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

by

John Ryan Zeeman
B.S., Chemical Engineering
Northwestern University, 2021

Submitted to the MIT Sloan School of Management and
Department of Chemical Engineering
in partial fulfillment of the requirements for the degrees of

Master of Business Administration

and

Master of Science in Chemical Engineering

in conjunction with the Leaders for Global Operations program
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
May 2026

Author .....
MIT Sloan School of Management and
Department of Chemical Engineering
May 8, 2026

Certified by .....
Mohammad Fazel Zarandi, Thesis Supervisor
Senior Lecturer and Research Scientist, Operations Management

Certified by .....
J. Christopher Love, Thesis Supervisor
Raymond A. (1921) and Helen E. St. Laurent Professor of Chemical Engineering

Accepted by .....
Hadley Sikes
Willard Henry Dow Professor of Chemical Engineering

Accepted by .....
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

THIS PAGE INTENTIONALLY LEFT BLANK

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

by

John Ryan Zeeman

Submitted to the MIT Sloan School of Management and
Department of Chemical Engineering
on May 8, 2026, in partial fulfillment of the
requirements for the degrees of
Master of Business Administration
and
Master of Science in Chemical Engineering

Abstract

Process characterization (PC) studies are important for demonstrating the robustness of biologics manufacturing processes, but they are resource-intensive, requiring extensive factorial experiments that consume laboratory capacity and extend timelines. As AMGEN's biologics pipeline grows, laboratory and workforce capacity cannot scale proportionally, creating urgency for more data-efficient approaches. This thesis investigates whether hybrid mechanistic machine learning models can reduce the experimental burden of PC studies while maintaining predictive accuracy. Using hybrid models developed via the DataHowLab software, we benchmark against a standard JMP regression approach across multiple product quality attributes (PQAs). The central finding was that hybrid models trained on approximately 35 PC experiments, augmented with 24 upstream commercial process development 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 behavior varied by PQA. Additional findings include that measured time-series inputs outperformed idealized setpoint perturbations, space-filling sampling strategies outperformed boundary-focused designs, and the shared mechanistic backbone was the dominant performance lever over PQA-specific tuning. Even conservatively, a 30% reduction in experimental workload could translate to meaningful cost savings and freed up capacity across AMGEN's network.

Thesis Supervisor: Mohammad Fazel Zarandi
Title: Senior Lecturer and Research Scientist, Operations Management

Thesis Supervisor: J. Christopher Love
Title: Raymond A. (1921) and Helen E. St. Laurent Professor of Chemical Engineering

Acknowledgments

I would like to thank my MIT thesis advisors, Dr. Mohammad Fazel-Zarandi and Dr. J. Christopher Love, for their support, guidance, flexibility, and technical expertise throughout both my internship and the thesis writing process. I am also grateful to Dr. Harini Narayanan and Dr. Connor Coley at MIT for their valuable input and encouragement.

I am deeply appreciative of AMGEN and my manager, Elçin Içten Gençer, for her guidance and the opportunity to work on this project. I also thank the Process Development Transformative Digital Capabilities (TDC) team and the AMGEN Massachusetts process development team for their collaboration, feedback, and day-to-day support, in particular, Miguel Valderrama Gomez, Austin Xiong, Bill Heyman, Juliane Glaser, Ivy Huang, Fabrice Schlegel, and Pablo Rolandi. I would also like to recognize Chris Garvin, Michaela Murr, Phillipp Simons, and the LGO community at AMGEN for their support and guidance.

To my fellow AMGEN LGO interns: Isabel, Kendall, and Jeremy, our weekly check-ins and continuous encouragement were such a joy, and one of the true highlights of this experience. I am forever grateful to be part of the LGO family and wish to acknowledge all 44 of my LGO classmates, as well as my peers in the MBA program at the Sloan School of Management, for their inspiration.

To all of my professors, teachers, and mentors (both academic and professional), I express my thanks for continually encouraging me to pursue my interests, to ask questions, and to push myself. In particular, my undergraduate research advisor, Dr. Jeffrey Richards, deserves recognition for encouraging me to pursue research early in my academic career.

To my friends in Chicago and from Northwestern: thank you for years (and in some cases, decades) of friendship. Moving away was hard, and I miss you all. I'm also grateful for the community that made Boston feel like home, especially fellow Northwestern alumni (go 'Cats!), and my local extended family. It has been a privilege to spend more time together.

Finally, I owe my deepest gratitude to my parents, Martha and Greg Zeeman, my siblings, Caroline and Brian, my honorary big sister Katie, my extended family, and all my dear friends. Their unwavering support and love have sustained me in every pursuit. It is an incredible privilege to be at MIT, and I would not be here without the many people who have believed in me along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter motivates the research undertaken in this thesis by situating it within the operational and strategic context of AMGEN's growing biologics pipeline. It begins by describing the business pressures driving the need for more efficient process development, then formalizes the core problem: that current Process Characterization study designs require a large number of bioreactor experiments whose value diminishes as more are conducted. The chapter defines the scope of the project, which focuses on retrospective evaluation of hybrid machine learning–mechanistic models applied to upstream cell culture PC data, and articulates the central hypothesis that these models can maintain predictive accuracy comparable to established statistical approaches while training on substantially fewer experiments. The chapter concludes with a summary of the four project goals and an overview of the thesis organization.

1.1 Business Context: Growing Capacity and Pipeline

Over the past decade, AMGEN and the broader biopharmaceutical industry have expanded both clinical pipelines and manufacturing networks, increasing the number of molecules and modalities that must be advanced from development into reliable commercial supply. For mammalian cell culture products, this growth increases demand on process development organizations to generate robust process understanding, establish control strategies, and ensure consistent product quality across scales and

sites. According to publicly available investor guidance, AMGEN projects revenue growth at a mid-single-digit CAGR through 2030, with IND filings expected to increase substantially by 2032, driving a corresponding rise in process development and characterization demand.[1]

In parallel, AMGEN has made major capital investments to expand capacity and improve operational performance (e.g., throughput, yields, and supply resilience) in support of pipeline growth and indication expansion. Recent examples include a $1B expansion to establish a second drug substance manufacturing facility in Holly Springs, North Carolina,[2] a $900M manufacturing expansion in Ohio,[3] and a $650M expansion of its Puerto Rico biologics manufacturing campus with stated integration of advanced technologies.[4] These investments underscore a strategy of growing capacity while improving manufacturing performance, which in turn elevates the value of approaches that accelerate late-stage process understanding and de-risk scale-up.

Process Characterization (PC) studies are a critical component of this transition from late-stage development into commercialization. They support the definition of proven acceptable ranges (PARs), inform regulatory filings, and de-risk manufacturing performance. However, PC studies also consume scarce experimental capacity: bioreactor time, analytical throughput, and engineering effort. As pipelines grow, the opportunity cost of long, resource-intensive PC campaigns increases, motivating approaches that preserve rigor while reducing experimental burden. A typical PC campaign requires on the order of 90 bioreactor runs when accounting for replicates and supplementary one-factor-at-a-time runs. This represents a process that spans months and consumes a significant share of available laboratory capacity.

[1] https://www.amgen.com/newsroom/press-releases/2026/02/amgen-reports-fourth-quarter-and-full-year-2025-financial-results

[2] https://www.AMGEN.com/newsroom/press-releases/2024/12/AMGEN-announces-%241-billion-manufacturing-expansion-in-north-carolina

[3] https://www.AMGEN.com/newsroom/press-releases/2025/04/AMGEN-announces-900-million-manufacturing-expansion-creation-of-350-new-jobs-in-ohio

[4] https://www.AMGEN.com/newsroom/press-releases/2025/09/AMGEN-announces-650m-expansion-of-us-manufacturing-creating-hundreds-of-new-jobs

1.2 Problem Statement

Process Characterization is typically executed using classical Design of Experiments (DoE) methods intended to map the relationship between controllable process inputs (CPPs) and process outputs (e.g., performance metrics and product quality attributes). In practice, the number of candidate experiments grows quickly with the number of factors and levels.

For example, consider a central composite design (CCD) over four variables of interest. A CCD includes a center point, axial points ($+a$ and $-a$), and factorial points ($+1$ and -1) for each variable. If one were to enumerate all combinations of five levels per factor, the full grid would contain $5^4 = 625$ combinations. CCDs strategically sample this space and typically reduce a four-factor design to on the order of $\sim$30 runs. In industrial settings, these designs are often augmented with replicates (to manage experimental failures and improve uncertainty estimates) and supplemented with one-factor-at-a-time (OFAT) runs for troubleshooting, edge-case exploration, and follow-up questions. In aggregate, a single PC study may require on the order of $\sim$90 bioreactor runs.

This experimental burden constrains throughput. The core technical problem addressed in this thesis was how to reduce the number of experiments required to achieve comparable predictive and decision-making value—while maintaining the level of rigor expected for process understanding and, ultimately, regulatory justification. This thesis therefore asks: can hybrid machine learning–mechanistic models, trained on a systematically selected subset of experiments, match the predictive accuracy of current statistical approaches for cell culture process characterization?

1.3 Project Scope

This project focuses on in-silico modeling of mammalian cell culture processes within the Process Characterization (PC) phase of commercialization. The research leverages historical data from AMGEN's bio-manufacturing processes to develop and evaluate

hybrid machine learning–mechanistic models, and to study how alternative training-set selection strategies impact predictive performance.

Specifically, the project scope includes:

- Data ingestion and standardization: Aggregation and parsing of structured datasets from prior PC studies and Commercial Process Development (CPD) campaigns, with an emphasis on consistency, traceability, and alignment with FAIR (Findable, Accessible, Interoperable, and Reusable) principles.
- Model development: Construction of hybrid models using the DataHowLab framework, integrating first-principles constraints (e.g., mass balances) with data-driven learning to predict bioreactor performance and product quality attributes.
- Experimental design evaluation: Comparative analysis of training-set selection strategies, including Random Sampling, Sobol sequences, and D-Optimal designs, to quantify how predictive accuracy scales with the number of training experiments.
- Transfer learning application: Evaluation of whether upstream development data (CPD) can reduce the data requirements for late-stage characterization by informing model training and generalization.

This work was computational and retrospective; it does not include the execution of new wet-lab experiments. Instead, historical datasets were used to simulate and assess alternative experimental design and modeling strategies.

1.4 Project Goals and Hypothesis

The performance of hybrid ML models was unknown in this use case, as was its ability to meet an acceptance threshold and be implemented in practice. Ultimately, we sought to answer the question posed in Figure 1-1. Does machine-learning enable us to move more quickly in Process Characterization? This question can be further broken down into a few sub-questions:

- For existing PC studies, how many experiments are needed to accurately predict product quality attributes (PQAs)?
- How does the model error scale with the size of the training set?
- Does the model ever meet a threshold where this approach could be used proactively instead of retrospectively?

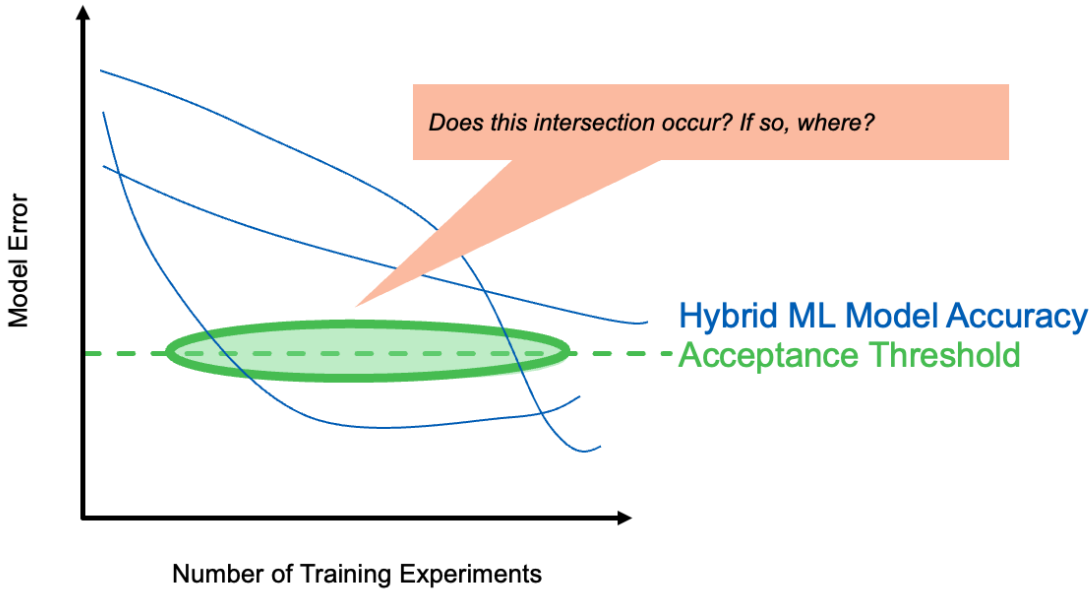

Figure 1-1: Potential performance curves of hybrid ML models in Process Characterization.

1.4.1 Hypothesis

The central hypothesis of this research was that machine learning–augmented hybrid models can reduce the experimental burden of Process Characterization studies while maintaining predictive rigor comparable to current DoE-based approaches, specifically, achieving product quality attribute prediction error within an acceptable margin of the established JMP multiple linear regression baseline. In particular, by incorporating prior knowledge through mechanistic structure and leveraging transfer learning across related processes, predictive models can be trained on a subset of the experiments typically required to support PC decision-making.

1.4.2 Project Goals

To test this hypothesis and deliver business value to AMGEN, this project pursued the development of four objectives:

1. Quantifiable efficiency gains: Demonstrate a measurable reduction in experimental workload for future PC studies. The target was to identify a framework enabling a 30% decrease in experimental resources (bioreactor runs) while preserving predictive performance comparable to established statistical baselines.

2. A sustainable digital framework: Develop a reusable and reproducible software framework to support model training, evaluation, and adaptive training-set selection. This includes robust data pipelines and modeling utilities aligned with agile software engineering practices, enabling practical adoption within Process Development.

3. Increased predictive power and knowledge transfer: Use hybrid ML + mechanistic models to better forecast product quality attributes (such as Titer, metabolic attributes, charge variant profiles, size variant/aggregation profiles) earlier while applying transfer learning so insights from past molecules inform new ones, which cuts redundancy across products and scales.

4. A path toward regulatory innovation: Providing recommendations to a modernized regulatory strategy by using ML-aided DoE to build robust, evidence-based process validation.

1.5 Thesis Organization

The remainder of this thesis is organized into five chapters, structured to guide the reader from foundations to implementation and results, and finally to business and organizational implications.

Chapter 2 provides background and a literature review. It introduces the role of PC within biopharmaceutical development, summarizes classical DoE approaches

used in practice, and reviews modeling methods relevant to this work, including hybrid semi-mechanistic models, transfer learning, and active learning / adaptive experimental design.

Chapter 3 presents the methodology. It describes the end-to-end data pipeline, software and reproducibility practices, training-set selection strategies (Random, Sobol, and D-Optimal), and the hybrid modeling approach implemented within the DataHowLab framework.

Chapter 4 reports results. It benchmarks hybrid-model performance against statistical baselines, quantifies how predictive accuracy varies with training-set size and selection strategy, and evaluates transfer learning using CPD data to reduce experimental requirements in later-stage characterization.

Chapter 5 discusses implications. It interprets scientific insights from model successes and limitations, proposes practical directions for future PC study design, estimates potential economic and capacity impact at AMGEN, and addresses change management and regulatory considerations for implementation.

Chapter 6 concludes with a summary of findings and recommendations for AMGEN's process characterization strategy, and identifies future research directions, including extensions to other unit operations and potential applications to bioreactor monitoring and control.

Chapter 2

Background and Literature Review

This chapter provides the technical and industry context necessary to understand the methods and results presented in later chapters. It begins with an overview of the biopharmaceutical industry and biologics manufacturing, with particular attention to the upstream cell culture process, scale-up challenges, and AMGEN's position within the landscape. It then describes the role and structure of Process Characterization studies, followed by a review of classical and alternative experimental design approaches used in bioprocessing. The chapter surveys the modeling methods most relevant to this work, including linear regression baselines, mechanistic mass-balance models, and hybrid architectures that combine first-principles structure with data-driven learning, and reviews adjacent topics in transfer learning, active learning, Bayesian experimental design, and sequential optimization. The chapter closes by identifying the specific research gap this thesis addresses: the lack of empirical evidence on how predictive performance scales with training set size in industrial PC settings, and whether informed experiment-selection strategies can meaningfully reduce the experimental burden while preserving decision-making utility.

2.1 A Brief Overview of the Biopharmaceutical Industry

2.1.1 Industry Overview

The biopharmaceutical industry plays a central role in global public health and is characterized by long development timelines, high technical risk, and substantial capital requirements. Developing a new prescription drug can cost more than $2 billion and take over 10 years to reach the market [1]. As pipelines grow and modalities diversify, manufacturers face increasing pressure to improve development throughput and reduce the marginal cost of advancing each program. In parallel, the industry is actively pursuing digital transformation and advanced process development strategies to unlock manufacturing cost reductions and accelerate timelines [2]. Competitive advantage increasingly comes from enabling therapeutic innovation while manufacturing at scale in a financially sustainable manner [3].

From a process development perspective, these pressures are expressed through constrained laboratory capacity and the need to generate robust process understanding on tighter timelines. This thesis focuses on one specific bottleneck within that context: PC studies for mammalian cell culture processes.

2.1.2 Biologics Manufacturing

Biologics manufacturing produces therapeutic proteins using living cells, resulting in processes that are high-dimensional, time-varying, and sensitive to both operating conditions and biological variability. The manufacturing train is commonly divided into upstream (cell culture) and downstream (purification) processing [4]. Upstream fed-batch cell culture is particularly challenging because the system evolves over time (cell growth and death, metabolite accumulation, and productivity shifts), while measurements are often sparse, noisy, and heterogeneous across studies.

2.1.3 The Production Bioreactor

The core unit operation for upstream processing is the bioreactor, typically a stirred-tank vessel equipped with impellers for agitation and spargers for gas delivery. Within this vessel, the physical environment must be strictly controlled to support cell growth; critical variables including pH, dissolved oxygen (DO), temperature, and agitation rate are regulated via feedback loops. However, the physical dynamics of these vessels—specifically fluid dynamics, shear stress distributions, and gas mass transfer ($k_L a$)—are highly dependent on vessel geometry and size.

2.1.4 Scale-Up Challenges

A major challenge in process development is scaling a cell culture process from early-stage vessels (e.g., cryovials, shake flasks, and bench bioreactors) to production-scale bioreactors. This progression is typically staged through a seed train, where cells are expanded across a sequence of increasing volumes, before inoculating the production bioreactor for a fed-batch run. At each scale, engineers must ensure that the process remains robust and that performance and product quality remain within acceptable bounds.

In practice, the operational risk of scale-up is not limited to physics alone. Mammalian cell culture processes are sensitive to both controllable operating conditions and biological variability, and small changes in the bioreactor environment can influence growth, metabolism, productivity, and final product quality. As a result, late-stage development places significant emphasis on de-risking manufacturing operations by explicitly testing process robustness across a feasible operating region.

2.1.5 AMGEN Inc.

AMGEN Inc. is one of the early pioneers of the modern biotechnology industry. The company was founded in 1980 (originally as Applied Molecular Genetics, Inc.) and has been headquartered in Thousand Oaks, California since its founding.[1] Today,

[1] https://www.AMGEN.com/about/AMGEN-history

AMGEN operates a global manufacturing network and multiple upstream production platforms, including fed-batch processes in stainless-steel bioreactors and single-use technologies [1, 5].

In parallel with broader industry trends, AMGEN is entering a period of sustained portfolio expansion, supported by a growing biologics pipeline and increasing expected demand across indications. To support this growth, AMGEN has made significant investments in manufacturing capacity and capability, including major expansions across multiple sites. These investments reflect a strategy of scaling supply while improving manufacturing throughput and reliability, which elevates the value of approaches that accelerate late-stage process understanding and de-risk operations. This context aligns with the motivation described in Chapter 1: as the number of programs increases, laboratory and workforce capacity cannot expand proportionally, and efficiency gains in late-stage characterization become increasingly consequential.

2.1.6 Machine Learning at AMGEN

AMGEN has explored machine learning to improve operations across the business, including applications in process development. Within the Process Development Transformative Digital Capabilities (TDC) team in particular, there is a goal to create digital twins across the upstream manufacturing process, connecting data infrastructure, predictive models, and simulation capabilities. This vision is summarized in Figure 2-1.

Prior LGO theses and academic literature describe related applications of ML and hybrid modeling across process development, including:

- **Scale-Up Prediction:** Using hybrid models to better predict commercial-scale performance from small-scale data, reducing the reliance on purely heuristic scaling rules (like P/V) [4].
- **Knowledge-Constrained ML:** Strategies to build predictive process models even when datasets are limited by incorporating domain knowledge into learning [6].

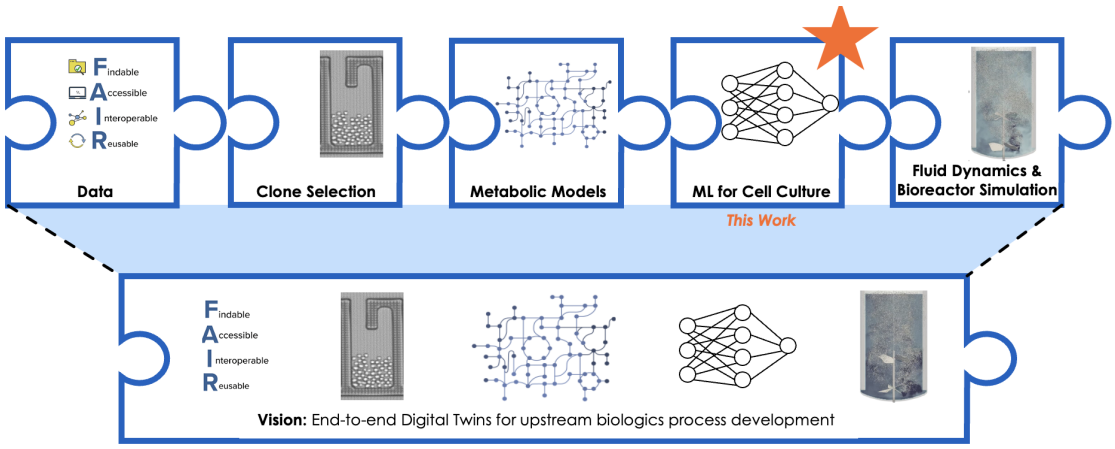

Figure 2-1: Illustrative view of the upstream digital-twin stack pursued within AMGEN Process Development, linking data accessibility (FAIR principles), mechanistic and metabolic modeling, machine-learning-based prediction for cell culture, and higher-fidelity simulation. The focus of this thesis is on predictive modeling for process characterization and the experimental-design question of how many PC runs are required to achieve acceptable prediction accuracy.

- **Model Predictive Control (MPC):** ML-based MPC (including neural network and Gaussian process controllers) to optimize feeding strategies and improve cell growth and metabolite profiles [5].

- **Multi-Scale Modeling:** Frameworks integrating genome-scale representations with dynamic simulation (e.g., COSMIC-dFBA) to predict metabolic dynamics in bioreactors [7].

- **Clone Selection and Anomaly Detection:** Using computer vision models (autoencoder, CNN, and SVM) applied to fluorescent cell imagery, Albright (2025) automated anomaly identification during clone screening in Cell Line Development, reducing manual review burden and standardizing candidate selection ahead of downstream development [8].

This thesis is differentiated by its emphasis on PC study efficiency: rather than optimizing a control policy or scale-up prediction alone, it evaluates whether hybrid predictive models can maintain decision-relevant accuracy when trained on fewer PC experiments, and how training-set selection policies affect learning curves.

2.2 Process Characterization in Biologics Manufacturing

Process Characterization (PC) is a late-stage process development activity intended to demonstrate process robustness and support definition of acceptable operating ranges. In practice, PC studies quantify relationships between critical process parameters (CPPs) and critical quality attributes (CQAs) to support control strategies and regulatory filings [1]. Classical PC execution often emphasizes statistically powered designs for estimating factor effects and interactions, but these designs can be resource-intensive when multiple factors, levels, and follow-up questions are required.

Regulatory guidance has increasingly been explicit that advanced modeling, including AI/ML, may support regulatory decision-making if model credibility is established using an appropriate, risk-based framework and the model is shown to be fit-for-use [9]. In parallel, emerging approaches such as calibration design propose that PC can be streamlined by transferring knowledge from historical products, potentially reducing experimental effort substantially relative to traditional screening designs [10]. These developments motivate the central question of this thesis: what experimental budget is required to obtain predictive utility comparable to current PC approaches, and what design policies best allocate that budget?

A generally accepted PC study design used by AMGEN can be visualized in Figure 2-2. As mentioned in Section 1.2, this includes a large volume of experiments to conduct. In practice, these campaigns are executed in blocks and are coupled to the broader scale-up timeline, since bioreactor runs must be scheduled around seed-train capacity, analytical throughput, and development priorities. Each block is intended to both (i) stress the process across a range of controllable inputs and (ii) generate the statistical evidence needed to justify operating ranges and control strategies. When outcomes indicate unexpected sensitivities or failure modes, follow-up experiments are frequently required, further extending timelines. Bayer et al. (2020) proposed hybrid modeling paired with intensified DoE as a direct approach to accelerating upstream process characterization, demonstrating on lab-scale *Escherichia*

coli data that mechanistic structure can reduce the number of experiments required to achieve comparable predictive coverage. Their work establishes proof-of-concept for the efficiency gains this thesis evaluates on industrial CHO PC data [11].

This structure highlights why PC is both essential and expensive: it is a de-risking activity that translates process understanding into defensible operating condition bounds, but it consumes scarce bioreactor and analytical capacity at a time when the organization is simultaneously advancing many programs toward commercialization. Consequently, any approach that preserves the decision-making value of PC while reducing the required number of runs can have outsized impact on overall process development throughput.

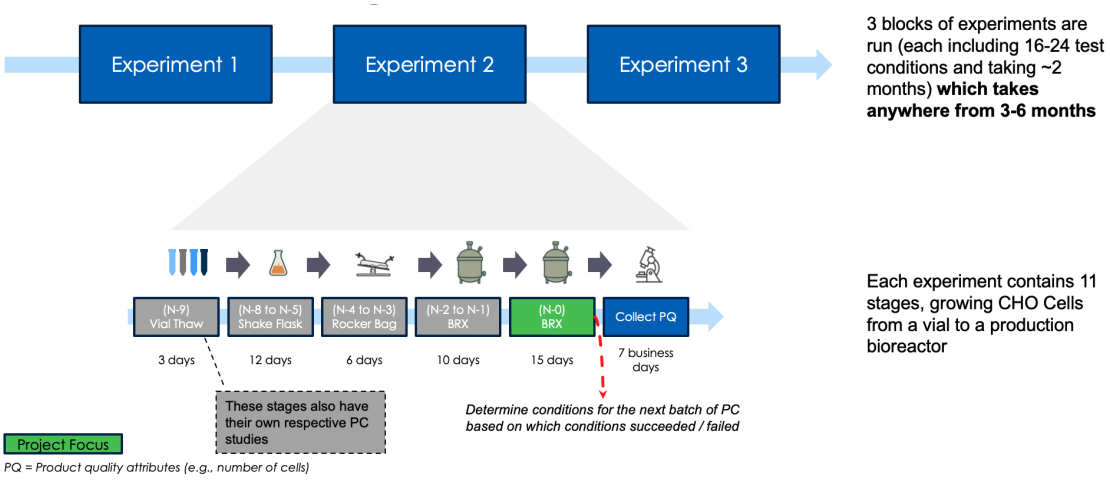

Figure 2-2: Depiction of the Generally Accepted PC Study Design in Biopharma, with example stages of expansion and approximate timelines. This shows the blocked nature of experiments, where 3 large studies are run, each with a set of conditions that are sampled.

