LYGOS

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

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

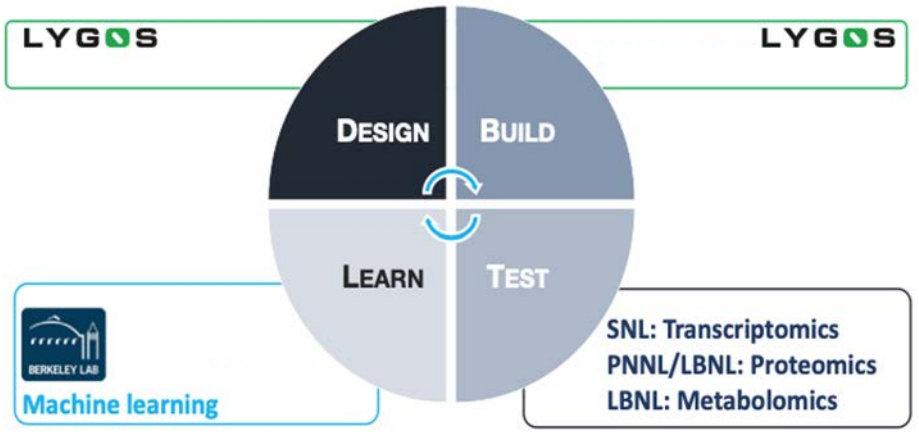
2021-03-11

Presenter: Mark Held
Associate Director of Strain Operations and Systems Biology
Lygos, Inc.

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)

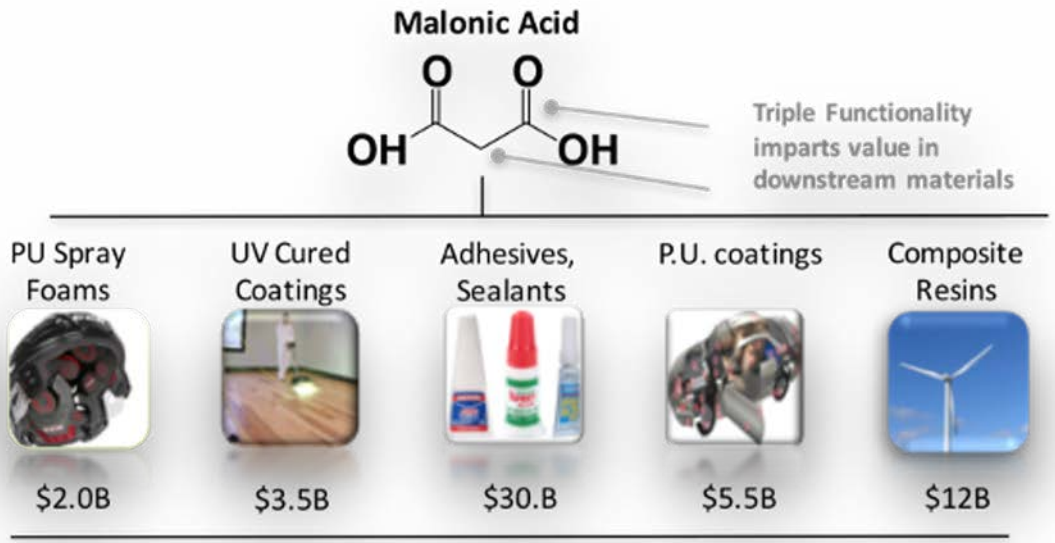


Tool development and modelling:

