Bayesian optimization of microbiomes using a tailored machine learning model

Jaron Thompson[1], Victor Zavala[1], Ophelia Venturelli[1,2,3]

[1]Department of Chemical and Biological Engineering

[2]Department of Biochemistry

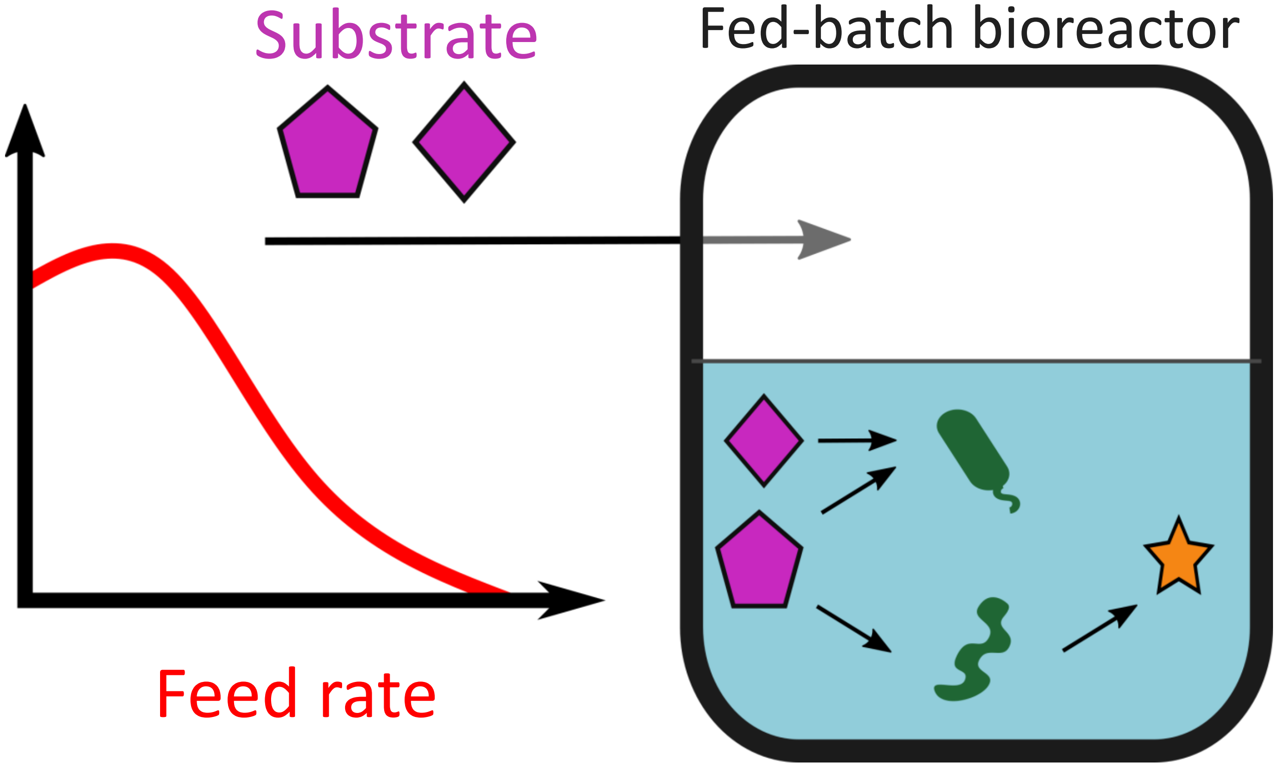
[3]Department of Bacteriology

University of Wisconsin-Madison

2023 BIRS Workshop: Emerging mathematical challenges in synthetic biological network design

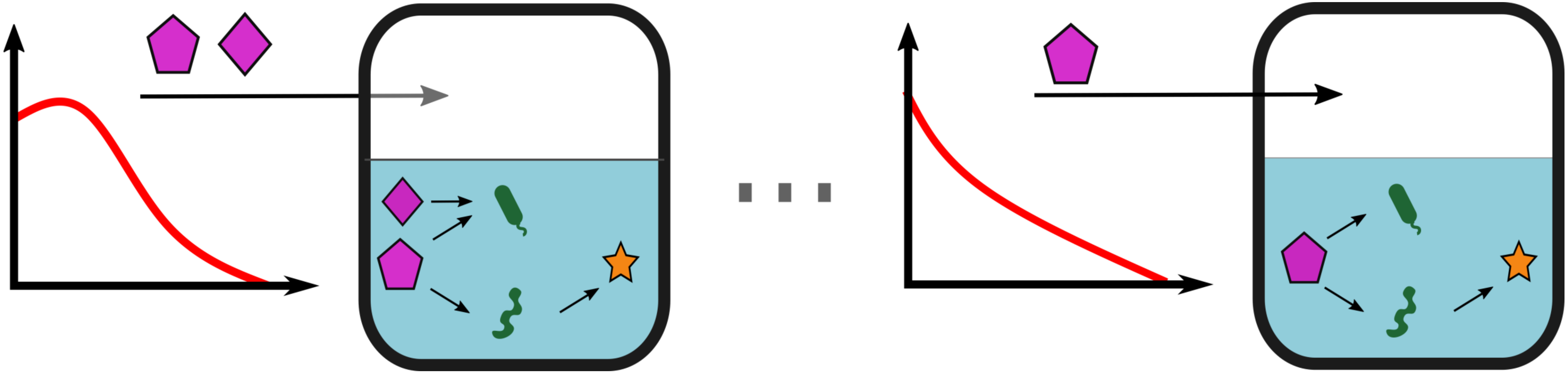
Bioprocessing applications of mixed microbial community bioreactors

- Waste valorization
- Chemical production
- Bioplastics production



Maximizing product requires optimizing feed rate and substrate composition

Experimental design case study:



Reactor 1

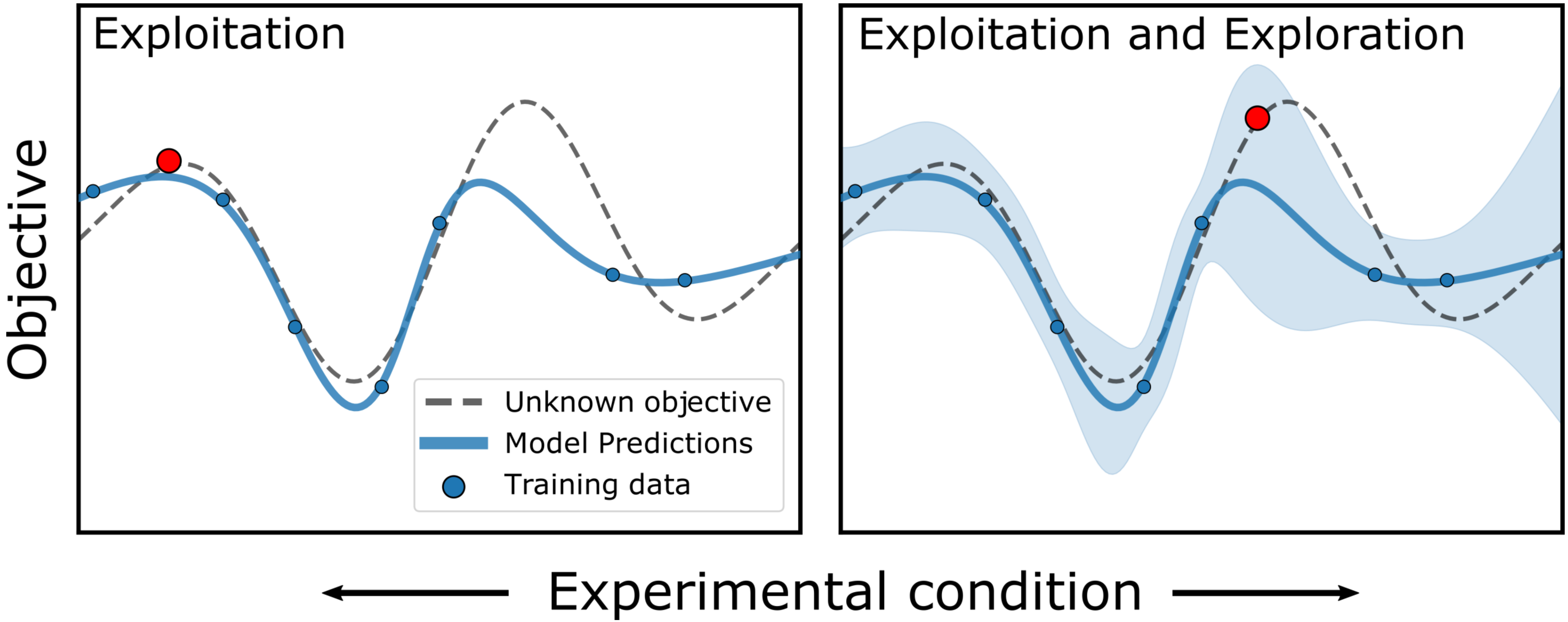
Reactor 5

We can run 5 reactors in parallel to improve throughput, each with a choice of

- Selection of 7 possible substrates
- Selection among 20 possible time-dependent feed rates

