

# IWBDA 2026

June 18-20th Boulder, Colorado

**18<sup>th</sup> Annual International Workshop on Bio-Design Automation**

**Boulder, Colorado, USA**

**June 18<sup>th</sup>-20<sup>th</sup> 2026**

## Foreword

We aim to bring together academic researchers and industry partners to push the field of biodesign automation for synthetic biology forward. The field of biology is still lagging behind the industry when it comes to implementation and adoption of computational and digital solutions. This is primarily because it is extremely difficult to offer a one-size-fits-all solution to engineer biology. Some BDA solutions serve as templates for a larger problem; however, to make them usable in a practical setting, the solutions often have to be redesigned for bespoke biological workflows. Without understanding the requirements of the broader synthetic biology community, it is hard to design a solution that can have a wider impact. Increasing the efficiency of biological processes and workflows with biodesign automation is a continuous journey that requires fresh exchange of ideas. It is crucial for the BDA community to have a forum where they can collaborate, gather new perspectives, and understand the requirements of the broader synthetic biology community. IWBD provides this forum for cross-disciplinary discussion, with the aim of seeding and fostering collaboration between the researchers from the synthetic biology, systems biology, and design automation communities.

## **Organizing Committee**

Chris Myers: General Chair

Aaron Adler: Finance Chair

Carolus Vitalis: Program Chair

William Mo: Program Chair

Daniel Fang: Tutorials Chair

Gonzalo Vidal: Tutorials Chair

Hatem Abdelrahman: Tutorials Chair

Kimberly Low: Arrangements Chair

Chunxiao Liao: Publicity Chair

## **Program Committee**

Lukas Bücherl (Utah State University)

David J. Ross (NIST)

Aaron Adler (BBN)

Martín Gutiérrez (Universidad Rey Juan Carlos)

Caleb Bashor (Rice University)

Zhen Zhang (Utah State University)

Harrison Steel (University of Oxford)

Daisuke Kiga (Waseda University)

Jake Beal (BBN)

Tae Seok Moon (J. Craig Venter Institute)

Eric Young Worcester (Worcester Polytechnic Institute)

Prashant Vaidyanathan (Oxford Biomedica)

Brian Munsy (Colorado State University)

Jeanet Mante (Army Research Lab)

Nicholas Roehner (BBN)

Pete Carr (BBN)

Traci Haddock (Asimov)

Christian Hochberger (TU Darmstadt)

Joshua Chan (Colorado State University)

# Agenda

## Tutorials: June 13<sup>th</sup>-14<sup>th</sup>

### Saturday, June 13, 2026 - Location: KOBL 352

**08:30-10:00** Chris Myers, Introduction to the SynBioSuite Data Management System

10:00-10:30 Break

#### **10:30-12:00** Resources

10:30-11:30 Carolus Vitalis, Encoding Resources with Excel2Sbol

11:30-12:00 Daniel Fang, Introduction to the SynBioHub Repository

12:00-13:30 Lunch Break (Not catered)

#### **13:30-15:00** Design/Modeling

13:30-15:00 Hatem Abdelrahman, Design and Modeling of Genetic Circuits with SBOLCanvas

15:00-15:30 Break

#### **15:30-17:00** Build/Test

15:30-16:00 Chunxiao Liao, Plasmid Annotation with SeqImprove

16:00-17:00 Gonzalo Vidal, Capturing Experimental Data and Metadata with XDC

### Sunday, June 14, 2026 - Location: KOBL 352

#### **08:30-10:00** Repositories

08:30-09:30 Carolus Vitalis, Analysis of Plate Reader Data Using Flapjack

09:30-10:00 Daniel Fang, SynBioHub Installation

10:00-10:30 Break

#### **10:30-12:00** SynBioHub

10:30-12:00 Daniel Fang, Usability Evaluation of SynBioHub for Synthetic Biology Data Management

12:00-12:30 Closing Remarks (Chris Myers)

## Workshop: June 18<sup>th</sup>-20<sup>th</sup>

**Thursday, June 18, 2026 - Location: KOBL 352**

15:30-16:00 Welcome

**16:00-17:30 Panel Session (Douglas Densmore): Opportunities and Challenges for Bio-Design Automation**

The panel opens with a retrospective on the two-decade trajectory of bio-design automation, identifying which bottlenecks we've cleared and which remain stubbornly in place. We will pivot quickly to the future: examining how AI and emerging technologies can break current deadlocks. Our discussion will focus on the cross-disciplinary talent and large-scale community initiatives required to accelerate progress while rigorously ensuring biosecurity in an increasingly automated landscape.

**17:30-17:45 Allan Kuchinsky Scholarship Presentation**

17:45 Reception

**Friday, June 19, 2026 - Location: ECCS 201**

08:30-09:00 Breakfast

**09:00-10:00 Invited Speaker: Domitilla Del Vecchio - Modular Machine Learning for Genetic Circuit Engineering**

10:00-10:30 Coffee Break

**10:30-11:45 Session 1: Design Rules for Programmable Genetic Systems (Chairs: Lukas Buecherl / Daniel Fang)**

10:30-10:45 Masayuki Yamamura, Rule based generation of synthetic genetic circuits - towards v 3.0

10:45-11:00 Ayush Pandey, Formal specifications for the compositional design of synthetic biological circuits

11:00-11:15 Zoila Jurado, Geoffrey Taghon and Samuel W. Schaffter, Towards interoperable modeling of toehold mediated strand exchange circuits across DNA nanotechnology and engineering biology

11:15-11:30 Zachary Hastings, Zoila Jurado, Bingqing Hu, Zhen Zhang, Samuel W. Schaffter and Lukas Buecherl, Host Aware Design of Cellular ctRSD Circuits

11:30-11:45 Stephen Chiu and Tae Seok Moon, Dynamic control of metabolic flux by RNA based genetic circuits designed, built, and tested through automation

11:45-13:00 Lunch

**13:00-14:00 Session 2: AI, Curation, and Closed Loop Biological Discovery (Chairs: Zhen Zhang / Chunxiao Liao)**

13:00-13:15 Niloofar Arazkhani and Natasa Miskov Zivanov, BELL: Biomodel Evidence and LLM based Logic

13:15-13:30 Haomiao Luo, Niloofar Arazkhani, Difei Tang and Natasa Miskov Zivanov, VIOLIN WEB: a web application for reconciliation and curation of molecular interaction graphs

13:30-13:45 Robin Henry, Harrison Steel and Jean Baptiste Lugagne, Myriad: A GPU Accelerated Platform for In Silico Prototyping of Closed Loop Single Cell Experiments

13:45-14:00 Difei Tang and Natasa Miskov Zivanov, CELESTA Web: An Interactive Tool for LLM Enhanced Biomedical Context Annotation

14:00-14:30 Coffee Break

14:30-16:00 Nona Talks (Douglas Densmore / Chris Krenz)

16:00-17:00 DevCell Discussion (Richard Murray)

**Saturday, June 20, 2026 - Location: ECCS 201**

08:30-09:00 Breakfast

**09:00-10:00 Invited Speaker: Emma Chory - Navigating the Evolution Multiverse: Real-time evolution and open-source robotic platforms for engineering biology**

10:00-10:30 Coffee Break

**10:30-12:00 Session 3: Predictive Models of Emergent Biochemical Dynamics (Chairs: Harrison Steel / Hatem Abdelrahman)**

10:30-10:45 Colin Yancey, EmergeX: A computational framework for designing emergent dynamics in chemical reaction networks

10:45-11:00 Quang Luan Dang Tran, Nguyen H.N. Tran, Derrick Osei and Ania Ariadna Baetica, Performance of Metaheuristic Algorithms in Finding Tradeoffs in One and Two Species Biological Feedback

11:00-11:15 Nguyen Tran, Ania Ariadna Baetica and Samuel Schaffter, Predictive Analysis of Amplification in Biochemical Feedback Circuits

11:15-11:30 Payton Thomas, Johanna Cao, Caleb Wong and Jeff Hasty, Engineered Critical Point Dynamics in Genetic Circuits

11:30-11:45 Zachary Hastings and Lukas Buecherl, Physics Informed Inference of Gene Expression Kinetics Using bioCRNpyler

11:45-12:00 Dimitris Papamichail, Bang Chiem and Georgios Papamichail, Codon Pair Optimization Web Tool

12:00-13:00 Lunch

**13:00-14:15 Session 4: Measurement, Microphysiology, and Living Experimental Platforms (Chairs: Jeanet Mante / Gonzalo Vidal)**

13:00-13:15 Vicente Trelles Fernandez and Harrison Steel, Novel Microfluidic Chips for Massively Parallelised Biological Data Acquisition

13:15-13:30 Nona Hashemi, Mahdi Hasanzadeh Hesar and Samuel M.D. Oliveira, Microfluidic Physics Constrained Agent Based Modeling of Artificial Microbial Biofilms

13:30-13:45 Harman Mehta, Yusuf Ugurluoglu, Meng Zhang, Martyn Dade Robertson and Harrison Steel, Microbial 3D Bioprinter for Novel Living Materials

13:45-14:00 Niall McIntyre and Nick Brooks, Manufacturing a high throughput, label free platform for rapid screening of drug permeation across in vitro plasma membrane models

14:00-14:15 Evan Holbrook, Elizabeth Codd, Ron Weiss and David Kong, TransfectionWizard: An Integrated Open Source Platform for Genetic Circuit Design and Automated Mammalian Cell Liposome Transfection for the Lab and Classroom

14:15-14:45 Coffee Break

**14:45-15:45 Session 5: From Standards to Buildable Biology (Chairs: Tae Seok Moon / Brian Bartley)**

14:45-15:00 Doug Densmore, Chris Krenz, and Guzman Vigliecca, Biodesign Metadata Exchange for Use in Biosecurity

15:00-15:15 Gonzalo Andrés Vidal Peña et al., Standardized Synthetic Biology Data Curation Effort

15:15-15:30 Travis Uhrig and Chris Myers, Adding SBML Modeling Support to SBOLCanvas

15:30-15:45 Jeanet Mante, From Design to Build: Sequence Domestication, Robotic Assembly, and Integrated SBOL Data Capture

15:45-16:00 Closing Remarks (Including Nona Awards)

# Rule-based generation of synthetic genetic circuits - towards v 3.0 -

**Masayuki Yamamura\*\***  
Higashi Nippon International  
University, Fukushima, Japan  
myamamura@m.tonichi-kokusai-u.ac.jp

**Masahiro Takinoue**  
**Ryoji Sekine**  
Institute of Science Tokyo  
Tokyo, Japan  
takinoue@cls.iir.isct.ac.jp  
ryoji\_722@msn.com

**Kazuteru Miyazaki**  
Bunkyo University  
Kanagawa, Japan  
k.miyazaki@bunkyo.ac.jp

**Naoki Kodama**  
Tokyo University of Science  
Tokyo, Japan  
nkodama@rs.tus.ac.jp

**Daisuke Kiga**  
kiga@waseda.jp  
Waseda University  
Tokyo, Japan

## 1 INTRODUCTION

In this abstract, we propose the final form of our genetic circuit design system. To design cellular behavior based on the synthetic biology approach, the appropriate selection of network motifs, as well as the parts in a motif, is essential. In contrast to manual design, design based on logic programming allows the extensive generation of combinations of parts[1]. Model parameterization and numerical calculation are also required to evaluate the cellular behavior specifications. An automation tool designed for Boolean networks assigns biological parts to nodes within these networks and determines those parts' dynamic behavior to verify the designs' feasibility[2]. Tools calculating other types of networks are also developed[3][4], and databases for the models or simulation results are provided[5][6].

These five years, we have developed an automation tool for genetic circuit design by using Prolog inference engine. Table1 shows a brief summary of the improvement history of our system. Our IWBD 2022 abstract referred to our previous paper, which describes a synthetic genetic circuit for reprogramming and diversifying the gene-expression status of living cells[7]. In the manual design procedure in the paper, we initially combined the toggle switch and gene overexpression motifs in a cell because a manual phase-space analysis of the toggle switch with and without the overexpression shows bifurcation between bistability and monostability, and both are required for the reprogramming and diversification process. For design automation of the circuit using an inference engine, we developed a combination of Inside Prolog and C++ codes available in Zenodo[8].

This v1.0 implements full function into one software system. Reading the configuration files for genetic circuit design

rules and the required specification, our program will enumerate candidates of a network design with a combination of reaction parameters. It can produce an explanation of the inference process by a proof tree. The numerical simulation module equipped a cache mechanism for simulation results. We found Prolog inference engine has enough flexibility to connect with external numerical simulations, and also has enough scalability to extend a program by just adding new inference rules.

Our other modification of the toggle switch in the IWBD 2023 code was the addition of intercellular communication, which governs autonomous cell-type diversification on an epigenetic landscape showing bifurcation from monostability to bistability. The diversification of cell states on the epigenetic landscape was the specification for an automated design of the cell-population behavior regulated by cell-cell communication. This diversification has been achieved by using our synthetic circuit in living cells[9]. In the inference-engine-based design flows for the communication-dependent cell-type diversification, parameterized models were modified to incorporate cell-cell communication, and the modified model was examined for the specification of cell state movements on the epigenetic landscape.

This v1.1 achieves a parallel execution of inference rules for multi cellular system with cell-cell communication. Based on a server/client model, we added explicit parallel inference primitives in the inference rule. Inference rules can communicate with each other by inter-thread communications. We extended the cache function for numerical simulation results and the pretty print function for proof trees.

In IWBD 2024, we designed a multi-layer system consisting of a receiver cell and two sender cells, each of which can produce its unique cell-cell communication molecule. The inner states of the two sender cells are inputs for a logic gate in the receiver cell. After an incubation of each sender cell, each supernatant containing the communication molecule

\*All authors were supported by JST, CREST Grant Number JPMJCR21N4, Japan

is transferred to the other liquid culture of the receiver cell. The two inner states of the sender cells thus become an input pattern for the logic gate. With an adequate parameter set, only one of the four input patterns produces GFP efficiently.

This v1.2 achieved a complete form of parallelism. We extended parallel design rules by adding request forms to other processes. In order to optimize parallel calculation in given computational environment, users can specify the group of parallel computation for each parallel execution, and also users can specify a certain server to be assigned.

After IWBD 2024, we achieved major version up to v2.0, which realizes a generalization on motif representation. We assumed a simple typical formula format for a specific chemical reaction for v1s. We extended to allow reaction formula which includes multiple productive terms in following three types: (1) multiplicative, (2) additive and (3) enzymatic form.

$$\alpha \cdot \frac{\left(\frac{x_1}{K_1}\right)^{n_1}}{1 + \left(\frac{x_1}{K_1}\right)^{n_1}} \cdot \frac{\left(\frac{x_2}{K_2}\right)^{n_2}}{1 + \left(\frac{x_2}{K_2}\right)^{n_2}} \cdots \quad (1)$$

$$\alpha \cdot \frac{\left(\frac{x_1}{K_1}\right)^{n_1} + \left(\frac{x_2}{K_2}\right)^{n_2} + \cdots}{1 + \left(\frac{x_1}{K_1}\right)^{n_1} + \left(\frac{x_2}{K_2}\right)^{n_2} + \cdots} \quad (2)$$

$$\alpha \cdot \left(\frac{x_1}{K_1}\right)^{n_1} \cdot \left(\frac{x_2}{K_2}\right)^{n_2} \cdots \quad (3)$$

We restricted the maximum number of genes to 20, and the maximum number of productive terms to 10. We changed the interface specification for numerical simulation programs according to the general representation. We also prepared several primitive functions to improve the computational performance for a numerical simulation. For example, users can set the grid size for the initial value to search steady states, and can select growth models like logistic curves other than exponential ones assuming low cell density.

For v2.0, we have tested our system by implementing the mutual repression toggle switch with communication molecules, which is the test bed for v1.3, by new representation. Then we examined new design examples for logic gate systems whose description requires this update of the design tool. Figure 1 shows (A) the circuit design, (B) the truth table and (C) a schematic diagram for a typical time series expected in numerical simulations.

## 2 RESULTS AND CONCLUSION

Towards v3.0, we have two plans to enhance utility and performance of our system.

Firstly, we are developing a new local LLM specialized for our use. We found some of publicly available generative AIs can build programs even for Prolog with complicated compound terms like our rulebase. We also tried to extract mathematical conditions for the functionality of a certain

genetic circuit from related technical document data in Synthetic Biology field. Surprisingly, our LLM can translate such mathematical conditions into Prolog programs for some of well known genetic circuits. We now investigate how we can validate such an automatically generated rules effectively.

Secondly, we are developing APIs for Prolog logic engine to improve total search performance. Until v2.0, we implemented simple exhaustive grid search for genetic parameters such as enzymatic reaction coefficients that can be modified by protein engineering technology. After an exhaustive search, you can pick up the best combination of the genetic parameters. Actually, we don't need exhaustive search for normal use. We can improve the total search performance by examining the most plausible candidates first. Please remark that you can expect similar results when you combine similar components. We now accumulate simulation results to analyze them by simple Machine Learning technology.

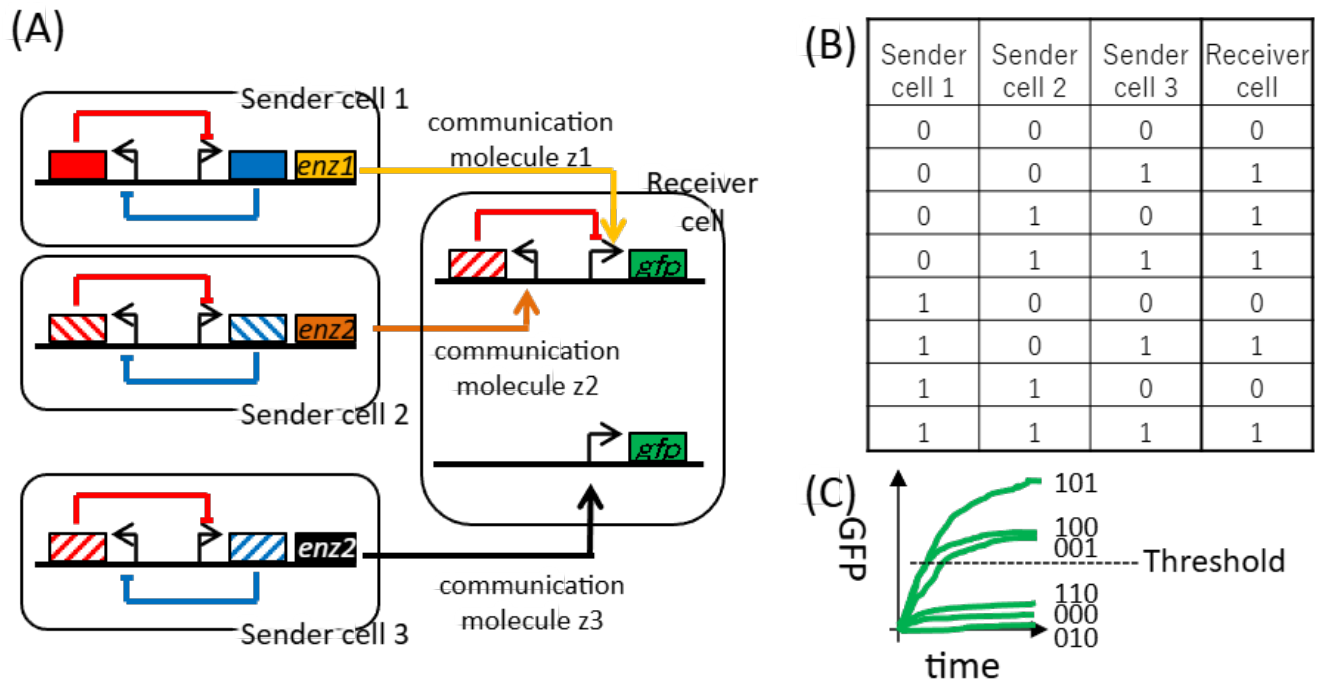
This project was motivated by a criticism upon black-box AI approaches with high performance but no explanation such as Deep Learning for genetic circuit design in Synthetic Biology. We have achieved a white-box approach to satisfy given specifications with logical inference and numerical simulation. We expect that the combination of logical inference and LLMs become a breakthrough for the design issues in Synthetic Biology.

## REFERENCES

- [1] Pederson, M. and Phillips, A. Towards programming languages for genetic engineering of living cells. *J. R. Soc. Interface* 2009, 6S437–S450.
- [2] Jones, T. S.; Oliveira, S. M. D.; Myers, C. J.; Voigt, C. A.; Densmore, D. Genetic circuit design automation with Cello 2.0. *Nat. Protoc.* 2022, 17, 1097–1113.
- [3] Tas, H.; Grozinger, L.; Goni-Moreno, A.; de Lorenzo, V. Automated design and implementation of a NOR gate in *Pseudomonas putida*. *Synth. Biol. (Oxf)* 2021, 6, ysab024.
- [4] Boada, Y.; Reynoso-Meza, G.; Vignoni, A. Multi-objective optimization framework to obtain model-based guidelines for tuning biological synthetic devices: an adaptive network case. *BMC Syst. Biol.* 2016, 10, 27–0.
- [5] McLaughlin, J. A.; Myers, C. J.; Zundel, Z.; Misirli, G.; Zhang, M.; Ofiteru, I. D.; Goni-Moreno, A.; Wipat, A. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synth. Biol.* 2018, 7, 682–688.
- [6] Yanez Feliu, G.; Earle Gomez, B.; Codoceo Berrocal, V.; Munoz Silva, M.; Nunez, I. N.; Matute, T. F.; Arce Medina, A.; Vidal, G.; Vitalis, C.; Dahlin, J.; Federici, F.; Rudge, T. J. Flapjack: Data Management and Analysis for Genetic Circuit Characterization. *ACS Synth. Biol.* 2021, 10, 183–191.
- [7] Ishimatsu, K.; Hata, T.; Mochizuki, A.; Sekine, R.; Yamamura, M.; Kiga, D. General applicability of synthetic gene-overexpression for cell-type ratio control via reprogramming. *ACS Synth. Biol.* 2014, 3, 638–644.
- [8] <https://doi.org/10.5281/zenodo.8148662>.
- [9] Sekine, R.; Yamamura, M.; Ayukawa, S.; Ishimatsu, K.; Akama, S.; Takinoue, M.; Hagiya, M.; Kiga, D. Tunable synthetic phenotypic diversification on Waddington's landscape through autonomous signaling. *Proc. Natl. Acad. Sci. U. S. A.* 2011, 108, 17969–17973.

**Table 1: A summary of the system improvement**

ver	date	progress	simulator	example circ
1.0	2023.9	total system starts - network -> param set - proof tree pretty pr - cache sim results		mutual repression toggle switch
1.1	2023.12	parallelize (thread comm) - server/client model - parallel pretty print - parallel caching	precision improvement - convergence test - equilibria search	toggle switch (parallelized)
1.2	2024.6	parallelize (process comm) - divide process - assign server		10 elements logic circuit
1.3	2024.12	remote up/down-load - via web	add functions - SingleCellTimeCourse - configure initial values	toggle switch (with comm chem)
2.0	2025.6	expand representation - new interface to simulator	select 4 formula types - multiplicative - additive - enzymatic	toggle switch (rewrite) 10 elements logic circuit (rewrite)
3.0	2027.3	- Automatic rulebase programming by LLM from technical papers - Bestfirst genetic parameter search by Machine Learning from simulation data		multilayer logic gate (alternative) genetic operation amplifier



**Figure 1: Genetic circuits designed by the inference engine based on Prolog**

- (A) A multilayer genetic system composed of three sender cells and one receiver cell implementing a logic gate. Each sender cell is capable of synthesizing its distinct cell-cell communication molecule. The internal states of the sender cells serve as inputs to the logic gate embedded in the receiver cell. After individual incubation of the sender cells in liquid culture for a defined period, their supernatants—potentially containing communication molecules—are transferred into the culture medium of the receiver cell.
- (B) The truth table and the ideal time course of GFP expression in the receiver cell. The implemented logic gate follows the expression  $(z1 \text{ AND } (\text{NOT } z2)) \text{ OR } z3$ . Communication molecule  $z1$  activates GFP expression, whereas  $z2$  induces a repressor that inhibits GFP production from a  $z1$ -regulated hybrid promoter. Molecule  $z3$  independently drives GFP expression from a separate promoter. Under appropriately tuned parameters, the receiver cell produces GFP levels that exceed the threshold required for logic evaluation.

# Formal specifications for the compositional design of synthetic biological circuits

Ayush Pandey

University of California, Merced  
ayushpandey@ucmerced.edu

## 1 ABSTRACT

Design of synthetic biological circuits often relies on model simulations to verify whether desired properties are met. These models typically encode guarantees on state variables, leaving assumptions about environment and nominal conditions implicit. As a result, existing approaches are limited to small-scale designs or prediction of logical behaviors. In this paper, we consider open chemical reaction network (CRN) models and show how monomial dominance relations in these CRNs induce polyhedral regime assumptions in logarithmic coordinates, which yields useful polyhedral guarantees on subsystem behavior. Further, we show how order-of-magnitude dominance assumptions can be written as polyhedral constraints to obtain assume-guarantee specifications for biocircuits. Therefore, these results provide a systematic path to derive formal specifications from CRNs.

## 2 RELATED WORK

Current synthetic biological circuit design is partially heuristic driven. Often, ODE models are simulated to validate proposed hypotheses and support predictions about fixed points, oscillations, resource competition, and inducer levels needed for desired behavior [5]. However, this does not scale well with the number of components. On the other hand, from a design automation perspective [1], significant progress has been made in designing larger systems that achieve desired logical behavior. For example, Cello [4] uses high-level descriptions of logic behavior and compiles biocircuit implementations using libraries of characterized parts. Despite this progress, achieving desired dynamic properties at scale remain limited.

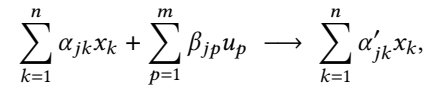
Context-dependence of each component and modularity are major challenges that are still difficult to address with model-by-model reasoning. Reduced models are usually derived under assumptions such as excess of a species, saturation, or time-scale separation [2]. These assumptions often remain implicit in the model. As a result, a model may correctly capture a subsystem in one operating regime, but it is difficult to use compositionally because the conditions under which the model is valid are not written explicitly. So, it is difficult to determine whether a composed design of multiple subsystems will work as expected. This is the central problem that we address in this paper.

## Main contributions

We present a method for deriving formal specifications for synthetic biological circuits from CRN descriptions. Specifically, we prove that mass-action CRNs can be translated into formal assume-guarantee contract specifications using reaction order polytopes. This enables compositional reasoning in biocircuit designs. These results provide a bridge between two previously disparate approaches: polytope-based methods for biomolecular analysis and stability [7], and assume-guarantee frameworks for formal specification [3].

## 3 PROBLEM FORMULATION

Consider a mass-action CRN



with  $\gamma_j = \alpha'_j - \alpha_j$ , and reaction rate for  $k_j > 0$   $r_j(x, u) = k_j x^{\alpha_j} u^{\beta_j}$ . In production-degradation form, for each  $x_k$

$$\dot{x}_k = f_k^{(p)}(x, u) - f_k^{(d)}(x, u), \quad k = 1, \dots, n,$$

$$f_k^{(p)}(x, u) = \sum_{j \in J_k^{(p)}} \gamma_{jk} k_j x^{\alpha_j} u^{\beta_j}, \quad f_k^{(d)}(x, u) = \sum_{j \in J_k^{(d)}} (-\gamma_{jk}) k_j x^{\alpha_j} u^{\beta_j}.$$

where  $J_k^{(p)} := \{j : \gamma_{jk} > 0\}$ ,  $J_k^{(d)} := \{j : \gamma_{jk} < 0\}$ . Additionally, for brevity, we introduce the notation  $c_{kj} = \gamma_{jk} k_j$ . Our goal is to obtain a higher-level abstraction of the CRN by defining the specification model. For this, we restrict our attention to a class of formal specifications called polyhedral assume-guarantee contracts, which are widely used in formal methods and design automation. An assume-guarantee contract  $C = (a, g)$ , where  $a, g$  are the set of polyhedral constraints describing the assumptions and guarantees, respectively. Therefore, for  $\xi$ , the variables retained at the abstraction level, a polyhedral assume-guarantee contract for a subsystem  $\Sigma$  is a pair  $C = (a, g)$ , where the assumption set is  $a := \{\xi : A_a \xi \leq b_a\}$ , and the guarantee set is  $g := \{\xi : A_g \xi \leq b_g\}$ , for matrices  $A_a, A_g$  and vectors  $b_a, b_g$ .

## Preliminaries

LEMMA 1 (ADAPTED FROM [7]). *For two monomials  $m_j$  and  $m_{j'}$  in the CRN, if we have the dominance relation  $m_j(x, u) \geq m_{j'}(x, u)$ , then it is equivalent to an affine inequality in the*

log variables  $z_x := \log x$  and  $z_u := \log u$ ,

$$\sum_{\ell=1}^n (\alpha_{j\ell} - \alpha_{j'\ell}) z_{x,\ell} + \sum_{p=1}^m (\beta_{jp} - \beta_{j'p}) z_{u,p} \geq \log c_{j'} - \log c_j,$$

Using this, for a positive sum of monomials,  $q(x, u) = \sum_{j \in I} c_j x^{\alpha_j} u^{\beta_j}$ , we can write convex polytopes, that are required by polyhedral contracts.

**THEOREM 1 (FROM [7]).** *For the deterministic mass-action CRN, the log-derivative vector of the positive sum of monomials,  $q(x, u)$  satisfies*

$$H_q(x, u) := \left( \frac{\partial \log q}{\partial \log x}, \frac{\partial \log q}{\partial \log u} \right) = \sum_{j \in I} \lambda_j(x, u) (\alpha_j, \beta_j), \quad (1)$$

$$\lambda_j(x, u) := \frac{c_j x^{\alpha_j} u^{\beta_j}}{\sum_{i \in I} c_i x^{\alpha_i} u^{\beta_i}} \geq 0, \quad \sum_{j \in I} \lambda_j(x, u) = 1. \quad (2)$$

Thus,  $H_q(x, u)$  is a convex combination of the exponent vectors.

## 4 RESULTS

We present a summary of our main results that describe how to derive polyhedral contracts from CRNs. For a deterministic mass-action CRN model of a subsystem  $\Sigma$ , there exists a polyhedral assume-guarantee contract with assumptions given by the by the affine inequalities in Lemma 1 and guarantees given by the convex hulls in Theorem 1. It is common to use assumptions such as excess of a species or saturation in reduced models. Our result makes these explicit by using monomial dominance margin relation, modeled using  $\eta$ .

**THEOREM 2.** *Suppose that for a regime  $\mathcal{R}$  of a deterministic mass-action CRN there is a dominant subset  $D_k \subseteq J_k^{(p)}$  s.t.*

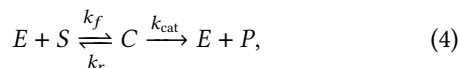
$$c_{kj} x^{\alpha_j} u^{\beta_j} \geq \eta c_{k\ell} x^{\alpha_\ell} u^{\beta_\ell}, \quad \eta > 1, \quad (3)$$

for all  $j \in D_k$  and  $\ell \in J_k^{(p)} \setminus D_k$ . Then there exists a polyhedral assume-guarantee contract  $\mathcal{C}_{\mathcal{R}, \eta} = (a_{\mathcal{R}, \eta}, g_{\mathcal{R}, \eta})$ , where the assumption set is obtained by conjoining the logarithmic form of the dominance inequalities to the regime inequalities. So, in conjunction to the inequality in Lemma 1, we also have  $(\alpha_j - \alpha_\ell)^\top z_x + (\beta_j - \beta_\ell)^\top z_u \geq \log \eta + \log c_{k\ell} - \log c_{kj}$ . Similarly, for the guarantees we have a nondominant-weight bound, in conjunction to the convex hull constraints in Theorem 1,  $\sum \lambda_{k\ell} \leq \rho_k(\eta)$ , with constants

$$\rho_k(\eta) := \frac{n_k}{\eta d_k + n_k}, \quad d_k := |D_k|, \quad n_k := |J_k^{(p)} \setminus D_k|.$$

## 5 APPLICATIONS: BIOCIRCUIT DESIGN

For an enzymatic reaction system, we show how to use the CRN description of the system to derive assume-guarantee contracts as a set of formal specifications. The CRN is



where  $E$  is an enzyme,  $S$  is the substrate,  $C$  is the enzyme-substrate complex, and  $P$  is the product (see Figure 1a). We define total resources as  $E_{\text{tot}} := E + C$ ,  $S_{\text{tot}} := S + C$ . We define  $x := S_{\text{tot}}$ ,  $u := E_{\text{tot}}$ ,  $y := P$ . From [7], the log-derivative vector of  $C$  with respect to  $(x, u)$  is

$$H(x, u) = \frac{K}{K + E + S} (1, 1) + \frac{E}{K + E + S} (1, 0) + \frac{S}{K + E + S} (0, 1),$$

so  $H(x, u) \in \text{conv}\{(1, 1), (1, 0), (0, 1)\}$  (see Figure 1b): The result gives a polyhedral description of the admissible log-derivative behavior. However, we notice that the output relation between  $(z_x, z_u)$  and  $z_y$  is still nonlinear. Here, we use our result from Theorem 2. We introduce  $z_x := \log x$ ,  $z_u := \log u$ ,  $z_y := \log y$ , and we treat  $\kappa := \log K$ ,  $\kappa_{\text{cat}} := \log k_{\text{cat}}$  as fixed constants for a chosen parameter instance. For the  $S$ -dominant contract, we have the assumptions  $A_S := (z_x - z_u \geq \delta_{xu}) \wedge (z_x - \kappa \geq \delta_{xK})$ , where  $\delta_{xu} > 0$  and  $\delta_{xK} > 0$  are fixed margins that encode the order-of-magnitude differences.  $G_S := (z_u + \kappa_{\text{cat}} - \eta_S \leq z_y) \wedge (z_y \leq z_u + \kappa_{\text{cat}})$   $\eta_S := -\log((1 - e^{-\delta_{xu}})/(1 + e^{-\delta_{xK}}))$ . Similarly, we have the enzyme-saturated contract and another for the special condition that leads to the Michaelis-Menten kinetic model (see Figure 1d). We fix the constants  $K$  and  $k_{\text{cat}}$  and create each contract in the log variables  $(z_x, z_u, z_y)$ . We check contract satisfaction and numerically verify whether the model simulations lie within the guarantees when the assumption conditions are met (and not met).

**Compositional design:** We apply this approach to reason about a compositional design problem (see Figure 2). The full composed contract is  $\mathcal{C}_{\text{full}} = \mathcal{C}_1 \parallel \mathcal{C}_2 \parallel \mathcal{C}_3$ . Using Pacti [3], we obtain which designs are compatible and which meet the top-level specification (see Table 1 and Figure 2). The code is available at [6].

## REFERENCES

- [1] Evan Appleton, Curtis Madsen, Nicholas Roehner, and Douglas Densmore. 2017. Design automation in synthetic biology. *Cold Spring Harbor Perspectives in Biology* 9, 4 (2017), a023978.
- [2] Andras Gyorgy and Domitilla Del Vecchio. 2014. Modular composition of gene transcription networks. *PLoS Computational Biology* (2014).
- [3] Inigo Incer, Apurva Badithela, Josefine B Graebener, Piergiuseppe Mallozzi, Ayush Pandey, Nicolas Rouquette, Sheng-Jung Yu, Albert Benveniste, Benoit Caillaud, Richard M Murray, et al. 2025. Pacti: Assume-guarantee contracts for efficient compositional analysis and design. *ACM Transactions on Cyber-Physical Systems* 9, 1 (2025), 1–35.
- [4] Timothy S Jones, Samuel MD Oliveira, Chris J Myers, Christopher A Voigt, and Douglas Densmore. 2022. Genetic circuit design automation with Cello 2.0. *Nature protocols* 17, 4 (2022), 1097–1113.
- [5] Cameron D McBride and Domitilla Del Vecchio. 2021. Predicting composition of genetic circuits with resource competition: demand and sensitivity. *ACS Synthetic Biology* 10, 12 (2021), 3330–3342.
- [6] Ayush Pandey. [n. d.]. Code for paper. [Online].
- [7] Fangzhou Xiao, Mustafa Khammash, and John C Doyle. 2021. Stability and control of biomolecular circuits through structure. In *2021 American Control Conference (ACC)*. IEEE, 476–483.

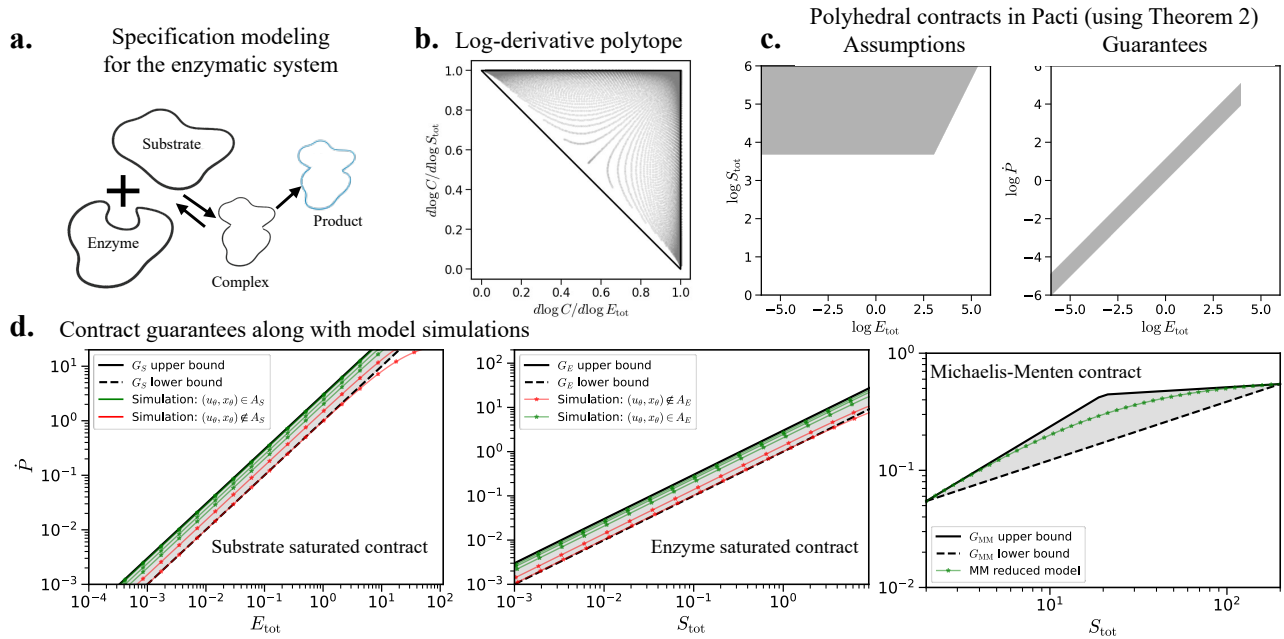


Figure 1: Contract specifications for the enzymatic system. (a) System illustration. (b) Structural polytope of the log-derivative vector showing the admissible regions. (c) Demonstration of polyhedral assumptions and guarantees using Pacti for the substrate-saturated contract. (d) Numerical verification of the reduced model simulations against the contract guarantees for each regime: enzyme-dominant, substrate-dominant, and Michaelis-Menten.

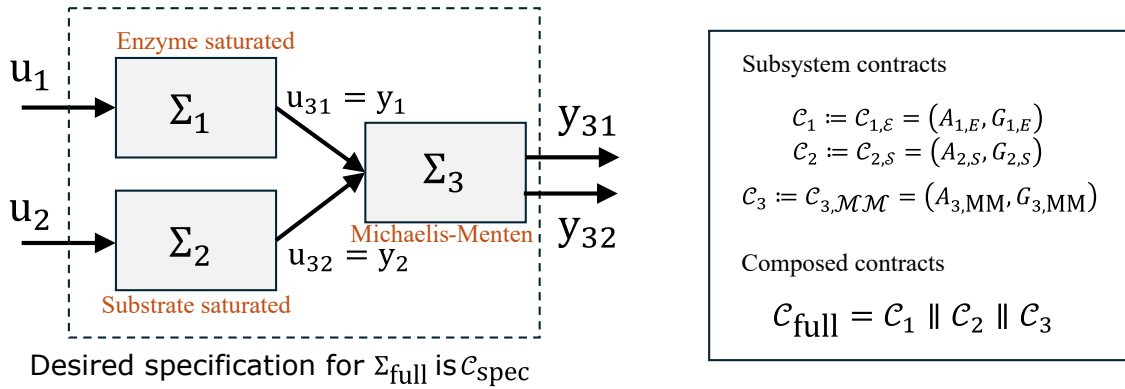


Figure 2: Compositional design of a cascade of three enzymatic subsystems. (left) System illustration. (right) Contract specifications for the three subsystems and the composed system.

#	Comp.	$A_{spec} \subseteq A_{full}$	$G_{cond}$	$C_{full} \preceq C_{spec}$ (refinement)
D1	Yes	No	Yes	No
D2	Yes	No	Yes	No
D3	Yes	Yes	Yes	Yes

Table 1: Compositional design using for contract composition and refinement checking of three designs (design #1 with  $K_{31} = 1$ ,  $K_{32} = 1$ , design #2 with  $K_{31} = 0.9$ ,  $K_{32} = 1.1$ , and design #3 with  $K_{31} = 1.2$ ,  $K_{32} = 0.9$ ). Here  $G_{cond}$  is  $(G_{full} \mid A_{spec}) \subseteq (G_{spec} \mid A_{spec})$ . All parameters not shown, see code [6].

**APPENDIX**

For completeness of the abstract submission, it is imperative that at least a proof sketch is provided. This appendix provides just that. We request that reviewers consider this appendix as a part of the “figures and tables” section of the abstract.

**Proof sketch for contract construction from polytopes**

PROOF. We only provide a short proof sketch in this abstract. The monomial dominance inequalities are affine in  $(z_x, z_u)$  (from Lemma 1), which gives a polyhedral assumption set for a regime. The production and degradation terms are positive sums of monomials, so Theorem 1 implies that  $H_k^{(p)}$  and  $H_k^{(d)}$  lie in the convex hulls of their exponent vectors. Since these are polytopes, the guarantees are polyhedral.  $\square$

**Proof sketch for dominance margin-based contract construction**

PROOF. We present a short proof sketch. Define the sums of the dominant and nondominant monomials  $S_D$  and  $S_N$ .

$$S_D := \sum_{j \in D_k} c_{kj} x^{\alpha_j} u^{\beta_j}, \quad S_N := \sum_{\ell \in J_k^{(p)} \setminus D_k} c_{k\ell} x^{\alpha_\ell} u^{\beta_\ell}.$$

Summing the dominance inequalities over dominant and nondominant terms gives  $d_k S_N \leq \eta^{-1} n_k S_D$ , and therefore

$$\sum_{\ell \in J_k^{(p)} \setminus D_k} \lambda_{k\ell} = \frac{S_N}{S_D + S_N} \leq \frac{n_k}{\eta d_k + n_k} = \rho_k(\eta).$$

This yields the refined guarantee and the contract  $C_{\mathcal{R}, \eta}$ .  $\square$

Finally, please find the contract derivations for the enzymatic system in the code repository [6].

# Towards interoperable modeling of toehold-mediated strand exchange circuits across DNA nanotechnology and engineering biology

**Zoila Jurado**

National Institute of Standards and  
Technology  
Gaithersburg, MD  
zoila.jurado@nist.gov

**Geoffrey Taghon\***

National Institute of Standards and  
Technology  
Gaithersburg, MD  
geoffrey.taghon@nist.gov

**Samuel W. Schaffter**

National Institute of Standards and  
Technology  
Gaithersburg, MD  
samuel.schaffter@nist.gov

## 1 INTRODUCTION

A major goal of engineering biology is the ability to rationally design genetic circuits from first principles that process molecular information in biological systems. Paramount to this goal are mathematical models that simulate the dynamics of the biomolecular components that make up these engineered systems [4, 6, 7]. Great strides have been made in this direction, with genetic circuit modeling and design software packages gaining broader adoption, particularly for transcription-factor-based genetic circuits [9, 19]. Not only are these models guiding how we design genetic circuits to achieve a desired outcome, they are also broadening our understanding of these systems by identifying undesigned secondary reactions [5].

