LYGOS

DOE Bioenergy Technologies Office (BETO) 2023 Project Peer Review

DE-EE0008489: Accelerating engineered microbe optimization through machine learning and multi-omics datasets

2023-04-04

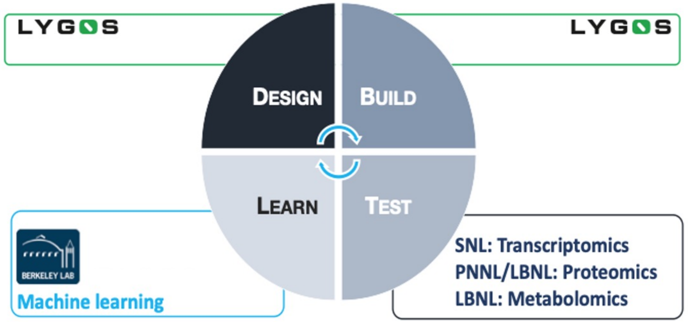
Presenter: Rebecca Lennen
Title: Director of Strain Engineering
Lygos, Inc.

This presentation does not contain any proprietary, confidential, or otherwise restricted information

Agile BioFoundry

Project overview

- Leverage multi-omics datasets to populate machine learning networks
- Make predictions on how to engineer *P. kudriavzevii* (Pk) strains to improve malonic acid production
- Iterate on Design-Build-Test-Learn cycles (6 total, >80,000 data points per cycle)

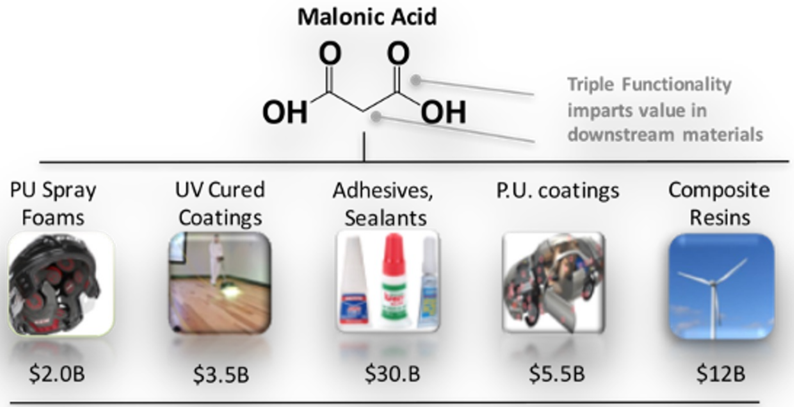


<u>**Tool development and modelling:**</u>

- Expand promoter diversity to enable better tuning of gene expression levels
- Construct a genome-scale metabolic model to assess our carbon capture within the system

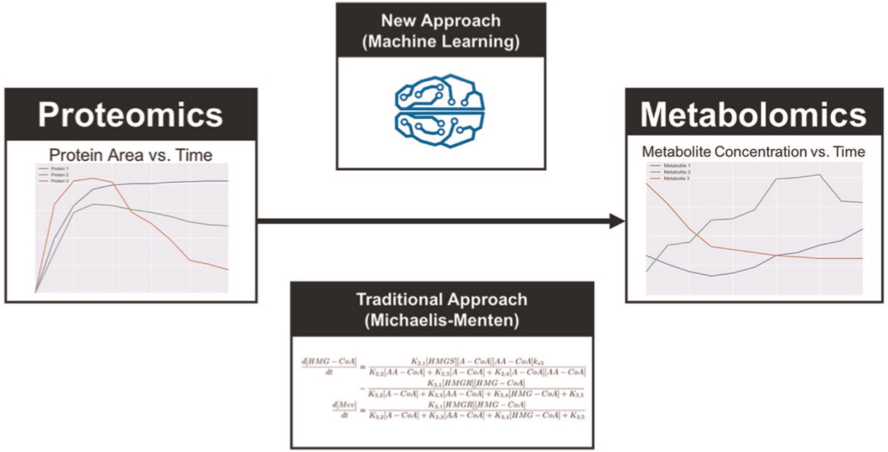
Why malonic acid?

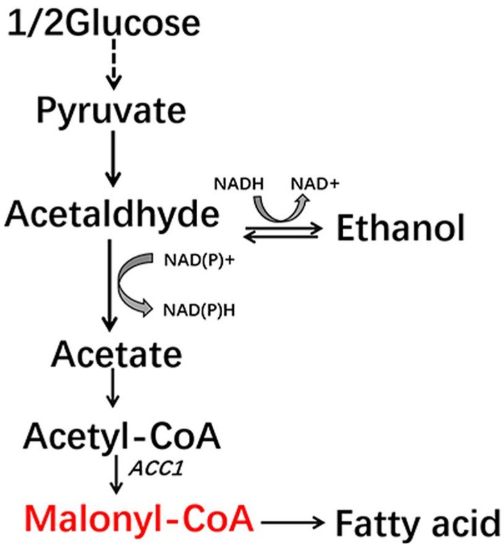
- Over 150-years of use in synthetic chemistry
- Difficult to produce from petrochemistry (<75% yields)
- Production largely driven by foreign suppliers



Why machine learning?

