Efficient Screening of Strain Collections with Bayesian Inference and Thompson Sampling

Osthege, M., Helleckes, L.M., Wiechert, W., Oldiges, M.

Institute of Bio- and Geosciences: IBG-1, Forschungszentrum Jülich, Jülich, Germany

m.osthege@fz-juelich.de   github.com/michaelosthege   github.com/JuBiotech

SPONSORED BY THE
Federal Ministry of Education and Research

(031B0463A)

JÜLICH
Forschungszentrum

ABSTRACT

Gene editing, cloning or mutagenesis techniques can deliver large numbers of **candidate strains** from which high-performers must be identified. Such strain collections can easily saturate the throughput of cultivation and characterization techniques, in particular those with fine process control and production scale comparability. It is therefore desireable to **characterize high-performers** well, without wasting experimental resources on under-performing strains. This task of exploiting high-performing candidates while **minimizing the resources spent on under-performers** is a prime example for the application of Bayesian optimization techniques.

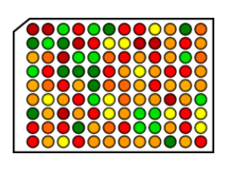
On this poster we present how probabilistic generative models of automated microbioreactor (MBR) processes can be combined with the Thompson sampling algorithm to characterize high-performing strains from a mutagenesis collection in few rounds of experimentation.

Prerequisites

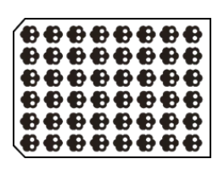
experimental

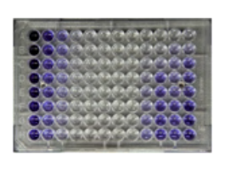
Collection of histidine-producing Corynebacterium glutamicum
- 96 mutant strains provided by SenseUP Biotechnology GmbH
- Growth-coupled product formation
- Productivity unknown beforehand

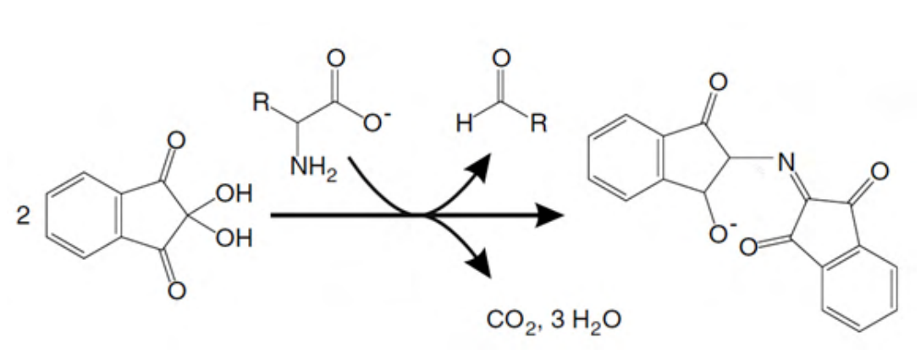
SenseUP BIOTECHNOLOGY



Autonomous MBR cultivation + sampling + assays
- Inoculation from cryo MTPs
- Batch cultivation on CGXII+glucose parallelized 48x
- Time-based harvesting, centrifugation & storage
- Hexokinase assay for substrate quantification
- Ninhydrin assay for product quantification