Beyond genetic circuits, i.e., circuits that rely on gene expression to process and transduce information, nucleic acid-only circuits that operate via toehold-mediated strand exchange (TMSE) reactions [17, 20] (Figure 1) offer an attractive medium for programming biology. TMSE circuits rely entirely on nucleic acid base pairing interactions, making them easy to rationally program [2, 3, 14], and versatile for use across different application spaces [10]. Despite the increasing utilization of TMSE in engineering biology, a modeling package that supports flexible integration of TMSE reactions with genetic circuitry is lacking. To bridge the modeling needs that span the fields of DNA nanotechnology and engineering biology, we develop a suite of software tools to support the compilation and simulation of TMSE reactions alongside genetic circuitry. This work provides a foundational resource for the rational design of molecular information processing systems, paving the way towards establishing a common language — or at least establishing interoperability — within and across DNA nanotechnology and engineering biology models. We anticipate this work will significantly enhance the design-build-test-learn (DBTL) cycle for TMSE circuits in biological contexts.

\* Author also is affiliated with Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland.

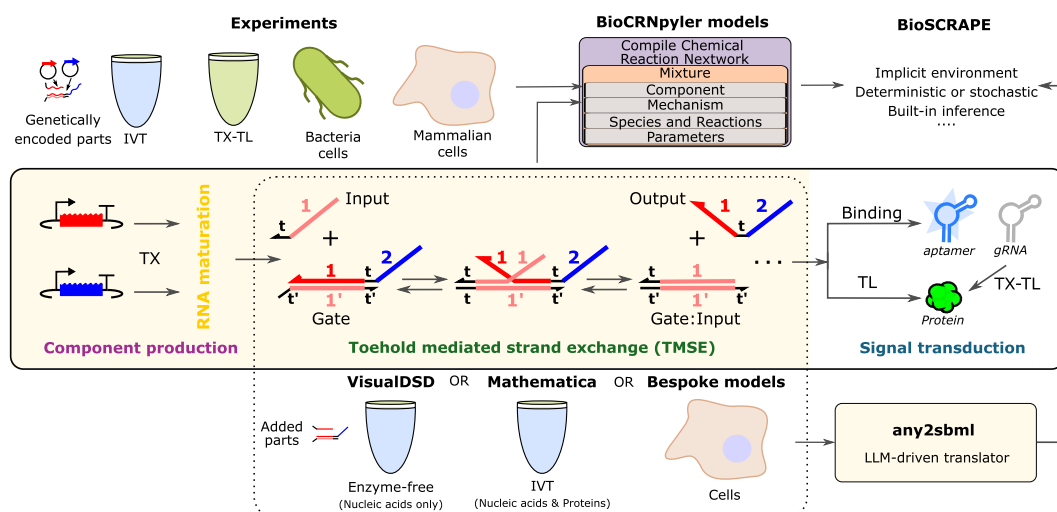
## 2 RESULT AND DISCUSSION

Our suite of software tools are developed for BioCRNpyler — a Python package for generating chemical reaction networks (CRNs) [13]— and BioSCRAPE — a corresponding Python package for the flexible modeling and simulation of CRNs [12] (Figure 1). We demonstrate the utility and flexibility of our TMSE modules by compiling and simulating >10 examples from the literature. These examples include incorporating TMSE circuits with upstream transcription processes and downstream gene expression pathways to showcase applications where TMSE is used across various sub-fields of DNA nanotechnology and engineering biology.

The compiled models from BioCRNpyler are simulated using BioSCRAPE [12] or exported in Systems Biology Markup Language (SBML) format. To enable the reuse and repurposing of existing models unavailable as in SBML format, we develop any2sbml — an LLM-driven translator — to translate reaction descriptions from a PDF or TXT file to SBML version 3 XML file. As part of this work, we demonstrate the conversion of model descriptions from literature into SBML format, using any2sbml, for simulation and modification in BioSCRAPE. The software tools we develop, along with the examples we present to illustrate their utility, will help bridge modeling across DNA nanotechnology / DNA computing and engineering biology to open up a myriad of new applications.

The literature examples we model in Figure 2 showcase our TMSE tool’s ability to instantiate circuits utilizing existing BioCRNpyler capabilities — such as simulating different environments and leveraging gene regulation, TX, and TL models. These examples include incorporating TMSE circuits with upstream transcription processes and downstream gene expression pathways to showcase applications spanning DNA computing, cell-free biosensors, and bacterial and mammalian cell engineering.

In enzyme-free conditions, DNA-based circuits that operate entirely on TMSE have been scaled up to hundreds of components to execute digital calculations, molecular pattern recognition, and supervised learning [2, 3, 14]. Thus, we



**Figure 1: Schematic overview of BioCRNpyler modeling workflow, TMSE circuits, and current modeling platforms. A high-level depiction of the experiment, model, and simulation workflow encompassing BioCRNpyler and BioSCRAPE (top). An illustration of compiling signal transduction circuit using TMSE circuits in BioCRNpyler (middle). The new components for strand-exchange circuits are highlighted in yellow: component production, RNA maturation, and TMSE. The dotted box encapsulated current TMSE circuits modeling platforms (bottom) in their respective experimental conditions. The any2sbml tool enables cross platform integration to BioSCRAPE or other SBML simulators.**

first illustrate how to use the TMSE modules we developed to implement a DNA-based circuit that implements OR logic in BioCRNpyler (Figure 2B(i), Figure 2C(i), and Figure 2D(i)).

A straightforward route for connecting TMSE information processing circuits to non-nucleic acid inputs is through introduction of allosteric transcription factors [11]. Figure 2C(ii) demonstrates how our TMSE tool and BioCRNpyler can be used to combine transcription and DNA-based TMSE reactions for a circuit that executes OR logic on molecules A and B that regulate transcription (Figure 2B(ii)) with simulation results depicted in Figure 2D(ii). The model we present here differs slightly from the model described in [11], which explicitly models transcription factor and RNA polymerase binding and unbinding. However, our any2sbml tool can convert the reactions described in the Supporting Information of [11] into SBML format that can be simulated or expanded to incorporate translation in BioSCRAPE (Figure 3).

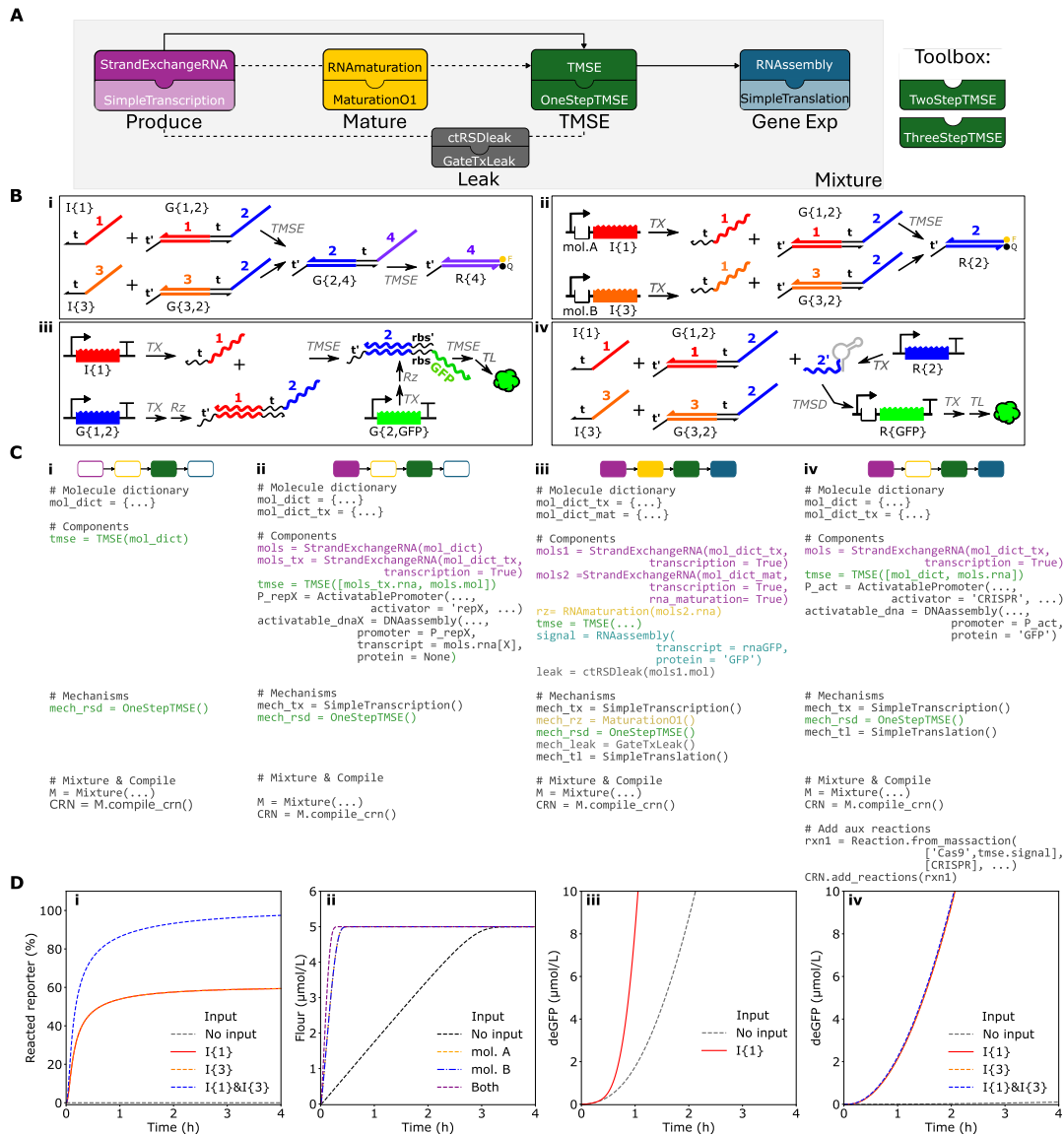
A promising new direction for TMSE circuits is the development of versions composed entirely of RNA components that can be genetically encoded for continuous operation in cells [1, 15, 16]. Figure 2C(iii) illustrates how our TMSE modules and BioCRNpyler can implement these RNA circuits that regulate protein expression in cell-free-TX-TL environments (Figure 2B(iii)).

In addition to using genetically encoded TMSE circuits in bacteria, fixed concentrations of TMSE components have been transfected into mammalian cells and used to process information and regulate gene expression [8, 18]. Figure

2B(iv) shows a system in which TMSE is used to regulate the activity of a CRISPR guide RNA, which in turn activates gene expression. Figure 2C(iv) illustrates how our TMSE modules and BioCRNpyler can be implemented to compile this model.

### 3 CONCLUSION

We anticipate that this work will significantly enhance the DBTL cycle for TMSE circuits in biological contexts. The TMSE modules allow for the customization of components, mechanisms, and parameters of desired TMSE reactions. The implementation in BioCRNpyler enables the integration of TMSE reactions with genetic circuitry, facilitating the design, analysis, and predictive modeling of complex biochemical networks within and across DNA nanotechnology and engineering biology. To demonstrate adaptability of the modules, we validated the modules through the reconstruction and simulation of ten published TMSE applications in BioSCRAPE. To enable interoperability of existing modeling platforms in BioSCRAPE and other SBML simulators, we developed a LLM-driven translator, any2sbml, to convert models described in plain text to SBML Level 3 XML file. By providing a common language and framework for TMSE circuits, we envision accelerating the development of innovative engineering biology applications of nucleic acid computing. Ultimately, this work provides a foundational resource for the rational design of molecular information processing systems — bridging DNA nanotechnology and engineering biology.



**Figure 2: Comprehensive overview of the design, components, and simulation of transcriptional machinery-based synthetic gene expression (TMSE) circuits. (A) Illustration of primary TMSE modules, including their respective interactions and connections, with dashed lines indicating optional pathways. Each component (upper block) is highlighted, along with the corresponding mechanism (lower block). Additional mechanisms are displayed in the toolbox. (B) Schematics of four distinct TMSE circuit types: (i) DNA seesaw circuits [2, 3, 14], (ii) transcription-factor-based biosensors [11], (iii) genetically encoded RNA-based TMSE circuits [1, 15, 16], and (iv) DNA-based TMSE circuits designed to regulate gene expression in mammalian cells [18]. (C) Pseudocode outlining the modeling approach and (D) the respective simulated results for each of the four circuit types.**

**REFERENCES**

[1] BAE, W., STAN, G.-B. V., AND OULDRIDGE, T. E. In situ Generation of RNA Complexes for Synthetic Molecular Strand-Displacement Circuits in Autonomous Systems. *Nano Letters* 21, 1 (Jan. 2021), 265–271.

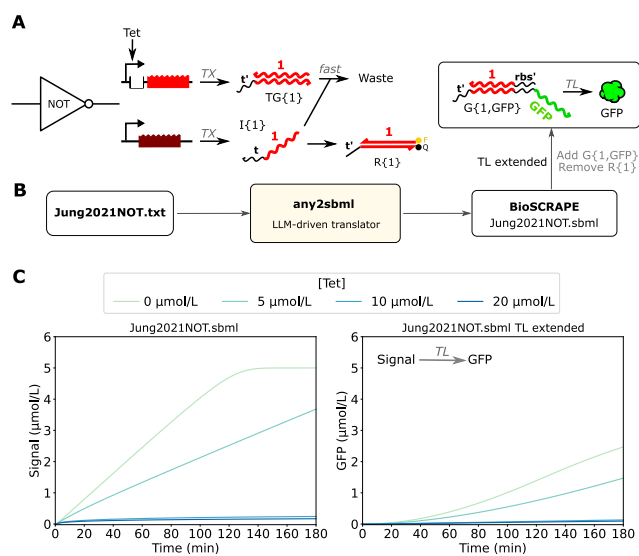
[2] CHERRY, K. M., AND QIAN, L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature* 559, 7714 (2018), 370–376.

[3] CHERRY, K. M., AND QIAN, L. Supervised learning in DNA neural networks. *Nature* 645, 8081 (2025), 639–647.

[4] ELWITZ, M. B., AND LEIBLER, S. A synthetic oscillatory network of transcriptional regulators. *Nature* 403, 6767 (Jan. 2000), 335–338.

[5] ESPAH BORUJENI, A., ZHANG, J., DOOSTHOSSEINI, H., NIELSEN, A. A. K., AND VOIGT, C. A. Genetic circuit characterization by inferring RNA polymerase movement and ribosome usage. *Nature Communications* 11, 1 (Oct. 2020), 5001.

[6] GARCIA-OJALVO, J., ELWITZ, M. B., AND STROGATZ, S. H. Modeling a synthetic multicellular clock: Repressilators coupled by quorum



**Figure 3: Simulating transcription-factor-based biosensors [11] in BioSCRAPER. (A) Workflow of converting transcription-factor-based biosensors model to SBML using any2sbml tool. (B) Simulation results of the original model that produces a Signal RNA (left) and the integration of translation of Signal RNA to GFP (right).**

sensing. *Proceedings of the National Academy of Sciences* 101, 30 (July 2004), 10955–10960.

- [7] GARDNER, T. S., CANTOR, C. R., AND COLLINS, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403, 6767 (2000), 339–342.
- [8] GROVES, B., CHEN, Y.-J., ZURLA, C., POCHEKAILOV, S., KIRSCHMAN, J. L., SANTANGELO, P. J., AND SEELIG, G. Computing in mammalian cells with nucleic acid strand exchange. *Nature Nanotechnology* 11, 3 (2016), 287–294.
- [9] JONES, T. S., OLIVEIRA, S. M. D., MYERS, C. J., VOIGT, C. A., AND DENSMORE, D. Genetic circuit design automation with Cello 2.0. *Nature Protocols* 17, 4 (2022), 1097–1113.
- [10] JUNG, H., COLLINSON, E., HAWES, A. P., AND FELLERMANN, H. From the Test Tube to the Cell: A Homecoming for DNA Computing Circuits? *Intelligent Computing* 4 (2025), 0112.
- [11] JUNG, J. K., ARCHULETA, C. M., ALAM, K. K., AND LUCKS, J. B. Programming cell-free biosensors with DNA strand displacement circuits. *Nature Chemical Biology* 18, 4 (2022), 385–393.
- [12] PANDEY, A., POOLE, W., SWAMINATHAN, A., HSIAO, V., AND MURRAY, R. M. Fast and flexible simulation and parameter estimation for synthetic biology using bioscrape. *Journal of Open Source Software* 8, 83 (2023), 5057.
- [13] POOLE, W., PANDEY, A., SHUR, A., TUZA, Z. A., AND MURRAY, R. M. BioCRNpyler: Compiling chemical reaction networks from biomolecular parts in diverse contexts. *PLOS Computational Biology* 18, 4 (2022), e1009987.
- [14] QIAN, L., AND WINFREE, E. Scaling Up Digital Circuit Computation with DNA Strand Displacement Cascades. *Science* 332, 6034 (2011), 1196–1201.
- [15] SCHAFFTER, S. W., AND STRYCHALSKI, E. A. Cotranscriptionally encoded RNA strand displacement circuits. *Science Advances* 8, 12 (2022), eab14354.
- [16] SCHAFFTER, S. W., WINTENBERG, M. E., MURPHY, T. M., AND STRYCHALSKI, E. A. Design Approaches to Expand the Toolkit for Building Cotranscriptionally Encoded RNA Strand Displacement Circuits. *ACS Synthetic Biology* 12, 5 (2023), 1546–1561.
- [17] SRINIVAS, N., OULDRIDGE, T. E., ŠULC, P., SCHAEFFER, J. M., YURKE, B., LOUIS, A. A., DOYE, J. P. K., AND WINFREE, E. On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research* 41, 22 (2013), 10641–10658.
- [18] SUN, J., XIONG, X., LAI, W., WU, Z., WANG, H., YANG, L., XUE, N., YAO, Q., SONG, G., ZHAO, Y., LI, L., WANG, F., FAN, C., AND PEI, H. Implementing complex nucleic acid circuits in living cells. *Science Advances* 11, 18 (2025), eadv6512.
- [19] WATANABE, L., NGUYEN, T., ZHANG, M., ZUNDEL, Z., ZHANG, Z., MADSEN, C., ROEHNER, N., AND MYERS, C. iBioSim 3: A Tool for Model-Based Genetic Circuit Design. *ACS Synthetic Biology* 8, 7 (2019), 1560–1563.
- [20] ZHANG, D. Y., AND WINFREE, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *Journal of the American Chemical Society* 131, 47 (2009), 17303–17314.

# Host-Aware Design of Cellular ctRSD Circuits

Zachary Hastings<sup>1</sup>, Zoila Jurado<sup>2</sup>, Bingqing Hu<sup>3</sup>, Zhen Zhang<sup>3</sup>, Samuel W. Schaffter<sup>2</sup>, Lukas Buecherl<sup>1</sup>

<sup>1</sup> Biological Engineering, Utah State University, USA

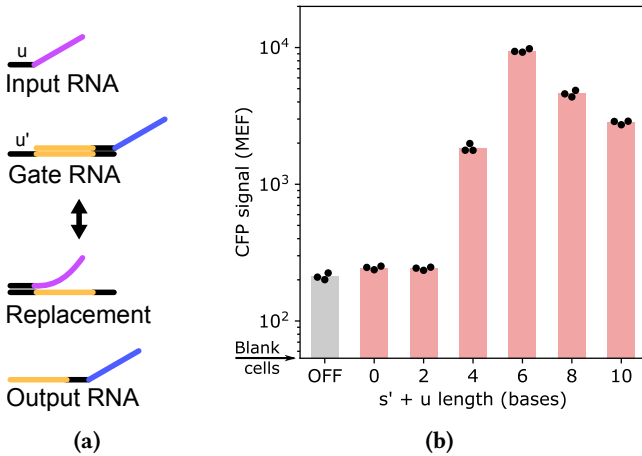
<sup>2</sup> National Institute of Standards and Technology, USA

<sup>3</sup> Electrical and Computer Engineering, Utah State University, USA

lukas.buecherl@usu.edu

## INTRODUCTION

Co-transcriptional RNA strand-displacement (ctRSD) circuits are a modular framework for programmable molecular computation [1]. In these systems, an Input RNA (purple) binds an exposed sequence on a Gate RNA (yellow), called the toehold  $u'$ , initiates strand exchange, and displaces a previously blocking strand to release a sequestered Output strand (blue), thereby activating the circuit output (Figure 1a) [2, 3].



**Figure 1: (a) An input RNA binds a gate toehold, initiates strand exchange, and releases a sequestered output strand. (b) *In vivo* output varies non-monotonically with input  $s' + u$  length [2, 3].**

Because ctRSD function is encoded directly in RNA sequence and structure, circuit behavior depends strongly on design choices. This makes predictive modeling important for more systematic and automatable circuit design. In particular, toehold length is often treated as a key design parameter because it modulates strand-displacement efficiency and thereby influences circuit output. Here, we focus on an unexpected toehold-length-dependent trend in cellular ctRSD circuits. Experimentally, signal increased as input toehold length increased, but only up to a point; beyond that, further increased length reduced output strength (Figure 1b). This non-monotonic behavior is inconsistent with the hypothesis that longer toeholds should improve intended circuit interactions as observed in *in vitro* systems [2, 3]. We therefore ask whether endogenous host transcript interactions contribute to the observed *in vivo* behavior. In this work, *host-aware*

refers to models and design workflows that account for unintended interactions with the host transcriptome in addition to intended circuit behavior. Simply put, this work asks why increasing the input binding region improves circuit performance in cell-free settings but can reduce performance in cells, and whether endogenous RNA interactions explain this shift.

Beyond explaining this specific trend, our results raise a broader challenge for bio-design automation. Candidate RNA circuit components may need to be evaluated not only for intended circuit function, but also for interactions with the host transcriptome.

## METHODS

To investigate the discrepancy between expected and observed toehold-length-dependent behavior in cellular ctRSD circuits, we combined prior experimental measurements [2, 3], transcriptome-scale interaction analysis, and *chemical reaction network* (CRN) modeling. We used *cyan fluorescent protein* (CFP) output to compare ctRSD designs with different  $s' + u$  input lengths, where  $s' + u$  represents the effective input binding region, and measured how output expression changed as input length increased.

To test whether endogenous RNAs contribute to the observed trend, we analyzed predicted interactions between input RNAs and the *E. coli* transcriptome using IntaRNA [4]. We then compared predicted interaction energies across input lengths to assess whether longer inputs exhibited stronger or greater number of candidate off-target interactions.

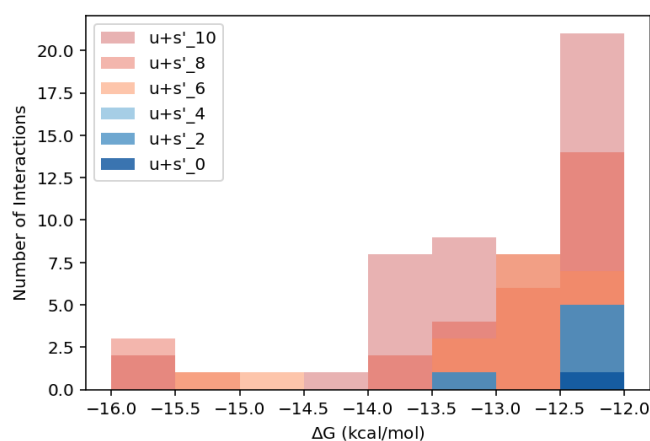
We modeled ctRSD circuit function and extended the baseline model with interference reactions representing endogenous RNAs that bind circuit components and reduce effective circuit function. Model behavior was evaluated by simulating the output across input designs and parameter regimes.

Because transcriptome interaction predictions provide binding energies rather than direct kinetic parameters, we compared different approaches for translating predicted interaction strengths into effective reaction rates, including a standard thermodynamic relationship between  $\Delta G$  and equilibrium behavior [5] and an alternative approximation with a stronger dependence on binding energy [6]. To further isolate host-context effects, we used a simplified model

with manually defined interfering transcripts to vary interferer abundance, binding strength, and strand-displacement kinetics and examine their effects on predicted CFP output.

## RESULTS

Transcriptome-wide interaction analysis supported the hypothesis that endogenous host RNAs contribute to the observed non-monotonic toehold-length-dependent behavior. Predicted interactions between the input RNAs and the *E. coli* transcriptome shifted toward stronger binding energies as input length increased (Figure 2). Longer inputs also exhibited more candidate off-target interactions, consistent with the idea that increasing the accessible binding region improves intended circuit interactions while also increasing opportunities for unintended interactions with host transcripts. These results provide a plausible mechanistic explanation for why longer inputs may eventually reduce, rather than improve, circuit output *in vivo*.



**Figure 2: Predicted interactions between input RNAs and the *E. coli* transcriptome. Longer inputs show stronger off-target interactions, suggesting greater endogenous interference.**

Comparison of alternative energetic-to-kinetic mappings showed that the predicted impact of host interference depends strongly on how binding energies are translated into effective reaction rates. Under a standard thermodynamic relationship between  $\Delta G$  and equilibrium behavior, changes in interference strength had relatively modest effects over the relevant range of predicted interaction energies. Under the alternative approximation, stronger interactions produced a larger shift in predicted CFP output and more pronounced suppression of expression. This indicates that host-aware ctRSD models are sensitive not only to which endogenous interactions are included, but also to the assumptions used to convert sequence-derived energetics into kinetic effects.

## DISCUSSION AND FUTURE WORK

Together, these results support the hypothesis that endogenous host transcript interactions contribute to the discrepancy between expected and observed toehold-dependent

ctRSD behavior *in vivo*. Rather than being determined solely by intended strand-displacement design rules, cellular RNA circuits may also be shaped by sequence-dependent interactions with the host transcriptome. This suggests that host-context effects should be treated as a design constraint in bio-design automation, rather than as a secondary complication considered only after experimental failure.

At the same time, the predicted impact of host interference depended strongly on how sequence-derived interaction energies were translated into effective reaction rates, indicating that the qualitative conclusion is more robust than any single quantitative parameterization. Future work will focus on experimentally validating the proposed interference mechanism and incorporating these insights into a host-aware workflow that screens candidate ctRSD designs for likely endogenous transcriptome interactions before cellular implementation. These results also motivate updates to bio-design software so that host-aware screening can be incorporated alongside intended circuit analysis during sequence design.

## ACKNOWLEDGMENTS

This work was supported by the BE Dept. at Utah State University (USU) for Z.H. and L.B., and by the ECE Dept. at USU for B.H. and Z.Z. B.H. was supported in part by National Science Foundation (NSF) grant No.2422206. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. Z. Jurado is supported by the National Research Council Postdoctoral Fellowship.

## REFERENCES

- [1] Begüm Sarac, Seydanur Yücer, and Fatih Ciftci. Synthetic biology-driven biosensors for healthcare applications: A roadmap toward programmable and intelligent diagnostics. *Biosensors and Bioelectronics*, 291:118036, January 2026. doi:10.1016/j.bios.2025.118036.
- [2] Samuel W Schaffter, Olga B Vasilyeva, Molly E Wintenberg, John M Hurley, and Nina Y Alperovich. Genetically encoded RNA strand exchange circuits for programmable protein expression and computation in cells, September 2025. doi:10.1101/2025.09.24.678369.
- [3] Fernanda Piorino, Eugenia Romantseva, and Samuel W Schaffter. Cell-free systems enable mechanistic characterization of genetically encoded RNA strand exchange circuits for programmable protein expression, September 2025. doi:10.1101/2025.09.25.678518.
- [4] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: Enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Research*, 45(W1):W435-W439, July 2017. doi:10.1093/nar/gkx279.
- [5] Dokyun Na, Seung Min Yoo, Hannah Chung, Hyegwon Park, Jin Hwan Park, and Sang Yup Lee. Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nature Biotechnology*, 31(2):170-174, February 2013. doi:10.1038/nbt.2461.
- [6] Olivia Gallup and Harrison Steel. Generative design of synthetic gene circuits for functional and evolutionary properties. *npj Systems Biology and Applications*, March 2026. doi:10.1038/s41540-026-00683-6.

# Dynamic control of metabolic flux by RNA-based genetic circuits designed, built, and tested through automation

Stephen Chiu  
Tae Seok Moon\*

schiu@jcvl.org  
tsmoon7@gmail.com

J. Craig Venter Institute  
La Jolla, CA, USA

## 1 INTRODUCTION

Dynamic controllers in metabolic engineering enable autonomously tuned gene expression in genetically engineered microbes (GEMs), minimizing unnecessary gene expression, optimizing cellular resource allocation, and enhancing target product yields [1]. Quorum sensing and RNA-based genetic circuits are important tools in synthetic biology, offering autonomous and just-in-time gene regulation in GEMs (Fig. 1). However, building such genetic circuits is still challenging due to the required optimization and lack of design rules even after over two decades of advancements in synthetic biology.

## 2 CONTENTS OF THE ABSTRACT

In this work, we leverage standardized genetic parts well-characterized by automation devices for culturing, liquid handling, and measurement to build genetic circuits and metabolic pathways (Fig. 2). As a proof of concept, we integrate the LuxI/LuxR quorum sensing system with a selected RNA-based genetic circuit to achieve completely autonomous regulation in metabolic pathways. Using our predictive model-based small RNA (sRNA) design [2], we downregulated the *E. coli pfk-1* gene and combined it with the small transcription activating RNAs (STARs) to construct repressor and activator modules as well as logic circuits (Fig. 3, 4). Additionally, an autoinducer production module was added to the system to enable the engineered strain to autonomously rewire the metabolic flux from the glycolytic pathway to the target production pathway. This circuit design was first applied to optimize the heterologous Myo-inositol production pathway, improving the titer significantly (Fig. 5). Subsequently, the circuit was extended to regulate the glucaric acid pathway [3], incorporating three orthogonal STAR variants to balance metabolic flux by controlling the expression level and further optimizing the production. These results demonstrate that autoregulated RNA-based circuits can effectively enhance dynamic gene

expression control and improve the overall productivity of metabolic pathways in GEMs.

## 3 REFERENCES

- [1] A Hoynes-O'Connor and TS Moon. Programmable genetic circuits for pathway engineering. *Curr. Opin. Biotechnol.* 36, 115-121 (2015)
- [2] YJ Lee, SJ Kim, MB Amroffell and TS Moon. Establishing a multivariate model for predictable antisense RNA-mediated repression. *ACS Synth. Biol.* 8, 45–56 (2019)
- [3] TS Moon, S-H Yoon, A Lanza, J Roy-Mayhew and KJ Prather. Production of Glucaric Acid from a Synthetic Pathway in Recombinant *Escherichia coli*. *Appl. Environ. Microbiol.* 75, 589-595 (2009)

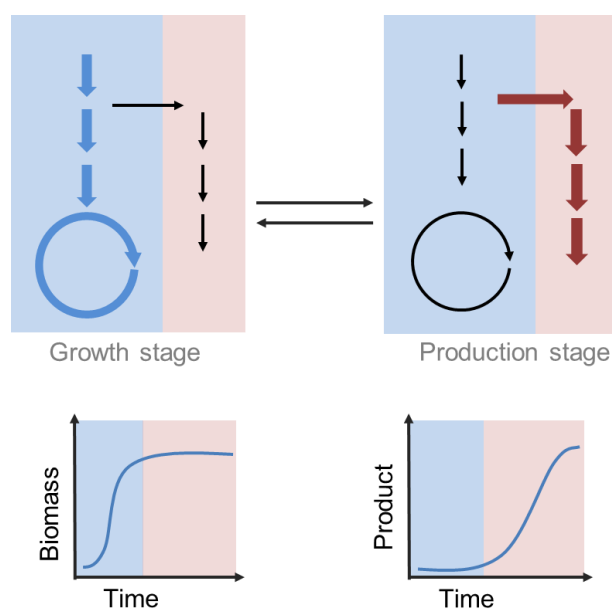


Fig. 1. Quorum sensing-based metabolism control as a function of cell population density.

\*The research was funded by the National Science Foundation (EF-2222403, MCB-2001743, and OISE-2435184 to T.S.M.)

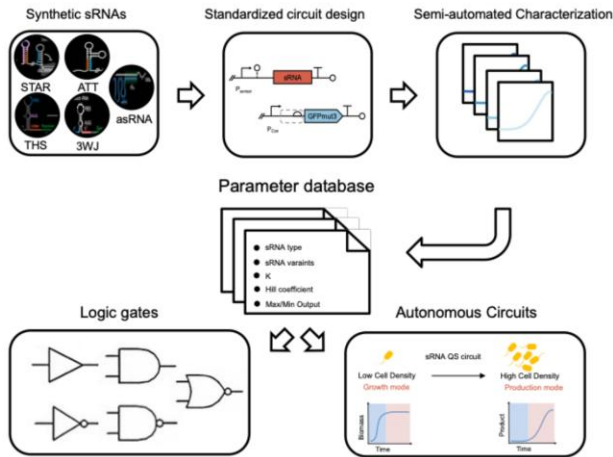


Fig. 2. Scheme of our workflow for standardized genetic part design, automation-assisted characterization, and rapid genetic circuit building and testing.

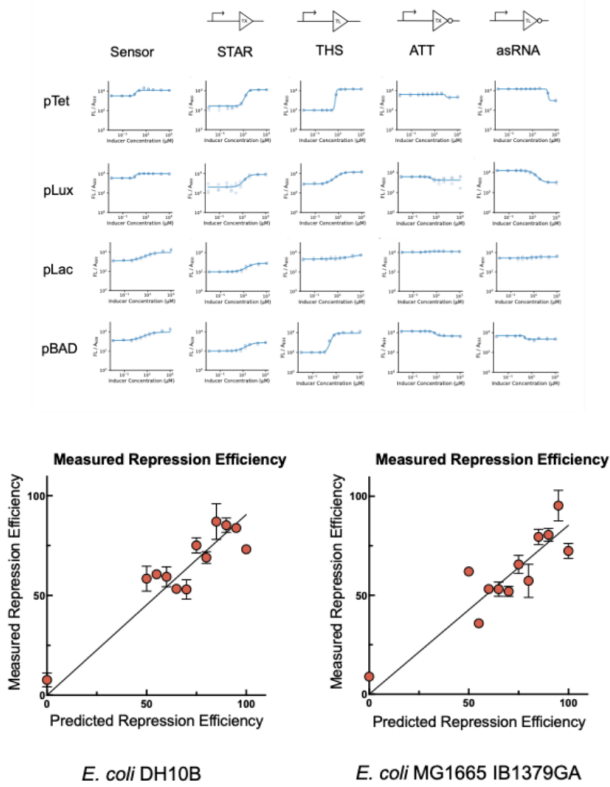


Fig. 3. Automation-assisted characterization of genetic parts and devices for predictable genetic circuit construction. (Top) Four different promoter-protein regulator sensors were combined with RNA-based regulators for gene activation or repression. (Bottom) *pfk-1* targeted asRNAs (antisense RNAs) show predictable gene repression, allowing fine-tuned metabolic control.

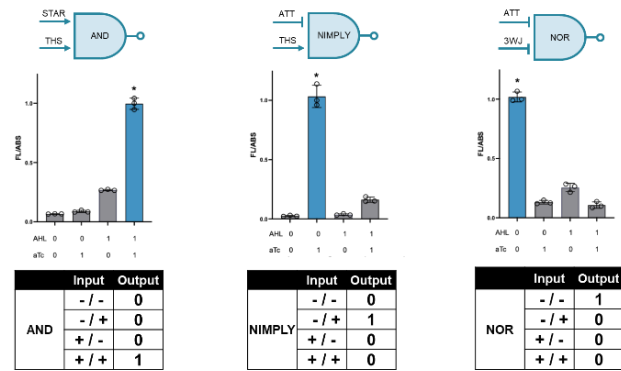
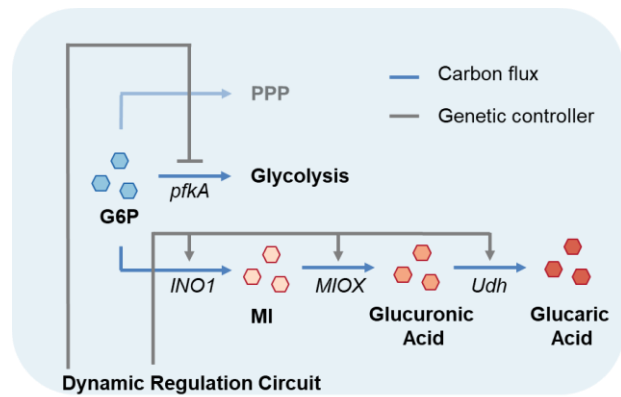


Fig. 4. Three representative genetic logic circuits designed, constructed, and tested using the well-characterized genetic parts and devices using our automation machine for culturing, liquid handling, and measurement. STAR, small transcription activating RNA; ATT, attenuator; THS, toehold switch; 3WJ, 3-way junction-based repressing RNA.



Repression and Activation

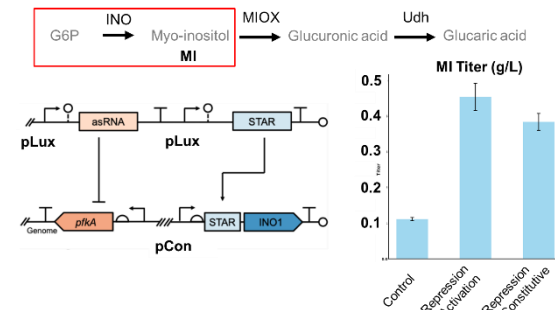


Fig. 5. Proof-of-concept demonstration of quorum-sensing-based genetic circuits that control the heterologous Myo-inositol production pathway as a function of cell population density, improving the titer significantly. Blunt arrow, repression. Pointed arrow, activation.

# BELL: Biomodel Evidence and LLM-based Logic

Niloofar Arazkhani  
Department of Electrical and Computer Engineering  
*University of Pittsburgh*  
Pittsburgh, PA, USA

Natasa Miskov-Zivanov  
Department of Electrical and Computer Engineering  
*University of Pittsburgh*  
Pittsburgh, PA, USA

## I. INTRODUCTION

Natural language processing (NLP) has expanded the scale at which interactions can be extracted from the literature, but extraction errors remain common and difficult to detect without manual review. While recent advances in large language models (LLMs) offer an additional layer of reasoning over the extracted knowledge, these models are still prone to hallucinations. At the same time, large scale knowledge extraction is critical for creating comprehensive mechanistic knowledge graphs and executable models of biological systems, and the inaccuracies in extraction tools carry real consequences for reasoning about these systems. For example, a single misrepresented or absent interaction can alter predicted network dynamics and produce misleading biological conclusions. Therefore, verification of these networks is critical, as it confirms that individual interactions are supported by biological evidence, a distinct concept from validation, which compares predicted dynamics to experimental outcomes.

Since much of the human interactome is already documented in publicly accessible databases that can be programmatically queried, this offers an opportunity for scalable network verification [1-3]. Systematic assessment of interactions extracted by NLP-based and LLM-assisted tools requires cross-referencing each interaction against multiple databases that differ in coverage, curation standards, and evidence representation, a process that is both time-consuming when done manually and difficult to apply consistently at scale. Existing tools such as FLUTE [4] address this by downloading information from databases for offline verification, but this approach requires periodic data updates to remain current, which is impractical as a long-term curation strategy.

To address this gap, we present BELL (Biomodel Evidence and LLM-based Logic), a biocuration support tool that automates the evidence retrieval, scoring, and explanation steps of interaction-level verification. Recent advances in LLMs offer an additional layer of reasoning over the retrieved evidence, enabling automated interaction-level verification at scale. We evaluated the use of BELL on a corpus containing interactions extracted from published literature. Our results confirm that its user-friendly interface substantially reduces the manual effort required for post-extraction curation while providing the evidence transparency needed for informed and reproducible annotation decisions.

## II. METHODS

**Interaction Processing Pipeline.** BELL processes each interaction through a five-step pipeline (Figure 1). The first four steps, namely entity grounding, database ranking, evidence

retrieval, and scoring, are fully deterministic ensuring transparency and reproducibility. In the fifth step, Llama 3.1 [5] receives the pre-computed scores as read-only input and generates a natural language chain-of-thought explanation along with a recommended curator action (Accept, Accept with enrichment, Flag for review, or Reject). Every step of the evaluation is visible to the curator; the resolved entity names and identifiers, the evidence returned by each database, the computed scores, and the full LLM prompt and response are all displayed in the interface alongside the interaction record.

**Interaction Representation Format.** For interactions, BELL uses the BioRECIPE interaction list format (BioRECIPE-IL)[6]. This format includes entity attributes such as name, type, subtype, unique ID, and compartment for the regulator and regulated entity in a directed interaction, as well as interaction attributes, namely sign, mechanism, connection type, context attributes cell line, cell type, tissue type, and organism, and provenance attributes for the source of interaction such as paper ID and statement. Detailed documentation for this representation format is available in BioRECIPE ReadtheDocs [7]. BioRECIPE-IL is a tabular format where each interaction is included as a row in the table and columns are dedicated to entity and interaction attributes.

**Entity grounding.** The first step in the pipeline ensures assignment of unique identifiers to entities. When the HGNC Symbol and the Database ID attributes are already populated for both the regulator and regulated entities in the BioRECIPE-IL row, BELL uses those identifiers directly for querying. For entities without pre-existing identifiers, Gilda [8] resolves the entity name to a canonical identifier using the source statement for disambiguation. The canonical identifier (e.g. HGNC:3236) and approved gene symbol (e.g. EGFR) are both displayed to the curator in the interface, while the approved gene symbol is also used for database queries.

**Database selection and evidence retrieval.** We selected seven databases to use with BELL: INDRA [9], OmniPath [10], STRING [3], SIGNOR [11], Reactome [1], BioGRID [12], and PCnet [13]. These databases are publicly accessible, widely used in the cell signaling community, and collectively cover the major evidence types relevant to interaction-level verification. Together they span manually curated causal relationships, experimentally detected physical interactions, pathway co-participation, and functional associations, ensuring that evidence is retrieved from complementary sources. Importantly, all seven databases provide free programmatic access via REST APIs, making BELL easily deployable. For every interaction in the input interaction lists, BELL queries all databases uniformly

and uses the Mechanism attribute to rank them and reorder the display, prioritizing the databases most likely to contain relevant evidence for that mechanism type. This priority mapping is based on the known specialization of each database: SIGNOR is ranked first for phosphorylation and other post-translational modifications, Reactome for transcriptional regulation, and BioGRID and STRING for physical binding interactions.

**Interaction scoring.** Four scores (S1-S4) are computed programmatically from the aggregated evidence (Table 1). S1 measures source coverage as a weighted sum of the points from databases that found the interaction. Tier 1 databases (INDRA, SIGNOR) contribute 3 points each as they are based on direct manual curation from primary literature with explicit mechanism annotations. Tier 2 databases (OmniPath, Reactome, BioGRID) contribute 2 points each as they provide high-quality but less specific evidence. Tier 3 databases (STRING, PCnet) contribute 1 point each as they rely partly on computational inference rather than direct experimental evidence. The scores are normalized by dividing with the maximum possible sum of 14.

S2 measures attribute match across found sources for three interaction attributes, direction, sign and mechanism: 1.0 when confirmed by the source, 0.5 if the database found the interaction but does not report that attribute, and 0.0 if the attribute is contradicted. For each database where the interaction is found, the three attribute scores are averaged to produce a per-database attribute score. The final S2 score is calculated as the mean of the per-database attribute scores, considering only databases that found the interaction; databases that returned no result are excluded from this calculation entirely.