} 2,540 configurations

Bayesian optimization: model guided design



Acquisition function captures both the expected objective and prediction uncertainty

Bayesian optimization using a Gaussian process

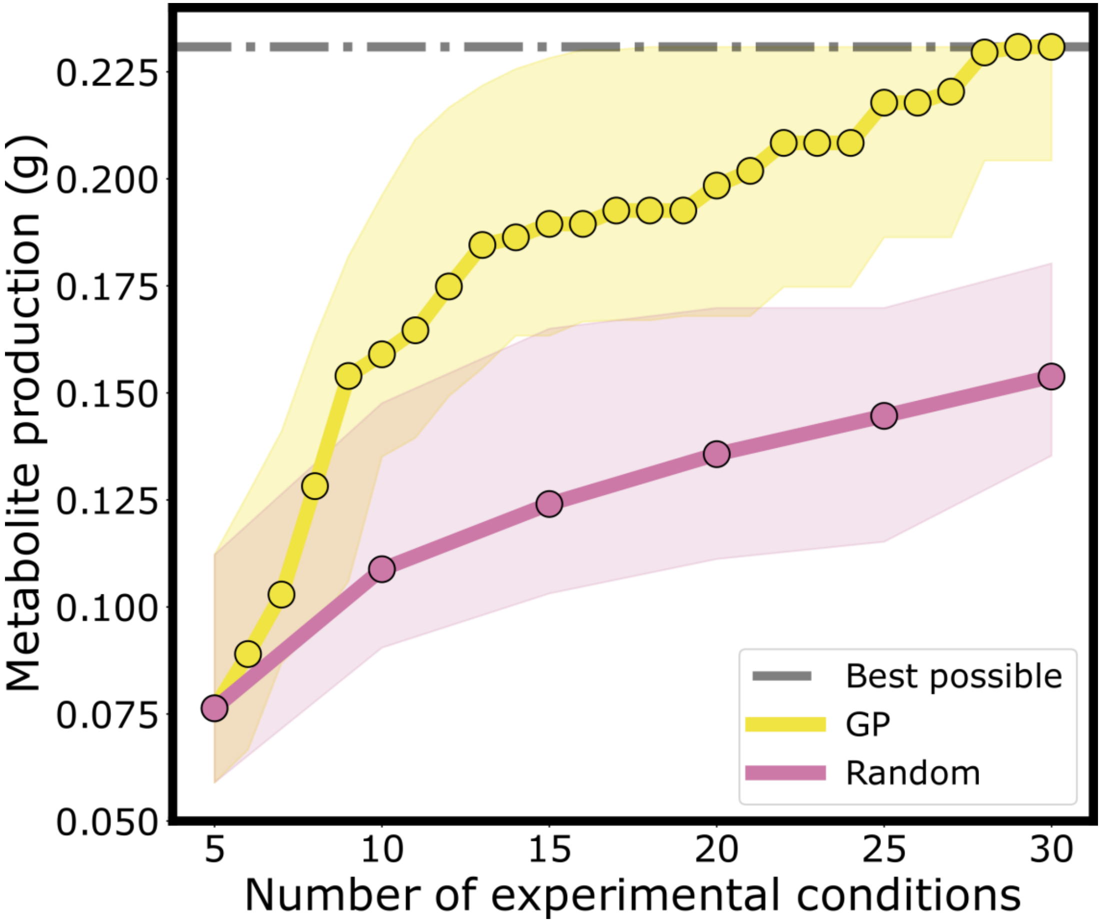
$$p(y|q_i) = \mathcal{N}(\mu_{GP}(q_i),\ \sigma^2_{GP}(q_i))$$

$y :=$ Production

$q_i :=$ Choice of substrates and feed rate

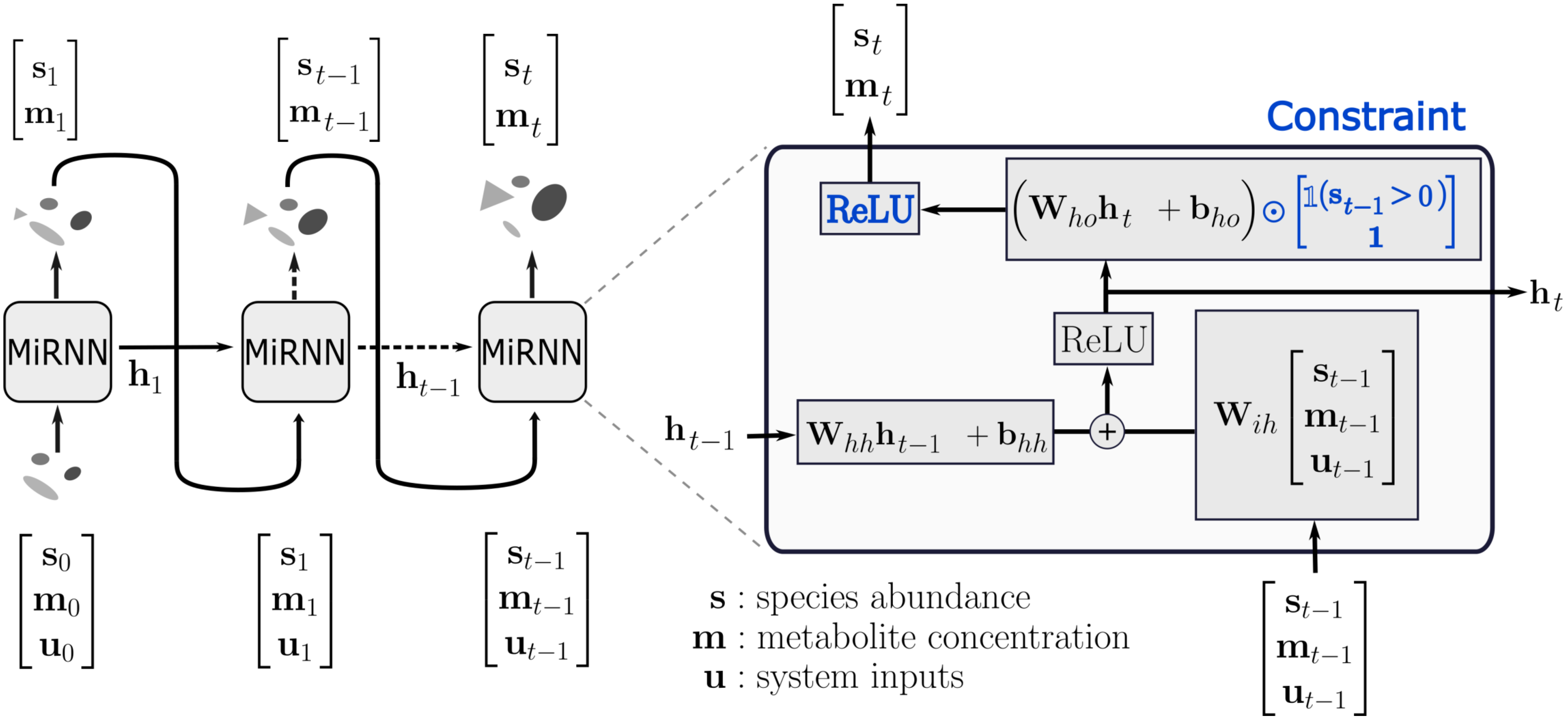
Upper confidence bound (UCB) sampling:

$$q^* = \underset{q}{\operatorname{argmax}} \quad \underbrace{\mu_{GP}(q) + \kappa \cdot \sigma_{GP}(q)}_{\text{acquisition function}}$$



Standard GP Bayesian optimization ignores ability to perform experiments in parallel and the model structure is not tailored to characterize system dynamics

