a median value of 25.9%. B. A boxplot comparison of perfect oligos and oligos with mutations. C. The distribution of oligos showed a Gini coefficient of 0.53, with a rank ordering as shown in D.



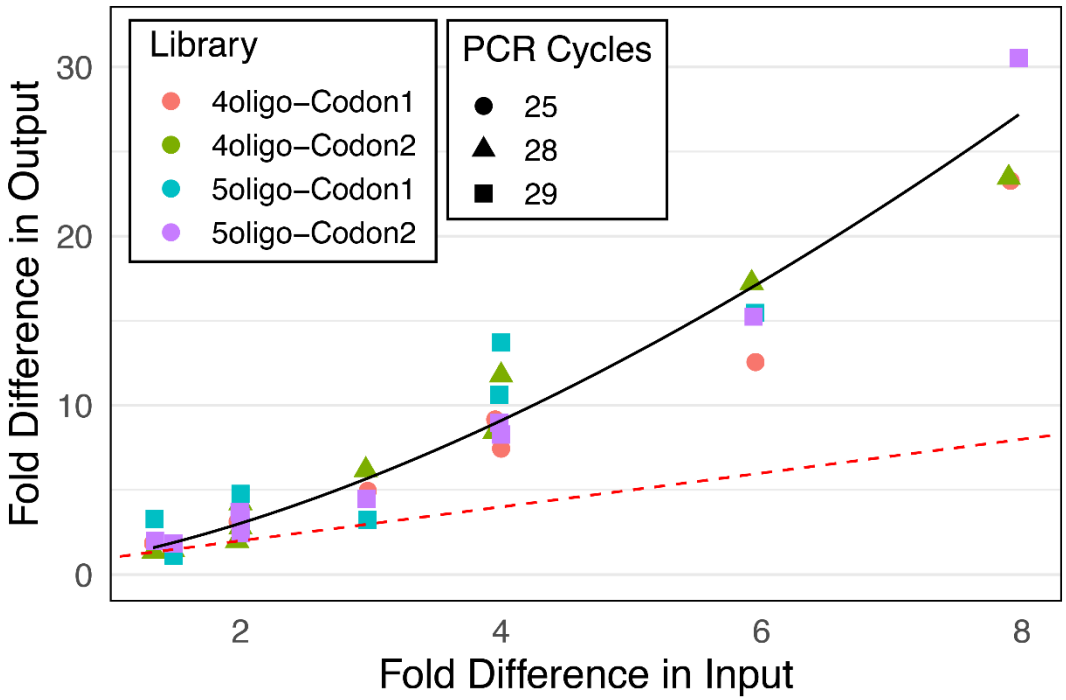
Appendix Figure A10 - Observed Barcode Fraction Relative to Design Fraction. The fraction of barcodes observed, normalized by the total fraction of designs. For each library, to calculate the fraction of observed barcodes, we determined the sum of all observed unique gene barcodes at a particular degeneracy level and divided it by the total number of observed unique gene barcodes in the library. We determined the fraction of designs at each degeneracy level by dividing the total number of variants at that degeneracy level by the total number of variants in the entire library. We see a strong decay in the observed barcodes as the degeneracy level is increased.