2.3 Experimental Design in Bioprocessing

2.3.1 Classical DoE and Factorial Designs

The industry standard for quantifying relationships between CPPs and CQAs relies on Design of Experiments (DoE), including fractional factorial designs and definitive screening designs (DSD) [10]. These designs are effective for structured screening and

for estimating main effects and interactions under classical modeling assumptions. However, as the number of factors increases, or as nonlinearities and time-dependent dynamics become important, experimental requirements can become large. In practical PC settings this can translate into extended timelines, high analytical load, and limited flexibility for iterative follow-up [12].

2.3.2 Alternative Designs

To address limitations of classical DoE—particularly when the objective is predictive modeling rather than factor-effect estimation—alternative designs are increasingly adopted. Many of these designs are motivated by improved space coverage (for model training) or improved information efficiency (for parameter estimation), and several are directly relevant to the subset-selection strategies evaluated in Chapter 3.

D-optimal Space Filling Designs

Model-based DoE often uses optimality criteria such as D-optimality to maximize information content. In classical settings, D-optimality maximizes the determinant of the Fisher Information Matrix (FIM), minimizing the volume of the parameter confidence ellipsoid and improving identifiability [12, 13]. While derived under linear or parametric assumptions, D-optimal selections often yield empirically strong training sets for flexible models by promoting input diversity and reducing redundancy, which motivates their inclusion in this thesis as a pragmatic selection heuristic.

Space-Filling and Latin Hypercube Sampling

Space-filling designs are commonly used to initialize data-driven and hybrid models by encouraging broad coverage of the feasible input region [14]. Latin Hypercube Sampling (LHS) enforces stratification across each dimension, reducing clustering and improving coverage relative to purely random draws [10, 15, 16]. In the context of this thesis, space-filling motivation supports the use of low-discrepancy sequences (e.g., Sobol) and informs interpretation of why certain training subsets may yield better

learning curves at small training budgets.

2.4 Statistical Modeling Approaches: Linear Regression and PLS

Multiple Linear Regression (MLR) and Partial Least Squares (PLS) are widely used in practice as baseline "black-box" models for process understanding and prediction [17]. These approaches can be robust for interpolation within the executed design and remain attractive due to interpretability and established workflows. However, they often fail to capture the time-dependent nonlinearities characteristic of cell culture processes and can generalize poorly outside the calibration range [18]. Empirical studies have shown that PLS models may produce unrealistic predictions when tested outside their calibration range, motivating model architectures that incorporate mechanistic structure or more flexible nonlinear learning [17].

This thesis uses statistical baselines to contextualize hybrid-model performance and to translate predictive error into an operational decision question: whether a reduced-experiment PC campaign could plausibly maintain comparable predictive and decision-making utility.

2.5 Mechanistic & Hybrid Modeling

2.5.1 Mass-Balance Models in Cell Culture

Mechanistic models represent cell culture using differential equations describing rates of change for cell density, metabolites (e.g., glucose, lactate), and product titer. The primary benefit is interpretability and constraint: mass balances encode physically meaningful structure and prevent certain classes of unrealistic behavior. However, establishing a universal mechanistic model for mammalian cell culture remains challenging due to incomplete understanding of metabolic networks, context dependence across cell lines and molecules, and the difficulty of parameterizing kinetics under lim-

ited data [18]. As a result, purely mechanistic approaches can be brittle or expensive to maintain across programs.

2.5.2 Hybrid Gaussian Processes & Physics-Informed Neural Nets

Hybrid models aim to combine mechanistic structure with machine learning flexibility. A proven architecture is the "Combined Hybrid Model," which links a Propagation Model (predicting process dynamics) with a Historical Model (predicting final quality attributes) [17]. This decomposition mirrors the practical separation between (i) dynamic evolution of bioreactor states and (ii) endpoint quality outcomes conditioned on process history.

- Hybrid Gaussian Processes: Gaussian Processes (GPs) can be used to learn unknown kinetic rate functions (e.g., growth rate μ) embedded within mass balance equations. This structure has been shown to outperform purely black-box alternatives (e.g., PLS) for predicting CQAs, with reported error reductions on the order of $\sim$50% in some settings [18, 19]. Bayer et al. (2020) extended this approach explicitly to a process characterization context, pairing hybrid modeling with intensified DoE to demonstrate data-efficiency gains over classical full-factorial designs, albeit on academic-scale fermentation data. This thesis asks whether equivalent or greater gains are achievable on industrial CHO PC data collected under regulatory-grade conditions [11].

- Physics-Informed / Knowledge-Constrained ML: By incorporating domain knowledge as mathematical constraints or inductive biases, models can learn efficiently from sparse datasets and produce physically consistent predictions in regimes where purely data-driven models may fail [6].

- Industrial-Scale Application: Recent work has demonstrated that these architectures translate beyond academic settings. Cao et al. (2024) applied a physics-informed neural network to large-scale pilot CHO fed-batch data at

Pfizer, modeling viable cell density, metabolite consumption, and titer across manufacturing-scale campaigns [20]. That study establishes that deep hybrid models can be trained on data with the noise characteristics and operational variability of industrial production — a precondition for the analysis undertaken here.

These hybrid approaches motivate the modeling choices in Chapter 3 and the evaluation focus in Chapter 4: if mechanistic structure improves data efficiency, then learning curves under reduced training budgets should degrade more slowly than classical baselines.

2.6 Transfer Learning

Transfer learning leverages information from historical molecules (source tasks) to improve modeling for a new molecule (target task). This is particularly relevant in process development settings where datasets are small for any single program but large in aggregate across a portfolio.

- Entity Embeddings: Categorical variables (e.g., cell line identity) can be mapped to continuous vector representations, enabling models such as hybrid Gaussian Processes to learn similarity structure and transfer information across related entities [21]. This approach can reduce the amount of new data required to build models for a new cell line or program.

- Meta-Learning: Methods such as PACOH (PAC-optimal hyper-posterior) learn a prior distribution from historical tasks. Benchmarking indicates that meta-learning GPs can reduce test error substantially relative to task-local models in low-data regimes (e.g., 2–4 experiments) [10].

- Clone Selection: Transfer learning has been applied to accelerate clone selection by leveraging historical screening data, reducing the number of new measurements required to identify top-performing clones.

In this thesis, transfer learning is operationalized through the use of upstream commercial process development data (CPD) as prior information when modeling late-stage characterization outcomes, and the impact is evaluated in Chapter 4.

2.7 Advanced Optimization & Active Learning

Active learning couples modeling and experiment selection by choosing new experiments that are expected to be maximally informative or maximally improve an objective. In media development, Bayesian optimization-based active learning has demonstrated order-of-magnitude reductions in experimental burden relative to traditional DoE in identifying optimal media compositions [15]. More broadly, iterative model-based process development strategies can converge to optima in fewer steps than static approaches by incorporating new data as it becomes available [22].

While this thesis is retrospective and evaluates fixed training subsets (rather than conducting prospective closed-loop optimization), the active-learning literature motivates the core framing: experiment selection policy matters, and efficiency gains are possible when selection is guided by model needs rather than by uniform grids alone.

2.8 Bayesian Experimental Design and Batched Bayesian Optimization

Bayesian Optimal Experimental Design (BOED) formalizes the selection of experiments to maximize expected information gain (EIG), supporting principled tradeoffs between experimental cost and learning value [13, 23]. In process development, BOED connects naturally to hybrid modeling because mechanistic structure enables parameter- and uncertainty-aware decision criteria.

- **Model-Based DoE (MBDoE):** Tools such as `Pyomo.DOE` enable optimization of experimental conditions to improve parameter precision under criteria such

as A-optimality or D-optimality for nonlinear dynamic models [12].

- **Measurement Optimization:** Recent frameworks also optimize which measurements to take ("measure this, not that"), explicitly trading off measurement cost against information gain. This is particularly relevant for digital-twin maintenance, where the measurement plan can dominate operational burden [12].

This literature provides a principled backdrop for why reduced-experiment PC may be possible. However, the implementation focus of this thesis is a practical subset-selection evaluation over existing executed experiments, enabling comparability across selection strategies and direct benchmarking against current-state baselines.

2.9 Multi-Armed Bandits, Asynchronous Optimization

Multi-Armed Bandit (MAB) methods address efficient resource allocation under uncertainty and have been applied in contexts such as clinical dose-finding (e.g., Thompson Sampling) [24]. These principles transfer to process development when experiments are expensive and feedback arrives sequentially: policies can be designed to allocate runs adaptively as evidence accumulates.

A related operational idea is asynchronous or iterative optimization, where data are incorporated as soon as they become available rather than waiting for complete blocks of experimentation. This can accelerate development cycles by enabling earlier learning and earlier replanning [22]. Although not implemented prospectively here, the MAB and asynchronous optimization literature motivates the longer-term vision for integrating hybrid models into future PC planning workflows.

2.10 Summary and Research Gap

The literature suggests a clear opportunity to reduce experimental burden in late-stage process development by combining (i) hybrid modeling architectures that improve data efficiency through mechanistic structure [18, 17, 6] with (ii) informed experiment-selection policies grounded in optimal design, space-filling principles, and sequential decision-making [12, 10, 13, 23]. Regulatory guidance further indicates that AI/ML methods may support decision-making when a risk-based credibility assessment demonstrates that a model is fit for its intended use [9]. Together, these threads motivate the hypothesis that predictive, decision-support models for Process Characterization can be developed using fewer bioreactor runs than classical DoE-centric campaigns. Additionally, it is worth noting that prior hybrid modeling studies have demonstrated strong predictive accuracy in bioprocess settings; they have largely evaluated models trained on fixed, complete datasets—leaving open the question of how performance degrades as training data are progressively reduced and which experiment-selection strategies best preserve accuracy at smaller budgets.

However, it remains unclear how these ideas translate into practical, measurable efficiency gains in industrial PC settings. In particular, there is limited empirical evidence on (i) how predictive performance scales with the number of PC experiments available for training, (ii) whether different training-set selection strategies produce meaningfully different learning curves at small experimental budgets, and (iii) the extent to which earlier-stage upstream development data can reduce the number of late-stage PC runs required for comparable predictive accuracy. These questions are especially relevant in cell culture, where dynamics are nonlinear and time-dependent, measurements are heterogeneous, and dataset sizes for any single molecule are often modest.

Accordingly, this thesis evaluates a retrospective, data-driven framework for reducing PC experimental burden. A key premise is that existing historical PC datasets were not collected with machine learning as the primary objective; they reflect operational constraints, classical DoE goals, and practical idiosyncrasies of execution.

Demonstrating predictive utility on this "non-ML-optimized" data therefore serves as a conservative test: if the approach works under these conditions, it is likely to perform at least as well—and plausibly better—when future data collection is designed explicitly to support hybrid modeling and adaptive experiment selection.

A recent systematic review of 270 hybrid modeling publications from 1990–2024 provides useful context for situating this work: roughly half of published studies used laboratory-scale data, approximately one quarter used pilot-scale data, and just over one fifth used industrial-scale data [25]. This distribution confirms that industrial applications exist but remain a minority of the literature, and that formal process characterization, as distinct from process development or optimization, remains an underrepresented application domain. The present study addresses this gap by evaluating hybrid models on PC data governed by ICH Q8 process validation principles, generated by GMP-qualified analytical methods, and linked directly to commercial manufacturing conditions.

This framing is also intended to reduce organizational risk associated with changing long-held PC study design approaches. Pharmaceutical organizations are appropriately cautious about modifying established development workflows, particularly when those workflows support regulatory filings and manufacturing readiness. By benchmarking hybrid modeling and subset-selection strategies on executed studies, this work provides evidence that can help AMGEN assess the transition risk: the proposed methods can be evaluated against current baselines using existing data before any prospective changes are made.

Chapter 3 describes the data pipeline, reproducible software framework, subset-selection strategies (Random Sampling, Sobol sequences, and D-optimal selection), and the hybrid modeling approach implemented in DataHowLab. Chapter 4 then benchmarks hybrid-model performance against statistical baselines and quantifies how error varies with training-set size and selection method, including an assessment of transfer learning using Commercial Process Development data to improve data efficiency in late-stage characterization.

Chapter 3

Methodology

This chapter details the approach undertaken to develop and execute a generalized optimization and learning framework aimed at reducing the experimental burden of PC studies at AMGEN. It will describe the software engineering approaches utilized, the sources of bioprocess data and how the data were processed, how experiments were selected from the design space to train the models, the architecture of the models used, how models were tuned and iterated on, and finally, how models were evaluated.

3.1 General Approach

This thesis develops and evaluates a generalized optimization-and-learning workflow, visualized in Figure 3-1, to reduce the experimental burden of process characterization (PC) studies while maintaining predictive utility for downstream process understanding and decision-making. The workflow was designed to be implementable prospectively (i.e., prior to observing outcomes) and to support comparisons across alternative subset-selection strategies and model classes.

At a high level, the methodology follows three sequential stages: (i) ingesting and standardizing historical bioprocess data into a common format suitable for model training, (ii) selecting candidate training subsets and fitting hybrid mechanistic–ML models on each, and (iii) evaluating model performance across a range of training-set sizes to quantify how predictive accuracy scales with experimental budget. The

remainder of this chapter details each stage.

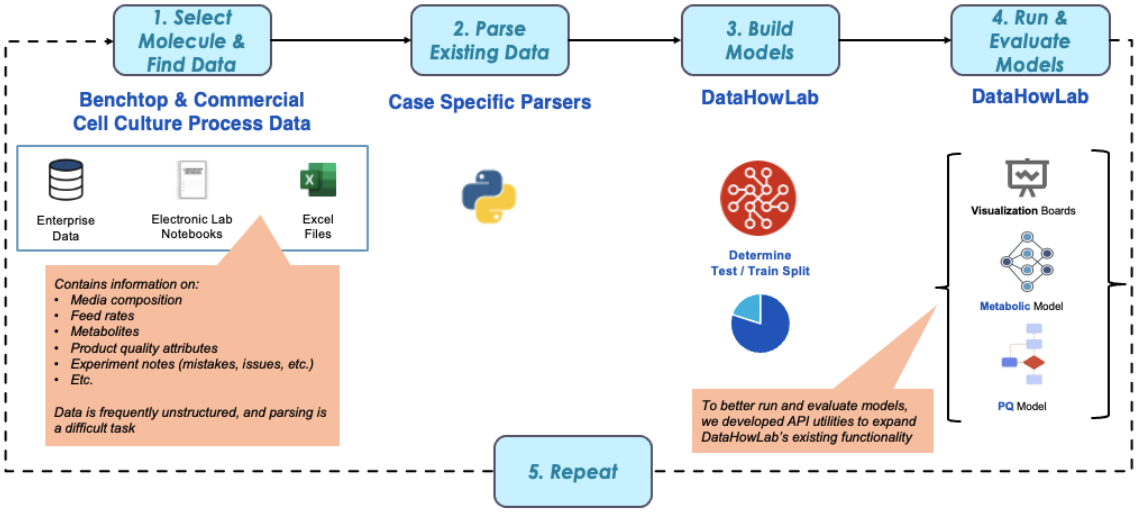

Figure 3-1: Closed-loop "design–learn–evaluate" approach to gathering and using data from PC studies to build models.

In practice, each iteration of this workflow began by selecting a candidate molecule in AMGEN's portfolio that had previously undergone a PC study, gathering the associated data, and building the necessary parsers to collate it into a standard template for model construction. A Python development environment was set up to enable going through this process in an efficient and repeatable manner, because, as alluded to in Subsection 2.1.6, the TDC team at AMGEN embraces the agile software engineering methodology. This was the main approach taken in this work, and one of the core learnings the author took away from this experience, so will be elaborated upon below in Section 3.2.

3.1.1 Primary outputs of the methodology

The methodology yields three outputs used throughout Chapter 4: (i) *learning curves* that quantify how predictive performance scales as additional PC experiments are added to the training set, (ii) a *comparative evaluation* of training-subset selection strategies across PQAs and training-set sizes, and (iii) an empirical *training-sufficiency analysis* that identifies the region of diminishing returns in the learning curves, in-

terpreted relative to the JMP regression baseline and biological replicate error as practical reference points.

3.2 Agile Software Engineering Practices

The modeling and data infrastructure developed in this thesis were built within the context of an agile, team-based software development environment. Given the iterative nature of the work, spanning data parsing, feature engineering, model training, and evaluation, development practices were chosen to prioritize rapid feedback, correctness, and reproducibility.

The project was executed within a cross-functional team operating under an agile cadence. The core coordination mechanism consisted of three recurring standup meetings per week, during which team members shared progress updates, identified blockers, and aligned on near-term priorities. These standups provided a regular synchronization point across modeling, data engineering, and domain experts, ensuring that development remained closely coupled to evolving technical and business objectives.

In addition to scheduled standups, ad hoc individual check-ins and collaborative hackathon-style working sessions were held throughout the week as needed. These sessions were typically convened to jointly debug code, implement new features, or rapidly prototype modeling approaches. This structure enabled concentrated periods of pair- or group-programming when complexity or uncertainty was highest, while avoiding unnecessary coordination overhead during periods of independent work.

3.2.1 Unit Testing

To ensure correctness and maintainability of the codebase, the project emphasized small, composable functions with complete unit test coverage. Rather than testing large, monolithic scripts, functionality was decomposed into narrowly scoped functions whose behavior could be exhaustively specified and validated through unit tests.

Each such function was written with the explicit goal of achieving 100% unit test

coverage, including tests for edge cases and failure modes. This approach reduced the likelihood of silent errors during data processing and model evaluation, particularly important in a context where data inconsistencies at the time of collection could otherwise propagate undetected into future analyses.

This testing strategy was intentionally aligned with industry best practices: small batch sizes, high test coverage, and fast feedback loops all serve as key enablers of both software quality and development velocity [26]. By constraining complexity at the function level and enforcing comprehensive test coverage, the codebase supported rapid iteration without sacrificing reliability.

3.2.2 API Utilities and Codebase Design

To enable reproducible *in-silico* experimentation and repeated model evaluation, a modular Python codebase was developed to interface with the DHL modeling platform and orchestrate *in-silico* simulations. The codebase separates low-level platform interactions from higher-level experimental logic, allowing changes to modeling assumptions or evaluation strategies without altering data access primitives.

While the DataHow graphical user interface (GUI) is well suited for interactive, experiment-level analysis and model inspection, it is not optimized for large-scale, programmatic simulation. The analyses in this thesis required repeated forward propagation, robustness sweeps, and systematic evaluation across many candidate experimental designs. Executing these studies manually through the GUI would have been time-intensive and operationally brittle. The development of an API-based utility layer therefore enabled scalable, automated model execution and ensured reproducibility across high-throughput *in-silico* experiments.

A dedicated API utility layer abstracts interactions with the DHL client, including project discovery, dataset lookup, experiment retrieval, and model execution. This layer standardizes timestamp handling, model input formatting, and result extraction across all experiments. In particular, model execution is wrapped in reusable functions that distinguish between propagation models—used to simulate time-evolving state variables—and historical models—used to predict final product quality attributes

(PQAs).

This separation of concerns enables consistent reuse of model execution logic across multiple studies while minimizing duplication and reducing the likelihood of silent data handling errors. All downstream analysis and simulation scripts rely on these utilities to ensure uniform execution semantics across experimental conditions.

3.3 Data Collection and Parsing

3.3.1 Overview of Available Data

As can be seen in Figure 3-1, data from PC studies are often generated across heterogeneous laboratory systems, analytical platforms, and documentation formats. This includes enterprise-level data (SOPs, media composition documents, data from automated bioreactor systems, etc.), lab notebook data that is manually entered or collated by technicians, Excel files, and more. This data includes information on bioreactor conditions (e.g., temperature, pH, various chemical concentrations), media composition, media feed rates, and final PQAs. While these data may be technically accessible, they are frequently difficult to locate programmatically, inconsistently structured across studies, and tightly coupled to local context or institutional knowledge. Such characteristics limit their downstream utility for data-driven modeling and significantly increase the marginal cost of each new analysis.

3.3.2 FAIR Data

A central objective of the data collection and parsing effort in this work was to move historical process characterization (PC) study data toward compliance with FAIR data principles—specifically, that data be Findable, Accessible, Interoperable, and Reusable. In the context of biopharmaceutical process development, these principles are not an abstract ideal but a practical requirement for enabling scalable modeling, cross-study learning, and reproducible analysis.

Within this project, FAIR data principles were interpreted operationally as fol-

lows. Findability refers to the ability to systematically identify relevant datasets and experiments without manual, study-specific investigation. Accessibility denotes the ability to retrieve data in a consistent and permissioned manner suitable for automated workflows. Interoperability requires that variables, units, and identifiers be standardized such that data from different studies can be combined or compared without bespoke transformations. Reusability implies that datasets retain sufficient context, provenance, and structural consistency to support future analyses beyond the immediate scope of the original study.

The data parsing and standardization work described in this thesis represents an incremental step toward these goals rather than a complete realization. At the outset of the project, data ingestion relied primarily on manually curated exports and case-specific parsers which will be discussed in the next section. Approximately halfway through the project, a separate team within the organization initiated parallel efforts to automate data extraction directly from electronic lab notebooks and AMGEN's internal IT infrastructure. While this automation effort was not fully integrated into the workflow described here, it reflects a broader organizational commitment to reducing manual data handling and improving data standardization at the source.

Importantly, the FAIR data objective is shared across the organization and extends beyond the scope of this thesis. The modeling work conducted here benefited from, and in turn helped motivate, these broader initiatives by illustrating how data accessibility and interoperability directly affect the feasibility and scalability of advanced modeling approaches. As automation and infrastructure maturity improve, future applications of similar modeling frameworks are expected to require substantially less bespoke data engineering effort, thereby accelerating iteration and enabling more systematic reuse of historical development data.

3.3.3 Parsing Approach

Process characterization (PC) study data were collected from multiple structured sources associated with each experimental run, including task plans, calculated feed rates, glucose and permeate consumption estimates, and measured process variables.

These data sources differ in sampling frequency, duration, and naming conventions, necessitating a unified preprocessing workflow prior to model training or evaluation.

For each experimental run, raw data tables were aligned to a common process timeline measured in run days. Tables exceeding the effective experiment duration were truncated, while missing values were explicitly retained to preserve temporal alignment. Initial conditions were extracted separately from dynamic time-series variables to support models requiring distinct initialization inputs.

Final product quality attributes were treated as end-point measurements rather than time-series outputs. For each PQA, the final valid observation within the run window was selected and stored as the response variable. In cases where measurements were below detection limits or missing, values were explicitly encoded as missing rather than imputed, allowing downstream models and evaluations to handle uncertainty transparently. Following collection, all experimental data were parsed into a canonical representation designed to support both model training and *in-silico* simulation. Each experiment was represented by three primary components: (1) dynamic input time-series, (2) static initial conditions, and (3) final-day output variables.

Dynamic inputs were stored as fixed-length vectors aligned to run day indices, ensuring consistent dimensionality across experiments. Static inputs were encoded separately to avoid unintended temporal broadcasting. Output variables consisted of final-day PQAs, extracted using a uniform rule that selects the last non-missing value within the experiment duration.

This parsing strategy enforces a strict schema that simplifies model interfacing and enables batch execution across heterogeneous experimental conditions. By standardizing all experiments into a single data structure, downstream scripts can iterate over experiments without special-case logic for individual runs, while preserving traceability back to the original raw measurements.

3.3.4 Data Cleaning

As can be seen in the left portion of Figure 3-2, one of the challenges with this work was the consistency of the data available to analyze. Data was frequently littered

with small outliers, which arose from a variety of possible sources, including but not limited to:

- Automated Measurement Error: stochastic error inherent to automated measurement equipment;

- Sensor Detection Limits: some machines, such as those used to measure lactate, have upper limits on the concentration they are able to detect, and samples must be titrated manually to record the true value. This is what is shown in the left side of Figure 3-2, specifically the regions circled in red;

- Lab Technician Error: measurements that are flawed because of human error (e.g., pipetting the wrong volume of liquid out of a bioreactor);

- Inconsistent Data Collection, Recording Approach, or Annotation: some measurements like the volume of base added to a bioreactor, are recorded in multiple ways across different experiments. In some cases, the base volume was recorded as cumulative base added to a bioreactor, but in others it was recorded as the volume of base in the base storage vessel (a measurement that decreases as the base is added to the bioreactor). This often takes the form of missing or inappropriate recording of the data collection approach.

In these cases, it became necessary to manually correct and manipulate the data to help ensure models built on said data performed as well as possible. In the example of Lactate concentration, on the right side of Figure 3-2, the values that were clearly running into the sensor detection limits were removed, and replaced with a first-order polynomial fit between known accurate data points. In this case, in the parsed data, we found examples where measurements hit the detection limit (e.g., 14.01 g/L of Lactate, which is recorded in data tables as "$>$ 14.01") and were later sampled manually by a lab technician via dilution.

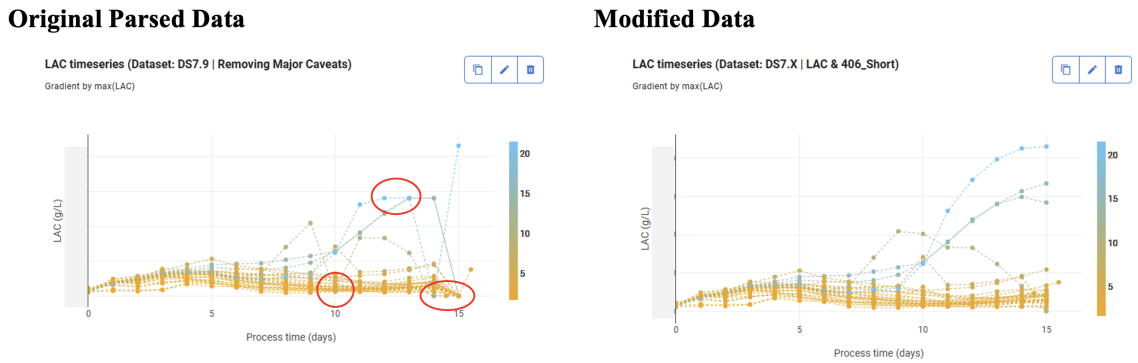

Figure 3-2: Example data-cleaning case for lactate: saturation at an assay detection limit and subsequent manual dilution measurements motivate removal of saturated points and local interpolation to restore a usable trajectory.

3.4 Training and Test Subset Selection Methods

A central objective of this work is to quantify how predictive performance improves as additional process characterization (PC) experiments are incorporated into the training set. To enable systematic evaluation, multiple strategies were used to select nested subsets of experiments from the full PC dataset for model training, with the remaining PC experiments reserved for testing. To ensure that subset-selection policies are implementable for prospective study planning, training-set selection is performed using only the controllable inputs (CPPs) and feasibility constraints, and does not use measured outcomes (PQAs or other responses). This avoids target leakage and mirrors the real decision context in which experiments must be selected before results are observed.

Let the full set of available PC experiments be denoted as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \tag{3.1}$$

where $x_i \in \mathbb{R}^d$ represents the vector of *standardized* process conditions for experiment i, and y_i denotes the associated response variables (e.g., final product quality attributes). For a given training budget $k < N$, the goal is to select a subset

$$\mathcal{D}_{\text{train}} \subset \mathcal{D}, \quad |\mathcal{D}_{\text{train}}| = k, \tag{3.2}$$

such that models trained on $\mathcal{D}_{\text{train}}$ generalize well to the held-out set $\mathcal{D}_{\text{test}} = \mathcal{D} \setminus \mathcal{D}_{\text{train}}$.