Microbiome recurrent neural network (MiRNN) prevents physically unrealistic predictions



$\begin{bmatrix} \mathbf{s}_1 \\ \mathbf{m}_1 \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_{t-1} \\ \mathbf{m}_{t-1} \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_t \\ \mathbf{m}_t \end{bmatrix}$
MiRNN
$\mathbf{h}_1$
MiRNN
$\mathbf{h}_{t-1}$
MiRNN
$\begin{bmatrix} \mathbf{s}_0 \\ \mathbf{m}_0 \\ \mathbf{u}_0 \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_1 \\ \mathbf{m}_1 \\ \mathbf{u}_1 \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_{t-1} \\ \mathbf{m}_{t-1} \\ \mathbf{u}_{t-1} \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_t \\ \mathbf{m}_t \end{bmatrix}$
Constraint
ReLU
$\left(\mathbf{W}_{ho}\mathbf{h}_t + \mathbf{b}_{ho}\right) \odot \begin{bmatrix} \mathbb{1}(\mathbf{s}_{t-1} > 0) \\ \mathbf{1} \end{bmatrix}$
$\mathbf{h}_t$
ReLU
$\mathbf{h}_{t-1}$
$\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_{hh}$
+
$\mathbf{W}_{ih} \begin{bmatrix} \mathbf{s}_{t-1} \\ \mathbf{m}_{t-1} \\ \mathbf{u}_{t-1} \end{bmatrix}$
$\begin{bmatrix} \mathbf{s}_{t-1} \\ \mathbf{m}_{t-1} \\ \mathbf{u}_{t-1} \end{bmatrix}$

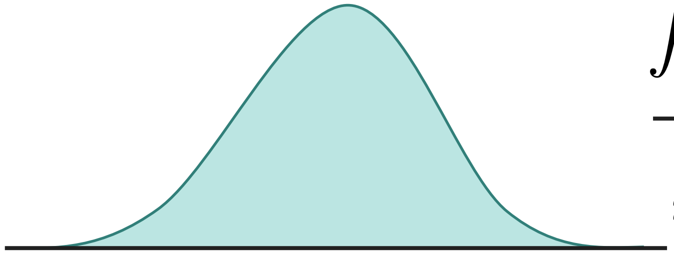
$\mathbf{s}$: species abundance
$\mathbf{m}$: metabolite concentration
$\mathbf{u}$: system inputs

Bayesian inference of parameter and model prediction distributions

A model, $\mathcal{M}$, predicts outcomes, $\mathbf{y}$, using parameters, θ, under condition, q_i

$$\mathbf{y}(q_i) = \mathcal{M}(\theta, q_i) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_y)$$

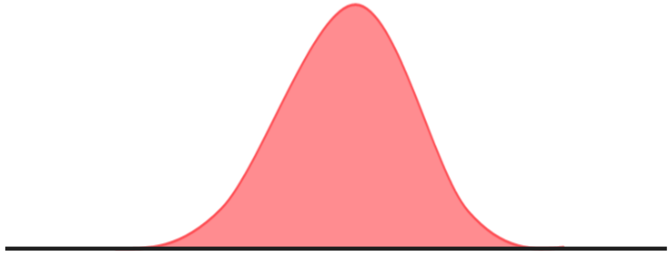
A set of experimental conditions $\mathbf{q} = \{q_1, ..., q_n\}$ provides data $\mathcal{D}(\mathbf{q}) = \{\mathbf{y}(q_1), ..., \mathbf{y}(q_n)\}$



$\theta|\mathbf{q}$

Posterior parameter distribution

$$\xrightarrow[\approx \mathcal{N}(\mathbf{y}|\mathcal{M}(q_i,\theta), \Sigma_y(q_i))]{\int_\theta p(\mathbf{y}|q_i,\theta)\; p(\theta|\mathcal{D}(\mathbf{q}))d\theta}$$

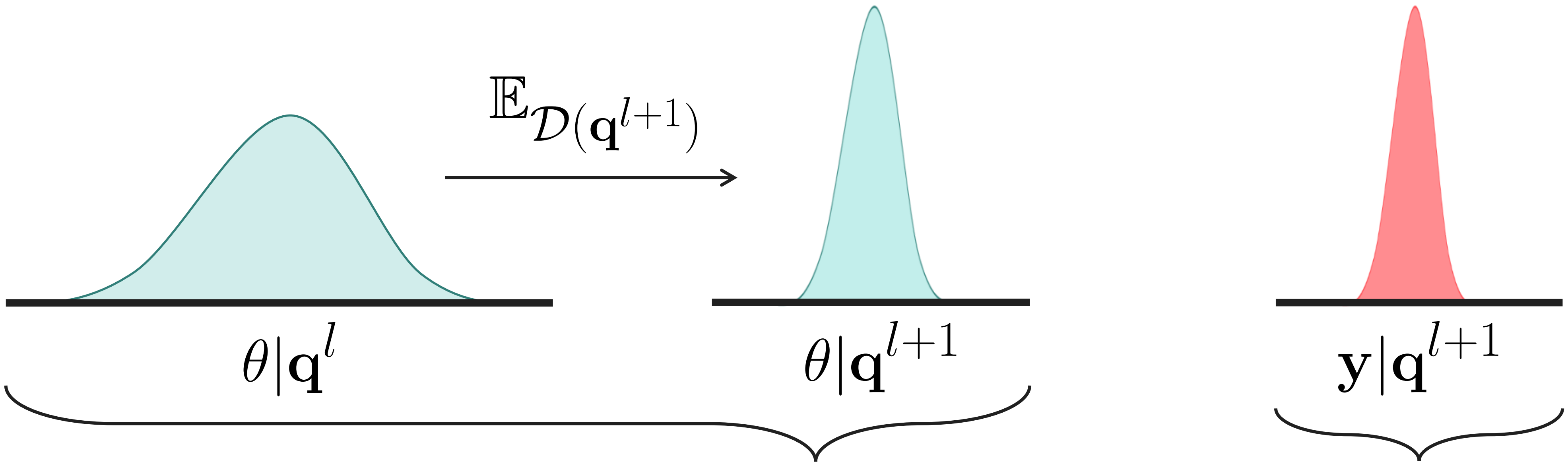


$\mathbf{y}|q_i, \mathbf{q}$

Posterior predictive distribution

Acquisition function balances information content and predicted outcomes to rank experimental designs

$$\text{Given previous data } \mathcal{D}(\mathbf{q}^{l}),\ \text{find next design } \mathbf{q}^{l+1} = \{q_1^{l+1}, \dots, q_n^{l+1}\}$$


$\xrightarrow{\mathbb{E}_{\mathcal{D}(\mathbf{q}^{l+1})}}$
$\theta|\mathbf{q}^{l}$
$\theta|\mathbf{q}^{l+1}$
$\mathbf{y}|\mathbf{q}^{l+1}$

$$\underset{\text{Acquisition function}}{f\left[\mathbf{q}^{l+1}\right]} = \underset{\text{Information function}}{f_I\left[\theta|\mathbf{q}^{l+1}\right]} + \underset{\text{Profit function}}{f_P\left[\mathbf{y}|\mathbf{q}^{l+1}\right]}$$

Bayesian experimental design using the MiRNN outperforms a conventional GP approach

