A brief overview of Machine Learning and its application to Metabolic Engineering

Kenny Workman, undergrad at Berkeley

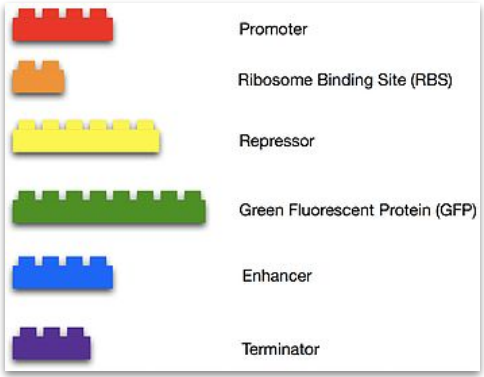
Revisiting Synthetic Biology

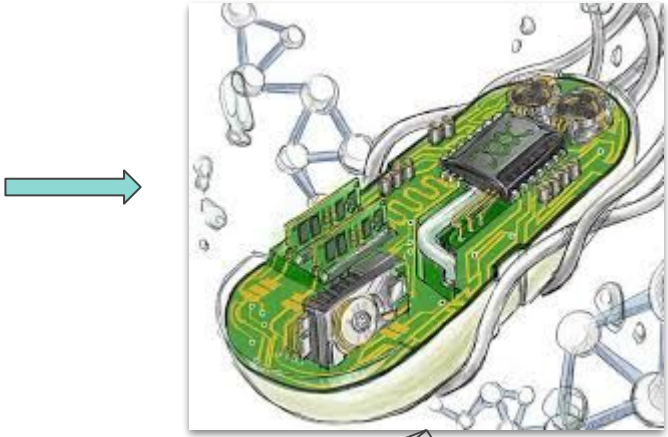
To reduce biology to reusable and modular components.

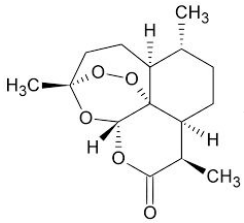
Put the pieces back together to make something new.

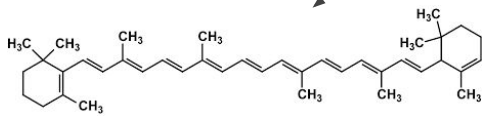
BioBricks FOUNDATION

iGEM









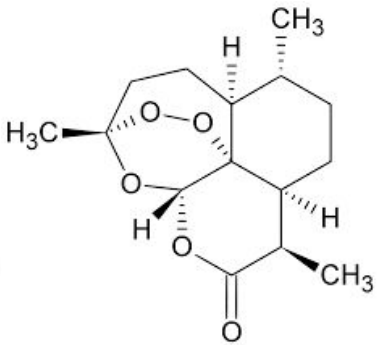
Modern Bottlenecks in Metabolic Engineering

Synthetic biology "remains a collection of elegant demonstrations rather than a systematic discipline".