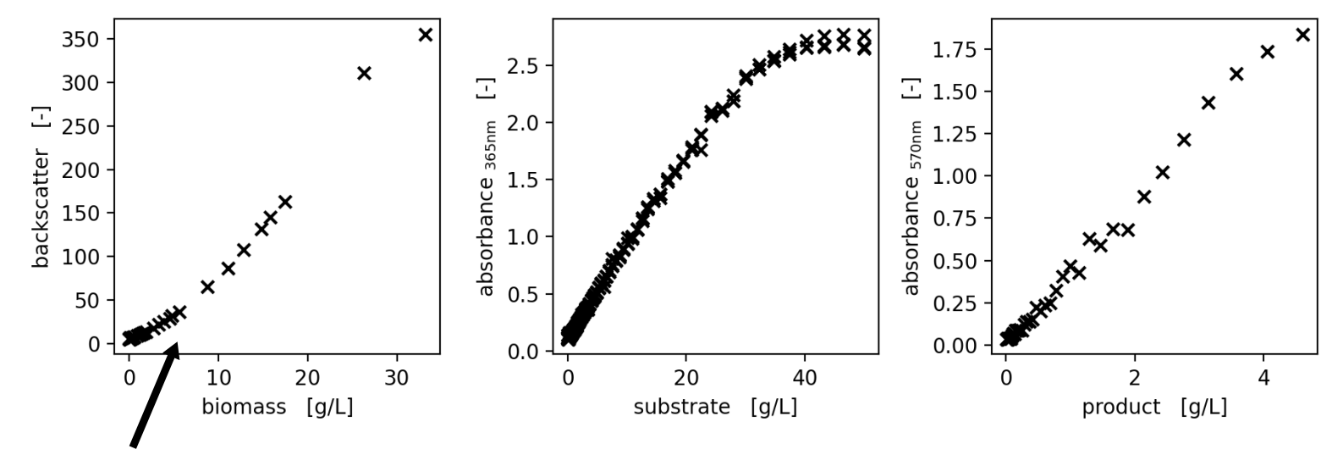






calibration

Needed to translate between...
- BioLector backscatter vs. biomass conc.
- 365 nm absorbance vs. substrate conc.
- 570 nm absorbance vs. product conc.



⚠ Non-linear relationships in most measurement procedures
- ▶ Need empirical model of measurement uncertainty
- ▶ Built Python package **`calibr8`** for calibration modeling
- ▶ Enables probabilistic machine learning with real data

pypi v6.0.0 | pipeline passing | codecov 93% | docs passing | DOI 10.5281/zenodo.4651250

process model

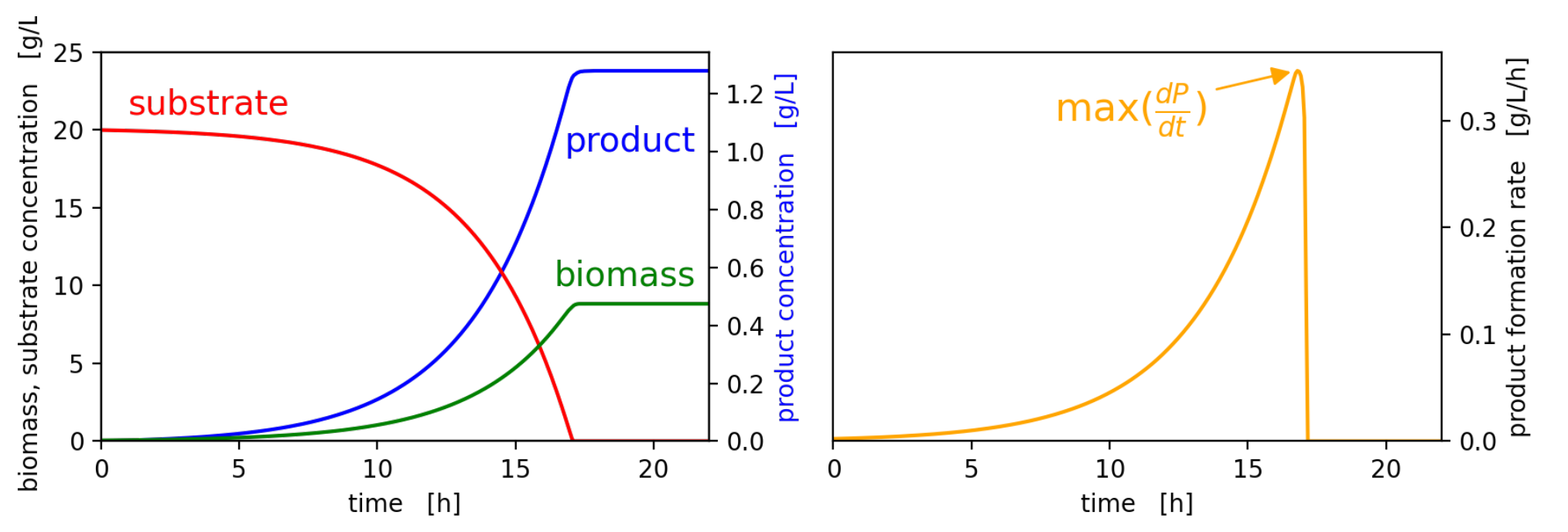
Mechanistic bioprocess model
- Monod-like differential equations (ODE)
- Growth-coupled product formation
- Screening metric predicted by model under standardized conditions
- Lag phase explained by simple fraction of adapted cells