- Traditional, kinetic modelling often depends on having information that is difficult or impossible to get
- Machine learning affords us a way to circumvent this need



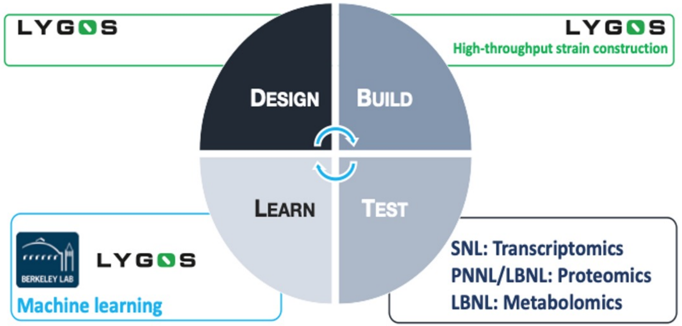


Chen et. al., 2018
Costello and Martin, 2018

Project Structure

Lygos *(Design, Build, Test)*

- Domain knowledge in the host strain
- Genetic engineering tools in hand
- Robust fermentation capabilities (Ambr250)



ABF *(Test and Learn)*

- Omics pipelines for high throughput analysis of multi-omics data
- Machine learning pipelines and super-computing resources

Key risks and mitigation strategies

Risk: Complex, interdependent workflows

- Strain build, AMBR fermentation, and sample collection/extraction at Lygos
- Sample processing, omics data analysis, and ML at ABF

Mitigations

- Good communication between Lygos and ABF

Key risks and mitigation (new)

Risk: Recommendation downselection and build

- Thousands of recommendations are generated, not all of which can be built
- While mostly shared high-ranking strategies were selected, some manual curation was needed
- Build difficulties (was the case for many) in our diploid non-model yeast (only 34 could be built of 59 attempted; only one knockout target could be built)
- Strain build may not actually achieve the desired outcome with increased expression of a gene

Mitigations

- Attempted more builds than would be tested
- Proteomics performed on new strains to learn from this cycle

Key risks and mitigation (new)

Risk: Operations and Staffing

- Operational and staffing issues at both Lygos and ABF caused significant project delays
- Resources became available at Lygos to continue project work in Q4 2022

Mitigations

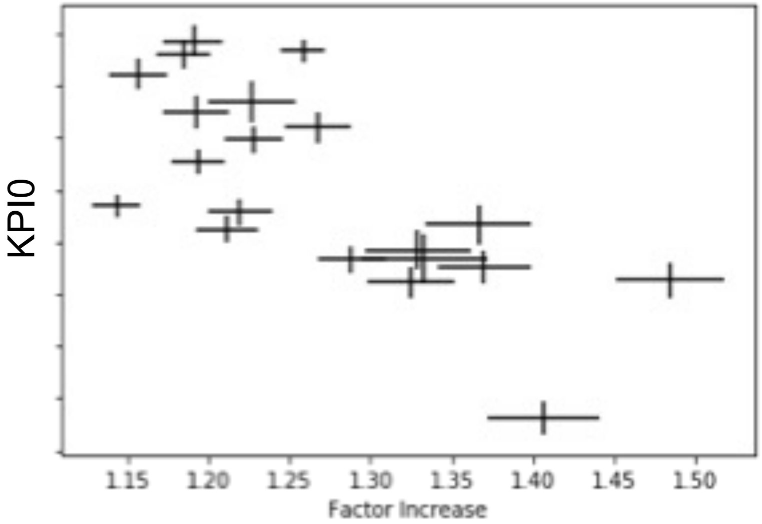
- Jointly decided with DOE and ABF to conclude the program after one completed DBTL cycle