amyris

150 Person-Years

Roosth, Sophia; *Synthetic*

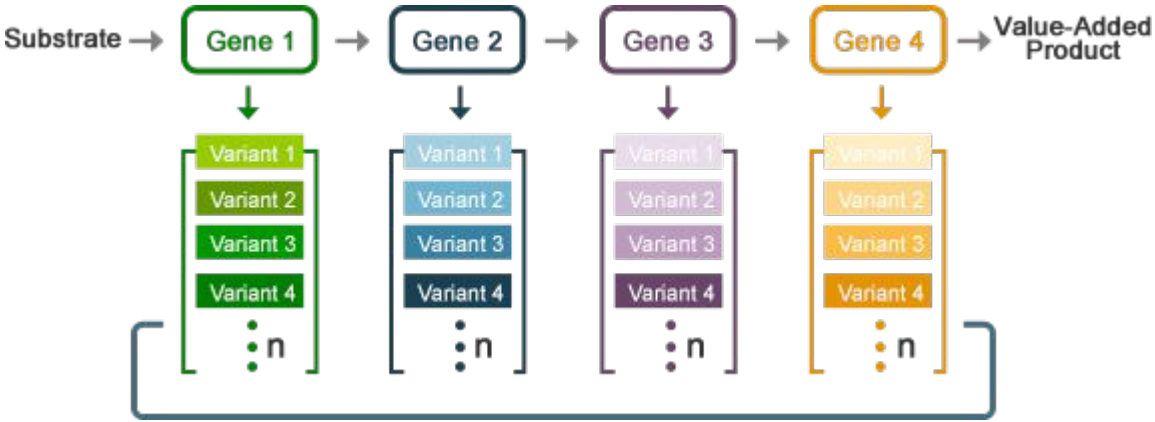


DUPONT

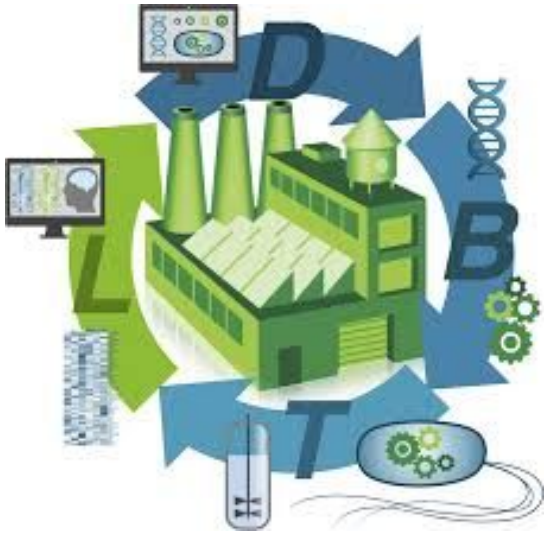
575 Person-Years

Roosth, Sophia; *Synthetic*

An Approach: Combinatorial Libraries and High-Throughput Assay



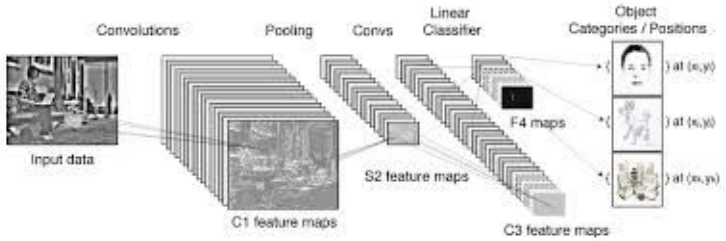
Alternative Approach: Machine Learning and multiple DBTL cycles



Use a fraction of your library constructs.

Emphasize the **learning** to extrapolate from the data.

Traditional Machine Learning Problem Domains



Novel Solution Generation w/ GANs and NLP

Look closely, but you won't catch it. This woman *does not exist.*

Legolas and Gimli advanced on the orcs, raising their weapons and shouting in unison. The orcs swung their weapons at them, but the two heroes were ready. They leapt from their horses and charged the orcs, slashing and slashing until their blades were dull and their weapons were broken. They were the first to fall, but the orcs were not done yet. They rushed the two heroes, and they fell in a hail of arrows.

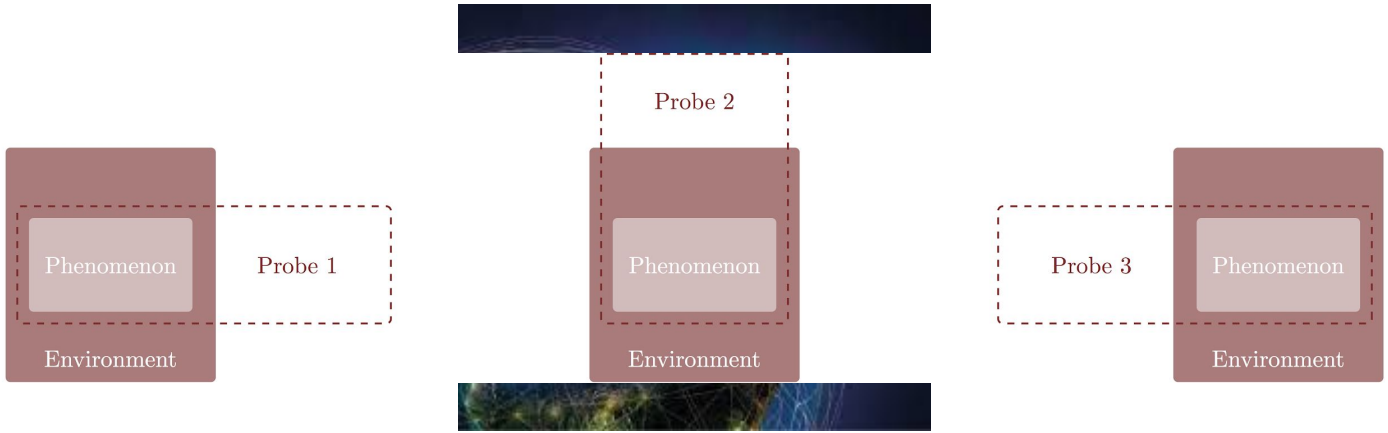
The orcs were not finished yet, however. They charged the two heroes again, and they fell in a hail of arrows. The orcs were not finished yet. They charged the two heroes again, and they fell in a hail of arrows.