To quantify data sufficiency, models were trained using the following numbers of PC experiments:

$$k \in \{15, 25, 35, 45, 55, 65, 76\}. \tag{3.3}$$

In addition, each model was trained with a fixed set of 24 upstream Commercial Process Development (CPD) points, which were included in every training run to represent available prior-development data and keep that contribution constant across k. Across subset-selection methods that generate *synthetic* target conditions, a

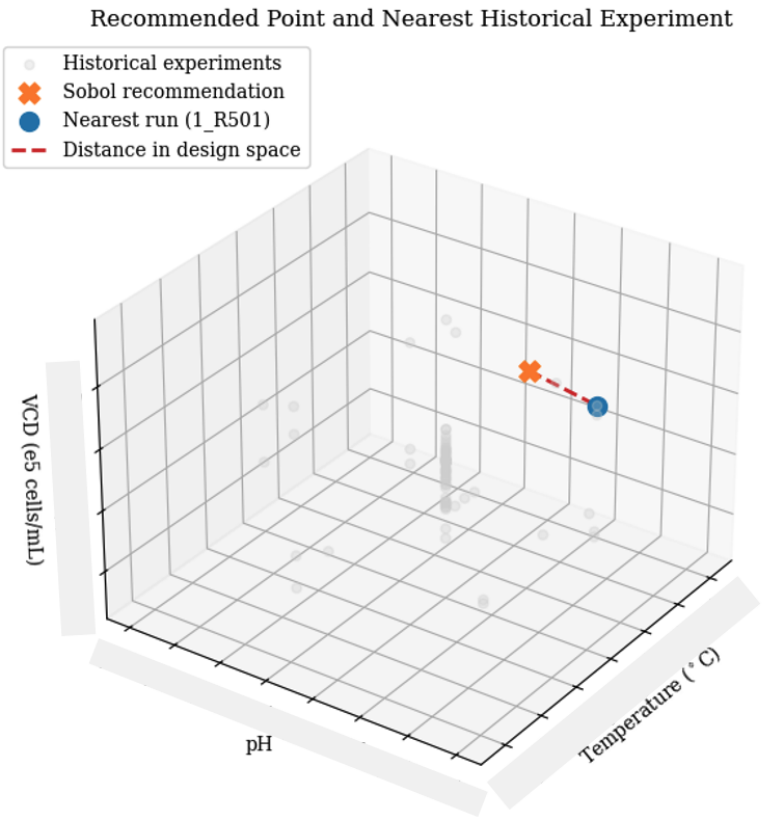

Figure 3-3: Depiction of how a synthetic data point (e.g., a random point generated in the design space) is mapped to its nearest existing PC experiment (in this case, the conditions from bioreactor number 501 from block number one of experimentation, which was named as "1_R501").

consistent mapping step was used to ensure that training data remained restricted to experiments that were actually executed. Specifically, for each synthetic point $s_j \in \mathbb{R}^d$ defined in standardized variable space, the nearest available experimental condition

was selected by Euclidean distance (as shown in Figure 3-3):

$$i^* = \arg \min_{i \in \{1,\ldots,N\}} \|x_i - s_j\|_2. \tag{3.4}$$

The corresponding experimental condition x_{i^*} was then selected and added to $\mathcal{D}_{\text{train}}$. The time-series process data and the associated response data y_{i^*} were not considered in the selection criteria. If multiple synthetic points mapped to the same experiment, duplicates were removed and additional synthetic points were generated (or the next-nearest experiments were selected) until $|\mathcal{D}_{\text{train}}| = k$.

Four subset selection strategies were evaluated: random sampling, Sobol sequence sampling, a D-optimal greedy algorithm, and exterior point selection. Each approach embodies a different tradeoff between simplicity, space-filling behavior, and information efficiency. An example of one of these methods can be seen in Figure 3-4 and Figure 3-5.

3.4.1 Random Sampling

Random sampling serves as a baseline against which more structured selection strategies can be compared. In this approach, the training subset is selected by drawing k experiments uniformly at random without replacement from $\mathcal{D}$:

$$\mathcal{D}_{\text{train}} \sim \text{Uniform}(\mathcal{D}, k), \tag{3.5}$$

where sampling is performed over experimental conditions x_i (inputs) only. To enable reproducible comparisons across methods, a fixed random seed was used for each sampling run. This method makes no assumptions about the geometry of the design space or the underlying process behavior. While random sampling is unbiased in expectation, it does not explicitly promote coverage of extreme operating conditions or low-density regions of the design space. As a result, its performance can exhibit high variance across different random draws, particularly when k is small.

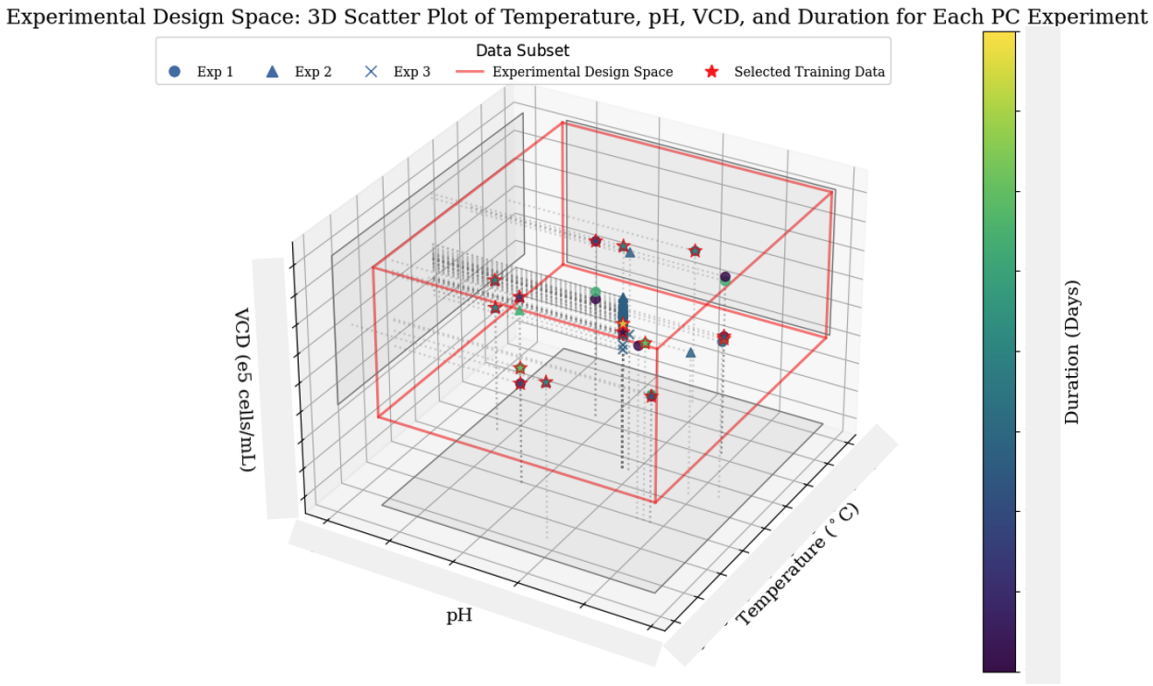

Figure 3-4: 3D scatter plot of available PC experiment setpoint conditions, where the four variables were pH, initial VCD, Temperature, and Duration. This is a combination of experiments run across three "blocks" labeled "Exp 1", "Exp 2", and "Exp 3". The red box and the shaded regions in the axial planes represent the limits of the design space covered in the PC study used. The points with red outlines represent points selected by one of the selection methods described in this section.

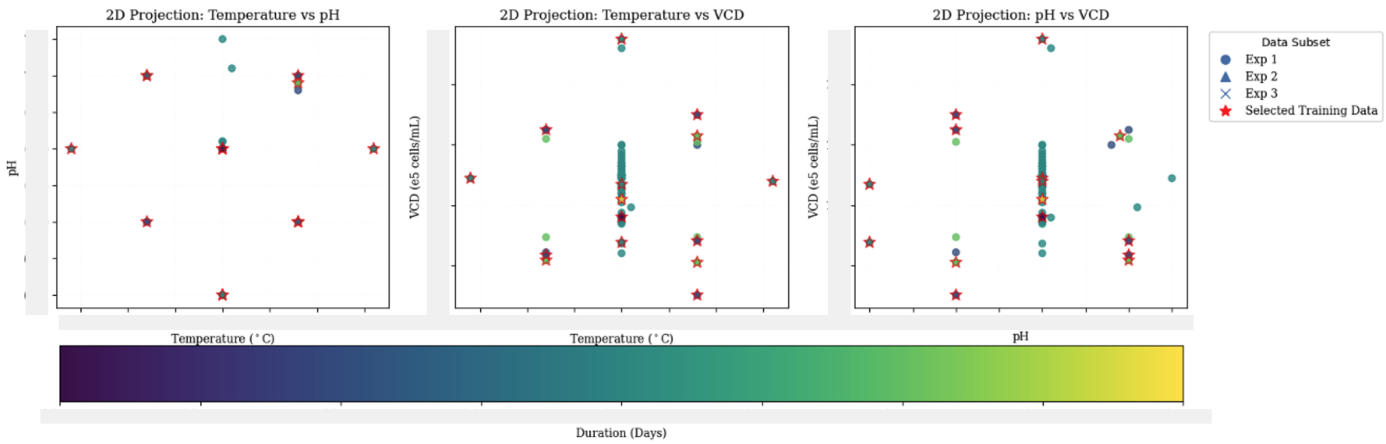

Figure 3-5: 2-dimensional projections of the design space, showing the training set selection. Here the central composite design is clearly visible (a clear centerpoint with variables perturbed up and down around that value).

3.4.2 Sobol Sequence Sampling

Sobol sequences are low-discrepancy quasi-random sequences designed to uniformly fill a high-dimensional hypercube. To apply Sobol sampling to an existing PC dataset, the minimum and maximum values for each process variable were first determined from the available experimental data. A Sobol sequence $\{u_j\}_{j=1}^{k}$ was then generated in the unit hypercube $[0,1]^d$ (with scrambling applied for improved uniformity) and linearly scaled to the experimental bounds to yield synthetic points $\{s_j\}_{j=1}^{k}$. Each scaled Sobol point s_j was then mapped to the nearest available experimental condition in standardized variable space using Euclidean distance:

$$i^* = \arg \min_{i \in \{1,\ldots,N\}} \|x_i - s_j\|_2. \tag{3.6}$$

The selected experiments $\{x_{i^*}\}$ form the training subset. This approach encourages space-filling coverage while remaining constrained to previously executed experiments. Compared to purely random sampling, Sobol-based selection reduces clustering and improves coverage of the interior of the design space, especially in moderate to high dimensions. Given the same seed, this procedure produces reproducible point selections.

3.4.3 D-Optimal Greedy Point Selection Algorithm

D-optimal experimental design seeks to maximize the information content of selected experiments by minimizing the uncertainty in estimated model parameters. For linear models, this corresponds to maximizing the determinant of the Fisher information matrix:

$$M = X^\top X, \tag{3.7}$$

where X is the design matrix formed from the selected experiments. Formally, the D-optimal subset maximizes

$$\det(X^\top X). \tag{3.8}$$

Because exhaustive optimization over all subsets is computationally intractable, a greedy approximation was employed. At each iteration, the next experiment was selected as the point that produced the largest incremental increase in the determinant:

$$x_{\text{next}} = \arg \max_{x_i \in \mathcal{D} \setminus \mathcal{D}_{\text{train}}} \det\left(X_{\text{train} \cup i}^{\top} X_{\text{train} \cup i}\right). \tag{3.9}$$

For numerical stability, this maximization was implemented using the log-determinant (e.g., via `slogdet`) rather than computing $\det(\cdot)$ directly. This procedure favors points that are linearly independent and complementary to the existing subset, resulting in designs that efficiently span the input space. Although the criterion is derived under linear modeling assumptions, D-optimal designs often provide strong empirical performance for nonlinear and hybrid models by promoting diverse excitation of inputs.

3.4.4 Exterior Point Selection

Exterior point selection prioritizes experiments located near the boundaries of the design space, where model extrapolation error is typically largest. For each standardized input dimension, lower and upper bounds were identified, and experiments closest to these extrema were selected. More generally, exterior points were identified by maximizing the distance from the center of the design space:

$$x_{\text{ext}} = \arg \max_{x_i \in \mathcal{D}} \|x_i - c\|_2, \tag{3.10}$$

where $c = (0, \ldots, 0)$ denotes the center of the standardized feature space. Once exterior points were selected, the remaining training slots were filled using one of the other selection strategies (e.g., random or Sobol sampling) to ensure sufficient interior coverage. This hybrid approach reflects common industrial PC practice, where boundary conditions are emphasized to establish operating limits while maintaining representative coverage of nominal conditions.

3.5 Hybrid ML-Mechanistic Model Implementation

The core modeling efforts in this work utilized DataHowLab (DHL), a hybrid modeling platform developed by DataHow AG. This platform integrates first-principles engineering knowledge with data-driven machine learning techniques, allowing for the generation of predictive models even in data-sparse environments typical of process characterization (PC). Unlike purely statistical approaches, which treat the bioreactor as a black box, or purely mechanistic models, which require extensive kinetic parameterization, the hybrid framework leverages the strengths of both: it enforces mass balance and biological constraints while using machine learning to capture complex, nonlinear process dynamics that are difficult to describe mechanistically. The practical consequence for data efficiency was that the mechanistic structure acts as a form of built-in regularization: by constraining the model to solutions that are physically consistent (e.g., concentrations cannot be negative, mass is conserved), the ML component needs fewer observations to learn the residual process behavior. In other words, the model does not need to rediscover basic biology from data, instead it arrives with that knowledge encoded, and uses the available experiments to learn only what cannot be derived quickly from first principles.

3.5.1 Data Representation in DataHowLab

To enable consistent training, simulation, and deployment across heterogeneous bioprocess datasets, DHL requires that process data be represented using a standardized set of variable roles. In practice, this work leveraged DHL's template-based schema, in which each column is assigned to one of five categories: mandatory metadata fields, X variables, W variables, Z variables, and Y variables. This can be seen in Figure 3-6

- Mandatory variables capture experiment identifiers and bookkeeping fields (e.g., experiment name, dates, and required keys used to join time-series tables and endpoint tables).

- X variables represent *observed* process measurements (e.g., glucose, lactate,

Variable Class	Mandatory Input	Mandatory Input	Mandatory Input	User Field	Mandatory Input	Z Variable	Mandatory Input	X Variable	W Variable	Bolus Feed Volume	Y Variable
Variable Code	Date	ExpID	ExpName	TrainTest	Time	IngredientA	VCD	Glc	temp	BasalFeed	Aggr
Variable Description	Date	ExpID	ExpName	TrainTest	Time	IngredientA	VCD	Glucose	Temperature profile	BasalFeed	Aggregate
Variable Unit	N/A	N/A	N/A	N/A	Days	-	10^6 cell/ml	g/L	degC	L	mg/L
Lower Limit	N/A	N/A	N/A	N/A	0	N/A	0	0	34	N/A	0
Upper Limit	N/A	N/A	N/A	N/A	15	N/A	N/A	N/A	40	N/A	N/A
Measurement Std.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Variable type	string	string	string	string	numeric	logical	numeric	numeric	numeric	numeric	numeric
		Run01	Run01	0	0	1					
					1						
					2						
					3						
					4						
					5						
					6						
					7						
					8						
					9						
					10						
					11						
					12						
					13						
					14						

Figure 3-6: Redacted DataHow template with some example variables.

ammonia, viability, and other in-process analytes) that describe the evolving bioreactor state over time.

- W variables represent *dynamically controlled* trajectories that are not the primary design factors of the study (e.g., realized pH and temperature profiles). These variables can be included as time-varying inputs to better reflect the executed control strategy rather than idealized setpoints.

- Z variables capture *categorical variables and design factors* (e.g., block identifiers, feed strategy categories, or engineered factor settings). In this thesis, Z includes the core designed PC conditions (such as pH, temperature, initial VCD, and duration) as well as optional categorical descriptors when available.

- Y variables represent the *target outputs* that the models are trained to predict, including final product quantity and quality attributes (PQAs) such as titer and product quality readouts (e.g., charge variants).

This explicit typing is not merely a data-formatting requirement; it defines the causal directionality of the modeling task and enables DHL to compose chained simulations. In particular, propagation models are trained to forecast time-evolution for relevant X variables (and selected W/Z inputs), and historical (quality) models are trained to map the resulting trajectories (or their summaries) alongside process conditions to endpoint Y attributes.

3.5.2 Hybrid ML User Interface and Workflow

The interaction with the modeling framework was designed to accommodate both high-level user oversight and automated, scalable execution. The workflow began with the ingestion of the standardized data structures described in Subsection 3.3.3.

Data were uploaded into the platform and organized into a hierarchical structure of *Projects* (corresponding to specific molecules or programs) and *Datasets* (corresponding to specific experimental campaigns, e.g., PC or CPD studies). While the platform provides a web-based Graphical User Interface (GUI) for manual inspection, variable selection, and individual model training, the scale of this study—requiring the training of hundreds of models across varying data subsets—necessitated a programmatic approach.

Consequently, we utilized the API utilities described in Subsection 3.2.2 to interface directly with the backend. This allowed us to programmatically define training sets, trigger model training jobs, and retrieve performance metrics without manual intervention. The GUI remained a critical tool for sanity-checking model inputs, visualizing training curves in real-time, and conducting ad-hoc "what-if" simulations to validate model behavior before deploying them for bulk analysis.

3.5.3 Propagation and Historical Models

As introduced in Section 2.5, the hybrid modeling architecture in this work consists of two distinct but coupled model types: *Propagation Models* and *Historical Models*. This separation of concerns mirrors the biological reality of the process, distinguishing between the dynamic evolution of the cell culture and the resulting final product quality.

Propagation Models

Propagation models are designed to predict the time-evolution of state variables within the bioreactor, (e.g., Viable Cell Density (VCD), glucose concentration, lactate, ammonia, titer). These models function as dynamic systems, predicting the state of

the reactor at time $t+1$ given the state at time t and the operating conditions (inputs). These correspond to the hybrid Gaussian Process and physics-informed neural network architectures reviewed in Section 2.5.

Mathematically, these are implemented as hybrid differential equations. A mechanistic layer enforces fundamental mass balances (e.g., accumulation = input - output + reaction), while a machine learning component (typically an ensemble of neural networks or Gaussian processes) estimates the specific kinetic rates (e.g., specific growth rate μ or specific productivity q_p) that drive the reaction terms. This structure ensures that predictions physically "make sense" (e.g., concentration cannot be negative, mass is conserved) while allowing the model to learn complex metabolic shifts from the data.

Historical Models

Historical models (also referred to as Quality Models) are designed to predict static, end-point Product Quality Attributes (PQAs) such as glycosylation profiles, charge variants, or aggregation levels. Historical models correspond to the regression-based approaches described in Section 2.4, specifically state-space regression and PLS. Unlike propagation models, which evolve over time, historical models map a vector of process summary statistics to a final quality output.

These models utilize inputs from two sources:

1. Process Conditions: Static setpoints (e.g., pH, Temperature) and initial conditions.

2. Time Course Conditions: The dynamic trajectories of different process parameters such as metabolite concentrations or media consumption. These conditions can either be provided from actual data or from the outputs of propagation models.

By chaining these models together, using the Propagation model to simulate the process dynamics and feeding those results into the Historical model, we established

a complete "digital twin" capable of predicting how upstream process parameters influence final drug quality. This process can be seen in Figure 3-7.

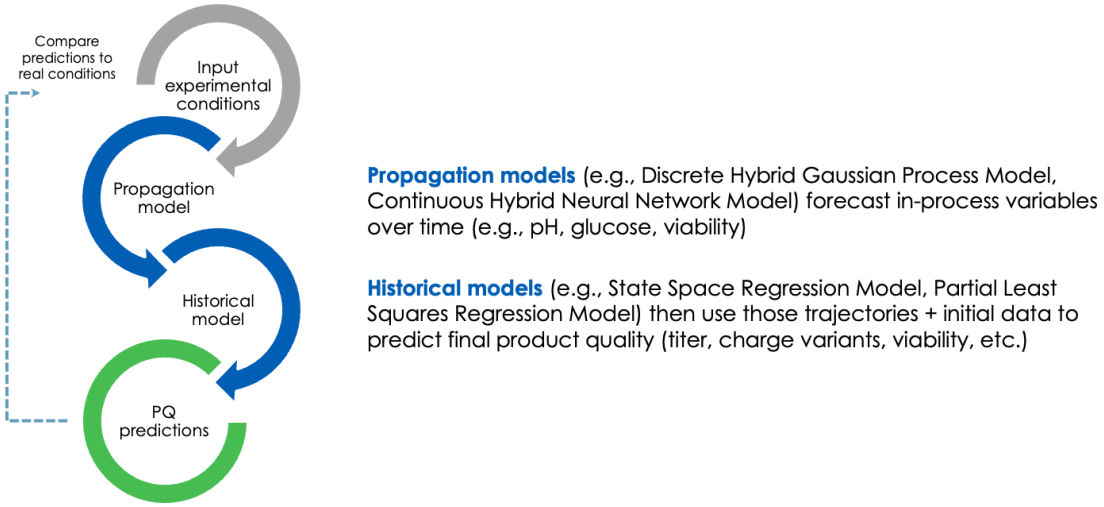

Figure 3-7: Diagram describing how input experimental conditions are utilized to generate full *in-silico* simulations of bioreactor runs.

3.5.4 Model Configurability, Ensembles, and Cross-Validation

A practical advantage of DHL for PC settings is that both propagation and historical models are configurable by selecting a model family and adjusting a small set of hyper-parameters. In this project, propagation models included hybrid Gaussian-process and hybrid neural-network formulations, and historical models included regression-based formulations (e.g., state-space regression and partial least squares regression), consistent with the objective of balancing flexibility with interpretability and stability in low-data regimes. DHL additionally supports ensemble training, where multiple sub-models are fit and aggregated to improve robustness. Key configuration parameters include the number of sub-models in the ensemble and the percentage of observations included in each sub-model, enabling bagging-style variance reduction in settings where PC datasets are small and heterogeneous. To further mitigate overfitting risk, DHL provides automatic cross-validation during training to estimate generalization error and support model selection across candidate configurations. In this thesis,

these built-in cross-validation outputs were used as primary signals when comparing candidate models, while recognizing that hyperparameter tuning remains a partially manual workflow that requires training and comparing multiple candidate models; the tuning process and its limitations are discussed in Subsection 3.6.2.

3.6 Iterative Model Development and Hyperparameter Tuning

This section describes how candidate hybrid models were iteratively constructed and compared in DHL, including (i) how explanatory variables were selected and adjusted across model iterations, (ii) how hyperparameters were tuned to balance flexibility and generalization, (iii) the practical workflow used to train and version models, and (iv) the quantitative metrics used to compare models, including replicate-error baselines and multi-PQA performance aggregation.

3.6.1 Selection of Variables

Variable selection was guided by two objectives: maximizing predictive performance while maintaining explainability and robustness appropriate for a regulated process-development setting. In practice, we prioritized variables that (i) were mechanistically interpretable and defensible as plausible drivers of process performance and product quality, (ii) were consistently available across runs (minimizing missingness and inconsistent sampling), and (iii) exhibited high data quality across the full set of experiments used for training and evaluation.

A key practical constraint in PC datasets is that many candidate variables exist in principle, but only a subset are collected consistently and reliably across all experimental runs. Therefore, rather than exhaustively searching over all possible predictors, we began from baseline DHL projects that had already undergone an initial round of variable selection and hyperparameter tuning within the organization. From these baselines, variables were added, removed, or modified in a controlled,

hypothesis-driven manner, with each iteration evaluated against the same set of model comparison metrics described in Subsection 3.6.4.

Variable roles and accessibility by model type. As described in Subsection 3.5.1, DHL requires each variable to be assigned to a role (X, W, Z, or Y). These roles determine which variables can be used as inputs or targets for different model components.

- Propagation model inputs. Propagation models simulate the time-evolution of state variables and therefore can use (i) static design factors and initial conditions (Z), (ii) time-varying controlled trajectories (W), and (iii) the current (or lagged) process state (X). In training, their targets are typically the next-step state variables (or increments), and during simulation they generate full trajectories for selected X variables as outputs.

- Propagation model outputs. Propagation models produce predicted trajectories for modeled X variables (and any additional state variables included in the propagation framework). These outputs can be used directly as intermediate results or passed forward to historical models.

- Historical model inputs. Historical (quality) models predict endpoint attributes and therefore take as inputs (i) static process conditions and design factors (Z) and (ii) time-course information derived from X and/or W. In practice, time-course variables may enter either as measured trajectories, as simulated trajectories from propagation models, or as summary statistics derived from those trajectories (e.g., extrema, slopes, integrals, end-of-run values).

- Historical model outputs. Historical models predict endpoint response variables (Y), including PQAs and other final-day outcomes.

This schema helped enforce explainability: variables designated as explanatory were traceable to either controlled conditions (Z, W) or measured process state (X), while targets were restricted to endpoint attributes (Y) and/or state trajectories being

explicitly modeled. Variables that were not consistently collected or were difficult to interpret in a mechanistic context were deprioritized unless they provided clear, repeatable improvements in generalization metrics.

3.6.2 Hyperparameter Tuning

Hyperparameter tuning was performed iteratively to balance model flexibility with generalization. DHL exposes a small set of hyperparameters that govern model complexity and robustness, particularly for ensemble-based training. These can be seen in Figure 3-8. The principal hyperparameters adjusted in this work were:

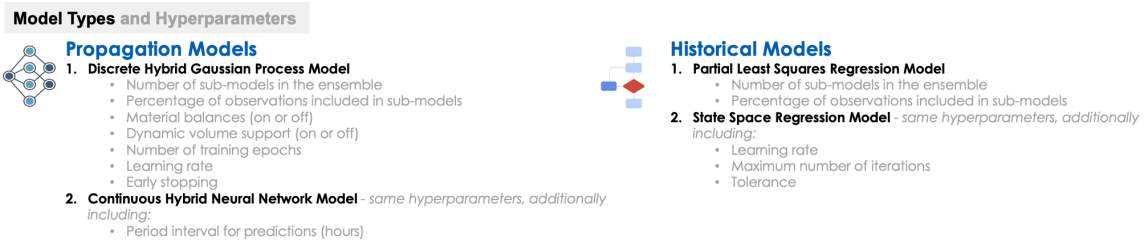

Figure 3-8: Hyperparameters accessible by model type within DHL.

- Sub-sampling percentage. The fraction of observations included in each ensemble member during training. Lower sub-sampling can reduce variance and increase robustness at the cost of higher bias; higher sub-sampling increases fidelity to the training data but can increase overfitting risk.

- Number of models in the ensemble. The number of sub-models trained and aggregated. Larger ensembles can stabilize predictions and reduce sensitivity to idiosyncrasies in training subsets, but increase training time and can obscure overfitting if not paired with cross-validation checks.

- Input feature selection. The specific set of $X/W/Z$ variables made available to each model component (propagation and historical). Because feature choice can dominate model behavior, this was treated as a first-class tuning dimension alongside numerical hyperparameters.

We explicitly avoided including variables that would trivially encode experimental identity without causal meaning (e.g., experiment index, run ID, block labels used solely as identifiers) unless they were being used as legitimate categorical descriptors with a process rationale. This constraint was important to prevent leakage-like behavior in which a model learns to associate an identifier with outcomes rather than learning transferable relationships.