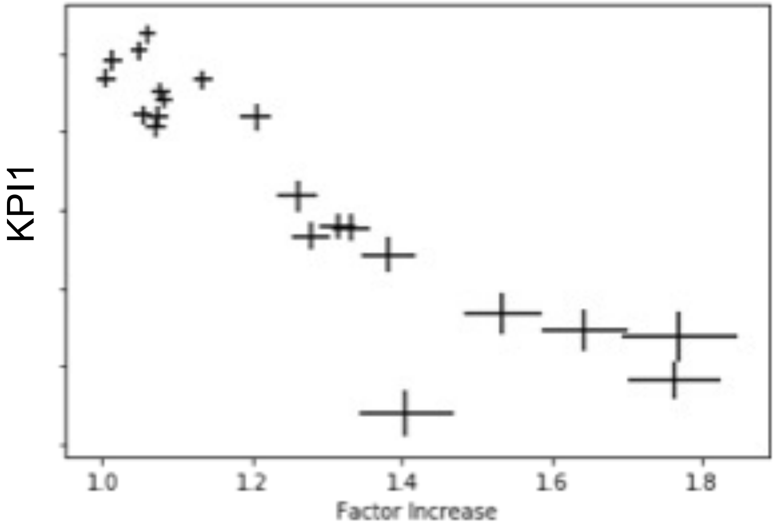
Approach

- Lygos ran a set of 24 strains in fermentation to provide a training set for machine learning
- A set of proposed, genetic modifications that are predicted to satisfy the KPI of interest were recommended by ABF
- The recommendations were based on the trends observed in metabolomics, proteomics, TCA cycle metabolites and malonic acid production values
- Strains were ranked by an ‘Improvement Factor’ which is the predicted increase in the KPI that can be expected from the changes requested
- From the training set recommendations, Lygos successfully built 34 strains, of which 22 strains were selected for a set of 24 fermentation runs (with 2 controls)
- Two key performance metrics (KPI1 and KPI2) were considered as the main criteria to evaluate strain improvement, compared to a control strain

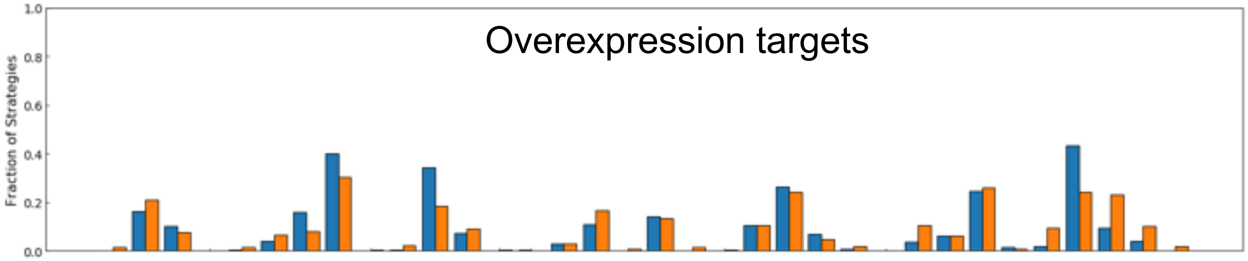
Approach (con’t)

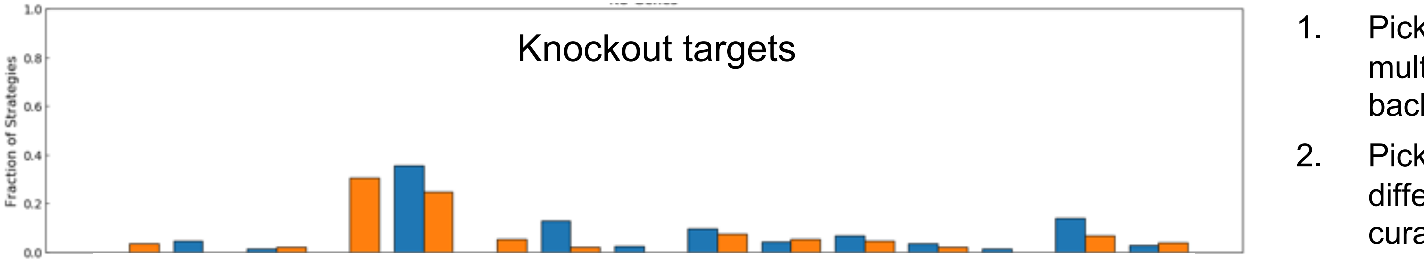
- It was decided to switch from an original KPI0 to KPI1 based on statistical analysis of the variance of predicted metrics





Approach (con't)