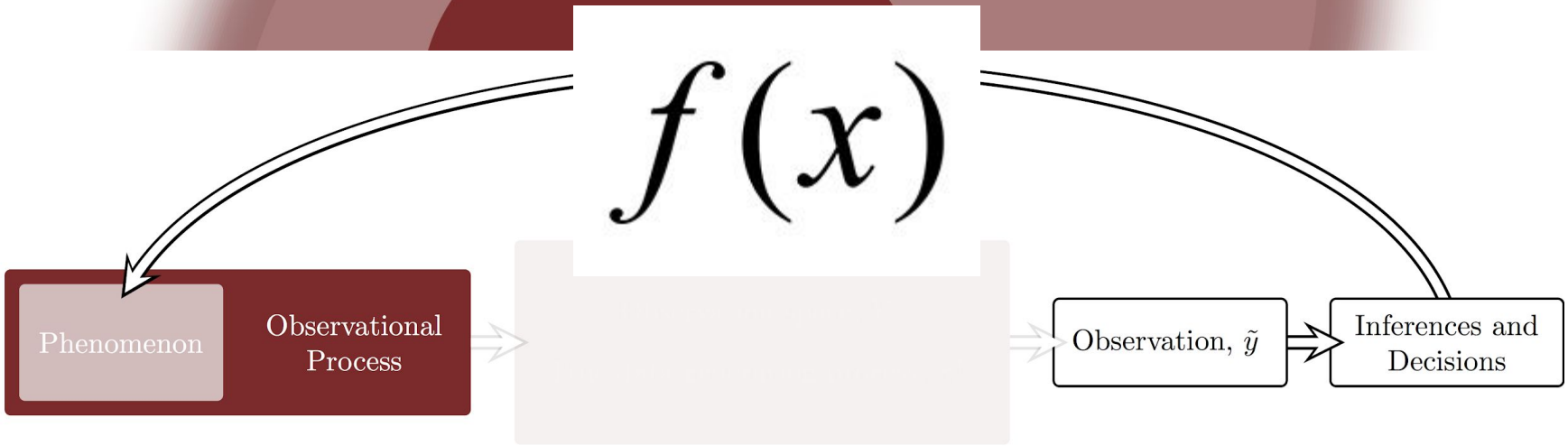
The orcs were not finished yet, however. They charged the two heroes again, and they fell in a hail of arrows. The orcs were not finished yet.. The orcs were not far off, and the two of them were not far from the edge of the cliff.

Gimli turned to face the orcs, who were now close enough to see him. He raised his hands and shouted, "I'm Gimli, son of Thráin, and I'm here to help you!"

The orcs were confused. They had seen Gimli before, but he was not the same man who had saved them from the orcs in the cave. Gimli was a tall and powerful man, and he had a beard and a moustache. He was also a dwarf, and he had a strong build, and he was covered in tattoos. He was not a man who looked like a hobbit.

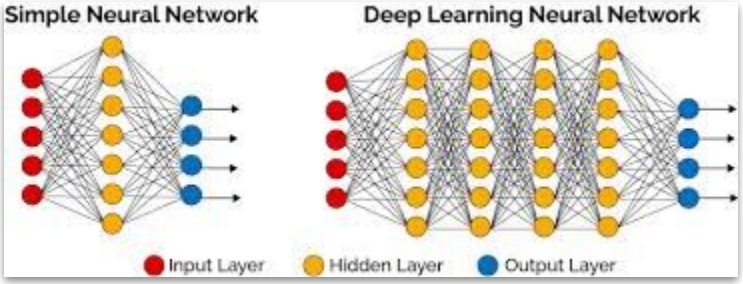
Fundamentals of Probability Theory

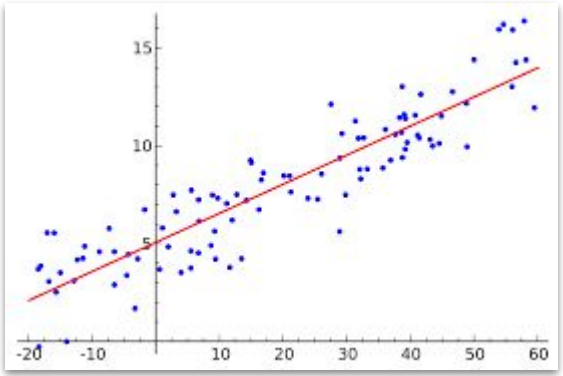



$f(x)$
Phenomenon
Observational Process
Observation, $\tilde{y}$
Inferences and Decisions

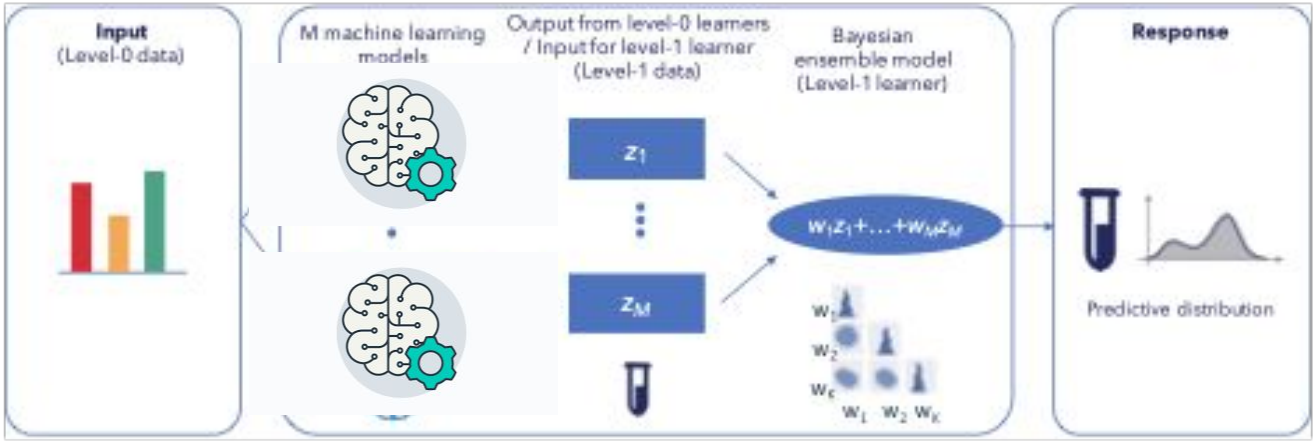
$\mathcal{P}$

How do we model latent phenomena?