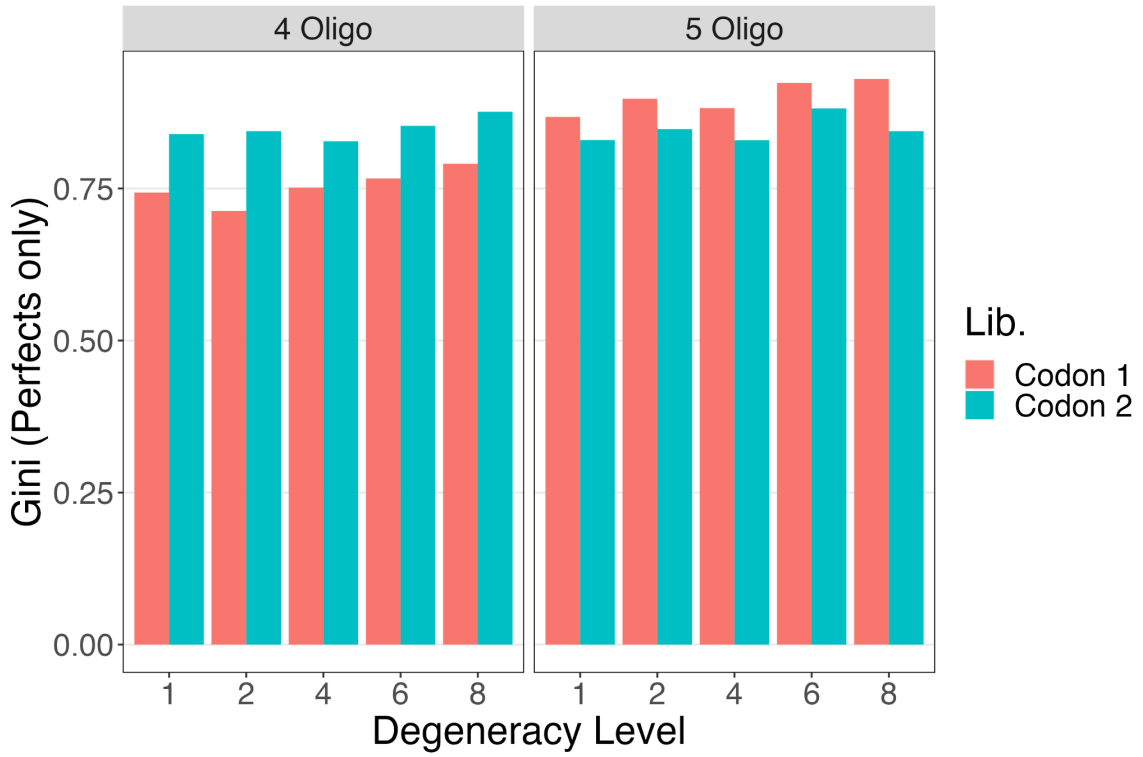
Appendix Figure A11 - Barcode Observation Scaled by Degeneracy Level. The fraction of barcodes observed, divided by the total fraction of designs, scaled by degeneracy level. In other words, we take the data from Appendix Figure A9 on the fraction of observed barcodes and multiply it by the degeneracy level. If we assume the amount of DNA for variants at the end of assembly is inversely proportional to the degeneracy level, we would expect roughly similar numbers (eg. a variant from a degeneracy level of 8 has ⅛ the amount of DNA as a variant from a degeneracy level of 1). Since we still see a strong decay, this implies that other factors such as PCR amplification contribute to the effect.



Appendix Figure A12 - PCR Amplification Model Across Libraries. A model of PCR amplification applied to all four libraries. The y-axis values are log transformed barcodes observed per variant, while the x-axis is the expected variant concentration given by the total number of barcoded beads with a given degeneracy level divided by the total number of variants at that level.