3.6.3 Model Training Process

Model development proceeded through repeated cycles of (i) selecting variables and hyperparameters, (ii) training candidate models in DHL, and (iii) evaluating them using a consistent set of quantitative metrics. While DHL provides automated cross-validation outputs during training, model development remained partially manual because variable selection and model-family selection require domain judgment and careful interpretation of tradeoffs (e.g., improved fit for one PQA versus degraded performance for another).

Two additional workflow practices were used to support reproducibility and institutional learning:

- Model comparison: Candidate models were compared using a consistent decision logic that incorporated multi-PQA summary statistics, and generalization indicators. An example visual used in this decision logic can be seen in Figure 3-10.
- Model lineage documentation: The sequence of model variants, including their origin, key configuration changes, and motivations, was documented in a Miro board to ensure traceability. A simplified representation of this lineage is provided in Figure 3-9.

Stopping criteria Iteration was stopped when improvements exhibited diminishing returns and/or began to increase estimated generalization error. Specifically, we treated rising cross-validation error (reported by DHL) as a strong indicator of overfitting and deprioritized configurations that improved training error but worsened cross-validation

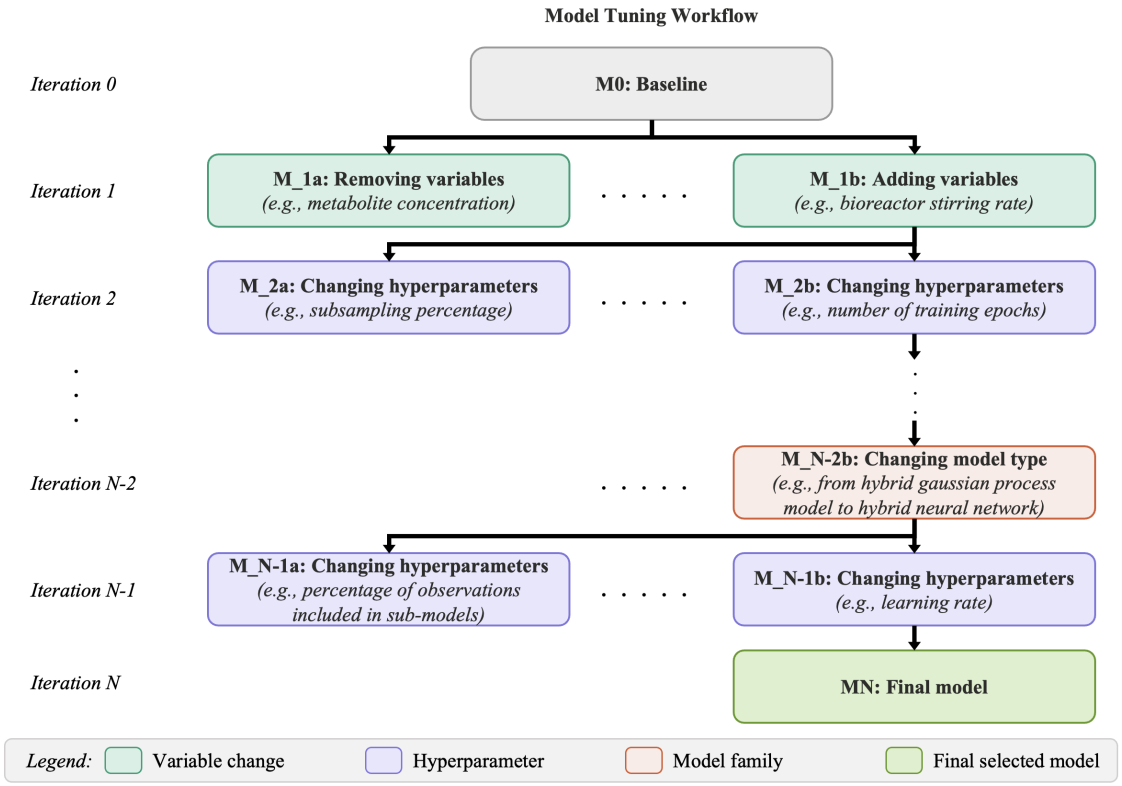

Figure 3-9: Illustrative example of model lineage tracking during iterative hybrid model development, as documented on a collaborative Miro board. Each node represents a candidate model variant, with parent-child relationships showing how models evolved through variable adjustments (green), hyperparameter updates (blue), and model-family changes (orange). Branches represent parallel exploration of alternative configurations at a given iteration; when a change did not improve model performance, that branch was abandoned (indicated by the absence of outgoing arrows). In some cases, a configuration change produced no improvement, causing iterations to be skipped (e.g., a model from Iteration 1 directly producing a child at Iteration 3). In practice, the lineage tree exhibited greater breadth and depth than shown here. Maintaining this structured lineage record served multiple purposes beyond individual model selection: it enabled collaborators to understand the rationale behind each modeling decision, supported reproducibility across teams, and facilitated transfer learning practices — for example, by allowing process characterization models to be initialized from earlier commercial process development (CPD) models rather than built from scratch.

performance. This stopping rule was important given the limited size of PC datasets and the high flexibility of hybrid model families.

3.6.4 ML/Hybrid Model Comparison Metrics

Candidate models were compared using a consistent set of metrics that reflect both prediction accuracy and robustness across multiple PQAs. Metrics were computed per-PQA, and then aggregated across PQAs to support selection of models that generalized well for the full set of predicted quality attributes (rather than optimizing one attribute at the expense of others).

Multi-PQA aggregation

Let $\mathcal{Q}$ denote the set of PQAs modeled, and let E_q denote a scalar error metric (e.g., RMSE) for PQA $q \in \mathcal{Q}$. For each candidate model, we computed:

$$\overline{E} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} E_q, \tag{3.11}$$

and the across-PQA variability:

$$\sigma_E = \sqrt{\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} (E_q - \overline{E})^2}. \tag{3.12}$$

The objective was to identify models with low average error ($\overline{E}$) and low dispersion (σ_E), reflecting both strong overall accuracy and consistent performance across PQAs. These can be plotted against one another to compare models visually, with "better" models being those in the upper-right hand corner of Figure 3-10.

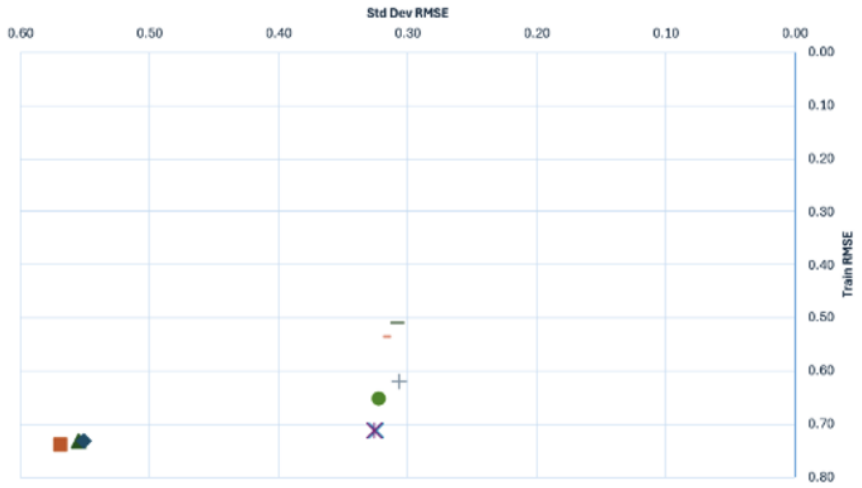

Figure 3-10: Plot showing how models can be compared to one another visually, in this case plotting σ_E versus $\overline{E}$.

RMSE

For a given PQA with true values $\{y_i\}_{i=1}^n$ and predictions $\{\hat{y}_i\}_{i=1}^n$, the root mean squared error (RMSE) is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3.13}$$

RMSE penalizes larger errors more heavily due to the squared term, which was useful when large deviations are particularly undesirable for quality prediction.

MAE

The mean absolute error (MAE) is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{3.14}$$

MAE provides a complementary view to RMSE that is less sensitive to outliers and can be easier to interpret in the original units of the PQA.

Replicate Mapping

To contextualize achievable predictive performance, we estimated replicate error as a practical lower bound driven by experimental variability and analytical noise. Rather than defining replicate groups programmatically, we leveraged DHL's built-in replicate mapping functionality via the GUI, which is designed to enable the grouping of repeated experimental runs.

In this work, replicates were defined as experiments executed at the same input conditions within the same PC "block." This definition reflects how the underlying studies were structured: each block frequently contained repeated runs at key operating points (most notably the centerpoint condition), enabling direct empirical estimation of within-condition variability without confounding across blocks.

For each PQA, DHL reports the resulting replicate error by comparing dispersion of measured outcomes across replicate groups. Conceptually, let $\mathcal{R}$ denote the set of replicate groups identified by DHL, where each group $r \in \mathcal{R}$ contains experiments that share identical input conditions within the same block. For a given PQA with measured values $\{y_i\}$, the replicate-group mean is:

$$\bar{y}_r = \frac{1}{|r|} \sum_{i \in r} y_i, \tag{3.15}$$

and the replicate dispersion within the group is:

$$s_r^2 = \frac{1}{|r| - 1} \sum_{i \in r} (y_i - \bar{y}_r)^2. \tag{3.16}$$

A pooled replicate standard deviation can then be defined as:

$$s_{\text{rep}} = \sqrt{\frac{\sum_{r \in \mathcal{R}} (|r| - 1) s_r^2}{\sum_{r \in \mathcal{R}} (|r| - 1)}}. \tag{3.17}$$

We used this replicate-error estimate as a benchmark when interpreting model performance: as model errors approached s_{rep}, further reductions in RMSE/MAE were treated as increasingly likely to be limited by irreducible experimental variability rather

than model structure. This also informed our stopping criteria in Subsection 3.6.3, since pursuing marginal gains below replicate error risked overfitting to noise.

Error Distributions

In addition to scalar summary metrics, we examined distributions of residual errors to assess bias, heavy tails, and failure modes that are not captured by RMSE/MAE alone. For a given PQA, residuals are defined as:

$$e_i = y_i - \hat{y}_i. \tag{3.18}$$

We analyzed:

- **Central tendency and spread** (e.g., median, interquartile range) to understand typical error magnitude and variability;
- **Skewness and bias indicators** by inspecting whether residual distributions were centered near zero;
- **Tail behavior** to identify whether a model occasionally failed catastrophically under specific conditions.

These distributional checks were particularly important for selecting models intended to be robust across multiple PQAs and across diverse regions of the design space, where average metrics can hide localized weaknesses.

3.7 Benchmarking Against Current Statistical Approaches

A central goal of this thesis is to quantify how predictive performance improves as additional PC experiments are incorporated into the training set. However, the current-state industrial workflow for PC studies is not primarily optimized around predictive accuracy on held-out data. Instead, PC studies are designed to achieve

statistical power for estimating factor effects and interactions, and then analyzed using regression-based tools (commonly JMP) fit using the full study dataset. As a result, a direct apples-to-apples comparison between "traditional ML" evaluation practices (train/test splits with fixed holdout sets) and current PC statistical analysis is not straightforward.

This section describes (i) how current approaches are typically used, (ii) what benchmarking baselines were feasible given dataset constraints, and (iii) which evaluation metric was selected to support comparison across training-set sizes while remaining aligned with how models would be used in practice.

3.7.1 Current-State Practice: JMP Regression Fit on the Full PC Dataset

In current PC workflows, regression models are typically fit using all available PC experiments to estimate main effects and interactions over a pre-specified design (e.g., central composite or factorial). These models are used to support interpretability, establish operating ranges, and ensure that the study has sufficient power to detect effects. Predictive accuracy on unseen runs is generally not the primary objective, and the analysis is often performed on the same data used to fit the model.

Therefore, JMP-based regression can serve as a *directional* baseline for performance, but it should not be interpreted as a fully evaluative benchmark in the conventional machine learning sense. In particular, because JMP models are fit on the full dataset, their reported fit metrics are best interpreted as descriptive of how well the regression surface explains observed outcomes within the executed design, rather than as an estimate of out-of-sample generalization.

3.7.2 Constraints on Train/Test Splits and the Need for an Alternative Benchmarking Metric

A natural approach to benchmarking would be to (i) train models on an increasing subset of PC experiments and (ii) evaluate them on a fixed held-out test set. In this

project, the size of the available dataset and the multi-PQA nature of the problem made such a split unstable: holding out a sufficiently large test set materially reduced the number of experiments available for model construction, while small test sets produced high-variance estimates that changed meaningfully with which experiments happened to be held out.

Additionally, because the primary question in this thesis is "how many experiments are needed," any approach that uses a different test set at each training budget k confounds performance changes due to additional training data with performance changes due to a changing evaluation set.

To address this, we adopted a comparison strategy that separates (i) the subset-selection logic used to construct training sets (described in Section 3.4) from (ii) a consistent evaluation approach that remains comparable across k.

3.7.3 Selected Benchmarking Approach: Whole-Set Error as a Consistent Evaluation Metric

For benchmarking across different training-set sizes, we selected **whole-set error** as the primary metric. Under this approach, models are trained on a specified training subset (which varies with k and subset-selection strategy), but evaluated on the *entire* dataset, including both training experiments and experiments not included in that training subset.

Concretely, for each training budget k (and each training set selection), a model was trained using the corresponding training subset, then predictions were generated for every available experiment in the dataset. Error metrics (e.g., RMSE, MAE; see Subsection 3.6.4) were computed over this full evaluation set.

This choice was motivated by two considerations:

1. **Comparability across k.** Using the same evaluation set for all values of k ensures that performance changes reflect the information added by additional training experiments rather than shifts in the composition of the test set.

2. **Alignment with intended use.** In practice, once a PC dataset exists, the

most common deployment pattern is to train on all available experiments to support prediction and interpretation for future decisions. Whole-set performance therefore reflects model behavior in a usage regime closer to deployment than strict holdout evaluation.

This does not eliminate the risk of overfitting. For that reason, whole-set error was interpreted in conjunction with generalization indicators reported by DHL (cross-validation error) and the stopping criteria described in Subsection 3.6.3. Configurations that improved whole-set error but increased cross-validation error were treated as likely overfit and de-prioritized.

3.7.4 Alternative Generalization Metrics: Leave-One-Out and Cross-Validation Error

Whole-set error provides a consistent basis for comparing models across different training-set sizes k, but it is not, by itself, a pure measure of out-of-sample generalization because the evaluation set includes the training points. For that reason, we also tracked generalization-oriented error estimates.

First, we computed **leave-one-out (LOO) error** for select baseline comparisons. Under LOO evaluation, a single experiment is held out, the model is trained on the remaining $N - 1$ experiments, and the error is computed on the held-out point. Repeating this procedure for all experiments yields an out-of-sample error estimate that is maximally data-efficient for small datasets, albeit computationally expensive and potentially high-variance in highly heterogeneous designs.

Second, we used **cross-validation (CV) error** as reported by DHL during model training. DHL's CV outputs provided a practical signal of whether changes that improved whole-set error were also improving estimated generalization performance, and therefore served as a guardrail against overfitting as model complexity increased.

In the Results chapter, whole-set error is used as the primary metric for comparing performance as a function of k, while LOO and CV error are used to interpret generalization behavior and to identify configurations that appear to overfit. The

implications of these metric tradeoffs—and the limitations of each approach in small, structured PC datasets—are discussed in Chapter 4.

3.7.5 Lower-Bound Baseline: Replicate Error

In addition to JMP regression as a directional comparison point, we used replicate error as an empirical lower bound on achievable predictive accuracy (see Section 3.6.4). Because replicate error reflects within-condition variability across repeated runs at the same settings (within the same block), it provides a practical estimate of irreducible noise in the data. Model performance approaching replicate error indicates that further reductions may be limited by measurement and biological variability rather than model structure.

Accordingly, replicate error was used as a benchmark to contextualize whether improvements in model accuracy were meaningful and to reinforce the stopping rule of avoiding excessive complexity when performance gains were small relative to the inherent variability.

3.7.6 Summary of Benchmarking Interpretation

Taken together, the benchmarking framework in this thesis provides three complementary reference points:

- **JMP regression (directional).** A representation of current-state descriptive modeling fit within the executed PC study design, typically fit using all PC experiments and optimized for interpretability and factor-effect estimation rather than out-of-sample prediction.

- **Whole-set error (comparability across k).** A consistent evaluation set used to quantify how predictive performance scales as additional training experiments are included, enabling direct comparison of learning curves across subset-selection strategies.

- **Generalization and noise context (CV/LOO and replicate error).** Cross-validation and leave-one-out error estimates are used to detect overfitting and assess generalization, while replicate error provides a practical lower bound on attainable accuracy given experimental variability.

This structure motivates the Results chapter evaluation approach and sets up the Discussion chapter, where we interpret the tradeoffs between comparability, generalization, and experimental noise, and explain how these considerations affect conclusions about "how many experiments are enough" in practice.

Chapter 4

Results

This chapter presents the results of applying hybrid mechanistic–machine learning models to retrospective process characterization data from AMGEN. It begins by establishing the comparative framework and benchmarking approach used to evaluate model performance, including the rationale for whole-set error as the primary learning-curve metric and the use of JMP regression accuracy and biological replicate error as reference baselines. The chapter then compares two simulation approaches for representing process conditions and demonstrates the advantage of using measured time-series inputs over idealized setpoint perturbations. Next, it examines the value of upstream commercial process development (CPD) data as a prior, showing how augmenting CPD data with progressively larger subsets of PC experiments improves prediction accuracy. The central analysis follows: a training data sufficiency study that quantifies how model error decreases as experiments are added, revealing distinct convergence behaviors across product quality attributes and identifying an inflection point around 35 training experiments beyond which diminishing returns are observed. Finally, the chapter evaluates sampling strategy performance and feature importance, providing evidence that space-filling designs outperform boundary-focused approaches and that variation in PQA prediction quality is linked to how well the propagation model captures the intermediate state variables most relevant to each attribute.

4.1 Comparative Frameworks, Benchmarking, and Validation

As alluded to in previous sections, benchmarking in this thesis is complicated by (i) the relatively small dataset size, (ii) the multi-PQA nature of the modeling problem, and (iii) the central question of how predictive performance scales as additional training experiments are added. A conventional fixed hold-out test set is undesirable in this setting because the evaluation set would change with training budget k, confounding changes due to added training data with changes due to a shifting test set.

Accordingly, this chapter emphasizes evaluation approaches that remain comparable across k while still providing guardrails for generalization and overfitting. Four metrics are referenced throughout, and shown in Figure 4-1. Test-set error evaluates performance on a held-out set; however, in this project, changing the training budget necessarily changes the held-out composition unless an unrealistically large test set is reserved. Cross-validation (CV) error is used as a generalization check and as an overfitting signal. In practice, CV error differs substantially by PQA and is best interpreted as a guardrail rather than the primary learning-curve metric. Whole-set error is used as the primary metric for learning-curve comparisons: models are trained on a specified subset but evaluated on the entire dataset, enabling consistent comparison across k and matching the common operational pattern of training on all available data once a PC dataset exists. Finally, leave-one-out (LOO/L1O) error provides a more out-of-sample estimate, but is computationally expensive at scale; in this work, L1O was computed selectively (and averaged across repeated model fits) to contextualize whole-set error.

4.1.1 Baseline: Current PC Methods

The primary directional benchmark used in this thesis is the accuracy of a JMP regression model fit to the full PC dataset. In current practice at AMGEN, JMP is often used at the outset of a PC study to assess the statistical power of the experimental

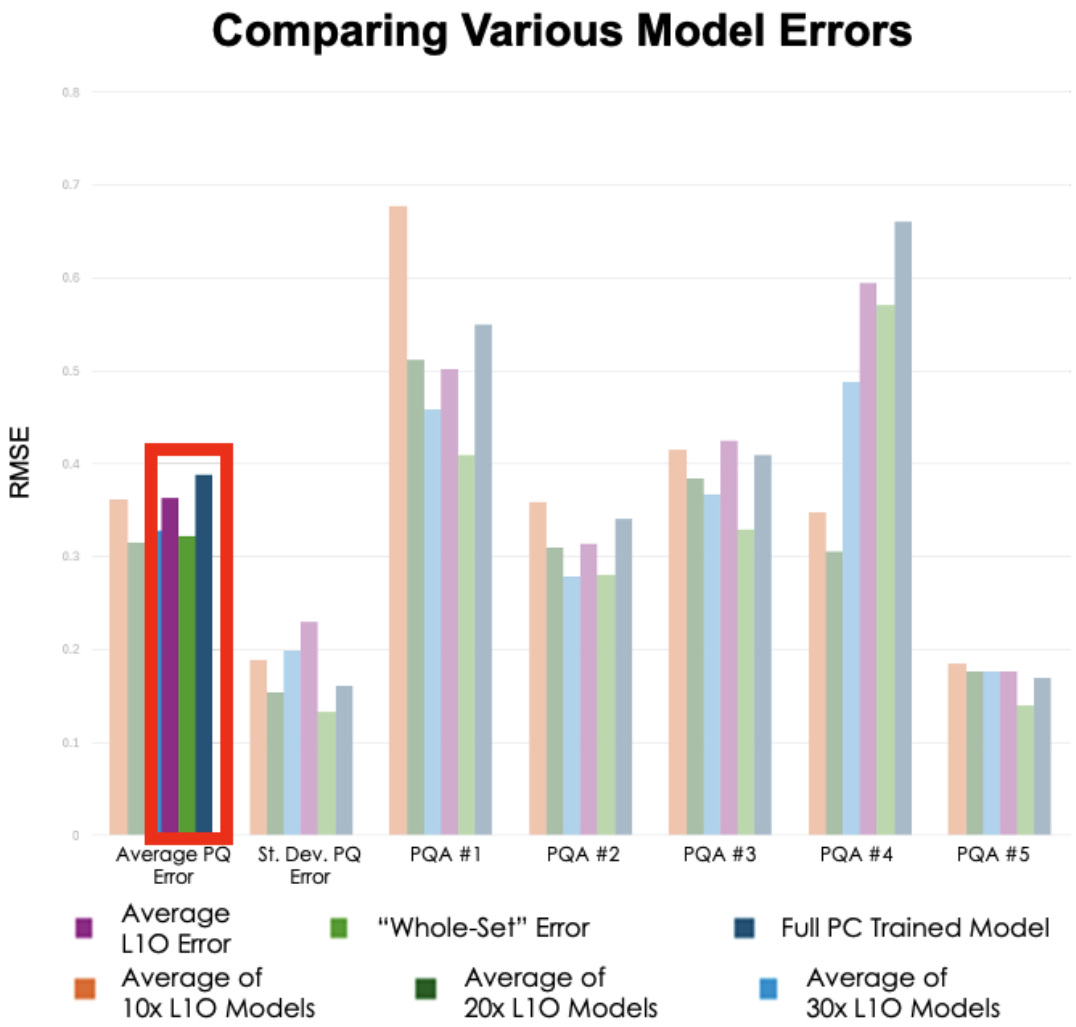

Figure 4-1: Comparison of whole-set error, cross-validation error, and leave-one-out (L1O) error estimates across PQAs. L1O error was computed by holding out a single experiment at a time and retraining, and was additionally averaged across repeated fits to quantify variability due to stochastic training effects. These results were used to contextualize whole-set error and verify that whole-set learning curves were not driven by pathological overfitting.

design, and again after all data have been collected to fit a response surface model relating process parameters to each PQA. The output is a per-PQA R^2 value (and associated RMSE) that summarizes how well the linear and quadratic model explains the observed variation. This value is computed routinely and is already familiar to PC scientists, making it a natural reference point.

However, it is important to understand what this baseline represents and what it does not. The JMP R^2 is not formally used today as a criterion for whether a PC study has "succeeded"; it is better described as a byproduct of the analysis—something that falls out of the standard workflow rather than something that drives decisions. In practice, the success of a PC study is judged by whether the results support defensible operating ranges and a credible control strategy, not by whether the regression model achieves a particular R^2 threshold. Nonetheless, the JMP RMSE for each PQA provides a concrete, already-computed number against which hybrid model accuracy can be compared directionally, and it is shown as a horizontal reference line in the learning curves throughout this chapter.

The JMP model also has a well-understood structural framework, which is substantially different from that of DHL hybrid modeling. In this framework, JMP captures only linear and quadratic main effects and two-factor interactions, it takes as inputs a small number of initial-condition setpoints rather than full time-series trajectories, and it builds a separate model for each PQA independently. By contrast, the DHL hybrid model ingests measured time-series data across the full bioreactor run, encodes mechanistic mass-balance constraints in its propagation component, and can learn shared structure across PQAs through its multi-output architecture (Figure 4-2). These architectural advantages mean that, given appropriately constructed input data, DHL would be expected to match or exceed JMP's accuracy, and the central question of this thesis is how many experiments are needed to realize that expectation.

The comparison between the two models carries an important asymmetry: JMP is always evaluated on its fit to the full PC dataset, while the hybrid model is trained and evaluated on progressively smaller subsets. This means the JMP baseline is never stress-tested under the same data constraints imposed on the hybrid model. In an

idealized study design, JMP would be retrained on each subset as well, generating a parallel learning curve that reveals where response surface regression breaks down. Given that a central composite design requires a minimum number of runs to estimate main effects, quadratic terms, and two-factor interactions, one would expect JMP to become under-determined more quickly (at a larger number of experiments) than a hybrid model degrades—making the comparisons presented here conservative with respect to hybrid modeling's relative advantage. Generating matched JMP learning curves is recommended for follow-on work (see Section 6.3).

The comparison between the two is therefore useful both as a practical benchmark for stakeholders who are familiar with JMP and as a conceptual illustration of the structural differences between the approaches, though the models are fundamentally different in architecture and should not be interpreted as a controlled apples-to-apples comparison.

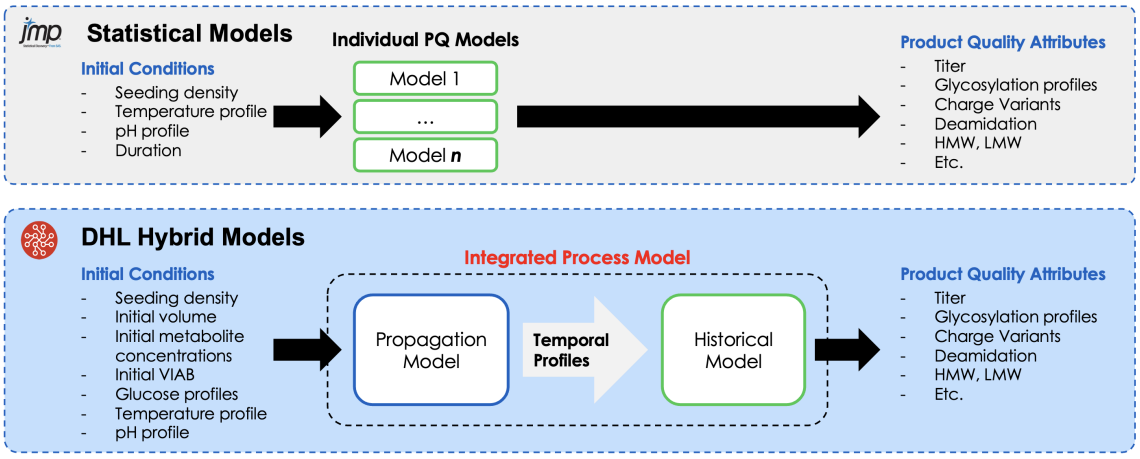

Figure 4-2: Structural comparison of the JMP regression baseline and the DHL hybrid model. JMP fits a separate linear-plus-quadratic response surface per PQA from a small set of initial-condition setpoints, while DHL integrates mechanistic propagation of full time-series inputs with a shared multi-output prediction architecture.