$$F: \quad y = \boldsymbol{w}^T \mathbf{f}(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$



ART: an Automated Recommendation Tool

ART as both a predictive tool and a recommender

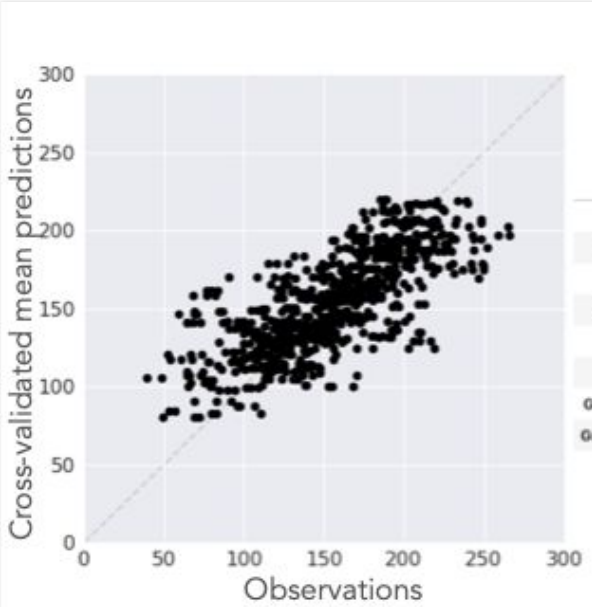
Generates recommendations to fit some objective:

- Maximization
- Minimization
- Specification

Quantifies Likelihood of the prediction being true:

There is a 15% chance that protein D will be expressed at x concentration.

Bayesian Ensemble approach can extrapolate from limited data



Mean Absolute Error

Name	MAE	MRAE(%)	StdAE	MSE	R^2
Neural Regressor	28.138447	20.361717	23.362420	1337.574858	0.280271
Random Forest Regressor	22.179447	16.572849	20.586204	915.719671	0.507265
TPOT Regressor	22.129173	16.830512	18.869021	845.740245	0.544920
Support Vector Regressor	31.538757	25.082998	23.121450	1529.294651	0.177109
Kernel Ridge Regressor	67.255258	40.803989	48.364739	6862.417693	-2.692567
K-NN Regressor	30.869251	22.341597	24.013975	1529.581659	0.176954
Gaussian Process Regressor	24.741041	18.732631	20.313311	1024.749716	0.448597
Gradient Boosting Regressor	23.318813	17.875105	20.019813	944.559977	0.491746
Ensemble Model	21.375663	16.256596	18.653554	804.874046	0.566909

The math behind ART

Training Model Parameters

$$F: \quad y = \boldsymbol{w}^T \mathbf{f}(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

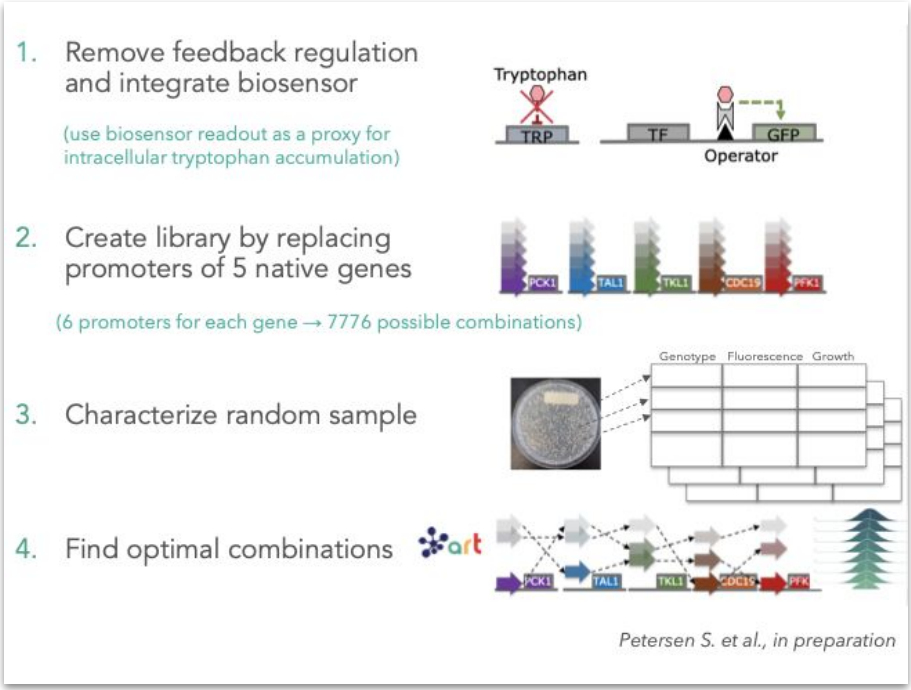
Producing a distribution of confidence in prediction

$$p(y|\mathbf{x}^*, \mathcal{D}) = \int p(y|\mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta} = \int \mathcal{N}(y; \boldsymbol{w}^T \mathbf{f}, \sigma^2) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta}.$$