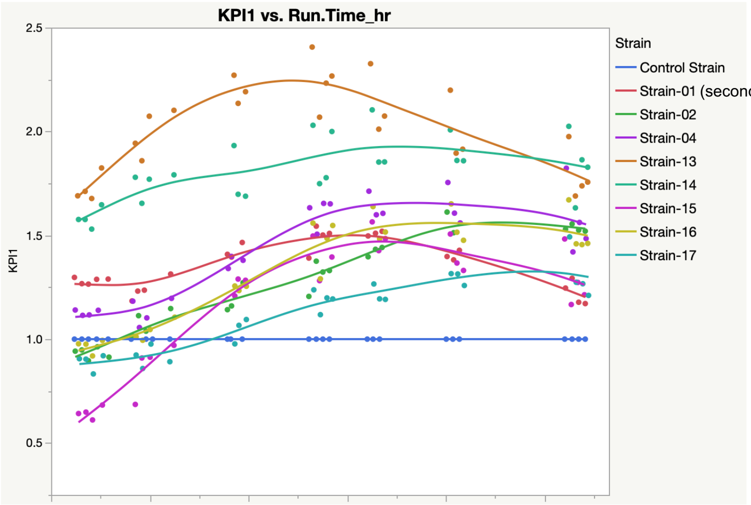


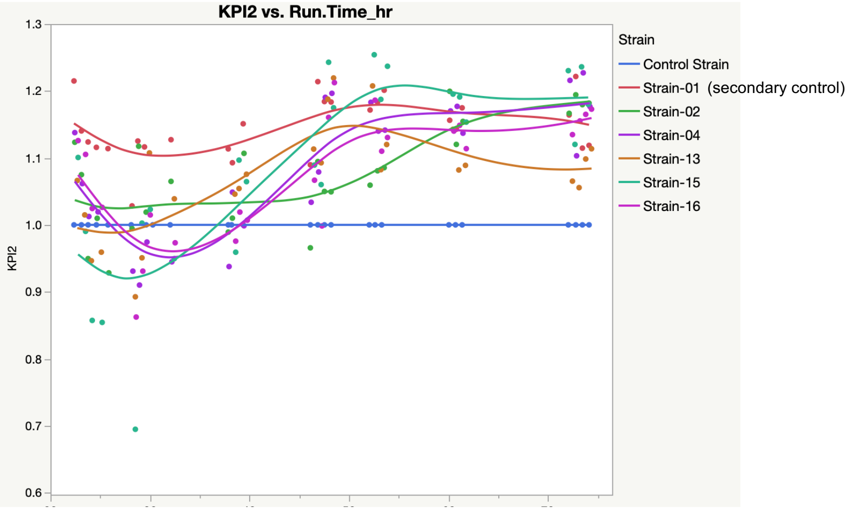
From the highest ranked recommendations for KPI1:

1. Picked 3-5 "shared" strategies in multiple selected strain backgrounds
2. Picked 1-2 "wildcard" strategies in different backgrounds (manually curated unusual combinations; note none were successfully built)
3. Picked non-shared strategies in different backgrounds (note only one strain successfully built)

Progress and Outcomes

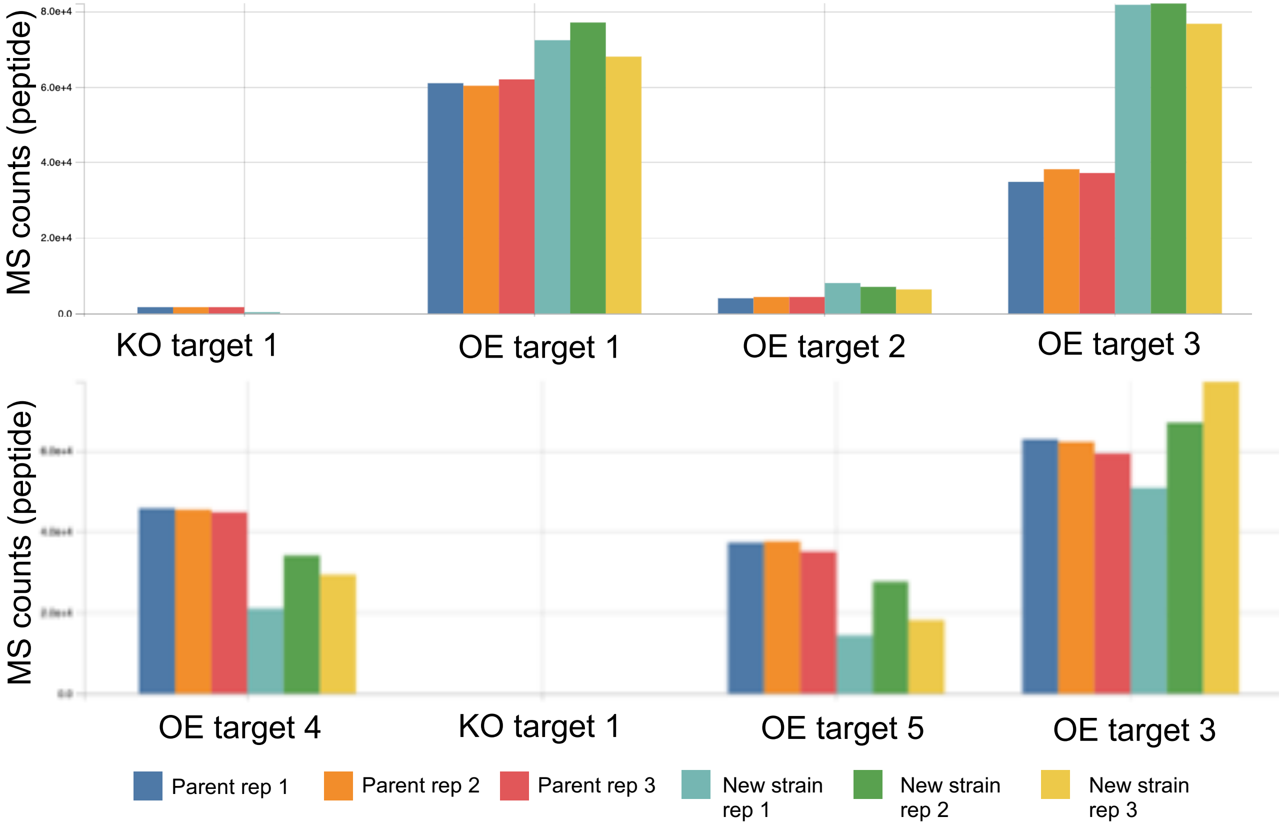
- 7 strains showed improvement in KPI1 and 5 of those strains also showed improvement in KPI2, compared to the original best performer ("Control strain")
- 2 of 3 strains made with Strain-01 parent (secondary control) showed KPI1 improvement vs. parent but not KPI2
- Original project metrics were productivity and KPI2, neither of which were achieved in this cycle (~60% and ~64% of targets, respectively), however more cycles were originally planned