S3 is the INDRA aggregated belief score for retrieved statements, reflecting probabilistic confidence derived from literature frequency and source reliability; it is set to 0.0 if INDRA returns no results. S3 is reported as a supplementary metric but excluded from the confidence threshold (described below), as a score of 0.0 reflects only that INDRA did not return results rather than a lack of supporting evidence overall. Finally, S4 is a contradiction flag set to 2 if at least two sources disagree on the sign of the interaction, and 0 otherwise.

BELL outputs final confidence scores using the following reasoning: HIGH when  $S1 \geq 0.7$  and  $S2 \geq 0.5$  and  $S4 = 0$ ; MEDIUM when  $S1 \geq 0.4$  or  $S2 \geq 0.3$ ; and LOW otherwise. To direct curator's attention to specific aspects of the evidence, BELL also generates three qualitative flags: low\_source\_coverage ( $S1 < 0.4$ ); attribute\_match\_partial ( $S2 < 0.7$ ); and mechanism\_unconfirmed (no source confirms the declared mechanism).

**LLM-based Explanation.** While all the previous computations are deterministic and reproducible, in the final step BELL generates chain-of-thought explanations using Llama 3.1. The model receives a structured prompt (Figure 2) containing three sections: the interaction record with source statement, per-database evidence with direction, sign, and mechanism annotations, and pre-computed S1-S4 scores with interpretive labels. As shown in Figure 2, the model is instructed to reason through six explicit steps including source coverage, attribute match, INDRA belief, contradictions, flags, and

recommendation, before producing its output. The model is explicitly constrained from modifying the numeric scores and must conclude with exactly one of four curator actions: Accept, Accept with enrichment, Flag for review, or Reject. Figure 3 shows an example of the LLM's response suggesting to 'Accept with enrichment'.

### III. RESULTS AND DISCUSSION

In the BELL interface (Figure 4), a curator uploads a BioRECIPE interaction list and runs BELL on each interaction. By reviewing the confidence level and recommended action assigned to each interaction, the interactions marked Accept or Accept with enrichment can be approved quickly, while those flagged for review are brought to attention for closer inspection. The qualitative flags direct the curator to the specific aspect of the evidence that is weak without requiring the curator to read the full explanation for every interaction. For interactions requiring closer inspection, the curator can expand the detail panel to review the full evidence trail and use it to inform their final annotation decision. The curator then annotates each interaction independently as Accepted, Flagged, or Rejected. The fully annotated list can be exported as a spreadsheet for direct use in downstream model construction.

### REFERENCES

- [1] M. Milacic *et al.*, "The Reactome Pathway Knowledgebase 2024," *Nucleic Acids Res*, vol. 52, no. D1, pp. D672–D678, Jan 5 2024, doi: 10.1093/nar/gkad1025.
- [2] D. Pratt *et al.*, "NDEx, the Network Data Exchange," *Cell Syst*, vol. 1, no. 4, pp. 302–305, 2015, doi: 10.1016/j.cels.2015.10.001.
- [3] D. Szklarczyk *et al.*, "The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest," *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, Jan 6 2023, doi: 10.1093/nar/gkac1000.
- [4] E. Holtzapple, C. A. Telmer, and N. Miskov-Zivanov, "FLUTE: Fast and reliable knowledge retrieval from biomedical literature," *Database*, vol. 2020, 2020, doi: 10.1093/database/baaa056.
- [5] A. e. a. Grattafiori, "The Llama 3 Herd of Models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>.
- [6] E. Holtzapple *et al.*, "The BioRECIPE Knowledge Representation Format," *ACS Synth Biol*, vol. 13, no. 8, pp. 2621–2624, Aug 16 2024, doi: 10.1021/acssynbio.4c00096.
- [7] M. Lab. "BioRECIPE Documentation." <https://melody-biorecipe.readthedocs.io/en/latest/> (accessed April 2025).
- [8] B. M. Gyori, C. T. Hoyt, and A. Steppi, "Gilda: biomedical entity text normalization with machine-learned disambiguation as a service," *Bioinform Adv*, vol. 2, no. 1, p. vbac034, 2022, doi: 10.1093/bioadv/vbac034.
- [9] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger, "From word models to executable models of signaling networks using automated assembly," *Molecular systems biology*, vol. 13, no. 11, pp. 954–954, 2017, doi: 10.15252/msb.20177651.
- [10] D. Turei *et al.*, "Integrated intra- and intercellular signaling knowledge for multicellular omics analysis," *Mol Syst Biol*, vol. 17, no. 3, p. e9923, Mar 2021, doi: 10.15252/msb.20209923.
- [11] P. Lo Surdo, A. Calderone, G. Cesareni, and L. Perfetto, "SIGNOR: A Database of Causal Relationships Between Biological Entities-A Short Guide to Searching and Browsing," *Curr Protoc Bioinformatics*, vol. 58, pp. 8 23 1–8 23 16, Jun 27 2017, doi: 10.1002/cpbi.28.
- [12] R. Oughtred *et al.*, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Sci*, vol. 30, no. 1, pp. 187–200, Jan 2021, doi: 10.1002/pro.3978.
- [13] I. Rodchenkov *et al.*, "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data," *Nucleic Acids Res*, vol. 48, no. D1, pp. D489–D497, Jan 8 2020, doi: 10.1093/nar/gkz946.

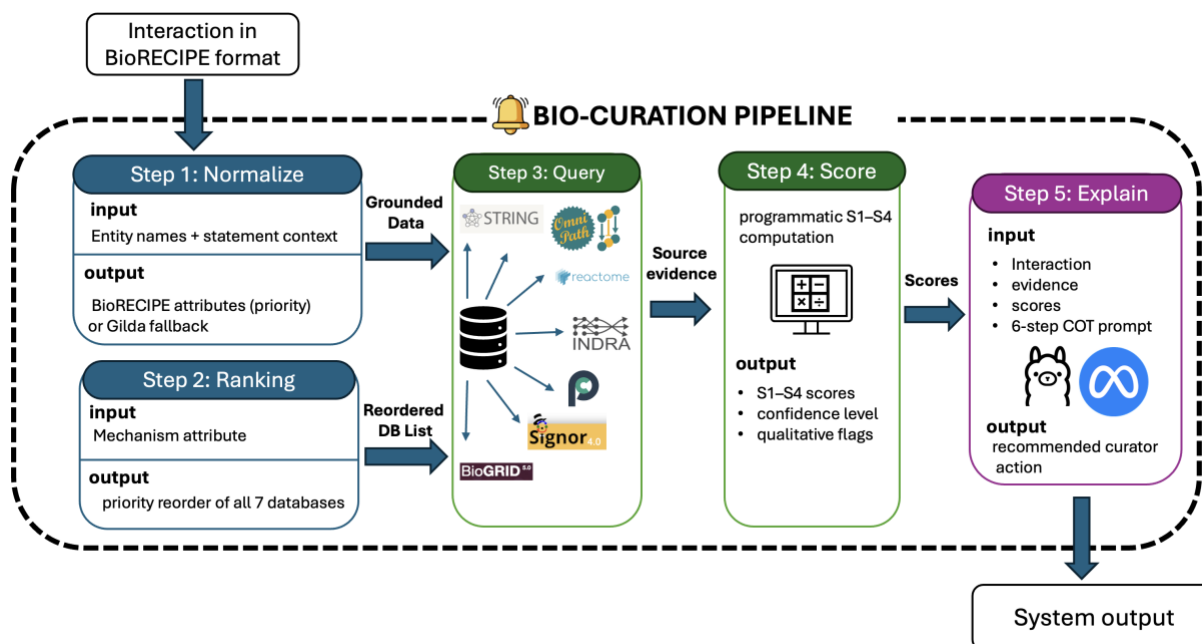


Figure 1. BELL pipeline.

Table 1. Scoring evidence.

Score	Name	Definition	Interpretation
S1	Source coverage	Weighted sum of found sources / max possible weight (14). Tier 1 (INDRA, SIGNOR) = 3 pts; Tier 2 (OmniPath, Reactome, BioGRID) = 2 pts; Tier 3 (STRING, PCnet) = 1 pt.	0 = no database corroborates; 1.0 = all databases corroborate
S2	Attribute match	Mean across found sources of (directionality + sign + mechanism) / 3. Each attribute: 1.0 = confirmed by source, 0.5 = not reported by source, 0.0 = contradicted by source.	Quality of evidence relative to declared interaction attributes
S3	INDRA belief	INDRA aggregated belief score for retrieved statements; 0.0 if INDRA returns no results.	Probabilistic confidence derived from literature frequency and source reliability
S4	Contradiction	0 = no sign conflict across sources; 2 = at least two sources disagree on sign.	Flags curator attention when databases disagree on activation vs. inhibition

```

● QUERY SENT TO LLM

=== INTERACTION ===
AKT1 -> PTEN | Sign: positive | Mechanism: Phosphorylation

Source statement: "Since 2011, several therapies have been Food and Drug Administration (FDA) approved for melanoma, but brain metastases are often the major cause of treatment failure. Melanoma patients with brain metastases have a dismal prognosis and median overall survival is only about 1 year from the time of diagnosis.1 Given this grim prognosis, effective therapeutic approaches are urgently needed for these patients. A major challenge in developing effective treatments for melanoma brain metastases has been the lack of relevant preclinical models that mimic metastatic patterns observed in patients. Using data obtained from human melanoma samples, which demonstrated increased levels of phosphorylated AKT (P-AKT) and decreased levels of phosphatase and TENsin homolog deleted on chromosome 10 (PTEN) in brain metastases.2,3,4 we generated a mouse model of melanoma with hyperactivation of AKT1 signaling that develops lung and brain metastases similar to the human disease. We used this model to delineate the mechanisms by which AKT promotes metastasis and evaluated whether this could be exploited therapeutically.5"

=== DATABASE EVIDENCE ===
[FOUND] SIGNOR score=1.00 dir=yes sign=unknown mech=confirmed | 11 SIGNOR interaction(s). Effect: down-regulates activity, up-regulates activity. Mechanism: , dephosphorylation.
[FOUND] INDRA score=0.55 dir=yes sign=unknown mech=confirmed | Belief 0.54. 39 statements. Types: Activation, Complex, DecreaseAmount.
[FOUND] STRING score=0.00 dir=unknown sign=unknown mech=unknown | STRING combined score: 0.001.
[FOUND] Reactome score=0.59 dir=unknown sign=unknown mech=confirmed | 3 shared Reactome pathway(s).
[FOUND] BioGRID score=0.60 dir=unknown sign=unknown mech=unknown | 2 interaction(s). Systems: Affinity Capture-Western, FRET. 2 publication(s).
[FOUND] PCnet score=0.55 dir=yes sign=pos mech=confirmed | 1 Pathway Commons interaction(s). Types: controls-expression-of.
[NOT FOUND]: OmniPath

=== SCORES ===
S1 (source coverage): 0.86 (HIGH)
S2 (attribute match): 0.54 (PARTIAL)
S3 (INDRA belief): 0.55 (MODERATE)
S4 (contradiction): 0 (NONE)
Flags: attribute_match_partial

=== REASONING STEPS ===
Step 1 — Source coverage (S1): How many databases found this interaction? Are high-tier sources (SIGNOR, INDRA) among them? Is S1 high (>=0.7), moderate (0.4-0.7), or low (<0.4)?
Step 2 — Attribute match (S2): Do the found databases confirm the direction, sign, and mechanism stated in the BioRECIPE-IL row? Are there gaps (unknown fields)?
Step 3 — INDRA belief (S3): Is the belief score strong (>=0.8), moderate (0.5-0.8), or weak (<0.5)? If S3=0, note that INDRA found nothing.
Step 4 — Contradictions (S4): If S4=2, sources disagree on sign — explain the conflict. If S4=0, confirm no contradictions.
Step 5 — Flags: Interpret each flag. What does it imply for curation?
Step 6 — Recommendation: Based on Steps 1-5, choose exactly one of: Accept / Accept with enrichment / Flag for review / Reject. Justify in one sentence.

```

Figure 2. Example chain-of-thought prompt sent to Llama 3.1 for the interaction AKT1→PTEN (positive, Phosphorylation).

```

● LLM RESPONSE

**Step 1 — Source coverage (S1)**:
The interaction AKT1 -> PTEN is found in multiple databases (SIGNOR, INDRA, Reactome, BioGRID, and PCnet), including high-tier sources (SIGNOR and INDRA). S1 is high (0.86), indicating good source coverage.

**Recommendation:** Accept

**Step 2 — Attribute match (S2)**:
The found databases confirm the direction and mechanism stated in the BioRECIPE-IL row, but the sign is unknown in most databases. There are gaps in the sign field, which is partially matched (0.54).

**Recommendation:** Accept with enrichment (to gather more evidence on the sign)

**Step 3 — INDRA belief (S3)**:
The INDRA belief score is moderate (0.55), indicating a reasonable level of confidence in the interaction.

**Recommendation:** Accept (no change)

**Step 4 — Contradictions (S4)**:
There are no contradictions (S4=0), confirming that the sources agree on the interaction.

**Recommendation:** Accept (no change)

**Step 5 — Flags**:
The attribute_match_partial flag implies that the sign field is not fully matched, which may indicate a need for further evidence or clarification.

**Recommendation:** Accept with enrichment (to gather more evidence on the sign)

**Step 6 — Recommendation**:
Based on the analysis, the interaction AKT1 -> PTEN is well-supported by multiple databases, including high-tier sources. Although there are gaps in the sign field, the overall evidence is strong enough to accept the interaction with enrichment to gather more evidence on the sign.

**Recommendation:** Accept with enrichment

```

Figure 3. Example of prompt response in which LLM suggests accepting with enrichment for AKT1→PTEN interaction.

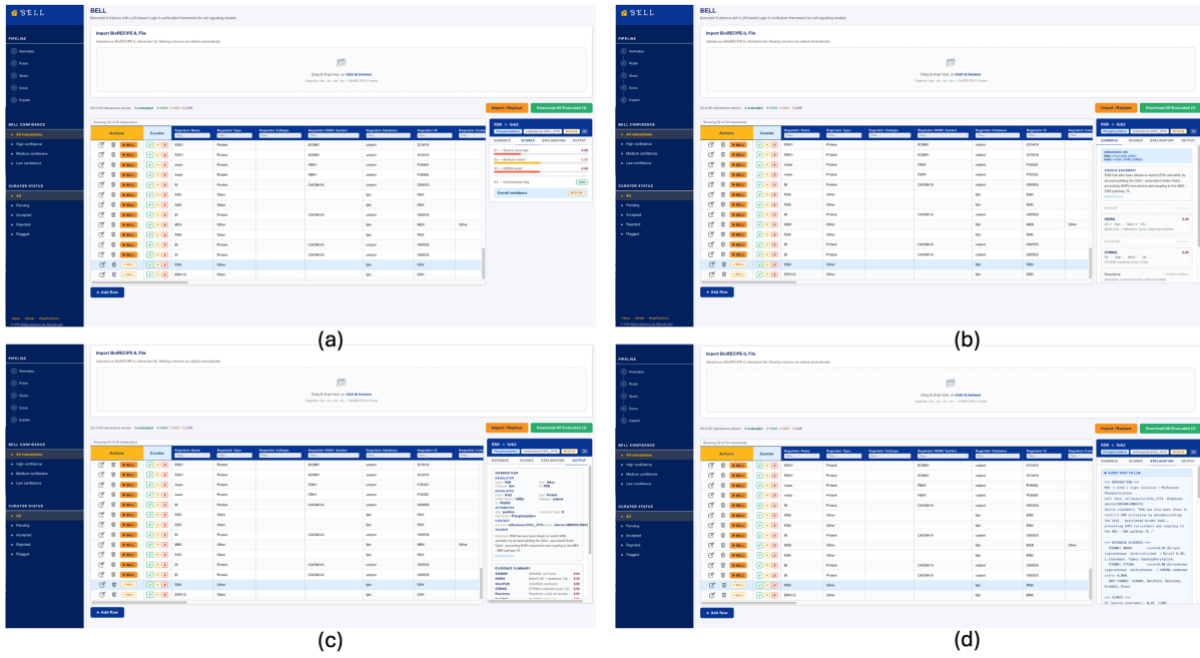


Figure 4. BELL curator interface. (a) Scores tab showing  $S1$ – $S4$  metrics and overall confidence level for a selected interaction. (b) Evidence tab displaying per-database query results retrieved from the seven integrated biological databases. (c) Output tab summarizing the full interaction record, evidence, scores, and curator status. (d) Explanation tab showing the Llama 3.1 chain-of-thought reasoning and recommended curator action. The table supports row-level curator annotation (Accept, Flag, or Reject), inline editing and deletion of individual interactions, and batch export of all evaluated interactions as a downloadable spreadsheet.

# VIOLIN-WEB: a web application for reconciliation and curation of molecular interaction graphs

Haomiao Luo<sup>1</sup>  
Niloofar Arazkhani<sup>1</sup>

Difei Tang

Natasa Miskov-Zivanov<sup>2</sup>

{hal357, nia129, dit18, nmzivanov}@pitt.edu  
University of Pittsburgh, Pittsburgh, PA

## 1 INTRODUCTION

Current approaches to updating mechanistic models<sup>1-3</sup> typically rely on manual curation or ad hoc filtering rules<sup>4, 5</sup>. While database-backed verification tools can remove unsupported interactions, they do not formally distinguish structural corroboration from contextual disagreement, nor do they expose alternative semantic interpretations of mismatches<sup>6-10</sup>. Conversely, literature extraction tools can be used for generating candidate interactions but do not provide mechanisms for evaluating those interactions relative to a target knowledge graph or executable model<sup>11-13</sup>. The absence of a configurable, attribute-aware reconciliation engine limits the scalability and reproducibility of literature-to-model workflows.

To mitigate this problem, we recently introduced VIOLIN (Versatile Interaction Organizing to Leverage Information in Networks)<sup>14</sup>, a configurable, attribute-aware reconciliation framework that compares newly extracted interaction lists against structured baseline graphs. To facilitate the use of VIOLIN by domain experts and enable interpretations of its graph-based reconciliation results, we developed a web-based user interface. The main contributions of this work are: 1) web-based integration of VIOLIN including file upload, parameter configuration, and interactive visualization of reconciliation results on the baseline graph; 2) two matching metrics that address false negatives inherent to the original exact-matching (EM) criterion; 3) user-adjustable threshold that lets users control the precision-recall trade-off directly from the interface. The interface of VIOLIN can be accessed through (<https://boheme.pitt.edu/violin>)

## 2 VIOLIN FUNCTIONALITY

VIOLIN reconciles a baseline mechanistic knowledge graph with an interaction graph, which is extracted or retrieved from knowledge sources such as literature and databases. VIOLIN is developed as a decision tree that classifies the

interactions within the interaction graph into four categories, relative to the baseline graph: corroboration, contradiction, flagged, and extension. We also defined several sub-categories considering the path, directionality, sign, physical contact, and additional attributes such as cell line, cell type, and entity location.

## 3 INTERACTIVE WEB INTERFACE

**Compatibility with existing curation pipelines.** VIOLIN provides an extensible reconciliation framework as a modular tool compatible with diverse extraction pipelines and data formats. This modularity is enabled by the translator interface for converting between the BioRECIPE and other commonly used formats<sup>15</sup> (Figure 1). The interface converts the output from diverse data acquisition pipelines, including large language models (LLMs) with in-context learning, rule-based NLP systems<sup>11</sup>, and curated knowledge bases such as INDRA<sup>13</sup> and PathwayCommons<sup>3</sup>.

**Configurability and modeling objectives.** Another central feature of the framework is its configurability. Through the control panel (Figure 2), users can select different attributes to be included in the comparison of interaction context, thereby tailoring the reconciliation criteria for specific modeling objectives. For instance, inclusion of contextual attributes may be critical in highly specific cell-type models, whereas broader pathway validation may rely primarily on structural sign and direction agreement.

**Visual interpretation of results.** This interface features a two-window layout: one displays the interaction list, while the other window shows the baseline graph topology (Figure 2). The interface allows users to review the extracted interactions, their classifications, and the mapping of classified interactions onto the baseline graph.

**Entity match confidence.** Integrating results from heterogeneous resources introduce several challenges: cross-referencing conflicts, ID misalignment, and extraction errors. We perform entity matching based on three attributes: ID,

---

<sup>1</sup> Both authors contributed equally to this research.

<sup>2</sup> Corresponding author

type, and name. Since the accuracy and computational speed of entity matching are often competing objectives, the interface offers multiple metrics to calculate similarity  $s_{A,B}$  (Table 1). Specifically, ‘Jaccard’ similarity enables faster processing with moderate precision, whereas ‘Edit\_sim’ provides more precise results at the cost of slower processing speed. Users can also adjust the threshold to control the matching rigor, so that the match condition could be expressed as a function:

$$\text{Match}(A, B) = \begin{cases} 1, & \text{if } s_{A,B} > \delta \\ 0, & \text{otherwise} \end{cases}$$

$A$  and  $B$  are entity identifiers (names, IDs, or HGNC symbols) from the interaction graph and baseline graph, respectively. ‘1’ indicates  $A$  and  $B$  are matched, while ‘0’ means they are different.  $\delta$  represents a user-defined threshold between 0 and 1. A higher value for  $\delta$  sets stricter requirements for string pattern similarity, ensuring higher precision in entity matching. The overall entity matching score is calculated as the product of metrics score for regulated and regulator elements. For example, extension interactions are assigned scores of zero due to complete mismatches in one or both entity names.

**Application runtime.** We benchmarked the responsiveness of the interface using public web services, with two model files and two interactions list serving as the inputs to test the interactions processed per second. The results are reported in Table 2.

**Implementations.** VIOLIN is implemented as a Python package with dependencies on Pandas, Numpy, and NetworkX. The framework operates on BioRECIPE-formatted interaction files and supports configurable classification schemes and attribute inclusion strategies. Path search within baseline graph is performed using NetworkX API. The architecture of the interface is based on a client-server model. Specifically, the frontend server is implemented with React framework, and backend server is built with FastAPI to serve the VIOLIN endpoints. Both frontend and backend servers are hosted on an AWS EC2 instance for online usage.

## 4 CONCLUSION

VIOLIN is a configurable, attribute-aware reconciliation framework for integrating heterogeneous, literature-derived interaction graphs with curated baseline knowledge graphs. We developed a user-friendly interface to enhance accessibility for the systems and synthetic biology community. As automated extraction technologies continue to expand the volume of structured interaction data, reconciliation frameworks such as VIOLIN will play a critical role in maintaining coherence between evolving literature and curated knowledge representations.

## 5 ACKNOWLEDGEMENTS

This research was funded by National Science Foundation (NSF) EAGER award CCF-2324742 and CAREER award CCF-2442884.

## 6 REFERENCES

- (1) Malik-Sheriff, R. S.; Glont, M.; Nguyen, T. V.; Tiwari, K.; Roberts, M. G.; Xavier, A.; Vu, M. T.; Men, J.; Maire, M.; Kananathan, S. BioModels—15 years of sharing computational models in life science. *Nucleic acids research* **2020**, *48* (D1), D407-D415.
- (2) Milacic, M.; Beavers, D.; Conley, P.; Gong, C.; Gillespie, M.; Griss, J.; Haw, R.; Jassal, B.; Matthews, L.; May, B. The reactome pathway knowledgebase 2024. *Nucleic acids research* **2024**, *52* (D1), D672-D678.
- (3) Cerami, E. G.; Gross, B. E.; Demir, E.; Rodchenkov, I.; Babur, Ö.; Anwar, N.; Schultz, N.; Bader, G. D.; Sander, C. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research* **2010**, *39* (suppl\_1), D685-D690.
- (4) Holtzapple, E.; Cochran, B.; Miskov-Zivanov, N. Context-aware query design combines knowledge and data for efficient reading and reasoning. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021; pp 238-246.
- (5) Holtzapple, E.; Telmer, C. A.; Miskov-Zivanov, N. FLUTE: Fast and reliable knowledge retrieval from biomedical literature. *Database* **2020**, *2020*. DOI: 10.1093/database/baaa056.
- (6) Hoyt, C. T.; Domingo-Fernández, D.; Aldisi, R.; Xu, L.; Kolpeja, K.; Spalek, S.; Wollert, E.; Bachman, J.; Gyori, B. M.; Greene, P.; et al. Re-curation and rational enrichment of knowledge graphs in Biological Expression Language. *Database: The Journal of Biological Databases and Curation* **2019**, *2019*. DOI: 10.1093/database/baz068 (accessed 2021/06/01/16:23:12). From PubMed Central.
- (7) Ahmed, Y.; Telmer, C. A.; Zhou, G.; Miskov-Zivanov, N. Context-aware knowledge selection and reliable model recommendation with ACCORDION. *Frontiers in Systems Biology* **2024**, *4*, 1308292.
- (8) Sayed, K.; Bocan, K. N.; Miskov-Zivanov, N. Automated Extension of Cell Signaling Models with Genetic Algorithm. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018/07//, 2018; IEEE: Honolulu, HI, pp 5030-5033. DOI: 10.1109/EMBC.2018.8513431.
- (9) Ahmed, Y.; Butchy, A. A.; Sayed, K.; Telmer, C.; Miskov-Zivanov, N. New advances in the automation of context-aware information selection and guided model assembly. *arXiv preprint arXiv:2110.10841* **2021**.
- (10) Liang, K.-W.; Wang, Q.; Telmer, C.; Ravichandran, D.; Spirtes, P.; Miskov-Zivanov, N. Methods to Expand Cell Signaling Models Using Automated Reading and Model Checking. In *Computational Methods in Systems Biology*, Feret, J., Koepl, H. Eds.; Vol. 10545; Springer International Publishing, 2017; pp 145-159.
- (11) Valenzuela-Escárcega, M. A.; Hahn-Powell, G.; Surdeanu, M.; Hicks, T. A domain-independent rule-based framework for event extraction. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, 2015; pp 127-132.
- (12) Novichkova, S.; Egorov, S.; Daraselia, N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* **2003**, *19* (13), 1699-1706. DOI: 10.1093/bioinformatics/btg207 (accessed 2021/06/01/12:19:10). From DOI.org (Crossref).
- (13) Gyori, B. M.; Bachman, J. A.; Subramanian, K.; Muhlich, J. L.; Galescu, L.; Sorgger, P. K. From word models to executable models of signaling networks using automated assembly. *Molecular systems biology* **2017**, *13* (11), 954.
- (14) Luo, H.; Hansen, C.; Arazkhani, N.; Telmer, C. A.; Tang, D.; Zhou, G.; Spirtes, P.; Miskov-Zivanov, N. VIOLIN: A modular framework for scalable reconciliation of heterogeneous interaction graphs. *bioRxiv: the preprint server for biology* **2026**, 2024.2007. 2021.604448.
- (15) Holtzapple, E.; Zhou, G.; Luo, H.; Tang, D.; Arazkhani, N.; Hansen, C.; Telmer, C. A.; Miskov-Zivanov, N. The BioRECIPE knowledge representation format. *ACS Synthetic Biology* **2024**, *13* (8), 2621-26

## 7 FIGURES AND TABLES

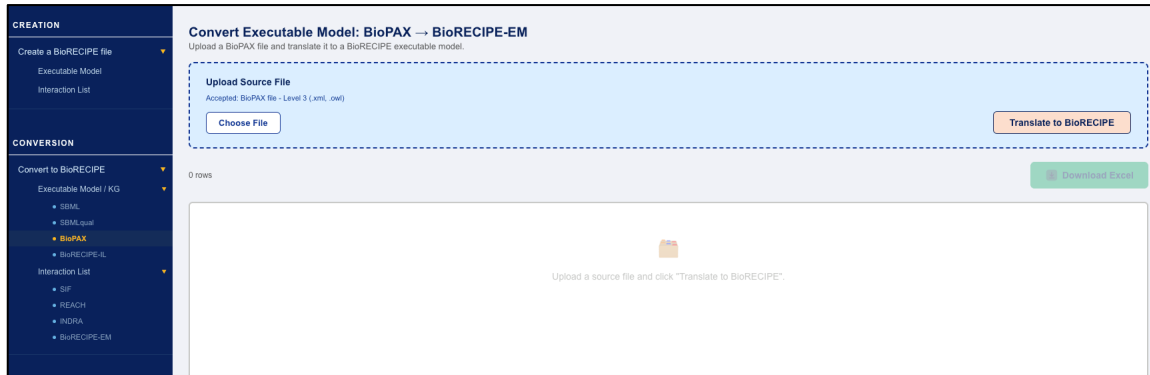


Figure 1 The BioRECIPE translators' interface

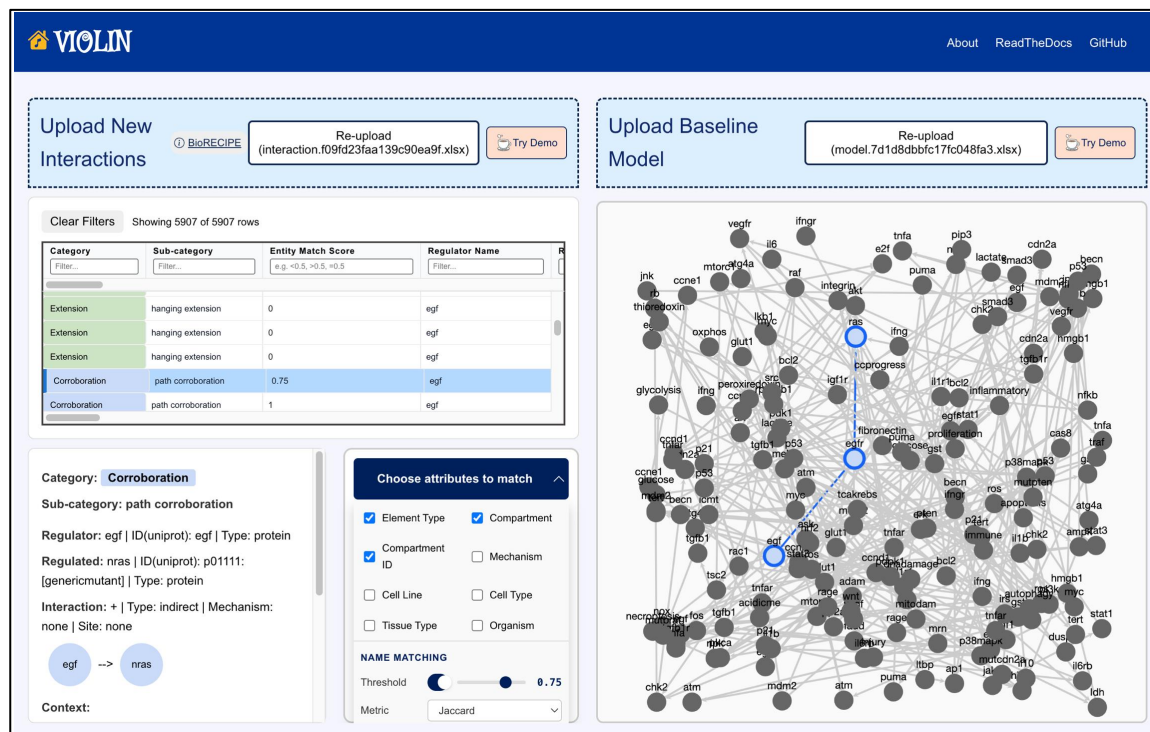


Figure 2. Classification visualization from interaction list to baseline graph.

Table 1 The metrics used to assess the matching condition.

Option	Score
'Jaccard'	$s_{A,B} = \frac{ A \cap B }{ A \cup B }$
'Edit_sim'	$s_{A,B} = 1 - \frac{D(A,B)}{\max(L_A, L_B)}$

Note:  $D(\cdot, \cdot)$  denotes the Lenvenshtein distance, while  $L_A$  and  $L_B$  represent the number of characters for  $A$  and  $B$ .

Table 2 The runtime of VIOLIN testing on real-time web service. In the “No. interactions” column, the value on the left indicates the number of interactions in interaction list and the value on the right denotes the number of edges in the baseline graph. In the “Metric” column, Edit\_sim@0.9 and Jaccard@0.9 indicate that the threshold of 0.9 is applied to the corresponding metrics. Each test was repeated twice, and the “Reconciliation time” column reports the average runtime for reconciliation.

Test scenarios	No. interactions	Metric	Reconciliation time (sec)	No. interactions/sec
<b>A</b>	5907/266	Edit_sim@0.9	>60	-
	5907/266	Jaccard@0.9	46.21	127.82
	5907/266	Exact Match	30.53	193.48
<b>B</b>	1124/70	Edit_sim@0.9	19.52	57.58
	1124/70	Jaccard@0.9	4.21	266.98
	1124/70	Exact Match	3.53	318.41

# Myriad: A GPU-Accelerated Platform for In-Silico Prototyping of Closed-Loop Single-Cell Experiments

**Robin Henry**  
University of Oxford  
Oxford, United Kingdom  
robin.henry@eng.ox.ac.uk

**Harrison Steel**  
University of Oxford  
Oxford, United Kingdom  
harrison.steel@eng.ox.ac.uk

**Jean-Baptiste Lugagne**  
University of Oxford  
Oxford, United Kingdom  
jean-baptiste.lugagne@eng.ox.ac.uk

## 1 INTRODUCTION

Single-cell microfluidic platforms, such as mother machines, combined with per-cell optogenetic actuation now enable simultaneous observation and independent actuation of up to  $N = 10^6$  individual cells in real time [1, 6]. Because each cell follows its own stochastic trajectory under an independently chosen input, this effectively replaces sequential experimental iterations with  $N$  parallel experiments, compressing wall-clock time. Our recent work has begun to explore this *in-silico*, with reinforcement learning (RL) agents that learn control policies for bistable gene circuits in hours rather than weeks of simulated lab time [3], and active feedback control used to resolve fundamental identifiability bottlenecks in the automated characterization of multi-stable systems [4].

In developing these methods, we encountered bottlenecks that slow progress across the field. Single-cell control, system identification, and optimal experimental design sit at the intersection of biology, machine learning, statistics, and control theory, and would interest researchers in all four communities. However, the barrier to entry is high: working with stochastic gene circuit models requires domain-specific knowledge, substantial boilerplate code for simulation and experiment tracking, and no standardized benchmarks exist at the scale of mother-machine experiments. Algorithms in RL and control also rarely consider  $N \sim 10^6$  independent systems or treat wall-clock time rather than sample count as the primary cost, so existing tools are not designed to exploit GPU acceleration at this scale.

We developed Myriad<sup>1</sup>, a JAX-native platform for massively parallel stochastic simulation of single-cell experiments, to lower these barriers and enable *in-silico* validation of algorithms before lab deployment. Myriad provides exact Gillespie SSA simulation of up to  $10^6$  parallel cells in a single compiled GPU kernel, a layered environment abstraction that separates physics, task definition, and algorithm implementation, and handles GPU optimization, training loops, hyperparameter tracking, and logging out of the box (Fig. 1). The platform grew out of our experience building RL-based controllers for bistable gene circuits [3], where we

initially used Gymnax [5] but found it difficult to adapt to the specifics of stochastic biological simulations at this scale.

## 2 PLATFORM DESIGN

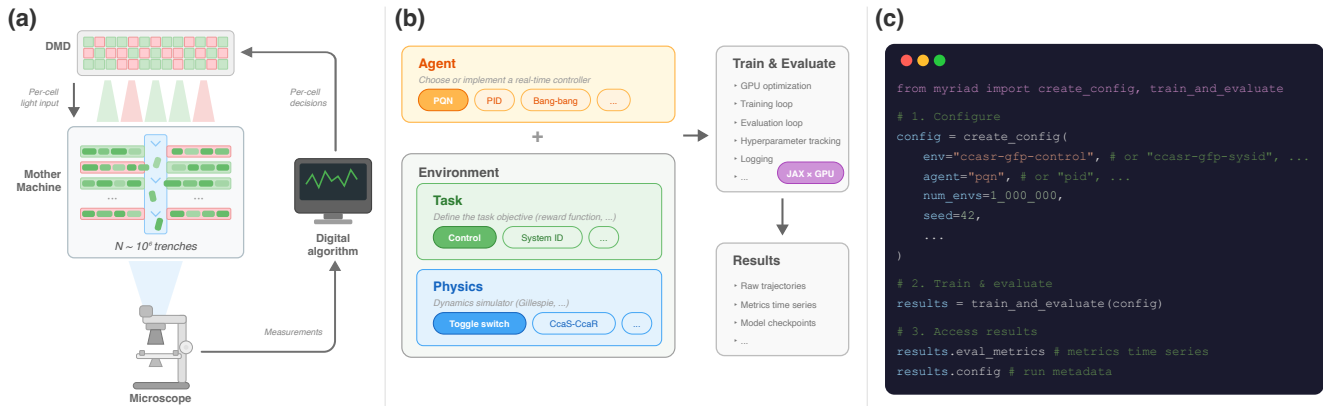
Myriad’s architecture lets algorithm developers work on biological control and inference problems without reimplementing simulation infrastructure. The platform separates the physics (stochastic dynamics of gene circuits), the task (the objective: control the behavior of cells, infer their parameters, or learn a new model from data), and the algorithm (classical controller, RL agent, or inference method) into independent layers. This separation means, for example, that a new control algorithm can be tested on an existing task with few modifications, and a new biological system can be added without rewriting the training loop (Fig. 1b). Myriad implements stochastic dynamics via the Gillespie algorithm, preserving the discrete, asynchronous molecular events that drive qualitative behavior in gene circuits such as bistable switching and rare transitions between cell states. All simulation and training code is vectorized across cells via JAX, so that  $10^6$  independent trajectories run in a single GPU kernel within.

At the time of writing, our flagship environment models a CcaS-CcaR optogenetic circuit in *E. coli*, a bistable self-activating system controllable by green/red light input (Fig. 1a). This circuit has been used in real mother-machine experiments [6] and has well-characterized kinetics [3], making it a natural starting point. We defined two tasks on the same physics: a control task, where an agent is asked to steer GFP expression to track a setpoint, and a system identification task, where kinetic parameter distributions must be inferred from observed trajectories of a heterogeneous cell population.

## 3 CASE STUDIES AND ONGOING WORK

In a recent work, we validated Myriad’s control task on the CcaS-CcaR circuit [3]. Parallelised Q-Networks (PQN) [2], a recent RL algorithm designed for massively parallel training, learn to track fluorescence setpoints in the bistable regime, outperforming bang-bang, PI, and deep model-predictive controllers. We showed that PQN converges in approximately 7 hours of simulated lab time with  $2^{16}$  parallel cells, fast enough

<sup>1</sup><https://github.com/robinhenry/myriad-jax>



**Figure 1: Overview of the Myriad platform. (a) Closed-loop experimental setup: a digital micromirror device (DMD) projects per-cell light patterns onto a mother-machine microfluidic chip containing up to  $10^6$  cells; a microscope captures fluorescence measurements; a digital algorithm processes observations and selects the next light input for each cell. (b) Myriad’s modular architecture: users select an agent and an environment (task + physics), and Myriad handles GPU-accelerated training, evaluation, and logging. (c) Python API: a complete experiment requires only a few lines of code, with environments, algorithms, and custom configuration options easily swappable.**

to be practical for real-time deployment. Separately, we used Myriad to study system identification on a genetic toggle switch and showed that passive observation cannot identify key parameters of multi-stable circuits regardless of population size, because cells settle into stable states and never visit the information transition regions between them [4].

We are now using Myriad to tackle a harder and largely unexplored question: can active algorithms recover regulatory network structure, not just parameters, within a single mother-machine experiment? Starting from the CcaS-CcaR circuit, we are investigating a progression of inference tasks: recovering kinetic parameters for single cells, characterizing parameter distributions across heterogeneous populations, and inferring circuit topology without assuming a fixed model structure. Myriad’s layered architecture makes it easy to swap inference algorithms and input-design strategies at each stage, allowing rapid comparison of approaches.

## 4 OUTLOOK

We have open-sourced Myriad and continue to develop it in public. The algorithms we validate *in-silico* are intended for deployment on real mother-machine experiments in our lab, and we hope the platform will also serve as an entry point for researchers in RL, control, and Bayesian inference who want to work on automated characterization and design of biological systems. PQN is one example of an algorithm designed for GPU parallelism that turns out to suit this setting well, but the space of algorithms adapted to this regime is largely unexplored. Standardized biological benchmarks at experimental scale, with clear metrics tied to wall-clock time,

could catalyze the same kind of cross-disciplinary progress that game and robotics benchmarks brought to RL.

## REFERENCES

- [1] EATON, D. S., SÁNCHEZ, C., GUTIÉRREZ-LÓPEZ, L., SHENKER, J. Q., GOULEV, Y., WATSON, B. R., HENRIOT, V., GARNER, E. C., JUSTMAN, Q. A., MOFFITT, J. R., AND PAULSSON, J. Essentialome-wide multigenerational imaging reveals mechanistic origins of cell growth laws. *bioRxiv* (2025).
- [2] GALLICI, M., FELLOWS, M., ELLIS, B., POU, B., MASMITJA, I., FOERSTER, J. N., AND MARTIN, M. Simplifying deep temporal difference learning. In *International Conference on Learning Representations (ICLR)* (2025).
- [3] HENRY, R., AND LUGAGNE, J.-B. Control of a bi-stable genetic system via parallelized reinforcement learning. In *2025 IEEE 64th Conference on Decision and Control (CDC)* (2025), IEEE, pp. 2898–2904.
- [4] HENRY, R., PEARSON, J., STEEL, H., AND LUGAGNE, J.-B. Active control for system identification of multi-stable genetic circuits. Submitted to the 2026 IEEE Conference on Decision and Control (CDC), 2026.
- [5] LANGE, R. T. gymnax: A JAX-based reinforcement learning environment library, 2022.
- [6] LUGAGNE, J.-B., BLASSICK, C. M., AND DUNLOP, M. J. Deep model predictive control of gene expression in thousands of single cells. *Nature Communications* 15 (2024), 2148.

# CELESTA Web: An Interactive Tool for LLM-Enhanced Biomedical Context Annotation

Difei Tang  
University of Pittsburgh  
dit18@pitt.edu

Natasa Miskov-Zivanov  
University of Pittsburgh  
nmzivanov@pitt.edu

## 1 Introduction

Understanding the relationships between biological entities such as genes, proteins, and chemicals is fundamental to studying complex biological processes. While natural language processing (NLP) methods have achieved notable success in biomedical relation extraction (RE), extracted relations frequently lack contextual information, including cell type, cell line, intracellular location, disease, and organism. This contextual information is critical for interpreting the biological significance of extracted interactions, as the same molecular interaction can behave differently across cellular environments [1].

CELESTA (Context Extraction through LEarning with Semi-supervised multi-Task Architecture) was developed to address this gap, using a fine-tuned BioBERT model with semi-supervised multi-task learning and out-of-distribution (OOD) detection to classify biomedical contexts from textual evidence [1]. While CELESTA achieves strong classification performance, it operates as a command-line tool that requires users to prepare structured inputs and interpret raw model outputs. This limits its accessibility for biomedical researchers who need to annotate interaction lists interactively. Meanwhile, large language models (LLMs) have demonstrated strong capabilities in biomedical entity and relation extraction [2], but lack the domain-specific fine-tuning that CELESTA provides for context classification. Neither approach alone offers a complete annotation solution.

In this work, we present CELESTA Web, an interactive web-based tool that combines an LLM-based relation extraction pipeline with CELESTA's fine-tuned context classifier in a unified annotation workflow. Users can input biomedical text or upload BioRECIPE interaction lists [3], and the tool produces structured annotations with entity highlights, relation arcs, and context labels, all validated against predefined biomedical taxonomies.