$$\frac{dX}{dt} = \mu_{\max} \cdot S \cdot X \frac{X}{K_S + S}$$

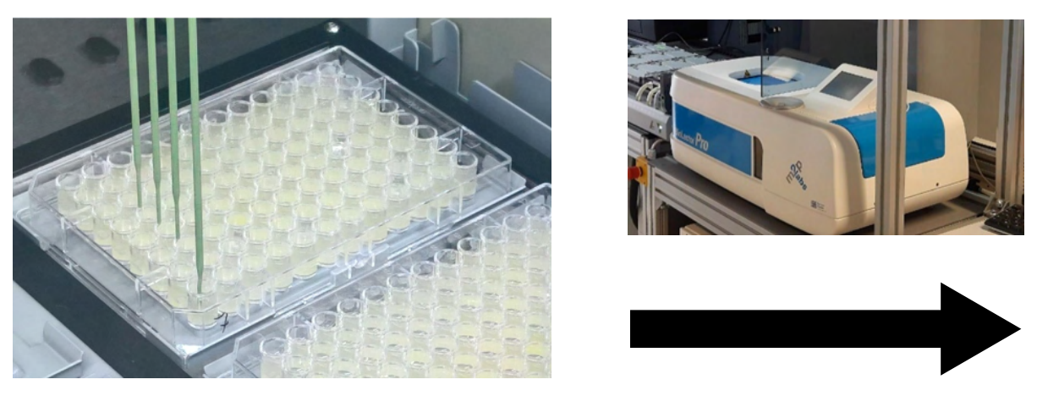
$$\frac{dP}{dt} = q_{P,max} \cdot S \cdot \frac{X}{K_P + S}$$

$$\frac{dS}{dt} = -\frac{1}{Y_{XS}} \cdot \frac{dX}{dt} - \frac{1}{Y_{PS}} \cdot \frac{dP}{dt}$$

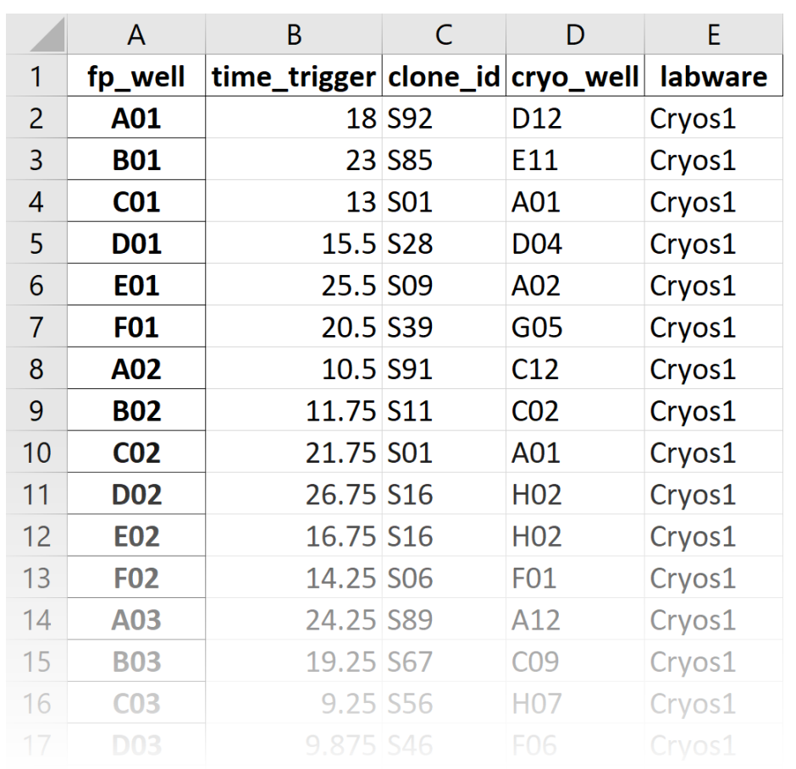
$$X_{0,\text{effective}} = X_{0,\text{alive}} + X_{0,\text{dead}}$$



