Machine Learning Optimization of Photosynthetic Microbe Cultivation and Recombinant Protein Production

Caitlin Gamble, Drew Bryant, Damian Carrieri, Eli Bixby, Jason Dang, Jacob Marshall, David Doughty, Lucy Colwell, Marc Berndl, James Roberts, Michael Frumkin

Addie Chambers & Erin Wilson

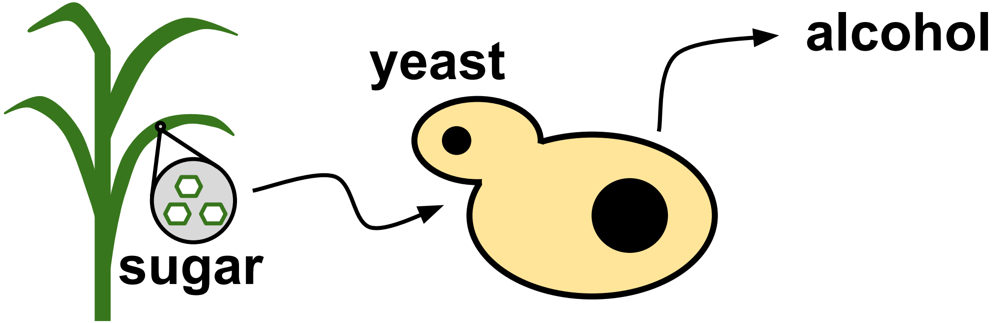
CompBio Seminar

October 25, 2021

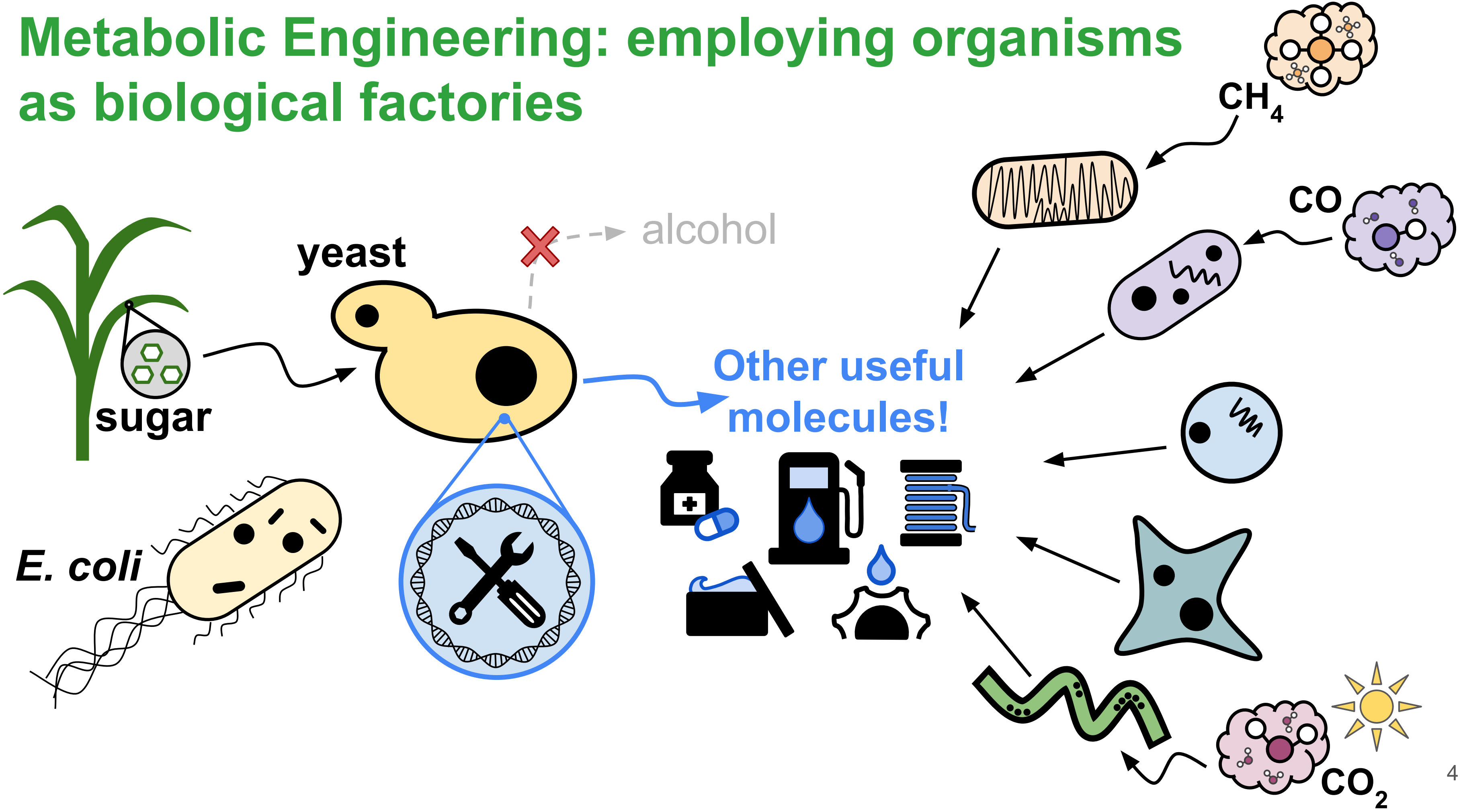
Overview

- Background
  - Metabolic Engineering + Lumen Biosciences
  - Bayesian Optimization (Gaussian Process - BUCP)
- Goals of this paper
  - Experimental set up + measurements
- Results
  - Preliminary optimization outcomes
  - Validation of top configurations
  - Biological interpretation + scale up
- Key takeaways
  - Discussion questions!

Metabolic Engineering: employing organisms as biological factories



Metabolic Engineering: employing organisms as biological factories



yeast

sugar

alcohol

E. coli

Other useful molecules!

CH_4

CO

CO_2

Benefits of working with *Arthrospira platensis* (Spirulina)

- **Cyanobacterium**
  - Photosynthetic metabolism

- **FDA:** "Spirulina is source of protein and contains several vitamins and minerals"

- **GRAS:** Generally regarded as safe

This paper: a partnership between Lumen Bioscience and Google!

LUMEN BIOSCIENCE

+

Google Research

Lumen's biotech platform:

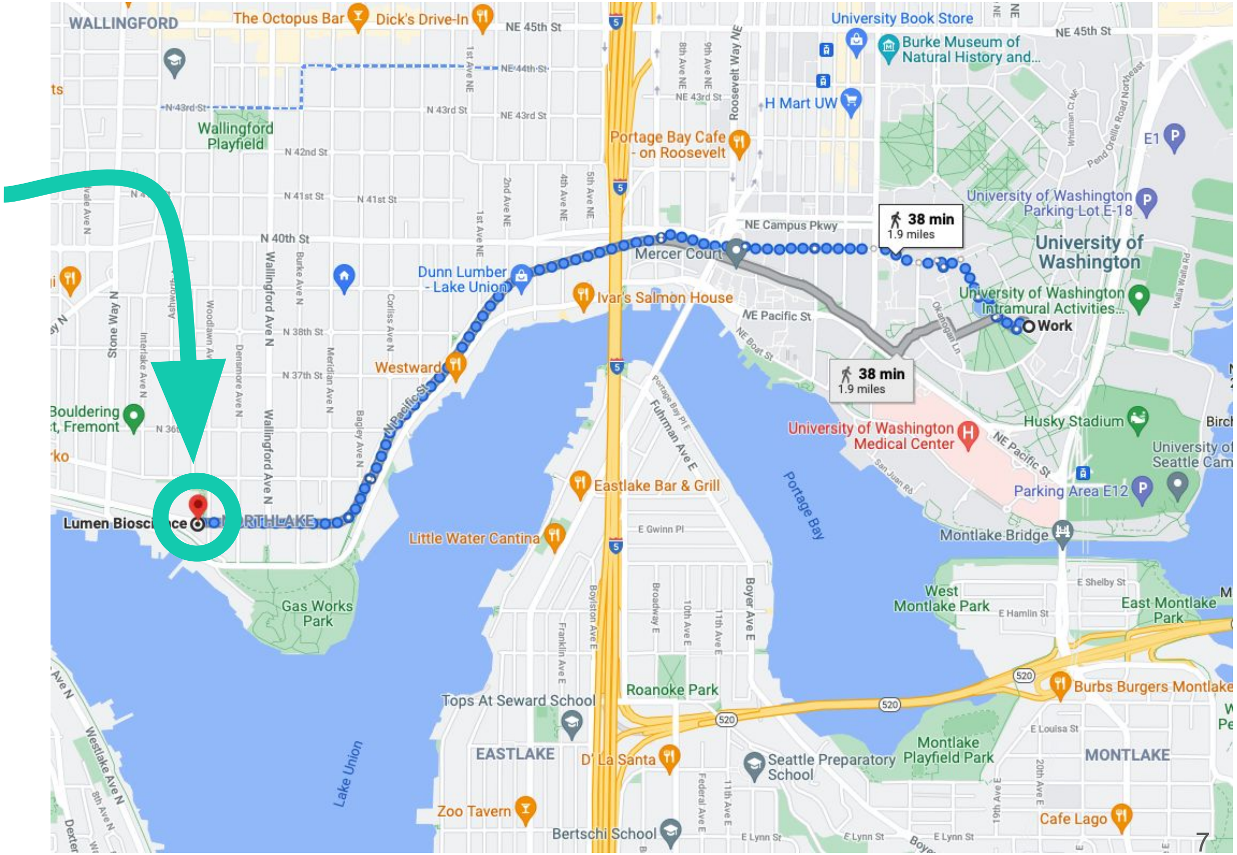
- Manufacture biopharmaceuticals, antibodies, therapeutic proteins
- "Orally delivered biologics"
- Scale up production by engineering Spirulina
    - "Cheap" inputs: water, salt, CO2, light

This paper: a partnership between Lumen Bioscience and Google!

LUMEN BIOSCIENCE

+

Google Research



This paper: a partnership between Lumen Bioscience and Google!

LUMEN

Lumen Bioscience Expands Biologics Manufacturing Capacity with Lease of Historic Seattle Bakery

Research

Metabolic Engineering “performance” is measured in biomass, titer, yield, and productivity

Biomass

Can your organism grow?

Titer

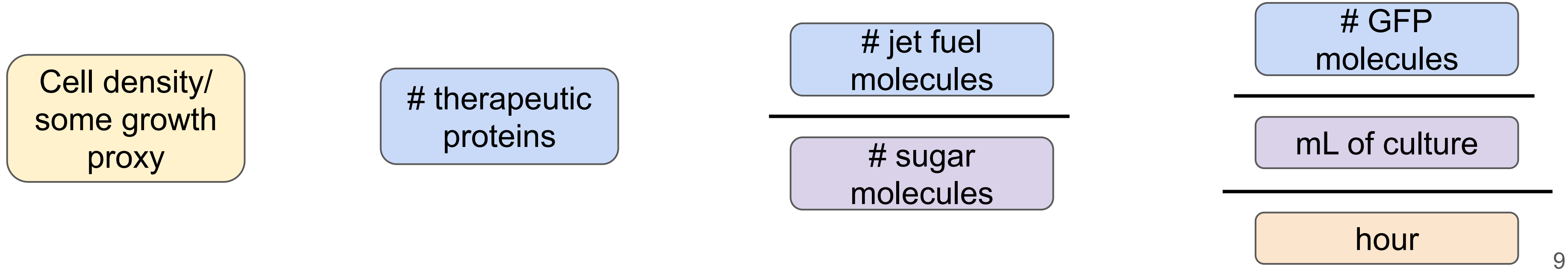
final concentration of product

Yield

units of product synthesized per unit of raw material consumed

Productivity

amount of product formed per unit of time (rate)



Cell density/ some growth proxy

# therapeutic proteins

# jet fuel molecules

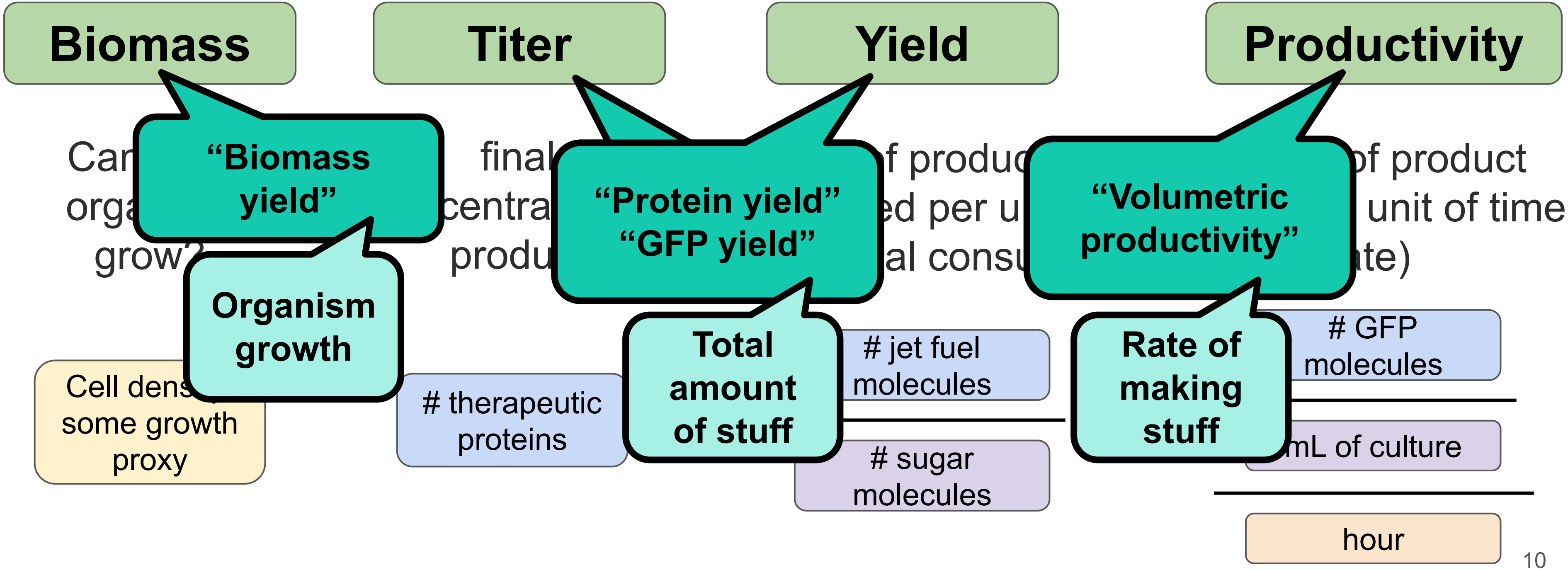
# sugar molecules

# GFP molecules

mL of culture

hour

This paper:



Biomass

Titer

Yield

Productivity

Car
orga
grow?

final
centra
produ

f produc
ed per u
al consu

of product
unit of time
ate)

“Biomass yield”

Organism growth

“Protein yield”
“GFP yield”

Total amount of stuff

“Volumetric productivity”

Rate of making stuff

Cell dens
some growth
proxy

# therapeutic proteins

# jet fuel molecules

# sugar molecules

# GFP molecules

mL of culture

hour

How can scientists improve performance?