## 2 METHODOLOGY

CELESTA Web is built as a web application with a React frontend and a Python (FastAPI) backend, integrated into

the DySE (Dynamic System Explanation) framework [4]. The annotation pipeline operates in a hybrid mode that leverages both an LLM and the CELESTA classifier, as illustrated in Figure 1. The pipeline processes each interaction statement through the following steps:

**Step 1: Entity and Relation Extraction.** Given a biomedical text, the system first determines whether entity and relation information is already available (e.g., from an uploaded BioRECIPE interaction list). If entities and relations are present, the LLM refines and verifies the existing annotations. If absent, the LLM performs full extraction, identifying regulator and regulated entities, their types (e.g., protein, gene, chemical), and the interaction relation (e.g., Activation, Inhibition, Phosphorylation).

**Step 2: Context Classification with CELESTA.** The text is passed to the CELESTA backend, which runs five context classification tasks in parallel: location, cell line, cell type, organ, and disease. For each task, CELESTA first applies an OOD detector based on Maximum Softmax Probability (MSP) [5] to determine whether the text contains an in-distribution context for that category. If the text is classified as in-distribution, the fine-tuned BioBERT model predicts the context label from a predefined taxonomy.

**Step 3: LLM Verification and Trigger Identification.** The LLM verifies each CELESTA prediction. For in-distribution predictions, the LLM checks whether the predicted label is consistent with the text and identifies whether the context is explicit (a trigger word appears in the text) or implicit (inferred from domain knowledge). For OOD predictions (i.e., CELESTA reports no known context), the LLM checks whether any trigger words in the text suggest a context that CELESTA's taxonomy may not cover.

**Step 4: Validation and Output.** All outputs are validated against CELESTA's predefined label taxonomy. Invalid labels (those not in the allowed list) are discarded. The final annotation includes entity spans with types, a relation label, and a list of context labels classified as explicit with trigger word positions or implicit. Results are mapped to BioRECIPE format [3], a standardized knowledge representation for biological interactions that

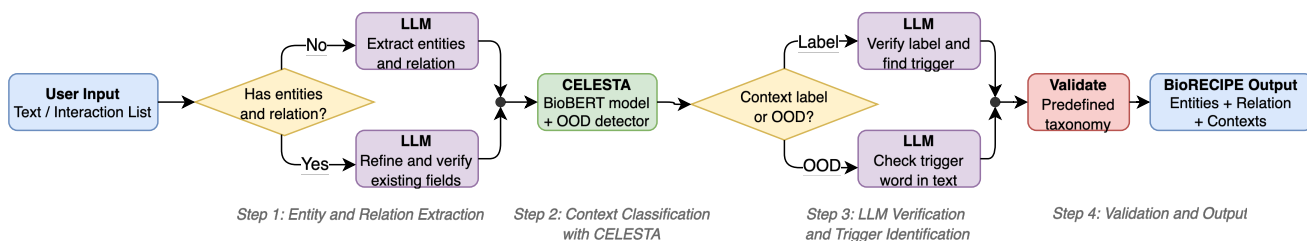


Figure 1: Overview of the CELESTA Web annotation pipeline with color-coded components: LLM (purple), CELESTA (green), I/O (blue), and validation (red).

captures entities, mechanisms, and contextual attributes. Specifically, relation maps to Mechanism and Sign, location maps to Regulator/Regulated Compartment, and other contexts map to their corresponding columns.

### 3 User Interface

The tool provides an interface where users can upload a BioRECIPE interaction list, view the interaction table with editable columns, and select individual rows for annotation. Annotation results are displayed as text annotation cards alongside the table, using brat-style visualization [6] with entity type tags, relation arcs, and context annotation indicators. Users can toggle column visibility, save changes, and export the final table in the BioRECIPE interaction list format.

### 4 Results and Discussion

CELESTA Web provides an end-to-end annotation workflow that combines the strengths of two complementary approaches: LLMs for flexible entity and relation extraction, and CELESTA for accurate, taxonomy-constrained context classification. The hybrid pipeline addresses key limitations of each approach used independently: LLMs may generate context labels outside predefined taxonomies, while CELESTA alone cannot perform relation extraction or identify trigger words for explicit contexts. Figure 2 demonstrates the web interface applied to a set of biomedical interaction statements. The left panel displays the BioRECIPE interaction table with context columns automatically populated and modified cells highlighted. The right panel shows the corresponding text annotation cards, where entities are highlighted with type labels, relation arcs connect regulator and regulated entities, and context annotations are displayed as colored indicators. Explicit contexts, where a trigger word is identified in the text (e.g., "nucleus" for location, "THP-1 cells" for cell line), are shown as inline highlights with labels. Implicit

contexts inferred from domain knowledge (e.g., macrophage and Homo sapiens from the mention of THP-1 cells) are displayed below the text with parenthetical markers.

### 5 Future Directions

A current limitation is that context labels are constrained to CELESTA's predefined taxonomy, which may not cover all relevant categories. We plan to address this by enabling interactive curation through the web interface, where domain experts can correct predictions and contribute new annotations. These curated annotations can expand the training data for CELESTA's classifiers, supporting the addition of new context types. We also plan to further integrate CELESTA Web with other DySE framework tools to enable end-to-end workflows from literature mining to model construction.

### REFERENCES

1. Tang, D., et al., An open-set semi-supervised multi-task learning framework for context classification in biomedical texts. *Journal of Biomedical Informatics*, 2025: p. 104886.
2. Jahan, I., M.T.R. Laskar, C. Peng, and J.X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 2024. 171: p. 108189.
3. Holtzapple, E., et al., The BioRECIPE knowledge representation format. *ACS Synthetic Biology*, 2024. 13(8): p. 2621-2624.
4. Telmer, C.A., et al. Dynamic system explanation: DySE, a framework that evolves to reason about complex systems-lessons learned. in *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*. 2019.
5. Hendrycks, D. and K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
6. Stenetorp, P., et al. BRAT: a web-based tool for NLP-assisted text annotation. in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012.

CELESTA About ReadTheDocs GitHub

Annotate All Annotate Selected (1) Upload .xlsx Export

Statements	Regulator Name	Regulator Compartment	Regulated Name	Regulated Compartment	Sign	Mechanism	Cell Line	Cell Type	Organism	Disease
1	Protein levels of RFX5 inc	IFN-gamma	RFX5	Cell Nucleus	pos	Activation		Homo sa		
2	Whereas hyperglycemia an	angiotensin II	MIF	Cell Nucleus	pos	Activation	THP-1 c	macroph	Homo sa	
3	activation of AKT/PKB by v	CD28	AKT	Cell Membrane	pos	Activation				
4	We and others have shown	OSM	ectoxin-1	Cell Membrane	pos	Activation		fibroblat	Mus mus	
5	experimental emphysema c	IL-13	MMP-9	Cell Membrane	pos	Activation			Homo sa	lung ca
6	TIMP-3 can be anti-angioge	TIMP3	angiogenesis		neg	NONE				
7	Treatment with L-arginine e	L-arginine	HMOX1		pos	NONE				
8	We reported that OxLDL sti	Oxidized Low t	NFAT		pos	NONE				
9	It has also been demonstra	TERF2	cell growth		pos	NONE				
10	Moreover, risk factors for th	Hypertension	Atherosclerosis		pos	NONE				
11	Due to the protection again	CCR2	Atherosclerosis		pos	NONE				
12	pilavastatin, a novel HMG e	RHGA	APOB		pos	NONE				
13	The stimulatory effect of nic	alpha-Bungaro	nicotine		pos	NONE				
14	Coumarin was used as a m	coumarin	Sic27a2		pos	NONE				
15	Increased production of pei	PP2A Complex	Capillary Perme		pos	NONE				
16	Transient exposure to FGF	FGF2	cellular extra		pos	NONE				
17	dietary intake of PUFAs in I	polyunsaturate	F3		pos	NONE				
18	Similar effects of cdk1 depl	CDK1	cisplatin		pos	NONE				
19	Oxidized low-density lipoo	Oxidized Low t	TFPI		neg	NONE				
20	VEGFR1 does appear to m	FLT1	chemotaxis		pos	NONE				
21	Human cardiac microvascu	IL18	NFkappaB		pos	NONE				
22	Generation of reactive oxyg	serum glucose	NFkappaB		pos	NONE				
23	Nitric oxide (NO) has a key	vasodilation	angiogenesis		pos	NONE				
24	IL4Ralpha expression was	TNF	IL4R		pos	NONE				
25	annexin V-propidium iodidi	Tlx4	Apoptosis		neg	NONE				
26	ectopic expression of MyoC	MYOY1	myoblast differ		pos	NONE				
27	The reduced lipid content i	lipid	Insulin Resist		pos	NONE				
28	in rat hepatocyte primary c	l6	Orn1		pos	NONE				
29	Evidence indicates that sut	response to ox	Telomere		neg	NONE				

Quick Workspace

**Text Annotation** regulator regulated location cell line cell type organ disease species

#1  

 Protein levels of RFX5 increase in the nucleus during IFN-gamma treatment.

#2  

 Whereas hyperglycemia and insulin had no effect on MIF expression, angiotensin II induced MIF secretion in THP-1 cells via the angiotensin 1- receptor [58].

#3  

 activation of AKT/PKB by various growth and survival factors, including TCR and CD28, involves a phosphatidylinositol 3-kinase (PI3K)-dependent membrane translocation of AKT/PKB

#4  

 We and others have shown that OSM can induce ectoxin-1 in mouse lung fibroblasts (MLF) (28) and lung smooth muscle cells (29).

#5  

 experimental emphysema caused by lung overexpression of IL-13 or deletion of surfactant protein D involved both MMP-9 and MMP-12 (173, 348)

#6



CELESTA About ReadTheDocs GitHub

Annotate All Annotate Selected (1) Upload .xlsx Save Export

Statements	Regulator Name	Regulator Compartment	Regulated Name	Regulated Compartment	Sign	Mechanism	Cell Line	Cell Type	Organism	Disease
1	Protein levels of RFX5 inc	IFN-gamma	RFX5	Cell Nucleus	pos	Activation		Homo sa		
2	Whereas hyperglycemia an	angiotensin II	MIF	Cell Nucleus	pos	Activation	THP-1 c	macroph	Homo sa	
3	activation of AKT/PKB by v	CD28	AKT	Cell Membrane	pos	Activation				
4	We and others have shown	OSM	ectoxin-1	Cell Membrane	pos	Activation		fibroblat	Mus mus	
5	experimental emphysema c	IL-13	MMP-9	Cell Membrane	pos	Activation			Homo sa	lung ca
6	TIMP-3 can be anti-angioge	TIMP3	angiogenesis		neg	NONE				
7	Treatment with L-arginine e	L-arginine	HMOX1		pos	NONE				
8	We reported that OxLDL sti	Oxidized Low t	NFAT		pos	NONE				
9	It has also been demonstra	TERF2	cell growth		pos	NONE				
10	Moreover, risk factors for th	Hypertension	Atherosclerosis		pos	NONE				
11	Due to the protection again	CCR2	Atherosclerosis		pos	NONE				
12	pilavastatin, a novel HMG e	RHGA	APOB		pos	NONE				
13	The stimulatory effect of nic	alpha-Bungaro	nicotine		pos	NONE				
14	Coumarin was used as a m	coumarin	Sic27a2		pos	NONE				
15	Increased production of pei	PP2A Complex	Capillary Perme		pos	NONE				
16	Transient exposure to FGF	FGF2	cellular extra		pos	NONE				
17	dietary intake of PUFAs in I	polyunsaturate	F3		pos	NONE				
18	Similar effects of cdk1 depl	CDK1	cisplatin		pos	NONE				
19	Oxidized low-density lipoo	Oxidized Low t	TFPI		neg	NONE				
20	VEGFR1 does appear to m	FLT1	chemotaxis		pos	NONE				
21	Human cardiac microvascu	IL18	NFkappaB		pos	NONE				
22	Generation of reactive oxyg	serum glucose	NFkappaB		pos	NONE				
23	Nitric oxide (NO) has a key	vasodilation	angiogenesis		pos	NONE				
24	IL4Ralpha expression was	TNF	IL4R		pos	NONE				
25	annexin V-propidium iodidi	Tlx4	Apoptosis		neg	NONE				
26	ectopic expression of MyoC	MYOY1	myoblast differ		pos	NONE				
27	The reduced lipid content i	lipid	Insulin Resist		pos	NONE				
28	in rat hepatocyte primary c	l6	Orn1		pos	NONE				
29	Evidence indicates that sut	response to ox	Telomere		neg	NONE				

Quick Workspace

**Text Annotation** regulator regulated location cell line cell type organ disease species

#1  

 (Protein levels of RFX5 increase in the nucleus during IFN-gamma treatment.)

#2  

 (Whereas hyperglycemia and insulin had no effect on MIF expression, angiotensin II induced MIF secretion in THP-1 cells via the angiotensin 1- receptor [58].)

#3  

 activation of AKT/PKB by various growth and survival factors, including TCR and CD28, involves a phosphatidylinositol 3-kinase (PI3K)-dependent membrane translocation of AKT/PKB

#4  

 We and others have shown that OSM can induce ectoxin-1 in mouse lung fibroblasts (MLF) (28) and lung smooth muscle cells (29).

#5  

 (experimental emphysema caused by lung overexpression of IL-13 or deletion of surfactant protein D involved both MMP-9 and MMP-12 (173, 348))

#6

Figure 2: CELESTA Web interface. Top: Interaction table and text annotation cards after uploading a BioRECIPE interaction list before annotation. Bottom: after running annotation, with auto-populated context columns highlighted in the table (left) and text annotation cards (right) displaying entity highlights with type labels, relation arcs, and context indicators. Explicit contexts appear as inline highlights, and implicit contexts are shown below the text.

# EmergeX: A computational framework for designing emergent dynamics in chemical reaction networks

Colin Yancey  
Johns Hopkins University  
Baltimore, USA  
cyancey7@jhu.edu

Cameron Kolisko  
Johns Hopkins University  
Baltimore, USA  
ckolisk1@jhu.edu

Rebecca Schulman  
Johns Hopkins University  
Baltimore, USA  
rschulm3@jhu.edu

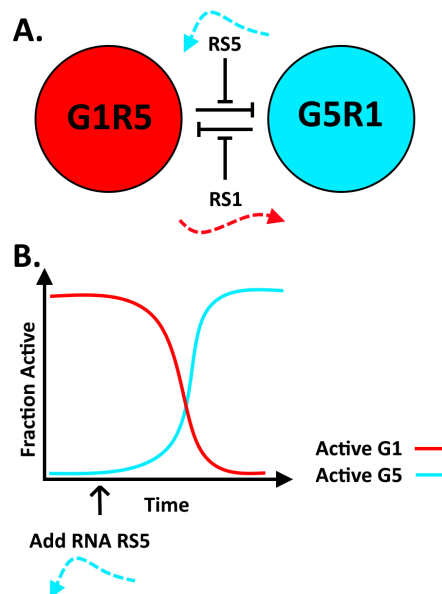
## 1 INTRODUCTION

Complex chemical reaction networks (CRNs) that give rise to emergent functions are ubiquitous in nature, from biological clocks such as circadian oscillators to gene expression pathways that regulate protein expression such as the lac operon of *E. coli*. Biological CRNs have inspired many synthetically engineered emergent functions, including feed-forward loops, bistable switches, and repressilators [3]. To create novel functions, a reasonable CRN can be designed and the initial conditions tuned in accordance with the desired behavior. Many biomolecular circuits have internal structures that allow the same set of components to exhibit rich dynamics under different initial component concentrations. However, there are no standard, general methods for identifying these concentrations; related design problems are often addressed using manual tuning or heuristic optimization methods [1].

Here we present *EmergeX*, a computational platform for CRN design and analysis that includes a generalized behavior optimizer. The behavior optimizer identifies initial conditions for CRNs specified by mass action kinetics that produce an emergent function described by a set of target behaviors. These behaviors encode desired component concentration trajectories over time. *EmergeX* uses the Python library JAX for rapid optimization, enabling efficient numerical integration and gradient-based optimization even for complex CRNs with widely varying reaction rates and concentration regimes that lead to stiff systems of ODEs. This approach allows for simultaneous optimization over large numbers of parameters and customizable event schedules, while supporting constrained optimization over parameter domains, including relationships between initial concentrations.

## 2 EXAMPLE SYSTEM AND OPTIMIZATION PROBLEM

As a representative nonlinear CRN, we consider a T7 RNA polymerase-driven (RNAP-driven) DNA/RNA bistable switch. Two transcriptional states mutually inhibit one another through RNA-mediated promoter blockade, and a transient externally supplied RNA input, labeled as RS1 and RS5 in Figure 1A, can induce switching from one state to the other. Figure 1B illustrates a desired behavior, in which an externally supplied RS5 RNA input induces a state switch.

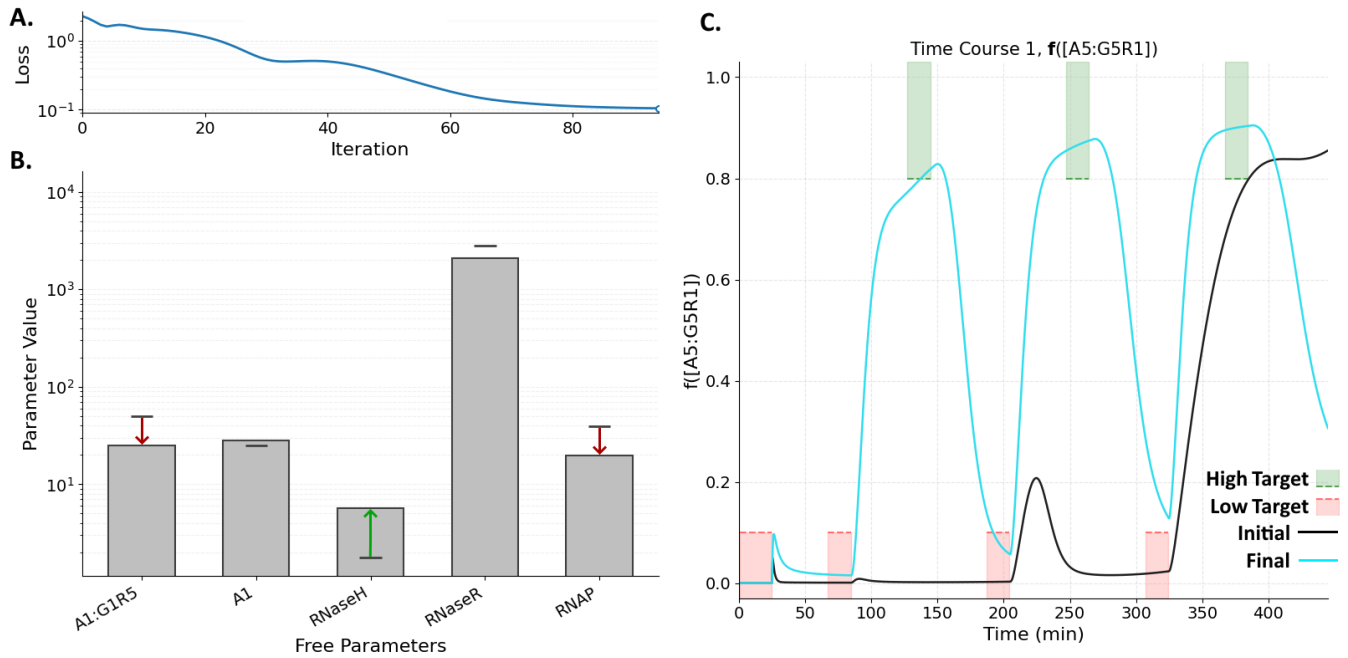


**Figure 1: Example bistable-switch topology and intended behavior.** (A) Two transcriptional states mutually inhibit one another through RNA-mediated blockade of each other’s template promoters through sequestration of a domain-specific activator. We refer to the activator of the G1 domain as A1, and the activator of the G5 domain as A5. (B) An example of a transient external RNA input inducing a state switch. While the model topology may theoretically allow for a bistable switch-like behavior to take shape, the concentrations of critical components to enter such a regime are unknown.

In this example, the optimizer searches experimentally controllable quantities including RNAP concentration, RNases H and R concentrations, template concentrations, and activator concentrations, while respecting user-defined bounds and parameter relationships.

## 3 BEHAVIOR OPTIMIZATION RESULTS

Figure 2 demonstrates optimization of the reaction network’s initial conditions to meet the desired behavior represented by the red and green intervals, target conditions in Figure 2C. Figure 2B shows the initial (represented by the starting lines) and final (represented by the bar positions) parameter values that allow for the trajectories in Figure 2C to occur. Aside from the first low target interval, the end of each low target



**Figure 2: Behavior optimization for the bistable-switch model. (A) Loss is minimized over iterations. (B) Free parameters are adjusted during optimization using the loss gradient found via auto-differentiation. (C) The normalized behavior of the species of interest is transformed from the initial trajectory to one that more closely satisfies the repeated high and low target windows.**

interval is when a fixed amount of RS5 is added to induce G5 dominance. Conversely, at the end of each high interval, a fixed amount of RS1 is added to re-induce G1 dominance. Starting from an initial parameterization that does not produce the desired switching pattern even when RNA is added, EmergeX changes free parameters spanning orders of magnitude, and rectifies the normalized trajectory of the species of interest so that it more closely satisfies repeated low and high target windows by efficiently navigating the loss landscape. Instead of requiring manual parameter sweeps, the user specifies desired temporal behavior and allows the optimizer to search the experimental space which the user defines as feasible.

#### 4 PARAMETER FITTING

EmergeX incorporates a parameter fitting module using optimization architecture that integrates well with behavior optimization tools within the library. Together, these tools support multiple steps in a model-guided design workflow: EmergeX can propose experimental conditions predicted to produce a desired behavior, while experimental measurements can be used to refine model parameters through parameter fitting. This integration enables iterative improvement of both model fidelity and CRN design toward achieving the desired functionality.

#### 5 AVAILABILITY

While the provided example is a specific case study of more niche interest such as within the genetic circuit community [2], the library is general-purpose, capable of accommodating any reaction network topology and any describable event and behavior schedules. The library is designed to accommodate Python users of all experience levels, and has built-in visualization functions that can create images and videos of any steps in the optimization process. These tools were used to create Figure 2.

EmergeX is available on PyPI for a quick and easy installation and on GitHub for code implementation and example use cases.

<https://pypi.org/project/emergex/>

<https://github.com/YanceyColin/EmergeX>.

#### REFERENCES

- [1] HUYNH, L., KECECIOGLU, J., KÖPPE, M., AND TAGKOPOULOS, I. Automatic design of synthetic gene circuits through mixed integer non-linear programming. *PLOS ONE* 7, 4 (2012), e35529.
- [2] MÜLLER, M. M., ARNDT, K. M., AND HOFFMANN, S. A. Genetic circuits in synthetic biology: broadening the toolbox of regulatory devices. *Frontiers in Synthetic Biology* 3 (2025), 1548572.
- [3] SANTOS-MORENO, J., TASIUDI, E., STELLING, J., AND SCHAERLI, Y. Multistable and dynamic CRISPRi-based synthetic circuits. *Nature Communications* 11, 1 (2020), 2746.

# Performance of Metaheuristic Algorithms in Finding Tradeoffs in One- and Two-Species Biological Feedback

Quang Luan Dang Tran, Nguyen Hoai Nam Tran, Derrick Appiah Osei, Ania-Ariadna Baetica  
Drexel University  
{qt49,nt625,da886,ab4778}@drexel.edu

## ABSTRACT

Biological feedback mechanisms play a vital role in helping organisms adapt and respond to their changing environments [2, 10]. Changes in external conditions such as temperature, pH, and nutrient availability can affect intracellular biochemical reaction rates, hence altering a cell's ability to survive, adapt, and respond [2]. Feedback regulation is used widely to buffer these changes, but how effectively it does so remains understudied. Similarly, feedback regulation has been used in synthetic biology to maintain homeostasis in engineered circuits and to ensure reliable performance under varying conditions [3, 5, 15].

In our work, we aim to design synthetic circuits that successfully mitigate changing external conditions. We do so by identifying the best biochemical parameter regimes. This can be a computationally demanding task [17, 20]. Our previous work [20] analyzed how variations in biochemical parameters affect circuits with feedback regulation and minimized these variations. This was previously done using a grid search method, which samples the biochemical parameter space at high resolution [20]. However, as circuit complexity increased, this brute-force approach became computationally prohibitive. In this work, we evaluate how suitable several algorithm families are for solving these problems. We identified several algorithm families with significantly lower runtime than grid search and high-quality solutions.

To formalize the problem of identifying circuit designs and biochemical parameter regimes that mitigate changes in external conditions, we use sensitivity analysis to quantify how changes in biochemical parameters affect a circuit's output. A synthetic circuit's output is commonly the steady-state concentration of a molecular species. This output is sensitive to changes in the underlying biochemical parameters [6]. To characterize how variations in these parameters affect the output, sensitivity analysis provides a quantitative measure of robustness [11, 14]. Robustness reflects the degree to which the output remains invariant under such changes [12]. Systems with lower sensitivity are considered more robust. Here, we seek designs whose outputs are minimally sensitive to variation in multiple biochemical rates.

Since the concentration of a molecular species can depend on multiple biochemical parameters, such as transcription and degradation rates [2], this gives rise to multiple sensitivity functions, each corresponding to sensitivity with respect to one parameter. Achieving robustness to simultaneous variation in multiple biochemical parameters therefore requires a multi-objective optimization problem that minimizes these sensitivities together. Trade-offs between sensitivities can arise, meaning that if the output becomes less sensitive with respect to one parameter, it may be more sensitive to the others. Such trade-offs can be illustrated by a Pareto front, which shows the set of optimal solutions that cannot be improved in one sensitivity without degrading another [7]. We aim to find the biochemical parameter values that represent the best possible trade-off when that occurs.

When solving multi-objective optimization problems, algorithms have to compromise between runtime, convergence, complexity, etc, and we aim to find algorithms that give us the best overall performance. Although grid search can give the perfect Pareto front, its runtime is significant, and it increases with circuit complexity. We initially tested several existing algorithms and frameworks, including AMOSA [4], TorchJD [16], and OpenMDAO [9], but found them unsuitable for our problems. AMOSA was prohibitively slow, TorchJD did not fit our multi-objective formulation, and OpenMDAO solves only convex problems. We therefore focused on the three algorithm families shown in Figure 1.

We benchmark algorithms known for solving multi-objective optimization problems from three major metaheuristic families - Sobol multi-objective simulated annealing (Sobol MOSA) from simulated annealing, Non-dominated Sorting Genetic Algorithm II (NSGA-II) from genetic algorithms, and Exact Pareto Optimization (EPO) from gradient-based method - on two canonical synthetic feedback architectures: negative autoregulation and coupled positive-negative feedback [2, 19]. We benchmark the algorithms on these two circuits because negative autoregulation is a common regulatory motif for synthetic and systems biology, positive-negative feedback architectures have been built experimentally in [19], and, importantly, our previous work in [20] has shown that both

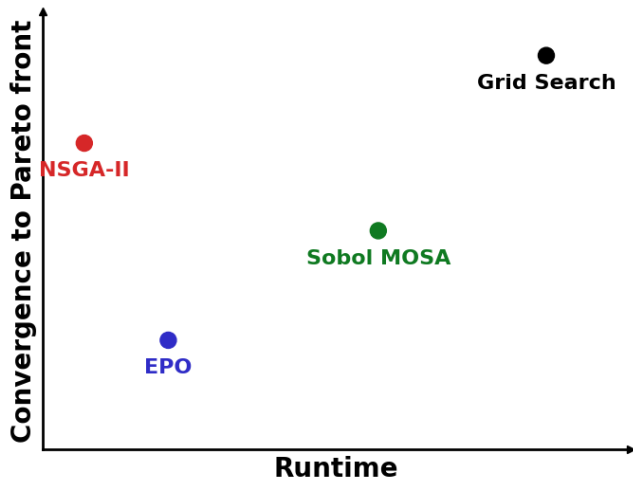


Figure 1: We compare the four algorithms on two circuits: negative autoregulation and positive-negative feedback. We rank their convergence using the inverted generational distance, which measures both how close the recovered Pareto front is to the true Pareto front and how well solutions are distributed along it. The grid search produces the exact Pareto front. It has the best convergence and the longest runtime. NSGA-II performs best among the approximate methods in both runtime and convergence to the true front. EPO has higher runtime and worse convergence than NSGA-II. Sobol MOSA has high convergence but is slower than NSGA-II.

circuits exhibit nontrivial Pareto fronts rather than collapsing to a single solution.

We briefly describe each of the three algorithms. Simulated annealing acts like a high-temperature system: it bounces around the landscape of possible solutions energetically, which allows it to hop out of local minima. As the "temperature" gradually cools, the search becomes more focused, eventually settling into the most optimal solutions [1]. We integrated a Sobol sequence which is a sampling method that spreads sampling points out more evenly across the parameter space [18] to ensure proper coverage of parameter space and reproducible results. NSGA-II is a genetic algorithm, and it maintains a diverse population of many candidate solutions at once. In each generation, the algorithm identifies the fittest individuals that best balance the trade-offs between competing objectives, while ensuring the population remains diverse to prevent the search from early convergence [8]. Lastly, EPO is a gradient-based approach, which functions like navigating down a hill by always moving in the direction of steepest descent. While this allows for very rapid convergence to a solution, it is prone to getting stuck in local minima [13]. We illustrate the algorithms in Figure 2.

Feedback is a set of biochemical processes, including production of output particles (involving a production rate),

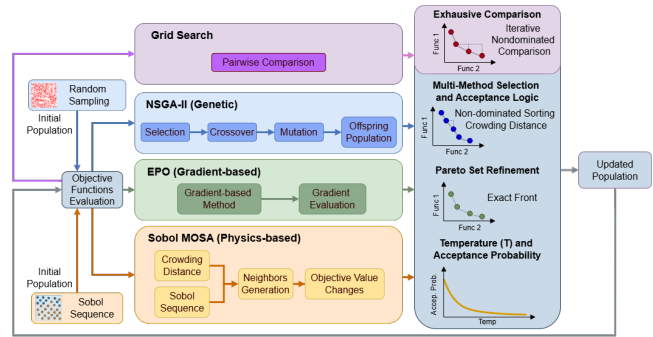


Figure 2: General architecture of the algorithms.

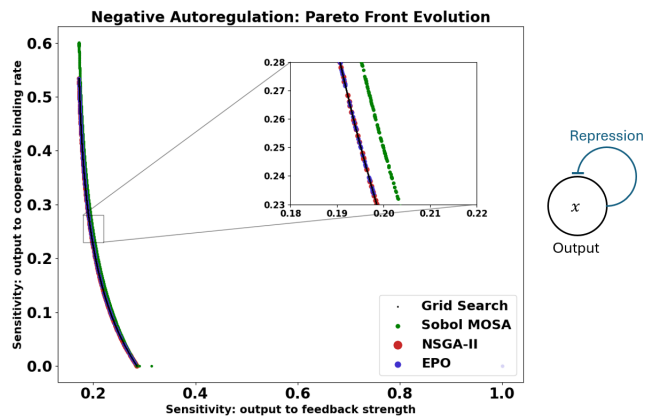
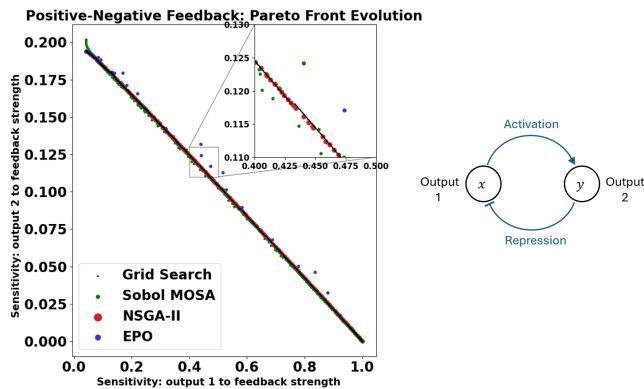


Figure 3: (Right): A single biochemical output under negative autoregulation. The repression mechanisms involves cooperative binding and feedback interaction. Therefore, our output is dependent on both feedback strength and cooperative binding, which the Pareto front on the left measures. (Left): Pareto fronts describing when our output is most robust to both feedback strength and cooperative binding. All algorithms converge closely to the true Pareto front found using grid search. However, Sobol MOSA and EPO still have some points on the right that their runs need to be more refined to remove.

binding to other particles (whose binding affinity we call the dissociation constant) and degradation of output particles (whose rate we call the degradation rate). The "feedback strength" is defined as the ratio of the production rate to the degradation rate per unit of dissociation constant. We use generational distance (GD) and inverted generational distance (IGD) which measure the average distance from any point in our obtained Pareto front to the closest point in the grid search's (true) Pareto front, and vice versa [21, 22], correspondingly, to quantify the convergence and diversity of solutions. Small GD and IGD values indicate better convergence and diversity. In both circuits (Figure 3 and Figure 4), the three algorithms recover Pareto fronts that are closely



**Figure 4: (Right):** Two biochemical outputs under positive-negative feedback. Both outputs are dependent on feedback strength. **(Left):** Pareto fronts describing when both outputs are most robust to feedback strength. NSGA-II and Sobol MOSA recover the true Pareto front found using grid search. EPO still needs more evaluations though it has the same running budget as NSGA-II’s.

aligned with the grid search method, while being up considerably faster. NSGA-II is the fastest, 21 hours faster compared to grid search method in positive-negative feedback, and also returns the closest Pareto fronts with very small GD and IGD values. Sobol MOSA is slightly slower than grid search in negative autoregulation, but faster in positive-negative feedback. EPO, even though is 20 hours faster than grid search in positive-negative feedback, retrieves much less Pareto points than the other two algorithms, and their Pareto points are not fully converged.

We also examine the impact of algorithm-specific hyperparameter choices on solution quality and performance to further determine the trade-offs between runtime and accuracy of each algorithm. Sobol MOSA’s convergence and diversity improve with increasing number of runs, solutions, and iterations. The number of solutions is the safest option to change, followed by number of runs without considerably increasing the runtime. NSGA-II solves both circuits with a budget of 600 evaluations per run, with close proximity and dense coverage across the true Pareto front in a single run. With the same budget, we can improve EPO’s coverage by increasing its step size. All simulations were run on a 64-bit workstation with an AMD Ryzen Threadripper 3960X 24-Core CPU, 126 GB RAM, and 48 logical cores.

Here we solve the multi-objective optimization problem of finding the most robust synthetic circuit designs for negative autoregulation and positive-negative feedback using three families of algorithms. All algorithms are faster and similarly accurate compared to grid search through the parameter space of the circuits. Other optimization algorithms have been tested, but these three stand out for their accurate

and quick convergence, while being easy-to-implement out of the box and easy-to-finetune. In the future, we plan to expand our case studies to synthetic circuits for metabolic engineering, as these systems are often subject to multiple competing demands and constraints. Moreover, mechanistic models of metabolic circuits contain biochemical parameters spanning many orders of magnitude, providing a valuable testbed for evaluating the scalability of multi-objective optimization algorithms. Our work aims to reduce the experimental trial-and-error and computational burden associated with identifying optimal parameter sets for synthetic circuits.

## REFERENCES

- [1] AARTS, E., AND KORST, J. *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing*. John Wiley & Sons, Inc., 1989.
- [2] ALON, U. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2019.
- [3] AOKI, S. K., LILLACCI, G., GUPTA, A., BAUMSCHLAGER, A., SCHWEINGRUBER, D., AND KHAMMASH, M. A universal biomolecular integral feedback controller for robust perfect adaptation. *Nature* 570, 7762 (2019), 533–537.
- [4] BANDYOPADHYAY, S., SAHA, S., MAULIK, U., AND DEB, K. A simulated annealing-based multiobjective optimization algorithm: Amosa. *IEEE transactions on evolutionary computation* 12, 3 (2008), 269–283.
- [5] BRIAT, C., GUPTA, A., AND KHAMMASH, M. Antithetic integral feedback ensures robust perfect adaptation in noisy biomolecular networks. *Cell systems* 2, 1 (2016), 15–26.
- [6] BROPHY, J. A., AND VOIGT, C. A. Principles of genetic circuit design. *Nature methods* 11, 5 (2014), 508–520.
- [7] CENSOR, Y. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization* 4, 1 (1977), 41–59.
- [8] DEB, K., PRATAP, A., AGARWAL, S., AND MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.
- [9] GRAY, J. S., HWANG, J. T., MARTINS, J. R. R. A., MOORE, K. T., AND NAYLOR, B. A. OpenMDAO: An open-source framework for multidisciplinary design, analysis, and optimization. *Structural and Multidisciplinary Optimization* 59, 4 (April 2019), 1075–1104.
- [10] HOLYST, R., BUBAK, G., KALWARCZYK, T., KWAPISZEWSKA, K., MICHALSKI, J., AND PILZ, M. Living cell as a self-synchronized chemical reactor. *The Journal of Physical Chemistry Letters* 15, 13 (2024), 3559–3570.
- [11] INGALLS, B. Sensitivity analysis: from model parameters to system behaviour. *Essays in biochemistry* 45 (2008), 177–194.
- [12] KITANO, H. Biological robustness. *Nature Reviews Genetics* 5, 11 (2004), 826–837.
- [13] MAHAPATRA, D., AND RAJAN, V. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 6597–6607.
- [14] PANTAZIS, Y., KATSOLAKIS, M. A., AND VLACHOS, D. G. Parametric sensitivity analysis for biochemical reaction networks based on pathwise information theory. *BMC bioinformatics* 14, 1 (2013), 311.
- [15] PURNICK, P. E., AND WEISS, R. The second wave of synthetic biology: from modules to systems. *Nature reviews Molecular cell biology* 10, 6 (2009), 410–422.
- [16] QUINTON, P., AND REY, V. Jacobian descent for multi-objective optimization. *arXiv preprint arXiv:2406.16232* (2024).

- [17] SHOPERA, T., HENSON, W. R., NG, A., LEE, Y. J., NG, K., AND MOON, T. S. Robust, tunable genetic memory from protein sequestration combined with positive feedback. *Nucleic acids research* 43, 18 (2015), 9086–9094.
- [18] SOBOL, I. M. Distribution of points in a cube and approximate evaluation of integrals. *USSR Computational mathematics and mathematical physics* 7 (1967), 86–112.
- [19] STRICKER, J., COOKSON, S., BENNETT, M. R., MATHER, W. H., TSIMRING, L. S., AND HASTY, J. A fast, robust and tunable synthetic gene oscillator. *Nature* 456, 7221 (2008), 516–519.
- [20] TRAN, N. H. N., NGUYEN, A., RAHMAN, T. W., AND BAETICA, A.-A. Fundamental trade-offs in the robustness of biological systems with feedback regulation. *ACS Synthetic Biology* 14, 4 (2025), 1099–1111. PMID: 40198741.
- [21] VAN VELDHUIZEN, D. A., AND LAMONT, G. B. Multiobjective evolutionary algorithm test suites. In *Proceedings of the 1999 ACM symposium on Applied computing* (1999), pp. 351–357.
- [22] WANG, Z., XIAO, C., AND ZHOU, A. Exact calculation of inverted generational distance. *IEEE Transactions on Evolutionary Computation* 29, 5 (2025), 1966–1975.

# Predictive Analysis of Amplification in Biochemical Feedback Circuits

Nguyen HN Tran  
Drexel University  
Philadelphia, USA  
nt625@drexel.edu

Ania-Ariadna Baetica  
Drexel University  
Philadelphia, USA  
ab4778@drexel.edu

Samuel W. Schaffter  
National Institute of Standards and Technology  
Maryland, USA  
samuel.schaffter@nist.gov

## 1 INTRODUCTION

Biochemical circuits require amplification mechanisms to propagate signals and process information. At the molecular scale, however, unintended reactions, or “leak,” are difficult to eliminate and can generate output even in the absence of input. As a result, leak interferes with measurement and complicates circuit design. [1,2,3,4]

A central challenge is to understand how amplification depends on system parameters in the presence of leak, in order to identify regimes of effective amplification. This is particularly difficult due to the large number of parameters and tunable reaction rates involved in these systems. There is therefore a need for general approaches and computational tools to systematically characterize and optimize amplification in biochemical circuits.

In this work, we present a computational approach to this problem, illustrated with a case study of a circuit called the seesaw element. We define amplification in a way that explicitly accounts for leak and use it to make predictions of optimal amplification design. This prediction is based on a mechanistic model that has been fit to experimental data.

## 2 BACKGROUND

### Seesaw element

In RNA strand displacement systems, amplification can be implemented using the seesaw element – a biochemical feedback circuit. In the forward strand displacement pathway (Fig. 1 green), an input RNA strand reacts with a DNA gate to produce an output strand while leaving behind a residual gate. In the reverse gate regeneration pathway (Fig. 1 purple), a fuel RNA strand recycles this residual gate to regenerate the input, enabling repeated cycles of signal production. [3,5,6]

However, effects from thresholding and spurious interactions between non-complementary strands can introduce unintended reactions. In particular, interactions between fuel strands and gates can generate output even in the absence of input, forming a leak pathway. How fast this leak occurs, the rate of fuel strand displacement, and how much fuel template is present, therefore, affect the efficiency of this feedback loop and thus output amplification. The key challenge is to balance the cost of driving this loop with the amplification gain it produces. [4,6]

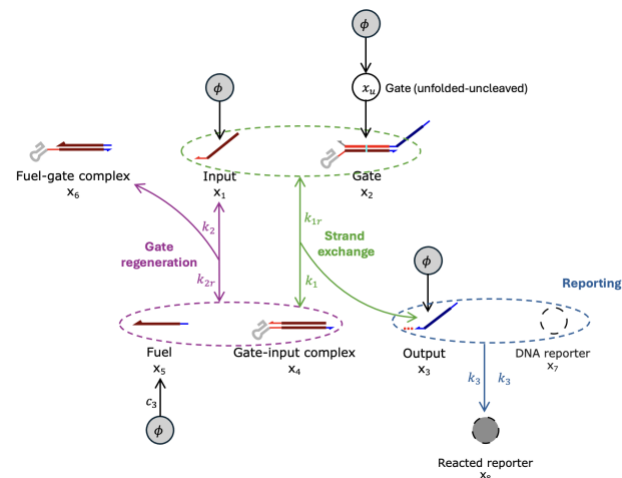


Figure 1: Seesaw element. Process begins with strand exchange (green) and restarts with gate regeneration (purple). Reporting (blue) allows us to observe outputs. Figure adapted from [7].

### Amplification

Amplification refers to the ability of a system to generate more output than would be expected from a single-pass, stoichiometric reaction, typically through catalytic or feedback-driven reuse of input species.

In the seesaw element, amplification arises from the ability of a single input RNA strand to generate multiple output strands through catalytic reuse, rather than one-to-one stoichiometric conversion. This behavior is enabled by a fuel strand, which regenerates the input after output release and allows repeated rounds of signal

production. Our output is observed via a fluorescent reporter.

To distinguish contributions to the observed output, we consider two conditions:

- *ON: input template present; output is observed through both input and leak pathways*
- *OFF: input template absent; output is observed solely through leak*

We define amplification at time  $t$  and fuel template  $f$  as the ratio of output production rates between ON and OFF conditions:

$$A_0(t, f) = \left( \frac{dx_8}{dt}(ON, f) \right) / \left( \frac{dx_8}{dt}(OFF, f) \right)$$

To isolate the effect of fuel on amplification, we define the normalized amplification relative to its zero-fuel baseline:

$$A(t, f) = A_0(t, f) / A_0(t, 0)$$

Values of  $A(t, f) > 1$  indicate enhancement relative to the zero-fuel baseline, while values below 1 indicate suppression.