- 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

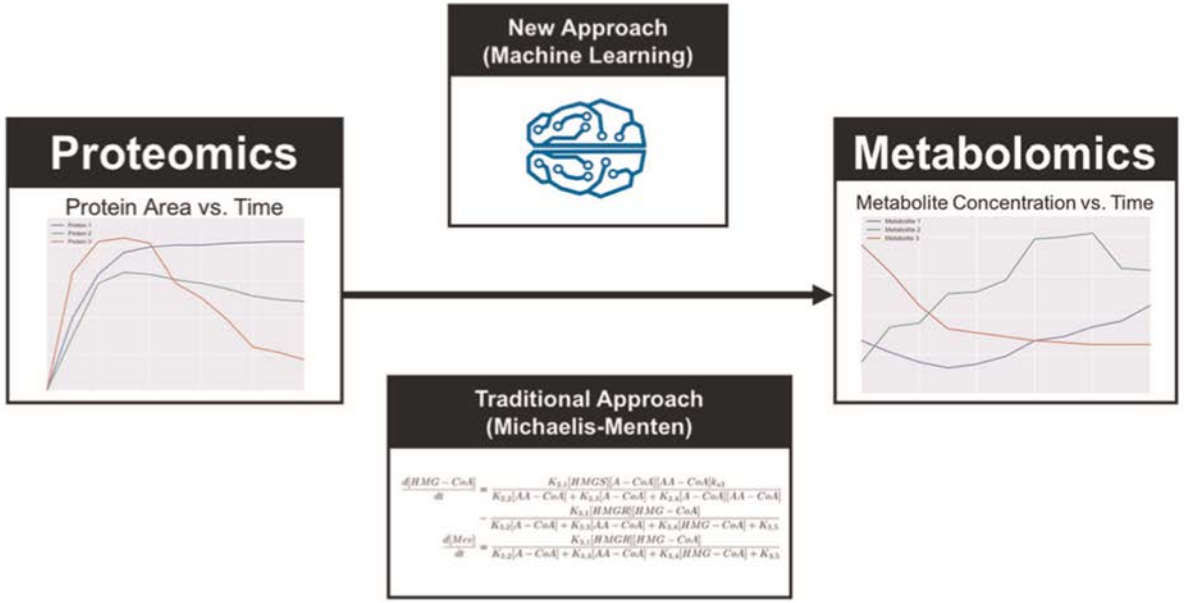
Why malonic acid?

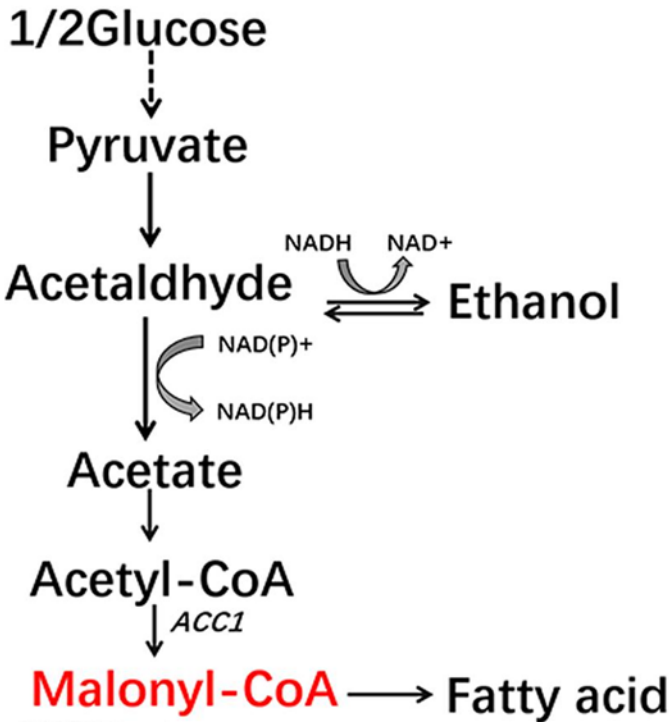
- Over 150-years of use in synthetic chemistry
- Difficult to produce from petrochemistry (<75% yields)
- Production largely driven by foreign suppliers



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

Complimentary skill sets

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.

Confidential Information of Lygos, Inc.

1 - Management

- Several, overlapping lines of communication are used to ensure alignment on progress and challenges.
  - Monthly meetings between the ABF and Lygos
  - Monthly meetings between the DOE and Lygos
  - Off-cycle meetings to coordinate sample drop-off, etc.
  - Quarterly progress and financial reports

Key stakeholders for this project:

Chris Petzold (LBNL) - targeted proteomics and metabolomics

Hector Garcia Martin (LBNL) - Machine learning

Kristin Burnum-Johnson/Jon Magnuson (PNNL) - Global proteomics

Jon Gladden (SNL) - Targeted transcriptomics

Key risks and mitigation strategies

1. Complex, interdependent workflows

Solutions:

- Communication, communication, communication
- Gantt charts
- Numerous one-off meetings and phone calls

2 - Approach

- Iteration on this multi-omics/machine learning approach is key to success.
- Go/no-go decisions provide logical, step-wise expectations throughout this process.

