|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| JUNE 2023 |  |  |  | A picture containing drawing  Description automatically generatedA picture containing drawing  Description automatically generated |

|  |
| --- |
| RAPID DESIGN AND ENGINEERING OF  SMART AND SECURE MICROBIOLOGICAL SYSTEMS |
|  |
| Argonne National Laboratory’s Secure Biosystems Design project |
| **FINAL REPORT** |

|  |  |  |
| --- | --- | --- |
| **Participating  National Laboratories** | **Participating  Universities** | **Laboratory  Research Manager** |
| Argonne National Laboratory (lead) | University of Chicago | Dionysios Antonopoulos |
| Lawrence Berkeley National Laboratory | Northwestern University | Office phone: (630) 252-3935 |
|  |  | Email: dion@anl.gov |

Title: RAPID DESIGN AND ENGINEERING OF SMART AND SECURE MICROBIOLOGICAL SYSTEMS

Authors:

Project Principal Investigator - Dionysios Antonopoulos1

Co-Principal Investigators - Christopher Henry1, Michael Jewett2, Arvind Ramanathan1

Investigators - Gyorgy Babnigg1, Michael Fonstein1, Mark Mimee3, Yasuo Yoshikuni4

Affiliations:

1Argonne National Laboratory, Lemont, IL, USA; 2Northwestern University, Evanston, IL, USA; 3University of Chicago, Chicago, IL, USA; 4Lawrence Berkeley National Laboratory, Berkeley, CA, USA

**Abstract**

The design and application of successfully engineered biosystems requires an understanding of how engineered microbes will interact with other organisms – either as one-on-one competitors or in the context of microbial consortia. Engineering microorganisms from first principles for non-laboratory, environmental, applications is inherently challenging because: (1) engineered systems tend to quickly revert back to their wild-type behaviors; and (2) these systems typically pay a price in reduced fitness making them uncompetitive against invasive contaminating species (i.e., metabolic burden). ~~A key question is how do sensing, signaling, and metabolism contribute to the stabilization and destabilization of these interactions?~~ Here, the organization, control, stabilization, and destabilization of natural and engineered microbes was investigated through synthetic biology. This approach enabled the development of (1) single-strain systems capable of detecting and responding to target organisms in the environment; (2) a pipeline for refining and engineering biological constructs in new non-model host organisms; and (3) improved systems for the rapid designing, engineering, and assaying of new biological modules. This approach sought to predictably safeguard system design across bacterial species, whioe specifically focusing on microbes that benefit the plant microbiome. A long-term goal beyond the proposed research is to enable the rational engineering of microbial communities based on first principles that mimic the performance of natural communities. ~~This will enable a new vision of biosecurity and biocontainment that harnesses the underlying mechanisms of resource management occurring within and between organisms.~~

**Acknowledgements**

[PLACEHOLDER TEXT]

**Table of Contents**

Title [title page – page i]

Authors [title page – page i]

Abstract [title page – page i]

Acknowledgements [page ii]

Table of Contents [page iii]

List of Figures [page iv]

List of Tables [page v]

Executive Summary [page 1]

*Summarizes problem, results, conclusions, recommendations*

Introduction [first page of text – page 2]

*States subject, purpose, scope and plan for developing report*

Project Results

Products Delivered from Project

Conclusions [final paragraph of text]

List of Acronyms and Abbreviations

Distribution List

**List of Figures**

**Fig. 1.** Building a biosensor containing strain capable of sensing another organism and then responds.

**Fig. 2.** Designing and engineering probabilistic decision.

**Fig. 3.** Switch gRNAs and GC bias.

**Fig. 4.** Overview of the CRISPR-ActML model.

**Fig. 5**. Construction of pJZ3 for CRAGE using lambda red recombination system.

**Fig. 6.** Developingcell free methods for engineering biosensors.

**List of Tables**

[PLACEHOLDER TEXT]

**Executive Summary**

~~Long-term objectives. The long-term objective for this project is to elaborate secure biodesign strategies for microbial systems that operate in the dynamic abiotic and biotic conditions of natural environments, thus enabling systems-level and rational biological design for field use.~~ The advent of synthetic biology tools provides an opportunity for systems-level approaches to design microorganisms and their communities for a range of custom applications or enabling their deployment in other contexts, including the natural environment. However, there are practical and ethical concerns for deploying these systems in natural settings, including unintended effects as well as how to track and discern what has been engineered from what has not. One approach for designing and engineering of microorganisms for non-industrial contexts is to implement design safeguards in cellular mechanisms that can controllably curtail or kill – effectively secure – an organism when needed.

There are several key challenges to incorporate safeguard systems at the design stage, including: (1) a lack of knowledge for how well safeguards operate across the diverse environmental and physiological conditions that organism experience; (2) a need to integrate the safeguards with other cellular designs so that it can sense and recognize specific signals from the intracellular or extracellular environment, and mediate a response; and (3) a need for rapid and reliable methods to engineer and optimize the biological components for safeguard construction and functional integration. Presently, the engineering and optimization of biological processes requires time-consuming construction of hundreds or even thousands of unique cell lines, each with a single genetically encoded design for a protein, a biosensor, or a biosynthetic pathway (e.g., a set of enzymes that yield a biochemical product of interest). The deliberate design, construction, and integration of numerous biological parts using current practices is cumbersome and resource-intensive using current practices. We propose that AI combined with robotics can resolve this primary challenge by powerfully assisting researchers in exploring large combinatorial spaces and by accelerating the design, build, test, learn (DBTL) cycle to develop new biological parts.

~~This project will demonstrate that AI and automation can be coupled to enable model-driven design of biosensors, metabolic pathways, and genetic circuits and assemble these designs into synthetic programs in virtually any microbe. We apply this coupled approach to design and build safeguard systems for intrinsic biocontainment that are predictable and portable across bacterial species, focusing on microbes that are part of the beneficial plant microbiome. Finally, we will integrate designed synthetic programs and safeguards together with cellular metabolism to develop smart and secure biological systems that sense their environment, recognize specific signals, respond by a complex change of phenotype, and can then be eliminated on-demand. The outcome of this project will be an integrated and generalizable framework to design synthetic programs and safeguard systems that predictively control phenotypic behavior of an engineered microbe within the context of a multi-species community.~~

~~Hypotheses. The external and internal factors a microbial cell faces in a natural environment are far broader than those found within the isolated bioreactor where the cell was engineered. This poses new questions: what are the design rules for engineering organisms that are competitive in natural communities, such as the rhizosphere, and while still carrying out the desired synthetic activity (e.g., promoting plant growth, or combating invasion by specific microbes)? These rules likely depend heavily on how the engineered organism interacts and compete with the native microbes in the environment.~~

In this project, we propose that recent technological advances in the fields of synthetic biology, artificial intelligence, and automation are creating the conditions for a paradigm shift in our understanding of the ways that cellular function can be designed at the level of bacterial communities. Uniting these advances, we hypothesize, will enable us to rationally design synthetic genetic programs (e.g., degrade specific nutrients, displace competitors, or produce plant hormones) that can be integrated into a microbial strain while conserving competitive and resilient attributes and the engineered safeguards for programmed elimination. This primary driving hypothesis gives rise to three principal challenges and associated sub-hypotheses:

*H3. Can we design smart and secure microbiological systems that recognize specific signals, respond by complex phenotypic changes and ultimately adapt their ecological behavior?*

We propose that genotype-to-phenotype predictions from mechanistic models can be leveraged to design an organism that can manipulate its environment in response to a specific stimulus to change the structure of a microbiome in which it resides.

*H2. Can we design safeguards for biocontainment that are predictable and portable across species so that engineered microbes can be securely eliminated if released in the environment?*

We hypothesize that the cell-killing activity of CRISPR systems can be predicted across species using machine learning techniques. This will enable design of efficient safeguards that are portable and can be programmed to prevent the release of engineered DNA in the environment.

*H1. Can we accelerate design and engineering of cellular functions to the point where synthetic genetic programs can be built in virtually any microbial system?*

We hypothesize that recent advances in AI and laboratory automation can be combined to create a closed-loop system that will provide an unprecedented capability to design, modify and control biological systems.

Proposed experimental design. The fundamental research to test the hypotheses above has been organized into three Aims:

**Aim 1**. **Develop a resilient single-strain system capable of detecting and responding to a target organism based on its secreted metabolites.**

In this aim, we will apply a model-driven design strategy to develop a platform strain that is: (1) capable of detecting and responding to the presence of another strain; (2) capable of eliminating another strain; and (3) retains its own internal safeguards so the platform strain can also be eliminated on demand. The development of a platform strain that is capable of targeted destruction of an invasive strain, followed by triggered self-destruction will have many practical applications of relevance to DOE, from plant growth health to biomanufacturing, as well as providing the base toolkit for ultimately engineering more complex microbial community interactions. Work will be divided into subtasks that increase the complexity of the proposed platform strain step-wise, by integrating new modules and regulatory responses. The final set of subtasks will develop the modeling approaches for predicting the overall system behavior arising from each tested platform design. Then we will apply these modeling methods to design optimal combinations of biosensor properties, response system activation, and the response itself to achieve a desired overall effect. More importantly, this system establishes many foundation stones for the development of more complex synthetic biology-based systems.