### 3 METHODS

Experimental data was collected from cotranscriptionally encoded toehold-mediated strand displacement (ctrSD) experiments, in which fluorescent reporter concentration was measured over time under input template concentrations of 0, 1.25, and 2.5 nM, and fuel template concentrations of 0 and 25 nM. [7]

Using a mechanistic model of the seesaw element, we iteratively fit the system parameters  $\theta$  to the experimental data using nonlinear least squares optimization by minimizing the sum of squared residuals between experimental measurements  $y_i$  and model predictions  $f$  at matching times  $t_i$ :

$$S(\theta) = \sum_i (y_i - f(t_i, \theta))^2$$

We evaluated two candidate models for leak: (i) direct interaction between fuel and active gate state, and (ii) interaction between fuel and the inactive (unfolded-uncleaved) gate state. The second model was found to provide a better fit (see Figure 2).

We additionally evaluated three candidate models for transcription rate: constant, linearly decreasing in time, and exponentially decreasing in time. The time-dependent transcription models did not improve fit quality and were therefore excluded from subsequent analysis (see Figure 3).

Using the fitted model, we then simulated amplification profiles across varying fuel strand displacement rates and leak rates to identify parameter regimes that yield optimal amplification, within physically reasonable parameter ranges. [4,8]

### 4 RESULTS

Figure 2 illustrates the ability of both models to predict the experimental condition (input, fuel) = (2.5,25) nM using parameters obtained from fitting the remaining four experimental conditions: (1.25,0), (2.5,0), (0,25), and (1.25,25) nM. The inactive gate model achieved both a lower fitting error and a lower prediction error than the active gate model. These errors are listed in Table 1.

	<b>Fitting RMSE:</b>	<b>Prediction RMSE:</b>
<b>Model 1</b>	$5.0 \times 10^{-3}$ (nM)	$1.0 \times 10^2$ (nM)
<b>Model 2</b>	$2.8 \times 10^{-3}$ (nM)	$4.1 \times 10^1$ (nM)

*Table 1: Global root mean square error (RMSE) values for both models. Column 1 shows the RMSE for fits to the four selected experimental conditions, while Column 2 shows the RMSE for prediction of the held-out experimental condition.*

Figure 4 presents amplification profiles across varying fuel concentrations, fuel strand displacement rates, and leak rates. Increasing the fuel strand displacement rate leads to higher peak amplification, but shifts the peak to earlier times, indicating a trade-off between amplification magnitude and persistence of this magnitude.

Increasing fuel concentration similarly enhances peak amplification, with higher fuel levels producing larger but earlier responses. This is expected, as fuel is the driver of the feedback loop.

In contrast, variations in leak rate have a negligible effect on amplification profiles within the explored parameter range.

Overall, these results reveal a design trade-off: conditions that maximize amplification also reducing the duration over which amplification is sustained.

## 5 FURTHER DIRECTIONS

This work provides experimentally testable predictions for future seesaw circuit designs, particularly for systems operating at input and fuel template concentrations up to 50 nM.

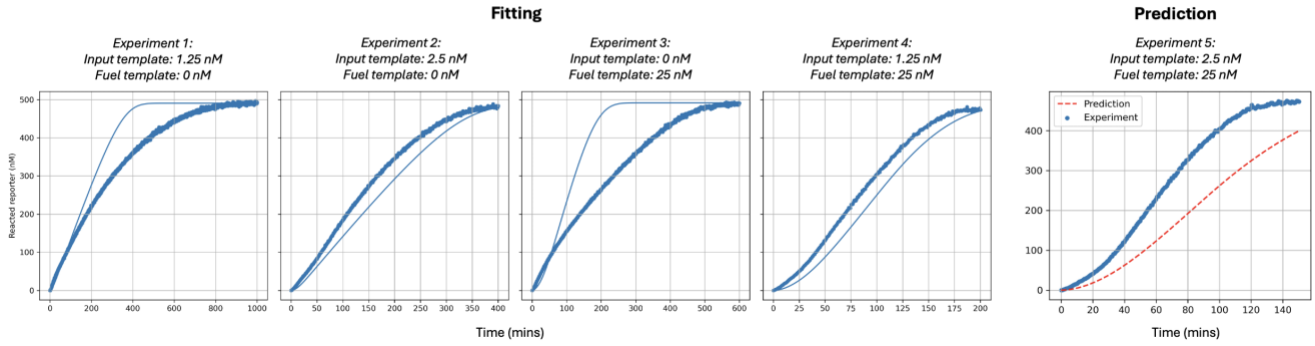
More broadly, this work presents a computational workflow for analyzing and optimizing amplification in biochemical feedback circuits in the presence of leak. By combining mechanistic modeling, parameter estimation, and amplification analysis, the framework may be extended to other architectures as seen in biosensors, transduction circuits, molecular classifiers, and molecular switches. [9,10,11,12,13, 14].

## REFERENCES

- [1] Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC.
- [2] Antebi, Y. E., Nandagopal, N., & Elowitz, M. B. (2017). An operational view of intercellular signaling pathways. *Current Opinion in Systems Biology*, 1, 16–24. <https://doi.org/10.1016/j.coisb.2016.12.001>
- [3] Zhang, D. Y., Turberfield, A. J., Yurke, B., & Winfree, E. (2007). Engineering entropy-driven reactions and networks catalyzed by DNA. *Science*, 318(5853), 1121–1125. <https://doi.org/10.1126/science.1148532>
- [4] Zhang, D. Y., & Winfree, E. (2009). Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47), 17303–17314. <https://doi.org/10.1021/ja906987s>
- [5] Qian, L., & Winfree, E. (2011). Scaling up digital circuit computation with DNA strand displacement cascades. *Science*, 332(6034), 1196–1201. <https://doi.org/10.1126/science.1200520>
- [6] Qian, L., & Winfree, E. (2011). A simple DNA gate motif for synthesizing large-scale circuits. *Journal of the Royal Society Interface*, 8(62), 1281–1297. <https://doi.org/10.1098/rsif.2010.0729>
- [7] Schaffter, S. W., & Strychalski, E. A. (2022). Cotranscriptionally encoded RNA strand displacement circuits. *Science Advances*, 8(12), eabl4354. <https://doi.org/10.1126/sciadv.abl4354>
- [8] Srinivas, N., Ouldrige, T. E., Šulc, P., Schaeffer, J. M., Yurke, B., Louis, A. A., Doye, J. P. K., & Winfree, E. (2013). On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research*, 41(22), 10641–10658. <https://doi.org/10.1093/nar/gkt801>
- [9] Li, Y., et al. (2025). A cell-free biosensor signal amplification circuit with polymerase strand recycling. *Nature Chemical Biology*, 21, 949–958. <https://doi.org/10.1038/s41589-024-01816-w>
- [10] Wan, X., et al. (2019). Cascaded amplifying circuits enable ultrasensitive cellular sensors for toxic metals. *Nature Chemical Biology*, 15(5), 540–548. <https://doi.org/10.1038/s41589-019-0244-3>
- [11] Liu, B., Samaniego, C. C., Bennett, M. R., et al. (2023). A portable regulatory RNA array design enables tunable and complex regulation across diverse bacteria. *Nature Communications*, 14, 5268. <https://doi.org/10.1038/s41467-023-40785-x>
- [12] Mihailovic, M. K., Vazquez-Anderson, J., Li, Y., et al. (2018). High-throughput in vivo mapping of RNA accessible interfaces to identify functional sRNA binding sites. *Nature Communications*, 9, 4084. <https://doi.org/10.1038/s41467-018-06207-z>
- [13] Kieffer, C., Genot, A. J., Rondelez, Y., & Gines, G. (2023). Molecular computation for molecular classification. *Advanced Biosystems*. <https://doi.org/10.1002/adbi.202200203>
- [14] Subsoontorn, P., Kim, J., & Winfree, E. (2012). Ensemble Bayesian analysis of bistability in a synthetic transcriptional switch. *ACS Synthetic Biology*, 1(8), 299–316. <https://doi.org/10.1021/sb300018h>

## Comparison of leak models

### Model 1: leak from fuel and active gate



### Model 2: leak from fuel and inactive gate

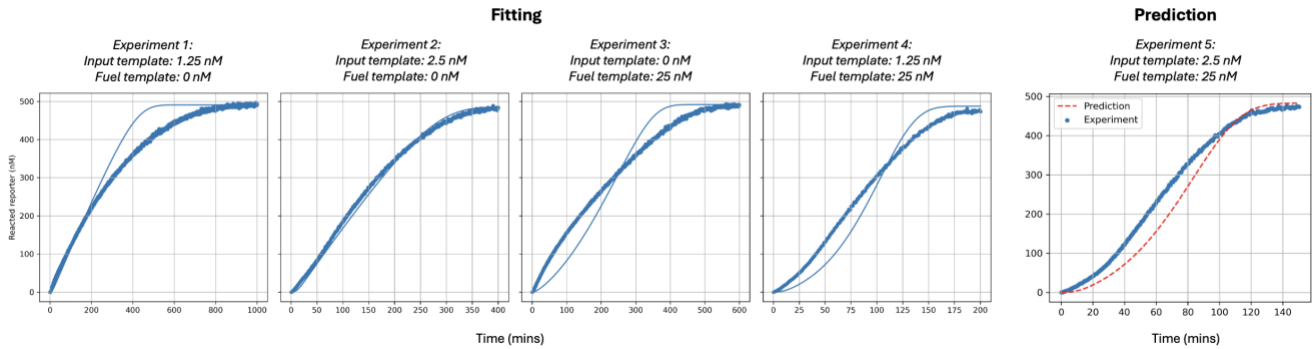


Figure 2:

Comparison between predictions of models 1 and 2, by fitting four experimental timecourses and predicting the fifth. Model 2 shows better fitting and prediction.

(Left) Plots of all four experimentally measured reacted reporter timecourses (dotted) against best fits from each model (lined).  
 (Right) Plots of the fifth experiment (dotted) against prediction from the corresponding model (dashed).

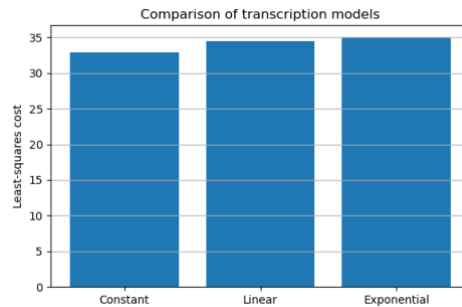


Figure 3: Fitting cost of different transcription rate models. Constant transcription rate was found to be the best mechanistic model.

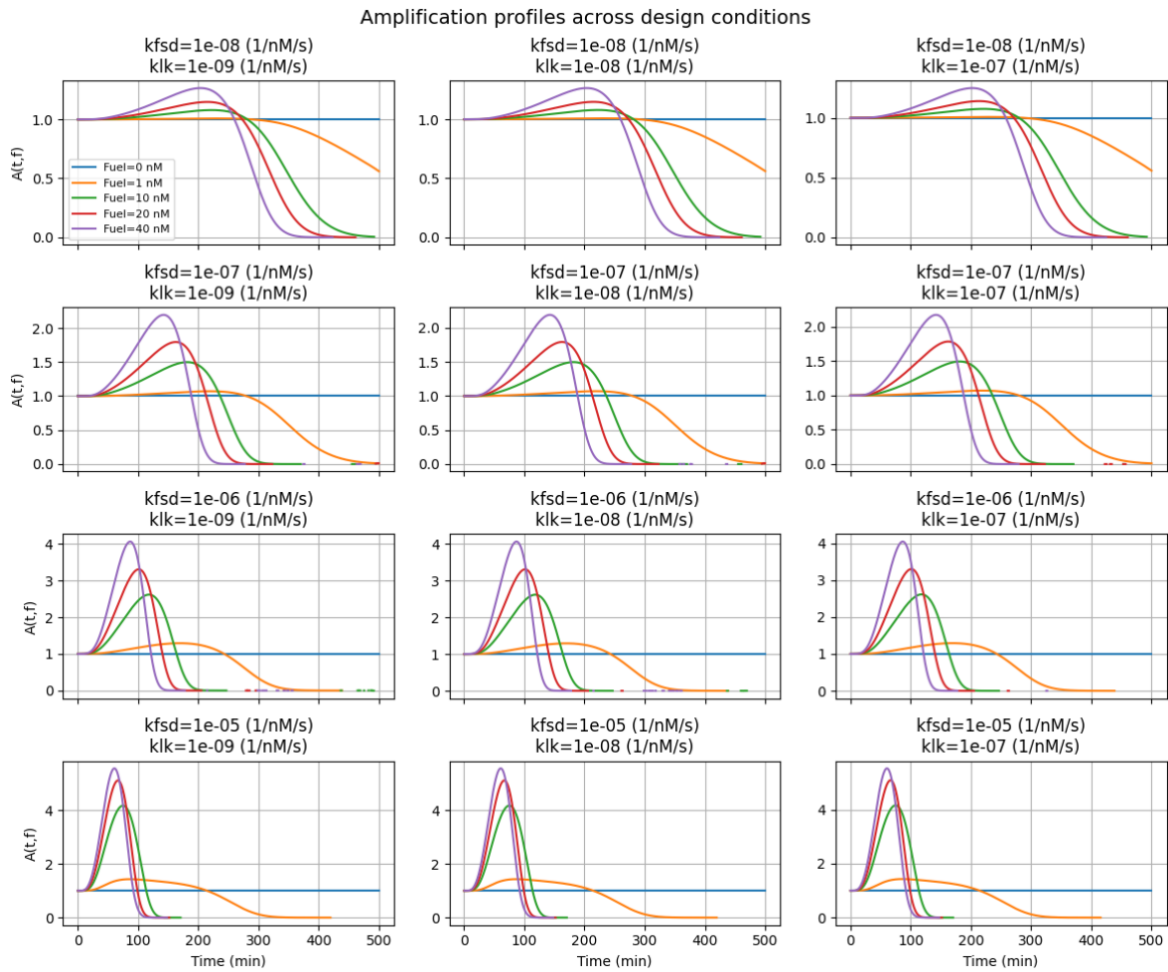


Figure 4: Prediction of amplification profiles across different fuel strand displacement rates ( $k_{fsd}$ ) and leak rates ( $k_{lk}$ ).

# Engineered Critical Point Dynamics in Genetic Circuits

Payton J Thomas\*, Johanna Cao, Caleb Wong, Jeff Hasty

UC San Diego Synthetic Biology Institute

\*paytonco@ucsd.edu

## 1 SELF-ORGANIZED CRITICALITY IN PHYSICAL SYSTEMS

Many physical systems which operate near a bifurcation point or phase transition tend to exhibit scale-invariant dynamics, long-range correlations, and spontaneous organization into critical states without fine-tuning of external parameters. This phenomenon, called self-organized criticality (SOC) [2], has been studied in avalanches [3], earthquakes [1, 12], and forest fires [9], among other phenomena. Such dynamical systems exhibit power-law frequency distributions, scale-invariance, and long-range correlations, all hallmarks of SOC.

We consider snow avalanches as a prototypical SOC system. Avalanches occur when snowfields evolve to the neighborhood of a critical point and are then perturbed, leading to the avalanche event. The system is left in a lower energy state, from which it continues to evolve and build to another critical state under a stochastic drive. The gradual accumulation of snow acts as a slow input of energy, driving the system toward a metastable configuration near a bifurcation point, where the system is maximally sensitive and small perturbations can produce cascading avalanches spanning a wide range of sizes. Similarly, the most devastating wildfires and earthquakes arise from perturbations after the system has reached a critical state. This interplay between slow energy accumulation and rapid, scale-free release gives rise to long-range correlations and power-law frequency distributions, often observed as  $1/f$  (pink) noise. Away from criticality, minor disturbances do not produce avalanche-like events, as the system is either too stable to propagate perturbations or too far from the threshold required for large-scale reorganization.

Consider as a toy model of SOC dynamics the supercritical pitchfork bifurcation (subject to noise),

$$dX = (\lambda X - X^3)dt + \sigma dB,$$

and its deterministic part  $\dot{x} = \lambda x - x^3$ , which undergoes a supercritical pitchfork bifurcation at  $\lambda = 0$  (Figure 1). For  $\lambda > 0$  and small  $\sigma$ , the stochastic system has two metastable states at  $x^* = \pm\sqrt{\lambda}$  and randomly switches between them, giving rise to a fast time scale of local fluctuations and a slow time scale of switching (Figure 2).

Near the bifurcation point, the Jacobian  $J = \lambda - 3x^2$  satisfies  $J \sim O(\lambda)$  at the metastable states, so perturbations decay extremely slowly,  $\varepsilon(t) \approx \varepsilon(0)$ , exhibiting critical slowing

down. This produces long-range correlations, quantified by the autocorrelation function  $\rho_{XX}(\tau)$ ,

$$\rho_{XX}(\tau) \approx A\tau^{-\beta}$$

and consequently the power spectral density,

$$S_{XX}(\omega) = \mathcal{F}\{\rho_{XX}(\tau)\}(\omega) \sim \frac{1}{\omega},$$

i.e., pink noise (Figure 3).

The slow time scale is governed by noise-driven transitions across an energy barrier defined by the potential function

$$U(x) = -\int (\lambda x - x^3)dx = \frac{1}{4}x^4 - \frac{\lambda}{2}x^2,$$

with barrier height  $\Delta U = \lambda^2/4$  (Figure 4). By Kramers' rate theory,

$$T \sim \exp\left(\frac{\Delta U}{D}\right) = \exp\left(\frac{\lambda^2}{2\sigma^2}\right),$$

sp that  $\lambda^2 \sim 2\sigma^2 \ln T$ . Differentiating gives

$$\frac{d\lambda}{dT} \sim \frac{1}{\lambda T} \sim \frac{1}{T\sqrt{\ln T}}.$$

Assuming  $\lambda$  is approximately Gaussian,  $P(\lambda) \sim \exp(-\lambda^2/2\sigma_\lambda^2)$ , we obtain

$$P(\lambda) \sim T^{-\sigma^2/\sigma_\lambda^2}.$$

By change of variables,

$$P(T) = P(\lambda) \left| \frac{d\lambda}{dT} \right| \sim T^{-(1+\beta)} \frac{1}{\sqrt{\ln T}}, \quad \beta = \frac{\sigma^2}{\sigma_\lambda^2}.$$

For sufficiently large  $T$ , the logarithmic correction is subdominant, yielding the approximate power-law

$$P(T) \approx T^{-\alpha}, \quad \alpha = 1 + \beta.$$

This system therefore exhibits the well-known hallmarks of SOC (long-range correlations, pink noise, scale-invariant power laws), all of which arise from the dynamics near a bifurcation point.

## 2 SELF-ORGANIZED CRITICALITY IN GENETIC CIRCUITS

We aim to investigate the emergence of self-organized criticality in gene regulatory networks using stochastic bifurcation theory. We will do so by designing and building a genetic circuit which realizes self-organized criticality on a hysteresis curve and verifying its behavior using simulation techniques and *in vivo* fluorescence microscopy utilizing

a microfluidic platform. Potential biomedical applications include targeted drug delivery, where engineered bacteria could transiently release chemotherapy agents in response to systemic induction, enabling precise dosage control within tumor microenvironments.

*Design.* To realize this kind of system as a genetic circuit, we have a handful of design requirements. We must include the positive autoregulation motif so that the system includes the necessary saddle-node phase transitions and exhibits hysteresis. We also require that the autoregulatory species be tunably degraded, and that the species that degrades it equilibrate slowly relative to the positive feedback loop (Figure 6) (Figure 11). We can achieve this with LuxI/AHL as the autoregulatory species and AiiA as the degradase.

Using Hill-Langmuir kinetics, we can write a (nondimensionalized) dynamical model for this system:

$$\begin{cases} \dot{x} = \frac{\beta(1+\alpha x^2)}{1+x^2} - (y + \gamma)x \\ \dot{y} = C(\text{IPTG}) - y \end{cases},$$

where  $x = [\text{AHL}]$  and  $y = [\text{AiiA}]$ . This model gives us the classical hysteresis curve we're looking for when  $y$  is at equilibrium (Figure 7). We move on to the SDE model of the system given by the Chemical Langevin Equation, introducing noise [7]. Simulations of this system exhibit stochastic avalanching (Figure 8), a power law-like avalanche time distribution (Figure 9), and long-range correlations and pink noise (Figure 10). These properties are essentially the 'goal-posts' for demonstrating SOC in the context of the double saddle-node hysteresis curve, as we established in the previous subsection. We can be confident that the engineered system exhibits self-organized criticality *in silico*.

*Build.* We selected well-characterized parts from the iGem Registry [11] with which to implement the circuit design. Part choices are summarized in Table 1. We selected iGem vector pSB1A2 [8] as a backbone, which contains a pUC19-derived pMB1 [4] origin and Carbenicillin resistance for selection. We selected the pMB1 origin because its copy number on the order of  $10^2$  will allow deterministic drift terms to be much larger than noise. Additionally, we designed a  $P_{\text{lux}}$  reporter plasmid in order to measure the activity of the lux promoter, a stand-in for the state variable [LuxI]. The reporter has origin of replication ColE1 [5] and relies on Kanamycin resistance for selection.

We constructed the genetic circuit and reporter plasmid using Gibson assembly [6] to achieve a scarless product (Figure 12). We confirmed cloning products via an Oxford Nanopore sequencing platform [10]. We constructed our deployment strain by co-transforming both engineered plasmids in K12 MG1655 *E. coli*.

*Test.* We first tested whether IPTG-induced AiiA modulation is sufficient to place the circuit parameters in the bistable region (Figure 7). We did so by measuring the cell density-normalized GFP fluorescence of our deployment strain at steady state at varying levels of IPTG induction (Figure 13). The expected outcome of this experiment is for the cell populations to lie along the lower branch of the hysteresis curve for large IPTG concentrations until a critical induction concentration, below which they undergo the cusp catastrophe [13] (Sec. 3.7) and switch irreversibly to high GFP expression. All cell trajectories should begin on the lower hysteresis branch because AHL production is proportion to cell density, so expression always begins low. Instead, we see no catastrophe event.

Measuring time-series GFP per OD as the cells grow to carrying capacity (Figure 14), we see that the catastrophe event (quorum) occurs around OD 0.1 across a wide range of IPTG concentrations. This suggests that LuxI expression must be weakened and/or carrying capacity must be reduced for the cells to exhibit bistability once the population reaches steady state.

*Learn.* We are in the process of constructing an RBS-variant library of SOC plasmid variants using the weaker Weiss RBS sequences (B0032, B0033) in the LuxI transcription unit [14]. Additionally, we are developing less nutrient-dense cell media to modulate carrying capacity (Figure 15). Both of these changes are expected to move the quorum threshold of the system closer steady-state population density.

Once we have identified parameters under which the steady-state population density of the cells is in or near the bistable region, we will optimize IPTG concentration of SOC behavior by testing many concentrations in parallel using a gradient-generating microfluidic device (Figure 16).

### 3 FUTURE WORK AND THERAPEUTIC APPLICATIONS

With appropriate tuning, the circuit may be used in biomedical applications where a transient behavior is desired. For example, bacteria transformed with the circuit could couple the delivery of a chemotherapy drug to LuxI expression. Delivering the correct concentration of IPTG, which can be delivered systemically, to a patient would allow the cells to evolve to the critical point, where they will deliver the drug transiently and stop once every cell has avalanched. In a strain that colonizes the tumor microenvironment, such as *E. coli Nissile*, this would allow for precise dosage control inside the tumor via systemic delivery of a harmless sugar.

**REFERENCES**

213				
214	[1]	BAK, P., AND TANG, C.	Earthquakes as a self-organized critical phenomenon. <i>Journal of Geophysical Research: Solid Earth</i> 94, B11 (1989), 15635–15637.	
215				
216	[2]	BAK, P., TANG, C., AND WIESENFELD, K.	Self-organized criticality. <i>Physical review A</i> 38, 1 (1988), 364.	
217				
218	[3]	BIRKELAND, K. W., AND LANDRY, C. C.	Power-laws and snow avalanches. <i>Geophysical Research Letters</i> 29, 11 (2002), 49–1–49–3.	
219				
220	[4]	BOLIVAR, F., RODRIGUEZ, R. L., GREENE, P. J., BETLACH, M. C., HEYNEKER, H. L., BOYER, H. W., CROSA, J. H., AND FALKOW, S.	Construction and characterization of new cloning vehicle. ii. a multipurpose cloning system. <i>Gene</i> 2, 2 (1977), 95–113.	
221				
222				
223	[5]	CLEWELL, D. B., AND HELINSKI, D. R.	Supercoiled circular dna-protein complex in escherichia coli: purification and induced conversion to an open circular dna form. <i>Proceedings of the National Academy of Sciences</i> 62, 4 (1969), 1159–1166.	
224				
225	[6]	GIBSON, D. G., YOUNG, L., CHUANG, R.-Y., VENTER, J. C., HUTCHISON III, C. A., AND SMITH, H. O.	Enzymatic assembly of dna molecules up to several hundred kilobases. <i>Nature methods</i> 6, 5 (2009), 343–345.	
226				
227				
228				
229				
230				
231				
232				
233				
234				
235				
236				
237				
238				
239				
240				
241				
242				
243				
244				
245				
246				
247				
248				
249				
250				
251				
252				
253				
254				
255				
256				
257				
258				
259				
260				
261				
262				
263				
264				
265				
	[7]	GILLESPIE, D. T.	The chemical langevin equation. <i>The Journal of Chemical Physics</i> 113, 1 (2000), 297–306.	266
	[8]	KNIGHT, T.	Idempotent vector design for standard assembly of bio-bricks.	267
	[9]	MALAMUD, B. D., MOREIN, G., AND TURCOTTE, D. L.	Forest fires: An example of self-organized critical behavior. <i>Science</i> 281, 5384 (1998), 1840–1842.	268
	[10]	MIKHEYEV, A. S., AND TIN, M. M.	A first look at the oxford nanopore minion sequencer. <i>Molecular ecology resources</i> 14, 6 (2014), 1097–1102.	269
	[11]	SMOLKE, C. D.	Building outside of the box: igem and the biobricks foundation. <i>Nature biotechnology</i> 27, 12 (2009), 1099–1102.	270
	[12]	SORNETTE, A., AND SORNETTE, D.	Self-organized criticality and earthquakes. <i>EPL (Europhysics Letters)</i> 9, 3 (1989), 197–202.	271
	[13]	STROGATZ, S. H.	<i>Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering</i> , 3 ed. CRC Press, Boca Raton, FL, 2024. See Section 3.7.	272
	[14]	WEISS, R., KNIGHT JR, T. F., AND SUSSMAN, G. J.	Cellular computation and communication using engineered genetic regulatory networks., 2004.	273
				274
				275
				276
				277
				278
				279
				280
				281
				282
				283
				284
				285
				286
				287
				288
				289
				290
				291
				292
				293
				294
				295
				296
				297
				298
				299
				300
				301
				302
				303
				304
				305
				306
				307
				308
				309
				310
				311
				312
				313
				314
				315
				316
				317
				318

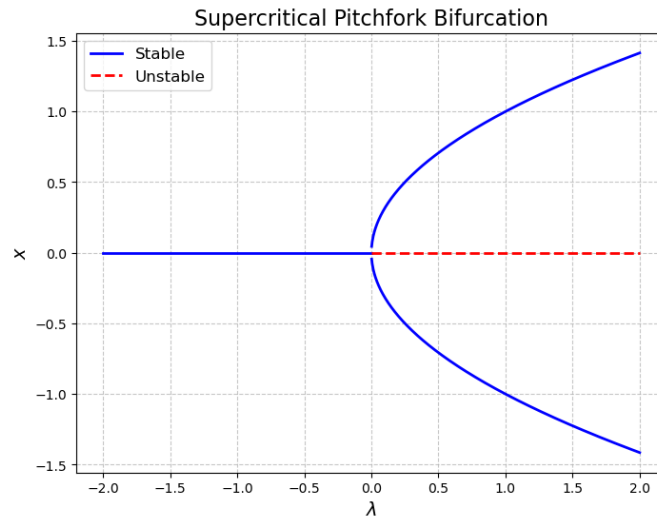


Figure 1: Bifurcation diagram for the supercritical pitchfork normal form  $\dot{x} = \lambda x - x^3$ . At  $\lambda = 0$ , a supercritical pitchfork bifurcation occurs, where a single stable fixed point splits into two stable fixed points and an unstable fixed point. Near this bifurcation point, the Jacobian  $J = \lambda - 3x^2$  becomes small at the stable states, giving rise to critical slowing down and enhanced sensitivity to perturbations, which underlies long-range correlations in the stochastic system.

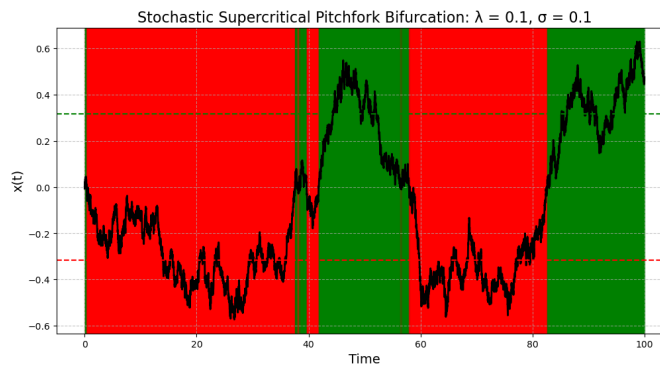
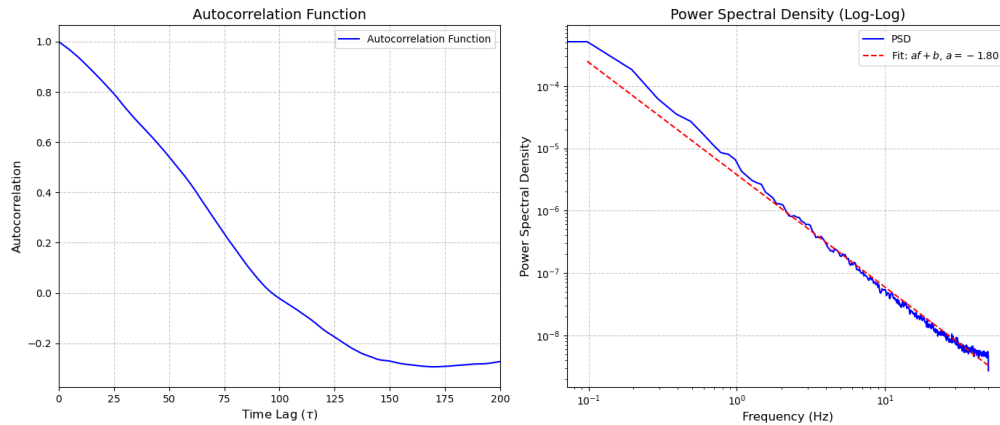


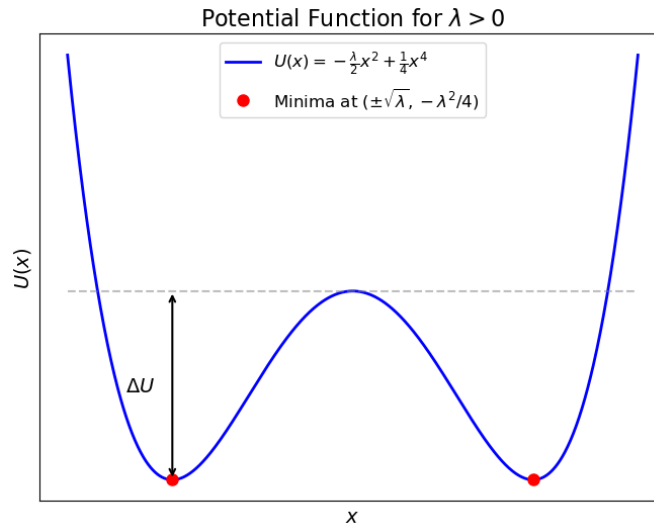
Figure 2: Stochastic switching in the noisy supercritical pitchfork system with  $\lambda > 0$  and small  $\sigma$ . The system exhibits two metastable states at  $x^* = \pm\sqrt{\lambda}$  and randomly switches between them. Fast fluctuations occur within each metastable well, while rare noise-driven transitions produce a slow time scale. This separation of time scales is characteristic of systems near a bifurcation point and contributes to scale-invariant dynamics.

Table 1: Table Heading

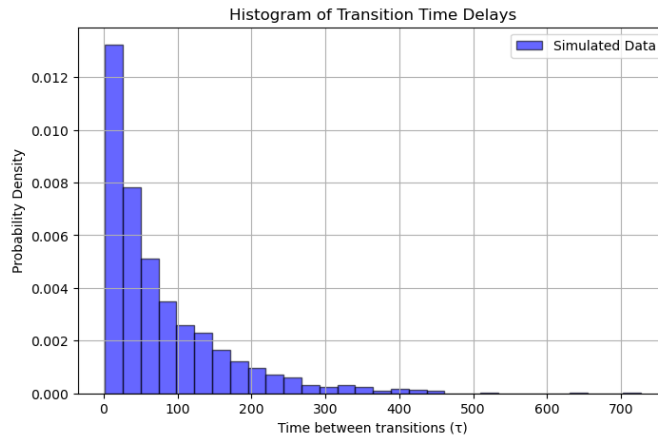
Transcription Unit	Promoter	RBS	CDS	Terminator
LuxI	R0062	B0030	C0061+LAA	B0010
AiiA	R0010	B0034	C0160	B0010
LacI+LuxR	J23102	B0032, B0034	C0012, C0062	B0015
sfGFP	R0062	B0034	E0040+LAA	ECK120029600



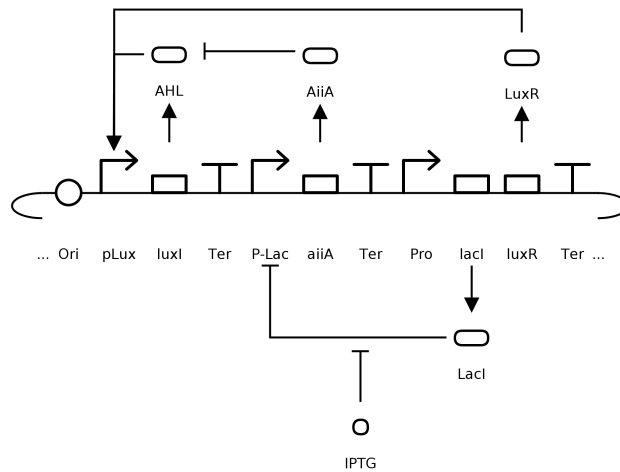
**Figure 3: Autocorrelation  $\rho_{XX}(\tau)$  and corresponding power spectral density  $S_{XX}(\omega) = \mathcal{F}\{\rho_{XX}(\tau)\}(\omega)$  for the stochastic pitchfork system near criticality. The autocorrelation decays slowly, approximately as  $\rho_{XX}(\tau) \sim \tau^{-\beta}$  with  $\beta \approx 0$ , indicating long-range temporal correlations. The Fourier transform yields  $S_{XX}(\omega) \sim 1/\omega$ , i.e., pink noise, a hallmark of self-organized criticality.**



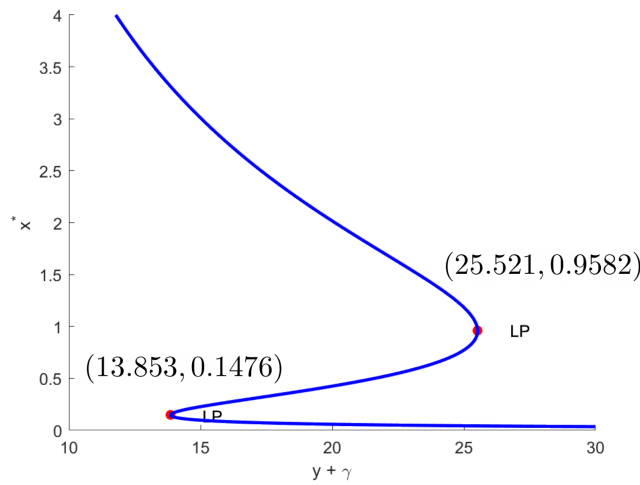
**Figure 4: Potential function  $U(x) = \frac{1}{4}x^4 - \frac{\lambda}{2}x^2$  corresponding to  $\dot{x} = \lambda x - x^3$ . The two wells correspond to metastable states, while the unstable fixed point at  $x = 0$  defines an activation energy barrier  $\Delta U = \lambda^2/4$ . Noise-driven transitions across this barrier govern the slow time scale of the dynamics and lead to broadly distributed switching times.**



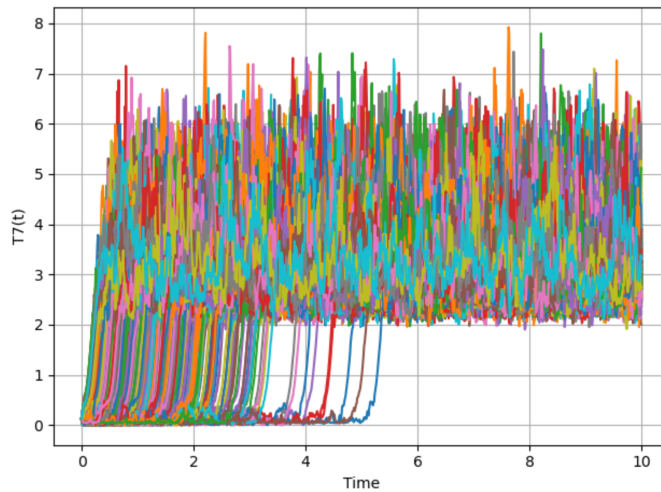
**Figure 5: Distribution of switching times  $T$  between metastable states. From Kramers' rate theory,  $T \sim \exp(\lambda^2/2\sigma^2)$ , which implies a heavy-tailed distribution  $P(T) \sim T^{-\alpha}$  over a wide range of scales. This approximate power-law behavior reflects scale invariance in the temporal dynamics and is another hallmark of self-organized criticality.**



**Figure 6: SBOL Visual diagram of our circuit. We express LuxI on the luxI promoter with constitutive expression of LuxR. To make degradation tunable, we express AiiA on the inducible Lac promoter; by varying the concentration of the inducer molecule IPTG, we can vary the steady-state concentration of AiiA. We also ensure that the AHL dynamics are faster than the AiiA dynamics by marking LuxI for degradation with an LAA tag.**



**Figure 7: Circuit bifurcation diagram. Using continuation software, we find that the saddle-node bifurcations occur when  $(y + \gamma, x) = (13.853)$  and  $(y + \gamma, x) = (25.521, 0.9582)$  for  $\beta = 1$  and  $\alpha = 50$ . To cause SOC, we conduct subsequent simulations with  $\gamma = 10$  and  $C = 3.853$ , with the same  $\alpha$  and  $\beta$  values.**



**Figure 8: Stochastic avalanche behavior in the biochemical system. Behavior reflects the expected time-distributed avalanche events from low to high expression. These events are irreversible due to hysteresis.**

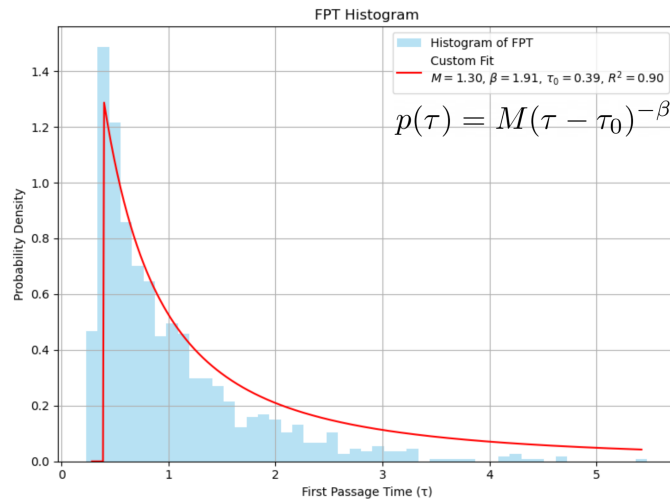


Figure 9: Distribution of avalanche times in the biochemical model (in simulations starting from the critical point). We observe a power law distribution with  $R^2 = 0.90$ .

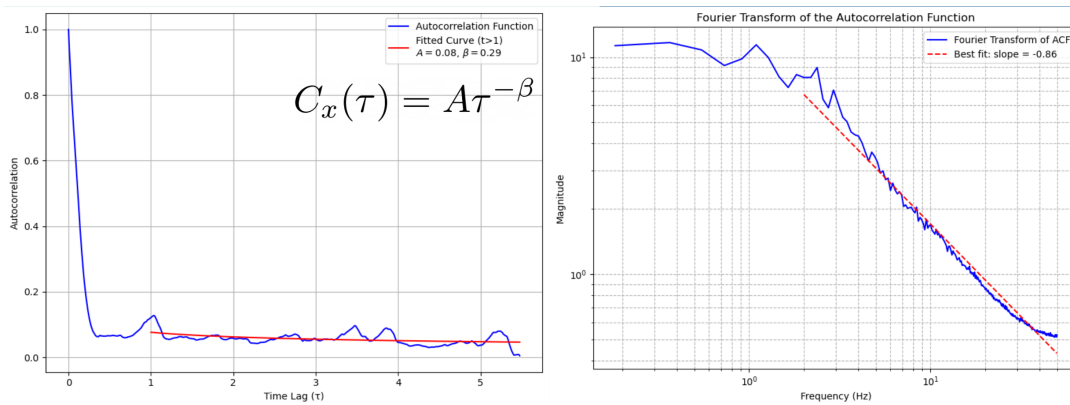


Figure 10: Long-range correlations and pink noise in the SOC genetic circuit. Correlation up to 5 time units is long-range in this context because the system avalanches in 5 time units or fewer in almost all cases—5 time units borders on the slower time scale of the stochastic switching. The power spectral density decays like  $f^{-\alpha}$  for  $0 < \alpha = 0.86 < 2$ , so the PSD is a pink noise.