Modification of the host organism

- **Overexpression** of key enzymes
- **Deletion** of pathways to “waste products”
- Optimize **codon usage**
- Metabolic **flux balancing**

Modification of the culture conditions

- Feed **rate**, feed **type**
- **Concentrations** of input
- Temp., pH, O_2 flow, etc
- *All the buttons you can press on the bioreactor machine*

Lower the cost of biologic manufacturing

Gaussian Process - Batched Upper Confidence Bound

- **Goal:** find input *x* that maximizes *f(x)* for some unknown function of interest *f*

Gaussian Process - Batched Upper Confidence Bound

- **Goal:** find input *x* that maximizes *f(x)* for some unknown function of interest *f*
- **Given:**
    - Input space D
    - Gaussian process prior: μ_0, σ_0, k
    - Ability to sample $y = f(x) + \epsilon$
        - Oftentimes, assume that these samples are in some way expensive to procure

Gaussian Process - Batched Upper Confidence Bound

- **GP-UCB (no batching) algorithm:**

```
for $t = 1, 2, \ldots$
    $x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \sqrt{\beta_t}\sigma_{t-1}(x)$
    $y_t \sim f(x_t) + \epsilon$
    Bayesian update to obtain $\mu_t, \sigma_t$
```

Gaussian Process - Batched Upper Confidence Bound

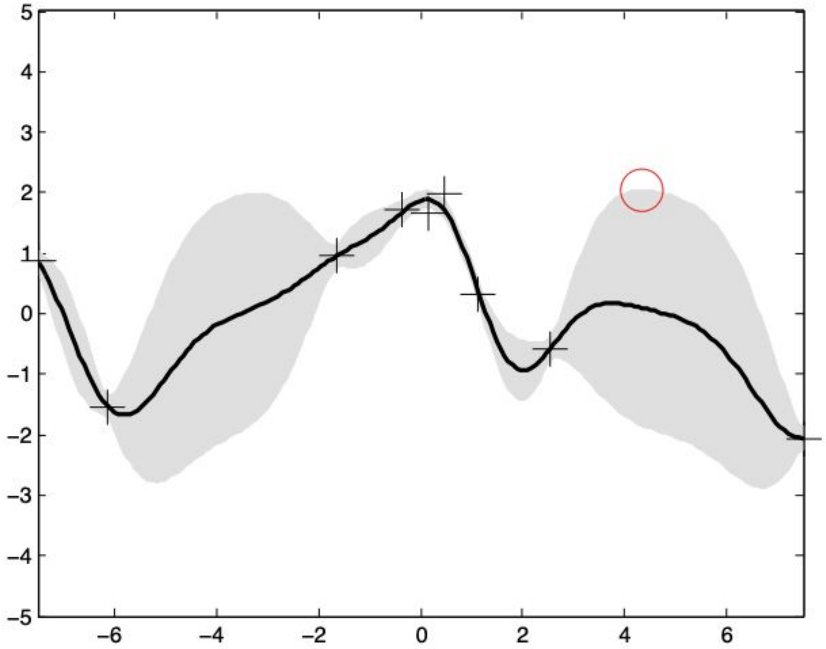
- **GP-UCB (no batching) algorithm:**

```
for $t = 1, 2, \ldots$
    $x_t = \arg\max_{x \in D} \mu_{t-1}(x) + \boxed{\sqrt{\beta_t}}\sigma_{t-1}(x)$
    $y_t \sim f(x_t) + \epsilon$
    Bayesian update to obtain $\mu_t, \sigma_t$
```

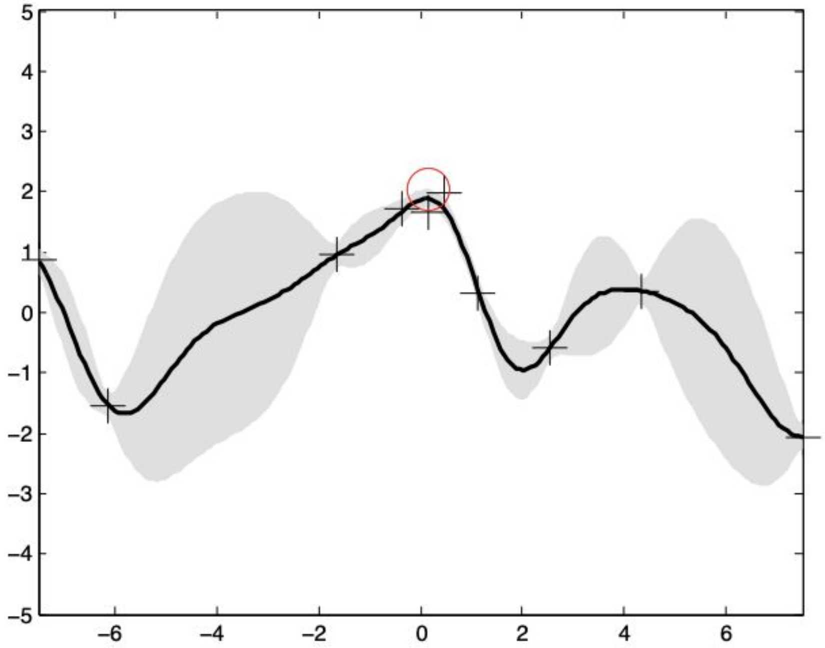
Tradeoff between exploration and exploitation in reward function with confidence level:

- Smaller β -> biased towards *x* where $\mu_{t-1}(x)$ is large (so *f(x)* is thought to be large)
- Larger β -> biased towards *x* where $\sigma_{t-1}(x)$ is large (so *f(x)* is uncertain)

Gaussian Process - Batched Upper Confidence Bound



(b) *Iteration t*



(c) *Iteration* $t+1$

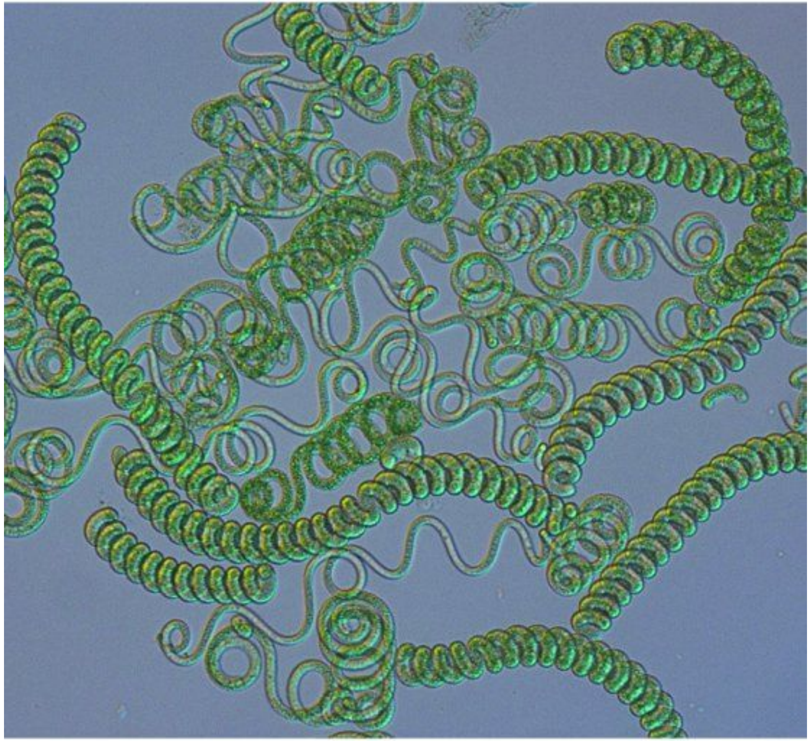
From Srinivas et. al., “Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design.”

Gaussian Process - Batched Upper Confidence Bound

- Don’t want to be limited to sampling one *x* at a time -> batching
    - Simulate posterior given previous *x* in batch -> pessimistic assumption of outcome
    - Re-apply selection policy on posterior
    - Repeat until batch size reached
    - Used Google Vizier with relatively limited available batch sizes

Overview

- Background
  - Metabolic Engineering + Lumen Biosciences
  - Bayesian Optimization (Gaussian Process - BUCP)
- **Goals of this paper**
  - **Experimental set up + measurements**
- Results
  - Preliminary optimization outcomes
  - Validation of top configurations
  - Biological interpretation + scale up
- Key takeaways
  - Discussion questions!



Goal of this paper

Optimize **culture conditions** for the spirulina-based production of therapeutic proteins.

Goal of this paper

Optimize **culture conditions** for the spirulina-based production of ~~therapeutic proteins.~~ **GFP.**

- Environmental “hyperparameters”
  - Intensity, color, cycle of light
  - pH
  - Temperature
  - Etc.
- Reward
  - Volumetric productivity -> measured by GFP fluorescence
  - C = Labor cost (empirically set to 200)
  - Reward function: $R(g) = \max_t g(t) = \max_t \frac{F(t) - F(0)}{t + C}$

Reward Function

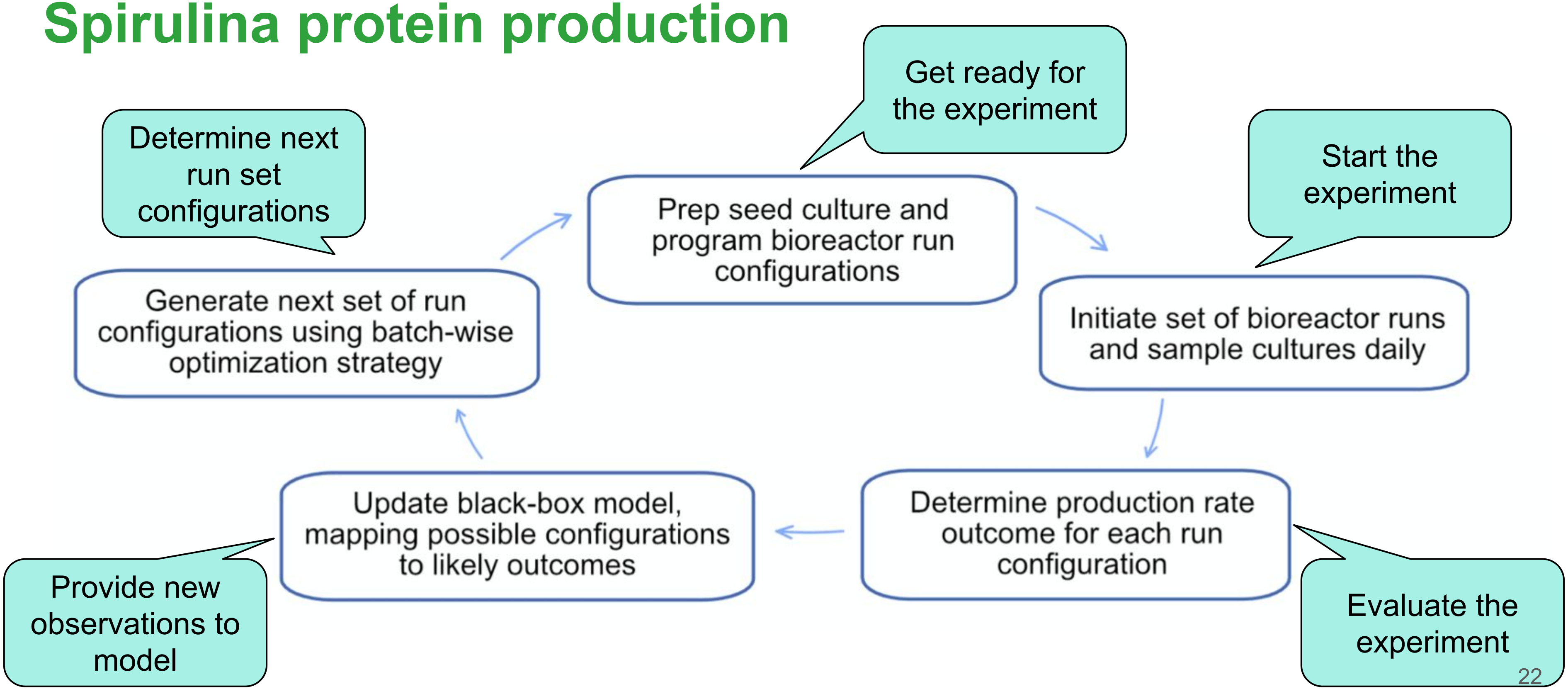
- “Run set / Batch”: multiple bioreactors seeded with common starting culture
- “Standard conditions”: common spirulina culture conditions
    - e.g., pH in [9.75, 9.95]
- Inter- and intra-batch variance estimated using control condition replicates at standard conditions
- Reward: Adjust for batch effect and normalize by standard conditions to get:

$$R(g) = \max_t g(t) = \max_t \frac{F(t) - F(0)}{t + C}$$

$$R'(g) = \frac{R(g) - \mu_{batch}^{(std)} + \mu_{global}^{(std)}}{\mu_{global}^{(std)}}$$

“Performance”

Gaussian Process Algorithm as implemented for Spirulina protein production



Prep seed culture and program bioreactor run configurations

Get ready for the experiment

Initiate set of bioreactor runs and sample cultures daily

Start the experiment

Determine production rate outcome for each run configuration

Evaluate the experiment

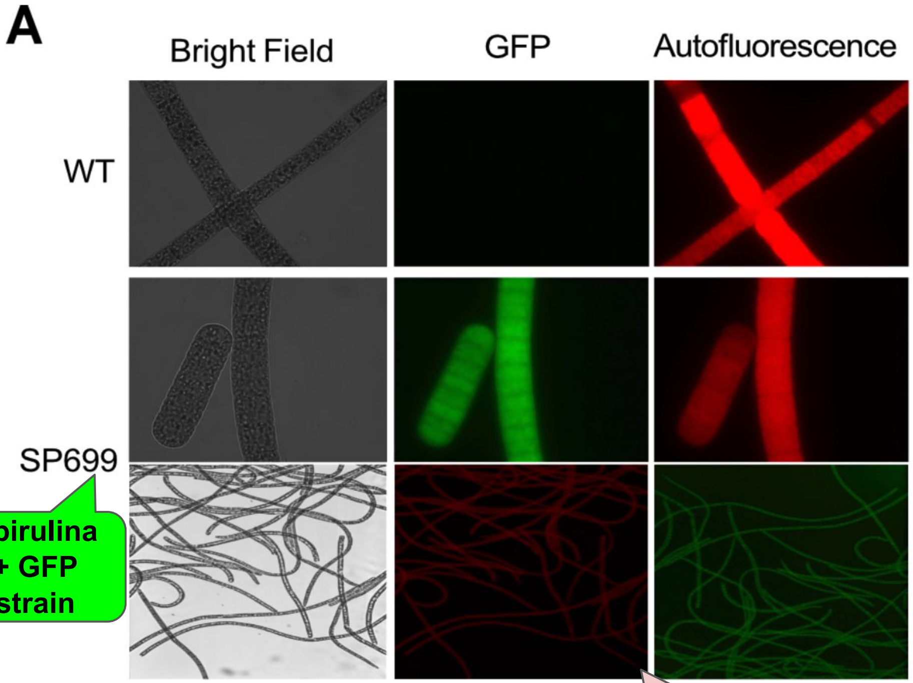
Update black-box model, mapping possible configurations to likely outcomes