**Aim 2. Secure design of engineered model and non-model microbes**

Predictably designing safeguard systems that are portable across bacterial species is a key challenge for generalizing secure biosystems design. Safeguard systems must guarantee that genetically engineered cells can be eliminated if they deviate from assigned task or once the task is performed, and that engineered DNA constructs are not released and potentially transferred to other microbes. Here, we propose an AI-guided design of safeguards based on CRISPR systems which are potent cell killing mechanisms. We will apply this approach to design and build effective safeguard systems in plant-associated bacterial species, enabling us to develop strategies for securely engineering microbes that are part of the beneficial plant microbiome.

**Aim 3. Develop systems for designing, engineering, and assaying of new biological modules that enable stability in engineered microbial systems**.

Rationally and predictably designing cellular functions remains a complex and formidable challenge. To overcome this challenge, this aim will develop AI-driven, multiplexed cell-free methods that enable the rapid testing of large combinations of biological modules for engineering smart and secure microbiological systems. We will apply these methods to address three key needs for secure biosystems design: biosensors, metabolic pathways, and genetic programs. The result will be a generalizable framework for studying, engineering, and designing the synthetic programs that we will use in the project.

~~Investigators and affiliations. The Argonne Secure Biosystems Design project adopts a multidisciplinary approach requiring an integration of the scientific expertise and technical capabilities at Argonne and partnering institutions including the University of Chicago, Northwestern University, and Lawrence Berkeley National Laboratory (LBL). Key investigators include: Argonne – Dionysios Antonopoulos, Gyorgy Babnigg, Michael Fonstein, Chris Henry, and Arvind Ramanathan; UChicago – Mark Mimee; Northwestern – Michael Jewett; and LBL – Yasuo Yoshikuni.~~

~~The project team has been assembled according to their expertise and organized so as to complement and build opportunities for long-lasting scientific interactions. The opportunity to leverage local institutional connections between Argonne, UChicago, and Northwestern offers a tremendous opportunity to invest in an integrated postdoctoral pool. We anticipate that this postdoctoral group will serve not only as the developers of the research approaches and science embodied in this project, but will serve as future thought leaders in the area of secure biosystems design.~~

**Introduction**

Engineered microbial systems are central to various industrial processes such as bioproduction, biomaterials synthesis, and other applications such as bioremediation. Individual bacteria and communities are also being engineered to confer phenotypic traits that are beneficial to eukaryotic organisms hosting the microbiome, including plants and mammalian guts. Additionally, genome editing techniques such as CRISPR (clustered, regularly interspaced, short palindromic repeats)-associated systems (Cas) are paving the way for high-throughput engineering leading to new microbial phenotypes with novel functions. Taken together there is an opportunity to not only radically rework the laboratory process to customize microorganisms for environmental applications, but to do so in an automated and highly coordinated way that can enable more complex systems to be created, including full-scale microbial communities.

The rational design of microorganisms to be ultimately deployed in environmental settings faces several immediate practical challenges that this project aims to address: (1) the need for the rapid generation and optimization of biological parts encoding specific functions to be used in environmental contexts, (2) the requirement to generate intrinsic biocontainment mechanisms for multiple tasks of multiple organisms, and (3) an understanding of how to cohesively integrate these parts into an organism with resiliency to dynamic environments.

Engineering microorganisms from first principles for non-laboratory, environmental, applications is inherently challenging because engineered systems (1) tend to quickly revert to wild-type and (2) typically lose uncompetitive fitness against invasive contaminating species. Tackling these challenges is part of the longer-term objective for this project to elaborate secure biodesign strategies for microbial systems and enable systems-level and rational biological design for field use.

A key question underlying these challenges is how do sensing, signaling, and metabolism contribute to the stabilization and destabilization of natural and engineered microbial function? Biological systems gain their inherent complexity from layered checks and balances that collectively control mechanisms for the optimal competitive fitness in a given environment. Reverse engineering to experimentally tease apart these layers by the typical single-investigator laboratory to understand the underlying set of biological design principles is laborious and time-consuming. For example, optimizing expression of a 6 gene pathway, by changing just 4 nucleotides in the promoter of each gene to all possible permutations (256 possibilities per gene), leads to more than 1014 possibilities, which is far greater than typical transformation efficiencies.

Experimental strategies that synergize laboratory automation and machine-learning approaches to analyze data are being used to constrain the experimental space and more effectively utilize laboratory resources, especially bench time. But what if we went a step further and could generate a fully “closed-loop” laboratory platform that not only executes the laboratory assays required at scale (using automation) and analyze the resultant data (using machine-learning strategies) but automatically drives the subsequent round of experimentation? This type of an experimental vehicle make the aforementioned combinatorial space (1014 possibilities) tractable for experimental exploration.

**Project Results**

**Results from Aim 1.** *In vivo* apps - developing “smart” single-strain systems capable of detecting and responding to target organisms in the environment.

We applied a model-driven design strategy to develop a platform strain that is capable of detecting and responding to the presence of another strain. Ultimately, this would serve as a platform to eliminate another strain while retaining its own internal safeguards so that the platform strain can also be eliminated on demand. The platform strain of *Escherichia coli* was being initially designed in the context of a metabolic competition with *Pseudomonas fluorescens*. We chose these two organisms because *E. coli* had been identified as being capable of growing on maltose while *P. fluorescens* cannot, and conversely *P. fluorescens* can grow on 4-hydroxybenzoate (4-HB) but *E. coli* cannot (see **Fig. 1**). We planned to engineer *E. coli* to constitutively express a degradation pathway targeting 4-HB.

**Fig. 1.** Building a biosensor containing strain capable of sensing another organism and then responds.



**Specific tasks for Aim 1.**

1-1. Develop constitutively expressed metabolic competition against a target strain.

1-2. Develop mechanism for detection of a target strain and response activation.

1-3. Develop probabilistic response to mitigate metabolic burden of expensive responses.

1-4\*. Integrate toxin production pathway to attack the target strain.

1-5\*. Demonstrate programmed cell death.

1-6. Develop mechanistic modeling to support strain design and troubleshooting.

1-7. Develop spatial dynamic modeling of organism detection to support detection system design.

\*Based on the revised project scope (September 2020), Tasks 4 and 5 (regarding a strain capable of eliminating another strain) were de-emphasized, so that the focus would be on generating a strain capable of detecting and responding to the presence of another.

**Overview of Aim 1 activities.** Work to date focused on: (1) validating substrate utilization capabilities by bacterial strains to be used in metabolic competition studies, (2) developing quantitative methods to evaluate metabolite-dependent coculturing, (3) revealing secondary metabolite cross-feeding using a parallelized chemostat platform, (4) predicting interspecies metabolic dynamics using metabolic modeling, (5) developing a probabilistic decision circuit to mitigate the metabolic burden of expensive responses, and (6) designing metabolic modeling approaches to understand system impacts of engineering and predict strain adaptations.

Validating substrate utilization capabilities by bacterial strains to be used in metabolic competition studies. We described in our Science Plan a strategy to eliminate the target strain *P. fluorescens* SBW25 by an engineered *E. coli* MG1655 strain via metabolic competition. A set of potential unique carbon sources for each strain had been reported previously in the literature (Keseler *et al.*, 2017, Loper *et al.*, 2012). We initially measured carbon source utilization by our strains using BIOLOG phenotype microarrays (PM1 and PM2) and verified in liquid cultures that 4-HB is a unique carbon source for *P. fluorescens* SBW25, while maltose is uniquely utilized by *E. coli* MG1655.

Developing quantitative methods to evaluate metabolite-dependent coculturing. We then focused on developing quantitative methods to distinguish and track multiple organisms in a high-throughput manner using assays that are amenable to integration with laboratory automation. The key driving question here is, can we mimic more environmentally relevant contexts in our laboratory-based evaluations? We chose to visualize and enumerate a mixed community of *E. coli* MG1655 and *P. fluorescens* SBW25 via a plasmid expressing either mNeonGreen (*P.* *fluorescens* SBW25) or mRuby2 (*E. coli* MG1655) fluorescent protein. This plasmid has been shown to be stable in microbes associated with the rhizosphere (Wilton *et al.*, 2018). We have demonstrated the applicability of this plasmid to environmental contexts by showing that in the absence of the selection pressure (tetracycline), not only the individual cells, but also the mixture of them, constituting a minimal community, maintain the plasmid while colonizing a plant root (Noirot-Gros *et al.*, 2020). In co-culture experiments the labeled strains can be tracked by high throughput, automated methods, such as fluorometry and flow cytometry, as well as confocal microscopy and temperature dependent agar growth assays.

Revealing secondary metabolite cross-feeding using a parallelized chemostat platform. We sought to answer the question – can we better analyze metabolic signals within a co-culture while also better simulating an environmental context like the rhizosphere? Chemostat-based co-cultures meet these criteria due to their fixed media dilution rates, and thus fixed metabolic and growth states, as well as being able to establish a carbon-limited environment with constant nutrient turnover. We first assessed *E. coli* and *P. fluorescens* co-culture composition with, again, a fixed D-maltose concentration (10 mM) and different concentrations of 4-HB (0-6 mM) to determine whether different carbon source concentrations predictably tune co-culture composition. We found that there was a negative linear relationship (R2 = 0.97) between 4-HB concentration in the media and percentage *E. coli* in the co-culture. This relationship validated our proposed strategy to impact community composition via altered metabolite availability. Chemostats also provided us with tunable control of metabolic steady states for *E. coli* and *P. fluorescens* co-cultures, allowing us to further study the cross-feeding previously observed in bulk-liquid and microfluidic culture. Chemostats lacking 4-HB revealed that *P. fluorescens* persists at low levels (2-5% of co-culture) despite the absence of their primary carbon source.