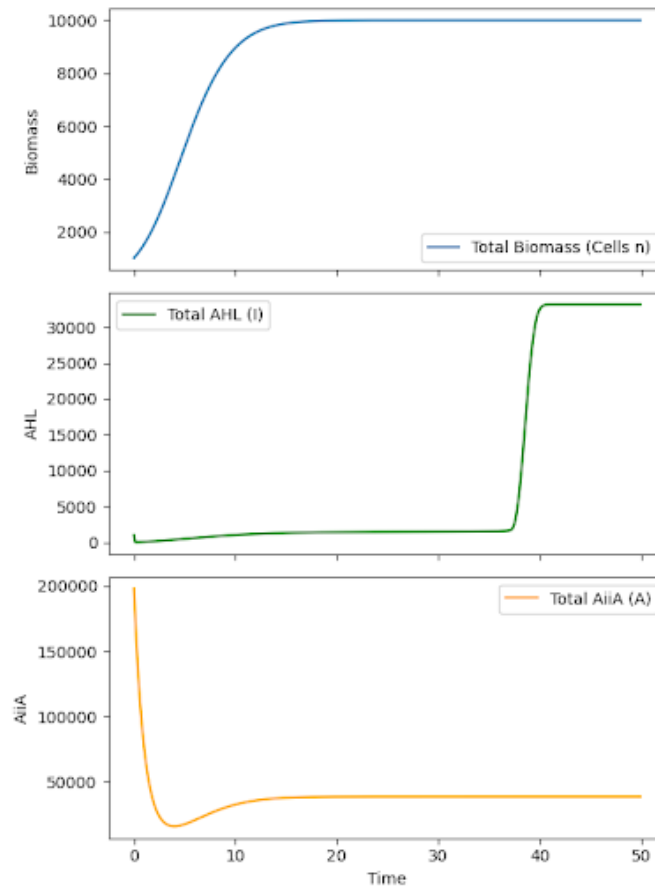
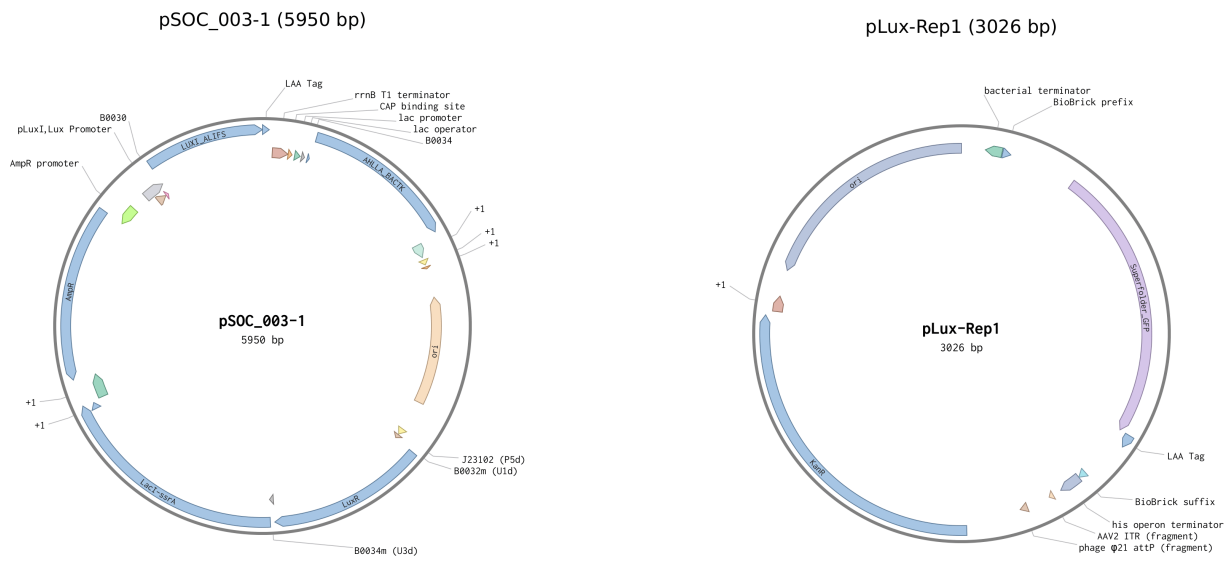
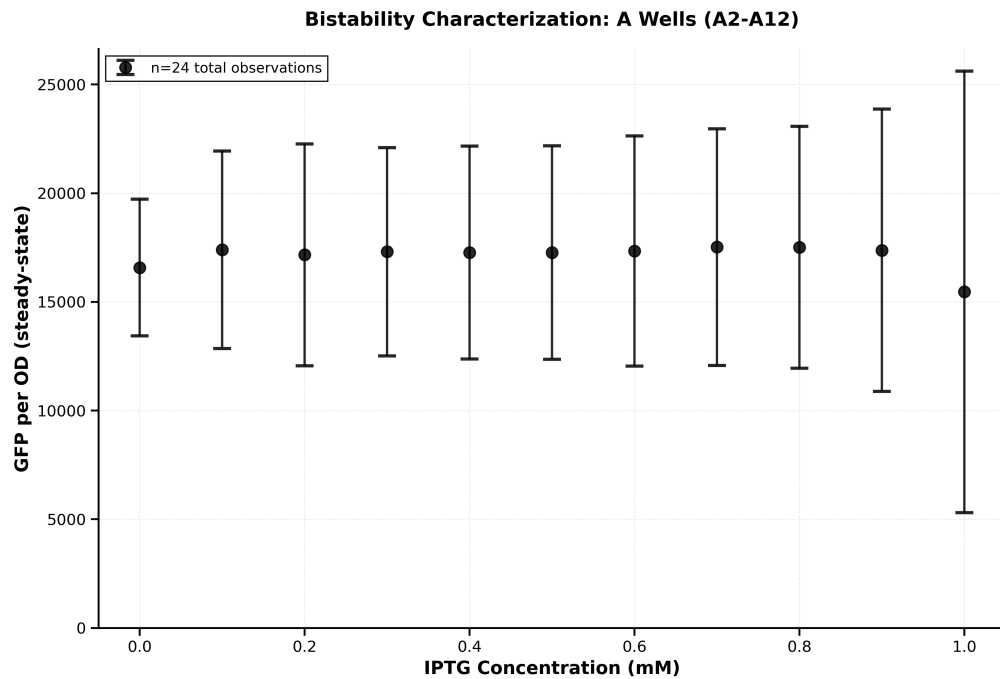


Figure 11: Simulated SOC circuit.



**Figure 12: Plasmid maps for the genetic circuit and reporter constructs. (Left) Circuit plasmid assembled in the pSB1A2 backbone, containing a pMB1 origin of replication and Carbenicillin resistance cassette. The circuit comprises the selected BioBrick parts (Table 1) arranged to implement the designed regulatory architecture. (Right)  $P_{lux}$  reporter plasmid used to quantify promoter activity, containing a ColE1 origin and Kanamycin resistance cassette, with a fluorescent reporter downstream of the lux promoter.**



**Figure 13: Steady-state GFP response across IPTG induction levels. Cell density-normalized GFP (GFP/OD) at steady state versus IPTG concentration. The design predicts a cusp catastrophe below a critical IPTG threshold, but no such transition is observed; GFP varies smoothly instead. Cultures were grown to carrying capacity and averaged across replicates. The lack of a discontinuity indicates the system does not reach the bistable regime.**

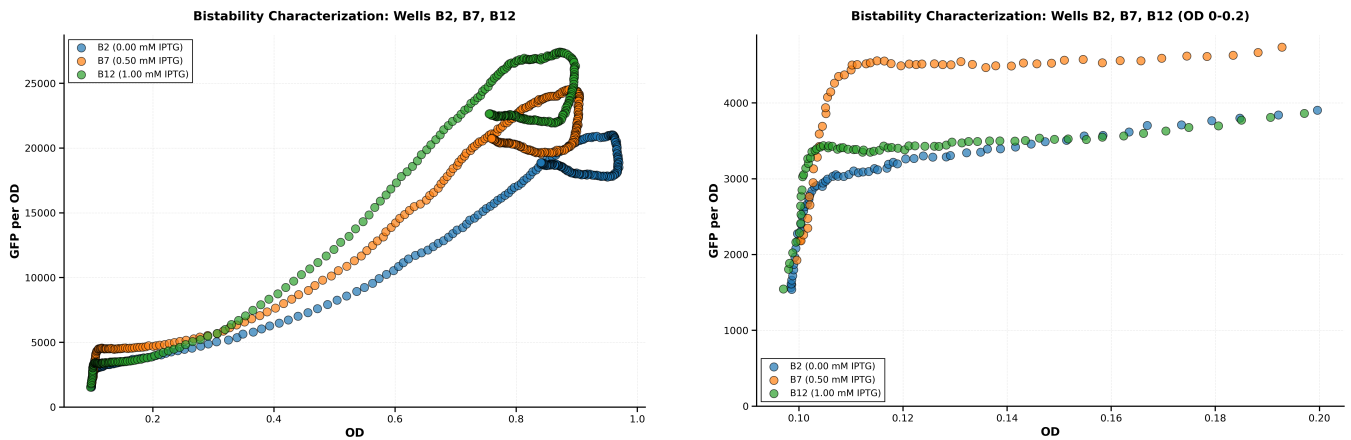


Figure 14: Time-series GFP/OD during population growth. (Left) GFP fluorescence normalized by OD as a function of time as cultures grow to carrying capacity across a range of IPTG concentrations. All trajectories exhibit a sharp increase in GFP/OD as the population approaches quorum. (Right) Zoomed-in view of the transition region highlighting the onset of quorum sensing behavior, which occurs around OD ~ 0.1 across all tested IPTG concentrations.

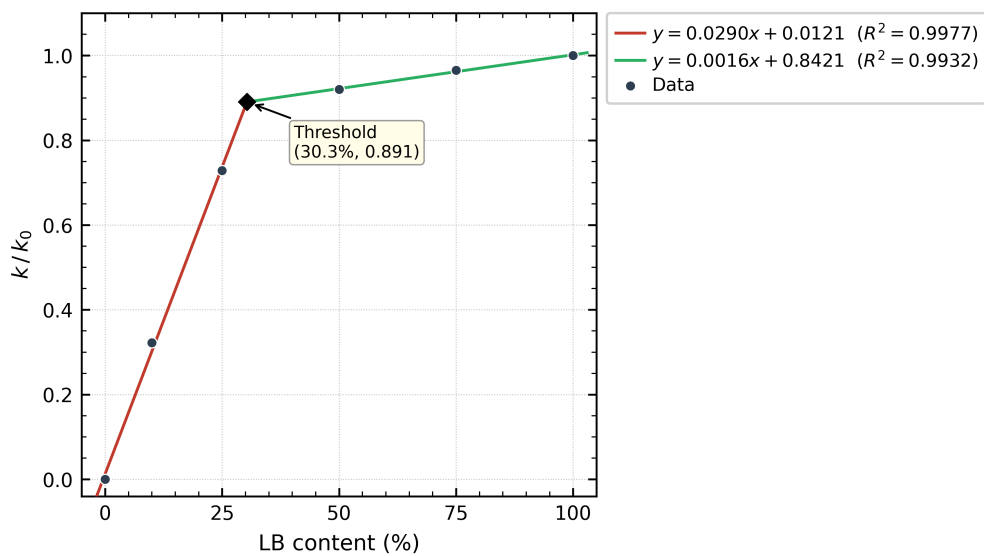
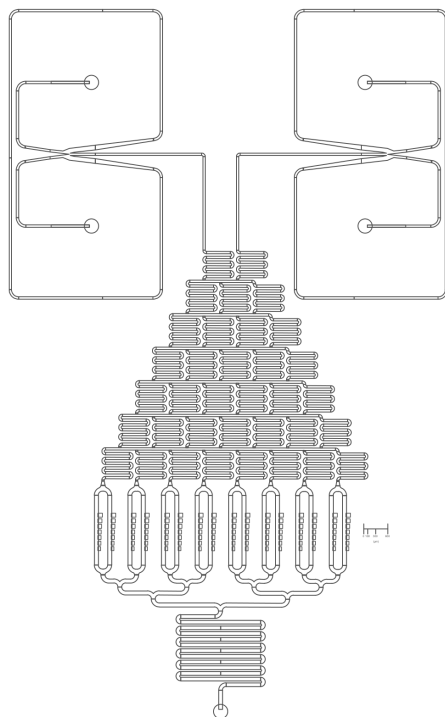


Figure 15: Normalized carrying capacity as a function of percent LB in media. We diluted LB in osmolarity-matched NaCl solution and measured carrying capacity for varying LB concentrations. Osmolarity matching ensures that dilute LB solutions do not become hypotonic. We see that carrying capacity varies linearly with LB concentration up to a threshold around 30.3% LB. Beyond the threshold, nutrient concentration maintains a positive relationship with carrying capacity, but it is weaker. We postulate that this change is due to cell crowding, which does not occur at lower maximum  $OD_{600}$ .



**Figure 16: Gradient-generating Microfluidics Chip for IPTG induction tuning.** Inlets (top) are mixed together binomially to produce a gradient in the traps (bottom-middle, center). Media flow out the outlet (bottom, center). The four inlets allow for mid-experiment media swaps as well as gradient generation. Traps are  $1.2\mu\text{m}$  in height to allow only a monolayer of cells. Channels are  $16\mu\text{m}$  in height to no impede flow.

# Physics-Informed Inference of Gene-Expression Kinetics Using bioCRNpyler

Zachary Hastings<sup>1</sup> and Lukas Buecherl<sup>1</sup>

<sup>1</sup> Biological Engineering, Utah State University, USA

lukas.buecherl@usu.edu

## INTRODUCTION

Predictive modeling remains a major bottleneck in the *design-build-test-learn* (DBTL) cycle for synthetic biology, particularly when circuit behavior depends on reaction kinetics that are difficult to measure directly from experimental data. In this context, reaction kinetics refer to the rates governing processes such as transcription, translation, and molecular dilution or degradation that shape the dynamic behavior of synthetic gene-expression systems. Better computational approaches for inferring these parameters could reduce experimental trial-and-error, improve model-guided design, and support the engineering of more complex and reliable synthetic biosystems.

*Physics-informed neural networks* (PINNs) [1] provide a promising approach because they combine flexible function approximation with mechanistic constraints by incorporating governing differential equations directly into training. This is especially attractive in synthetic biology, where measurements such as fluorescence often provide only indirect information about the underlying kinetics.

In this work, we investigate whether physics-informed learning can recover mechanistically meaningful kinetic parameters from simulated gene-expression data generated in bioCRNpyler [2]. Simulated data provide an initial test case because the ground-truth parameter values are known, enabling direct evaluation of parameter recovery. Using an IPTG-inducible transcriptional unit, we infer the translation and dilution rate constants,  $k_{tl}$  and  $k_{dil}$ , from simulated expression outputs. We compare a baseline PINN with a *physics-informed autoencoder* (PIAE), which uses a constrained latent representation to improve parameter inference, and a *physics-informed long short-term memory network* (PILSTM), which leverages time-series structure in simulated trajectories to recover kinetic parameters. Together, these models test whether neural networks can learn not only circuit outputs, but also the hidden parameters governing synthetic gene-expression behavior.

## METHODS

Simulated gene-expression data were generated in bioCRNpyler using an IPTG-inducible transcriptional unit composed of a promoter, ribosome binding site, coding sequence, and terminator. Simulations were performed across multiple IPTG

concentrations to produce final mRNA, protein, and YFP outputs, yielding a synthetic benchmark dataset for evaluating physics-informed approaches to kinetic parameter inference.

The primary learning task was to infer the translation and dilution rate constants,  $k_{tl}$  and  $k_{dil}$ , from the simulated expression outputs. For the initial PINN model, 40 simulated conditions were randomly divided into training and test sets using an 80/20 split. The PINN was implemented in Python using PyTorch [3] as a feedforward neural network with two hidden layers of 64 neurons each and trained for 10,000 epochs. The training objective combined data fitting with a physics-based loss term derived from the governing system dynamics.

We next developed a PIAE to improve parameter recovery. In this formulation, the encoder mapped simulated outputs into a lower-dimensional latent representation, and the decoder reconstructed the original data while incorporating physics constraints to encourage the latent variables to encode mechanistically meaningful kinetic information. For the PIAE, simulated data with artificial noise were normalized prior to training, and optimization was performed using Adam with a mean squared error loss augmented by a physics-based loss term.

To further evaluate sequence-aware architectures for kinetic inference, we developed a PILSTM. Unlike the PINN and PIAE experiments, the PILSTM was evaluated on a separate dataset generated from a simple constitutive transcriptional unit. This model consisted of four LSTM layers with 64 hidden units each and was trained using Adam optimization with a mean squared error objective. In addition to inferring the rate constants  $k_{tl}$  and  $k_{dil}$ , the PILSTM was trained to predict mRNA and DNA concentrations from simulated data.

Model performance was evaluated based on two criteria: agreement between predicted and simulated YFP responses, and accuracy of the inferred kinetic parameters relative to the known ground-truth values from the bioCRNpyler simulations.

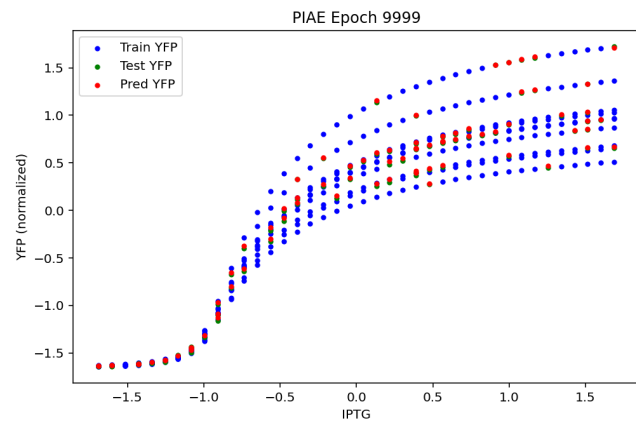
## RESULTS

The baseline PINN successfully learned the simulated IPTG-dependent gene-expression response and produced accurate predictions of the observable YFP output across both training and test conditions. Training and test losses decreased substantially over the course of optimization, indicating that the

model was able to capture the overall nonlinear input–output behavior of the simulated transcriptional unit.

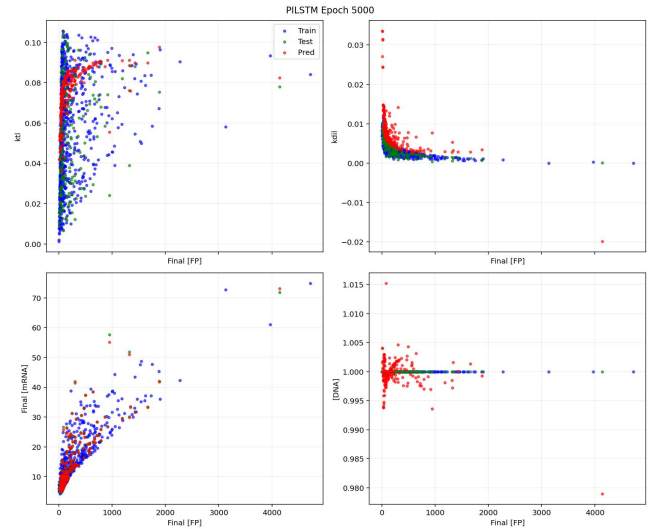
However, recovery of the underlying kinetic parameters lagged behind improvement in output prediction. While the PINN reproduced the YFP response relatively early in training, convergence of the inferred translation and dilution rate constants required substantially longer optimization. The final PINN recovered  $k_{tl}$  to within approximately 10–20% of the ground-truth value, suggesting that accurate prediction of observable outputs does not necessarily imply equally accurate recovery of the underlying mechanism. To address this limitation, we next evaluated a PIAE.

Compared with the baseline PINN, the PIAE showed moderate improvement in recovering the target kinetic parameters while maintaining agreement with the simulated data. In the noisy simulated setting, the PIAE recovered  $k_{tl}$  and  $k_{dil}$  with errors in the range of approximately 2–20%. As shown in Figure 1, the PIAE maintained close agreement with the simulated YFP response while improving recovery of mechanistically meaningful kinetic parameters. Together, these results suggest that latent-space physics constraints may improve mechanistic identifiability beyond what is achieved by output fitting alone.



**Figure 1: Predicted and simulated normalized YFP responses across IPTG conditions for the physics-informed autoencoder (PIAE). The model maintained close agreement with the simulated dose–response data while improving recovery of mechanistically meaningful kinetic parameters.**

Preliminary results with the PILSTM further suggest that sequence-aware architectures can capture relationships between kinetic parameters and simulated species trajectories (Figure 2). Using 1,000 noisy simulations with perturbations in  $k_{tl}$  and  $k_{dil}$ , the PILSTM recovered kinetic parameters and species concentrations with good overall agreement. Models trained without the physics-based loss also achieved strong predictive performance, but occasionally produced nonphysical outputs such as negative values, highlighting the role



**Figure 2: Predicted and simulated kinetic rates and species concentrations across fluorescent protein concentrations for the physics-informed long short term memory network (PILSTM). The model recovered mechanistically meaningful kinetic parameters and accurate species concentrations.**

of physics constraints in maintaining mechanistic plausibility. The PILSTM also converged in fewer training epochs than the baseline feedforward PINN, consistent with the suitability of recurrent architectures for time-series inputs.

## DISCUSSION AND FUTURE WORK

These results suggest that physics-informed learning can recover mechanistically meaningful kinetic parameters from simulated gene-expression data, while also highlighting an important limitation: accurate output prediction does not necessarily imply accurate parameter recovery. In particular, the baseline PINN reproduced the simulated YFP response before converging to the ground-truth kinetic values, indicating a gap between output fitting and mechanistic identifiability. The improved parameter recovery observed with the PIAE suggests that physics constraints in a structured latent space may help address this challenge.

Future work will extend this framework to experimental measurements. *E. coli* cells will be transformed with simple transcriptional units encoding fluorescent proteins, and fluorescence trajectories will be collected using a plate reader. These data will be used as inputs to the proposed models, and inferred kinetic parameters will be evaluated by forward simulation in bioCRNpyler and related Python tools. Additional directions include BINN [4] formulations, more complex circuit architectures, and robustness under increased biological noise and experimental variability.

## ACKNOWLEDGMENTS

Z.H. and L.B. were supported by the Department of Biological Engineering at Utah State University.

## REFERENCES

- [1] Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. Finite basis physics-informed neural networks (FBPINNs): A scalable domain decomposition approach for solving differential equations. *Advances in Computational Mathematics*, 49(4):62, August 2023. doi:10.1007/s10444-023-10065-9.
- [2] William Poole, Ayush Pandey, Andrey Shur, Zoltan A. Tuza, and Richard M. Murray. BioCRNpyler: Compiling chemical reaction networks from biomolecular parts in diverse contexts. *PLOS Computational Biology*, 18(4):e1009987, April 2022. doi:10.1371/journal.pcbi.1009987.
- [3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. <https://arxiv.org/abs/1912.01703v1>, December 2019.
- [4] Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, Taylor E. Arnoff, Jihye Park, William C. Hahn, and Eliezer M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, October 2021. doi:10.1038/s41586-021-03922-4.

# Codon Pair Optimization Web Tool

Dimitris Papamichail  
Bang Chiem  
papamicd@tcnj.edu  
The College of New Jersey  
Ewing, PA, USA

Georgios Papamichail  
New York College  
Athens, Greece

## 1 INTRODUCTION

Synonymous codon selection is far from random, it varies systematically across organisms and tissues, exerting a measurable influence on translation kinetics, protein folding, mRNA stability, and ultimately gene expression levels [7, 8, 13]. Measures such as the codon adaptation index (CAI), introduced by Sharp and Li (1987), provided a quantitative framework for measuring the extent to which a given gene's codon usage mirrors that of highly expressed reference genes, and it became a widely used predictor of recombinant protein expression levels. However, codon usage bias considers each codon position independently, ignoring the context imposed by its neighbors.

Codon pair bias (CPB) refers to the statistically significant over- or under-representation of specific codon pairs in coding sequences relative to what would be predicted from the individual frequencies of the constituent codons alone. Initially observed in *Escherichia coli* by Gutman and Hatfield (1989) [6], it was brought into focus by Coleman et al. in 2008[3], who introduced the concept of codon pair deoptimization (CPD) as a rational strategy for virus attenuation. By systematically replacing wild-type codon pairs with synonymous but statistically under-represented pairs – without altering the amino acid sequence or overall codon usage – the authors generated attenuated poliovirus variants that elicited robust immune responses in animal models while displaying dramatically reduced pathogenicity. The approach, termed Synthetic Attenuated Virus Engineering (SAVE), demonstrated that CPB exerts an independent and substantial effect on translational output that cannot be attributed to individual codon frequencies alone. This work established CPD as a promising platform for live-attenuated vaccine development and has since been applied to influenza virus, respiratory syncytial virus, and other clinically relevant pathogens [10].

Beyond vaccine design, CPB has significant implications for recombinant protein production. Optimizing codon pair usage in heterologous expression constructs has been reported to improve protein yield, and there is growing evidence that the rate and rhythm of translation elongation – influenced by the local codon pair context – affects co-translational protein folding and the acquisition of correct tertiary structure [8].

Numerous computational tools have been developed to aid the design of synthetic coding sequences that optimize codon usage for a target expression host. Codon context optimization tools are less common, and platforms for multi-objective sequence optimization that included codon context such as EuGene [5] and COOL [2] are now defunct. To our knowledge, the field now lacks a tool that enables the optimization of protein coding sequences for codon context preferences, and we have created a simple web tool to address this specific need 1.

## 2 METHODS

*Web Tool Architecture and Accessibility.* Out Codon Pair Bias Optimization tool is implemented as an online web application. The interface accepts a protein-coding DNA sequence as input and returns an optimized synonymous sequence designed to maximize codon pair bias (CPB) score for the selected target organism. Output is displayed directly within the browser interface, and the optimized sequence can be downloaded in FASTA format for further processing or immediate use in downstream applications such as gene synthesis ordering or expression vector design.

*Codon Pair Usage Tables.* Codon pair preferences for the two supported target organisms – *Homo sapiens* and *Mus musculus* – were derived from the Consensus Coding Sequence (CCDS) dataset [4, 12], a high-confidence, manually curated set of protein-coding regions that are consistently annotated. For each organism, a codon pair usage table was constructed by exhaustively enumerating all adjacent codon pairs across the coding sequences in the CCDS set, recording the observed frequency of each of the 3,721 possible sense codon pairs ( $61 \times 61$ , excluding stop codons). From these observed frequencies, a codon pair score (CPS) was calculated for each codon pair following the formulation introduced by Coleman et al [3]. This formulation isolates the codon pair preference that is independent of amino acid composition and individual codon usage frequency. The overall CPB score of a coding sequence is then computed as the arithmetic mean of the CPS values across all adjacent codon pairs in the sequence. More information about CPB scores and their interpretation can be found in [11].

## Codon Pair Bias Optimization Tool

### Required Parameters

Input File

Output File

### Optional Parameters

Target Organism

### STDOUT

```
Setting organism to default (human)
Setting simulated annealing rounds to default (500,000)

>>>INPUT INFORMATION

Sequence to be optimized:
ATGGGTGCTCAGGTTTCATCACAGAAAGTGGGGCCACATGAAAACTCAAATAGAGCGTATGGTAGTCTACCATTAATTACACCACCTAATTATTATA
GAGATTCAAGTACAGTAAAGCGGCTTCCGAAACAGGACTTCTCTCAAGACCCTTCCAAAGTTACACGAGCCCATCAAGGATGCTCTGATAAAAAAGCCCAAT
GCTAAACTCGCAAACATAGAGGCTTGGGGTATAGGATAGAGTACTGCAATTAACACTGGGAACTCCACTATAACACACAGAGGGCGGTAATTCGA
GTAGTGCTTATGGGCGTTGGCCTGAATATCTGAGGGACAGCGAAGCAATCCAGTGACACCGCAGAACCCAGCTCGCTGCATGACAGTTTATA
CGCTAGACACCGTGTCTTGGACGAAAGAGTGGGAGGGTGGTGGGAAG

Sequence length (DNA bases): 450
Sequence cpb before optimization: -0.0914270796308769

>>> OUTPUT INFORMATION

Sequence cpb after optimization: 0.234726663744111

>Optimized Sequence -- cpb = 0.234726663744111
ATGGGGCGCAGGTGCTGTCGCAGAAAGTGGTGTCTCATGAAAAACAGCAACAGAGCTTATGGTCTTCAACAATAAATTATACACCATTAAATTATTATC
GAGACAGTGTCTCAAATGCTGCTTCAAAAACAAGATTTTCTCAAGACCCTGCAAGTTACACAGAGCCAAATAAAGATGCTCTAATAAAAAAGCACCACAT
GCTAAATCTCAAACATTAAGAGCTGCGGGTATTGACAGGGTCTCGAGCTGACGCTGGGAAACAGCACCATCACACCAGGAGGCTGCAAAATTCGA
GTAGTAGCTTATGGAAGATGGCAGAAATTTAAGAGACTCGGAGGCCAACCTGTGGACAGCAACAGAGCTGATGTGGCGGCTGCAGGTTCTACA
CGCTGGACACTGTCTCTGGACCAAGGAAAGTGGTGGTGGGAAG
```

### STDERR

### Output Files

- [output.fa](#)

Drag to outliner or Upload  
Close

Figure 1: Codon Pair Bias Optimization Tool interface

**Optimization Algorithm: Simulated Annealing.** Our tool utilizes simulated annealing, a metaheuristic technique with demonstrated effectiveness on large combinatorial search problems that are Travelling Salesman Problem (TSP)-like [1, 9]. Synonymous codons are selected randomly to be swapped according to a probabilistic criterion governed by a temperature parameter that decreases over the course of the optimization. Swapping of synonymous codons guarantees that both the amino acid encoding and codon bias of the original input sequence are strictly preserved.

The number of simulated annealing iterations is user-configurable via a parameter in the web interface, with a permitted range of 50,000 to 3,000,000 iterations, limits that accommodate a spectrum of practical input sequence lengths.

**Output.** Upon completion of the optimization run, the tool displays the optimized nucleotide sequence directly in the browser alongside the CPB score of the input sequence and the CPB score of the optimized output, allowing the user to assess the magnitude of improvement achieved. The optimized sequence is additionally available for download in standard FASTA format.

The codon pair optimization tool can be accessed at: [https://compbio.hpc.tcnj.edu/cpb\\_max\\_tool](https://compbio.hpc.tcnj.edu/cpb_max_tool)

### 3 ACKNOWLEDGEMENTS

The authors acknowledge use of the ELSA high performance computing cluster at The College of New Jersey for housing the tools and supporting the research reported in this paper. This cluster is funded in part by the National Science Foundation under grant numbers OAC-1826915 and OAC-2320244.

## REFERENCES

- [1] ČERNÝ, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications* 45, 1 (1985), 41–51.
- [2] CHIN, J. X., CHUNG, B. K.-S., AND LEE, D.-Y. Codon optimization onLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* 30, 15 (2014), 2210–2212.
- [3] COLEMAN, J. R., PAPAMICHAIL, D., SKIENA, S., FUTCHER, B., WIMMER, E., AND MUELLER, S. Virus attenuation by genome-scale changes in codon pair bias. *Science* 320, 5884 (2008), 1784–1787.
- [4] FARRELL, C. M., GOLDFARB, T., WESOLOWSKI, M., ASTASHYN, A., CUMMINS, B., WALLIN, C., PRUITT, K. D., AND RIDDICK, L. D. New developments in the consensus coding sequence (CCDS) project. *Nucleic Acids Research* 52, D1 (2024), D315–D321.
- [5] GASPAR, P., OLIVEIRA, J. L., FROMMLET, J., SANTOS, M. A. S., AND MOURA, G. EuGene: Maximizing synthetic gene design for heterologous expression. *Bioinformatics* 28, 20 (2012), 2683–2684.
- [6] GUTMAN, G. A., AND HATFIELD, G. W. Nonrandom utilization of codon pairs in *Escherichia coli*. *Proceedings of the National Academy of Sciences* 86, 10 (1989), 3699–3703.
- [7] IKEMURA, T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology* 146, 1 (1981), 1–21.
- [8] KIMCHI-SARFATY, C., OH, J. M., KIM, I. W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V., AND GOTTESMAN, M. M. A “silent” polymorphism in the *MDR1* gene changes substrate specificity. *Science* 315, 5811 (2007), 525–528.
- [9] KIRKPATRICK, S., GELATT, C. D., AND VECCHI, M. P. Optimization by simulated annealing. *Science* 220, 4598 (1983), 671–680.
- [10] MUELLER, S., COLEMAN, J. R., PAPAMICHAIL, D., WARD, C. B., NIMNUAL, A., FUTCHER, B., SKIENA, S., AND WIMMER, E. Live attenuated influenza virus vaccines by computer-aided rational design. *Nature Biotechnology* 28, 7 (2010), 723–726.
- [11] PAPAMICHAIL, D., LIU, H., MACHADO, V., GOULD, N., COLEMAN, J. R., AND PAPAMICHAIL, G. Codon context optimization in synthetic gene design. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15, 2 (2018), 452–459.
- [12] PRUITT, K. D., HARROW, J., HARTE, R. A., WALLIN, C., DIEKHANS, M., MAGLOTT, D. R., SEARLE, S., FARRELL, C. M., LOVELAND, J. E., RUEF, B. J., HART, E., SUNER, M.-M., LANDRUM, M. J., AKEN, B., AYLING, S., BAERTSCH, R., FERNANDEZ-BANET, J., CHERRY, J. L., CURWEN, V., AND HAUSSLER, D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* 19, 7 (2009), 1316–1323.
- [13] SHARP, P. M., AND LI, W.-H. The codon adaptation index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15, 3 (1987), 1281–1295.

# Novel Microfluidic Chips for Massively Parallelised Biological Data Acquisition

Vicente Trelles Fernandez  
Harrison Steel  
vicente.trellesfernandez@wolfson.ox.ac.uk  
harrison.steel@eng.ox.ac.uk  
University of Oxford

## 1 INTRODUCTION

### Throughput and miniaturisation

Large-scale single-cell datasets generated by massively parallelised measurements are essential to understand complex cellular physiology, perform high throughput screens, and increasingly, train machine-learning models. Because laboratory space and resources cannot scale with throughput, such measurements must be performed in miniaturised platforms such as microfluidic systems.

A widely used microfluidic device is the mother machine [6], which consists of a long main channel lined with short perpendicular single-cell-width “trenches”. The “mother cell” is stably positioned at the end of the trench and can be imaged over long time frames. Multiple generations of its progeny can also be transiently imaged as they populate the trench before escaping into the main channel (Figure 1).

### Mother machine microscopy

Mother cells are commonly imaged with phase-contrast and fluorescence microscopy, but other setups such as confocal and Raman microscopy are possible. Crucially, mother machines produce temporally-connected data, unlike alternative high throughput methods such as flow cytometry or droplet microfluidics, which yield single-time-point data. Mother cells can be actuated in real time: chemical inputs can be switched using valves or inlets [4]; temperature can be regulated through dedicated setups; magnetic stimuli can be applied externally [1]; and light can be delivered to individual cells using a digital micromirror device (DMD) [5].

All this has made the mother machine an increasingly attractive platform for training machine-learning models. In 2024, Lugagne *et al.* used a deep neural network to predict the response of an optogenetic system in *E. coli* and control it via DMD-based actuation [5]. In 2025, Hallström *et al.* used vision transformers and convolutional neural networks for rapid identification of clinically relevant bacteria [3].

### Backport mother machine designs

Standard mother machine designs suffer three key limitations. First, trench loading requires centrifugation or highly

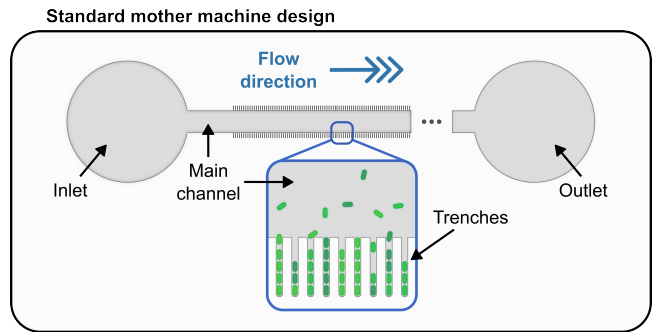


Figure 1: Mother machine devices consist of a long main channel lined with trenches. Mother cells at the dead end of trenches can be stably imaged over time. Not to scale.

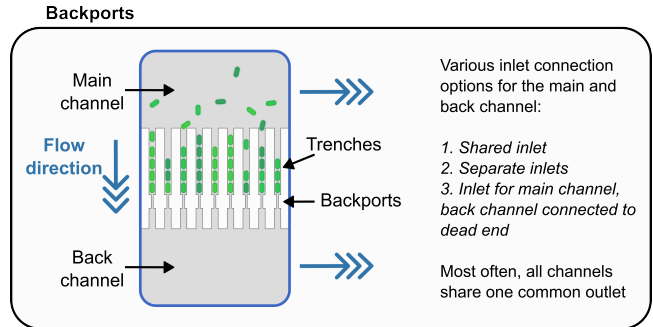


Figure 2: Backports enable active medium flow through trenches by connecting them to a back channel.

concentrated cultures. Second, motile cells cannot be stably entrapped. Third, nutrients and chemicals rely on diffusion to reach the mother cell, resulting in a lower effective concentration and delayed exposure. These limitations stem from the dead-end geometry of trenches and can be overcome by adding a “backport” that connects them to a “back channel” (Figure 2). Backports are too narrow to allow cell to pass but enable media to flow, removing dependency on diffusion and loading and confining cells, even motile ones.

Backport designs were first published in 2017 [2], but their adoption in academic research has been incredibly limited. Informal discussions suggest several groups have attempted

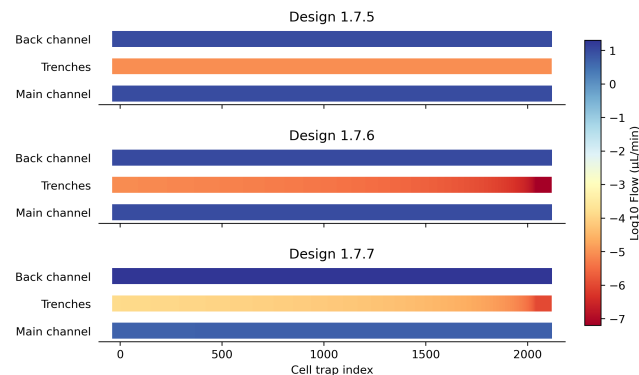
to fabricate such devices but abandoned them due to substantial challenges in modelling, fabrication, and operation. Here, we propose a design and implementation aimed to enable the widespread adoption of backport mother machines, allowing researchers to produce high-quality single-cell datasets.

## 2 RESULTS

We have designed, modelled, and optimised the fabrication and operation of backport mother machines. We perform proof-of-concept experiments to showcase the unique advantages of our designs and protocols.

### Device flow modelling

Understanding the flow profile through trenches is critical, as large variations could lead to uneven growth conditions and confound experimental results. Computational fluid dynamics (CFD) is the standard approach for modelling flow in microfluidic devices, but is poorly suited to mother machines due to their large number of trenches. We instead leverage their repetitive geometry to develop an analytical model. Using the Navier–Stokes equations and conservation of mass and energy, we reduce the problem to a system of  $3n_{\text{trench}} + 4$  linear equations which can be efficiently solved in Julia using a linear solver. We find that homogeneous trench flow can only be achieved under certain parameter constraints. Poor choices can generate inter-trench flow variations of several orders of magnitude. We also show that flow rates can be adjusted through other design parameters (Figure 3).

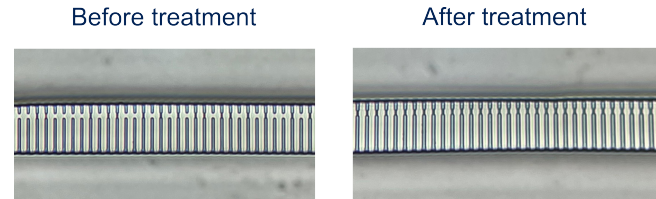


**Figure 3: Flow simulations for three backport devices. Only design 1.7.5 yields homogeneous trench flow. Design 1.7.7 parameters results in higher trench flow than 1.7.6.**

### Device fabrication

Mother machine devices are fabricated in PDMS from silicon moulds, but the small width of backports makes them prone to collapse in the PDMS replica. We developed a protocol that reliably restores collapsed backports using isopropanol to

expand PDMS followed by rapid drying immediately before bonding (Figure 4). In addition, we show that backport width can be increased by 50% relative to that proposed in the 2017 study, making them substantially more resistant to collapse while still reliably preventing cell passage.



**Figure 4: Recovery of collapsed backports by IPA treatment.**

### Device operation

Operational failures of backport devices often result from backflow or bubbles, which can unload cells from trenches. Through the use of pressure-based pumps, a simplified tubing system, and a custom loading protocol, we prevent backflow and the introduction of bubbles into our setup.

### Biological applications

As a proof-of-concept, we utilise our backport mother machines to study cell differentiation in *B. subtilis*, a motile bacterium with important applications in industrial biotechnology. We aim to use the real time data generated to predict transitions between different cell states. Future work will extend their application to perform large scale screens and directed evolution of proteins.

## REFERENCES

- [1] ABRAHAMS, G., ŠTUHEC, A., SPRENG, V., HENRY, R., KEMPF, I., JAMES, J., SECHKAR, K., STACEY, S., TRELLES-FERNANDEZ, V., ANTILL, L. M., TIMMEL, C. R., MILLER, J. J., INGARAMO, M., YORK, A. G., TETIENNE, J.-P., AND STEEL, H. Quantum spin resonance in engineered proteins for multimodal sensing. *Nature* 649, 8099 (Jan. 2026), 1172–1179.
- [2] BALTEKIN, Ö., BOUCHARIN, A., TANO, E., ANDERSSON, D. I., AND ELF, J. Antibiotic susceptibility testing in less than 30 min using direct single-cell imaging. *Proceedings of the National Academy of Sciences of the United States of America* 114, 34 (Aug. 2017), 9170–9175.
- [3] HALLSTRÖM, E., KANDAVALLI, V., WÄHLBY, C., AND HAST, A. Rapid label-free identification of seven bacterial species using microfluidics, single-cell time-lapse phase-contrast microscopy, and deep learning-based image and video classification. *PLOS ONE* 20, 9 (Sept. 2025).
- [4] KAISER, M., JUG, F., JULOU, T., DESHPANDE, S., PFOHL, T., SILANDER, O. K., MYERS, G., AND VAN NIMWEGEN, E. Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software. *Nature Communications* 9, 1 (Jan. 2018).
- [5] LUGAGNE, J.-B., BLASSICK, C. M., AND DUNLOP, M. J. Deep model predictive control of gene expression in thousands of single cells. *Nature Communications* 15, 1 (Mar. 2024), 2148.
- [6] WANG, P., ROBERT, L., PELLETIER, J., DANG, W. L., TADDEI, F., WRIGHT, A., AND JUN, S. Robust growth of *Escherichia coli*. *Current biology : CB* 20, 12 (June 2010), 1099–1103.

# Microfluidic Physics-Constrained Agent-Based Modeling of Artificial Microbial Biofilms

Nona Hashemi<sup>†</sup>, Mahdi Hasanzadeh<sup>2†</sup>, and Samuel M.D. Oliveira<sup>1\*</sup>

<sup>1</sup> Nanoengineering Department, Joint School of Nanoscience and Nanoengineering, North Carolina A&T State University, Greensboro, NC, USA

<sup>2</sup> Department of Computer Systems Technology, North Carolina A&T State University, Greensboro, NC, USA

## 1. INTRODUCTION

Microfluidic systems have become valuable tools in synthetic biology because they enable precise control over the key factors regulating biofilm development, including flow, mass transport, and physical confinement. Rather than relying on bulk-averaged conditions, these platforms enable the application of defined shear stresses, controlled concentration gradients, and direct observation of structural evolution in well-characterized environments, thereby supporting the systematic design and analysis of biofilm systems [1,2]. Existing models do not yet provide a unified, predictive framework that captures biofilm behavior under the coupled transport and hydrodynamic constraints characteristic of microfluidic systems.

iDynoMiCS 2.0 is widely used for biofilm modeling and captures cell-level dynamics through a combination of transport equations (partial differential equations, PDEs), bulk processes (ordinary differential equations, ODEs), and agent-based rules defining growth, division, and detachment [3]. Its mechanical framework operates under low-Reynolds-number ( $Re$ ) assumptions, which are appropriate for microbial systems ( $Re < 1000$ ). However, from a biological modeling perspective, the tool has not yet been designed to represent microfluidic environments with laminar flow ( $Re \sim 0.1-100$ ). For example, their tool lacks a Navier–Stokes solver, which is required to account for channel geometry and to consider the effects of shear stress on nutrient supply and potential surface interactions [4], both of which are central constraints on the dynamics of microfluidic systems. Namely, shear influences detachment, velocity profiles regulate the delivery of carbon and oxygen to the biofilm, and axial transport generates spatial gradients that cannot be captured under well-mixed assumptions [5]. Therefore, many iDynoMiCS simulations depend on default or empirically adjusted parameters that are not intrinsically connected to the underlying flow conditions. This work explicitly links those parameters to the underlying system physics. Continuum-scale simulation tools, such as COMSOL Multiphysics, enable detailed characterization of fluid flow, shear stress, and transport processes in microfluidic systems [6].

Here, we introduce a workflow that integrates COMSOL-derived quantities, including shear stress, near-wall transport, and substrate availability, and translates them into iDynoMiCS inputs that are compatible with the model’s structure (Figure 1). This work represents Stage 1 of a three-step roadmap toward full coupling between the two modeling frameworks. Rather than estimating parameters heuristically, we derive them precisely from the fundamental physics of the microfluidic system. Next, this approach is demonstrated using an aerobic *Escherichia coli* K-12 biofilm model that includes both carbon and dissolved oxygen. Bacterial parameters obtained from the literature [7].