4.1.2 Validation Approach: Leave One Out Cross Validation

Whole-set error was selected as the primary decision metric because it replicates the operational reality of training on all available data to predict future batches while

remaining comparable across training budgets. While leave-one-out (L1O) error is more rigorous as an out-of-sample estimate, it is computationally expensive to compute at scale for repeated subset-selection experiments. For this reason, L1O was calculated selectively and averaged over repeated model fits; these comparisons indicated that the L1O error bounds were consistent with whole-set error, supporting the use of whole-set error for broader learning-curve analysis. Finally, the practical lower bound for model error in this setting is biological replicate error, defined as variability between nominally identical runs. The objective is therefore to approach this floor rather than to drive error toward zero, and the replicate mapping used for this estimate is shown in Figure 4-3.

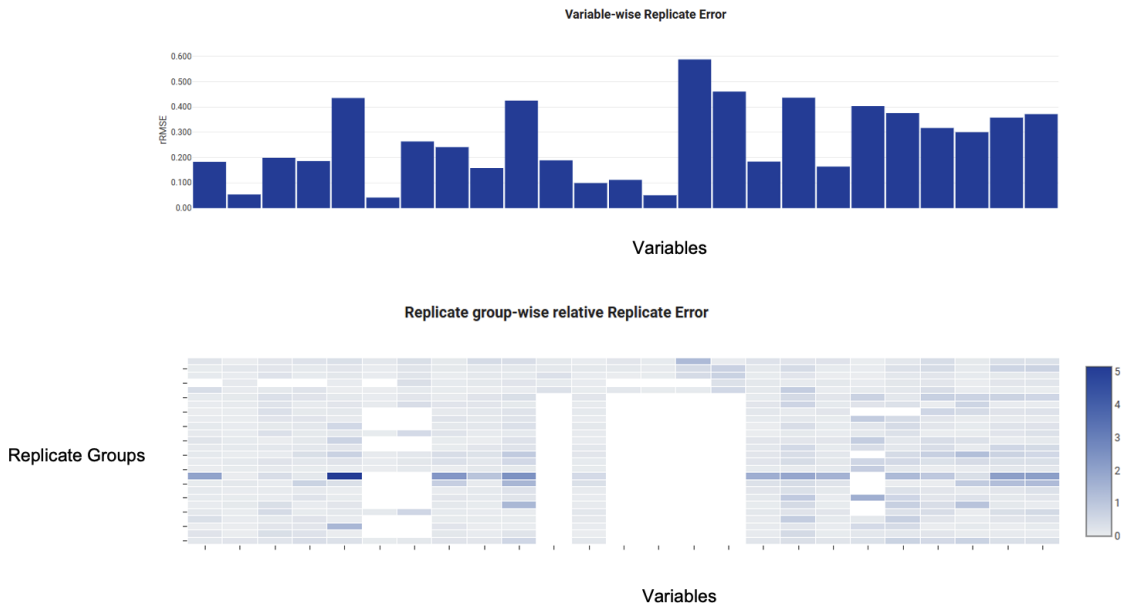

Figure 4-3: Replicate mapping showing variable-wise relative RMSE and how that relative RMSE scaled by variable across replicate groups. This estimate provides a practical lower bound on achievable accuracy given experimental and analytical variability.

4.1.3 Robustness, Limitations, and Reproducibility Considerations

Biological noise sets a limit on achievable predictive accuracy, with the effective noise floor typically falling in the range of 5 to 10 percent depending on the quality attribute and the measurement method. Hybrid models approached this floor for some outputs, such as Titer, but remained higher for more complex attributes. In addition, measurement artifacts and data quality limitations affected robustness. For example, metabolite concentrations occasionally exceeded detection limits, requiring manual data cleaning, and overall robustness was sensitive to the quality of upstream parsing as well as missingness and censoring in time-series inputs.

4.2 Simulation Approaches

Two simulation approaches were evaluated for generating in-silico representations of process characterization conditions. The intent of this comparison was to determine whether setpoint-based "what-if" simulations were sufficient, or whether using measured time-series inputs (including real control variability) was necessary for accurate prediction.

4.2.1 Simulation Approach 1: Perturbing Reference Experiments

In the first approach, process characterization conditions were simulated using a reference (centerpoint) experiment as the base trajectory, and then manually adjusting the relevant input conditions to the target setpoints. For example, to simulate a "high pH" run, the centerpoint experiment was loaded and pH was set to the "high pH" setpoint value for all time periods, while all other input trajectories were held constant.

4.2.2 Simulation Approach 2: Full In-Silico Simulations

In the second approach, process characterization conditions were simulated using the measured time-series values of the input conditions for each experiment, preserving the realized trajectories rather than idealized setpoints. This approach is closer to how the system behaved in the lab, including the run-to-run control variation and transient disturbances that occur in practice.

4.2.3 Comparing Results Between Approaches

In practice, setpoint-based simulations can produce "sticky" predictions: multiple distinct runs with the same nominal setpoints collapse to identical or near-identical simulated inputs, even though real bioreactors exhibit control variation (for example, a setpoint of X realized as $X \pm 10\%$). This effect is amplified when the selected reference vessel is not perfectly representative of a true centerpoint, leading many otherwise distinct runs to map to effectively the same simulated trajectory. As shown in Figure 4-4, the measured-trajectory approach better traces the parity line by preserving the run-to-run variability present in the inputs.

The performance difference is quantified in Figure 4-5. Across the eight PQAs evaluated, Approach 2 (measured trajectories) achieved a mean RMSE of 0.87 compared to 1.72 for Approach 1 (averaged across both reference vessels)—a reduction of approximately 49%. The improvement was most pronounced for PQA 6 (RMSE of 0.79 vs. 2.88) and PQA 8 (1.38 vs. 3.08), where the setpoint-based approach produced errors roughly 2–4× larger; it was smallest for PQA 7 (0.09 vs. 0.19) and PQA 4 (0.87 vs. 1.16), where both approaches performed comparably. Mean absolute bias, by contrast, was similar between approaches (0.48 for Approach 2 vs. 0.51 for Approach 1), indicating that the dominant advantage of measured trajectories is a reduction in prediction scatter rather than a correction of systematic offset. The choice of reference vessel within Approach 1 (R404 vs. R501) had minimal effect on average RMSE (1.73 vs. 1.71), confirming that the performance gap is structural rather than an artifact of a single reference selection. On the basis of these results, Approach 2 was adopted for

all downstream analyses.

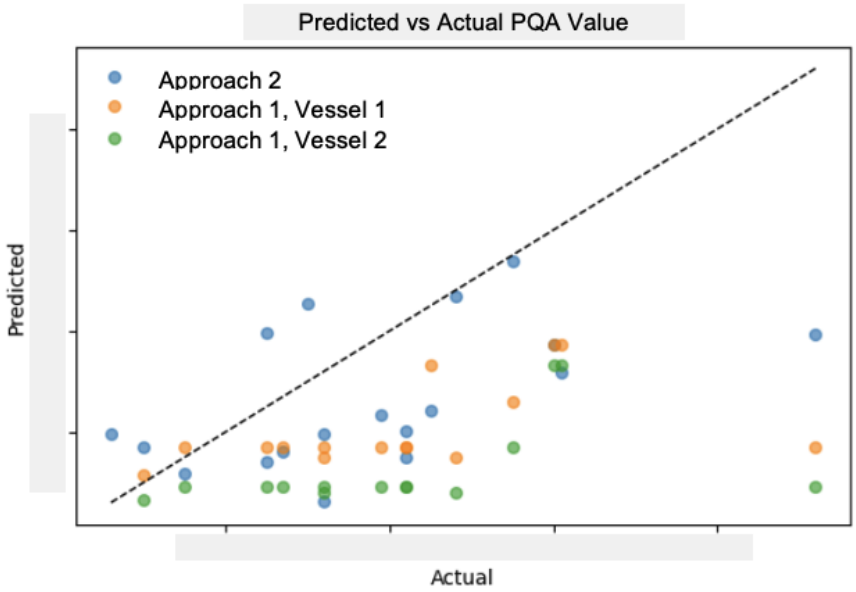

Figure 4-4: Parity plot comparing simulation approaches. The reference-vessel, setpoint-perturbation approach (Approach 1) yields clustered, "sticky" predictions because multiple experiments share identical simulated inputs; using measured time-series inputs (Approach 2) preserves real control variability and improves agreement with the parity line.

The advantage of Approach 2 is not merely that it is "closer to reality"—it is that the ±10% control variability around each setpoint provides the model with local input diversity that is essential for learning. Under Approach 1, experiments that share the same nominal setpoints are presented to the model as identical inputs, even though their outputs differ because the bioreactor actually operated at slightly different conditions. The model cannot explain this output variation from identical inputs; the result is inflated residuals, systematic bias, and, for attributes that are sensitive to small input perturbations, artificially flat response surfaces near cluster centers. Approach 2 resolves this by supplying the realized trajectories, giving the model the input variation it needs to distinguish between nominally similar but operationally distinct runs. The measured-trajectory representation can therefore be understood as providing the ML component of the hybrid model with the local gradient information it requires, whereas the idealized representation collapses that gradient to zero.

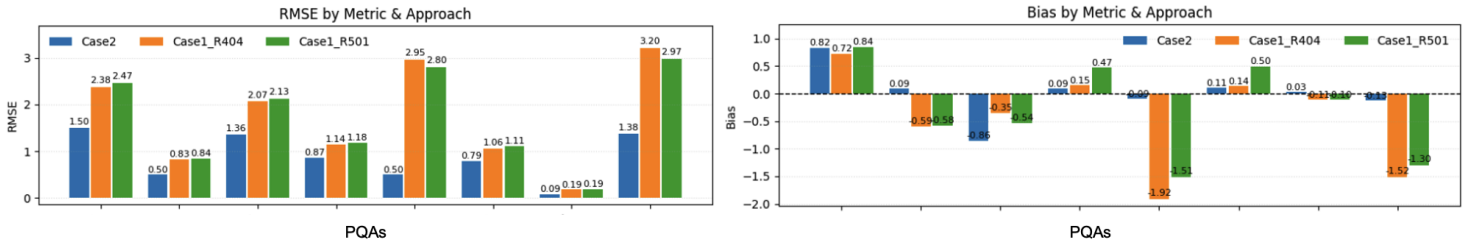

Figure 4-5: RMSE and bias by PQA for the two simulation approaches. Approach 2 (measured trajectories), called "Case 2" in the figure, reduces both average error and systematic offset relative to Approach 1 (setpoint perturbation), called "Case 1" in the figure for the majority of PQAs, with the largest gains in attributes sensitive to within-run trajectory variability. The Approach 1 results are shown for two different reference vessels as setpoints, vessel "R404" and "R501" which both had the same set-points.

4.3 Utilizing Prior Knowledge with Upstream Development Data (Commercial Process Development)

4.3.1 Transferability of Data

Commercial Process Development (CPD) data covered a narrower operating space than the process characterization (PC) data used in this thesis. In particular, CPD experiments lacked variation in process duration and exhibited smaller ranges for pH and temperature than the characterization design space. A PCA comparison shown in Figure 4-6 indicated that the two datasets were measurably distinct. As a result, CPD data can serve as a useful prior, but it cannot replace characterization data because it does not sufficiently stress system boundaries, including failure modes at high pH, that are typically required for regulatory characterization.

4.3.2 Upstream Data Trained Models

Models trained only on CPD data captured the directionality of several major effects when evaluated against later-stage conditions, such as increased pH being associated with reduced titer. However, absolute predictions were systematically offset, indicating a magnitude calibration gap. Without exposure to boundary-stressing PC experiments, these models lacked the information required to predict exact concentrations and quality attributes across the broader operating space. The results of these models can

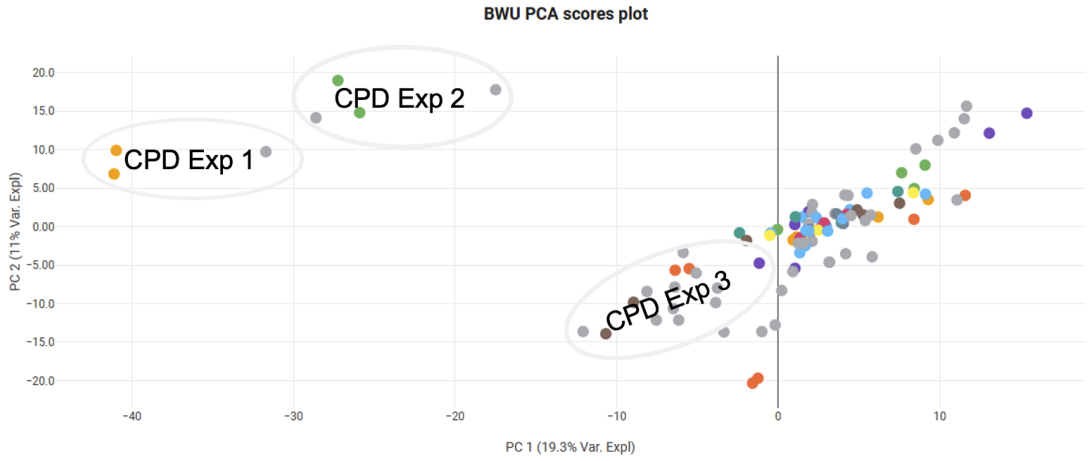

Figure 4-6: PCA mapping of experimental runs, with CPD experiments circled and labeled. CPD occupies a narrower region of the operating space than PC, supporting its role as a prior rather than a substitute for characterization.

be seen in Figure 4-7, alongside the intermediate propagation model outputs which can be seen in Figure 4-8.

4.3.3 Augmenting Upstream Data with New Experiments

Integrating CPD data with a subset of characterization data materially reduced prediction error. When CPD was combined with progressively larger PC training sets, parity behavior tightened and systematic offsets decreased across many PQAs. Under the evaluation conventions used in this work, models trained on CPD plus the full characterization set achieved substantially improved agreement with the parity line, with multiple attributes approaching $\pm 10\%$ error bands. The results of these models can be seen in Figure 4-9.

4.3.4 Model improvement with exposure to all PC data

Moving from CPD-only training to CPD plus PC training resulted in a universal reduction in RMSE, as can be seen in Figure 4-10 and Figure 4-11. These results support the conclusion that CPD is a valid prior that accelerates convergence toward a desired accuracy threshold, but it cannot stand alone for characterization. In practice,

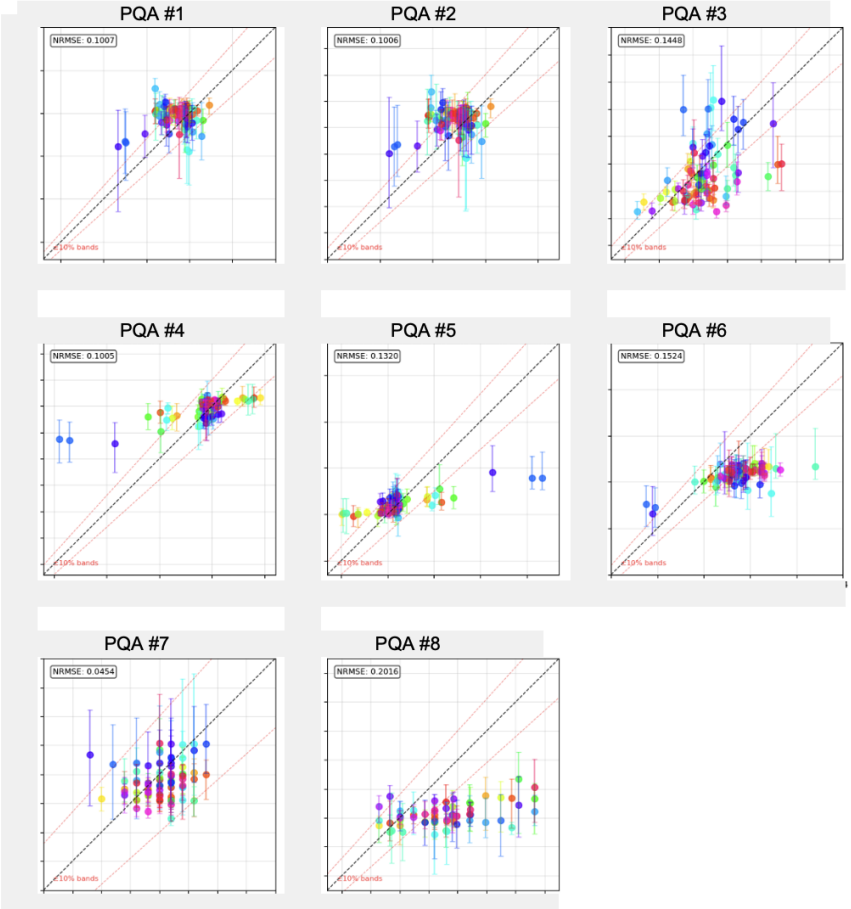

Figure 4-7: Parity plots (predicted vs. actual) for a CPD-trained model evaluated across PQAs. The model captures broad directional behavior but shows systematic offset and reduced accuracy when applied to the broader PC operating space.

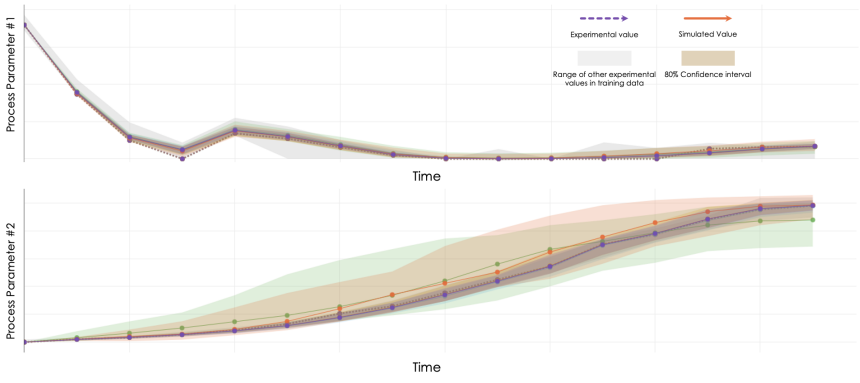

Figure 4-8: Example propagation-model outputs for intermediate state variables. Simulated trajectories are compared to experimental time-series measurements; shaded bands summarize uncertainty and/or the range of trajectories implied by the model ensemble, illustrating how propagation performance underpins downstream PQA prediction quality.

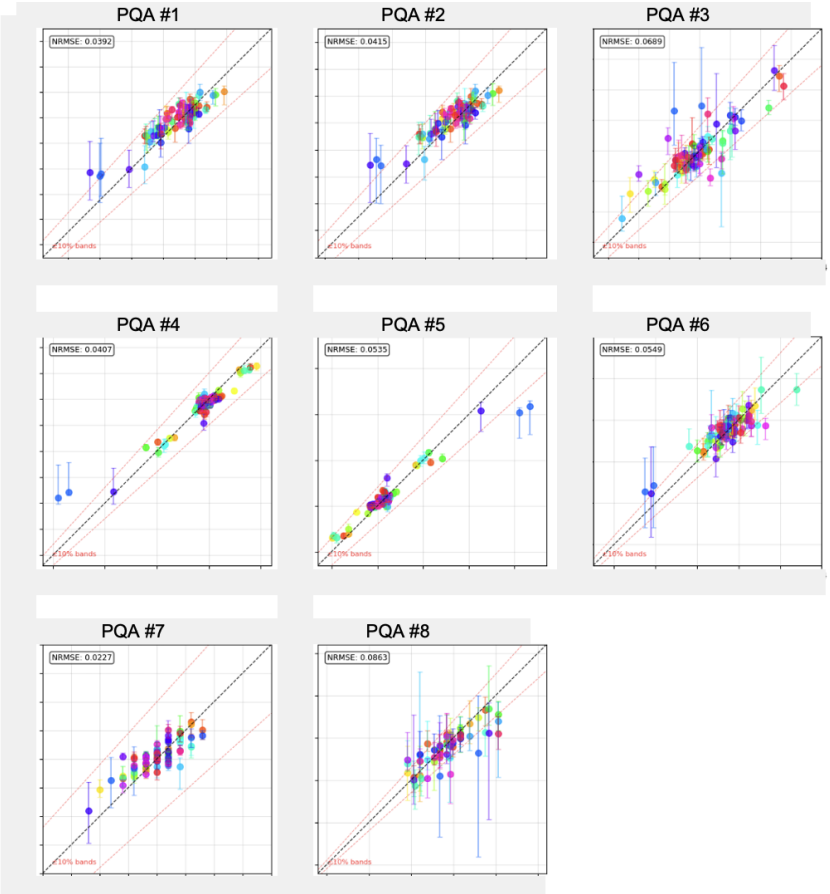

Figure 4-9: Parity plots (predicted vs. actual) for a model trained with CPD data augmented by characterization data. Adding PC exposure reduces systematic offset and improves agreement with the parity line across PQAs relative to CPD-only training (cf. Figure 4-7).

upstream priors can reduce the number of experiments required in a subsequent characterization study, but they do not eliminate the need for PC experiments that cover the broader design space.

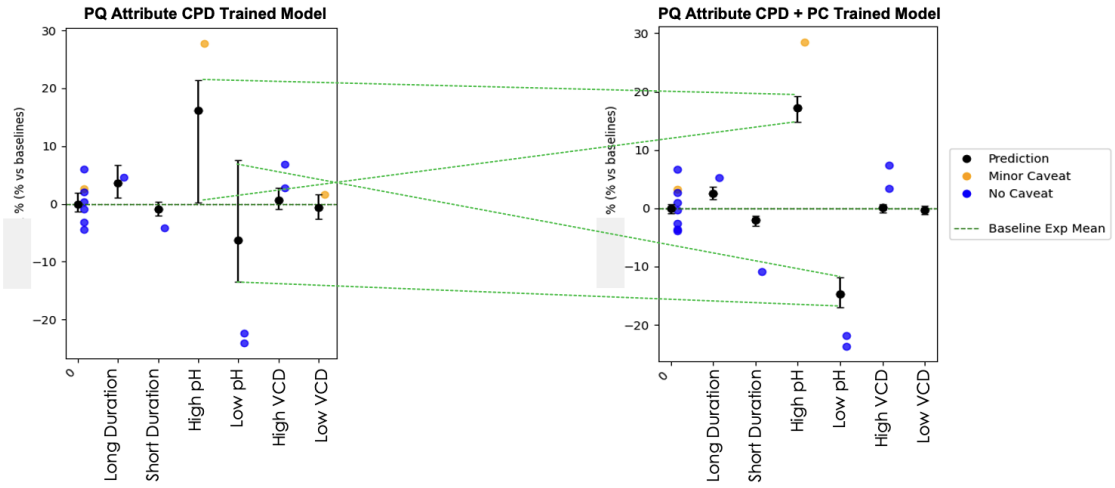

Figure 4-10: Illustrative single-PQA comparison showing how adding PC data improves calibration across boundary-stressing conditions. The CPD-only model captures qualitative shifts but misestimates magnitude under conditions not represented in CPD; the CPD+PC model reduces this gap and better matches the baseline experimental mean across conditions.

4.4 Training Data Sufficiency Analysis

This section quantifies how predictive performance scales with the number of PC experiments used for training and compares subset selection strategies under a consistent whole-set evaluation framework. Learning curves are interpreted relative to two practical benchmarks: current-state regression performance (JMP) as a directional threshold, and replicate error as an empirical lower bound driven by experimental variability.

4.4.1 "General" Models

The average model error as a function of training set size can be seen in Figure 4-12.

Across PQAs, learning curves generally flattened around approximately 35 training

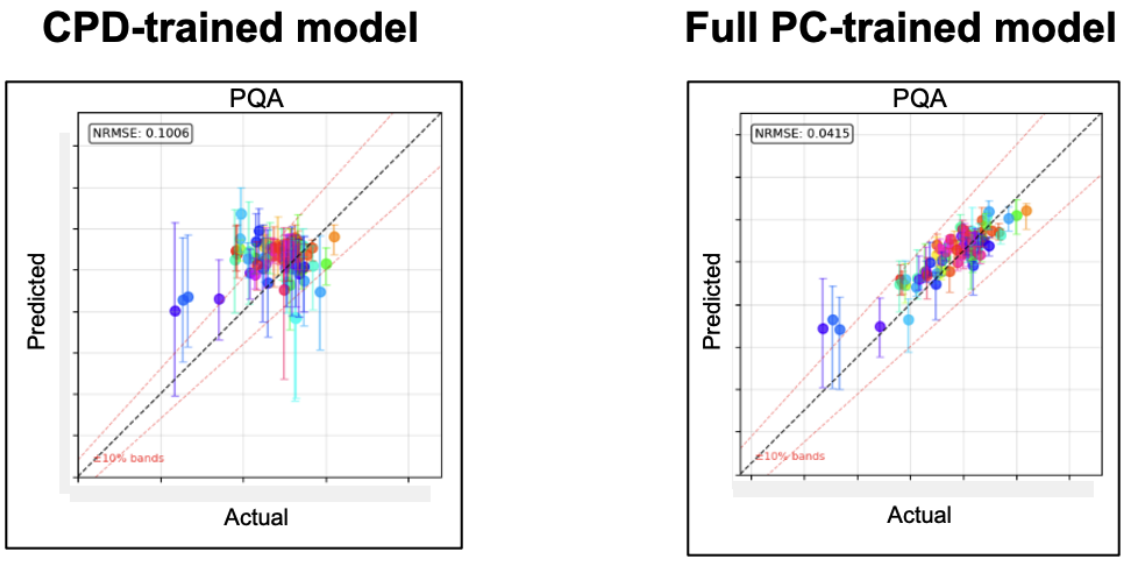

Figure 4-11: Representative parity comparison for a single PQA: CPD-trained model versus full PC-trained model. Training on the full characterization dataset improves parity behavior and reduces RMSE, highlighting the need for late-stage boundary coverage even when strong upstream priors exist.

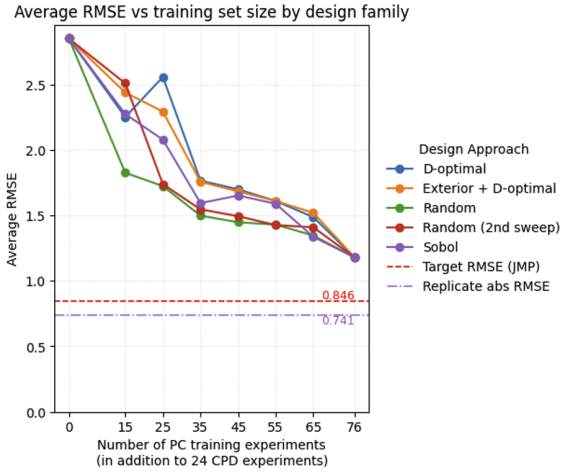

Figure 4-12: Average RMSE as a function of training set size for different subset selection strategies (design families). Horizontal reference lines indicate a directional JMP regression baseline and replicate error. Curves flatten around the mid-30s in training experiments, suggesting diminishing returns beyond this region for average multi-PQA performance.