Predicting interspecies metabolic dynamics using metabolic modeling. The discovery that *E. coli* is producing large enough quantities of secondary metabolites to support *P. fluorescens* growth (i.e., acetate and lactate) provided us with an opportunity to use flux balance analysis models to predict how this cross-feeding relationship might be disrupted. According to the models, complete elimination of acetate production by *E. coli* as a byproduct of growth has two negative consequences: (1) *E. coli* growth is detrimentally affected by up to 11%; and (2) alternative byproducts will be produced instead (e.g., ethanol, lactate, pyruvate), which *P. fluorescens* can also subsist on. Therefore, increasing the consumption of acetate by *E. coli* remains a more viable strategy to prevent cross-feeding. Indeed, an example in the literature showed *E. coli* cells in a carbon-limited chemostat evolved mutations that increase expression of acetyl-CoA synthetase and in turn allow cells to increase consumption of acetate (Treves *et al.*, 1998). To study these interactions further using models, we devised a modeling scheme whereby FBA is applied to predict normalized yields, uptake kinetics, phenotype expressivity, and byproduct production over time for *E. coli* and *P. fluorescens* from a range of primary carbon sources. These predictions for each nutrient consumption phenotype (e.g., *E. coli* growing on maltose, *E. coli* growing on acetate) were derived solely from genome sequences and by fitting phenotype biomasses, kinetics parameters, and bioproduction to optimally recapitulate the forementioned fluorescence growth data for each examined condition. Our model uniquely resolved the interconversion of species phenotypes -- *E. coli* cells switching from feeding on maltose to feeding on acetate – to intimately understand the community dynamics that govern the community interaction. This analysis confirmed that increasing acetate synthase activity in *E. coli* can improve *E. coli* competition with *P. fluorescens* when acetate is the primary syntrophic agent.

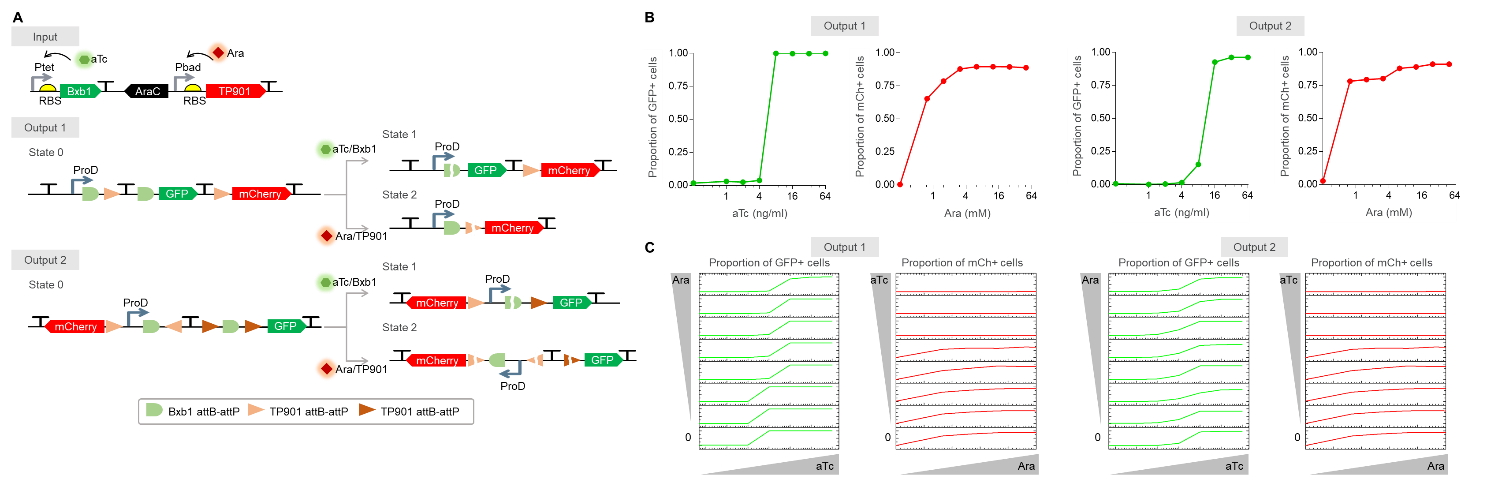
Applying novel metabolic models to investigate metabolic cross-feeding. The development of a phenotypic coculture model requires a robust experimental dataset to train and validate model predictions. We employed 96-well growth assays that continuously measure engineered *E. coli* and *P. fluorescens* fluorescent protein signals along with optical density, enabling deconvolution of each organism’s growth profile from mixed cocultures. With these assays we measured and analyzed over 200 unique combinations of initial coculture ratio (*E. coli* to *P. fluorescens*) and primary carbon source (i.e., D-maltose, 4-HB, and acetate). These experiments revealed several traits of *E. coli* and *P. fluorescens* metabolism that suggested acetate cross feeding is a key metabolic interaction between *E. coli* and *P. fluorescens*. First, we found that *P. fluorescens* consumes acetate much more rapidly than *E. coli* as a primary carbon source, confirming that if acetate is produced it will be mostly consumed by *P. fluorescens*. We next observed that *E. coli* growth on maltose is inhibited by increasing concentrations of acetate. These data help explain subsequent results that showed *E. coli* grow to a similar, or even greater, degree when cocultured with *P. fluorescens* compared to monocultured. We thus hypothesized that *E. coli* growing on maltose and cocultured with *P. fluorescens* expels acetate that typically inhibits growth but is instead consumed by *P. fluorescens*.

To investigate the molecular mechanism of this hypothesis, we deployed an *E. coli* Pta knockout strain that cannot produce acetate and instead excretes lactate (Castaño-Cerezo *et al.*, 2009). We repeated coculture assays with this new strain and found that *P. fluorescens* grew to a greater extent when cocultured with the knockout strain than wildtype *E. coli*. This is expected given lactate is a richer carbon source than acetate and suggests that the acetate pathway is directly involved in cross feeding. Armed with a wealth of kinetic data including knockout strain growth profiles, we next applied the metabolic model described above to this dataset. The model generated several qualitative and quantitative predictions that could be experimentally tested and validated by metabolomics. This approach and the modeling predictions enable us to assess the model’s accuracy via metabolomics and suggest model parameters to alter that improve the model’s accuracy.

The model predicts specific primary and secondary metabolic concentrations over time; thus, to validate the model, we measured primary and secondary small molecules via metabolomics. We analyzed coculture metabolite concentrations in 20 conditions that consisted of five timepoints, with either maltose alone or combined maltose and 4-HB as a primary carbon source, and with either a wildtype or Pta KO *E. coli* strain. These experiments provided a wealth of information and validated the quantitative and qualitative predictions made by the metabolic model. Specifically, the metabolomics definitively demonstrated that acetate is the syntrophic agent, where it is produced and immediately consumed when PF grows. Moreover, adding 4-HB, which is only consumed by *P. fluorescens*, does not impact the time of acetate production, but does delay the time of consumption of acetate, suggesting that PF is the one consuming acetate. This trend is also observed in the knockout cells that produce lactate instead of acetate. Finally, the broad-spectrum metabolomics panels did not discover any other metabolites that (1) changed at the same timepoint that *P. fluorescens* grows and (2) were produced at a high enough concentration to support the observed *P. fluorescens* growth. These data indicate that *E. coli* produce high concentrations of short chain fatty acids at later growth stages that enable *P. fluorescens* to grow despite lacking a suitable primary carbon source.

Developing a probabilistic decision circuit to mitigate the metabolic burden of expensive responses. To mitigate the metabolic burden of expensive cellular responses, we implemented a probabilistic decision circuit to differentiate the engineered cell population into distinct states with defined proportions (ultimately, “producer” and “defender” cells) upon exposure to an environmental cue**.**As shown in **Fig. 2A**, the probabilistic decision circuit contains a combination of serine integrases that catalyze precise rearrangement of DNA determined by a cognate pair, attP (derived from a phage) and attB (derived from its bacterial host). Integrases excise the DNA between them if a cognate attP-attB pair are aligned with the same polarity. In contrast, integrases invert the DNA between them if a cognate attP-attB pair are anti-aligned with opposite polarity. The input module in our circuit contains two serine integrases under inducible promoters - Bxb1 under Ptet induced by anhydrotetracycline (aTc) and TP901 under Pbad induced by arabinose (Ara). Two output modules are distinctively ordered with constitutive promoter ProD, the cognate pairs of Bxb1 and TP901 attP-attB sites, terminators, and fluorescent protein genes (GFP, mCherry). The expression of each fluorescent protein represents the differentiated population in different states induced by aTc or Ara. Without any inducers, GFP and mCherry expressions are prevented by terminators before genes or opposite polarity of ProD (state 0). However, in the presence of chemical inducers, integrases are activated and trigger the rearrangement of DNA, resulting in the differentiation of cell populations into different states (state 1/2). In the case of output1 modules, activation of Bxb1 leads to the excision of a terminator between the Bxb1 attP-attB pair, which drives GFP expression (state 1). Activation of TP901 leads to the excision of GFP gene and terminators between the TP901 attP-attB pair, which drives mCherry expression (state 2). In the case of output2 modules, activation of Bxb1 leads to the excision of a terminator between the Bxb1 attP-attB pair, which drives GFP expression (state 1). Activation of TP901 leads to an inversion of ProD between the TP901 attP-attB pair, which drives mCherry expression (state 2).