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.

3 - Impact

- Complexity of the dataset is increased significantly in this grant
  - Process data - CO2, 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) that has ever been employed for this sort of machine learning and strain improvement.

Impact (con’t)

- The production goals outlined in this grant would greatly advance Lygos’ ability to further commercialize our malonic acid platform.

- 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).

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

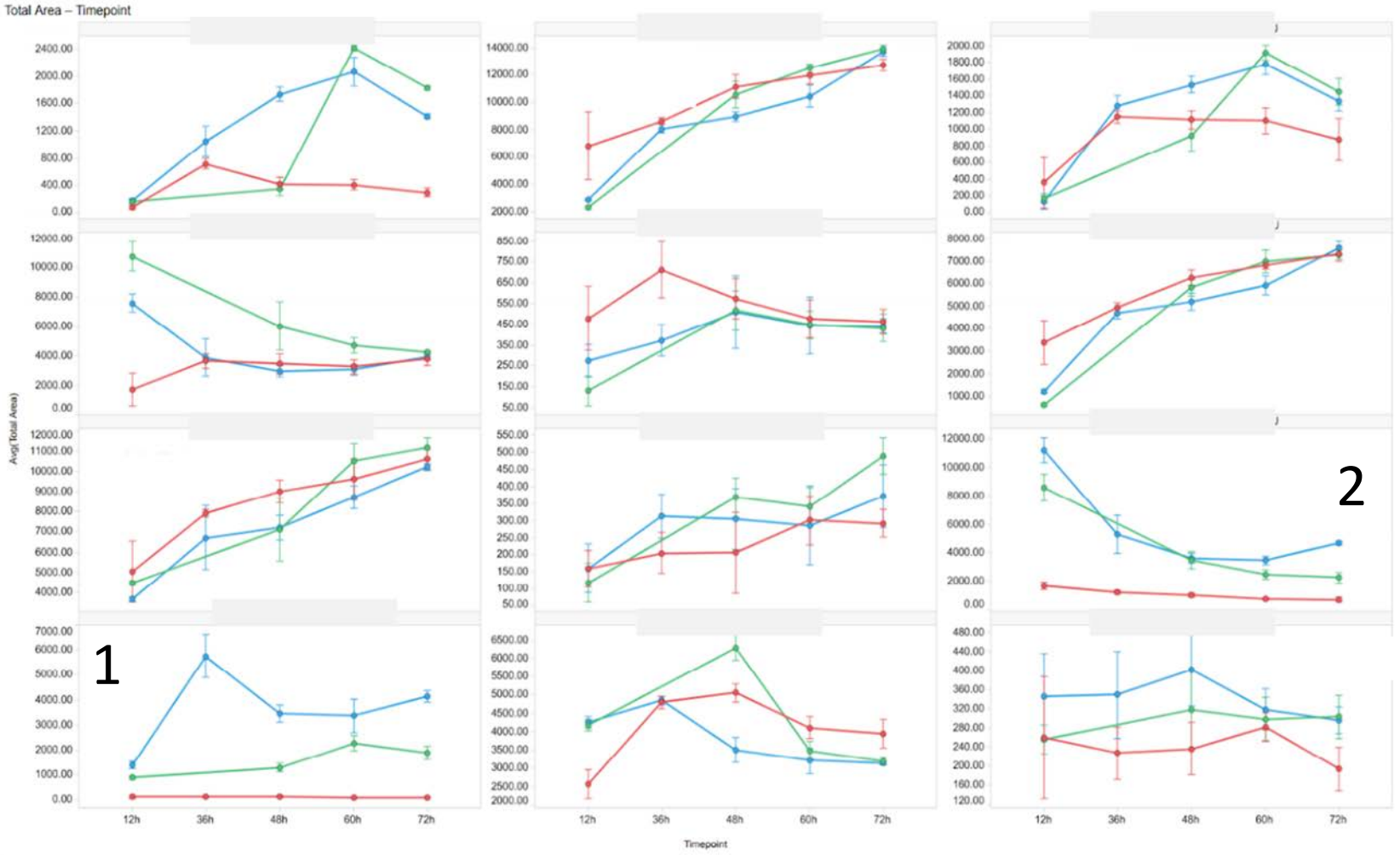
Status – Complete

Targeted Proteomics output

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

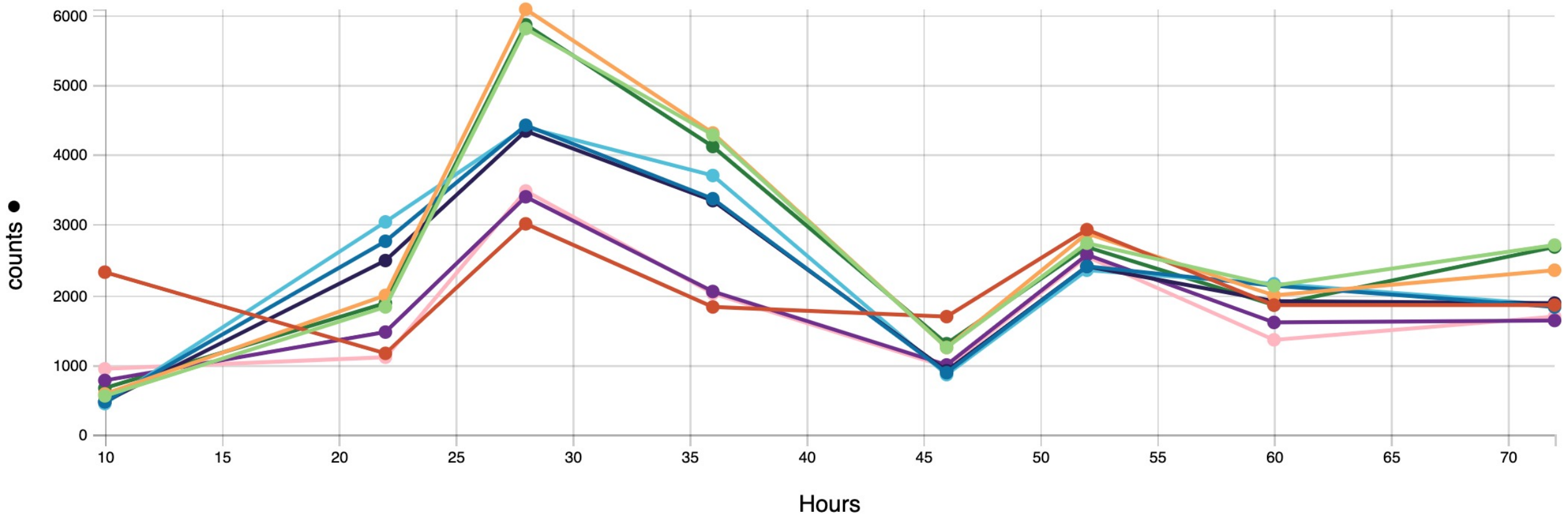
1. A protein lacking from the ‘red’ strain, introduced in ‘green’, and modified in ‘blue’