MBR Batch

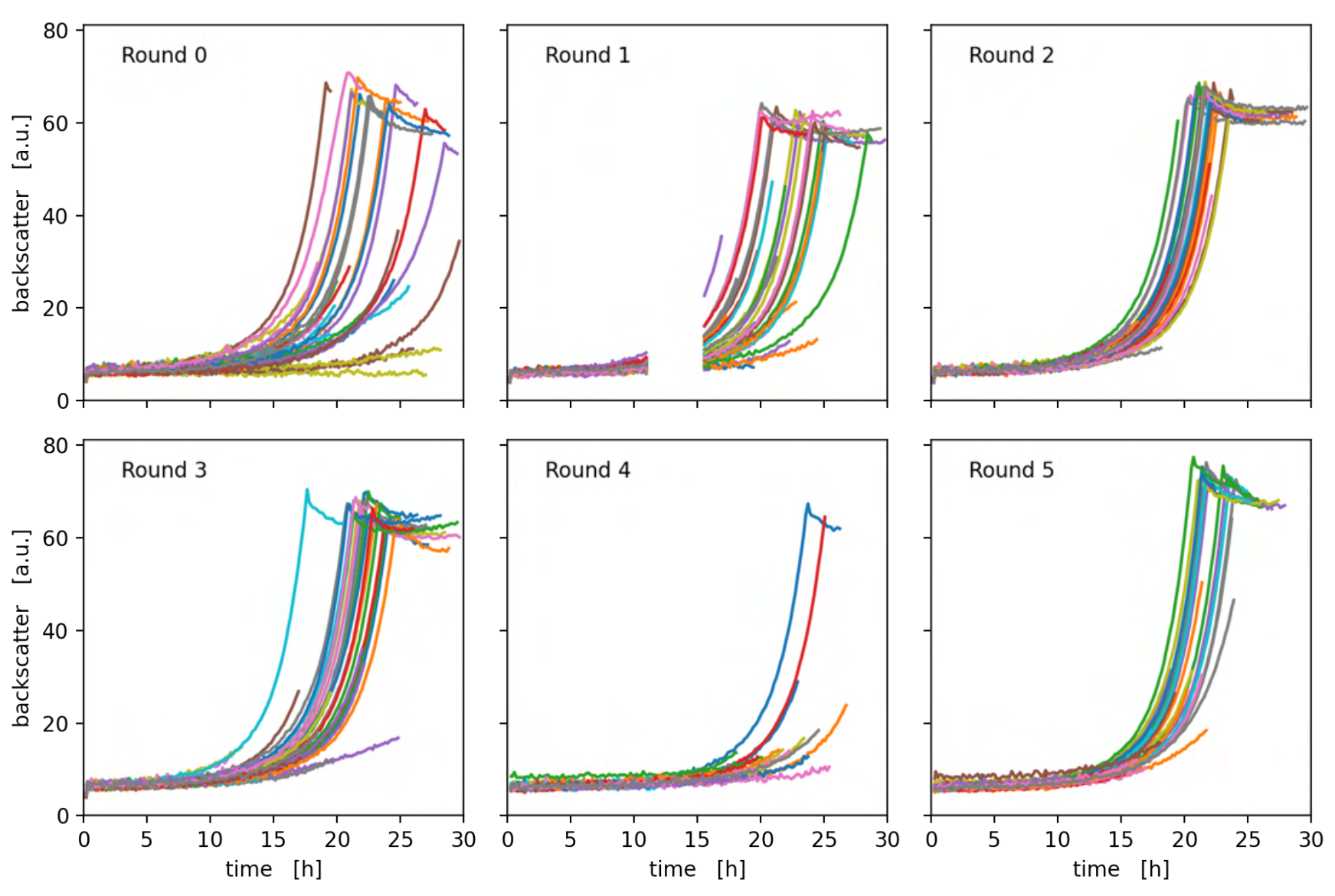


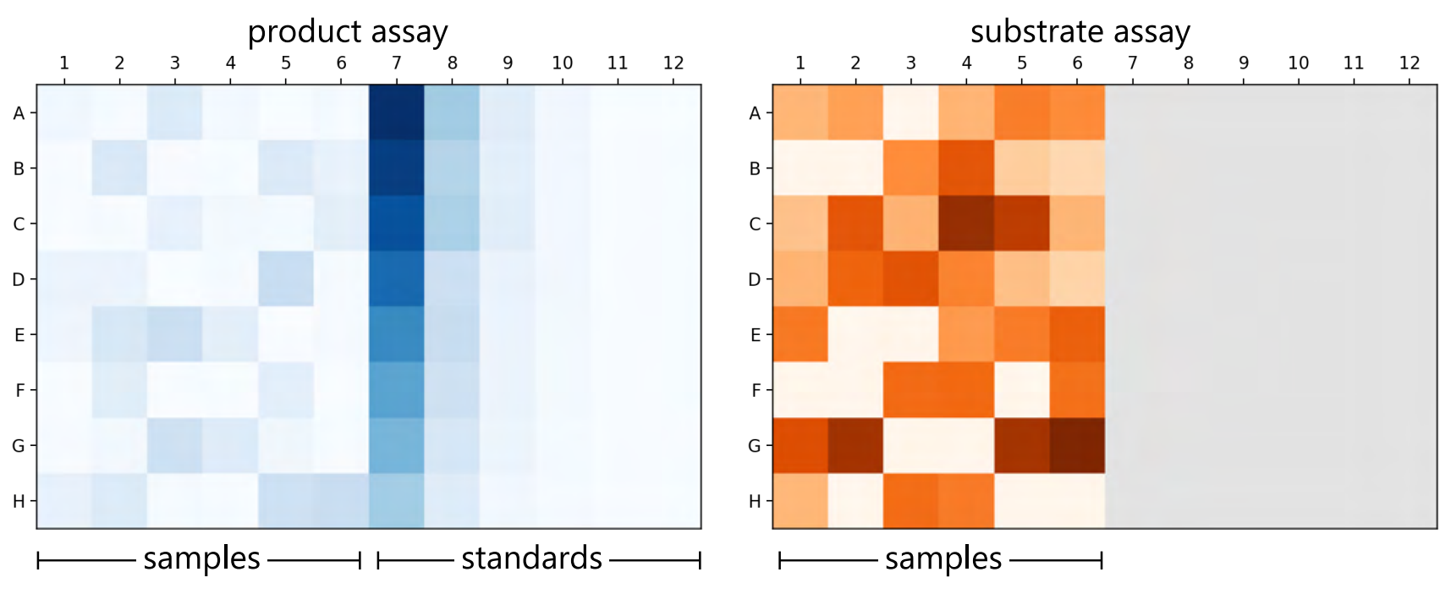
Robotic Inoculation from cryos based on AI-generated experiment design