**Fig. 2.** Designing and engineering probabilistic decision. (A) A detailed state diagram of an input module and two output modules demonstrating the results of DNA rearragenment induced by aTc or Ara. (B) The proportion of fluorescence-positive cells after treatment of aTc (0, 1, 2, 4, 8, 16, 64 ng/ml) or Ara (0, 0.78, 1.56, 3.125, 6.25 mM). *E. coli* co-transformed with input-output1 or input-output2 were cultured at 37°C for 18 hours after treatment of aTc or Ara. (C) The proportion of fluorescence-positive cells after simultaneous treatment of inducers.



As shown in **Fig. 2B**, with both output modules, the proportion of GFP-positive(+) cells increased as the concentration of aTc increased. Likewise, the proportion of mCherry+ cells increased as the concentration of ATC increased. GFP+ cells and mCherry+ cells were hardly detected when Ara and ATC, respectively, were treated (data unshown). To test if this circuit can induce differentiation by two different inducers in the environment, we treated aTc and Ara simultaneously and analyzed the proportion of fluorescence+ cells (see **Fig. 2C**). Because Bxb1 and TP901 rearrange DNA in output modules competitively, the proportion of fluorescence+ cells negatively affected the opposite proportion, except in a high concentration of aTc (>16 ng/ml) when mCherry+ cells were hardly detected regardless of Ara concentration. The distribution of differentiated cells in the population depends on the ratio of Bxb1 to TP901 activity. The biased activity of one integrase can interrupt the recognition of environmental cues. Therefore, to minimize the biased activity of one integrase, we will optimize the ratio of integrase activity by screening for appropriate RBS strengths, degradation tags, and integrases with different catalytic efficiencies. With the initial proof-of-concept using separate inducible gene circuits, we will embed the ratio of integrase activity into a singular input such that a single small molecule inducer can trigger the differentiation into defined proportions of distinct states.

Designing metabolic modeling approaches to understand system impacts of engineering and predict strain adaptations. Our strategies for designing *smart* strains will require significant engineering efforts. Our 4-HB pathway engineering progress demonstrates our ability to design and enhance specific enzyme activities, but equally important is demonstrating an ability to integrate enhanced pathways into host strains, as well as rearranging strain metabolism to meet specific design needs (e.g., modifying substrate preference and byproduct profiles as described in our metabolic competition experiments). Rational engineering of strain behavior remains a grand challenge, but we have developed an integrated approach to accomplish this task by combining three technologies: (1) automated laboratory systems (see Aim 3 section below) to rapidly generate thousands of measurements from distinct combinations of strains grown under different conditions; (2) metabolic models capable of predicting strain response to small engineered perturbations through approaches like minimization of metabolic adjustment (MOMA/ROOM; Segrè *et al.*, 2002); and (3) machine learning which is capable of combining complex experimental data to predict higher-order strain responses. Specifically, we applied a sophisticated existing metabolic model of *E. coli* (Monk *et al.*, 2017) to predict a wide range of single gene knockouts, knockdowns, or inductions that would likely result in overproduction of threonine. Next, we applied our automated laboratory systems (see Aim 3) to assay the performance of 15 of these model-designed single interventions, as well as testing a selection of hundreds of combinations of multiple interventions. We note that while our metabolic models were >90% successful at predicting the impact of single interventions, accuracy declined dramatically when attempting to predict successful combinations. However, we then applied machine learning to analyze the results from omics data collected from a small random sampling of combined strain modifications to then predict more sophisticated ideal combinations. Based on these ML-proposed designs, we were able to combine 4-5 strain modifications to produce strains with threonine production that is two to three times higher than production from commercial strains available today (Lee *et al.*, 2017). We are continuing to improve this integrated approach by increasing the cycling between model and ML driven design, as well as streamlining and expanding the experiments that can be performed and the data that can be collected by our automated laboratory systems.

**Results from Aim 2.** Biocontainment – a pipeline for refining and engineering biological constructs in new non-model host organisms.

For this aim, we are focused on predictably designing safeguard systems that are portable across bacterial species, which is a key challenge for generalizing secure biosystems design. Safeguard systems must guarantee that genetically engineered cells can be eliminated if they deviate from assigned tasks or after completing a specific task, and that engineered DNA constructs are not released and potentially transferred to other microbes. One such safeguard is using controlled activation of the CRISPR nuclease Cas9 with a self-targeting guide RNA (gRNA) to break the bacterial chromosome and kill the cell. However, like gRNAs employed to engineer specific changes to a genome, gRNAs directing Cas9 cleavage for a kill-switch vary significantly between screens. This variability is both a challenge, in that previously developed gRNA screens are wholly inadequate for building transferable kill-switches, and an opportunity to discover underlying causes of this variability and develop prediction tools that are adaptable and robust. We hypothesize that dynamic gene expression responses in varying physiological conditions influences the cell killing activity of the CRISPR/Cas9 system.

**Specific tasks for Aim 2.**

Aim 2 – Task 1: Genome-wide profiling of CRISPR nuclease-gRNA cell killing activity in various organisms.

Aim 2 - Task 2: Leveraging AI-PED to predict gRNA activity and gRNA design rules.

Aim 2 - Task 3: High throughput *in vivo* characterization of genetic parts and inducible systems.

Aim 2 - Task 4: Assemble and test safeguards in various host species.

Aim 2 - Task 5: Safeguarding engineered hosts and synthetic constructs.

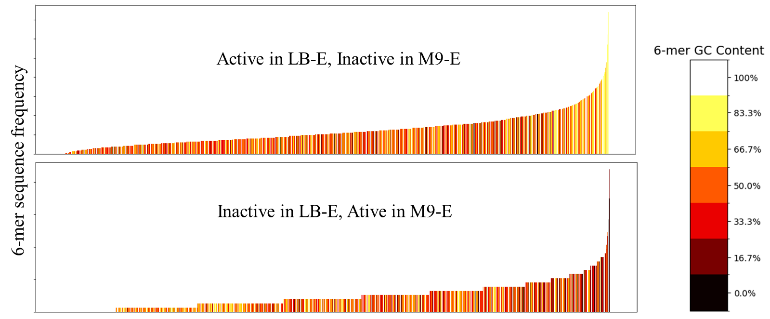
~~\*Based on the revised project scope (September 2020), Tasks 4 and 5 (regarding a strain capable of eliminating another strain) were de-emphasized, so that the focus would be on generating a strain capable of detecting and responding to the presence of another.~~

**Overview of Aim 2 activities.** ~~Work to date focused on: (1) validating substrate utilization capabilities by bacterial strains to be used in metabolic competition studies, (2) developing quantitative methods to evaluate metabolite-dependent coculturing, (3) revealing secondary metabolite cross-feeding using a parallelized chemostat platform, (4) predicting interspecies metabolic dynamics using metabolic modeling, (5) developing a probabilistic decision circuit to mitigate the metabolic burden of expensive responses, and (6) designing metabolic modeling approaches to understand system impacts of engineering and predict strain adaptations.~~

Large scale unbiased test of gRNA function reveals physiology specific behavior. To test our hypothesis, we developed an initial library of gRNAs to generate a substantial dataset to train our gRNA activity prediction model, CRISPR-ActML. Our library consisted of ~180,000 gRNAs spanning the *E. coli* MG1655 chromosome, as well as negative control guides (~20,000) that did not have a matching location in the chromosomal sequence. We hypothesized that different bacterial growth conditions would produce changes in cellular behavior (e.g., gene expression) resulting in varying gRNA activity between the conditions. Using this library, we conducted screens in 3 physiological conditions: rich (LB) media in exponential phase growth (LB-E), rich media in stationary phase growth (LB-S), and defined (M9 with glucose) media in exponential growth (M9-E). We induced Cas9 expression with anhydro-tetracycline and then sequenced the gRNA expression plasmids in the population remaining over time. In this system, plasmid depletion serves as a proxy for gRNA activity. We also collected transcriptome data from each of the physiological conditions, allowing us to correlate gRNA function with gene expression.

Our screens revealed that many guides cut well across physiological conditions, yet we identified ~6,000 gRNAs that were substantially overrepresented for physiology-specific function via ANOVA. Sequence analysis of guides active in M9-E and inactive in LB-E revealed that they were enriched for specific motifs (e.g. “GGCGGG”) and overall biased towards higher GC content (see **Fig. 3**). A small subset of gRNAs (174) exhibited extreme behavioral changes across conditions. These “outlier switch” guides had outstanding killing activity in one or two physiological conditions, but no activity in at least one other condition. 55% of outlier switches were active in M9-E and inactive in LB-E.

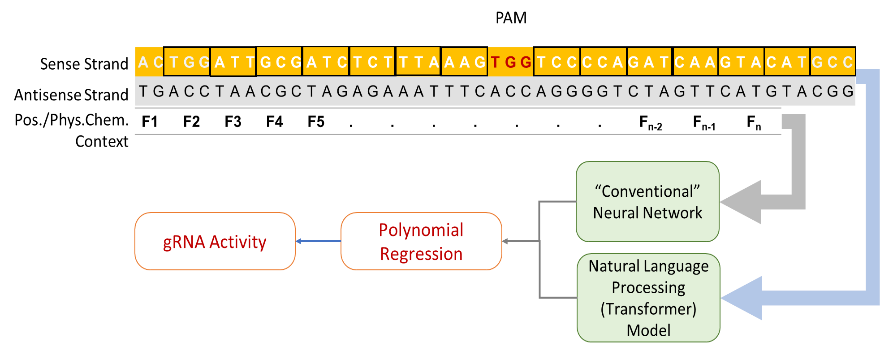
**Fig. 3.** Switch gRNAs and GC bias. Displayed here are the frequencies of each possible 6-mer within the population of gRNAs. 6-mer combinations were colored according to GC content, with higher GC percent indicated by lighter color. GC content is higher among guides that cut well in M9-E but inactive in LB-E.