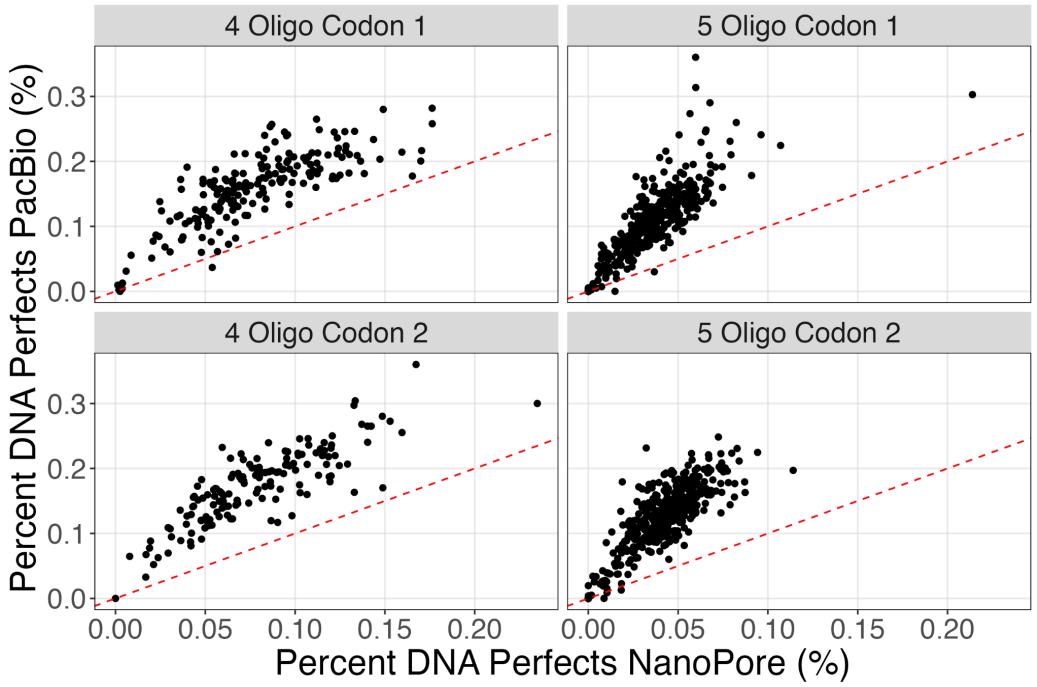


Appendix Figure A13 - Fold Differences in DNA Output vs Input by Degeneracy. The fold difference in DNA output between different degeneracy levels in each sample was plotted against the corresponding fold difference in DNA input. The dotted red line is the unity line, while the black line is a log-log fit to the experimental data [log10(output) = 0.00425+1.585997*log10(input)].

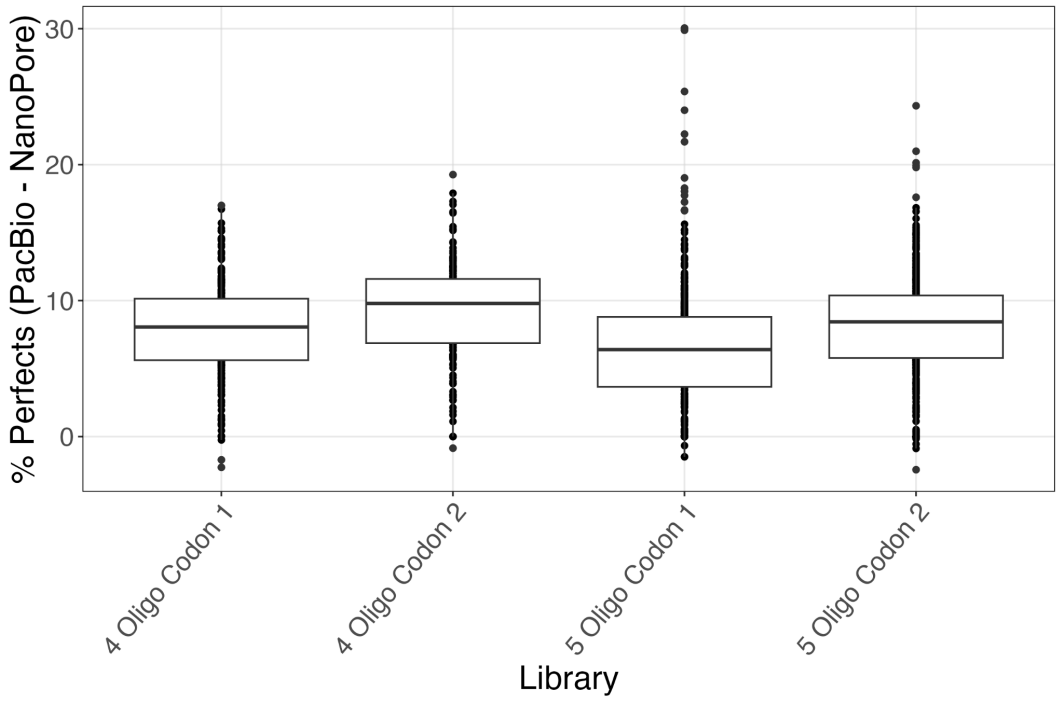


Appendix Figure A14 - Gini Coefficients Indicating Library Distribution Uniformity. The Gini coefficient is a measure of the inequality among the distribution of library members. The values observed are consistent with previous libraries assembled with DropSynth.