Progress and Outcomes

- Early proteomics analysis indicates some strains have protein expression changes as desired, while others do not
- This information will help inform any future ML efforts and understanding the underlying genotype to phenotype relationships



Impact

- Complexity of the dataset is increased significantly in this grant
  - Process data - CO_2, feed rate, pH, base additions, etc.
  - Intra- and extracellular metabolomics
  - Targeted proteomics
  - OD, DCW
- Equates to more than 80,000 data points per DBTL cycle
- To our knowledge, this is the largest dataset (containing real data) per cycle that has ever been employed for this sort of machine learning and strain improvement in the public domain

Impact (con’t)

- The ABF will also demonstrate the ability to generate this type of dataset, which is expected to generate significant and future investment
- A high-impact publication will be generated as part of this grant (Milestone 5.2)
- New strains from cycle #1 are Lygos’ highest performing malonate producers to date and may positively impact process commercialization

Summary

- We have shown that we can generate reliable workflows and data collection schemes that generate, large, multi-omic, datasets to inform the learning of complex neural networks

- These networks can generate actionable recommendations and have been successful in improving malonic acid production in our engineered strains

- We are excited by the results of the ML in this DBTL cycle and look forward to further collaboration with the DOE, ABF, and their partners

Quad chart

Timeline

- 12/31/2021-12/31/2022

	FY23 Costed	Total Award
DOE Funding	$00,000	$2,000,000
Project Cost Share		

Project Goal

Accelerating engineered microbe optimization through machine learning and multi-omics datasets

Funding Mechanism

- BioEnergy Engineering for Products Synthesis (BEEPS)
- DE-FOA-0001916

Project Partners

- Sandia National Lab
- Lawrence Berkeley National Lab
- Pacific Northwest National Lab

Additional Slides

Milestone 2 - Omics pipeline dev

Milestone	Description	Glucose Type	Month	Date
2.1	Completion of *P. kudriavzevii* global proteomics analysis	N/A	9	Sept. 30, 2019
2.2	Completion of *P. kudriavzevii* targeted proteomics analysis (50 proteins)	N/A	9	Sept. 30, 2019
2.3	Completion of *P. kudriavzevii* targeted metabolomics analysis (50 metabolites)	N/A	9	Sept. 30, 2019
2.4	Completion of *P. kudriavzevii* targeted transcriptomics analysis (50 genes)	N/A	9	Sept. 30, 2019