	A	B	C	D	E
1	fp_well	time_trigger	clone_id	cryo_well	labware
2	A01	18	S92	D12	Cryos1
3	B01	23	S85	E11	Cryos1
4	C01	13	S01	A01	Cryos1
5	D01	15.5	S28	D04	Cryos1
6	E01	25.5	S09	A02	Cryos1
7	F01	20.5	S39	G05	Cryos1
8	A02	10.5	S91	C12	Cryos1
9	B02	11.75	S11	C02	Cryos1
10	C02	21.75	S01	A01	Cryos1
11	D02	26.75	S16	H02	Cryos1
12	E02	16.75	S16	H02	Cryos1
13	F02	14.25	S06	F01	Cryos1
14	A03	24.25	S89	A12	Cryos1
15	B03	19.25	S67	C09	Cryos1
16	C03	9.25	S56	H07	Cryos1
17	D03	9.875	S46	F06	Cryos1

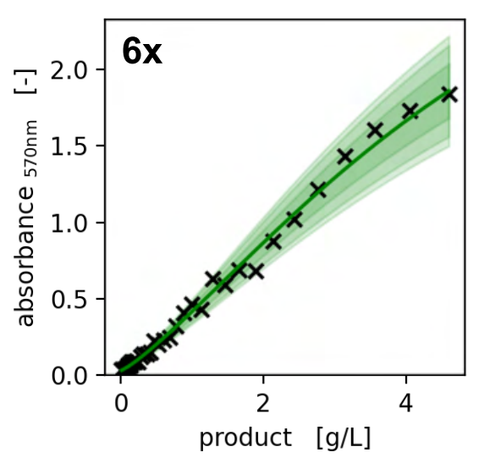
Dataset grows by 48 replicates every round

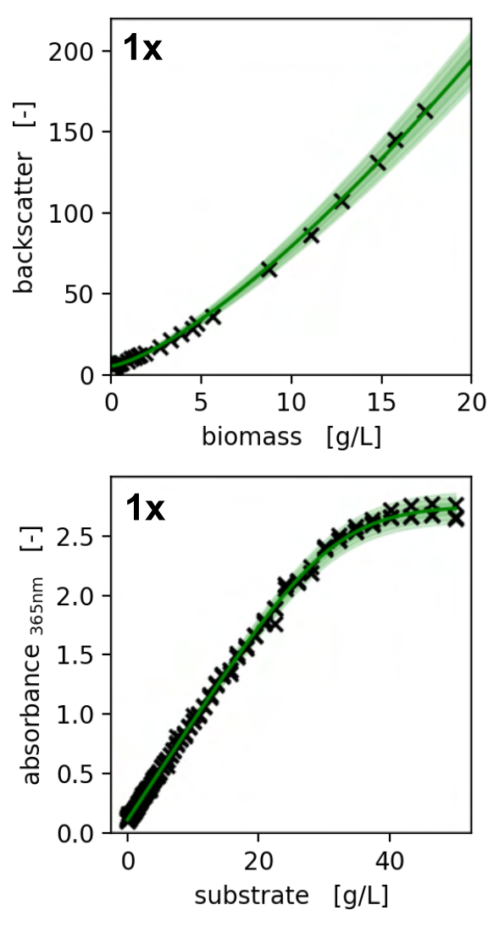




Bayesian Inference

I
- ▶ Aggregate data **from all previous batches**
- ▶ Fit new calibration models