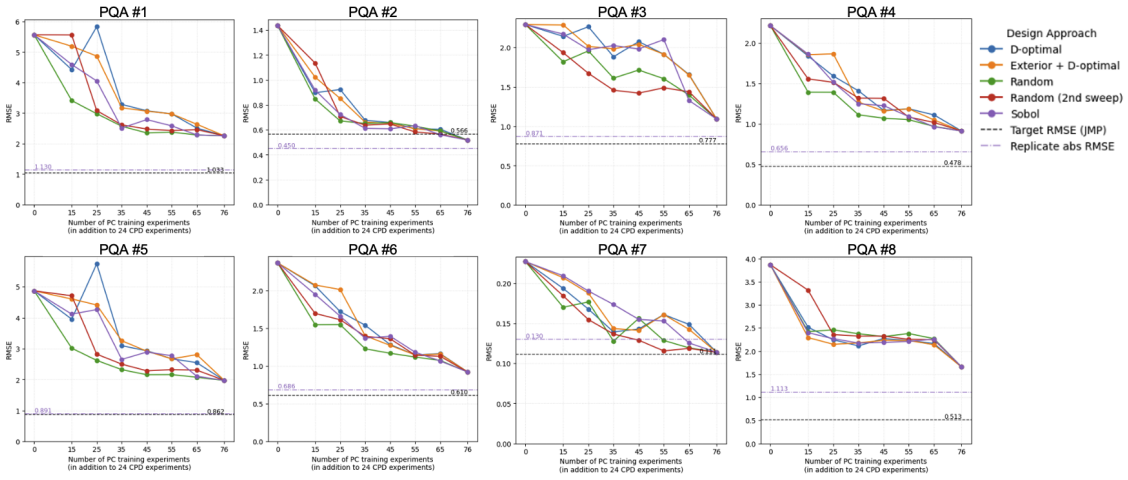

Figure 4-13: Learning curves (RMSE vs. training set size) for each PQA, comparing subset selection strategies. While many attributes show rapid early improvements, convergence behavior differs meaningfully by PQA, motivating the categorization discussed in Subsection 4.4.1.

Table 4.1: Whole-set RMSE of the hybrid model (median across sampling strategies) at selected training set sizes, compared to the JMP regression baseline and biological replicate error for each PQA. Training set sizes indicate the number of PC experiments added to 24 CPD experiments. Shaded cells indicate where the hybrid model RMSE is within 10% of, or below, JMP RMSE.

	Number of PC Training Experiments (+24 CPD)									
PQA	**0**	**15**	**25**	**35**	**45**	**55**	**65**	**76**	**JMP**	**Repl.**
PQA 1	5.5	5.5	4.5	3.0	2.7	2.7	2.5	2.3	1.033	1.130
PQA 2	1.4	0.9	0.7	0.65	0.62	0.60	0.58	0.56	0.566	0.450
PQA 3	2.2	1.8	1.5	1.2	1.1	1.1	1.0	0.9	0.777	0.871
PQA 4	2.2	1.5	1.3	1.15	1.05	1.0	0.9	0.85	0.478	0.656
PQA 5	4.9	4.5	3.5	3.0	2.8	2.5	2.2	2.1	0.862	0.891
PQA 6	2.3	1.7	1.5	1.3	1.2	1.15	1.0	0.9	0.610	0.686
PQA 7	0.22	0.15	0.13	0.12	0.11	0.11	0.12	0.10	0.111	0.130
PQA 8	3.9	2.5	2.4	2.3	2.3	2.2	2.1	1.9	0.513	1.113
Avg.	2.8	2.4	2.3	1.7	1.5	1.4	1.25	1.15	0.846	0.741

Notes: All values are absolute RMSE (native PQA units). Column "0" represents CPD-only models (24 CPD experiments, 0 PC experiments). "JMP" is the RMSE of the JMP regression fit to the full CCD dataset (read from labeled annotations on each subplot). "Repl." is the biological replicate error (noise floor). Hybrid model values represent the median across sampling strategies; individual strategies may differ substantially (see Figure 4-13).

experiments, suggesting that the remaining experiments in a traditional ∼90-run campaign yield diminishing returns for predictive accuracy under this evaluation framework. Within this overall pattern, three learning behaviors were observed in Figure 4-13 and Table 4.1. For some outputs such as PQA number 2, models reached the regression accuracy threshold with substantially fewer experiments than the full set. For other outputs, such as PQA number 8, models often converged quickly but stalled above the baseline, suggesting structural limitations rather than data scarcity. Finally, a subset of attributes, like PQA number 6, showed more consistent improvements with additional training experiments, indicating that they would likely benefit from expanded datasets and/or targeted experimentation.

This pattern (quick model convergence around 35 experiments) is consistent with the hybrid model's mechanistic structure providing an effective form of regularization: by encoding mass-balance constraints and biological priors, the model requires fewer observations to learn the residual data-driven component, such that the most informative experiments (those that span distinct metabolic regimes or stress key parameters) are captured early in the training sequence. The remaining experiments in a traditional campaign contribute increasingly redundant information, reflected in the diminishing slope of the learning curves beyond this region. A more detailed interpretation of these convergence behaviors, including PQA-specific differences, is provided in Section 5.1.

4.4.2 Marginal Efficiency of Additional Experiments

To quantify the diminishing-returns pattern visible in the learning curves, Table 4.2 compares the marginal RMSE reduction per experiment in the first 35 PC training runs against the subsequent 41 runs. Across PQAs, the first 35 experiments captured between 68% and 89% of the total RMSE improvement observed over the full 76-experiment training sequence, despite representing fewer than half of the runs. On average, each of the first 35 experiments reduced mean RMSE by 0.031 units, whereas each of the remaining 41 experiments contributed only 0.013 units—a 2.3× difference in marginal efficiency. Notably, as seen in Figure 4-13 and Table 4.1, the point selection

methodology did not materially affect this result, which speaks to the robustness of the modeling framework itself.

Table 4.2: Marginal efficiency of PC training experiments: RMSE improvement per experiment for the first 35 versus the remaining 41 runs, by PQA. "% in first 35" indicates the fraction of total RMSE improvement (from k=0 to k=76) captured by the first 35 experiments.

	ΔRMSE		**RMSE / expt**		
PQA	**First 35**	**Last 41**	**First 35**	**Last 41**	**% in first 35**
PQA 1	2.50	0.70	0.071	0.017	78%
PQA 2	0.75	0.09	0.021	0.002	89%
PQA 3	1.00	0.30	0.029	0.007	77%
PQA 4	1.05	0.30	0.030	0.007	78%
PQA 5	1.90	0.90	0.054	0.022	68%
PQA 6	1.00	0.40	0.029	0.010	71%
PQA 7	0.10	0.02	0.003	0.0005	83%
PQA 8	1.60	0.40	0.046	0.010	80%
Avg.	1.10	0.55	0.031	0.013	67%

Notes: "First 35" spans k=0 (CPD-only) to k=35; "Last 41" spans k=35 to k=76 (full PC set). RMSE values are medians across sampling strategies from Table 4.1. The average row uses the average RMSE across PQAs rather than the average of per-PQA ratios.

This asymmetry has direct operational implications. If a future PC study were designed around the hybrid modeling workflow, a campaign of approximately 35 experiments (plus available prior information from CPD experiments or other programs, in this case 24 experiments) would capture the large majority of attainable predictive improvement for most PQAs.[1] The additional 40+ experiments in a conventional ∼90-run campaign contribute marginal accuracy gains that are, for several attributes, comparable in magnitude to biological replicate noise. This finding underpins the workload-reduction estimate discussed in Chapter 5.

[1]This estimate is conditioned on the 24 CPD experiments available as prior data in the present study. The CPD dataset occupied a narrower region of the operating space than the PC design (see Figure 4-6), lacking variation in process duration and covering smaller ranges for pH and temperature. If richer upstream data were available—for example, from a CPD campaign that more broadly stressed system boundaries or from a completed PC study on a related molecule—the prior would provide stronger initial calibration, potentially shifting the inflection point to fewer than 35 PC experiments. Conversely, if no upstream prior were available, the model would need to learn both directionality and magnitude entirely from PC data, likely requiring additional experiments before the same diminishing-returns behavior emerged. This sensitivity to prior data quality and coverage is a specific instance of the broader transfer learning dynamics discussed in Subsection 6.4.3.

4.4.3 PQA-specific models

It was hypothesized that training separate models for each PQA, rather than one multi-output model, would improve predictive accuracy. These models were built by re-optimizing the hyperparameters from the general historical models, focused on one specific PQA at a time. The propagation models used were unchanged. In practice, as shown in Figure 4-14, attribute-specific models did not show faster learning or lower final error compared to general multi-output models. This result suggests that shared latent structure, and in particular the propagation model that predicts biomass and metabolites over time, is a primary driver of performance. Improving this shared backbone appears to be more impactful than tuning individual PQA-specific heads.

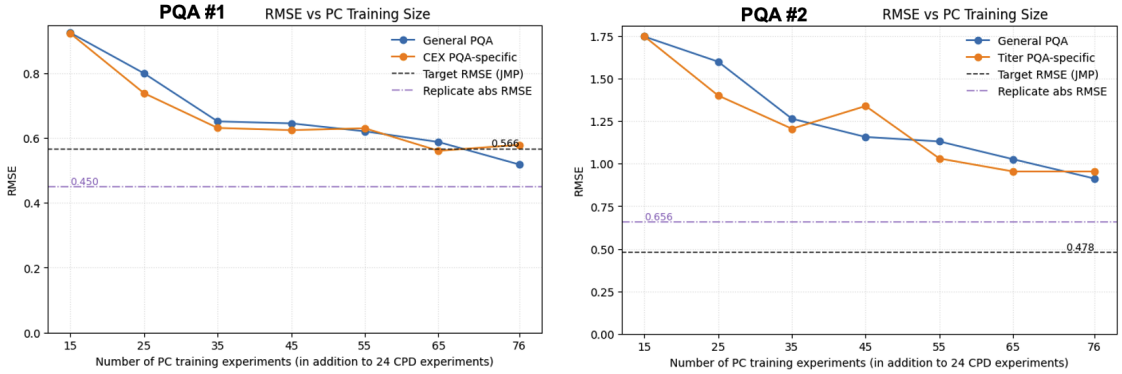

Figure 4-14: Comparison of general multi-output models versus PQA-specific models across training set sizes. PQA-specific heads did not consistently outperform the shared multi-output configuration, suggesting the dominant performance lever is the shared propagation backbone rather than attribute-level specialization.

4.5 Generalizability, Transfer Learning Findings, and Model Performance Across Programs

4.5.1 Sampling Strategy Performance

To isolate the effect of training-set composition from training-set size, the learning curves in Figure 4-12 and Figure 4-13 were disaggregated by subset-selection strategy.

Several patterns emerged that were not anticipated at the outset.

Exterior point selection—which prioritizes experiments near the boundaries of the design space and fills remaining slots with interior points—was expected to perform well, given that boundary coverage is emphasized in conventional PC practice to establish proven acceptable ranges. In practice, however, exterior-first selection was among the weaker strategies at most training budgets and for most PQAs. At small training set sizes (15–25 PC experiments), exterior selection frequently produced the highest RMSE of any strategy, and it rarely produced the lowest. This pattern was consistent across PQAs and was most pronounced for attributes where the propagation model's accuracy depended on capturing nonlinear metabolic dynamics in the interior of the operating space rather than at its edges.

Sobol sequence sampling exhibited the most consistently strong performance across PQAs and training budgets. While it was not always the single best strategy for any individual PQA, it was rarely the worst, producing learning curves that tracked near or below the median across strategies at every value of k. This consistency is notable because Sobol sequences are deterministic given a fixed seed, which was kept constant across different k values, meaning the result is not sensitive to random-draw variability.

Random sampling produced more variable results, as expected for a stochastic method. To assess whether observed performance differences were artifacts of a single draw, two independent random sweeps were conducted using different seeds. The two sweeps produced qualitatively similar learning-curve shapes but differed meaningfully in absolute RMSE at small training budgets, confirming that random selection introduces non-negligible variance into the evaluation. For certain PQAs, one or both random draws outperformed all structured strategies at intermediate training set sizes—a result that appears paradoxical but likely reflects cases where the random draw happened to sample informative interior regions that the structured methods under-weighted.

D-optimal selection performed comparably to Sobol at moderate-to-large training budgets but showed more PQA-to-PQA variability at small k, likely because its greedy

construction can over-emphasize directions of high linear information content that do not align with the nonlinear features most relevant to certain quality attributes.

4.5.2 Feature Importance and Physical Interpretation

Feature importance analysis revealed distinct drivers for different PQAs, which can be seen in the circled regions of Figure 4-15. These differences suggest that performance variation across PQAs is linked to how well the propagation model predicts the intermediate state variables most relevant to each attribute.

To illustrate why this matters, consider a hypothetical example. For a PQA such as final titer (total protein produced), one would expect the most influential features to be related to nutrient availability and feeding strategy (variables like glucose concentration, feed rate profiles, and media composition over time) because these directly govern how much metabolic energy the cells can allocate to protein production. By contrast, for a PQA such as final cell viability, the dominant features are more likely to be initial cell density and early-stage growth conditions, because viability at harvest is largely determined by whether the culture entered the decline phase during the run, which in turn depends on how the population was established in the first few days. If the propagation model accurately predicts glucose consumption dynamics but poorly captures early growth kinetics, one would expect strong predictive performance for titer-related PQAs but weaker performance for viability-related PQAs, even though both are trained on the same data and use the same model architecture.

This pattern is consistent with what was observed. The PQAs for which the hybrid model performed best tended to be those whose dominant features aligned with the intermediate state variables that the propagation model predicts well (e.g., metabolite trajectories). The PQAs where performance lagged tended to depend on features that the propagation model captures less accurately or that are influenced by unmeasured inputs. This connection between feature importance and propagation model fidelity reinforces the finding from Section 4.4.3: improving the shared mechanistic backbone is likely to be more impactful than tuning individual PQA-specific output heads, because the backbone determines how well the features most relevant to each PQA

are represented.

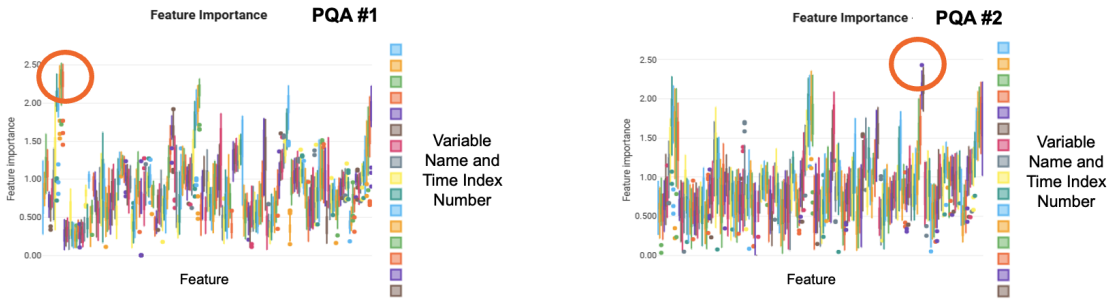

Figure 4-15: Feature importance (time-indexed variables) for two representative PQAs. Differences in dominant features suggest that variation in PQA performance is tied to which intermediate state variables are most influential for each downstream prediction.

Chapter 5

Discussion

This chapter discusses the scientific, operational, and strategic implications of applying hybrid machine learning to process characterization at AMGEN. The findings demonstrate that hybrid models can accurately predict certain product quality attributes using significantly fewer experiments than traditional regression methods, though performance varies by attribute. This discussion explores the underlying reasons for these variations, proposes new frameworks for future study designs, and outlines the regulatory and change management steps required to integrate these tools into standard biopharmaceutical process development.

5.1 Scientific Insights

The results presented in Chapter 4 reveal a nuanced picture of hybrid model performance: for some product quality attributes, our models achieved accurate predictions with as few as 35 experiments, representing a 54% reduction relative to the 75 experiments required by JMP regression models to reach comparable accuracy. For other PQAs, however, the hybrid models never matched the accuracy of JMP, even when trained on the full dataset.

Several factors may explain this variability. First, the experimental designs used in the retrospective studies were central composite designs originally optimized for response surface regression, not for the sequential, information-dense sampling that

hybrid models favor. While our hybrid models should, in principle, be able to replicate JMP performance given the same data, the mismatch between the sampling strategy and the model's learning dynamics may have limited convergence for certain outputs. Second, noise in the underlying PQA measurements, whether from analytical variability, biological stochasticity, or batch-to-batch heterogeneity, likely affects the hybrid ML and JMP models differently. JMP's low-order polynomial fits can smooth over measurement noise, whereas neural-network-based models may attempt to fit it, particularly when data are scarce. Third, the propagation (hybrid-mechanistic) and historical (data-driven) components of the DHL hybrid modeling architecture were tuned independently in this work; joint optimization of both components could potentially improve performance on the more difficult PQAs. It is hypothesized that in an "apples-to-apples" comparison between DHL and JMP, which would involve a historical model trained only on input setpoints as Z variables and outputs as Y variables, DHL would surpass JMP performance. A complementary test of this hypothesis would be to retrain JMP models on the same data subsets used for the hybrid model learning curves, rather than on the full dataset as done here — a comparison expected to reveal the point at which JMP's response surface regression becomes under-determined and degrades, potentially well before (at a larger number of experiments) hybrid model performance does (see Section 3.7). This would require future analysis, as discussed in Section 6.3.

While prior industrial work has demonstrated that hybrid models can be applied to process development data from major manufacturers [17, 20], process characterization represents a qualitatively different and more demanding test case. PC studies are the evidentiary backbone of biologics regulatory submissions: they are guided by ICH Q8 process validation principles, produce data under GMP-qualified analytical methods, and must support defensible claims about proven acceptable ranges and process design space. To the author's knowledge, no published study has previously evaluated hybrid model performance in this specific context. This work addresses that gap, and does so under conservative conditions, applying models to experimental designs that were not optimized for machine learning and to a dataset that reflects the full heterogeneity

of industrial PC execution rather than curated academic benchmarks. Critically, this thesis reports both successes and failures with equal fidelity. The willingness to characterize the conditions under which these models fall short is itself a contribution: it replaces the selective optimism that characterizes much of the hybrid modeling literature with an empirical foundation on which realistic deployment decisions can actually be made.

5.1.1 Model Performance and Applications

The central finding of this thesis is that a hybrid modeling approach can, for certain PQAs, dramatically accelerate the path to predictive accuracy. In the best cases, models trained on approximately 35 experiments achieved prediction errors comparable to or better than JMP models trained on the full 76-experiment design. As quantified in Table 4.2, these first 35 experiments captured 68–89% of total RMSE improvement per PQA, with each experiment contributing roughly 2.3× more error reduction than those in the remaining campaign. This corresponds to a potential 54% reduction in experimental burden for those attributes, which would translate directly into shorter study timelines and reduced consumable costs (see Section 5.3).

However, performance was not uniform across all PQAs. For attributes with strong, relatively linear relationships to the process parameters, JMP's response surface models performed well and the hybrid models offered limited incremental benefit. For more complex, nonlinear responses, the DHL models showed advantages only after a sufficient number of training points had been accumulated. For a third category of PQAs, the hybrid models always underperformed JMP. These cases tended to involve attributes with high measurement noise or weak sensitivity to the varied process parameters, suggesting that the mechanistic component of the hybrid model did not provide enough inductive bias to overcome the data limitations.

The finding that PQA-specific models did not outperform general multi-output models (Subsection 4.4.3) was contrary to our initial hypothesis and warrants explicit interpretation. Because both model types share the same propagation backbone, the result implies that the bottleneck for most PQAs is the quality of the intermediate state

predictions — not the flexibility of the output mapping. This is consistent with the feature importance analysis in Subsection 4.5.2, which showed that the PQAs with the weakest predictions were those most dependent on state variables that the propagation model captures poorly. In this architecture, tuning the output head independently cannot compensate for upstream prediction errors, which is why joint optimization of the propagation and historical components (discussed in Subsection 5.1.3) is a more promising path than PQA-level specialization.

The retrospective nature of this analysis is both a strength and a limitation. It is a strength because it tests models against real, production-grade data rather than idealized simulations. It is a limitation because the experimental designs were not tailored to hybrid model learning. A prospective study using a sequential, model-guided design (see Section 5.2) would likely show larger and more consistent improvements.

5.1.2 Interpreting Sampling Strategy Results

The finding that exterior point selection underperformed space-filling strategies challenges a deeply held intuition in process characterization: that boundary-stressing experiments are the most informative for building process understanding. This intuition is well-founded for the purpose exterior points actually serve in classical PC—namely, demonstrating that the process remains within specification at the edges of the operating range, which is fundamentally a testing objective. However, when the objective shifts from testing to learning (training a model that must capture nonlinear, time-dependent dynamics across the full operating space) boundary experiments contribute disproportionately less information per run than interior points that sample the transition regions where metabolic behavior shifts. The hybrid model's propagation component relies on learning rate functions (e.g., growth rate, specific productivity) that vary continuously across the design space; experiments clustered at the extremes provide limited signal about the shape of these functions in the regions where most future manufacturing operation will occur. This interpretation is further supported by the observation that Sobol sequences, which distribute points quasi-uniformly throughout the space, produced the most reliable performance across

PQAs and training budgets—exactly the behavior expected when the bottleneck is capturing interior nonlinearity rather than confirming boundary behavior. For future PC studies employing hybrid models, this evidence supports designing initial experimental blocks around space-filling criteria rather than boundary-first strategies, reserving targeted boundary experiments for later blocks once the model has learned the dominant internal dynamics.

5.1.3 Remaining Open Questions

Several open questions emerge from this work that merit investigation in future studies.

Impact of experimental design on model accuracy. The retrospective datasets used here were generated from central composite designs. As discussed in Section 5.2, space-filling designs such as Latin Hypercube sampling are expected to be more compatible with hybrid model learning. Quantifying how much model accuracy improves under purpose-built experimental designs is a critical next step.

Role of categorical variables. The current models were trained on continuous process parameters only. Incorporating categorical variables such as raw material lot, experiment batch number, etc. could improve robustness across batches and reduce unexplained variance that currently manifests as prediction error. That said, adding categorical variables is a double-edged sword: each new category fragments the training data, reducing the effective sample size available per group and potentially requiring additional experiments to maintain coverage. Where possible, a preferable approach may be to replace categorical variables with continuous surrogates that capture the underlying variation; for example, encoding raw material lot identity through measured lot properties recorded on the batch ticket, rather than as an opaque category label.

Transferability across molecules. This work focused on a single molecule. A key question for enterprise deployment is whether trained models, or components

thereof, can transfer to new molecules expressed in the same cell line. In the current framework, molecule identity is treated essentially as a categorical variable, which limits the model's ability to generalize across products. A more principled approach would develop continuous representations of each molecule, derived from physicochemical descriptors, entity embeddings, or protein language model features, that encode structural similarity and allow the model to weight historical data from related molecules accordingly [21]. Narayanan et al. explored this direction in a Takeda collaboration, using *in silico* molecular descriptors as GP inputs for cross-molecule prediction of chromatographic performance, though the limited molecular diversity in their training set (five mAbs) constrained what could be learned [27]. Applying analogous representation strategies to upstream cell culture models, where molecule-specific growth kinetics and product quality responses may correlate with biophysical properties, is a natural extension that should be tested as cross-molecule datasets are assembled from AMGEN's large existing, and growing, portfolio of characterized molecules (see Section 5.5).

Joint tuning of model components. The DHL architecture combines a mechanistic propagation model with a historical data-driven model. In this work, these components were tuned separately. As discussed in Subsection 4.5.2 and Subsection 4.4.3, there may be some benefits to be derived from tuning a pair of models together to predict specific PQAs. This could help address our failure to prove the hypothesis that PQA specific models would outperform the general models used in this study.

Alternative model architectures. This thesis evaluated a single hybrid model architecture. Other architectures deserve comparison, including multi-layer perceptrons, Gaussian process models, and simpler explainable models such as gradient-boosted trees, which may work well with incorporating batches as a categorical variable. It is also worth investigating whether existing tools like JMP's neural network platform or other commercial software could be configured to achieve similar results. Although we

believe DHL offers the best combination of mechanistic interpretability and predictive power for this application, a rigorous head-to-head comparison across model families is essential before committing to a single platform for enterprise deployment.

Direct input–output comparison with DHL and JMP As mentioned earlier in Section 5.1, a more controlled comparison between JMP and DHL would involve training a DHL historical model using only Z variables as inputs and Y variables as outputs, omitting the propagation component and the batch-wise unfolded time-series data entirely. This configuration, effectively a PLS model mapping setpoints to product quality within the DHL platform, represents the true DHL analog of the JMP regression approach, as both would operate on the same information set. Performing this comparison in future studies would isolate the contribution of model architecture from the contribution of richer input data, providing a cleaner signal for stakeholders evaluating whether DHL's predictive advantage stems from its hybrid structure, its access to time-series process history, or both.

5.2 Future PC Study Design

The evidence presented in this thesis: that DHL models can predict PQAs accurately with substantially fewer experiments, motivates a fundamental rethinking of how process characterization studies are designed and executed. As discussed in Chapter 1 and Chapter 2, upstream PC teams face growing demand driven by an expanding pipeline and increasing regulatory expectations for process understanding. Meeting this demand with the current experimental paradigm, which relies on large central composite designs and power-based sample size calculations, is unsustainable. A new design philosophy is needed: one that replaces rigid factorial designs with targeted exploration, static designs with adaptive learning, and power analysis with predictive accuracy as the primary stopping criterion.

5.2.1 Alternative DoE Explorations

We recommend replacing central composite factorial designs with Latin Hypercube sampling (LHS) as the initial experimental design for future PC studies. LHS provides superior coverage of the multidimensional parameter space with fewer points than factorial approaches, and its space-filling properties align naturally with the data requirements of hybrid models. The initial design can be generated either through JMP's built-in LHS tools or through the design optimizer module within the DHL platform.

It is important to recognize that this represents an "either-or" decision with respect to the modeling framework. Central composite designs are optimized for estimating the coefficients of second-order polynomial models and will not unlock the full benefit of hybrid modeling. Conversely, a Latin Hypercube design may not provide the structure needed for traditional JMP response surface analysis to perform well. The choice of experimental design and modeling approach must therefore be made jointly at the outset of a study.

The recommended workflow is as follows:

1. Use Latin Hypercube sampling to select an initial set of experiments (e.g., 20–30 runs) that spans the design space.

2. Execute those experiments and train a DHL model on the resulting data.

3. Based on model predictions and uncertainty estimates, identify regions of the design space where additional data would most improve prediction accuracy.

4. Conduct a second block of experiments targeting those regions.

5. Reassess model accuracy and repeat as needed until the desired level of predictive performance is achieved.

The expected benefits of this approach include: reduced total experimentation required to achieve a given level of process understanding; the potential to identify more optimized process conditions (e.g., higher titer setpoints) by exploring the design

space more thoroughly; and improved characterization of the relationships between process variables and PQAs, including nonlinear and interaction effects that quadratic regression models cannot capture.

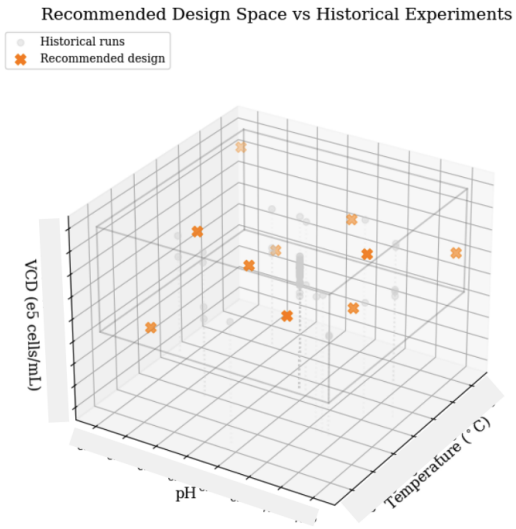

Figure 5-1: Comparison of central composite and Latin Hypercube design space coverage. Latin Hypercube sampling achieves more uniform coverage of the multi-dimensional parameter space, enabling hybrid models to learn more efficiently from fewer experiments.