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

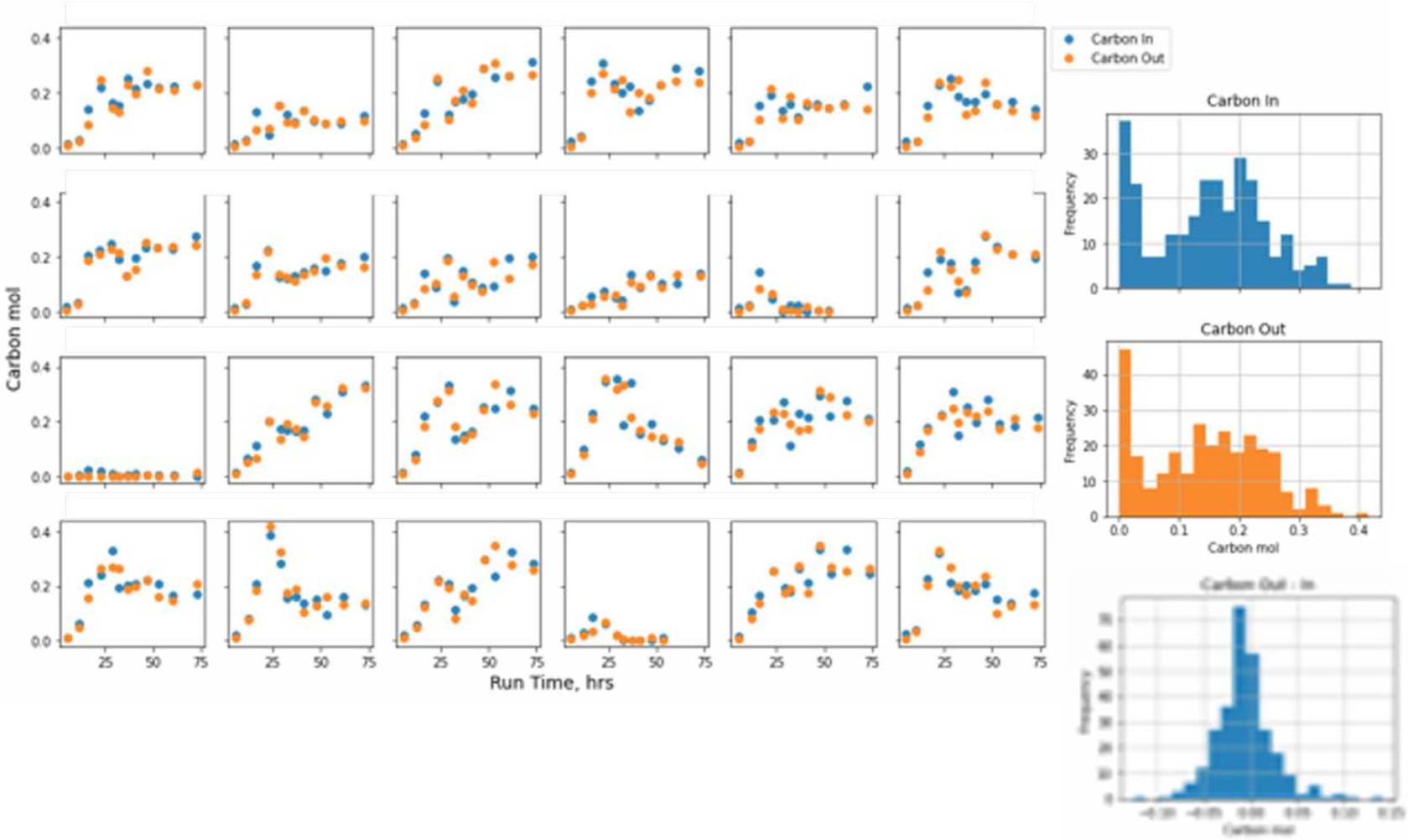
Purpose - to complement our omics and machine learning with an accurate genome-scale metabolic model.