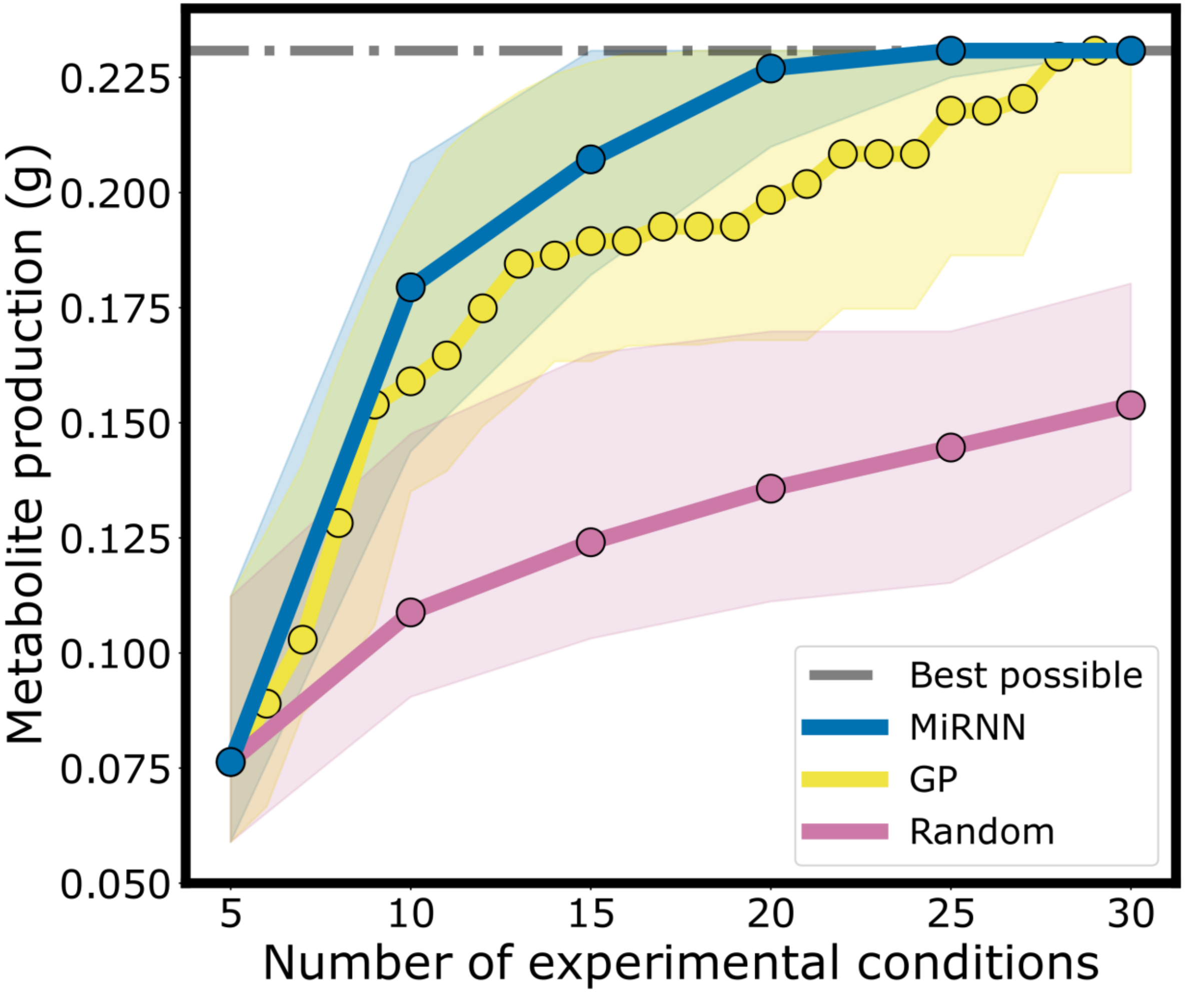
Given $\mathcal{D}(\mathbf{q}^l)$, infer $p(\theta|\mathcal{D}(\mathbf{q}^l))$

Optimize next design:

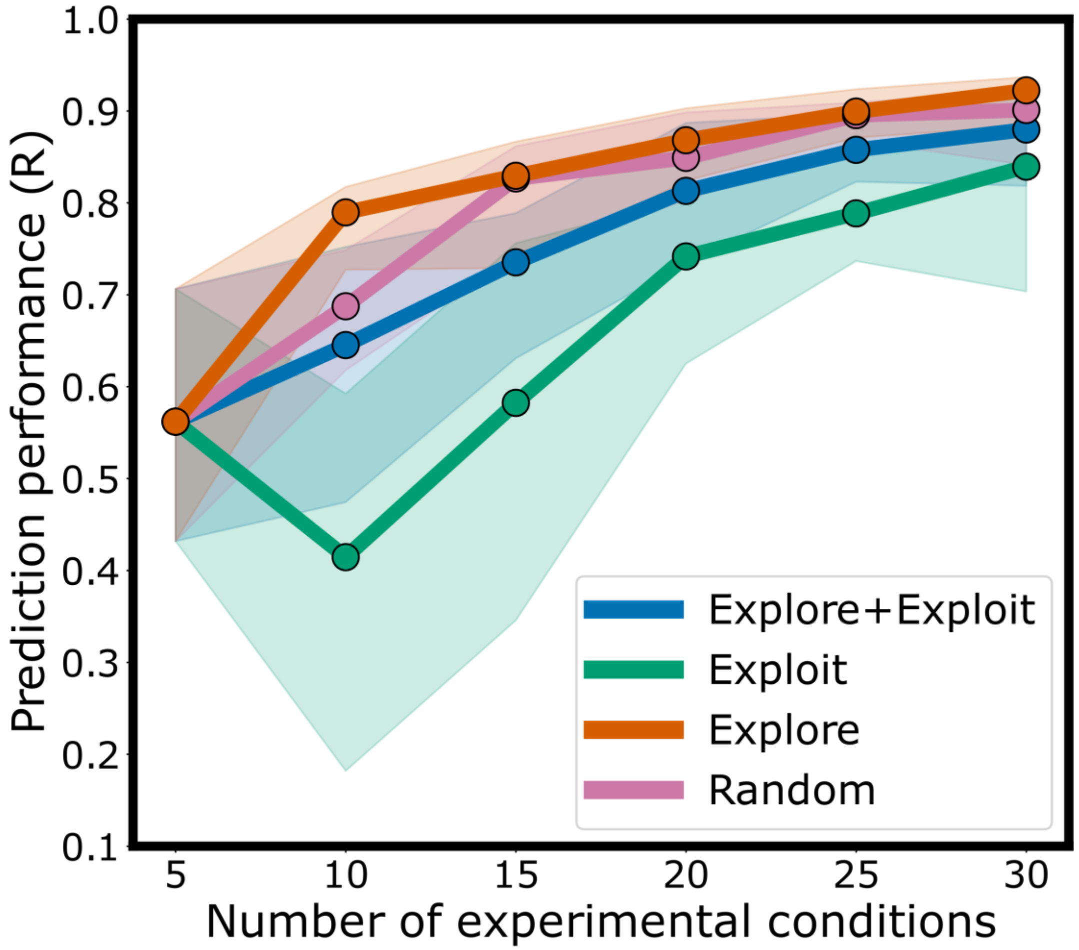
$$\mathbf{q}^{l+1} = \underset{\mathbf{q}}{\text{argmax}} \quad f_I\left[\theta|\mathbf{q}\right] + f_P\left[\mathbf{y}|\mathbf{q}\right]$$

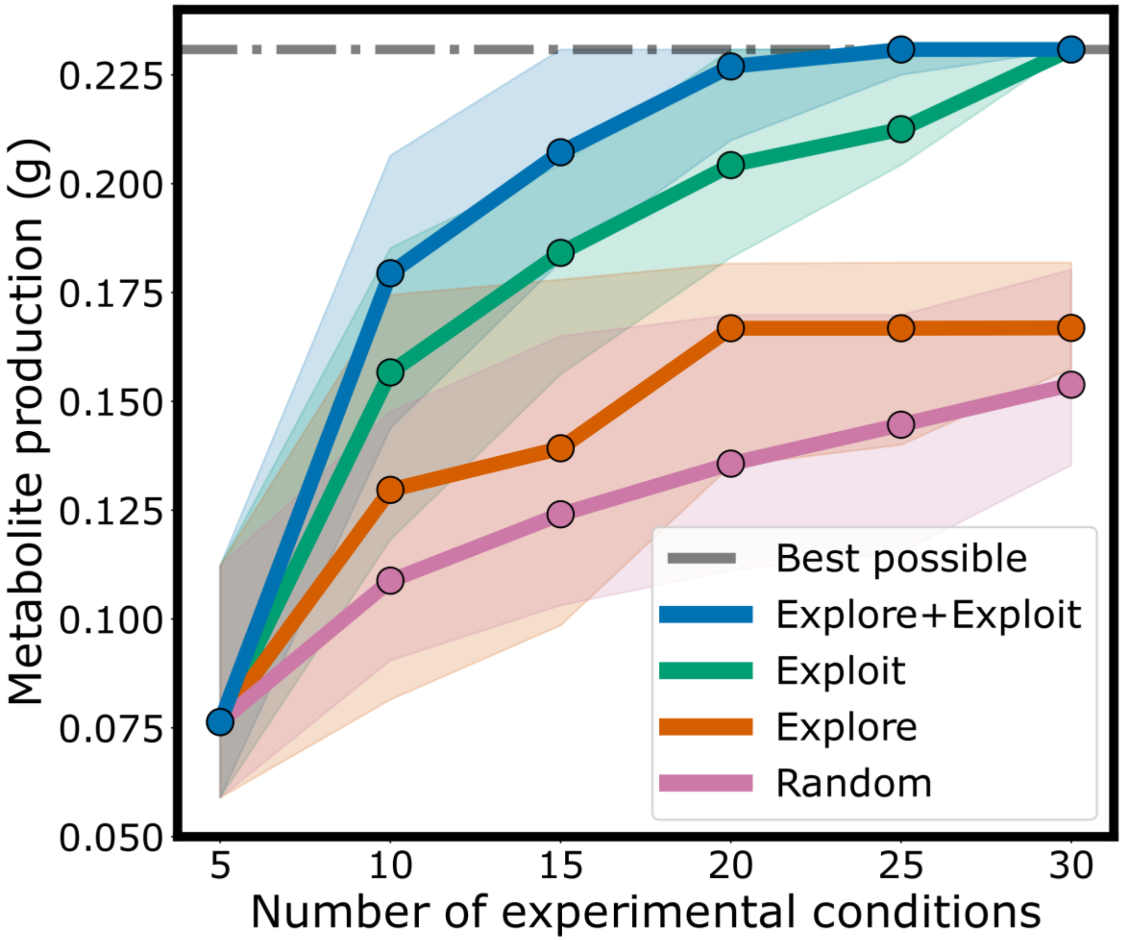
Update and repeat:

$$\mathbf{q}^l \leftarrow \mathbf{q}^l \cup \mathbf{q}^{l+1}$$



Combining exploration with exploitation outperforms either approach alone (MiRNN)

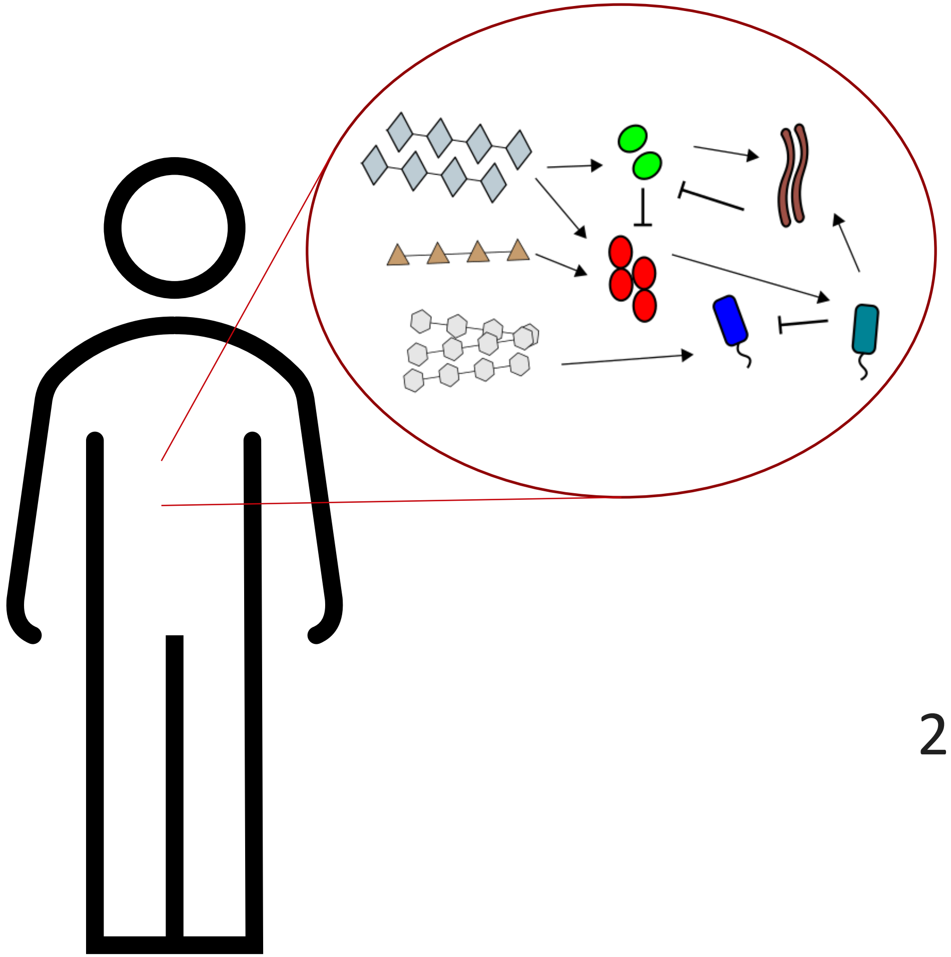




Bayesian optimization of a synbiotic

In collaboration with

Motivation to develop a defined microbial consortia synbiotic

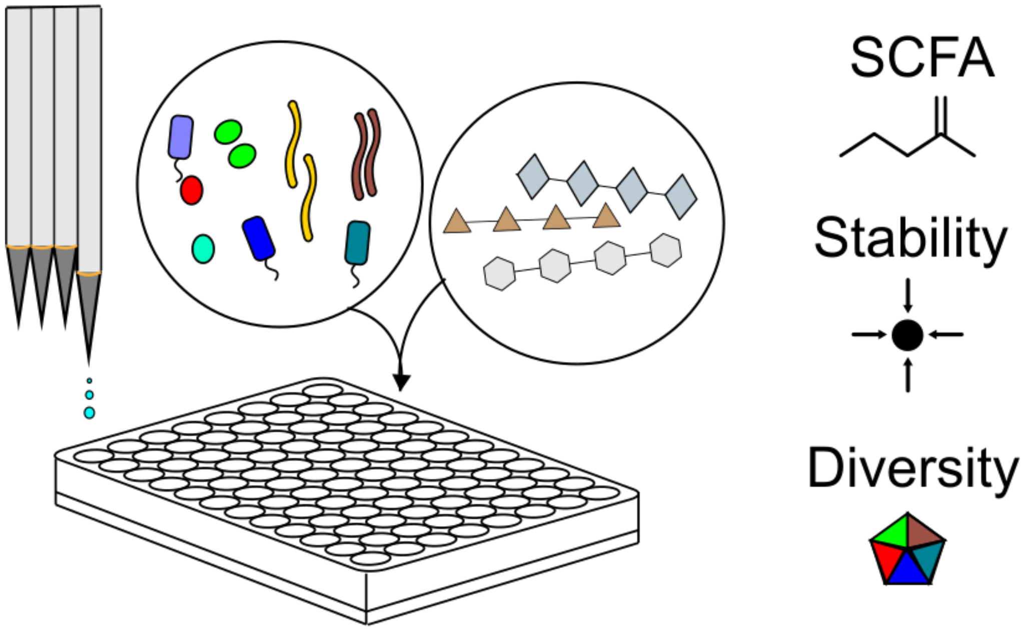


1. As a therapeutic
   - Over 450,000 annual C.diff infections in U.S.
   - FMTs are effective treatments for CDI but suffer significant limitations
   - Emerging treatments use defined consortia (Vedanta)

2. Over the counter supplements
   - Global market >$50 billion
   - Projected >$88 billion by 2026

Experimental design space for synbiotic engineering

- 15 representative gut “probiotic” bacteria
- 6 “prebiotic” fibers
- $(2^{15} - 1)*(2^{6} - 1) > 2$ million possible “synbiotics”
- Design objective: Select 288 conditions that balance exploration and exploitation of
    - Short chain fatty acid (SCFA) production
    - Stability (over time)
    - Diversity (representation of all species)