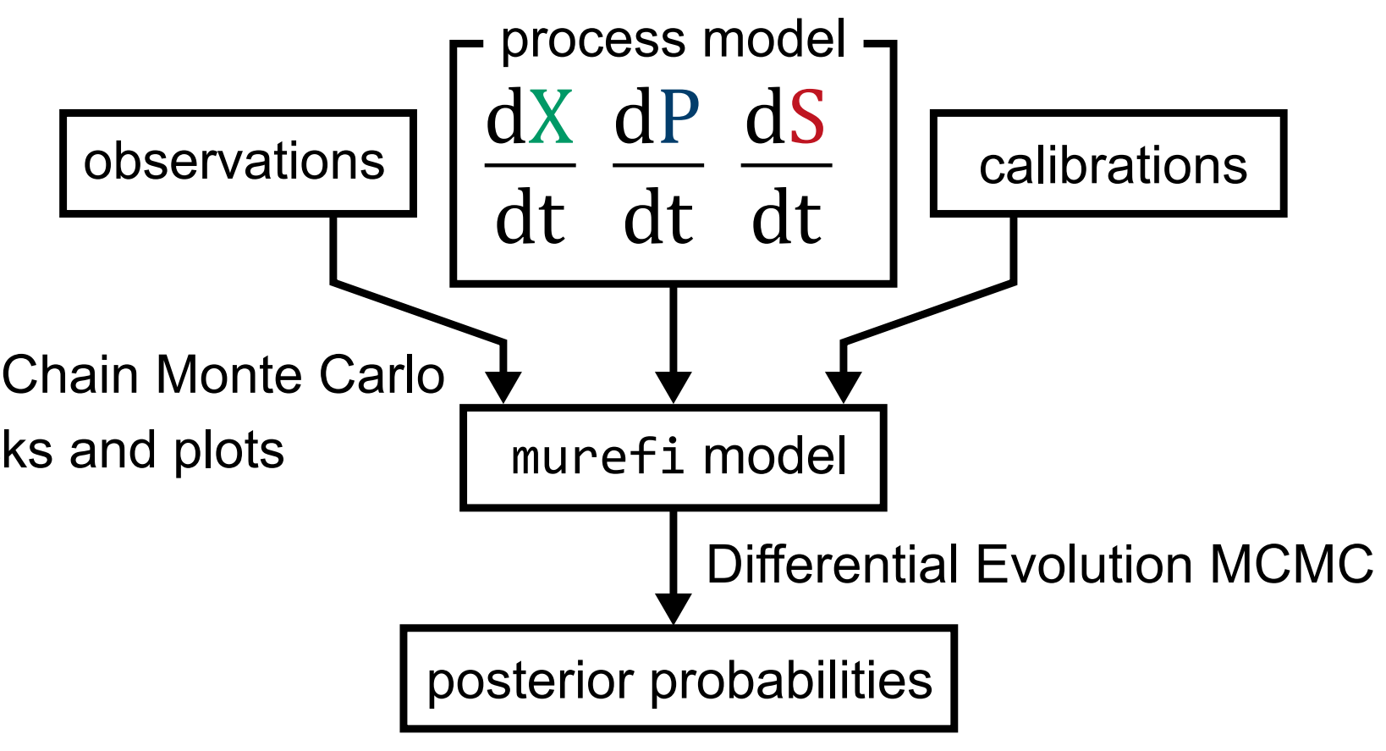




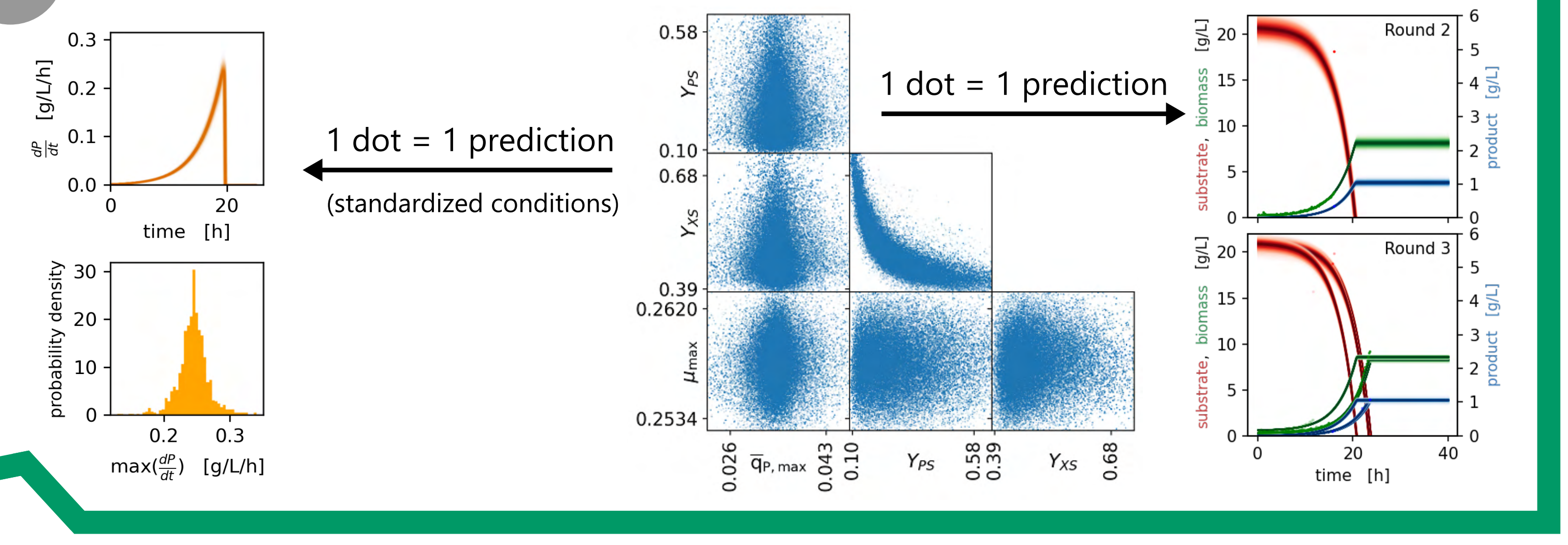
96x

II
- ▶ Build ODE models for all replicates
- ▶ Multi-replicate modeling with **`murefi`**

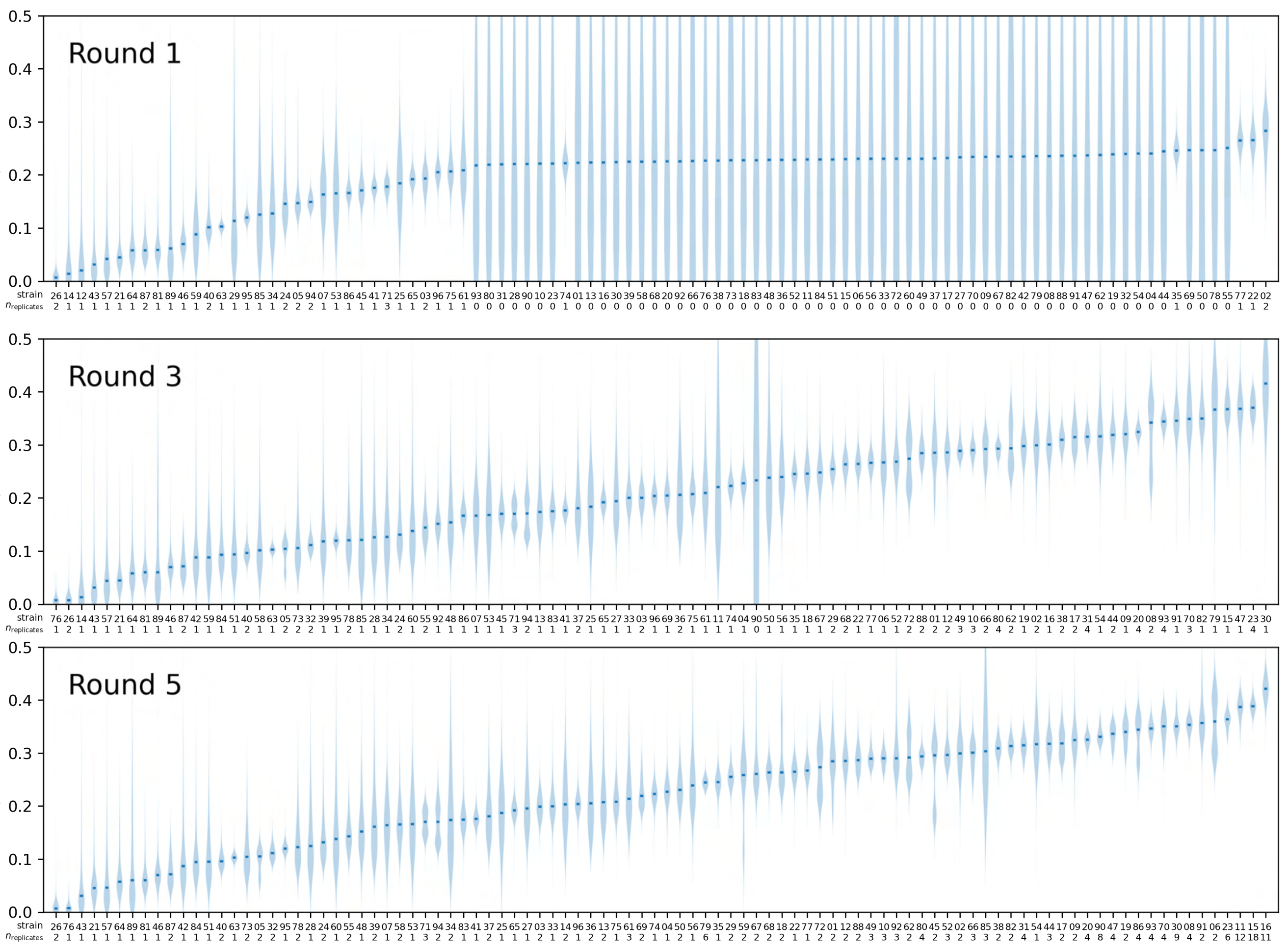
III
- ▶ Fit with Markov-Chain Monte Carlo
- ▶ Diagnostic checks and plots



IV
- ▶ Predict screening metric under standardized conditions



Thompson Sampling



Model **predicts** with high uncertainty for yet unobserved strains.