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

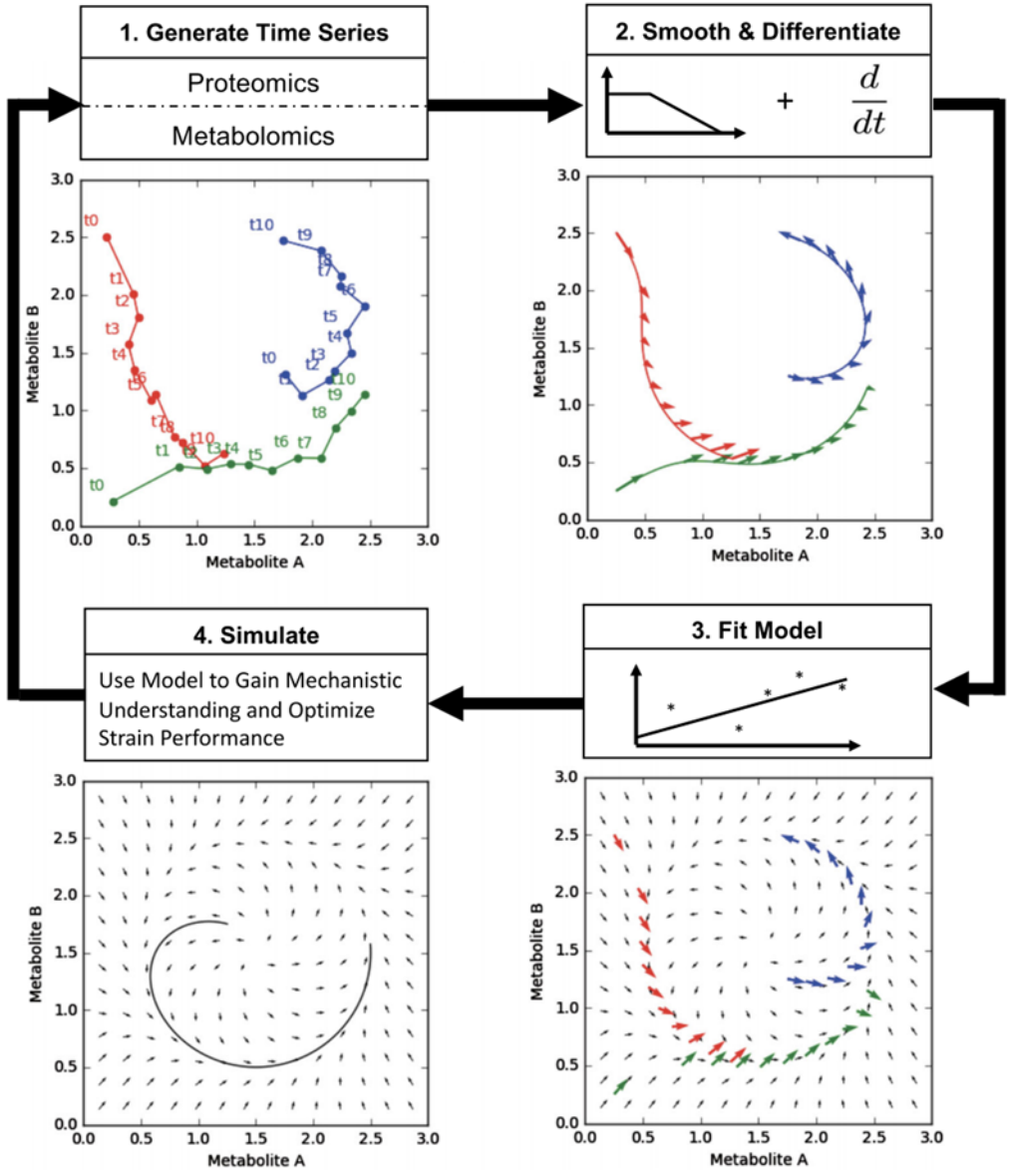
- First, full DBTL cycle!

- 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.

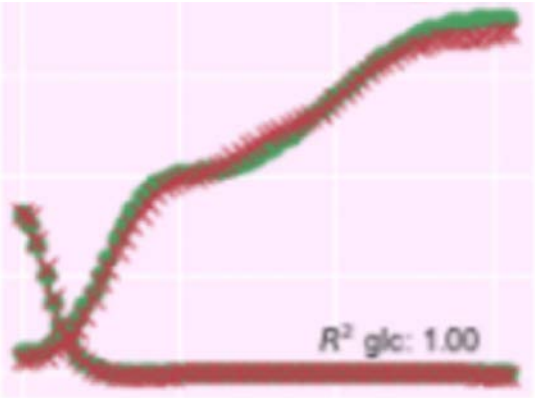


Prediction validation

- 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).