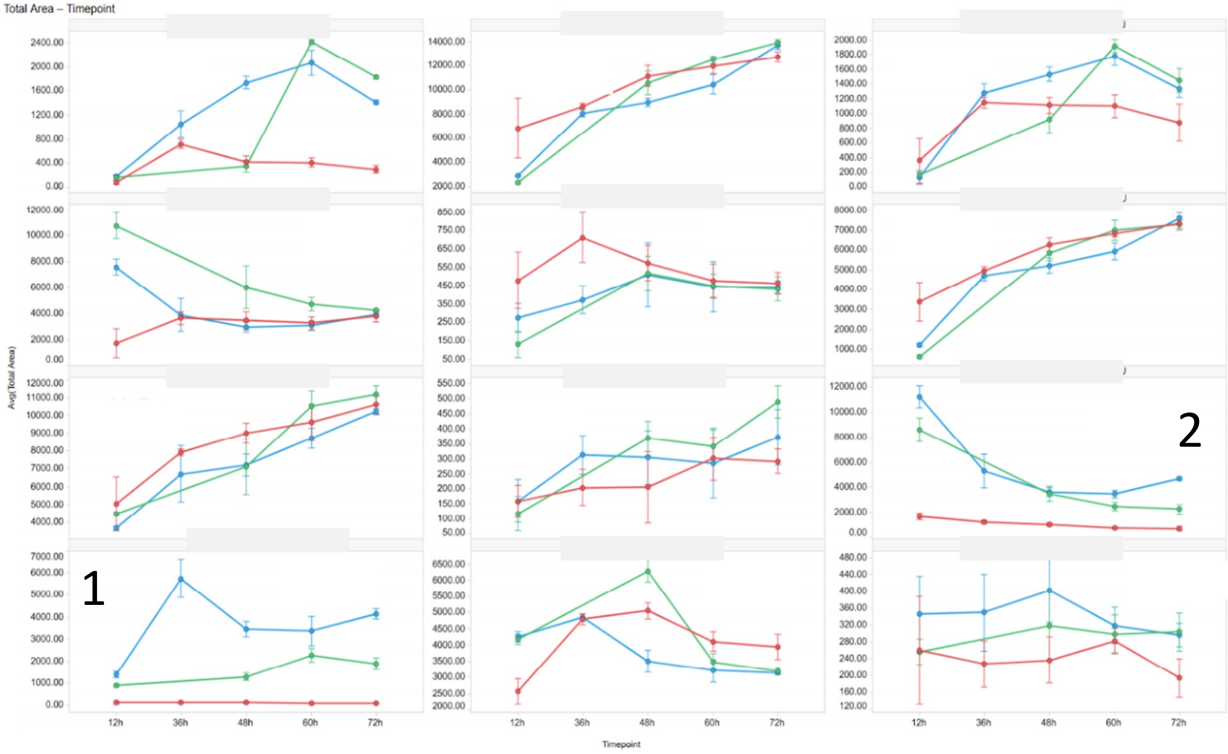
Status – Complete

Targeted Proteomics output

- Changes in expression over time yield valuable insight into strain performance

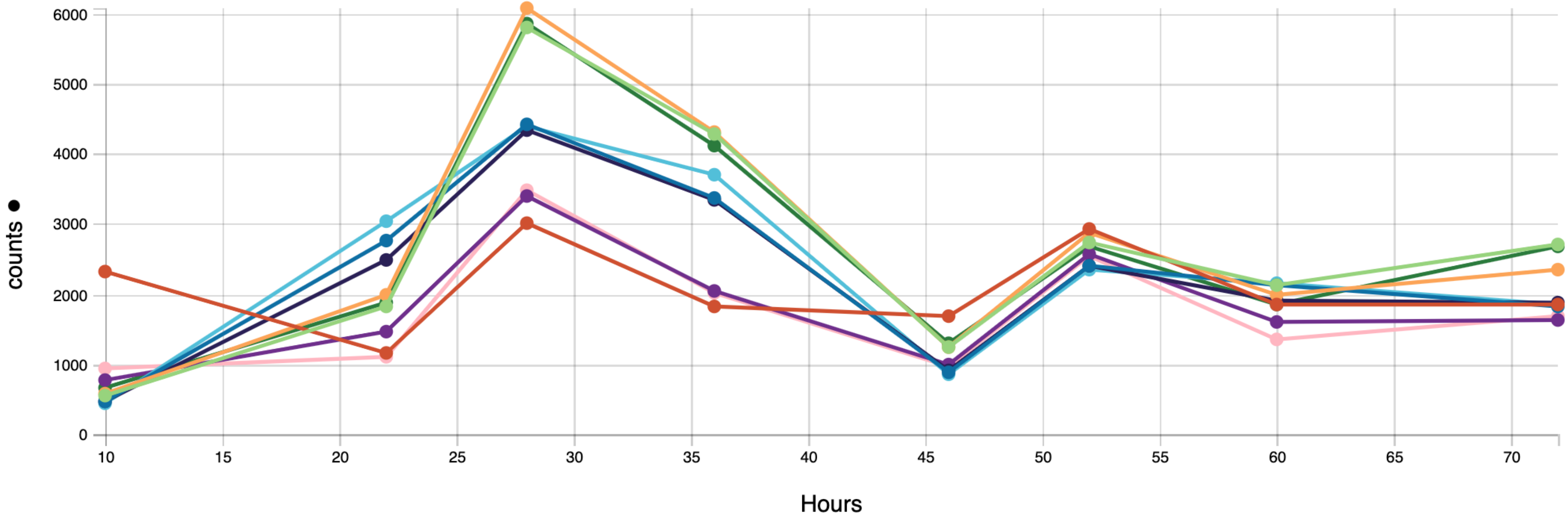
1. A protein lacking from the ‘red’ strain, introduced in ‘green’, and modified in ‘blue’

2. An important protein showing significant decline throughout the fermentation.



Targeted metabolomics output - Experimental Data Depot

- TCA metabolite shown for 3 different strains
- Allows researchers to visualize the impact of strain engineering on carbon flux



Milestone 3 - Promoter diversity

Milestone	Description	Glucose Type	Month	Date
3.1	Identify at least 15 native *P. kudriavzevii* promoters that demonstrate RFP expression between the 50 – 3,000 RFU/OD range	N/A	12	Dec. 31, 2019
3.2	Generate and characterize at least 1,000 mutant *P. kudriavzevii* promoters	N/A	27	March 31, 2021
3.3	Generate a *P. kudriavzevii* promoter library that exhibit 10,000-fold dynamic range in RFP expression levels	N/A	27	March 31, 2021

Purpose - to generate new promoter variants that allow for more range in gene expression.

Status – Complete/On-time

- Cap Analysis of Gene Expression (CAGE) was completed to map the transcriptional start sites for the Pk promoters. Subset of native, Pk promoters identified (52).
- Error-prone PCR used to generate diversity.