Provide new observations to model

Generate next set of run configurations using batch-wise optimization strategy

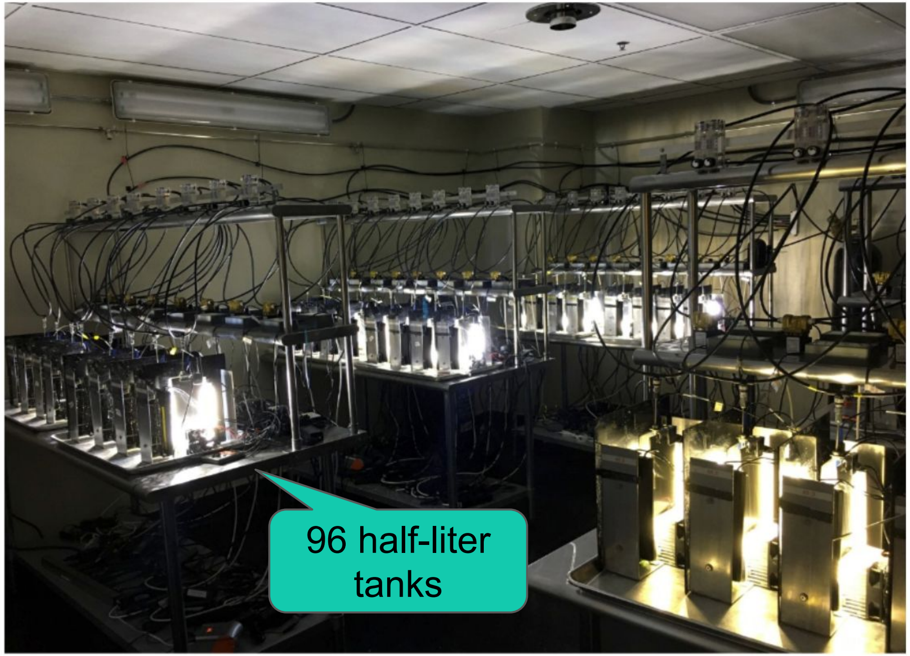
Determine next run set configurations

Figure 1A,B: Obligatory pretty biology pictures :)



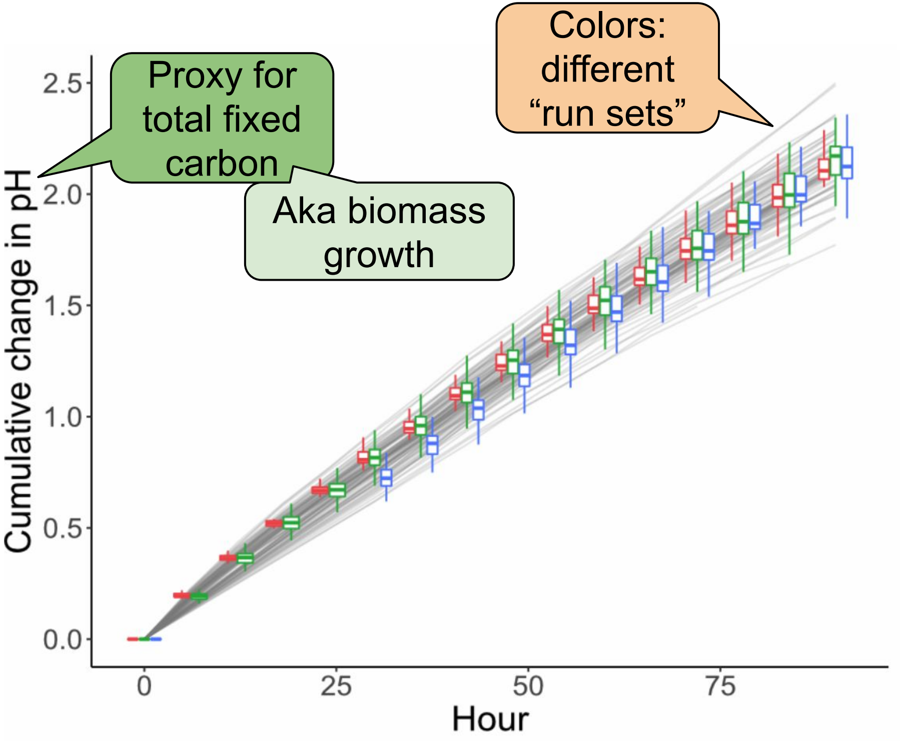
Spirulina + GFP strain

backwards...?



96 half-liter tanks

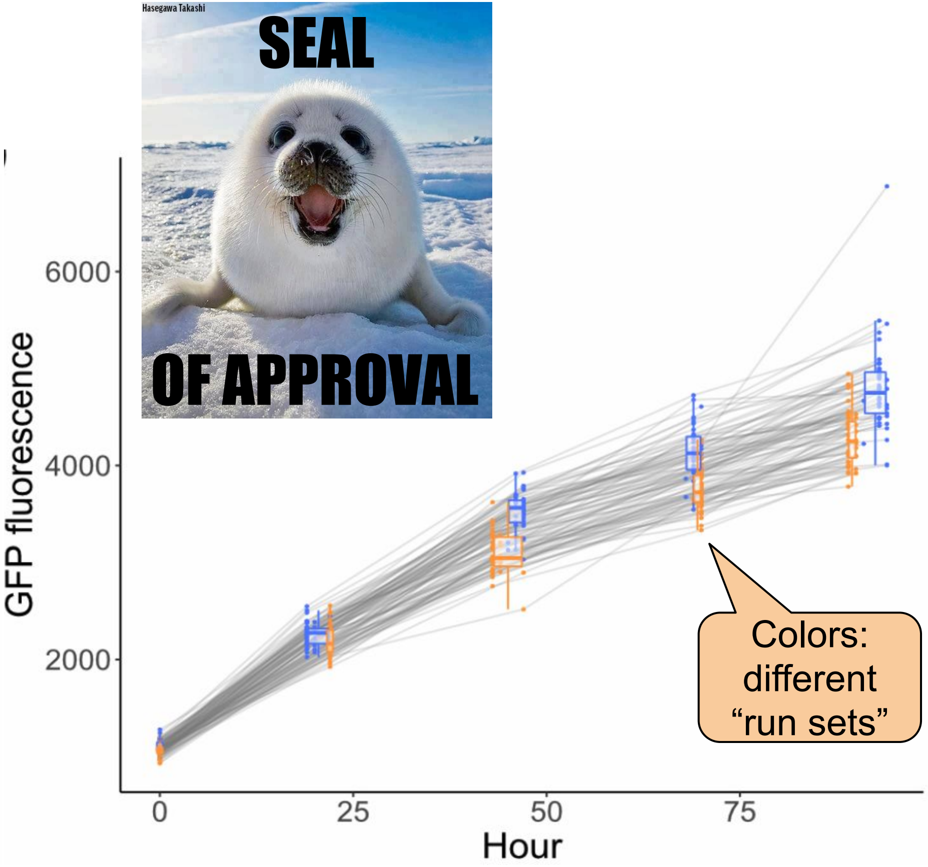
Figure 1C,D: “Commissioning” (preliminary equipment test for reproducibility)



Proxy for total fixed carbon

Aka biomass growth

Colors: different “run sets”

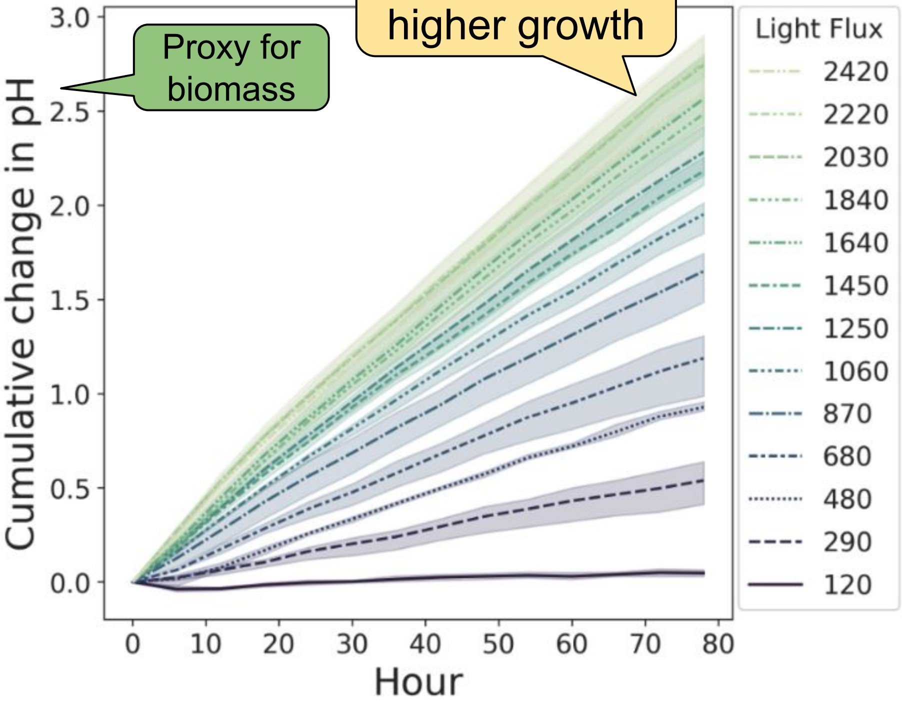


Colors: different “run sets”

Overview

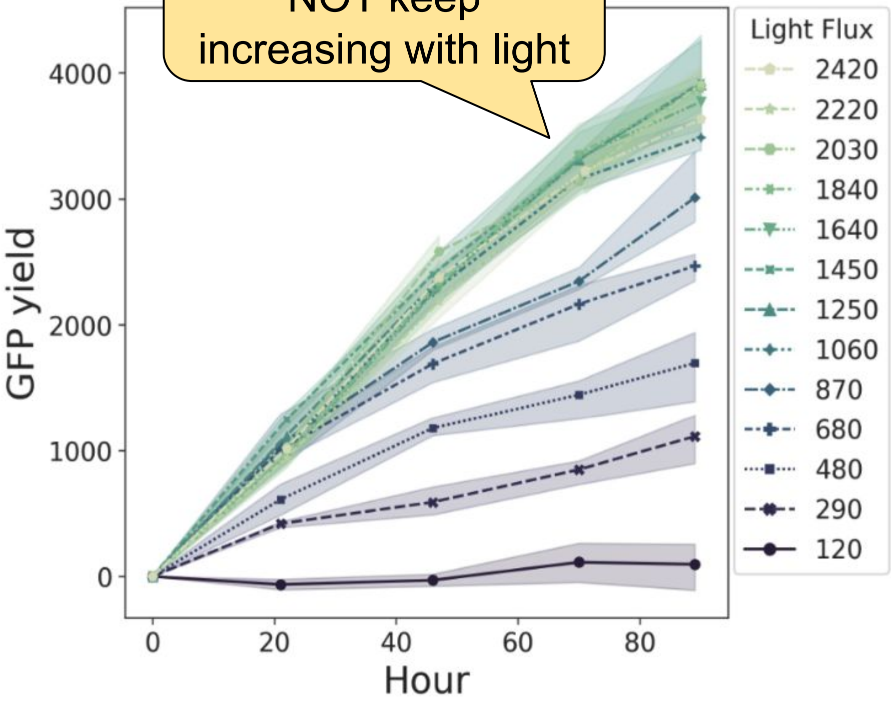
- Background
  - Metabolic Engineering + Lumen Biosciences
  - Bayesian Optimization (Gaussian Process - BUCP)
- Goals of this paper
  - Experimental set up + measurements
- **Results**
  - **Preliminary optimization outcomes**
  - **Validation of top configurations**
  - **Biological interpretation + scale up**
- Key takeaways
  - Discussion questions!

