A Machine Learning Framework for Optimized and Reduced Experimentation in Cell Culture Process Characterization

BUSINESS PROBLEM

Process Characterization (PC) studies are important for demonstrating biologics manufacturing robustness and supporting regulatory filings, but a standard campaign requires ~90 bioreactor runs and months of laboratory capacity. As Amgen's biologics pipeline grows, lab and workforce capacity cannot scale proportionally. This thesis investigates whether hybrid mechanistic machine learning models can reduce the experimental burden of PC studies while maintaining predictive accuracy comparable to Amgen's established JMP regression baseline across multiple product quality attributes (PQAs).

DATA SOURCES

Process Characterization (PC) Dataset – Historical bioreactor run records from prior upstream PC studies, including experimental conditions (CPPs), full time-series trajectories, and product quality attribute (PQA) measurements

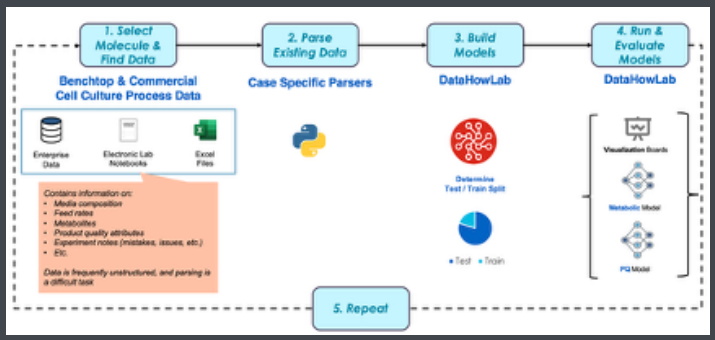
Commercial Process Development (CPD) Dataset – Upstream development experiments (24 runs) used as transfer learning priors, covering a narrower operating range than PC

Data Types and Format

Boolean, Categorical (nominal), Count (integer), Numeric, Static, Time series (Dynamic)

APPROACH

Historical PC and upstream Commercial Process Development (CPD) datasets were ingested via a FAIR data pipeline. Hybrid mechanistic ML models were developed in DataHowLab, integrating mass-balance structure with data-driven learning to predict PQAs. Four training-set selection strategies were compared: Random, Sobol, D-Optimal, and Exterior Point. Transfer learning applied 24 CPD experiments as a prior to reduce PC data requirements. Learning curves were benchmarked against JMP regression. An economic model translated technical findings into site-level and enterprise cost and capacity impact.



John Ryan Zeeman
MIT Leaders for Global Operations, Class of 2026

Leaders for Global Operations

IMPACT

Hybrid models trained on approximately 35 PC experiments, augmented with 24 upstream CPD experiments as a prior, matched or exceeded JMP accuracy for the best-performing PQAs — a 54% reduction relative to the 75-experiment design analyzed, and a 61% reduction relative to a standard 90-run campaign. Learning curves flattened around 35–45 experiments on average, though convergence varied by PQA. Measured time-series inputs outperformed idealized setpoint perturbations, and space-filling sampling strategies (Sobol) outperformed boundary-focused designs. The shared mechanistic propagation backbone was the dominant performance lever over PQA-specific tuning. Conservatively modeled at a 30% experimental reduction, the economic impact at a single Amgen site is approximately $145k in direct annual execution savings and 3.85 FTE-quarters of unlocked scientific capacity. Scaled across 5 major global sites, the estimated enterprise impact is ~$725k/year and ~19 FTE-quarters. The time-to-market NPV benefit is estimated at ~$15M per molecule under conservative assumptions, given the value of accelerating regulatory filing by approximately 30–45 days.

DRIVERS

Amgen's biologics pipeline is growing with IND filings expected to rise substantially by 2032, increasing PC study demand. Standard campaigns require ~90 bioreactor runs, and lab capacity cannot scale proportionally. The high opportunity cost of resource-intensive PC created urgency for more data-efficient approaches.

BARRIERS

PC datasets are inherently small (30–90 runs per molecule), limiting ML training data. Regulatory requirements demand rigorous process justification, raising the bar for any reduced-experiment approach. Historical data was generated using central composite designs rather than space-filling designs optimized for hybrid model learning.

ENABLERS

Amgen's historical PC and CPD datasets provided a rich retrospective foundation. The DataHowLab platform offered an integrated hybrid modeling environment. Embedding within the TDC team provided domain expertise and stakeholder access. Agile software practices enabled reproducible, modular model development and lineage tracking.

ACTIONS

Built reusable Python API utilities for data parsing, model training, and evaluation. Evaluated four training-set selection strategies (Random, Sobol, D-Optimal, Exterior Point) to generate per-PQA learning curves. Developed and tuned hybrid ML-mechanistic models in DataHowLab with iterative lineage tracking. Applied transfer learning from CPD data and benchmarked against JMP regression. Built a parameterized economic model of business impact.

INNOVATION

This is the first empirical evaluation of hybrid ML-mechanistic models for reducing PC experimental burden on industrial CHO cell culture data collected under regulatory-grade conditions. Prior work was limited to academic-scale fermentation. The framework jointly addresses experimental design strategy, transfer learning from upstream data, and learning curve quantification — grounded in real process data rather than simulated results.

IMPROVEMENT

Hybrid models trained on ~35 PC experiments plus 24 CPD experiments matched or exceeded JMP accuracy for the best-performing PQAs — a 54% reduction vs. a 75-experiment design and 61% vs. a standard 90-run campaign. A conservative 30% workload reduction yields ~$145k in annual direct savings and 3.85 FTE-quarters of unlocked capacity at a single site, scaling to ~$725k and ~19 FTE quarters across five global sites.

BEST PRACTICES

Ask questions to surface domain knowledge not captured in data. Track model lineage throughout iterative development to support reproducibility and knowledge transfer. Build modular, well-tested code from the start to enable flexible reuse across configurations. Align modeling and experimental design choices jointly at project outset.

OTHER APPLICATIONS

Applicable to other biopharma PC activities where experimental capacity is constrained, and historical data exists, including downstream purification and scale-up prediction. Transfer learning approach extends to new molecules within the same cell line platform. Scheduling framework applies to organizations running multiple concurrent PC campaigns.