Digital Twin Development for Pichia pastoris Precision Fermentation

Algocell develops hybrid mechanistic-ML digital twins to accelerate bioprocess development. This Case study presents our work on a *Pichia pastoris* process for recombinant protein production, emphasizing design-space definition via *iDoE* (intensified DoE), model calibration/validation, and the biological/technological elements behind the optimized protocols.

- **Objective:** Demonstrate the process of creating and applying a digital twin for yeast cells (*Pichia pastoris*).
- **Target outcome:** Faster, more efficient bioprocess development and optimization

Project Overview

- **Biological system:** *Pichia pastoris* expressing a recombinant protein under AOX1 promoter (methanol induction; mixed carbon strategy considered during exploration and optimization).
- **Application focus:** Precision fermentation protein production in lab scale bioreactors with a view to scale up.
- **Goal:** Calibrate and validate a digital twin that proposes robust, actionable operating protocols (feeding strategies) subject to biological and engineering constraints.

Short Outline of Project Stages

Historical Runs Analysis → *iDoE*-Calibration Experiment Plan → Exploration & Calibration Experiments → Model Calibration & Validation → Optimization Problem Definition → Constraints Definition → Optimal Protocol(s) Generation → Optimized Protocol(s) Execution

Model Calibration and Validation

Biology

- Biomass growth kinetics
- Substrate uptake rates
- AOX1 (induction/repression logic glycerol repressing)
- Secreted protein expression kinetics
- Specific productivity
- Stress responses
- Oxygen limitation/overflow
- Stress impact on per-cell productivity
- Toxic byproduct accumulation

Engineering

- Aeration/agitation capacity
- Harvest/refill dynamics
- Pump rate ranges

Operations

- Feed recipes
- Residual methanol control
- CER as a soft sensor for on-the-fly feed tuning

Assumptions & Bounds

- Methanol induction with minimized glycerol to avoid AOX1 repression

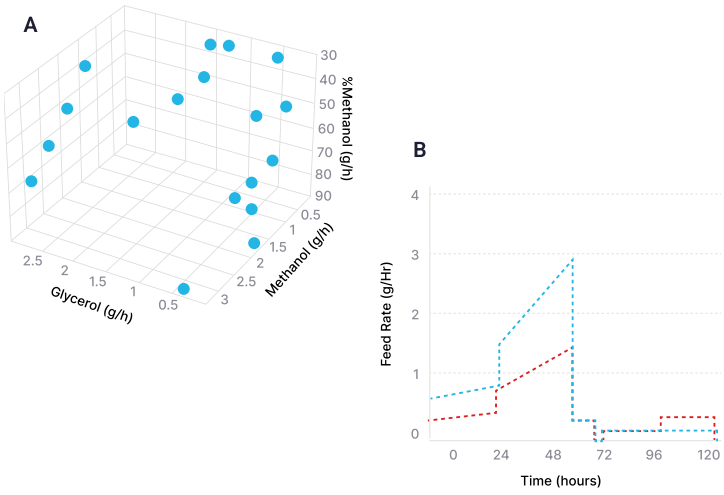

Figure 1: iDoE used during calibration *- methanol/glycerol ratios, Feed rates (mL/h), Feed amounts (g/h) and maintenance range (feed at stationary/ expression phase) were explored.* ***A)*** *In this example, the design space of three variables (%Methanol in feed and Methanol and Glycerol feed rates in g/h) is presented. In this graph, each data point represents an experimental segment from a separate calibration run.* ***B)*** *A graphical representation of the feeding regime (a mix of two carbon sources, Methanol and Glycerol) in one of the calibration experiments. Overall, four calibration experiments were executed, each one composed of four segments.*

Preliminary Data Analysis

Historical experiments were reviewed to select key readouts and parameter ranges for calibration, including biomass, specific productivity trends, and residual inducer control.

Design-Space Definition with *iDOE*

First, process variables such as the feed rate ($mL \cdot h^{-1}$), feed concentration ($g \cdot L^{-1}$), feed composition, and the practical ranges that span the operating envelope were defined. Then, using *iDoE*, we designed a short set of experiments that efficiently maps this entire design space (Fig.1).

Experimental Execution (Exploration/Calibration)

Lab-scale bioreactors with off-gas analysis were used. Proliferation phase preceded methanol-induced production phase. Data captured included pH, DO, off-gas, wet-cell weight (WCW), product titter, and residual methanol.

Within the calibration runs, controlled step-changes swept informative regions (Fig.2), allowing quantification of how cultivation conditions drive the critical process responses:

- Growth rate (μ) during proliferation and under AOX1 induction.
- Protein specific productivity (qP) under AOX1.

The resulting trajectories were used to calibrate the mechanistic layer of the digital twin. Calibrated parameters are validated on held-out segments prior to optimization.

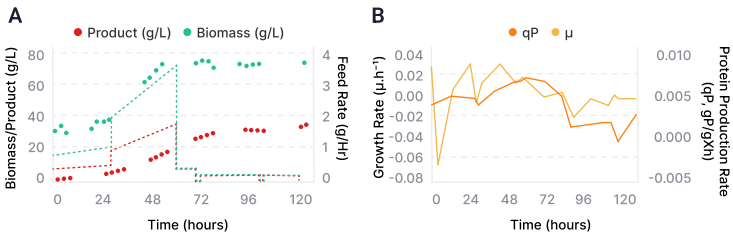

Figure 2: Model calibration (example run). (A) *Empirical measurements of biomass (WCW, g·L⁻¹) and product titter (g·L⁻¹) across the calibration experiment (diamonds); dashed lines trace the run trajectory of the feed schedule, as presented in Fig. 1B.* ***(B)*** *Derived kinetics used for fitting: computed specific growth rate μ (h⁻¹), and protein specific productivity qP (units per axis label), used to calibrate the mechanistic layer of the digital twin model.*

Model Calibration and Validation

We calibrated the mechanistic-ML hybrid model and validated against both an unseen calibration run (Fig.3) and customer historical data (Table 1) before optimization.