5.2.2 Future PC Scheduling to Increase Capacity

Beyond changing what experiments are run, there is an opportunity to change *when* they are run. The adaptive, block-sequential nature of hybrid-model-guided experimentation introduces scheduling flexibility that does not exist in the current paradigm. Two scheduling approaches are illustrated in Figure 5-2.

Approach A: Fixed design, accelerated execution. In this approach, the full experimental design is defined at the start of the study using LHS, and all runs are executed sequentially without pausing for model updates between blocks. This is operationally simpler and mirrors the current scheduling cadence, but it forgoes the benefit of adaptive learning between experimental blocks. The primary advantage is a reduction in total run count (and therefore calendar time) relative to the baseline

CCD approach, without requiring changes to the scheduling infrastructure.

Approach B: Adaptive design with interleaved studies. In this approach, an initial block of experiments is executed, after which the team pauses to collect PQA results, update the DHL model, and redesign the next experimental block based on model predictions. During this "learning" downtime—which may span several weeks while analytical results are generated—the bioreactor capacity and scientific staff can be redirected to execute runs for a different PC study. Once the updated design is ready, the team returns to the first study for its next block. This interleaving of studies is more complex to schedule but has two significant advantages: it leverages the adaptive learning capability of hybrid models more fully, and it increases effective throughput by utilizing equipment and personnel during what would otherwise be idle time.

The interleaved approach introduces uncertainty into scheduling because the number of adaptive iterations required to reach sufficient model accuracy is not known in advance. Establishing this through prospective pilot studies is essential before the approach can be adopted as a standard workflow (see Subsection 5.6.1).

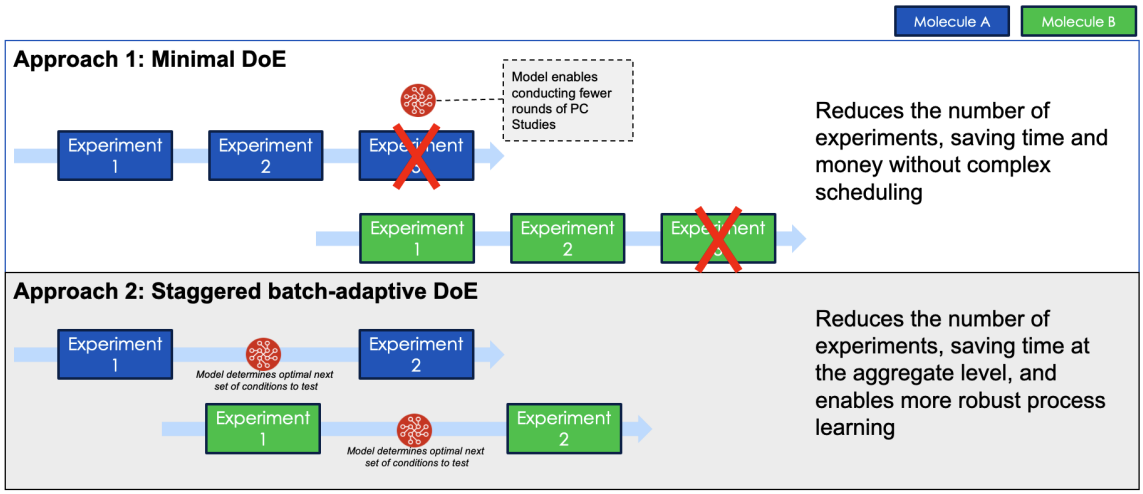

Figure 5-2: Two scheduling approaches for hybrid-model-guided PC studies. Approach A defines the full design upfront and executes without interim model updates. Approach B interleaves experimental blocks with model updates, using the intervening time to advance other studies in parallel.

5.3 Business Impact at AMGEN

This section translates the modeling results of Chapter 3 and Chapter 4 into an operational and economic lens for upstream process characterization (PC) at a hypothetical AMGEN site. The intent is not to produce a fully-loaded corporate financial model, nor to disclose any confidential site-specific costs. Instead, this section provides a transparent, parameterized *example* economic model that uses (i) high-level operational inputs elicited from stakeholders (e.g., typical study cadence and labor effort) and (ii) public, order-of-magnitude cost proxies for labor and consumables. Where public proxies are used, they are cited explicitly as footnotes and should be interpreted as illustrative, not representative of AMGEN internal accounting.

The conservative 30% reduction in experimental workload that was hypothesized in Subsection 1.4.1 and demonstrated in Chapter 4 is the entry point for a wider set of operational impacts. Fewer runs per study directly compresses execution timelines, accelerating CMC milestones and creating earlier opportunities for regulatory filing, while the freed bioreactor and labor capacity can be redeployed to absorb additional studies, increasing laboratory throughput without proportional headcount growth. The remainder of this section formalizes and quantifies these impacts.

5.3.1 Economic Model of Current State Process Characterization

Model scope. This economic model is scoped to the upstream PC study team at a single site. It focuses on incremental labor and variable run costs associated with executing bioreactor experiments and producing the associated reporting deliverables. Consistent with stakeholder feedback, material costs are not currently tracked in a uniform way across studies and are therefore modeled with a simple per-run proxy and separated from labor. All variable cost inputs in this section are illustrative proxies, not AMGEN costs.

Work volume and cadence. Let N_{PC} denote the number of upstream PC studies executed per year, and let T_{cycle} denote the end-to-end study cycle time in quarters (including reporting). Based on site experience, upstream PC is approximated by $N_{\mathrm{PC}} \approx 3$–$4$ studies/year and $T_{\mathrm{cycle}} \approx 4$ quarters, with a direct execution phase of $T_{\mathrm{exec}} \approx 2$ quarters.[1]

Labor model. Let F_{US} denote upstream PC staffing in FTE per quarter, and F_{ct} denote annual upstream PC effort in FTE-quarters:

$$F_{\mathrm{ct}} = 4\, F_{\mathrm{US}}. \tag{5.1}$$

Using stakeholder estimates, $F_{\mathrm{US}} \approx 3.5$ FTE/quarter, implying $F_{\mathrm{ct}} \approx 14$ FTE-quarters/year.[2]

Define C_{FTE} as an *illustrative* fully-loaded annual cost of one FTE:

$$C_{\mathrm{FTE}} = S\,(1 + B), \tag{5.2}$$

where S is a public salary proxy and B is a benefits load factor proxy.

Upstream labor per study (in FTE-quarters) is then approximated as:

$$F_{\mathrm{study}} = \frac{F_{\mathrm{ct}}}{N_{\mathrm{PC}}}. \tag{5.3}$$

Converting FTE-quarters to FTE-years (divide by 4), labor cost per study is:

$$C_{\mathrm{labor}}^{(\mathrm{base})} = \left(\frac{F_{\mathrm{study}}}{4}\right) C_{\mathrm{FTE}}. \tag{5.4}$$

[1] Stakeholder estimate from informal site discussion; values are used for high-level modeling only.

[2] Stakeholder estimate from informal site discussion; includes execution and reporting effort rather than lab time only.

Variable cost model (per run). Let R_{base} denote the number of bioreactor runs in a baseline PC design, and let c_{run} denote an illustrative variable cost per run:

$$C_{\text{var}}^{(\text{base})} = R_{\text{base}}\, c_{\text{run}}. \tag{5.5}$$

In the current model, c_{run} is decomposed into media, analytical, and disposables:

$$c_{\text{run}} = c_{\text{media}} + c_{\text{analytical}} + c_{\text{disp}}. \tag{5.6}$$

A simple proxy implementation is:

$$c_{\text{media}} = V_{\text{run}}\, P_{\text{media}}, \tag{5.7}$$

$$c_{\text{analytical}} = N_s\, P_{\text{analytical}}, \tag{5.8}$$

where V_{run} is run working volume (L), P_{media} is a catalog price proxy (\$/L), N_s is number of samples per run, and $P_{\text{analytical}}$ is a per-sample analytical price proxy. These public proxies are included to ground the example model and are not intended to reflect AMGEN procurement pricing.

Total cost per study. The baseline cost per upstream PC study is then:

$$C_{\text{study}}^{(\text{base})} = C_{\text{labor}}^{(\text{base})} + C_{\text{var}}^{(\text{base})}. \tag{5.9}$$

Baseline parameterization (illustrative example PD site). The following parameterization mirrors the structure of the spreadsheet model and is presented solely to demonstrate how the cost model can be instantiated with transparent assumptions:

- $N_{\text{PC}} = 4$ studies/year and $F_{\text{ct}} = 14$ FTE-quarters/year (stakeholder estimate)[3]
  $\Rightarrow F_{\text{study}} = 3.5$ FTE-quarters/study.

[3]Stakeholder estimate from informal site discussion; used to parameterize work volume.

- Salary proxy $S = \$103{,}650$ and benefits load $B = 0.295$[4] $\Rightarrow C_{\text{FTE}} = \$134{,}226.75$ per FTE-year.
- Baseline design size $R_{\text{base}} = 90$ runs/study (design assumption in this thesis).
- Media proxy $V_{\text{run}} = 0.25$ L and $P_{\text{media}} = \$169/\text{L}$[5] $\Rightarrow c_{\text{media}} \approx \$42.25$ per run.
- Analytics proxy $N_s = 15$ samples/run and $P_{\text{analytical}} = \$22/\text{sample}$[6] $\Rightarrow c_{\text{analytical}} \approx \$100$ per run.
- Disposables proxy $c_{\text{disp}} = \$5$ per run[7] $\Rightarrow c_{\text{run}} \approx \$147.25$ per run.

This yields the illustrative baseline:

$$C_{\text{labor}}^{(\text{base})} \approx \left(\frac{3.5}{4}\right) 134{,}226.75 = \$117{,}448 \tag{5.10}$$

$$C_{\text{var}}^{(\text{base})} \approx 90 \times 147.25 = \$13{,}252.50 \tag{5.11}$$

$$C_{\text{study}}^{(\text{base})} \approx \$130{,}701 \text{ per study.} \tag{5.12}$$

5.3.2 Economic Model of Possible Future State Process

Two levers: fewer runs and shorter execution. We model a hybrid-model-enabled future state as a reduction in (i) the number of executed runs and (ii) the execution duration required to reach a comparable level of process understanding.

Let α_r be the fractional reduction in runs per study and α_t be the fractional reduction in execution time (in quarters). The future-state run count is:

$$R_{\text{new}} = (1 - \alpha_r)\, R_{\text{base}}. \tag{5.13}$$

[4]Public salary proxy and benefits framing used for illustration: U.S. Bureau of Labor Statistics, Occupational Outlook Handbook (Biochemists and Biophysicists), https://www.bls.gov/ooh/life-physical-and-social-science/biochemists-and-biophysicists.htm.

[5]Illustrative catalog price proxy (example listing): Thermo Fisher product page, https://www.thermofisher.com/order/catalog/product/10743029.

[6]Illustrative third-party lab pricing proxy (example listing): University of Illinois Analytical Testing Laboratory, https://ibrl.aces.illinois.edu/analytical-testing/.

[7]Assumption used for illustration to represent a small set of consumables; not based on AMGEN internal purchasing.

Likewise, the execution duration becomes:

$$T_{\text{exec,new}} = (1 - \alpha_t)\, T_{\text{exec}}. \tag{5.14}$$

Labor reduction mapping. Labor does not necessarily scale one-to-one with run count or with execution time, because reporting and coordination persist. We therefore model labor reduction as a weighted combination of the two effects:

$$\alpha_f = \omega_r \alpha_r + \omega_t \alpha_t, \qquad \omega_r + \omega_t = 1, \tag{5.15}$$

where ω_r and ω_t represent the share of labor that scales with run execution versus calendar cycle time. These weights are modeling assumptions and can be stress-tested.

Future-state labor and variable cost per study are then:

$$C_{\text{labor}}^{(\text{new})} = (1 - \alpha_f)\, C_{\text{labor}}^{(\text{base})}, \tag{5.16}$$

$$C_{\text{var}}^{(\text{new})} = (1 - \alpha_r)\, C_{\text{var}}^{(\text{base})}, \tag{5.17}$$

and total cost per study:

$$C_{\text{study}}^{(\text{new})} = C_{\text{labor}}^{(\text{new})} + C_{\text{var}}^{(\text{new})}. \tag{5.18}$$

Future-state parameterization (conservative example). To align with the technical results, in particular the diminishing-returns analysis in Table 4.2, which shows that marginal accuracy gains per experiment decline sharply beyond approximately 35 PC runs, we set:

$$\alpha_r = 0.30, \tag{5.19}$$

$$\alpha_t = 0.25, \tag{5.20}$$

$$\omega_r = 0.50, \quad \omega_t = 0.50 \quad \Rightarrow \quad \alpha_f = 0.275. \tag{5.21}$$

This implies $R_{\text{new}} = 63$ runs/study and a proportional reduction in variable costs.

Using Eq. 5.18 with the illustrative baseline parameterization above:

$$C_{\text{var}}^{(\text{new})} \approx 0.70 \times 13{,}252.5 = \$9{,}276.8, \tag{5.22}$$

$$C_{\text{labor}}^{(\text{new})} \approx 0.725 \times 117{,}448 = \$85{,}150.1, \tag{5.23}$$

$$C_{\text{study}}^{(\text{new})} \approx \$94{,}427. \tag{5.24}$$

Therefore, the modeled savings per upstream PC study in this *example* are:

$$\Delta C_{\text{study}} = C_{\text{study}}^{(\text{base})} - C_{\text{study}}^{(\text{new})} \approx \$36{,}274, \tag{5.25}$$

$$\frac{\Delta C_{\text{study}}}{C_{\text{study}}^{(\text{base})}} \approx 27.8\%. \tag{5.26}$$

Under $N_{\text{PC}} = 4$ studies/year at a single site, this corresponds to $\sim$ \$145k per year in direct, local execution cost reduction in this illustrative model (exclusive of any value-of-time or downstream impact).

Capacity unlocked (site-level). The labor reduction fraction α_f also implies upstream PC effort freed per year:

$$\Delta F_{\text{ct}} = \alpha_f \, F_{\text{ct}} \approx 0.275 \times 14 = 3.85 \text{ FTE-quarters/year}. \tag{5.27}$$

This "capacity unlocked" can be used in two ways:

1. *Do more PC work:* execute additional studies per year (capacity expansion).

2. *Do different work:* reallocate scientists to higher-value development bottlenecks (portfolio acceleration).

Scaling beyond a single site. Because this model is scoped to a single site, enterprise impact is best expressed parametrically. Let N_{sites} be the number of sites with comparable upstream PC operations, and let $N_{\text{PC,site}}$ be the average studies/year

per site. Then enterprise-level annual savings in direct execution costs are:

$$\Delta C_{\text{AMG}} \approx N_{\text{sites}}\, N_{\text{PC,site}}\, \Delta C_{\text{study}}. \tag{5.28}$$

Similarly, enterprise-level capacity unlocked is:

$$\Delta F_{\text{ct,AMG}} \approx N_{\text{sites}}\, \Delta F_{\text{ct}}. \tag{5.29}$$

A practical next step to instantiate Eqs. 5.28–5.29 is to replace public proxies with an internally approved cost model and estimate $(N_{\text{sites}}, N_{\text{PC,site}})$ from portfolio planning tools or headcount proxies.

Strategic Scaling and Portfolio Value. While the site-level analysis provides a concrete baseline, the enterprise-level impact is best understood by scaling these metrics across AMGEN's global manufacturing network. Assuming five sites with comparable upstream PC operations, the organization could realize over $725k in annual direct execution savings and unlock approximately 19 FTE-quarters of scientific capacity. A comparison of these site-level and projected global impacts is summarized in Table 5.1.

Table 5.1: Comparison of Site-Level and Projected Enterprise Business Impact

Metric	**Site Level**	**Enterprise Estimate (Global)**
Annual PC Studies	4	20 (Assumed 5 major sites)
Annual Direct Savings	∼$145k	**∼$725k**
Scientific Capacity Unlocked	3.85 FTE-quarters	**∼19 FTE-quarters**
Time-to-Market Impact	∼25% reduction in T_{exec}	Earlier Regulatory Filings

Beyond these direct costs, the true business case is likely dominated by "value-of-time" and downstream manufacturing improvements. Monetizing the 25% reduction in execution cycle time through earlier regulatory filings or clinical transitions could outweigh direct labor savings by orders of magnitude. Furthermore, identifying more robust, higher-titer setpoints through hybrid modeling offers significant COGS improvements at commercial scale. By shifting from a purely descriptive statistical

approach to a predictive hybrid framework, AMGEN can effectively trade a reduction in experimental burden for an increase in portfolio velocity and manufacturing resilience. To place an order of magnitude on the value-of-time effect, consider a widely cited industry estimate that each day of delayed market entry for a blockbuster biologic costs approximately $C_{\text{delay}} \approx \$0.5\text{M}$ in net present value [28]. Let ΔT_{exec} denote the reduction in upstream PC execution time achieved through hybrid-model-guided experimentation. Under the conservative $\alpha_t = 0.25$ assumption (5.21), this corresponds to roughly one quarter of calendar acceleration in the PC phase:

$$\Delta T_{\text{exec}} = \alpha_t \cdot T_{\text{exec}} = 0.25 \times 2 \text{ quarters} = 0.5 \text{ quarters} \approx 45 \text{ days}. \tag{5.30}$$

If even a fraction β of this PC-phase acceleration translates to earlier regulatory filing (accounting for the reality that downstream activities, PPQ, filing preparation, and regulatory review, impose their own constraints), the implied NPV gain per molecule is:

$$\Delta V_{\text{time}} = \beta \cdot \Delta T_{\text{exec}} \cdot C_{\text{delay}}. \tag{5.31}$$

Under a conservative assumption of $\beta = 0.67$ (i.e., one month of the ∼45-day reduction reaches the filing date), this yields $\Delta V_{\text{time}} \approx 30 \times \$0.5\text{M} = \$15\text{M}$ per molecule. Across a portfolio of $N_{\text{mol}} = 3$–5 molecules in late-stage development at any given time, the aggregate time-value opportunity is on the order of \$45–75M, roughly two orders of magnitude larger than the direct execution savings estimated in Equation 5.26. This calculation is necessarily approximate: the \$0.5M/day figure reflects an industry average for blockbuster-class drugs, β will vary by program depending on whether PC is on the critical path, and the estimate assumes no offsetting delays elsewhere in the development timeline. Nonetheless, it illustrates a key strategic point: *the economic case for reduced PC experimentation is dominated by the value of time, not the cost of experiments.*

5.3.3 Enterprise Strategic Implications

The unlocked scientific capacity (∼19 FTE-quarters globally) provides a significant lever for AMGEN's broader Process Development and Operations strategy. This capacity can be reallocated to pursue a higher volume of potential drug candidates in parallel, effectively increasing the organization's shots-on-goal without a proportional increase in laboratory footprint.

Furthermore, the transition to model-aided DoE like Latin Hypercube or D-optimal sampling ensures a more thorough mapping of the design space. This minimizes the risk of discovering unexpected sensitivities during late-stage Tech Transfer or Process Performance Qualification (PPQ). By leveraging transfer learning to build "master models" for specific cell lines, AMGEN can continuously refine its process understanding across molecules, turning every PC study into a cumulative asset rather than an isolated experiment.

5.4 Regulatory Considerations

The adoption of hybrid machine learning models in process characterization must be considered within the regulatory frameworks that govern biopharmaceutical manufacturing. Both the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have signaled openness to advanced modeling and data-driven approaches, but neither has issued prescriptive guidance on the use of hybrid ML models in process validation filings. Navigating this landscape requires a deliberate engagement strategy that builds regulatory confidence incrementally.

5.4.1 Current regulatory landscape

The FDA's Process Analytical Technology (PAT) initiative and ICH Q8–Q12 guidelines collectively encourage the use of enhanced process understanding, design space characterization, and model-based control strategies [29, 30, 31, 32, 33]. ICH Q8(R2) explicitly supports the use of multivariate models to define design spaces, and ICH Q12

provides a framework for post-approval changes enabled by deeper process knowledge. The EMA has similarly endorsed Quality by Design (QbD) principles and has published reflection papers encouraging the use of advanced statistical and computational methods in process development [34, 35]. More recently, the FDA has released discussion papers on the use of artificial intelligence and machine learning in pharmaceutical manufacturing, signaling that the agency is actively developing its internal capacity to evaluate such submissions [9].

5.4.2 Phased regulatory engagement strategy

Given the absence of explicit regulatory precedent for hybrid ML-based PC studies, AMGEN should pursue a phased approach to regulatory engagement:

Phase 1: Internal validation and dual-track filing. The first regulatory touchpoint should be a PC study in which the hybrid model is run in parallel with the standard JMP-based analysis. The regulatory filing would be based entirely on the traditional approach, but the hybrid model results would be included as supplementary material. This demonstrates the model's capabilities without creating regulatory risk, and it provides a concrete artifact for pre-submission discussions with reviewers.

Phase 2: Pre-submission engagement. Following the dual-track filing, AMGEN should request a Type B or Type C meeting with the FDA (or a Scientific Advice meeting with the EMA) to discuss the use of hybrid models as a primary or co-primary analytical tool for future PC studies. These meetings would present the retrospective validation results from this thesis alongside the prospective dual-track results from Phase 1, framing the hybrid approach as an enhancement of—not a departure from—established QbD principles.

Phase 3: Model-primary filing. Once regulatory feedback has been incorporated and a second prospective study has been completed using the hybrid-model-guided design, AMGEN could submit a filing in which the hybrid model serves as the primary basis for design space characterization, with traditional regression analysis included as a confirmatory check. This inverts the relationship established in Phase 1 and positions the hybrid approach as the standard methodology going forward.

Key considerations for regulatory acceptance. Reviewers will likely focus on several aspects of the hybrid modeling approach: model interpretability (can the mechanistic component be explained in physicochemical terms?), validation methodology (how is prediction accuracy assessed on held-out data?), robustness to distributional shift (will the model remain accurate under manufacturing-scale conditions?), and change management (how will model updates be governed post-approval?). Proactively addressing these questions in the filing, through clear documentation of model architecture, training procedures, and validation metrics, will be essential for a favorable review.

5.5 Strategic Implications of Transfer Learning

Transfer learning, which is defined here as the practice of reusing knowledge from one trained model to accelerate learning on a related task, offers one of the most compelling long-term advantages of adopting hybrid ML for process characterization. In the context of biopharmaceutical development, transfer learning could fundamentally change the relationship between successive PC studies, transforming them from isolated experiments into nodes in a cumulative knowledge network.

Faster timelines for new molecules. When a new molecule enters process development, the PC team currently starts from a blank slate: a new experimental design is created, all runs are executed, and a new statistical model is built from scratch. With transfer learning, a model trained on a previous molecule expressed in the same CHO cell line could serve as a warm-start for the new study. The mechanistic component of the hybrid model, which encodes cell growth kinetics, nutrient consumption, and metabolite production, is largely cell-line-specific rather than molecule-specific, meaning that a substantial portion of the model's learned parameters should be transferable. Only the molecule-specific output layers (e.g., those predicting glycosylation patterns or charge variants) would need to be retrained, and this retraining could require significantly fewer experiments than training from scratch.

Master models for cell lines. Over time, as multiple molecules expressed in the same cell line are characterized, the accumulated data can be used to build a "master model" for that cell line. This master model would encode the shared biological behavior across molecules and would serve as the default starting point for any new PC study using that cell line. Each new study would then refine and extend the master model, contributing its data back to the shared knowledge base. This creates a flywheel effect: the more studies that are completed, the better the master model becomes, and the fewer experiments each subsequent study requires.

A key open architecture question is how deeply molecular identity should condition such a master model. At one extreme, the shared mechanistic backbone is frozen entirely and only the final product-quality layers are retrained for each new molecule. At the other, molecular identity conditions the model more deeply, for example, through a joint molecule–cell-line embedding that influences predicted cell response itself, not just the mapping from cell state to PQAs. The right depth likely depends on how strongly a given molecule perturbs cell behavior, an empirical question that will become answerable as cross-molecule datasets accumulate.

Enabling conditions. Realizing the benefits of transfer learning at enterprise scale requires several enabling conditions. First, data from PC studies must be stored in a consistent, structured format that allows models to be trained across studies. This is a non-trivial data engineering challenge, as current PC data are often stored in study-specific spreadsheets with inconsistent naming conventions and variable definitions. Second, the hybrid modeling platform must support transfer learning natively, allowing users to load a pre-trained model, freeze certain layers, and fine-tune others on new data. Third, a governance framework must be established to manage the lifecycle of master models, including versioning, revalidation triggers, and retirement criteria.

A flywheel for AI integration. The transfer learning paradigm creates a natural flywheel for broader AI adoption in process development. As data cleanup and integration efforts progress, driven initially by the requirements of hybrid modeling,

the resulting data infrastructure becomes available for other ML applications, such as predictive maintenance, real-time process monitoring, and automated deviation investigation. Each application generates demand for better data, which in turn improves the foundation for all applications. In this way, the investment in hybrid modeling for PC serves as a catalyst for a broader digital transformation of the development organization.

5.6 Change Management and Adoption

The technical merits of hybrid modeling are necessary but not sufficient for successful adoption. The transition from established JMP-based workflows to a model-guided paradigm represents a significant organizational change that will affect the daily work of scientists, the planning processes of managers, and the review expectations of quality and regulatory teams. Managing this transition effectively requires attention to human factors, incentive structures, and organizational learning.

Start with quick wins. The most effective way to build organizational support for a new tool is to demonstrate immediate, tangible value in a context that matters to the end users. Rather than leading with a top-down mandate to adopt hybrid modeling, the implementation strategy should identify specific pain points in the current PC workflow, such as long turnaround times for a particular PQA, or a recurring need to add unplanned experimental runs, and demonstrate that the hybrid model can alleviate them. These quick wins build credibility and create internal champions who can advocate for broader adoption.

Listen before prescribing. A common failure mode in technology adoption is the "hammer looking for a nail" problem: a team with a powerful new tool attempts to apply it everywhere, regardless of whether the application is well-suited. The hybrid modeling team should invest significant time in understanding the needs, constraints, and frustrations of the PC scientists and study directors who would be the primary users. Some of the most valuable applications may not be the ones the modeling team

anticipated. For example, scientists may value the model's ability to identify which process parameters are most influential for a given PQA more than its ability to reduce experiment count, because that information helps them write better characterization reports.

Embrace iterative deployment. Hybrid modeling tools do not need to be perfect before they are useful. A minimum viable deployment, perhaps a dashboard that displays DHL predictions alongside JMP results for an ongoing study, without requiring any changes to the experimental design, can provide value and generate feedback while the platform continues to mature. Subsequent iterations can introduce adaptive design recommendations, automated model retraining, and integration with laboratory information management systems (LIMS). This iterative approach aligns with established best practices in technology deployment and avoids the risk of a prolonged development phase that delivers a polished tool nobody asked for.

Organizational alignment. Successful adoption also requires alignment across functional boundaries. The PC team, the data science team, the quality organization, and regulatory affairs must all understand the capabilities and limitations of the hybrid approach. Cross-functional working sessions—not just presentations, but collaborative problem-solving exercises using real study data—can build shared understanding and identify potential concerns early. Establishing a formal governance body (e.g., a Modeling Center of Excellence) that includes representatives from each function can provide ongoing oversight and decision-making authority for model deployment decisions.