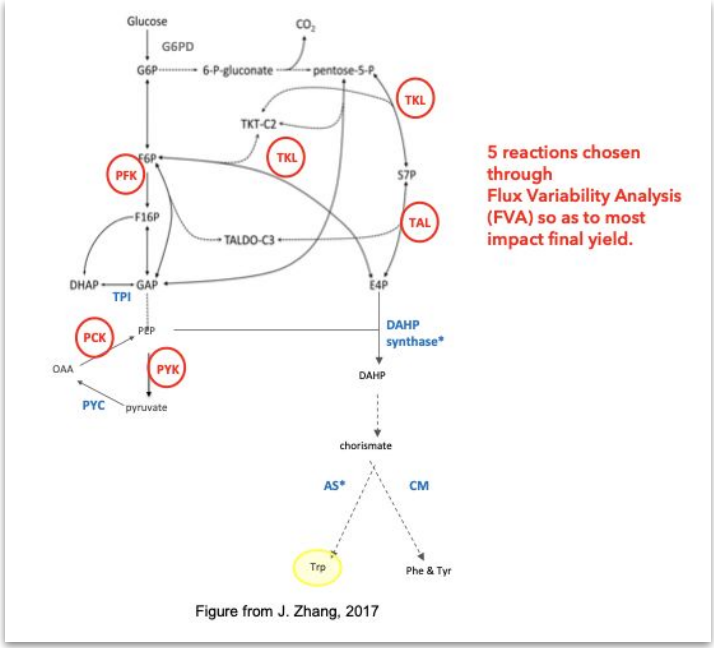
Optimizing this parameter space to some objective

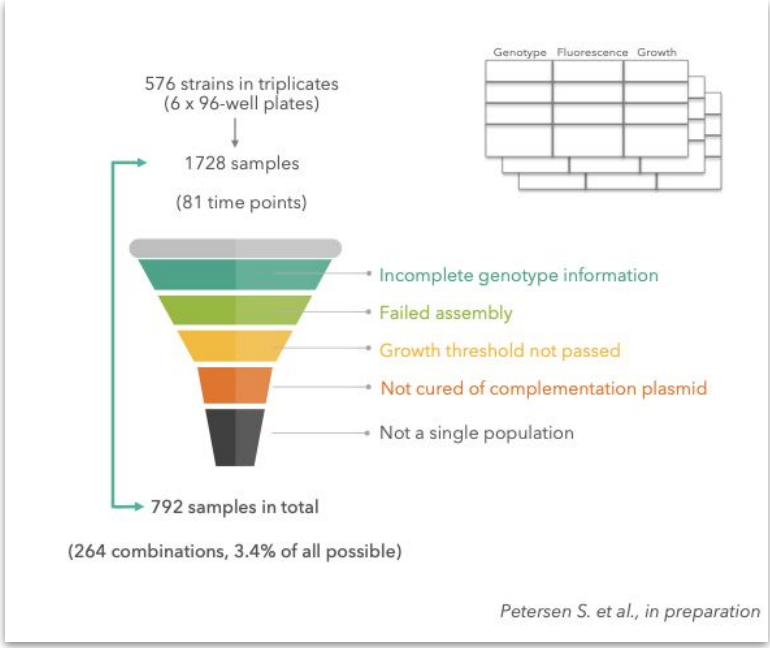
$$G(\mathbf{x}) = \begin{cases} (1-\alpha)\mathbf{E}(y) + \alpha \mathrm{Var}(y)^{1/2} & \text{(maximization case)} \\ -(1-\alpha)\mathbf{E}(y) + \alpha \mathrm{Var}(y)^{1/2} & \text{(minimization case)} \\ -(1-\alpha)||\mathbf{E}(y) - y^*||_2^2 + \alpha \mathrm{Var}(y)^{1/2} & \text{(specification case)} \end{cases}$$

A general strategy for applying Machine Learning to pathway design.