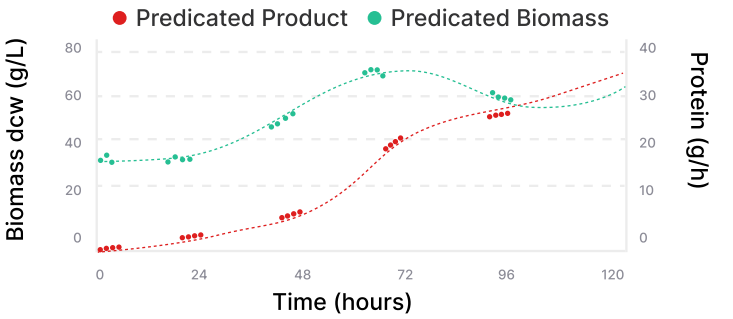

Figure 3: Predicted vs. measured *- biomass and expressed protein predicted by the model, calibrated with data from three calibration runs (dashed lines) aligned with empirical biomass and protein expression measurements (diamonds) from the remaining, un-seen experiment. For statistical accuracy of the model, see Table 1 below.*

Experiment #	Product RMSE [g/L]	Product MAPE	Biomass RMSE [g/L]	Biomass MAPE
Calibration Exp. 1	0.603	9.2%	1.78	4.0%
Calibration Exp. 2	2.24	10.2%	2.86	4.7%
Calibration Exp. 3	1.14	7.25	7.35	10.8%
Historical Data F1	2.03	4.8%	2.56	2.1%
Historical Data F2	1.63	3.8%	0.46	0.3%
Historical Data F3	3.05	8.1%	7.16	7.2%
Historical Data F4	2.69	7.7%	8.73	9.0%
Historical Data F5	7.14	11.2%	9.76	7.2%
Historical Data F6	3.34	8.3%	7.97	8.2%

Table 1: *Model prediction accuracy of the calibrated digital twin on unseen calibration runs (Exp. 1-3) and historical fermentations (F1-F6): Product and biomass RMSE (g/L) and MAPE (%) are presented.*

Optimization Problem Definition

Objectives: Maximize protein titer/ total output and time-normalized productivity while maintaining induction stability and healthy biomass.

Constraints:

- Biological: AOX1-compatible induction (avoid glycerol repression/toxicity), plausible μ bounds from *iDoE*.
- Engineering: pump/feed capabilities; working-volume ceilings.
- Operational: Harvest cadence, sampling workload, and planned duration.

Optimal Protocol Generation

The platform produced two actionable protocols (Fig. 4):

- **Fed-Batch** with stepwise exponential feed transitioning to a stationary hold.
- **Continuous/Harvest** with repeated exponential steps interleaved with partial harvests/refills to sustain growth-production coupling.

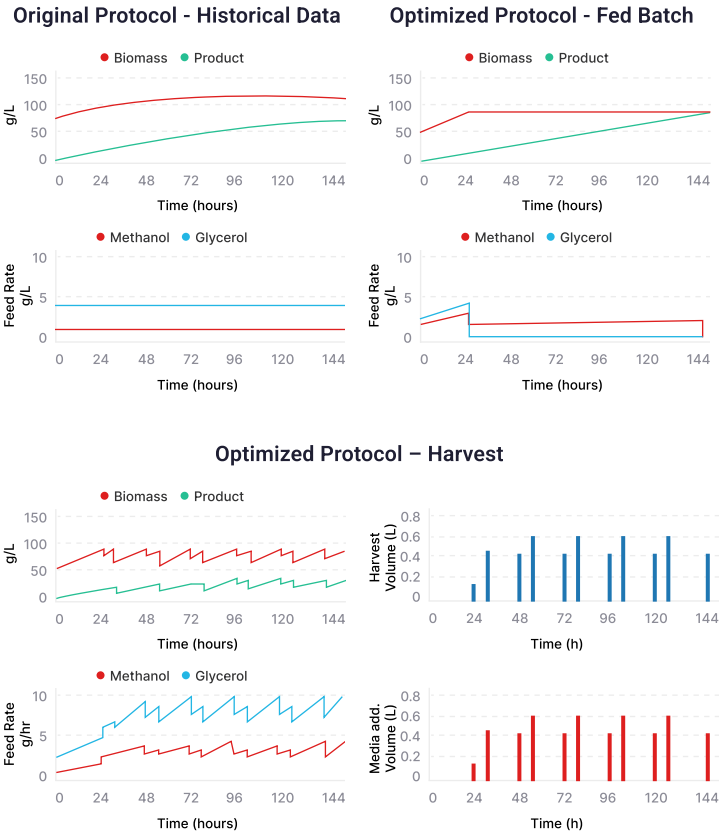

Figure 4: *Side-by-side timelines for Control vs. Fed-Batch vs. Continuous/ Harvest. Predicted protein expression per run increase: 20% in optimized fed-batch, and 40% in harvest protocols relative to control.*

Expected Outcomes

- **Higher yield:** Predicted protein expression per run increase: 20% in Optimized fed-batch, and 40% in harvest protocols relative to control.
- **Lower experimental burden** via *iDoE*, Hybrid modelling, and in-silico simulation.
- **More robust induction/productivity** and smoother scale-up/ tech-transfer.