Figure 2: Varying light intensity shows tradeoffs in biomass growth and GFP yield



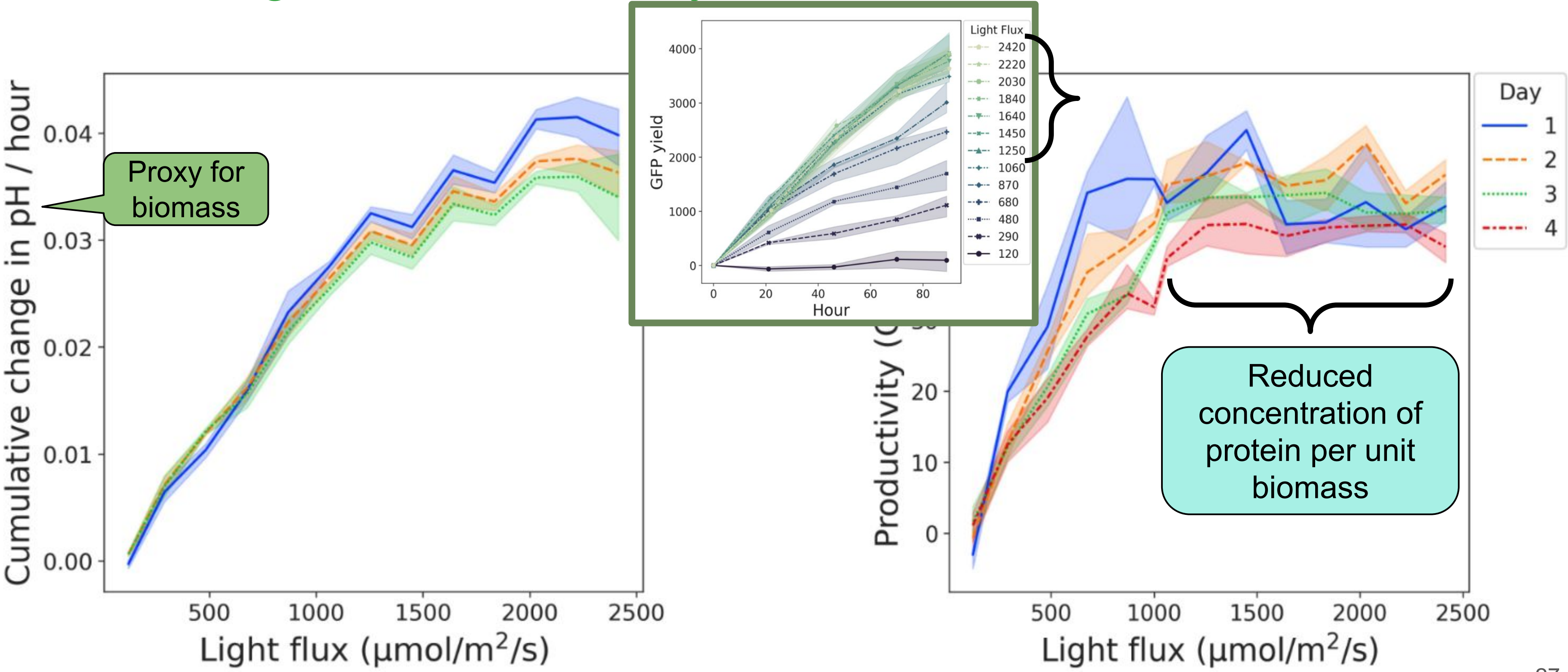
Proxy for biomass

More light, higher growth



GFP yield does NOT keep increasing with light

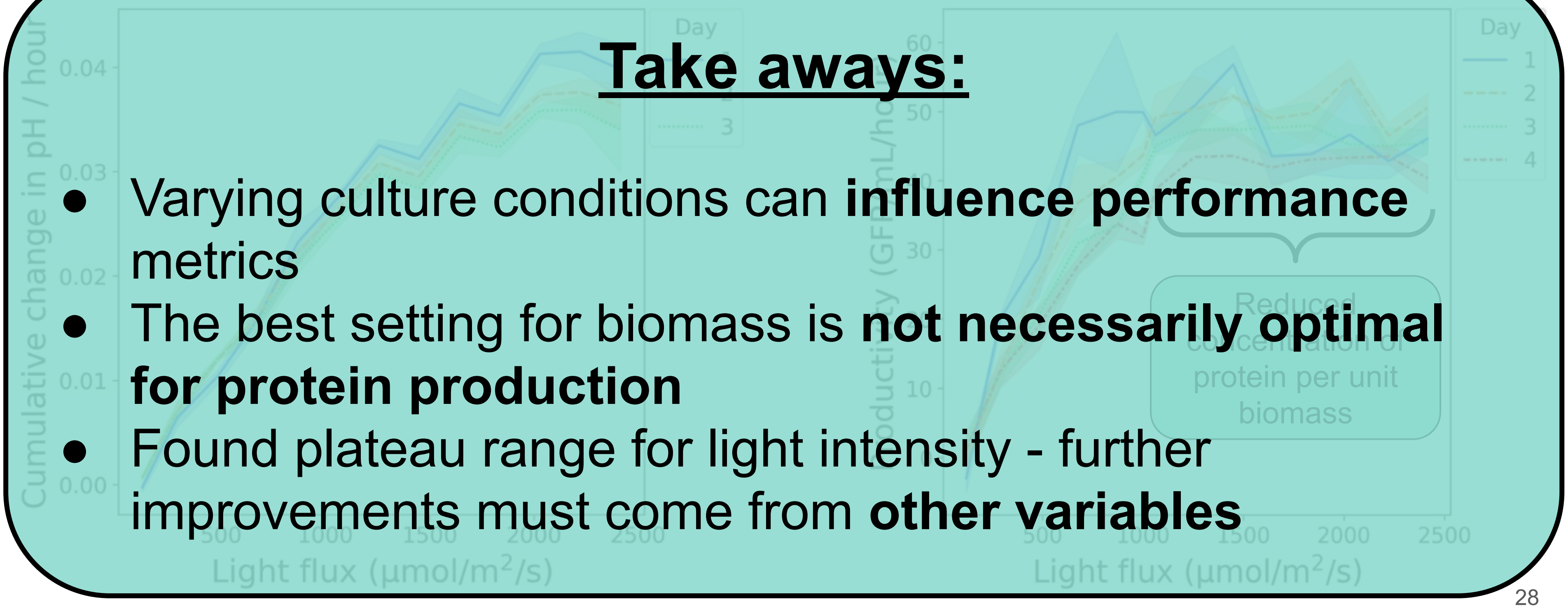
Figure 2: Varying light intensity shows tradeoffs in biomass growth and GFP yield



Proxy for biomass

Reduced concentration of protein per unit biomass

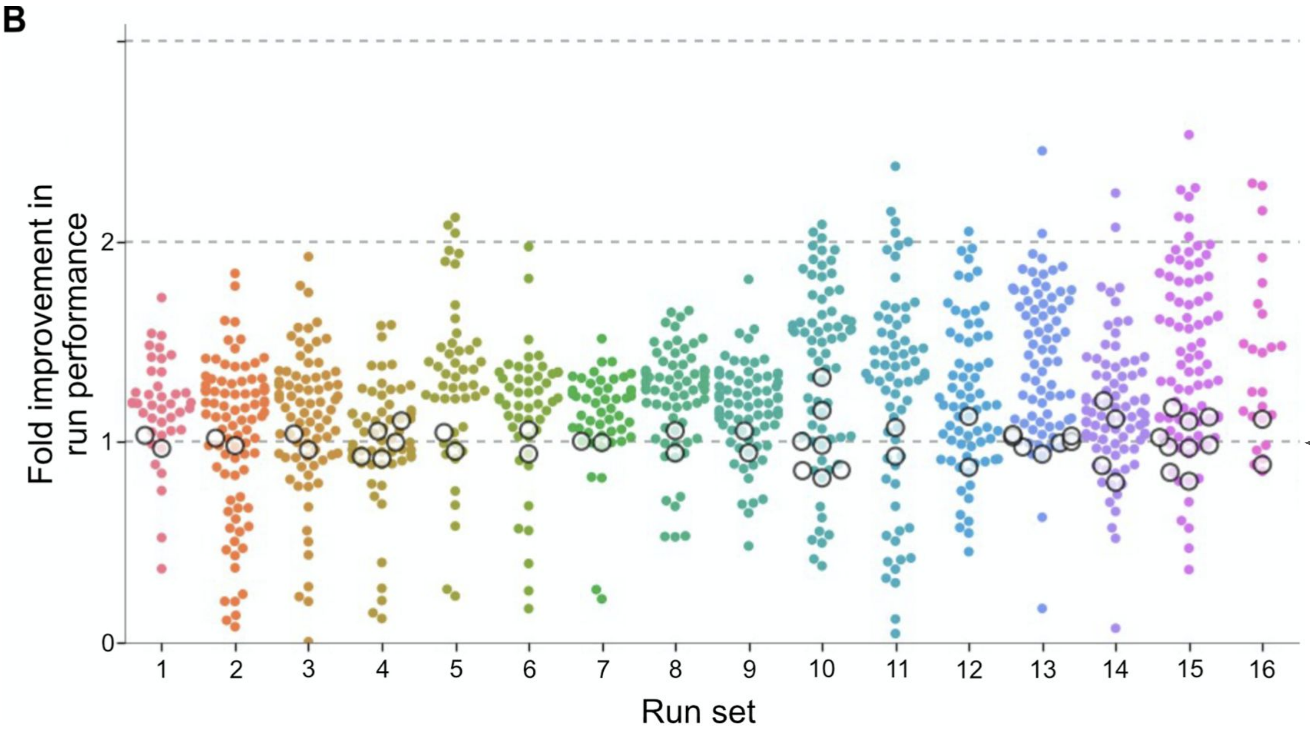
Figure 2: Varying light intensity shows tradeoffs in biomass growth and GFP yield



Take aways:

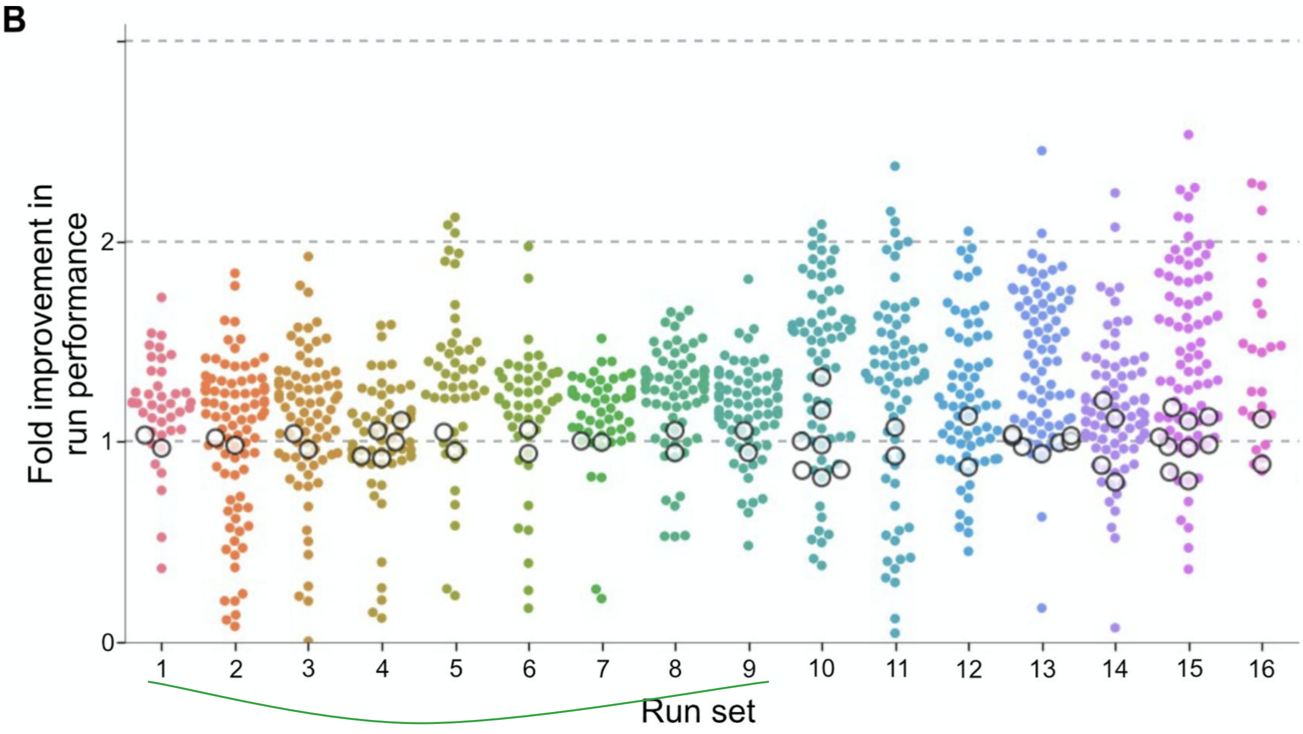
- Varying culture conditions can **influence performance** metrics
- The best setting for biomass is **not necessarily optimal for protein production**
- Found plateau range for light intensity - further improvements must come from **other variables**

Figure 3b: Run sets improve over iterations



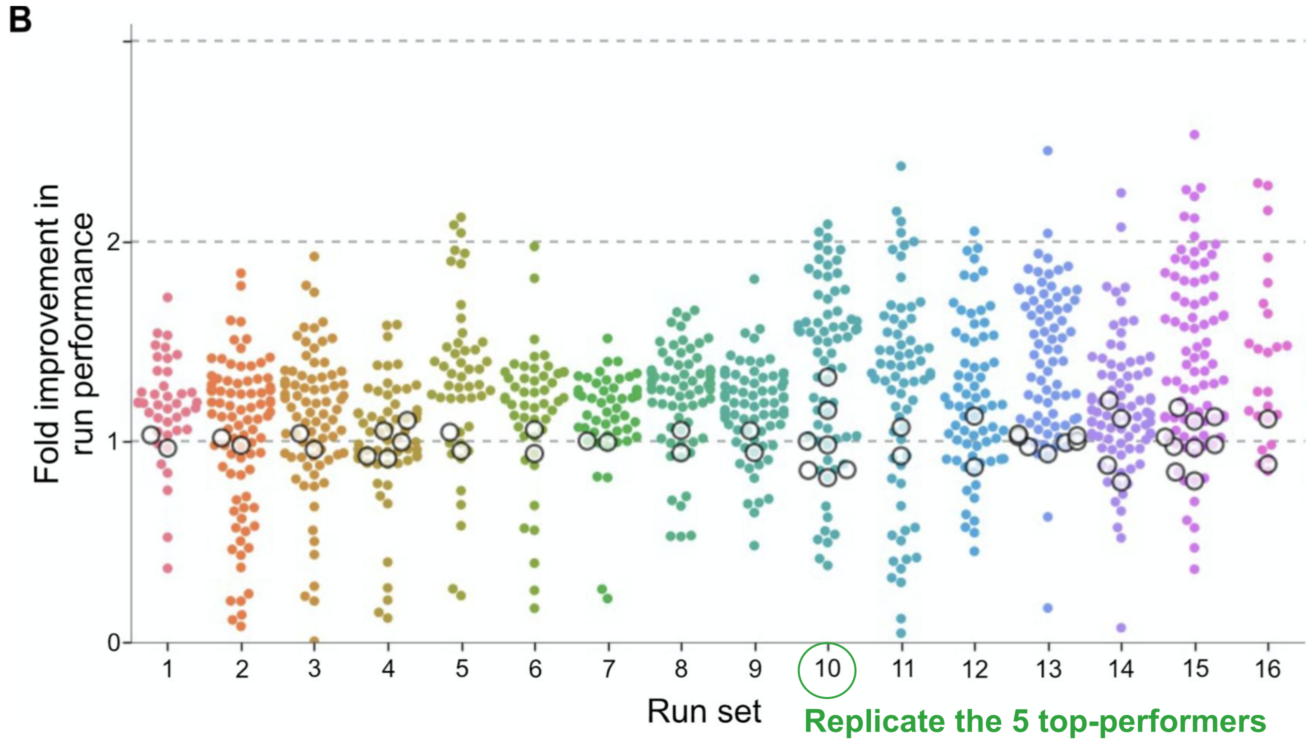
Adjusted for batch effect to mimic the global standard mean