5.6.1 Human Workflow and PC Planning Implications

Process characterization planning is a structured, well-established process at AMGEN, with defined timelines, resource allocations, and deliverable milestones. Introducing hybrid-model-guided experimentation would disrupt this process in ways that must be carefully managed.

A new planning workflow. Under the proposed paradigm, the human workflow for a PC study would proceed as follows:

1. **Initialize:** Load the relevant master model (if available) for the cell line and molecule class. Use the model, in conjunction with prior process knowledge, to inform the initial experimental design and define the minimal starting set of experiments.

2. **Execute Block 1:** Run the initial set of experiments (e.g., 20–30 runs defined by Latin Hypercube sampling).

3. **Update:** Collect PQA results, retrain or fine-tune the hybrid model on the new data, and evaluate prediction accuracy against predefined acceptance criteria.

4. **Redesign:** Based on model predictions and remaining uncertainty, generate an updated set of experiments targeting the regions of the design space where additional data would most improve the model.

5. **Execute Block 2+:** Run the next block of experiments. Return to Step 3.

6. **Terminate:** End experimentation once the model meets accuracy thresholds for all PQAs of interest, or once a predefined maximum number of iterations has been reached.

The scheduling challenge. The iterative nature of this workflow introduces uncertainty into study scheduling, because the number of experimental blocks required is not known at the outset. In the current paradigm, a study director can commit to a fixed number of runs and a predictable timeline. Under the adaptive approach, the total run count and calendar duration depend on how quickly the model converges for each PQA.

To make this operationally viable, the team must establish a maximum number of adaptive iterations (e.g., three blocks) and a maximum total run count (e.g., 60 runs) as planning guardrails. These maxima should be set conservatively based on the retrospective results from this thesis and refined as prospective experience

accumulates. With these guardrails in place, study directors can plan scheduling and resource allocation with bounded uncertainty, even if the exact stopping point within the bounds is determined adaptively.

Proving the workflow. Before the adaptive workflow can be adopted as standard practice, it must be validated through one or more prospective pilot studies. These pilots should be conducted on molecules where the traditional PC study has already been completed (or is being completed in parallel), allowing a direct comparison of outcomes. The pilots will establish empirical distributions for the number of iterations required, the total run count at termination, and the resulting model accuracy—data that are essential for calibrating the scheduling guardrails described above.

Chapter 6

Conclusions and Future Work

This thesis investigated whether hybrid mechanistic machine learning models can reduce the experimental burden of upstream process characterization studies at AMGEN while maintaining the predictive accuracy required for regulatory filings. This chapter summarizes the key findings, distills their implications for AMGEN's PC strategy, outlines concrete next steps to advance the work begun here, and identifies broader applications of hybrid ML and AI across cell culture development.

6.1 Summary of Findings

The central result of this thesis is that hybrid machine learning models, which couple a mechanistic propagation model of cell culture dynamics with a data-driven historical model, can predict certain product quality attributes with substantially fewer experiments than the traditional JMP regression approach. Additionally, it is worth noting that, unlike JMP, the hybrid-mechanistic approach not only predicts PQAs, but also predicts the temporal evolution of key signals such as cell density, viability, titer & the trajectories of other key metabolites; a useful tool for scientists at AMGEN. For the best-performing PQAs, models trained on approximately 35 experiments matched or exceeded the accuracy of JMP models trained on the full ∼75-experiment design, a 54% reduction in experimental burden. On average across PQAs, learning curves flattened around 35-45 training experiments, suggesting diminishing returns from the

remaining runs in a conventional ∼90-run campaign.

However, performance was not uniform. Three distinct learning behaviors were observed. For some PQAs (e.g., PQA 2), hybrid models converged rapidly and crossed the JMP accuracy threshold well before the full dataset was consumed. For others (e.g., PQA 8), models converged quickly but plateaued above the JMP baseline, suggesting structural limitations in the model or irreducible measurement noise rather than data scarcity. For a third category (e.g., PQA 6), accuracy improved steadily with additional data, indicating that these attributes would benefit from expanded datasets or targeted experimentation.

Several additional findings merit emphasis. First, using measured time-series inputs (Simulation Approach 2) rather than idealized setpoint perturbations (Approach 1) produced meaningfully better predictions by preserving run-to-run control variability. Second, commercial process development (CPD) data served as a useful prior that accelerated model convergence when combined with PC data, but could not substitute for characterization experiments that stress the boundaries of the design space. Third, space-filling sampling strategies such as Latin Hypercube and Sobol sequences outperformed exterior-point designs for training hybrid models, consistent with the intuition that capturing internal nonlinear dynamics matters more than boundary coverage for predictive learning. Fourth, PQA-specific models did not consistently outperform general multi-output models, suggesting that the shared propagation backbone, rather than attribute-level specialization, is the dominant performance lever, and that improving this backbone should be the primary focus of future model development.

6.2 Implications for AMGEN's Process Characterization Strategy

These findings carry direct implications for how AMGEN designs, executes, and resources its PC studies. As discussed in Section 5.3, even a conservative 30%

reduction in experimental burden translates to approximately $36k in direct cost savings per study and $145k annually at a single site, with the potential for over $725k in annual savings across five comparable sites. Perhaps more importantly, the associated 25% reduction in execution cycle time and the ∼19 FTE-quarters of unlocked scientific capacity globally could enable AMGEN to increase the number of molecules it characterizes each year without proportional increases in headcount or laboratory infrastructure.

The evidence is sufficient to justify continued investment in hybrid modeling as a complement to, and eventually a partial replacement for, JMP-based regression in PC workflows. The economic model presented in Chapter 5 demonstrates that the minimal value case is robust even with conservative assumptions about the degree of experimental reduction achievable. As data infrastructure, model maturity, and organizational familiarity improve, the realized benefits are likely to exceed the conservative estimates presented here.

6.2.1 Recommendations for Future PC Study Design

Based on the findings of this thesis and the analysis in Subsection 5.2.1, it is recommended that AMGEN adopt a new experimental design paradigm for future PC studies. The preferred approach is Approach B from Subsection 5.2.2: an adaptive, block-sequential design in which an initial Latin Hypercube experiment set is executed, the hybrid model is trained and evaluated, and subsequent experimental blocks are designed based on model predictions and remaining uncertainty. Between blocks, bioreactor capacity and scientific staff can be redirected to other PC studies, increasing effective throughput through interleaving.

We recognize that Approach B requires significant changes to scheduling infrastructure, analytical turnaround workflows, and study director planning practices that may not be feasible immediately. If operational constraints preclude the adaptive interleaved approach in the near term, AMGEN should begin with Approach A: defining the full experimental design upfront using Latin Hypercube sampling and executing all runs without interim model updates. This approach still captures the

benefit of reduced total experimentation relative to central composite designs, and it can be implemented within the existing scheduling framework. As prospective pilot studies validate the adaptive workflow and establish empirical distributions for the number of iterations required (see Subsection 5.6.1), the organization can transition from Approach A to Approach B with confidence.

In either case, the critical first step is to commit to the joint decision of experimental design and modeling framework at the outset of each new study. Central composite designs paired with JMP and Latin Hypercube designs paired with hybrid modelling represent two coherent paradigms; mixing them (e.g., applying DHL models to data from a CCD) yields suboptimal results, as this thesis has demonstrated.

6.3 Next Steps for This Work

The research presented in this thesis opens several concrete avenues for follow-on work, each of which would strengthen the evidence base and accelerate practical deployment.

Joint tuning of propagation and historical models. As discussed in Subsection 5.1.3, the propagation and historical components of the DHL architecture were tuned independently in this work. Joint optimization could improve coherence between the two model components and boost performance on the PQAs where neither component alone was sufficient. This is particularly motivated by the finding that PQA-specific models did not outperform general models (Subsection 5.1.1); joint tuning may be the mechanism through which PQA-specific performance gains are realized.

Comparison with alternative model architectures. This thesis evaluated a single hybrid model type. A rigorous head-to-head comparison with alternative approaches (including multi-layer perceptrons, Gaussian process models, gradient-boosted trees, and JMP's built-in neural network platform) is essential before committing to DHL as the enterprise standard. This comparison should be conducted on the same datasets used here, with the same evaluation metrics, to ensure a fair assessment.

Larger datasets and improved train-test splits. The relatively small dataset size in this work (∼75–90 experiments) constrained the ability to perform conventional held-out test set evaluations. Identifying or generating larger PC datasets, either from molecules with more extensive experimental histories or by pooling data across related programs, would enable more statistically rigorous train-test splits and provide stronger evidence of generalization performance.

Direct input-output comparison with JMP. The comparison between hybrid modeling and JMP in this thesis was directional rather than exact, because the two frameworks use different input representations and model structures. A more controlled comparison, as discussed in Subsection 5.1.3, in which the hybrid model is configured to use the same input variables and predict the same output variables as the JMP regression model would isolate the effect of model architecture from the effect of input representation. This direct comparison is expected to favor hybrid modeling by removing the variable noise introduced by time-series inputs, and would provide a cleaner signal for stakeholder decision-making. A separate natural extension of this work would also retrain JMP on the same data subsets used in the learning-curve analysis, generating a parallel JMP learning curve that reveals where response surface regression degrades with reduced data—a comparison expected to further favor hybrid modeling and one that is not possible to draw from the current retrospective design (see Subsection 4.1.1).

Continued roll-out to AMGEN teams. Beyond the technical work, continued engagement with PC study teams at AMGEN is essential. Providing scientists with access to DHL tools in a low-stakes, exploratory mode, without requiring changes to the study design, builds familiarity, surfaces practical usability issues, and identifies PQAs or molecules where hybrid models are most likely to add value. The API tools referenced in Subsection 3.2.2 enable this rollout in a much more straightforward way. Additionally, these tools can support one-off problem solving efforts that are common to the work done by the process development and data science teams at AMGEN's

sites.

Transfer learning across programs. The CPD-to-PC transfer results in Section 5.5 demonstrated that upstream priors accelerate model convergence. The logical next step is to test transfer learning across molecules that share the same cell line, training a model on one completed PC study and evaluating how much it accelerates convergence on a second molecule. This would provide the first empirical evidence for the "master model" concept discussed in Chapter 5.

Prospective pilot study. Finally, the most impactful next step is to identify an upcoming PC study where the hybrid-model-guided workflow can be applied prospectively. This pilot should run in parallel with the traditional approach (dual-track) to enable a direct comparison of outcomes, timelines, and costs. The pilot would serve as both a technical validation and a regulatory proof point, providing the concrete evidence needed to justify broader adoption.

6.4 Future Work: Other Use Cases of AI and Hybrid ML for Cell Culture at AMGEN

While this thesis focused on production bioreactor PC, the hybrid modeling framework and the organizational capabilities developed through this work have natural extensions to several adjacent areas of cell culture development. These represent opportunities to compound the value of AMGEN's investment in hybrid ML infrastructure.

6.4.1 Application of Hybrid ML to Earlier PC Studies (Expansion Bioreactors)

Expansion bioreactor process characterization studies, covering the seed train stages from N−3 through N−1, represent a compelling early target for hybrid-model-guided experimental design, and may in fact offer a lower barrier to entry than production

bioreactor PC. This is visualized in Figure 2-2.

Expansion PC studies focus on ensuring robust cell growth in the early seed train stages. The critical outputs are viable cell density (VCD) and viability, rather than the complex product quality attributes (glycosylation, charge variants, aggregation) that dominate production PC. This narrower output space simplifies the modeling problem considerably: the hybrid model needs to predict growth kinetics only, without the additional complexity of product quality prediction.

Despite this relative simplicity, expansion PC currently faces significant experimental burden. Different feed methods—batch, fed-batch, and perfusion—may be used in series across seed train stages for certain cell lines, and a full multivariate characterization can require 90–120 experiments. Because this burden is perceived as unsustainable, teams often resort to lean, univariate experimental designs that test one variable at a time. While operationally efficient, univariate designs miss interaction effects between process parameters, leaving potential failure modes undiscovered until tech transfer or PPQ.

Hybrid ML-guided DoE could address this gap by enabling multivariate exploration with a fraction of the experimental load that a full factorial design would require. Furthermore, the hybrid framework's mechanistic component is well suited to modeling the serial nature of expansion bioreactors: linking stage-to-stage models would allow scientists to assess how suboptimal conditions at N−3 propagate through N−2 and N−1 to affect production bioreactor inoculation quality. This stage-linking capability, which is already conceptually practiced in manual form today, could be formalized and made quantitative through hybrid modeling.

A successful model-aided DoE for expansion bioreactors would materially reduce reactor burden, accelerate PC timelines, fulfill regulatory expectations for seed train robustness, and critically serve as a lower-risk proving ground that de-risks the broader adoption of hybrid DoE in production PC.

6.4.2 Utilizing Prior Knowledge Assessment Information to Inform PC DoE

AMGEN maintains Prior Knowledge Assessments (PKAs) for its cell culture processes: structured documents that capture expert intuition, historical experimental evidence, and mechanistic reasoning about which process variables are most likely to affect specific quality attributes. These PKAs represent a valuable but underutilized trove of institutional knowledge.

Today, PKAs inform PC study design qualitatively: experienced scientists draw on them when selecting which variables to include in a characterization study and what ranges to test. However, this translation from documented knowledge to experimental design is informal and dependent on individual expertise. There is an opportunity to formalize and systematize this process by pairing PKA information with computational tools.

In the near term, PKA content could be used as a structured input to the adaptive DoE framework proposed in this thesis: helping to select which variables to vary, defining initial ranges, and prioritizing regions of the design space where prior knowledge suggests sensitivity is highest. This would make the initial experimental block more informative and reduce the number of adaptive iterations required.

Looking further ahead, large language models (LLMs) could be applied to extract, standardize, and synthesize PKA information across molecules and cell lines. An LLM-assisted workflow could parse PKA documents, identify consensus patterns (e.g., "pH is consistently flagged as a critical variable for charge variants across CHO cell lines"), and generate draft variable selection recommendations for new PC studies. This approach would be highly explainable to regulators, as it grounds model-guided decisions in documented prior knowledge rather than opaque algorithmic choices.

The standardization effort required to make PKAs machine-readable would itself be valuable, creating a more consistent and auditable knowledge base regardless of whether AI tools are applied to it.

6.4.3 Transfer Learning

As discussed in Section 5.5, transfer learning offers the potential to accelerate model convergence for new molecules by reusing knowledge from previously characterized processes. Two primary transfer directions are relevant:

Across molecules within a cell line: When multiple molecules are expressed in the same CHO cell line, the underlying cell growth kinetics, nutrient consumption patterns, and metabolite production dynamics are largely shared. A model trained on Molecule A can provide a strong initialization for Molecule B, with only the product-specific output layers requiring retraining. This is the most immediately actionable form of transfer learning and should be prioritized.

Across cell lines for a single molecule type: When similar therapeutic modalities (e.g., monoclonal antibodies) are produced in different cell lines, some high-level process–quality relationships may transfer. This form of transfer is more speculative and would require empirical validation, but could become valuable as AMGEN's hybrid modeling dataset grows.

As AMGEN's data quality and integration efforts continue to mature, the feasibility of both transfer directions will increase. The practical recommendation is to begin building cross-molecule models for a single well-characterized cell line and empirically assess how much each additional molecule's data improves the shared model. Over time, this cumulative dataset will become one of the organization's most valuable digital assets for process development.

6.4.4 Bioreactor Control Strategies

The most forward-looking application of hybrid modeling is in real-time bioreactor control at manufacturing scale. The models developed in this thesis predict how time-course process data (temperature, pH, dissolved oxygen, metabolite concentrations) influence final product quality attributes. This predictive capability, if deployed in real time, could fundamentally change how manufacturing deviations are detected and managed.

In a model-aided control paradigm, in-process sensor data from a production bioreactor would be fed continuously into a hybrid model. The model would project the current trajectory forward, estimating the expected final PQA values given the data observed so far. If the projected trajectory deviates from acceptable ranges, the system could alert operators early—potentially days before the deviation would be detected by conventional end-of-batch testing—and suggest corrective actions (e.g., adjusting feed rates, modifying temperature profiles) to bring the batch back on track.

This capability would enable faster problem-solving at manufacturing scale by identifying problems earlier in the batch and providing quantitative guidance on corrective interventions. It would also support more sophisticated batch release decisions: rather than relying solely on end-of-batch analytical testing, quality teams could incorporate the model's continuous trajectory assessment as supplementary evidence of process control.

In the longer term, this architecture could form the basis for highly automated "lights-out" manufacturing, in which the control system autonomously adjusts process parameters in real time to optimize product quality and yield. While fully autonomous bioreactor control remains a distant goal that would require extensive regulatory engagement and validation, the hybrid models developed through PC studies provide the essential predictive foundation on which such a system would be built. Each PC study that trains and validates a hybrid model contributes to the eventual realization of this vision.

6.4.5 Bioreactor Scale Up

A related application is scale-up prediction. Translating process conditions from bench-scale (e.g., 250 mL) to pilot (e.g., 50 L) and manufacturing scale (e.g., 2000 L+) remains one of the most resource-intensive and failure-prone steps in biologics development. Scale-dependent variables such as mixing time, oxygen transfer rate, and CO_2 stripping efficiency change in ways that are partially understood mechanistically but difficult to predict quantitatively. Hybrid models are well positioned to address this: the mechanistic backbone can encode known scaling relationships (e.g., $k_L a$

correlations, power-per-volume), while the data-driven component learns residual scale-dependent effects from historical multi-scale data. If trained on paired small-scale and manufacturing-scale runs from prior programs, such models could predict how a new molecule's process will behave at larger scale before committing to expensive manufacturing campaigns. Prior LGO work at AMGEN by Wolszon demonstrated the feasibility of hybrid modeling for scale-up prediction, and the data infrastructure investments discussed in Section 5.5 would directly enable this extension [4].

Closing Remarks

This thesis has demonstrated that hybrid mechanistic-machine learning models can meaningfully reduce the experimental burden of upstream process characterization while maintaining predictive accuracy comparable to established regression methods. For certain product quality attributes, the reduction exceeded 50%, and the associated economic and capacity benefits are substantial even under conservative assumptions. While the retrospective nature of this work leaves important questions open, particularly around prospective validation, alternative model architectures, and cross-molecule transferability, the evidence presented here establishes a clear foundation for continued investment. The path from proof-of-concept to operational deployment will require sustained collaboration between data scientists, process development engineers, quality organizations, and regulatory agencies. The opportunity, however, is significant: by embedding predictive modeling into the fabric of process characterization, AMGEN can accelerate its pipeline, deepen its process understanding, and strengthen its competitive position in an industry where speed, quality, and efficiency are increasingly inseparable.

Appendix A

LLM Use Acknowledgment

Any use of generative AI in this manuscript adheres to ethical guidelines for the use and acknowledgment of generative AI in academic research. The author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of his contributions [36].

The use cases of generative AI tools included the following:

- Gathering information and ideas, explaining concepts, parsing primary sources;
- Drafting, refactoring, and debugging code;
- Structuring, revising, and reviewing language for clarity and grammatical correctness.

Bibliography

[1] Or Dan. "Improving Prior Knowledge Assessment in Process Characterization". Master's Thesis. Massachusetts Institute of Technology, 2020.

[2] Fabian Feidl. "Assessing the economic impact of digital process development within the biopharmaceutical industry". MA thesis. University of St. Gallen, Jan. 2024.

[3] Michael Sokolov. *Radical Bioprocessing Efficiencies and Cost Reductions — The Next Wave of Biopharmaceutical Innovation*. Tech. rep. Zurich: DataHow AG, 2024. URL: www.datahow.com.

[4] Zoë Wolszon. "Improving Predictability of Cell Culture Processes During Biologics Manufacturing Scale-Up through Hybrid Modeling". Master's Thesis. Massachusetts Institute of Technology, 2020.

[5] Mohammad Rashedi et al. "Machine learning-based model predictive controller design for cell culture processes". In: *Biotechnology and Bioengineering* 120.8 (Aug. 2023), pp. 2144–2159. ISSN: 10970290. DOI: 10.1002/bit.28486.

[6] Daniel Griffin, Behnam Partopour, and Seth Huggins. "Knowledge-Constrained Machine Learning: A Strategy for Producing Predictive Process Models in the Absence of Mechanistic Understanding and Large Data Sets". In: *Proceedings of the Foundations of Process Analytics and Modeling (FOPAM)*. Amgen Inc. Trondheim, Norway, 2019. URL: https://skoge.folk.ntnu.no/prost/proceedings/FOPAM_2019/%20FOPAM%5C%20Contributed%5C%20Papers/60_FinalAbstract.pdf.

[7] Saratram Gopalakrishnan et al. "COSMIC-dFBA: A novel multi-scale hybrid framework for bioprocess modeling". In: *Metabolic Engineering* 82 (Mar. 2024), pp. 183–192. ISSN: 10967184. DOI: 10.1016/j.ymben.2024.02.012.

[8] Jackson A. Albright. "Computer Vision for Cell Line Development". Master's Thesis. Massachusetts Institute of Technology, 2025.

[9] U.S. Food and Drug Administration. *Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry and Other Interested Parties DRAFT GUIDANCE*. Tech. rep. U.S. Food and Drug Administration., Jan. 2025. URL: https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances.

[10] Laura M. Helleckes et al. “Novel calibration design improves knowledge transfer across products for the characterization of pharmaceutical bioprocesses”. In: *Biotechnology Journal* 19.7 (July 2024). ISSN: 18607314. DOI: 10.1002/biot.202400080.

[11] Benjamin Bayer, Gerald Striedner, and Mark Duerkop. “Hybrid Modeling and Intensified DoE: An Approach to Accelerate Upstream Process Characterization”. In: *Biotechnology Journal* 15.9 (Sept. 2020). ISSN: 18607314. DOI: 10.1002/biot.202000121.

[12] Jialu Wang et al. “Measure this, not that: Optimizing the cost and model-based information content of measurements”. In: *Computers and Chemical Engineering* 189 (Oct. 2024). ISSN: 00981354. DOI: 10.1016/j.compchemeng.2024.108786.

[13] D. E. Ricciardi et al. “Bayesian optimal experimental design for constitutive model calibration”. In: *International Journal of Mechanical Sciences* 265 (Mar. 2024). ISSN: 00207403. DOI: 10.1016/j.ijmecsci.2023.108881.

[14] Luc Pronzato and Werner G. Müller. “Design of computer experiments: Space filling and beyond”. In: *Statistics and Computing* 22.3 (May 2012), pp. 681–701. ISSN: 09603174. DOI: 10.1007/s11222-011-9242-3.

[15] Harini Narayanan et al. *Accelerating Cell Culture Media Development Using Bayesian Optimization-Based Iterative Experimental Design*. Nov. 2024. DOI: 10.1101/2024.10.29.620971. URL: http://biorxiv.org/lookup/doi/10.1101/2024.10.29.620971.

[16] M. D. McKay, R. J. Beckman, and W. J. Conover. “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code”. In: *Technometrics* 42.1 (2000), pp. 55–61. DOI: 10.1080/00401706.2000.10485979.

[17] Jakub Polak et al. “An innovative hybrid modeling approach for simultaneous prediction of cell culture process dynamics and product quality”. In: *Biotechnology Journal* 19.3 (Mar. 2024). ISSN: 18607314. DOI: 10.1002/biot.202300473.

[18] Harini Narayanan et al. “A new generation of predictive models: The added value of hybrid models for manufacturing processes of therapeutic proteins”. In: *Biotechnology and Bioengineering* 116.10 (Oct. 2019), pp. 2540–2549. ISSN: 10970290. DOI: 10.1002/bit.27097.

[19] M. Nicolás Cruz-Bournazou et al. “Hybrid Gaussian Process Models for continuous time series in bolus fed-batch cultures”. In: *IFAC-PapersOnLine*. Vol. 55. 7. Elsevier B.V., 2022, pp. 204–209. DOI: 10.1016/j.ifacol.2022.07.445.

[20] Shu Yang et al. “Hybrid Modeling of Fed-Batch Cell Culture Using Physics-Informed Neural Network”. In: *Industrial and Engineering Chemistry Research* 63.39 (Oct. 2024), pp. 16833–16846. ISSN: 15205045. DOI: 10.1021/acs.iecr.4c01459.

[21] Clemens Hutter et al. “Knowledge transfer across cell lines using hybrid Gaussian process models with entity embedding vectors”. In: *Biotechnology and Bioengineering* 118.11 (Nov. 2021), pp. 4389–4401. ISSN: 10970290. DOI: 10.1002/bit.27907.

[22] Riccardo De Luca et al. “Comparison of strategies for iterative model-based upstream bioprocess development with single and parallel reactor set-ups”. In: *Biochemical Engineering Journal* 191 (Feb. 2023). ISSN: 1873295X. DOI: 10.1016/j.bej.2023.108813.

[23] Tom Rainforth et al. “Modern Bayesian Experimental Design”. In: (Nov. 2023). URL: http://arxiv.org/abs/2302.14545.

[24] Masahiro Kojima. “Application of multi-armed bandits to dose-finding clinical designs”. In: *Artificial Intelligence in Medicine* 146 (Dec. 2023). ISSN: 18732860. DOI: 10.1016/j.artmed.2023.102713.

[25] Juan Federico Herrera-Ruiz, Javier Fontalvo, and Oscar Andrés Prado-Rubio. *Hybrid Modeling for Bioprocesses: Architectures, Applications, and Perspectives*. Dec. 2025. DOI: 10.1002/eng2.70502.

[26] Derek DeBellis et al. *2024 Accelerate State of DevOps Report*. Tech. rep. DORA, Oct. 2024. URL: https://dora.dev/research/2024/dora-report/.

[27] Harini Narayanan et al. “Gaussian Processes for Predictive QSAR Modeling of Chromatographic Processes”. In: *Biotechnology and Bioengineering* (2026). ISSN: 10970290. DOI: 10.1002/bit.70168.

[28] Francesco Destro and Massimiliano Barolo. *A review on the modernization of pharmaceutical development and manufacturing – Trends, perspectives, and the role of mathematical modeling*. May 2022. DOI: 10.1016/j.ijpharm.2022.121715.

[29] U.S. Food and Drug Administration. *Guidance for Industry Process Validation: General Principles and Practices Guidance for Industry*. Tech. rep. 2011, pp. 800–835.

[30] U.S. Food and Drug Administration. *Guidance for Industry Q8(R2) Pharmaceutical Development*. Tech. rep. 2009. URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

[31] U.S. Food and Drug Administration. *Guidance for Industry: Q10 Pharmaceutical Quality System*. Tech. rep. 2009, pp. 20852–1448. URL: http://www.fda.gov/cder/guidance/index.htmhttp://www.fda.gov/cber/guidelines.htm.

[32] U.S. Food and Drug Administration. *Q12 Technical and Regulatory Considerations for Pharmaceutical Product Lifecycle Management Guidance for Industry*. Tech. rep. 2021. URL: https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances.

[33] U.S. Food and Drug Administration. *Q9(R1) Quality Risk Management Guidance for Industry ICH-Quality*. Tech. rep. 2023. URL: https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances.

[34] European Medicines Agency. *Committee for Medicinal Products for Human Use (CHMP) Committee for Medicinal Products for Veterinary Use (CVMP) Guideline on process validation for finished products-information and data to be provided in regulatory submissions*. Tech. rep. 2016. URL: www.ema.europa.eu/contact.

[35] European Medicines Agency. *Questions and Answers on Design Space Verification*. Tech. rep. 2013. URL: www.ema.europa.eu.

[36] Sebastian Porsdam Mann et al. *Guidelines for ethical use and acknowledgement of large language models in academic writing*. Nov. 2024. DOI: 10.1038/s42256-024-00922-7.