We observed that incorporating COMSOL-derived parameters leads to substantial differences in predicted biofilm growth compared to the default configuration (Figure 2). These differences arise from the enforcement of realistic transport and hydrodynamic constraints characteristic of microfluidic environments. This framework provides a step toward bridging continuum-scale

transport modeling with agent-based biofilm simulations under experimentally relevant conditions.

## 2. METHODS

### 2.1 Microfluidic geometry and analytical framework

The system consists of a realistic straight rectangular PDMS microchannel [8] with a length of  $L = 10$  mm, a width of  $w = 2$  mm, and a height of  $h = 150$   $\mu\text{m}$ . The flow rate is ped at  $Q = 5$   $\mu\text{L}/\text{min}$  using LB medium, approximated as water ( $\rho \approx 1000$   $\text{kg}\cdot\text{m}^{-3}$ ,  $\mu \approx 10^{-3}$   $\text{Pa}\cdot\text{s}$ ). The mean velocity is calculated from  $u_{\text{mean}} = Q/(wh)$ , yielding  $u_{\text{mean}} = 0.278$   $\text{mm}\cdot\text{s}^{-1}$  after unit conversion. The hydraulic diameter is  $D_h = 0.279$  mm, and the Reynolds number is  $Re \approx 0.078$ , confirming creeping (Stokes) flow conditions. This supports the force-balance assumptions used in iDynoMiCS. An analytical estimate of wall shear stress gives  $\tau_w \approx 0.011$  Pa; however, relevantly, all transport and detachment parameters used in this work are obtained from COMSOL results rather than analytical approximations.

### 2.2 COMSOL Multiphysics simulation

The geometry was implemented in COMSOL 6.1 using the Laminar Flow and Transport of Diluted Species modules. The inlet was defined with a fully developed velocity profile, the outlet at zero pressure, and all walls as no-slip boundaries. The bottom surface (biofilm interface) was assigned a carbon consumption flux following Michaelis–Menten kinetics,  $J = V_{\text{max}}C/(K_m + C)$ . A swept mesh with wall-refinement (minimum element size of 2  $\mu\text{m}$ ) was used. The model was solved at steady state. Outputs exported to CSV included wall shear stress along the biofilm surface, carbon concentration at the interface, and near-wall velocity. These values were used for parameter translation (Table 1).

### 2.3 Parameter translation workflow

COMSOL outputs were translated into iDynoMiCS inputs, as summarized in Table 1. Wall shear stress was converted to a decay coefficient using a linear relation  $b = \alpha\tau_w$ , with  $\alpha = 0.020$   $\text{h}^{-1}\cdot\text{Pa}^{-1}$ , giving  $b = 0.027$   $\text{h}^{-1}$ . The carbon concentration at the biofilm surface was used as the boundary condition (984  $\text{mg}\cdot\text{L}^{-1}$ ). The boundary layer thickness was reduced to 8.2  $\mu\text{m}$  based on near-wall transport, and carbon diffusivity was adjusted to 560  $\mu\text{m}^2\cdot\text{s}^{-1}$  using a porosity factor of 0.75. All values were directly implemented in the iDynoMiCS XML configuration file.

### 2.4 iDynoMiCS simulations

Biofilm growth was simulated for *E. coli* K-12 using parameters consistent with the XML configuration:  $\mu_{\text{max}} = 0.69$   $\text{h}^{-1}$ ,  $K_s = 0.00427$   $\text{mg}\cdot\text{L}^{-1}$ ,  $K_{O_2} = 0.15$   $\text{mg}\cdot\text{L}^{-1}$ ,  $m_s = 0.05$   $\text{h}^{-1}$ , and density 1.105  $\text{g}/\text{mL}$ .

Two cases were simulated: (A) default parameters and (B) COMSOL-constrained parameters (Table 1). The domain was 200  $\mu\text{m} \times 100$   $\mu\text{m}$  (2D) with cyclic lateral boundaries. Simulations were run for 24 hours with five replicates per condition. Carbon and oxygen concentrations in the constrained case were 984  $\text{mg}\cdot\text{L}^{-1}$  and 6.4  $\text{mg}\cdot\text{L}^{-1}$ , respectively.

<sup>†</sup> Both authors contributed equally to this research.

\* Corresponding authors (smdoliveira@ncat.edu)

### 3. PRELIMINARY RESULTS

To evaluate the COMSOL–iDynoMiCS coupling workflow, two simulations were performed using the same *E. coli* K-12 biofilm model, differing only in transport and detachment parameters. Simulation A used default iDynoMiCS values, while Simulation B used COMSOL-derived parameters obtained through the workflow described in the Methods (Table 1).

The translated parameters introduce several key changes. The COMSOL-derived wall shear stress ( $\tau_w \approx 1.35$  Pa) was mapped to a decay coefficient of  $b = 0.027 \text{ h}^{-1}$  using a linear relation  $b = \alpha\tau_w$  with  $\alpha = 0.020 \text{ h}^{-1}\cdot\text{Pa}^{-1}$ . The effective boundary layer thickness decreased from  $32 \mu\text{m}$  to  $8.2 \mu\text{m}$ , reflecting enhanced near-wall transport. The effective boundary layer thickness decreased from  $32 \mu\text{m}$  to  $8.2 \mu\text{m}$ , reflecting enhanced near-wall transport. This behavior is consistent with shear-driven transport in laminar microfluidic systems [9]. Carbon diffusivity was reduced from  $700 \text{ }\mu\text{m}^2\cdot\text{s}^{-1}$  using a porosity factor of 0.75, and the effective boundary carbon concentration decreased from  $1200$  to  $984 \text{ mg}\cdot\text{L}^{-1}$  due to axial depletion. Despite the reduction in boundary carbon, the COMSOL-constrained simulation produced substantially greater biofilm growth. As shown in Figure 2, biofilm thickness reached  $4.20 \mu\text{m}$  in Simulation B compared to  $2.48 \mu\text{m}$  in Simulation A, corresponding to a 69% increase. The divergence becomes noticeable after approximately 8 hours and continues to increase over the remainder of the simulation.

The primary driver of this difference is the reduced decay rate. Lowering  $b$  from  $0.050$  to  $0.027 \text{ h}^{-1}$  reduces biomass loss due to detachment, thereby increasing net accumulation over time. Because growth follows exponential dynamics, even a modest change in the net growth rate ( $\mu_{\text{net}} = \mu_{\text{max}} - b$ ) leading to a significant difference in total biomass after 24 hours

A quantitative comparison of all parameters is provided in Table 2. Together, these results indicate that incorporating hydrodynamic constraints affects predicted biofilm development and that the translation workflow provides a consistent method for transferring COMSOL-derived physics into iDynoMiCS simulations.

### 4. PLATFORM ROADMAP AND FUTURE WORK

The long-term goal is to move from a one-way parameter translation to a fully coupled simulation framework in which iDynoMiCS and COMSOL exchange information during runtime. This will allow biofilm growth and fluid flow to evolve together under consistent physical constraints. The overall roadmap is shown in Figure 1. Stage 1, presented in this work, establishes the parameter translation workflow and demonstrates that incorporating COMSOL-derived inputs changes biofilm predictions in a measurable way.

Stage 2 introduces an iterative coupling scheme. In that, iDynoMiCS periodically exports the biofilm geometry, which is then used to update the channel boundaries in COMSOL. The flow and concentration fields are re-solved, and the updated transport conditions are fed back into iDynoMiCS. This creates a quasi-steady coupling between growth and transport without requiring full runtime integration.

Stage 3 extends this approach to fully dynamic co-simulation. Using the COMSOL LiveLink API together with a custom geometry export module from iDynoMiCS, both models operate simultaneously and exchange information at each time step. This enables direct coupling between biofilm morphology and the evolving hydrodynamic field.

Experimental validation will be carried out alongside all stages using PDMS microfluidic devices [8]. Biofilm structure will be quantified using confocal laser scanning microscopy, allowing direct comparison between simulation predictions and measured three-dimensional morphology.

### 5. CONCLUSION

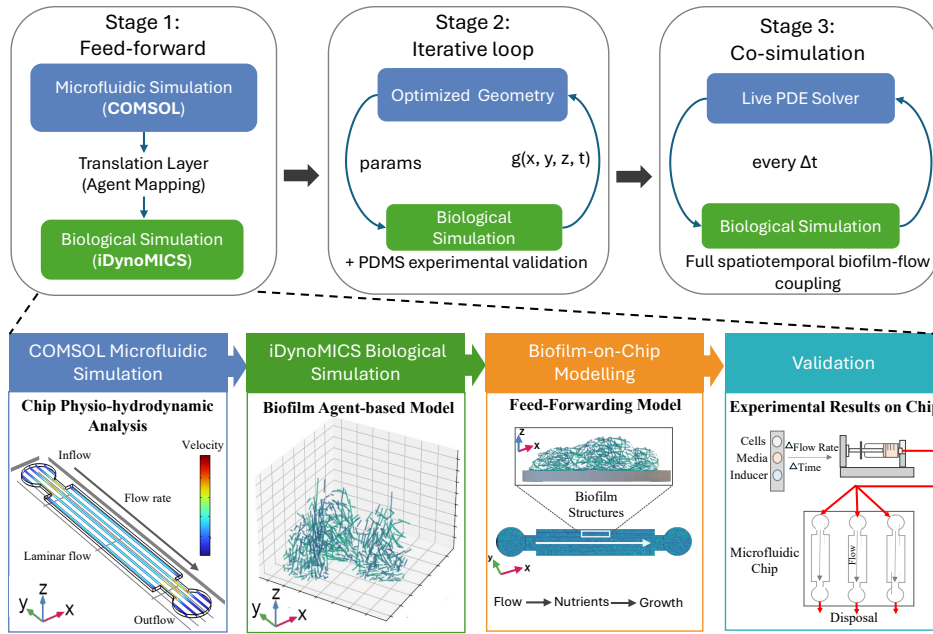
In this work, we studied a direct link between microfluidic physics and agent-based biofilm modeling by introducing a parameter translation workflow between COMSOL and iDynoMiCS. Instead of relying on default or tuned values, the model parameters are derived from the actual flow and transport conditions in the channel. Applied to a representative microfluidic system, our simulations demonstrated significant differences in predicted biofilm growth, compared to the default setup (as shown in Figure 2), from incorporating physically consistent transport and detachment conditions. Finally, this work corresponds to Stage 1 of a broader roadmap toward fully coupled biofilm–flow simulation, which provides a practical way to make iDynoMiCS simulations consistent with microfluidic environments. Although, in Stage 1, the workflow is based on general transport relationships and COMSOL outputs, other future stages have the promise to extend the platform to support varying channel geometries and biological systems without modifying its core modeling framework. This makes it usable as a design tool for connecting simulation and experiment in microfluidic biofilm studies.

### ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) Awards 4500004978, 2327906, and ECCS-1542174; the Defense Advanced Research Projects Agency (DARPA) Award D24AP00330-00; and the North Carolina Collaboratory Award C676-26-KHMSI.

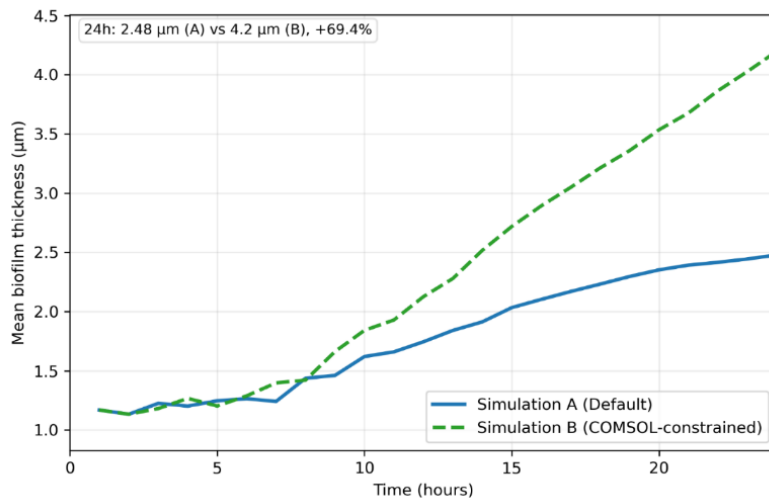
### REFERENCES

- [1] Yuan, L., Straub, H., Shishaeva, L., and Ren, Q. *Microfluidics for biofilm studies*. Annual Review of Analytical Chemistry, 16, 1, 2023.
- [2] Oliveira, S. M. D. and Densmore, D. *Hardware, software, and wetware codesign environment for synthetic biology*. BioDesign Research, 2022, 2022.
- [3] Cockx, B. J. R., Foster, T., Clegg, R. J., et al. *Is it selfish to be filamentous in biofilms? Individual-based modeling links microbial growth strategies to morphology using the new, modular iDynoMiCS 2.0*. PLoS Computational Biology, 20, 2, 2024.
- [4] Karimifard, S., Li, X., Elowsky, C., and Li, Y. *Modeling the impact of evolving biofilms on flow in porous media inside a microfluidic channel*. Water Research, 2021.
- [5] Thomen, P., Robert, J., Monmeyran, A., Bitbol, A.-F., Douarche, C., and Henry, N. *Bacterial biofilm under flow: First a physical struggle to stay, then a matter of breathing*. PLoS ONE, 12, 4, 2017.
- [6] Mahdavi, R., Hashemi-Najafabadi, S., Ghiass, M. A., et al. *Microfluidic design for in-vitro liver zonation—a numerical analysis using COMSOL Multiphysics*. Medical & Biological Engineering & Computing, 62, 2024.
- [7] Smaluch, K., Wollenhaupt, B., Steinhoff, H., Kohlheyer, D., Grünberger, A., and Dusny, C. *Assessing the growth kinetics and stoichiometry of Escherichia coli at the single-cell level*. Engineering in Life Sciences, 23, e2100157, 2023.
- [8] Blanco-Cabra, N., López-Martínez, M. J., Arévalo-Jaimes, B. V., et al. *A new BiofilmChip device for testing biofilm formation and antibiotic susceptibility*. npj Biofilms and Microbiomes, 7, 2021.
- [9] Wittig, C., Wagner, M., Vallon, R., Crouzier, T., van der Wijngaart, W., Horn, H., and Bagheri, S. *The role of fluid friction in streamer formation and biofilm growth*. npj Biofilms and Microbiomes, 11, 1, 2025.



**Figure 1. Multi-stage framework for integrating microfluidic physics with agent-based biofilm modeling.** Three-stage roadmap for coupling COMSOL-based microfluidic simulations with iDynoMiCS biofilm modeling and experimental validation. Stage 1 (this work) implements a one-way, feed-forward workflow that translates hydrodynamic and transport outputs from COMSOL into iDynoMiCS parameters. Stage 2 introduces an iterative loop in which biofilm growth updates the channel geometry and transport conditions. Stage 3 extends this to fully coupled co-simulation with bidirectional exchange between flow and biofilm dynamics at each time step. The lower panel illustrates the full pipeline from microfluidic simulation and agent-based modeling to biofilm-on-chip prediction and experimental validation.

**Biofilm thickness comparison: Default vs COMSOL-constrained parameters**



**Figure 2. *In silico* simulation of biofilm growth and mean biofilm thickness over 24 hours.** Simulation A (default parameters, blue solid line) vs. Simulation B (COMSOL-constrained parameters, green dashed line). At 24h: 2.48 µm (A) vs. 4.20 µm (B), representing a 69% increase in biofilm thickness when using COMSOL-derived transport parameters.

Table 1. Parameter translation: COMSOL Multiphysics outputs → iDynoMiCS 2.0 protocol inputs

COMSOL Output	Translation	Assumption	iDynoMiCS Input	Value
Wall shear stress $\tau_w$ [Pa]	$b = \alpha \cdot \tau_w$	$\alpha = 0.020 \text{ h}^{-1}/\text{Pa}$ (from COMSOL calibration)	Decay rate $b$ [ $\text{h}^{-1}$ ]	0.027
Carbon at wall $C_{\text{wall}}$ [mg/L]	$\text{CBC} = C_{\text{wall}}$	Accounts for axial depletion (Péclet effects)	Boundary carbon [mg/L]	984
Near-wall velocity $u_{\text{wall}}$ [m/s]	$\delta = D / \text{km}$	Laminar mass transfer correlation	Boundary layer $\delta$ [ $\mu\text{m}$ ]	8.2
Boundary layer $\delta$ [ $\mu\text{m}$ ]	$D_{\text{eff}} = \varepsilon \cdot D_w$	$\varepsilon = 0.75$ (biofilm porosity)	Carbon diffusivity [ $\mu\text{m}^2/\text{s}$ ]	560
Mean velocity $\bar{u}$ , geometry	$\text{Re} = \rho \bar{u} D_h / \mu$	Confirms creeping-flow regime	Validation only	$\text{Re} \approx 0.078 \checkmark$

Table 2. Quantitative comparison: default (A) vs. COMSOL-constrained (B) at 24 hours

Parameter	Sim. A (Default)	Sim. B (COMSOL)	Change
Biofilm thickness [ $\mu\text{m}$ ]	2.48	4.20	+69%
Decay rate $b$ [ $\text{h}^{-1}$ ]	0.050	0.027	-46%
Boundary carbon [mg/L]	1200	984	-18%
Boundary layer $\delta$ [ $\mu\text{m}$ ]	32	8.2	-74%
Carbon diffusivity [ $\mu\text{m}^2/\text{s}$ ]	700	560	-20%

# Microbial 3D Bioprinter for Novel Living Materials

Harman Mehta University of Oxford Harman.mehta@eng.ox.ac.uk	Yusuf Ugurluoglu Northumbria University yusuf.ugurluoglu@northumbria.ac.uk	Meng Zhang Northumbria University meng.zhang@northumbria.ac.uk	Martyn Dade-Robertson Northumbria University martyn.dade-robertson@northumbria.ac.uk	Harrison Steel University of Oxford harrison.steel@eng.ox.ac.uk
---	--	--	--	---

## 1 INTRODUCTION

Bacterial cellulose (BC) is a versatile and sustainable biomaterial with high mechanical strength, chemical stability and excellent biocompatibility. It offers several advantages, including a nanofibrous architecture, high water retention capacity and the ability to be readily functionalised, making it suitable for advanced applications [1]. BC is produced in an ultrapure form as an extracellular matrix by cellulose-synthesising bacteria, particularly species within the *Komagataeibacter* genus, resulting in a highly crystalline material [1].

Its unique structural and physicochemical properties have enabled applications across diverse fields, including wound dressings, tissue scaffolds, filtration systems and sustainable textiles. In the textile industry, BC can be processed into bio-based leather alternatives and flexible fabrics, offering a biodegradable and renewable substitute for conventional materials [2].

Furthermore, synthetic biology enables the programming of cellulose-producing strains and engineered co-cultures to generate BC materials with tunable colours, textures and compositions. In parallel, BC can be functionalised through both in-situ and post-synthetic (ex-situ) strategies, including the incorporation of antimicrobial agents, conductive nanoparticles, polymers or other bioactive compounds. Together, these approaches allow the development of hybrid materials with antimicrobial, conductive and stimuli-responsive properties, significantly expanding the potential of BC in smart materials, sensors and wearable technologies [3]. However, its structure and properties remain highly sensitive to culture conditions, limiting reproducibility and scalability. While industrial production typically relies on shallow tray cultures or agitated systems [1], recent work has demonstrated increasingly complex and

patterned BC materials, though largely restricted to the lab scale. Current fabrication techniques still lack precise spatial and temporal control, making it difficult to reliably produce complex, patterned or functionalised materials at scale.

In this project, we develop an automated microbial 3D bioprinter to modulate bacterial cellulose biosynthesis in static cultures with high spatial and temporal precision. The system integrates programmable positioning with multi-channel liquid handling to deliver small molecules, proteins, media or secondary cultures at defined locations and time points, alongside optogenetic control using a projector to pattern light inputs. This enables in situ patterning during growth. This platform provides an automated reproducible approach to fabricating spatially organised, functional living materials by combining synthetic biology, optogenetics and robotic control.

## 2 3D Bioprinter

For reliable, reproducible, automated and scalable production of complex BC pellicles, we are developing a microbial 3D bioprinter that enables precise spatial and temporal control over biosynthesis and pattern formation.

### Framework

The printer is housed within a temperature-controlled, sterile chamber equipped with an integrated filtration system that maintains positive pressure, ensuring that only filtered air enters the enclosure and reducing the risk of contamination. The interior surfaces are finished with a matte black coating to minimise light reflection, providing a controlled optical environment suitable for optogenetic applications (Figure 1A).

The culture sits on a stabilised base that is mechanically isolated from the main frame. This design minimises

disturbances caused by vibrations from the printer or nearby equipment, which is critical for maintaining uniform pellicle formation in static cultures.

The system incorporates a programmable XYZ positioning stage, adapted from an open-source 3D printer, allowing precise and repeatable spatial delivery of liquids, cells or signals to defined locations. This enables controlled patterning and localised modulation of bacterial cellulose growth throughout the culture (Figure 1B).

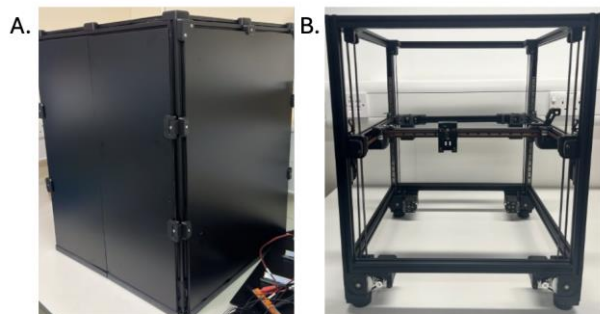


Figure 1. 3D Bioprinter Framework A. Enclosure B. XYZ Positioning system

### Liquid Handling System

The printer incorporates a custom liquid handling system with a dual design, enabling the automated delivery of both low-volume and high-volume inputs (Figure 2A). Small-volume compounds, such as inducers, purified proteins or other chemical modulators, can be dispensed with high precision to defined locations. To minimise reagent loss, these channels use air-driven delivery, avoiding the need to prime the entire line with valuable compounds. In parallel, the system supports larger-volume delivery of media and cell cultures, allowing sustained growth and dynamic modulation of the environment.

The platform also enables the introduction of secondary cell strains at any stage of the culture, supporting co-culture approaches such as the incorporation of pigment-producing bacteria or yeast. Integrated Chi.Bio bioreactors allow these cells to be cultured and grown to a desired phase before being precisely introduced into the bacterial cellulose producing culture (Figure 2A).

To maintain sterility and prevent biofilm formation, the printer includes a cleaning system, flushing channels of the liquid handling system with 70% ethanol or bleach followed by water, ensuring reliable and repeatable operation for long running experiments.

### Projector

The printer also incorporates a mounted projector to enable optogenetic control for pattern formation (Figure

2B). This allows spatially defined light inputs to be projected onto the culture, enabling precise, non-contact modulation of cellular behaviour. By growing co-cultures of BC-producing bacteria with light-responsive bacteria or yeast, pigment production can be locally induced in response to light, enabling the formation of high-resolution patterns during pellicle growth.

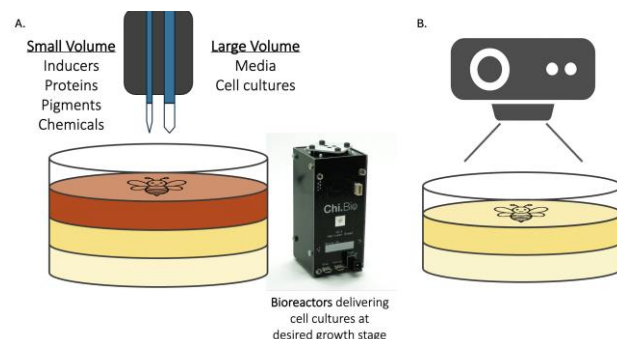


Figure 2. A. Liquid Handling system capable of delivering small and large volume inputs. Secondary cultures can be grown in Chi.Bio bioreactors and delivered during the culture run. B. Optogenetics system projecting image on the BC pellicle

### 3 Concluding Remarks

Overall, this platform shows how combining synthetic biology and robotic control can enable the construction of increasingly complex engineered living materials. By introducing greater control, reproducibility and flexibility into bacterial cellulose production, it becomes possible to fabricate more complex and functionalised BC structures. We see this as a step toward bridging the gap between small-scale experimental systems and more reliable, scalable approaches, opening up new possibilities for translating functionalised BC-based materials into real-world applications.

### References

- [1] Chunyan Zhong. 2020. *Industrial-scale production and applications of bacterial cellulose*. *Frontiers in Bioengineering and Biotechnology* 8 (2020), 605374. <https://doi.org/10.3389/fbioe.2020.605374>
- [2] Rajkishore Nayak, Donna Cleveland, Giang Tran, and Frances Joseph. 2024. *Potential of bacterial cellulose for sustainable fashion and textile applications: A review*. *Journal of Materials Science* 59, 16 (2024), 6685–6710. <https://doi.org/10.1007/s10853-024-09577-6>
- [3] Ahmed K. Saleh, Julie Basu Ray, Mohamed H. El-Sayed, Adel I. Alalawy, Noha Omer, Mahmoud A. Abdelaziz, and Ragab Abouzeid. 2024. *Functionalization of bacterial cellulose: Exploring diverse applications and biomedical innovations: A review*. *International Journal of Biological Macromolecules* 264 (2024), 130454. <https://doi.org/10.1016/j.ijbiomac.2024.130454>

## IWBDA Poster / Talk Abstract

### Manufacturing a high-throughput, label-free platform for rapid screening of drug permeation across *in vitro* plasma membrane models

Niall McIntyre <sup>a</sup>, Joachim Dickhaut <sup>b</sup>, Oscar Ces <sup>a</sup> and Nicholas J Brooks <sup>\*a</sup>

a. Imperial College London, London, United Kingdom

b. BASF, Ludwigshafen, Germany

First and corresponding authors email: nm1916@imperial.ac.uk, n.brooks@imperial.ac.uk\*

#### Abstract

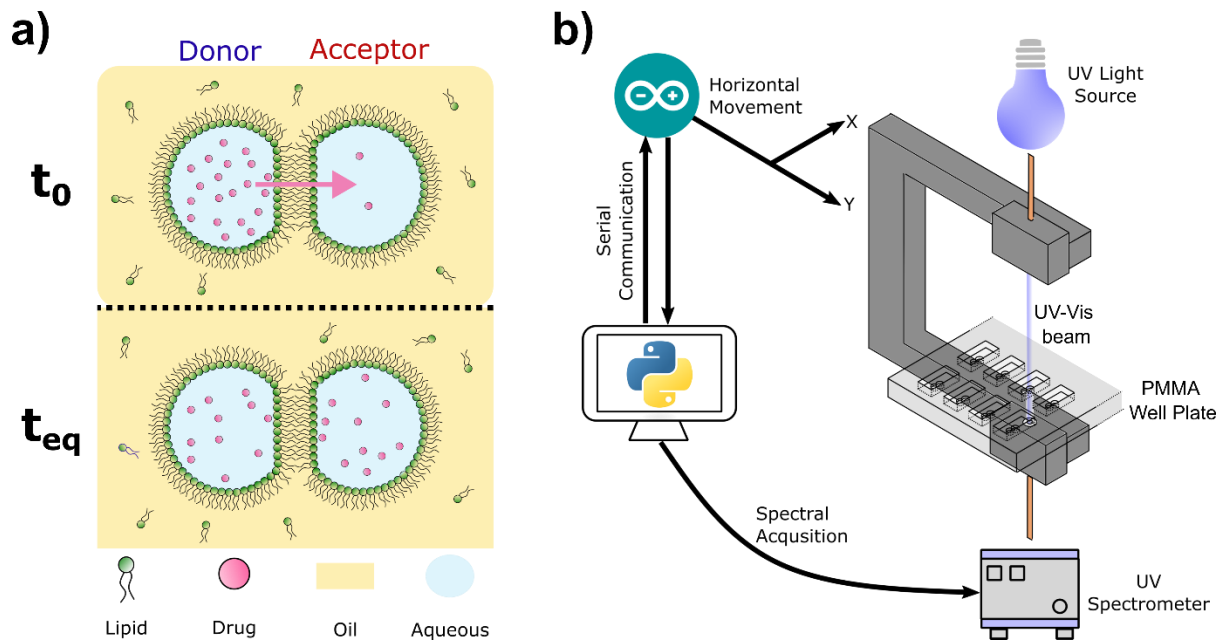
Drug transport across cellular membranes is a critical determinant of drug absorption, however many commercial *in vitro* permeability assays fail to capture the structural complexities of plasma membranes.<sup>1</sup> Droplet interface bilayers (DIBs) have emerged as a powerful bottom-up synthetic biology platform for modelling cell membranes, due to forming a phospholipid bilayers with tuneable membrane composition separating nanolitre volume aqueous compartments.<sup>2</sup> While preliminary work using UV-Vis spectroscopy has been promising as a label-free method to quantify drug permeation in DIBs, its low-throughput has significantly limited its transferability to a commercial setting in early stage drug discovery.<sup>3</sup>

This work presents a custom, automated, high-throughput screening platform to quantify drug permeation across arrays of DIB networks (**Figure 1**). The setup combines CAD design, rapid prototyping, motorised actuators, microcontrollers and custom software with UV-Vis absorption spectroscopy, leading to an 80-fold increase in throughput compared to previous iterations.<sup>3</sup> In addition to screening compound libraries, the setup allows systematic exploration of *in vivo* conditions that strongly influence drug diffusion, including membrane composition and pH gradients, both of which significantly altered diffusion properties in DIBs. Additionally, the modular design of the platform enables measurements to be extended beyond simple two droplet networks, allowing permeant transport across sequential lipid membranes to be modelled in extended one and two dimensions.

Through combining a high-throughput, label-free detection platform with synthetic plasma membranes, the setup enables the production of large datasets of drug permeation across biomimetic cell membranes, which is invaluable for use in predictive models in early-stage drug discovery.

#### References

1. S. Bachler, M. Ort, S. D. Krämer and P. S. Dittrich, *Anal. Chem.*, 2021, **93**, 5137–5144.
2. H. Bayley, B. Cronin, A. Heron, M. A. Holden, W. L. Hwang, R. Syeda, J. Thompson and M. Wallace, *Mol Biosyst*, 2008, **4**, 1191–1208.
3. R. Strutt, F. Sheffield, N. E. Barlow, A. J. Flemming, J. D. Harling, R. V. Law, N. J. Brooks, L. M. C. Barter and O. Ces, *Lab Chip*, 2022, **22**, 972–985.



**Figure 1- Overview of the high-throughput drug permeability screening platform in droplet interface bilayers (DIBs).** a) Schematic of the DIB drug permeability assay. The phospholipid bilayer separating water-in-oil droplets serves as a cell plasma membrane model. If membrane permeable, the drug will diffuse from the donor across the lipid bilayer into the acceptor droplet. b) Cartoon schematic showing how the custom measurement platform is used for high-throughput drug permeability screening in DIBs. The UV detection pathway is integrated with a motorised actuator through custom 3D printed components. The platform is then controlled using a microcontroller and custom software, enabling drug concentration measurements in arrays of sub  $\mu\text{L}$  droplets in the horizontal plane.

# TransfectionWizard: An Integrated Open-Source Platform for Genetic Circuit Design and Automated Mammalian Cell Liposome Transfection for the Lab and Classroom

Evan Holbrook<sup>1,2</sup>, Elizabeth Codd<sup>3</sup>, Ron Weiss<sup>1</sup>, David Kong<sup>2</sup>

<sup>1</sup>MIT Department of Biological Engineering, USA, <sup>2</sup>MIT Media Lab, USA, <sup>3</sup>MIT Department of Computational and Systems Biology, USA

## 1 MOTIVATION

Transfecting nucleic acids into cells is a foundational technique in mammalian synthetic biology, enabling the construction of synthetic genetic circuits, production of valuable biomolecules and medicines, and as a basic research tool as in CRISPR screens. However, generating high throughput workflows is hindered by between-study variability and lack of automated protocol options.

While automated liquid handlers such as the Opentrons OT-2 and the Labcyte Echo have been used to partially address between-study variability and increase throughput [1, 2], existing approaches are time consuming to set up and lack an intuitive graphical interface. Furthermore, no existing tool integrates the full workflow from circuit design and protocol generation through to circuit behavior prediction into a single platform.

Here we report the TransfectionWizard, an integrated open-source platform that addresses these gaps by combining computer-aided design (CAD), automated OT-2 protocol generation, genetic circuit prediction, and automated liposome-based mammalian cell transfection into one accessible tool suitable for both research labs and educational settings.

## 2 RESULTS & DISCUSSION

### 2.1 TransfectionWizard overview

The TransfectionWizard is built around two modular tabs: the Design tab and the Predict tab. The Design tab provides a csv-based interface through which users specify the minimal inputs required to define a transfection experiment,

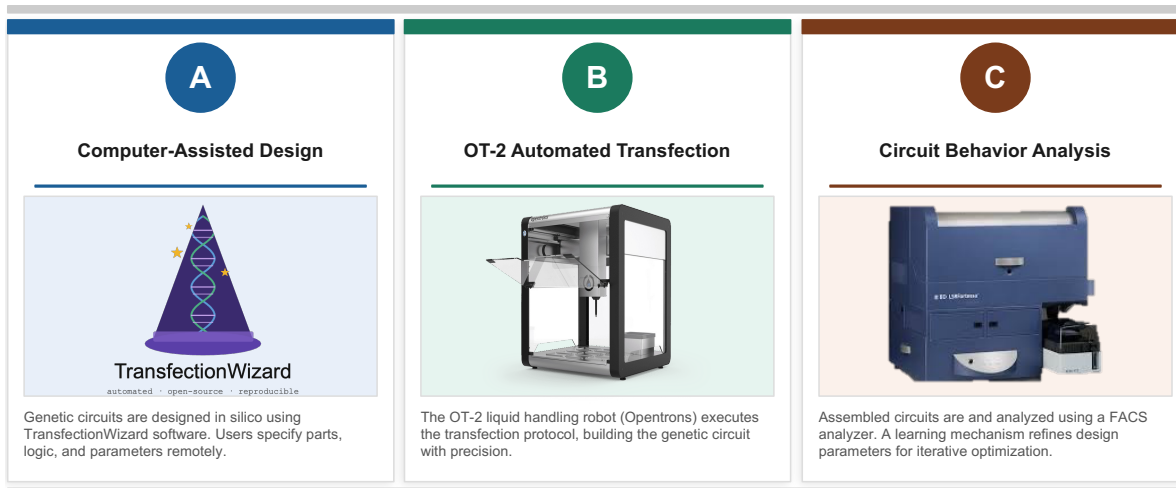
including which genetic parts they'd like to transfect, the mass required, and the concentration of each part. With a single click, the tool generates OT-2 protocol instructions and a formalized protocol capture of the OT-2 deck. The Predict tab accepts the circuit specifications from the Design tab as an input and outputs predicted genetic circuit behavior using any user-specified model. An overview of the TransfectionWizard workflow is shown in Figure 1.

### 2.2 Reproducibility and comparison with manual transfection

To validate the automated protocol, we compared the strength of synthetic upstream open reading frames (uORFs) from those transfected manually to those transfected in sextuplicate using the TransfectionWizard to assess the accuracy and technical variance of the automated protocol. Automated transfections were highly reproducible and consistent with manual transfections, shown in Figure 2, marked by low technical variance and low mean absolute error (MAE) across technical replicates. The synthetic uORFs with IDs 'uORF' and 'wuORF' were taken from [3].

### 2.3 Global distribution and classroom deployment

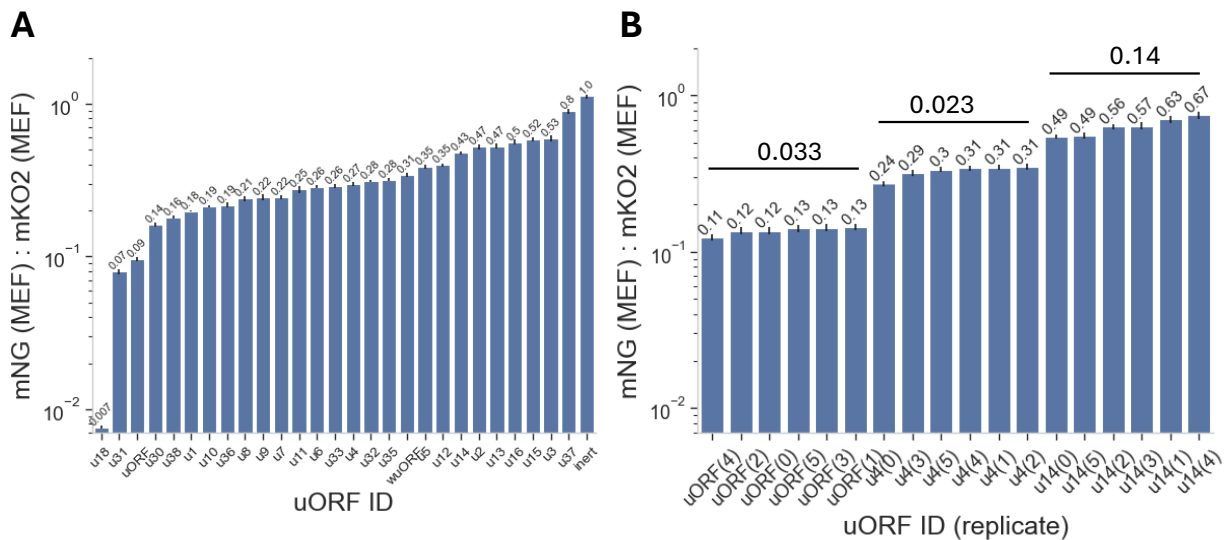
To demonstrate open-source accessibility, the TransfectionWizard was used in the global How to Grow (Almost) Anything (HTGAA) class, where the software was distributed to 14 HTGAA nodes across 4 continents. Across these 14 nodes, 40 students with no prior experience in mammalian cell transfection used the TransfectionWizard to successfully design their own genetic circuits and used the OT-2 for the automated transfection of



**Figure 1.** Schematic overview of the TransfectionWizard automated transfection platform. (A) Genetic circuits are designed in silico. (B) The Opentrons OT-2 liquid handling robot autonomously transfects the circuit from user specifications. (C) Transfected circuits are assayed for function, and a learning module integrates results to iteratively optimize designs.

their genetic circuits within the span of 2 days, demonstrating that the tool can be deployed with minimal onboarding across diverse environments. Post-module interviews with students revealed that students gained confidence in both genetic

circuit design and transfection and rated the tool as intuitive and effective at reducing experimental complexity.



**Figure 2.** Human vs. OT-2 experimental data showing that the OT-2 is highly consistent and reproduces human-transfected results. (A) Quantification of the strength of synthetic upstream open reading frames (uORFs) transfected by hand. Plotted are 27 synthetic uORFs whose strengths were assessed using a dual reporter plasmid where a synthetic uORF was placed upstream of the mNeonGreen ORF and the strength of each uORF, the number above each bar, was quantified by taking the ratio of mNeonGreen and mKO2 in molecules of equivalent fluorescence (MEF) units and are relative to an inert sequence normalized to a value of one. (B) Three synthetic uORFs from panel A were selected to assess the accuracy and technical variance of the TransfectionWizard’s automated transfection protocol. Each synthetic

uORF was transfected in sextuplicate and its strength is shown above each bar. Above the strength score is an error metric, the mean absolute error, for the whole group of sextuplicates relative to human data.

## 2.4 Discussion

The TransfectionWizard addresses a key gap in the mammalian synthetic biology toolkit by unifying genetic circuit design, behavior prediction, and automated transfection in a single open-source platform. Compared to prior automated approaches, the tool reduces setup time and lowers the expertise barrier for new users. Limitations of the current implementation include OT-2 deck setup time (e.g., putting tubes in racks), throughput limitations from OT-2 deck constraints, and use of BSL2 organisms (e.g., human cell lines) that limit which labs from around the globe can run the complete TransfectionWizard workflow from design to FACS analysis. Future work will focus on decreasing OT-2 deck setup time and increasing throughput by using 96 well plates for transfection rather than 24-slot tube racks.

## 3 ACKNOWLEDGEMENTS

We gratefully acknowledge DARPA (036099-00001) and NSF ERFI BEGIN OI (035795-00002) for funding this research.

## 4 REFERENCES

- [1] NAIR, V.S. AND WATSON, K. 2024. A novel, fully automated, and reagent-agnostic transient transfection protocol. *Current Protocols* 4, 2.
- [2] COLIN, B., DEPREZ, B., AND COUTURIER, C. 2019. High-throughput DNA plasmid transfection using acoustic droplet ejection technology. *SLAS Discovery* 24, 4, 492–500.
- [3] FERREIRA, J.P., OVERTON, K.W., AND WANG, C.L. 2013. Tuning gene expression with synthetic upstream open reading frames. *Proceedings of the National Academy of Sciences* 110, 28, 11284–11289.

# Biodesign Metadata Exchange for Use in Biosecurity

Doug Densmore, PhD

Chris Krenz

Guzman Vigliecca

doug@latticeautomation.com

christopher.krenz@latticeautomation.com

guzman@latticeautomation.com

Lattice Automation

Boston, Massachusetts, United States

## 1 INTRODUCTION

The rapid advancement of biodesign tools, particularly those leveraging artificial intelligence (AI), enables researchers to engineer novel biological systems with unprecedented speed and creativity. However, this progress introduces a growing tension with existing biosecurity infrastructure. Current screening frameworks, such as those employed by the International Gene Synthesis Consortium (IGSC), primarily rely on comparing DNA synthesis orders against databases of known hazardous sequences [2]. While essential for identifying orders that closely resemble known pathogens or toxins, this approach becomes increasingly inadequate for biodesign tools generating novel sequences that diverge substantially from anything found in nature [1]. For DNA synthesis and other service providers, assessing the risk of designs that bear little resemblance to known organisms presents a fundamental challenge — one that threatens to slow the design-to-synthesis automation pipeline as providers face growing uncertainty about the orders they receive.

## 2 THE BIODESIGN METADATA EXCHANGE (BMDE)

To address this gap, NTI | bio has partnered with Lattice Automation to design and pilot the Biodesign MetaData Exchange (BMDE), a standard for capturing and transmitting metadata alongside DNA and protein sequences. Rather than relying solely on the final sequence for biosecurity evaluation, the BMDE captures the provenance of a design — the series of operations, sources, and tools used throughout the design process. These details can be used to reconstruct the intermediate states of the design and automatically flag particular operations that may be of concern (based on a customizable rule set). This contextual information can substantially improve the ability of biosecurity decision-makers — such as screeners at DNA synthesis companies — to assess risk, and in turn, reduce friction in the design-to-synthesis pipeline for legitimate research.

The standard defines a structured format for recording this metadata — including design tool, operation type (e.g., insertion, paste, etc.), timestamps, and other details — which accompanies the sequence when submitted to a service.

The core principle is straightforward: a sequence whose design history is transparent and traceable presents a fundamentally different risk profile than an identical sequence of unknown provenance. By making this distinction explicit, the BMDE enables more targeted and efficient screening: orders with complete, low-risk provenance can be processed with greater confidence, while those with gaps or concerning patterns can be prioritized for deeper review.

In Phase 1 of this partnership, NTI and Lattice solicited feedback from a range of industry, academic, and government stakeholders, holding a workshop in May of 2025. These discussions culminated in a white paper that proposed the basic principles of the metadata standard [3]. Now in Phase 2 of this partnership, the team is developing the software libraries (and associated documentation/infrastructure) that will allow design tool and service providers to implement the standard, as well as working with partners such as Twist Bioscience (a DNA synthesis company) and Aclid (a biosecurity screening company) to perform the first end-to-end tests with real tools and providers.