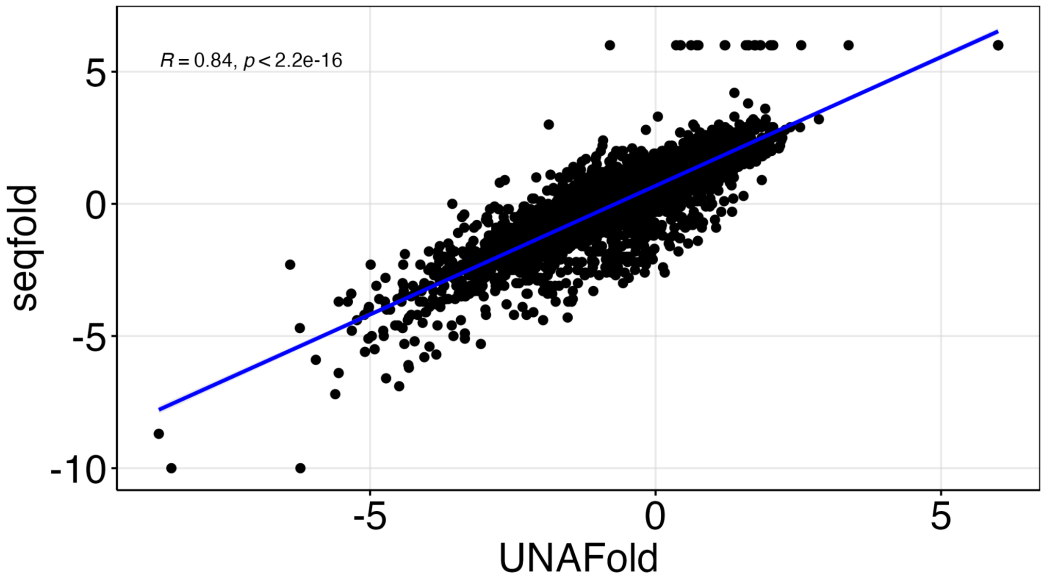
A.



B.



Appendix Figure A15 - Comparison of Perfect Rates: PacBio vs Nanopore sequencing. A. The percentage of perfects determined from the PacBio data (y-axis) plotted against the percentage of perfects determined using Oxford Nanopore data (x-axis). Data using only degeneracy level of 1. B. The distribution of delta percentage perfects shows a consistent (median) 7.8% higher rate for PacBio data, highlighting its lower error rate in sequencing.



Appendix Figure A16 - Correlation of Folding Energy Estimates Between seqfold and UNAfold. The folding energy of 4000 random 20 bp sequences determined using both seqfold and UNAfold shows a 0.84 correlation.

Appendix Table A1 - Primer Sequences Used in this Study.