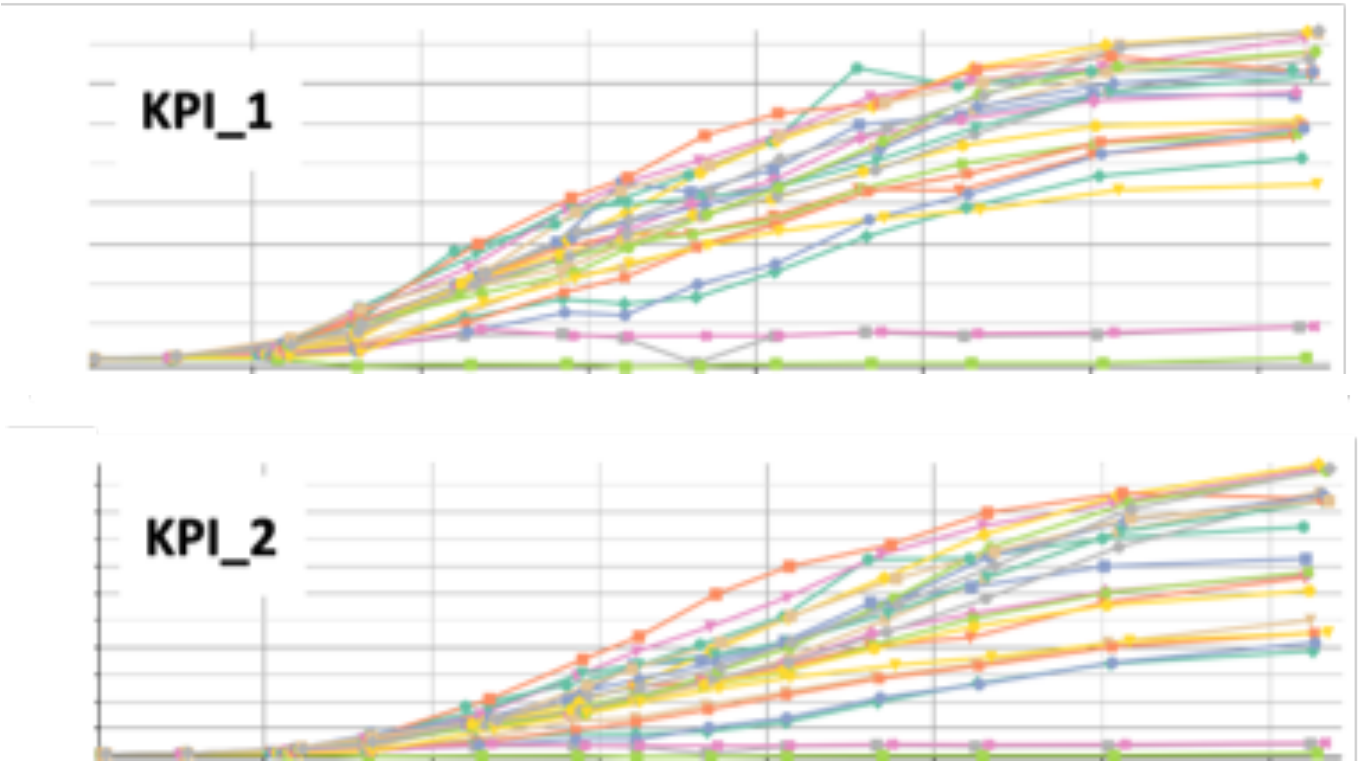
- A good fit indicates that the model is capturing the dynamics of the system well.

- 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.