## 3 IMPLEMENTATION

The BMDE pipeline essentially consists of two complementary components:

- Metadata capture at the design tool stage
- Metadata evaluation at the service provider stage

See Figure 1 for a high level diagram outlining these steps.

For metadata capture, we have developed a software library (available in Python, TypeScript, and Java) for capturing design operations, converting them to the BMDE standard format, and exporting the metadata as a JSON file alongside the design. As a first proof of concept, we have integrated this library into a fork of the Open Vector Editor (OVE) [4], an open-source sequence design tool. The integration automatically records operations as the user works — tracking edits, pastes, imports, and other actions — without requiring manual annotation or disrupting the designer's workflow. We plan to extend this integration to a broader range of biodesign tools.

On the evaluation side, we have used our underlying libraries to develop and deploy the Interpreter webapp, a tool for parsing, visualizing, and evaluating BMDE metadata. The Interpreter, which runs entirely in the user's browser (i.e. there is no backend server or database to which the data is transferred), provides several key capabilities:

- (1) It verifies the signature and checksum on the metadata to both authenticate the tool/user that created it (feature still in progress) and ensure it matches the provided design.
- (2) It computes the intermediate versions/diffs of the design based on the operation details in the metadata file. This design history is then rendered as an interactive provenance graph, enabling reviewers to trace the lineage of any portion of the final sequence.
- (3) A configurable rule engine automatically analyzes the design history to flag potential concerns — for example, identifying paste/import operations originating from unknown or uncharacterized sources, or detecting clusters of operations performed in rapid succession that suggest AI-driven/automated design processes.
- (4) Based on this evaluation (and rule configuration), the Interpreter recommends certain flagged intermediate sequences for additional biosecurity screening, enabling a targeted review process.
- (5) Finally, the tool also generates a simple summary of these evaluation results.

Together, these components establish an end-to-end pipeline: metadata is captured transparently during design, transmitted alongside the sequence, and evaluated systematically upon receipt — augmenting, rather than replacing, existing sequence-based screening. Because the rules are customizable, service providers can tune the sensitivity of the evaluation to their needs. The goal is not only increased security but also increased efficiency, with metadata enabling greater automation in the screening process.

#### 4 EXAMPLE CASE

Figure 2 shows a screenshot of a portion of the Interpreter tool. The graph displayed showcases a simple example of sequence obfuscation: a sequence of unknown origin was divided into smaller chunks, one of which was then pasted into a destination sequence submitted for synthesis. Although the final, submitted sequence was not flagged by standard screening, the pasted sequence may have been pulled from a potentially hazardous source. In this case, the customizable rules are set to flag PASTE operations from external sequences (highlighted in the graph in yellow). Other examples might include an IMPORT operation (the origin of which may or may not be known) or an AI-driven operation (where metadata about model/prompt used may be available).

Flagging these operations presents a few opportunities:

- The flagged revision could be ignored. If there are few other indicators of concern or if the customer is known and trusted, no additional action may be needed.
- The intermediate sequence (or a precursor to it) could be screened, potentially revealing malicious origins that were obfuscated in the final sequence.
- If sufficient data is not available, the service provider could request additional details from the customer.

In this way, the metadata provides an additional resource services can use to augment their screening practices.

#### 5 CHALLENGES AND FUTURE DIRECTIONS

A primary challenge for the BMDE is driving adoption across a fragmented ecosystem of biodesign tools and service providers. We believe the best way to drive adoption is to ensure that the design and implementation of the BMDE reflect the needs of its users. Designers need clear incentives — such as expedited screening for orders accompanied by complete and transparent metadata, or other useful features, like the ability to share a replayable design history with a colleague or generate a methods section of a paper — while providers need confidence that the metadata is reliable and actionable. Establishing these feedback loops will be essential to building a self-reinforcing adoption cycle.

We are currently working with partners like Twist Bioscience and Aclid to execute the first end-to-end pilot of the complete BMDE pipeline, from metadata capture through provider-side evaluation. In parallel, we are expanding the Interpreter's standard rule set, implementing identity validation and implementation certification systems, and adding a variety of other features based on partner feedback. Looking ahead, broader tool integration and provider engagement will be critical to establishing the BMDE as a practical layer of the biodesign-to-synthesis infrastructure — one that maintains the pace of innovation while strengthening the biosecurity safeguards needed to support it.

#### REFERENCES

- [1] CARTER, S. R., WHEELER, N., OLONOFF, M., AND YANG, H. Developing guardrails for AI biodesign tools. Tech. rep., Nuclear Threat Initiative, 2024.
- [2] INTERNATIONAL GENE SYNTHESIS CONSORTIUM. Harmonized screening protocol. <https://genesynthesisconsortium.org/harmonized-screening-protocol/>, 2023. Version 2.0.
- [3] NTI | BIO, AND LATTICE AUTOMATION. Biodesign metadata exchange: A standard for capturing and transmitting biodesign metadata for use in biosecurity. <https://www.nti.org/analysis/articles/white-paper-a-proposal-for-biodesign-metadata-exchange-for-use-in-biosecurity/>, 2025.
- [4] TESELAGEN BIOTECHNOLOGY. Open vector editor. <https://github.com/TeselaGen/openVectorEditor>, 2024.

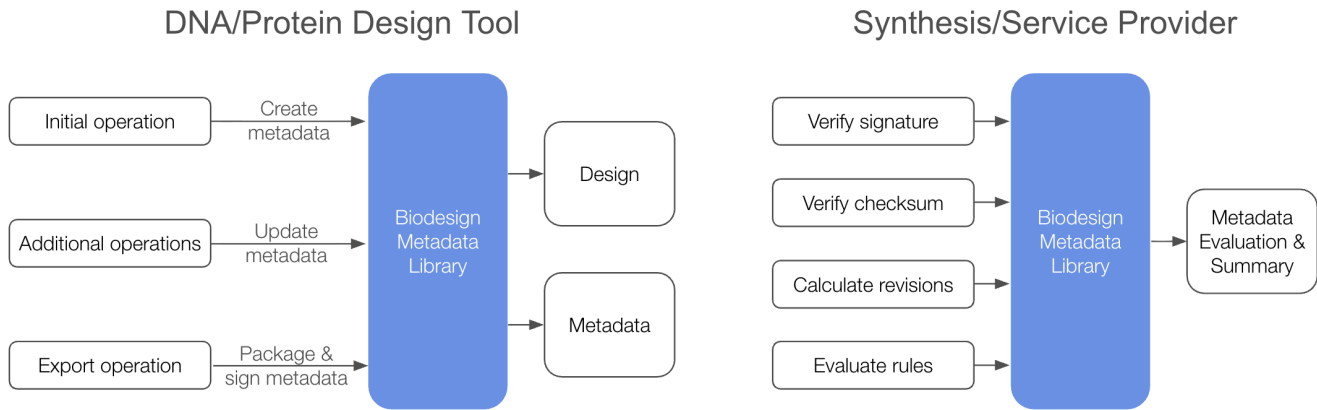


Figure 1: High level diagram of the BMDE flow, from design tool metadata capture to service evaluation.

The screenshot shows the 'Interpreter' webapp interface. At the top, there are two panels: 'Design Overview' and 'Metadata Overview'. The 'Design Overview' panel shows details for 'synpuc19.gb', including LOCUS, DEFINITION, ACCESSION, VERSION, SOURCE, and ORGANISM. The 'Metadata Overview' panel shows details for 'metadata\_synpuc19v.json', including Id, Parent Id, Design Name, Author, and Last Updated. Below these is a 'Version Graph' showing a sequence of revisions for 'sequence-obfuscation' and 'synpuc19v'. The 'Revision Details' panel is open for 'Revision 3 - PASTE', showing the operation details with a JSON object: 

```
{
  "paste_position": 239,
  "pasted_text": [
    "aatgaattgggtaa"
  ],
  "copied_from": "26da574f-3af1-4c85-b76e-be17b353f9cc"
}
```

Figure 2: Screenshot of the Interpreter webapp, showing the design history graph and operation details of an example design.

# Standardized Synthetic Biology Data Curation Effort

Gonzalo Vidal, Hatem Abdelrahman, Saanika Fadia,  
Jackson Fairborn, Daniel Fang, Ryan Greer,  
Conner Groth, Kerem Gurkan, Peter Hindes,  
Ashlyn Huynh, Indiana Kretschmar, Chunxiao Liao,  
William Mo, James Nguyen, Zane Perry,  
Oscar Rodriguez, Kenzo Schwab, Travis Uhrig,  
Carolus Vitalis, Chris Myers  
University of Colorado Boulder  
Boulder, United States of America  
chris.myers@colorado.edu

Eric Young  
Worcester University  
Worcester, United States of America

Georgie Hau Sørensen  
University of Bristol  
Bristol, United Kingdom

Prashant Vaidyanathan  
Oxford Biomedica  
Oxford, United Kingdom

## 1 INTRODUCTION

Synthetic biology projects generate diverse, interdependent artifacts across a project life cycle (genetic designs, sequences, models, protocols, images, and time-series measurements). In practice, these assets are frequently scattered across spreadsheets, tool-specific exports, and custom code, making it difficult to preserve provenance and to keep data, metadata, and analysis code coherently linked over time—thereby limiting reuse, automation, compliance, and scale-up. The *Synthetic Biology Open Language* (SBOL) [7] directly targets this problem by providing a community-developed, ontology-backed standard that captures biological designs in a machine accessible representation. The FAIR Guiding Principles (Findable, Accessible, Interoperable, Reusable) [12] further clarify that the goal is not just “open data,” but machine accessible research objects across data, software, and workflows. The need of FAIR data is even more evident in the advent of data-driven *machine learning* (ML) workflows. This work reports concrete data curation advances from the SBOL community and frames it as a community effort grounded in a published standardized data-driven workflow framework [10].

## 2 RESULTS

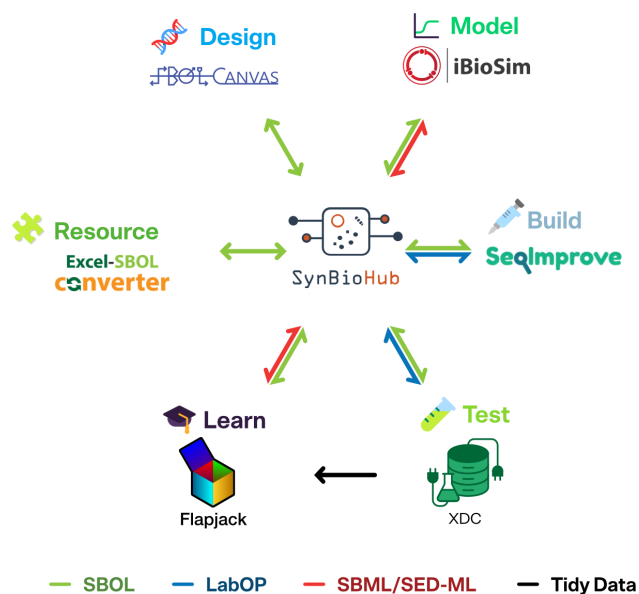
We applied the standardized data management workflow shown in Figure 1 to curate publication-derived synthetic biology data into linked, machine accessible artifacts spanning resource, design, model, build, test, and learn stages. Our initial community effort focuses on three representative studies [1, 5, 6]. Together, given the significant supplemental materials provided, these publications provide rich use cases for reconstructing biological parts, circuit structure, sequence-level build artifacts, and associated experimental data within an SBOL-centered workflow. The curation effort is organized through public GitHub repositories that

preserve version history and stage-specific artifacts, while the broader framework maps these artifacts to interoperable standards and repositories across the synthetic biology workflow.

For the Cello case study [5], this effort began from the *User Constraint Files* (UCFs) and generated an initial collection of genetic parts and interactions. We then used SBOLCanvas [9] to reconstruct the reported circuit designs by selecting components from that curated collection and exporting the resulting constructs as SBOL. These SBOL designs were subsequently used in iBioSim [11] to generate executable models and simulations, enabling direct comparison between curated designs and expected circuit behavior. In parallel, plasmid sequences associated with the implementation were curated with SeqImprove [3] and annotated against the same part library, thereby linking design intent to physical build artifacts. For the test stage, experimental measurements and metadata were captured using the *eXperimental Data Connector* (XDC) [8], which supports standardized linkage between experimental data and repository metadata. The 0xF6 workflow example illustrates this end-to-end pattern clearly, with linked directories for designs, models, plasmids, and resources.

For the scalable dynamic characterization of synthetic gene circuits case study [1], this effort used Excel-to-SBOL [2] to capture publication resources, including genetic parts, chemicals, and chassis information, and then used SBOLCanvas [9] to reconstruct the corresponding designs. The curated package already separates design and resource artifacts, making it a useful scaffold for progressive community completion and refinement.

For 2-bit MD5 hash function case study [6], this effort similarly captured resources using Excel-to-SBOL [2], reconstructed designs in SBOLCanvas [9], and generated model



**Figure 1: Software ecosystem used for the data curation workflow.** Resources are stored in SynBioHub using the Excel-SBOL Converter. The selected resources can then be fetched from SynBioHub using a design tool, such as SBOLCanvas. Once the design is complete, it can be modeled using analysis tools, such as iBioSim, that support SBOL, SBML, and SED-ML. Once the modeling step indicates that the design is correct, the built plasmid can be annotated using SeqImprove. Built plasmids and strains can be tested in the laboratory, the metadata and protocols can be stored in SynBioHub, and the measurement data can be stored in Flapjack using a test tool, such as XDC. Finally, the test data can be analyzed using a learn tool, such as Flapjack.

artifacts compatible with downstream simulation workflows. Build-stage curation relies on SeqImprove [3] to annotate the plasmid sequences provided by the publication. Examples from this case study are shown in Figure 2. This case study also exposed an important representational challenge for the community: how best to encode plasmids that are later integrated into genomic loci while preserving provenance between plasmid intermediates and chromosomal implementations.

## Discussion

Taken together, these results show that SBOL-based community curation can transform publication supplements into connected digital research objects that are easier to inspect, reuse, simulate, and extend. The workflow also reinforces the role of repositories such as SynBioHub [4] as the dissemination layer for standardized biological design information and associated metadata. We view these curated examples not

as finished endpoints, but as seed contributions for a growing community effort. We therefore make an open call for others to join this work by contributing new papers, reviewing curated artifacts, and helping define best practices for unresolved cases such as incomplete metadata, ambiguous construct descriptions, and genome-integration workflows.

## Data availability

<https://github.com/SynBioDex/SBOL-Data-Curation>

<https://github.com/MyersResearchGroup/SynBioWorkflowExamples>

## REFERENCES

- [1] DALCHAU, N., GRANT, P. K., VAIDYANATHAN, P., SPACCASASSI, C., GRAVILL, C., AND PHILLIPS, A. Scalable dynamic characterization of synthetic gene circuits. *bioRxiv* (2019), 635672.
- [2] MANTE, J., ABAM, J., SAMINENI, S. P., POTZSCH, I. M., BEAL, J., AND MYERS, C. J. Excel-sbol converter: Creating sbol from excel templates and vice versa. *ACS synthetic biology* 12, 1 (2023), 340–346.
- [3] MANTE, J., SENTS, Z., BRITT, D., MO, W., LIAO, C., GREER, R., AND MYERS, C. J. Seqimprove: Machine-learning-assisted curation of genetic circuit sequence information. *ACS Synthetic Biology* 13, 9 (2024), 3051–3055.
- [4] McLAUGHLIN, J. A., MYERS, C. J., ZUNDEL, Z., MISIRLI, G., ZHANG, M., OFITERU, I. D., GONI-MORENO, A., AND WIPAT, A. Synbiohub: a standards-enabled design repository for synthetic biology. *ACS synthetic biology* 7, 2 (2018), 682–688.
- [5] NIELSEN, A. A., DER, B. S., SHIN, J., VAIDYANATHAN, P., PARALANOV, V., STRYCHALSKI, E. A., ROSS, D., DENSMORE, D., AND VOIGT, C. A. Genetic circuit design automation. *Science* 352, 6281 (2016), aac7341.
- [6] PADMAKUMAR, J. P., SUN, J. J., CHO, W., ZHOU, Y., KRENZ, C., HAN, W. Z., DENSMORE, D., SONTAG, E. D., AND VOIGT, C. A. Partitioning of a 2-bit hash function across 66 communicating cells. *Nature Chemical Biology* 21, 2 (2025), 268–279.
- [7] ROEHNER, N., BEAL, J., CLANCY, K., BARTLEY, B., MISIRLI, G., GRÜNBERG, R., OBERORTNER, E., POCOCK, M., BISSELL, M., MADSEN, C., NGUYEN, T., ZHANG, M., ZHANG, Z., ZUNDEL, Z., DENSMORE, D., GENNARI, J. H., WIPAT, A., SAURO, H. M., AND MYERS, C. J. Sharing structure and function in biological design with SBOL 2.0. *ACS Synthetic Biology* 5, 6 (June 2016), 498–506. doi: 10.1021/acssynbio.5b00215.
- [8] SAMINENI, S. P., VIDAL, G., VITALIS, C., FELIU, G. Y., RUDGE, T. J., MYERS, C. J., AND MANTE, J. Experimental data connector (xdc): integrating the capture of experimental data and metadata using standard formats and digital repositories. *ACS Synthetic Biology* 12, 4 (2023), 1364–1370.
- [9] TERRY, L., EARL, J., THAYER, S., BRIDGE, S., AND MYERS, C. J. Sbolcanvas: a visual editor for genetic designs. *ACS Synthetic Biology* 10, 7 (2021), 1792–1796.
- [10] VITALIS, C., VIDAL, G., SAMINENI, S. P., FONTANARROSA, P., AND MYERS, C. J. A framework for a standard-enabled fair data management workflow for synthetic biology. *ACS Synthetic Biology* (2026).
- [11] WATANABE, L., NGUYEN, T., ZHANG, M., ZUNDEL, Z., ZHANG, Z., MADSEN, C., ROEHNER, N., AND MYERS, C. ibiosim 3: a tool for model-based genetic circuit design. *ACS synthetic biology* 8, 7 (2018), 1560–1563.
- [12] WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J.-W., DA SILVA SANTOS, L. B., BOURNE, P. E., ET AL. The fair guiding principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.



# Adding SBML Modeling Support to SBOLCanvas

Travis Uhrig

University of Colorado Boulder  
Boulder, Colorado  
travis.uhrig@colorado.edu

Chris J. Myers

University of Colorado Boulder  
Boulder, Colorado  
chris.myers@colorado.edu

## 1 INTRODUCTION

In genetic circuit design, building a structural representation of a circuit and then modeling and simulating its behavior are essential steps in the design process (see Fig. 1A). iBioSim [13] supports structural design, modeling, and simulation using the *Systems Biology Markup Language* (SBML) [4], but it is a standalone Java application requiring local installation. SBOLCanvas [12] provides a web-based editor for structural design using the *Synthetic Biology Open Language* (SBOL) [8] and SBOL Visual [1], but does not support modeling or simulation. SynBioSuite [10] was developed as a web-based platform that integrates SBOLCanvas for structural design with iBioSim for simulation, converting the SBOL designs to SBML models using iBioSim's converter [9]. However, this conversion lacks modeling options and instead uses predefined defaults.

We address this by adding modeling support to SBOLCanvas. Users can now configure modeling options, such as kinetic constants, species initial amounts, and discrete events to change species amounts, alongside their structural design (see Fig. 1B). With SBOLCanvas now handling both structural design and modeling, iBioSim is only needed as a simulation backend, enabling a complete web-based workflow from design to simulation within SynBioSuite.

## 2 SYSTEM DESIGN

### Visual Interface

The SBOLCanvas visual interface was expanded to support modeling. Existing molecular species and interaction glyphs now include user inputs for model elements such as initial amounts and kinetic constants. A new Event glyph was added to the palette so users can place and configure time-based discrete simulation events on the design canvas (see Fig. 2, top).

### Export Architecture

The exporter takes SBOLCanvas's internal MXGraph representation, which now includes the user-configured model elements, and converts it into SBML using the Java SBML library (JSBML) [7]. The export follows a *Transcriptional Unit* (TU) based approach, divided into four phases:

**Phase 1 – Species Creation.** The SBML exporter scans the visual model to find all molecular species (proteins, small

molecules, complexes) and maps them to SBML Species. Additionally, iBioSim "Promoter" species are created to represent TUs. Since SBOLCanvas represents TUs visually as backbones with glyphs, each backbone is mapped to a separate promoter species. Currently, each TU must be on its own backbone; detecting multiple TUs within a single backbone is not yet supported.

**Phase 2 – Reaction Creation.** Reactions are generated based on the interaction edges in the design. *Genetic Production:* all production edges from a single backbone are collected into one TU-based reaction. The kinetic law varies by regulation type (constitutive, repressed, activated, or mixed) using Hill-function rate equations. *Complex Formation:* association nodes bind multiple reactants into a single complex product using a reversible mass-action kinetic law. *Degradation:* degradation arrows are converted into irreversible mass-action reactions.

**Phase 3 – Visual Layout.** The SBML Layout Extension is used to store the positions of all species and reaction glyphs. When the exported SBML is opened in tools that support this extension, the visual layout from the original design is preserved.

**Phase 4 – Events.** Time-based discrete events, such as adding or removing small molecules at a specified time, are read from the new Event glyphs and exported as SBML Events.

The mapping from SBOLCanvas visual elements to SBML concepts is shown in Table 1.

Table 1: SBOLCanvas to SBML Mapping

SBOLCanvas Element	SBML Concept
Backbone	Promoter Species (SBO:0000590)
Molecular Species Glyph	Species (Protein, Complex, etc.)
Production Arrow	Product in TU Reaction
Inhibition Arrow	Modifier (Repressor)
Stimulation Arrow	Modifier (Activator)
Association Node	Complex Formation Reaction
Degradation Arrow	Degradation Reaction

### Kinetic Laws

The exporter computes kinetic law equations from user-configured parameters and interactions. For constitutive and

repressed production, the kinetic law is as follows:

$$\frac{P \cdot k_o \cdot K_o \cdot n_r}{1 + K_o \cdot n_r + \sum_i (K_{r,i} \cdot R_i)^{n_{c,i}}} \quad (1)$$

where  $K_o$ ,  $K_{r,i}$  are binding equilibrium constants and  $n_{c,i}$  the Hill coefficient. With no repressors, this reduces to constitutive expression. For activated and mixed-regulation production, the kinetic law is as follows:

$$\frac{P \cdot (k_b \cdot K_o \cdot n_r + \sum_j k_a \cdot K_{a,o} \cdot n_r \cdot (K_{a,j} \cdot A_j)^{n_{c,j}})}{1 + K_o \cdot n_r + \sum_j K_{a,o} \cdot n_r \cdot (K_{a,j} \cdot A_j)^{n_{c,j}} + \sum_i (K_{r,i} \cdot R_i)^{n_{c,i}}} \quad (2)$$

where repressor terms in the denominator handle mixed regulation. Degradation and complex formation use mass-action kinetics. All formulas were validated against iBioSim's reaction compiler [2, 5].

### Data Persistence

With this data available, SBOLCanvas exports complete SBML models. The model elements are stored as SBOL annotations on their corresponding ComponentDefinitions, Interactions, and GenericTopLevels, so designs with modeling data can be saved, shared, and reopened in SBOLCanvas without losing any of their configuration.

### SynBioSuite Integration

SBOLCanvas is embedded inside SynBioSuite as an iFrame. When the user edits a design, SBOLCanvas automatically exports the SBML model and sends it to SynBioSuite via a postMessage. SynBioSuite saves a companion SBML file alongside the SBOL design file. The user can then switch to SynBioSuite's Model panel to find the exported model ready for simulation, with iBioSim running the analysis as a headless backend (see Fig. 2).

## 3 RESULTS

We validated the exporter using the Genetic Toggle Switch [3], a well-known genetic circuit with two mutually repressing TUs. The toggle switch was designed in SBOLCanvas within SynBioSuite, with TetR repressing pTet and LacI repressing pLac, a reporter (GFP), and small molecule triggers (IPTG, aTc) added via the events system.

Figure 2 shows this workflow in practice. The circuit is designed and model elements are configured in the embedded SBOLCanvas editor (top). Switching to SynBioSuite's Model panel (bottom), the exported model is immediately available for simulation. The results show the expected toggle switch behavior, confirming that the correct reaction equations, parameters, and events were exported.

## 4 FUTURE WORK

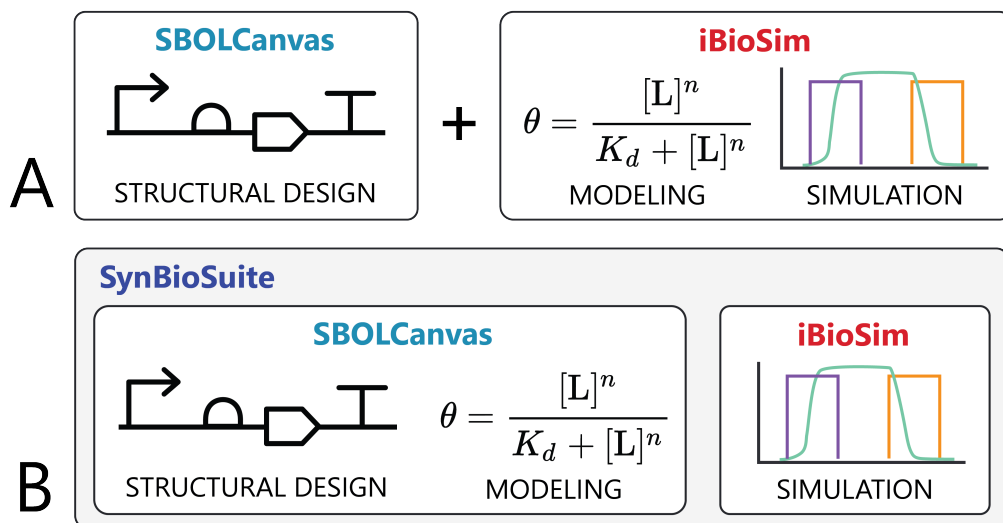
While the core export functionality is complete and integrated into SynBioSuite, several areas remain to reach full parity with iBioSim's modeling capabilities:

- **Rules and Constraints:** Assignment rules, rate rules, algebraic rules, and simulation constraints.
- **Functions:** Named functions usable in kinetic laws, rules, and events.
- **Custom Kinetic Laws:** User editing of the auto-generated formulas.
- **Expanded Events:** Conditional triggers, multiple assignments per event, delay, and priority.

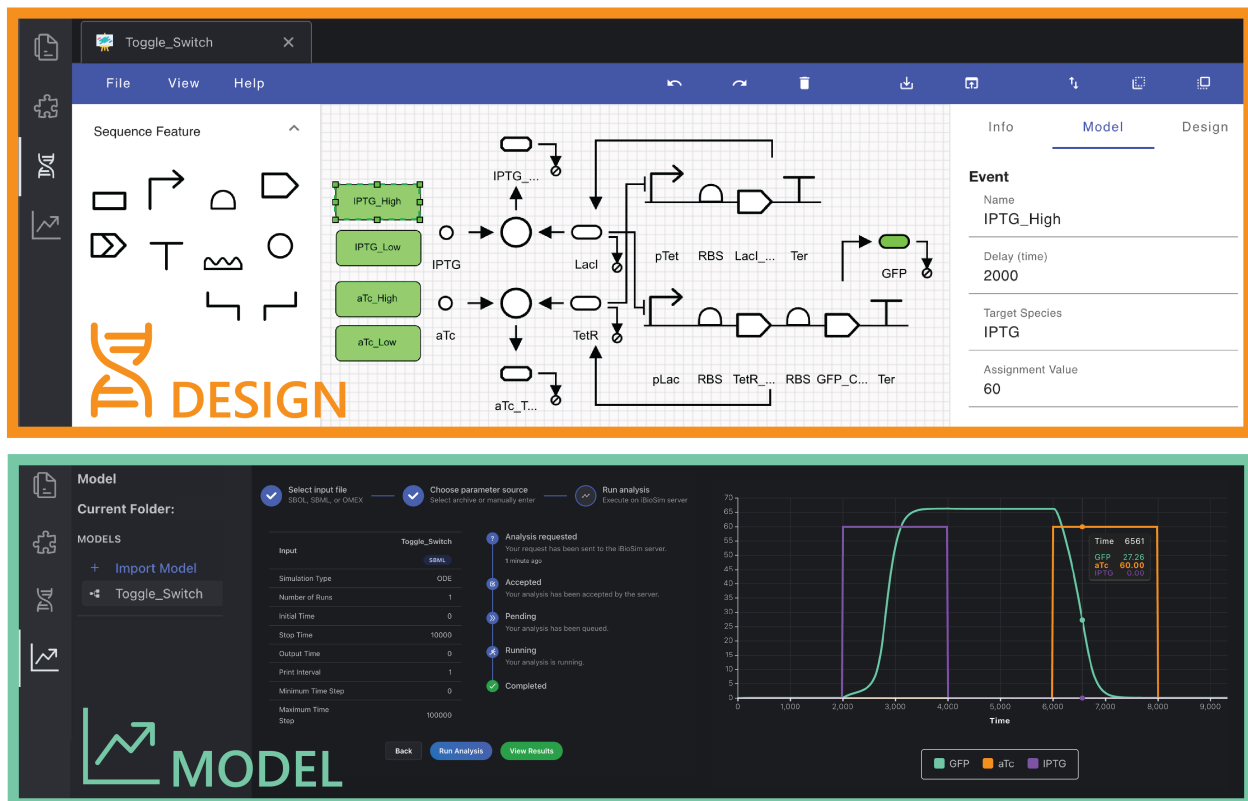
All source code is available in the SBOLCanvas [11] and SynBioSuite [6] repositories.

## REFERENCES

- [1] BEAL, J., NGUYEN, T., GOROCHOWSKI, T. E., GOÑI-MORENO, A., SCOTT-BROWN, J., MCLAUGHLIN, J. A., MADSEN, C., ALERITSCH, B., BARTLEY, B., BHAKTA, S., BISSELL, M., CASTILLO HAIR, S., CLANCY, K., LUNA, A., LE NOVÈRE, N., PALCHICK, Z., POCOCK, M., SAURO, H., SEXTON, J. T., TABOR, J. J., VOIGT, C. A., ZUNDEL, Z., MYERS, C., AND WIPAT, A. Communicating structure and function in synthetic biology diagrams. *ACS Synthetic Biology* 8, 8 (2019), 1818–1825. PMID: 31348656.
- [2] FONTANAROSA, P., DOOSTHOSSEINI, H., BORUJENI, A. E., DORFAN, Y., VOIGT, C. A., AND MYERS, C. Genetic circuit dynamics: Hazard and glitch analysis. *ACS Synthetic Biology* (2020).
- [3] GARDNER, T. S., CANTOR, C. R., AND COLLINS, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403, 6767 (2000), 339–342.
- [4] KEATING, S. M., ET AL. SBML level 3: an extensible format for the exchange and reuse of biological models. *Molecular Systems Biology* 16, 8 (2020), e9110.
- [5] MISIRLI, G., NGUYEN, T., MCLAUGHLIN, J. A., VAIDYANATHAN, P., JONES, T. S., DENSMORE, D., MYERS, C., AND WIPAT, A. A computational workflow for the automated generation of models of genetic designs. *ACS Synthetic Biology* 8, 7 (2019), 1548–1559.
- [6] MYERSRESEARCHGROUP. SynBioSuite source code. GitHub Repository. Accessed: 2026.
- [7] RODRIGUEZ, N., ET AL. JSBML 1.0: providing a smorgasbord of options to encode systems biology models. *Bioinformatics* 31, 20 (2015), 3383–3386.
- [8] ROEHNER, N., BEAL, J., CLANCY, K., BARTLEY, B., MISIRLI, G., GRÜNBERG, R., OBERORTNER, E., POCOCK, M., BISSELL, M., MADSEN, C., NGUYEN, T., ZHANG, M., ZHANG, Z., ZUNDEL, Z., DENSMORE, D., GENNARI, J. H., WIPAT, A., SAURO, H. M., AND MYERS, C. J. Sharing structure and function in biological design with SBOL 2.0. *ACS Synthetic Biology* 5, 6 (June 2016), 498–506. doi: 10.1021/acssynbio.5b00215.
- [9] ROEHNER, N., ZHANG, Z., NGUYEN, T., AND MYERS, C. J. Generating systems biology markup language models from the synthetic biology open language. *ACS Synthetic Biology* 4, 8 (2015), 873–879.
- [10] SENTS, Z., STOUGHTON, T. E., BUECHERL, L., THOMAS, P. J., FONTANAROSA, P., AND MYERS, C. J. SynBioSuite: A tool for improving the workflow for genetic design and modeling. *ACS Synthetic Biology* 12, 3 (2023), 892–897.
- [11] SYNBIODEX. SBOLCanvas source code. GitHub Repository. Accessed: 2026.
- [12] TERRY, L., EARL, J., THAYER, S., BRIDGE, S., AND MYERS, C. J. SBOLCanvas: A visual editor for genetic designs. *ACS Synthetic Biology* (2021).
- [13] WATANABE, L., NGUYEN, T., ZHANG, M., ZUNDEL, Z., ZHANG, Z., MADSEN, C., ROEHNER, N., AND MYERS, C. iBioSim 3: A tool for model-based genetic circuit design. *ACS Synthetic Biology* 8, 7 (2019), 1560–1563.



**Figure 1: Workflow comparison. (A) Previously, SBOL files from SBOLCanvas lacked model elements, so users had to recreate their design in iBioSim to build a complete SBML model. (B) SBOLCanvas now accepts model elements and directly exports SBML, so iBioSim is only needed as a simulation backend, enabling the full workflow within SynBioSuite.**



**Figure 2: The integrated design-to-simulation workflow in SynBioSuite. Top (Design): a genetic toggle switch designed in the embedded SBOLCanvas editor, with model elements such as kinetic constants, initial amounts, and discrete events configured in the Model tab. Bottom (Model): the exported model is immediately available for simulation with configurable simulation type, time, and run settings. The results show the expected toggle switch behavior.**

# From Design to Build: Sequence Domestication, Robotic Assembly, and Integrated SBOL Data Capture

**Jeanet Mante**

jvm836@utexas.edu

Oak Ridge Associated Universities

Knoxville, Tennessee

**Randi Pullen**

**Randy Hughes**

Army Research Laboratory

Adelphi, Maryland

## 1 INTRODUCTION

Standards such as the Synthetic Biology Open Language (SBOL) [2] and repositories such as SynBioHub [7] provide a foundation for machine-readable biological design, including API access, extensibility, and deployment in secure, local environments with federated querying across distributed instances. These systems enable structured storage and exchange of genetic designs and support the development of downstream computational tooling. Building on this foundation, the Synthetic Biology Knowledge System (SBKS) established the feasibility of populating repositories through automated extraction from the literature, linking sequences to publication context at scale [5]. However, this work also demonstrated that data quality depends strongly on input format, with unstructured sources such as PDFs introducing fragmentation and limiting annotation fidelity. These findings motivated moving from retrospective reconstruction toward capturing structured data at the point of design.

Subsequent work introduced SeqImprove, an author-in-the-loop curation system that combines machine learning-assisted annotation with ontology-grounded metadata and direct integration with SynBioHub [6]. By enabling standardized, machine-readable design records at submission time, SeqImprove improves interoperability while reducing the burden of manual curation. However, despite these advances, deposition remains decoupled from experimental execution, requiring researchers to engage in curation as a separate and often secondary task.

In parallel, experimental practice has increasingly shifted toward automated execution via liquid-handling robots. Platforms such as Formulatrix Mantis, Tecan Fluent, and Beckman Coulter Echo operate on machine-readable worklists that encode liquid transfers and plate layouts, enabling high-throughput combinatorial design and systematic exploration of experimental conditions.

However, the pipeline connecting standardized design representations to automated experimental execution remains incomplete. This disconnect is particularly evident in sequence domestication, a routine but labor-intensive step in the design-build-test-learn (DBTL) cycle. In practice, this process is performed through a patchwork of specialized

tools and manual intervention. Commercial platforms provide narrow functionality, such as primer design or codon optimization, but lack flexibility for custom assembly standards or synthesis constraints [1, 3, 9]. On the other hand, general-purpose software requires manual identification of restriction sites, introduction of mutations, and formatting of sequences, a process that is slow, error-prone, and difficult to scale across large design spaces. As a result, these design steps remain disconnected from both standardized representations and automated execution systems.

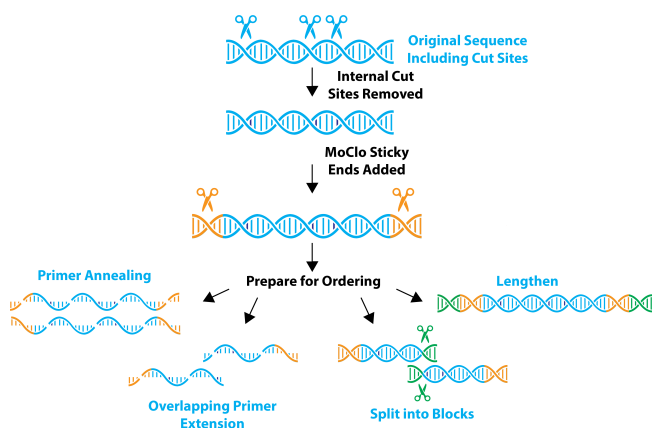
To address this, we developed a sequence domestication tool that automates the full workflow, including restriction site removal, addition of modular cloning (MoClo) ends, synthesis ordering sheets, and robotic assembly instructions. This workflow is directly integrated with the automated generation of structured data records, capturing sequence design in machine-readable SBOL. By embedding data capture and curation within the design process itself, this approach shifts curation from a separate obligation to a byproduct of productive work, reducing burden while improving data quality and interoperability across the DBTL cycle.

## 2 RESULTS

The sequence domestication and assembly planning pipeline (Figure 1) was developed to automate the full design-to-execution workflow. The system accepts structured input tables defining sequences with associated metadata, allowable codon substitutions, cloning templates with flanking regions, empirical overhang-ligation fidelity data, and source plate files and composition tables specifying reagents and desired assemblies. Source plate files may be provided after synthesis by commercial vendors, reflecting the physical layout, concentrations, and volumes of synthesized DNA upon delivery.

For each sequence, the pipeline performs initial cleaning and validation, removing whitespace, enforcing uppercase format, and rejecting sequences containing invalid bases or lengths below six nucleotides. Then it scans for internal restriction enzyme recognition sites from a user-defined list. When sites are detected, the program applies codon-based synonymous substitutions prioritized to minimize sequence

disruption and avoid introducing new restriction sites, or direct base complementation for non-coding sequences.



**Figure 1: The sequence domestication tool works by removing internal cut sites, adding MoClo sticky ends for a toolkit or standard specified by the user, and then splitting sequences based on synthesis-aware constraints.**

Following domestication, standardized flanking sequences are incorporated by substituting placeholders into user-defined cloning templates, supporting MoClo [10], CIDAR [4], or custom assembly syntaxes. The system then determines an appropriate synthesis strategy based on sequence length and user-defined thresholds, including oligo synthesis, sequence extension, or fragmentation into synthesis blocks. Candidate split points are selected to satisfy size constraints, preserve unique 4-bp overlaps, and avoid reuse of overhangs, with NEB ligation fidelity data used to reject low-specificity overlaps [8]. Recursive splitting and extension PCR primer design are applied as needed to meet all constraints.

The domesticated sequences are then used in assembly planning, where composition tables are expanded into combinatorial reaction sets and validated against source plate inventories. For restriction-ligation workflows, *in silico* digestion is performed on all parts and vectors, and resulting fragments are evaluated across all permutations to confirm overhang compatibility and correct assembly order. Each reaction is annotated with assembly validity, part arrangement, and predicted sequence output.

Next, liquid-handling instructions are generated for acoustic dispensing, with reagent volumes computed from molarity constraints and simulated dispensing from source plates. The system outputs Echo-compatible transfer files, updated plate maps, and reaction logs summarizing composition and predicted sequences. SBOL records are generated at both the sequence and assembly levels, preserving structured design information across all stages of the pipeline.

### 3 DISCUSSION

The pipeline produces synthesis-ready sequence designs, SBOL design records, and liquid-handling instruction files for automated execution. By integrating sequence domestication, synthesis planning, assembly validation, and robotic workflows, the system bridges the gap between design and execution. This approach reduces manual effort, improves data standardization, and provides a scalable framework for automated, high-throughput synthetic biology.

### REFERENCES

- [1] CHIN, J. X., CHUNG, B. K.-S., AND LEE, D.-Y. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* 30, 15 (Aug. 2014), 2210–2212.
- [2] GALDZICKI, M., CLANCY, K. P., OBERORTNER, E., POCOCK, M., QUINN, J. Y., RODRIGUEZ, C. A., ROEHNER, N., WILSON, M. L., ADAM, L., ANDERSON, J. C., BARTLEY, B. A., BEAL, J., CHANDRAN, D., CHEN, J., DENSMORE, D., ENDY, D., GRÜNBERG, R., HALLINAN, J., HILLSON, N. J., JOHNSON, J. D., KUCHINSKY, A., LUX, M., MISIRLI, G., PECCOUD, J., PLAGAR, H. A., SIRIN, E., STAN, G.-B., VILLALOBOS, A., WIPAT, A., GENNARI, J. H., MYERS, C. J., AND SAURO, H. M. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology* 32, 6 (June 2014), 545–550.
- [3] GAO, W., RZEWSKI, A., SUN, H., ROBBINS, P. D., AND GAMBOTTO, A. UpGene: Application of a Web-Based DNA Codon Optimization Algorithm. *Biotechnology Progress* 20, 2 (2004), 443–448. [\\_eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1021/bp0300467](https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1021/bp0300467).
- [4] IVERSON, S. V., HADDOCK, T. L., BEAL, J., AND DENSMORE, D. M. CIDAR MoClo: Improved MoClo Assembly Standard and New E. coli Part Library Enable Rapid Combinatorial Design for Synthetic and Traditional Biology. *ACS synthetic biology* 5, 1 (Jan. 2016), 99–103.
- [5] MANTE, J., HAO, Y., JETT, J., JOSHI, U., KEATING, K., LU, X., NAKUM, G., RODRIGUEZ, N. E., TANG, J., TERRY, L., WU, X., YU, E., DOWNIE, J. S., MCINNES, B. T., NGUYEN, M. H., SEPULVADO, B., YOUNG, E. M., AND MYERS, C. J. Synthetic Biology Knowledge System. *ACS Synthetic Biology* 10, 9 (Sept. 2021), 2276–2285. Publisher: American Chemical Society.
- [6] MANTE, J., SENTS, Z., BRITT, D., MO, W., LIAO, C., GREER, R., AND MYERS, C. J. SeqImprove: Machine-Learning-Assisted Curation of Genetic Circuit Sequence Information. *ACS Synthetic Biology* 13, 9 (Sept. 2024), 3051–3055. Publisher: American Chemical Society.
- [7] MCLAUGHLIN, J. A., MYERS, C. J., ZUNDEL, Z., MISIRLI, G., ZHANG, M., OFITERU, I. D., GOÑI-MORENO, A., AND WIPAT, A. SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synthetic Biology* 7, 2 (Feb. 2018), 682–688. Publisher: American Chemical Society.
- [8] PRYOR, J. M., POTAPOV, V., KUCERA, R. B., BILOTTI, K., CANTOR, E. J., AND LOHMAN, G. J. S. Enabling one-pot Golden Gate assemblies of unprecedented complexity using data-optimized assembly design. *PLOS ONE* 15, 9 (Sept. 2020), e0238592. Publisher: Public Library of Science.
- [9] VINCZE, T., POSEAI, J., AND ROBERTS, R. J. NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Research* 31, 13 (July 2003), 3688–3691.
- [10] WEBER, E., ENGLER, C., GRUETZNER, R., WERNER, S., AND MARILLONNET, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* 6, 2 (Feb. 2011), e16765.