Oligo Name	Sequence (annealing region highlighted)	Purpose
pSEVA121_AB_CARB_FWD	GAGAACGGTCTCCgtaaattagtagcccgcctaa	pSR348 → pSR348_Carb
pSEVA121_AB_CARB_REV	GAGAACGGTCTCCgggtcgtccaaaaaaaaagg	
pSR348_AB_CARB_FWD	GAGAACGGTCTCCacccccccagtatcagcccgtca	
pSR348_AB_CARB_REV	GAGAACGGTCTCCttacgaaacgatcctcatcc	
pSR348_AB_CARB_FWD	GAGAACGGTCTCCacccccccagtatcagcccgtca	pSR348_Carb → EnvZ_pSR348_Carb
pSR348_AB_CARB_REV	GAGAACGGTCTCCttacgaaacgatcctcatcc	
EnvZ_WT_FWD	GTCATCGGTCTCCacatgAGGCGATTGCGCTTC	
EnvZ_WT_REV	GTCATCGGTCTCCttaCCCTTCTTTTGTCGTGC	
pSR348_delKpnI_FWD	aggggatcctctagagtcgac	KpnI site directed mutagenesis
pSR348_delKpnI_REV	gtacctataaacgcagaaaggc	
FrgC_HKBC2_FWD	CACCTCGGTCTcAAGAAGAGCgcacGACGTcaCgtCgcaGAATTCcttttcggggaaatgtgcgcggaacccgggtaataataatctagaccaggcatc	Make Fragment C
FrgC_HKBC2_REV	GGCACTGGTCTcatAtgagttgtgcgatttaaattaggagag	
frgB_bH_envZ_FWD1	gcCACCTCGGTCTcggaCgaccgcacgctgctga	Make Fragment B - PCR 1
frgB_bH_envZ_REV1	gctGTTCgGCAGCgACACttaGGTACCTTATTGGGAGGTttacccttcttttgtcgtgc	
frgB_bH_envZ_FWD2	gcCACCTCGGTCTcggaC	Make Fragment B - PCR 2 - Add barcode
frgB_bH_envZ_REV2	tgcGCTGGTCTcCTTCTcctNNHHNDBBVVHHDDBBVVHHVVNNgctGTTCgGCAGCgACAC	
HK_PB_02_FWD	CTACACGACGCTCTTCCGATCTACACACAGACTGTGAGCACACAGCACTCTCCTAATTTAAATCGCACAACTCATATG	PacBio lib. prep.
HK_PB_02_REV	AAGCAGTGGTATCAACGCAGAGCTCACAGTCTGTGTGTCGTGACGTCGTGCGCTCTTCT	
HK_PB_03_FWD	CTACACGACGCTCTTCCGATCTACACATCTCGTGAGAGCACACAGCACTCTCCTAATTTAAATCGCACAACTCATATG	
HK_PB_03_REV	AAGCAGTGGTATCAACGCAGAGCTCTCACGAGATGTGTCGTGACGTCGTGCGCTCTTCT	
HK_PB_06_FWD	CTACACGACGCTCTTCCGATCTCATATATATCAGCTGTCACACAGCACTCTCCTAATTTAAATCGCACAACTCATATG	