Replicates for the next round are randomly selected according to their **probability of being the best performer**.

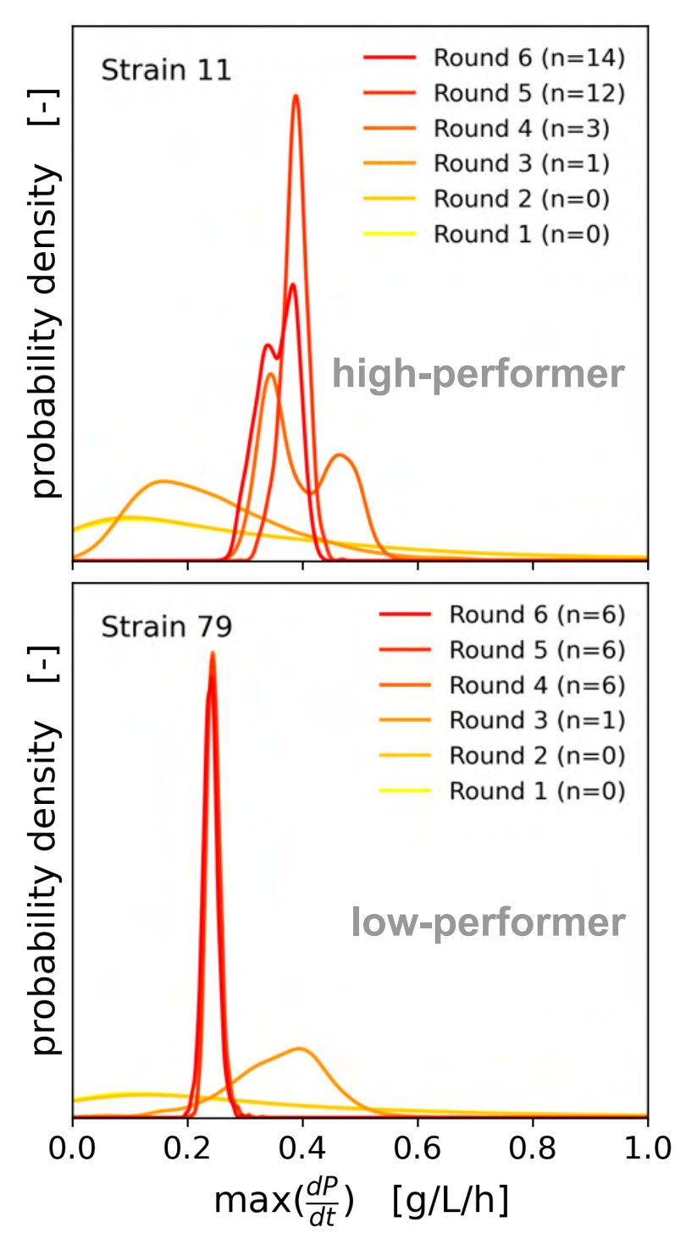
Few replicates are not enough to **distinguish top performers**.

After 5 rounds, the **top performers** were cultivated **~10x more often**.

Few **experimental resources were wasted** on low-performers.

Conclusions

- ✓ **Bayesian optimization characterizes top-performers with more replicates in fewer experiments.**
- ✓ **Human subjectivity in picking candidates for subsequent characterization was removed.**
- ✓ **Thorough quantification of experimental uncertainty enables process modeling with big data sets.**
- ✓ **Generative process modeling delivers predictions of relevant screening metrics.**
- ✓ **Our Python packages `calibr8` + `murefi` enable modelers to scale ODE process models across many replicates and experiments.**


Strain 11
Round 6 (n=14)
Round 5 (n=12)
Round 4 (n=6)
Round 3 (n=1)
Round 2 (n=0)
Round 1 (n=0)
high-performer
Strain 79
Round 6 (n=6)
Round 5 (n=6)
Round 4 (n=6)
Round 3 (n=1)
Round 2 (n=0)
Round 1 (n=0)
low-performer
probability density [-]
max(dP/dt) [g/L/h]