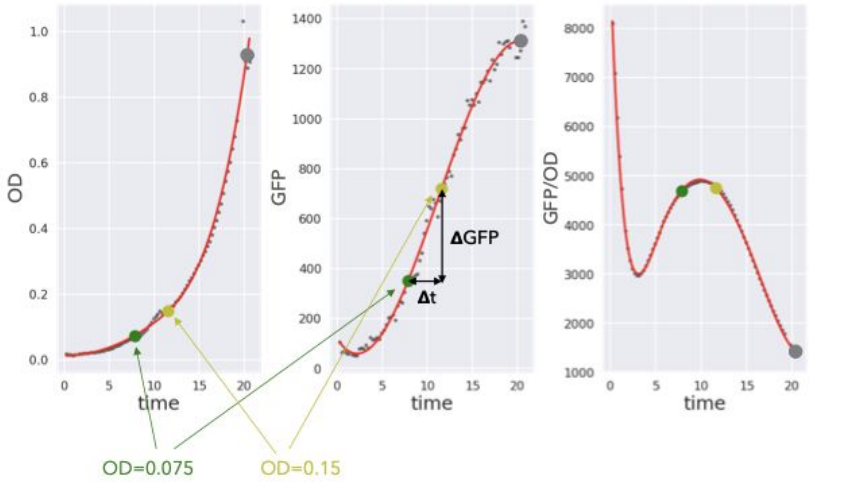


A Case Study: Optimizing Tryptophan production in *S. cerevisiae*

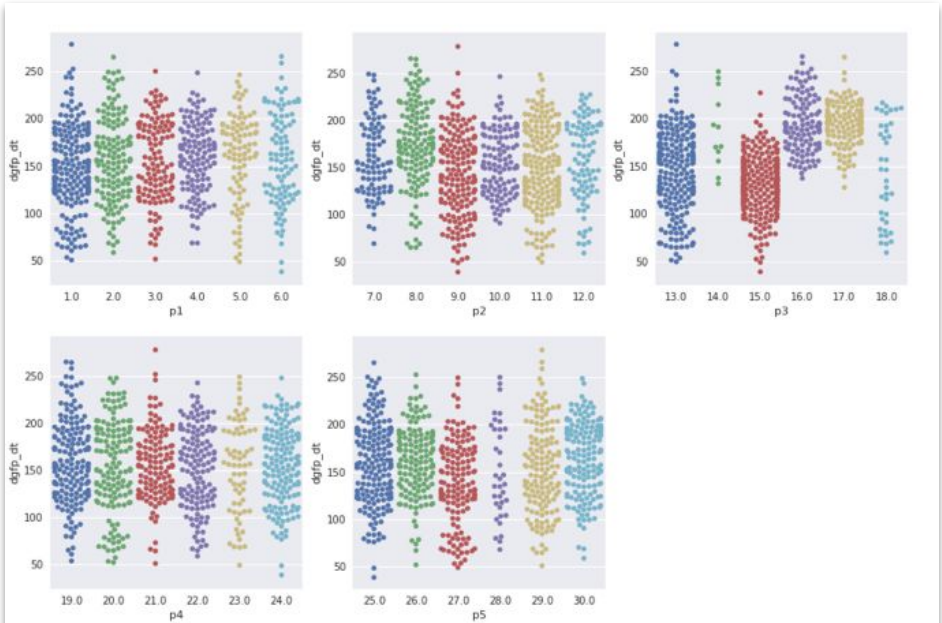




The relationship between production and metabolomics data is unclear.



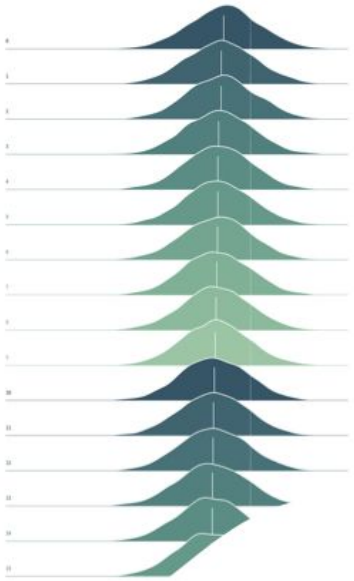
Petersen S. et al., in preparation



Petersen S. et al., in preparation

Constructs are currently being tested

	p1	p2	p3	p4	p5	dgfp_dt
0	3	8	16	22	28	242.419
1	3	8	16	19	28	240.314
2	3	8	16	22	29	240.233
3	3	8	16	19	29	238.07
4	3	8	16	22	25	237.457
5	3	8	16	19	26	237.453
6	3	8	16	22	26	237.2
7	6	8	16	22	29	236.44
8	3	8	16	19	25	235.903
9	2	8	16	19	25	235.204
10	2	8	16	22	28	234.39
11	5	8	16	22	29	233.748
12	2	8	16	19	28	233.315
13	2	8	16	22	25	232.858
14	2	8	16	23	29	232.68
15	2	8	16	22	29	