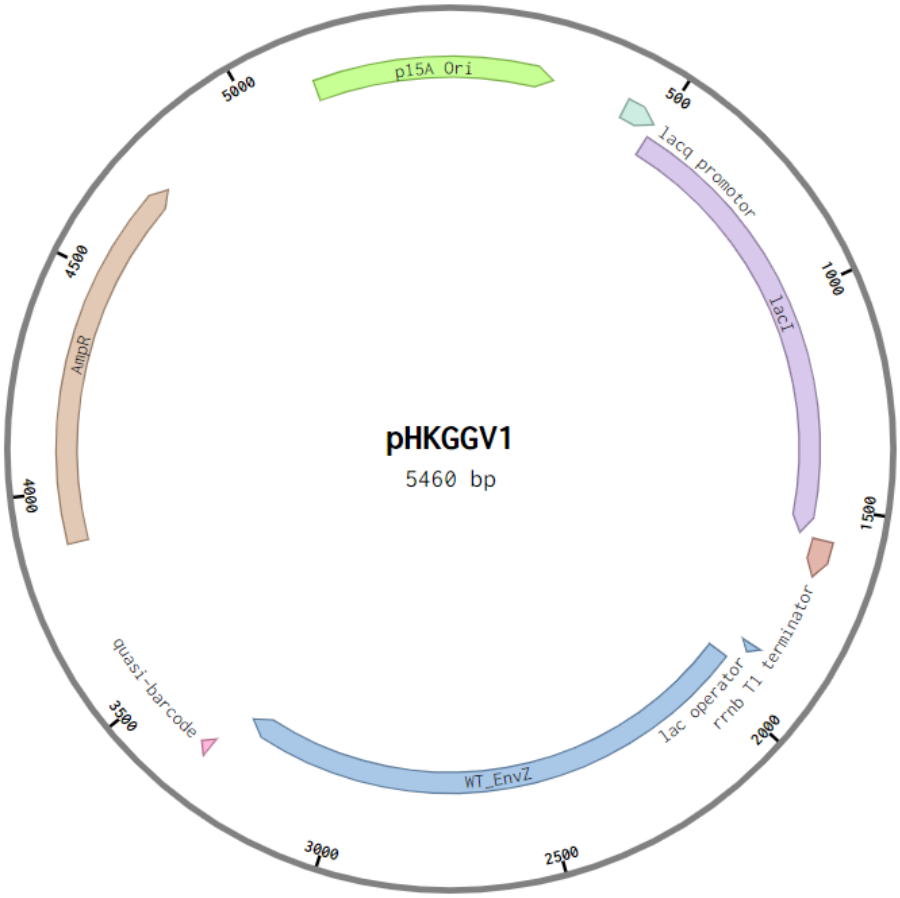
HK_PB_06_REV	AAGCAGTGGTATCAACGCAGAGACAGCTGATATATATGCGTGACGTCGTGCGCTCTTCT	
HK_PB_07_FWD	CTACACGACGCTCTTCCGATCTTCTGTATCTCTATGTGCACACAGCACTCTCCTAATTTAAATCGCACAACTCATATG	
HK_PB_07_REV	AAGCAGTGGTATCAACGCAGAGCACATAGAGATACAGACGTGACGTCGTGCGCTCTTCT	

Appendix Table A2 - Subpool Amplification Primer Sequences.

Library	**Oligo Name**	**Sequence**
4 Oligo Codon 1	skpp15-9-F filt15-453	Biotin-CGATCGTGCCCACCT
4 Oligo Codon 1	skpp15-9-R filt15-1189	GTGCGGGCTCCAACT
4 Oligo Codon 2	skpp15-13-F filt15-286	Biotin-GGGTTCGAGCGGGAG
4 Oligo Codon 2	skpp15-13-R filt15-11	TAGCGCGCAGAGAGG
5 Oligo Codon 1	skpp15-26-F filt15-16376	Biotin-GCGGCACCACAAACT
5 Oligo Codon 1	skpp15-26-R filt15-327	CGTGGCCTCTGTCCT
5 Oligo Codon 2	skpp15-28-F filt15-295	Biotin-GACTGCGGCGTTGGT
5 Oligo Codon 2	skpp15-28-R filt15-2129	TACGCCCGGGACAGA

Appendix Table A3 - CFU Counts After Transformation per Library.

Library	**Total CFUs**
4 Oligo Codon 1	2.49 x 10^6
4 Oligo Codon 2	2.1 x 10^6
5 Oligo Codon 1	14.1 x 10^6
5 Oligo Codon 2	66 x 10^6



Appendix Figure A17 - Map of Plasmid pHKGGV1. Our libraries are cloned into the N-terminal region of EnvZ. This plasmid is a derivative of plasmid pSR348 from Dr. Jeffrey Tabor's lab.

Appendix Table A4 - Sequences of Fragments Used for Golden Gate Assembly. The sequences of the three fragments used in Golden Gate to make libraries in plasmid pHKGGV1. The BsaI-HF-V2 recognition seq is shown in blue and overhang seq in red.