Figure 3b: Run sets improve over iterations



Initial run sets: more biased towards exploration

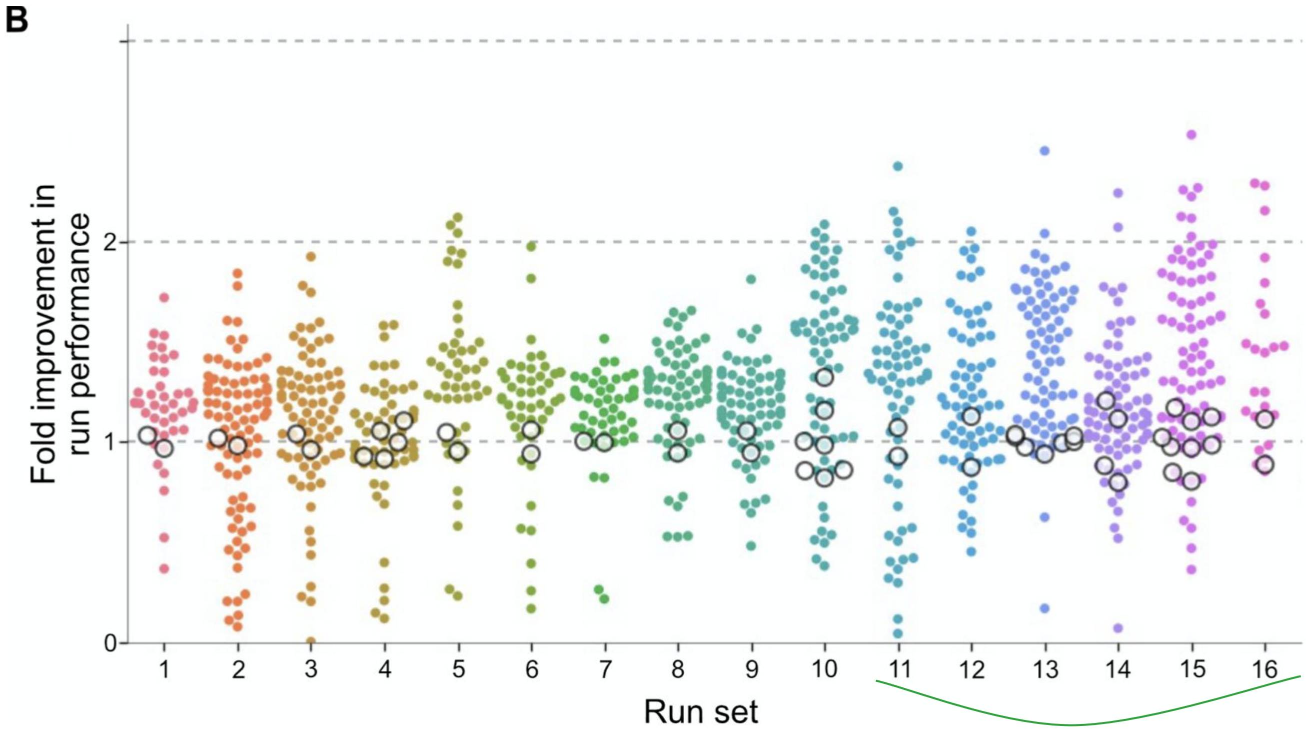
Figure 3b: Run sets improve over iterations



Replicate the 5 top-performers from run sets 1-9

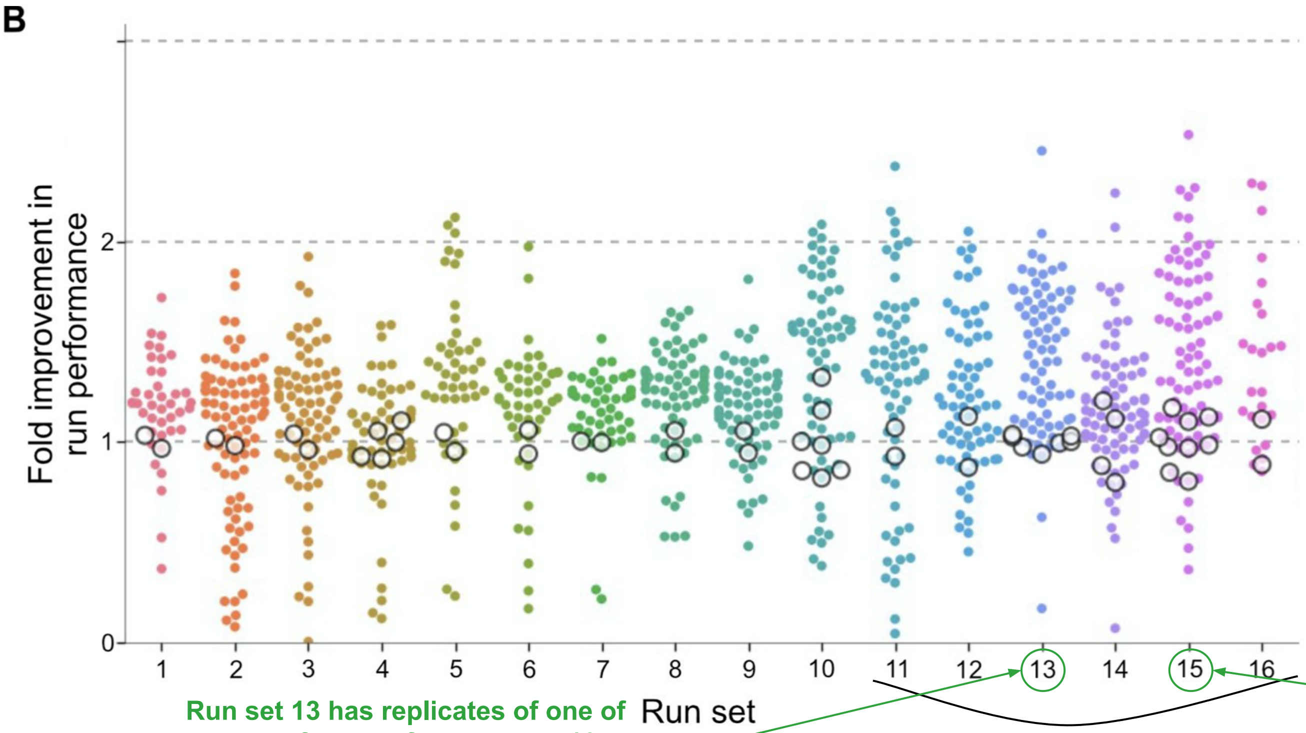
Run set 10:
Group mean fold improvement: 1.8
Std. dev: 0.25
T-test p-value: 3.3e-12

Figure 3b: Run sets improve over iterations



Second group of run sets

Figure 3b: Run sets improve over iterations



Run set 15 (?):
Group mean fold improvement: 1.8
Std. dev: 0.14
T-test p-value: 1.2e-6

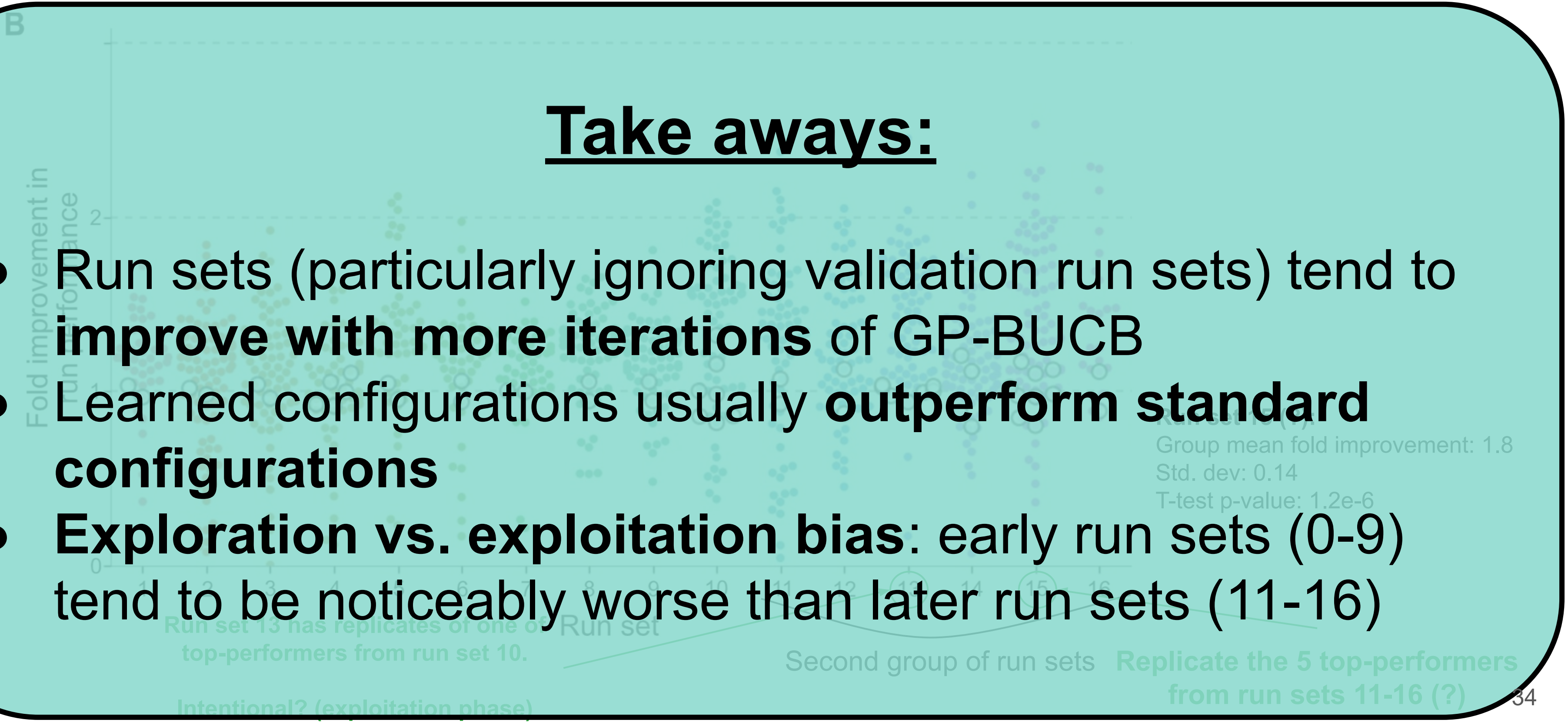
Run set 13 has replicates of one of top-performers from run set 10.

Intentional? (exploitation phase)

Second group of run sets

Replicate the 5 top-performers from run sets 11-16 (?)

Figure 3b: Run sets improve over iterations



Group mean fold improvement: 1.8
Std. dev: 0.14
T-test p-value: 1.2e-6

Run set 13 has replicates of one of top-performers from run set 10.

Intentional? (exploitation phase)

Second group of run sets

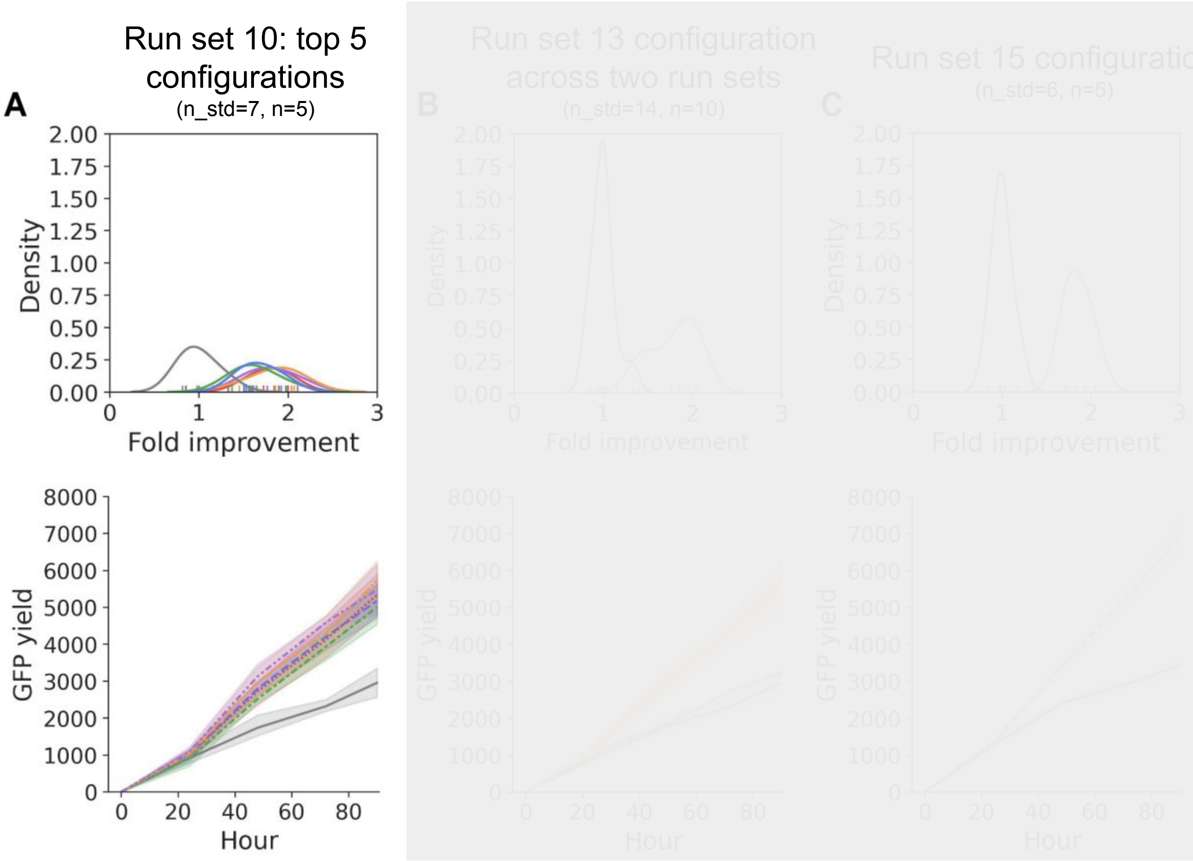
Replicate the 5 top-performers from run sets 11-16 (?)