We also examined gRNA activity as a function of chromosomal location. Active guides are distributed throughout the chromosome, but there are regions in which inactive guides tend to be found more often. These locations are not identical across physiological conditions. Both LB-E and LB-S have broadly similar and substantial regions of inactive guides, whereas in the M9-E physiological context, the low gRNA activity regions are less prominent. Furthermore, we found that LB-E, LB-S, and M9-E activity highly correlated with close proximity to certain nucleoid-associated protein binding sites within the genome in a strand-specific manner, including histone-like nucleoid structuring protein, Fis protein, and macrodomain Ter protein. Either greater access of Cas9 to the entire chromosome or more limited repair of double stranded breaks in M9-E could explain these data. Finally, we examined the chromosomal location of guides that showed physiologically specific function. We found that switch guides (guides that are active in one condition but inactive in another) are largely clustered in six specific regions of the genome. While the reasons for this localization are unclear, these data collectively demonstrate that gRNA behavior varies in different physiological conditions.

CRISPR-ActML model captures underlying features that result in physiology specific gRNA function.Our genome spanning dataset captured physiologically-dependent gRNA behavior in ~530,000 data points used to develop CRISPR-ActML. CRISPR-ActML combines a natural language processing (NLP) model based on the Google AI ALBERT architecture (Lan *et al.*, 2019) with a conventional neural network (NN) model to predict the efficiency of gRNAs. The NLP model treats genomic DNA as a “language” and is pre-trained on a representative set of 12 *E. coli* genomes (Abram *et al.*, 2020) divided into “sentences” of 7 “words” that each consist of three bases; this model is then fine-tuned on the gRNA activity dataset. The NN model uses a set of 428 features including gRNA positional and physicochemical properties (Guo *et al.*, 2018), as well as others derived from experimental parameters. The results of these models are then combined through a polynomial regression to predict the gRNA fold change as a proxy for bacterial killing (see **Fig. 4** on next page). CRISPR-ActML achieves a Mean Square Error of 0.13 and Spearman Correlation Coefficient of 47.31% on a test dataset of 71,228 gRNAs. By leveraging physiological conditions, CRISPR-ActML offers a resilient basis for transfer learning in novel organisms and environmental conditions without undergoing extensive costly retraining.

**Fig. 4.** Overview of the CRISPR-ActML model. The model consists of two parts: (1) a natural language processing (NLP) model that uses the DNA sequence upstream and downstream of the gRNA target’s PAM, and (2) a neural network utilizing physicochemical features of that context. These separate predictions are then combined through a polynomial regression to predict the percent fold change in cell population as a proxy for gRNA activity.



Assessment of gRNA activity reproducibility provides targets for prediction.Since the ultimate goal of CRISPR-ActML is to be able to harness our gRNA function data and readily available genome sequences to predict gRNA behavior in diverse new physiologies, we needed a way to assess the robustness of measurement of guide behavior to infer the limits of guide prediction. To this end, we performed a second round of examination of the genome spanning guides. We found that the majority of guides behaved consistently between trials with an overall Pearson correlation coefficient of 0.57 and mean absolute error of 0.62. However, there is a marked tendency for gRNAs in the 2021 experiments to have less cutting activity overall in comparison to the 2019 experiments (mean 2019 gRNA activity of -0.99 log2 fold change vs -0.53 in 2021). These datasets lay the foundation for creating models not just to predict gRNA activity, but to predict if a gRNA will behave consistently across trials, and also set bounds for reasonable performance expectations for activity predictors – i.e., the deviation between trial sets provides a theoretical upper bound on prediction accuracy.

Development of a framework for fine-tuning models based on organism specific features.We also investigated the potential for our model to transfer between species through a comparative survey of ~16,200 *E. coli*, ~450 *Shigella*, ~14,600 *Salmonella*, and ~7,800 *Pseudomonas* genomes retrieved from the PathoSystems Resource Integration Center (PATRIC) database (Davis *et al.*, 2020). We used Mash (Ondov *et al.*, 2016) to gauge similarity between whole genomes within each species, in addition to comparing the level of similarity of each genome’s set of CRISPR targetable sites for 5 different Cas9 species and AsCas12a. These comparisons demonstrate that while the genomes of the four genera differ significantly in characteristics like length and GC content, within each genus certain commonalities emerge – particularly that Cas-targetable sites within each genus’ genomes are as well-conserved as the surrounding genomic context and are even more highly conserved within core genes. This extends to conserved patterns in the relationship between the frequency of Cas-targetable sites within a genome and that genome’s GC content. We found that Cas-targetable sites with GC rich PAMs occur less frequently than expected (based on a genome’s A, C, G, and T contents), while AT-rich PAMs tended to occur more frequently than expected. Cas-targetable sites with roughly equal GC/AT content tended to occur far less frequently than expected. This conservation of genomic architecture is a strong indicator that patterns our CRISPR-ActML model learns in one species’ (or genus’) context will with proper fine-tuning of the model transfer to other prokaryotic species or genera. Using Mash distances, we compared accessible *E. coli* strains to identify a strain as genomically distinct from MG1655 as possible (ATCC35218).

Our pilot studies revealed intriguing physiological differences between the *E. coli* strains, with ATCC35218 growing much more rapidly in M9-E compared to MG1655. They also suggested performance of gRNA activity can vary according to osmotic stress. Therefore, we established an equivalent scale (200k) library, and conducted screens in the three prior physiological contexts (LB-E, LB-S and M9-E) as well as LB media with 0.6M NaCL (LB-O, for osmotic shock).

*Environmental factors impact Cas9 kill switch activity.* We hypothesize that environmentally relevant parameters, such as pH and salinity, modulate Cas9 kill switch efficiency. Research to date has been focused on optimizing Cas9 kill switch efficacy by targeting different genes for double strand cleavage and adding multiple copies of the Cas9 protein and gRNA in the target organism’s genome. These exciting works have advanced the feasibility of Cas9 kill switches as viable biocontainment tools, but overlooked whether environmental factors that would be present in real-world deployment of Cas9 kill switches modulate biocontainment efficacy. Our study is the first published research to fill this knowledge gap. We sought to determine the impact of various environmental signals on killing efficiency of a Cas9 kill switch engineered into Escherichia coli. We initially screened dozens of environmentally relevant conditions to determine which may impact Cas9 killing efficiency. We next performed a series of combinatorial experiments to parse how environmental factors alter different aspects of Cas9 kill switch efficacy including cell killing and cell recovery.

Our initial screen to assess Cas9 kill switch efficacy after exposure to environmental factors employed Biolog GenIII plates, which possess several dozen environmentally relevant carbon sources and stressors. We selected the Cas9-gRNA system EC20 due to its demonstrated efficacy in multiple growth media and phases of cell growth. We grew EC20 in Biolog GenIII plates and induced the kill switch with aTc approximately when cells entered log phase. We observed numerous instances where fundamental environmental factors, such as salinity or pH, modulated the efficacy or dynamics of the Cas9 kill switch. We also reproducibly observed increased or decreased kill switch efficacy depending on which antibiotics were present prior to kill switch activation. Finally, Cas9 kill switch efficacy as measured by difference in yield ranged significantly depending on carbon source availability. For example, cells with and without aTc addition grew with nearly identical growth dynamics when consuming α-D-lactose as a carbon source, but cells consuming D-galactose were severely inhibited by kill switch induction. Together these results demonstrate that individual preexisting environmental factors prepare cells, such as pH or metabolites, to increase or decrease their resistance to Cas9 induced genomic cleavage.