Fragment	Sequence
Frg_A - variable region	GGCACTGGTCTcacatatg NNNNNNNNNN gacgtgAGACCGAGGTGgc
Frg_B - C terminal portion of envZ and barcode	gcCACCTCGGTCTcggaCgaccgcacgctgctgatggcgggggtaagtcacgacttgc gcacgccgctgacgcgtattcgcctggcgactgagatgatgagcgagcaggatggctatctggca gaatcgatcaataaagatatcgaagagtgcaacgccatcattgagcagtttatcgactacctgcgcac cgggcaggagatgccgatggaaatggcggatcttaatgcagtactcggtgaggtgattgctgccga aagtggctatgagcgggaaattgaaaccgcgctttaccccggcagcattgaagtgaaaatgcaccc gctgtcgatcaaacgcgcggtggcgaatatggtggtcaacgccgcccgttatggcaatggctggat caaagtcagcagcggaacggagccgaatcgcgcctggttccaggtggaagatgacggtccggga attgcgccggaacaacgtaagcacctgttccagccgtttgtccgcggcgacagtgcgcgcaccatta gcggcacgggattagggctggcaattgtgcagcgtatcgtggataaccataacgggatgctggagc ttggcaccagcgagcggggcgggctttccattcgcgcctggctgccagtgccggtaacgcgggcg cagggcacgacaaaagaagggtaaACCTCCCAATAAGGTACCtaaGTGTcG CTGCcGAACagcNNBBDDBBVVHHDDBBVVHNDDNNaggAGA AGgAGACCAGCgca
Frg_C - Backbone	CACCTCGGTCTcAAGAAGAGCgcacGACGTcaCgtCgcaGAATTCc ttttcggggaaatgtgcgcggaacccgggtaataataatctagaccaggcatcaaataaaacgaaag gctcagtcgaaagactgggcctttcgttttatctgttgtttgtcggtgaacgctctctactagagtcacac tggctcaccttcgggtgggcctttctgcgtttataggtaccaggggatcctctagagtcgacctgcag gcatgcaagcttagcaagcgaaccggaattgccagctggggcgccctctggtaaggttgggaagc cctgcaaagtaaactggatggctttcttgccgccaaggatctgatggcgcaggggatcaagatctga tcaagagacaggatgaggatcgtttcgtaaattagtagcccgcctaatgagcgggctttttttaattcc cctatttgtttatttttctaaatacattcaaatatgtatccgctcatgagacaataaccctgataaatgcttca ataatattgaaaaaggaagagtatgagcattcagcattttcgtgtggcgctgattccgtttttgcggcgt tttgcctgccggtgtttgcgcatccggaaaccctggtgaaagtgaaagatgcggaagatcaactggg tgcgcgcgtgggctatattgaactggatctgaacagcggcaaaattctggaatcttttcgtccggaag aacgttttccgatgatgagcacctttaaagtgctgctgtgcggtgcggttctgagccgtgtggatgcg ggccaggaacaactgggccgtcgtattcattatagccagaacgatctggtggaatatagcccggtga ccgaaaaacatctgaccgatggcatgaccgtgcgtgaactgtgcagcgcggcgattaccatgagcg ataacaccgcggcgaacctgctgctgacgaccattggcggtccgaaagaactgaccgcgtttctgc ataacatgggcgatcatgtgacccgtctggatcgttgggaaccggaactgaacgaagcgattccga acgatgaacgtgataccaccatgccggcagcaatggcgaccaccctgcgtaaactgctgacgggt gagctgctgaccctggcaagccgccagcaactgattgattggatggaagcggataaagtggcggg tccgctgctgcgtagcgcgctgccggctggctggtttattgcggataaaagcggtgcgggcgaacg tggcagccgtggcattattgcggcgctgggcccggatggtaaaccgagccgtattgtggtgatttata ccaccggcagccaggcgacgatggatgaacgtaaccgtcagattgcggaaattggcgcgagcctg attaaacattggtaaaccgatacaattaaaggctccttttggagccttttttttggacgacccccagta