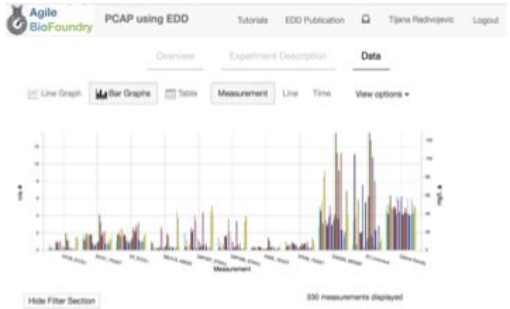


art

Petersen S. et al., in preparation

A Contrived Case Study: Improving Limonene production in *E. coli*

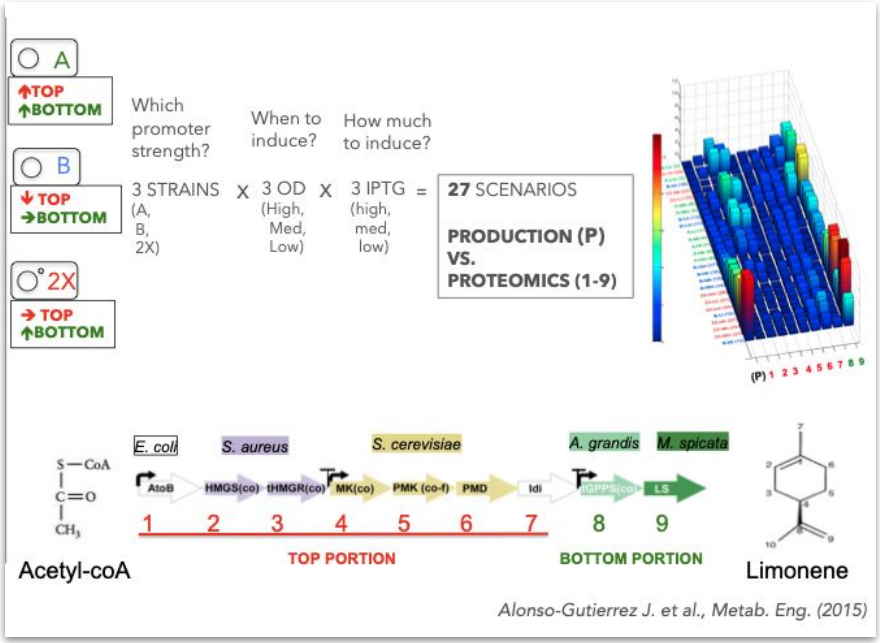
- **Data**: collection of proteomics and limonene production data

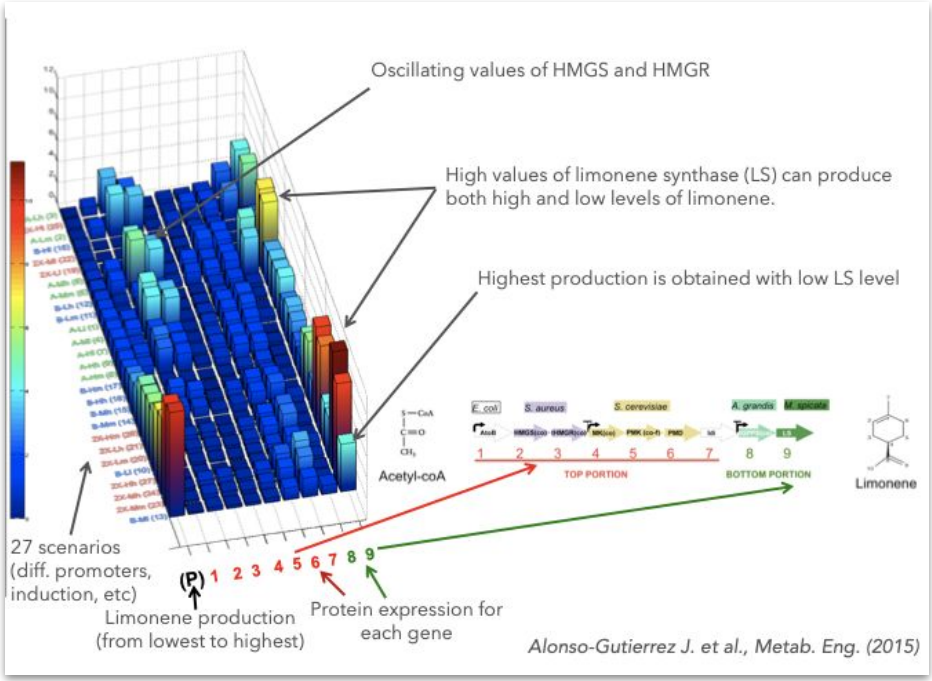


- **Objective:** identify proteins that need to have their expression level adjusted to maximize limonene production

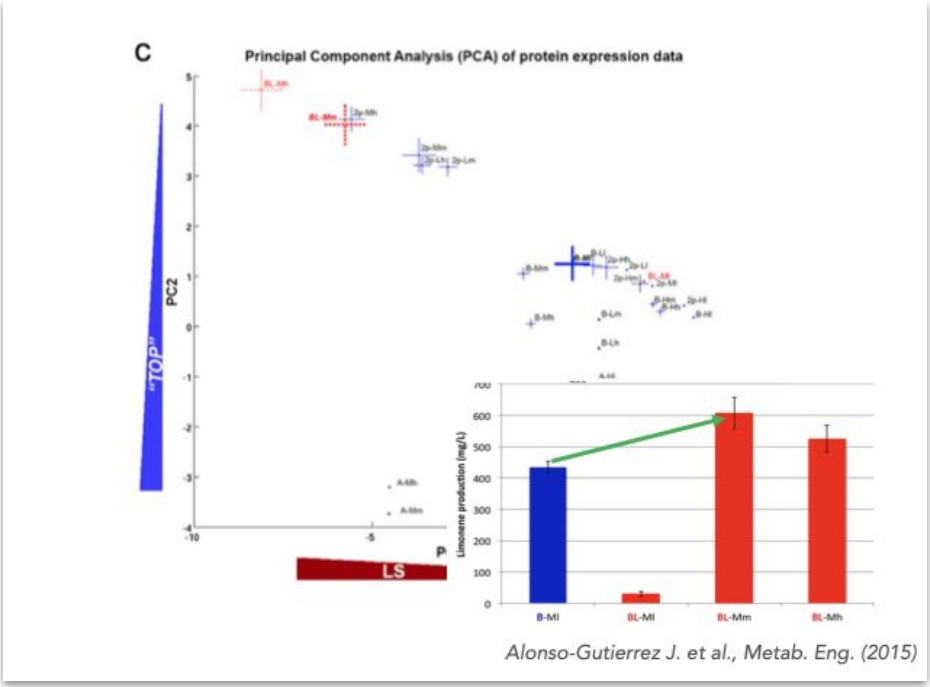
Alonso-Gutierrez J. et al., Metab. Eng. (2015)

Combinatorial libraries yield only 24 constructs from which to extrapolate.