Milestone 4 - Metabolic model

Milestone	Description	Month	Date
4.1	Completion of initial *P. kudriavzevii* metabolic network (capturing >80% of carbon flux)	15	Sept. 30, 2020
4.2	Completion of final *P. kudriavzevii* metabolic network (capturing >90% of carbon flux)	33	March 31, 2021

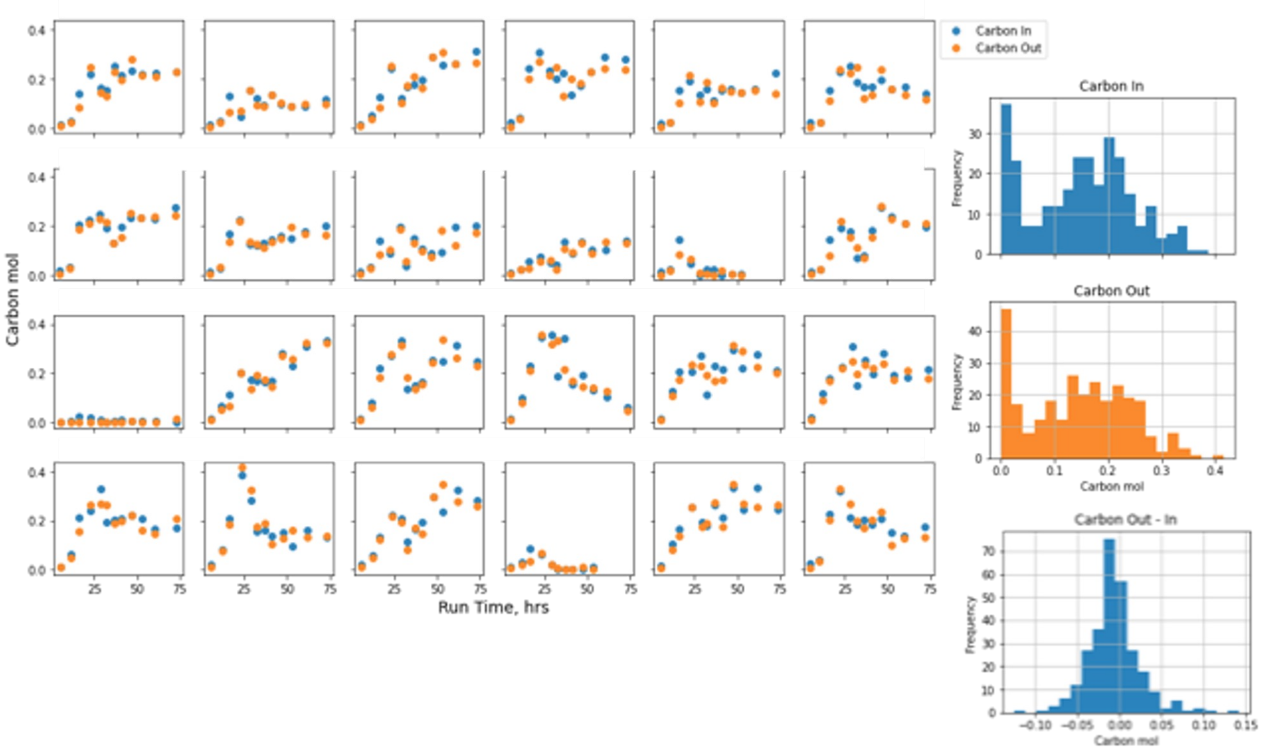
Purpose - to complement our omics and machine learning with an accurate genome-scale metabolic model.

Work led by Joonhoon Kim

Status – Complete/On-time

Milestone 4 - Metabolic model

- Carbon in and out (y-axis) over time (x-axis)
- Illustrates the dynamics of the carbon flux and highlights abnormalities where carbon capture is reduced.



The training set

Milestone	Description	Glucose Type	Month	Date
5.1	Complete 24-member machine learning training set	Crystalline	12	Dec. 31, 2019

Status – Complete/Delayed

- First, full DBTL cycle!

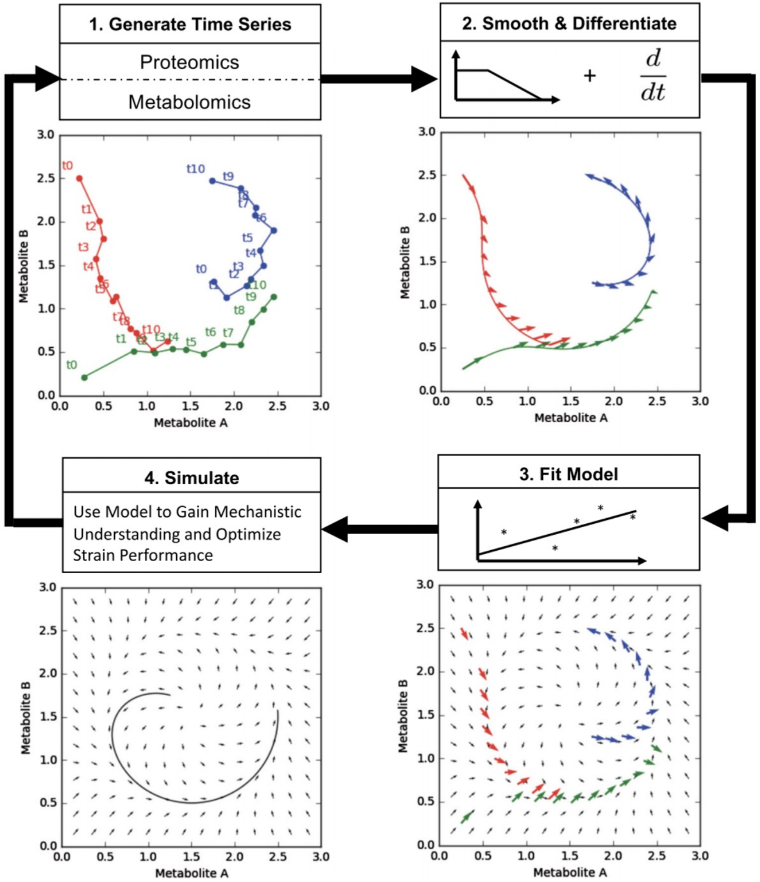
- Delayed due to several, compounding factors:

1. A fire in the lab at Emeryville Station East
2. Loss of lead data scientist to another position (6mth NCTE awarded)
3. COVID (6mth NCTE awarded)

Glimpse into machine learning

- Populate machine learning networks with multi-omics data collected from strains with different performance.

- These networks can then optimize strain performance by understanding how different concentrations of metabolites and proteins correlates with differences in production of malonate.

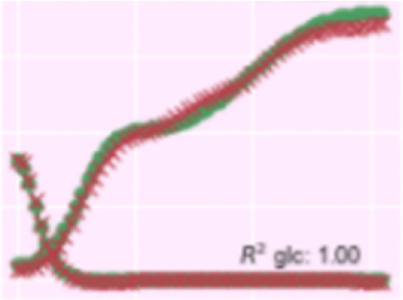


Prediction validation

- We can assess how good the neural network is at predicting performance by plotting the predicted vs. actual concentrations of a given factor (metabolite or protein level).

- A good fit indicates that the model is capturing the dynamics of the system well.

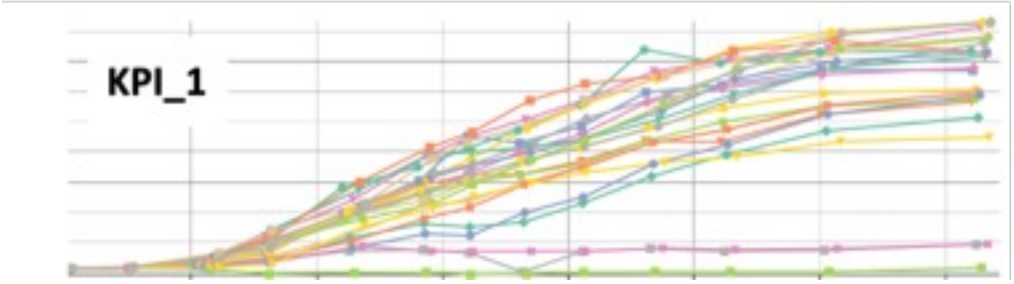
- Green = Experimental data for 2 metabolites
- Red = Model forecast

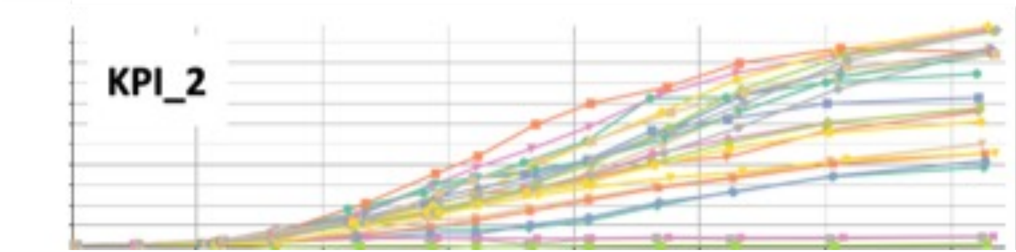


Focusing on malonate production

- Two different, malonate-related key performance indices (KPIs) were used to assess strain performance

- These same metrics are the key data used to generate recommendations.





Risks in our approach

1. **Data quality**
   - We must be able to generate high quality data with relatively low variation to ensure we can have confidence in the recommendations we are making.

Example: extraction of CoA species

- High priority, but very labile.
- Significant time on method development to optimize their extraction while also extracting others at high efficiency.
- Cast the net wide (cover the metabolomics space adequately to ensure confidence).
- 72 metabolites tracked, from 24 strains, at 8 timepoints per cycle.

Risks in our approach

2. **Interconnected workflows**
    - Each part of the DBTL cycle depends on the others.

Example: Sample prep at Lygos/LBNL

- During the training set, we needed to be able to rapidly prep ~600 samples for proteomics and another ~600 for metabolomics before collecting data.
- Significant time spent to optimize and operationalize workflows to ensure consistency and rapid turnaround.
- A lot of dry runs, trial and error, communication, and feedback.

Publications / Presentations:

- Accelerating engineered microbe optimization through machine learning and multi-omics dataset. Poster presentation at 2019 BETO Peer Review Meeting, March 4-7, 2019, Denver, CO.

- Accelerating engineered microbe optimization through machine learning and multi-omics dataset. Virtual oral presentation at the 2021 BETO Peer review, March 11th, 2021.