	tcagcccgtcatacttgaagctagacaggcttatcttggacaagaagaagatcgcttggcctcgcgc gcagatcagttggaagaatttgtccactacgtgaaaggcgagatcaccaaggtagtaggcaaataa gagctcgcttggactcctgttgatagatccagtaatgacctcagaactccatctggatttgttcagaac gctcggttgccgccgggcgtttttttattggtgagaatccaagcactagtaacaacttatatcgtatggg gctgacttcaggtgctacatttgaagagataaattgcactgaaatctagtaatattttatctgattaataag atgatcttcttgagatcgttttggtctgcgcgtaatctcttgctctgaaaacgaaaaaaccgccttgcag ggcggtttttcgaaggttctctgagctaccaactctttgaaccgaggtaactggcttggaggagcgca gtcaccaaaacttgtcctttcagtttagccttaaccggcgcatgacttcaagactaactcctctaaatca attaccagtggctgctgccagtggtgcttttgcatgtctttccgggttggactcaagacgatagttaccg gataaggcgcagcggtcggactgaacgggggttcgtgcatacagtccagcttggagcgaactgc ctacccggaactgagtgtcaggcgtggaatgagacaaacgcggccataacagcggaatgacacc ggtaaaccgaaaggcaggaacaggagagcgcacgagggagccgccaggggaaacgcctggt atctttatagtcctgtcgggtttcgccaccactgatttgagcgtcagatttcgtgatgcttgtcaggggg gcggagcctatggaaaaacggctttgccgcggccctctcactttcctgttaagtatcttcctggcatctt ccgggaaatctccgccccgttcgtaagccatttccgctcgccgcagtcgaacgaccgagcgtagcg agtcagtgagcgaggaagcggaatatatcccgaagcggcatgcatttacgttgacaccatcgaatg gtgcaaaacctttcgcggtatggcatgatagcgcccggaagagagtcaattcagggtggtgaatgtg aaaccagtaacgttatacgatgtcgcagagtatgccggtgtctcttatcagaccgtttcccgcgtggtg aaccaggccagccacgtttctgcgaaaacgcgggaaaaagtggaagcggcgatggcggagctga attacattccaaaccgcgtggcacaacaactggcgggcaaacagtcgttgctgattggcgttgccac ctccagtctggccctgcacgcgccgtcgcaaattgtcgcggcgattaaatctcgcgccgatcaactg ggtgccagcgtggtggtgtcgatggtagaacgaagcggcgtcgaagactgtaaagcggcggtgca caatcttctcgcgcaacgcgtcagtgggctgatcattaactatccgctggatgaccaggatgccattg ctgtggaagctgcctgcactaatgttcaggcgttatttcttgatgtctctgaccagacaccaatcaacag tattattttctcccatgaagacggtacgcgactgggcgtggagcatctggtcgcattgggtcaccagc aaatcgcgctgttagcgggcccattaagttctgtcTCGgcgcgtctgcgtctggctggctggcata aatatctcactcgcaatcaaattcagccgatagcggaacgggaaggcgactggagtgccatgtccg gttttcaacaaaccatgcaaatgctgaatgagggcatcgttccaactgcgatgctggttgccaacgat cagatggcgctgggcgcaatgcgcgccattaccgagtccgggctgcgcgttggtgcggatatctcg gtagtgggatacgacgatacagaagacagctcatgttatatcccgccgttaaccaccatcaaacagg attttcgcctgctggggcaaaccagcgtggaccgcttgctgcaactctctcagggccaggcggtgaa gggcaatcagctgttgcccgtctcactggtgaaaagaaaaaccaccctggcgcccaatacgcaaac cgcctctccccgcgcgttggccgattcattaatgcagctggcacgacaggtttcccgactggaaagc gggcagtgaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgttgtt tgtcggtgaacgctctcctgagtaggacaaatccgccgccctagacctagggcgttcggctgcggc gagcggtatcagctcactcaaaggcggtaatacgtaaatcactgcataattcgtgtagctcaaggcg cactcccgttctggataatgtttttttgcgccgacatcataacggttctggctaatattctgaaatgagctg ttgacaattaatcatcggctcgtataatgtgtggaattgtgagcggataacaatttcacacagcactctc ctaatttaaatcgcacaactcaTatgAGACCAGTGCC