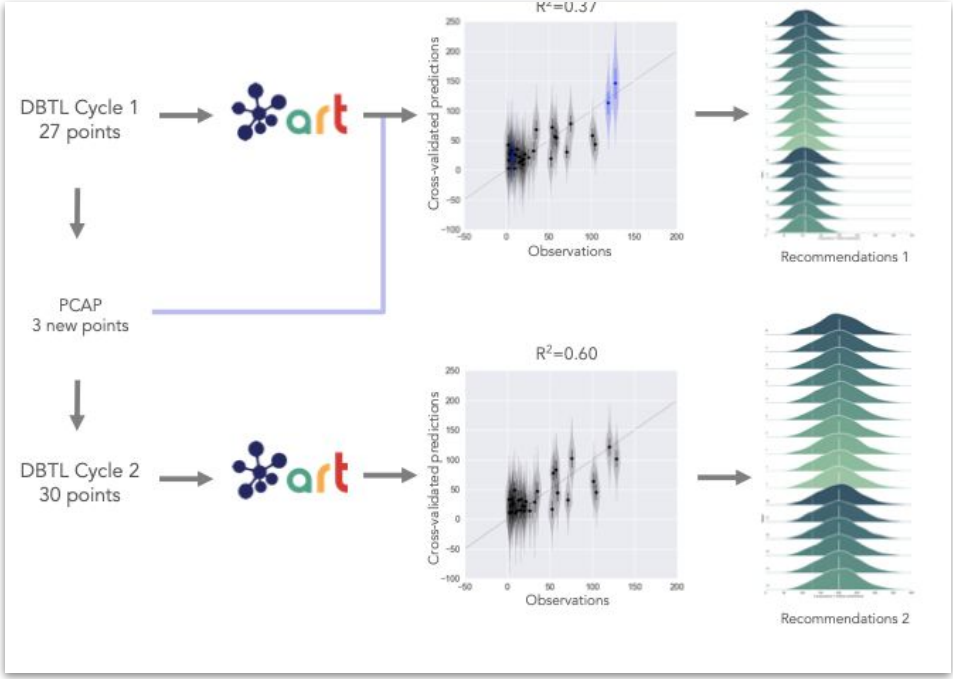


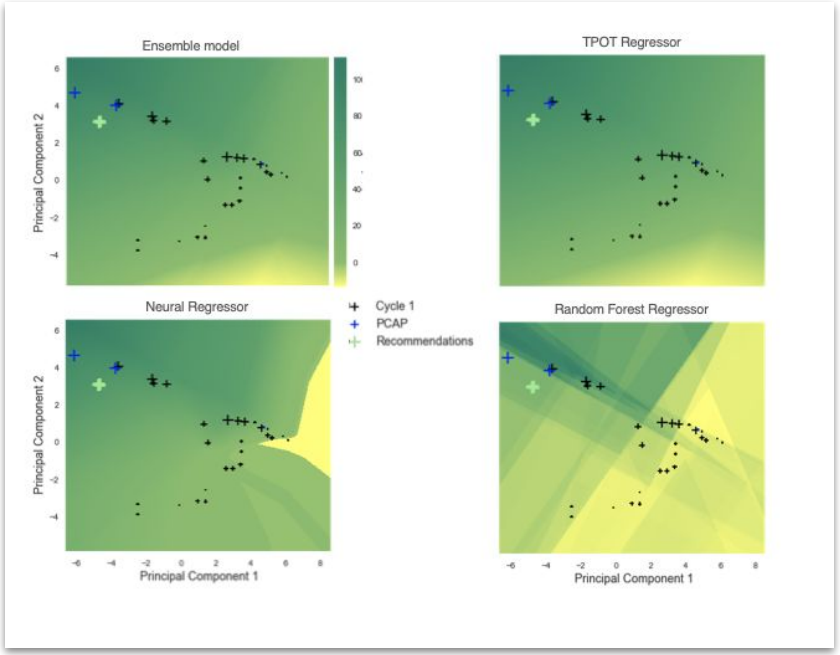


PCAP (Principal Component Analysis of Proteomics) suggests ~40% improvements.



ARTs predictions match and expand upon experimentally validated PCAP.





The importance of generalizable Machine Learning in Synthetic Biology

Opaque data; Pipeline not reproducible.

GINKGO BIOWORKS
THE ORGANISM COMP

Arzeda.

zymergen

Easy integration with virtually any repeatable experiment