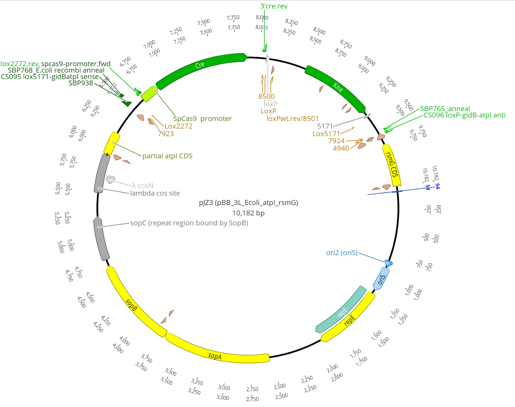
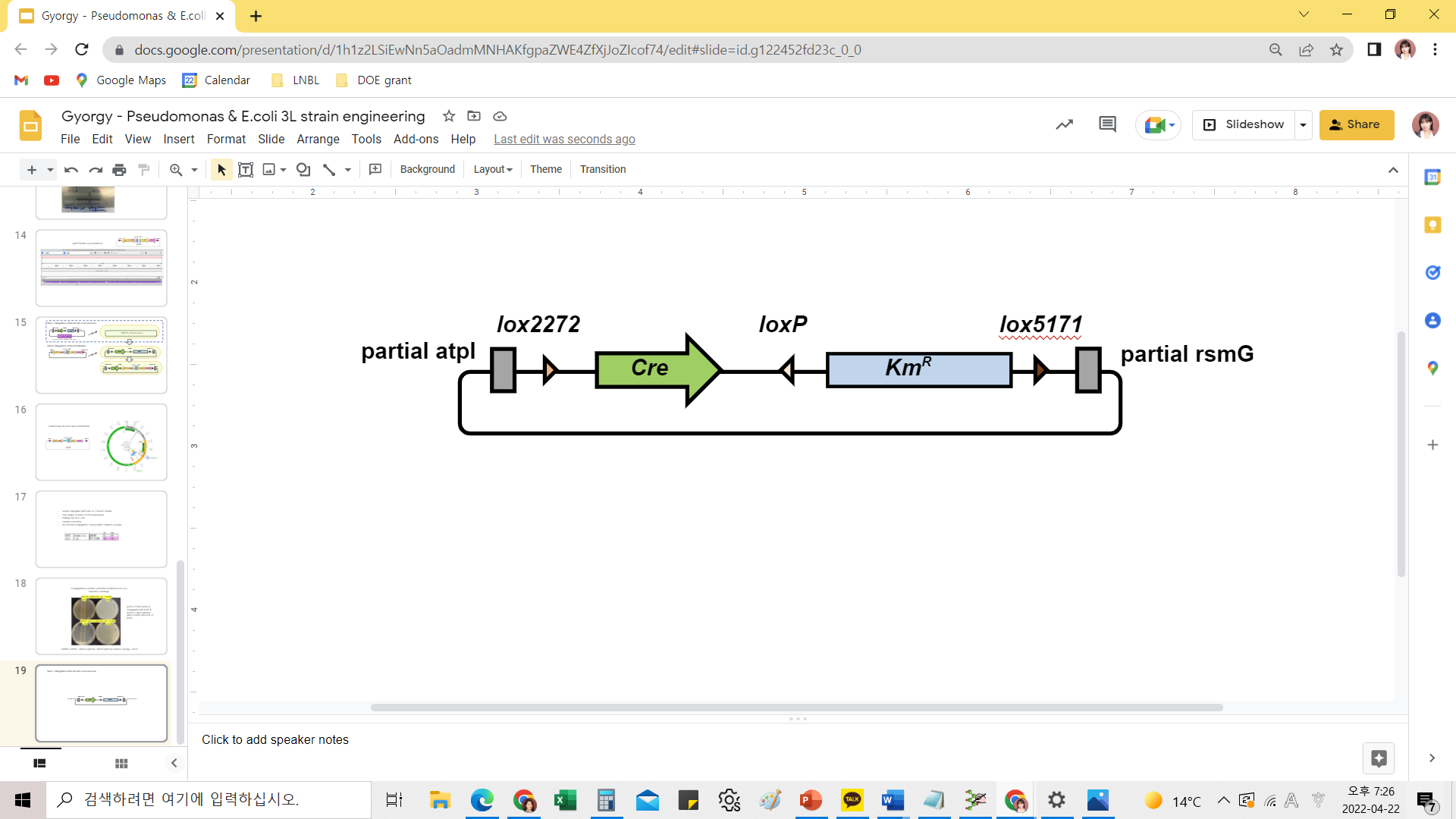
One caveat to using GenIII plates is that their exact components and concentrations are proprietary. Thus, we pursued several of these conditions in experiments with fully controlled media compositions to investigate the effects of several environmental factors on kill switch efficacy. We began by measuring not only the growth but also the percent cell killing of the EC20 kill switch at different pH. We inoculated 96-well plates containing M9 media at different pH with EC20 cells, activated the kill switch after 5h, and measured kill switch efficacy by cell plating. As expected, we observed decreased cell growth at lower pH for unactivated EC20 cells. We also found that the kill switch’s impact on cell growth remained similar to unmodified M9 at pH 6, but decreased significantly at pH 5. We suspected that cell killing may be dramatically lower at pH 5 and thus diminish the kill switch’s impact on cell growth. We did observe a >5 fold decrease in cell killing at pH 5, but >99% of cells were still killed. It’s possible, though unlikely, that cells with DNA damage recover more quickly at lower pH. To assess cell recovery after kill switch activation we inoculated different pH media with prekilled EC20 cells. As expected, we observed significantly diminished cell recovery at pH 5 and 6 compared to unmodified M9. Thus, there was no increase in cell recovery with decreasing pH. These data, although somewhat unexpected, confirm that environmentally relevant pH modulates kill switch efficacy to a significant degree.

Like with pH, salinity altered kill switch efficacy in GenIII plate experiments, warranting followup investigation in a controlled setting. As expected, we observed decreased cell growth with increasing salinity for EC20 cells without aTc, with 8% wt added NaCl resulting in no growth. We also found that the kill switch’s impact on cell growth either remained the same or increased slightly with increasing salinity. Surprisingly, the percent cell killing did not match observed differences in cell growth, with 4% wt added NaCl or greater causing nearly complete inhibition of cell killing. Furthermore, although 8% wt added NaCl prevented all cell growth, cells that were removed and placed in unmodified M9 media were able to recover completely.

Together, these preliminary results emphasize the previously uninvestigated impact of environmental factors on microbial Cas9 kill switches and may be relevant when deploying other engineered organisms in environmentally complex settings such as the rhizosphere or lignocellulosic bioproduction streams.

Harnessing versatile CRAGE genome editing to build strains containing kill-switches in new strains. In order to integrate the optimized 4-HB degradation pathway into *E. coli* (as described in the progress section for Aim 1) the CRAGE-Duet system has been used to integrate landing pads which facilitates the modular assembly of biological systems. In parallel, landing pads were introduced into the *E. coli* ATCC35218 strain to introduce Cas9 into the genome. Combined with a kanamycin-resistant version of the pgRNA plasmid for expression of gRNA, this parallel kill-switch equipped strain facilitates testing model predictions using a novel genomic context. The chassis-independent recombinase-assisted genome engineering (CRAGE)-Duet system has been previously implemented in *Pseudomonas fluorescens* SBW25. The plasmid pJZ8 was constructed for the first integration site. This plasmid contains the module for a kill switch, the Cas9 gene under the control of the Tet promoter. Next, we conjugally transformed pJZ8 into *P. fluorescens* SBW25 using *E. coli* BW29427 harboring pJZ8 as a conjugal donor strain. After conjugation, the strain mixture was spread on LB agar containing apramycin as a selection marker. To screen for integration, we counter-selected with kanamycin since successful integration makes the recipient strain, *P. fluorescens*, also sensitive to kanamycin. Upon integration of sgRNA, this system functions as a kill switch.

Construction of a landing pad (LP) targeting a specific locus.We aimed to integrate the CRAGE-Duet LP into a specific locus by using a lambda red recombination system instead of transposase (which results in integration of the LP into random locations). We designed and constructed a pJZ3 plasmid containing the LP flanked by ~500 bp homology arms to target a specific locus (an intergenic region between *atpI* and *rsmG*) in *E. coli* (see **Fig. 5**). This location is closer to the *E. coli* origin of replication (OriC) and allows a higher expression level of the payloads. In pJZ3, the first integration site contains a Km resistance gene flanked by *loxP* and *lox5171*, and the second integration site comprises the Cre recombinase gene flanked by *lox2272* and *loxP*. Upon integration of the LP, we integrated the *lux* operon into the first integration site via a cassette exchange. We confirmed the utility of this LP by luminescence activity and counter-selection using Apr and Km. This construct can be used to integrate the CRAGE-Duet LP into other *E. coli* strains.



**Fig. 5**. Construction of pJZ3 for CRAGE using lambda red recombination system.

**Results from Aim 3.** Biological modules - improved systems for the rapid designing, engineering, and assaying of new biological modules.

Rationally and predictably designing cellular functions remains a complex and formidable challenge. To overcome this challenge, work in this aim is focused on developing AI-driven, multiplexed cell-free methods that enable the rapid testing of large combinations of biological modules for engineering smart and secure microbiological systems. We will apply these methods to address three key needs for secure biosystems design: biosensors, metabolic pathways, and genetic programs. Initial work has focused on biosensors, specifically the transcriptional regulator HosA for cell free prototyping of 4-hydroxybenzoate sensing and degradation.

**Specific tasks for Aim 3.**

3-1. Developing an AI-PED framework to support DBLT experiments.

3-2. Develop a closed-loop, automated cell-free system for optimizing biosensors.

3-3. Developing a closed-loop, automated cell-free system to optimize metabolic pathways.

3-4. Develop a closed-loop, automated cell-free system for optimizing genetic parts and circuits.

3-5. Building metadata and data versioning for experiments and simulations.