Take aways:

- Run sets (particularly ignoring validation run sets) tend to **improve with more iterations** of GP-BUCB
- Learned configurations usually **outperform standard configurations**
- **Exploration vs. exploitation bias**: early run sets (0-9) tend to be noticeably worse than later run sets (11-16)

Figure 4: Learned configurations outperform the standard

- Gray is standard run
- Colors show configurations of interest
- GFP yield includes 95% confidence intervals



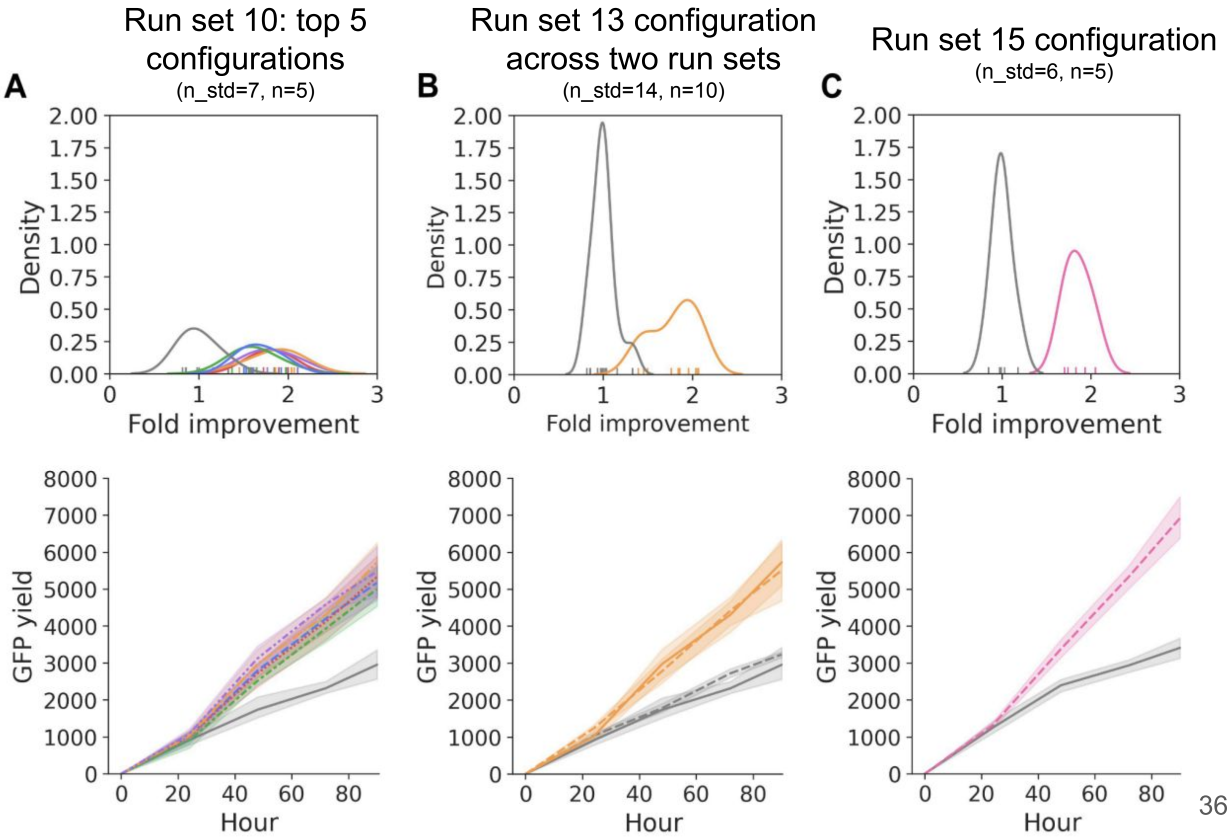
Run set 10: top 5 configurations (n_std=7, n=5)

Run set 13 configuration across two run sets (n_std=14, n=10)

Run set 15 configuration (n_std=6, n=5)

Figure 4: Learned configurations outperform the standard

- Gray is standard run
- Colors show configurations of interest
- GFP yield includes 95% confidence intervals



Run set 10: top 5 configurations (n_std=7, n=5)

Run set 13 configuration across two run sets (n_std=14, n=10)

Run set 15 configuration (n_std=6, n=5)

So this process seems to be able to improve performance...

Which parameters
(and which values)
were most important for success?

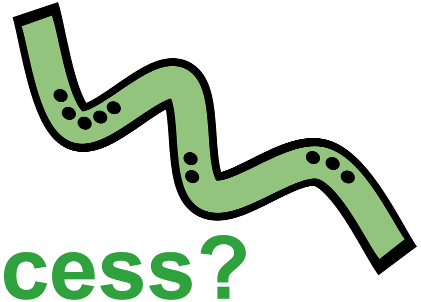


Figure 5:

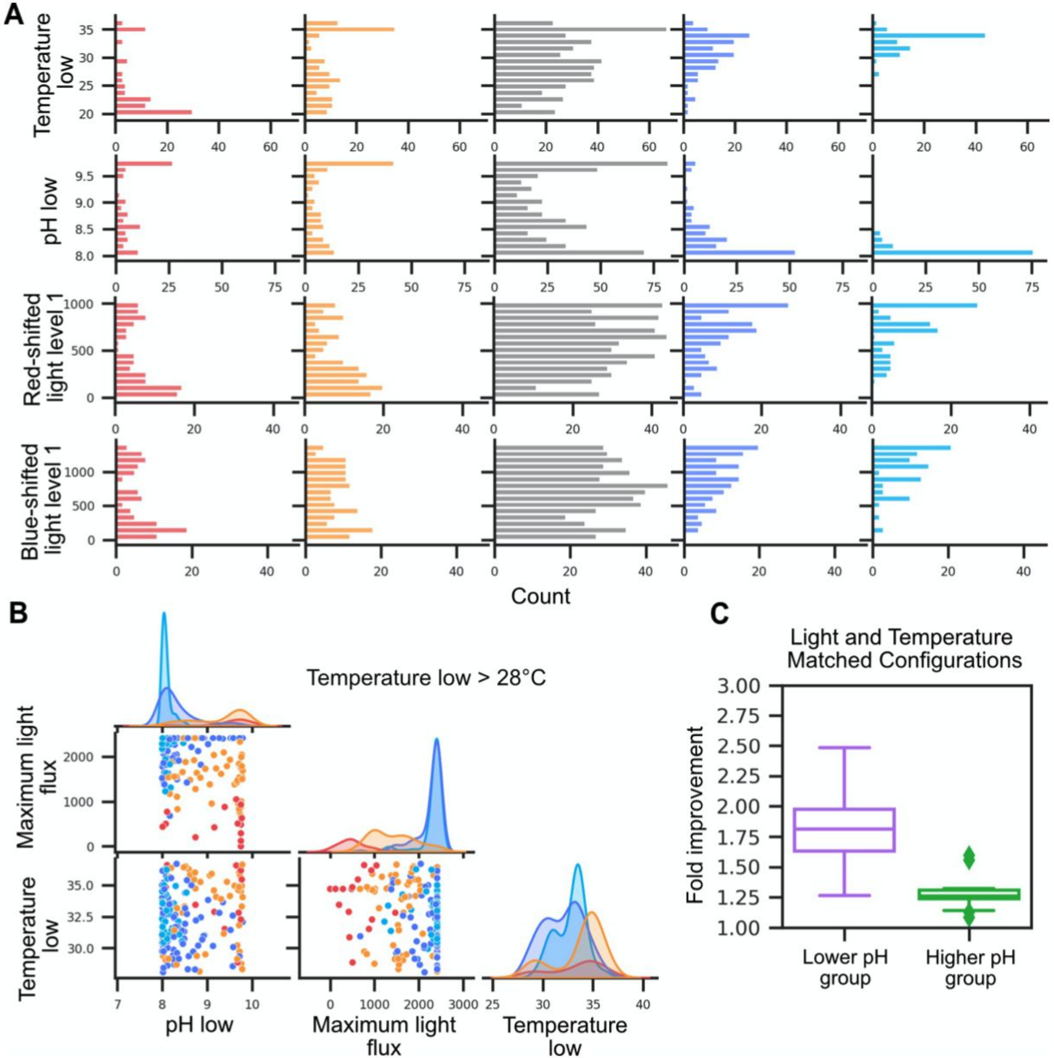

A
Temperature low
35
30
25
20
0
20
40
60
pH low
9.5
9.0
8.5
8.0
0
25
50
75
Red-shifted light level 1
1000
500
0
0
20
40
Blue-shifted light level 1
1000
500
0
0
20
40
Count
B
Temperature low > 28°C
Maximum light flux
2000
1000
0
Temperature low
35.0
32.5
30.0
7
8
9
10
pH low
0
1000
2000
3000
Maximum light flux
25
30
35
40
Temperature low
C
Light and Temperature Matched Configurations
Fold improvement
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
Lower pH group
Higher pH group

Figure 5A: Biased distributions of parameter values for top configurations

Bottom 10% Bottom 25% IQR Top 25% Top 10%

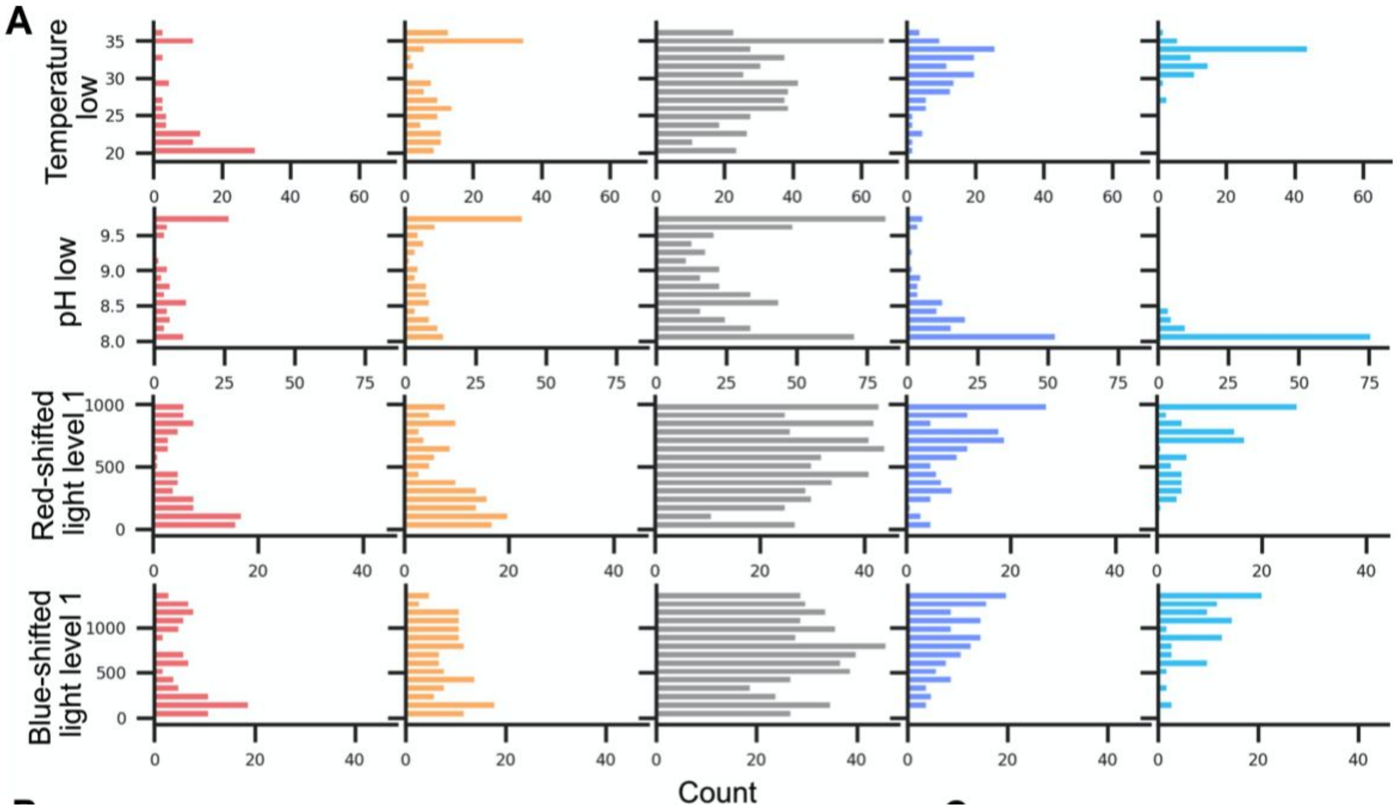
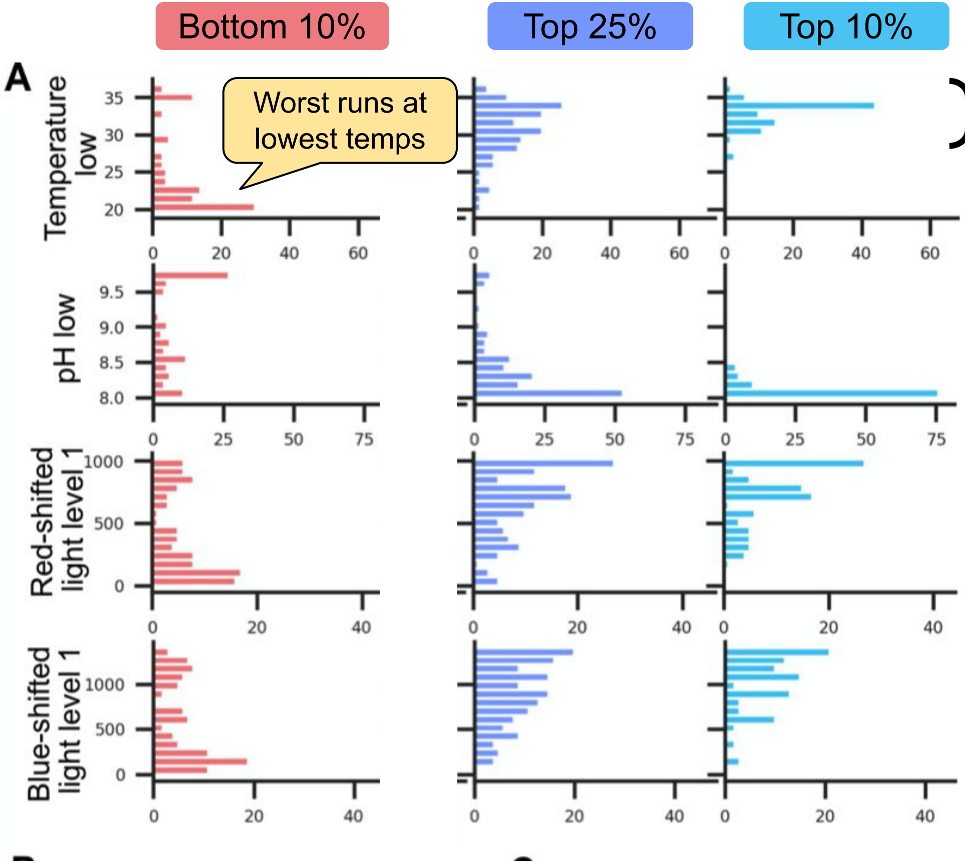