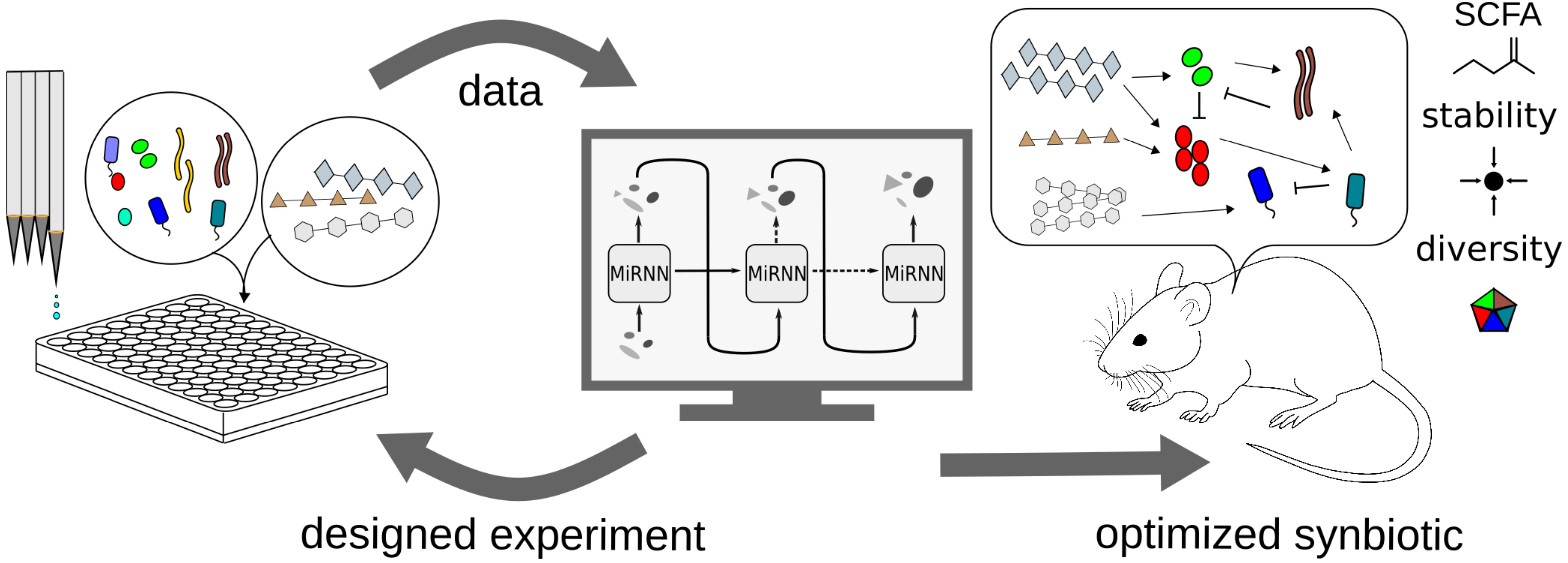


SCFA

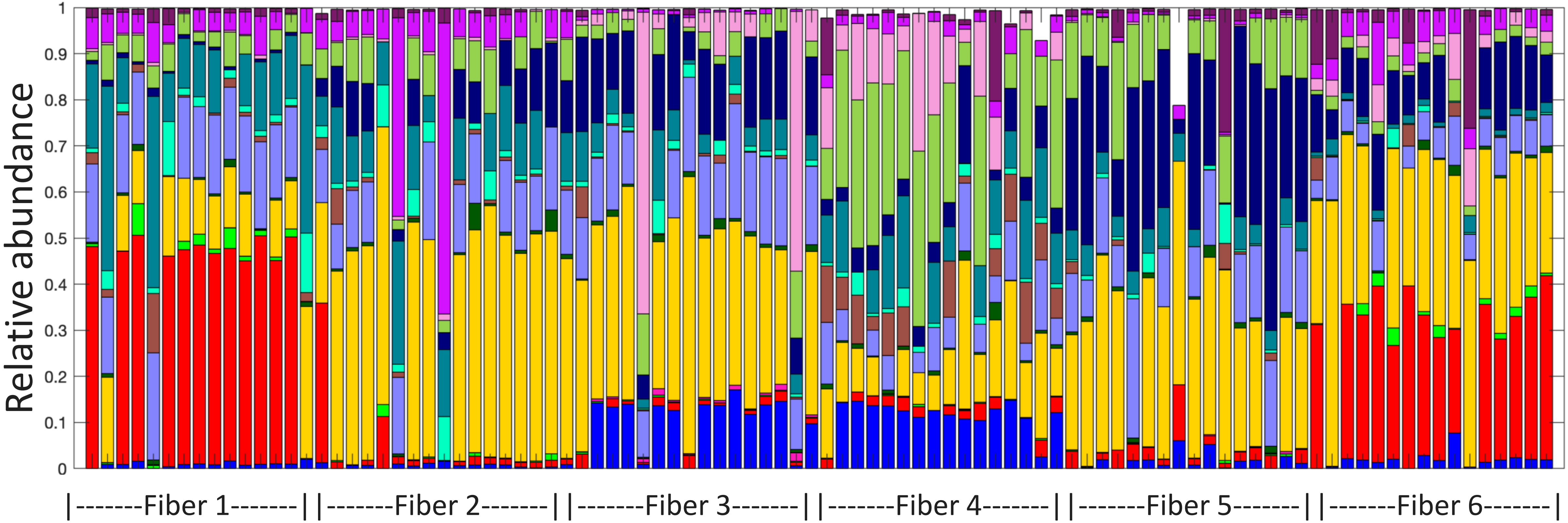
Stability

Diversity

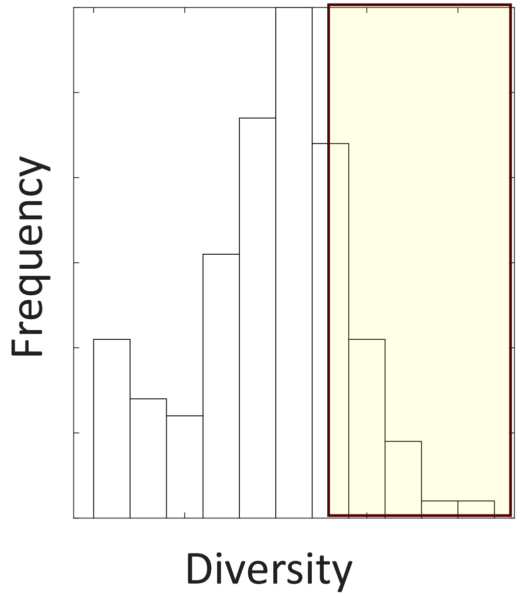
Model guided experimental design to optimize a synbiotic