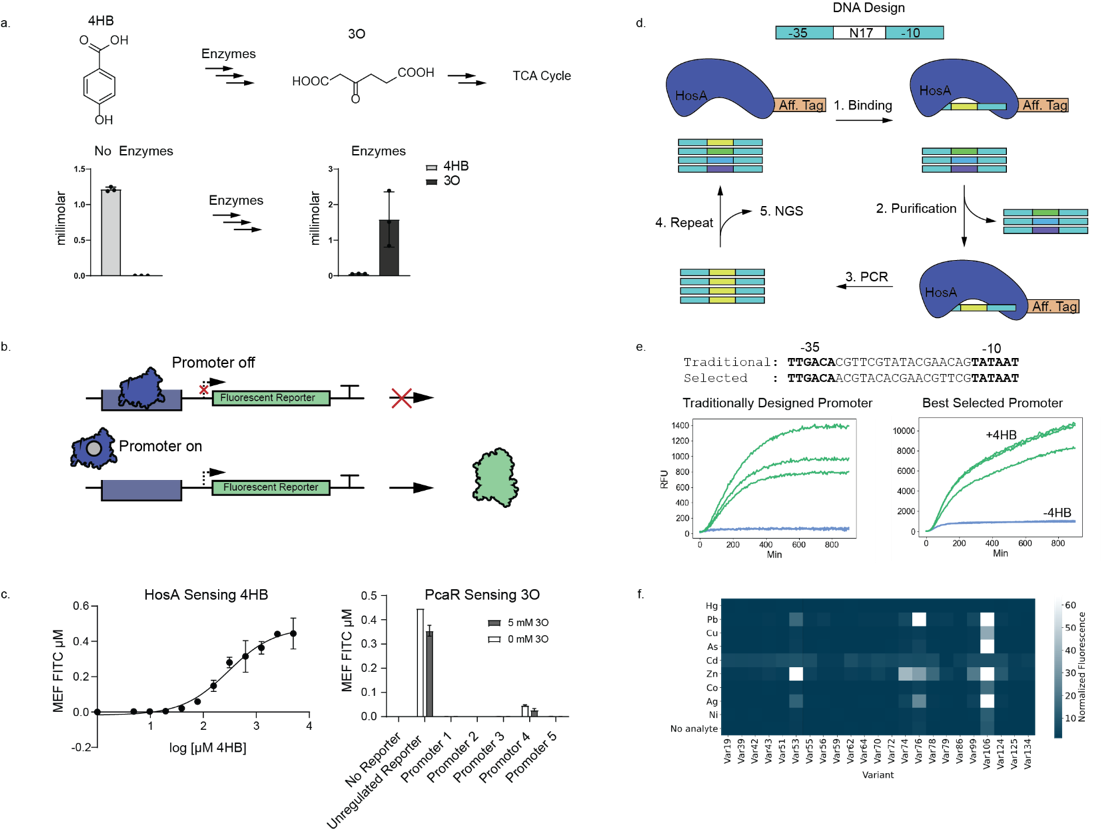
~~\*Based on the revised project scope (September 2020), Tasks 4 and 5 (regarding a strain capable of eliminating another strain) were de-emphasized, so that the focus would be on generating a strain capable of detecting and responding to the presence of another.~~

**Overview of Aim 3 activities.** ~~Work to date focused on: (1) validating substrate utilization capabilities by bacterial strains to be used in metabolic competition studies, (2) developing quantitative methods to evaluate metabolite-dependent coculturing, (3) revealing secondary metabolite cross-feeding using a parallelized chemostat platform, (4) predicting interspecies metabolic dynamics using metabolic modeling, (5) developing a probabilistic decision circuit to mitigate the metabolic burden of expensive responses, and (6) designing metabolic modeling approaches to understand system impacts of engineering and predict strain adaptations.~~

Assembly of the 4-HB degradation pathway. Pathway genes for the degradation of 4-HB by *E. coli* were identified and several orthologs were used for optimization. For reference, nine enzymes are required to breakdown 4-HB into succinyl-CoA and acetyl-CoA. Biosensors for the starting substrate in the pathway (4-HB) and for an intermediate (beta-ketoadipate) were used initially for optimization in a cell-free setting. Our plan is to then transfer the optimized pathway into *E. coli* and test its performance.

To accomplish this sub-task,we are developing a closed-loop system to optimize the breakdown of 4-hydroxybenzoic acid (4-HB) into TCA cycle intermediates using *in vitro* methods. A base case grouping of enzymes was chosen from *Pseudomonas* and contains several chemical steps (**Fig. 6A**). To rapidly improve this catabolic process, we require biosensors for both substrates and products (**Fig. 6B**). Biosensors are devices that can measure the presence or absence of biochemical metabolites, and they can be used *in vitro* and *in vivo*. We were able to demonstrate the successful employment of a 4-HB biosensor (HosA) *in vitro* and were developing PcaR, a biosensor for the stable 4-HB breakdown product 3-oxoadipate (3O). Unfortunately, traditional methods have so far not led to a successful 3O biosensor (**Fig. 6C**). To overcome this problem, we are developing two new strategies to broadly discover and engineer new allosteric transcription factor (aTF) biosensors. The first strategy is a promoter selection assay where aTF homologs are immobilized, and an *E. coli*-friendly promoter is selected from a random pool of DNA sequences using SELEX. The second strategy is the rapid and controlled mutational scanning of aTFs to understand and engineer their function.

**Fig. 6.** Developingcell free methods for engineering biosensors. (A) We can measure 4-HB degradation and 3O biosynthesis by GC/MS to show that our base case of enzymes work. (B) Our transcription factor-based biosensors synthesize the fluorescent reporter GFP when a specific small molecule (4-HB or 3O) is present. (C) We have an aTF biosensor (HosA) for 4-HB, but PcaR does not sense the presence of 3O using standard methods and promoter design. (D) We designed a promoter selection library to discover promoters that work with *E. coli* machinery. (E) The designed promoter selection assay selects novel, functional promoters for HosA. (F) By mutationally scanning the aTF MerR we have found mutational “hot spots” for aTF engineering. (Each box is normalized to WT sensing of Hg).



Because HosA is working well in our hands, we are using HosA as a model transcription factor to develop a promoter selection assay (**Fig. 6D**). To select *E. coli-*friendly promoters, we have designed an oligo library containing 17 randomized nucleotides between the -35 and -10 sites of the strong promoter J23119 from the Anderson promoter catalog. The library is incubated with immobilized HosA, and bound DNA sequences are purified. After multiple rounds of enrichment only high-affinity sequences remain and are identified by next generation sequencing. After optimizing selection conditions with a traditional promoter, we performed the selection with the randomized library and discovered novel HosA promoters that respond to 4-HB (**Fig. 6E**). This protocol was recently applied to 96 new HosA homologs, and we are currently in the process of validating that the selected promoters bind their cognate transcription factor and validating that the aTF / promoter pair is regulated by 4-HB or the related hydroxybenzoic acid. The best 4-HB sensor / promoter pair will be further used for this project. We plan to expand this assay to PcaR and related homologs to develop a 3O biosensor.

To mutationally scan aTFs we further designed a well-by-well assay to analyze single aTF variants by optimizing a high-throughput 384-well plate-based workflow that leverages liquid handling robotics to reduce time and labor. We used this assay to comprehensively analyze how an alanine mutation at every position along an aTF can alter function. We performed this screen with three metal-sensing aTFs (MerR, PbrR, and CadR) for sensing eight different metal ligands. The resulting data identified mutational “hot spots” for analyte promiscuity. See (**Fig. 6F**) for data concerning MerR that we are using to guide engineering efforts to make new aTFs with varied dynamic range, sensitivity, and changed specificity profiles. We expect to use this assay to optimize 4-HB and 3O biosensors for *in vivo* activity via iterative rounds of *in vivo* testing and *in vitro* optimization.

In parallel to developing a biosensor for 3O, we have also started analyzing the activity of our base case enzymes for 4-HB breakdown with GC/MS. Using only 1 µM of each catalyst and 1 mM of substrate, we see robust conversion of 4-HB to 3O (**Fig. 6A)**. However, PobA, the first enzyme in the pathway is inhibited by 4-HB above 1 mM. Removing such substrate inhibition will be a focus of our AI-driven optimization approach as we screen multiple homologs or by direct mutational scanning of PobA. In case developing a 3O biosensor remains elusive for *in vitro* applications, we are also exploring high-throughput MS techniques, such as SAMDI-MS, to enable AI-driven pathway optimization.

Closed loop experimental design with automated laboratory techniques. The engineering and optimization of biological processes generally requires the time-consuming construction of hundreds or even thousands of unique cell lines, each with a single genetically encoded design of a protein or biosynthetic pathway (e.g., a set of enzymes that yield products of interest). For even a small protein (say 100 amino-acid residues), modifying just 10 positions within the protein can result in nearly 2010 variants, which would be impossible to screen via traditional approaches. Directed evolution (and its variations) can be used to engineer protein(s) or other biomolecules of interest, however, multiparameter optimization eliciting specific response along diverse biological pathways can be limiting when relying solely on traditional bio-engineering approaches. We posit that these challenges can be overcome by the design of intelligent closed-loop experimental strategies.

Argonne has embarked on an ambitious path for automating biological experiments by leveraging recent advances in artificial intelligence (AI) and machine learning (ML), robotics and automation, and cell-free approaches to constructing discrete synthetic programs (i.e., transcriptional regulation, metabolic pathways) through modular assembly of cell-free lysates containing enzyme components produced by overexpression in the lysate chassis strain or by cell-free protein synthesis (CFPS). Through strategic investments, Argonne has acquired an array of production and test robotic platforms that will enable us to build, test, and improve how biological experiments are designed and executed in a high-throughput yet reproducible manner. The ability to write end-to-end biological experimental “programs” that can be executed on automated platforms provides an exciting opportunity to improve the design of experiments, while significantly reducing the combinatorial complexity (via AI/ML techniques). This strategy aligns with the recent investments within DOE Advanced Computing for Scientific Research (ASCR) and addresses key requirements within the AI for Science report for how the future of scientific experiments will evolve, especially in the context of Exascale computing and upgrades to almost all of DOE’s user facilities, including the Advanced Photon Source (APS) at Argonne.

Our production-scale automated lab environment consists of an integrated Hudson Robotics-based platform including a plate crane (robotic arm), liquid handler, plate reader, plate peeler and sealer, barcode reader, and colony picker. This environment lets us perform several liquid handling experiments, DNA transformation steps as well as protein/peptide screening experiments. We have developed an open-source Python-based library that allows remote operation of the Hudson Robotics platform, while designing several experimental protocols. This library enables us to implement and rapidly scale experiments on the production environment.