Figure 5A: Biased distributions of parameter values for top configurations

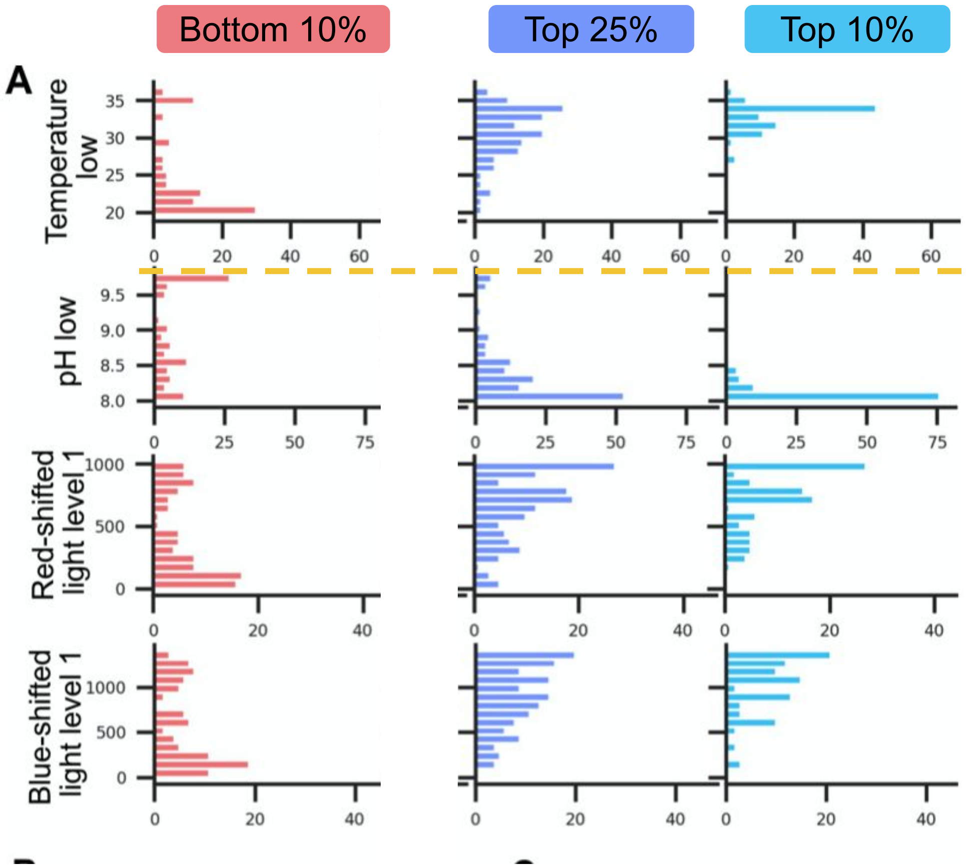


Worst runs at lowest temps

Best temps between 27-36° (IQR 33-34°)

Below max temp (37°) and “standard” (35°)

Figure 5A: Biased distributions of parameter values for top configurations



Opposite end from "standard" (9.75-9.95)

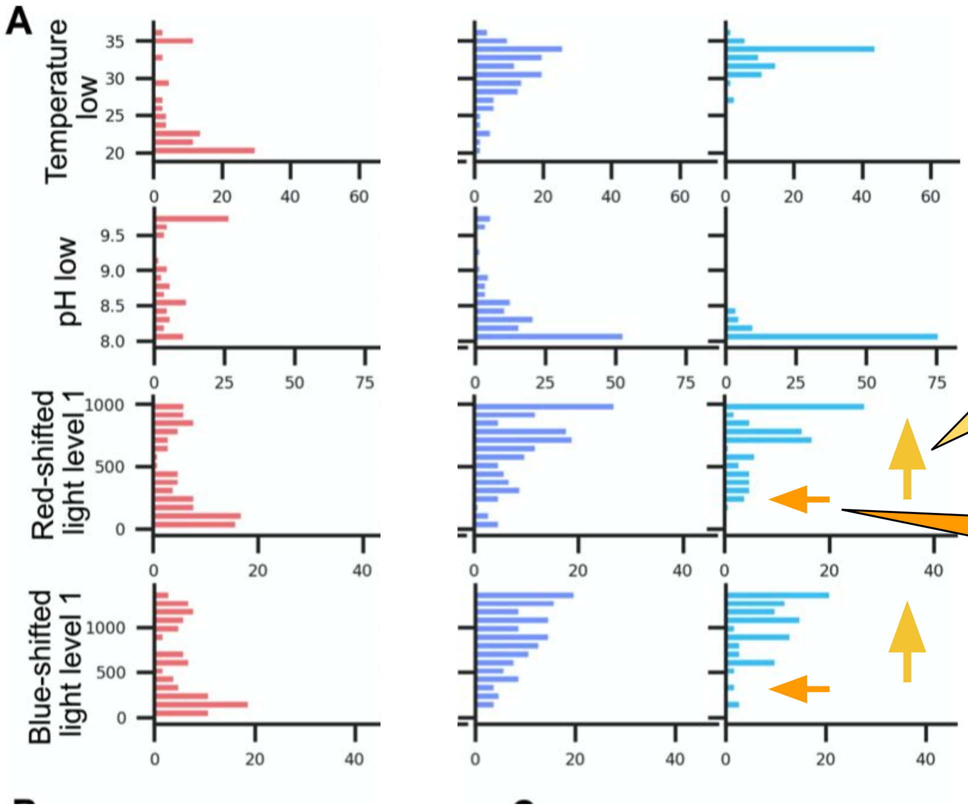
Strong bias towards lowest possible pH (8.06)

Figure 5A: Biased distributions of parameter values for top configurations

Bottom 10%

Top 25%

Top 10%



Trend towards higher light intensities

But high performance possible at lower levels too

Figure 5B: Some nuances exist in parameter combinations

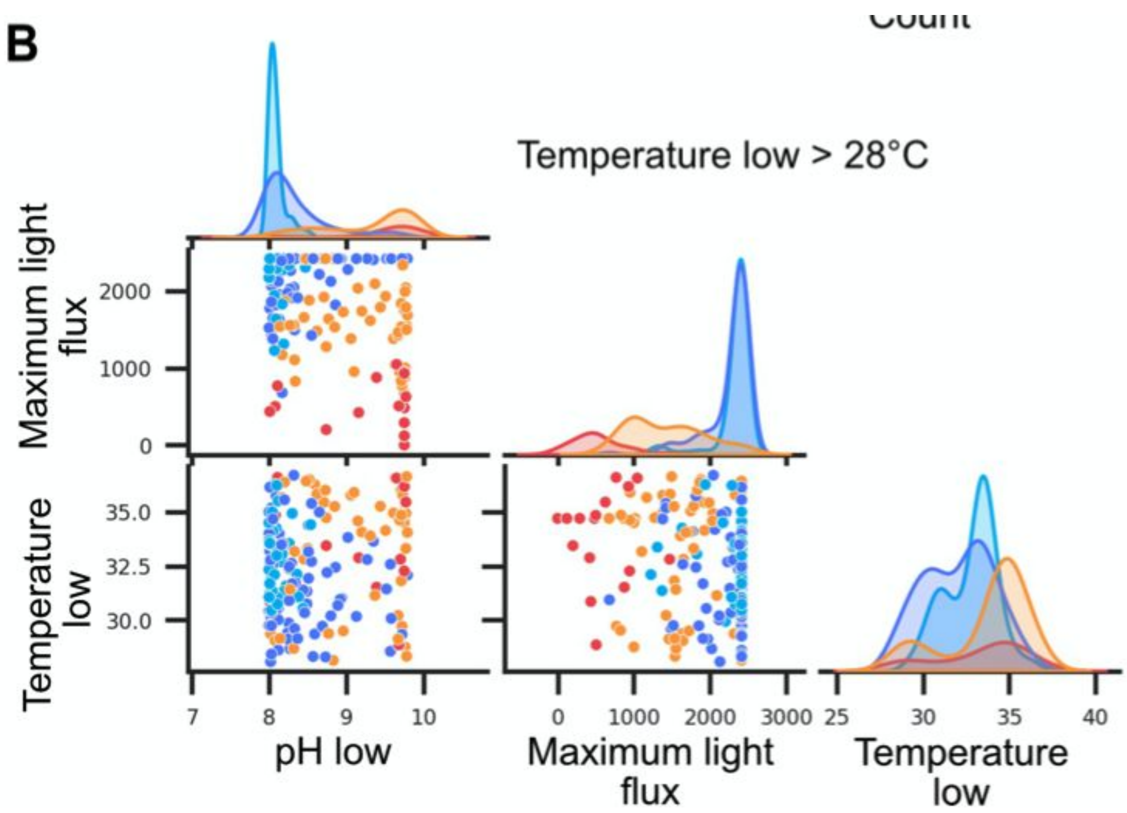


Figure 5B: Some nuances exist in parameter combinations

Strong bias towards lowest possible pH (8.06)

Bias towards high light flux

Preferred temp around 33-34°

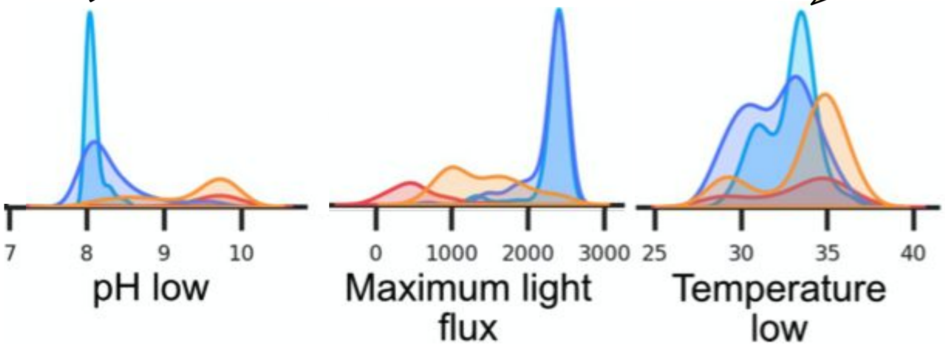
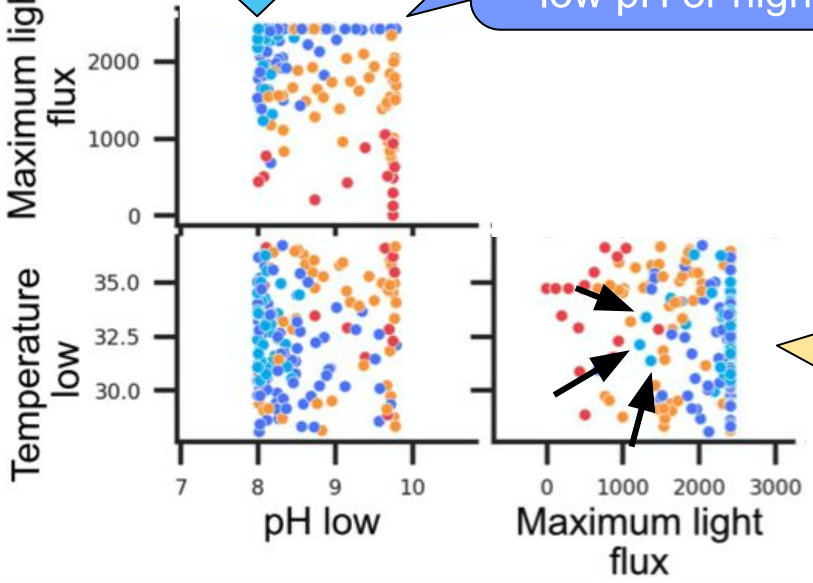


Figure 5B: Some nuances exist in parameter combinations

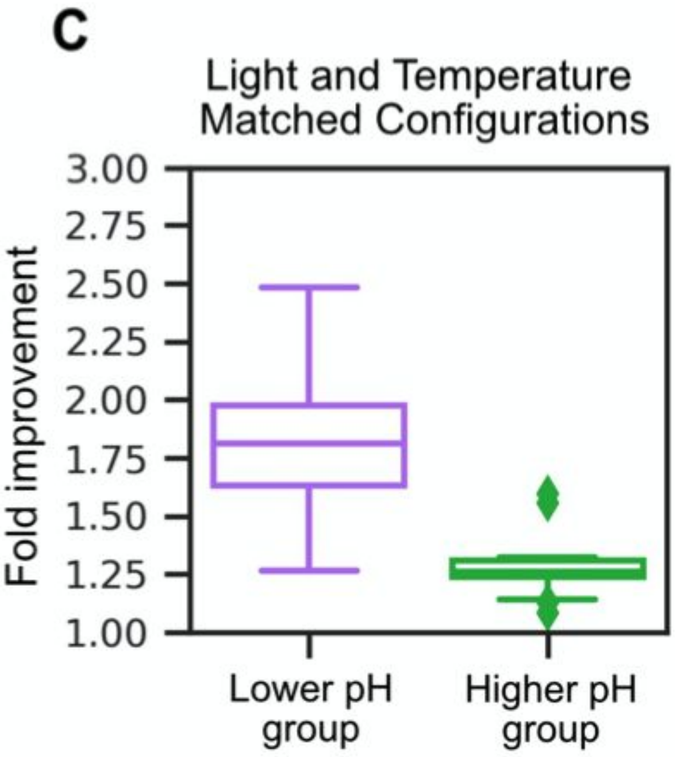
Top 10% configs tended towards **both**

Top 25% configs tended towards **either** low pH or high light



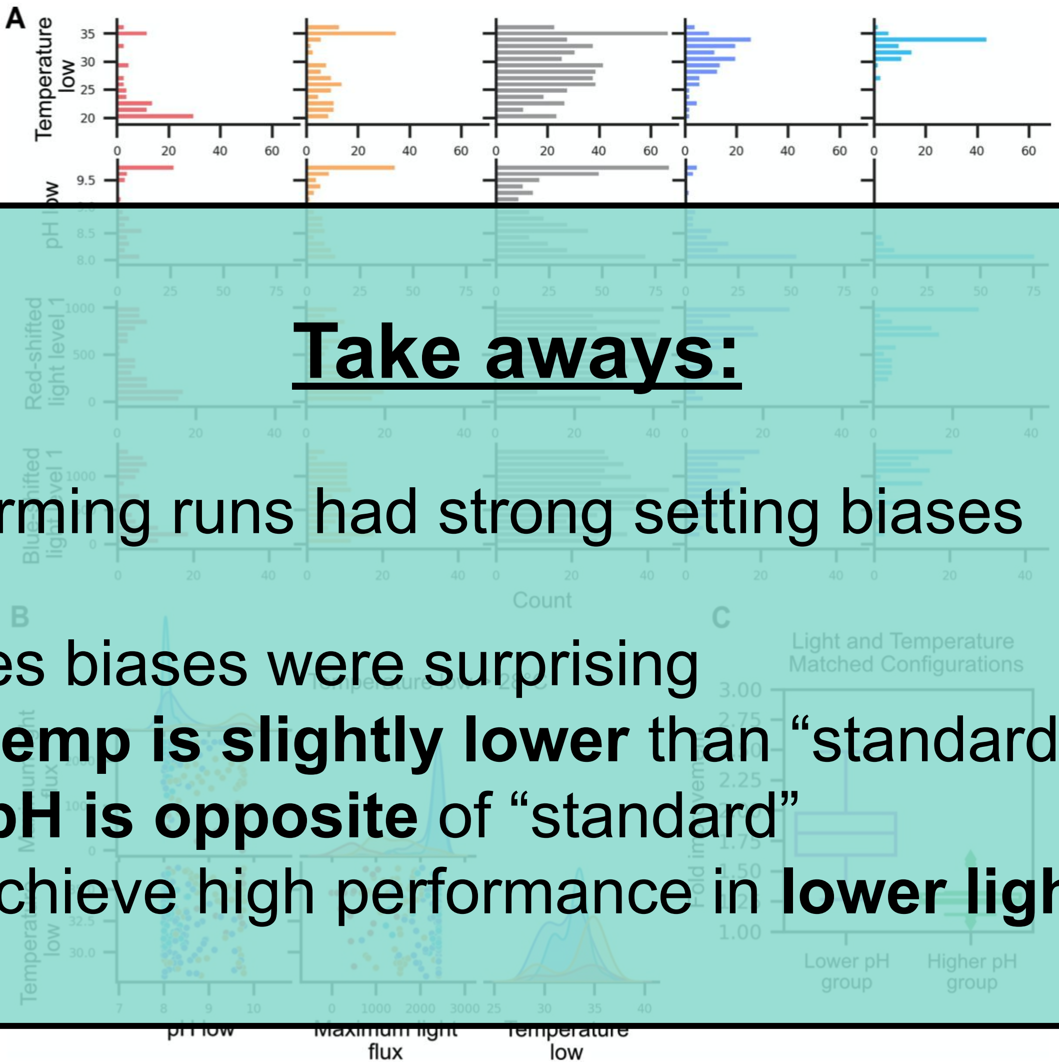
Some top configs had lower light

Figure 5B: Some nuances exist in parameter combinations



When light and temp are equal, **lower pH is clearly better**

Figure 5:



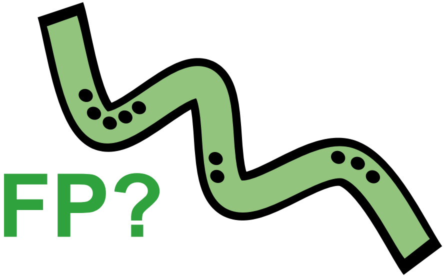
Take aways:

- Top performing runs had strong setting biases
- Sometimes biases were surprising
  - Ideal **temp is slightly lower** than “standard”
  - Ideal **pH is opposite** of “standard”
  - Can achieve high performance in **lower light** regimes

So ML discovered some promising new spirulina culture configurations...

Does it work:

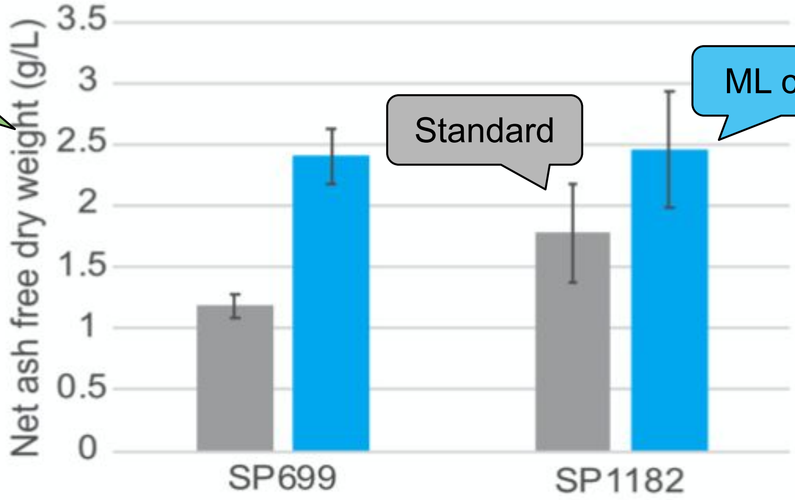
- At larger scales?
- For a protein other than GFP?



VHH

Name	Value
Air flow	0.8
Number of light levels	2
Number of light periods	9.27
Light level 1 fraction	0.16
Blue-shifted light level 1	1307
Blue-shifted light level 2	1399
Red-shifted light level 1	1003
Red-shifted light level 2	282
Blue-shifted light gradient	0.49
Red-shifted light gradient	0.37
Number of temperature levels	1
Number of temperature periods	
Temperature level 1 fraction	
Temperature level 1	33.85
Temperature level 2	
pH lower bound (Φ_{lower})	8.01
pH upper fraction (*f*)	0.045

Figure 6A: Biomass growth is better with ML config



Proxy for biomass

Standard

ML config

GFP

VHH

Fig S8: but not VHH protein 😬

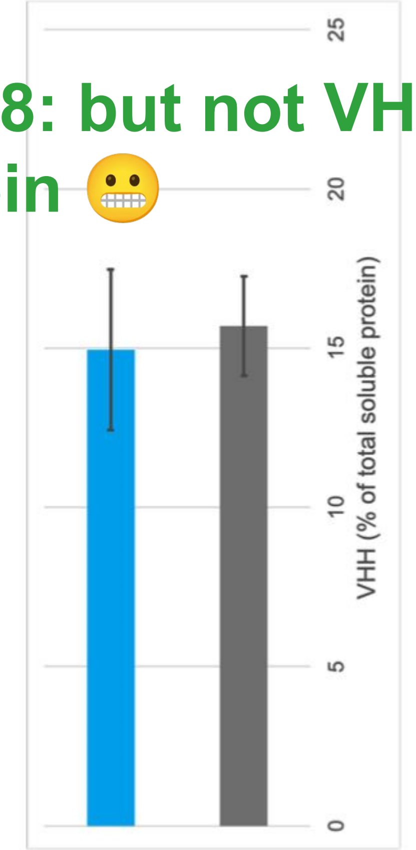
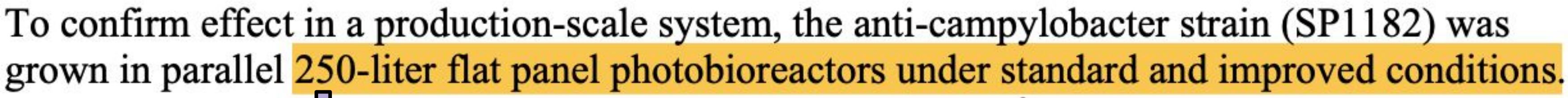
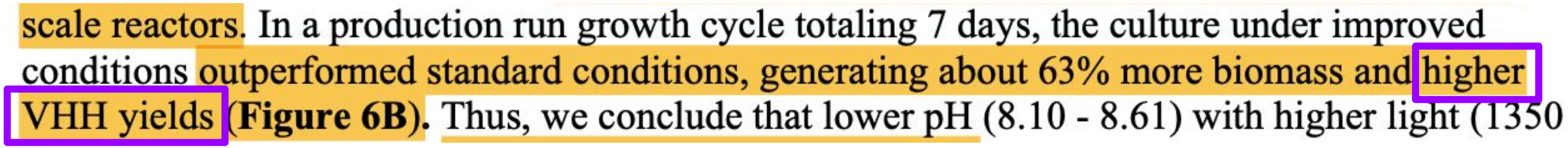


Figure 6B: a bit of a mystery...

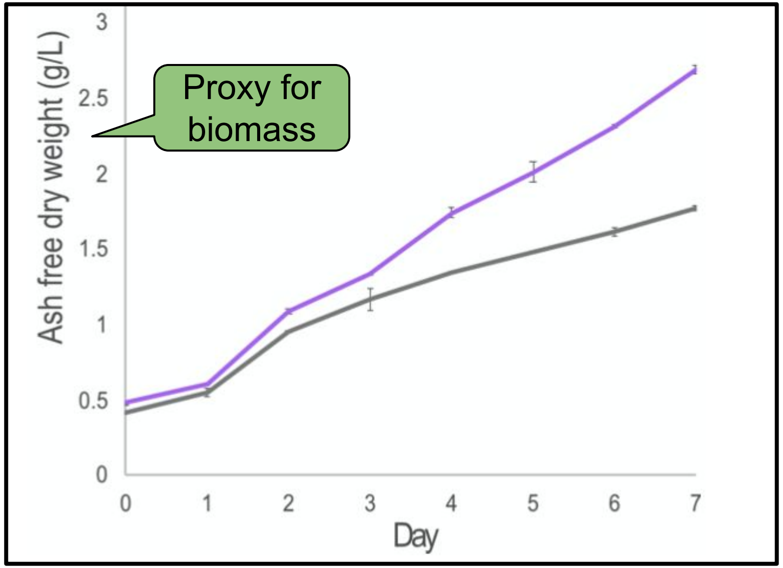


To confirm effect in a production-scale system, the anti-campylobacter strain (SP1182) was grown in parallel 250-liter flat panel photobioreactors under standard and improved conditions.

500x bigger



scale reactors. In a production run growth cycle totaling 7 days, the culture under improved conditions outperformed standard conditions, generating about 63% more biomass and higher VHH yields (**Figure 6B**). Thus, we conclude that lower pH (8.10 - 8.61) with higher light (1350



Proxy for biomass

B) Biomass growth of an anti-campylobacter antibody strain (SP1182) in 250 L reactors. Improved condition based on ML-guided experimentation (orange) and initial standard condition (blue). Error bars represent standard deviation of AFDW measurements.

Is the figure axis mislabeled?

Did they forget to put in the VHH graph?

Did they forget to edit out the VHH claim from the text?

Overview

- Background
  - Metabolic Engineering + Lumen Biosciences
  - Bayesian Optimization (Gaussian Process - BUCP)
- Goals of this paper
  - Experimental set up + measurements
- Results
  - Preliminary optimization outcomes
  - Validation of top configurations
  - Biological interpretation + scale up
- **Key takeaways**
  - **Discussion questions!**

Summary of Key Takeaways

- Spirulina culture **conditions are tunable** and have sizeable **impact on performance**

- **Existing computational methods** can be applied to this problem

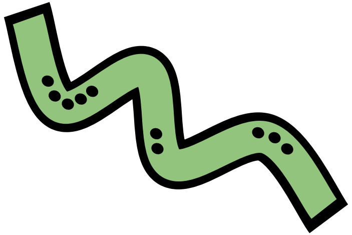
- Previously used “standard conditions” may be suboptimal for therapeutics production
    - **ML optimization** can provide a route to **improved efficiency** for biologic manufacturing

Discussion questions

- Why not **VHH whole time**?
  - Cost of GFP measurements?
- What if they repeated this process but **starting from GFP prior** but for VHH measurement
  - Maybe get there faster?
- Cool application of algorithm for **“hyperparameter” search**
  - Experimental settings instead of genetic changes
- **Figure** composition/usefulness?
- **Statistical robustness** of conclusions

- What was the **goal of paper**?
  - To tell people **actual optimal** experimental set up for spirulina?
  - To **advertise** that this company is doing ML?
  - **Required** by funding/Google collab?
  - **Encourage BO** in general?

- If you were a reviewer, what kinds of **feedback would you give**?



Thanks for listening!

Second Beach, Olympic Peninsula, WA

- Comparison of how they “wielded” BO
  - What settings were actually used?
  - Batching methodology

Fig 1: here are our machines - they make good data

- Fairly reproducible
- Ooooh lights
- Green vs red flip?
- How did 1c get to 2.5?

Fig 2: not yet doing opt but look at the difference 1 variable can make

- Hyperparams CAN be optimized
- Also, tradeoffs up to a certain point
  - More light does not always mean more protein
- Discussion: in addition to protein gathering cost, what’s the cost of running the machines
  - More light more expensive? (more energy expended)
  - More time = more expensive
  - Hyperparams themselves have costs
- Data viz - which version of fig more useful?
  - Showing the “plateau”
  - Confusing to understand
  -

Fig 3: mini sys diagram + look: configs get better over iterations

- Did they “explore” enough in the early run sets?
    - Sounds like a parameter you can tune
- Call out which runs sets are “special”
- Which samples are replicates vs diff config
    - Explain in detail 1 run set
- Run set 10, 13, 15 are confirmations
    - 10 - top 5 from early group
    - 13 = one of those top 5 again
    - 15 - top ever from second set (run 15 → fig 4C)

Fig 4: specific dives into best configs from fig 3

- A: results from runset 10
    - All engineered envs usually outperform standard
- B: took one of those 5, did it again
    - Week to week reproducibility
    - Run 13
- C: took top from second batch (11-16 (-13))
    - Top point on 15 run, rerun
- Gap between B and C is bigger - BO is still learning
- Did they update between 5-6? 7-8? Or just between 1-10, 11-16?
    - Are B-C between 1 update?

Fig 5: showings of where the best configs were

- Interpretability section
  - With no stats :(
- A: Mostly care about teal columns
  - Temp low: red and teal look very different
  - Maybe get rid of the middle ranges
  - Because of BO, fewer points at lower temps
  - Dark blue kind of mimics the teal
- B: max light flux convincing
  - Same with low ph
  - Call out dark blue: ph vs light flux - must have one. Teal - has both
- C: most clear part of this figure
  - When all else is equal, have a lower ph

Fig 6: did this work real protein (VHH)

- A: biomass at 450mL - higher in ML config
  - No p-value!
  - Not super strong stat power + overlap of error bars
  - Supp Fig 8 - shows no difference in VHH production :(
- B: in text it says VHH protein production was higher, but in fig, only shows growth
  - AH!
  - Maybe there was a mix up?
  - Growth vs protein - Correlated but not exact
  - If plot was actually VHH, that’d be a nice end to the story
  -

Background

- Metabolic engineering
  - Metrics you care about (yield vs productivity)
  - Challenges growing photo orgs
  - A few fun facts about spirulina
- Bayesian optimization
  - When to apply? When can you apply?
  - Upper confidence bound - borrow figures about narrowing in on certain regions
- Their goal: iteratively guide exp settings
  - What the standard conditions actually are
- >> then to figures
- >> discussion points

Slide flow?

- Background
  - Metabolic engineering + protein production measurements/proxies; spirulina + photosynthetic org systems
  - Bayesian optimization; when to apply/when can apply
  - Paper's goal - optimize bioreactor culture conditions
- Figure 1 - preliminary data collection set up
- Figure 2 - initial evidence that optimization tradeoffs are possible
- Figure 3 - evidence of configs getting better
- Figure 4 - confirmation/validation of specific configs relative to standard
- Figure 5 - interpreting best config settings
- Figure 6 - scale up + actual VHH protein run
- Summary of our take aways, lingering questions, complaints
- Discussion Questions + open to the audience

Erin?

Addie?