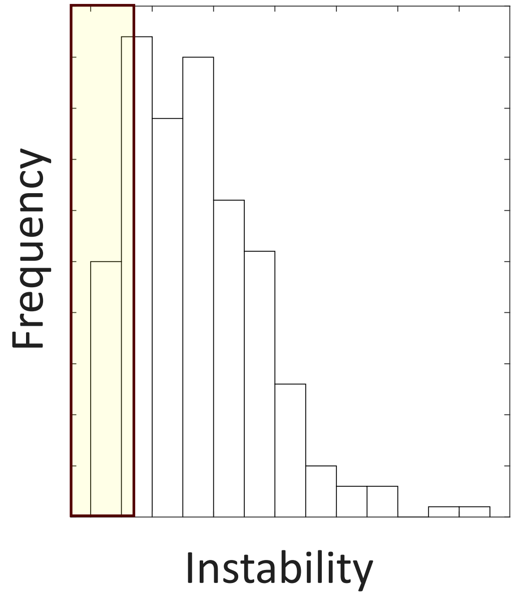


Cycle 0: Fiber determines community outcomes

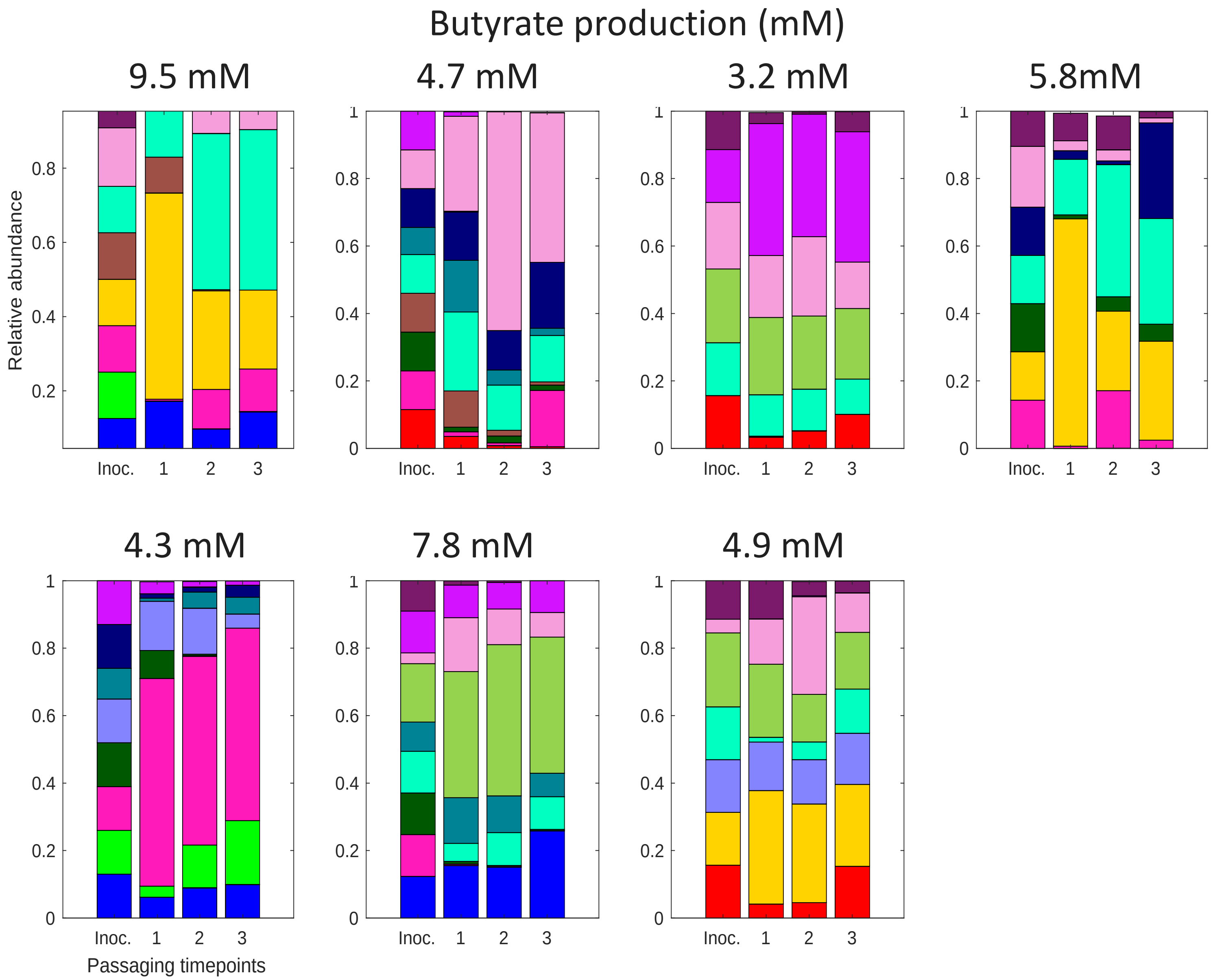


Cycle 1: Exploration finds diverse, stable communities





Intersection of comms w/ >80th percentile diversity and < 20th percentile instability

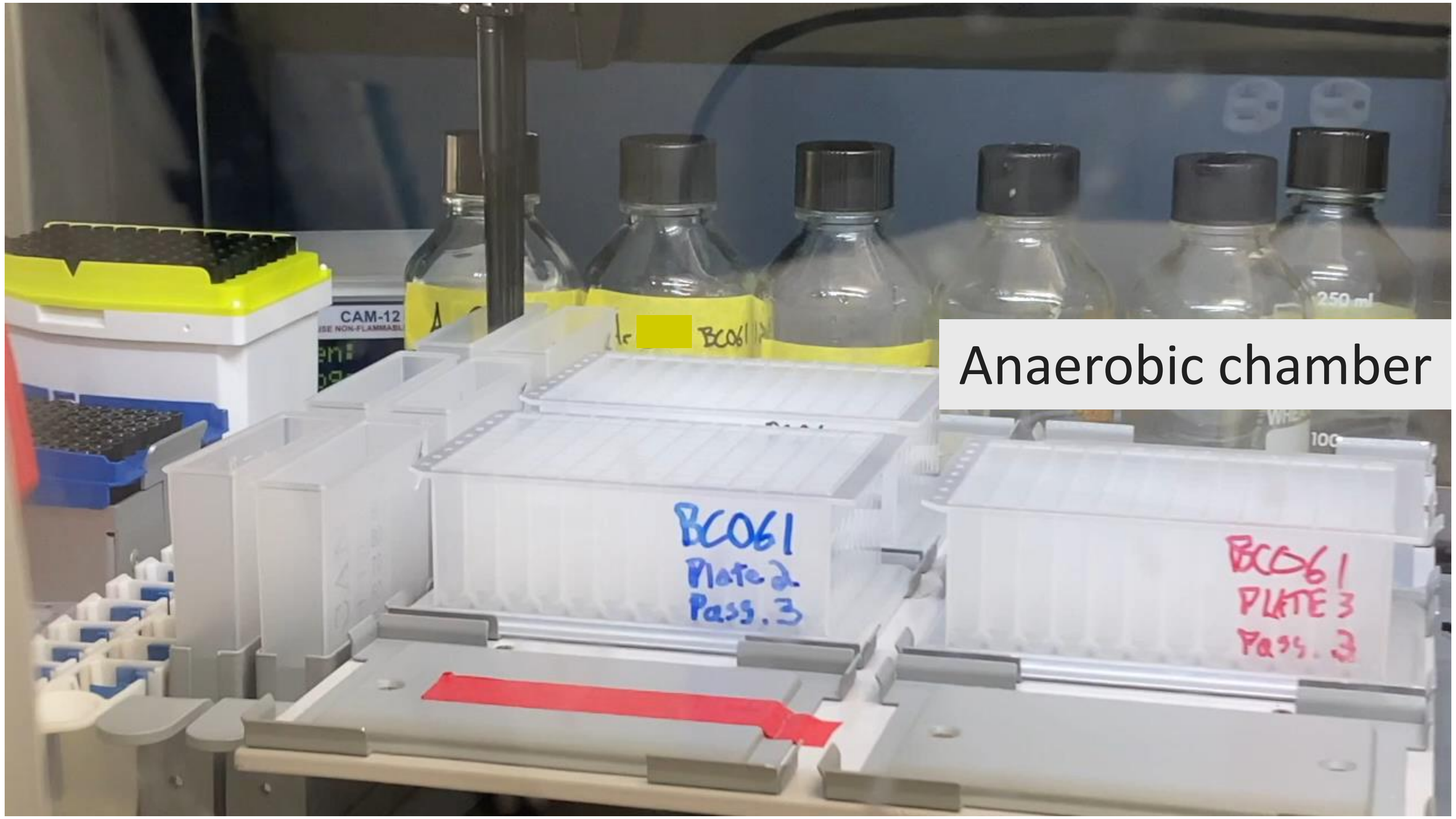


Future direction: Combine exploration with exploitation

- Preliminary data comprises about .02 % of conditions

- Design objective:
  - SCFA production
  - Stability
  - Diversity

- Convergence criteria:
  - Model performance
  - Measured objective



Anaerobic chamber

Manuscript, data, and code availability:

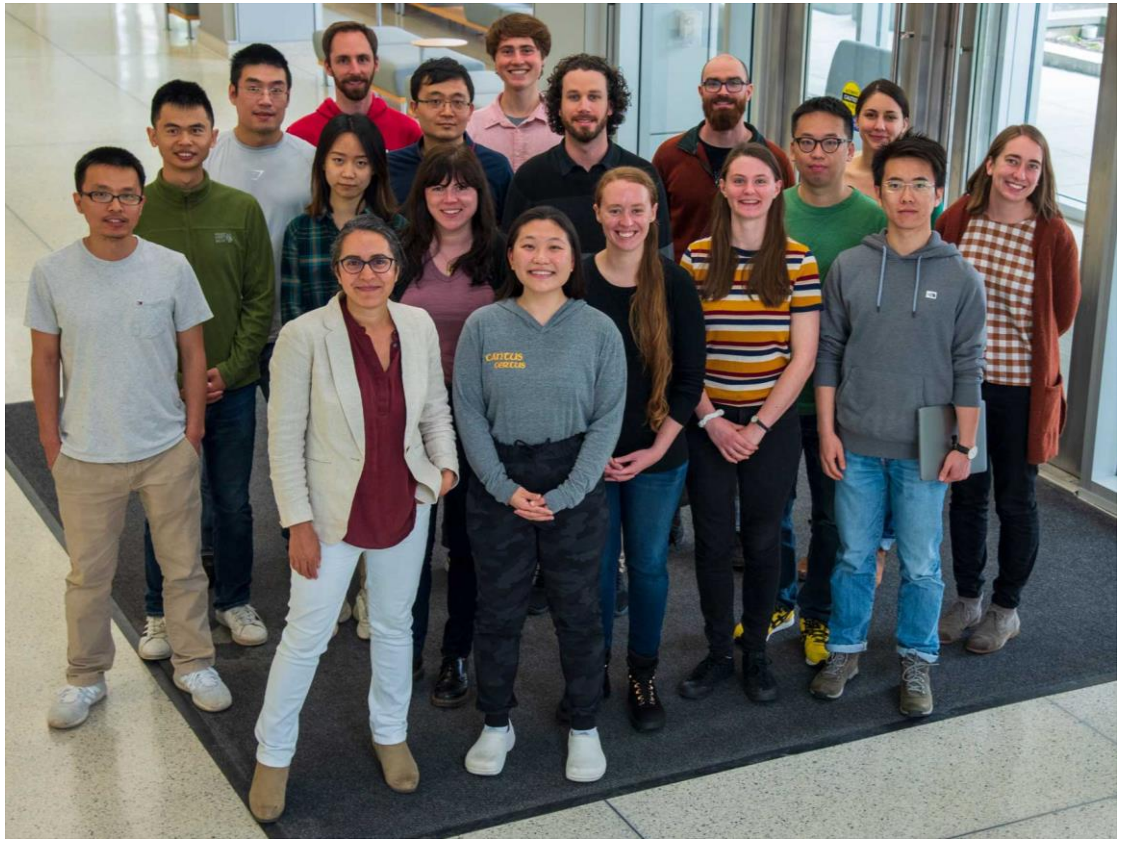
Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions Accepted in PLoS Comp Biology
bioRxiv: https://doi.org/10.1101/2022.11.12.516271

Installable Python package + Jupyter notebook tutorials:
https://github.com/VenturelliLab/Thompson_et_al_2023

Please email me at: jcthompson5@wisc.edu

Acknowledgements: Venturelli Lab

Yu-Yu Cheng, PhD
Freeman Lan, PhD
Erin Ostrem Loss, PhD
Claire Palmer, PhD
Yili Qian, PhD
Jordy Suliaman, PhD
Eloi Martinez-Rabert, PhD
Job Grant, PhD
Bryce Connors
Julie DuClos
Madeline Hayes
Wenbo Lu
Yiyi Liu
Yifei Ren
Tyler Ross
Tyson Wheelwright



Ophelia Venturelli

Army Research Office

National Institute of Biomedical Imaging and Bioengineering

Acknowledgements: Zavalab

Aurora Munguia

Daniel Laky

Bruce Jiang

Weiqi Zhang

Bo-Xun Wang

Amy Qin

Jiaze Ma

Leo Gonzalez

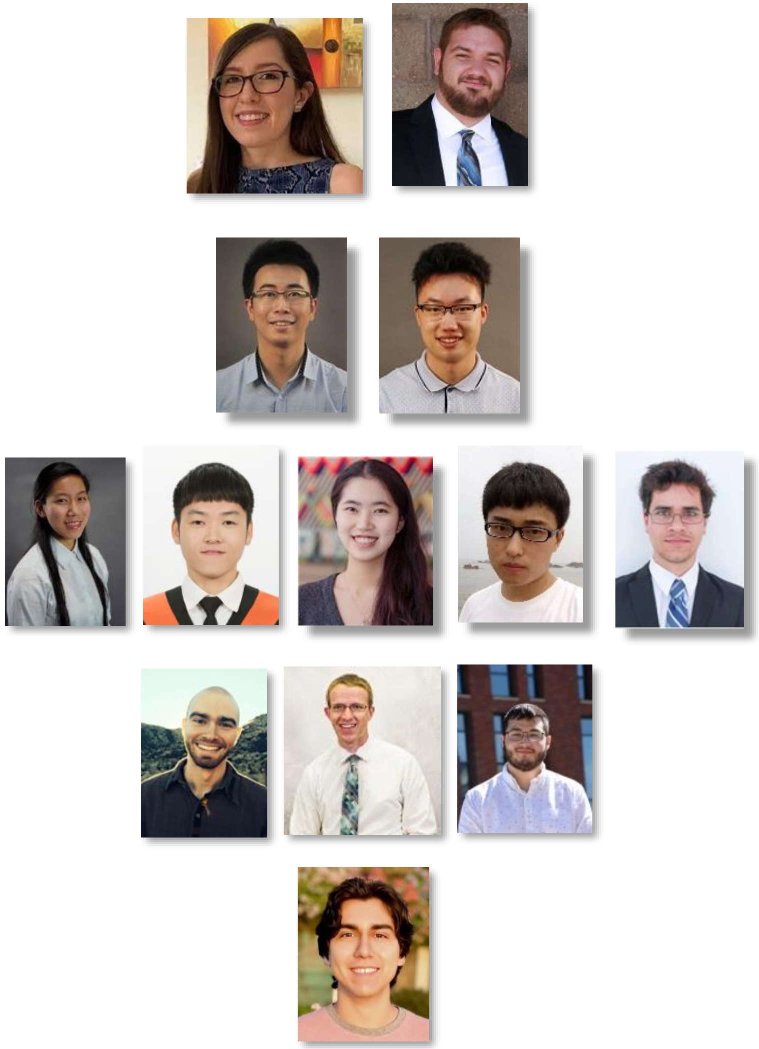
Lisa Je

Jaron Thompson

David Cole

Blake Lopez

Elvis Umana



Victor Zavala

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Approximate Bayesian inference

Assume a Gaussian sampling distribution for each of $j = 1, \ldots, n_y$ outcomes

$$y_j(q_i) = \mathcal{M}_j(q_i, \theta) + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$$

Assume a Gaussian prior distribution for each of $k = 1, \ldots, n_\theta$ parameters

$$p(\theta_k|\alpha_k) = \mathcal{N}(\theta_k|0, 1/\alpha_k)$$

Precision of sampling noise and parameter prior are model hyper-parameters $\xi = \{\alpha_1, \ldots, \alpha_k, \sigma_1, \ldots, \sigma_{n_y}\}$

Bayesian inference objective is to determine an approximate posterior $z(\theta|\phi) \approx p(\theta|\mathcal{D}(\mathbf{q}))$

$$\log p(\mathcal{D}(\mathbf{q})|\xi) = \underbrace{\int_\theta \log\left(\frac{p(\mathcal{D}(\mathbf{q}), \theta|\xi)}{z(\theta|\phi)}\right) z(\theta|\phi) d\theta}_{\mathcal{L}(z(\theta|\phi),\xi)} + \underbrace{\int_\theta -\log\left(\frac{p(\theta|(D)(\mathbf{q}), \xi)}{z(\theta|\phi)}\right) z(\theta|\phi) d\theta}_{\mathrm{KL}}$$

Function to quantify information content

A set of experimental conditions $\mathbf{q} = \{q_1, ..., q_n\}$ provides data $\mathcal{D}(\mathbf{q}) = \{\mathbf{y}(q_1), ..., \mathbf{y}(q_n)\}$

$$f_I(\mathbf{q}) := \mathbb{E}_{\mathcal{D}(\mathbf{q})}\left[\mathrm{KL}\left(p(\theta|\mathcal{D}(\mathbf{q}))||p(\theta)\right)\right]$$

$$\approx \ln\det\ \left(\mathbf{\Sigma}_\theta{}^{-1} + \textstyle\sum_{i=1}^n \mathbf{G}(q_i)^T \mathbf{\Sigma}_y^{-1} \mathbf{G}(q_i)\right) - \ln\det\ \left(\mathbf{\Sigma}_\theta{}^{-1}\right)$$

$$\mathbf{G}(q_i) := \nabla_\theta \mathcal{M}(q_i, \theta)$$

$$= \sum_{i=1}^n \ln\det\ \left(\mathbb{I}_{n_y} + \mathbf{\Sigma}_y^{-1} \mathbf{G}(q_i) \mathbf{A}_{i-1}^{-1} \mathbf{G}(q_i)^T\right)$$

$$\mathbf{A}_i = \mathbf{A}_{i-1} + \mathbf{G}(q_i)^T \mathbf{\Sigma}_y^{-1} \mathbf{G}(q_i)$$

$$\mathbf{A}_0 = \mathbf{\Sigma}_\theta^{-1}$$