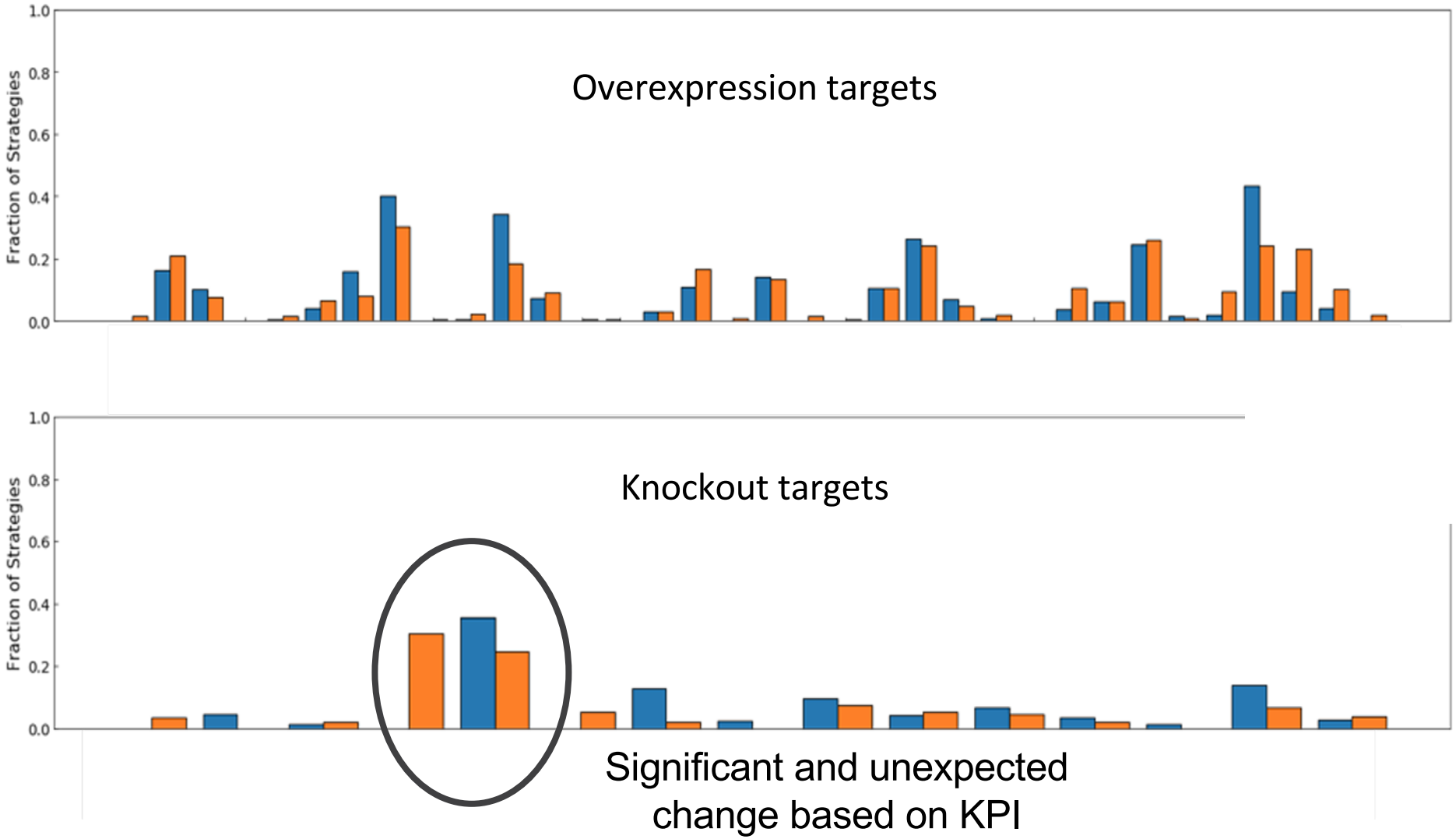
Recommendations

What are they?

- A set of proposed, genetic modifications that are predicted to satisfy the KPI of interest.
- They are ranked by an ‘Improvement Factor’ which is the predicted increase in the KPI that can be expected from the changes requested.

Comparing recommendation sets

- Similarities and differences noted when comparing recommendations.
- Some made rationale sense, others were very unexpected.



What is left to do

- Iterate, iterate, iterate

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 that are expected to improve malonic acid production in our engineered microbial strains.
- This is a very exciting and challenging project that Lygos is glad to be a part of
- The data from this work will truly change the industry
- We look forward to further collaboration with the DOE, ABF, and their partners.

Quad chart

Timeline

- 6/1/2019
- 12/31/2021

	FY20 Costed	Total Award
DOE Funding	$312,884	$2,000,000
Project Cost Share		$857,143

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