**Products Delivered from Project**

Publications

1. Dewey JA, Delalande C, Azizi S-A, Lu V,Antonopoulos D, Babnigg G. Molecular glue discovery: current and future approaches. *J Med Chem.* 2023 In review (revision).
2. Hanke PD, Parrello B, Vasieva O, Akins C, Chlenski P, Babnigg G, Henry C, Foflonker F, Brettin T, Antonopoulos D, Stevens R, Fonstein M. Engineering of increased L-threonine production in bacteria by combinatorial cloning and machine learning. *Metab Eng Comm.* 2023 Accepted for publication.
3. Freiburger AP, Dewey JA, Foflonker F, Babnigg G, Antonopoulos DA, Henry C. A microbial community growth model for dynamic phenotype predictions. *bioRxiv*2022 doi: 10.1101/2022.12.15.520667
4. Seshadri R, Roux S, Huber KJ, Wu D, Yu S, Udwary D, Call L, Nayfach S, Hahnke RL, Pukall R, White JR, Varghese NJ, Webb C, Palaniappan K, Reimer LC, Sardà J, Bertsch J, Mukherjee S, Reddy TBK, Hajek PP, Huntemann M, Chen IA, Spunde A, Clum A, Shapiro N, Wu ZY, Zhao Z, Zhou Y, Evtushenko L, Thijs S, Stevens V, Eloe-Fadrosh EA, Mouncey NJ, Yoshikuni Y, Whitman WB, Klenk HP, Woyke T, Göker M, Kyrpides NC, Ivanova NN. Expanding the genomic encyclopedia of *Actinobacteria* with 824 isolate reference genomes. *Cell Genom.* 2022 2(12):100213. doi: 10.1016/j.xgen.2022.100213. PMID: 36778052; PMCID: PMC9903846.
5. Han SW, Yoshikuni Y. Microbiome engineering for sustainable agriculture: using synthetic biology to enhance nitrogen metabolism in plant-associated microbes. *Curr Opin Microbiol.* 2022 68:102172. doi: 10.1016/j.mib.2022.102172. PMID: 35717707.
6. Sasaki Y, Yoshikuni Y. Metabolic engineering for valorization of macroalgae biomass. *Metab Eng.* 2022 71:42-61. doi: 10.1016/j.ymben.2022.01.005. PMID: 35077903.

Presentations

*The following four posters were presented at the DOE-SC/BER 2022 Genomic Science Program (GSP) Annual Principal Investigator (PI) Meeting:*

- Rapid Design and Engineering of Smart and Secure Microbiological Systems

Gyorgy Babnigg, Jeffrey Dewey, Fatima Foflonker, Michael Fonstein, Sara Forrester, Stephanie Greenwald, Christopher S. Henry, Michael Irvin, Jessica L. Johnson, Peter E. Larsen, Filipe Liu, Carla M. Mann, Sarah Owens,Arvind Ramanathan, Rebecca Weinberg, Marie-Francoise Gros, Philippe Noirot, Tomoya Honda, Zhiying Zhao, Yasuo Yoshikuni, Steven R. Fleming, Ashty S. Karim, Brenda Wang, Michael C. Jewett, Jaehyun Lee, Mark Mimee, and Dionysios A. Antonopoulos

- CRISPR-Act: AI-guided Prediction of a CRISPR Kill-switch Across Physiological Contexts

Rebecca Weinberg, Carla M. Mann, Gyorgy Babnigg, Sara Forrester, Stephanie Greenwald, Peter E. Larsen, Sarah Owens, Marie-Francoise Gros, Philippe Noirot, Arvind Ramanathan, and Dionysios A. Antonopoulos

- Guiding Data-Driven Integrative Design of Secure Biological Systems with Artificial Intelligence Techniques

Carla M. Mann, Michael Irvin, Rebecca Weinberg, Gyorgy Babnigg, Christopher Henry, Dionysios A. Antonopoulos,and Arvind Ramanathan

- Developing “Smart” Single-strain Systems Capable of Detecting and Responding to Target Organisms in the Environment

Gyorgy Babnigg, Jeffrey Dewey, Fatima Foflonker, Michael Fonstein, Christopher S. Henry, Jessica L. Johnson, Arvind Ramanathan, Tomoya Honda, Zhiying Zhao, Yasuo Yoshikuni,Steven R. Fleming, Ashty S. Karim, Brenda Wang, Michael C. Jewett, Jaehyun Lee, Mark Mimee, and Dionysios A. Antonopoulos

*The following poster was presented at the Fall 2023 American Chemical Society Conference, March, 2023:*

- Linear fitting model for chemical parameters of microbial communities

Andrew P Freiburger, Jeffrey Dewey, Fatima Foflonker, Gyorgy Babnigg, Dionysios Antonopoulos, and Christopher Henry

*The following poster was presented at the 2022 American Institute of Chemical Engineering, November, 2022:*

- CommPhitting: A fitting model for investigating the kinetics and cross-feeding of microbial communities

Andrew P Freiburger, Jeffrey Dewey, Fatima Foflonker, Gyorgy Babnigg, Dionysios Antonopoulos, and Christopher Henry

*The following poster was presented at the 2021 World Microbe Forum, June 20-24, 2021:*

- Predicting CRISPR Kill Switch Behavior Under Different Physiological Conditions Using CRISPR-Act, an AI-guided Approach

Rebecca Weinberg, Carla M. Mann, Gyorgy Babnigg, Sara Forrester, Stephanie Greenwald, Peter E. Larsen, Sarah Owens, Marie-Francoise Gros, Philippe Noirot, Arvind Ramanathan, and Dionysios A. Antonopoulos

*The following three posters were presented at the DOE-SC/BER 2021 Genomic Science Program (GSP) Annual Principal Investigator (PI) Meeting:*

- Rapid Design and Engineering of Smart and Secure Microbiological Systems

Gyorgy Babnigg, Fatima Foflonker, Michael Fonstein, Sara Forrester, Stephanie Greenwald, Christopher S. Henry, Peter E. Larsen, Filipe Liu, Carla M. Mann Sarah Owens, Arvind Ramanathan, Rebecca Weinberg, Marie-Francoise Gros, Philippe Noirot, Tomoya Honda, Yasuo Yoshikuni, Steven R. Fleming, Ashty S. Karim, Brenda Wang, Michael C. Jewett, Mark Mimee, and Dionysios A. Antonopoulos

- CRISPR-Act: AI-guided Prediction of a CRISPR Kill Switch Under Diverse Physiological Conditions

Rebecca Weinberg, Carla M. Mann, Gyorgy Babnigg, Sara Forrester, Stephanie Greenwald, Peter E. Larsen, Sarah Owens, Marie-Francoise Gros, Philippe Noirot, Arvind Ramanathan, and Dionysios A. Antonopoulos

- Grammar and Language of CRISPR/Cas-Targetable Sites in *Escherichia coli*, *Shigella*, *Pseudomonas*, and *Salmonella*: A Comprehensive Survey

Carla M. Mann, Rebecca Weinberg, Gyorgy Babnigg, Peter E. Larsen, Marie-Francoise Gros, Philippe Noirot, Dionysios A. Antonopoulos, and Arvind Ramanathan

**Conclusions**

[PLACEHOLDER TEXT]

**List of Acronyms and Abbreviations**

ACC 1-aminocyclopropane-1-carboxylate

AI artificial intelligence

AI-PED AI-driven strategy for probabilistic experimental design

AMP antimicrobial proteins

ANOVA analysis of variance

AT adenine and thymine

aTc anhydrotetracycline

aTF allosteric transcription factor

attB bacterial attachment site

attP phage attachment site

CFPS cell free protein synthesis

COMETS Computation Of Microbial Ecosystems in Time and Space

CRAGE Chassis-independent Recombinase-assisted Genome Engineering

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

CRISPR-ActML CRISPR-Activity Machine Learning algorithm

crRNA CRISPR ribonucleic acid

DBTL Design Build Test Learn cycle

DSB double stranded break

FACS Fluorescence Activated Cell Sorting

FBA Flux Balance Analysis

GC guanine and cytosine

GC/MS Gas Chromatography / Mass Spectrometry

GFP Green Fluorescent Protein

gRNA guide RNA

4-HB 4-hydroxybenzoate

HPC high performance computing

HPLC High-Performance Liquid Chromatography

HTP high-throughput

LB Lysogeny Broth

LB-E LB media, exponential phase

LB-S LB media, stationary phase

LC-MS Liquid Chromatography-Mass Spectrometry

LP Landing Pad

M9-E M9 media, exponential phase

MANOVA multiple analysis of variance

ML machine learning

MOMA Minimization of Metabolic Adjustment

MS mass spectrometry

NAP nucleoid associated protein

NLP natural language processing

NN neural network

3O 3-oxoadipate

OriC origin of replication

PGPR plant growth promoting rhizobacteria

PAM Protospacer Adjacent Motif

qPCR quantitative polymerase chain reaction

RMI-Chip root microbe interaction chip

RNA-seq ribonucleic acid sequencing

SAMDI-MS Self-Assembled Monolayers for Matrix-Assisted Desorption/Ionization

SELEX Systematic Evolution of Ligands by Exponential Enrichment

sfGFP superfolder GFP

TATB triaminotrinitrobenzene

TCA tricarboxylic acid

**Distribution List**

[PLACEHOLDER TEXT]